Index

1. Business Conclusions

1.1. Summary

The model had a good performance even though the results were substantially different than what expected. It's global success rate and detection of the successful operations was better than expected. However, this came at the cost of giving green light to more 23% search operations than intended.

Regarding the requirements, they were partially met as expected.
In summary:

- discrimination between Genders and Ethnicities: the performance was better than expect, but there was still stations above the 5% limit.
- minimum of 10% success rate per station per objective: Success.
- maximum variation of 10% between station's success rate: failed for 1 of the 4 stations.

The reason for the discrepancy is thought to be in part due to model performance, and in part due to bias in the new data.

In relation to the API, it worked as expected and described in the report 1.

2. Results Analysis

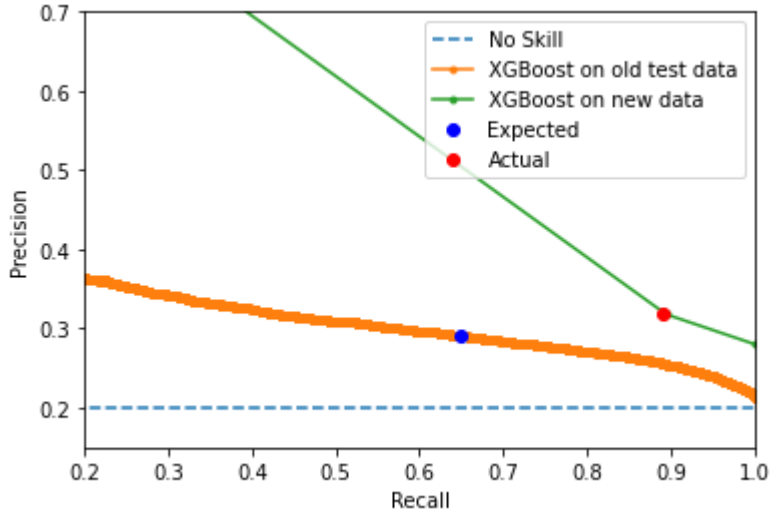
2.1. Model Performance

The model performance was assessed by comparing the ROC AUC score, PR AUC score and confusion matrix values between the model predictions on the previous test data (expectation) and the model predictions on the new data(actual).

The ROC AUC score shows how well our model is at distinguishing between positive and negative classes by calculating the probability that a randomly chosen positive class is ranked higher than a randomly chosen negative class.
In this metric there was a drop from 0.68 to 0.65, which is still within the 0.65-0.67 range values that resulted from cross validation.

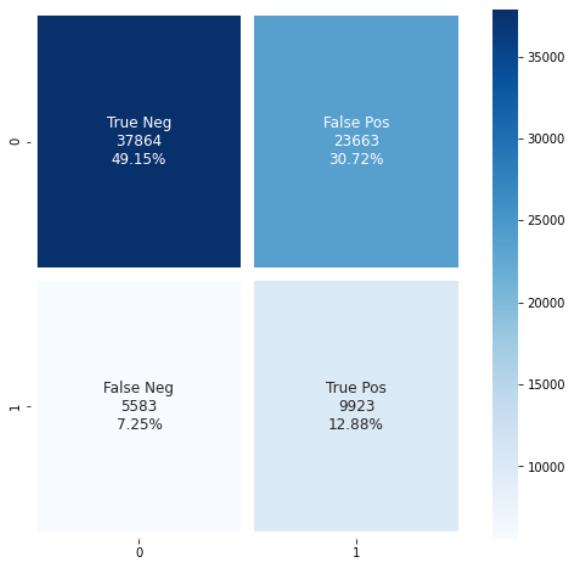
A more relevant metric in our case is PR AUC score (or Average Precision score), because our focus is more on making sure that every positive prediction is correct (precision) and that we get as many of the positives predicted as positives as possible (recall), and not so much in how it performs on the negative class or having an equal performance on both classes.
The PR AUC score had little variation, with 0.319 on the old test data and 0.314 on the new one. But looking at the Precision-Recall Curve, it seems that we had a better performance.

