# 

[**Business Conclusions**](#_u71r55nwzx40) **1**

[Summary](#_9m7touwwv6pf) 1

[**Results Analysis**](#_rhz120fomue5) **2**

[Model Performance](#_juq669imx7de) 2

[Success on requirements](#_qaopfmb30cz7) 4

[Population Analysis](#_mkal2eah8yld) 6

[**Next Steps**](#_dy4d8b7g9s91) **9**

[Next Steps](#_tdgyr8blrsjw) 9

[**Deployment Issues**](#_thagbmcdkz0) **10**

[Redeployment](#_lu2zwdxqoddn) 10

[Unexpected problems](#_5phooh4xivzc) 11

[What would you do differently next time](#_iti1vfthh97j) 12

[**Annexes**](#_2t6hs9vqljyg) **12**

[New model details](#_5n3pntq3iuft) 12

# Business Conclusions

## Summary

The expectation on the proof of concept was to be able to increase the overall success rate, having a low diminishing of overall discovery rate and having some discrepancy between protected classes, while aligned with the variation observed during the model development.

Regarding the technical measures, Table 1, the global precision was higher than expected, which is a good result once it means that the stop operations were more successful. On the other hand, the overall discovery rate diminished more than expected.

|  |  |  |
| --- | --- | --- |
| **Technical Measures** | | |
|  | **Deployed Model** | **Proof of concept (Testing set 1)** |
| AUROC | 0.61 | 0.56 |
| Global precision | 0.25 | 0.32 |
| Global recall | 0.85 | 0.67 |

*Table 1 -Technical measures of Deployed Model and Proof of concept*

The overall precision of the whole proof-of-concept, which is “Testing set 1” and “Testing set 2”, was 24.0%.

In regards to the fulfillment of requirements, Table 2, it was possible to reach more than 10% of success rate for all the stations, but none of them fulfills the search object requirement. The success rate variation between stations per Ethnicity and Gender are higher than 5 p.p. but acceptable once they are aligned with the deployed model expectation.

|  |  |  |  |
| --- | --- | --- | --- |
| **Requirements** | **Subpopulation** | **Deployed Model**  *(38 stations evaluated)* | **Proof of concept**  *(4 stations evaluated****)*** |
| Minimum 10% success rate for searches per station and search objective | Station | 37 stations fulfill the requirement | all stations fulfill the requirement |
| Search objective | 20 stations fulfill the requirement | 0 stations fulfill the requirement |
| Discrepancy between stations would not be larger than 10 percentage points | Station | 30 p.p. of difference | 42 p.p. of difference |
| No police station should have a discrepancy bigger than 5% for search success rate between protected classes ethnicity and gender | Ethnicity | 8 stations fulfill the requirement  Average difference: 8.9% | 0 stations fulfill the requirement  Average difference: 8.6% |
| Gender | 20 stations fulfill the requirement  Average difference: 8.0% | 1 station fulfill the requirement  Average difference: 10.0% |

*Table 2 - Requirements fulfillment of Deployed Model and Proof of concept*

# Results Analysis

## Model Performance

As briefly discussed and shown in the previous section through Table 1, the precision of the model in the new observations (first part of the proof of concept named here as “Testing set 1”) was higher than the expectation, but the recall was lower than expected.

Analyzing the model performance per Station, Table 3, it is possible to see that City-of-London and Durham had the lowest recall. Nottinghamshire had a good overall performance and Cambridgeshire had only 31 observations which does not allow further analysis and conclusions.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Proof of concept** | **Nottinghamshire** | **City-of-london** | **Durham** | **Cambridgeshire** |
| Number of Observations | 4,000 | 2,203 | 1,272 | 494 | 31 |
| Precision | 0.32 | 0.29 | 0.27 | 0.64 | 0.22 |
| Recall | 0.67 | 0.75 | 0.61 | 0.56 | 1.00 |

