Stalking and Harassment

and

Risk of Post Harm Escalation

Data Analytics Lab

December 2024

Contents

Table of Figures

Table of Tables

1 Executive Summary

This project presents four predictive models designed to support the Stalking Triage Clinic in triaging offenders and victims involved in stalking and harassment cases. These models aim to identify offenders likely to escalate to high-harm offenses and victims at risk of experiencing high-harm incidents within the next 12 months. Two of the models provide probabilities for escalation to high-harm following a stalking or harassment offense, while the other two predict the number of days until an offender or victim is likely to be involved in a severe crime, such as murder, rape, or assault with grievous bodily harm.

The analysis relied on data from WMP systems, encompassing various features tied to offenders' criminal histories and victims' past experiences. Key features include the Cambridge Crime Harm Index scores, age at offence, gender, and several engineered features derived from keywords and free-text incident logs. To extract relevant information from free text – particularly indicators linked with risk escalation in stalking and harassment cases – a large language model was used alongside prompt engineering to focus on key aspects of interest.

The approach was shaped by a thorough exploratory analysis and a rigorous model selection process. The initial exploratory analysis offered insight into feasible modelling approaches and expected performance. Model selection involved setting up baselines, assessing the impact of different features, exploring hyperparameters, conducting error analysis, and finalising the models to deploy. This iterative process ensured optimal predictive accuracy and relevance.

The project also addresses challenges in modelling stalking and harassment due to the complexity of these phenomena and the limitations of the associated feature space. The practical applications for the stalking models in offender and victim triage are discussed, with illustrative examples. Finally, alternative modelling approaches tested during exploration but found ineffective are also reviewed.

Following the exploratory data analysis and model training and testing phases, a final XGBoost offender model was trained using hyperparameters optimised determined through Bayesian optimisation. A similar model was built based on victim history and associated time to event models have also been developed

2 Introduction

In the UK, stalking is defined under the Protection from Harassment Act 1997 [1], which was amended in 2012 to specifically address stalking. According to the law, stalking involves a pattern of repeated and unwanted behaviour that is intended to cause alarm, distress, or fear of violence. The behaviour can include, but is not limited to, the following actions:

- Following a person
- Contacting, or attempting to contact, a person by any means
- Publishing any statement or other material relating or purporting to relate to a person, or purporting to originate from a person
- Monitoring the use by a person of the internet, email, or any other form of electronic communication
- Loitering in any place (whether public or private)
- Interfering with any property in the possession of a person
- Watching or spying on a person

The law focuses on the impact of the behaviour on the victim rather than the specific acts themselves. For an act to be considered stalking, there needs to be a pattern of behaviour that causes the victim serious alarm or distress, making them fear that violence could be used against them.

Harassment is also defined under the Protection from Harassment Act 1997 in the UK. Harassment refers to a course of conduct that causes alarm or distress to another person. It involves repeated and unwanted behaviour that is perceived as oppressive, threatening, or abusive.

Key elements of harassment include:

- A “course of conduct” which means the behaviour must occur on at least two occasions.
- The conduct must cause the victim alarm or distress.
- The behaviour must be such that a reasonable person in possession of the same information would think it amounts to harassment.

Examples of harassment can include:

- Unwanted communication, such as emails, phone calls, or text messages.
- Repeatedly showing up at someone’s home or place of work.
- Threatening behaviour or gestures.
- Damaging someone’s property.
- Following someone.

It is important to note that harassment can include both physical acts and non-physical acts, such as verbal abuse or persistent unwelcome contact.

2.1 Stalking and harassment in West Midlands

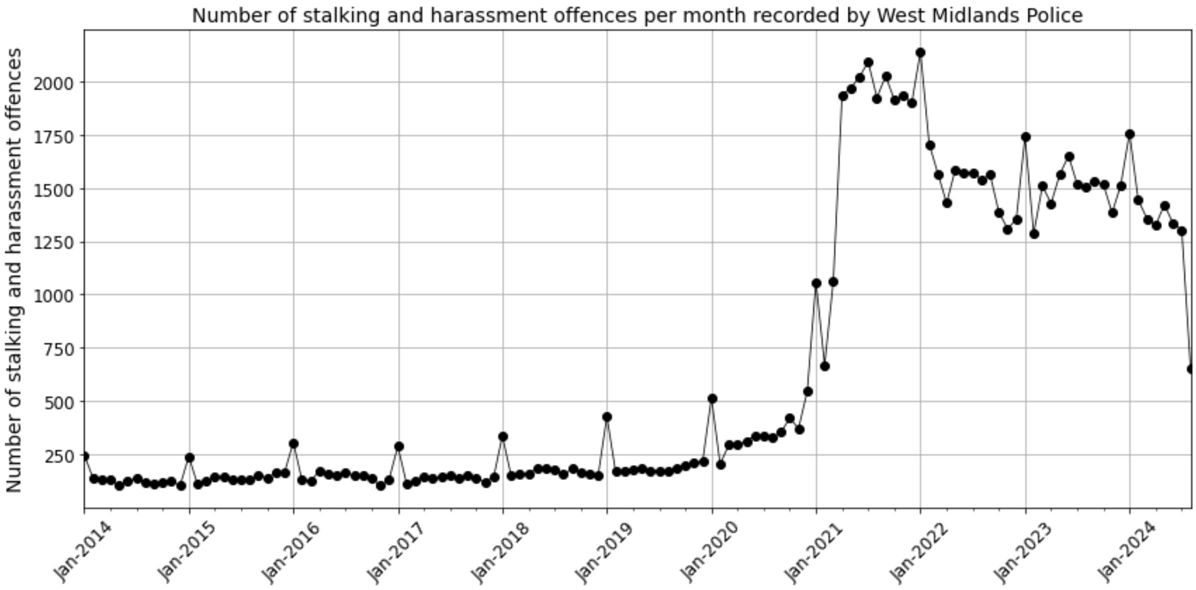


Figure 1. Number of stalking and harassment offences per month recorded by West Midlands Police between January 2014 and January 2024.

The data in **Figure 1** highlights a significant increase in reported offences beginning around early 2020, with a notable peak reaching over 2,000 offences per month in 2021.

There was a sharp rise in the number of recorded stalking and harassment offences starting in early 2020. **These patterns can be attributed to changes in reporting systems, counting rules, and crime recording inconsistencies rather than behavioural reasons**.

Reasons for the sudden spike from 2020-2021:

- **Implementation of ControlWorks:** The introduction of ControlWorks, a new system for recording incidents and managing command and control operations, greatly enhanced the ability of West Midlands Police to link information across various records. This system allowed for more detailed categorisation of crimes, better linking of offences to offenders, and improved integration between intelligence systems involving people, places, events, and objects. As a result, the quality and accuracy of crime recording significantly improved, leading to a noticeable increase in the recorded offences of stalking and harassment.
- **Changes in Home Office counting rules:** In April 2020, the Home Office introduced new counting rules for recording offences between victims and their partners. These changes mandated that any historical data involving stalking or harassment occurring on two or more occasions be categorised specifically under stalking and harassment. Previously, such incidents might have been recorded under broader categories like malicious communications or general harassment, leading to an apparent rise in offences once the new rules took effect.

Explanation for recurring spikes at the start of each year:

- The recurring spikes seen in January of each year can be explained by a recording practice used by West Midlands Police. When the exact date of a crime is unknown, but it is reported to have occurred within a certain month, the incident is recorded as having occurred on the 1st of that month. This practice often results in artificial spikes at the beginning of each year, as historical or undated offences are logged on January 1st.

2.2 Demographics of stalking and harassment offenders and victims

Stalking and harassment primarily involve male perpetrators and female victims, with the average victim being slightly over 32 years old, with small variations reported among studies [2] [3]. The majority of stalkers are male, though female stalkers also exist but are less likely to escalate from threats to physical violence. A significant proportion of stalking cases involve former intimate partners, where the emotional and relational history intensifies the risk of violence [2] [3]. Ex-intimates are at the highest risk for physical harm, including grievous bodily harm and murder, due to the strong emotional attachment and unresolved feelings from the terminated relationship [3]. The demographics of stalkers frequently include younger individuals, particularly those under 30, with lower education levels, minority racial backgrounds, and histories of criminal behaviour or substance abuse [2] [3].

2.3 Stalking and harassment and escalation to serious harm

Stalking and harassment encompass a range of intrusive and threatening behaviours, including unwanted communications, following, property damage, and direct threats of violence [4] [5]. However, the escalation to serious harm, such as grievous bodily harm or murder, is most strongly associated with specific types of behaviours. The use of weapons in stalking, even primarily for intimidation, is a significant predictor of serious violence. Weapons commonly used include knives, firearms, and blunt objects, which are often employed in a manner designed to instil fear rather than cause immediate physical injury [6] [7]. However, their presence significantly increases the potential for fatal outcomes [8].

The most severe forms of violence, such as attempted or actual homicide, are frequently preceded by threats to kill, direct assaults, and specific behaviours like non-fatal strangulation, which is strongly linked to an increased risk of lethal violence [9] [10] [11]. Non-fatal strangulation, in particular, is a critical indicator, as victims who have experienced it are significantly more likely to be killed or face attempted murder [12] [13]. Other serious behaviours include attempts to gain entry to the victim’s home, early and repeated confrontations, and the use of coercive control tactics, such as threats involving children or financial manipulation [14].

Several key predictors are associated with the escalation from stalking and harassment to serious crimes like grievous bodily harm and murder. A combination of ex-intimacy, explicit threats, and property damage were reported to be effective in predicting stalking violence [15]. Explicit threats, particularly those involving direct harm or weapon use, are consistent predictors of future violence [16]. Verbal threats alone are significantly correlated with physical violence, and their presence should be treated as a serious risk factor [16].

Rapid escalation of behaviours, such as early appearances at the victim's home, is also a critical predictor of severe violence [15] [16]. The timing of these actions indicates a higher propensity for serious harm, especially when they occur early in the stalking period. The duration of stalking is another factor, where shorter stalking periods often indicate a greater risk of serious violence, suggesting that rapid escalation is a critical warning sign [15].

Substance abuse, including drug and alcohol use, is another significant predictor of injury during stalking incidents [15] [14]. The impairment caused by substance abuse can exacerbate aggressive behaviours and increase the likelihood of spontaneous and severe assaults. Prior criminal history, particularly with violent offences, also heightens the risk of serious harm, as it reflects a pattern of aggressive and antisocial behaviour [15] [14].

Moreover, coercive control is a pervasive predictor in cases that escalate to intimate partner homicide [15] [17] [18]. This involves behaviours designed to dominate and manipulate the victim, such as restricting their movements, controlling finances, and imposing psychological pressure. Coercive control, even in the absence of overt physical violence, can escalate to lethal outcomes when the stalker perceives a loss of control, such as during relationship termination or other significant personal crises [15] [19].

Restraining orders are commonly issued to protect victims; however, their effectiveness varies significantly based on the stalker's perception of consequences. For some stalkers, particularly those with histories of criminal behaviour or domestic abuse, the issuance of a restraining order can provoke anger or a sense of defiance, leading to an escalation in violent behaviours [20]. Research indicates that breaches of restraining orders are not uncommon, with many stalkers disregarding legal boundaries, which can lead to increased aggression and potentially severe outcomes [21]. In some cases, the issuance of a restraining order may re-ignite a volatile situation, particularly if the stalker perceives it as a challenge or as further rejection, thereby intensifying the risk of serious harm, including grievous bodily harm or even homicide [23] [1].