	Expected	Actual
Precision	29%	32%
Recall	65%	89%

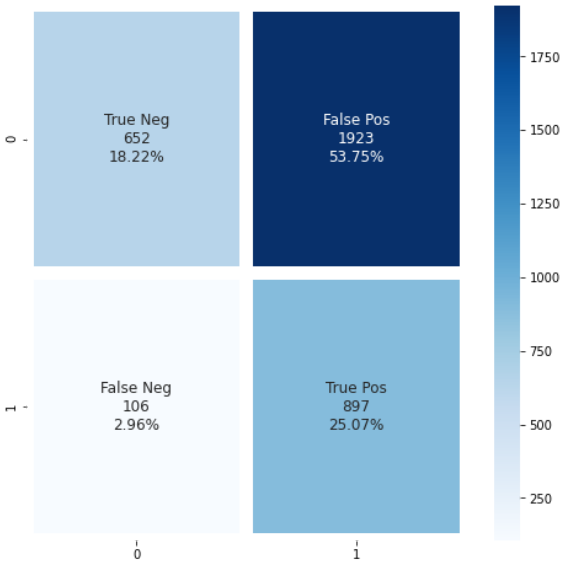


Comparing the confusion matrix, the model on the new data seems to be biased towards positive cases, which resulted in 23% more search operations than what was expected and intended.

Model on previous test data



Model on the new test data



The discrepancy could be in part due to bad model performance when generalizing to the new data, but the similar ROC/PR AUC scores on both data suggests that it might mainly come from the characteristics of new data. As we will see on the Population Analysis there isn't new unseen data but there is an overall higher success rate mainly coming from the Durham station. This bias on the new data might in part explain the model's performance variation. See Population Analysis for more details.

2.2. Success on requirements

There were 4 requirements considered in the project.

The first one and main objective was a 10% minimum success rate per station per objective, and it was successful on the new data as we can see on the following table:

Station	Object of search	Success rate
Nottinghamshire	Controlled drugs	28%
	Offensive weapons	20%
Durham	Controlled drugs	63%
	Offensive weapons	54%
	Article for use in theft	27%
	Stolen goods	65%
	Game or poaching equipment	25%
	Articles for use in criminal damage	50%
City of London	Controlled drugs	27%
	Offensive weapons	16%
	Article for use in theft	16%
	Stolen goods	25%
Cambridgeshire	Controlled drugs	25%

Secondly, regarding the 10% maximum difference in success rate between stations, it was successful in 3 of the 4 stations.

Station	Success rate
Nottinghamshire	27.9%
Durham	55.4%
City of London	26.5%
Cambridgeshire	25.0%

The Durham station was expected to have a higher success rate around 37%, but it vastly surpassed it. See Population Analysis for more information on a possible justification for this discrepancy.

The third one was related to the degree of discrimination between Genders and Ethnicities. As expected there were stations above 5% but the degree of discrimination was significantly less.

Gender		
Station	Expected	Reality
Nottinghamshire	1.24%	6.6%
City of London	14.23%	2.7%
Durham	17.29%	4.7%

Ethnicity		
Station	Expected	Reality
Nottinghamshire	19.23%	7.5%
City of London	9.39%	9.9%

The fourth and final requirement was to reduce the number of search operations at the cost of missing some in order to exclude the “obvious” search operation that shouldn’t be done. The original data had a global success rate of 21% on the search operations, which means that there were 79% False Positives. We were expecting a reduction to 30.72% based on the model performance on the old test data but it was 53.75% on the new data. See the Population Analysis for more information about a possible cause for this discrepancy.

