Client Requirements

Summary

Awkward Problem Solutions™ has been hired by the United Kingdom Department of Police to help evaluate and optimize their stop and search process, due to press accusations of discrimination towards certain minorities, not only on the stop and search frequency but also on the further need of undressing the subject to conclude about the result of the intervention.
This project aims to:

1. Attest if the claims are true
2. Create a service for authorizing the search

This service is then to be integrated in their own internal approval system and used by police officers as an evaluation system for probable cause. Based on facts and the analysis made it should improve the results of the stop and search process and eliminate the discrimination.

Requirements Clarifications

The main objective with this API is that searches are performed only when there is more than 10% likelihood that the search will be successful, reducing with this value the oversearch. This establishes our minimum threshold for the predicted probability of our model to predict a successful operation and by being so low implies that the precision does not need to be very high and the maximization of recall is possible.

Maximization of recall is important because we do not want to allow criminality so we want to maximize the number of prevaricators being caught. Therefore we need to minimize the False Negatives which are subjects who are committing an illegal activity and are not searched/stopped. This is technically achieved by maximizing Recall which is calculated by dividing the same True Positives stated above by the sum of True Positives with False Negatives.

To tackle the discrimination issue, we need to make sure that precision is similar across minorities. This ensures that the criteria for probable cause is always the same and the ratio prevaricators/suspects is not dependent on ethnicity or sex.

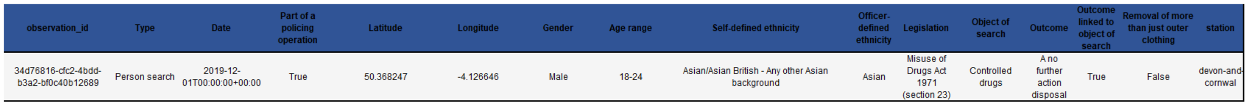
The same is applied about the suspect undressing. The precision for stops/searches where suspects are asked to undress has to be similar across minorities and even age ranges.

Important to state that for an operation to be considered successful, its outcome, in addition to discovering prevaricators, has to be related to the search motives which originated the suspicion in the first place.

Dataset Analysis

General Analysis

The dataset consists of a list of search/stop occurrences identified by unique ids on a determined , available time and place, either done as part of a police operation or not and with the goals of searching persons, vehicles or both. The police station responsible for the operation is also available and the reason why the person was stopped is stated both as an object of search and as the law which is suspectedly being broken. Each occurrence has the suspect's characteristics such as age range, gender, self and officer defined ethnicity and the outcome of search. Here is an occurrence sample example:



observation_id	Type	Date	Part of a policing operation	Latitude	Longitude	Gender	Age range	Self-defined ethnicity	Officer-defined ethnicity	Legislation	Object of search	Outcome	Outcome linked to object of search	Removal of more than just outer clothing	station
34d76816-cfc2-4bdd-b3a2-bf0c40b12689	Person search	2019-12-01T00:00:00+00:00	True	50.368247	-4.126646	Male	18-24	Asian/Asian British - Any other Asian background	Asian	Misuse of Drugs Act 1971 (section 23)	Controlled drugs	A no further action disposal	True	False	devon-and-cornwal

Fig 1. Occurrence Sample

Regarding the quality of the supplied data, there are no duplicated occurrences and a little part of the features available have missing values. (More details on the annexes, Fig 7. Data analysis)

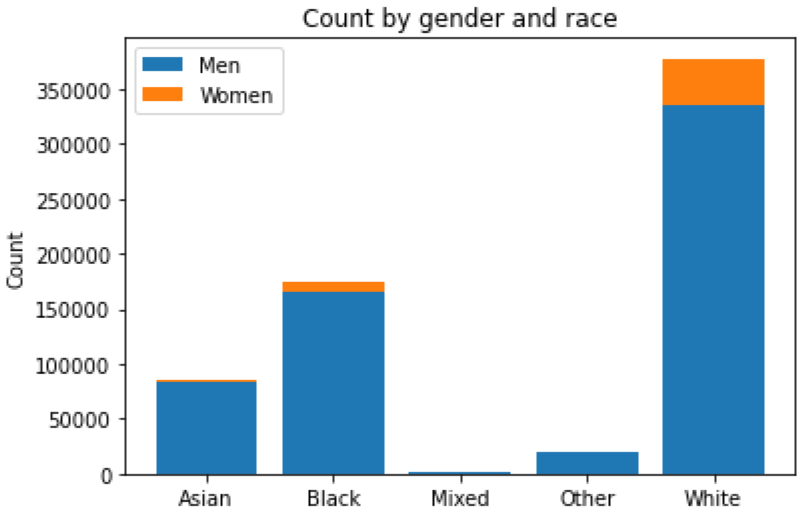


Fig 2. Officer-defined ethnicity distributions

By gender, 91.6% of suspects are male, 8.3% are female and 0.1% other. By ethnicity we have 57.3% White subjects, 26.4% are Black, 13% are Asian, 3.3% have other etnies. These are the distributions by age range:

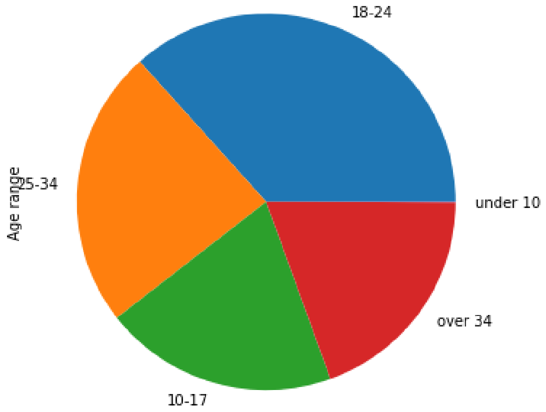


Fig 3. Age range distributions

Regarding Police stations present on the data we have 42 departments and regarding the Outcome of the occurrences, these are the counts of each one:

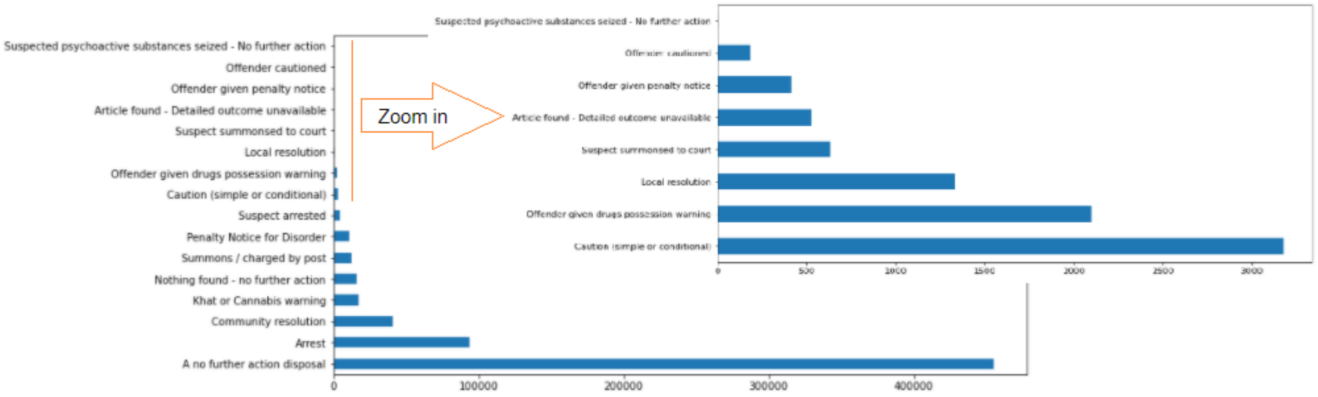


Fig 4. Outcome distributions

The data quantity seems to be increasing over time which suggests the concern of keeping data for analysis and quality control is increasing.

