IOM Model Update

Data Analytics Lab

January 2020

This report provides a brief update on the IOM model.

1 Table of Contents

2 Changes to the model

The original IOM model (see previous Committee briefs) was built upon the whole of the data within the Crimes system.

We have since re-built the model as a result of now having the ability (due to having developed the necessary business and technical logic) to filter the data to be in line with the requirements of the Management of Police Information (MoPI).

This report therefore relates to updates of the IOM model with a refinement in the selection of the hyper-parameters for the final models following the reduction in the data set to conform with the MOPI requirements. The changes are associated with the removal of data that is no longer eligible and the use of only Stop and Search with positive finds.

Details of the re-built model are in the Appendix.

At this stage we will be looking to move towards beta testing the model (whereby the model is productionised and predictions produced solely for the purpose of testing the accuracy of the model on new data).

However, as a part of this beta testing we are also looking to have 2 Local Offender Manager Units (LOMUs) use the resulting dashboard and outputs from the model.

It is considered that this will enable:

1. An assessment as to the use of the model's outputs by Offender Managers
2. A comparison by LOMUs of their currently managed offenders to the RFSDi for an assessment of the necessity of retaining their currently managed offenders.
3. Policy development amongst LOMUs for any 'surprises' found within the RFSDi / model.

At present, the beta testing is envisioned as running for an initial three months after which an assessment would be made as to whether to continue the beta testing for another three months.

3 Appendix – Model Rebuild

The process used mirrored that of the original work. The starting point was the model parameterisation of that work and the data approach was a direct parallel. In cases where there is no positive stop and search, the value for the relevant data is set to 0.

The original model was specified as follows:

XGBoost model, trained on pre_transition_score>50 imbalanced dataset with top 50 most important variables, algorithm specific parameters: eta = 0.3, max.depth=7, colsample_bytree=0.7, nrounds=80

As previously, the data only included those nominals with an RFSDi score of 50 or more. The data included in the model after the screening is approximately 40,000 nominals. The pre-screened and post-screened data distribution is presented below.

The work here builds upon the previous work of December 2018 using the same fundamental approach with a simplification of the model selection using only the XGBoost model (Chen et al. (2018)). Using the initially selected model parameters, a repeated subsampling of the data was used to verify the consistency of the top 50 variables, which were selected using the base model.

Table 1: Pre-screened (RFSDi >= 0)

Dependent Variable	ABT Date	Count
1	2012-11-01	440
1	2013-11-01	719
1	2014-11-01	676
1	2015-11-01	472
1	2016-11-01	241
1	2017-11-01	554
1	2018-11-01	581
0	2018-11-01	182645
0	2019-08-30	94
1		**3683**
0		**182739**

Table 2: Post-screened (RFSDi >50)

Dependent Variable	ABT Date	Count
1	2012-11-01	440
1	2013-11-01	712
1	2014-11-01	673
1	2015-11-01	472
1	2016-11-01	241
1	2017-11-01	553
1	2018-11-01	581
0	2018-11-01	39256
0	2019-08-30	13
1		**3672**
0		**39269**

The previous data had a total of 458, 366 observations of whom 1.8% were HHOs. The current data has 1.91% HHOs in the raw data. Once screening has been implemented the proportions are approximately 8.5% HHOs in the data.

The model initially uses the defaults for the XGBoost model and extracts the top 50 factors. It uses a 70-30 training- testing data set split.

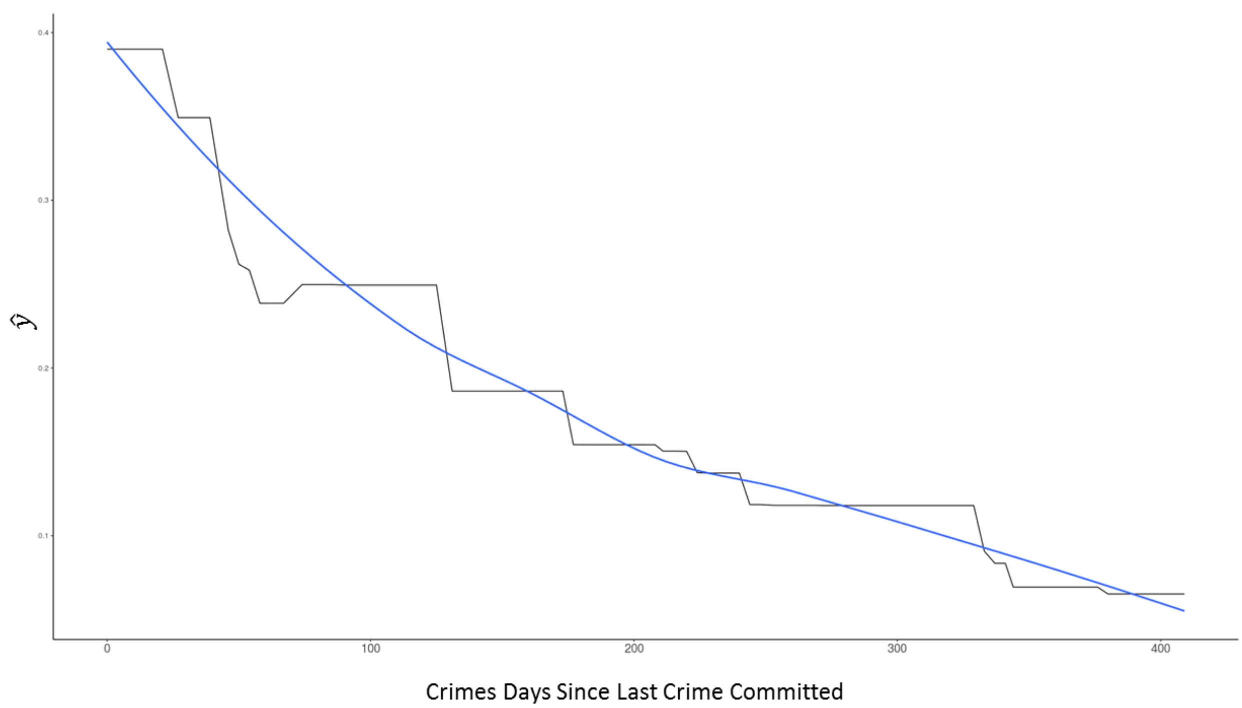
The top fifty variables are to be considered for the final models. In order to consider the robustness of the base model, a bootstrap was used to consider the mean and standard deviation of these variables in terms of their Gain. Of the top 50 in the base, 39 of the variables are in each of the boot-strapped versions though perhaps in a different order. It is therefore reasonable to use the base as a foundation for the modelling. The variables included are listed below.

