Harmful Stalking and Harassment offenders and the Estimation of future risk – INTERIM REPORT

Data Analytics Lab

September 2023

This project aims to estimate the probability that a nominal goes on to commit high harm crimes given the pattern of offending amongst offenders who have a criminal history of stalking and harassment. Part of a wider approach to tackle Violence Against Women and Girls (VAWG), this project aims to identify individuals who are likely to escalate in their offending as a means to create intervention opportunities, thus offering opportunities to protect potential victims from future high harm offences.

All available data has been considered ranging from the full criminal history of offenders, text information contained in crime logs, to the number of calls a victim has placed with WMP before the most recent stalking offence.

The main body of this report focusses on the *Stalking* crime as a subset of the *Stalking and Harassment* crime group. Results for *Stalking and Harassment* as a whole are contained in the Appendix.

The data in this report is based on victims, offenders and suspects. Offenders and suspects are considered the same, and are referred to as nominals. In order to identify patterns of behaviour, it is also necessary to include non-crimes as well as crimes in the historical feature set so these together are referred to throughout as incidents. These decisions reflect the sources of information officers would use to undertake assessment of risk.

The optimal model is chosen based on metrics provided through model training and feedback from practitioners who would use the model. The 'optimal' model reports an accuracy of 84% across the two most important classes in identifying the most harmful stalking offenders within 12 months before they escalate in offending, while balancing the false positive rate for operational use.

Due to changes in the Home Office counting rules regarding stalking and harassment, the analyses will need to be rebuilt due to associated changing definitions, so this report is provided to highlight the basic framework of the modelling approach.

Contents

Table of Figures

Table of Tables

Glossary of Terms

Term	Definition
Non-Crime	Non-Crime denotes incidents that are recorded on the WMP crime system that do not meet the threshold of a substantive offence. Incidents of Non-Crime can include domestic incidents that are below the threshold of domestic abuse and anti-social behaviour. The recording and analysis of these Non-Crimes are used routinely throughout policing to build an understanding of the vulnerability in victims' and nominals' lives. External reviews of policing point to the need to utilise the data within our systems to manage potential risk and threat in order to perform our duty to protect the public. For example, the recent review of Operation Soteria Bluestone (which focuses on the work to improve outcomes for victims of rape and serious sexual offences) states that, "*while there are important ethical considerations around the inclusion of unconvicted suspects in samples for research purposes, on the basis of ensuring that guilt is not assumed in these circumstances, the premise of police work is based on the collation of intelligence, which is predicated on allegation as opposed to conviction data. The police, by virtue of recording criminal allegations, have a wealth of information which can be used to explore repeat offending and repeat suspects. The police, therefore, have an opportunity to draw on the intelligence contained within their own records to better understand the nature of repeat offending and repeat suspects and to make more informed decisions about how to tackle this type of offending*." Stanko et al Operation Soteria Bluestone Year 1 Report 2021 – 2022, p.99 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1124704/Operation_Soteria_Bluestone_Year_1_Report-_FINAL.v3.pdf
Stalking day zero	This is the time reference point used for every nominal and victim in the dataset. Stalking day zero is the date where the most recent stalking incident occurred. From this relative date, day counts and other features are calculated by looking backwards and forwards in time. Using this normalises the incident recency for all nominals and victims.
CCHI	Cambridge Crime Harm Index – This is a harm score index which associates a harm score to crimes in England and Wales. The score is based on the minimum recommended custodial sentence for a first offence where the score is the number of days that sentence carries. If a crime warrants 30 days in prison then the harm score is 30. For crimes that do not carry a custodial sentence, the score is calculated based on the time required to carry out a community order. https://www.crim.cam.ac.uk/research/thecambridgecrimeharmindex
Incident	This denotes the row level information contained in the WMP crime system. An incident can be a crime or a Non-Crime whereby the incident was substantial enough to warrant recording on the crime system. Not every incident results in a crime, see Non-Crime definition.
Nominal	This is the title given to offenders and suspects used in this dataset.
Document Term Matrix	A matrix that describes the frequency of terms that occur in documents. Each row would represent a document, each column would represent a word.
Stemming	Stemming removes affixes of words. For example; running, runs and run all become run. A predefined vocabulary is usually used for this task.

Lemmatisation	Lemmatisation reduces words to their base meaning and root word. For example; the word better would reduce down to the word good. A predefined vocabulary is usually used for this task.
Bi-grams	Bi-grams counts and links words together which commonly occur together. For example; ‘the best’’ and ’best performance’ would be two bi-grams. Evaluation of these parings and the frequency to which pairing appear can be useful in language modelling.
Tri-grams	The same as b-grams but with three words. It is possible for any length of word combinations, this is known as n-grams.

1. Stalking and Harassment

1.1. Stalking and Harassment in England and Wales

Stalking and harassment is when someone repeatedly behaves in a way that makes people feel scared, distressed or threatened (POLICE.UK, 2023). The Protection from Harassment Act of 1997 introduced harassment offences into law, with Stalking offences defined in law in The Protections of Freedoms Act 2012 (CPS, 2023).

The Office for National Statistics (ONS) collects data on crime across the country; stalking offending is detailed in the Crime Survey for England and Wales (CSEW)[1] (ONS, 2022). The findings show that through the 1 year period since the last survey 9.5% of men and 23.3% of women report that they have been a victim of stalking at least once since they have been above the age of 16. Furthermore, it is estimated that there have been 1,170,000 female stalking victims in the year ending March 2022.

Looking at the published Police Recorded Crime Outcomes for March 2021 – March 2022 (Home Office, 2023), there were a total of 718,480 crimes recorded as sub category stalking and harassment, with 117,973 specifically crimed as stalking. This therefore shows that there is a stark difference between the actual number of recorded crimes, and the number of crimes estimated to have been experienced in the country; suggesting that the vast majority of incidents of this offence type are not reported to the police. Figure 1 shows how the number of recorded offences has changed in recent years, from a low of 2,252 in 2015 (first year recorded) to more than 117,000 in 2022, and increase of more than 5000%. The magnitude of the change in the number of stalking offences suggests an enormous societal issue, however there are more factors affecting this figure, mainly the Home Office recording practices of such crimes, as will be detailed in Section 2.1.

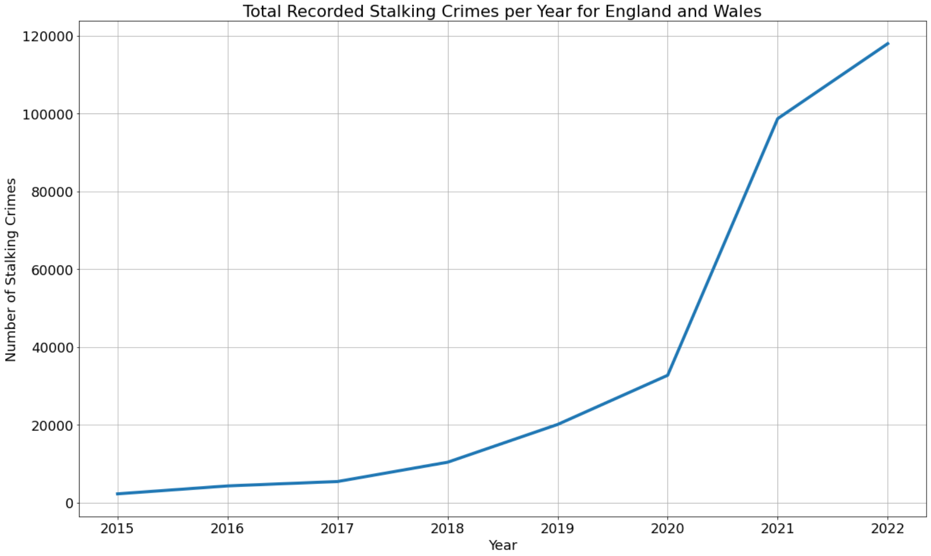


Figure 1 - Total recorded stalking crimes per year in England and Wales

1.2. Stalking and Harassment Characteristics

Defining stalking in a clinical sense can be defined as "the wilful, malicious, and repeated following or harassing of another person that threatens his or her safety" (Meloy and Gothard, 1995). Stalking as a behaviour can be categorised into 4 main groups, (1) Surveillance (eg. Vehicle tracking, online tracking etc.), (2) impact to life with unwanted forms of contact (letters,

[1] CSEW is a crime survey and does not relate to the raw number of recorded crimes in England and Wales.

messages, gifts), (3) intimidation (threatening behaviour), and (4) assault (physical and sexual) (Logan & Walker, 2019). These acts are usually carried out by the individual in question, or can be carried out by proxy; provision in UK law protects against stalking by proxy. Regular incidents of this nature result in lasting effects on victims' physical and mental health, behaviour, and general quality of life. The regular and repetitive incidents are one of the defining characteristics of stalking that indicates an obsession and fixation on a victim.

The evaluation of reoffending is a useful tool to measure stalking behaviour as it provides an analysis of harmful behaviour and the effectiveness of police interventions. There is evidence to suggest that restraining orders are regularly violated, which suggests that such interventions are not cogent to protecting the victim (Häkkänen, Hagelstam & Santtila, 2003). One study suggests that amongst stalking offenders, there is a 49% re-offending rate, where 80% of such offenders reoffend within one-year (Rosenfeld 2003). Furthermore, high-harm offenders are up to four times more likely to repeat offences against the original victim than lower-harm offenders of stalking (McEwan et al, 2018).

Research suggests that 79% of stalkers are male, while stalking victims are 75% female, and that offenders are known to the victim 75% of the time (Baum et al, 2009). Cases where the offender/victim relationship is known can be of higher concern due to the duration of the crime being over a long period of time, typically 13 months on average up to in excess of 26 months in some cases (Tjaden and Thoennes, 2000).

Traditionally, research around stalking behaviour focussed on obsessive, excessive, delusional love which is known as erotomania and is classified as a mental illness (DSM V, APA, 2013). However, the majority of stalking cases in the general population are not solely focussed around the romantic or intimate aspect. Research suggests that 70% of stalking victims felt they were experiencing behaviours around retaliation, spite and control, which came from feelings of anger from their perpetrator (Baum et al, 2009).

2. Overview of Stalking and Harassment in the West Midlands

2.1. Summary of Data

In the West Midlands Police (WMP) Force area, there has been a significant number of reported incidents pertaining to stalking and harassment (S&H). Specifically, a total of 168,850 S&H incidents have been recorded, perpetrated by 102,906 distinct nominals. Notably, the majority of these incidents have occurred subsequent to the implementation of the Protection from Harassment Act 1997 in March of that year. However, a small number of offences, specifically 231, have been recorded with incident dates prior to January 1997, indicating that the period of offending in these cases extended from before the aforementioned date. The 102,906 nominals are identified in 1,103,267 other incidents of any crime type. Figure 2 highlights the escalation in the volume of S&H incidents observed in the most recent 5 years. There was a total of 7,956 recorded stalking incidents in WMP in the year March 2021-March 2022.

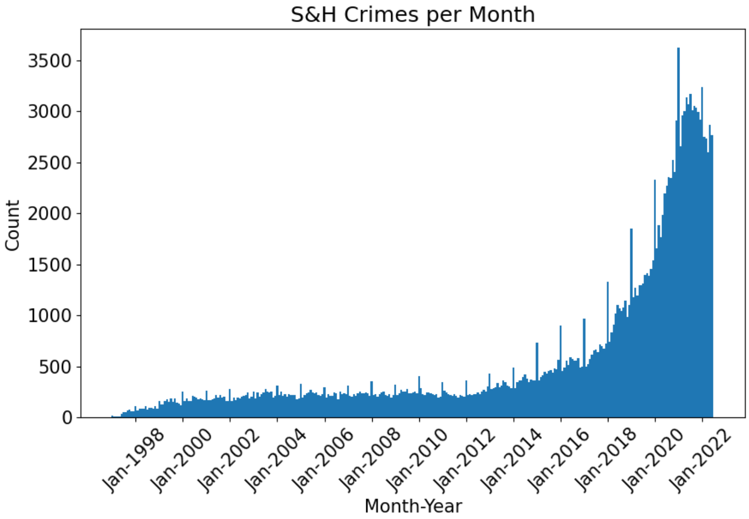


Figure 2 - S&H incident volume

Focussing specifically on Stalking incidents, it was found that there were 37,108 reported incidents of stalking. These incidents were perpetrated by 28,279 identified nominals. Of these, 15,097 incidents were specifically recorded as stalking, while an additional 13,203 incidents, originally recorded as harassment, were reclassified as stalking for the purposes of this project. This reclassification was implemented in cases where there existed a current or previous intimate relationship between the victim and offender, in alignment with current practices for recording stalking incidents. The inclusion of these reclassified incidents serves to expand the available data on the prevalence of stalking. The 28,279 nominals are identified in 420,901 other crimes. Figure 3 highlights the escalation in the volume of S&H crimes observed in the most recent 5 years.

It is evident when looking at Figure 2 and Figure 3 that S&H incidents have experienced a sharp increase in the most recent 5 years. There are numerous reasons behind this, one of which relates to recording practices prescribed by the Home Office[2]. In 2018, the counting rules and recording practices for S&H offences required police officers at the scene of an incident to record more harassment-based offending. The specific counting rules related to the recording of S&H incidents in addition to another substantive offence that might have been committed. In April 2020, the Home Office updated guidance such that any domestic abuse harassment offence should be recorded as stalking or controlling and coercive behaviour. Furthermore, in late 2020 the Force underwent a crime data integrity exercise to ensure domestic cases post 1st April 2020 were

[2] https://www.gov.uk/government/publications/counting-rules-for-recorded-crime

correctly recorded. These recording practices somewhat explain the escalation in the number of recorded incidents observed today. This is not to suggest that there is not a problem with the volume of S&H incidents, more to suggest that there is a better grasp of the scale of the problem.

The increase in the number of S&H incidents recorded is not solely attributed to changes in counting rules, but rather likely reflects the true scale of offending that was previously unobserved. It is important to note that the regular spikes observed in January of each year are primarily due to historical reporting practices, where offences that occurred on an unknown date are recorded as occurring on January 1st of that year. This highlights the significance of considering the limitations and potential biases of recording practices when interpreting crime and non-crime data.