Additionally, child-related factors are significant predictors of the escalation to severe violence in stalking and harassment cases. Threats involving children, such as child abduction or threats to take children away, are used by stalkers as powerful tools of psychological manipulation and control [16] [24] [17]. These threats are especially distressing to victims and can be part of a broader strategy of coercive control, increasing the victim's fear and sense of helplessness. The involvement of children complicates the dynamics of stalking, as stalkers may use access to or custody of children to maintain contact with the victim or exert control.

[1] Understanding and Countering Stalkers - TorchStone Global

Table 1. Key predictors for serious harm identified in the relevant literature.

Ex-intimacy (separation from partner)	Following and other surveillance behaviours
Explicit threats (direct harm or weapon use)	Attempts to gain entry to victim's home
Verbal threats	Breach of court orders
Rapid escalation of behaviours (early appearances at victim's home)	Prior physical violence such as strangulation, choking, hair pulling, punching
Unwanted communications	Age under 30 (considering that most reported ages in the literature hover around this range)
Substance abuse (drug and alcohol use)	Prior criminal history (especially violent offences)
Coercive control (domination and manipulation tactics)	Psychological abuse such as intimidation, spreading false rumors, blackmailing, impersonation, etc.
Child-related factors (threats involving children, child abduction)	Unwanted communications

2.4 Defining the scope of analysis

The primary aim of this project is to assess the likelihood that individuals, referred to as "nominals" with a history of stalking and harassment, will escalate their behaviour to commit high harm crimes, such as grievous bodily harm, homicide, or sexual assault. The project is divided into three main objectives. First, it aims to develop a model that identifies high-risk offenders likely to escalate their criminal behaviour. Second, a similar model will be trained to classify whether victims of stalking and harassment are likely to be re-victimised by serious offenses within the next 12 months. Finally, the third objective is to create two time-to-event models, one for offenders and one for victims, to estimate when these nominals are likely to commit or fall victim to high harm crimes. These models will provide valuable insights to a Stalking Triage Clinic team, enabling them to prioritise potential interventions early with both high-risk offenders and potential victims.

3 The Data

The data utilised in this analysis was compiled from three distinct sources and merged into a final dataset:

1. **Crime dataset:** Contains detailed information about each nominal's criminal history.
2. **Custody dataset:** Provides records of the number of times a nominal has been in custody.
3. **Victim call dataset:** Tracks the number of times each nominal has called the police as a victim.

3.1 Data cleaning and selection

During the data preprocessing stage, the following steps were implemented:

- **Removing duplicates:** Duplicate rows were identified and removed.
- **Column renaming:** Columns were renamed to enhance clarity and interpretability.
- **Simplifying crime role flags:** The roles in crime flags were reduced to the categories from "offender", "suspect"[2], and "victim" to just "offender" and "victim."
- **Date filtering:** The dataset was filtered to include records from 2014 onwards.
- **Text cleaning and formatting:** Incident summaries recorded by the police were cleaned and formatted to improve feature extraction from free text.
- **Age filtering:** Only nominals between 10 and 100 years old were kept in the dataset.

The dataset was further refined by applying specific filters to retain only those nominals who met the following conditions:

- **Stalking and harassment offence as offender:** The nominal must have at least one offence related to stalking and harassment where their role in the crime was classified as "offender". While nominals can have a mixed crime history where their role is either as a victim, an offender, or both, this analysis focuses on those offences committed as an offender.
- **Prior offending behaviour:** The nominal must have committed at least one offence as an offender prior to the first stalking and harassment offence. This criterion ensures a historical context of offending behaviour that precedes the stalking and harassment incident.
- **Subsequent offending behaviour:** The nominal must have committed at least one offence as an offender after the stalking and harassment offence. This condition ensures that there is data available to evaluate the escalation and quantify the harm caused by the nominal in their subsequent offending behaviour.

[2] It will be noted from the previous version of this analysis (and its associated report) that not including crimes with suspect status (not including eliminated) severely reduced the performance of modelling.

These conditions are critical for the analysis as they ensure the following:

- **Historical context of offending:** By including nominals with prior offending history, the data provides the necessary context around the stalking and harassment offence. This approach allows for the identification of behavioural patterns such as escalation and a deeper understanding of the circumstances in which the stalking and harassment offence occurred. This contextual information is important for both model training and empirical inference.
- **Quantification of harm post-offence:** The inclusion of at least one offence after the stalking and harassment incident enables the assessment of the nominal's potential for further harm as an offender. Without subsequent offences, the data lacks the necessary outcomes to inform the modelling process, making such records unsuitable for predictive analysis.
- **Feature engineering for harm metrics:** The requirement for at least one offence prior to the stalking and harassment incident also facilitates the calculation of new features, such as harm rate and harm momentum, derived from the Cambridge Crime Harm Index (CCHI) scores. These features necessitate a minimum of two historical offences to capture trends in the severity and progression of offending behaviour.

3.2 Offence sequence splitting for exploratory data analysis and predictive modelling

In the data preparation process, the offence history for each nominal was split into two distinct sequences, as exemplified in **Figure 2**: **(1)** offences that occurred before and up to the stalking and harassment incidents, and **(2)** offences that occurred after these incidents. For each nominal, sequences of events were constructed based on the number of stalking and harassment offences where the individual acted as an offender, and only if there was at least one crime as an offender both before and after the stalking and harassment offence.

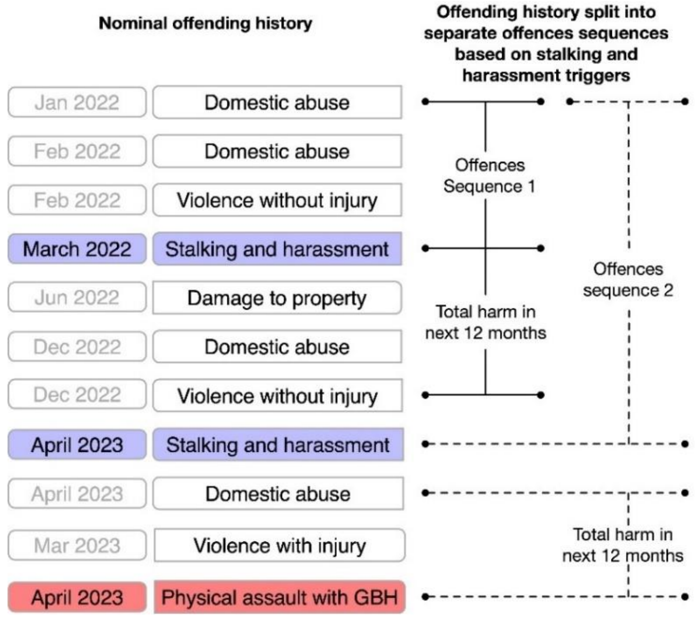

Nominal offending history
Offending history split into separate offences sequences based on stalking and harassment triggers
Jan 2022 Domestic abuse
Feb 2022 Domestic abuse
Feb 2022 Violence without injury
March 2022 Stalking and harassment
Jun 2022 Damage to property
Dec 2022 Domestic abuse
Dec 2022 Violence without injury
April 2023 Stalking and harassment
April 2023 Domestic abuse
Mar 2023 Violence with injury
April 2023 Physical assault with GBH
Offences Sequence 1
Total harm in next 12 months
Offences sequence 2
Total harm in next 12 months

Figure 2. Example of splitting offending history into sequences based on the occurrence of stalking and harassment offence.

Rationale for creating offending history sequences:

- **Trigger events for harm escalation evaluation:** Stalking and harassment offences serve as key trigger events for assessing the likelihood of harm escalation for each nominal. Each occurrence of a stalking and harassment offence acts as a point at which the predictive model can be applied to evaluate the risk of future serious harm. The model uses the most recent stalking and harassment offence along with all prior offences to make this prediction. This approach allows the model to continuously update its assessment as new data becomes available, reflecting the evolving criminal behaviour of the nominal.
- **Enhanced model training through contextual offending history:** A model trained only on the first occurrence of stalking and harassment offences and the preceding crimes assumes a static pattern of offending behaviour. This limitation would hinder the model's ability to predict further escalation when subsequent stalking and harassment offences occur. For example, if a nominal commits multiple offences, such as domestic violence, property damage, and assault with injury, after the first stalking and harassment incident and then reoffends with other crimes and another stalking and harassment offence, a model trained solely on the initial event would not account for these additional patterns of escalation. By including subsequent stalking and harassment offences and all prior incidents, the model learns from a broader and more dynamic range of offending behaviour, enhancing its predictive capability across different stages of a nominal's criminal history.
- **Data augmentation through sequential offence analysis:** This approach effectively acts as a form of data augmentation. By treating each new stalking and harassment offence as a separate trigger event, the model is repeatedly trained and tested using all available historical data leading up to each offence. This repeated analysis of sequential offending allows the model to capture patterns that might be missed if only the first offence was considered. It ensures that the model can provide updated and contextually relevant predictions about the risk of harm escalation over the next 12 months, considering the most recent and comprehensive history of the nominal's offending behaviour.
- **Improved exploratory data analysis (EDA):** Splitting the offence history into pre- and post-stalking and harassment sequences also enhances exploratory data analysis. By examining the sequences separately, it becomes easier to identify trends and patterns in offending behaviour relative to the stalking and harassment offences. This approach allows for a more granular analysis of how offending behaviours evolve before and after key trigger events, providing valuable insights into escalation patterns and the impact of specific offences on subsequent behaviour. It also facilitates the identification of common pathways that lead to more serious harm, aiding in the development of targeted interventions and strategies for risk mitigation.

Following each stalking and harassment offence, all subsequent offences within a defined time frame (**next 12 months**) are utilised for labelling purposes in the analysis. This approach involves examining the total harm caused by the subsequent offences committed by the nominal after each stalking and harassment incident.

4 Feature Engineering

The feature sources listed in **Table 2** were utilised for feature engineering. For each nominal, the offence history was split into sequences as outlined in the previous section. As a result, a nominal may have multiple offence sequences, depending on the number of stalking and harassment incidents they committed, and whether these incidents were both preceded and followed by at least one other offence. Feature engineering is applied solely to the data up to and including each stalking and harassment offence. As previously shown in **Figure 2**, offences occurring after each stalking and harassment incident, within the subsequent 12 months, are used to create the labels (e.g., calculate total CCHI score and label data as 1 if total CCHI >= threshold and 0 otherwise).

Table 2. Feature sources used for feature engineering.

Feature source	**Description**
Nominal unique ID	The unique identifier for each nominal.
Cambridge Crime Harm Index (CCHI)	Measures total harm from crime based on sentencing guidelines.
Incident date	Date when the incident was reported to the police.
Date of birth	Nominal’s date of birth.
Gender	Nominal’s gender.
Role in crime	Whether the nominal was recorded as suspect or offender at the time of offending.
Suspect-victim relationship	The relationship between the victim and the nominal at the time of offending (e.g., partner, ex-partner, brother, etc.)
Number of custodies	The number of times the nominal was in custody.
Number of victim calls	The number of times a nominal who committed stalking and harassment as an offender had previously called the police as a victim for other crimes.
Keywords	Keywords extracted from incident logs. These are indicators for crimes related involving alcohol/drug abuse, weapons, threats, coercion and control, and other life endangering crimes.
LLM extracted features	Features extracted from incident logs using a large language model (LLM) with one-shot prompting.