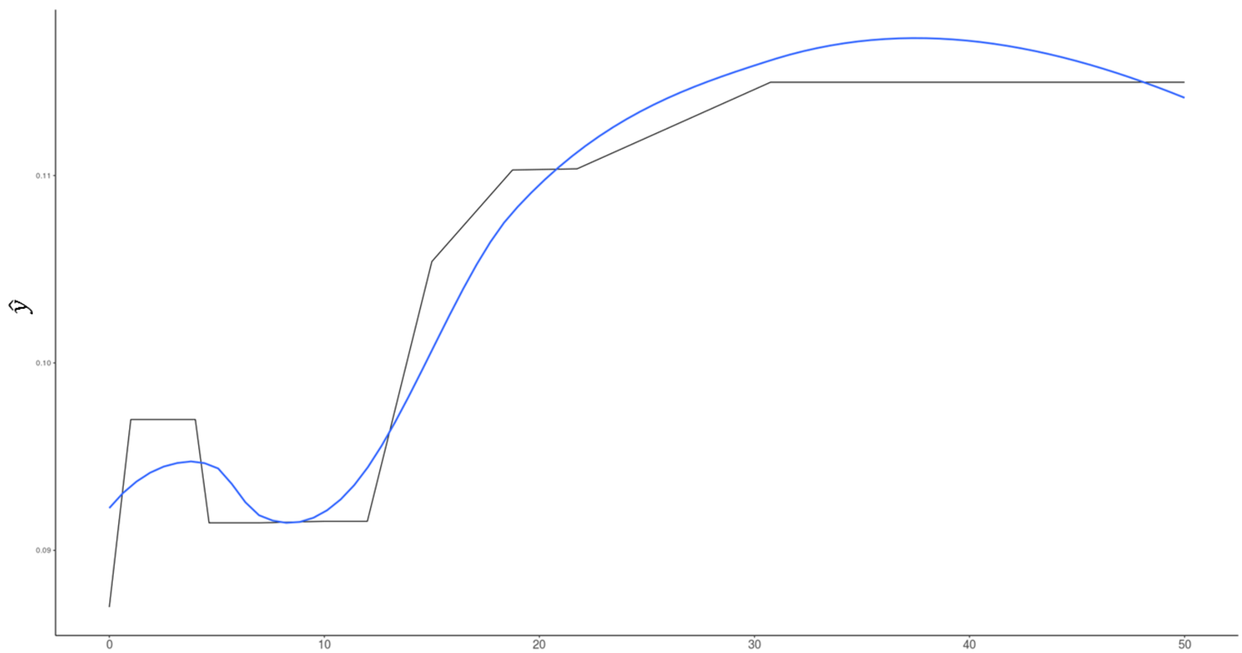
As was found previously, the networks of the nominals are seen to be important as are the changes in the various variables.

