Ethics Committee Briefing Note

Project Reference: DAL_2018_0001_IOM Model

Purpose of data analysis:

The ultimate aim is to make predictions at the level of individuals as to the probability that they will move from committing low / middling levels of harm (via criminal activity) to perpetrating the most harmful offending. ‘Harm’ is as defined in appendix A.

It is envisaged that the resulting harm score and, separately, the predictions arising from the model will be used by Offender Managers enabling them to concentrate their resources in putting interventions in place in order to dissuade individuals from progressing further criminal activity.

Any potential interventions are designed according to the nature of the activities undertaken by individuals and can follow a number of routes. For those who come under the Offender Management process a management plan is put in place and potential interventions can range from providing help with finding training, employment or housing, entering into drug rehabilitation programmes through to risk management of certain offenders, setting license conditions with probation, etc. For individuals arising from the results of the predictive modelling element it is envisaged that the former types of interventions (rehabilitation) would be entered into.

Source of analytical question / hypotheses to be examined:

This project stems from the aims of the Force Delivery Plan and has been requested via the FET.

Data to be used:

The dataset includes underlying data from Crimes (crimes committed), IMS (intelligence), ICIS (custody), PINS (prison notification system), Corvus (intelligence and tasking system), OCG (organised crime group data), OASIS (the event logging system), SAS (stop and search) and DiP (drug intervention programme data). The view of individuals (the nominal view) has been undertaken following the GID (golden ID) creation process. This requires matching and merging of individuals who may initially appear to be separate individuals due to errors between systems, typos or other data quality issues. This should, as far as is possible, avoid the incorrect matching of crimes / intelligence to individuals, ensure that individual records are not duplicated and that individuals are not allocated to the incorrect area of residence, etc.).

The crimes dataset available covers the last 20 year period, with the target variable calculated over the last 8 years (to ensure that there are a reasonable number of observations of those moving from low/middle levels of harm creation to the most harmful offending). The definition of the differing levels of harm has been derived from the crimes data, intelligence logs, data from the drug intervention programme (DiP) and the Cambridge Crime Harm Index (see appendix A for details).

Level of analysis:

☑ Individual
  Individuals aggregated?
    ☐ Yes
    ☑ No
☐ Specific Area:
    ☐ Output Areas
    ☐ Super Output Areas - Lower
    ☐ Super Output Areas - Mid
    ☐ Wards
    ☐ Districts
☐ West Midlands
☐ Other

Reliability of data:

The data are sourced from WMP systems as noted earlier. A large part of this project has involved making an assessment of the quality of the data, the robustness of the various systems, etc. Any data quality issues have been noted and where applicable have been incorporated into the project (e.g. by excluding some data from a system if it is felt to be unreliable). These data, as part of the analytical project life cycle, have been assessed for missing values, outliers and potential biases (see the potential for bias section and methodology sections below).

These systems are those currently used by WMP in their day-to-day business. Specifically in the case of intelligence data, these have been examined as to their veracity, source, etc. prior to inclusion (i.e. only intelligence considered to be credible from credible sources has been used). The intelligence data has been given a lower weight in the construction of the harm score (the RFSDi) (essentially given lower importance) than actually observed crime incidents.

Method of collection:

Data regarding crimes, etc. are held on the current WMP systems and extracts have been used to build this initial version of the model. When put into the productionised stage, data will again come from the relevant WMP systems on the Hadoop infrastructure.

Is the proposed use of the data compatible with the use for which it was originally intended?

A large part of the remit of WMP is the prevention of crime as well as its detection. Through the appropriate analyses, the crime data held by WMP will ultimately be used to prevent crime and so harm. In these circumstances it is considered that the use of the data will be compatible with the purpose for which they were originally intended.

Population involved (does this include victims, vulnerable adults, children, etc.):

The requisite dataset used for defining the target variable and the features contained in the analytical base table (ABT) and used in the training and testing of the model come from the core 9 systems of Crimes, IMS, ICIS, Oasis, Compact, DiP, SAS, PINS and OCG. The dataset does include vulnerable adults and children if they were perpetrators, have been reported missing or a victim of crime.

Age of data:

The crimes data cover a 20 year period (8 years in the case of the target variable), however some of the core systems have data going back 5 years. The datasets used in building the model covered a period up to June 2018. In the interest of producing a model that will be robust, the overall time period is necessary to ensure that there are enough observations of individuals moving from low/middle levels of harm to the most harmful levels whilst also allowing a period for data prior to movement into the most harmful offenders category (see appendix B – Exploratory Data Analysis for further details).

Sample or entirety:

Entirety (all incidents of applicable offences over the last 20 years).

If sample:

NA

Method of sampling:

NA

Method of choosing sample size:

NA

Sample size:

NA

Potential for bias in data:

There is potential for bias to be present in the underlying dataset in terms of the recorded incidents of harmful / most harmful offences and within the intelligence reports. Such biases may have arisen from a "self-fulfilling prophecy" process in terms of the allocation of resources to locales / individuals previously noted for certain offences.

A number of steps have been taken to reduce the potential for bias:

1. Intelligence has been extracted according to its veracity; this should help ameliorate the potential for bias. The effect of intelligence in the construction of the underlying RFSDi has also been down weighted (thereby reducing its contribution to the overall RFSDi score).
2. An extensive exploratory data analysis (EDA) stage prior to model build has not highlighted any other areas of concern in relation to potential biases (although see below and the methodology section).

Following data quality checking and ID allocation, as part of the initial EDA phase, the spatial and temporal nature of the target variable has been examined to ascertain if there are any clusters of the most harmful. Sensitive attributes (such as ethnicity) have been examined. This initial sift consisted of building an initial model using an ensemble method (XGBoost) in order to ascertain the most important features with the greatest amount of predictive power. This approach has the advantage of taking potential interactions amongst the different features into account whilst assessing feature importance. During the EDA stage, an initial assessment of the effect of a sensitive attribute (ethnicity in this instance) has been undertaken; further details are in appendix B. These attributes did not make it into the final list of features used to build the model.

Inaccuracies / missing data points:

The EDA stage did include an extensive assessment of data quality issues, missing observations, etc. It was found that for the most part data were not missing at random and this therefore informed the approach taken to missing values (further details are in appendix B).

Proposed means of addressing missing values:

In the final ABT, some values were missing as they were logically not present for some individuals (e.g. number of solo crimes committed when there have been no solo crimes). Where data were missing due to data quality issues, etc. then missing has been treated as a category in its own right (see appendix B for details).

Type of analysis:

☐ Exploratory
☐ Explanatory
☑ Predictive
☐ Optimisation

Methodology:

Following data extraction, quality checking, parsing and uploading an EDA process was undertaken to examine the nature of the data, the potential for differentiation over sensitive attributes, the presence of missing values, outliers, etc. The target variable has been defined as individuals who moved from the low and middling groups of the RFSDi into the high / super high harm RFSDi groups; this therefore becomes a binary classification problem (see appendix A for details). Following this, the potential for a feature to be related to the target variable (i.e. have some predictive power) has been assessed on a multivariate basis via running all features through an XGBoost model and ordering of variable importance based on the gain.

In order to answer the predominant question (estimation of probability of individuals moving from low/medium harm to high levels of harm) a number of predictive models were built based around three types, namely XGBoost, random forest and a support vector machine (SVM). These model types were chosen due to the nature of the data and the known robustness and accuracy of the model types. The number of models increased above three due to models using all of the data and models built excluding observations whereby observations with RFSDi scores below 50 were excluded in conjunction with including the full circa 500 features and those using the top 50 ascertained via the initial sift of features. For the list of features used in the final model, see the end of appendix C.

The overall ABT once constructed was split 70% / 30% into training and test sets. Each of the models built was then tested against the test data by way of making predictions as to who did and did not move into the most harmful offender categories. For this binary dependent variable an initial cut-off point of a probability of >= 0.5 shall be taken as the point at which an offender shall be assessed as being within the most harmful offender category (although see below). Comparison of the predictions to the actuals on the test dataset has formed the basis of the decision as to which of the models is taken forward to the beta production phase. This proximity (i.e. how good or otherwise a model is) has been assessed by way of:

1. Accuracy
2. Overall error
3. Area under the (ROC) curve (AUC)
4. Sensitivity
5. Specificity
6. Precision
7. F1 score over sensitivity and specificity

(see the glossary at the end of the document for an explanation of these terms).

As part of this process, the performance of the models over a range of cut-off points have also been assessed to gain a picture of the predictive accuracy of the models over a range of probability cut-off points (partly to assess the distributional properties of the estimated probabilities from the models and partly to ensure that should WMP wish to concentrate on those potential offenders with higher estimated probabilities that they could allocate resources to, the final model will also be robust at these points).

The beta production phase will consist of using the winning model to make predictions for a period of between 1 and 3 months (these predictions not to be used for operational purposes). At the end of this period, the predictions shall be compared to the actual occurrences of offending in order to ensure that the model can make predictions within reasonable tolerance levels for operational purposes. Tolerable in this instance shall mean similar occurrences of predictions and comparison with the actual movement of individuals into perpetrating high levels of harm that is at least 20% better than choosing occurrences at random (including random over the whole offending population and over those individuals who appear to be on a trajectory of becoming high harm offenders).

Should the beta production phase be passed, the model will then go into the production phase (envisaged at present to be run daily) and the predictions monitored on a regular basis to ensure that there are no issues with changes to the underlying data, the model or the quality of the predictions made. This monitoring shall include the performance measures (as noted above) on recent weeks’ predicted and actual movements into high harm levels.

Discussions with subject matter experts (SMEs) revealed that, for reporting purposes, as well as the propensity for individuals to become high harm offenders, it would be highly useful information if the potential type of crime they may commit was also included. To this end, a model has been built that estimates the probability that an individual, should they fall into the potential high harm category (and not those who don't) will commit any of 9 crime types. The two crime types with the highest estimated probabilities are reported (see appendix D).

Potential benefits for the public:

The main potential benefit to the public is likely to be the deployment of WMP resources in a targeted manner to reduce occurrences of the most harmful offending / attempt to dissuade individuals from entering into behaviours that generate the higher levels of harm.

Expected level of intrusion:

As the analysis / modelling will be at an individual level there is potential for intrusion. However, the model will involve the use of WMP data currently available which are accessible by WMP on a statutory basis and does not otherwise involve the use of external datasets or special permissions.

Potential for identifiability of individuals:

The model will be built at the individual level and therefore individuals will be identified (otherwise no operational actions could be undertaken as a result of the modelling). The potential risks of this level of identifiability will be mitigated by way of ensuring that information from the modelling process shall be contained in the appropriate sandbox environment and shall not be used for purposes (or access granted) other than those intended.

A list of the stakeholders who will use the outputs from the model shall be compiled to ensure that the data are only provided to them (by way of a Qlik dashboard), notably senior officers and offender managers. Changes to these applicable personnel shall be notified to the Principal Data Scientist who will keep a centralised list of operational stakeholders in order to mitigate the risk of data being provided to stakeholders when it is no longer appropriate to do so.

If not aimed at identifying individuals, how will anonymisation be undertaken:

N/A

Will the project eventually be automated:

☑ Yes
☐ No

If automated, what will be the level and nature of human oversight:

Once in full production phase, the model outputs will be checked on a monthly basis for distinct changes (in terms of the distribution of estimated probabilities) and comparisons to occurrences of harmful / most harmful offending shall be made at each run (and assessments made as to performance via the measures noted above). This will ensure that there are no issues with the underlying data and that the deviations from actual levels are within tolerable limits.

This will also ensure that the model is provided with real-world feedback to ensure that there are no instabilities and that it's outputs continue to be in line with expectations.

Any decision as to which nominals are interacted with by offender managers and the nature of any such interactions will be wholly determined by offender managers and as such there are no automated decisions.

Potential risks of automation:

The main risk is in the misallocation of resources. This risk can be mitigated via the governance process following production.

Potential for unintended consequences:

Due to the level of analysis being at the individual (and subsequent interventions), there is a potential for unintended consequences, particularly in terms of setting up a self-fulfilling prophecy cycle.

Means of mitigating potential unintended consequences:

The risk of unintended consequences shall be mitigated by:

1. Ensuring the final model is as robust as possible. In building the model, we have placed greater weight on the specificity as opposed to sensitivity so that the probability of mis-identifying individuals as potential high harm offenders is minimised.
2. Keeping track of the model's predictive performance as noted above once in the production phase.
3. Keeping track, through an intervention monitoring process, of the occurrence or otherwise of unintended consequences.
4. Receiving information through intervention management regarding the SMEs' assessment of the usefulness / accuracy of the model.

Are the public likely to agree with the purpose of the project and means of undertaking?

Given concerns from the public regarding the ability to reduce, deter and prevent crime, it is likely that this model and the operational allocation of resources following therefrom would be viewed by the majority of the public as being beneficial.

Risk of challenge from affected individuals, groups or organisations:

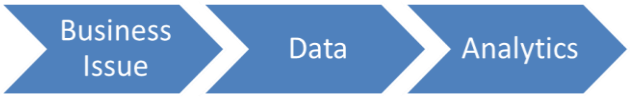
There is a potential for challenge from individuals as to their inclusion within the model. This will be mitigated firstly by ensuring that the underlying data are as robust as possible and the probability of including individuals who should not be reduced to as a low a level as possible.

Should it be discovered that an individual has been included in error, this individual shall be removed from future rounds of the model run.

This risk shall further be mitigated by way of the model outputs ONLY being used for the purposes for which they are intended – unless a part of any planned intervention, the model's outputs (or indeed whether an individual is included in datasets used for prediction) will NOT be provided to any other parties or agencies.

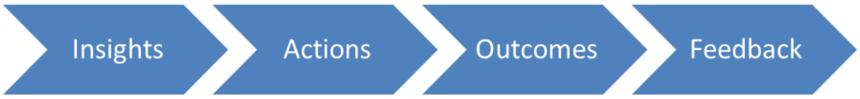
Governance and oversight process for the project (historical):

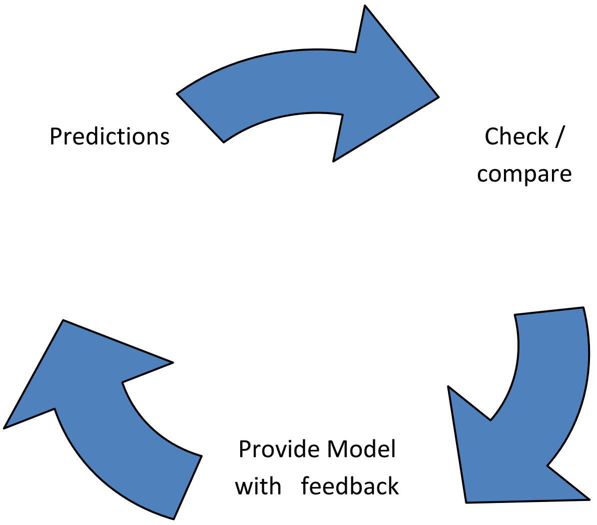
This project has been subject to the project inception process as outlined in the Data Analytics Lab operational procedures:



Governance and oversight process for the project (operational):

If approval is recommended by the Ethics Committee / Chief Constable, the Governance process in line with the operational procedures of the lab will be adhered to:





Data security measures:

The underlying data will be held wholly within the secure computing environment of WMP / the Data Analytics Lab with no sharing of those data with other parts of the business other than stakeholders who require the outputs for resource allocation purposes.