2.3. Population Analysis

There are 3578 new observations with known outcomes divided between the 4 stations:

Station	Number of observations
Nottinghamshire	1959
City of London	1155
Durham	438
Cambridgeshire	26

Compared with the old data, the new data had a global success rate of 28% instead of 20%. Looking at the success rate on each station we can see that the new data on Durham station is especially biased towards successful cases.

Station	Old data	New data
Nottinghamshire	25.5%	24.2%
City of London	27.8%	24.2%
Durham	35.7%	55.4%
Cambridgeshire	18.0%	15.3%

This might in part explain the unintended 23% increase in False Positives referred to on Model Performance and the requirement discrepancy on Success on requirements.

In relation to discrimination between Ethnicities, assuming a minimum of 30 observation for each subpopulation, the following table shows a comparison between the old and new data:

Ethnicity	Old data				New data			
Station	Ethnicity	Success rate	Diff%	Number of observations	Ethnicity	Success rate	Diff%	Number of observations
Nottinghamshire	Asian	29.37	**10.37**	698	White	26.09	**6.84**	1238
	Mixed	19.00		458	Mixed	19.25		161
City of London	White	30.46	**5.05**	1894	Black	26.42	**8.12**	299
	Black	25.41		791	Asian	18.30		224

There was a reduction in Nottinghamshire station from 10.37% to 6.84%, and an increase in the City of London station from 5.05% to 8.12%.

In the case of Gender we have the following:

Gender	old data				new data			
Station	Gender	Success rate	Diff%	Number of observations	Gender	Success rate	Diff%	Number of observations
Nottinghamshire	Male	25.98	**6.92**	6609	Male	24.71	**5.53**	1813
	Female	19.06		467	Female	19.18		146
City of London	Male	28.65	**8.24**	3218	Male	24.26	**0.22**	1051
	Female	20.41		343	Female	24.04		104
Durham	Male	36.39	**6.66**	2468	Male	54.40	**4.78**	375
	Female	29.73		259	Female	59.18		49

The biggest variation happened in the City of London station, with a reduction in the success rate between Genders of 8.02%.
It's also interesting to note the variation between Genders on Durham station. Whereas in the old data it was 6.66% higher for Male Gender, in the new data there was a reversal to 4.78% higher for the Female Gender.

3. Next Steps

3.1. Next Steps

It would be helpful to better clean the data so that the model doesn't learn from bad examples. In the analysis of report 1, various problems of data input and outliers were identified. Investigating further bad data and them removing them from the training data will reduce noise and wrong influences.

We could also improve our feature engineering and selection by:

- more sophisticated methods in filling the missing values
- adjust the clusters on each station
- improve the use of the Date features, test encoding them cyclical features and allowing the model to better capture possible seasonal patterns
- create a new feature to give higher importance to more recent data

After that it would help to do further disambiguations on the requirements and reframe the problem, defining our goals as perfectly as possible.
One of the objectives of the first report was to give a clear view of the trade-off between the requirements so that the client could decide what direction to go, but this wasn't clearly done. The next step could then be do to a better job at defining the trade-offs between :

- minimize the search operations performed in order to eliminate obviously wrong ones
- minimum difference of 5% on success rate between Gender/Ethnicities
- minimum 10% search success per station per objective
- maximum 10% variation between stations success

Assuming that it's not possible to meet all these criteria, the falling of meeting some should be worse than the others(cost more). Taking into account all metrics, there is an overall cost associated with all the failed ones. Our objective should be to lower this predicted cost.
If we could clearly define the costs, we could take them into account directly during the classifier training by inducing bias in the data using the appropriate resampling.
We could then use regularization to restrict the model variations and grid search to find the best parameters. This should improve the model in the intended direction and theoretically find the best model for what was accorded and assumed.
The data is what it is and the classifier can be near-optimally tuned. If this is done correctly and the results still aren't satisfactory, what is left is to better clarify the goals and adjust the metrics so that the model can be perfectly tuned in that direction.

4. Deployment Issues

4.1. Redeployment

The redeployment was meant to improve and fix some issues of the first model.
The new one is able to convert Latitude/Longitude into clusters and use the feature when making predictions.