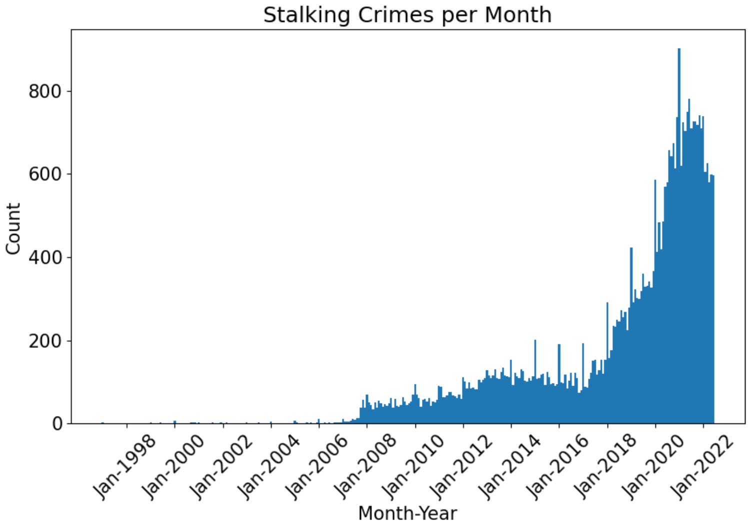


Figure 3 - Stalking incident volume

2.2. Demographics

It is important to have knowledge and understanding of the different demographics of both the nominals and the victims. Figure 4 shows the distribution of gender per S&H crime for nominals (the final two crimes displayed are *Controlling or Coercive Behaviour* and *Racial or Religious Aggravated Harassment)*.

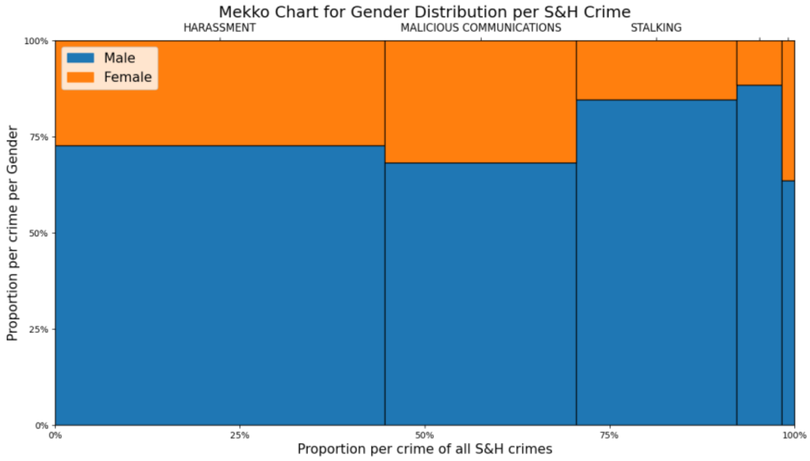


Figure 4 - Gender Distribution per S&H Crime

It is clear when looking at Figure 4 that Harassment has the greatest number of crimes in the S&H crime group, accounting for 44.6%, while Malicious Communications accounts for 25.9%, Stalking accounts for 21.7%, Controlling and Coercive Behaviour accounts for 6.1%, and Racial or Religious Aggravated Harassment accounts for 1.7%. It is also clear that males account for the majority of incidents in S&H.

It is also important to consider the ethnicities contained within the dataset, to mitigate the potential for biases carried forward into the modelling aspect of this project. Figure 5 shows the relative likelihood of each ethnicity compared to one another of a stalking crime. There are some instances where the relative likelihood between ethnicities indicate a higher potential to be a stalking nominal, where Any Other Black/African/Caribbean background vs White British shows a high relative risk of 9.01. Ethnicity is not used as a feature in any modelling, which will be discussed further in Section 2.3. The vast majority of these higher probabilities are related to the aggregation of ethnicities. A reason behind these numbers are the relatively low number of samples of some ethnicities compared to others. When comparing ethnicities to the most populous ethnicity ((white) English, etc.). There is little difference in the relative likelihood between ethnicities. Ethnicity is tested against the selected features in Section 3.2 to show no correlation with ethnicity.

An accompanying table containing the raw incident numbers can be found in Appendix 3. Ethnicity Demographics, along with relative likelihood information for victims.

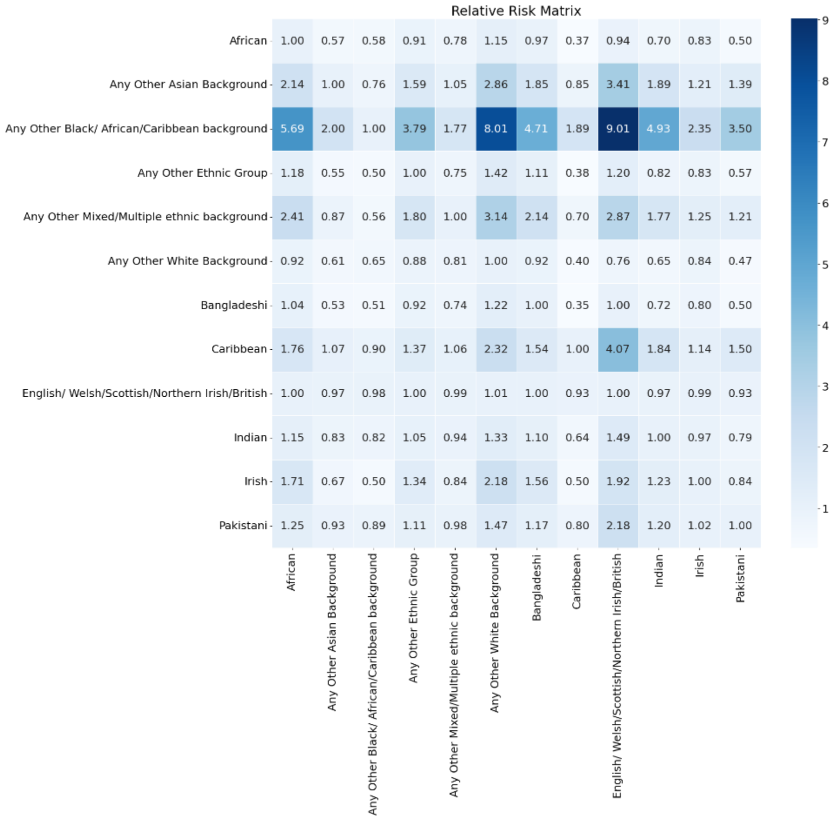


Figure 5 - Relative Likelihood of Stalking Offender per Ethnicity

Another important consideration for stalking nominals and victims is the age of the nominal at the time of the offence. As seen in Figure 6, the age of the nominal for both male and female follows an unsurprising profile, being similar to the profile of general offending. However, in this case, the peak between the ages of 20 and 40 is particularly prominent. This chart also illustrates the disparity between male and female nominals, with males making up an overwhelming majority.

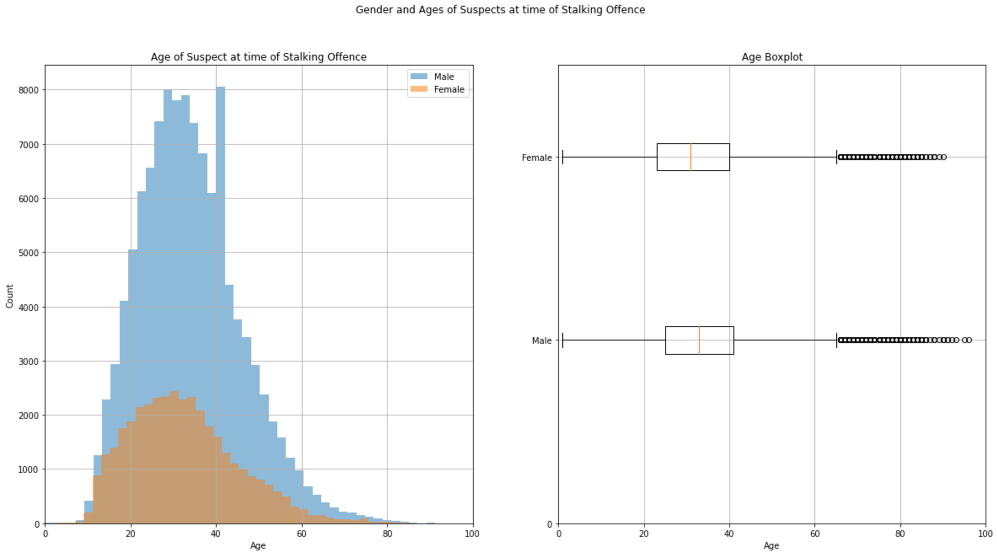


Figure 6 - S&H Nominal age distribution by Gender

Furthermore, Figure 7 shows the rate at which a female is a victim of stalking compared to the general population. It clearly shows that for the ages of 20 to 35, a female is over three times more likely to be a victim of stalking than at any other point during their lives. This peaks between the ages of 25 and 30, where they are nearly four times more likely.

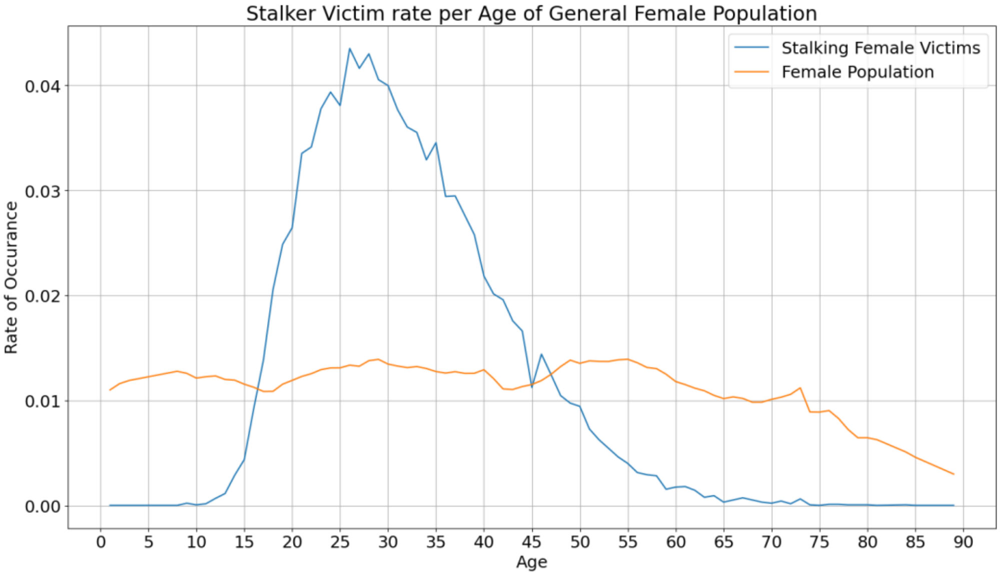


Figure 7 - Stalker Victim rate per age vs General Female Population

2.3. Victim and Nominal Harm

From this section onwards, analysis will focus only on stalking and not S&H as a whole for brevity.

The Cambridge Crime Harm Index (CCHI) allows a harm score to be assigned to each crime / incident, providing an understanding of the harm inflicted by nominals who commit stalking incidents. This includes not only the harm caused by the nominal, but also takes into account any harm a victim might experience.

Figure 8 illustrates the total harm inflicted by all stalking nominals when considering all other crimes / incidents they have committed, ranked from the lowest to highest harm score. The victim harm is also included, showing that a greater number of victims experience some level of harm and that the highest level of harm experienced is higher than the highest harm inflicted by an offender. This highlights the importance of considering both the offender's harm and the victim's harm in understanding the full impact of stalking crimes.

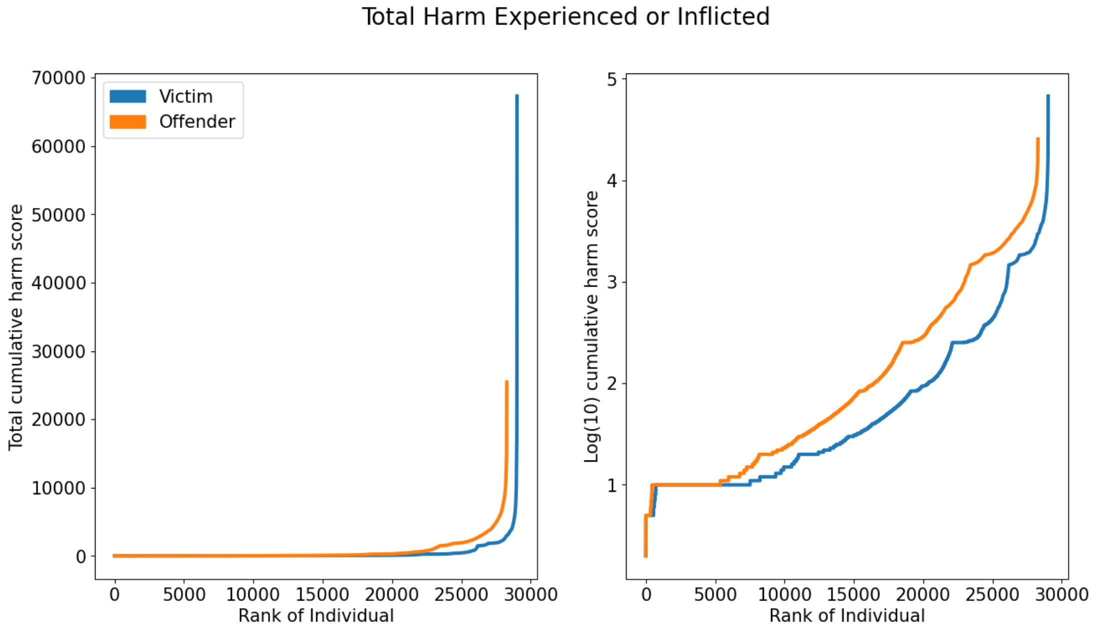


Figure 8 - Total Harm Victims vs Offenders

2.4. Victim Calls

Another important variable to consider in the case of stalking offences is the number of times a victim will reach out and contact WMP (see Figure 9 below). Since it is known that victims experience harm very differently to one another, with the top 10% experiencing significantly higher harm than the other 90%, the data presented in Figure 9 has been separated into deciles. The deciles are defined by the amount of harm a victim experienced in the 12 months following stalking day zero (stalking day zero is used throughout this report as a reference point in time for each nominal and represents their most recent stalking offence, all other events in time are measured in reference to this point). The number of calls presented is the average number of calls each victim made to WMP **before** stalking day zero (over differing periods of time). It is clear that all the victims shown are experiencing significant enough harm to need to reach out to WMP, however it is also noted that the top 20% are contacting WMP roughly twice that of any other decile.