Balance between privacy and likely benefits of the project:

Given the information noted above as to the security of the data, ensuring the robustness of the data, the model build and run procedures and the uses to which the model outputs will be subject, it is considered that there is a balance between privacy and the likely benefits of the project.

Level of stakeholder engagement:

All relevant stakeholders (including relevant SMEs and via the governance and commissioning process) have been informed and the project has been discussed with them.

Legal Perspective: To follow

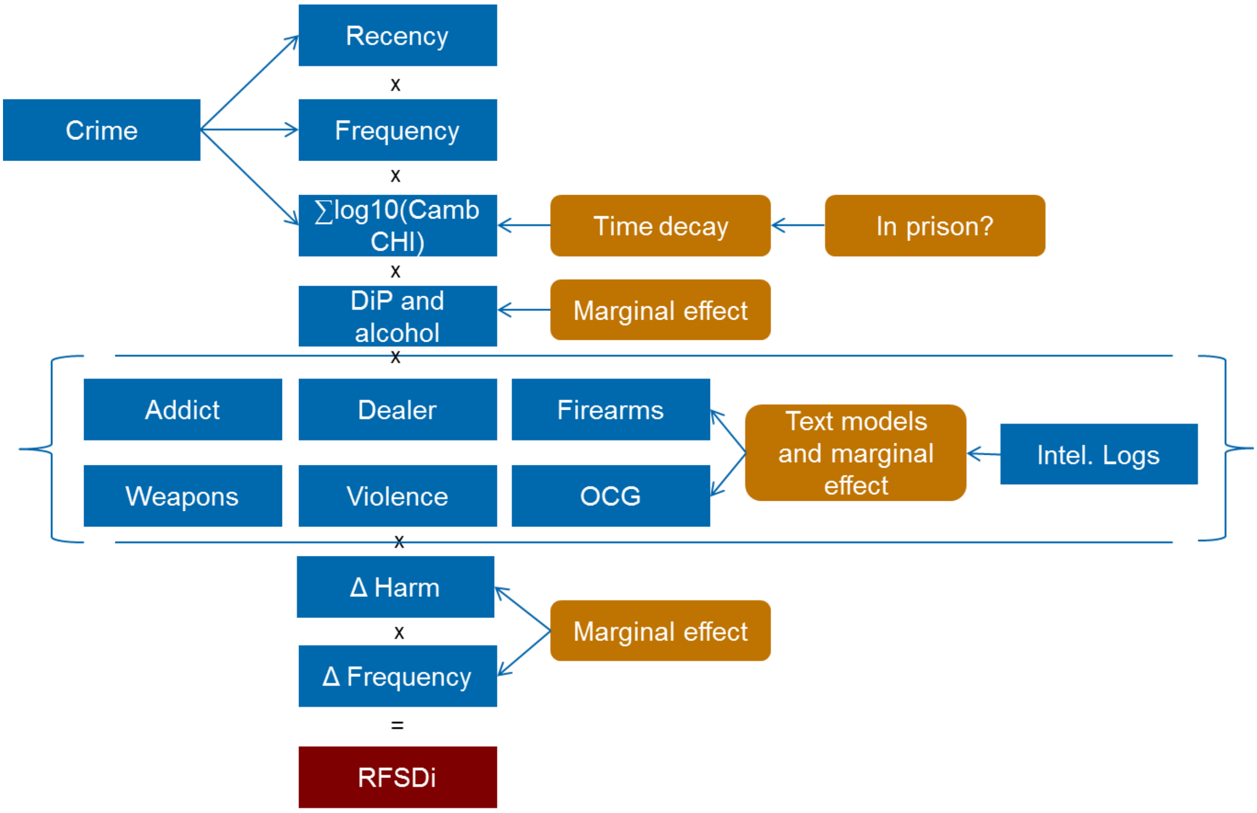
APPENDIX A:

Definition of Harm (and the target variable):

The target variable in this instance is a binary classification whereby 0 = have not transitioned from low / medium harm into high harm and 1 = have transitioned.

This classification is based upon the harm created by a nominal as measured by the RFSDi (Recency, Frequency, Severity, Drugs and intelligence).

The broad process for the construction of the RFSDi is outlined in the diagram below:



The crime history of nominals is ordered and nominals are allocated to 1 of 10 bins in terms of the recency of their crime and the frequency (essentially the average number of crimes per year where a frequency of 0.001 is entered if there is only one crime, etc.). These measures are relative over the nominals, so recency is in terms of the recency over the whole dataset rather than some hard break point.

Each of the crimes is then matched to a figure from the Cambridge Crime Harm Index (CCHI). The log to the base 10 is then actually used in order to reduce the potential for volatility and skewing of the score due to the orders of magnitude differences between the different crimes. This is then summed over each nominal following the application of an exponential weight decay to reflect the passage of time (so, for example, a robbery committed 2 years ago scores lower than a robbery committed yesterday). SMEs considered that it would be important to reflect the harm of those who are currently serving custodial sentences without this being reduced due to time effects. For this reason, for those who are currently in prison, the time decay starts from the point at which they entered prison (for the current sentence), essentially the RFSDi is calculated for these nominals as if their sentence start date were the current date. For any crimes committed post the start of their sentence, the RFSDi is calculated as for everyone else (the start date for the weight decay is the current day).

Other measures of crime harm / severity such as the Crime severity score published (experimentally) by the Office for National Statistics were also investigated. The ONS score, when applied to nominals, correlates highly with the CCHI (spearman correlation of circa 0.8 and a maximum information criterion of 0.8). It was decided, at this stage, to use the CCHI due to its democratic accountability.

When consulting with SMEs it was also considered that the use of drugs and alcohol were also important considerations in the committing of crime and so in creating harm. Whether a crime was related to alcohol or whether a nominal is addicted to alcohol has been taken from the crimes system. Drug usage (and whether a nominal has tested positive for a particular substance) has been taken from data arising from the DiP data. These have then formed weights to be applied as part of the RFSDi. The weights have been determined by regressing the centred $\log_{10}$(CCHI) on the types of crimes committed, years active in crime and the feature of interest. The resultant weights therefore are essentially some fraction of the standard deviation of the $\log_{10}$(CCHI) which are then applied.

SMEs also considered it important to include information from the intelligence logs as to drug dealing, addiction, firearms, violence, weapons (other than firearms) and affiliation with OCGs. Whether a nominal is mentioned in an intelligence log that refers to these subjects has been ascertained via the building of naïve Bayes models (with document type frequency as a prior) using document term matrices as the features which identifies whether a log refers to these issues. Logs have been linked to nominals via the GID creation process and so weights are applied (and calculated in the same fashion as for alcohol addiction, etc. with a further weight (reducing the initial weight's value) then being applied.).

Changes in behaviour are also accounted for in the RFSDi by way of applying weights (ascertained via a similar method as for drug addiction, firearms, etc.) in relation to the trajectory of a nominal's behaviour in terms of harm and frequency – if the trajectory is upwards, a positive weight is applied, if downwards, a negative weight.

The different elements are then multiplied to produce the final RFSDi.

Appendix B:

Exploratory Data Analysis

This appendix explains and conducts the process of exploratory data analysis (EDA) for the IOM model and presents the findings.

The overall purpose of the model is to predict whether a nominal will become a High Harm Offender (the target group) in a year from the moment he's scored. In order to account for transition into a target group, criminal behaviour paths for years 2010 - 2018 were examined. Each year, every nominal was assigned an RFSDi score and depending on the value, the nominal was classified into one of the following groups:

- not known - when nominal had no crime history prior to this year, the score was 0.
- Low, low-medium, medium-high - all classified as 0's
- High, Super High – classified as 1's

Target group definition and analysis

Firstly, we check the number of zeros and ones (target group) in the final full analytics base table (ABT) and the "ABT date" variable (date for which the ABT was calculated/used) for each observation. We aggregate these into a frequency matrix.

target variable	frequency	Percentage
0	449982	98.2%
1	8384	1.8%

target variable	ABT date	frequency
0	08/05/2017	449982
1	08/05/2012	3060
1	08/05/2013	1321
1	08/05/2014	1301
1	08/05/2015	954
1	08/05/2016	1100
1	08/05/2017	648

The dataset created contains 458,366 observations. It is an ABT with one distinct nominal per row (the observations), and each nominal is classified into one of 2 groups:

- Ones (High harm offenders - HHO) – 8,384 nominals: All nominals who transitioned into HHO in the years 2013 - 2018 were selected. A "transition" occurs when a nominal changes their

RFSDi group from any of the (Low, low-medium, medium-high) harm group in a previous year into any of the (High, Super High) harm group in the current year. If many transitions happen during the observation period the first one is taken. Transitions into HHO were calculated for years 2010 - 2018, but only years 2013 - 2018 were considered an observation period, resulting in ABT dates from years 2012 - 2017 (Offender's history is calculated one year before transition). The main reason for this year range is data availability – it was desired to ensure we have the possibility of at least 2 years of history of a nominal from every system.

- Zeros (Non -HHO) – 449,982 nominals: all nominals who did not transition into HHO and were never classified as high or super high were selected.

Due to the above classification, some nominals were not selected either as 0 or 1. This includes nominals who were unknown throughout the whole period and committed their first offence in the year 2017/2018, or those who follow the path of "Not known" -> {High, Super High} (as their history one year before transition is not known).

Target group analysis. Comparison with non-HHO group

Score distribution for transitioning nominals the year when transitioned:

```
 Min.    1st Qu.   Median  Mean    3rd Qu.  Max.
 282.2   310.1    335.2    353.5    377.5    788.2
```

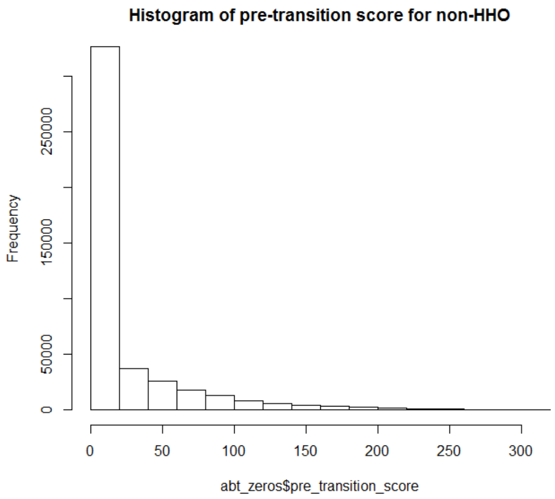
and one year before transition:

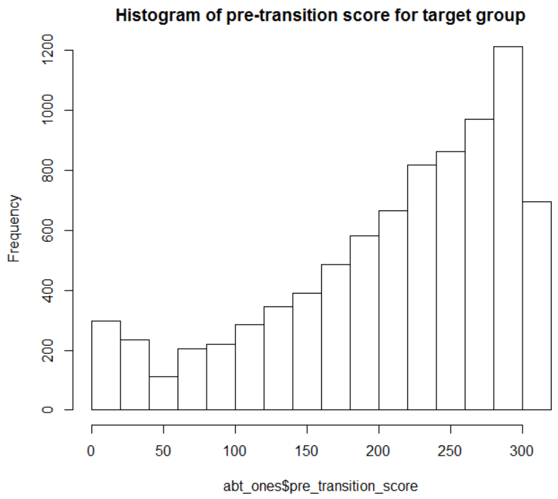
```
Min.     1st Qu.   Median   Mean      3rd Qu.   Max.
0.2423   160.3492 229.4245 208.7078 276.3008 313.2896
```

change of score for the year when transitioning:

```
 Min.    1st Qu.   Median   Mean     3rd Qu.  Max.
 -24.62  61.06     134.60    144.77   215.53    631.25
```

Some nominals have a score decrease that has led to a change in groups (transition to higher group). These are due to the fact that the threshold for the transition is relative and depends on the distribution of the score for a given year. In most cases however, a nominals' score increased, with an average increase of 145 points.





As we can see, the target and non-target groups have different distributions of pre-transition scores (pre-transition score for non-transitioning nominals means the score from 2017). The majority of the target group have relatively high scores prior to transition, which means they have high crime frequency, harm or drug usage history (or all of these combined). The score itself could therefore be a good predictor, but due to non-interpretability and very high correlation with the target variable it will be excluded as a feature.

Other Features:

- Age:
  - Target group (1's)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.00	22.00	28.00	29.55	35.00	73.00	1

  - non-HHO (0's)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
-1.0	29.0	37.0	39.6	48.0	113.0	6024

With some of these demographic variables especially, it can be seen that there are some data quality issues, however, in the data available, the proportion of nominals with an age less than 5 was 0.016% and with an age less than 11 was 0.039%. With no means of ascertaining their real age and given the need to keep age as a quantitative feature coupled with the number of other features available to the model, these data quality issues are not considered material to the performance of the modelling.

Looking in more detail into crime prolificacy (measured by crime committed count, total harm, and days since last crime) and drug history in both groups, it can be seen that the target group generally has higher values (on average) than the non-target group with the exception of recency:

- Total number of committed crimes:
    - Target group (1's)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	5	11	17	21	349

-
    - non-HHO (0's)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	1.000	1.000	2.295	2.000	309.000

- Total harm (as measured by CCHI of crimes committed; note that this is the raw CCHI and not the RFSDi):
    - Target group (1's)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	124.4	594.3	1175.1	1674.3	39569.5

-
    - non-HHO (0's)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.00	2.00	6.00	95.95	18.75	31955.50	3673

- Crime recency (measured as days since last crime committed)
    - Target group (1's)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.0	260.8	728.0	1053.1	1538.0	6309.0

-
    - non-HHO (0's)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
1	1983	3796	3768	5626	39233	7036

- Drug addiction (measured as total number of positive drug tests in DIP data):
    - Target group (1's)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.000	0.000	1.000	1.801	2.000	47.000	1929

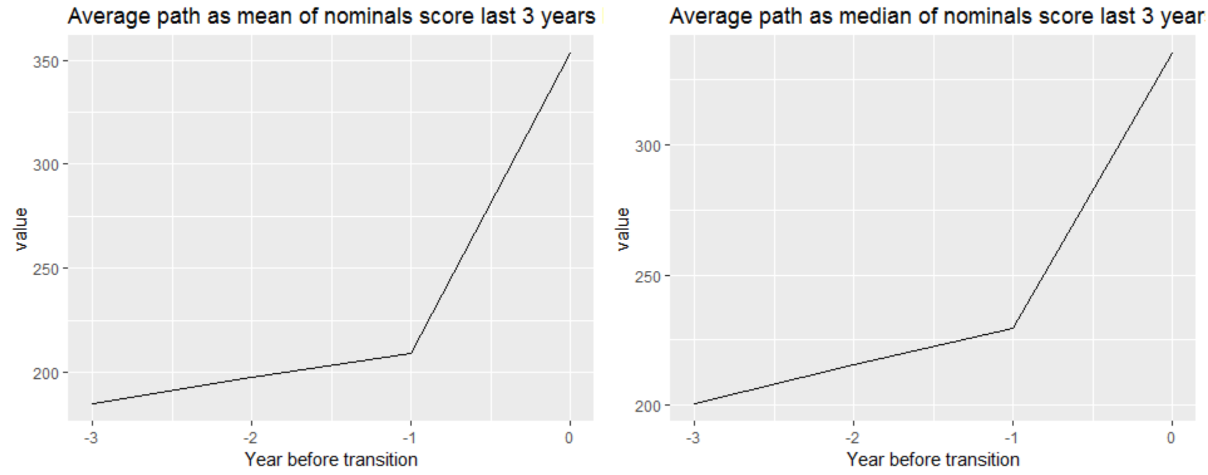
-
    - non-HHO (0's)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.0	0.0	0.0	0.5	1.0	41.0	395354

All of these variables appear to have completely different distributions in the target and non-target groups. They are likely to be good predictors for modelling.

