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™

***28 of May 2023***

# Table Of Contents

# 1. Summary

The goal of our project was to assist the United Kingdom Department of Police in enhancing and standardizing their stop and search policy across all UK police stations. To achieve this, we developed an API platform that allows officers to validate stop and search requests in real-time. The API only validates requests if the predicted success probability for a search is greater than 10%. Its primary objectives were to improve the discovery rate (Precision) while maintaining a high level of overall offense detection (Recall), which were evaluated using the F5 score.

In the initial model, trained with 70% of the data as the training set, the performance metrics were as follows:

- Precision: 0.21
- Recall: 0.957
- F5 score: 0.842

After analyzing the first round of requests, we discovered that our API rejected all requests with missing values, except for the latitude and longitude fields. In total, we retrieved 2691 requests from which only 722 were updated with the true outcome (see unexpected problems). However, we were able to update and redeploy our API with the first 1000 requests, leading to significant improvements in the updated model:

- Precision: 0. 244
- Recall: 0.987
- F5_score: 0. 884

While the new model performed exceptionally well overall (see metrics above), it also exhibited great performance when evaluated per search objective (see model performance). However, it should be noted that a slight under-performance was observed for the Cambridgeshire station, as indicated in annex.1.

Regarding the business requirements, our analysis was limited to discrimination criteria 1, which focused on gender and ethnicity (see success on requirements), as the necessary features for criteria 2 and 3 were not available in the new dataset. Nonetheless, the new test set revealed an overall improvement in the search rate for discrimination criteria 1 compared to the training dataset (see success on requirements).

In summary, our updated model demonstrated significant performance improvements, and the quality of the more recent test dataset was notably enhanced.

# 2. Results Analysis

## 2.1 Model Performance

When assessing the performance of our initial model, we used the validation set from the original data to measure the success rate or discovery rate (Precision) and the overall ability to detect offenses (Recall). The F5 score, which prioritizes recall to maximize the identification of offenders, was used to select and fine-tune our model. Upon evaluating the new data with the test set and our updated model, we observed significant overall improvements compared to the validation set and our initial model, as seen in Figure 2.

Figure 2: Precision, Recall and F5-score metrics in validation and test set

| | precision | recall | F5_score |
|---|---|---|---|
| **Avg val set** | 0.210 | 0.957 | 0.842 |
| **Avg test set** | 0.244 | 0.988 | 0.884 |

However, it is important to note that the Cambridgeshire station exhibited poorer performance than the average metrics in the validation set (Figure.3). This observation should be interpreted with caution since these analyses were conducted on a reduced sample size of 1058 observations, which could potentially explain the observed poorer performance for the Cambridgeshire station.

Figure 3: Precision, Recall and F5 scores for each station of the new data (test set)

| | observations | precision | recall | F5_score |
|---|---|---|---|---|
| **bedfordshire** | 225 | 0.272727 | 0.982759 | 0.893309 |
| **cambridgeshire** | 117 | 0.196429 | 0.956522 | 0.832606 |
| **city of london** | 142 | 0.177778 | 1.000000 | 0.848980 |
| **devon and cornwall** | 298 | 0.225806 | 1.000000 | 0.883495 |
| **durham** | 83 | 0.385542 | 1.000000 | 0.942242 |
| **nottinghamshire** | 193 | 0.254144 | 0.978723 | 0.882006 |

Furthermore, when examining the model's performance across ethnicities, we found a generally balanced distribution and improvement in all metrics compared to the validation set (annex.2). However, it is worth mentioning that our model slightly underperformed for black and mixed ethnicities, which could potentially be attributed to the reduced sample size in our testing.

Overall, our updated model outperforms the initial model, which could be attributed to changes made in our model pipeline during re-deployment (see re-deployment section). Additionally, the utilization of a more recent dataset (test set) may reflect the improved performance of the police department over time, potentially due to changes in agent conduct.

## 2.2 Success on requirements

In our previous report, we have identified and analysed three different criteria to investigate if discrimination claims made to the press were supported by the data. However, based on the information provided by the new dataset (test set), we can only continue investigating one of these criteria (as “Removal of more than just outer clothing” and “Self-defined ethnicity” features were not available in the request) :

- ***Criteria_1***: gender and ethnicity discrimination in stop and search (using success rate metric)

We created sub-groups for each combination of ´Gender’ + ‘station’ + ‘Officer-defined ethnicity’ with more than 30 occurrences. This approach reduced our test set from 36 to 10 subgroups representing 77.88% of searched cases. As in report 1, we analysed the discrimination criteria 1 locally within each station (with +/- 5% discrepancy) and globally between all stations (with +/- 10% discrepancy).