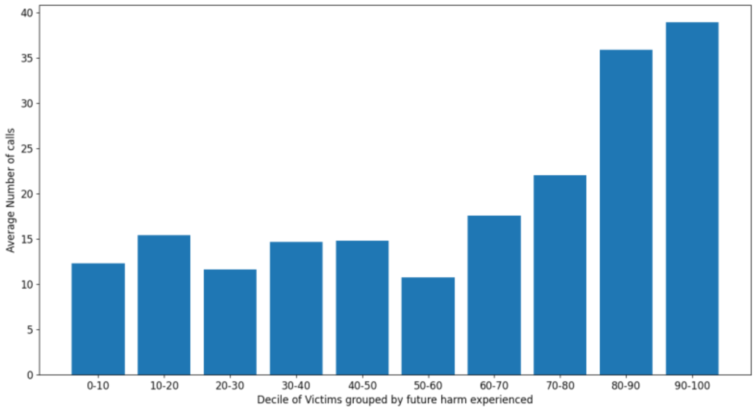


Figure 9 - Average calls made by victims grouped by decile of future harm experienced

It is worth noting, that the call volumes displayed in Figure 9 include 7,348 records where there were no recorded calls. This does not mean that these victims never contacted WMP (although it is possible), it is simply because the ControlWorks system (where records of contact (RoC) are recorded) and CONNECT (crime system) are not concretely linked, so where a confident link between the two was not possible, that victim would have zero associated calls. Also, the count for the calls is indiscriminate as to what the nature of the call is about, this is related again to the systems not linking and the fact that not every call leads to a crime, non-crime or recordable incident for attendance.

2.5. Escalation of Offending

The aim of this project is to use stalking incidents as a point from which to predict for escalation of offending, and therefore act as an indicator for intervention to protect victims. Therefore, it is important to develop an understanding of both harm leading up to a stalking incidence, and the harm inflicted or experienced following a stalking incidence.

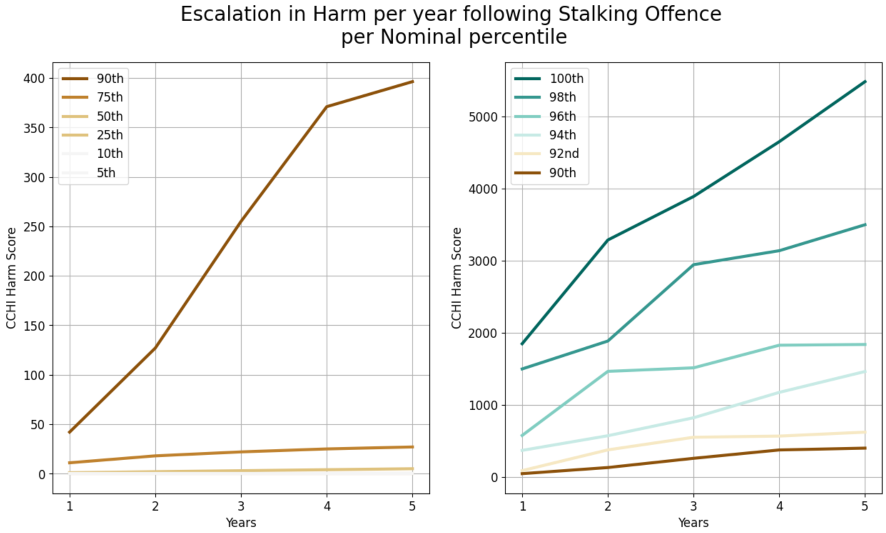


Figure 10 - Harm Escalation in years following Stalking Incident

The data presented in Figure 10 highlights the issue of escalating offending among those who perpetrate stalking. The majority of nominals commit no additional incidents or only a small number of low-harm incidents, as evidenced by the 75th percentile in the left chart (~5,600 out of ~7,500 in 2021). However, as the offender's percentile increases, the harm inflicted also escalates, as seen in the right chart, where the 90th percentile is included to demonstrate the scale of harm caused by the top 10% of offenders; with the 98th percentile contributing 12,800% more harm than the 75th percentile. It is clear that identifying these high-risk offenders at the point of their initial stalking offence would potentially provide an opportunity to intervene and prevent further harm.

2.5.1. Escalation Case Study

The data presented in Figure 11 shows an example of offending history of an individual with at least one stalking crime (the data represented in Figure 11 is entirely synthetic and does not relate back to an individual, however it is similar to patterns observed in nominals in the dataset). The red dotted line shows the day of their most recent stalking offence, which acts as a reference point to measure all other crimes from, this is stalking day zero.

Figure 11 shows it is clear that there is a pattern of offending within an intimate relationship, where there have been several assaults and harassment offences recorded before a stalking offence was recorded. Following this, there is a an escalation of offending where an aggravated burglary occurs, followed by harassment and threats to kill. The offending finally culminates in a rape of a female aged 16 or over. This is an example of someone who would be in the top 10% of future harm offenders in this dataset. The continual escalation in offending is clear to see, where the stalking crime prompted an escalation in further offending, and an escalation of harm.

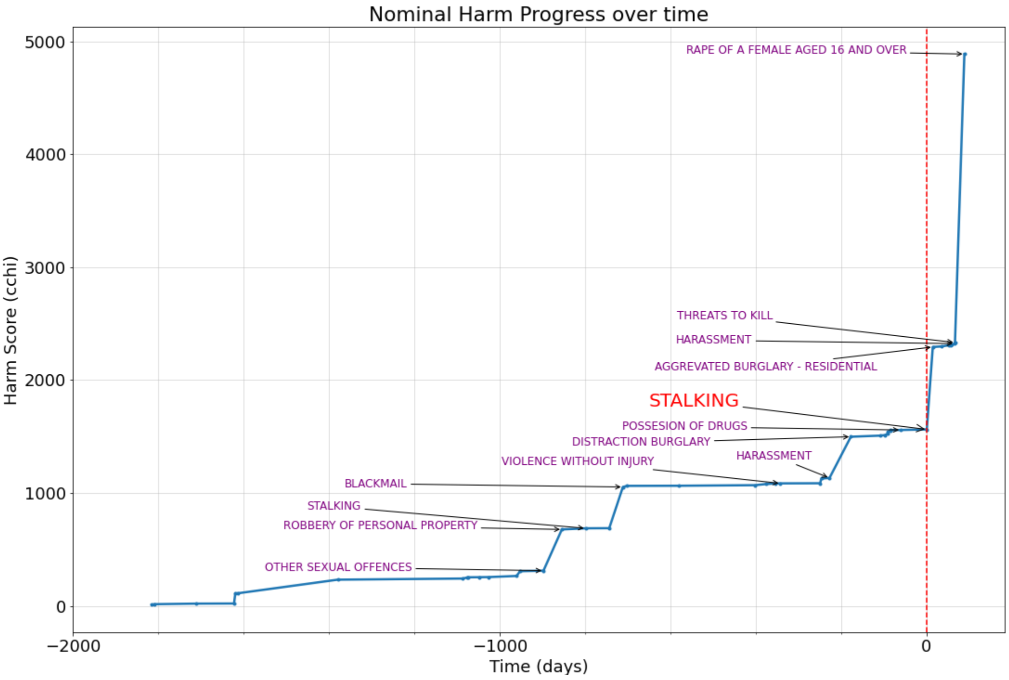


Figure 11 - Harm Escalation Case Study

3. Modelling Approach

3.1. Harm Score for Domestic Non-Crime

WMP records data on domestic incidents that they are called to respond to as part of their day to day recording activity. which has been normal practice since circa 2007, and was introduced to capture the entirety of Domestic Abuse (DA) reporting to give WMP an opportunity to analyse the volume of reports for DA victims, identify repeat victims, help risk assess and support them. These incidents are categorised as Non-Crime, meaning that they are recorded like a crime, but do not carry any financial or criminal justice outcomes. (Note: There are substantive offences for DA and Domestic Violence, this section accounts for the incidents where the threshold has not been met for the substantive offence). The nature of domestic incidents provides important context for understanding victim-offender relationships, making it a crucial aspect of this study. The data presented in Figure 11 demonstrate that domestic incidents can make up a significant portion of an individual's offending history, offering a deeper insight into patterns of offending. It was decided to include non-crime data in the model to reflect the sources of information used by officers in their assessment of risk.

However, due to the non-crime nature of some domestic incidents, they do not have an associated Cambridge Crime Harm Index (CCHI) score. Therefore, they do not contribute to the overall harm inflicted or experienced by an individual. It is recognized that any domestic incident requiring police attendance would have likely caused harm to the victim. Therefore, it was deemed necessary to develop a method to infer a harm score for these domestic incidents in order to provide a more comprehensive understanding of the harm inflicted.

The CCHI harm score is based on the minimum recommended custodial sentence in days (or equivalent working hours cost) for any given crime. The scores are broad ranging reflecting the number of different possible crimes, where a *Murder* has a score of 5,475 and *Theft from Shops and Stalls* has a score of 1.

It was decided to limit any harm score inferred to a maximum of 10, as this represents the lowest harm score for an *Assault with Injury* crime type. The reasoning behind this decision is that if the domestic incident was severe enough to cause injury, then it would likely have been recorded as an actual crime, with assault being an example. This approach ensures that any domestic incident recorded as non-crime is assessed as less severe than the lowest harm score for an *Assault with Injury* crime type.

The harm score is not a linear scale, with possible values of 1, 2, 3, 5, and 10; this lends the inference model towards a classification model. A dataset was cultivated from crimes where the harm score was less than 10, and all textual information from these incident logs were extracted. These logs were then cleaned and pre-processed to extract only the key words, which were used to train a classification model. This approach allows for the inference of a harm score for domestic incidents based on the severity of the incident as described in the key words extracted from the crime logs.

A simple six layer fully connected neural network was trained with the keywords from the crime records as the features, and the harm score as the target. This achieved an accuracy of 72.6% across the five classes. Following training, the keywords were extracted from the domestic non-crime incident logs and then processed on the trained neural network to infer a CCHI harm score for each non-crime incident.

Specific metrics and modelling regarding the harm score inference can be found in Appendix 8. Non-Crime Domestic Harm Score.

3.2. Main Model Feature Engineering

This project aimed to utilise all available sources of data to model the future harm caused by a stalker, including harm score, time between offences, and victim calls. Initially, it was believed that information extracted from the text-based crime records (that is records for all crimes and non-crimes committed by the nominals) would be useful for the model, however, during the modelling process it was determined that this information added confusion to the model rather than providing additional insight. As a result, the information extracted from the text logs was removed from the model. At the peak number of features, there were over 700 features, but the final set of features used for the model consisted of 19 features per incident per offender. This final set of features was determined to provide the best balance between accuracy and false positives of the models' predictions.

The use of prior crimes to gain a picture of potential future offending is suitable for (multivariate) time series modelling. There were two approaches that could be taken in this project. The first approach was to aggregate the offending history of an offender into set time intervals, while the second approach was to include the offending history at a per crime level for a defined number of offences.

The approach of aggregating crimes into set time windows showed promise in early results. This approach was tested by aggregating crimes into 4 x 6-month intervals until 24 months before the stalking day zero, with a fifth aggregate group for all crimes before that. This method of aggregation allows for the analysis of the offender's history of crimes in relation to the stalking day zero, providing insight into patterns of offending over time.

The second approach of using crime level data involved taking an offender's most recent stalking offence, and then the previous 19 crimes before that, resulting in a total of 20 crimes. For offenders who did not have 20 total crimes, this information was padded with zeros, so that all the information for every offender in the model was the same dimension. This approach allows for a more detailed analysis of the offender's history of crimes, specifically in relation to their most recent stalking offence. The final features used in the modelling can be found in Table 1. The definitions for each feature can be found in Appendix 2. Model Feature Definitions.

Approach two delivered better results when modelling on the data, and therefore was the approach taken forward in this project.

Table 1 - Model Features

Age at offence	Harassment offence (binary)
Harm Score (CCHI)	Violent offence (binary)
Time delta from most recent S&H offence	Sexual offence (binary)
Cumulative harm score	Stalking offence (binary)
Days between offences	Domestic offence (binary)
Harm rate	Offender (binary)
Harm momentum	Suspect (binary)
Harm cumulative with decay	Victim call count (where available)
Harm momentum with decay	Gender (factorised)
Custody appearances	

All the features denoted *(binary)* are given a 0 or 1 for the model to have context of what crime type they are. For *Offender* and *Suspect* this identifies whether the individual was the offender or suspect in each crime. **Note, all crimes where the individual is *suspect eliminated* are not included in any of the datasets.**

A multinomial logistic regression model was trained to understand whether ethnicity was correlated with any other features in the dataset. The regression found that there were no features which had substantial coefficients where ethnicity could have been indirectly included in the model. When calculating feature importance, the offender age had the greatest importance, but at 28%, was not substantial enough to inform the main model of ethnicity. **Ethnicity was not included in the modelling dataset**.

3.3. Classification Modelling

3.3.1. Dataset Preparation

The aim of this work was to use prior crimes to predict an escalation of offending following a stalking offence. The concept of stalking day zero is important in this context, as during model development, it is possible to look forward from stalking day zero to see what the offender went on to do. However, in deployment, this would not be possible. By using the data available, it is possible to use previous data to understand some indicators that may lead to an offender escalate in their harm.

As shown in Figure 8 and Figure 10, it is clear that the harm perpetrated by the top 10% of nominals is significantly larger than that of the other 90%. Therefore, this group of nominals would be the main target group. However, it would not be sufficient to simply predict if someone would be in the top group or not. Even if an offender was in the 89th percentile, it would still be useful to know about them and not ignore them. Therefore, the offenders were divided into 4 distinct unbalanced groups, 0-50%, 50-80%, 80-90%, 90-100%.

The definition of belonging to any of these groups was determined by the harm committed in the first 12 months following the most recent stalking incident, known as stalking day zero. The data regarding the harm in those 12 months was collected for every individual in the dataset, and then they were ranked in percentiles to decide which of the four groups they belonged to. This approach allowed for the identification of nominals who may be at a higher risk of committing harm and may lead to the prevention of the most harm if intervention and prevention efforts were successful

The usable dataset includes 28,279 nominals with a stalking incident. The distribution of nominals in each subgroup for classification is presented in Table 2. Additionally, Table 2 includes the training/testing split for the model training. The 0-50% group accounts for more than 50% of the total number because when specifically considering the 0th percentile, there are more than 15,174 offenders who have no offending history in the 1 year following stalking day zero.