RFDSi score path for last years before transition

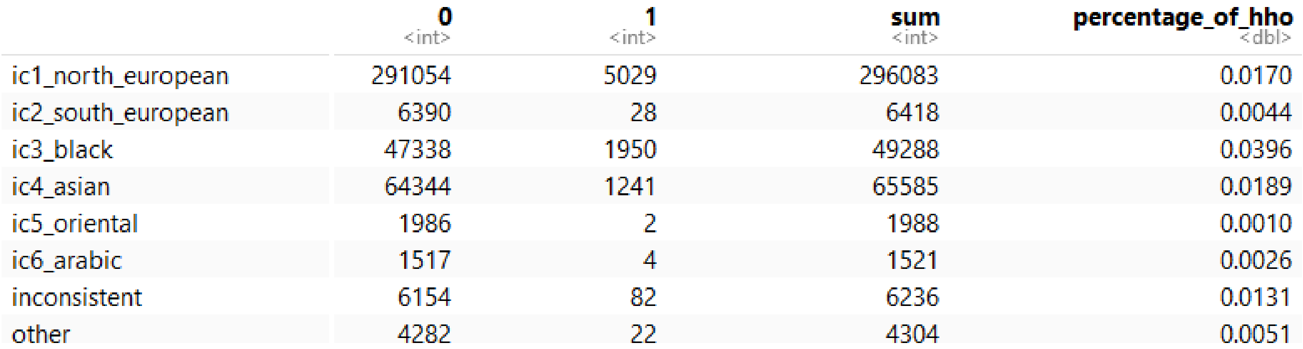
As can be seen from the charts below the mean and median score for transitioning nominals is on the upward trend for the last 3 years before transition. However, the largest spike (measured as year-to-year difference) is visible in the transition period. This is due to the fact that some nominals transition from relatively low scores, thus skewing averages over previous years.



Analysis of potentially sensitive attributes – ethnicity

As part of the EDA process for the IOM model, we have investigated the presence of differentiation between (ethnic) groups resulting from the target variable (i.e. whether there are different degrees of representation of different ethnic groups within the target group of the high harm offenders).

Within the data, the following probabilities were found:



	0 <int>	1 <int>	sum <int>	percentage_of_hho <dbl>
ic1_north_european	291054	5029	296083	0.0170
ic2_south_european	6390	28	6418	0.0044
ic3_black	47338	1950	49288	0.0396
ic4_asian	64344	1241	65585	0.0189
ic5_oriental	1986	2	1988	0.0010
ic6_arabic	1517	4	1521	0.0026
inconsistent	6154	82	6236	0.0131
other	4282	22	4304	0.0051

The degree of differentiation has been measured as:

$$P(y = 1 \mid S = a) - P(y = 1 \mid S = b)$$

Where:

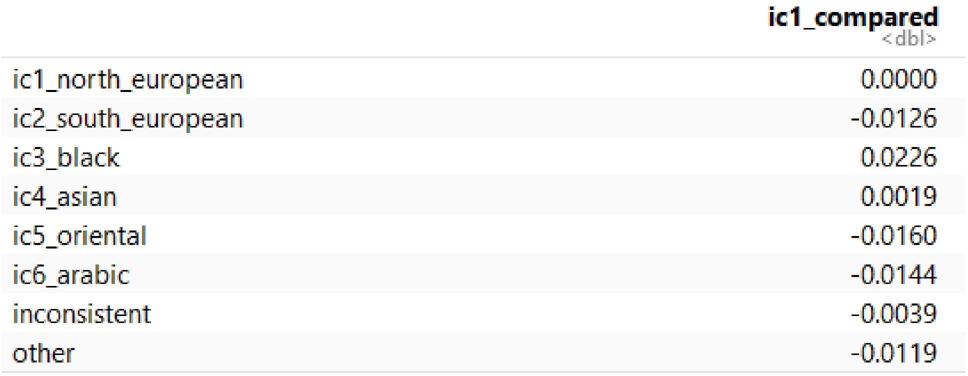
P = probability (as present within the data)

y = the target state of interest

S = a potentially sensitive attribute (in this case ethnicity)

a and b = particular states of the sensitive attribute

If the reference group is taken as IC1 (north European), the above differentiation measure would lead to:



	ic1_compared <dbl>
ic1_north_european	0.0000
ic2_south_european	-0.0126
ic3_black	0.0226
ic4_asian	0.0019
ic5_oriental	-0.0160
ic6_arabic	-0.0144
inconsistent	-0.0039
other	-0.0119

It can be seen therefore that the majority of groups, when compared to IC1, have very little difference or are represented to a lower degree than the reference group. The exception is for IC3. From this initial analysis, it could be ascertained that a person in the IC3 group is 2.4 times more likely to be in the high harm group than a person in the IC1 group (relative risk of 2.33 and an odds ratio of 2.38).

Are the differences large?

Whilst the majority of the differences are small, there will always likely be differences and therefore the question arises as to whether the differences are meaningful.

This was tested for by way of applying chi-square tests (and differences in proportions tests), however due to the large numbers involved, almost any difference would result in a finding of very strong statistical evidence in favour of the finding (a p-value of less than 0.01).

- Pearson's Chi-squared test with Yates' continuity correction

  ```
  data:  2 ethnicities: ic1_north_european and ic3_black
  X-squared = 1086.8, df = 1, p-value < 2.2e-16
  ```

- 4-sample test for equality of proportions without continuity correction

  ```
  data:  ethnicities: ic1_north_european, ic2_south_european, ic3_black, ic4_asian
   X-squared = 1196.5, df = 3, p-value < 2.2e-16
  alternative hypothesis: two.sided
  sample estimates:
     prop 1     prop 2     prop 3     prop 4
  0.01698510 0.00436273 0.03956338 0.01892201
  ```

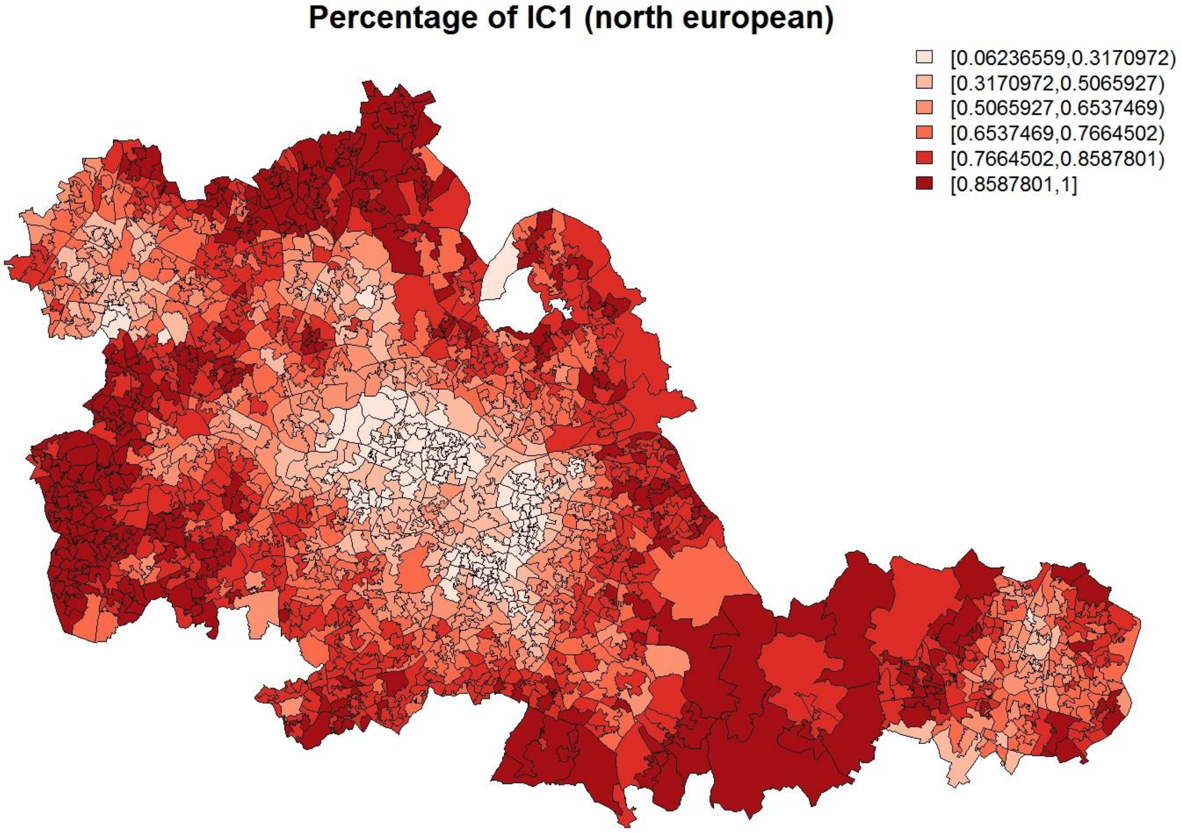
For this reason, the data were randomly sampled to provide tests with no more than 1000 in each state of the target variable (i.e. high vs not high group) for the main groupings (IC1 – IC6). A thousand such tests were run.

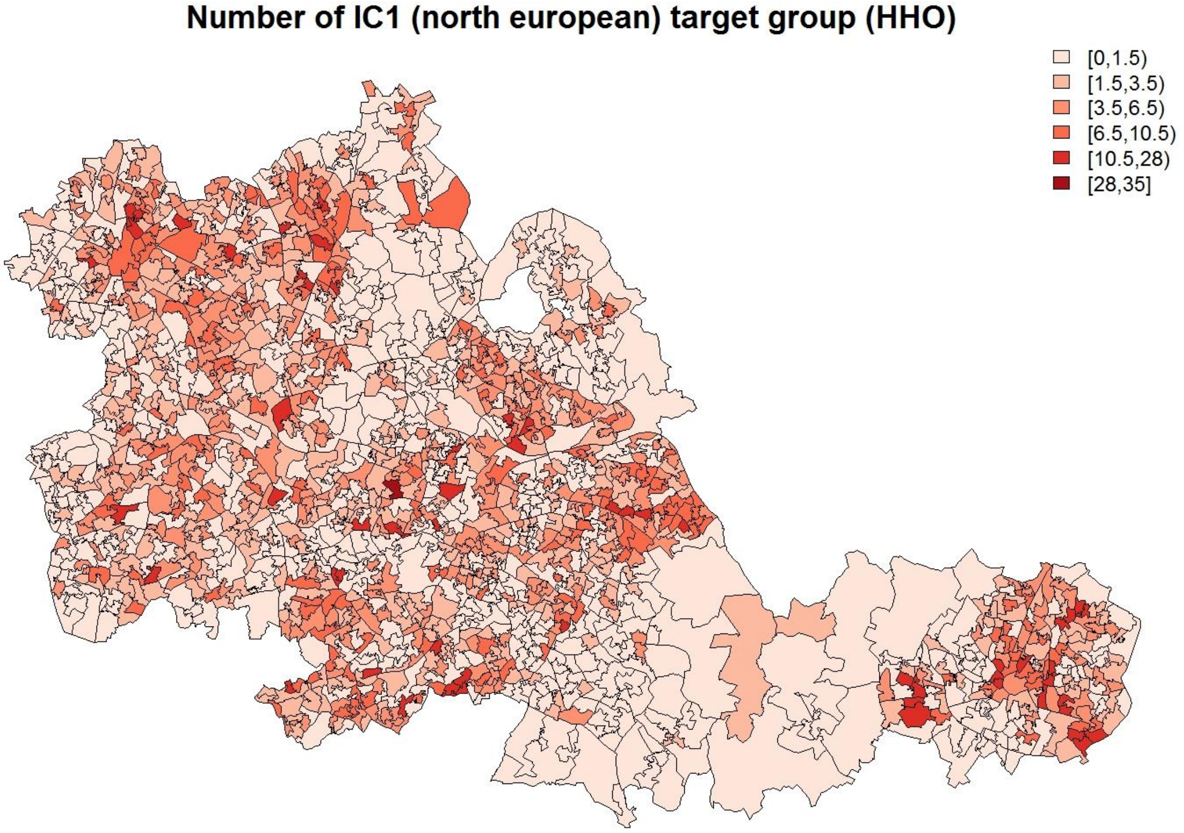
Using this approach, we both adjusted the resulting p-values (using the false discovery rate) and examined the p-values via the local false discovery rate. The first approach resulted in 87.2% of the tests having adjusted p-values lower than 0.05 and the second approach having 99.8% of the tests resulting in an (l)fdr of less than or equal to 0.2.

These findings would suggest that the differences found were meaningful and therefore, should the IC grouping be found to have predictive capacity in the final model, this feature would be taken into account. In the end, ethnicity did not make it through to the final list of features and therefore was not included in the final model. See appendix C for further details of the modelling and the features used in the final model.