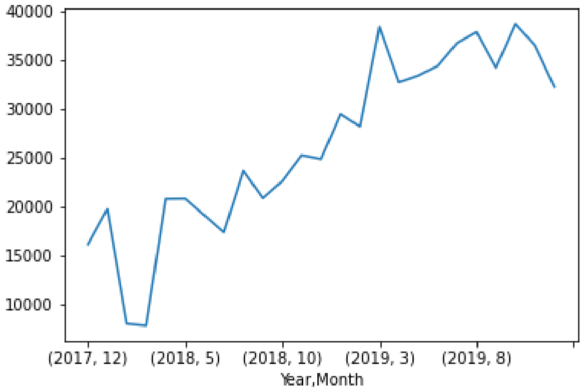


Fig 5. Evolution of number of occurrences over time

The majority of data is from the Metropolitan Police department, with more than 50% of occurrences. Here we have a count of occurrences per station without it so we can have an idea of the unbalancement of data regarding this feature.

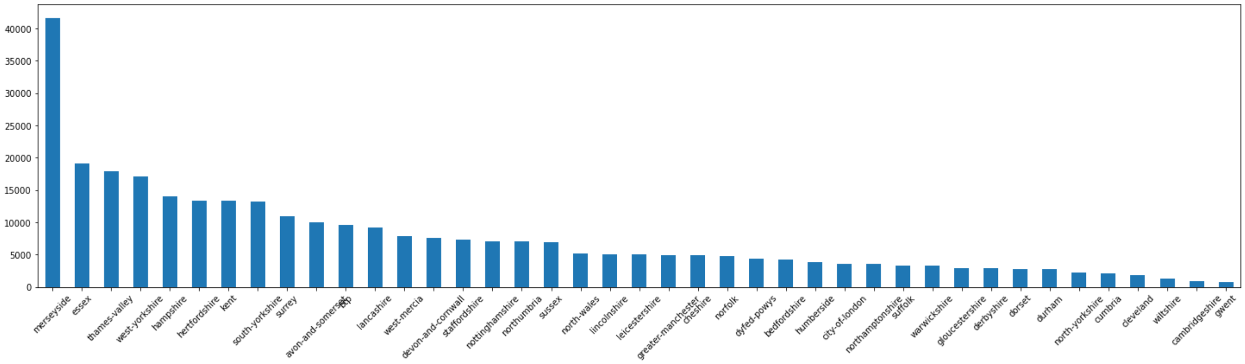


Fig 6. Stations distribution

Business Question Analysis

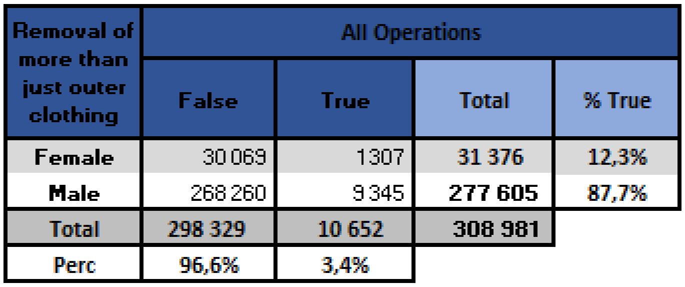
Considering the business definition of successful operation implying that the operation outcome has to be related with the search motive and the Metropolitan Police department does not have data on this subject, this department was not considered on the analysis. For the rest of the missing values on this column it was considered that the data missing implied an unsuccessful operation (see annexes, Fig. 11).

Regarding ethnicity, the differences between the officer defined and the self defined are very big. In reality each group of officer-defined ethnicity has every kind of self-defined ethnicity in it which can bias the study results. It is the object of the study, the officers perception of the situation and way of acting but if the data is analysed with real ethnicities, results may be different and this bias difficult to explain to the press.

This being said, as we initially analyse the remaining, treated data, we verify that Btp, Northamptonshire, and the City-of-london police departments have an average of Black people searches 10 percentual points above the national average, and Bedfordshire, Thames-valley, City-of-london and West-yorkshire departments for Asian people.

When breaking this data analysis in age ranges, the incidence of this behavior seems to happen to subjects between 10 and 24 years old for the Btp department and 18 to 24 in Bedfordshire and West-yorkshire in this high tendency to suspect etnic young people.

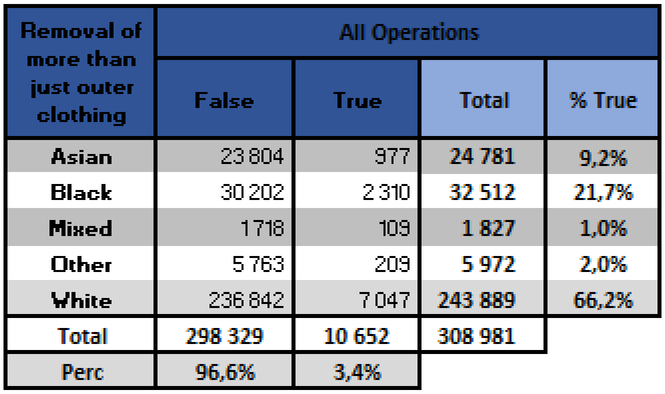
Regarding the unnecessary undressing of suspects, we see that in 12.3% of the cases where the subject is asked to remove more than just outer clothing, this subject is a woman, which is higher than the percentage of female suspects being stopped/searched.



Removal of more than just outer clothing	All Operations			
	False	True	Total	% True
Female	30 069	1307	31 376	12,3%
Male	268 260	9 345	277 605	87,7%
Total	298 329	10 652	308 981	
Perc	96,6%	3,4%		

Fig7. Numbers of subjects asked to undress more than outer clothes by gender

When breaking this data by ethnicity the percentage of black people asked to undress is 21.7% of all undressed subjects which more than doubles the average of black people of the population.

