LISBON DATA SCIENCE ACADEMY

UK Stop and Search Policy

A Data-Driven Approach For Fairness

Prepared for:
The United Kingdom Department of Police

Prepared by:
Fabien Guegan
Data scientist
Awkward Problem Solutions™

30 of April 202

Table Of Contents

1. Client requirements

1.1 Summary

The aim of this report is to assist the United Kingdom Department of Police in enhancing and standardizing their stop and search policy across all UK police stations. Our data-driven analysis will help to identify and quantify potential instances of discrimination against minority groups, including discrimination in requesting individuals to remove clothing. Our analysis will evaluate the current claims made to the press and provide recommendations for improvement on the 42 police stations available.
Based on our analysis, we will also develop an API platform that officers can use to validate stop and search requests in real-time. This approach will enable the Department to monitor a greater number of occurrences and react quickly to new situations. The overall aim is to defend the Department's public image, improve their performance, and ensure fairness and justice for all.

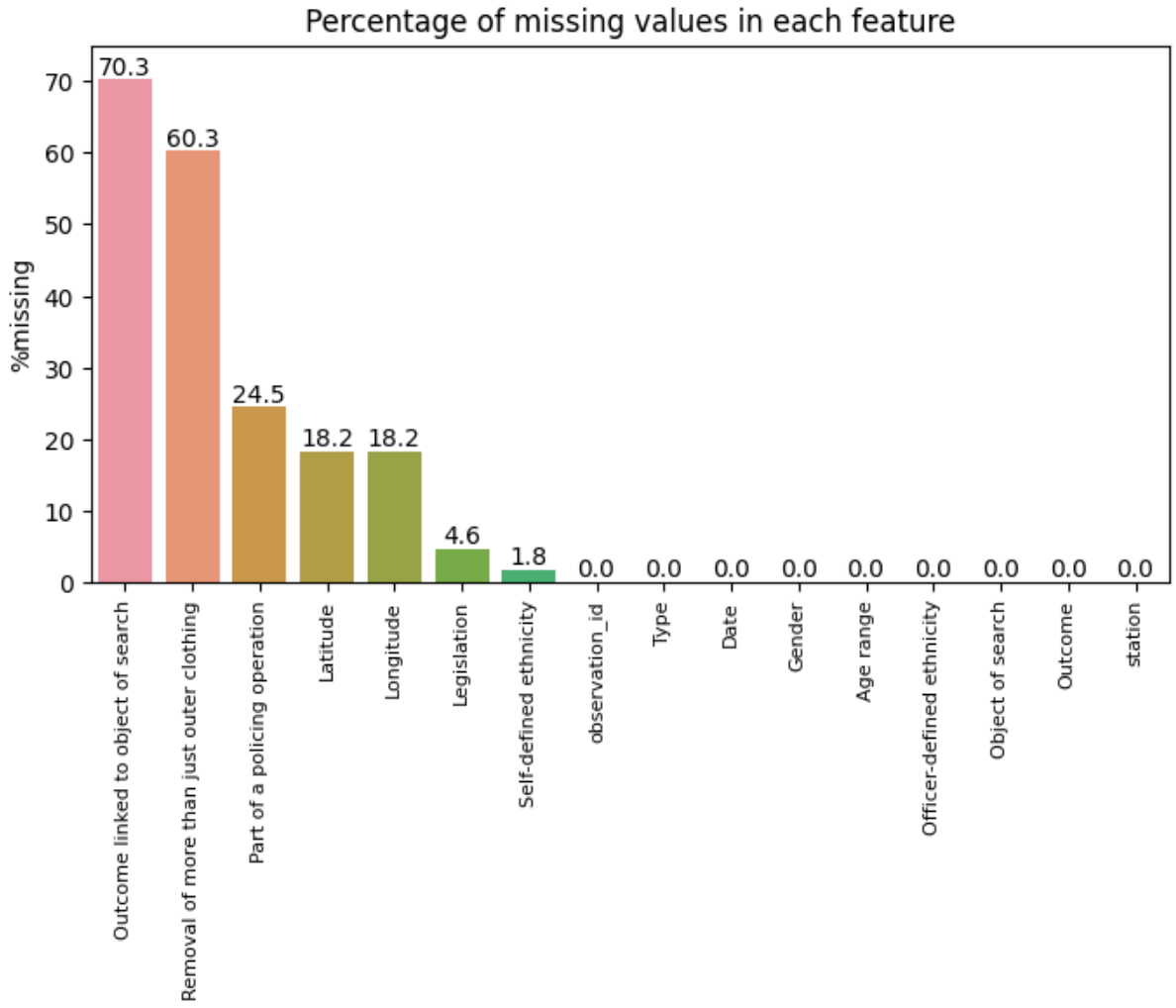
1.2 Requirements clarifications

Our API will only validate stop and search requests from police officers if the predicted success probability for a search is greater than 10%. Its primary objective will be to level the discovery rate or success rate (measured by the Precision metric) while maintaining a high level of overall offences detection (measured by the Recall metric). With the development of our API, we aim to ensure consistency in decision-making across stations, search objectives, and ethnicities to prevent discrimination.
To assess claims of discrimination, we will consider any sub-group of station/gender/ethnicity or station/gender/ethnicity/age with a minimum of 30 occurrences in the initial dataset. We will focus on three discrimination criteria: 1) gender and ethnicity; 2) gender/ethnicity/age when police officers ask individuals to remove clothing; and 3) mismatches between officer-defined ethnicity and self-defined ethnicity. We will use the search success rate metric to evaluate discrimination for criteria 1 and 2, and the percentage of mismatches in defined ethnicity for criterion 3. We will evaluate these three discrimination criteria both locally within each station (accepting a +/- 5% discrepancy) and globally between stations (with a +/- 10% discrepancy). We define a search as successful if its outcome is positive and related to the search object. Positive outcomes include community resolutions, khat or cannabis warnings, cautions (simple or conditional), penalty notices for disorder, arrests, summons/charges by post, suspect arrests, and suspect summonses to court. Stations with no data to establish their success rate will be excluded from the analysis.

2. Dataset analysis

2.1 General analysis

The current dataset comprises ~850,000 observations (rows) and 16 features (columns). Three of these features are not suitable for our API (Outcome, Outcome linked to the object of search and Self-defined ethnicity). To begin, we will analyze the content of each of these features and then determine our strategy for addressing any missing values. The "observation-id" feature shows 100% of unique values (annex.1), which indicates that our dataset has no duplicate entries. However, we observe an important number of missing values in at least 5 of the dataset features ("Outcome linked to object of search", "Removal of more than just outer clothing", "Part of a policing operation", "Latitude" and "Longitude" - annex.2, figure.1).
figure.1:



Our strategy for addressing missing values:

- ***Outcome linked to object of search****:* Metropolitan, Lancashire, Humberside stations have no recorded data for this feature (annex.3). As "outcome linked to object of search" is a mandatory feature to define the success rate of each station (target variable), these stations will not be included in most of our analysis. The rest of the missing values will be replaced by False, as we consider that officers tend to write when they find something and forget to report when the outcome is negative.