4.1 Engineered numerical features

Derived metrics from the Cambridge Crime Harm Index (CCHI)[3] such as cumulative CCHI, change in CCHI, CCHI momentum, and decayed CCHI provide valuable insights into the severity and progression of offending behaviour over time. These metrics help in quantifying the harm associated with each offence and understanding how it evolves, which can facilitate predicting the risk of harm escalation and identifying individuals at risk of committing high-harm crimes. Here is why each metric is important and what it tells us:

[3] The Cambridge Crime Harm Index (CCHI) | Institute of Criminology

1. **Cumulative CCHI:** This metric accumulates the harm values from each offence over time, providing a running total of the severity of offences committed by an individual. It reflects the overall harm trajectory, helping to identify offenders who consistently engage in high-harm activities. High cumulative CCHI scores indicate a persistent pattern of severe offending, which is a red flag for potential future harm.

2. **Change in CCHI:** By measuring the difference in cumulative CCHI between consecutive offences, this metric highlights fluctuations in offending severity. It can identify whether the harm associated with offences is escalating or de-escalating, providing a dynamic view of an offender's behaviour. Sudden increases in harm can signal an emerging risk of severe offences, prompting timely interventions.

3. **CCHI Momentum:** CCHI momentum measures the rate of change in harm severity relative to the time between offences. It is calculated as:

$$\text{CCHI Momentum} = \frac{\text{Change in CCHI}}{\text{Days Since Previous Incident}}$$

   This metric captures how quickly the harm level is changing, offering insights into the acceleration of harmful behaviour. High momentum indicates a rapid increase in harm, suggesting an urgent need for intervention to prevent further escalation.

4. **Decayed CCHI:** This metric applies a decay function to the CCHI values to account for the time elapsed since each offence. The idea is that the relevance of an offence diminishes over time, especially if there has been a significant gap since the last incident. The decay function used is:

$$\text{Decayed CCHI} = \text{CCHI} \times \mathrm{e}^{-0.01\times(\text{Days Since Previous Incident}-30)}$$

   for days greater than 30. For days less than or equal to 30, the CCHI remains unchanged. This formula applies an exponential decay with a rate of 0.01, meaning that offences that occurred a longer time ago contribute less to the current harm level. The decay factor helps in emphasising recent offences more, which are likely more predictive of future behaviour.

5. **Decayed CCHI Momentum:** Similar to CCHI momentum, decayed CCHI momentum measures the rate of change in decayed CCHI over time, highlighting the adjusted rate at which an individual's harm potential is evolving. It combines the insights from decayed CCHI and momentum to provide a nuanced view of the escalation pattern, factoring in both the recency and severity of offences.

6. **Day Sin and Cos Transformations:** These transformations capture the cyclical nature of days within a month, allowing the model to recognise patterns related to specific days. The formulas used are:

$$\text{Day Sin} = \sin\left(\frac{2\pi \times \text{Day}}{31}\right), \text{Day Cos} = \cos\left(\frac{2\pi \times \text{Day}}{31}\right)$$

   These transformations help the model understand periodic trends, such as increased offending on certain days of the month.

7. **Month Sin and Cos Transformations:** Similar to day transformations, month sin and cos transformations account for the cyclical pattern of months in a year, aiding in the detection of seasonal trends in offending behaviour. The formulas are:

$$\text{Month Sin} = \sin\left(\frac{2\pi \times \text{Month}}{12}\right), \text{Month Cos} = \cos\left(\frac{2\pi \times \text{Month}}{12}\right)$$

These transformations allow the model to incorporate seasonal variations in the analysis, such as spikes in offending during certain times of the year.

Table 3. Summary of engineered numerical features.

Engineered feature name	**Description**
Age at offence	The number of days between incident date and nominal's date of birth.
Days since previous incident	Number of days between each incident and previous incident.
Cumulative CCHI	CCHI scores are cumulatively summed from one offence to the next within each sequence, capturing the total harm over time.
Change in cumulative CCHI	The change in cumulative CCHI is determined by calculating the difference in cumulative CCHI scores between consecutive offences.
CCHI rate	CCHI rate is calculated by dividing each CCHI value by the total CCHI sum for each crime history sequence, representing the relative contribution of each offence to the overall harm.
CCHI decayed	CCHI decayed values are adjusted using a decay factor based on the number of days since the previous incident, with decay applied if the interval exceeds 30 days.
Cumulative CCHI decayed	Cumulative CCHI decayed is calculated by cumulatively summing the decayed CCHI values within each sequence, reflecting the total adjusted harm over time.
Change in decayed CCHI	Change in decayed CCHI is determined by calculating the difference between consecutive decayed CCHI values within each sequence, highlighting variations in adjusted harm between offences.
CCHI momentum	CCHI momentum is computed by dividing the change in cumulative CCHI by the number of days since the previous incident, capturing the rate of harm escalation over time.
Decayed CCHI momentum	Decayed CCHI momentum is calculated by dividing the change in decayed CCHI by the number of days since the previous incident, capturing the rate of change in adjusted harm over time.
Day $\boldsymbol{sin}$ and $\boldsymbol{cos}$ transform	Day sin and cos transformations are applied to the extracted day of the incident date to capture the cyclical nature of days within a month. These transformations convert the day into two continuous features, representing the cyclical pattern of daily occurrences.
Month $\boldsymbol{sin}$ and $\boldsymbol{cos}$ transform	Month sin and cos transformations are applied to the extracted month from the incident date, converting the month into two continuous features. These transformations account for the cyclical nature of months within a year, helping to model seasonal patterns in the data.

4.2 Engineered binary features

The “gender” feature was transformed into two separate binary columns, representing male and female, to distinctly capture gender information in the data[4]. Additionally, a binary feature was created to indicate whether the nominal is under 30 years of age. This threshold was selected based on research findings in stalking and harassment, which suggest that individuals under 30 are at a higher risk of escalation.

The victim-nominal relationship was also encoded into binary features. Six new binary columns were generated to represent the relationship categories as outlined in **Table 4**, allowing for a more detailed analysis of these relational dynamics. The new relationship categories were derived by manually categorising the existing relationships in the data.

Table 4. Relationships between offenders and victims.

New relationship category	Relationships identified in the data
Partner	partner, spouse, legacy husband (inc common law), legacy boyfriend, legacy wife (inc common law), legacy estranged husband/wife, legacy girlfriend
Ex-partner	ex partner, legacy ex husband, legacy ex wife
Family member	son, brother, daughter, father, mother, sister, grandchild, relative – other, brother in law, step parent, step son, cousin, nephew, uncle, sister in law, grandparent, aunt, step daughter, niece, ex foster parent, legacy mother in law, legacy son in law, legacy step brother, legacy child, legacy father in law, legacy foster child, legacy daughter in law, appropriate adult, legacy step sister, great grandchild, foster parent, dependent, legacy extended family relation, legacy pupil
Professional	associate, colleague / work, business associates, lodger, landlord, customer, carer, overseen by, employer, employee, doctor / medical, legacy tenant, solicitor
Other/unknown	none, unknown, legacy other, relationship type refused, victim is crown, stranger, described person
Neighbour/friend/co-habitee	neighbour, friend, co-habitee

The keywords extracted from the incident logs for each offence were utilised to create ten additional binary feature columns. These features correspond to categories that represent key predictors and indicators identified as common or significant among stalking and harassment offenders who escalated to high harm offences, including life-threatening behaviours, as discussed in **Subsection 1.4**. These newly derived binary features are detailed in **Table 5**.

[4] Of course, this introduces the potential for the dummy variable trap, but this is not consequential for the methods used in these analyses.

Table 5. Crime type indicators derived from keywords extracted from incident logs.

Created crime type indicators	Keywords from incident logs
Physical violence/assault	actualbodilyharm, adultabuse, adultviolence, aggravate, aggravated, assalted, assault, assault occasioning, assaulted, assaulting, assaultoccasioning, batter, battered, battering, battery, beat, beating, bite, biting, bitten, brokenbone, brokenbones, commonassault, fracture, g.b.h., grevous, harming, hurt, inflict, inflict gbh, injured, intimateviolence, malicious, maliciously, serious harm, violent, wound, woundinggreivous
Sexual assault and abuse	ag fac child sexual abuse, ag fac child sexual exploitation, ag fac sex-based hostility, anus, attemptrape, buggery, child abuse, childviolence, clitorisrape, coercive, genitals, intimateabuse, molestation, molesting, partnerabuse, penetrate, penetrated, penetrating, penetration, penis, rape fem, rape female 16, sex ass, sexual, sexualgratification, sexualintercourse, sexualoffencesact2003, vagina
Threats and intimidation	malicious coms, ag fac alcohol, blackmail, coerce, conspiracy, conspire, conspired, conspiring, frighten, frightening, intimidating, intimidation, loiter, loitered, mailing, provocation, spy, spyed, spying, threat, threat endanger, threat letter, threaten dam, threatened, threatening, threateningwords, threats
Weapons and dangerous objects	acid, ag fac corrosive based offence, ag fac use of knife or other sharp instr, blade, bladed, corrosive, firearm, firearm fear, firearms, gun, knife, knives, machete, off weapon, poss firearm, poss gun, sword, weapon
Property damage and theft	burg res, dwelling, burg, burg dwell, burglar, burglary, damage, damaged , destroy, destroyed, destroying, dwelling, house, maliciousdamageact1861, personinherhome, rob, robbery, stealing, stolen, theft, theft dwell, theft from, theft mot veh, theft oth, theft ped, theft person, theftact1968
Coercion and control	forced, ag fac modern slavery, againstwill, circumcision, coercing, coercive, coercive control, control, controlled, controlling, femalecastration, femalegenitalmutilationact2003, forced marriage, forcedmarriage, infibulated, infibulation, intimidation, labiaminora, labour, manipulating, manipulation, modernslavery, modernslaveryact2015, oppress, oppression, perform labour, prostitute, prostituted, prostituting, prostitution, restrain, servile , sexualexploitation
Psychological abuse	malicious communications, malicious communications act 1988, distress , fear, fright, intimidation, malicious, malicious coms, psychological, threat, threat kill, threaten dam, threatened
Child-related crime	child, child abuse, childprotection, children, children abuse, children molest, childrenandyoungpersonsact1933, childviolence, fem pen 13, female 16+ no penetration, male 13+, male 16 or over, molest, molested, over16, under13, under16, undersixteen
Life endangering	actual bodily harm, attempt 18 gbh, death, deathriding, endangeringlife, homicide, kill, killed, killing, murder, mutilate, mutilated, strangle, strangulation, suffocate, suffocation
Breach of court order	breach, breachcriminalbehaviourorder, breached, breachofcriminalbehaviourorder

4.3 Engineered binary features using a Large Language Model

The features engineered from keywords found in the incident logs provide a straightforward, although rigid, approach to identifying potential signs of criminal behaviour. However, without proper context, it is difficult to determine whether these keywords truly reflect the behaviour or events they seem to indicate. For example, terms like “beat”, “beating,” “knife”, or “control” do not necessarily mean that these actions took place. The presence of these words in the logs could simply indicate a threat or mention of such actions, rather than their actual occurrence. For instance, a statement like "*I was threatened with being knifed*" suggests the mention of a knife but does not confirm that a knife was physically involved. The keyword approach, while useful for flagging certain terms, lacks the depth needed to discern whether these words genuinely signify the behaviours they imply.

To address this limitation, a large language model (LLM) was used to extract more nuanced features that better capture the true occurrence of behaviours such as the use of weapons, physical assaults like choking or beating, and other aggressive actions. The LLM excels at understanding the context within which an event is described, allowing it to distinguish between mere mentions of harmful actions and their actual occurrence. By considering the context in which terms appear, the LLM is able to identify behavioural patterns that are more aligned with what is documented in relevant research as indicators of serious harm, particularly in cases involving stalking or harassment. This contextual understanding allows the LLM to provide a more accurate assessment of escalating violence or controlling behaviours, thus offering a richer and more reliable set of features than what keywords alone can offer.