Table 2 - Samples for model training and testing

Group	Total	Training	Testing	Proportion in Training
0-50%	21,595	1,743	1,593	52.2%
50-80%	3,932	1,743	457	79.2%
80-90%	1,310	1,052	258	80.3%
90-100%	1,442	1,159	283	80.4%

As previously discussed, the classes defined lead to an imbalanced dataset. The approach taken was to train the model on a *more* balanced dataset so that the model had the greatest chance of success in identifying differences between the classes, while the testing was conducted on an imbalanced dataset to better reflect the distribution of offending seen in the data.

3.3.2. Modelling Approach

The modelling approach for this project focused on developing a multi-class classification model that would receive the time-series crime data as input and would output the estimated offender class as the output. This approach meant that there were several different modelling approaches that could be used to suit this problem.

The main model types used were *Support Vector Machine (SVM), Multi-class (multinomial) Logistic Regression, Random Forest, Gradient Boosting Machine (GBM), XGBoost, Basic Neural Network (feed forward), Long-Short Term Memory (LSTM) Neural Network*, and *Convolutional Neural Network (CNN).* All of the models were trained and tested on their own; however, it was found that the results of all the models were improved by either creating model ensembles or by stacking the models and training a third stacked model. Ensemble methods combine the predictions of multiple models to produce more accurate results, while stacking combines the predictions of multiple models to create a new, more powerful model. This approach takes advantage of multiple models' strengths and improve upon their weaknesses. The baseline results for the individual models can be seen in Appendix 4. Baseline Model Results.

For the ensemble models, all the models were trained individually, and the probabilities of each class output. The probabilities of each class for each model was then evaluated two ways in order to select a class for the final prediction. Firstly, the probabilities of each class would be averaged across both models, with the greatest probability class gaining the prediction. Secondly, the maximum classification across the two models would be taken (e.g. if Random Forrest predicted 80-90% class, and XGBoost predicted 90-100% class, the prediction used would be 90-100%). The basic neural network was not used in the ensemble modelling, which resulted in 7 model types and 42 pairwise model combinations. See Figure 12 for more information.

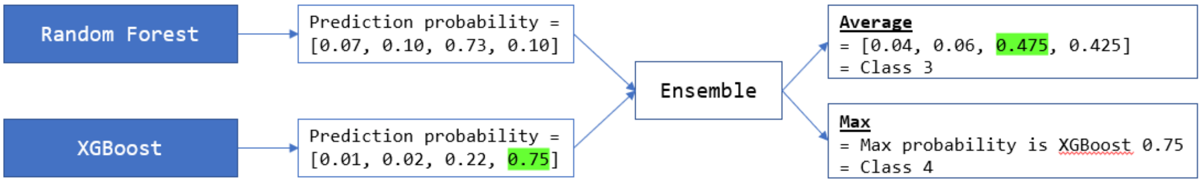

Random Forest
XGBoost
Prediction probability = [0.07, 0.10, 0.73, 0.10]
Prediction probability = [0.01, 0.02, 0.22, 0.75]
Ensemble
Average
= [0.04, 0.06, 0.475, 0.425]
= Class 3
Max
= Max probability is XGBoost 0.75
= Class 4

Figure 12 - Example of Ensemble process with Random Forest and XGBoost

For the stacked models, the process was similar. Models were trained individually to extract the prediction probabilities for each class (dimension of 4). Following this, comparing two models at a time, the prediction probabilities for each model was concatenated together to create a variable with a dimension of 8. This variable was then used as the input to train a new model in a second stage, with the output remaining as the prediction of the class. As before, the first stage of model training contained the same 7 model types used in the ensemble. For the second stage, the new model, trained on the output of the first stage, was any of the previously mentioned models (with the exception of LSTM and CNN as the data at this stage is no longer time-series). See Figure 13 for more information. The total possible combinations of these stacked models were 126.

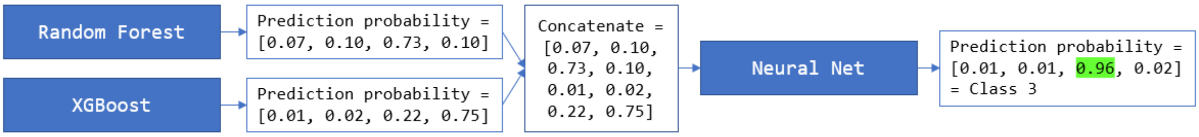

Random Forest
XGBoost
Prediction probability = [0.07, 0.10, 0.73, 0.10]
Prediction probability = [0.01, 0.02, 0.22, 0.75]
Concatenate = [0.07, 0.10, 0.73, 0.10, 0.01, 0.02, 0.22, 0.75]
Neural Net
Prediction probability = [0.01, 0.01, 0.96, 0.02]
= Class 3

Figure 13 - Example of Stacked model process with Random Forest, XGBoost and NN

Following the training of all these models, they were all evaluated in the same way to be able to select the most optimal model for the task at hand. More information regarding the specific model definitions can be found in Appendix 1. Model Definitions.

4. Results

The method for evaluating the results in this project needed to be balanced and bespoke around the level of harm captured by any model and the level of harm missed (when comparing to the test dataset). Due to testing the models on an imbalanced dataset, accuracy metrics for the overall results would not suffice or capture the nuances of the model correctly. The results for the baseline models before being used in ensemble and stacked models can be found in Appendix 4. Baseline Model Results.

As previously mentioned, the class of 90-100% was the most important to identify accurately, due to the level of harm this group perpetrate. Therefore, any model selection can be built around not only the performance of the model as a whole, but also the specific performance of the 90-100% class relating to the sensitivity and specificity in this class.

The chosen model was evaluated on the amount of harm that class 90-100% missed through false negatives, versus the amount of harm potentially *prevented* by false positives. The false negatives would be defined as having a true classification of 90-100% but a predicted class of 0-50% or 50-80%. The false positives would be defined as having a prediction class of 90-100% but a true class of 0-50% or 50-80%. An adjusted F1 score only taking the results in class 90-100% was also calculated for all models based on these defined false positives and false negatives. The results have been evaluated in this way as it is known that the 90-100% class contribute more harm than any other class, and is therefore the most important class (see Figure 10).

The harm calculation comes from a cumulative sum of what each nominal went on to commit in the 12 months after stalking day zero.

4.1. Classification results

Following the training and testing of 168 different model combinations, the best model selected was a stacked model consisting of an XGBoost and LSTM stacked into an SVM. Figure 14 shows the confusion matrix for the best model results. It is worth remembering that the model was tested on an imbalanced dataset to mimic the real-world distribution of classes.

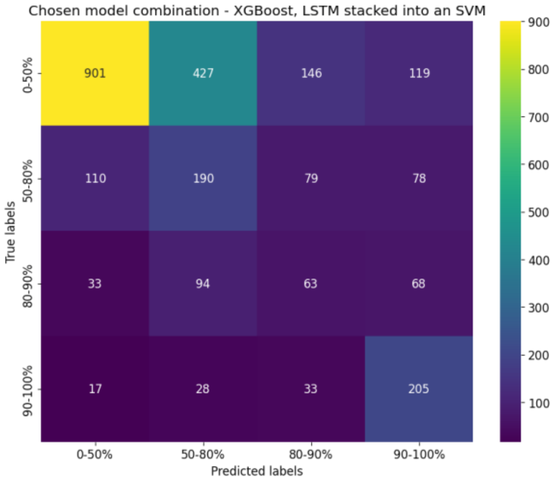


Figure 14 - Chosen Model Confusion Matrix

The measures chosen for model performance in this instance reflect the most likely situation of WMP being able to examine those estimated to be in the top harm decile (bearing in mind that the second top decile could also be the source of high harm). Looking to Figure 14, the bottom right corner is the most important aspect. This shows that out of 283 nominals who are known to have gone on to be high harm, the model correctly identified 205 (72%) of them, with a further

33 offenders flagged in the next level of harm, this gives a top 2 accuracy of 84%. This also means that 16% of high harm offenders are *missed* by the model, assuming that the lower two classes would receive little attention. The selected model returned an adjusted F1 score of 0.682. For the matrix above, the calculation would be[3]:

$$Precision = \frac{True\ Positives}{Predicted\ Positives}$$

$$Precision = \frac{68 + 205}{119 + 78 + 68 + 205} = 0.581$$

$$Sensitivity = \frac{Correct\ True\ Positives}{Actual\ Positives}$$

$$Sensitivity = \frac{63 + 68 + 33 + 205}{33 + 94 + 63 + 68 + 17 + 28 + 33 + 205} = 0.682$$

$$F1\ Score = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity}$$

$$F1\ Score = 2 * \frac{0.581 * 0.682}{0.581 + 0.682} = 0.627$$

The model also estimates a number of false positives, where 119 nominals were identified as belonging to the highest harm group, while the crimes they went on to commit put them in the lowest harm group.

The challenge with the final model arose from choosing a model which had an acceptable balance of false positives to false negatives. There was one model combination which provided a top 2 accuracy of 87%, however, there were more false positives which would begin to be an unmanageable number of cases for practitioners to work through. The F1 score for that model was 0.626, showing that the balance of results is worse than the chosen model.

Therefore, by relating the false positives and false negatives back to the harm that the individuals went on to commit would create a system where the model was selected on the net harm inflicted in the 12 months after stalking day zero in conjunction with accuracy measures. Referring back to Figure 14, the total harm *missed* by the model is the total harm inflicted by nominals in the bottom two left most cells (17+28) covering the bottom row. The additional total harm *prevented* by the model is the total harm inflicted by nominals in the top two right most cells (119+78) covering the first two rows (*prevented* under the assumption that the false positive would have led to some intervention due to the high harm prediction). The harm scores were calculated for each nominal, looking forward 12 months from stalking day zero, and totalling the harm each nominal committed in that period. The total harm was then calculated for each group by adding the total nominal harm scores together. In the selected model, the total harm missed totalled 47,456.5, and the total prevented harm totalled 33,837.5, resulting in a net cost of 13,619 harm. In general, this measure shows the balance between false positive and false negatives so the smaller the difference, the more in balance the model is in terms of the false determinations

[3] The precision is calculated over the column of the "positive" predictions whilst the sensitivity is calculated over both the bottom rows (as the two top groups are included); for the sensitivity therefore, this is the equivalent of changing the confusion matrix from 4 x 4 to 2 x 2.

regarding harm; however, if this difference was to be minimised by way of increasing the values in the additional prevented harm, the greater the probability of unwarranted extra workload for WMP. A full table of all model results can be found in Appendix 5. All results for Stalking Nominal Model.

4.2. Analysis of Missed Crimes

While it is useful to quantify the *missed* and *prevented* crimes by way of using the harm score, this does not offer full context in terms of the actual crimes which were missed. Furthermore, it does not take into consideration as to whether the offenders of the *missed* crimes would have realistically been identified through other methods. This section will look into a few examples of what is present in the *missed* offenders in the red rectangle in Figure 14.

Firstly, looking at the missed crimes, the total test set of 283 high harm offenders contained 50 rape crimes. Of these crimes, 40 nominals would have been correctly identified and classified as being in the highest harm group, theoretically before they committed the rape offence, creating an intervention opportunity. 10% (5) of the rape offenders would have been classified as either 0-50% or 50-80%, theoretically resulting in a *missed* intervention.

It is worth noting, that in most cases of rape in this dataset, the victim of the subsequent rape following stalking day zero is **not** the same person who was the victim of stalking.

When looking deeper in to the *missed* nominals who went on to rape, it is easy to see why the model didn't identify them. All five individuals who were not identified and went on to commit rape had **zero** previous crimes, other than the stalking crime which highlighted them to be included in the dataset in the first instance. It could therefore be argued that there is very little likelihood that a subject matter expert would be able to identify these individuals either.

This pattern was similar with all of the *missed* offenders, where there wasn't sufficient criminal history of the offenders to make a reliable prediction. Other than rape and murder, the type of offences which would categorise an individual in the highest harm group would be a single serious assault offence, multiple minor assaults, or other high frequency, medium harm offences.

In the whole test set, there were only two offenders who went on to commit a murder in the 12 months following stalking day zero. The model correctly identified both of these offenders as belonging to the highest harm group (before the murders were committed).

5. Time-to-Event Modelling

Identifying the most potentially harmful nominals is only part of the problem, it would also be useful to be able to understand when these nominals are likely to escalate in terms of time. The classification modelling identifying the nominal harm group is based on the estimated harm in the subsequent 12 months following their most recent stalking offence. In this section a survival model to estimate the probability of a nominal committing a crime on any given day in the 12 months following the most recent stalking offence is discussed.

The survival model used was a Random Forrest Survival model. The data used for training and testing was an imbalanced dataset of nominals who did not go on to commit further crimes, and nominals who belong to the highest harm group in the subsequent 12 months. The time scale was censored on 365 days, therefore anything above this meant no crime occurred in the time frame, while the count in days between stalking and the highest harm crime for the high harm group was provided. This count in days provided the target variable for the survival modelling.

There were 1,601 training samples with 1,038 samples censored and 563 samples where a crime did occur inside 12 months. There were 687 testing samples with 460 censored samples and 227 non-censored samples.

In terms of features, each nominal had much the same feature set as used in the main stalking harm model, whereby the features for the 20 previous crimes were included. Not all the features used in the main harm model were used in the survival model, this was due to trial and error when tuning the survival model. The features in the survival model are seen in Table 3.

Table 3 - Survival Model Features

Age at Offence	Harm momentum
CCHI Harm score for the crime	Cumulative harm decay
Time delta	Harm momentum decay
Harm score cumulative sum	Victim call count
Days between offences	Custody count
Harm rate	

The random forest survival model achieved a score of 70.2% on the test set using the Concordance Index. The Concordance Index is a measure of *how in order* a set of predictions are. Since the stalker harm model will be used as a prioritisation tool, this is an important measure. Therefore, this means that of the 687 samples in the test set, it predicted the majority of samples to be in the correct order.

Some examples of the survival model output can be seen in Figure 15. In this figure, there are 10 samples from the test set plotted, where the plot for each sample is the survival function for each nominal. This graph is to be interpreted as the probability that a nominal does **not** commit a further crime on any given day in the 12 months following their most recent stalking offence; therefore, the higher the line is on the chart, the less likely there are to commit an additional offence in the given time frame.