- ***Removal of more than just outer clothing***: missing values will be replaced by False, except when officers only search the vehicle (missing values will remain).
- ***Part of a policing operation***: since less than 3% of total observations are part of a policing operation, we will consider them to be rare occurrences. Therefore, we will replace missing values with False
- ***Latitude/Longitude***: missing values will be replaced with the mode (most frequent location) of each corresponding station, except for South Yorkshire and Nottinghamshire stations, which have no recorded data (annex 4). For these stations, the overall average location will be used instead.
- ***Legislation***: missing values will be replaced by the mode (most frequent legislation) of each corresponding object of search, as these two features are highly related.
- ***Self-defined ethnicity***: only the officer-defined ethnicity feature will be available for our API. Therefore, missing values will be addressed only later on, in our mismatches analysis between self and officer-defined ethnicity.

2.2 Business questions analysis

We will now investigate if discrimination claims made to the press are supported by the data. We will look at three discrimination criteria :

- ***Criteria_1***: gender and ethnicity discrimination in stop and search (using success rate metric)
- ***Criteria_2***: gender, ethnicity and age discrimination in removing clothing (using success rate metric)
- ***Criteria_3***: mismatches between Self- and Officer-defined ethnicity (using percentage of mismatch metric)

For our analysis, we consider only sub-groups with more than 30 occurrences (=260 subgroups in all dataset with the most represented subgroup: metropolitan/male/white). Station with null success rate (Metropolitan, Humberside, Lancashire) were removed from the analysis of criteria 1 and 2. We analyse the three discrimination criteria locally within each station (with +/- 5% discrepancy) and globally between all stations (with +/- 10% discrepancy).

Criteria_1:

- **locally** within each station (annex.5) :
  - 34 stations have at least one subgroup with more than 5% discrepancy
  - 24 stations have all their subgroups with more than 5% discrepancy
  - White Female is the most represented subgroup with more than 5% discrepancy (annex.6 ; see “dis” metric annex.5)
  - Black Female is most discriminated subgroup (annex.7 ; see “first” metric annex.5)
  - **Good station : Merseyside and Thames-valley**
    (no discrepancy with other subgroups and success rate station)
    (all station subgroups success rate > 10%)

- **globally** between all stations (annex.8):
  - **Bad stations : Leicestershire, west-midlands, Dyfed-powys, Lincolnshire**
    (less 10% success rate)
    (more than 10% discrepancy with mean success rate of all stations)

- **Good stations:** thames-valley, devon-and-cornwall, suffolk, btp, Wiltshire
  (> 10% success rate)
  (10% discrepancy with only the 4 bad stations)

Criteria_2:
- **locally** within each station (annex.9):
  - 35 stations have reliable data (no missing data in sub-groups)
  - Bad stations: Durham, Wiltshire, Thames-valley, Northamptonshire
    (more 50% of their subgroups have +5% discrepancy)

- **globally** in all stations (annex.10):
  Top 3 of most discriminated subgroups:
  - Female/white/18-24 is discriminated in 15 stations out of 33 (45,45%)
  - Female/white/10-17 is discriminated in 12 stations out of 19 (63,16%)
  - Male/Asian/over 34 is discriminated in 11 stations out of 23 (47,83%)

Criteria_3:
- **locally** within station (annex.11, annex.12):
  - **Bad stations : North-yorkshire, City-of-london, Dorset, Btp, Metropolitan**
    (+5% discrepancy with the average discrimination of all stations - 21,42%)

  - **Good stations** : 24 stations
    (-5% discrepancy with average discrimination of all stations)
    (**Gloucestershire, Northumbria, Durham, Cleveland, Humberside, Devon-and-cornwall, West-midlands** – have less than 5% discrimination rate)

- **globally** between all stations (annex.13):
  - Mixed and other are the most discriminated ethnicities (>40% of the cases)
  - White is the least discriminated ethnicity (~13% of cases)

2.3 Recommendations

Our data analysis suffered from a large amount of missing values present in the current dataset. Therefore, we highly recommend each police station to instruct officers to double-check and verify their request for missing values before sending to the API. This is especially important for the following fields:

- **"Outcome linked to object of search"**: as this field is mandatory to establish the success of the search, many cases from the current dataset had to be removed from our analysis. We urge the chief of UK Police department to correct this issue, particularly at three police stations where no data were recorded: **Metropolitan, Lancashire, and Humberside** stations. These three stations represent more than 53% of the entire dataset.
- **"Removal of more than just outer clothing"**: missing data in this field impaired our analysis of investigating discriminatory behavior by police officers who ask individuals to remove more clothing in certain population subgroups (station, gender, ethnicity, age). This issue is especially concerning for five police stations - **Metropolitan, Lancashire, Surrey, Cleveland, and North Yorkshire** - where no data was recorded for this field, representing more than 54% of the entire dataset.

- **“Latitude” and “Longitude”**: as we will show in the next section, latitude and longitude are important features for our model to accurately predict whether a search will be successful or not. Therefore, we recommend that police officers record these coordinates from the stop and search requests, particularly at stations where no coordinates were recorded, such as **South Yorkshire and Nottinghamshire** stations.

Other concerns:

- 439 individuals **under the age of 10** were stopped and searched by the UK police. We recommend that the Chief of the UK Police Department review this stop and search policy as it may not be well-received by the general public.
- In 45153 observations, although the “Outcome linked to object of search” were recorded as True (object of search found), none of this individuals were found to be involved in a criminal offense. This highlights **data entry issues** that should be corrected.
- There are 39747 missing values in the **legislation field**, which should never be missing in terms of legal jurisdiction. We would like to draw your attention to this issue.

Discrimination claims in the stop and search policy:

- **Criteria_1** : our analysis has identified 4 problematic stations when comparing sub-group search rates: **Leicestershire, west-midlands, Dyfed-powys, Lincolnshire**.

- **Criteria_2** : discrimination analysis for Criteria_2 was inconclusive due to a high number of missing values in the "Removal of more than just outer clothing" features.