*Table 3 - Proof of concept Technical measures per Station*

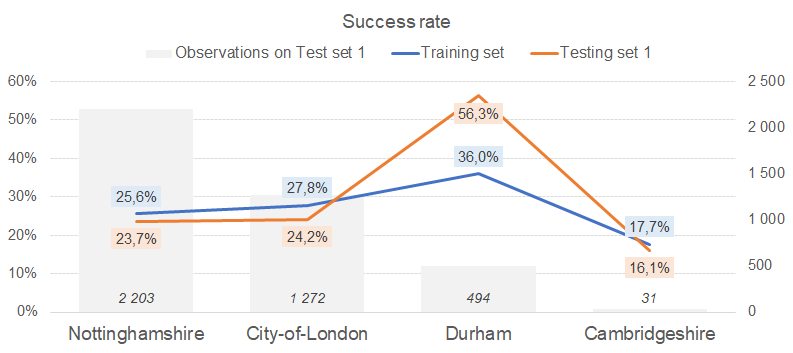
This drop of performance can be explained due to population changes over time, and also due to the fact that the subset of stations in the proof-of concept is reasonably different from the set of stations used to estimate the model (only 4 of the 38 stations used during the model development).

Starting looking at the success rate, Table 4, it is possible to see that the subset of stations used in the proof-of-concept has a higher success rate than the observed on the training set:

|  |  |  |
| --- | --- | --- |
|  | **Training set** | **Testing set 1** |
| **Success rate** | 20,7% | 27,8% |

*Table 4 - Success rate*

Detailing this analysis per station, Graphic 1, Durham stands out having a success rate on “Testing set 1” around 50% higher than in training set.



*Graphic 1 - Success rate per station*

Population changes were also detected between training and testing sets, and they will be deeper analyzed in section [“Population Analysis”](#_mkal2eah8yld).

Unfortunately, the drop of performance for the proof-of-concept is also explained due to a [deployment issue](#_thagbmcdkz0). As reported in [Report 1](https://docs.google.com/document/u/0/d/1GBnU7DYedkmGVtF9aS43TefMgo9qklh8C3xC2KVnk_4/edit), the threshold used to define if a search would be performed was 0.414. However, during this analysis of “Testing set 1”, it was discovered that the threshold was deployed as 0.5.

When the performance of the model is calculated using the correct threshold, Table 5, it is possible to see that performance is aligned with the expectation:

|  |  |  |  |
| --- | --- | --- | --- |
| **Technical Measures** | | | |
|  | **Final Model (deployed)** | **Proof of concept** | **Proof of concept with correct threshold** |
| AUROC | 0.61 | 0.56 | 0.56 |
| Global precision | 0.25 | 0.32 | 0.31 |
| Global recall | 0.85 | 0.67 | 0.87 |

*Table 5 -Technical measures of Deployed Model, Proof of concept and Proof of concept with correct threshold*

## Success on requirements

During this section it will be evaluated and discussed the [business requirements](https://docs.google.com/document/u/0/d/1GBnU7DYedkmGVtF9aS43TefMgo9qklh8C3xC2KVnk_4/edit) fulfillment, defined on Report 1, for the new population ("Testing set 1").

Regarding requirements fulfillment, Table 2, it was possible to have more than 10% of success rate for all the stations and the success rate variation between stations per Ethnicity and Gender are aligned with the deployed model and expectations. However, none of the stations fulfills the requirement of having more than 10% of success rate per search object.

