Solvability – Some Remarks

Siddhartha Bandyopadhyay, Anindya Banerjee and Tom Olphin

26 June 2017

Introduction

An understanding of solvability is an understanding of the optimal use of limited investigative resources:

How can PFAs maximize the use of their resources to reach an optimal outcome, usually clearance rates?

This requires us to define an *optimal outcome*: Is it solving the maximum number of crimes?; maximum number of crimes weighted by severity of harm? Other?

This implicitly has a resource constraint because the optimal outcome is subject to the resources available

Simple solvability models try to identify how different factors correlate with successful outcomes (usually measured by clearance rates)

These models do not typically account for unobservable factors (e.g. effort investigators put; ability of investigators, risk-taking preference of criminals; ability of criminals)

Current models are fairly simplistic in their approach when attempting to separate a correlation from a true causal effect between a solvability factor and clearance rate:

- Include solvability factors which are correlated with each other; e.g., the presence of a forensic expert is correlated with the existence of forensic evidence. While both can affect solvability presence it creates multi collinearity problems in the empirical analysis
- Do not distinguish between factors that are outside of the control of the police (e.g. day versus night-time crime) and those that are policy determined (e.g. sending a same sex/race officer for a report of sexual violence)

- **Do not take care of reverse causality running between perception of the effectiveness of a certain solvability factor and its subsequent effect on solvability (e.g. self-fulling prophecy)**

Key Issue: How focusing on solvable crimes can free up resources

A Wish list of Objectives:

- Step 1: Define an objective function e.g. solving the maximum number of crimes weighted appropriately (e.g., by severity of harm, lower reoffending)

- Step 2: Agree with forces the subset of crimes to focus on

- Step 3: Understand the outcome of any changes that have been implemented to any subset of crime where solvability factors have been taken into account

- Step 4: Develop an empirical model which overcomes problems typically not considered in the literature

- Step 5: Develop an experiment based on solvability factors previously identified (Step 4)

- Step 6: Use any resources saving from using the model to decide on what other crimes can be prioritized

Our study:

Data for burglary offences in Norfolk between April 2012 and May 2015 were used to build a statistical model capable of predicting whether offences would be solved or not, based on the evidence gained from the initial investigation.

The data were randomly split into two, with one half of it being used to build the predictive model, which was then tested on the other half of the data.

The purpose of this work was to develop a statistical model (the solvability model) that would use factors that are correlated with a crime being solved to predict whether crimes are likely to be solved from the evidence discovered during the primary investigation.

Dependent upon the **cut-off** used, the model is capable of identifying hundreds of cases where investigation is extremely unlikely to result in a positive outcome, allowing resources to be

freed up to concentrate on other cases which are more solvable, or on other demands.

The work allows for a rigorous model-based method of screening crimes and, by focussing on crimes that are likely to be solvable, frees up scarce police resources.

Research recognises that, with the evidence and resources available, it is not possible for police to solve all crimes.

Implicit use of solvability factors is already present in most forces but they are based on officer judgment (and are hence subjective), and lack external validity.

The building of an algorithmic model provides additional consistency and rigour to a process of screening that already exists.

Data Description

Approximately three years (from April 2012 to May 2015) of police recorded data on burglary in Norfolk was used to build the model. The data comprised both electronically recorded categories as well as officer free text that had to be coded manually. In all, data were compiled for 253 variables (68 manually coded from officer free text fields and 185 coded from downloads of Norfolk Constabulary systems). This initial list of 253 variables was narrowed down to 42 after removing those variables which were found to have very little explanatory power in terms of solvability or which were highly correlated with each other (the procedure is described under methodology).

Outcome Variable

It was originally intended that Home Office Outcome (as updated in 2014) was to be used as the outcome variable to indicate whether the offence was solved or not.

However, this was not possible as coding of some outcomes were not present prior to 2014.

Therefore, sanction detection has been used as the outcome variable to determine whether the case was solved (detected) or unsolved (filed undetected).

Some inconsistencies were present in the data pertaining to offence outcome due to change from using detections to using Home Office outcomes. This was for a small number of cases.

Offences where there are inconsistencies have been removed from the analysis. All cases which were cleared by TIC, or which were identified as having been solved during the initial investigation, were also removed.