- **Criteria_3** : our analysis has identified 5 problematic stations with regards for Self- and Officer-defined ethnicity mismatches: **North-yorkshire, City-of-london, Dorset, Btp, Metropolitan.**

Through our analysis of the three discrimination criteria, we have not only identified problematic police stations but also an exemplary one in **Merseyside**. This station sets a high standard for stop and search policies and can serve as a model for other UK police stations to enhance and standardize their practices.

Merseyside : criteria_1: best station ; crietria_3: only 11.6% mismatches

3. Modelling

3.1 Model specifications

Here comes the following steps to create our model:

- **First step: dataset pre-processing**
  - Remove all features that are not available in API requests
    (“Outcome linked to object of search”, “Outcome”, “Self-defined ethnicity” and “Removal of more than just outer clothing”)
  - Remove the stations with no data to build the target variable (successful search)

(Metropolitan, Humberside and Lancashire stations have no data in "Outcome linked to object of search")

- Remove 'Legislation' features as it is highly correlated to Object of search'

- **Second step: pipeline workflow (annex. 14)**
  - 3 groups of features with different transformers:
    Numerical : *{ Latitude , Longitude }*
    Categorical_1 *: { station }*
    Categorical_2 : *{ Part of a policing operation, Object of search, Officer-defined ethnicity, Age range, Gender, Type }*

  - 1 classifier: **GradientBoostingClassifier**

To select the best performing model, we use the F-score metric, which combines precision and recall using their harmonic mean. Our model prioritizes recall over precision by assigning a higher weight to recall using the F beta score with a beta value of 5, which maximizes the overall identification of offenders. By evaluating several classifiers, we found that GradientBoostingClassifier performed best (annex.15). We trained our model using 70% of the dataset as the training set and reserved 30% for evaluating and testing its performance. Additionally, to avoid missing any potential offenders, our model will only approve search requests with a probability of success above 10%.

3.2 Model performance and expected outcomes

In order to improve the efficiency of police officers in identifying offenders without conducting searches on every individual, our model implements a threshold of 10% probability of success for authorizing a search. Furthermore, the model prioritizes a high level of overall offense detection (recall) while simultaneously enhancing the discovery rate (precision). Understanding the importance of overall detection of offense in our model, we chose to use the F_beta score for evaluation and selection. By setting the beta value to 5, thereby placing a greater emphasis on recall compared to precision, our model achieved an impressive score of 84.2%.

In our current settings, our model has a recall of 95.7% and a precision of 21%, indicating that it accurately predicts only 21% of occurrences, but misses very few opportunities to search a suspect offender (only 4.3%). Our analysis has determined the optimal threshold for our model to be about 0.10000048627, and increasing it is expected to result in a reduction of the overall model performance, as shown by the accompanying graphs (annex.16). Moreover, we noticed that our success rate (precision) had increased from our previous analysis in the discrimination criteria_2 (annex.8) from 17.86% to 21%.

However, we also identified certain stations, which can be referred to as "bad" stations, that exhibit a significative lower overall ability to detect offenses (annex.17A) or/and stations that exhibit a significative lower success rate (annex.17B) when compared to the average of all UK stations. Upon examining the average recall and average precision between ethnicities, we observed a balanced distribution, indicating that our model generally does not exhibit discrimination based on ethnicity (annex.18). After comparing the overall ability of our model to detect offenses per station and object of search with the overall rates of the same object

across all stations, we have identified 38 stations where there is at least one object of search that exhibits a 10% discrepancy (example for the btp station in annex.19).

Finally, when examining the feature importance in our model (annex.20), we anticipate that our API will exhibit high sensitivity with respect to "object of search", "station", "age range", "latitude", and "longitude". Additionally, we found that ethnicity and gender have minimal impact on our model's performance.

3.3 Alternatives considered

During testing, we evaluated several classifiers (see annex.15), and found that both XGBClassifier and LGBMClassifier demonstrated similar performance to our chosen model. However, we decided to continue with GradientBoostingClassifier because it performs slightly better in a recall metric.

We attempted to improve the performance of our GradientBoostingClassifier through hyperparameter tuning. This involved adjusting 5 parameters, each with 2 potential values (annex.21).

- 'learning_rate': [0.01, 0.1],
- 'n_estimators': [50, 100],
- 'max_depth': [3, 5],
- 'min_samples_leaf': [0.1, 0.3],
- 'max_features': [9, None]

Despite tuning five hyperparameters with two values each, we were unable to improve both recall and precision scores simultaneously compared to the default parameters. As a result, we decided to stick with the current parameters.

4. Model Deployment

4.1 Deployment specifications

The objective here is to build and deploy an API that will validate stop and search police request only if its probability of success exceeds 10%. The API was developed using Flask application and deployed on the Railway hosting platform.

- The API uses our selected and optimized model (model) to predict the probability and outcome of each police stop and search request:
    - Load pickle files containing the characteristics of our trained model:
        - features name
        - feature type
        - pipeline steps
    - Use the characteristics of our trained model to compute the success probability of the requested search and predict the outcome of this search using the success probability and the optimal threshold (annex.16)

- The API have 2 endpoints to either predict the search outcome (endpoint: should_search/) or modify the true outcome of recorded result search (endpoint: search_result/):
    - should_search/
        - This endpoint handles information about the stop and search request
        - When submitting a request, information must be formatted according to specification or request is rejected, 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”}

        - This endpoint handles errors by rejecting the request and sending an error message to the police officer:
            - If observation_id field is not present in the request
            - If there are any missing or unrecognized field in the request
            - If latitude or longitude fields have invalid values. Latitude values must be between -90 and 90 (inclusive). Longitude values must be between -180 and 180 (inclusive)
            - If there are missing values in any of the request fields
        - If the observation ID already exists on the database, the API should return an error message. Otherwise, it should predict an outcome for that observation (whether the officer should search or not):

            {"outcome": Boolean}

    - 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 }

        - if the observation ID is not on the database, an error message should be displayed. If it is, the API should return an object with the observation ID, your predicted outcome and the true outcome given:

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

- The API records the result of each police stop and search request:
    - Creates a database table containing the following information:
        - observation_id of the police request (which should be a unique value)
        - search success probability calculated by our model
        - search predicted outcome delivered by our model
        - true outcome of the performed search
    - Connects to a PostGreSQL database integrated in our Railway hosting platform and stores the request search result table in it

4.2 Known issues and risks

Our model's 'GradientBoosting' method is sensitive to outliers, as each tree is built on the residuals/errors of previous trees. As a result, outliers and errors in the observed data labels may have a significant impact on our model's performance. One way to address these issues is to update the model regularly. However, our API currently rejects stop and search requests with missing values, which maximizes valid feature values for model prediction but also decreases the amount of new data available for model updates. Depending on the results of the first round of trials, we may need to change this setting to accept requests with missing values.