Despite the fact that certain behaviours, such as coercive control, may be identified by both the keywords and the LLM, it is advantageous from a predictive perspective to keep these two feature sets separate. Keywords give a direct, surface-level signal of behaviour, while the LLM provides a more sophisticated, contextually-driven interpretation. By maintaining the distinction between features extracted through keywords and those produced by the LLM, predictive models benefit from both approaches. The model can learn to weigh the importance of explicit mentions and context-based inferences differently, depending on the scenario. This separation ensures that the model has access to both raw, easily detectable signals and more subtle, complex insights, ultimately leading to improved predictive accuracy. Rather than diluting the strength of either method through merging, keeping them distinct enhances the model's flexibility and ability to detect patterns of criminal behaviour with greater precision.

4.3.1 LLM feature extraction with one-shot prompting

The large language model (LLM) used for advanced feature extraction in this case is the **Mistral-NeMo-12B-Instruct**[5], a cutting-edge, 12-billion parameter model developed in collaboration with NVIDIA. This model is notable for its substantial context window, which can process up to 128,000 tokens, allowing it to analyse large bodies of text at once. Its exceptional performance in areas such as reasoning, world knowledge, and coding accuracy positions it among the most capable models of its size. These capabilities make Mistral-NeMo particularly well-suited for understanding and processing complex narratives, such as those found in incident logs, where it must extract relevant features while making sense of nuanced contextual information. This model, as any LLMs are, was downloaded and used in the WMP computing environment.

In this specific application, the model was utilised to extract features that indicate key behavioural patterns, such as the use of weapons or physical violence, within the context of stalking and harassment. This process was guided by one-shot learning, a machine learning technique where the model is trained on just a single example to understand the task at hand. In the case of Mistral-NeMo, one-shot prompting involved providing the model with a carefully crafted example of how to extract features and justify decisions around flagging behaviours of

[5] Mistral NeMo | Mistral AI | Frontier AI in your hands

interest. By exposing the model to just one instance of this task, it understood how to generalise and apply similar reasoning to new, unseen data.

The one-shot prompting process started with a structured prompt based on a feature extraction template, which was derived from findings in the literature on predictors of escalating harm in stalking cases. This template defined the key behaviours to look for, such as signs of physical assault or controlling behaviours, and was informed by established research on what tends to escalate into more serious harm. To guide the model, a Data Scientist first manually performed the task of feature extraction and provided a rationale behind why specific behaviours were flagged. This human-generated example served as the reference for the model.

A list of 200 questions and answers, manually compiled from research on stalking and harassment, was used to evaluate the model's understanding of these behaviours, their patterns, and escalation triggers. The model was then tested and assessed to determine how well it could grasp the complexities of stalking and harassment. Given that the model was likely trained on relevant open-source literature, it was able to generate responses similar to those that had been manually compiled, demonstrating its capability to align with expert knowledge in this domain.

The template prompt was structured into several key components:

1. **User message**: The purpose of the user message is to outline the objective, offer context, and explain what needs to be done. The model can focus on achieving the desired outcome and apply its reasoning abilities to produce relevant, accurate results.

2. **Guidelines for feature extraction**: This section provides specific instructions on how to extract relevant features from the data. It offers criteria or principles that the model should follow to ensure consistency and precision in identifying the correct behavioural patterns.

3. **Sequence of events leading to stalking and harassment**: This part presents a chronological series of incidents related to stalking and harassment, giving the model the necessary context to analyse and extract the features of interest. It illustrates how behaviours escalate over time, providing a framework for understanding the context of the events.

4. **Example of expected output**: The final component shows an example of the desired output, demonstrating how the extracted features should be presented based on the provided sequence of events. This helps the model understand the format and structure of the results, ensuring the output aligns with the intended expectations.

User message:

```
**Purpose**: Your task is to use your knowledge of stalking and harassment behavioural research to extract relevant features from crime history logs for a given individual. Below, you are provided with an example that illustrates the feature extraction process. After reviewing the example, you must apply the same methodology to the provided incident logs to accurately extract and categorise the features.
```

```
# Guidelines for Feature Extraction

## 1. Relationship with Victim:
- Determine if shared child custody is present.

## 2. Threats, Violence, and Other Risk Factors
- **Feature Presence**:
  - For each feature, evaluate all incidents collectively.
  - If any incident contains evidence of a particular feature (e.g., physical assault, threats, stalking), mark the feature as present ("Yes").
  - **Example**: If one incident involved an explicit threat, mark "Explicit threats" as "Yes" even if other incidents did not include threats.
  - Do not separate incidents; instead, provide a single answer that encapsulates all historical context.

## 3. Aggregation of Information
- **Incident Review**:
  - Consider all incidents as a whole when determining whether a feature is present.
  - Look for patterns of behaviour across incidents that might indicate features such as coercive control, psychological abuse, or stalking.
  - **Example**: If multiple incidents suggest an escalation in behaviour (e.g., from verbal arguments to physical threats), this should be reflected in the extracted features.

## 4. Decision-Making Process
- **Binary Features**:
  - Each feature should be classified as "Yes" if any related evidence is found across the incidents.
  - If no evidence of a feature is found in any incidents, mark the feature as "No".
```

```
# Example of Incident Feature Extraction

**Series of incidents example**:

2017-01-02
Offender is related to the victim as ex-partner. Police was called about an altercation that occurred on a street. When Police arrived, both parties where questioned. The victim reported that she was stopped by her ex-partner on her way to her mother. No offences, assaults, or complaints were made. It was just an argument.

2017-02-01
Offender is related to the victim as ex-partner. The offender became verbally aggressive towards the victim …

2017-05-16
…

2018-10-22
…

2021-05-18
…
```

## Extracted Features:

## 1. Relationship with Victim
- **Shared child custody present** (No)
  Reason: There was no indication of shared child custody between the victim and the offender.

## 2. Threats
- **Explicit threats** (Yes)
  Reason: In the 2021-05-18 incident, the offender explicitly threatened the victim by saying, “watch your back” which the victim perceived as a threat to herself.

## 3. Physical Violence
- **Physical assault or beating** (No)
  Reason: There were no reports of direct physical assault or beating of the victim by the offender.
- **Sexual assault** (No)
  Reason: No incidents reported involved sexual assault.
- **Strangulation/Choking** (No)
  Reason: There were no reports of strangulation or choking.
- **Use of weapons** (No)
  Reason: No weapons were used directly against the victim.
- **Serious bodily harm (GBH/ABH)** (No)
  Reason: There were no reports of serious bodily harm such as GBH or ABH.
- **Evidence of persistent violent pattern** (No)
  Reason: While there were multiple verbal altercations, there was no evidence of a persistent pattern of physical violence.

## 4. Coercive and Controlling Behaviour
- **Coercive control** (Yes)
  Reason: In the 2018-10-22 incident, the offender tried to prevent the victim from speaking to the police, indicating controlling behaviour.
- **Retaliation for leaving** (Yes)
  Reason: In the 2016-11-20 incident, the offender could not accept that the relationship was over and refused to leave the victim’s address.
- **Evidence of wanting to regain control** (Yes)
  Reason: The offender’s refusal to leave the victim’s address and subsequent pleading in the 2016-11-20 incident showed an effort to regain control.
- **Spatially confining or restraining victim** (No)
  Reason: No incidents reported confining or restraining the victim.
- **Evidence of financial, social, or physical control** (Yes)
  Reason: The 2017-01-02 incident involved a financial dispute, indicating potential financial control.

## 5. Stalking and Surveillance
- **Following and surveillance behaviours** (No)
  Reason: There was no specific evidence of following or surveillance behaviours.
- **Use of tracking devices** (No)
  Reason: No tracking devices were reported.
- **Unwanted contact and communications** (Yes)
  Reason: The 2021-05-18 incident involved unwanted threatening communication towards the victim.
- **Unwanted intrusions** (Yes)
  Reason: The offender repeatedly intruded into the victim’s personal space in several incidents (e.g., 2016-11-20, 2017-05-16).

## 6. Property Damage
- **Vandalism or arson** (No)
  Reason: There were no incidents involving vandalism or arson.

```
## 7. Psychological Abuse
- **Spreading false rumors** (Yes)
  **Reason:** The 2021-05-18 incident involved the offender telling personal matters to neighbours, suggesting an intent to manipulate or harm the victim’s reputation.
- **Blackmailing** (No)
  **Reason:** No incidents of blackmail were reported.
- **Impersonating victim** (No)
  **Reason:** There were no reports of impersonation.
- **Evidence of fixation/obsession** (Yes)
  **Reason:** The 2016-11-20 incident where the offender begged the victim to take her back shows obsessive behaviour.
- **Public confrontation/arguments** (Yes)
  **Reason:** Multiple incidents involved public arguments, such as the 2017-02-01 street altercation.

## 8. Legal and Procedural
- **Court order issued** (No)
  **Reason:** No court orders were mentioned in the incidents.
- **Breach of legal orders** (No)
  **Reason:** There were no reported breaches of legal orders.

## 9. Opportunistic Factors
- **Accessibility to the victim** (Yes)
  **Reason:** The offender repeatedly accessed the victim at home and on the street (e.g., 2017-02-01, 2016-11-20).
- **Gathering personal information** (No)
  **Reason:** No incidents involved gathering personal information.
- **Increased frequency/intensity of pursuit** (No)
  **Reason:** The pursuit did not show a marked increase in frequency or intensity over time.
- **Standing/littering around victim’s home/school/work** (No)
  **Reason:** There was no mention of the offender loitering around the victim’s home, school, or work.

## 10. Other Risk Factors
- **Substance or alcohol abuse** (No)
  **Reason:** There was no direct mention of substance or alcohol abuse by the offender.
- **Documented mental health issues** (Not explicitly stated)
  **Reason:** There was no direct mention of mental health issues.
- **Evidence of social instability** (Yes)
  **Reason:** Multiple incidents involved unstable behaviour, such as outbursts and refusal to leave when asked.
- **Victim fear** (Yes)
  **Reason:** The 2021-05-18 incident explicitly states that the victim felt threatened for her safety.
- **History of domestic violence** (No)
  **Reason:** There was no direct history of physical domestic violence reported.
- **Presence of step-child** (No)
  **Reason:** There was no mention of a step-child in the incidents.
- **Victim’s pregnancy** (No)
  **Reason:** There was no mention of pregnancy in the incidents.
- **Separation from the perpetrator** (Yes)
  **Reason:** Several incidents occurred after the relationship had ended, demonstrating ongoing issues post-separation.

---END OF EXAMPLE---
```

The LLM will process the provided instructions, guidelines, series of incidents, and example along with the feature extraction rationale. Based on this information, it will generate its response in the exact same format as shown in the “Extracted Features” section below.

```
# **Series of incidents to analyse -- follow the exact template and order as seen in the example**:

{incidents}

## Extracted Features:
{{Add the extracted features here. Do not output anything else after the extracted features. End the response here.}}
```

In order to control the randomness and diversity of the model's output, the `temperature` and `top_p` hyperparameters were set to 0.2. The `temperature` parameter influences how confidently the model selects its next word; lowering it to 0.2 makes the output more focused and deterministic, reducing creative variation. The `top_p` parameter, which limits the selection to a subset of high-probability tokens, was also set to 0.2, ensuring that only the most likely options are considered. These settings help produce more objective and reproducible results by minimising unpredictability in the model's responses.[6] [7]

4.3.2 LLM extracted features

A total of 29 features were extracted from the LLM's outputs for each sequence of events leading up to and including the occurrence of stalking and harassment for each individual. These features are detailed in **Table 6**.