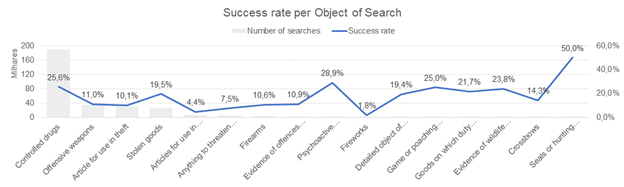
Table 6 shows the analysis of the requirements per station. In order to avoid atypical results due to low quantity of observations, the precision was calculated for the combination “station-protected” class with more than 50 observations. This way, as there are only 31 observations from Cambridgeshire, this station is not being considered in this analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Requirements** | **Protected classes** | | **Nottinghamshire** | **City-of- London** | **Durham** |
| Minimum 10% success rate for searches per station and search objective | Station | Success rate | 29.5% | 26.6% | 63.9% |
| Search objective | controlled drugs | 29.5% | 27.2% | 64.5% |
| offensive weapons | 0.0% | 0.0% | 0.0% |
| stolen goods | 0.0% | 0.0% | 0.0% |
| article for use in theft | 0.0% | 0.0% | 0.0% |
| No police station should have a discrepancy bigger than 5% for the search success rate between protected classes ethnicity and gender | Ethnicity | white | 32.4% | 30.9% | n/a |
| black | 28.2% | 25.6% | n/a |
| asian | 24.1% | 22.0% | n/a |
| Gender | female | 25.6% | 18.5% | 52.9% |
| male | 29.8% | 27.4% | 64.1% |

*Table 6 - Success on requirements (precision) per station*

The highlighted points of Table 6 are Durham with a high precision, which is justified per its high success rate observed on the “Testing set 1”, and also the fact that the model had precision equal to zero to all stations for search objectives different from “controlled drugs”.

Remembering an analysis from the [Report 1](https://docs.google.com/document/u/0/d/1GBnU7DYedkmGVtF9aS43TefMgo9qklh8C3xC2KVnk_4/edit) about success rate per search objective, Graphic 2, we can see that “offensive weapons”, “stolen goods” and “article for use in theft” have a lower success rate than “controlled drugs”.



*Graphic 2 - Success rate per Object of search on Training set*

This unexpected precision equal to zero is also explained by the deployment issue on the threshold used to define if a search would be performed. Once a higher threshold (0.5) was used instead of the correct one (0.414), these subsets of observations that have lower success rate were directly affected, as they were classified as “do not search” in situations that the response should be “search”.

When the requirements are calculated using the correct threshold, Table 7 and Table 8, a better fulfillment in the search of object requirement is observed:

|  |  |  |  |
| --- | --- | --- | --- |
| **Requirements** | **Subpopulation** | **Proof of concept** *(deployed threshold)* | **Proof of concept with correct threshold** |
| Minimum 10% success rate for searches per station and search objective | Station | all stations fulfill the requirement | all stations fulfill the requirement |
| Search objective | 0 stations fulfill the requirement | 1 stations fulfill the requirement |
| Discrepancy between stations would not be larger than 10 percentage points | Station | 42 p.p. of difference | 40 p.p. of difference |
| No police station should have a discrepancy bigger than 5% for the search success rate between protected classes ethnicity and gender | Ethnicity | 0 stations fulfill the requirement  Average difference: 8.6% | 0 stations fulfill the requirement  Average difference: 6.7% |
| Gender | 1 station fulfill the requirement  Average difference: 10.0% | 1 station fulfill the requirement  Average difference: 7.0% |

*Table 7 - Requirements fulfillment Proof of concept (deployed threshold) and Proof of concept using the correct threshold*

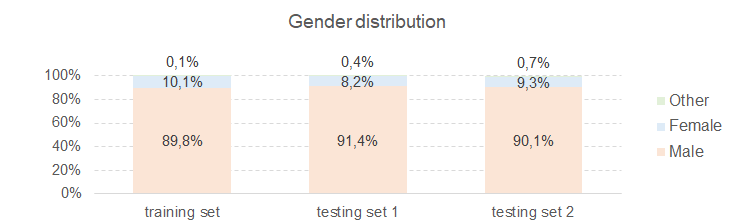
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Requirements** | **Protected classes** | | **Nottinghamshire** | **City-of-london** | **Durham** |
| Minimum 10% success rate for searches per station and search objective | Station | Success rate | 27.4% | 26.4% | 58.4% |
| Search objective | controlled drugs | 27.4% | 27.2% | 64.0% |
| offensive weapons | 23.8% | 12.9% | 55.6% |
| stolen goods | 0.0% | 33.0% | 72.6% |
| article for use in theft | 0.0% | 24.2% | 23.6% |
| No police station should have a discrepancy bigger than 5% for the search success rate between protected classes ethnicity and gender | Ethnicity | white | 29.4% | 29.0% | n/a |
| black | 25.8% | 26.0% | n/a |
| asian | 23.1% | 21.8% | n/a |
| Gender | female | 21.9% | 25.9% | 65.1% |
| male | 27.7% | 26.4% | 56.9% |