***Criteria_1***:
- **locally** within each station (annex.3):
    - 2 stations have at least one subgroup with more than 5% discrepancy
        Stations: **Bedfordshire and City-of-london**
    - 1 station has have all their subgroups with more than 5% discrepancy
        Station: **City-of-london**
    - Black Male is the most represented subgroup with more than 5% discrepancy (see “dis” metric annex.3)
    - Black Male is most discriminated subgroup with more than 5% discrepancy (see “first” metric annex.3)
    - **Good station : Cambridgeshire, devon-and-cornwall, durham, nottinghamshire**
        (no discrepancy with other subgroups and success rate station)
        ( all station subgroups success rate > 10%)

Compared to the discrimination criteria 1 results of the original data (annex.3B), the test data shows an overall improvement in search rate for 4 out of 6 stations. However, unlike the original dataset, the test set reveals that 2 stations have subgroups exhibiting a first-degree discrimination pattern (see “first” metric annex.3). This discrepancy could potentially be attributed to the smaller sample size of the new dataset, which may affect statistical representation and lead to variations in the subgroup patterns.

- **globally** between all stations (annex.4):
    - results for global discrimination criteria 1 are similar to the original dataset, except for Durham station.
    - Durham station exhibits a discrepancy of more than 10% compared to the mean success rate across all stations. However, this difference is associated with an increase in the success rate of the Durham station.
    - 4 out of 6 stations exhibit an improvement in their success rate compared to the original dataset.

Next, we analyse the performance of our model per object of search by comparing the new test set to the original dataset. This analysis is represented in Figure 4.

Figure 4: Model performance per object of search

| | Off | %R_test | %P_test | %R_val | %P_val | imp_R | imp_P |
|---|---|---|---|---|---|---|---|
| **Controlled drugs** | 200.0 | 100.0 | 26.74 | 99.40 | 22.68 | 1.0 | 1.0 |
| **Article for use in theft** | 18.0 | 100.0 | 24.66 | 63.45 | 12.74 | 1.0 | 1.0 |
| **Offensive weapons** | 15.0 | 80.0 | 12.63 | 75.28 | 12.30 | 1.0 | 1.0 |
| **Stolen goods** | 13.0 | 100.0 | 17.81 | 100.00 | 20.47 | 0.0 | -1.0 |
| **Articles for use in criminal damage** | 1.0 | 100.0 | 25.00 | 75.56 | 13.73 | 1.0 | 1.0 |
| **Firearms** | 0.0 | 0.0 | 0.00 | 74.36 | 10.70 | NaN | NaN |
| **Fireworks** | 0.0 | 0.0 | 0.00 | 40.00 | 9.30 | NaN | NaN |

Off : Numbers of Offenders
%R_test: recall for the test set (in %)
%P_test: precision for the test set (in %)
%R_val: recall for the original dataset (in %)
%P_val: precision for the original dataset (in %)
Imp_R: improved recall from original dataset to test set (1=increase; -1=decrease; 0=stable)
Imp_P: improved precision from original dataset to test set
(1=increase; -1=decrease; 0=stable)

When comparing the performance of our updated model with the new test set to our initial model with the original dataset, we observe that our updated model generally outperforms in terms of overall offense detection (recall) and discovery rate (precision) for most object of search categories (Figure 4). However, it should be noted that there is an exception in the category of Stolen goods, where our updated model may not show significant improvements. Overall, these improvements could be attributed to changes made in our model pipeline (see re-deployment) but also by the utilization of a more recent dataset (test set), which may reflect the improved performance of the police department over time.

Similarly, when analysing our model's performance per object of search and police stations (annex.5), we observed that our updated model with the new test set outperforms our initial model with the original dataset in terms of overall offense detection (annex 5A) and discovery rate (annex.5B). However, it is worth noting that our updated model underperforms only in terms of discovery rate for the Cambridgeshire and City-of-london stations (annex.5B).

## 2.3 Population Analysis

After cleaning the original dataset, we got 394183 observations:

- Removing stations with no data in “Outcome linked to object of search” (Metropolitan, Humberside and Lancashire stations)
- Filling the null values in “Outcome linked to object of search” by False
- Creating a successful search feature (when outcome is positive and Outcome linked to object of search is True)

On the other hand, our new test set retrieved by our API contains only 1058 observations (see unexpected problems). These two dataframes have the following characteristics (Figure 5):

Figure 5: Unique and missing values in original (A) and test (B) dataset,

A.