We tried to understand if the reason for the discrepancy between the Expected and Actual model performance was because of the model's weak power of generalization or because the data was inherently biased or having lots of new unseen values.
The conclusion was that the model could be a bit improved if it was calibrated. This means that the model's predicted percentage for each observation should correspond to the actual frequency of those observations.

The new model was trained without the focus on 29% global precision in mind than was previously assumed, which gave more freedom to fine-tune the classification threshold in the direction of the various relevant metrics. It was also thought that too much importance was given to reducing the number of search operations when that was never clearly stated as a goal. Lowering this down at the cost of a bit of global precision could greatly improve the other requirements, so the new model was tuned more in the direction of the new data.

In the end due to misunderstanding of the requirements (minimum 10% success rate per station per objective), the classification threshold was set too low which resulted in a much worse performance for the main requirement.

4.2. Unexpected problems

An unexpected problem was discovered right on the first received API request when the app responded with an error.
It was a simple problem due to a print that tried to access a key of the received JSON that didn't exist in the new observations (it existed on testing requests).
It got a bit more complicated because the model was inside a try/except and the error message wasn't very illuminating, I should have returned the error in the except. This allowed me to wrongly assume that the problem was in the prediction line, maybe there were new columns in the observations, or they had a different type than what the model was expecting.
I could reproduce the error locally and tried to debug it. First checking if the JSON was being correctly converted into a dataframe, then if the prediction line was able to process the dataframe observation. Then I found out the problem was in the lost print line after the model prediction.
This led to 48 minutes of downtime and around 70 observations lost.

Besides that everything seemed to work well, the server log was clean with only OK responses.

4.3. What would you do differently next time

Next time I would start by clearly understanding the business problem/requirements/goals. In our case we were interested in optimizing for:

A. maximum model's performance in finding successful search operations
B. a minimum success rate in each station/objective tuple
C. minimize the variation of success rate between each station
D. minimize the variation of success rate between each gender and each ethnicity

After that, I would try to understand how the business problem/requirements/goals can be translated into the best metrics to quantify the model's success on them.
If it's a straight forward optimization problem like in a Kaggle challenge, then the direction can be simply to maximize accuracy. In cases where the goal is more nuanced because there are requirements on different dimensions, the solution might need to be a conscious balance in the trade-offs on each metric, and the objective to present those trade-offs instead of choosing a specific solution.

In our case the model's success could be evaluated on the following metrics:

A. maximum PR AUC score for the overall model performance and F1 score to find a good range of possible threshold (general direction on classification problems)
B. minimum 10% precision in each station/objective (minimum model performance for this specific business problem)
C. maximum 10% precision difference between Stations
D. maximum 5% precision difference between Genders and Ethnicities

The metrics B and C are influenced by outliers which can lead to the precipitated invalidation of models. To deal with this, I would try to exclude them when evaluating the model, or evaluate the business problem in a more general way by creating a more flexible metric using a standard deviation instead of a specific cut-off number.

Having already a good understanding of the overall project direction, I would start the work of pre-processing data, feature selection/engineering and create a baseline. I would also create a file with all the metrics defined, containing the functions used to plot and evaluate them. This will organize the code and ensure that we will apply the same evaluation on all future models that might be on different files. It's a good idea to create a new file for each model iteration and name it appropriately instead of stacking everything in a big long file which can lead to a lot of confusion and up/down scrolling.

For the next model iterations, we should have an idea of what can be modified to optimize our model. In our case, for each metric the main focus can be on:

A. model and feature selection
B. probability threshold for classification
C. and D. data resampling, regularization, feature selection

All these different parameters work together and will influence the model's success on all metrics. Next time I would focus on understanding their importance on each metric and on finding a systematized way to test the various combinations. This will be important to evaluate and understand the trade-offs in a clear, efficient and organized way.