*Table 8 - Success on requirements (precision) per station using the correct threshold*

## Population Analysis

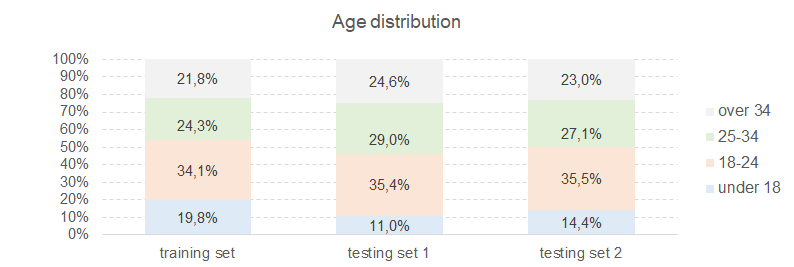
As shown on Graph 1 and Table 4, the success rate on “Testing set 1” is higher than the overall success rate observed on “Training set”, and the Durham station has an increase of around 50% on its success rate.

The gender distribution, Graphic 3, is stable between training and testing sets.



*Graphic 3 - Gender distribution*

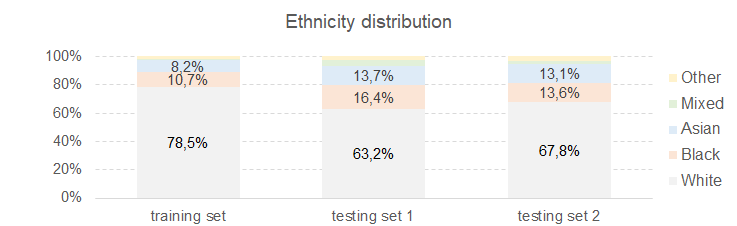
Age distribution, Graphic 4, is stable between testing sets, but “under 18” range is lower than observed in the training set.



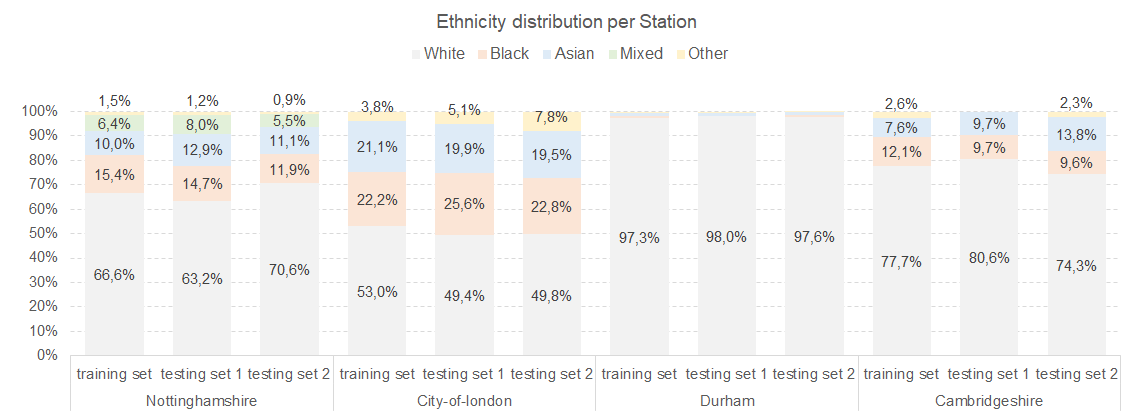
*Graphic 4 - Age distribution*

Ethnicity distribution, Graphic 5, is also stable between testing sets but different from the training set.