3.1 Partial Dependency Plots

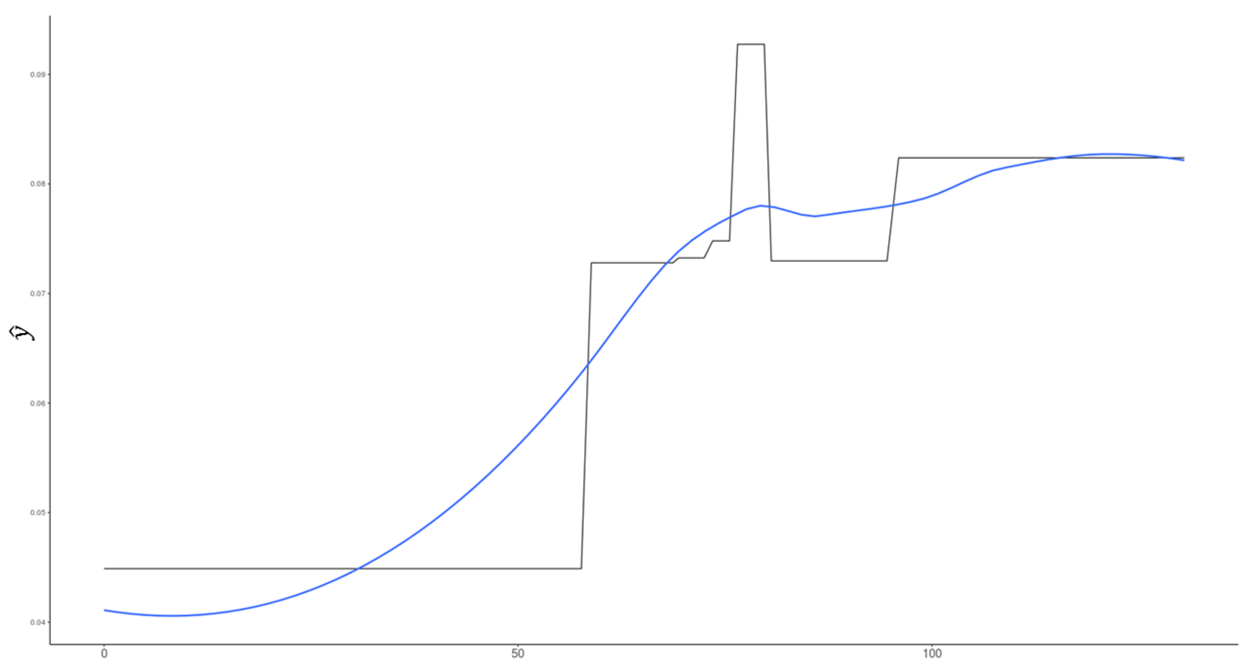
An explainer for the XGBoost base model is used using DALEX (Biecek (2018)). This allows us to isolate variables or data points to examine the net impact of changing a particular variable's value or how an individual was scored.

In order to consider the marginal effect (in regression terms), one uses a partial dependence plot. It demonstrates the impact of changing the variable of interest on the outcome variable. In the case of a linear regression, this would be a straight line with a slope equal to that of the coefficient. In more complex models this is not always the case, with potential for non-linearities and breaks being modelled. Plots for a number of top variables for the model are presented below with a smoothing line to demonstrate the overall trend or direction.

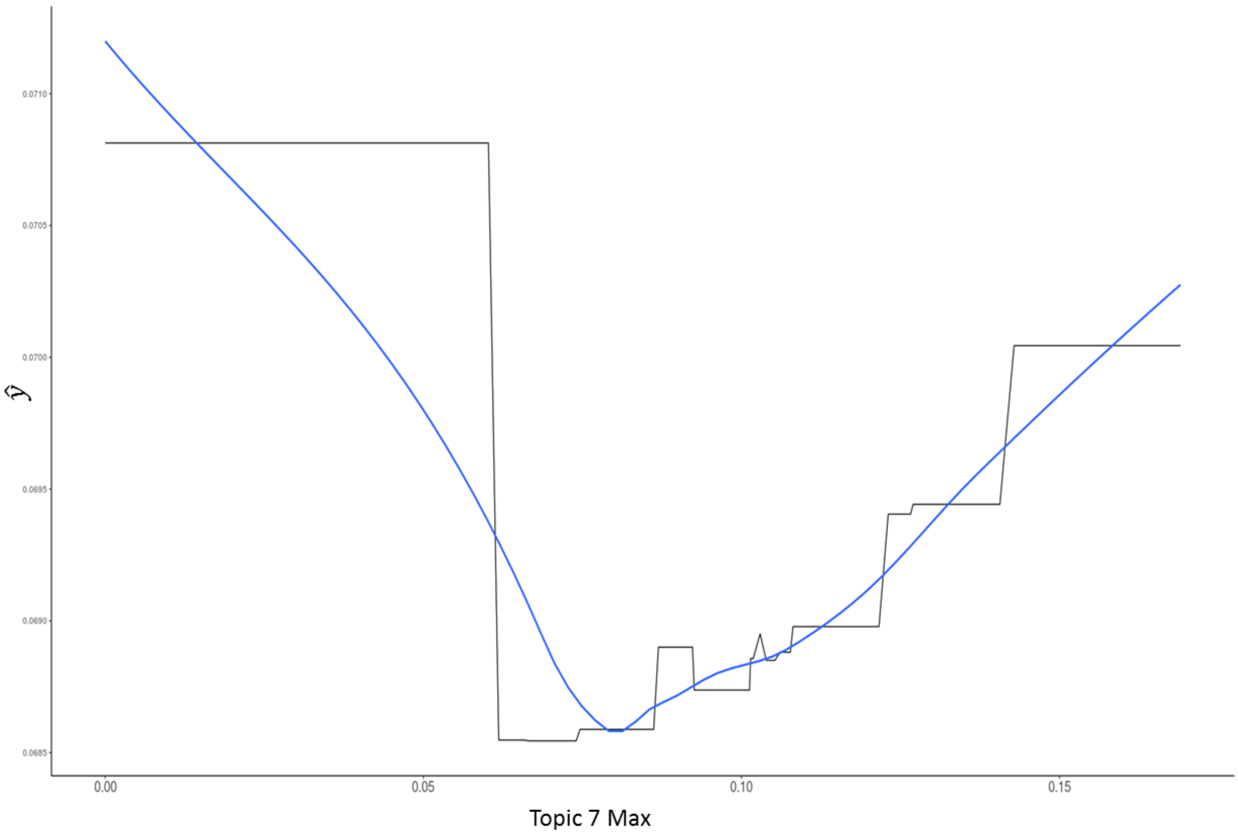




Crimes Cambridge Harm Index 12m



Page Rank New



These show the effect of changing the variable by a particular amount on the outcome.