Methodology

Factor Identification

Following data cleansing and coding, factors have been analysed using Chi-squared tests to show differences in their prevalence between solved and unsolved cases.

Factors were then assessed for close correlation so that factors which were essentially showing the same things as other factors could be removed.

Those factors shown as being more prevalent in solved cases (solvability factors) are shown in Table 1, and those factors which are significantly more prevalent in unsolved cases (case-limiting factors) are shown in Table 2.

Table 1: Solvability Factors (All significant at p<0.05)

Factor Description	χ^2 Score	Method Obtained
Arrest Made	338.216	Free-text Coding
There is Suspect Information Provided	222.793	Free-text Coding
Between Times Less than 1 hour	146.796	Downloaded Data
DNA Recovered	141.902	Downloaded Data
Prints Recovered	126.320	Downloaded Data
Burglary occurred in a Dwelling	124.673	Downloaded Data
Definitely Video (Downloaded) or CCTV is Evidential (Freetext)	123.769	Combination
Definite Victim Statement or ABE Interview	90.951	Free-text Coding
Report Time Less than 15 minutes	81.143	Downloaded Data
Witnessed by Officer	78.671	Free-text Coding
Are there witnesses other than the victim	70.040	Free-text Coding
Stolen Items Recovered	65.185	Free-text Coding
Shoe Mark Recovered	51.865	Downloaded Data
Officer states there are forensic opportunities (including request for CSI)	51.586	Free-text Coding
Suspect's Clothing Seized	42.578	Free-text Coding
Other Sample Recovered	37.655	Downloaded Data
Forceful Entry	35.643	Downloaded Data

Factor Description	χ^2 Score	Method Obtained
Entry Rear	35.276	Downloaded Data
Entry Window	31.283	Downloaded Data
Items Seized	29.131	Free-text Coding
NORSOC Ref	28.864	Downloaded Data
Domestic Indicator	27.951	Downloaded Data
Entry Door	27.798	Downloaded Data
Gaming Username Enquiries	17.039	Free-text Coding
Victim witnessed offence	15.599	Free-text Coding
Telecommunications Opportunities Available	15.218	Free-text Coding
Scene Photographed	13.208	Free-text Coding
Item located at second hand shop	10.363	Free-text Coding
Part of Series	8.313	Free-text Coding
Victim is Unemployed	8.142	Downloaded Data
Discovered By Police	6.969	Downloaded Data
VRM Provided	4.490	Free-text Coding
Financial Enquiries Card or Account Usage	4.058	Free-text Coding

Table 2: Case-limiting Factors (All significant at p<0.05)

Factor Description	χ^2 Score	Method Obtained
There is no CCTV available	146.779	Free-text Coding
Between Times Greater than 12 Hours	118.567	Downloaded Data
Offence Occurring in an Outbuilding (Shed/Stable/Allotment etc.)	69.374	Downloaded Data
No Witnesses Clearly Stated	50.666	Free-text Coding
Negative House to House Conducted	29.382	Free-text Coding
House to House not Conducted in Initial Investigation	22.238	Free-text Coding
Stolen Property - Industrial Equipment	20.647	Downloaded Data
Stolen Property - Cycle	15.707	Downloaded Data
Time to Report Greater than 18 Hours	11.494	Downloaded Data

Interesting Factors with No Significant Difference

The following factors were interesting due to being factors that one might assume would be important in determining solvability, but which had no difference in prevalence between solved and unsolved cases:

- Tool marks being recovered by Forensics
- A press release being completed
- Victim providing a negative statement or refusing to cooperate
- Stolen items being found for sale online
- Whether the offence was a distraction offence

Logistic Regression

Because it is important to include both solvability factors and case-limiting factors in any algorithmic prediction model, both sets of significant factors were then included in a logistic regression analysis in order to build a model for prediction of burglary solvability.

Logistic regression is a statistical method for analysing a dataset in which the outcome is a dichotomous variable (i.e. there are only two possible outcomes). It predicts the outcome based on a number of explanatory variables that are likely to determine the outcome.

In this case the dichotomous outcome variable is whether cases were solved or not solved, and we have included all the variables that are shown to be correlated with solvability from the individual chi-squared tests.