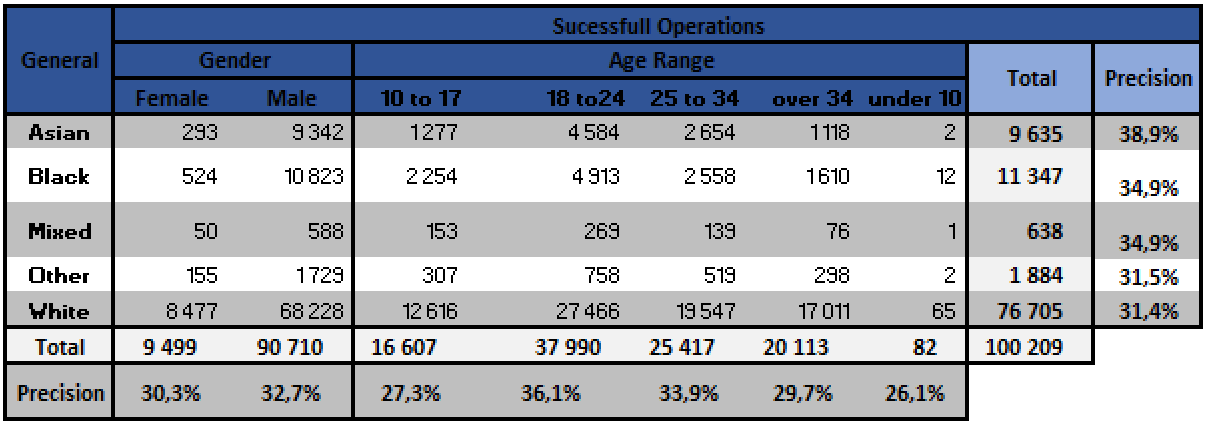


Removal of more than just outer clothing	All Operations			
	False	True	Total	% True
Asian	23 804	977	24 781	9,2%
Black	30 202	2 310	32 512	21,7%
Mixed	1 718	109	1 827	1,0%
Other	5 763	209	5 972	2,0%
White	236 842	7 047	243 889	66,2%
Total	298 329	10 652	308 981	
Perc	96,6%	3,4%		

Fig 8. Numbers of suspects asked to remove more than outer clothes by ethnicity

These types of analysis may be similar to the ones used by the press as a source of truth for their accusations but they are misleading. As stated before, these analysis represent likelihood but cannot be immediately interpreted as discrimination. It can happen that these groups represent indeed a higher rate of criminality, presenting incriminating proof which justifies the tendency to search them more frequently. To check for discrimination proof we must compare the success rate (aka Precision) of each case, already described above as the True Positives within each group divided by the total population searched in each group (True Positives + False Positives), and guaranteeing this way that the criteria for searches is the same across all suspects.

When analysing precisions across ethnic groups we realize that the bigger incidence on asian people is really explainable by results, as the success rate on this group is significantly higher. Black and Mixed people do not present problematic results and white people should actually be less stopped.

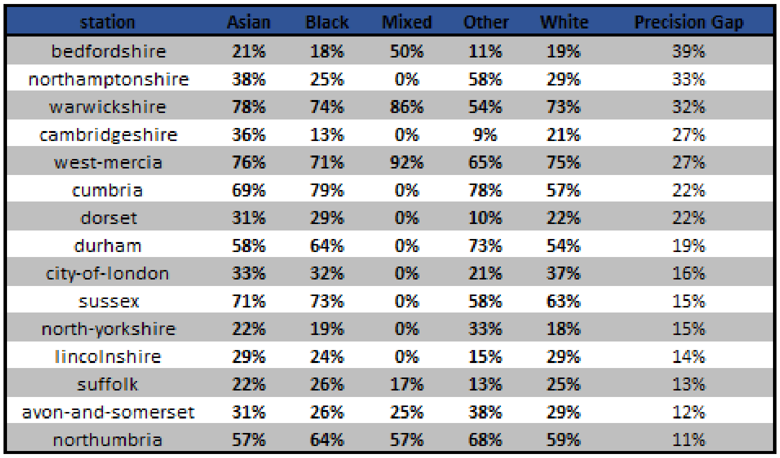


General	Sucessfull Operations								
	Gender		Age Range					Total	Precision
	Female	Male	10 to 17	18 to24	25 to 34	over 34	under 10		
Asian	293	9 342	1 277	4 584	2 654	1 118	2	9 635	38,9%
Black	524	10 823	2 254	4 913	2 558	1 610	12	11 347	34,9%
Mixed	50	588	153	269	139	76	1	638	34,9%
Other	155	1 729	307	758	519	298	2	1 884	31,5%
White	8 477	68 228	12 616	27 466	19 547	17 011	65	76 705	31,4%
Total	9 499	90 710	16 607	37 990	25 417	20 113	82	100 209	
Precision	30,3%	32,7%	27,3%	36,1%	33,9%	29,7%	26,1%		

Fig 9. Precisions analysis across sensible groups

Regarding gender, the success rate is pretty close between men and women and concerning age ranges, people between 0 and 17 years old have the lowest success rate and the gap to the max precision gets up to 10 percentage points.

When going over results by station, this bias is also observable and in some cases the differences between precisions for two different ethnicities can go up to 39%, as can be observed in the following table, when the objective was initially 5% and established at 10% after data analysis, as explained in Conclusions and Recommendations chapter.



station	Asian	Black	Mixed	Other	White	Precision Gap
bedfordshire	21%	18%	50%	11%	19%	39%
northamptonshire	38%	25%	0%	58%	29%	33%
warwickshire	78%	74%	86%	54%	73%	32%
cambridgeshire	36%	13%	0%	9%	21%	27%
west-mercia	76%	71%	92%	65%	75%	27%
cumbria	69%	79%	0%	78%	57%	22%
dorset	31%	29%	0%	10%	22%	22%
durham	58%	64%	0%	73%	54%	19%
city-of-london	33%	32%	0%	21%	37%	16%
sussex	71%	73%	0%	58%	63%	15%
north-yorkshire	22%	19%	0%	33%	18%	15%
lincolnshire	29%	24%	0%	15%	29%	14%
suffolk	22%	26%	17%	13%	25%	13%
avon-and-somerset	31%	26%	25%	38%	29%	12%
northumbria	57%	64%	57%	68%	59%	11%

Fig 10. Top 15 precision gaps across ethnicities per station

Conclusions and Recommendations

The main recommendation would be for the Metropolitan Police department to improve their data collection. This department is responsible for more than half of all stop and search operations records, as seen on the initial analysis, being therefore very relevant for every analysis made on the subject and very influential on the actual results and the image of the police force as a whole.

As far as conclusions are concerned, the first one we can take both from the data analysis and the modeling process is that a 5 percentage point difference between groups to define bias is very low and with such an ambitious objective, there is indeed bias in the police force stop/search policy. Regarding Ethnicity the success rates are between 31.4 and 38.9% and regarding age ranges between 26.1 and 36.1%. (See annexes, Fig. 20

As far as undressing more than outer clothes is concerned success rates in these operations regarding ethnicity vary from 35.9 to 50.5%, regarding age, between 35 and 43.9% and gender 34 and 41.8%. (See annexes, Fig. 21)

When analysing results per station, bias can go up to 39% gaps between groups. The results can be found in the annexes (See annexes, Fig.19)

The overall success rate for analysed data is 32%.

Modeling

Model Expected Outcomes Overview

The tradeoff between precision, recall and the elimination of discrimination was a difficult one to manage. An initial model, which could indeed eliminate discrimination, had precision of 84% but a recall of 8%, which means very little crime would actualy be discovered due to a very high number os false negatives or as explained previously, the people not stopped nor searched. Considering the low probability of success rate border given, it allowed us a big range of values to analyse and try to gain recall still eliminating discrimination.
Regarding bias across sensible groups, the less precision we seem to aim at, the bigger the discrepancy between precision cross sensible groups. Our results regarding this objective vary inversely with recall. It was chosen a compromise in the final model presented, which manages to eliminate discrimination, considering a 10% difference in opposition to the 5% asked, keep the recall at an acceptable level and maintain a high precision which is what allows us to control discrimination. It is expected to have a success rate for operations of around 82%, a slightly higher recall of 12% and the problems with bias across sensible groups are expected to get solved. This model was achieved by manipulating its results depending on its success probability.