| train | unique | missing | %unique | %missing |
|---|---|---|---|---|
| observation_id | 394183 | 0 | 100.00 | 0.00 |
| Type | 3 | 0 | 0.00 | 0.00 |
| Date | 254555 | 0 | 64.58 | 0.00 |
| Part of a policing operation | 2 | 184430 | 0.00 | 46.79 |
| Latitude | 96528 | 90769 | 24.49 | 23.03 |
| Longitude | 96998 | 90769 | 24.61 | 23.03 |
| Gender | 3 | 0 | 0.00 | 0.00 |
| Age range | 5 | 0 | 0.00 | 0.00 |
| Officer-defined ethnicity | 5 | 0 | 0.00 | 0.00 |
| Legislation | 19 | 39747 | 0.00 | 10.08 |
| Object of search | 17 | 0 | 0.00 | 0.00 |
| station | 38 | 0 | 0.01 | 0.00 |
| Successful_search | 2 | 0 | 0.00 | 0.00 |

B.

| test | unique | missing | %unique | %missing |
|---|---|---|---|---|
| observation_id | 1058 | 0 | 100.00 | 0.00 |
| Type | 2 | 0 | 0.19 | 0.00 |
| Date | 919 | 0 | 86.86 | 0.00 |
| Part of a policing operation | 2 | 335 | 0.19 | 31.66 |
| Latitude | 473 | 301 | 44.71 | 28.45 |
| Longitude | 473 | 301 | 44.71 | 28.45 |
| Gender | 3 | 0 | 0.28 | 0.00 |
| Age range | 5 | 0 | 0.47 | 0.00 |
| Officer-defined ethnicity | 5 | 0 | 0.47 | 0.00 |
| Legislation | 5 | 0 | 0.47 | 0.00 |
| Object of search | 7 | 0 | 0.66 | 0.00 |
| station | 6 | 0 | 0.57 | 0.00 |
| Successful_search | 2 | 0 | 0.19 | 0.00 |

Next, we analyse if the new dataset (test set) contains any missing and extra feature categories.

Figure 6: Missing and extra feature categories

| Categories | missing | extra |
|---|---|---|
| Type | 1 | 0 |
| Gender | 0 | 0 |
| Age range | 0 | 0 |
| Officer-defined ethnicity | 0 | 0 |
| Legislation | 15 | 0 |
| Object of search | 10 | 0 |
| station | 32 | 0 |
| Successful_search | 0 | 0 |

Missing categories:

- Type: “vehicle search”
- Object of search: 'Anything to threaten or harm anyone', 'Crossbows', 'Detailed object of search unavailable', 'Evidence of hunting any wild mammal with a dog', 'Evidence of offences under the Act', 'Evidence of wildlife offences', 'Game or poaching equipment', 'Goods on which duty has not been paid etc.', 'Psychoactive substances', 'Seals or hunting equipment'

We observe that in the new dataset (test set) some categories are missing in type, legislation, object of search and station features. Legislation and station lists are too long to show.

The remaining features have practically identical distributions (annex.6), with the exceptions:

- “Part of a policing operation”: more True cases (from 8.8 to 15.2%)
- “Officer-defined ethnicity”: less White and more all other categories
- “Successful search”: more True cases (from 17.9 to 23.3%)

These changes in the category distribution may have an impact on the results obtained.

# 3. Deployment Issues

## 3.1 Re-deployment

During the first round of requests, unexpected issues were encountered, resulting in a lower-than-expected number of requests retrieved by our API (see unexpected problems). To address this issue, we decided to update our API to accept requests with missing values (see unexpected problems) and subsequently proceeded to re-deploy it on the railway platform.

During the API redeployment, we made important modifications to our pipeline model. These changes include:

- Implementing under-sampling of the majority category of the target variable using RandomUnderSampler with a sampling strategy of 0.5 and a random state of 42.

- Extracting the "Hour" and "Weekday" features from the "Date" feature by converting it to datetime using pd.Datetime, and then using dt.hour and dt.day_name to extract the new features.

These modifications resulted in significant improvements in our updated model, as illustrated in Figure 2.

## 3.2 Unexpected problems

Upon analysing the first round of requests, we found that our API rejected all requests with missing values, except for the latitude and longitude fields where np.nan was considered as a float. As a result, only 2691 out of the initial 8000 requests were retrieved. Additionally, due to an involuntary disconnection of our API deployment, only 722 of these retrieved requests were successfully updated with the true outcome. However, we were able to update and redeploy our API with the first 1000 requests that were kindly provided to us.

Our updated API will now check and accept any request with the following contents:

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

The updated API now allows for missing values in all request fields except for the "observation_id" field, which remains mandatory for a valid request. This modification enhances the flexibility in handling requests with missing data.

## 3.3 Learnings and Future Improvements

Our updated model, trained on a new dataset (test set), showed significant improvements compared to our initial model using the original dataset (refer to Figure 2). One notable improvement was achieved by undersampling the majority category of the target variable, indicating that our model could perform even better if the police stop and search dataset were more balanced in terms of the number of positive outcome cases. Considering the superior performance of the updated model and the apparent improvement in the police department's success rate (see model performance and success requirements sections), regular re-training of our model with continually improved datasets will undoubtedly enhance its performance.