The model was built using a randomly assigned half of the available dataset, with the other half used for external validity.

Some variables were removed, and changes were made to mandatory allocation rules for the model throughout eight iterations to produce a model which most effectively predicts solvability, before the final model was produced using 29 variables.

Testing on a secondary data set is an important step as this allows for the predictive accuracy to be assessed in a manner which is not biased through testing on the same data set used to build the model. This externally validates the model and allows for a true assessment to be made of how the model will fare in practice.

The results from the final logistic regression are presented in Table 3.

Table 3: Results of Logistic Regression Model

Variable	B	Sig.
Gaming Username Enquiries	-4.1501993945	.000
Item located at second hand shop	-2.8212215962	.031
Discovered By Police	-1.6088851600	.041
DNA Recovered	-1.2432782388	.000
Domestic Indicator	-0.9613807375	.091
Financial Enquiries Card or Account Usage	-0.8676973571	.126
Prints Recovered	-0.7089646728	.000
Part of Series	-0.5828106700	.006
Stolen Items Recovered	-0.5781382287	.034
Telecommunications Opportunities Available	-0.5742982424	.526
Either CCTV measure is positive	-0.5509179989	.002
Unemployed	-0.5078199848	.058
Are there witnesses other than the victim	-0.5019885667	.028
Witnessed by Officer	-0.4490801532	.637
Time to Report Less than 15 minutes	-0.4286801654	.001
There is Suspect Information Provided	-0.4064605307	.022
Shoe Recovered	-0.3194626005	.016
Forceful Entry	-0.2954104448	.026
Between Times Less than 1 hour	-0.2037576893	.212
Items Seized	-0.1057116916	.619
Suspect Clothing Seized	-0.0495460319	.941
Other Sample Recovered	-0.0201816735	.903
Time to Report Over 18 hours	0.0941689671	.674
Negative House to House Conducted	0.2047767659	.143
No Witnesses Clearly Stated	0.2454850535	.173
House to House not Conducted in Initial Investigation	0.3131791432	.051
Between Times Over 12 hours	0.3424270166	.011
There is no CCTV available	0.5928572330	.000
Offence Occurred in an Outbuilding	0.7883362897	.001
Constant	2.5917734249	.000

In addition, the model is built using the following mandatory factors:

- Aggravated Burglaries are automatically allocated for investigation
- Burglaries for GBH (rather than theft) are automatically allocated for investigation
- Offences where a suspect has been arrested are automatically allocated for investigation
- Cases where the honesty of the victim was questioned were automatically filed

The Model

The Solvability Model can be seen below in Figure 1. The model works by taking the constant value (2.5917734249), and adding to this the sum of the relevant values for factors that are present in a case.

The higher the score, the lower the probability the case is solved (this is only a normalization).

Figure 1: Norfolk Solvability Model

Case Solvability Score (Higher Score is Less Solvable) = 2.5917734249 + the sum of the relevant values for the factors that are present as follows:

(If Gaming Username Enquiries is present **-4.1501993945**)
(If Discovered By Police is present **-1.6088851600**)
(If Domestic Indicator is present **-0.9613807375**)
(If Prints Recovered is present **-0.7089646728**)
(If Stolen Items Recovered is present **-0.5781382287**)
(If Either CCTV measure is positive is present **-0.5509179989**)
(If Are there witnesses other than the victim is present **-0.5019885667**)
(If Time to Report Less than 15 minutes is present **-0.4286801654**)
(If Shoe Recovered is present **-0.3194626005**)
(If Between Times Less than 1 hour is present **-0.2037576893**)
(If Suspect Clothing Seized is present **-0.0495460319**)
(If Time to Report Over 18 hours is present **+0.0941689671**)
(If No Witnesses Clearly Stated is present **+0.2454850535**)
(If Between Times Over 12 hours is present **+0.3424270166**)
(If Offence Occurred in an Outbuilding is present **+0.7883362897**)
(If Item located at second hand shop is present **-2.8212215962**)
(If DNA Recovered is present **-1.2432782388**)
(If Financial Enquiries Card or Account Usage is present **-0.8676973571**)
(If Part of Series is present **-0.5828106700**)
(If Telecommunications Opportunities Available is present **-0.5742982424**)
(If Unemployed is present **-0.5078199848**)
(If Witnessed by Officer is present **-0.4490801532**)
(If There is Suspect Information Provided is present **-0.4064605307**)
(If Forceful Entry is present **-0.2954104448**)
(If Items Seized is present **-0.1057116916**)
(If Other Sample Recovered is present **-0.0201816735**)
(If Negative House to House Conducted is present **+0.2047767659**)
(If House to House not Conducted in Initial Investigation is present **+0.3131791432**)
(If There is no CCTV available is present **+0.5928572330**)