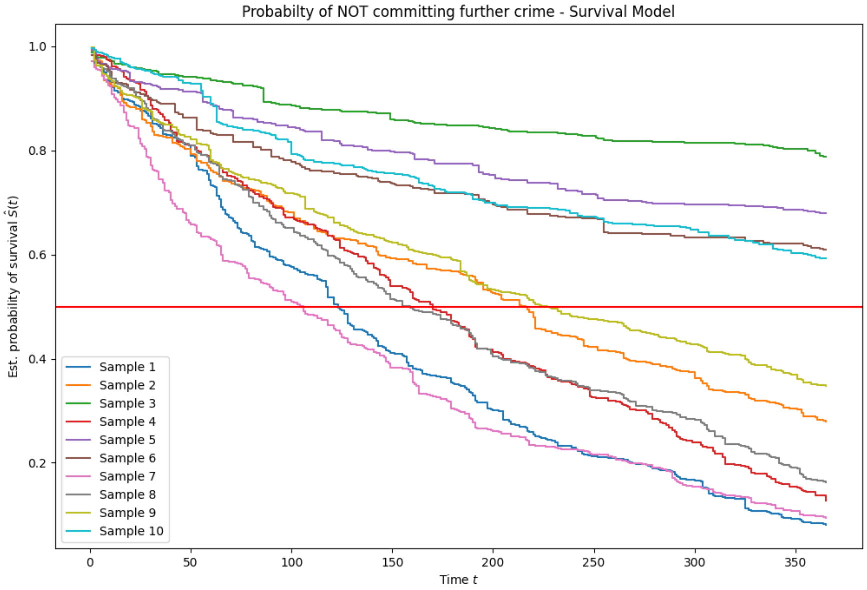


Figure 15 - Survival Model results

Looking at some particular examples in Figure 15, the green sample (sample 3) is a sample of an individual who belongs to the 0-50% future harm group, and the survival model reflects this where by the prediction shows they are very unlikely to commit any further crime in the time frame with their final survival probability being 0.8 (thus only 20% probabilty of committing further crime in time frame). Conversely, the pink sample (sample 7) shows a very different picture, whereby their final survival probability is 0.05 (thus 95% probability of committing further crime in time frame); furthermore, this sample has a 50% probability of committing further crime after 110 days.

6. Stalking Tool in Use

It is envisaged that this stalking and harassment harm escalation tool will be utilised by the Stalking Triage Clinic (STC). The STC was set up by the Force lead on Domestic Abuse and Domestic Abuse Stalking. The aim of the clinic is to provide a forum for officers and community partners to come together to discuss concerning stalking cases and assist in the decision-making process when dealing with dangerous individuals.

The partners who attend the clinic include the Crown Prosecution Service (CPS), Black Country Women's Aid, representatives from local borough councils, and academics with expertise in the field. Currently, the clinic is led by investigating officers voluntarily bringing their casework to the clinic when they are particularly concerned about the victim or nominal in their stalking case. Each case is then discussed by the clinic whereby decisions are made as to the best course of action, sometimes leading to arrest recommendations after consultation with CPS.

As the current working model of the STC is led by officers bringing cases themselves, it leaves the possibility that some of the most at-risk cases might not be seen or heard in the clinic, as it relies upon the investigating officer to identify that risk. This poses a risk both to the victim of the original stalking offence, but also the wider general public, as we know that subsequent crimes perpetrated by nominals are not always in the same nominal-victim pair.

Therefore, the proposal is to integrate the stalking escalation tool into the workings of the STC. The current working of the STC will remain the same, whereby concerned officers will be permitted to bring any case that concerns them to the clinic.

In addition to this, the results of the project will be used to assess and categorise all stalking incidents coming into the Force. The Public Protection Unit (PPU) will then use this to summon the investigating officer for each high-risk case to come to the clinic and comment on their case. This is where the professional judgement of the officer in charge and the expertise in the clinic will help to filter out false positives suggested by the model. Furthermore, this framework of working would ensure that 84% of the most high-risk stalkers who go on to commit high harm crimes within the next 12 months would have been discussed in the clinic, furthering the opportunity of intervention, and providing a better service to victims and the general public.

When considering the potential workload this model will introduce, a rough indication can be calculated as follows. There is a mean of 147 new stalking crimes recorded per week in the Force. Looking back to Figure 14, reading each predicted class vertically, the model would predict roughly 18% of all cases being high risk, equating to circa 27 crimes. Of these 27 crimes, 58% of them would belong to the two highest harm groups, accounting for roughly 16 genuinely high-risk offenders per week.

Due to changes in the Home Office counting rules regarding stalking and harassment, the analyses will need to be rebuilt due to associated changing definitions, so this report is provided to highlight the basic framework of the modelling approach.

Appendix 1. Model Definitions

This appendix will outline the models used in Section 3.3.2 in greater detail. The models used are *Support Vector Machine (SVM), Multi-class Logistic Regression, Random Forest, Gradient Boosting Machine (GBM), XGBoost, Basic Neural Network, Long-Short Term Memory (LSTM) Neural Network*, and *Convolutional Neural Network (CNN).*

The models used in the ensemble architecture (Figure 12) were SVM, *Multi-class Logistic Regression, Random Forest, Gradient Boosting Machine (GBM), XGBoost, LSTM and CNN.*

The same models were used in the first stage in the stacked architecture (Figure 13), however the second stage included the Basic Neural Network, but removed the LSTM and CNN.

SVM:

Support Vector Machine (SVM) is a type of supervised learning algorithm that can be used for classification and regression tasks. The algorithm works by finding the hyperplane in a high-dimensional space that maximally separates the different classes. Data points closest to the hyperplane are called support vectors and have the greatest impact on the position of the hyperplane. SVM can handle non-linearly separable data by transforming it into a higher dimensional space using a kernel.

Multinomial Logistic Regression:

Multinomial logistic regression is a variation of logistic regression, which is used for classification tasks where the outcome can take multiple class labels rather than just two. The model uses multiple binary logistic regression models, one for each class, to predict the probability of each class. The class with the highest probability is then chosen as the final prediction. The model can be trained using a one-vs-all or a softmax function. The one-vs-all method trains a separate binary classifier for each class, while the softmax function trains a single model for all classes.

Random Forest:

A random forest is an ensemble learning method for classification and regression. It creates multiple decision trees, and each tree makes a prediction independently. The final prediction is the majority vote of the predictions of all the trees. Random forests are useful in dealing with overfitting and high variance in decision trees.

GBM:

Gradient Boosting Machine (GBM) is a type of ensemble learning algorithm that can be used for both regression and classification tasks. The algorithm works by combining multiple decision trees in a way that minimizes the overall prediction error. It does this by training multiple weak learners in a sequential manner, with each new learner trying to correct the mistakes made by the previous learners. GBM uses an optimization technique called gradient descent to minimize the prediction error, which is why it is called "gradient boosting." GBMs have been shown to be very effective in many real-world applications.

XGBoost:

XGBoost is an optimised version of the GBM algorithm. It is a powerful and widely-used tool for both regression and classification problems. XGBoost is an implementation of GBM that is optimized for speed and performance, making it faster and more efficient than other GBM implementations.

LSTM:

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that is designed to handle sequential data and maintain a long-term memory. RNNs are neural networks that process sequential data by looping over the same parameters for each element in the sequence. However, the traditional RNNs have difficulty in retaining long-term information, known as the vanishing gradient problem. LSTMs solve this problem by introducing a memory cell, gates (input, output and forget gate) to control the flow of information into and out of the cell, and a hidden state. It enables the LSTM to selectively retain or forget information over time, making it particularly useful for tasks such as language modelling, speech recognition, and time series forecasting. A source for a more detailed explanation of how LSTMs work can be found in the references (Olah, 2015).

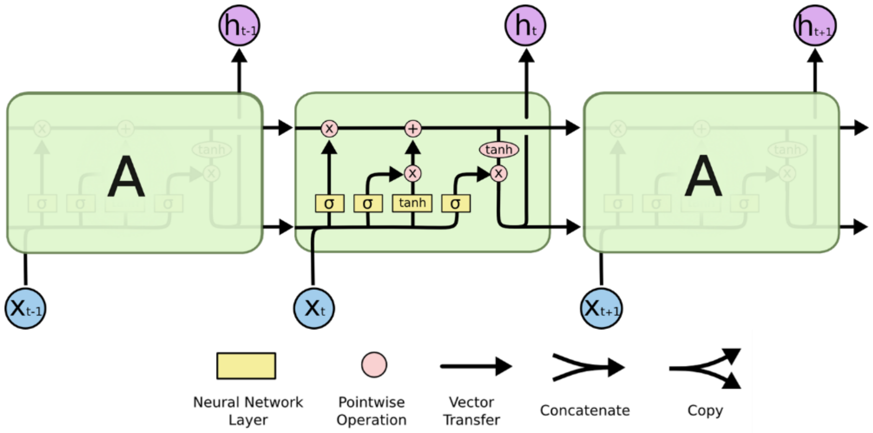


Figure 16 - Example of LSTM Architecture (Olah, 2015)

CNN:

A Convolutional Neural Network (CNN) is a type of deep learning neural network that is primarily used for image and video recognition tasks, however can be used for any pattern recognition task. CNNs are designed to process data that has a grid-like topology, such as an image. They have a unique architecture that includes layers such as convolutional layers, pooling layers, and fully connected layers. The convolutional layers are responsible for detecting features in the input data, the pooling layers are used for down-sampling, and the fully connected layers are used for classification. CNNs use a process called convolution, which involves applying a set of filters to the input data to extract features at different scales. The combination of these layers allows CNNs to automatically and adaptively learn spatial hierarchies of features from input data (in terms of the feature space, not geographical space).

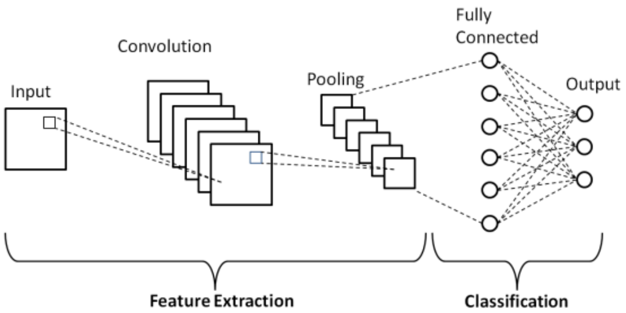


Figure 17 - Example of CNN architecture (Phung and Rhee, 2019)

Appendix 2. Model Feature Definitions

Table 4 - Feature Definitions

Feature	Definition
Age at offence	The age of each nominal at each incident.
Harm Score (CCHI)	The harm score associated with the crime the nominal was involved in.
Time delta from most recent S&H offence	The time in days between each incident for the nominal and the most recent S&H offence.
Cumulative harm score	The cumulative harm score at each incident measured from the start of offending from the nominal.
Days between offences	The time in days between individual incidents.
Harm rate	The proportion of harm score each incident contributed to the nominal's total harm score.
Harm momentum	The harm rate divided by the number of days between offending to give an indication of acceleration or deceleration in harm.
Harm cumulative with decay	Same as above with an exponential decay function applied. The exponential decay applies a reduction of 1% for every day after the incident, with a 30 day cooling off period before the decay starts.
Harm momentum with decay	Same as above with an exponential decay function applied. The exponential decay applies a reduction of 1% for every day after the incident, with a 30 day cooling off period before the decay starts.
Custody appearances	The number of times the nominal has been in custody. Nominal model only, not included in the victim model.
Harassment offence (binary)	A yes/no field to denote whether the incident was harassment.
Violent offence (binary)	A yes/no field to denote whether the incident was violence.
Sexual offence (binary)	A yes/no field to denote whether the incident was sexual.
Stalking offence (binary)	A yes/no field to denote whether the incident was stalking.
Domestic offence (binary)	A yes/no field to denote whether the incident was domestic.
Offender (binary)	A yes/no field to denote whether the nominal was the offender.
Suspect (binary)	A yes/no field to denote whether the nominal was the suspect.
Victim call count (where available)	The number of times the stalking victim on the stalking day zero crime has made contact to WMP.
Gender (factorised)	The gender of the nominal.

Appendix 3. Ethnicity Demographics

$$Relative\ Risk = \frac{A}{B} / \frac{C}{D}$$

Where A is the stalker population per ethnicity, B is the total stalking and harassment population, C is the ethnicity population for the total population, and D is the total population.

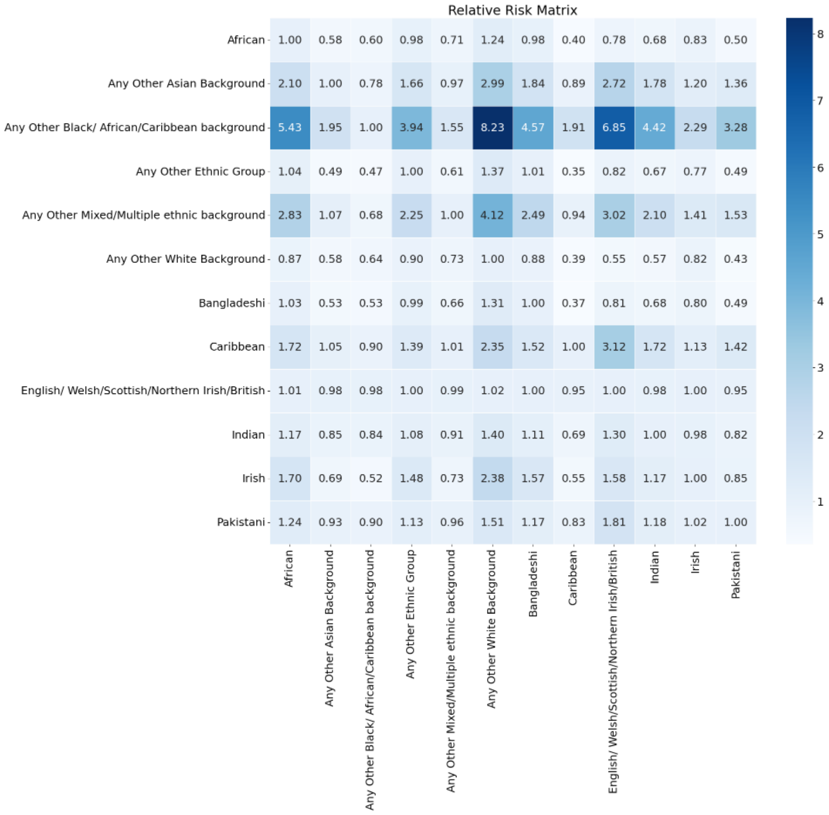
$$Relative\ Risk_{English} = \left(\frac{37365}{57283}\right) / \left(\frac{4471435}{5797300}\right) = 0.8387$$

$$Confidence\ Interval = e^{\ln(RR) \pm 1.96 \pm \sqrt{\frac{((C-A)/A)}{((C-A)+A)} + \frac{((D-C)/B)}{((D-C)+B)}}}$$

The relative risk for the matrix plot between ethnicities is calculated from the data contained in Table 5 and Table 6. The formula for relative risk is fundamentally the same with A being the stalker population per ethnicity for ethnicity i, B is the total stalking and harassment population, C is the stalker population per ethnicity for ethnicity j, D is the total stalking and harassment population.