3.2 Model Metrics for Base Model

The base model was assessed using the test (hold-out) data set. The standard model metrics were produced for the base model (pre-optimisation).

3.3 Optimization of the model

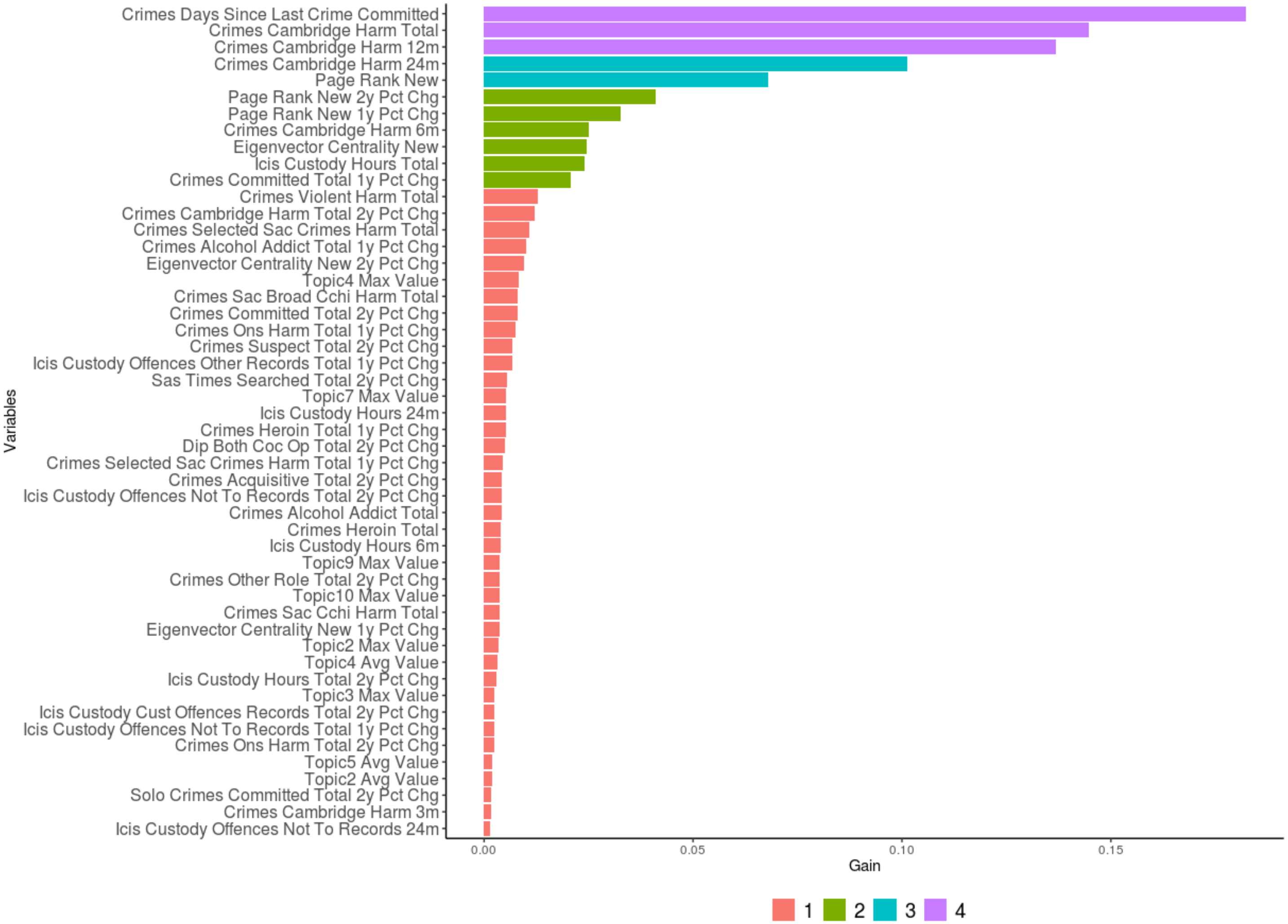
A grid for the hyper parameters was set up with searches across the following parameters

- η which determines the learning rate. The lower the value the more robust the approach is to overfitting, but there is a trade-off in terms of the speed.
- γ which represents the hurdle over which the loss reduction must pass. The larger the value of γ, the less likely a split is to occur
- *max_depth* determines the largest potential tree size in the algorithm.

A simple iterating search algorithm was written to search the parameter space. The algorithm used a 5- fold cross validation (for speed) and used the test AUCpr[i] mean as the metric for improvement. This found that the optimal parameters in the space of *max_depth*, η and γ was approximately 5, 0.19 and 0.75 with a test AUCpr mean of 0.9047744. It should be noted that the improvements are slight for many steps. The space was tested for local maxima, with no major problems. The original specification of the parameters *max_depth*=7, η = 0.3 and γ =0 and test . AUCpr mean of 0.8983. This is not a major improvement- there are limited improvements available.