Furthermore, analysing the original dataset across different collection periods revealed variations in the number of stop and search incidents and corresponding success rates during and after the COVID lockdown restrictions (annex.7). This highlights, once again, the importance of regular re-training to stay updated with the latest information and adapt to emerging patterns or changes over time.

Another noteworthy improvement in our updated model was achieved by incorporating new features such as "Hour" and "Weekday" extracted from the original "Date" feature. Building on this success, additional valuable features can be derived from the "Date" feature:

- Month/Season: Extracting the month or season from the datetime feature provides insights into seasonal patterns of criminal behaviour, such as increased thefts during holidays or specific patterns in certain months.

- Public Holidays: Identifying public holidays or special events helps understand variations in crime rates. Holidays introduce unique patterns of criminal activity, and incorporating this information enhances the model's accuracy in predicting crime trends during those periods.

Overall, by implementing regular re-training, balancing the dataset, and enriching our model with relevant features, we can significantly enhance its performance, adaptability, and accuracy in predicting outcomes and identifying trends in police stop and search incidents.

# 4. Learnings and Future Improvements

Throughout the development of our model, several key learnings have emerged that can guide future improvements and advancements in UK police stop and search performance and policy.

Firstly, enhancing the quality of input data by ensuring its accuracy, completeness, and representativeness of various demographic groups is essential to mitigate bias and improve overall fairness. It is crucial for police officers to be aware of issues such as missing values in latitude or longitude fields that can affect model performance.

Secondly, incorporating a wider range of features can significantly enhance our model's performance. For example, we observed that extracting new features from the "Date" field improved our predictions. Additional features to consider include:

- Weather conditions: Including weather data (temperature, precipitation, visibility) helps assess the influence of external factors on search outcomes.
- Officer experience: Collecting officers' experience level (years of service, specialized training) evaluates the impact of expertise on search outcomes.
- Recording officer's perception: Adding a field for subjective officer perceptions captures additional factors influencing search outcomes.
- Number of people in the car: Including the count of individuals in the vehicle provides insights into search dynamics and potential selectivity based on occupant numbers.
- Existence of a large event within a closed radius (5 or 10 km): This feature indicates significant nearby events (festivals, protests, sports) that impact search outcomes due to crowd dynamics, foot traffic, and security measure

These features provide valuable contextual information that can help account for external factors and better predict search outcomes.

Exploring different machine learning algorithms, such as ensemble methods or deep learning architectures, is another avenue for improvement. Experimenting with these algorithms can lead to better performance and more accurate predictions.

Furthermore, taking an exploratory approach to improve our model predictions can involve incorporating real-time data sources, such as social media or crowd-sourced information, to enhance accuracy and timeliness. Additionally, utilizing natural language processing techniques to analyse textual data, such as police reports or public feedback, can provide valuable insights into patterns and trends in stop and search practices.

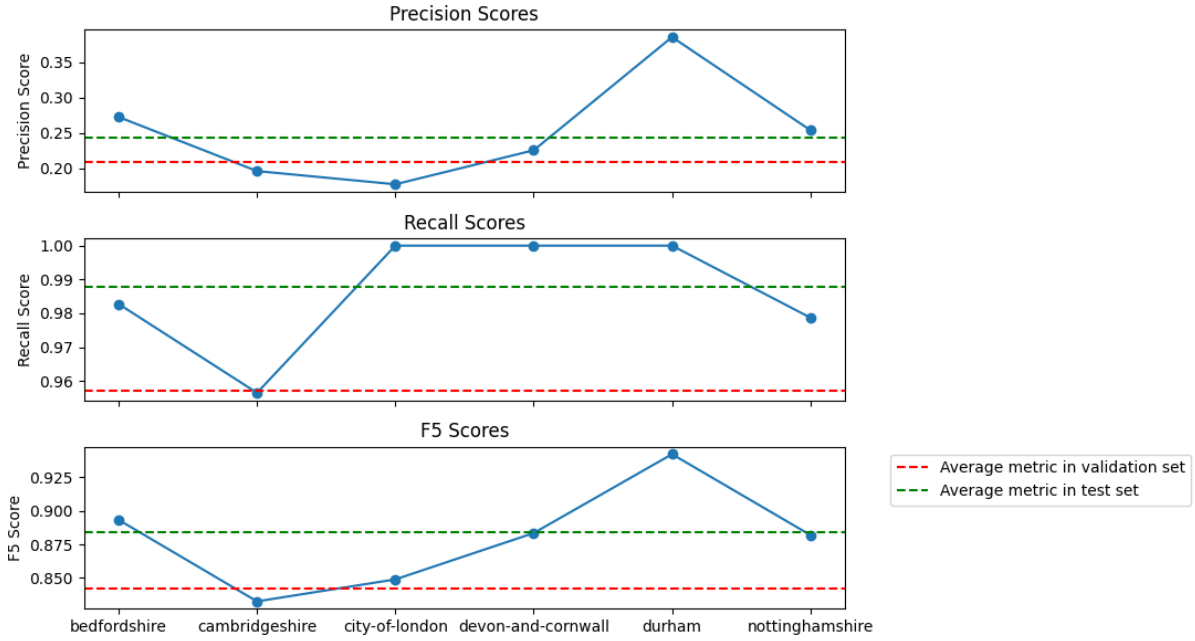
Considering implementing other tasks in our API, such as validating requests for suspects to remove clothes, can further expand the capabilities and usefulness of our model. Additionally, exploring projects that study zones and schedules with the majority of criminal activity to optimize the routes and schedules of patrol cars can support more efficient policing efforts.