Ethnicities: maps of offending population, target group and target group clusters by population

It is also informative to examine the spatial distribution of ethnicities and the high harm group:

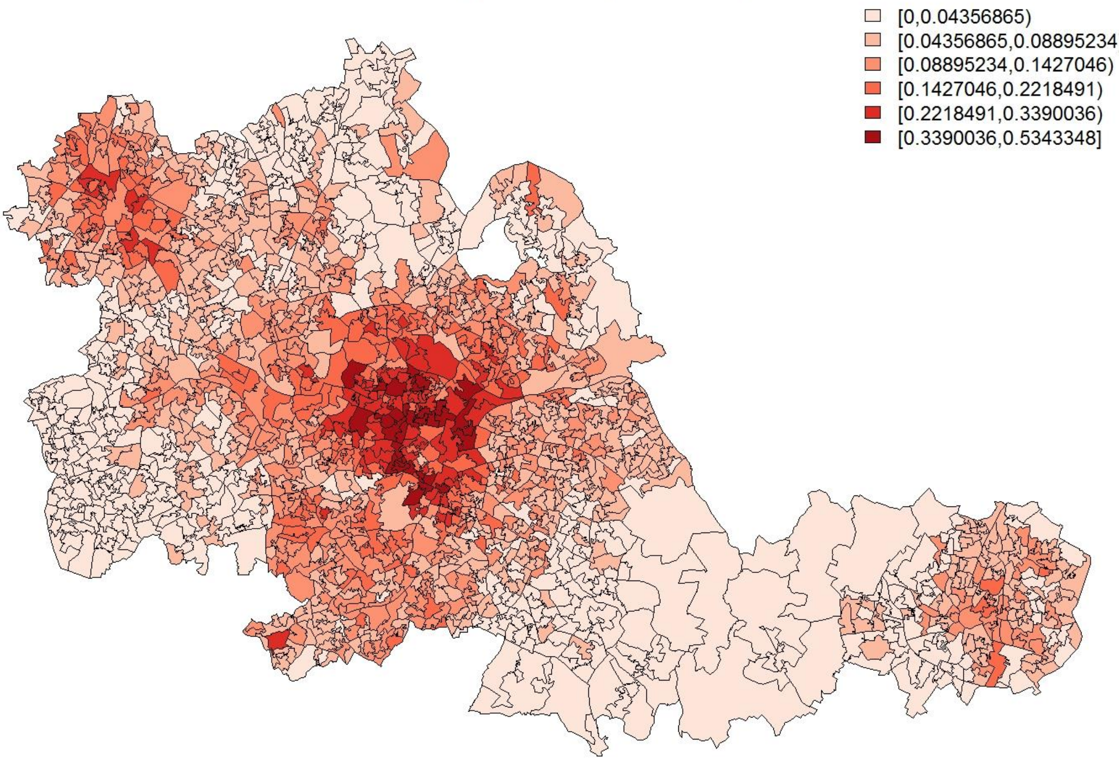




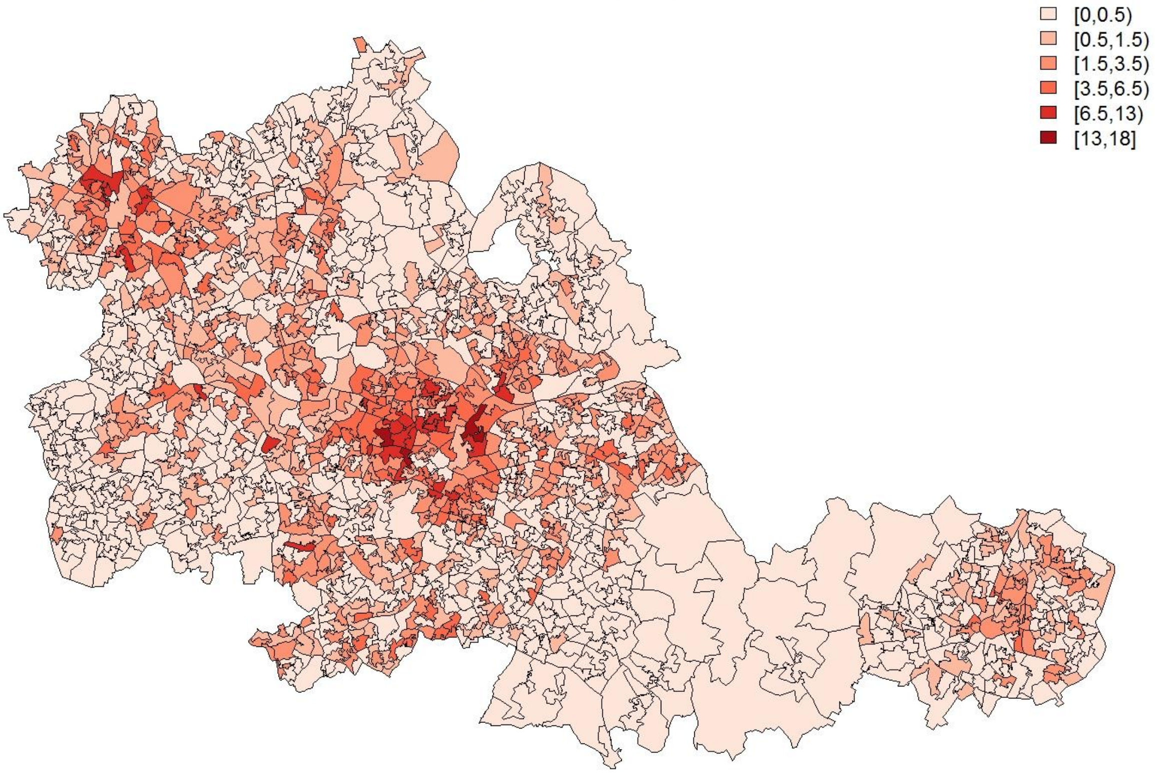
In terms of the population, the higher proportions of IC1 are within the more outerlying areas of the consituent districts and lowest within the central (city) areas. In terms of the count of those within the high harm group, this is somewhat more spatially dispersed, but tend to be higher in more central locations.

In the case of IC3, the spatial distribution of the (relative) population is different from that for IC1 in being more centralised. Counts within the high harm group follow a similar pattern compared to IC1 with central areas showing the highest concentration:

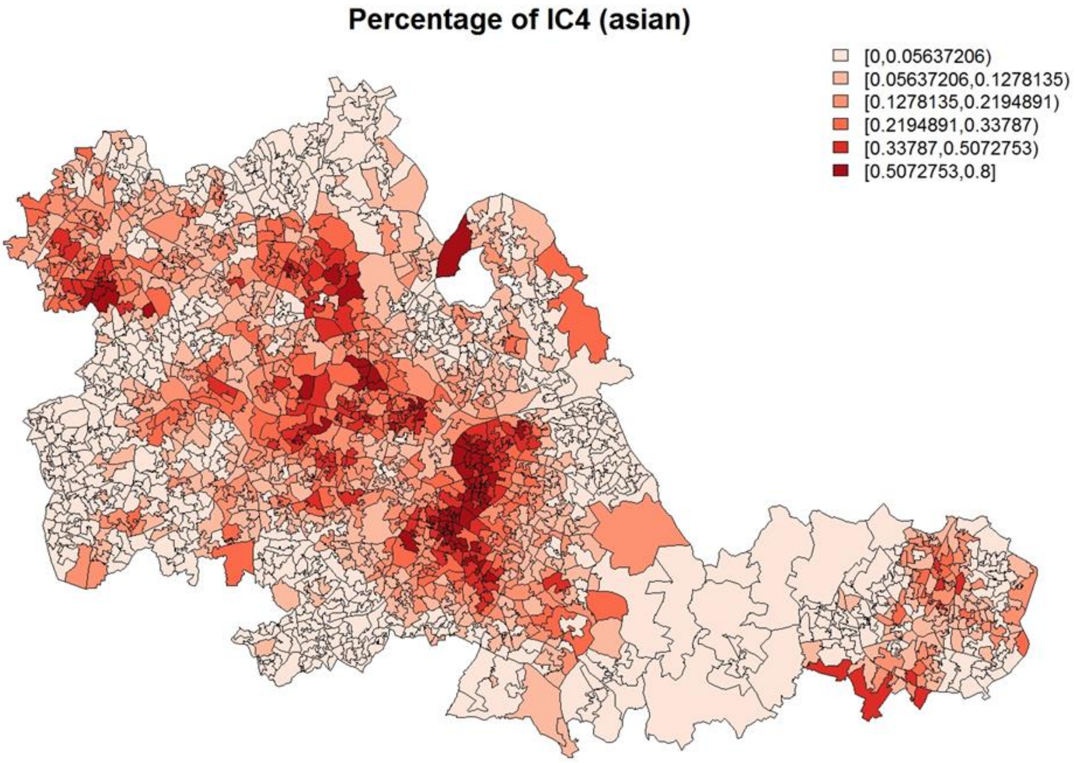
Percentage of IC3 (IC3 black)

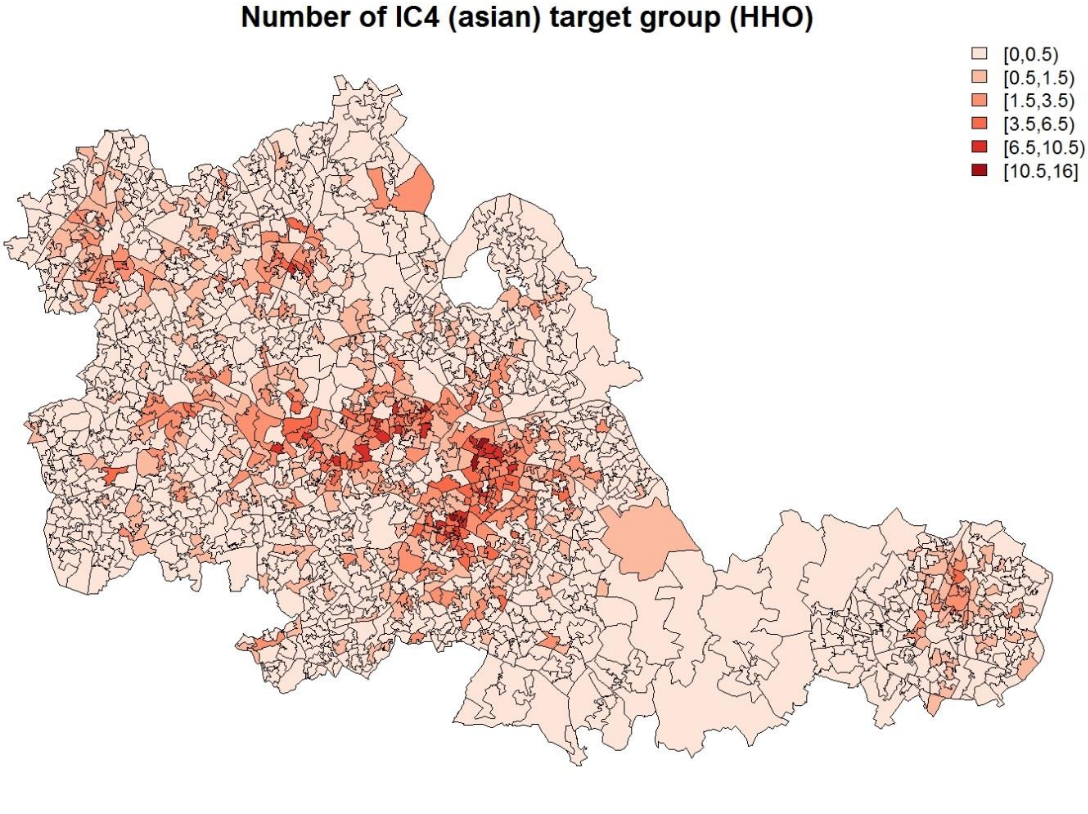


Number of IC3 (black) target group (HHO)



In the case of IC4, again compared to IC1, there is a greater concentration of the relative population in central areas and of counts within the high harm group also being in central areas, but to a lesser degree than is the case for IC3.

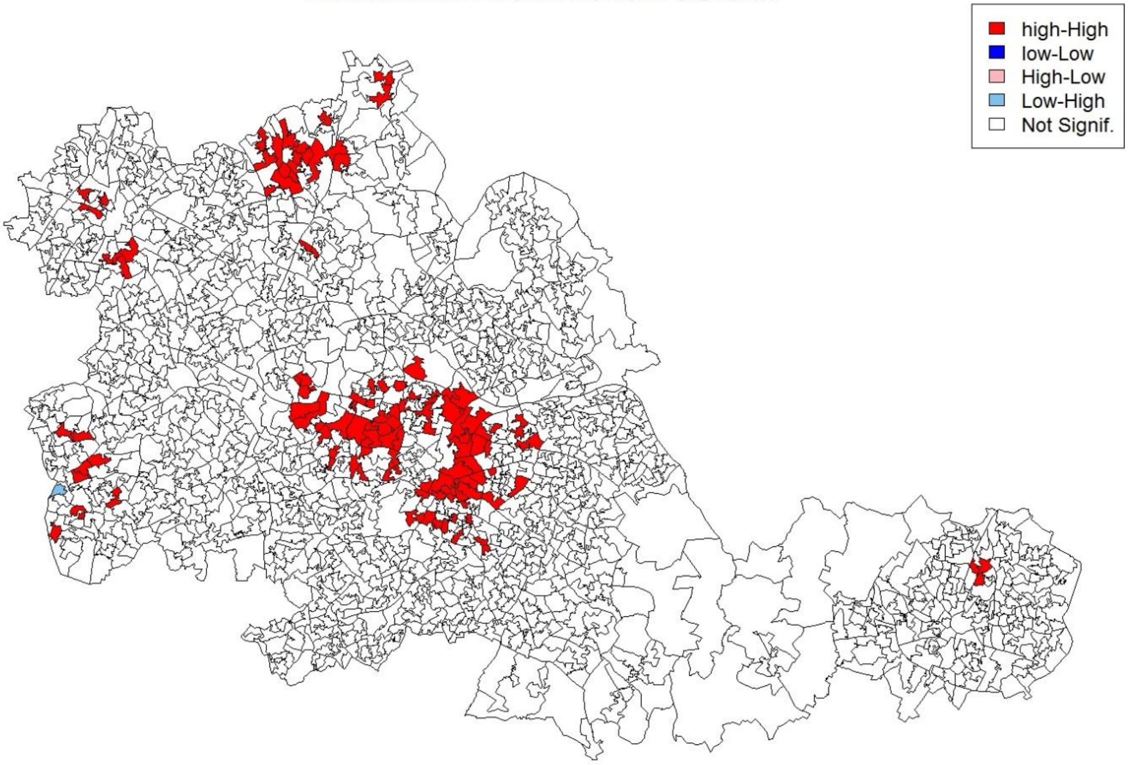




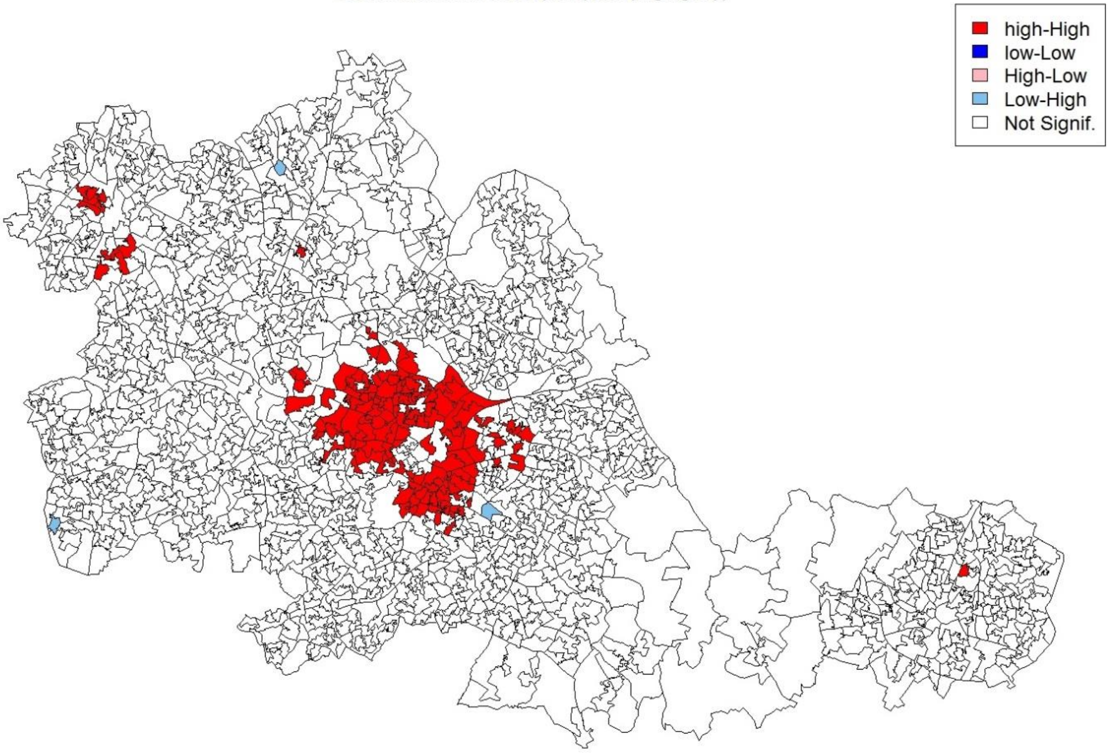
It is also of interest to ascertain if there is a degree of spatial clustering of the target group. This has been assessed by way of grouping the (adjusted) p-values resulting from localised Moran's I tests.