Model Specifications

The final chosen model is a Scikitlearn Random Forest Classifier, trained with Type, Gender, Age range, Officer defined ethnicity, Object of search and Station. Data features were added on the exploring phase but did not have a positive impact on our predictive power and in the end were not used in the final model. All features are categorical and therefore encoded with OneHotEncoder.
The hyperparameters used can be found in the (annexes Fig.24) and were tuned manually in order to allow the model to satisfy all requisites. The objective of avoiding bias across sex and gender in all stations was, as stated before, changed to 10% and all stations are expected to have discrepancies in precisions of sensible groups lower than this new objective.
As far as data is concerned, values from the Metropolitan police department were not used as the values which related the object of search with the outcome were not available and this turned out to be the target of our model. Information about possible undressing of more than outer clothes were also missing for this station and the data was considered insufficient and unreliable. The ‘Other’ gender was also dropped as there was very little data available.

By analysing the precision and recall curve, and our imposed threshold of 0.1 (green line) we can see that precision would be expected to be around 0.36. Considering that precision is what allows us to maintain discrimination at acceptable levels, we aimed at 0.82 (red line), which is the minimum value which guarantees bias across sensible groups is eliminated. This 0.82 corresponds to a threshold of 0.4490314298609562, above the minimum 0.1 asked.

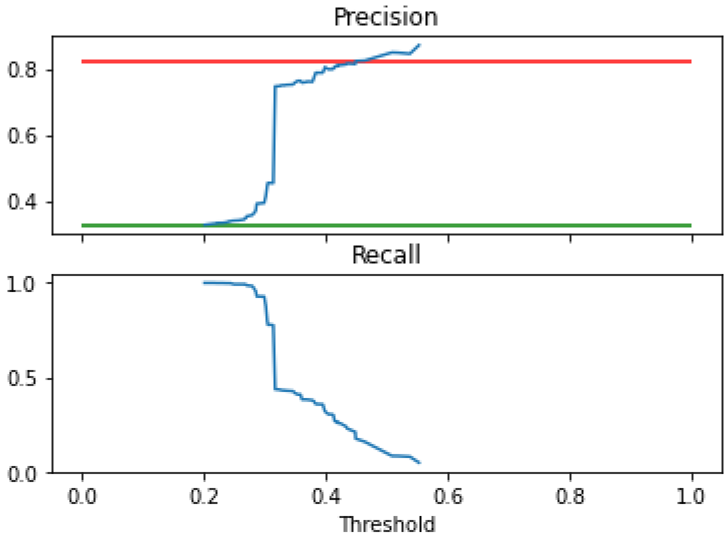


Fig 11. Precision and Recall curve, redline represents 0.82 precision and green line 0.36

The expected recall value after this adjustment is 17%. The chosen precision is much higher than the one being practiced at the datas studied and the recall cannot be compared as there is no data from non stopped/searched people.

Analysis of expected outcomes based on training set

Based on this training set and considering the results on the test set, bias across sensible groups such as ethnicity and gender are expected to decrease to a 10% difference gap. To achieve this, a very high precision was necessary in order to give us some control over the expected result. This precision is expected to be 82%. This was considered a priority over the maximization of recall due to the importance of this subject and its influence on the fairness and image of the police force. Recall is expected to be 12%.
The high success rate is a good thing per se but it implies a low recall. Therefore, the number of people which would not be considered a suspect and are indeed prevaricating is expected to have a relatively high value. More information about the expected performance of the model can be found in the annexes.

Alternatives considered

Several other alternatives were considered, such as the same model with very different hyperparameters, the disconsideration of non statistically relevant values, these being a combinations of department, ethnicity and gender values with less than 30 occurrences, and the use of a model simpler than the Random Forest Regressor, the Logistics Regressor, also manually tuned with several parameters. The expectation was that the model would follow less the individuality of each occurrence and would therefore present less bias regarding sensible groups. The current model was the one which presented better results, regarding all three objectives.
The oversampling of underrepresented ethnicities was also tried but it did not add value to our objectives neither recall nor precision across sensible groups.
Stratification on the train test split considering ethnicity was also a strategy which presented no results.

Known Issues and Risks

The production data may have classes of observations which are not known to the model. In this case the model will not be able to predict it. The encoder was set to ignore the unknown classes so we do not have an error in this situation.
Extra, unexpected features or features missing will also not produce a result and the presence of an unique id is mandatory for data storage, this can harm the data colectage process.
Also, if the production data differs a lot from the training data regarding sample distribution of each class the model can perform below its expected performance.

Model Deployment

Deployment Specifications

The API was deployed on Heroku http server and built with Flask. It has two endpoints, the '/should_search/', the '/search_result/'. The /should_search/ receives an observation, checks if it was properly introduced by checking if all columns are present and if all classes are known, meaning, there are no extra, unexpected columns. It should have the following content:

```
{
        "observation_id": <string>,
        "Type": <string>,
        "Date": <string>,
        "Part of a policing operation": <boolean>,
        "Latitude": <float>,
        "Longitude": <float>,
        "Gender": <string>,
        "Age range": <string>,
            "Officer-defined ethnicity": <string>,
                    "Legislation": <string>,
                "Object of search": <string>,
                    "station": <string>
}
```

If it passes the validation, a True or False response is sent to the user telling him if the stop/search should be performed if not, an error specifying what is wrong with it is sent:

```
{"outcome": <boolean>}
```

If the observation id was already used, an error will be generated, as observations are stored and the observation id should be unique.
The /search_result/ allows the user to compare the predicted result with the actual outcome of the operation and update that result. It expects to get an observation id and an outcome:

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

The response to this kind of request is the observation id, the real outcome, passed on the request and the previously predicted result:

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

The API receives the POST as a json and parses it. Using the columns and the dtypes, it builds a pandas dataframe and using the model, predicts the outcome by comparing the probability calculated with the threshold .

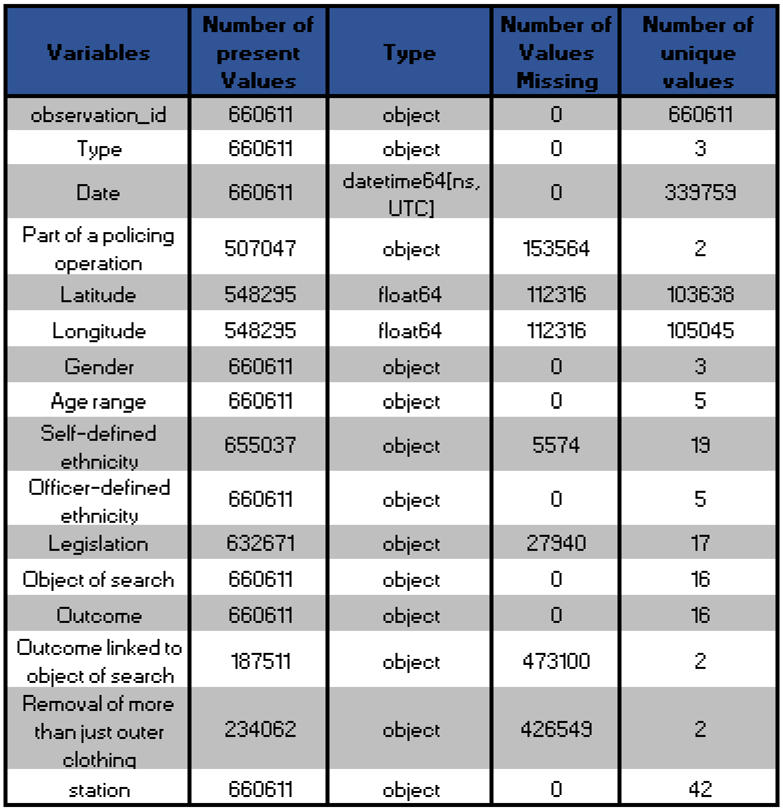
Known Issues and Risks

The Postgres database used as a resource on our Heroku http server only supports ten thousand observations. The risk of exhaustive use crashing the current API is real and if needed, a better alternative would have to be thought to better serve the users needs, which would certainly incur in financial costs.
The first requests made to the server after a period of inactivity may take more time than expected as the connection has to be reestablished. This can turn out to be a problem for the user.