Finally, from a business perspective, creating a platform to share real-time statistics and insights with each community (station/city) can serve as an awareness campaign and improve the Department's public image. By promoting transparency and proactive communication, the platform fosters trust, encourages dialogue, and ensures a better-informed community.

By incorporating these learnings and pursuing these future improvements, we can advance the effectiveness, fairness, and efficiency of police stop and search practices in the UK.

# 5. Annexes

Annex.1: Precision, Recall and F5-score metrics across stations in the test set




Annex.2: Precision, Recall and F5-score metrics across ethnicities in validation and test set

| | precision_val | precison_test | recall_val | recall_test | F5_score_val | F5_score_test | test_F5_improved |
|---|---|---|---|---|---|---|---|
| **White** | 0.210784 | 0.253835 | 0.957048 | 0.989130 | 0.842346 | 0.889976 | True |
| **Black** | 0.208377 | 0.168067 | 0.957955 | 0.952381 | 0.841526 | 0.807453 | False |
| **Asian** | 0.200357 | 0.271318 | 0.955319 | 1.000000 | 0.834394 | 0.906375 | True |
| **Mixed** | 0.231539 | 0.083333 | 0.978836 | 1.000000 | 0.870746 | 0.702703 | False |
| **Other** | 0.201850 | 0.272727 | 0.951282 | 1.000000 | 0.832413 | 0.906977 | True |

Annex.3: Tables summarizing analysis of discrimination criteria_1 (locally)

A. Test set:

| Station | Gender | ODE | obs | num_S | SR | stat_SR | dis | n_dis | dis_stat | first | second |
|---|---|---|---|---|---|---|---|---|---|---|---|
| bedfordshire | Male | Asian | 69 | 18 | 0.26 | 0.26 | True | 1 | False | False | False |
| | | Black | 31 | 3 | 0.10 | 0.26 | True | 2 | True | True | False |
| | | White | 95 | 25 | 0.26 | 0.26 | True | 1 | False | False | False |
| cambridgeshire | Male | White | 70 | 16 | 0.23 | 0.20 | False | 0 | False | False | False |
| city-of-london | Male | Black | 35 | 4 | 0.11 | 0.17 | True | 1 | True | True | False |
| | | White | 57 | 11 | 0.19 | 0.17 | True | 1 | False | False | False |
| devon-and-cornwall | Female | White | 51 | 9 | 0.18 | 0.21 | False | 0 | False | False | False |
| | Male | White | 226 | 46 | 0.20 | 0.21 | False | 0 | False | False | False |
| durham | Male | White | 67 | 28 | 0.42 | 0.39 | False | 0 | False | False | False |
| nottinghamshire | Male | White | 123 | 34 | 0.28 | 0.24 | False | 0 | False | False | False |

B. Original dataset:

| Station | Gender | ODE | obs | num_S | SR | stat_SR | dis | n_dis | dis_stat | first | second |
|---|---|---|---|---|---|---|---|---|---|---|---|
| bedfordshire | Male | Asian | 1670 | 377 | 0.23 | 0.21 | True | 2 | False | False | False |
| | | Black | 937 | 217 | 0.23 | 0.21 | True | 2 | False | False | False |
| | | White | 2631 | 549 | 0.21 | 0.21 | True | 1 | False | False | False |
| cambridgeshire | Male | White | 2214 | 523 | 0.24 | 0.24 | True | 1 | False | False | False |
| city-of-london | Male | Black | 996 | 243 | 0.24 | 0.25 | True | 1 | False | False | False |
| | | White | 1949 | 503 | 0.26 | 0.25 | True | 1 | False | False | False |
| devon-and-cornwall | Female | White | 1678 | 277 | 0.17 | 0.18 | True | 1 | False | False | False |
| | Male | White | 8560 | 1588 | 0.19 | 0.18 | True | 1 | False | False | False |
| durham | Male | White | 2854 | 794 | 0.28 | 0.28 | True | 1 | False | False | False |
| nottinghamshire | Male | White | 4685 | 1182 | 0.25 | 0.23 | True | 4 | False | False | False |