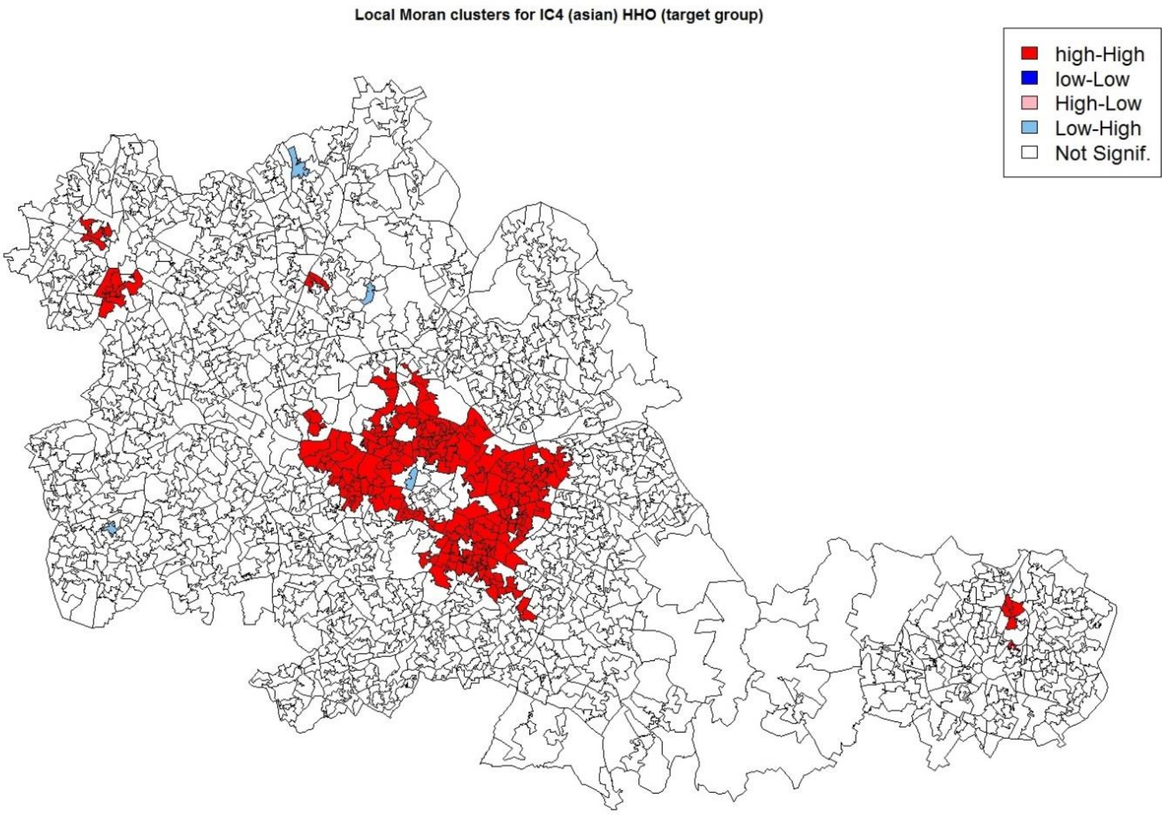
From the maps below, it can be seen that for IC1, IC3 and IC4 there is apparent spatial clustering of the target group in central areas, particularly in central Birmingham. There is a greater degree of clustering of the IC1 target group on the west and north sides of the force area compared to IC3 or IC4 and all the groups have some degree of clustering in the central Coventry area.

Local Moran clusters for IC1 (north european) HHO (target group)

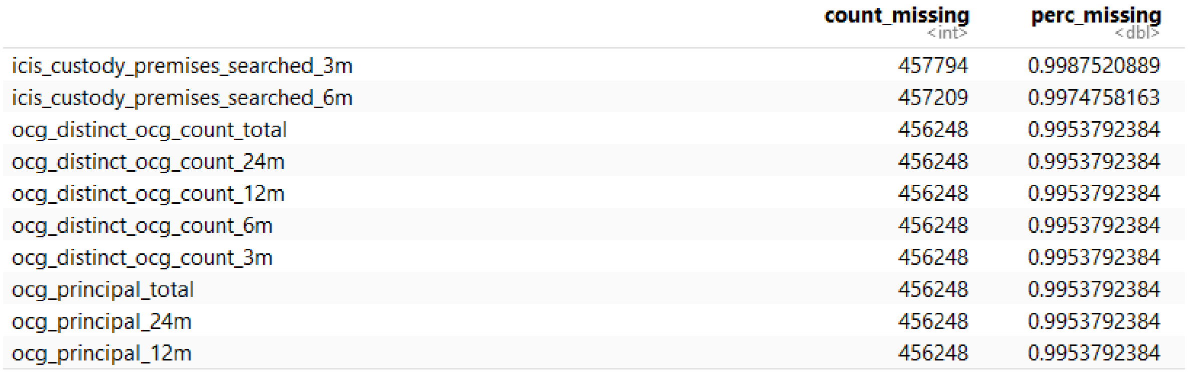


Local Moran clusters for IC3 (black) HHO (target group)





Missing values - types of missing values and ABT aggregation



	count_missing	perc_missing
	<int>	<dbl>
icis_custody_premises_searched_3m	457794	0.9987520889
icis_custody_premises_searched_6m	457209	0.9974758163
ocg_distinct_ocg_count_total	456248	0.9953792384
ocg_distinct_ocg_count_24m	456248	0.9953792384
ocg_distinct_ocg_count_12m	456248	0.9953792384
ocg_distinct_ocg_count_6m	456248	0.9953792384
ocg_distinct_ocg_count_3m	456248	0.9953792384
ocg_principal_total	456248	0.9953792384
ocg_principal_24m	456248	0.9953792384
ocg_principal_12m	456248	0.9953792384

Due to the methods by which the ABT was created most of the aggregating variables' missing values are not missing at random. They occur when a nominal has no history in a certain system at the time of ABT creation, thus resulting in no available data to create aggregates. This is especially common in systems that cover a small subset of the general offending population (like OCG or compact) where most nominals will have no history (eg. no data to calculate if a nominal is not a member of an OCG). The proposed way of filling in these missing variables is a constant value imputation of 0 (zero) - as this is the business meaning of these values. This will be applied to all aggregating variables.

A similar situation occurred with eigenvector_centrality and page_rank variables derived from social network analysis (SNA – note that this analysis is undertaken over WMP data only and does not include analysis of social media data), which were not calculated for standalone nodes (vertices with no neighbours). These will have an imputed constant value of 0 as well.

Other variables with missing values and proposed ways of imputation include:

- crimes_solo_min_age_committed, crimes_days_since_last_solo_committed (14% and 12% missing) - when no solo crimes were committed (and age of nominal was not known in case of crimes_solo_min_age_committed). These variables will be binned and treated as categorical with missing as a separate category.
- cluster_partition - 11% missing - as this is a categorical variable missing values will be assigned a separate category. Note that this "cluster_partition" relates to geographic areas that have been created by way of undertaking k-means, h-clustering using Ward distances and distance based clustering over socio-economic data (from the Census 2011) at the LSOA level. This produces 13 geographic areas whereby the socio-economic characteristics are more similar within them than between them (note that this feature was not included in the final model).
- ethnicity - we assume that data is missing at random (no independent source of ethnicity was available) and will be assigned the 'other' category (note that this feature was not included in the final model).
- ons_harm and _cambridge_harm_ variables (from 0.0001% to 4% missing) have missing values when offences for a nominal were not assigned a harm score. These will be imputed with constant value of 0.

As expected, GID, abt_date and target_variable have no missing values, so all observations can be used for modelling and scoring.

Multicollinearity

Due to the number of aggregate variables calculated, we can expect very high linearity (simple additivity in many cases) and collinearity within the same group of variables (eg. crimes_total_sum, crimes_24m_sum, crimes_12m_sum, crimes_6_m_sum etc).

A number of variables with very high correlation were diagnosed (using the variance inflation factor). It included variables with very low variability (mostly XXX_3m and XXXX_6m aggregate variables, which calculate a nominal's history in the last 3 and 6 months), as well as groups of variables which due to the calculation methodology brought the same results (SNA-related).

The following variables have therefore been removed from further analysis due to >.99% correlation:

- ocg_significant_3m
- sna_ocg_member_neighbors_sum_all

- sna_ims_drug_addict_neighbors_sum_all
- sna_ims_drug_dealrs_neighbors_sum_all
- sna_dip_drug_addict_neighbors_sum_all
- sna_crimes_2_drug_offences_neighbors_sum_all
- compact_avg_time_missing_in_days_total

A number of variables with perfect multicollinearity were diagnosed (not displayed in the table). Most of these are expected and have a property of perfect additivity (eg. sum of crimes for all categories separately = sum of all crimes).

They have not been removed as their collinearity will not impact model quality for the selected models.

Initial model for variable selection

The ABT was split into train (70%) and test (30%) datasets using stratified sampling over the target variable.

Both train and test datasets for non-HHO groups (0's) were split into 10 folds (separate subsamples). These subsamples were then iterated over and a separate XGBoost model was created for each of them, sampling zero's if necessary to keep the dataset balanced). All variables were included in a model as features. The importance of the variables was later averaged over these 10 models and presented below:

Top 30 most important variables as sorted by "Gain" measure include (however see appendix C):

Importance order	Variable name	Gain	Cover	Frequency
1	crimes_days_since_last_crime_committed	0.611	0.162	0.065
2	crimes_days_since_last_solo_committed	0.161	0.047	0.023
3	crimes_cambridge_harm_total	0.045	0.073	0.047
4	page_rank	0.025	0.022	0.030
5	eigenvector_centrality	0.022	0.045	0.033
6	crimes_committed_total	0.022	0.055	0.023
7	crimes_ons_harm_total	0.016	0.031	0.023
8	icis_custody_hours_total	0.012	0.024	0.021
9	crimes_days_since_last_coof_committed	0.010	0.022	0.024
10	crimes_ons_harm_24m	0.003	0.011	0.010
11	crimes_cambridge_harm_24m	0.003	0.014	0.013
12	crimes_violent_total	0.003	0.004	0.011
13	topic10_avg_value	0.003	0.020	0.015
14	crimes_ons_harm_12m	0.002	0.008	0.005
15	crimes_committed_24m	0.002	0.002	0.007

16	solo_crimes_committed_total	0.002	0.014	0.019
17	icis_custody_offences_other_records_total	0.002	0.015	0.009
18	nominals_age	0.002	0.003	0.019
19	icis_custody_cust_offences_records_total	0.002	0.001	0.006
20	topic5_max_value	0.002	0.003	0.016
21	topic3_avg_value	0.002	0.007	0.018
22	crimes_coof_min_age_committed_factornot_committed_coof	0.001	0.001	0.004
23	topic2_max_value	0.001	0.008	0.010
24	icis_custody_hours_24m	0.001	0.005	0.012
25	topic6_avg_value	0.001	0.010	0.018
26	topic5_entries_cnt	0.001	0.017	0.005
27	topic5_avg_value	0.001	0.007	0.016
28	topic7_max_value	0.001	0.018	0.009
29	topic8_avg_value	0.001	0.005	0.015
30	icis_custody_records_total	0.001	0.003	0.009

The top 10 variables focus on capturing the recency of crime and harm / number of crimes committed, which are a part of the target variable definition. These were important most likely because most HHOs have already been high (in terms of score, possibly medium-high in terms of group assigned) the year before transition, but didn't reach the threshold to classify them as "High". Moreover, zeros are randomly sampled from the whole offending population where the majority of offenders were not prolific/ harmful. Therefore, the model differentiates active and prolific offenders (high score, high harm, many crimes, recent crimes) from non-prolific (no recent crimes, not active, low score).

The above is further confirmed by a high AUC Statistic (0.999 on the training set and over 0.99 on the test set), regardless of the sample of zeros used for model training or test (as long as zeros are randomly sampled from the general offending population) or the imbalance of the set. By using the above model we could expect, that all nominals with high recency and harm (as measured by number of crimes / CCHI) would be scored as high, regardless of their propensity to transition to the high harm group. Potentially therefore the model would not be able to differentiate active and prolific offenders that have not transitioned from active and prolific offenders who did transition.

This hypothesis will be tested by creating non-random samples:

a) with distribution of crime harm (crimes_cambridge_harm_total variable) similar to the distribution of "ones" in the target group

AUC: 0.9885

Confusion Matrix and Statistics

```
                Reference
Prediction      0       1
```

```
         0                  2223      23
         1                  277       2492

           Accuracy : 0.9402
           Sensitivity : 0.9909
           Specificity : 0.8892
```

b) with distribution of crime recency (crimes_days_since_last_crime_committed variable) similar to the distribution of "ones" in the target group.

```
AUC: 0.9930
Confusion Matrix and Statistics

                            Reference
Prediction                  0         1
         0                  2509      23
         1                  241       2492

           Accuracy : 0.9499
           Sensitivity : 0.9909
           Specificity : 0.9124
```

c) with distribution of RFSDi scores in 2017 similar to distribution of RFSDi score the previous year before transition. Basic model fit statistics

```
AUC: 0.9517

                            Reference
Prediction                  0         1
         0                  1345      23
         1                  1155      2492

           Accuracy : 0.7651
           Sensitivity : 0.9909
           Specificity : 0.5380
```

All of the above predictions and model quality measures indicate that the model does not differentiate well between prolific offenders who transition from those that do not transition. Especially the third model indicates that the above model has very high sensitivity, but low specificity, misclassifying 0’s with high pre-transition scores (resulting in high a False Positive Rate). Despite a high AUC, accuracy had decreased largely, making it less useful to productionize than the initial model measures indicated.