One another sensitive issue to be addressed with our API and PostGreSQL database is the security with the type of information stored in the database. With the exception of "Observation_id", we consider that all information provided is confidential data and will therefore not be stored in the PostGreSQL database. However, our API will store all the request content information locally in our computer for future use in our model updates.

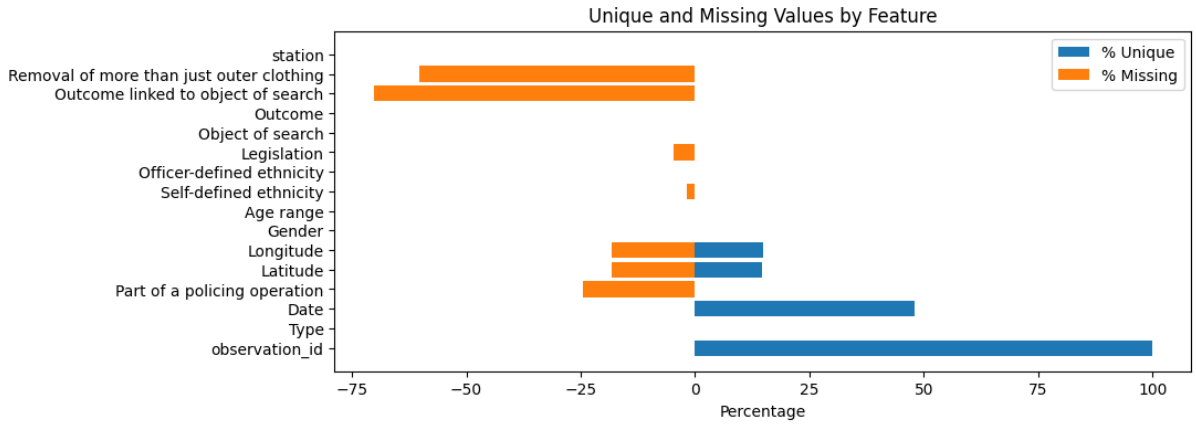
Finally, there are several limitations to be aware of when using free tier cloud server, such as railway:

- Limited storage (1 GB Disk)
- Limited computing power (512 MB RAM)
- Limited bandwidth (500 hours of usage per month)

These limitations can restrict the scalability and performance of our application.

5. Annexes

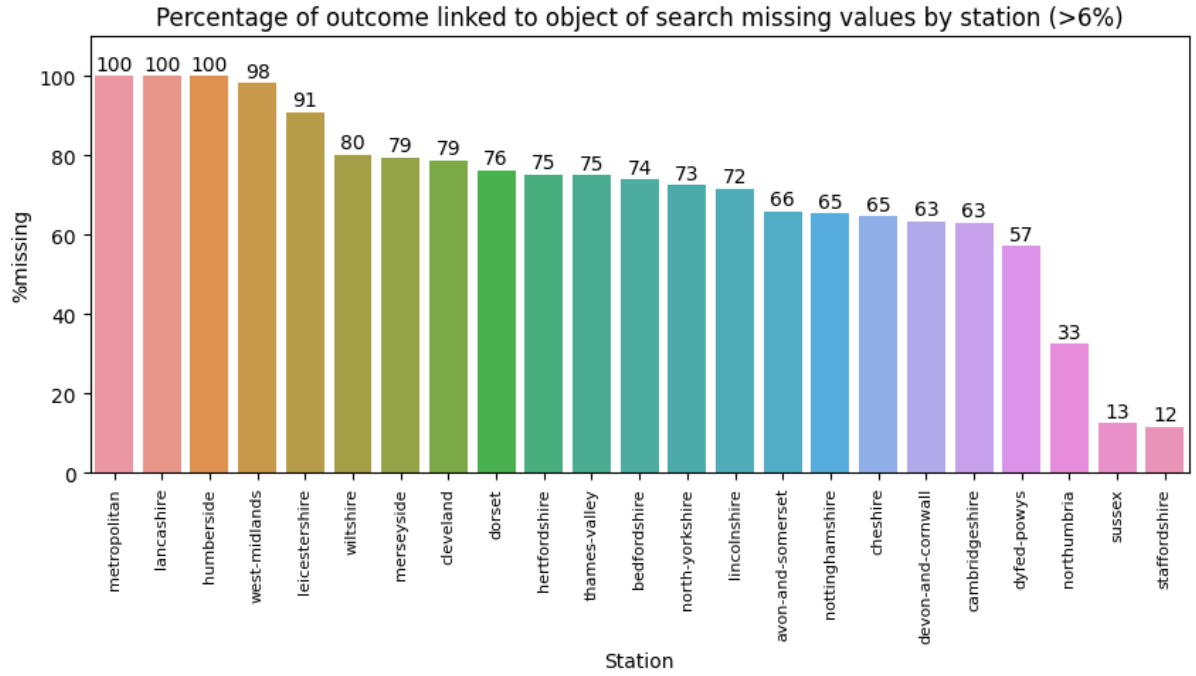
Annex.1: Percentage of unique and missing values in each feature



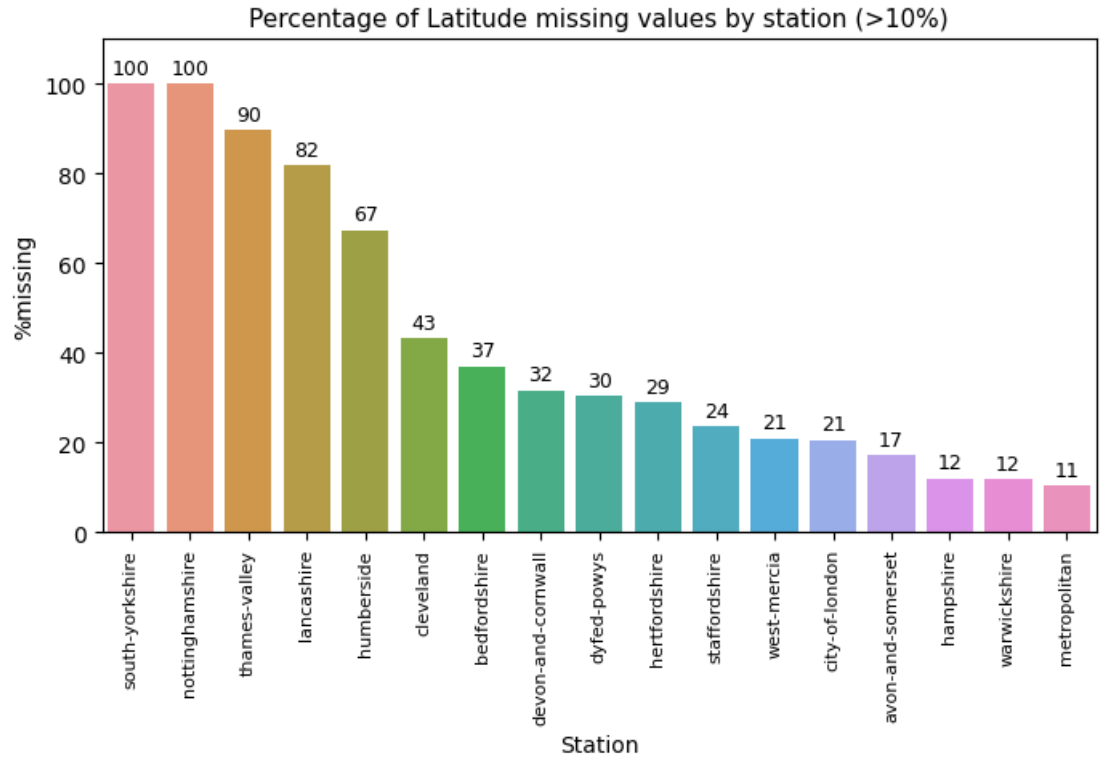
Annex.2: Table summarizing the number of unique and missing values in the dataset