- 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.4: Tables summarizing analysis of discrimination criteria_1 (globally)

A. Test set

| station | obs | num_S | SR | all_SR | n_dis | dis_global | all_SR_org | dis_global_org |
|---|---|---|---|---|---|---|---|---|
| **city-of-london** | 142 | 24 | 0.1690 | 0.2335 | 1 | False | 0.1786 | False |
| **cambridgeshire** | 117 | 23 | 0.1966 | 0.2335 | 1 | False | 0.1786 | False |
| **devon-and-cornwall** | 298 | 63 | 0.2114 | 0.2335 | 1 | False | 0.1786 | False |
| **nottinghamshire** | 193 | 47 | 0.2435 | 0.2335 | 1 | False | 0.1786 | False |
| **bedfordshire** | 225 | 58 | 0.2578 | 0.2335 | 1 | False | 0.1786 | False |
| **durham** | 83 | 32 | 0.3855 | 0.2335 | 5 | True | 0.1786 | True |

B. Original dataset

| station | obs | num_S | SR | all_SR | n_dis | dis_global |
|---|---|---|---|---|---|---|
| **city-of-london** | 4539 | 1144 | 0.2520 | 0.1786 | 7 | False |
| **cambridgeshire** | 3367 | 801 | 0.2379 | 0.1786 | 6 | False |
| **devon-and-cornwall** | 10906 | 1992 | 0.1827 | 0.1786 | 4 | False |
| **nottinghamshire** | 7523 | 1757 | 0.2336 | 0.1786 | 6 | False |
| **bedfordshire** | 5904 | 1258 | 0.2131 | 0.1786 | 5 | False |
| **durham** | 3585 | 987 | 0.2753 | 0.1786 | 10 | False |

- obs = number of observations per station
- num_S = number of successful search per station
- SR = success rate of the station
- all_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
- all_SR_org = average success rate of all stations in original dataset
- dis_global_org = 10% discrepancy with the average success rate of all station (from original dataset)

## Annex.5: Tables summarizing metrics improvements comparing updated-model/test-set and initial-model/original-model

A. overall ability to detect offenses (recall)

| | nottinghamshire | cambridgeshire | city-of-london | devon-and-cornwall | durham | bedfordshire |
|---|---|---|---|---|---|---|
| **Offensive weapons** | 1.0 | -1.0 | NaN | 1.0 | 0.0 | 1.0 |
| **Controlled drugs** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **Article for use in theft** | NaN | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 |
| **Stolen goods** | 0.0 | NaN | 0.0 | 0.0 | 0.0 | NaN |
| **Firearms** | NaN | NaN | NaN | NaN | NaN | NaN |
| **Articles for use in criminal damage** | NaN | NaN | NaN | NaN | NaN | 1.0 |
| **Fireworks** | NaN | NaN | NaN | NaN | NaN | NaN |
| **improved_R_sum** | 1.0 | 0.0 | 0.0 | 2.0 | 0.0 | 3.0 |

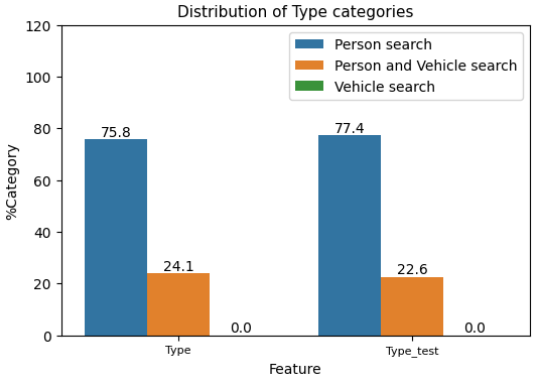
Recall improvements: 1 = increase ; -1 = decrease and 0 = stable
Improved_R_sum: the sum of recall improvement score per station

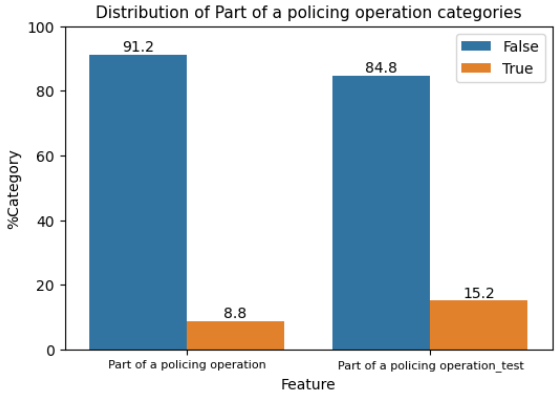
B. discovery rate (precision)