The stations of the proof-of-concept, especially City-of-London and Durham (Graphic 6), have an ethnicity distribution different from training set, which could impact the performance of the model and requirements fulfillment.

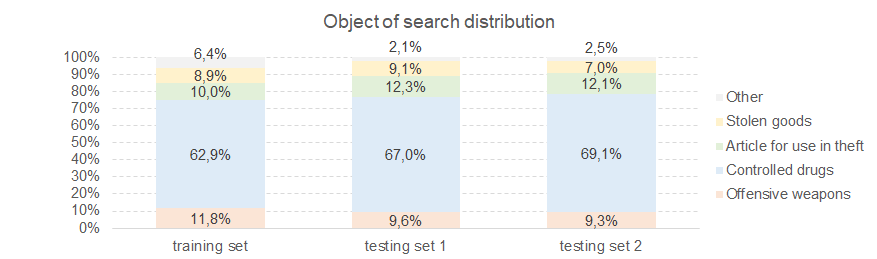


*Graphic 5 - Ethnicity distribution*

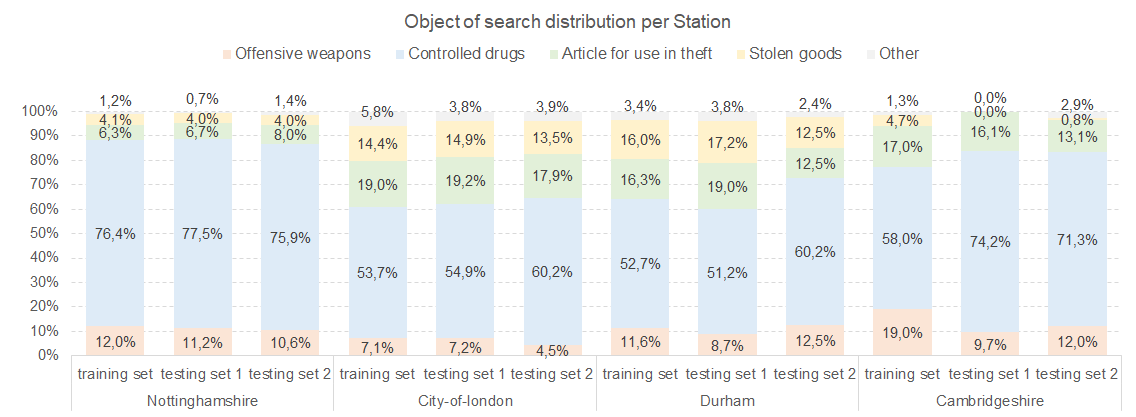


*Graphic 6 - Ethnicity distribution per station*

The object of search distribution, Graphic 7, is stable between training and testing sets, but the distribution of the object of search between stations, Graphic 8, is different. This could also affect the model performance and requirements fulfillment.



*Graphic 7 - Object of search distribution*



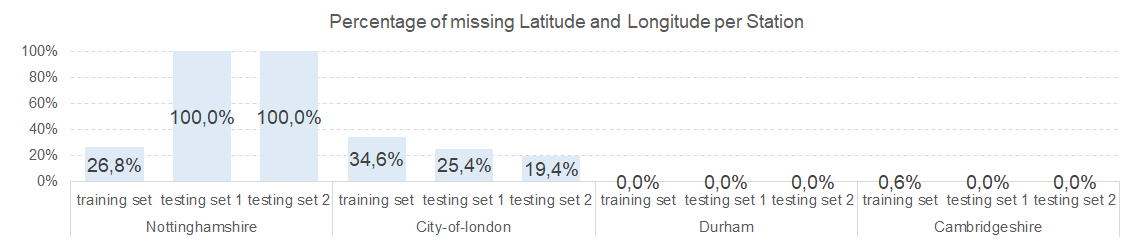
*Graphic 8 - Object of search distribution per station*

The most important population change observed was in regards to Latitude and Longitude features (Table 9), where, for some stations (Graphic 9), there are significantly more missing values than what was observed in the training set. As Longitude and Latitude were two of the five most important features of the model, it could have a significant impact in the model performance.