Table 6. Binary features extracted with the LLM.

Blackmailing	Impersonating victim	Spreading false rumors
Breach of legal orders	Physical assault or beating	Standing/littering around victim's home/school/work
Coercive control	Presence of step-child	Strangulation/Choking
Court order issued	Public confrontation/arguments	Substance or alcohol abuse
Documented mental health issues	Retaliation for leaving	Unwanted intrusions
Evidence of financial, social, or physical control	Separation from the perpetrator	Use of tracking devices
Evidence of persistent violent pattern	Serious bodily harm (GBH/ABH)	Use of weapons
Explicit threats	Sexual assault	Vandalism or arson
Following and surveillance behaviours	Shared child custody present	Victim's pregnancy
History of domestic violence	Spatially confining or restraining victim	

[6] Cheat Sheet: Mastering Temperature and Top_p in ChatGPT API - API - OpenAI Developer Forum

[7] Text generation - OpenAI API

4.4 Summary of all features used for exploratory data analysis and modelling

Table 7. Summary of selected features for initial exploratory data analysis and modelling.

Numerical features	**Binary features from incident log keywords**	**Binary features extracted from incident logs with the LLM**
Cambridge Crime Harm Index (CCHI)	Is Male	Blackmailing
Number of custodies	Is Female	Breach of legal orders
Number of victim calls	Is Partner	Coercive control
Age at offence	Is Ex-Partner	Court order issued
Days since previous incident	Is Family Member	Documented mental health issues
Cumulative CCHI	Relationship Is Professional	Evidence of financial, social, or physical control
Change in cumulative CCHI	Relationship Is Other/Unknown	Evidence of persistent violent pattern
CCHI rate	Relationship Is Neighbour/Friend/Co-Habitee	Explicit threats
CCHI decayed	Physical Violence/Assault	Following and surveillance behaviours
Cumulative CCHI decayed	Sexual Assault and Abuse	History of domestic violence
Change in decayed CCHI	Threats and Intimidation	Impersonating victim
CCHI momentum	Weapons and Dangerous Objects	Physical assault or beating
Decayed CCHI momentum	Property Damage and Theft	Presence of step-child
Day $\boldsymbol{sin}$ and $\boldsymbol{cos}$ transform	Coercion and Control	Public confrontation/arguments
Month $\boldsymbol{sin}$ and $\boldsymbol{cos}$ transform	Psychological Abuse	Retaliation for leaving
	Child-Related Crime	Separation from the perpetrator
	Life Endangering	Serious bodily harm (GBH/ABH)
	Breach of Court Order	Sexual assault
		Shared child custody present
		Spatially confining or restraining victim
		Spreading false rumors
		Standing/littering around victim's home/school/work
		Strangulation/Choking
		Substance or alcohol abuse
		Unwanted intrusions
		Use of tracking devices
		Use of weapons
		Vandalism or arson
		Victim's pregnancy

5 Findings from Exploratory Data Analysis

The complete exploratory data analysis, provided in **Appendix A**, begins by examining the types of crimes associated with different scores on the Cambridge Crime Harm Index (CCHI). This helps to clarify the actual harm quantified by these scores, providing a clearer understanding of the labeling process before the modeling stage. It then explores the distribution of harm scores and their connection to historical offences, focusing on escalation patterns. Additionally, it analyses other aspects, such as gender, the relationship between offenders and victims, and the types of crimes identified through keywords extracted from incident logs, with particular attention to harm levels before and after stalking and harassment offences. Lastly, it incorporates the analysis of features extracted by the LLM, investigating how these features correlate with harm scores, escalation patterns, and specific types of offenses. This provides further insights into the behaviours and actions leading to increased harm, enhancing the understanding of both pre- and post-stalking incidents.

5.1 Relationship between harm and tabular features

The findings revealed a consistent lack of strong linear correlations between the examined features and the severity of harm following stalking and harassment offences. For instance, while a significant portion of offenders had prior offences, there was no clear trend indicating that a higher number of prior offences led to increased harm post-offence. Similarly, gender analysis showed that male offenders generally had higher harm scores both before and after the stalking and harassment incidents; however, the variability was substantial, and no linear pattern emerged. Age also demonstrated a very weak and statistically insignificant correlation with post-offence harm levels, indicating that age alone is not a reliable predictor of future harm severity.

Further analysis of roles in crime and victim-offender relationships underscored the complexity of predicting harm based on these features. Offenders who were both perpetrators and victims exhibited a wider range of harm scores post-offence compared to those who were solely offenders, but again, without a clear linear association. Relationships categorised as ex-partners and current partners were associated with higher harm scores, yet the data did not support a linear relationship that could be utilised effectively in predictive models.

The implications for modelling are significant. The absence of strong linear associations suggests that predictive models relying solely on these features may have limited accuracy in forecasting future harm severity. This underscores the necessity for models to incorporate a broader range of variables, potentially including qualitative aspects of prior offences, behavioural patterns, mental health indicators, and other social factors (noting that there may still be scope for interactions between features to provide useful information). However, this does not rule out the potential effectiveness of models that can capture non-linear patterns.

5.2 Relationship between harm and features extracted with the LLM

Figure 3 illustrates the prevalence of various crimes in cases where harm levels before stalking and harassment (SH) were lower than 180, while harm levels after SH were equal to or greater than 180. The most significant differences in crime prevalence are observed in categories such as physical assault or beating, coercive control, retaliation for leaving, unwanted intrusions, vandalism and arson, public confrontations/arguments, court orders issued, breaches of legal orders, history of domestic violence, separation from the perpetrator, and shared child custody arrangements.

These crimes were notably more frequent in the offending history prior to the SH offence, compared to offending behaviour after the SH offence. This trend aligns with existing literature, which often highlights these crimes as associated with patterns of escalation leading up to stalking and harassment. It is clear that, in many cases, these behaviours intensify in the lead-up to SH offences.

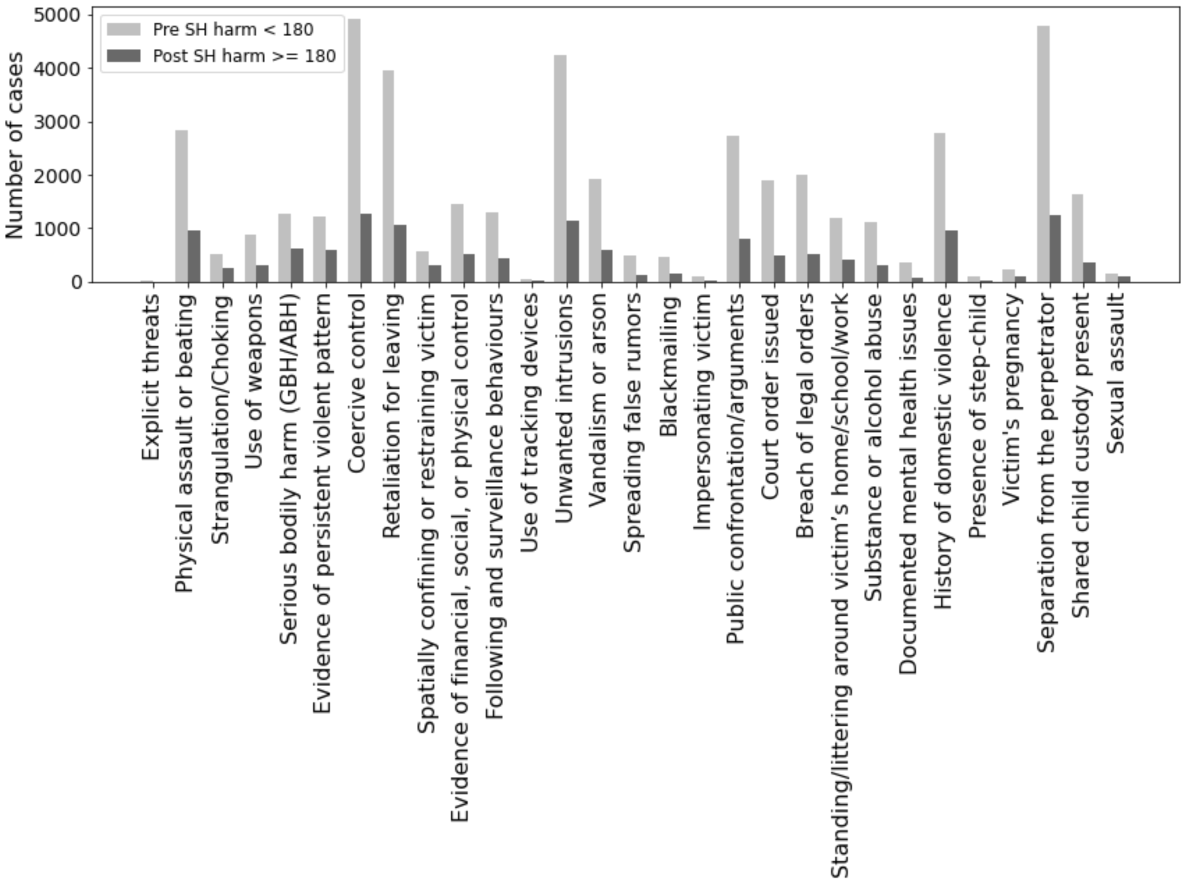


Figure 3. Prevalent crimes associated were total harm scores pre- and post-stalking and harassment was lower than 180 and equal/higher than 180, respectively.

What stands out, however, is the significant drop in the continuation of such criminal behaviours post-SH. A much smaller proportion of offenders maintained the same level of criminal escalation after the SH offence. This could suggest that interventions, whether legal, social, or relational, may play a role in reducing the intensity of harm after SH offences. Alternatively, it may reflect a natural de-escalation following the SH offence, possibly due to a culmination or breaking point in the offender's behaviour.

Furthermore, this pattern could indicate that the SH offence serves as a key moment in the offender's criminal trajectory, marking either the peak of their offending behaviour or a shift in their patterns of harm. The fact that crimes like breaches of court orders and physical violence are more prevalent pre-SH suggests that these factors could serve as early warning signs for law enforcement and social services in identifying cases at risk of significant harm escalation.

Figure 4 compares the prevalence of various crime categories where total harm before and after stalking and harassment (SH) was equal to or greater than 180. The analysis reveals that a higher proportion of offenders who exhibited significant harm (≥180) before SH also maintained similar levels of harm following the SH offence. This suggests that individuals with high pre-SH harm are more predictable in terms of their likelihood to continue causing elevated harm post-SH, compared to those who displayed lower levels of harm before the SH offence.