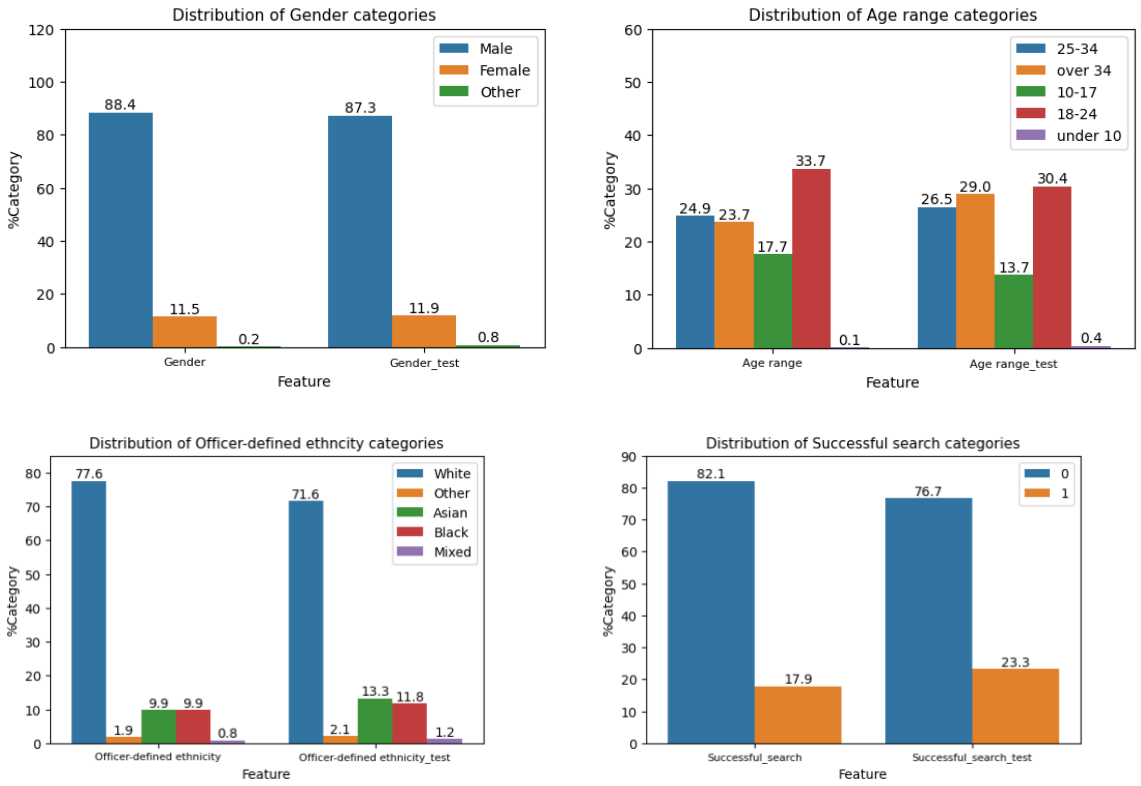
| | nottinghamshire | cambridgeshire | city-of-london | devon-and-cornwall | durham | bedfordshire |
|---|---|---|---|---|---|---|
| **Offensive weapons** | 1.0 | -1.0 | NaN | 1.0 | 1.0 | 1.0 |
| **Controlled drugs** | 1.0 | -1.0 | -1.0 | 1.0 | 1.0 | 1.0 |
| **Article for use in theft** | NaN | 1.0 | -1.0 | 1.0 | 1.0 | 1.0 |
| **Stolen goods** | -1.0 | NaN | -1.0 | -1.0 | 1.0 | NaN |
| **Firearms** | NaN | NaN | NaN | NaN | NaN | NaN |
| **Articles for use in criminal damage** | NaN | NaN | NaN | NaN | NaN | 1.0 |
| **Fireworks** | NaN | NaN | NaN | NaN | NaN | NaN |
| **improved_P_sum** | 1.0 | -1.0 | -3.0 | 2.0 | 4.0 | 4.0 |

Precision improvements: 1 = increase ; -1 = decrease and 0 = stable
Improved_P_sum: the sum of precision improvement score per station

