Business conclusion

Summary

The business problem stated that one of the objectives was to reduce overstops. The minimum success probability stated to accomplish this was 10% which allowed very low success rates technically called precisions. Considering the precision recall tradeoff and the implicit objective of maximizing recall, or in other words, minimizing the people not stopped but indeed prevaricating, this 10% probability of success makes sense. It reduces the probability of criminals not being caught. The problems arises when, following this logic, we start having discrimination across sensible groups. If we do not keep control on the criteria to stop and search subjects due to our low minimum precision accepted we end up hostages of the sampling. Everyone who passes a police officer is stopped and searched and the precisions across genders ethnicities and age ranges (tendentially low) will differ a lot from one group to the other.
We therefore conclude that the 10% criteria is not consistent with both reduce overstop and avoid discrimination objectives. This is why the threshold of 0.4490314298609562, which implied an overall precision of 82% was used instead.
When in production, the model failed to achieve its goals. Due to the fact that the population was relatively small, the test data unbalanced across stations and our precision too ambitious,

the model did not identify any subject as suspect and did not stop/search anyone. We expected a low recall and low bias across sensible groups and because we had no positives (neither true nor false) our recall was zero and the differences between precisions across genders and ethnicities were also zero. On the other hand, we expected a high precision and as a consequence of what was stated above, it was also zero. The maximum probability calculated for a successful operation in the test set was 0.39 which was not enough to trigger our model. After this test the model training data was balanced across stations and sensible groups to better improve the overall predictive capacity of it independently on the training set sampling. It was also considered, after seeing the model in action over such a small population that the chosen threshold resulted in an unacceptably low recall. Therefore it was chosen to change the compromise previously selected in order to improve it. This will lower our precision to an expected minimum of 50% and increase the bias across genders to around 15%. The recall is expected to increase to 25%.

Results Analysis

Model Performance

Considering the information from the previous report, a high precision value (82%) was expected. This would intentionaly increase our selectivity and reduce the number of stopped people as a consequence, solving the problem of overstop and helpping us to eliminate discrimination. Precision, the technical term for success rate, is calculated by dividing the number of successful operations (True Positives) by the total number of operations, successful or not (True Positives and False Positives).
The precision objective for the model was 82%, translated on a probability of success of 0.4 and it was too ambitious. Once set in production in a small sample of observations it showed itself too selective and ended up not appointing anyone as a suspect. The real precision was therefore 0% as we have no positive cases, neither true nor false.