	dtypes	missing	unique	%missing	%unique
Outcome linked to object of search	object	601927	2	70.27	0.00
Removal of more than just outer clothing	object	516513	2	60.30	0.00
Part of a policing operation	object	209990	2	24.51	0.00
Latitude	float64	156302	125553	18.25	14.66
Longitude	float64	156302	127197	18.25	14.85
Legislation	object	39747	19	4.64	0.00
Self-defined ethnicity	object	15183	19	1.77	0.00
observation_id	object	0	856610	0.00	100.00
Type	object	0	3	0.00	0.00
Date	object	0	410791	0.00	47.96
Gender	object	0	3	0.00	0.00
Age range	object	0	5	0.00	0.00
Officer-defined ethnicity	object	0	5	0.00	0.00
Object of search	object	0	17	0.00	0.00
Outcome	object	0	7	0.00	0.00
station	object	0	41	0.00	0.00

Annex.3: Percentage of missing values in “Outcome linked to the object of search” feature for station with more than 6% missing values



Annex.4: Percentage of latitude missing values for stations with more 10% missing values (percentage of longitude missing values are identical)



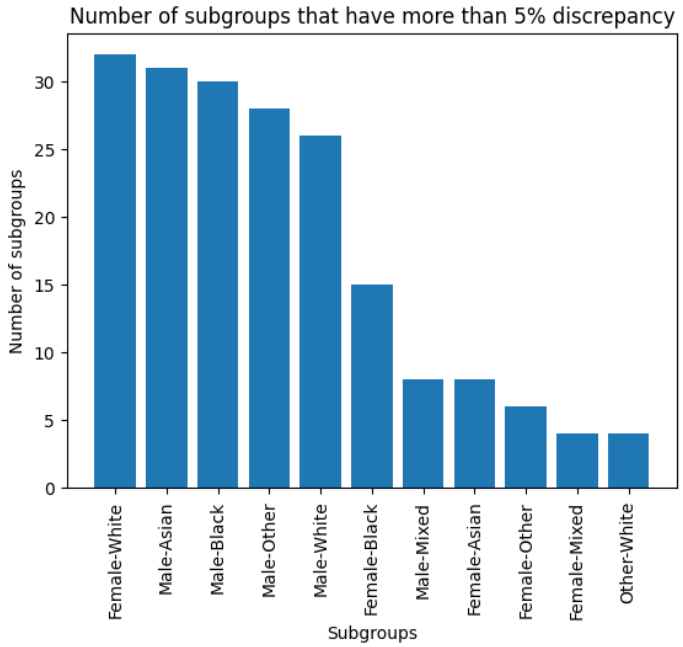
Annex.5: Table summarizing analysis discrimination criteria_1 (example of Cumbria station)

Gender	ODE	obs	num_S	SR	stat_SR	dis	n_dis	dis_stat	first	second
Female	**White**	430	61	0.14	0.21	True	3	True	True	False
Male	**Asian**	66	18	0.27	0.21	True	1	True	False	True
Male	**Black**	43	11	0.26	0.21	True	1	False	False	False
Male	**White**	2420	541	0.22	0.21	True	1	False	False	False

Table showing the results of criteria_1 analysis for Cumbria station:

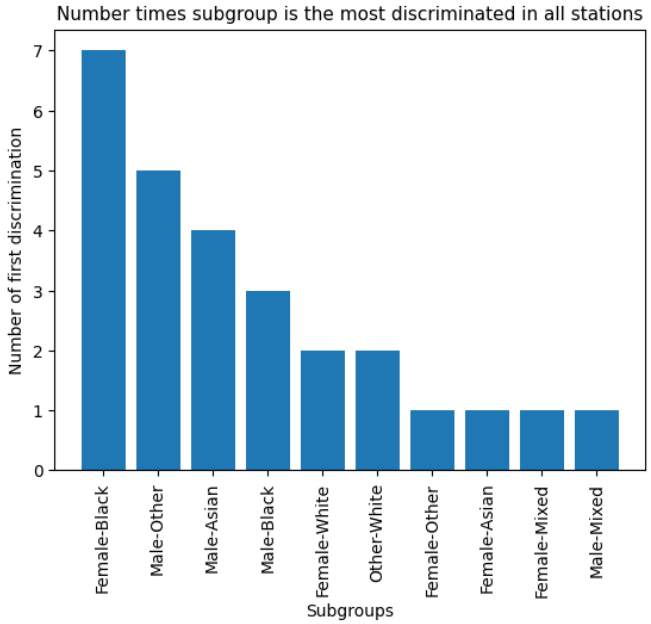
- Subgroup (>30 occurrences) = station/gender/Officer-defined ethnicity (ODE)
- obs = number of observations of the subgroup
- num_S = number of successful search of the subgroup
- SR = success rate of the subgroup (successful search rate)
- stat_SR = success_rate of the station
- dis = subgroup has 5% discrepancy in success rate with other subgroups
- n_dis = number of subgroups it has with 5% discrepancy in success rate
- dis_stat = 5% discrepancy with the success rate of the station
- first = first subgroup discriminated (dis == True & dis_stat == True & highest n_dis)
- second = second subgroup discriminated (dis == True & dis_stat == True)