## Annex.6: Categories distribution of test set features

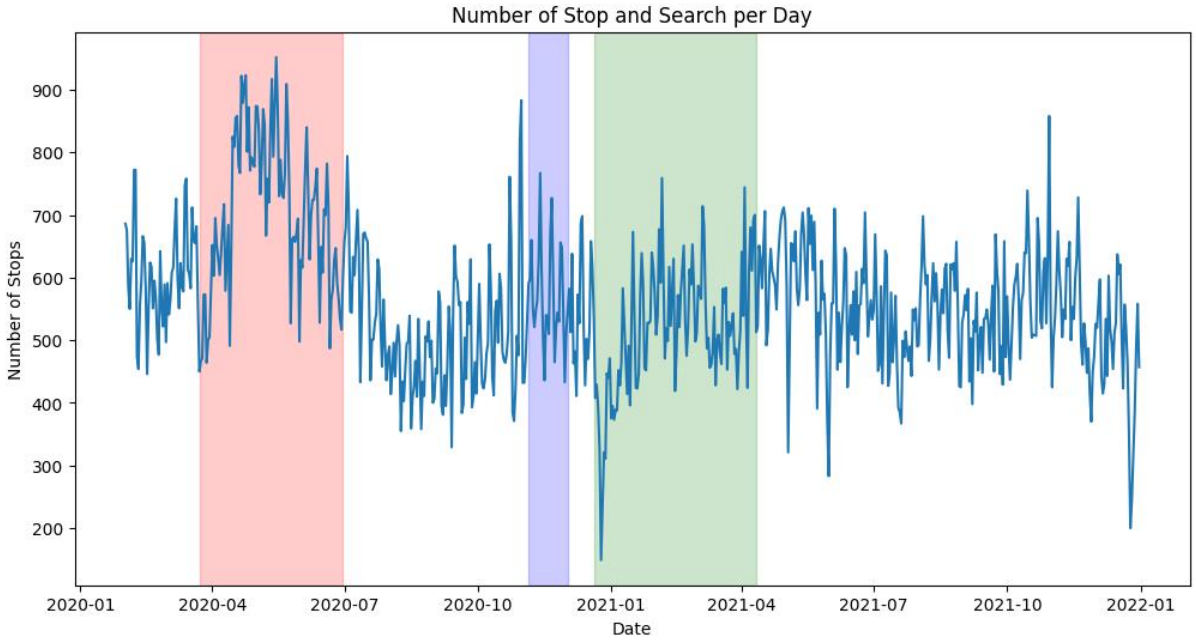


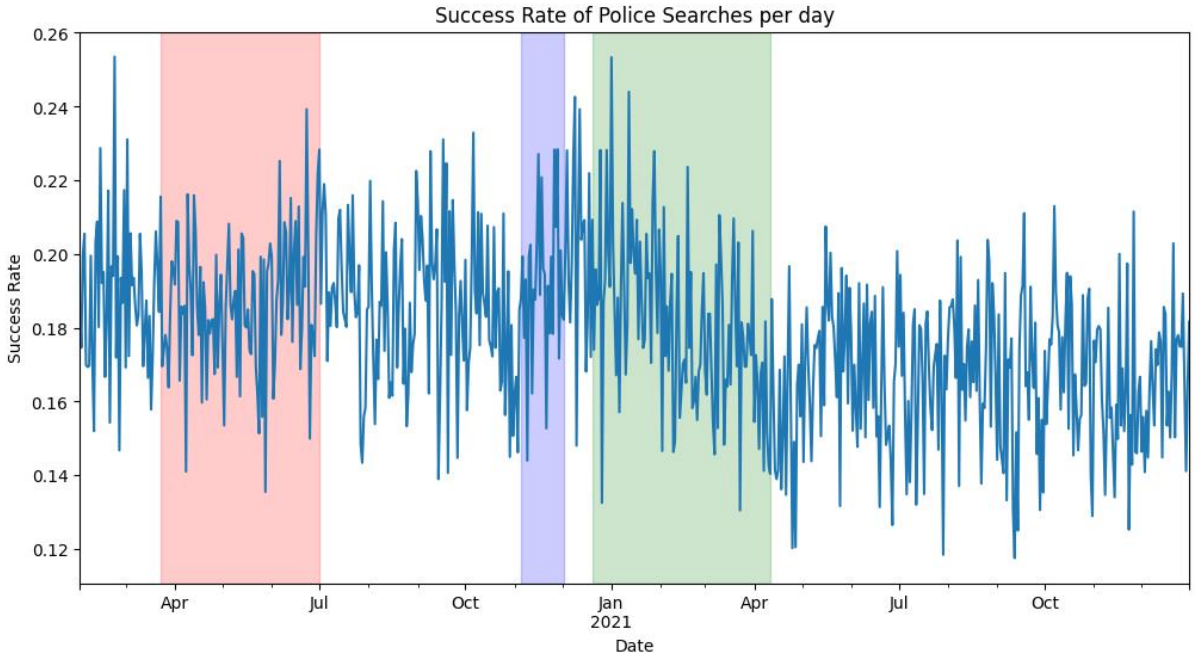



## Annex.7: Analysis of stop and search and success rate overtime

### A. Number of stop and search per day



B. Success rate per day




Covid lockdown periods:

- 1st lockdown period (in red): From 2020-03-23 to 2020-06-30
- 2nd lockdown period (in blue): From 2020-11-05 to 2020-12-02
- 3rd lockdown period (in green): From 2020-12-20 to 2021-04-11

End of COVID restriction period: From 2021-04-11 onwards