Therefore, for example, if a case was discovered by police, had DNA recovered, had items seized and had negative house to

house identified, the score for this case would be 2.5917734249 + SUM OF (-1.6088851600, -1.2432782388, -0.1057116916, +0.5928572330) = 0.2267555675.

It is then necessary to set a cut-off value to determine which cases will be allocated for investigation, and which will be filed.

Cut-Off Point

The solvability model works out the answer to the solvability problem as a sum which can then be compared to a cut-off value.

Cases are allocated if their score is less than or equal to the chosen cut-off value while they are filed if their score is higher than the chosen cut-off value.

To decide upon the appropriate cut-off value, it must be understood that there are two types of errors

(i) a case is allocated when it should have been filed and
(ii) a case is filed when it should have been allocated.

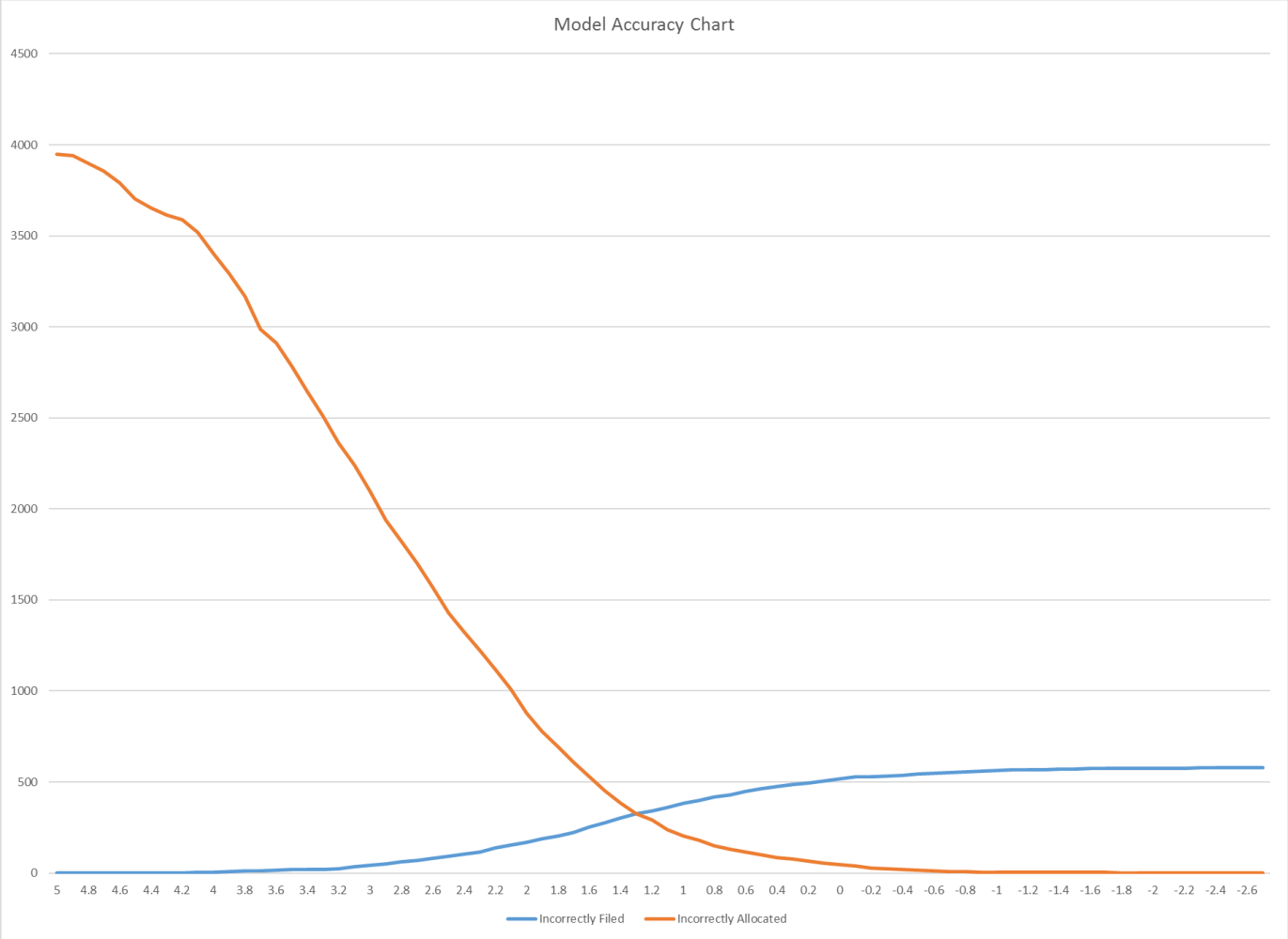
There is a trade-off between these, increasing one type of accuracy, increases the other type of error.