The model was re-fitted to the training data using these parameters and the variable importance and standard metrics calculated for the test set. The variable importances demonstrate some changes in rank. These are shown in the following graph. The networks become a little more important, however there are no major changes in the order of variables.

Figure 1 Variable Importance for Optimised Model

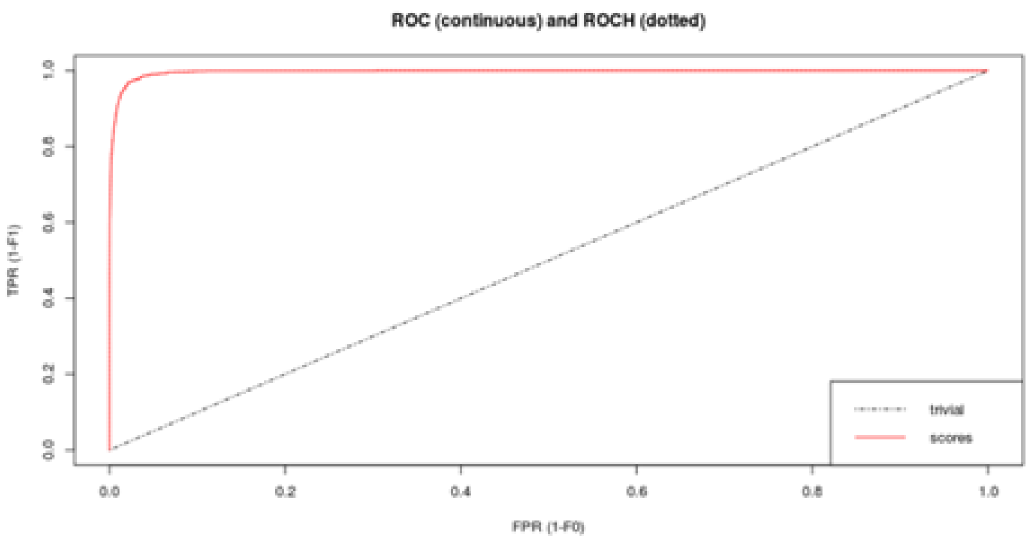


The statistics are presented below.

Table 3 Metrics for Optimised Model

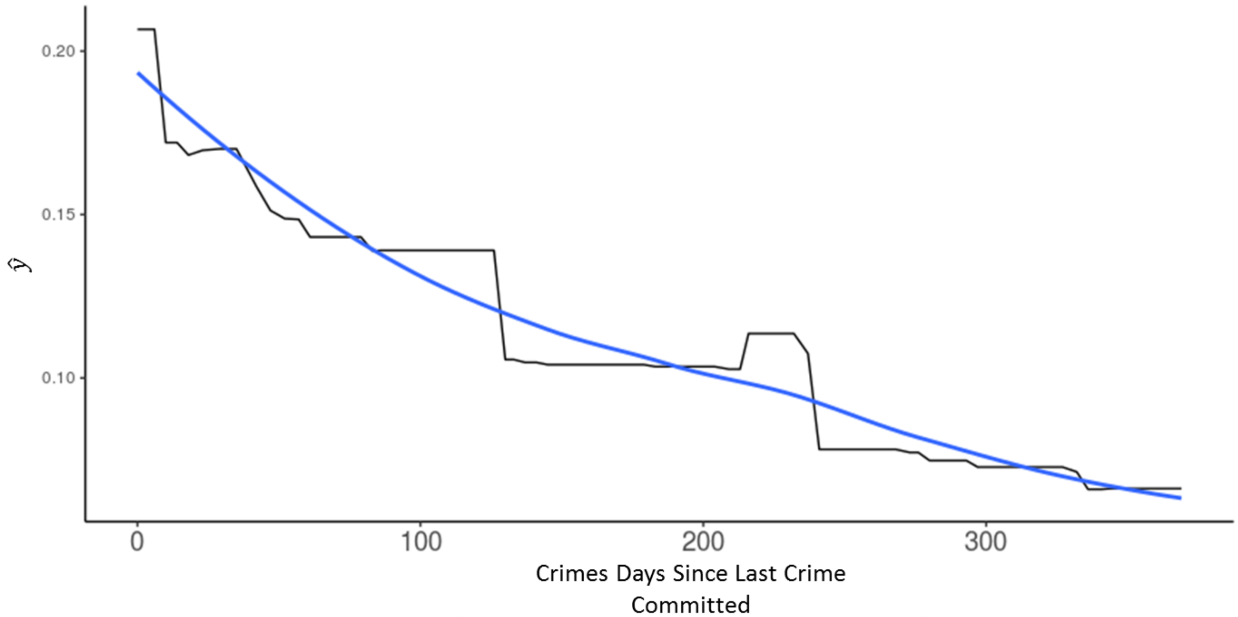
Measure	30%	40%	50%	60%	70%	80%	90%
Accuracy	0.992	0.993	**0.994**	0.993	0.993	0.992	0.991
Sensitivity	0.834	0.807	**0.774**	0.730	0.687	0.626	0.537
Specificity	0.996	0.997	**0.998**	0.999	0.999	0.999	1
Precision	0.790	0.837	**0.883**	0.913	0.935	0.957	0.98
F1 Sens Spec	0.908	0.892	**0.872**	0.843	0.814	0.770	0.698