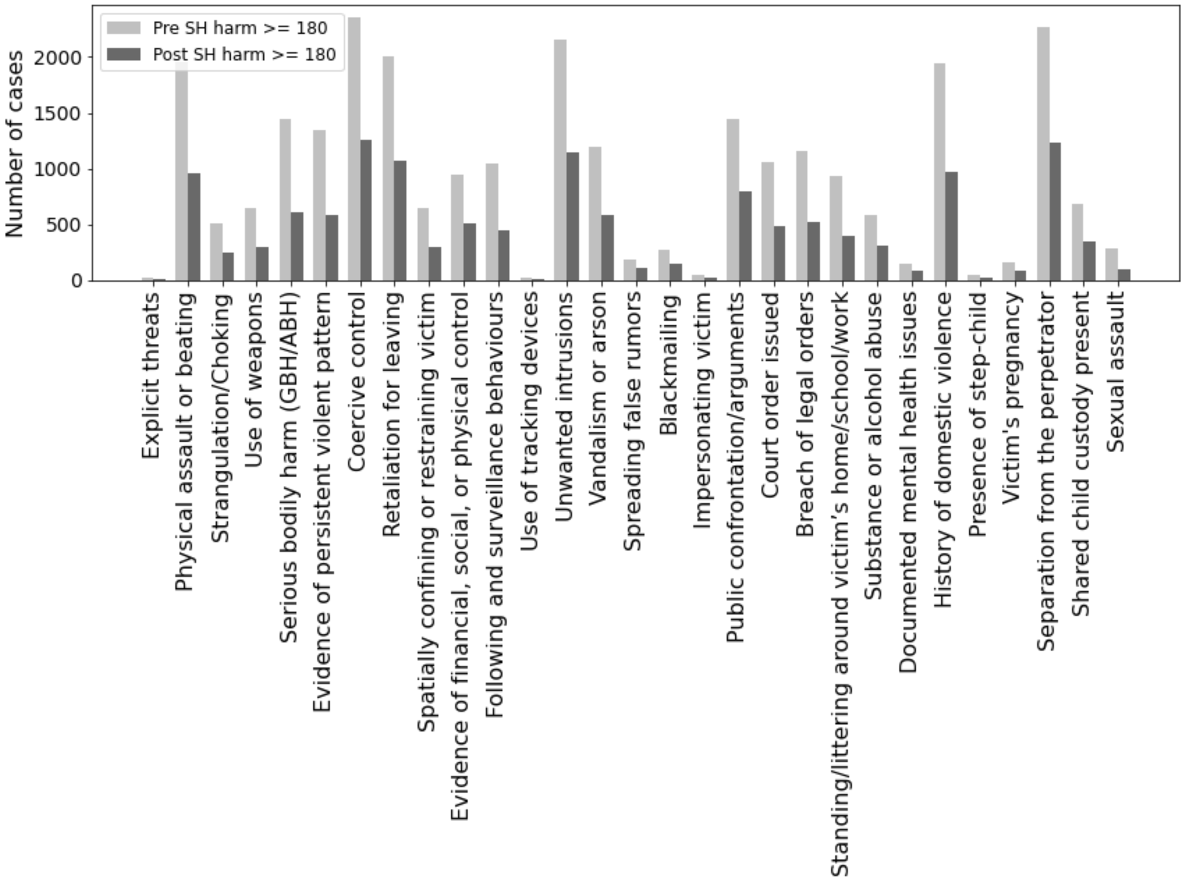


Figure 4. Prevalent crimes associated were total harm scores for both pre- and post-stalking and harassment were equal or higher than 180.

However, even in the most prominent crime categories – such as physical assault or beating, serious bodily harm (grievous bodily harm), coercive control, retaliation for leaving, unwanted intrusions, breaches of legal orders, court orders issued, history of domestic violence, and shared child custody – no more than half of the cases exhibited harm levels equal to or greater than 180 both before and after SH. Despite this, the ratio of cases with consistently high harm pre- and post-SH is notably higher than what was observed in **Figure 3**, indicating a stronger correlation between pre- and post-SH harm in these cases.

This pattern reinforces the idea that while high pre-SH harm may serve as an indicator of future risk, escalation to severe harm post-SH is not guaranteed, even among individuals with a history of significant violence.

Figure 5 highlights cases where the pre-stalking and harassment (SH) harm was equal to or greater than 180, but post-SH harm fell below 180. Interestingly, the same crime categories that were previously associated with higher pre-SH harm – such as physical assault or beating, coercive control, retaliation for leaving, unwanted intrusions, breaches of legal orders, court orders issued, history of domestic violence, and shared child custody – display a mixed pattern of harm escalation.

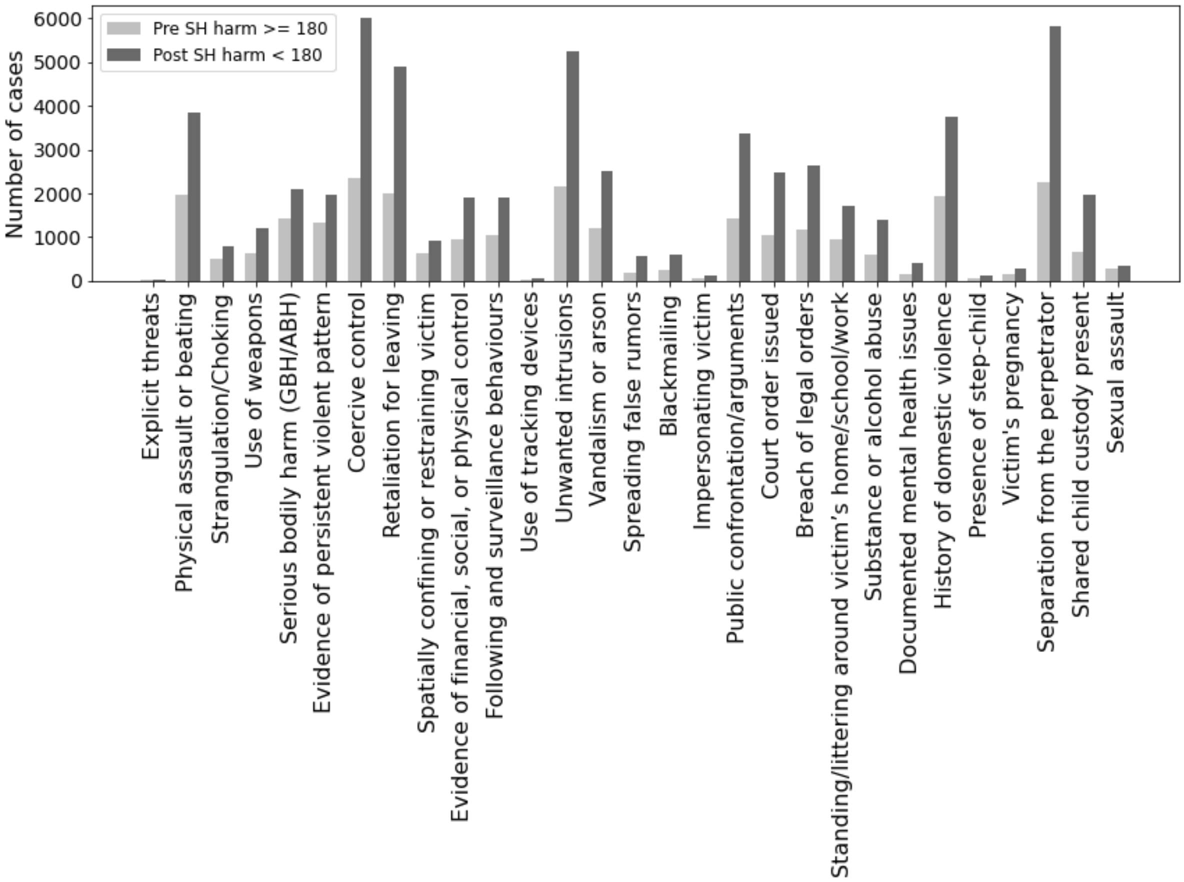


Figure 5. Prevalent crimes associated were total harm scores pre- and post-stalking and harassment were equal higher than 180 and lower than 180, respectively.

In approximately half of the cases, there is no consistent escalation in harm, with post-SH harm dropping below 180 despite a high pre-SH harm score. This lack of continued harm escalation suggests that, while high pre-SH harm is a strong indicator of potential risk, it does not universally translate into continued high harm post-SH.

The data from **Figure 5** also highlights that even high pre-SH harm individuals do not always maintain high harm levels post-SH. This introduces an element of variability in predicting post-SH behaviour. While high pre-SH harm is a useful predictor, it is variable, suggesting that there are other influencing factors, either external or situational, that could mitigate post-SH harm in some cases. This variability should be accounted for in any predictive model by incorporating additional features beyond just harm levels.

Figure 6 illustrates the prevalence of crimes in cases where both pre- and post-stalking and harassment (SH) harm levels remained below the threshold of 180. The data shows that, for most individuals (nominals), the levels of harm are relatively consistent, with minimal escalation across the pre- and post-SH periods. In these cases, the escalation in harm is marginal, and the overall harm levels remain below the threshold of 180 both before and after the SH offence.

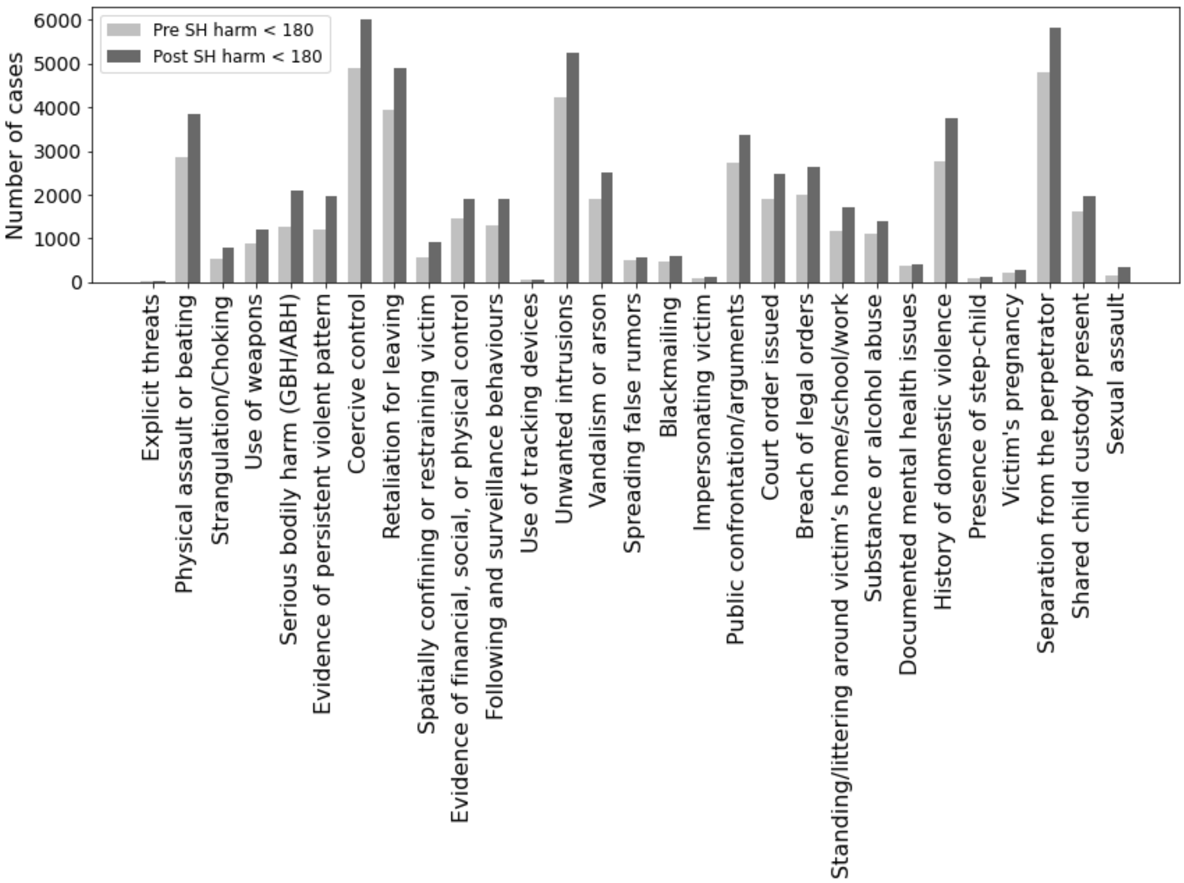


Figure 6. Prevalent crimes associated were total harm scores for both pre- and post-stalking and harassment were lower than 180.