This is intuitive, if we are to ensure that no case is incorrectly filed, we will end up with having a lot of incorrectly allocated

cases, while if we want to ensure we do not incorrectly allocate cases, we have to tolerate some error in incorrectly filing cases that would otherwise have been solved.

The cut-off can be optimised either for resourcing purposes, or based on a balance of solving crimes that are solvable vs wasting resources on unsolvable crimes. Figure 2 shows the error rates in each direction at each potential cut-off value.

Figure 2: Comparison of errors at a range of potential cut-off values for the build set



As can be seen from Figure 2 and Table 4, somewhere around 2.8 to 3.2 would maximise the case-filing potential of the model, whilst not losing many solved cases

Table 4: Successes and Errors at Cut-off values of 28-3.2 (Build Set)

Cut-off Value	3.2	3.1	3	2.9	2.8
Correctly Allocated	554	545	538	530	519
Correctly Filed	1586	1711	1850	2011	2126
Incorrectly Filed	25	34	41	49	60
Incorrectly Allocated	2364	2239	2100	1939	1824

External validity: Testing of the Model on a Separate Data Set

The remaining data were coded and cleansed to allow testing of the model on a separate data set which improves the external validity of the model, and allows an estimation to be made of the performance of the model when applied in real time, along with a prediction of the allocation levels that would be created through use of the model.

Figure 3 shows the error rates in each direction at each potential cut-off value when the model was applied to the testing set.

Figure 3: Comparison of errors at a range of potential cut-off values for the testing set