Specifically for Nottinghamshire, where there are no Latitude and Longitude values, this can be related to the precision problem (Table 8) of the objective of search requirement.

|  |  |  |
| --- | --- | --- |
| **Percentage of missing Latitude and Longitude** | | |
| **Training set** | **Testing set 1** | **Testing set 2** |
| 24,1% | 63,2% | 50,3% |

*Table 9 - Percentage of missing values in Latitude and Longitude features*



*Graphic 9 - Percentage of missing values in Latitude and Longitude features per station*

# Next Steps

## Next Steps

The discriminatory analysis that compares the station's situations with the general behavior can be improved by using Census information about Ethnicity and Gender distribution of each station area. It could help understanding if the stop operations are unequal compared to the population of that area.

Once the model will start to decide which searches will be performed, it could be useful to have a random percentage of searches where the response would always be “search”. With this kind of information it would be possible to analyze if those searches that the model are flagging as "do not search" continues to have a low incidence of success rate.

As the success rate can be very different between objects of search, the features that may better predict when a search will be successful for each of them can be also different. Reframing the problem in order to have one model per search object (or group of search objects) could bring improvement in the performance of the overall stop operations.

# Deployment Issues

## Redeployment

Between the first and second part of the proof-of-concept, the model was re-estimated and re-deployed due to the below reasons:

* **Deployment issue**

As reported in [Report 1](https://docs.google.com/document/u/0/d/1GBnU7DYedkmGVtF9aS43TefMgo9qklh8C3xC2KVnk_4/edit), the threshold used to define if a search would be performed was 0.414. In other words, if the calculated probability of success >= 0.414 the outcome is true (to search). However, during the analysis of “Testing set 1”, it was discovered that the threshold was deployed as 0.5, which is not correct and needed to be fixed.

* **Possibility of model improvement**

As was possible to see during the discussion in section [Population Analysis](#_mkal2eah8yld), the features Latitude and Longitude, for some stations, were sent with a higher percentage of missing values than what was observed in the training set. Once Longitude and Latitude were two of the five most important features of the model, it was decided to study if it was possible to build a new model version without these two features while keeping a similar performance.

Together with these tests, it was also applied an oversample technique to the Ethnicity feature, in order to analyze the possibility of improvement in the requirements fulfillment.

The re-estimation of the model was done using only the training set because the “Testing set 1” is small (4 thousand observations) compared to the training set size (around 300 thousand obseravitons), and also to be able to use the “Testing set 1” for an impartial evaluation of this new model performance and requirements fulfillment. Table 10 and Table 11 show the results of the new model.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Requeriments Fulfilment** | |
| **Requirements** | **Subpopulation** | **Model deployed** | **New Model** |
| Minimum 10% success rate for searches per station and search objective | Station | 37 stations fulfill the requirement | 37 stations fulfill the requirement |
| Search objective | 20 stations fulfill the requirement | 21 stations fulfill the requirement |
| Discrepancy between stations would not be larger than 10 percentage points | Station | 30 pp of difference | 32 pp of difference |
| No police station should have a discrepancy bigger than 5% for search success rate between protected classes ethnicity and gender | Ethnicity | 8 stations fulfill the requirement  Average difference: 8.9% | 21 stations fulfill the requirement  Average difference: 11.7% |
| Gender | 20 stations fulfill the requirement  Average difference: 8.0% | 21 stations fulfill the requirement  Average difference: 9.3% |
| No police station should have a discrepancy bigger than 5% for Removal of more than just outer clothing rate between protected classes (age, ethnicity, gender) | Gender | 31 stations fulfill the requirement  Average difference: 5.9% | 25 stations fulfill the requirement  Average difference: 7.4% |
| Age | 23 stations fulfill the requirement  Average difference: 8.4% | 19 stations fulfill the requirement  Average difference: 8.6% |
| Ethnicity | 18 stations fulfill the requirement  Average difference: 15.4% | 25 stations fulfill the requirement  Average difference: 14.3% |

*Table 10 - Requirements fulfillment for the deployed model and New model version*

|  |  |  |
| --- | --- | --- |
| **Technical Measures** | | |
|  | **Model deployed** | **New Model** |
| AUROC | 0.61 | 0.60 |
| Global precision | 0.25 | 0.25 |
| Global recall | 0.85 | 0.84 |
| Threshold to define True class | 0.41 | 0.43 |

*Table 11 - Technical Measures for the deployed model and New model version*

To see more details about the new model, please, consult [Annexes](#_2t6hs9vqljyg).