Therefore, the variable importance presented above may not be a good estimate of real variable importance in the final model.

To resolve this problem a number of different approaches were proposed:

- to build a model on a highly unbalanced dataset

- by sampling more (in terms of absolute values) nominals as zeros, who had high pre-transition score (prolific and harmful) and allow the model to better differentiate between 0's and ones.
- To create separate models for less active and more active nominals in the pre-transition period.

These approaches will be further developed, tested and assessed in the model building phase. See appendix C below.

Appendix C

Model Building and Selection

This appendix explains the process of model building and selection for the IOM model and presents the findings.

The overall purpose of the model is to predict whether a nominal will become a High Harm Offender (the target group) in a year from the moment they are scored. In order to account for transition into a target group, criminal behaviour paths for the years 2010 - 2018 were examined. Each year, every nominal's RFSDi score was calculated and depending on the value they were classified into one of the following groups:

- Not known - when nominal had no crime history prior to this year, and their score was 0.
- Low, low-medium, medium-high - all classified as 0's
- High, Super High

Target group definitions and analysis

Firstly, we check the number of zeros and ones (target group) in the final full analytics base table (ABT) table and the "ABT date" variable (date for which the ABT was calculated/used) for each observation. In order to maximise the number of observations falling within the target group (the 1's) those who transitioned over the period 2013 – 2018 were included as opposed, for example, just taking those who transitioned within a 2 year period starting in 2016. Because of the different periods of transition therefore, the features for each observation (nominal) are derived from data for periods prior to their transition and therefore cover different years. We aggregate these into a frequency matrix.

target variable	frequency	Percentage
0	449982	98.2%
1	8384	1.8%

target variable	ABT date	frequency
0	08/05/2017	449982
1	08/05/2012	3060
1	08/05/2013	1321
1	08/05/2014	1301
1	08/05/2015	954
1	08/05/2016	1100
1	08/05/2017	648

The dataset created contains 458,366 observations. It is an Analytical Base Table with one distinct nominal per row, and each nominal is classified into one of 2 groups:

- Ones (High harm offenders - HHO) – 8,384 nominals: all nominals who transitioned into HHO in years 2013 - 2018 were selected. A "transition" occurs when a nominal changes their RFSDi group from any of the (low, low-medium, medium-high) harm groups in a previous year into any of the (High, Super High) harm groups in the current year. If many transitions happen during the observation period the first one is taken. Transitions into HHO were calculated for the years 2010 - 2018, but only years 2013 - 2018 were considered for an observation period, resulting in ABT dates from years 2012 - 2017 (offenders' history is calculated one year before transition). The main reason for this year range is data availability - we wanted to ensure we have the possibility of at least 2 years of history of a nominal from every system.

- Zeros (Non -HHO) – 449,982 nominals: all nominals who did not transition into HHO (please look above), and were never classified as high or super high were selected.

Due to the above classification, some nominals were not selected either as 0 or 1. This includes nominals who were unknown throughout the whole period and committed their first offence in the year 2017/2018, or those who follow the path of "Not known" -> {High, Super High} (as their history one year before transition is not known).

Modelling approach

Different approaches have been tested to account for target variable class imbalance, model selection and feature selection. Due to different approaches not all of the models are directly comparable for all of the statistics. These will be explained in more detail over the next pages.

The general approach was to divide the whole dataset into two exclusive sets, retaining the proportions of the target variable in each set:

- Full training set, consisting of 70% of all observations,
- Full testing set, consisting of 30% of observations.

The full training set consists of 320 857 rows, out of which 5869 were 1's and the full test set consists of 137 509 rows, out of which 2,515 were 1's.

Class imbalance approach

Most modelling techniques do not handle imbalanced classes well. The following approaches were taken to resolve the problem of class imbalance:

- **Approach 1 (A1) - random sampling of zeroes to balance classes.**

10 different exclusive folds were randomly sampled for the non-target group for both training and test sets. Each fold consisted of approximately 31,500 'zeroes' in the train set and 13,500 in the test set. For each fold out of the train set 5,869 observations were randomly sampled and 2,515 observations were sampled for each fold of the test data. Samples from different folds were used for model training to test for variable importance robustness.

- **Approach 2 (A2) - RFSDi score based sampling**

Non-random sample of zeroes (the non-target group) were drawn based on the pre-transition score of ones (target group). The target group's pre-transition score was divided into 10 bins based on deciles, and for each bin 500 observations of zeroes (with similar scores that fall into the same bins) were drawn (or less, if fewer observations were available – especially in the top deciles of the score). The purpose of this sampling was to prepare a dataset, which would:

a) Mimic a possible scoring scenario where only more active and prolific offenders will be scored.
b) Allow the algorithm to differentiate prolific offenders who transition from prolific offenders that do not transition.

- **Approach 3 (A3)– Imbalanced dataset with no changes**

Full training set was used for modelling and full testing set for assessment. Due to large class imbalance this training set will be used with an XGBoost model only, as the other models cannot handle imbalanced training datasets as well.

- **Approach 4 (A4)– Pre-transition score>50 sample**

For this approach we include nominals with scores higher than 50 only - both for the training and test datasets, both zeroes and ones. There are 2 main reasons for this approach:

a) Including only nominals with scores above a certain threshold would exclude those who are generally non-active and non-prolific which is likely preferable.
b) Limiting the number of zeroes in the dataset – as most of the offenders are characterised by a low score.

The consequence of this approach is the following:

- Training set: the number of zeroes has decreased from 320, 857 to 50, 447 (84,3% decrease) and the number of ones has decreased from 5, 869 to 5, 458 (7% decrease)
- Testing set: the number of zeroes has decreased from 137, 509 to 21, 670 (84,2% decrease) and the number of ones has decreased from 2, 515 to 2, 341 (7% decrease)

Summarising, this approach decreases class imbalance based on business logic by largely decreasing the number of zeroes and minimizing the decrease of ones.

Test sets:

Three different test sets have been created for model assessment:

A. Non-random balanced testing set as described in approach 2 (A2) **(balanced)**,
B. Complete, highly imbalanced test set, as described in approach 3 **(full)**,
C. Reduced testing set with pre-transition score higher than 50, as described in approach 4 (**PTS>50)**

A random balanced test set as described in approach 1 (A1) was initially used for testing, but it did not prove to be useful for scoring. As an aim of the model is to differentiate people with very low crime prolificacy from those that are active and prolific it did not provide a good estimate of model quality for transition probability prediction. Therefore, it will not be included in the final results and will not be used to evaluate model quality.

Models trained

R software was used to train the models, score the data and calculate model performance. The following 3 types of models were trained in different configurations:

- **XGBoost** model, trained using grid search over the parameters.
- **Random Forest** model, trained using grid search over the number of trees.
- **SVM** model, trained using linear and radial kernals.

Model evaluation measures

Each model was assessed using the following model quality measures:

- Area Under Curve (AUC) for the training set and test sets A, B and C
- ROC Curve
- Prediction accuracy for the training set and test sets A, B and C
- The following statistics for the training set and test sets A, B and C with a prediction classification cutoff = 0.5:
  - Sensitivity,
  - Specificity,
  - Precision,
  - Confusion Matrix
- Lift chart and lift value for top 10% of predictions
- Kolmogorov Smirnov model fit test and chart.

Model descriptions are available below. A1 – A4 indicate the dataset used for training as described in the "Class imbalance approach" section above. Models are listed in training sequence and all changes in approach are the consequence of the analysis of previous models and their results.

- **model 1 (A2)**- XGBoost model trained on all variables, trained on score-based balanced sampled dataset, algorithm specific parameters: eta = 0.2, max.depth=10, colsample_bytree=0.7, nrounds=50
- **model 2 (A2)** – XGBoost model trained on all variables, trained on score-based balanced sampled dataset, algorithm specific parameters: eta = 0.2, max.depth=7, colsample_bytree=0.5 , nrounds=50
- **model 3 (A2)** – XGBoost model trained on all variables, trained on score-based, balanced sampled dataset, algorithm specific parameters: eta = 0.1, max.depth=10, colsample_bytree=0.7, early stopping based on test set, nrounds=80

- **model 4 (A3)** – XGBoost model trained on all variables, trained on full, highly imbalanced dataset, algorithm specific parameters: eta = 0.3, max.depth=10, colsample_bytree=0.7, early stopping based on test set, nrounds=80

- **model 5 (A4)** – XGBoost model trained on all variables, trained on pre_transition_score>50 imbalanced dataset, algorithm specific parameters: eta = 0.3, max.depth=10, colsample_bytree=0.7, early stopping based on test set, nrounds=80. 50 most important variables as measured by highest "Gain statistic" from this model were selected for Simplified versions of models (and termed 'Simpl.') – Please look at results summary below for further information.

- **model 6 (A2)** – Random Forest model trained on all variables, trained on score-based balanced sampled dataset, algorithm specific parameters: 50 trees,

- **model 7 (A4)** – Random Forest model trained on all variables, trained on pre_transition_score>50 imbalanced dataset, algorithm specific parameters: 50 trees, nodesize=8,

- **model 8 (A4, Simpl.)** – Random Forest Simplified, trained on pre_transition_score>50 imbalanced dataset with top 50 most important variables, algorithm specific parameters: 50 trees, nodesize=8,

- **model 9 (A4, Simpl.)** – Random Forest Simplified, trained on pre_transition_score>50 imbalanced dataset with top 50 most important variables, algorithm specific parameters: 200 trees,

- **model 10 (A4, Simpl.)** – Random Forest Simplified, trained on pre_transition_score>50 imbalanced dataset with top 50 most important variables, algorithm specific parameters: ntree=50, nodesize=10

- **model 11 (A2, Simpl.)** – SVM Linear model, trained on score-based balanced sampled dataset with top 50 most important variables, data centered and scaled as preprocessing step,

- **model 12 (A4, Simpl.)** – SVM Linear model, trained on pre_transition_score>50 imbalanced dataset with top 50 most important variables, data centered and scaled as preprocessing step,

- **model 13 (A2, Simpl.)** – SVM Radial model, trained on score-based balanced sampled dataset with top 50 most important variables, data centered and scaled as preprocessing step,

- **model 14 (A4, Simpl.)** – SVM Radial model, trained on pre_transition_score>50 imbalanced dataset with top 50 most important variables, data centered and scaled as preprocessing step,

- **model 15 (A3, Simpl.) –** XGBoost model, trained on full, highly imbalanced dataset with top 50 most important variables, algorithm specific parameters: eta = 0.3, max.depth=10, colsample_bytree=0.7, nrounds=80

- **model 16 (A4, Simpl.) –** XGBoost model, trained on pre_transition_score>50 imbalanced dataset with top 50 most important variables, algorithm specific parameters: eta = 0.3, max.depth=10, colsample_bytree=0.7, nrounds=80

- **model 17 (A4, Simpl.) –** 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

Model results are summarised in the table below. In total over 20 models were trained and tested. Only selected models and measures have been displayed to ensure readability and complete model performance assessment will be displayed only for the final model. Models trained on Approach 1 dataset (A1 - random sampling of zeroes to balance classes) did not explain the phenomenon well (substantially lower AUC, accuracy, sensitivity etc.) so they were removed from the table.

Model	Number of features	AUC train	AUC test balanced (A)	AUC test full (B)	AUC test PTS>50 (C)	Sensitivity train	Sensitivity test balanced (A)	Sensitivity test full (B)	Sensitivity test PTS>50 (C)	Specificity train	Specificity test balanced (A)	Specificity test full (B)	Specificity test PTS>50 (C)
model 1 (A2)	499	1.000	0.953	0.978	0.954	0.998	0.814	0.814	0.815	1.000	0.945	0.975	0.950
model 2 (A2)	499	0.997	0.954	0.979	0.956	0.938	0.822	0.822	0.824	0.997	0.942	0.974	0.947
model 3 (A2)	499	0.999	0.950	0.979	0.955	0.944	0.803	0.803	0.806	0.999	0.954	0.980	0.963
model 4 (A3)	499	0.999	0.940	0.993	0.975	0.964	0.730	0.730	0.753	1.000	0.979	0.999	0.994
model 5 (A4)	499	0.988	0.945	0.991	0.975	0.956	0.732	0.732	0.753	1.000	0.984	0.999	0.995
model 6 (A2)	499	1.000	0.946	0.975	0.953	1.000	0.793	0.793	0.794	1.000	0.955	0.979	0.966
model 7 (A3)	499	0.986	0.929	0.979	0.966	0.882	0.671	0.671	0.702	1.000	0.984	0.999	0.995
model 8 (A4, Simpl.)	50	1.000	0.931	0.976	0.965	0.911	0.684	0.684	0.712	1.000	0.985	0.999	0.995
model 9 (A4, Simpl.)	50	0.990	0.936	0.985	0.970	0.952	0.689	0.689	0.717	1.000	0.985	0.999	0.996
model 10 (A4, Simpl.)	50	1.000	0.923	0.976	0.965	0.852	0.663	0.663	0.697	1.000	0.985	0.999	0.996
model 11 (A2, Simpl.)	50	0.868	0.880	0.936	0.911	0.711	0.707	0.707	0.729	0.887	0.908	0.974	0.945
model 12 (A4, Simpl.)	50	0.850	0.874	0.969	0.932	0.518	0.522	0.522	0.547	0.964	0.973	0.997	0.990
model 13 (A2, Simpl.)	50	0.892	0.890	0.938	0.920	0.743	0.734	0.734	0.752	0.877	0.896	0.980	0.937
model 14 (A4, Simpl.)	50	0.876	0.866	0.880	0.922	0.619	0.574	0.574	0.608	0.965	0.964	0.998	0.989
model 15 (A3, Simpl.)	50	1.000	0.941	0.991	0.973	0.998	0.728	0.728	0.749	1.000	0.985	0.999	0.995
model 16 (A4, Simpl.)	50	1.000	0.943	0.991	0.974	1.000	0.727	0.727	0.800	1.000	0.982	0.999	0.987
model 17 (A4, Simpl.)	50	1.000	0.944	0.991	0.975	0.970	0.741	0.741	0.758	1.000	0.979	0.999	0.994