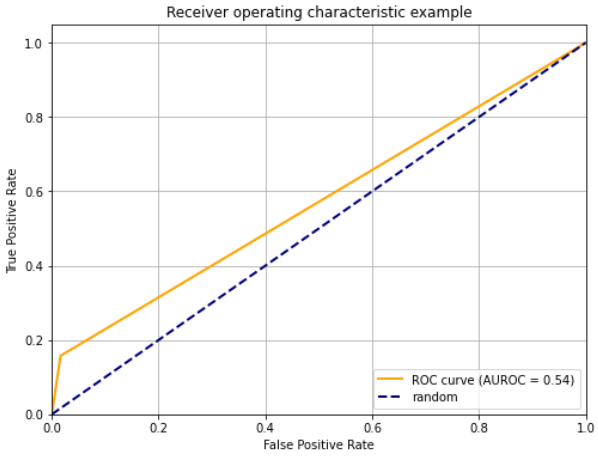

Fig 1. Model Roc_Auc curve Expected

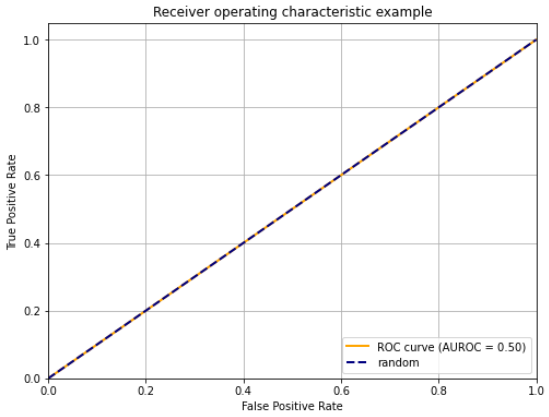

Fig 2. Model Roc_Auc curve Obtained

Discrimintation, or bias across sensible groups, was also described technically in the previous report as the difference across precisions for each individual group. The objective for this important metric was to keep it below 10%. In practice, as we did not have positive cases, all precisions calculated for each group were 0 and so there was no discrimination. As the sample size increased and because our model was being very cautious, the precisions would all be very close to 1 and very close to each other so discrimination was not expected.
Considering the recall, it was expected to be low. This was chosen to be the metric to sacrifice in order to improve the previously mentioned objectives, considered more important. Recall is calculated by dividing the number of True Positives by the sum of True Positive with False negatives. False Negatives, the only new concept here are the cases where the model decided not to stop/search and which were indeed prevaricating. Relating them with our situation, as our model was too precise, it would only decide to proceed when it was very sure of a successful result, the False Negatives number was expected to be high, bringing our recall to low levels. Again, because we had no True Positives, our recall was 0. From the 1.113 prevaricators in our test set we stopped/searched none so our False Negatives number was indeed high as expected.

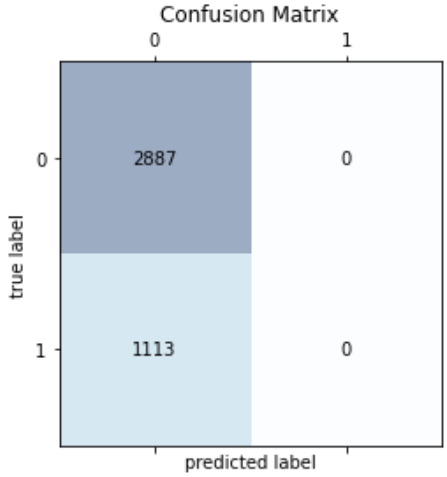

Fig 3. Confusion Matrix of the model set in production on the test set

Considering the data used in test one and the obtained results there is nothing indicating that the model did not perform as expected. It could have happened that the model would have predicted a higher probability for some observations, identifying some of the 1.113 guilty subjects and having therefore some True positive cases, making the results more interpretable and the chosen strategy more perceptible but considering the type of data in the test sample the model was not able to achieve this. As we will see on the Population Analysis chapter in more detail the data was not favourable to our model which was not trained for this particular sampling and that was probably the reason why it could not identify any of the prevaricators.
Regarding the app functioning, it was programmed to perform a certain number of verifications, not predicting the same observation id twice in order to preserve the quality of the database which stored the results, and it would not accept observations with missing columns, despite the fact that the model did not use all the available data for predicting the outcome.
The model needed only ‘Type’, ‘Date', 'Gender', 'Age range', 'Officer-defined ethnicity', 'Object of search' and 'station' but it did not accept observations which did not have all the columns which were part of the training set. The observation was expected to follow the pattern below:

```
{
        "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>
}
```

Despite this fact, the app would accept null values for the columns of the observations as requested. It was not acceptable to introduce data with extra columns as well.

Success on Requirements

The results of the model set in production during the first week are a total miss and in addition, they are not easy to analyse. Because of the low dimension of the test set and the fact that the sampling provided was not favourable, considering the sampling of the data set used to train our model, we ended up not having Positive cases identified and therefore the results are biased.
As can be seen on the following calibration plot, the model has good probability prediction power for low probabilities of success and starts predicting worse as probabilities increase.