As referred before, the observations go through a test before giving a response and if data does not satisfy the quality criteria expected an error will be generated. This can be frustrating for the user, but it helps ensure the quality of stored data and will be very beneficial in future analysis.

Annexes

Dataset Technical Analysis



Variables	Number of present Values	Type	Number of Values Missing	Number of unique values
observation_id	660611	object	0	660611
Type	660611	object	0	3
Date	660611	datetime64[ns, UTC]	0	339759
Part of a policing operation	507047	object	153564	2
Latitude	548295	float64	112316	103638
Longitude	548295	float64	112316	105045
Gender	660611	object	0	3
Age range	660611	object	0	5
Self-defined ethnicity	655037	object	5574	19
Officer-defined ethnicity	660611	object	0	5
Legislation	632671	object	27940	17
Object of search	660611	object	0	16
Outcome	660611	object	0	16
Outcome linked to object of search	187511	object	473100	2
Removal of more than just outer clothing	234062	object	426549	2
station	660611	object	0	42

Fig 12. Initial Data Analysis [Back]

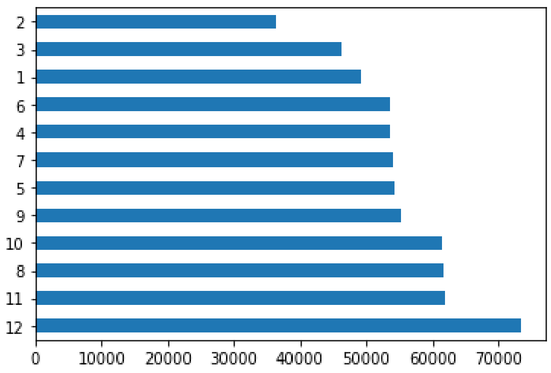


Fig 13. Distribution of results per month

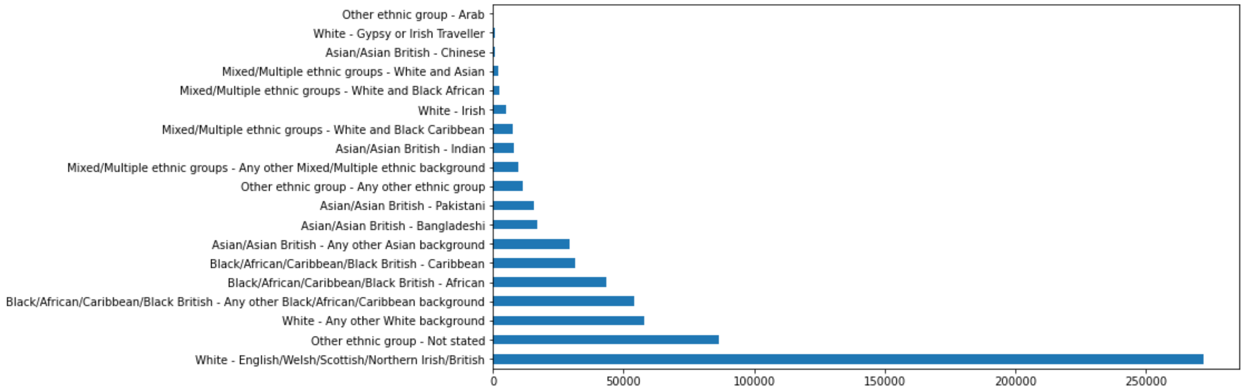
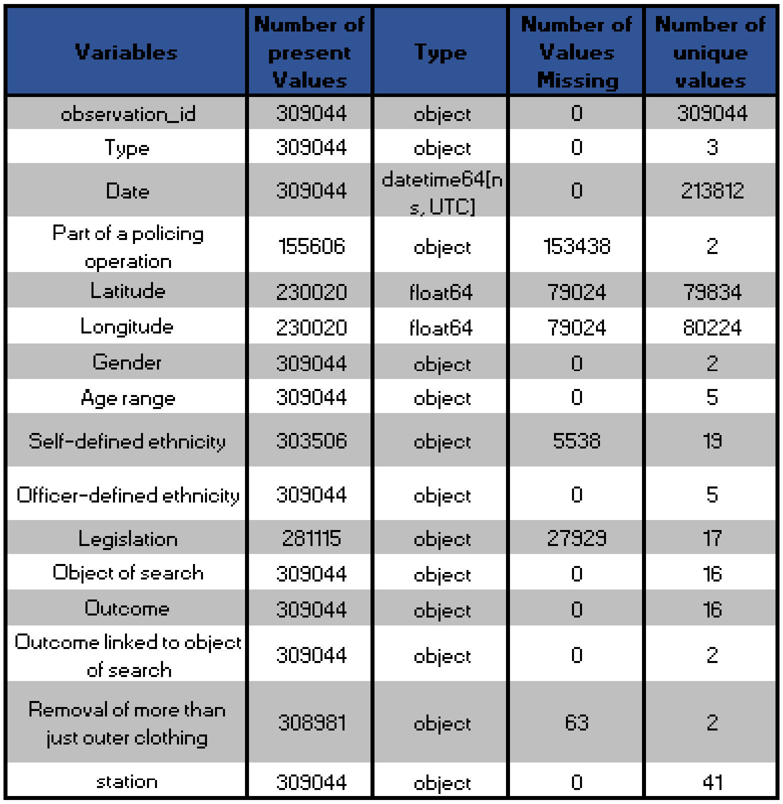
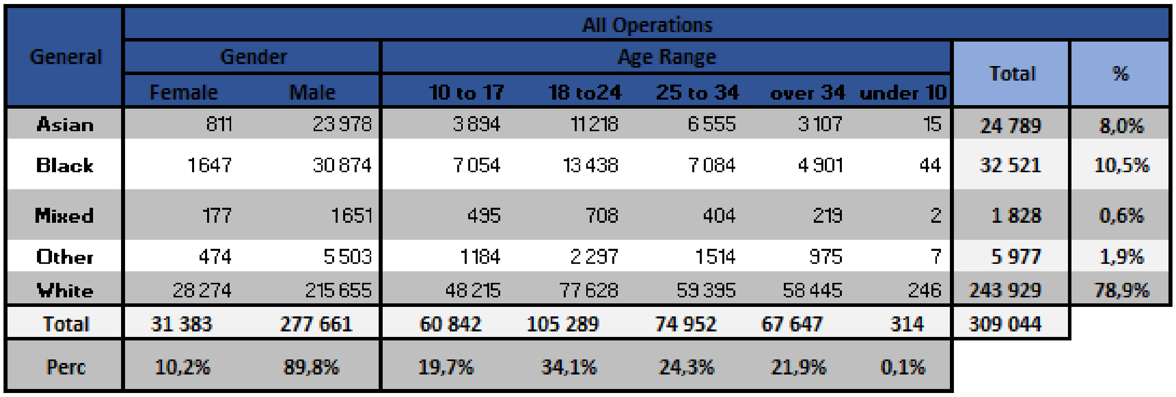