Model	Number of features	Precision train	Precision test balanced (A)	Precision test full (B)	Precision test PTS>50 (C)	Accuracy train	Accuracy test balanced (A)	Accuracy test full (B)	Accuracy test PTS>50 (C)	Lift 10% test balanced (A)	Lift 10% test full (B)	Lift 10% test PTS>50 (C)
model 1 (A2)	499	1.000	0.875	0.376	0.641	0.999	0.903	0.972	0.937	3.105	9.141	7.822
model 2 (A2)	499	0.996	0.871	0.369	0.629	0.970	0.904	0.971	0.935	3.105	9.225	7.779
model 3 (A2)	499	0.999	0.892	0.424	0.702	0.974	0.905	0.976	0.947	3.105	9.221	7.835
model 4 (A3)	499	1.000	0.944	0.934	0.935	0.984	0.899	0.994	0.971	3.105	9.793	8.342
model 5 (A4)	499	1.000	0.955	0.938	0.945	0.980	0.903	0.994	0.971	3.105	9.746	8.308
model 6 (A2)	499	1.000	0.894	0.418	0.718	1.000	0.903	0.976	0.949	3.105	9.014	7.724
model 7 (A3)	499	1.000	0.951	0.934	0.942	0.945	0.883	0.993	0.966	3.105	9.567	8.065
model 8 (A4, Simpl.)	50	0.999	0.955	0.938	0.944	0.991	0.888	0.993	0.968	3.105	9.467	8.107
model 9 (A4, Simpl.)	50	1.000	0.956	0.943	0.948	0.978	0.890	0.994	0.968	3.105	9.555	8.184
model 10 (A4, Simpl.)	50	0.998	0.955	0.946	0.947	0.997	0.882	0.993	0.966	3.105	9.630	8.005
model 11 (A2, Simpl.)	50	0.842	0.785	0.334	0.592	0.806	0.843	0.969	0.924	2.942	8.517	6.765
model 12 (A4, Simpl.)	50	0.925	0.902	0.792	0.851	0.759	0.828	0.989	0.946	2.970	9.082	7.084
model 13 (A2, Simpl.)	50	0.837	0.770	0.410	0.564	0.815	0.844	0.976	0.918	3.046	8.660	6.803
model 14 (A4, Simpl.)	50	0.937	0.882	0.845	0.853	0.805	0.838	0.990	0.951	3.054	7.984	7.272
model 15 (A3, Simpl.)	50	1.000	0.958	0.945	0.947	1.000	0.902	0.994	0.971	3.105	9.746	8.261
model 16 (A4, Simpl.)	50	1.000	0.952	0.941	0.872	1.000	0.900	0.994	0.969	3.105	9.714	8.261
model 17 (A4, Simpl.)	50	1.000	0.945	0.924	0.931	0.997	0.903	0.994	0.971	3.105	9.734	8.257

Results summary:

- Some of the models tend to overfit. Most of them produce accuracy/AUC very close to 1 on the training set, and a significantly lower statistic on any of the test sets (A, B or C). This problem was overcome by:
  - Limiting the tree depth / minimal node size considered for splitting.
  - Limiting the number of variables used for modelling – the top 50 variables with the highest Gain statistic from the XGBoost algorithm were selected and used as the only features for a number of models, where a complete model on all the features was not possible to estimate due to performance issues.
  - Using the test set for validation during model learning – this approach for XGBoost proved that for most cases, despite overfitting, the error on the validation (test) set kept decreasing over consecutive iterations of the algorithm. However, as this approach violates the rule of completely independent training and test sets it was used only to check for possible overfitting.
- Models trained on completely randomized balanced datasets tend to have, in general, lower performance when compared with models built on the full dataset (Approach 3, A3) or pre-transition score based subset of the data (Approach 4, A4). This is in particular true for XGBoost models, as the algorithm handles class imbalances well and –when trained on larger datasets – is able to produce better results.
- Increasing the number of trees in random forest does not improve model performance much. 50 trees are enough to explain the phenomenon relatively well, and increasing this number to 200 does not bring much improvement.
- XGBoost models, when compared with models in the same groups (ie. trained on the same datasets) have slightly better performance than random forest and much better than the SVM algorithms. They have higher threshold-independent statistics (AUC, top 10% lift) and higher sensitivity/specificity/precision for the selected cut-off point (p=0.5). Moreover, they are faster to train and to score the data.
- As different models have been built on different datasets, the training performance statistics cannot be compared directly between all of them. Due to overfitting most of them will indicate perfect fit, so these statistics will not be used to compare model performance.
- Different types of models are characterized by different distributions of predicted scores (estimated probabilities). Directly comparing specificities / sensitivities for cutoff point = 0.5 may therefore be misleading, so the above should be compared in pairs (sensitivity, specificity), bearing in mind that there will be a trade-off depending on the classification threshold.
- Two threshold-independent measures were computed: AUC (Area under curve) and top 10% Cumulative Lift measure (containing information of the percentage of all target observations captured in the top 10% of nominals with the highest predicted score (estimated probabilities) divided by 10, eg. Lift = 8 equals to 80% of real one's classified as one by the algorithm).

Final model selection

It is important to note that in assessing model performance we have placed greater weight on the specificity as opposed to sensitivity so that the probability of false positives is minimised.

The final model was selected based on a number of non-statistical criteria and model performance measures. The reduced testing set with pre-transition scores higher than 50, as described in approach 4 - **PTS>50 (C)** – was primarily used to assess model performance. This was due to:

- Ethical aspects of scoring – a threshold > 50 aims to remove from the scoring population non-prolific, non- active and non-harmful nominals, which make up the vast majority of the crimes database.
- Statistical – we want to teach the model to learn to differentiate between active and harmful offenders who transition into the target cohort from nominals that are active and harmful but do not transition. Including low-harm nominals will not improve this performance and most likely won't targeted for any pro-active actions due to low harm and prolificacy.

However, model performance statistics for the full dataset (full B) are highly correlated with the pre-transition-score>50 dataset (PTS>50 (C)) for most of the statistics, so any selected model should score the general offending population comparably well.

Model number 17 was initially selected as the final model.

This model is characterized by a very high AUC value (comparable with similar XGBoost models built on all features) both on the full test set (B) and PTS>50 test set (C). It has a PTS>50 test set lift of 8.257, which means that in the top 10% of highest probabilities the model recognizes 8.257 times more transitioning nominals than a random model, thus capturing 82.57 % of all transitioning nominals.

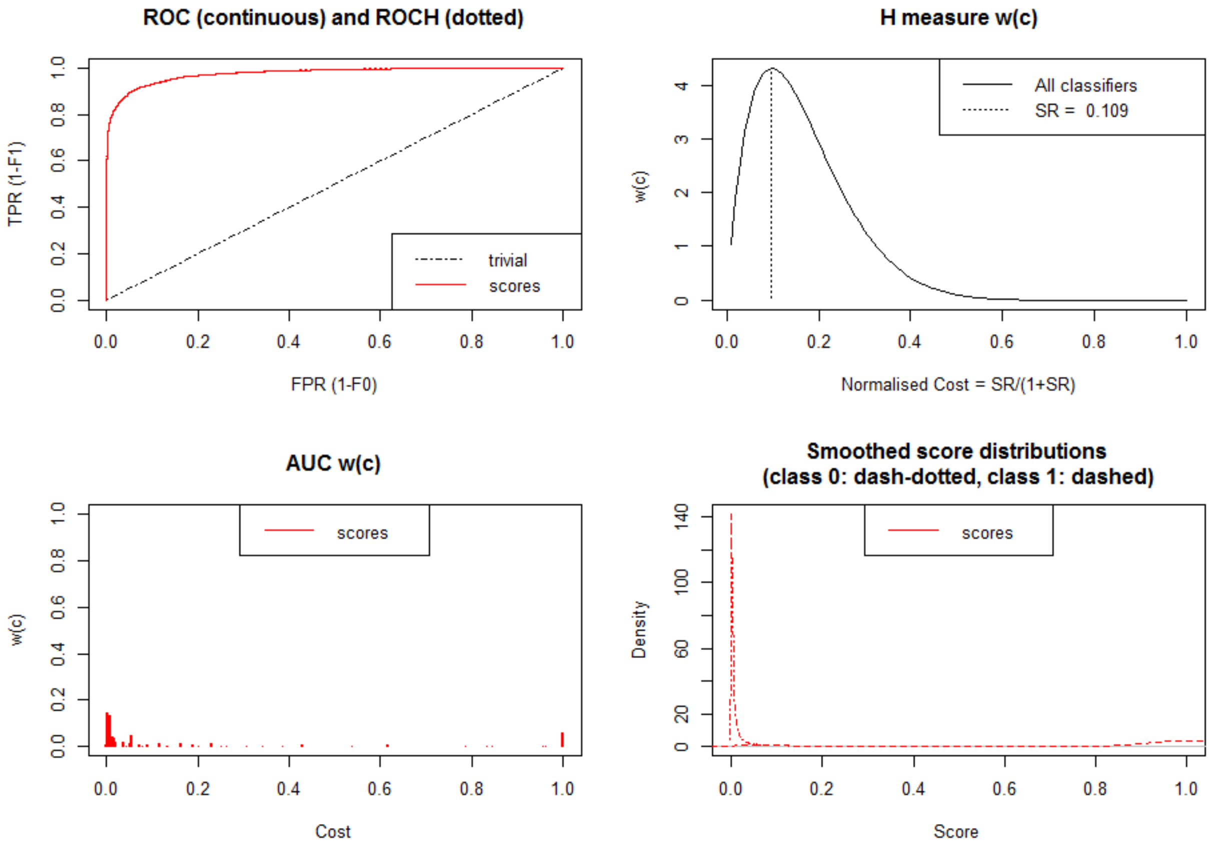
To further optimise the model a grid search was performed over the following parameters (and their values using k-fold cross-validation). Different XGBoost models were trained to test for different configurations of the following parameters:

- max.depth, values checked: 3,4,5,6,7,8,9,10
- eta, values checked: 0.1, 0.2,0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
- nrounds, values checked: 10,30,50,70,90,110
- colsample_bytree, values checked: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9

The results allowed for choosing the final model with the highest performance measures. **This final model, characterized by the highest PTS>50 test set (C) AUC (0.976) of all models and one of the highest PTS>50 test set (C) lift values (8.342) resulted in the following parameter values:**

Colsample_bytree = 0.5, max_depth =8, eta = 0.1, nrounds = 110.

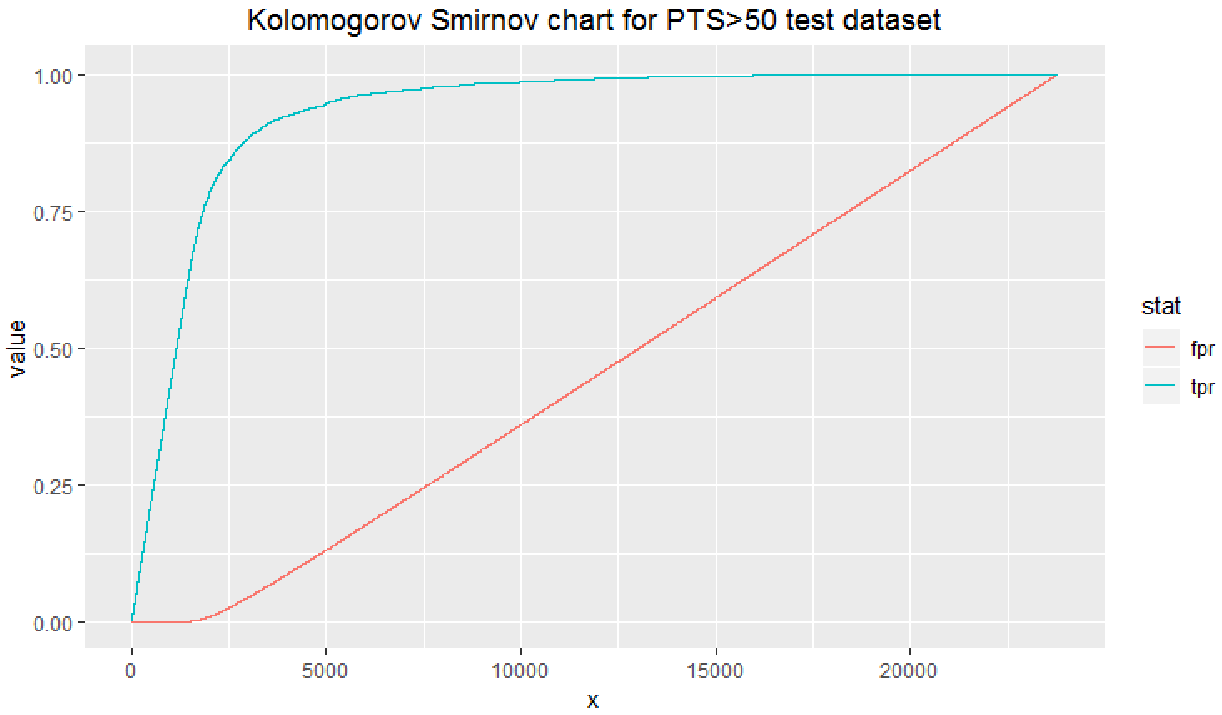
The below final model performance statistics were calculated for the reduced test set with pre-transition score higher than 50, as described in approach 4 - **PTS>50 (C)**



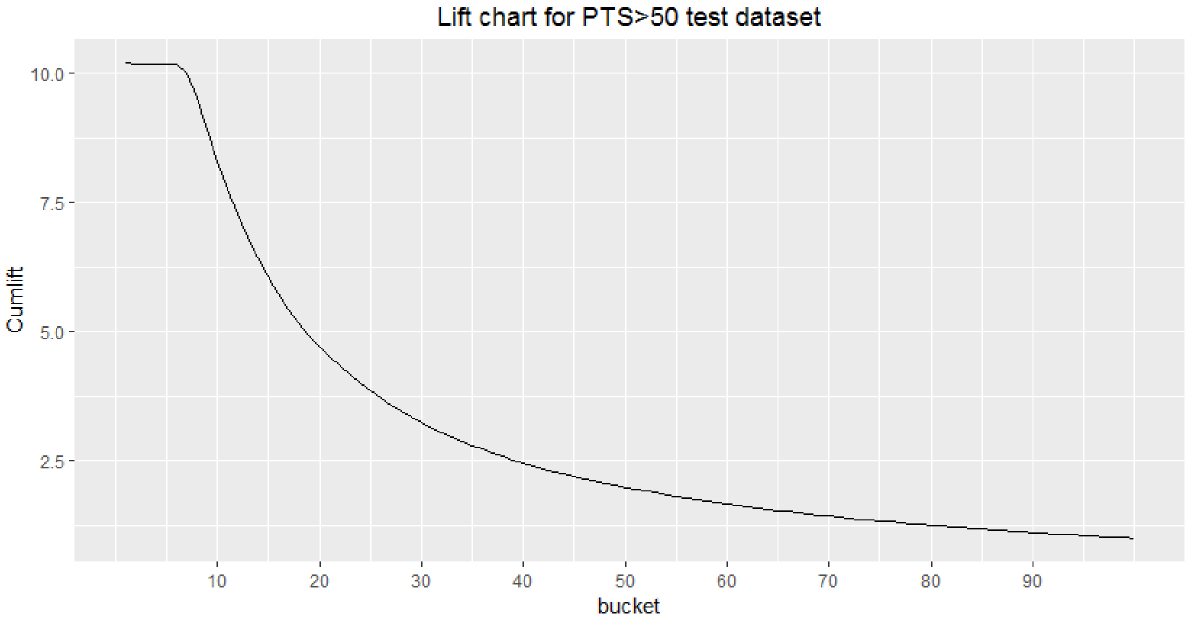
Final model performance statistics:

	Cutoff point						
Measure	0.3	0.4	0.5	0.6	0.7	0.8	0.9
accuracy	0.970	0.971	0.971	0.971	0.970	0.969	0.966
kappa	0.822	0.823	0.822	0.815	0.809	0.795	0.774
sensitivity	0.801	0.774	0.751	0.730	0.714	0.691	0.656
specificity	0.988	0.992	0.995	0.997	0.998	0.999	1.000
positive predicted value	0.881	0.917	0.946	0.963	0.978	0.985	0.997
negative predicted value	0.978	0.976	0.973	0.971	0.970	0.967	0.964
precision	0.881	0.917	0.946	0.963	0.978	0.985	0.997
recall	0.801	0.774	0.751	0.730	0.714	0.691	0.656
f1	0.839	0.839	0.837	0.830	0.825	0.812	0.791
prevalence	0.098	0.098	0.098	0.098	0.098	0.098	0.098
balanced accuracy	0.894	0.883	0.873	0.863	0.856	0.845	0.828
f1_sens_spec	0.885	0.869	0.856	0.843	0.832	0.817	0.792
AUC	0.976	0.976	0.976	0.976	0.976	0.976	0.976
AUCH	0.977	0.977	0.977	0.977	0.977	0.977	0.977
TP	1878	1815	1761	1712	1674	1620	1539
FP	253	165	100	66	37	24	5
TN	21241	21329	21394	21428	21457	21470	21489

FN	468	531	585	634	672	726	807



The maximum K-S statistic for model based on PTS>50 test dataset is 0.845

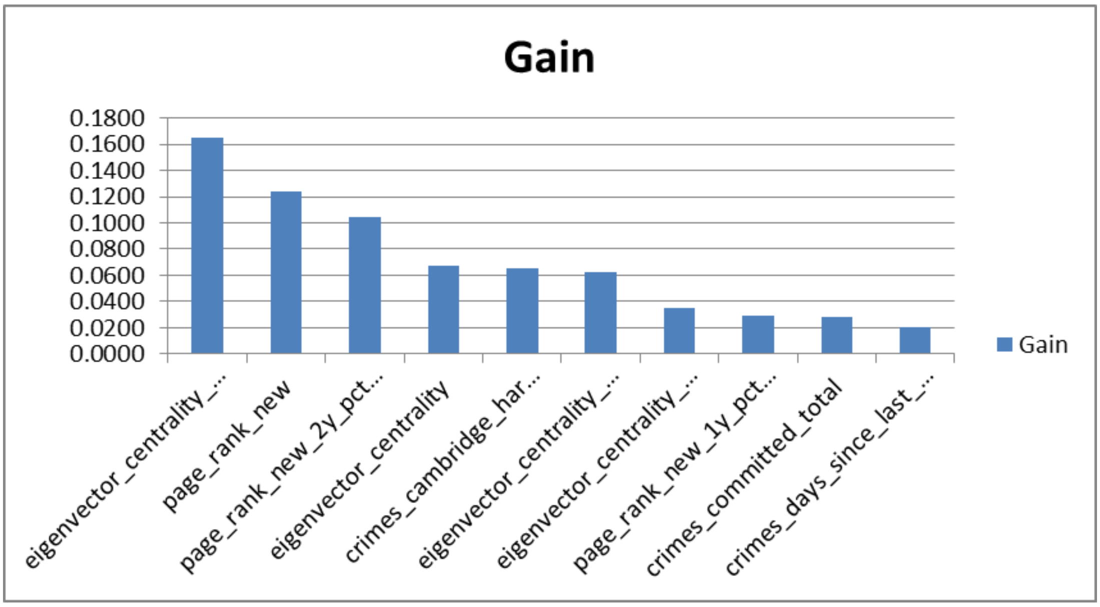


Feature Importance:

	Feature	Gain	Cover	Frequency
1	eigenvector_centrality_new_pct_chg	0.1651	0.0712	0.0315
2	page_rank_new	0.1236	0.0716	0.0416

3	page_rank_new_2y_pct_chg	0.1044	0.0539	0.0409
4	eigenvector_centrality	0.0677	0.0409	0.0321
5	crimes_cambridge_harm_total	0.0651	0.0779	0.0376
6	eigenvector_centrality_new_1y_pct_chg	0.0622	0.0603	0.0302
7	eigenvector_centrality_new	0.0354	0.0294	0.0332
8	page_rank_new_1y_pct_chg	0.0295	0.0541	0.0424
9	crimes_committed_total	0.0282	0.0351	0.0140
10	crimes_days_since_last_solo_committed	0.0208	0.0180	0.0256
11	crimes_ons_harm_total	0.0201	0.0221	0.0230
12	icis_custody_hours_total	0.0178	0.0265	0.0229
13	crimes_selected_sac_crimes_harm_total	0.0155	0.0236	0.0217
14	crimes_days_since_last_crime_committed	0.0151	0.0158	0.0237
15	nominals_age	0.0147	0.0460	0.0290
16	crimes_sac_broad_cchi_harm_total	0.0142	0.0196	0.0193
17	sas_times_searched_12m	0.0128	0.0289	0.0089
18	dip_both_coc_op_total	0.0097	0.0230	0.0129
19	crimes_days_since_last_solo_committed2	0.0094	0.0083	0.0129
20	icis_propery_searched_total	0.0094	0.0078	0.0113
21	crimes_days_since_last_coof_committed	0.0091	0.0134	0.0245
22	solo_crimes_committed_total	0.0084	0.0130	0.0185
23	icis_custody_hours_12m	0.0078	0.0165	0.0171
24	crimes_committed_24m	0.0077	0.0229	0.0082
25	crimes_min_age_committed	0.0068	0.0134	0.0196
26	icis_propery_searched_24m	0.0068	0.0196	0.0068
27	crimes_cambridge_harm_total_2y_pct_chg	0.0067	0.0093	0.0168
28	icis_custody_records_24m	0.0062	0.0217	0.0075
29	crimes_violent_harm_total	0.0061	0.0086	0.0179
30	crimes_cambridge_harm_24m	0.0060	0.0127	0.0137
31	topic10_avg_value	0.0055	0.0054	0.0197
32	sas_times_searched_total	0.0053	0.0048	0.0176
33	topic5_avg_value	0.0052	0.0053	0.0191
34	topic7_avg_value	0.0052	0.0060	0.0222
35	topic5_max_value	0.0052	0.0123	0.0175
36	icis_propery_max_cash_found_total	0.0049	0.0074	0.0221
37	crimes_ons_harm_24m	0.0048	0.0045	0.0149
38	topic7_max_value	0.0048	0.0083	0.0170
39	topic4_avg_value	0.0046	0.0057	0.0213
40	topic2_avg_value	0.0044	0.0063	0.0181
41	topic1_max_value	0.0044	0.0045	0.0183
42	topic9_avg_value	0.0043	0.0031	0.0194
43	topic8_avg_value	0.0040	0.0034	0.0168
44	topic6_avg_value	0.0039	0.0047	0.0164
45	icis_custody_cust_offences_records_total	0.0039	0.0053	0.0159

46	topic3_max_value	0.0038	0.0027	0.0171
47	topic9_max_value	0.0036	0.0049	0.0151
48	crimes_sac_broad_cnt_12m	0.0035	0.0122	0.0052
49	icis_custody_records_total	0.0035	0.0057	0.0102
50	crimes_selected_sac_crimes_total	0.0030	0.0024	0.0111

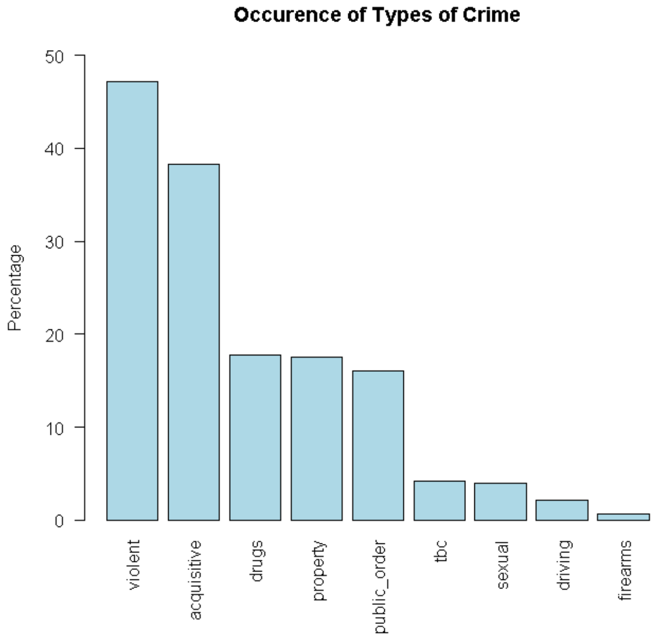


Appendix D

Potential Crime Types

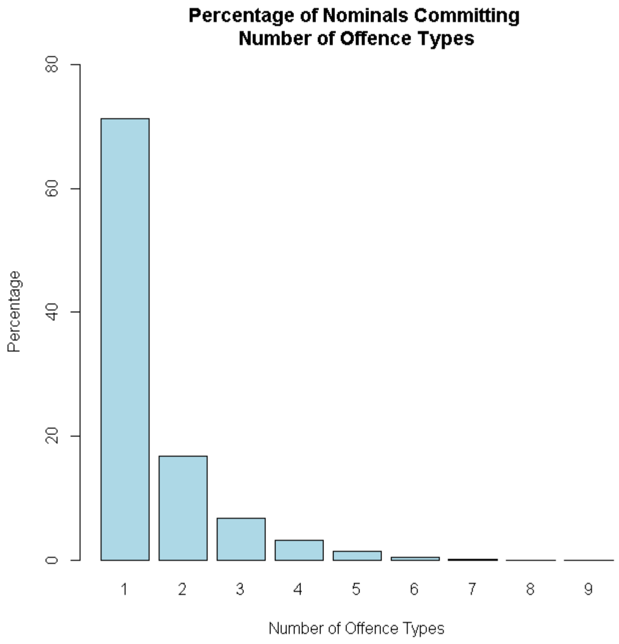
For the purposes of reporting (following discussions with subject matter experts (SMEs)), it was considered important to predict the next most likely crime type that may be committed by a nominal who is predicted to move into the high harm category.

There are 9 crime types that occurr in the proportions as shown in the chart below:



tbc is a category whereby the crime has not been allocated to any of the other types.

A nominal may have committed offences that fall into one or all nine of the categories with the majority having committed one crime type and a very small proportion having committed offences within all nine categories:



A random forest model (of 1,000 trees) has been built that can accommodate a multi-label classification problem. The model was built on a training set of 80,000 observations and tested on a test set of 20,000 observations. The resulting accuracy measures are:

Measure	Result
Hamming Loss	0.0362
Multilabel Subset	0.2182
Multilabel f1	0.9075
Multilabel Accuracy	0.8797

Two other methods were also tried, the classifier chains method using single classification trees as the base learner (whereby each label is trained in order with the features expanded to include information regarding previous labels in the chain and a similar, Bayesian approach (using C5.0 trees and Gibbs sampling).

The random forest performed slightly better than the classifier chain approach and the Bayesian method (whilst in some respects preferable) produces similar results to the random forest model but with much greater computational cost.

GLOSSARY:

Term	**Meaning**
Accuracy	How accurate, over both the positives and the negatives (‘yes’ and ‘no’) is a predictive model. The overall error rate can be calculated as 1 – the accuracy.
AUC	Area under the curve – this is the area enclosed by a ROC curve and is an indication of how much better than random guessing a model is in laking predictions. The higher this value the better.
AUCH	Similar to the above, but adjusted to enable better comparison between different models.
Centrality (in social network or graph analysis)	Various measures of how central a node is (e.g. a person) within a graph. There are various measures of centrality: eigenvector centrality is based on the concept that connections to high scoring nodes contribute more to the node in question than to low scoring nodes. Page rank is a variation of eigenvector centrality whereby nodes with many in-coming links (edges) are influential and nodes to which they are connected share some of that influence.
F1 (F score)	The weighted average of the either the precision and recall or sensitivity and specificity.
Features	These are other variables that may have some functional relationship with the target variable. Also known as explanatory variables.
Gain	The relative contribution of a feature to a model; a measure of the relative importance of a feature in the model (it is a measure of the improvement in accuracy brought by a feature to the branches that it is on).
Kappa	A measure of how well a predictive model performs compared to how well it performs simply by ‘chance’.
KS	Kolmogorov - Smirnov; the maximum difference between the true positive rate and the false

	positive rate. The associated chart is taken at different cutoff points.
Local false discovery rate - (l)fdr	A means of adjusting p-values (see below) in the face of running multiple tests.
Multicollinearity	The situation where one feature can be linearly predicted from the others. This can cause issues with some (but not all) model estimation methods and essentially means that the features that are correlated essentially provide the same information.
Negative Predicted Value	The proportion of negative results (‘no’) that are true negative.
Observation	Essentially the rows in a dataset to which a model is applied (these could be individuals, areas, etc.)
p-value	A result from statistical tests quantifying the probability of a statistic as large or larger being found ‘by chance’ (assuming a null hypothesis to be correct). These will only occur in the EDA stage of a project (if at all) and are not used in the building of predictive models.
Parameter(s)	Values of various constraints that are placed on predictive models during their training that are optimised to maximise their predictive capability.
Positive Predicted Value	The proportion of positive results (‘yes’) that are true positive.
Precision	When a predictive model predicts a ‘yes’, how often is it correct (i.e. the proportion of ‘yes’s’ correctly identified as such).
Random Forest	A predictive modelling methodology that, in the case of classification, uses multiple classification trees to determine the probability of an observation entering into a class (a ‘yes’ or a ‘no’).
ROC curve	The receiver operating characteristic curve; a

	graph that summarizes the performance of a predictive model over all possible thresholds. It is a plot of the true positive rate against the false positive rate as the threshold for assigning observations to a given class is varied.
Sensitivty	The true positive rate; when a data point was a 'yes', how often does a predictive model predict a 'yes'. The higher this proportion the better; although this needs to be balanced with the specificity. Also known as recall.
Specificity	The true negative rate; when a data point is a 'no', how often does a predictive model predict a 'no'. The higher this proportion the better; although this needs to be balanced with the sensitivity.
SVM – Support Vector Machine	A predictive modelling methodology that utilises seperation between attributes.
Target variable	The variable that is to be analysed / predicted. This could be binary (yes/no; 0/1), multinomial (red/blue/green), ordinal (first/second/third) or quantitative (where the variable is numeric). Also known as the dependent variable.
Variable	An item of interest; any characteristic, number or quantity that is measured or counted.
Variance inflation factor	A measure of the presence of multicollinearity by the features within a dataset.
XGBoost	A predictive modelling methodology that, in the case of classification, uses multiple models (usually trees) to determine the probability of an observation entering into a class (a 'yes' or a 'no').