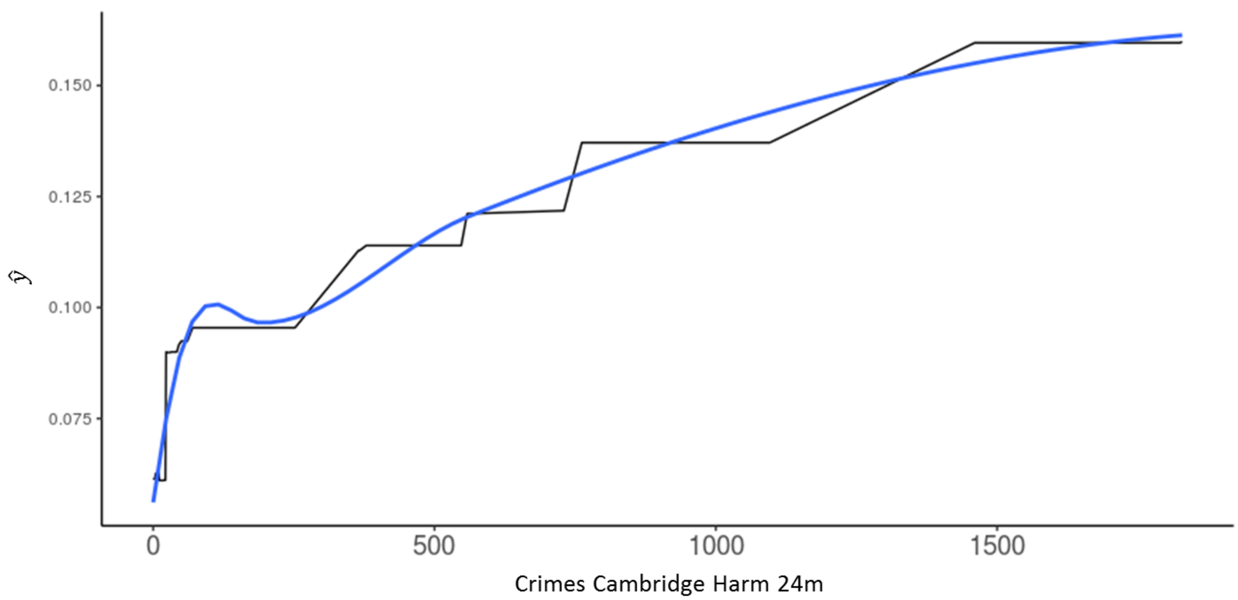
Figure 2 ROC plot for Optimised Model

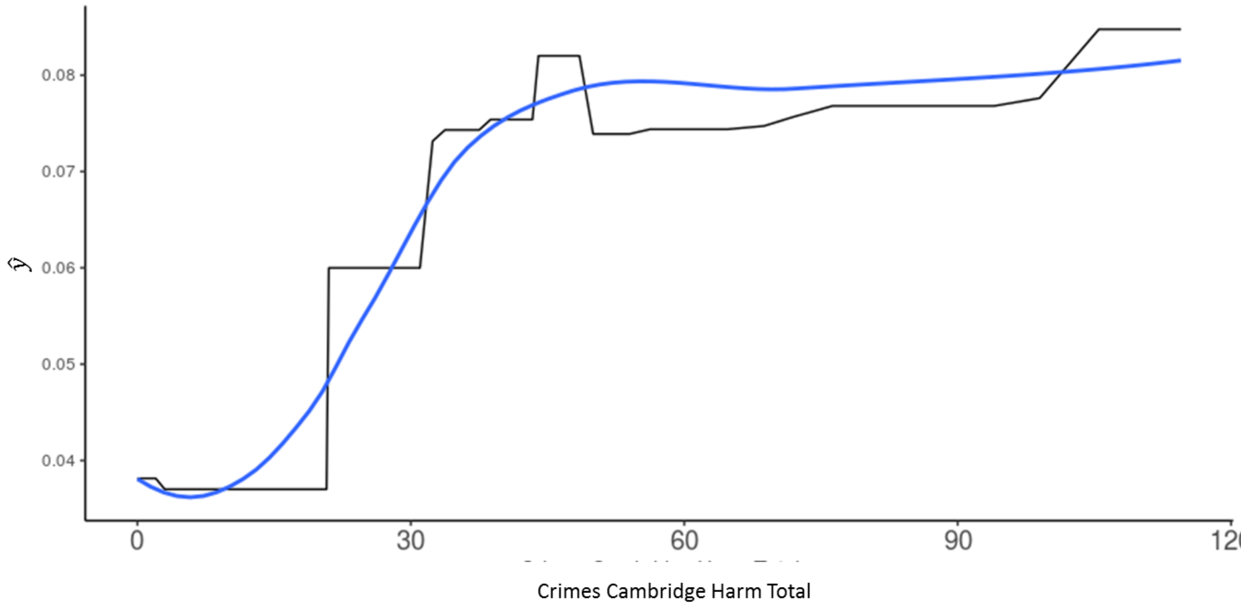


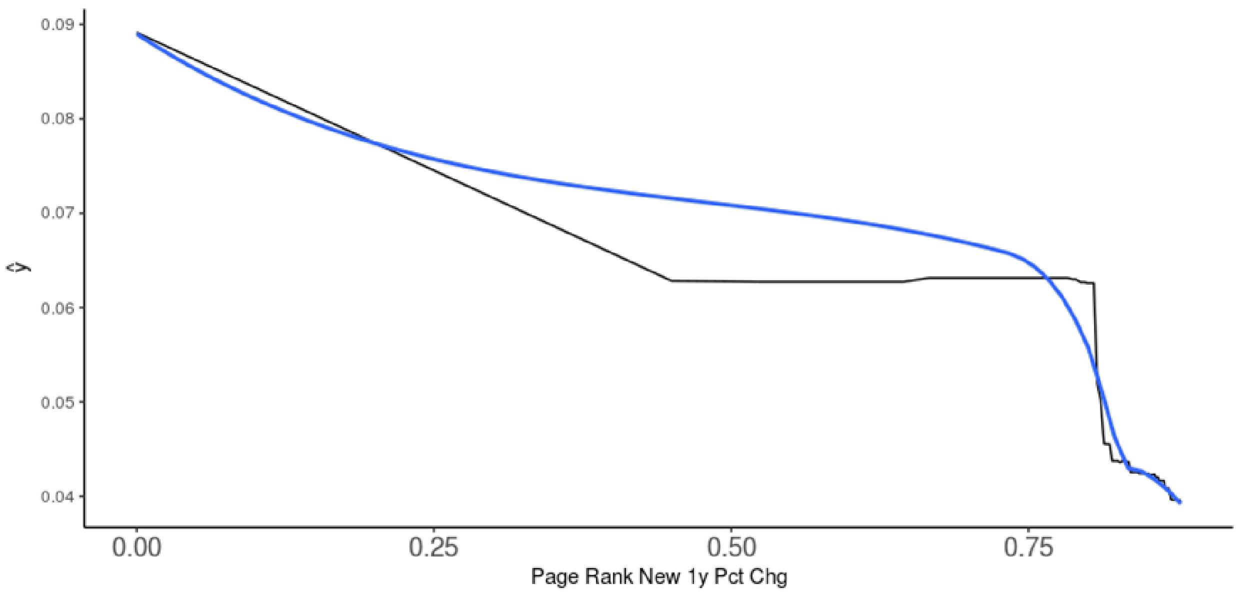
3.4 Partial Dependencies Plots

As before, the partial plots can give some idea about the impact of a particular variable.


0.20
0.15
0.10
ŷ
0
100
200
300
Crimes Days Since Last Crime
Committed


0.150
0.125
0.100
0.075
ŷ
0
500
1000
1500
Crimes Cambridge Harm 24m





As we can see, there is a general trend in some cases; though this tends to be non-linear with thresholds for a number of the variables. This suggests that there are ranges where the change in that variable, say the Cambridge Harm Index has no additional effect on the outcomes. Only when it passes a particular threshold again does it have any more impact.

3.5 Multi-label Classifier

Following the construction of the IOM model, a second step associated with the type of crime expected is used. The IOM predictions are linked to a multi-label classifier based on a random forest. This was a more direct approach than using binary classifiers whose linking can be problematic. Predictions from the IOM model for 2018 were used as an example. The Full multi-label random forest model was fitted against a subset of the 2018 data.