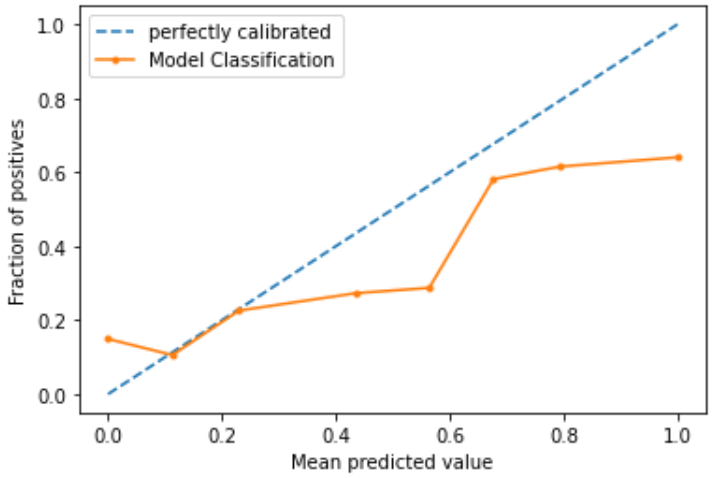

Fig 4. Reliability plot for the model in production during test 1

The decision to sacrifice Recall in order to improve success rate, stop discrimination and solve the problem of overstop ended up being a bad test strategy in that sense. Still, we ended up with a low recall, and no discrimination, as expected and a low precision which was not the objective at all.

Considering the requirements, the low probability of success needed as criteria to consider subjects as suspects implied a low success rate which was what our model attained. As far as the objective to eliminate discrimination is concerned, it was also attained as the model, because it did not suspect anyone, did not discriminate. Precision values across gender and ethnicity were zero for every station and as a global figure. This means there is no difference between success rates across sensible groups and therefore, no discrimination..

Another aspect asked to be held in consideration was the success rate across subgroups defined by combining the station, ethnicity, gender characteristics. Similarly to the success rates across sensible groups, the proposed objective was to keep success rate differences for these under 10%. As we had no positive cases, these were also zero and therefore no discrimination was made.

The high recall objective was implicit as it is the main objective of police operations to detect crime and also because if we aim at low values of precision we automatically imply high recall. Despite the fact that it was the chosen strategy to sacrifice this objective from the start in order to improve precision and eliminate bias across sensible groups, the value of zero for it due to our lack of positive cases was also not expected and explained above.

After observing the model in production, such a low recall value was perceived as unacceptable and risky and this was the main aspect to be corrected on the second model set in production.

To solve the above mentioned problems the model was retrained with a resampled training set across stations and sensible groups. This will tackle the risk of underperformance due to an unbalanced production sample.

To improve recall and minimize the risk of increased criminality a lower precision objective was also established. By adjusting the probability of successful operation needed to consider a subject a suspect from the initial 0.44 to 0.33, we expect a precision just above 50% and a recall

of around 25%. These new objectives are expected to maintain bias across sensible groups at controlled low levels of up to 15% difference.

	First Model		Second Model
	Expected	Obtained	Expected
Precision	0,82	0	0,51
Recall	0,12	0	0,25
Max precision difference across Gender	0,1	0	0,15
Max precision difference across Ethncity	0,1	0	0,15

Table 1. Requirements metrics

Population Analysis

Population analysis is important for the performance of the model. We want to make sure big enough samples of every kind of data are supplied to the model while on the training phase so it can balance them and find the right relationships between them and the target variable.
The training set population was as follows:

- **Considering Gender:** 10% Female, 90% Male
- **Considering Ethnicity:** 79% white, 10.5% Black, 8% Asian, 2% others and 0.5% Mixed race subjects
- **Considering Age Range:** 0.1% under 10, 20% 10 to 17, 34% 18 to 24, 24% 25 to 34 and 21.9% over 40 years old

Regarding the above mentioned distributions, the train set did not differ much from the test set. The test set distribution was:

- **Considering Gender:** 8% Female, 92% Male, 0.35% Other
- **Considering Ethnicity:** 62% white, 16.5 % black, 14% Asian, 4.5% Other, 2.4% Mixed
- **Considering Age Range:** 0.2% under 10, 11% 10 to 17, 35% 18 to 24, 29% 25 to 34 and 25% over 40 years old

The 14 observations from the ‘Other’ Considering the target, both populations were also similar:

- **Train set:** 32% successful operations
- **Test set:** 28% successful operations

And considering type:

- **Train set:** 24% Person and Vehicle search, 76% Person Search
- **Test set:** 25% Person and Vehicle search, 75% Person Search

On the other hand when analysing stations the case changes. The test set had only 4 stations a very different distant number from the one we had on the train set, 41:

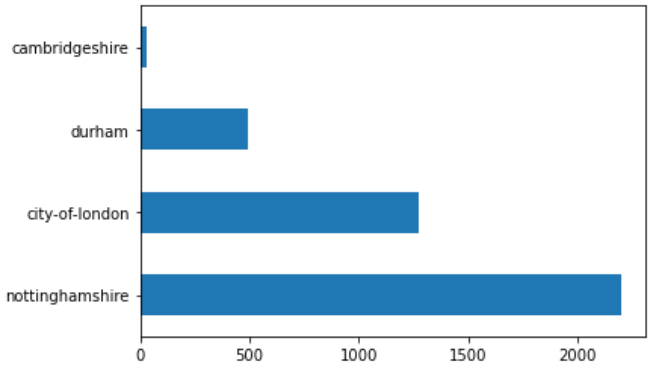

Fig 5. Station distribution from the test set

The percentages of observations per station on both train and test set can be compared on the following table:

Station	Train set		Test Set	
	Value	Percentage	Value	Percentage
Nottinghamshire	7 099	2%	2 203	55%
City-of-london	3 572	1%	1 272	32%
Durham	2 771	1%	494	12%
Cambridgeshire	876	0%	31	1%
SubTotal	**14 318**	5%	**4 000**	100%
Total	**309 044**	100%	**4 000**	100%

Table 2: Distribution of both train and test set across stations

This means that our model was mainly inquired on the data it knows less about as it did not have sampling enough on the training set. As it can be seen on the feature importance plot, stations represented nine out of the the top ten most important features meaning the model associates this feature with the outcome and low represented stations clearly have this relationship understudied. This helped the model to underperform in detecting the True Positives.

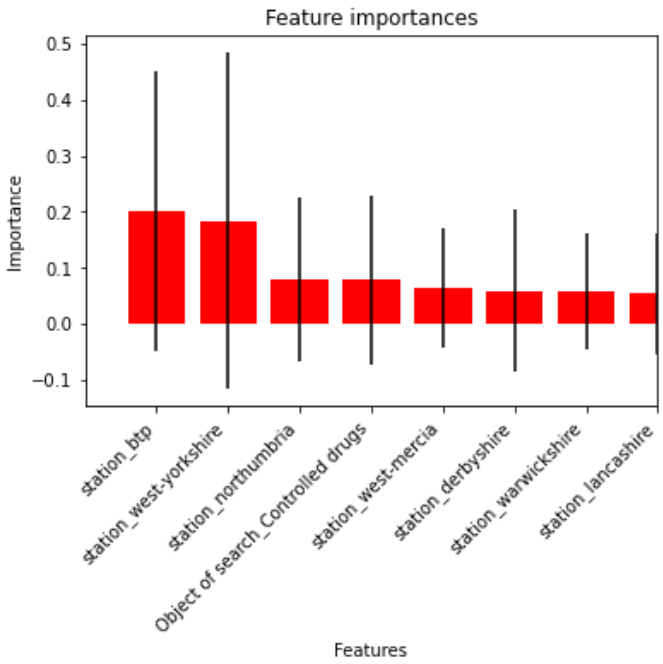

Fig 6. Feature importance of the model set in production

The same happened with the 'Other' Gender of the test data. As mentioned on the previous report, because we did not have enough samples on this gender we excluded it from the training data and as it is present on the test data, the model does not have the correlation between this variable and the target studied to better predict the result. From these 14 cases, 11 ended up counting as False Negatives as a consequence.
To solve this problem the model was retrained with resampled data across stations. This way we improve the models knowledge across stations and we promote an universal criteria for stopping/searching. This helps the bias across sensibles groups across stations to be reduced. Similarly, resampling over sensible groups was also performed to improve our predicting power across genders and prevent possible future problems.