Annex.6: Subgroups with 5% discrepancy



Total number of station/gender/ethnicity subgroups (>30 cases) in all dataset = 260
Subgroups most represented = metropolitan/Male/White

Annex.7: Number of times subgroup is the most discriminated



In total there are 27 most discriminated subgroups over the 41 station analysed

Annex.8: Global analysis of discrimination criteria 1 (only 15 stations shown)

station	obs	num_S	SR	all_stat_SR	n_dis	dis_global
leicestershire	8298	5	0.0006	0.1786	33	True
west-midlands	9271	123	0.0133	0.1786	33	True
dyfed-powys	4851	165	0.0340	0.1786	32	True
lincolnshire	5030	253	0.0503	0.1786	31	True
south-yorkshire	24300	2125	0.0874	0.1786	23	False
norfolk	8766	1142	0.1303	0.1786	11	False
cleveland	7907	1146	0.1449	0.1786	8	False
merseyside	38563	5922	0.1536	0.1786	7	False
west-yorkshire	25321	4195	0.1657	0.1786	6	False
north-wales	7044	1220	0.1732	0.1786	5	False
thames-valley	31437	5651	0.1798	0.1786	4	False
devon-and-cornwall	10906	1992	0.1827	0.1786	4	False
suffolk	6482	1195	0.1844	0.1786	4	False
btp	18557	3422	0.1844	0.1786	4	False
wiltshire	2867	530	0.1849	0.1786	4	False

Table showing the results of criteria_1 analysis:

- obs = number of observations per station
- num_S = number of successful search per station
- SR = success rate of the station
- all_stat_SR = average success rate of all stations
- n_dis = number of stations with 10% discrepancy the station has
- dis_global = 10% discrepancy with the average success rate of all station

stations with no data in Outcome linked to object of search or removal of more than just outer clothing were removed (Metropolitan, Humberside, Lancashire, Surrey, Cleveland, North-yorkshire)

(a) example for Durham station

(b) percentage of discriminated subgroups per station

a

Gender	ODE	Age	obs	SR_True	SR_False	SR	dis
Female	**White**	**over 34**	152	0.0	0.337748	0.335526	True
Male	**White**	**18-24**	845	1.0	0.303318	0.304142	False
		25-34	749	0.0	0.288591	0.287049	True
		over 34	863	0.0	0.265970	0.265353	True
Other	**White**	**18-24**	88	0.0	0.229885	0.227273	True
		25-34	36	0.0	0.257143	0.250000	True

b

station	%dis
durham	83.33
wiltshire	58.33
thames-valley	54.55
northamptonshire	53.33
northumbria	42.11
gloucestershire	41.67

ODE : Officer-defined ethnicity
Obs : number of cases
SR_True : success rate when Removal of more than just outer clothing is True
SR_False : success rate when Removal of more than just outer clothing is False
SR : success rate for the subgroup
dis : discriminated (SR_True has at least -5% discrepancy with SR)
%dis : percentage of discriminated subgroups per station (have dis == True)

Annex.10 : Global analysis of discrimination criteria 2

Gender	ODE	Age	obs_True	n_stat	%dis	dis_stat
Female	**White**	**18-24**	273.0	33	45.45	15
Female	**White**	**10-17**	42.0	19	63.16	12
Male	**Asian**	**over 34**	182.0	23	47.83	11
Male	**Black**	**over 34**	327.0	31	35.48	11
Female	**White**	**over 34**	603.0	35	31.43	11
Male	**Asian**	**18-24**	685.0	31	25.81	8
Male	**Asian**	**10-17**	78.0	21	33.33	7
Male	**Asian**	**25-34**	388.0	25	28.00	7
Male	**Other**	**18-24**	83.0	20	30.00	6
Male	**Black**	**10-17**	243.0	25	20.00	5
Male	**Other**	**25-34**	58.0	15	33.33	5
Male	**White**	**18-24**	2185.0	35	14.29	5
Male	**White**	**10-17**	536.0	33	15.15	5
Female	**White**	**25-34**	430.0	33	12.12	4
Male	**White**	**over 34**	2144.0	35	8.57	3
Male	**Black**	**25-34**	536.0	30	10.00	3
Male	**Black**	**18-24**	1162.0	31	6.45	2
Male	**White**	**25-34**	1940.0	35	5.71	2
Male	**Mixed**	**18-24**	67.0	7	28.57	2
Male	**Mixed**	**25-34**	37.0	6	33.33	2

ODE: Officer-defined ethnicity
obs_True: number of observations when Removal of more than just outer clothing is True
n_stat : number of stations in which the subgroup appears
%dis : percentage of stations in which the subgroup is discriminated
dis_stat : number of stations in which the subgroups is discriminated

Only subgroups with more than 30 occurrences in obs_True were analyzed

Annex.11: Table summarizing mismatches between Self- and Officer-defined ethnicity

	station	obs	ODE	SDE	dis
0	hampshire	1	White	Other ethnic group - Not stated	True
1	hampshire	1	Other	White - Any other White background	True
2	hampshire	1	White	White - English/Welsh/Scottish/Northern Irish/...	False
3	hampshire	1	White	Other ethnic group - Not stated	True
4	hampshire	1	Asian	Other ethnic group - Not stated	True

~85% of self-ethnicity missing values are imputed to one station (North-Yorkshire). For our analysis, we removed all Self-defined ethnicity missing values.

- obs = number of observations
- ODE = Officer-defined ethnicity
- SDE = Self-defined ethnicity
- dis = discrimination (mismatch between ODE and SDE)

Annex.12: Table summarizing mismatches between SDE and ODE by station

station	obs	%dis
north-yorkshire	537	51.2
city-of-london	4539	42.9
dorset	2807	30.9
btp	18498	29.5
metropolitan	436819	27.0
nottinghamshire	7523	24.4
thames-valley	28751	23.3
avon-and-somerset	12971	21.9
bedfordshire	5890	21.7

station	obs	%dis
gloucestershire	3542	4.8
northumbria	7876	4.2
durham	3585	3.8
cleveland	7541	3.2
humberside	9065	2.9
devon-and-cornwall	8089	2.3
west-midlands	8716	0.3

- obs = number of observations
- %dis = percentage of discrimination
- average discrimination in all stations = 21,42%

Annex.13: Global discrimination for criteria 3

ODE	obs	%dis
Mixed	2833	57.29
Other	28598	41.39
Black	195690	33.14
Asian	117572	27.96
White	496734	13.91

* obs = number of observations
* %dis = percentage of discrimination

Annex.14: Our pipeline workflow

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

1. Categorical_1

Feature :'station'