Interestingly, the pattern suggests that the prevalence of crimes slightly increases post-SH compared to pre-SH, but the difference is relatively small. This implies that while some offenders may exhibit a mild escalation in harmful behaviour following the SH offence, the majority do not cross the threshold into more severe or high-harm behaviours.

Across the figures, there is a clear trend showing that individuals with higher levels of pre-SH harm are more likely to continue exhibiting higher harm post-SH. In **Figure 4**, for example, we see that those with pre-SH harm scores lower than or equal to 180 tend to have higher harm levels pre-SH. However, **Figure 5** reveals that this is not a universal rule – approximately half of the offenders with high pre-SH harm exhibited lower harm post-SH. This variability suggests that while pre-SH harm is a useful indicator, it is variable, and other factors may contribute to whether harm escalates or de-escalates after an SH offence.

In contrast, **Figure 6**, which focuses on individuals with pre- and post-SH harm below the threshold of 180, highlights a much more stable and predictable pattern of low-level offending. Offenders in this category generally do not escalate to high harm post-SH, and the slight increase in crime prevalence post-SH suggests persistent, low-severity offences rather than significant harm escalation.

The findings across the figures indicate that pre-SH harm is indeed a valuable predictive feature for post-SH harm, but it should be used in conjunction with other factors. High pre-SH harm generally correlates with a higher likelihood of future harmful behaviour, as evidenced in

Figure 4. However, the inconsistency seen in **Figure 5**, where many high-harm offenders do not escalate post-SH, points to the need for additional variables in predictive models.

5.3 Challenges in identifying escalation to serious harm

In the exploratory analysis conducted on the relationship between pre- and post-stalking and harassment harm scores, as well as other derived harm derived, gender, age, and crime indicators (both keyword and LLM-extracted), no clear linear relationships or distinct patterns of harm escalation were found. Instead, the results revealed a variety of patterns, independent of harm scores, that showed both escalation and de-escalation in harm. This suggests that the relationships between criminal behaviour before and after stalking and harassment are likely complex and non-linear.

Due to these complexities, we manually sampled several offenders (nominals) to further investigate their incident histories pre- and post-SH. Key observations include the following:

- In cases where only one or two events occur pre-SH, followed by a very serious crime, such as grievous bodily harm, harm scores and derived metrics provide little insight into how harm accumulates over time. The short time frame and low pre-SH harm levels make it difficult to predict future escalation.
- The use of a 12-month window for evaluating escalation may lead to misclassification. The 12-month window was used since it reflects a period of time within which the risk of harm escalation tends to decrease significantly by the end of the 12th month (as also supported by some of the sources cited in this analysis). For example, some nominals commit serious offences just outside the 12-month period, meaning their escalation patterns are not captured within the model's time constraints. To address this, we developed multiple crime sequences at the occurrence of each SH offence, allowing us to track harm over a broader time frame for each SH event (as exemplified in **Subsection 3.2**).
- Nominals who escalate to serious offences just after the 12-month window pose a challenge for prediction. Although clear patterns of escalation may be evident pre-SH, the model's time frame restricts the ability to flag these individuals as high-risk.
- Certain escalations in harm appear to be coincidental, such as random public encounters between offender and victim that lead to serious incidents. These types of events, driven by chance, are inherently difficult to predict.
- There are extreme cases where low pre-SH harm leads to a very serious post-SH crime, and vice-versa. These unpredictable cases underscore the limitations of using static variables such as harm scores, age, and gender, suggesting that contextual information from incident logs might offer deeper insights into offender behaviour.

As seen in the earlier analysis, in nearly half of the cases, high harm pre-SH does not result in a high-harm post-SH offence, despite an escalation pattern that might suggest otherwise. This inconsistency could be influenced by external factors we cannot easily quantify, such as changes in the lives of offenders and victims. Examples include relocation, changes in employment, or shifts in personal relationships, such as offenders finding new partners, which may reduce their focus on the original victim.

The inability to consistently predict harm escalation post-SH highlights the complexity of criminal behaviour and the influence of unquantifiable (for the Police) life changes. While harm scores and other metrics provide valuable insights, they fall short in capturing the nuanced dynamics that drive harm escalation or de-escalation.

5.4 Examination of non-linear associations

This section explores the presence of non-linear relationships between harm scores and other features commonly cited in the literature (discussed in **Subsection 1.4**) as indicators of escalation to serious crime. Previous analyses did not reveal clear linear correlations, prompting us to investigate whether non-linear associations exist among features linked to each crime event. To achieve this, we use visualisation techniques such as t-SNE (t-Distributed Stochastic Neighbour Embedding) and UMAP (Uniform Manifold Approximation and Projection).

t-SNE excels at capturing complex patterns in high-dimensional data by mapping similar points closer together in a lower-dimensional space, making it effective at uncovering clusters or structures that are not immediately apparent. UMAP, on the other hand, preserves both the local and global data structures, providing a more balanced visualisation that can reveal broader relationships while maintaining local neighbourhood integrity. These complementary methods allow us to examine different aspects of non-linear relationships within the data.

Additionally, we use an autoencoder to encode non-linear relationships within the data before plotting with t-SNE and UMAP. Autoencoders, which are neural networks designed to learn efficient representations of input data, can capture intricate, high-dimensional non-linear patterns that might be missed by traditional analyses. By reducing the data into a compressed latent space, autoencoders can highlight underlying non-linear structures and relationships, potentially offering deeper insights into the factors driving escalation in harm.

For each crime event, harm scores and derived metrics – such as cumulative harm, change in cumulative harm, harm momentum, harm decayed, cumulative harm decayed, and harm momentum decayed – were scaled using log-transformations to manage skewness and improve interpretability. Additional features, including the number of victim calls, number of custodial sentences, age at offence, and days between previous incidents, were also log-transformed. Binary and cyclical features (e.g., sine and cosine transformations of day and month) were included in the final feature vector for each crime incident. Finally, each sequence of events associated with a nominal (e.g., nominalA_1, nominalA_2, nominal_B_1, etc.) are concatenated into a long feature vector and padded with zero to the length of the longest feature vector. This way, all inputs have the same length when using them for plotting and training the autoencoder model.

5.4.1 Visualisation of non-linear relationships among original features using t-SNE and UMAP

The t-SNE plot demonstrates the data distribution across two components, revealing clusters that suggest potential groupings or patterns within the features. The t-SNE algorithm is effective at capturing local similarities, grouping similar data points together based on the high-dimensional structure of the input data. However, the plot appears somewhat densely packed, with overlapping clusters, which may indicate that while local patterns exist, distinguishing broader trends or separations in the data is more challenging. The color scale, representing the total harm scores after stalking and harassment (log-transformed), shows a gradient of harm levels spread throughout the clusters, suggesting varied harm outcomes within similar feature groups. The relationship between pre-offense features and resulting harm is complex and not easily separable in lower dimensions.

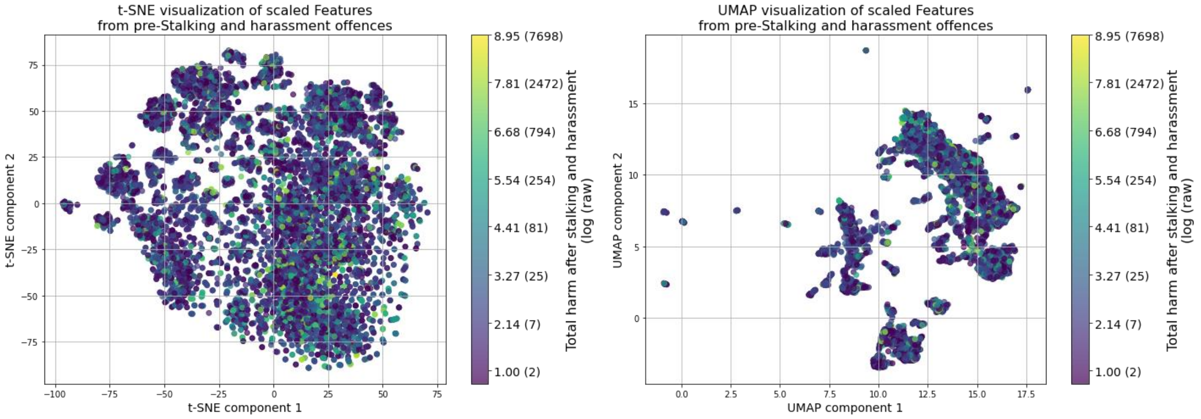


Figure 7. Non-linear relationships based on raw features using t-SNE and UMAP.

The UMAP plot in **Figure 7** presents a more distinct separation of data points into multiple clusters, providing a clearer view of both local and global relationships within the data. UMAP preserves both the local neighborhood structure and larger-scale relationships better than t-SNE, making it particularly useful for identifying broader patterns in complex datasets. The distinct clustering observed in the UMAP plot suggests that certain features or combinations of features may be strongly associated with specific harm outcomes. The color gradient indicates that harm scores vary across clusters, with some clusters associated with higher harm levels, potentially highlighting subsets of offences that contribute more significantly to harm escalation. There may be some underlying patterns or groupings in the data, but they are not strongly delineated by harm level alone.

5.4.2 Visualisation of non-linear relationships among encoded features using t-SNE and UMAP

Initially, the raw scaled features were visualised using dimensionality reduction techniques like t-SNE and UMAP to explore potential non-linear relationships, such as clusters or patterns within the data. These techniques are effective at mapping high-dimensional data into a lower-dimensional space for visualisation, which can reveal underlying structures that may not be evident from the raw data alone. However, t-SNE and UMAP directly applied to raw features often capture only certain aspects of the data’s inherent complexity, and the resulting visualisations may sometimes be noisy or ambiguous, especially when the original features are not optimised for revealing intricate relationships.

To further enhance the exploration of these non-linear relationships, the next step involves encoding the raw scaled features using an autoencoder and then reapplying t-SNE and UMAP on these encoded features. The autoencoder compresses the data into a lower-dimensional representation, which is designed to retain the most critical information while filtering out noise and less relevant variations. This encoding process can help in refining the data representation by capturing more complex patterns that are not readily accessible in the raw feature space.

Autoencoder architecture:

1. **Encoder**:
    - The encoder compresses the input data into a lower-dimensional space (encoding).
    - The architecture begins with an input layer that matches the dimensionality of the input data.
    - This input is passed through three fully connected (Dense) layers with ReLU (Rectified Linear Unit) activation functions:
        - The first layer has 256 neurons.
        - The second layer has 128 neurons.
        - The final encoding layer (bottleneck) has 64 neurons, which serves as the encoded representation or the bottleneck of the model.
2. **Decoder**:
    - The decoder reconstructs the input data from the encoded representation.
    - The decoder mirrors the encoder's structure but in reverse:
        - It begins with a layer of 128 neurons.
        - This is followed by a layer of 256 neurons.
        - The final output layer matches the original input dimension, using a linear activation function to allow a wide range of output values.
3. **Model structure**:
    - The encoder and decoder are combined into a single autoencoder model, where the input flows from the encoder through the bottleneck to the decoder, outputting a reconstructed version of the input data.

Mean Squared Error (MSE) as loss function:

- The loss function used in this autoencoder is Mean Squared Error, which measures the average of the squares of the errors, i.e., the average squared difference between the estimated values and the actual value.
- MSE is appropriate for autoencoders as it quantifies how well the model's output approximates the input, making it suitable for reconstruction tasks where the goal is to minimise the difference between the original and reconstructed data.