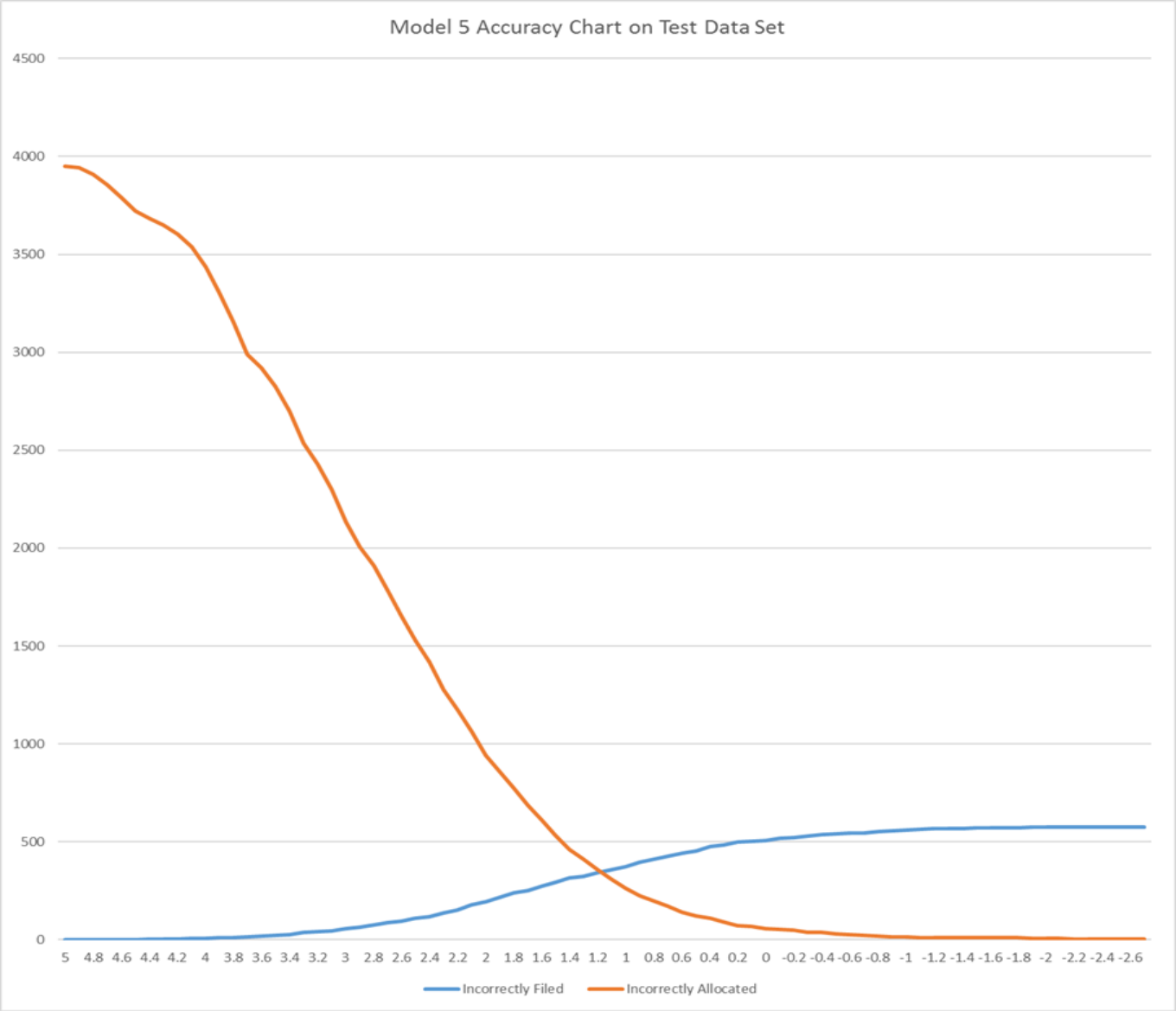


Table 5 shows the accuracy of the model at the same range of cut-off points used in Table 4 above.

Table 5: Successes and Errors at Cut-off values of 28-3.2 (Testing Set)

Cut-off Value	3.2	3.1	3	2.9	2.8
Correctly Allocated	536	531	520	513	502
Correctly Filed	1521	1650	1815	1946	2041
Incorrectly Filed	41	46	57	64	75
Incorrectly Allocated	2429	2300	2135	2004	1909

As can be seen, the model predicts with accuracy close to that obtained from using the ‘build’ dataset.

It is slightly less accurate than it was on the data it was built using, but due to the large dataset used to build the model, and the fact that cases were randomly assigned to each group, the model still performs well on a fresh dataset.

This is likely to be comparable to how the model would fare when used in real time.

Variations in the model

For the model to be applied in real time, there are choices which need to be made in relation to mandatory allocation rules, and the cut-off value which will be used to allocate cases.

As seen above, there are already some mandatory allocation rules built into the model, but these can be added to if required.

It is also possible that many of the cases which are found to be incorrectly filed would be found to be reopened due to forensic results or additional evidence being provided, therefore not incurring errors.

If this is the case, it may allow for the cut-off value to be lowered, with minimal impact upon cases that would be solved.

Conclusions

- Solvability models allow us to understand how the presence and absence of factors affect solvability od cases
- In our study, by using one part of the data to ‘build’ a model and the other part to test, we test for external validity of the model
- No statistical model has 100% accuracy, the question is whether or not it can aid human decision making as well as provide consistency on when to (based on a primary investigation) spend more (scarce) investigative resources
- As some screening post primary investigation is done, one can compare how such algorithmic decision-making compares with those made by officers
- This will enable one to see how much of a role it can play in directing police effort in an optimal way.