This gave nearly 40000 observations. The classification was generated from the variables in the (optimised) IOM model and combined with the predictions from the refreshed IOM model. The models were estimated on a training set of 70% of a data subset of 20000 observations.

Table showing the binary performance statistics for the Multi- Label Classifier based on a subsample for training of 7000 observations and 3000 for the test set give the following outcomes. The inference from this model should be carefully made as the sample omits firearms and driving offences in the sampling (these are relatively rare).

Table 4 Performance Statistics for Multi-label Model on hold-out set

	AUC test mean	F1 test mean	FNR test mean	FPR test mean	K test mean
acquisitive crime	0.993	0.989	0.021	0	0.984
driving crime	0.744	0	1	0	-Inf
drugs crime	0.921	0.670	0.434	0.038	0.591
firearm crime	0.891	0.011	0.995	0.0003	0.010
property crime	0.933	0.662	0.466	0.019	0.603
public order crime	0.817	0.468	0.653	0.044	0.368
sexual crime	0.878	0.479	0.668	0.007	0.443
other crime	0.859	0.243	0.860	0.001	0.224
violent crime	0.998	0.997	0.006	0	0.988

As is sensible, it is difficult to predict driving offences. The κ statistics record infinity due to the low numbers of these outcomes in the sample data.

The multivariate statistics are below. These are the Hamming Loss, Subset Accuracy, F-measure and Accuracy respectively.