** SimpleImputer with 'strategy'/'fill_value' equals to 'constant'/'None'*

** OneHotEncoder with 'handle_unknown' equals to 'ignore'*

2. Categorical_2

Features: 'Part of a policing operation', 'Object of search', 'Officer-defined ethnicity', 'Age range', 'Gender' and 'Type'

** SimpleImputer with 'strategy' equals to 'most_frequent'*

** OneHotEncoder with 'handle_unknown' equals to 'ignore'*

3. Numerical

Features:'Latitude' and 'Longitude'

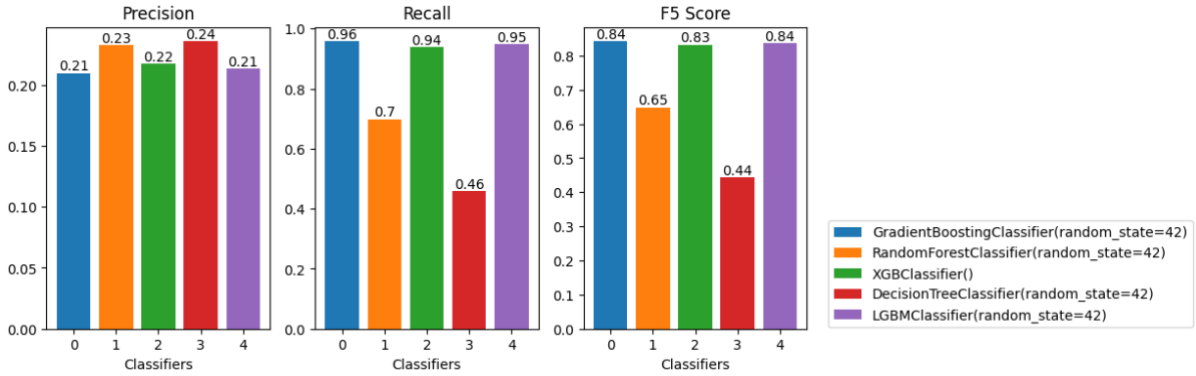
** SimpleImputer with 'strategy' equals to 'median*

** StandardScaler*

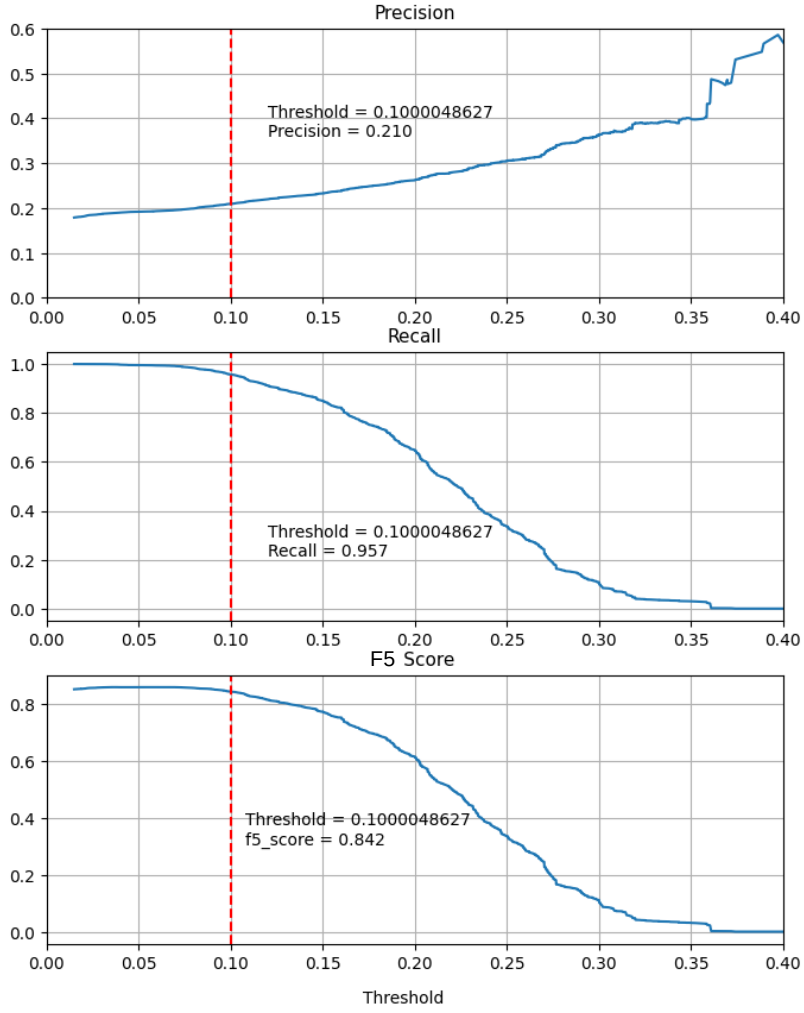
4. Classifier

a. GradientBoostingClassifier with 'random_state' equals to 42

Annex.15: Other classifiers test in our pipeline



Annex.16: Precision, recall and F5_score for our current model



Annex.17: Recall and precision per station

A. stations with more 10% discrepancy with mean recall's station

station	Offenders	%recall	%precision	%avg_recall	%avg_precision
leicestershire	2	0.00	0.00	95.71	20.96
west-midlands	43	0.00	0.00	95.71	20.96
dyfed-powys	41	17.07	17.50	95.71	20.96
lincolnshire	72	52.78	12.03	95.71	20.96

A. stations with more 5% discrepancy with mean recall's and precision's station

station	Offenders	%recall	%precision	%avg_recall	%avg_precision
leicestershire	2	0.00	0.00	95.71	20.96
west-midlands	43	0.00	0.00	95.71	20.96
lincolnshire	72	52.78	12.03	95.71	20.96
norfolk	337	90.50	15.32	95.71	20.96

Annex.18 Recall and precision between ethnicities

Officer-defined ethnicity	Offenders	%recall	%precision
Other	390	95.13	20.18
Asian	1880	95.53	20.04
White	16600	95.70	21.08
Black	2093	95.80	20.84
Mixed	189	97.88	23.15

