Notes:

Definitions:

Historical dataset: complete dataset collected by the police department and provided

Training set: Subset of the historical dataset containing observations for which feature *VehicleSearchedIndicator* equals ‘True’.

Test set: Subset of the training set that was used to assess model performance.

Production set: complete dataset resulting collected by the web app during the week the deployed model was in production.

Production test set:

# 

# Business Conclusions

## Summary

# Results Analysis

## Model Performance

“Based upon the performance that you expected and reported in report #1, how did you do on the observations for which you know the outcome.”

Confusion matrix for both methods?

In this section, we compare the model performance on the test set with its actual performance on production (naturally, only for those observations for which we know the true class - we have named this subset “production test set”). The results are presented in the table below and officers’ performance was also included in the last column as a baseline (i.e. what would have been the performance in case the model had not been deployed and all searches were performed). The recall is the most significant metric to measure the model performance, as defined in the third requirement of the briefing. The precision score is a byproduct of setting the classification threshold at 50%, according to our interpretation of the first requirement. We also present the percentage of searches performed, as a measure of the cost-effectiveness of the model, and accuracy, since it is a standard metric of common usage.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model performance | | Officers’ performance |
|  | Test set (40% of the training set) | Production test set | Production test set |
| Recall | 0.5144 | 0.4942 | 1.0 |
| Precision | 0.5998 | 0.6005 | 0.359 |
| Searches [% of total] | 29% | 30% | 100% |
| Accuracy | 0.7231 | 0.7004 | 0.359 |

As we have anticipated in report 1, performance in production remained very close to its figures on the test set. Two main reasons led us to believe the model performance wouldn’t significantly change in production:

1. Since the search selection process was going to be maintained: a first selection of subjects by the police officers, and only then a search decision provided by the model; we didn’t expect the population characteristics to change much.
2. The test set on which the model performance was assessed before deployment had a considerable size (40% of the training set, amounting to ). Accurate values

Given the size of the test set (40% of the training set, )

As we argued in report 1, since the first requirement of the briefing led us to define a specific classification threshold, the model dynamic assessment becomes less relevant. Nevertheless, we compare the ROC curve and True/False positives curves for our model on both the test set and production test set.

|  |  |
| --- | --- |
|  |  |

## Fairness

Comparison of the discrimination level between our model and the current process for the feature *SubjectRaceCode* on training and production sets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision score in the training set for feature *SubjectRaceCode* | | Precision score in the production set for feature *SubjectRaceCode* | |
|  | Current method | Our model | Current method | Our model |
| ‘White’ [%] | 34.68 | 61.99 | 36.75 | 60.48 |
| ‘Black’ [%] | 29.97 | 54.82 | 34.51 | 59.29 |
| ‘Asian/Pacific’ [%] | 29.11 | 54.76 | 26.09 | 57.14 |
| ‘Indian American’ [%] | 29.54 | 65.71 | 20.00 | 50.00 |
| Max. difference [p.p.] | 5.58 | 10.95 | 16.75 | 10.48 |
| Max. difference [%] | 16.05 | 16.67 | 45.58 | 17.32 |
| Std. Deviation [p.p.] | 2.59 | 5.45 | 7.73 | 4.69 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision score in the training set for feature *SubjectEthnicityCode* | | Precision score in the production set for feature *SubjectEthnicityCode* | |
|  | Current method | Our model | Current method | Our model |
| ‘Not applicable’ [%] | 34.99 | 60.64 | 37.98 | 61.38 |
| ‘Hispanic’ [%] | 27.71 | 57.64 | 29.84 | 55.75 |
| ‘Middle Eastern’ [%] | 27.97 | 50.00 | n.a. | n.a. |
| Max. difference [p.p.] | 7.28 | 10.64 | 8.14 | 5.63 |
| Max. difference [%] | 20.82 | 17.55 | 21.43 | 9.18 |
| Std. Deviation [p.p.] | 4.13 | 5.48 | 5.76 | 3.98 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision score in the training set for feature *SubjectSexCode* | | Precision score in the production set for feature *SubjectSexCode* | |
|  | Current method | Our model | Current method | Our model |
| ‘Male’ [%] | 33.54 | 59.73 | 36.97 | 60.55 |
| ‘Female’ [%] | 31.96 | 61.13 | 31.87 | 58.20 |
| Max. difference [p.p.] | 1.57 | 1.40 | 5.10 | 2.35 |
| Max. difference [%] | 4.70 | 2.29 | 13.80 | 3.88 |
| Std. Deviation [p.p.] | 1.11 | 0.99 | 3.61 | 1.66 |

## Population Analysis

Population on extended sense: not only individual characteristics for fairness assessment, but also the distribution of the features used by the model.

|  |  |  |
| --- | --- | --- |
|  |  |  |

# Next Steps

# Deployment Issues

## Re-deployment

## Unexpected problems

## What would you do differently next time