Fig 14. Distribution of results per self defined ethnicity



Variables	Number of present Values	Type	Number of Values Missing	Number of unique values
observation_id	309044	object	0	309044
Type	309044	object	0	3
Date	309044	datetime64[ns, UTC]	0	213812
Part of a policing operation	155606	object	153438	2
Latitude	230020	float64	79024	79834
Longitude	230020	float64	79024	80224
Gender	309044	object	0	2
Age range	309044	object	0	5
Self-defined ethnicity	303506	object	5538	19
Officer-defined ethnicity	309044	object	0	5
Legislation	281115	object	27929	17
Object of search	309044	object	0	16
Outcome	309044	object	0	16
Outcome linked to object of search	309044	object	0	2
Removal of more than just outer clothing	308981	object	63	2
station	309044	object	0	41

Fig 15. After treatment analysis [Back]



General	All Operations								
	Gender		Age Range					Total	%
	Female	Male	10 to 17	18 to24	25 to 34	over 34	under 10		
Asian	811	23 978	3 894	11 218	6 555	3 107	15	24 789	8,0%
Black	1647	30 874	7 054	13 438	7 084	4 901	44	32 521	10,5%
Mixed	177	1651	495	708	404	219	2	1 828	0,6%
Other	474	5 503	1 184	2 297	1 514	975	7	5 977	1,9%
White	28 274	215 655	48 215	77 628	59 395	58 445	246	243 929	78,9%
Total	31 383	277 661	60 842	105 289	74 952	67 647	314	309 044	
Perc	10,2%	89,8%	19,7%	34,1%	24,3%	21,9%	0,1%		

Fig 16. Data by gender, ethnicity and age range [Back]

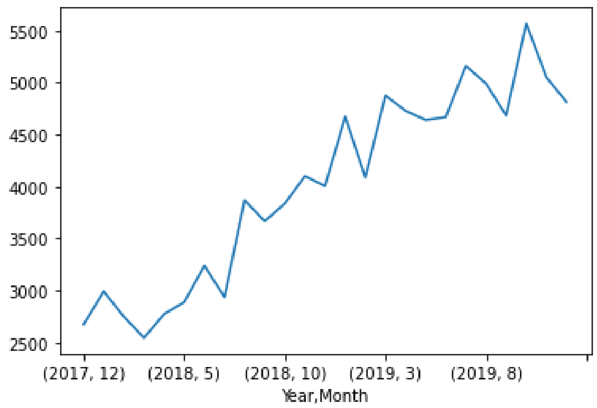


Fig 17. Evolution of positive outcomes over time

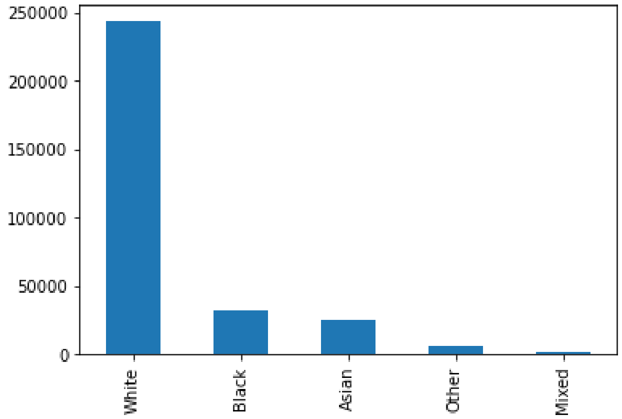
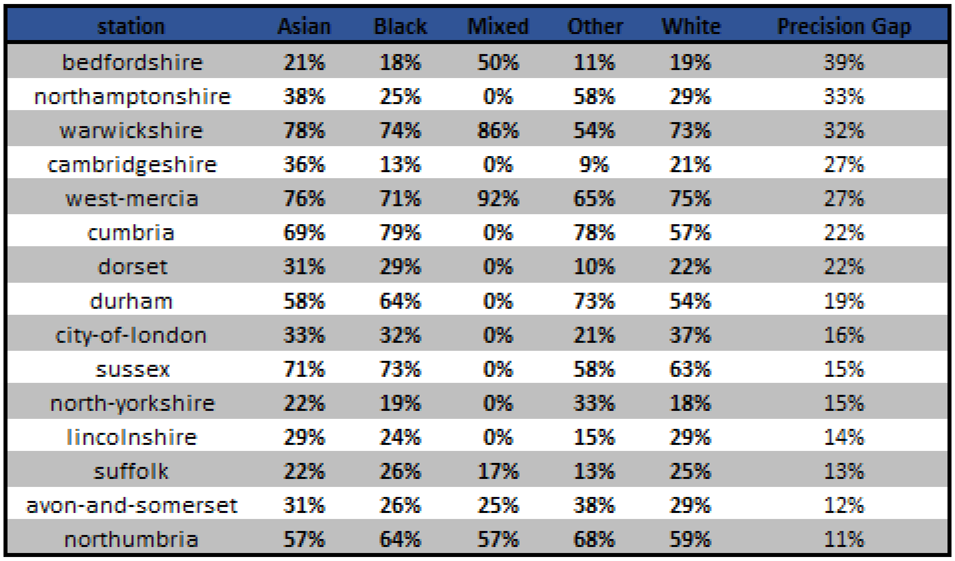