Table 5 – Stalker Nominal Ethnicity Populations

Ethnicity	Stalker Population	West Mids Population	Non Stalker pop	Relative Risk	Ci 5%	Ci 95%
English/ Welsh/Scottish/ Northern Irish/British	37365	4508800	4471435	0.8387	0.8279	0.8496
Pakistani	5258	270700	265442	1.9658	1.9115	2.0215
Caribbean	4014	110200	106186	3.6863	3.5722	3.8041
Indian	3293	259600	256307	1.2838	1.2397	1.3294
Any Other White Background	1502	241800	240298	0.6287	0.5974	0.6616
Any Other Asian Background	1467	50500	49033	2.9399	2.7935	3.0940
African	1036	133500	132464	0.7854	0.7388	0.8349
Any Other Black/ African/Caribbean background	916	12000	11084	7.7253	7.2553	8.2257
Bangladeshi	753	90900	90147	0.8384	0.7804	0.9006
Any Other Ethnic Group	661	66200	65539	1.0105	0.9363	1.0906
Irish	517	32200	31683	1.6249	1.4912	1.7707
Any Other Mixed/Multiple ethnic background	501	20900	20399	2.4260	2.2241	2.6462



	African	Any Other Asian Background	Any Other Black/ African/Caribbean background	Any Other Ethnic Group	Any Other Mixed/Multiple ethnic background	Any Other White Background	Bangladeshi	Caribbean	English/ Welsh/Scottish/Northern Irish/British	Indian	Irish	Pakistani
African	1.00	0.58	0.60	0.98	0.71	1.24	0.98	0.40	0.78	0.68	0.83	0.50
Any Other Asian Background	2.10	1.00	0.78	1.66	0.97	2.99	1.84	0.89	2.72	1.78	1.20	1.36
Any Other Black/ African/Caribbean background	5.43	1.95	1.00	3.94	1.55	8.23	4.57	1.91	6.85	4.42	2.29	3.28
Any Other Ethnic Group	1.04	0.49	0.47	1.00	0.61	1.37	1.01	0.35	0.82	0.67	0.77	0.49
Any Other Mixed/Multiple ethnic background	2.83	1.07	0.68	2.25	1.00	4.12	2.49	0.94	3.02	2.10	1.41	1.53
Any Other White Background	0.87	0.58	0.64	0.90	0.73	1.00	0.88	0.39	0.55	0.57	0.82	0.43
Bangladeshi	1.03	0.53	0.53	0.99	0.66	1.31	1.00	0.37	0.81	0.68	0.80	0.49
Caribbean	1.72	1.05	0.90	1.39	1.01	2.35	1.52	1.00	3.12	1.72	1.13	1.42
English/ Welsh/Scottish/Northern Irish/British	1.01	0.98	0.98	1.00	0.99	1.02	1.00	0.95	1.00	0.98	1.00	0.95
Indian	1.17	0.85	0.84	1.08	0.91	1.40	1.11	0.69	1.30	1.00	0.98	0.82
Irish	1.70	0.69	0.52	1.48	0.73	2.38	1.57	0.55	1.58	1.17	1.00	0.85
Pakistani	1.24	0.93	0.90	1.13	0.96	1.51	1.17	0.83	1.81	1.18	1.02	1.00

Figure 18 - Relative Risk of Stalking Victim per Ethnicity

Table 6 - Stalker Victim Ethnicity Populations

Ethnicity	Victim Population	West Mids pop	Non Victim Pop	Relative Risk	Ci 5%	Ci 95%
English/ Welsh/Scottish/Northern Irish/British	23610	4508800	4485190	0.9015	0.8866	0.9166
Pakistani	2703	270700	267997	1.719	1.6533	1.7874
Caribbean	1896	110200	108304	2.9619	2.8291	3.1010
Indian	1802	259600	257798	1.195	1.1399	1.2528
Any Other Asian Background	733	50500	49767	2.4988	2.3237	2.6871
Any Other White Background	686	241800	241114	0.4884	0.4529	0.5267
African	539	133500	132961	0.6951	0.6385	0.7567
Any Other Black/ African/Caribbean background	437	12000	11563	6.2693	5.7145	6.8779
Bangladeshi	383	90900	90517	0.7254	0.6560	0.8021
Any Other Mixed/Multiple ethnic background	334	20900	20566	2.7512	2.4722	3.0616
Any Other Ethnic Group	285	66200	65915	0.7411	0.6598	0.8326
Irish	267	32200	31933	1.4275	1.2662	1.6094

Appendix 4. Baseline Model Results

Table 7 - Baseline Model Results

Model Name	Hamming Loss	Adjusted F1 Score	Max Correct	Max False	Max Missed	Missed Harm	Positive Harm	Net Harm
XGB	0.453	0.606	192	105	21	59984.5	32786.5	27198.0
GBM	0.438	0.559	196	141	33	77286.0	33112.5	44173.5
RF	0.499	0.539	173	146	31	101022.5	27819.5	73203
LSTM	0.533	0.450	152	341	42	154844.5	45126.5	109718
LR	0.480	0.427	86	92	62	239433.0	27572.0	211861.0
CNN	0.496	0.419	87	133	72	221377.5	28709.5	192668
SVM	0.484	0.339	68	64	68	297635.5	12496.5	285139.0

Appendix 5. All results for Stalking Nominal Model

Table 8 - All Model Results for Stalking Nominal Model

Model Name	Model Type	Hamming Loss	Adjusted F1	Max Correct	Max False	Max Missed	Missed Harm	Positive Harm	Net Harm
[('XGB', 'LSTM'), 'max']	ensemble	0.552	0.527	233	392	16	34922	57925	-23003
[('GBM', 'LSTM'), 'max']	ensemble	0.547	0.507	227	416	22	48377.5	59680	-11302.5
[('XGB', 'CNN'), 'max']	ensemble	0.523	0.579	213	213	16	38951	45833	-6882
[('RF', 'XGB'), 'max']	ensemble	0.526	0.603	222	181	13	37274	40504.5	-3230.5
[('RF', 'XGB'), 'RF']	stacked	0.515	0.626	207	152	22	40151.5	37375.5	2776
[('RF', 'LSTM'), 'max']	ensemble	0.587	0.500	220	424	17	60843	56654	4189
[('GBM', 'XGB'), 'max']	ensemble	0.476	0.606	220	165	16	42895.5	37903.5	4992
[('LR', 'XGB'), 'max']	ensemble	0.499	0.605	212	166	18	46053.5	40358	5695.5
[('SVM', 'XGB'), 'max']	ensemble	0.494	0.594	212	154	17	50024	39249.5	10774.5
[('GBM', 'CNN'), 'max']	ensemble	0.508	0.543	205	239	27	59372	47956	11416
[('RF', 'GBM'), 'max']	ensemble	0.521	0.576	215	204	22	52037.5	39600.5	12437
[('GBM', 'XGB'), 'NN']	stacked	0.497	0.622	206	142	17	48826.5	35751.5	13075
[('XGB', 'LSTM'), 'SVM']	**stacked**	**0.475**	**0.627**	**205**	**119**	**17**	**47456.5**	**33837.5**	**13619**
[('XGB', 'CNN'), 'SVM']	stacked	0.474	0.623	208	133	18	48534	34854.5	13679.5
[('LR', 'XGB'), 'RF']	stacked	0.491	0.613	196	125	19	48653.5	34718.5	13935
[('XGB', 'CNN'), 'RF']	stacked	0.496	0.623	197	116	14	47858.5	33083.5	14775
[('RF', 'XGB'), 'NN']	stacked	0.542	0.626	194	129	12	47991.5	33150.5	14841
[('LR', 'XGB'), 'SVM']	stacked	0.474	0.627	199	116	18	48619.5	33190.5	15429
[('XGB', 'LSTM'), 'RF']	stacked	0.488	0.616	196	119	17	50887.5	34211.5	16676
[('GBM', 'XGB'), 'SVM']	stacked	0.482	0.633	197	112	18	49542	32716.5	16825.5
[('SVM', 'XGB'), 'RF']	stacked	0.488	0.612	194	121	17	51071	33925.5	17145.5
[('SVM', 'XGB'), 'SVM']	stacked	0.470	0.616	202	121	21	51806	34186.5	17619.5
[('GBM', 'XGB'), 'RF']	stacked	0.489	0.618	189	111	18	50936.5	33186.5	17750
[('LR', 'XGB'), 'GBM']	stacked	0.486	0.604	204	129	16	52172	33976.5	18195.5
[('XGB', 'CNN'), 'XGB']	stacked	0.487	0.611	202	123	17	52597	34298.5	18298.5

[('XGB', 'LSTM'), 'GBM']	stacked	0.492	0.607	205	131	17	53659	35182.5	18476.5
[('GBM', 'XGB'), 'GBM']	stacked	0.479	0.603	203	126	15	54710.5	34899.5	19811
[('RF', 'XGB'), 'SVM']	stacked	0.535	0.609	201	137	14	55413	34336.5	21076.5
[('SVM', 'XGB'), 'GBM']	stacked	0.474	0.610	201	119	16	55268	33903.5	21364.5
[('XGB', 'CNN'), 'GBM']	stacked	0.487	0.610	203	122	17	56225	34345.5	21879.5
[('SVM', 'XGB'), 'XGB']	stacked	0.482	0.607	199	126	18	56835.5	34196.5	22639
[('RF', 'CNN'), 'max']	ensemble	0.557	0.546	196	235	21	67051.5	44262	22789.5
[('RF', 'XGB'), 'LR']	stacked	0.514	0.607	202	128	15	58085	33523.5	24561.5
[('SVM', 'XGB'), 'LR']	stacked	0.478	0.621	193	104	16	57241.5	32491.5	24750
[('XGB', 'LSTM'), 'XGB']	stacked	0.494	0.608	197	120	16	58998	34225.5	24772.5
[('XGB', 'LSTM'), 'LR']	stacked	0.486	0.617	190	100	18	56718	31768.5	24949.5
[('GBM', 'XGB'), 'XGB']	stacked	0.488	0.602	197	122	16	59348	34155.5	25192.5
[('LR', 'XGB'), 'XGB']	stacked	0.497	0.600	195	126	17	59150.5	33713.5	25437
[('LR', 'XGB'), 'LR']	stacked	0.476	0.616	191	103	20	57967.5	32440.5	25527
[('RF', 'GBM'), 'RF']	stacked	0.525	0.605	143	120	32	50147.5	24490.5	25657
[('GBM', 'XGB'), 'LR']	stacked	0.475	0.620	191	103	21	58354.5	32299.5	26055
[('LR', 'GBM'), 'max']	ensemble	0.484	0.569	209	192	31	66774	40177	26597
[('SVM', 'RF'), 'RF']	stacked	0.525	0.600	158	138	29	57389	29441.5	27947.5
[('SVM', 'XGB'), 'NN']	stacked	0.490	0.621	189	103	14	60966	32121.5	28844.5
[('GBM', 'XGB'), 'avg']	ensemble	0.433	0.608	202	111	25	60437	30233.5	30203.5
[('SVM', 'GBM'), 'max']	ensemble	0.479	0.561	210	186	30	68363	37109.5	31253.5
[('RF', 'XGB'), 'avg']	ensemble	0.448	0.606	199	112	23	64713.5	32866.5	31847
[('RF', 'GBM'), 'NN']	stacked	0.498	0.616	184	122	26	65913.5	31904.5	34009
[('LR', 'XGB'), 'NN']	stacked	0.467	0.613	187	96	19	66564	31698.5	34865.5
[('XGB', 'CNN'), 'NN']	stacked	0.503	0.614	182	116	16	70259	35312.5	34946.5
[('XGB', 'CNN'), 'LR']	stacked	0.482	0.617	188	103	17	67162	31737.5	35424.5
[('LR', 'GBM'), 'NN']	stacked	0.471	0.574	190	145	25	74228	36274.5	37953.5
[('RF', 'LSTM'), 'RF']	stacked	0.536	0.602	146	116	22	64024.5	25443.5	38581
[('XGB', 'LSTM'), 'avg']	ensemble	0.435	0.605	200	128	25	74076	34334.5	39741.5
[('LR', 'GBM'), 'LR']	stacked	0.451	0.601	188	124	26	71991.5	31743.5	40248