Next Steps

Next Steps

The monitorization of the behaviour of our model on the short term should always be on our to-do list not only on a global level but also on specific subsamples of data. It is important to be able to perform well across every dimension of the problem and if some tuning is necessary in order to better adjust the end results to the field and business needs we are totally available to do so.
Very important for the overall results of police department operations regarding discrimination is to be able to accurately identify the ethnicity of the suspect. We introduced some measures to eliminate discrimination across officer-defined ethnicities but it was observed front the train set

that in each of the classes present in this feature, all different types of self defined ethnicities, the real ones, are mixed. This means that even if we succeed in eliminating discrimination in our model it would allways be dependent on the officer perception of the reality and not the reality itself and the police force can still be accused of discrimination if values across self defined ethnicities do not match with the officers perception. A better collection of data regarding the real ethnicity of each person will be necessary in order to analyse this reality and a better perception from each police station regarding ethnicity identification will help the model to better perform on the subject.
Still on the discrimination matter, it would be interesting to find alternative ways to achieve fairness. The elimination of data correlated with protected classes, which may help the model group subjects as belonging to sensible groups, would be an interesting approach. Intensive data study would be needed.
The criminality key performance indicators in each area are monitored and should be used to evaluate the influence of the use of the model on a macro level. This would be interesting complementary data to have in order to evaluate the influence of our model in the long run and to improve its tuning in the future. Reality is constantly changing and it is important to be able to react to these changes, putting our techniques to better use.

Deployment Issues

Redeployment

During the training of the initial model resampling was not considered beneficial. As stated in the previous report, it was tried but it did not improve our models results and so it ended up being discarded. In addition, the data used to train the model was only 75% of the available data. The remaining 25%, used for testing our model results during development, were forgotten. These facts allied with the strategy of going for a low recall led to very bad results and something had to be done to correct this situation and improve our models performance.
As a consequence, the model was redeployed as it was re-trained with a resampled, complete train set, regarding stations and sensible groups. When the initial model was set in production for test one, it became obvious that we were underperforming in some stations and this would imply a very low recall for these stations in particular and this would lower the overall recall score. As the strategy expected a low recall from the start, sacrificing this objective to improve the precision and lower the bias across sensible groups, the retraining of the model became even more important, the recall value for this test set was completely unacceptable.
In addition, when set in production we understood the practical effect of such a low recall on the real scenario and the threshold selected was also lowered, diminishing the penalization the model would bring to this metric. As stated above, the initial probability used was set in order to improve precision at a high value, 82%. By lowering this threshold we now expect a precision score, also known as success rate, of around 51% and an increase in recall to around 25%. As

a consequence, the difference between success rates across sensible groups is also expected to increase but as we have a tradeoff between all business metrics set as objective, this new strategy is considered a safer approach.

Unexpected Problems

Considering data, and as stated with more detail in the population analysis, it was not expected that train data was so unrepresentative of the test set leading to such an underperformance. Despite the difficulty in analysing results originated by this fact, as we ended up not having any positives neither false nor true and all metrics were zero, it ended up being very useful as a strategy validator, making perfectly clear that such a low recall had a very high probability of not working in the field. This allowed us to readjust the tradeoff between the expected results and provide a better second model, which will certainly adapt better to our clients needs.
It was not expected such an impact of a low recall on such small subsets of data which really compromised the test of the model as a whole. As a consequence the four police departments in charge of testing this model were in fact wasting time and resources as the model ended up not indicating anyone as a suspect.
As far as deployment is concerned everything ran smoothly and we got no unexpected surprises. The test sample was under the database limit of 10.000 observations and we were able to collect and analyse the results, despite the additional work involved in processing the json format saved on the database as a string instead of saving it as individual fields which would have been easier to use in the data frame.
Due to the lack of real outcomes in test number two, we are not able to evaluate the final results and the verification of the expected behavior is not possible this phase, as it happened on the first test. Again, we are totally available to answer any additional questions and make any extra tuning in order to guarantee our service is up to our clients expectations.

What would you do different next time

As stated above, the initial model was only trained with 75% of the available data and this was not intended. The first thing I would do differently would be to eliminate this mistake.
Making sure to have a similar number of occurrences across all variables in the dataset during the training of the model to be set to production is a very important issue and should not to be disregarded ever again. It helps save a lot of time and improve productivity.
The alignment of the whole team regarding test sampling and size would most certainly be a good improvement if there was another chance. Trying to gather information about where or when the tests will be conducted and by whom for example or any other additional insight on how our model is intended to be tested could be used to save time and effort and it is perfectly plausible that the client has some insight on the subject beforehand.
Better aligning expectations regarding minimum and maximum values with the client, who better understands the business and has, therefore, a better perception of what is acceptable and what is not to be set in production would also be an important contribution.