## Unexpected problems

As discussed in the previous sections, there was a deployment issue regarding the deployment of the wrong threshold used to classify the response as “search”/”no search”. This impacted the model performance and requirements fulfillment. The problem was addressed before the second part of proof-of-concept during the model redeployment.

During the first part of proof-of-concept, “Testing set 1”, 4,000 requests were made, and the API was able to respond to 95.8% (3.832) of them. As the API is prepared only to save valid requests, it is not possible to split the remaining 4.2% between “no response due to a bad request” and “no response due Heroku/database unavailability”.

Of the 3.832 API responses of the “Testing set 1”, the API was not able to save the true class for 168 of them. Once again, as only the valid requests are saved, it is not possible to know if the others were not saved because of bad requests or because of Heroku/database unavailability.

Regarding the second part of proof-of-concept, the API responded to 3,902 requests. At the same time, as the issue of not saving all the requests was not fixed, it was not possible to measure the response percentage and further analysis.

## What would you do differently next time

In order to have more information about the API, I would invest more time developing mechanisms to capture more information about the request and API usage, by doing the following:

* Saving all the requests, not only the valid ones. This would allow:
  + A deeper analysis and monitoring of the response rate of the API;
  + To analyze the failed requests in order to try to increase the response rate. This probably can be made by doing adjustments in the API and/or advising the client about adjustments that the users can make in the requests submission, aiming for a better experience and response rate of the API
* Saving the response time for each request, in order to be able to analyse if the API is respecting the client expectation of response in less than half a second. This also would provide the information if there were concentrations of requests in some parts of the day that can cause API to be unstable and unavailable.

# Annexes

## New model details

As discussed on subsection [Redeployment](#_lu2zwdxqoddn), a new version of the model was deployed before the second part of the proof of concept.

**Model specifications**

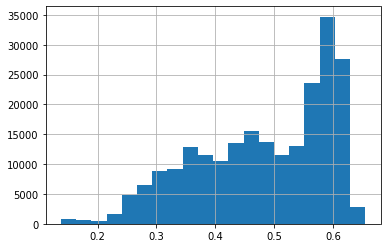
* Training set:
  + It was used only the “Training set”
  + Observations from stations Metropolitan, Humberside, Gwent and Dyfed-powys were not used because of the atypical successful rate
  + It was done an oversample of “Officer-defined ethnicity” in order to increase the representativeness of minority classes.
    - “Black” and “Asian” observations were replicated randomly until they reached the same number of “White” observations. RandomOverSampler was used from over\_sampling of imblearn version 0.7.0 with sampling\_strategy = 'minority' and remaining parameters used as default.
    - For “Other” and “Mixed”, all the observations were repeated only four times, because of its low quantity of observations a smaller oversample is preferable.
  + The data set resulted from oversample strategy was randomly split into two parts: a training set with a 70% of the data and a testing set with the remaining 30%
* Target:

A search is considered successful if the outcome is positive and is related to the search. That is, the “Outcome linked to object of search” (missing values are considered as False) must be True, and the Outcome has to be one of the following categories: “Local resolution”, “Community resolution”, “Offender given drugs possession warning”, “Khat or Cannabis warning”, “Caution (simple or conditional)”, “Offender given penalty notice”, “Arrest”, “Penalty Notice for Disorder”, “Suspected psychoactive substances seized - No further action”, “Summons / charged by post”, “Article found - Detailed outcome unavailable”, “Offender cautioned”, “Suspect arrested” or “Suspect summonsed to court”.