[('LR', 'RF'), 'max']	ensemble	0.541	0.548	187	192	27	82913	41006	41907
[('SVM', 'GBM'), 'SVM']	stacked	0.442	0.583	177	126	35	75695.5	33688.5	42007
[('LR', 'LSTM'), 'max']	ensemble	0.557	0.483	187	386	27	99591.5	56147	43444.5
[('SVM', 'GBM'), 'LR']	stacked	0.459	0.589	181	122	27	76945	32004.5	44940.5
[('GBM', 'LSTM'), 'avg']	ensemble	0.439	0.558	196	170	27	87103.5	41906	45197.5
[('RF', 'GBM'), 'avg']	ensemble	0.455	0.559	191	142	31	82896.5	35362.5	47534
[('SVM', 'GBM'), 'RF']	stacked	0.476	0.574	168	124	29	81162.5	33477.5	47685
[('LSTM', 'CNN'), 'max']	ensemble	0.572	0.472	180	405	25	103235.5	55240.5	47995
[('SVM', 'GBM'), 'GBM']	stacked	0.453	0.586	173	120	36	79682	31546	48136
[('XGB', 'LSTM'), 'NN']	stacked	0.484	0.610	189	105	13	80095.5	31742.5	48353
[('LR', 'GBM'), 'RF']	stacked	0.465	0.584	173	129	33	75343	26885	48458
[('GBM', 'LSTM'), 'RF']	stacked	0.467	0.599	171	111	31	83243	33315.5	49927.5
[('GBM', 'LSTM'), 'LR']	stacked	0.459	0.590	183	128	26	83793.5	32263.5	51530
[('RF', 'CNN'), 'GBM']	stacked	0.521	0.561	189	157	23	83824.5	30849	52975.5
[('GBM', 'LSTM'), 'SVM']	stacked	0.436	0.599	180	113	34	84826	31523.5	53302.5
[('SVM', 'GBM'), 'XGB']	stacked	0.479	0.580	170	120	28	83937.5	29549	54388.5
[('LR', 'GBM'), 'SVM']	stacked	0.433	0.593	183	114	38	82028	26937.5	55090.5
[('LR', 'GBM'), 'XGB']	stacked	0.460	0.587	169	120	31	84178	28828	55350
[('RF', 'CNN'), 'NN']	stacked	0.547	0.544	139	159	23	91556.5	35428	56128.5
[('LR', 'GBM'), 'GBM']	stacked	0.445	0.606	178	109	36	82111	25880.5	56230.5
[('RF', 'GBM'), 'GBM']	stacked	0.525	0.564	187	152	23	85301.5	28944	56357.5
[('RF', 'XGB'), 'GBM']	stacked	0.516	0.561	186	155	27	86435.5	29220	57215.5
[('RF', 'LSTM'), 'NN']	stacked	0.527	0.578	176	158	22	91585.5	34046.5	57539
[('RF', 'GBM'), 'LR']	stacked	0.514	0.561	180	151	23	88644	30425.5	58218.5
[('SVM', 'RF'), 'max']	ensemble	0.530	0.538	185	182	26	90083.5	31656.5	58427
[('SVM', 'LSTM'), 'max']	ensemble	0.555	0.481	187	377	23	113540	54908	58632
[('RF', 'XGB'), 'XGB']	stacked	0.513	0.557	183	150	23	90373	31139.5	59233.5
[('SVM', 'RF'), 'NN']	stacked	0.510	0.559	176	152	32	89611	28377.5	61233.5
[('SVM', 'RF'), 'GBM']	stacked	0.508	0.554	185	155	22	91111	29204	61907
[('LR', 'XGB'), 'avg']	ensemble	0.438	0.580	183	105	32	94220.5	32254.5	61966

[('RF', 'CNN'), 'XGB']	stacked	0.506	0.557	185	154	27	91111	28424.5	62686.5
[('LR', 'RF'), 'GBM']	stacked	0.511	0.561	184	149	25	91856	28570	63286
[('SVM', 'RF'), 'LR']	stacked	0.516	0.548	175	152	24	92105	28357.5	63747.5
[('RF', 'CNN'), 'SVM']	stacked	0.539	0.552	153	134	22	96145.5	31780.5	64365
[('LR', 'RF'), 'LR']	stacked	0.517	0.547	174	151	25	92492	27982.5	64509.5
[('RF', 'GBM'), 'XGB']	stacked	0.494	0.556	184	152	28	92957	28402.5	64554.5
[('GBM', 'LSTM'), 'XGB']	stacked	0.460	0.580	171	113	30	94630	29988.5	64641.5
[('RF', 'LSTM'), 'XGB']	stacked	0.504	0.556	183	153	27	93117.5	28414.5	64703
[('SVM', 'XGB'), 'avg']	ensemble	0.436	0.578	179	96	34	97497	31495.5	66001.5
[('SVM', 'RF'), 'XGB']	stacked	0.498	0.555	174	144	27	93702	27169.5	66532.5
[('GBM', 'LSTM'), 'GBM']	stacked	0.447	0.586	172	114	35	99312.5	31680.5	67632
[('RF', 'LSTM'), 'GBM']	stacked	0.514	0.538	187	157	28	96923	29265	67658
[('RF', 'LSTM'), 'LR']	stacked	0.516	0.551	174	146	25	95774	28005.5	67768.5
[('SVM', 'GBM'), 'NN']	stacked	0.418	0.600	158	83	47	97952.5	27985.5	69967
[('GBM', 'LSTM'), 'NN']	stacked	0.464	0.577	172	117	26	102998	31837.5	71160.5
[('LR', 'RF'), 'XGB']	stacked	0.492	0.552	173	140	31	98667.5	26917.5	71750
[('LR', 'RF'), 'NN']	stacked	0.513	0.550	163	143	30	105658	31954.5	73703.5
[('RF', 'LSTM'), 'avg']	ensemble	0.480	0.548	189	185	35	112889	39158	73731
[('RF', 'CNN'), 'LR']	stacked	0.530	0.528	151	142	24	116609	35022.5	81586.5
[('LR', 'RF'), 'SVM']	stacked	0.501	0.549	160	125	40	109219	24751.5	84467.5
[('XGB', 'CNN'), 'avg']	ensemble	0.450	0.562	161	98	36	114651	28579	86072
[('RF', 'CNN'), 'RF']	stacked	0.533	0.579	134	93	33	111531.5	22149.5	89382
[('LR', 'GBM'), 'avg']	ensemble	0.437	0.518	161	111	42	126123	31872.5	94250.5
[('RF', 'GBM'), 'SVM']	stacked	0.515	0.543	152	121	31	119023	23792.5	95230.5
[('RF', 'LSTM'), 'SVM']	stacked	0.512	0.544	157	121	32	121356.5	26000.5	95356
[('GBM', 'CNN'), 'RF']	stacked	0.512	0.511	127	125	40	130177	32485	97692
[('SVM', 'RF'), 'SVM']	stacked	0.504	0.529	149	111	37	126278	23502.5	102775.5
[('SVM', 'GBM'), 'avg']	ensemble	0.433	0.510	152	84	57	128946	19842.5	109103.5
[('LR', 'RF'), 'RF']	stacked	0.521	0.516	145	111	58	132966.5	23520.5	109446
[('GBM', 'CNN'), 'GBM']	stacked	0.501	0.505	124	125	43	145387	35476	109911

[('GBM', 'CNN'), 'SVM']	stacked	0.498	0.511	126	127	47	145851.5	34305	111546.5
[('GBM', 'CNN'), 'LR']	stacked	0.500	0.521	127	115	43	147447	34109	113338
[('GBM', 'CNN'), 'XGB']	stacked	0.519	0.505	119	124	46	149431.5	30112.5	119319
[('GBM', 'CNN'), 'avg']	ensemble	0.456	0.509	131	105	53	152885	30315.5	122569.5
[('LR', 'CNN'), 'max']	ensemble	0.528	0.488	123	178	42	168240	38680	129560
[('GBM', 'CNN'), 'NN']	stacked	0.475	0.522	118	86	52	158108.5	28027	130081.5
[('LR', 'RF'), 'avg']	ensemble	0.468	0.496	136	103	51	169801.5	30943.5	138858
[('RF', 'CNN'), 'avg']	ensemble	0.486	0.489	114	105	54	173345	26736.5	146608.5
[('LR', 'LSTM'), 'avg']	ensemble	0.494	0.458	135	188	57	190442.5	41236	149206.5
[('LSTM', 'CNN'), 'RF']	stacked	0.541	0.432	97	167	55	187281.5	35237.5	152044
[('LSTM', 'CNN'), 'XGB']	stacked	0.540	0.433	98	153	56	188531	36309.5	152221.5
[('LSTM', 'CNN'), 'GBM']	stacked	0.527	0.443	100	154	56	186887.5	33900	152987.5
[('LSTM', 'CNN'), 'SVM']	stacked	0.521	0.453	102	155	53	184757.5	31328.5	153429
[('SVM', 'CNN'), 'max']	ensemble	0.528	0.462	114	168	48	196184	37535	158649
[('LR', 'CNN'), 'SVM']	stacked	0.530	0.440	94	168	62	195736.5	36974.5	158762
[('SVM', 'CNN'), 'SVM']	stacked	0.528	0.431	90	160	63	196137.5	36527.5	159610
[('LSTM', 'CNN'), 'avg']	ensemble	0.490	0.450	110	169	57	194093.5	33646.5	160447
[('SVM', 'CNN'), 'XGB']	stacked	0.550	0.424	92	152	65	196844.5	32864	163980.5
[('LR', 'CNN'), 'RF']	stacked	0.551	0.423	92	163	59	195068.5	30986.5	164082
[('LR', 'LSTM'), 'GBM']	stacked	0.495	0.481	97	138	60	196312.5	31576.5	164736
[('SVM', 'LSTM'), 'SVM']	stacked	0.479	0.497	104	124	67	186802	22037.5	164764.5
[('SVM', 'LSTM'), 'XGB']	stacked	0.513	0.464	93	131	69	188912	24104.5	164807.5
[('SVM', 'LSTM'), 'GBM']	stacked	0.496	0.483	96	130	68	190615	25323.5	165291.5
[('LR', 'LSTM'), 'LR']	stacked	0.516	0.467	108	156	46	202108.5	35024	167084.5
[('LR', 'CNN'), 'XGB']	stacked	0.558	0.423	93	156	61	201735	33897	167838
[('SVM', 'CNN'), 'RF']	stacked	0.548	0.422	89	153	63	198742	30819.5	167922.5
[('LSTM', 'CNN'), 'LR']	stacked	0.525	0.424	90	152	60	207010	37135	169875
[('LR', 'CNN'), 'GBM']	stacked	0.534	0.419	90	157	61	203957.5	33705	170252.5
[('SVM', 'CNN'), 'LR']	stacked	0.531	0.428	91	151	61	205494	34396	171098
[('LSTM', 'CNN'), 'NN']	stacked	0.519	0.462	78	105	59	197838.5	26532.5	171306

[('SVM', 'CNN'), 'NN']	stacked	0.531	0.424	93	166	53	206237.5	34824.5	171413
[('LR', 'LSTM'), 'XGB']	stacked	0.514	0.465	94	140	62	197020.5	24758	172262.5
[('SVM', 'CNN'), 'GBM']	stacked	0.529	0.418	87	151	61	209450	35926	173524
[('SVM', 'LSTM'), 'RF']	stacked	0.529	0.469	96	131	68	196729	22430.5	174298.5
[('SVM', 'RF'), 'avg']	ensemble	0.468	0.469	130	88	62	195760.5	19597.5	176163
[('LR', 'LSTM'), 'RF']	stacked	0.531	0.471	93	136	48	204932.5	28027.5	176905
[('LR', 'CNN'), 'LR']	stacked	0.520	0.415	88	147	65	213112.5	33846.5	179266
[('LR', 'CNN'), 'NN']	stacked	0.539	0.424	81	127	57	211315.5	30863.5	180452
[('SVM', 'LR'), 'RF']	stacked	0.548	0.423	86	123	70	211786.5	27721.5	184065
[('SVM', 'LR'), 'SVM']	stacked	0.469	0.444	79	106	80	214208	28489	185719
[('SVM', 'LSTM'), 'LR']	stacked	0.497	0.467	101	132	56	212061.5	25600	186461.5
[('SVM', 'LSTM'), 'NN']	stacked	0.515	0.521	64	66	48	204080.5	12327	191753.5
[('LR', 'LSTM'), 'SVM']	stacked	0.495	0.451	90	124	62	221000.5	26057	194943.5
[('LR', 'CNN'), 'avg']	ensemble	0.483	0.427	90	114	66	223449.5	28484.5	194965
[('SVM', 'LR'), 'max']	ensemble	0.508	0.437	98	110	51	230180	32593	197587
[('SVM', 'LSTM'), 'avg']	ensemble	0.478	0.460	122	134	64	228009	27908	200101
[('SVM', 'LR'), 'GBM']	stacked	0.491	0.417	82	122	78	233268	30784	202484
[('SVM', 'LR'), 'XGB']	stacked	0.519	0.427	82	129	65	231650	27680.5	203969.5
[('SVM', 'CNN'), 'avg']	ensemble	0.470	0.419	82	96	76	241155	22554.5	218600.5
[('SVM', 'LR'), 'LR']	stacked	0.496	0.424	75	83	59	239546	20934	218612
[('LR', 'LSTM'), 'NN']	stacked	0.516	0.433	100	142	50	245576.5	26409	219167.5
[('SVM', 'LR'), 'avg']	ensemble	0.470	0.399	81	78	68	261274.5	18665	242609.5
[('SVM', 'LR'), 'NN']	stacked	0.481	0.376	85	87	58	273122.5	19449	253673.5

Appendix 6. Results for All Stalking and Harassment

The approach for the model that encompasses all Stalking and Harassment followed the same approach as outlined in Section 3.3.2. The dataset for all Stalking and Harassment was considerably larger, totalling more than 100,000 nominals and more than 1 million crimes. The dataset preparation approach was the same, by looking forward from Stalking and Harassment day zero, calculating the harm committed in the 1 year after the offence, then grouping into the same 4 classes of 0-50%, 50-80%, 80-90%, and 90-100%. The number of samples used for training and testing the models can be seen in Table 9.

Table 9 - Data Split for all Stalking and Harassment

Group	Total	Training	Testing
0-50%	62,765	1,800	1,635
50-80%	10,351	1,800	433
80-90%	3,796	1,800	452
90-100%	3,450	1,800	400

The same approach was used to select the most optimal model, considering the harm of the false negatives and the false positives. All 168 models were trained and the most optimal model in this case is a stacked model consisting of an XGBoost and an LSTM into a stacked third model of a Random Forest.

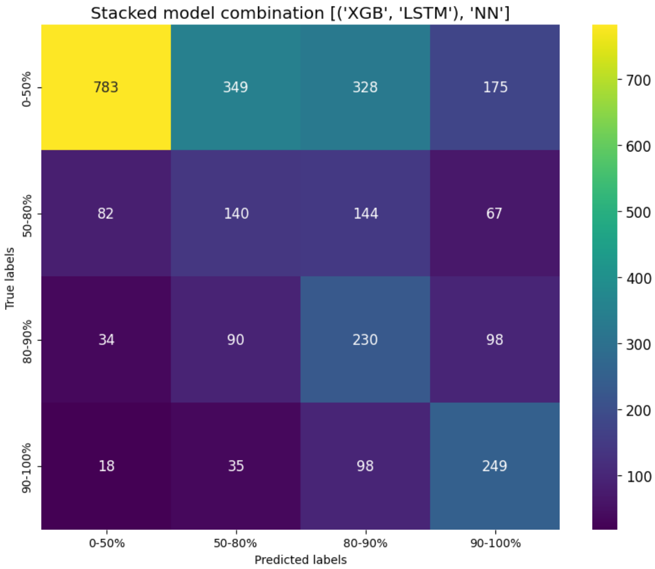


Figure 19 - All Stalking and Harassment confusion matrix

This model gave the results of a top 2 accuracy of 87% and adjusted F1 score of 0.72 . The model missed 67,353 harm, and prevented an additional 40,990 harm, resulting in a net harm cost of 26,364.