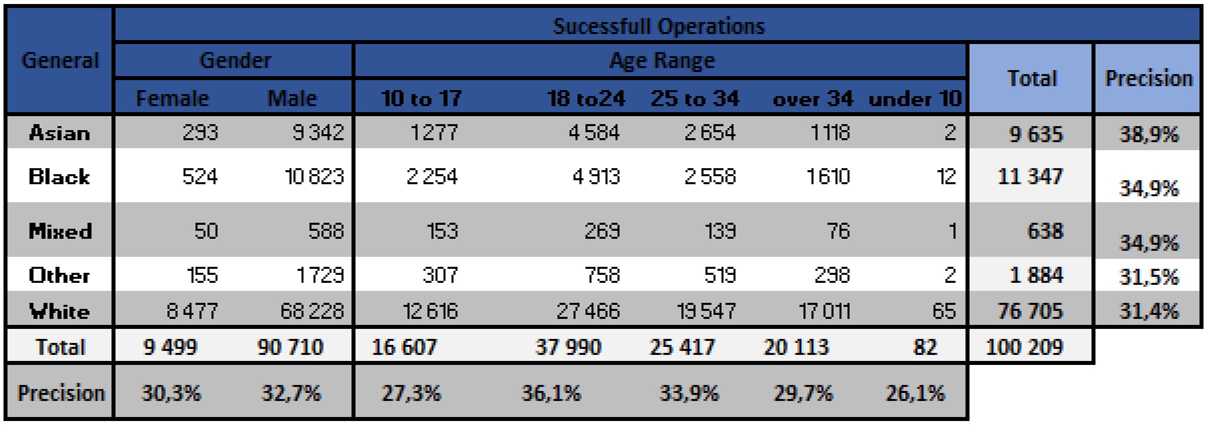
Fig 18. Used Ethnicities

Business Questions, Technical Support



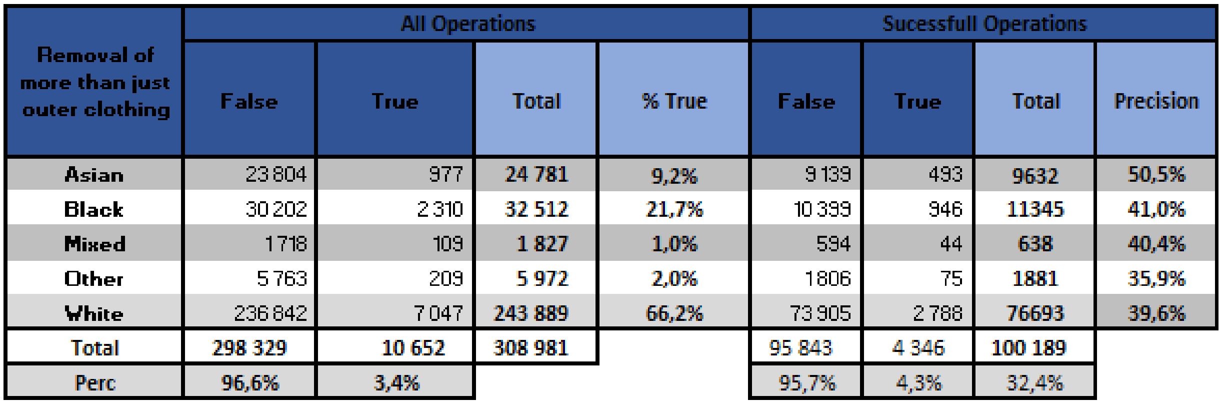
station	Asian	Black	Mixed	Other	White	Precision Gap
bedfordshire	21%	18%	50%	11%	19%	39%
northamptonshire	38%	25%	0%	58%	29%	33%
warwickshire	78%	74%	86%	54%	73%	32%
cambridgeshire	36%	13%	0%	9%	21%	27%
west-mercia	76%	71%	92%	65%	75%	27%
cumbria	69%	79%	0%	78%	57%	22%
dorset	31%	29%	0%	10%	22%	22%
durham	58%	64%	0%	73%	54%	19%
city-of-london	33%	32%	0%	21%	37%	16%
sussex	71%	73%	0%	58%	63%	15%
north-yorkshire	22%	19%	0%	33%	18%	15%
lincolnshire	29%	24%	0%	15%	29%	14%
suffolk	22%	26%	17%	13%	25%	13%
avon-and-somerset	31%	26%	25%	38%	29%	12%
northumbria	57%	64%	57%	68%	59%	11%

Fig 19. Top 15 precision gaps across ethnicities [Back]



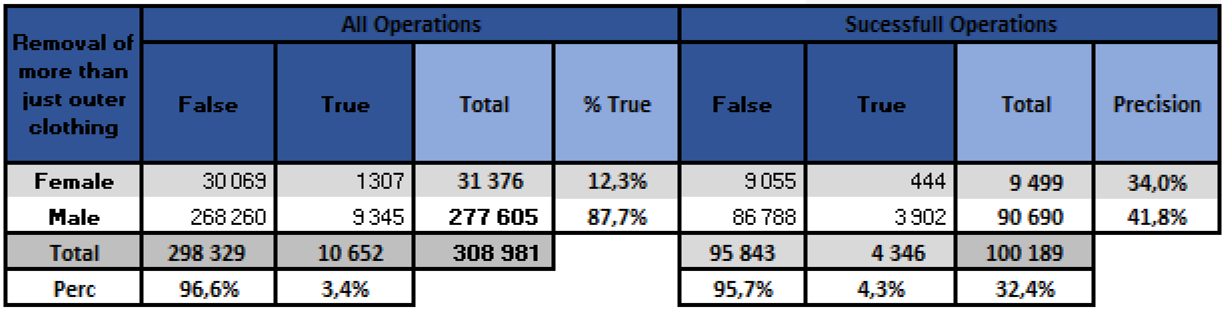
General	Sucessfull Operations								
	Gender		Age Range					Total	Precision
	Female	Male	10 to 17	18 to24	25 to 34	over 34	under 10		
Asian	293	9 342	1277	4 584	2 654	1 118	2	9 635	38,9%
Black	524	10 823	2 254	4 913	2 558	1610	12	11 347	34,9%
Mixed	50	588	153	269	139	76	1	638	34,9%
Other	155	1 729	307	758	519	298	2	1 884	31,5%
White	8 477	68 228	12 616	27 466	19 547	17 011	65	76 705	31,4%
Total	9 499	90 710	16 607	37 990	25 417	20 113	82	100 209	
Precision	30,3%	32,7%	27,3%	36,1%	33,9%	29,7%	26,1%		

Fig 20. Positive outcomes by gender, age range and ethnicity wit calculated precisions [Back]



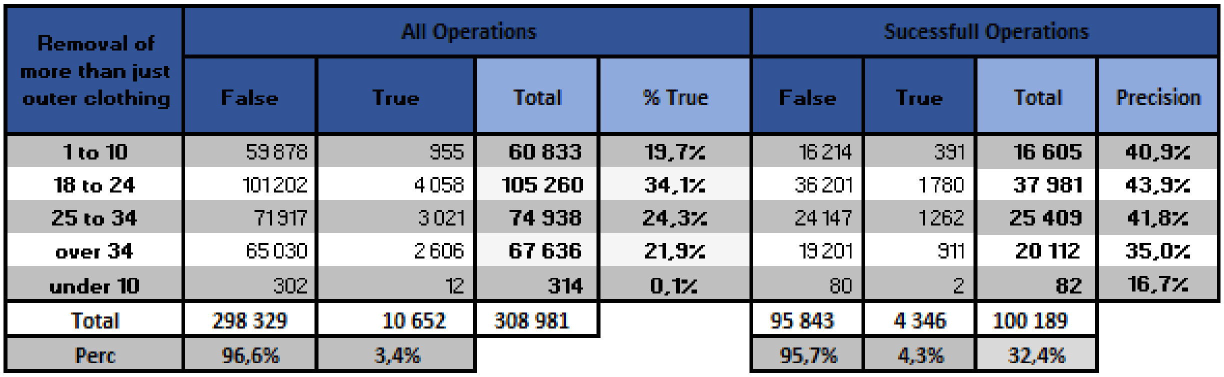
Removal of more than just outer clothing	All Operations				Sucessfull Operations			
	False	True	Total	% True	False	True	Total	Precision
Asian	23 804	977	24 781	9,2%	9 139	493	9632	50,5%
Black	30 202	2 310	32 512	21,7%	10 399	946	11345	41,0%
Mixed	1 718	109	1 827	1,0%	594	44	638	40,4%
Other	5 763	209	5 972	2,0%	1806	75	1881	35,9%
White	236 842	7 047	243 889	66,2%	73 905	2 788	76693	39,6%
Total	298 329	10 652	308 981		95 843	4 346	100 189	
Perc	96,6%	3,4%			95,7%	4,3%	32,4%	

Fig 21. Removal of more than outer clothes by ethnicity analysis [Back]