* Features:
  + Type
  + Part of a policing operation - missing values were filled with False
  + Age range - “under 10” and “10-17” were combined as a category “under 18”
  + Legislation - was recategorized to:
    - Missing values were assigned to a “missing information” category
    - All categories different than “misuse of drugs act 1971 (section 23)”, “police and criminal evidence act 1984 (section 1)”, “criminal justice and public order act 1994 (section 60)” or “firearms act 1968 (section 47)” were reclassified to “others”
  + Hour - hour of the search (Date feature)
  + Month - month of the search (Date feature)
  + Day\_of\_week - day of the week (Sunday, Monday, …, Saturday) of the search (Date feature)
* Numerical features:
  + Hour and Month
* Categorical features:
  + The remaining features
  + All categories from each feature were transformed in a dummy feature (one hot encoding)
  + The model is prepared to accept unknown categories
* Estimator:
  + Random forest with maximum depth of the tree = 300, and minimum number of samples required to be at a leaf node = 1% of the train dataset. The remaining parameters were used as the default definition according to scikit-learn 0.23.2
* Predicted class:
  + An observation is predicted as True when the predicted probability is equal or greater than 0.426.

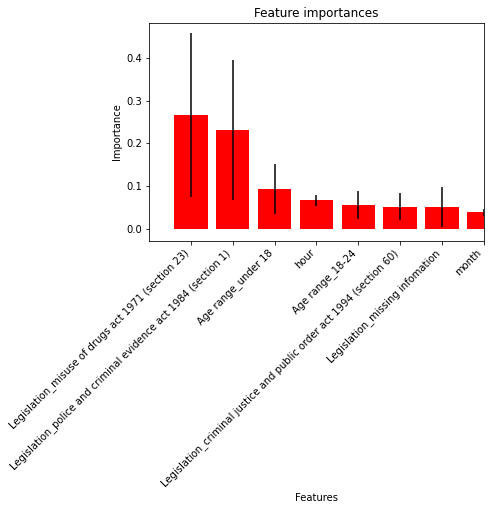
**Model results**

The new model continues to have a good probability distribution that allows the usage of a threshold of 0.426 and to avoid stations-objective search combinations with precision zero.



*Graphic 10 - Probability distribution of New Model*

The most important features for a successful search are some types of legislation, age of the person, and the hour and month of the search:



*Graphic 11 - Features importances*

**Back testing on “Testing set 1”**

As the new model only used the training set observations to be trained, it is possible to analyze how would be its performance on “Testing set 1”.

It is also possible to see that the new model has a better recall result on the “Test set 1”, and it was able to significantly diminish the success rate variation between stations per Ethnicity and Gender.

|  |  |  |
| --- | --- | --- |
| **Technical Measures** | | |
|  | **Proof of concept** *(deployed threshold)* | **Proof of concept using the New Model** |
| AUROC | 0.56 | 0.54 |
| Global precision | 0.32 | 0.30 |
| Global recall | 0.67 | 0.80 |

*Table 12 - Comparing Technical Measures*

|  |  |  |  |
| --- | --- | --- | --- |
| **Requirements** | **Subpopulation** | **Proof of concept**  *(deployed threshold)* | **Proof of concept using the New Model** |
| Minimum 10% success rate for searches per station and search objective | Station | all stations fulfill the requirement | all stations fulfill the requirement |
| Search objective | 0 stations fulfill the requirement | 1 stations fulfill the requirement |
| Discrepancy between stations would not be larger than 10 percentage points | Station | 42 p.p. of difference | 41 p.p. of difference |
| No police station should have a discrepancy bigger than 5% between the search success rate between protected classes ethnicity and gender | Ethnicity | 0 stations fulfill the requirement  Average difference: 8.6% | 0 stations fulfill the requirement  Average difference: 6.7% |
| Gender | 1 station fulfill the requirement  Average difference: 10.0% | 1 station fulfill the requirement  Average difference: 6.1% |

*Table 13 - Comparing Requirements fulfillment*