Table 5 Multivariate Performance Statistics

	Hamming Loss	subset01	F1	Accuracy
Model	0.073	0.461	0.840	0.772

4 Features used in Models

Rank in variable importance	Variable	Description
44	crimes_acquisitive_harm_total	Harm score for the nominal for acquisitive crime
34	crimes_cambridge_harm_24m	Cantab Harm Index for the Nominal over the last 2 years
6	crimes_cambridge_harm_total	Cantab Harm Index for the Nominal
49	crimes_cambridge_harm_total_2y_pct_chg	Percentage change in the Cantab Harm over 2 years
2	crimes_committed_total	Total number of crimes committed
26	crimes_coof_min_age_committed	Age at which nominal invovled in crime ass co-offender
9	crimes_days_since_last_coof_committed	Days since last crime committed alone
1	crimes_days_since_last_crime_committed	How long since the last crime
8	crimes_days_since_last_solo_committed	Days since last crime committed as a Co-offender
35	crimes_drug_addict_total	Drug addict related crime in total for nominal
46	crimes_drugs_harm_total	Harm associated with the nominal with regards to drug offences
27	crimes_min_age_committed	Lowest age of crime involvement
39	crimes_ons_harm_24m	Total measure of harm by nominal measured using ONS methodology
11	crimes_ons_harm_total	ONS crime score
45	crimes_ons_harm_total_1y_pct_chg	Percent change over last year for nominal using ONS methodology
38	crimes_ons_harm_total_2y_pct_chg	Percent change over last 2 years in nominal's ONS harm measure

37	crimes_other_role_total	Number of crimes nominal has been associated with in an 'other' role
31	crimes_property_harm_total	Harm score associated with property crime
42	crimes_public_order_harm_total	Public Order Harm score of the nominal
25	crimes_sac_broad_cchi_harm_total	Broad measure of SAC crimes in Cantab Harm Index
50	crimes_sac_broad_cnt_total	Broad measure count of SAC offences for the nominal
48	crimes_sac_cchi_harm_total	SAC total measure for the Cantab Harm Index
32	crimes_selected_sac_crimes_harm_total	Narrow SAC offence harm total
16	crimes_solo_min_age_committed	Earliest age of a crime committed by nominal as single offender
12	crimes_suspect_24m	Number of crimes for which the nominal has been a suspect in the last 2 years
36	crimes_suspect_6m	Nominal suspect for crimes in the last 6 months
18	crimes_suspect_total	Total number of crimes for which the nominal has been a suspect
21	crimes_suspect_total_1y_pct_chg	Change in the number of crimes nominal has been a suspect in over the last 12 months
10	crimes_suspect_total_2y_pct_chg	Percent change over last 2 years of crimes for which nominal is a suspect
23	crimes_victim_total	Number of crimes where the nominal has been a victim
29	crimes_victim_total_1y_pct_chg	Change in the number of crimes as a victim in the past year
41	crimes_victim_total_2y_pct_chg	Change in the number of crimes as a victim in the past 2 years
22	crimes_violent_harm_total	Harm score associated with violent crime
5	dip_both_coc_op_24m	Cocaine and Opiates within last 2 years
20	dip_opiates_12m	Optiate in last 12 months

3	eigenvector_centrality_new	Network centrality measure for nominal's network
17	eigenvector_centrality_new_1y_pct_chg	Change over the last 1 year in the network centrality measure
15	eigenvector_centrality_new_2y_pct_chg	Change over the last 2 years in the network centrality measure
33	icis_custody_cust_offences_records_24m	Number of records in the ICIS custody records in the last 2 years
7	icis_custody_cust_offences_records_total	Number of records in the ICIS custody records
28	icis_custody_offences_assault_records_total	Records of nominal involvement in assault in ICIS records
47	icis_custody_offences_other_records_24m	Other records of nominal in the ICIS system over the last 2 years
24	icis_custody_offences_other_records_total	Other records of nominal in the ICIS system
43	icis_custody_offences_theft_records_total	ICIS custody records for theft related crime
13	nominals_age	Age
4	page_rank_new	Importance in network of the nominal
19	page_rank_new_1y_pct_chg	Change over the last 1 year in the pagerank measure
14	page_rank_new_2y_pct_chg	Change over the last 2 years in the pagerank measure
40	solo_crimes_committed_24m	Number of crimes committed in the last 2 years alone
30	solo_crimes_committed_total	Total number of crimes alone

Further Explanation	**Details**
1. PageRank	A numeric weighting of relative importance of the nominal in their network. It is a 'vote' of how important a nominal is within their network. This vote is determined by the number of links to that nominal. The value is determined by a principal eigenvector of the linkage matrix.
2. Eigenvector Centrality	A numeric measure of the influence of the nominal inside their network based upon the adjacency matrices of the nodes. Nominals with a few highly connected links may have high eigenvector centrality despite not necessarily having many links themselves.
3. Latent Dirichalet Allocation (LDA)	LDA produces the probability of a document or sequence of words (here the OASIS log) being associated with a particular topic. There is no particular meaning of the topic such as Motor vehicles, rather they are linked in probability of co-occurrence. The probabilities give a characterisation of the log.
4. Cambridge Harm Index or Cambridge Crime Harm Index (CCHI)	A measurement of crime rates based on the 'harm' they do such that not all crimes are equal (Sherman, Lawrence; Neyroud, Peter William; Neyroud, Eleanor (3 April 2016). "The Cambridge Crime Harm Index: Measuring Total Harm from Crime Based on Sentencing Guidelines". Policing. 10 (3))

References

Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. 2018. *XGBoost: Extreme Gradient Boosting*. https://CRAN.R-project.org/package=XGBoost.

Saito, Takaya, and Marc Rehmsmeier. 2015. "The Precision-Recall Plot Is More Informative Than the Roc Plot When Evaluating Binary Classifiers on Imbalanced Datasets." *PloS One* 10 (3): e0118432.