Removal of more than just outer clothing	All Operations				Sucessfull Operations			
	False	True	Total	% True	False	True	Total	Precision
Female	30 069	1307	31 376	12,3%	9 055	444	9 499	34,0%
Male	268 260	9 345	277 605	87,7%	86 788	3 902	90 690	41,8%
Total	298 329	10 652	308 981		95 843	4 346	100 189	
Perc	96,6%	3,4%			95,7%	4,3%	32,4%	

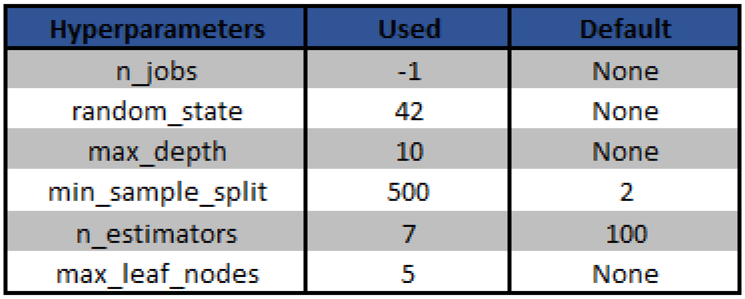
Fig 22. Removal of more than outer clothes by gender analysis [Back]



Removal of more than just outer clothing	All Operations				Sucessfull Operations			
	False	True	Total	% True	False	True	Total	Precision
1 to 10	59 878	955	60 833	19,7%	16 214	391	16 605	40,9%
18 to 24	101 202	4 058	105 260	34,1%	36 201	1 780	37 981	43,9%
25 to 34	71 917	3 021	74 938	24,3%	24 147	1 262	25 409	41,8%
over 34	65 030	2 606	67 636	21,9%	19 201	911	20 112	35,0%
under 10	302	12	314	0,1%	80	2	82	16,7%
Total	298 329	10 652	308 981		95 843	4 346	100 189	
Perc	96,6%	3,4%			95,7%	4,3%	32,4%	

Fig 23. Removal of more than outer clothes by age range analysis [Back]

Model Technical Analysis



Hyperparameters	Used	Default
n_jobs	-1	None
random_state	42	None
max_depth	10	None
min_sample_split	500	2
n_estimators	7	100
max_leaf_nodes	5	None

Fig 24. Random Forest Classifier Hyperparameters [Back]

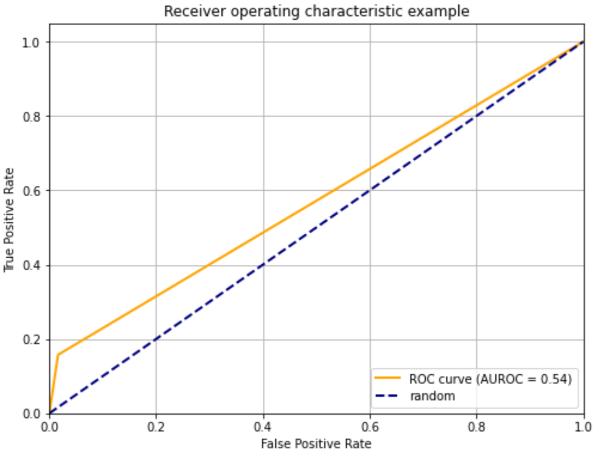
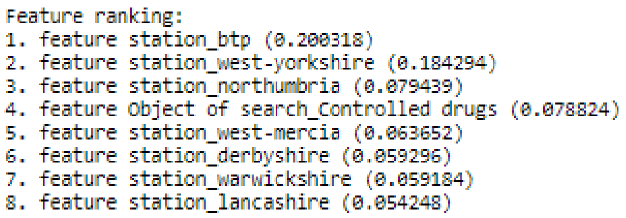


Fig 25. Model Roc_Auc curve [Back]



```
Feature ranking:
1. feature station_btp (0.200318)
2. feature station_west-yorkshire (0.184294)
3. feature station_northumbria (0.079439)
4. feature Object of search_Controlled drugs (0.078824)
5. feature station_west-mercia (0.063652)
6. feature station_derbyshire (0.059296)
7. feature station_warwickshire (0.059184)
8. feature station_lancashire (0.054248)
```

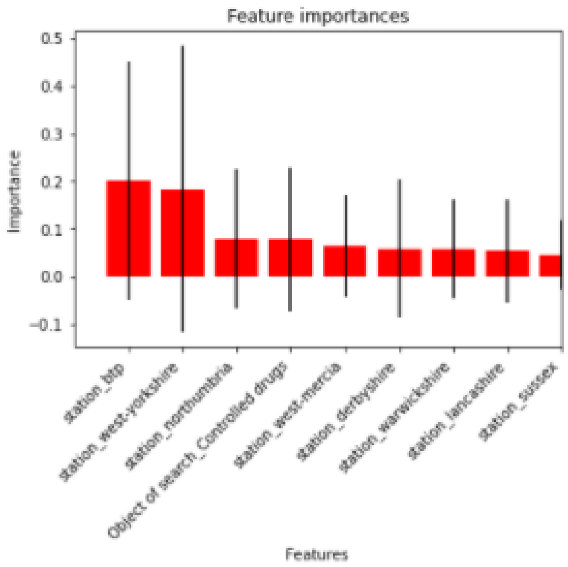


Fig 26: Feature importances[Back]

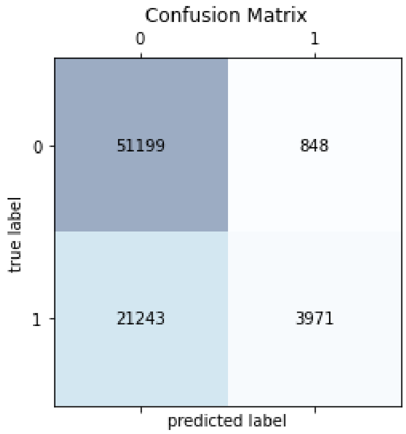
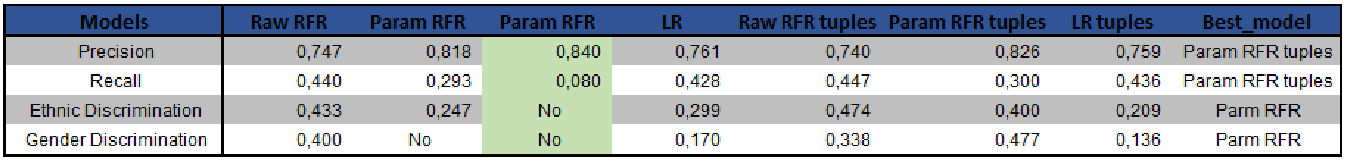


Fig 27. Model Confusion Matrix [Back]



Models	Raw RFR	Param RFR	Param RFR	LR	Raw RFR tuples	Param RFR tuples	LR tuples	Best_model
Precision	0,747	0,818	0,840	0,761	0,740	0,826	0,759	Param RFR tuples
Recall	0,440	0,293	0,080	0,428	0,447	0,300	0,436	Param RFR tuples
Ethnic Discrimination	0,433	0,247	No	0,299	0,474	0,400	0,209	Parm RFR
Gender Discrimination	0,400	No	No	0,170	0,338	0,477	0,136	Parm RFR

Fig 28. Alternative models results