Appendix 7. Results for Modelling from Victim Perspective

An identical modelling approach was taken to create a sister model to the nominal stalking model. While the nominal stalking model focussed on the nominal, this model was centred around the victim of the crime and their previous victim experiences. This would allow the prioritisation of nominals to be based not only on the nominal's crimes, but also give an estimation of how the harm experienced by a victim can escalate.

The use of this model would not change the way in which nominals are to be prioritised. If a nominal is categorised as high harm, and their victim is categorised as low estimated further harm, the nominal would still receive attention of investigative officers. However, if the nominal and their victim are both categorised as high harm, coupled with the time to event modelling, it would allow investigative officers to focus their efforts on individuals who are requiring police attention in the timeliest manner.

Table 10 - Data split for victim only stalking model

Group	Total	Training	Testing
0-50%	23,296	2,402	1,022
50-80%	3,365	2,402	598
80-90%	1,122	902	220
90-100%	1,234	1,002	232

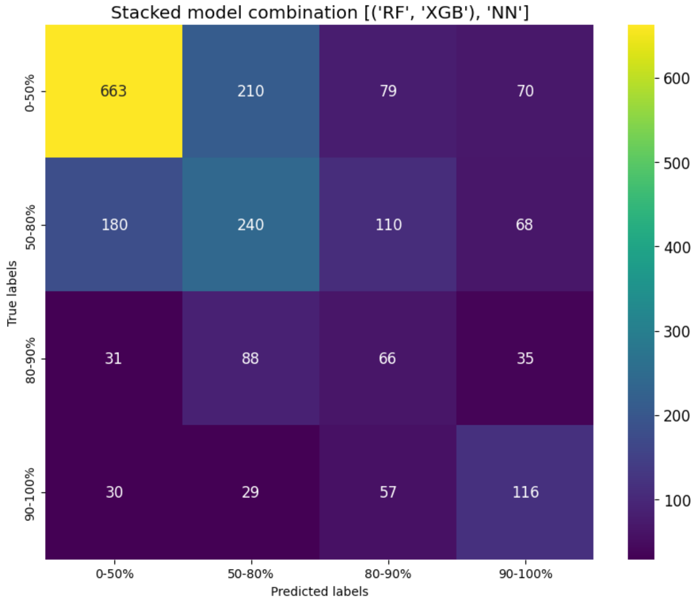


Figure 20 - Stalking Victim Harm Escalation Model

This model gave the results of a top 2 accuracy of 74.57% and adjusted F1 score of 0.7. The model missed 45,448 harm, and "prevented" an additional 27,717 harm, resulting in a net harm cost of 17,731.

Appendix 8. Non-Crime Domestic Harm Score

The dataset for the non-crime domestic harm score was built around text extracted from relevant crimes. The crimes excluded from this dataset were miscellaneous crimes against society, theft offences, non-crime, non-notifiable, fraud offences, and drug offences. Following this, crime records were only considered when the CCHI harm score was less than or equal to 10, as this was the cut-off point as explained in Section 3.1. This resulted in a total of 1,012,658 records.

The CCHI score is not a linear scale, between 0 and 10, a crime can only have a harm score of 1, 2, 3, 5, and 10; therefore, this lends this problem well to a (ordinal) classification model approach.

Following this, the dataset was reduced to 25,000 samples per class, totalling 125,000 samples to use for training and testing; the reduction was necessary due to the size of the textual data contained in more than 1 million records. Data preparation and text cleaning was conducted on the text logs contained in these crimes in order to make them representative and understandable for modelling. This included stemming and lemmatising the words to account for plurals, removing stop words, and calculating bi-grams and tri-grams of words that commonly occur together.

Finally, a document term matrix was calculated for all samples in the dataset. The words were only selected for the final document term matrix when the word occurred more than 100 times in the total dataset. This resulted in a final dataset of 125,000 samples with 2,106 unique words.

The training and testing split were 80:20 and balanced split across all classes.

The Neural Network selected consisted of 6 layers, with a 20% drop out function between each layer. The structure started with the input shape 2,106 leading to fully connected layers in the following order of 1028, 512, 256, 128, 32, and 5 being the output of 5 classes. Each layer had a ReLU activation, while the final layer had a softmax activation function.

The accuracy of the trained model on the test set was 72.6%. The confusion matrix of the results can be seen in Figure 21.

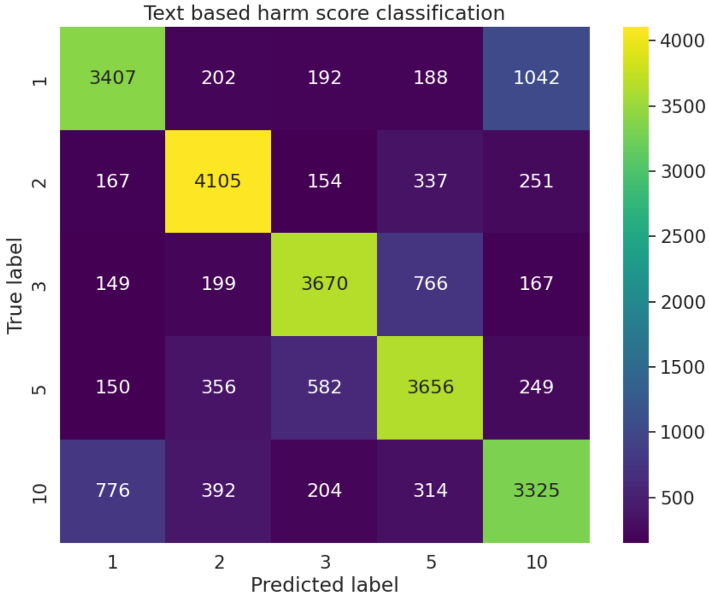


Figure 21 - Text based harm score confusion matrix

Appendix 9 – Stalker Harm Model without Suspect Data

This appendix will look at the results of the stalking only model when only offender data is used. This study is conducted to show the importance of using suspect data in a modelling process for crimes such as those in this report. The entire process of data preparation was identical to that of the model used in the main body of the report. The only difference to the data used in this section is that all crimes where the nominal is a suspect and is not ultimately charged are removed. This leaves a dataset where all crimes are those where the nominal was charged.

In the context of this model's use, an issue becomes immediately apparent, whereby using only offender data means any estimation in the escalation of harm an offender might go on to do would have to come after the charge has been brought, which given the timescales involved in some investigations, would mean that any estimation would be meaningless and offer no chance of early intervention for the victim or offender. However, for the avoidance of doubt, the following results are for a model trained on and predicted on nominals who are offenders.

Table 11 - Data Split for no suspect model

Group	Total	Training	Testing	Proportion in Training
0-50%	3,788	981	754	56.5%
50-80%	419	327	92	78.0%
80-90%	154	114	26	81.4%
90-100%	140	125	29	81.2%

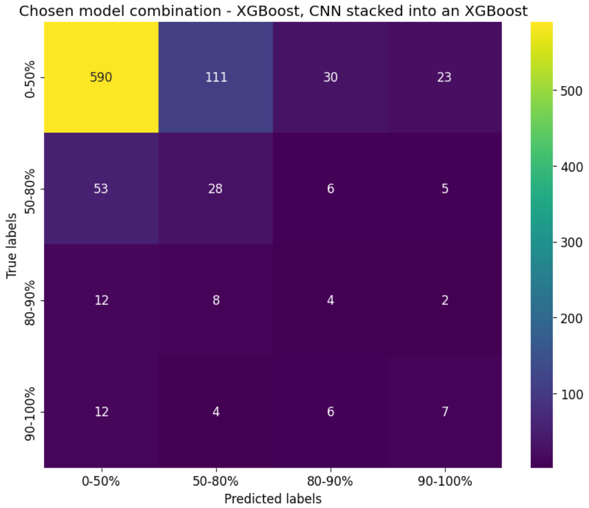


Figure 22 - Chosen best model results for no suspect model

This model gave the results of a top 2 accuracy of 45% and adjusted F1 score of 0.283. The model missed 16,294 harm, and "prevented" an additional 3,593 harm, resulting in a net harm cost of 12,701. However, there were 4,501 offenders in this cohort of nominals vs 28,279 in the cohort including suspects (the rate of harm cost per nominal in this model is 12691/4501 = 2.82, vs with suspect data, 15427/28279 = 0.55, meaning the model without suspect data results in a total harm cost 5 times higher than the model with suspect data over a normalised cohort). From these results, it is clear that the inclusion of suspect data is paramount to ensuring a usable model from both a data science perspective, but also from the perspective of providing a good level of service to the public.

Appendix 10 – Stalker Harm Model without Domestic Non-Crime

This appendix will look at the results of the stalking only model when domestic non-crime incidents are removed. This study is conducted to show the importance of using domestic non-crime incidents in the modelling to give additional context to some of the victim-offender relationships. The entire process of data preparation was identical to that of the model used in the main body of the report. The only difference to the data used in this section is that all incidents of domestic non-crime are removed. As previously discussed, domestic non-crime incidents are recorded by the force to be able to keep track of and to analyse the volume of reports for DA victims, identify repeat victims, help risk assess and support them. These incidents are categorised as Non-Crime, meaning that they are recorded like a crime, but do not carry any financial or criminal justice outcomes. To be clear, where an incident is recorded as domestic abuse, this is still included as this is a recordable offence type.

Table 12 - Data split for no domestic non crime model

Group	Total	Training	Testing	Proportion in Training
0-50%	21,595	1,768	1,609	52.4%
50-80%	3,932	1,768	432	80.4%
80-90%	1,310	849	191	81.6%
90-100%	1,442	899	245	78.6%

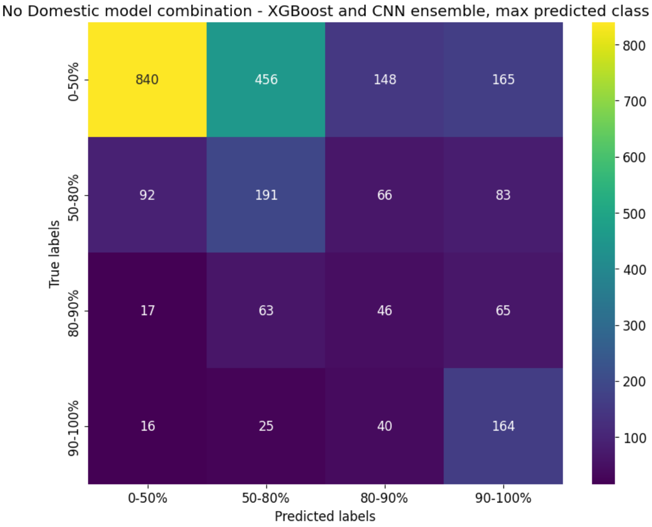


Figure 23 - Chosen best model results for no domestic non-crime model

This model gave the results of a top 2 accuracy of 83% and adjusted F1 score of 0.577. The model missed 53,296 harm, and "prevented" an additional 25,645 harm, resulting in a net harm cost of 27,651. From these results, it is clear that the inclusion of domestic non-crime data improves both the model accuracy and the model outcomes in terms of preventable harm. While the domestic non-crime incidents add little in terms of harm score to the various features (no greater than 10), the context and frequency of recorded domestic non-crime incidents obviously goes someway to offer context to a victim-nominal relationship.

References

Baum, K., Catalano, S., Rand, M., & Rose, K. (2009). National Crime Victimization Survey: Stalking victimization in the United States. *U.S. Department of Justice.*

CPS (2023) *Stalking and harassment, Stalking and Harassment*. The Crown Prosecution Service. Available at: https://www.cps.gov.uk/legal-guidance/stalking-and-harassment (Accessed: April 6, 2023).

Häkkänen, H., Hagelstam, C., & Santtila, P. (2003). Stalking actions, prior offender-victim relationships and issuing of restraining orders in a Finnish sample of stalkers*. Legal And Criminological Psychology*, *8*(2), 189-206.

Home Office (2023) *Police recorded crime and Outcomes Open Data Tables, Police recorded crime and outcomes open data tables*. Home Office. Available at: https://www.gov.uk/government/statistics/police-recorded-crime-open-data-tables (Accessed: April 6, 2023).

Olah, C. (2015) *Understanding LSTM networks*, *Understanding LSTM Networks -- colah's blog*. Available at: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (Accessed: April 11, 2023).

ONS (2022) *Domestic abuse in England and Wales Overview: November 2022, Domestic abuse in England and Wales overview.* Office for National Statistics. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/domesticabuseinenglandandwalesoverview/latest (Accessed: April 6, 2023).

Logan, T., & Walker, R. (2019). The Impact of Stalking-Related Fear and Gender on Personal Safety Outcomes. *Journal Of Interpersonal Violence, 36*(13-14), NP7465-NP7487.

McEwan, T., Shea, D., Daffern, M., MacKenzie, R., Ogloff, J., & Mullen, P. (2018). The Reliability and Predictive Validity of the Stalking Risk Profile. *Assessment, 25*(2), 259- 276.

Meloy, J.R. and Gothard, S., 1995. Demographic and clinical comparison of obsessional followers and offenders with mental disorders. *American journal of psychiatry*, *152*(2), pp.258-263.

Phung VH, Rhee EJ. A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets. *Applied Sciences*. 2019; 9(21):4500

POLICE.UK (2023) *What is stalking and harassment?,* POLICE.UK. Available at: https://www.police.uk/advice/advice-and-information/sh/stalking-harassment/what-is-stalking-harassment/ (Accessed: April 6, 2023).

Rosenfeld, B. (2003). Recidivism in stalking and obsessional harassment. *Law and Human Behavior, 27(3)*, 251-265.

Tjaden, P., & Thoennes, N. (2000). Prevalence, incidence, and consequences of violence against women: Findings from the national violence against women survey. *Violence Against Women.*