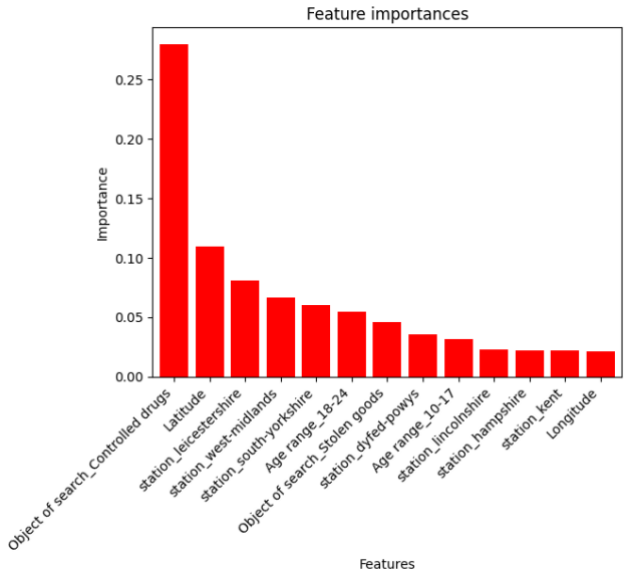
Annex.19: Recall and precision per station and object of search

example btp's station (object with >10% discrepancy with object mean recall for all station)

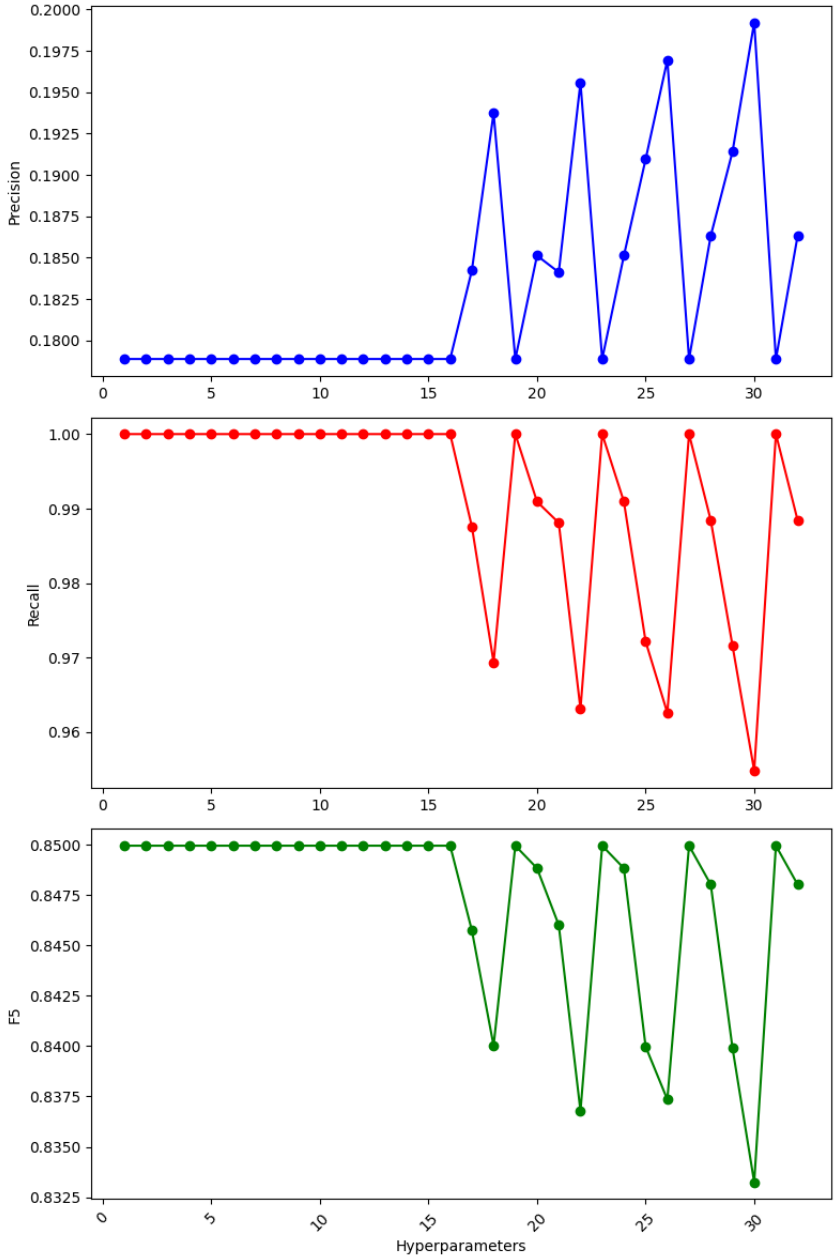
Object of search	obs	Offs	%R	%P	%D_tot	%P_tot
Anything to threaten or harm anyone	423	41	9.76	4.44	51.74	10.87
Articles for use in criminal damage	40	2	50.00	5.88	75.56	13.73
Offensive weapons	259	25	48.00	7.64	75.28	12.30

- obs = number of cases
- Offs = number of offenders
- %R = recall of the station/object subgroup
- %P = precision of the station/object subgroup
- %D_tot = recall of the object across all station
- %P_tot = precision of the object across all station

Annex.20: Features importance



Annex.21: Hyperparameters tuning



Hyperparameters combinations:
(learning_rate, n_estimators, max_depth, min_samples_leaf, max_features)
1: (0.01, 50, 3, 0.1, 9) ; 2: (0.01, 50, 3, 0.1, None), 3: (0.01, 50, 3, 0.3, 9), 4: (0.01, 50, 3, 0.3, None), 5: (0.01, 50, 5, 0.1, 9), 6: (0.01, 50, 5, 0.1, None), 7: (0.01, 50, 5, 0.3, 9), 8: (0.01, 50, 5, 0.3, None), 9: (0.01, 100, 3, 0.1, 9), 10: (0.01, 100, 3, 0.1, None), 11: (0.01, 100, 3, 0.3, 9), 12: (0.01, 100, 3, 0.3, None), 13: (0.01, 100, 5, 0.1, 9), 14: (0.01, 100, 5, 0.1, None), 15: (0.01, 100, 5, 0.3, 9), 16: (0.01, 100, 5, 0.3, None), 17: (0.1, 50, 3, 0.1, 9), 18: (0.1, 50, 3, 0.1, None), 19: (0.1, 50, 3, 0.3, 9), 20: (0.1, 50, 3, 0.3, None), 21: (0.1, 50, 5, 0.1, 9), 22: (0.1, 50, 5, 0.1, None), 23: (0.1, 50, 5, 0.3, 9), 24: (0.1, 50, 5, 0.3, None), 25: (0.1, 100, 3, 0.1, 9), 26: (0.1, 100, 3, 0.1, None), 27: (0.1, 100, 3, 0.3, 9), 28: (0.1, 100, 3, 0.3, None), 29: (0.1, 100, 5, 0.1, 9), 30: (0.1, 100, 5, 0.1, None), 31: (0.1, 100, 5, 0.3, 9), 32: (0.1, 100, 5, 0.3, None)