Learning rate as cosine decay with restarts for scheduling:

- The learning rate is controlled by a cosine decay schedule with restarts.
- This schedule starts with an initial learning rate (0.001) and decays it following a cosine function, resetting the learning rate at specified intervals/steps (1000) to help the model escape local minima and improve convergence.
- Parameters like time multiplication factor (set to 2.0), multiplier for the maximum learning rate (0.9), and alpha (0.0001) further fine-tune how the decay and restarts are

applied, promoting better learning dynamics by periodically increasing the learning rate.

- The Multiplication factor scales the period of each restart. Specifically, it multiplies the number of steps between each restart by this factor. A value of 2.0 means that each subsequent period (the number of steps before the next restart) will be twice as long as the previous one. This gradual increase in the period helps the learning rate decay more slowly over time, allowing longer training phases as the model stabilises.
- The multiplier for the maximum learning rate scales the learning rate.
- Alpha determines the minimum learning rate floor, ensuring that the learning rate does not decay to zero, which keeps the model learning at a minimal rate even in later stages of training.

Adam optimiser:

- The autoencoder uses the Adam optimiser, which is a popular choice for training neural networks due to its adaptive learning rate capabilities and robust performance across various tasks.
- Adam combines the advantages of two other popular extensions of stochastic gradient descent: Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp).

Baseline MSE vs. Autoencoder MSE for model evaluation:

- The performance of the autoencoder is evaluated against a baseline MSE, which represents the mean squared error if the input data were predicted by a simple mean of the features.
- The autoencoder achieved a lower MSE (0.073) compared to the baseline (0.77), indicating a 90.54% improvement, demonstrating the effectiveness of the autoencoder in learning and compressing data.

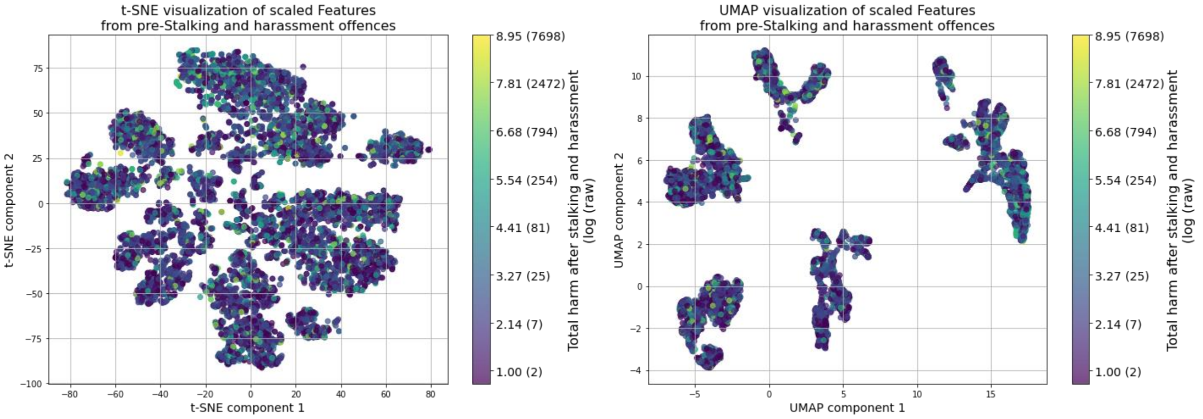


Figure 8. Non-linear relationships based on encoded features using t-SNE and UMAP.

The application of an autoencoder to the pre-stalking and harassment offense features has yielded significant improvements in data visualisation and potential insights, as per **Figure 8**. Both t-SNE and UMAP visualisations of the encoded features reveal clearer structures and distinct groupings compared to the previous plots of raw features. This suggests that the autoencoder has successfully learned to compress the data while preserving and emphasising important structural information.

The t-SNE visualisation now displays discernible clusters, indicating that the encoded features capture meaningful patterns in the data. Similarly, the UMAP plot exhibits even more pronounced structure, with distinct, separated clusters or "islands" of data points. This consistency between methods increases confidence in the observed patterns and underscores the effectiveness of the autoencoder in distilling discriminative features from the original dataset.

Interestingly, while the clusters in both visualisations show some internal variation in harm levels (as indicated by colour), there is not a clear separation of harm levels across clusters. This suggests a complex relationship between the encoded features and the total harm resulting from stalking and harassment. The clusters likely represent different types or patterns of offenses that do not directly correlate with harm levels but could be important for understanding the nature of these incidents.

5.5 Summary

- The analysis revealed that linear relationships between harm scores and commonly cited features (e.g., number of previous offences, types of crimes) were not evident. This suggests that harm escalation is driven by complex, non-linear interactions among features.
- Visualisation techniques like t-SNE and UMAP highlighted potential non-linear patterns within the data, with UMAP showing more distinct clusters that suggest broader groupings of offences and characteristics associated with different harm levels.
- The analyses emphasised that the count of previous offences alone (e.g., physical violence, threats, child-related crimes) was not a strong predictor of post-offence harm. Instead, the context, severity, and behavioural patterns of the offences play a more critical role.
- The data highlighted significant variability in harm outcomes across different crime types (e.g., physical violence, psychological abuse, breaches of court orders). Certain crime types, such as those involving ex-partners or coercion and control, displayed broader distributions of harm scores, underscoring the importance of understanding the specific dynamics of each offence.
- Simply counting the number of offences (e.g., breaches, violent offences, psychological abuse) was found to be insufficient in capturing the risk of harm escalation. The absence of clear linear relationships suggests that other factors, such as the timing, context, and interactions between offences, must be considered.

6 Methodology

This section covers the data labeling strategy and the various modeling approaches employed.

6.1 Establishing labelling criteria for the predictive model

This subsection aims to gain a deeper understanding of how harm scores translate into actual crime events. To achieve this, each unique harm score is analysed (see **Appendix B**) with a view to subsequently determine the appropriate threshold for defining the classes to be predicted (positive or negative). Based on the crimes associated with different CCHI scores, those with a **harm score of 180 (days) or higher** represent the most serious types of offences. This score threshold was also validated by a subject matter expert.

6.2 Deciding on the harm score threshold for labeling post-stalking and harassment risk

As per discussions in **Appendix B**, it was established that a total harm score post-stalking and harassment equivalent to 6 months in prison (corresponding to a CCHI score of 180) is the most appropriate benchmark for determining whether an offender (nominal) should be classified as low or high-risk.

First, it is important to understand what total harm can indicate. As demonstrated in **Figure 9**, a high total harm score can be linked to two main scenarios:

- **A single serious offence**: This could involve incidents resulting in severe physical injuries, where the seriousness of the offence justifies close monitoring.

A series of lower to medium-level offences: While individually these offences may not warrant immediate intervention, their cumulative effect could signal a pattern of escalating harm. Such patterns could potentially culminate in a serious offence, indicating the need for careful and ongoing monitoring.

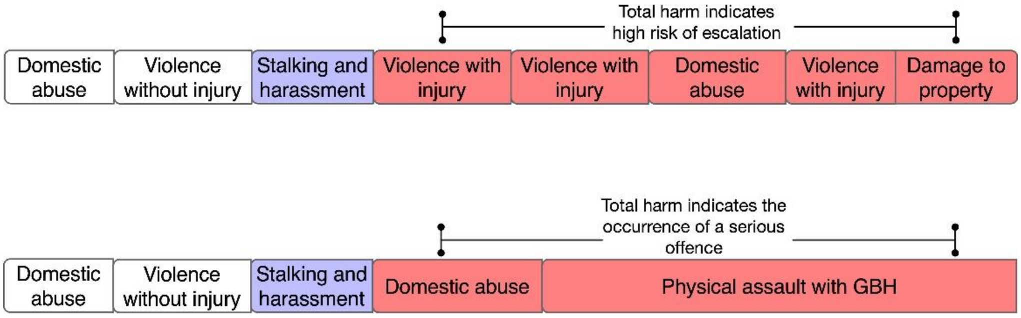


Figure 9. Meaning of total harm post-stalking and harassment based on offending pattern.

While we considered setting the harm score threshold at 180, we recognised that this might be too restrictive and could miss opportunities for early intervention to prevent escalation. A lower score threshold would allow a nominal to be identified as potentially high-risk sooner, rather than waiting until the point when a high-risk offense is imminent. Setting the threshold at 120 after stalking and harassment offenses, instead, enables an earlier identification of harmful behaviour, facilitating a more proactive response.

A CCHI score difference of 60 – equivalent to 60 days in prison – would suffice to signal high-risk individuals without flagging them excessively early. This lower threshold aims to catch escalating offenders before they reach severe harm levels, creating an opportunity to intervene before the risk becomes critical. Consequently, two classification groups emerge: the negative class, where the total post-stalking and harassment harm is below 120, and the positive class, where the total harm level meets or exceeds 120 in the 12 months following the occurrence of stalking and harassment.

6.3 Modelling approaches

This subsection outlines the model architectures used and the process of formatting input data prior to training. The dataset is highly imbalanced, containing 7,771 samples from the negative class and only 1,218 samples from the positive class. The approach to dealing with the class imbalance is discussed in **Appendix D**.

6.3.1 Ensemble methods, and meta-learner

The first model trained was an XGBoost. Following this, two ensemble models were developed: one using soft voting and the other utilising a shallow neural network meta-learner. Both ensembles were composed of the same underlying models: XGBoost, random forest, LightGBM, and SVM.

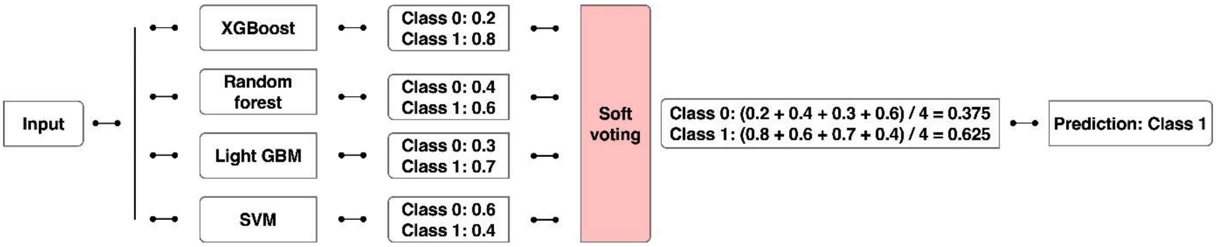


Figure 10. Ensemble model architecture with soft voting.

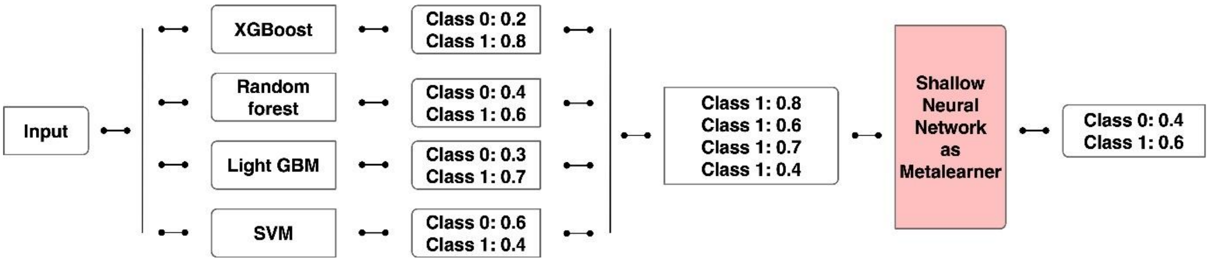
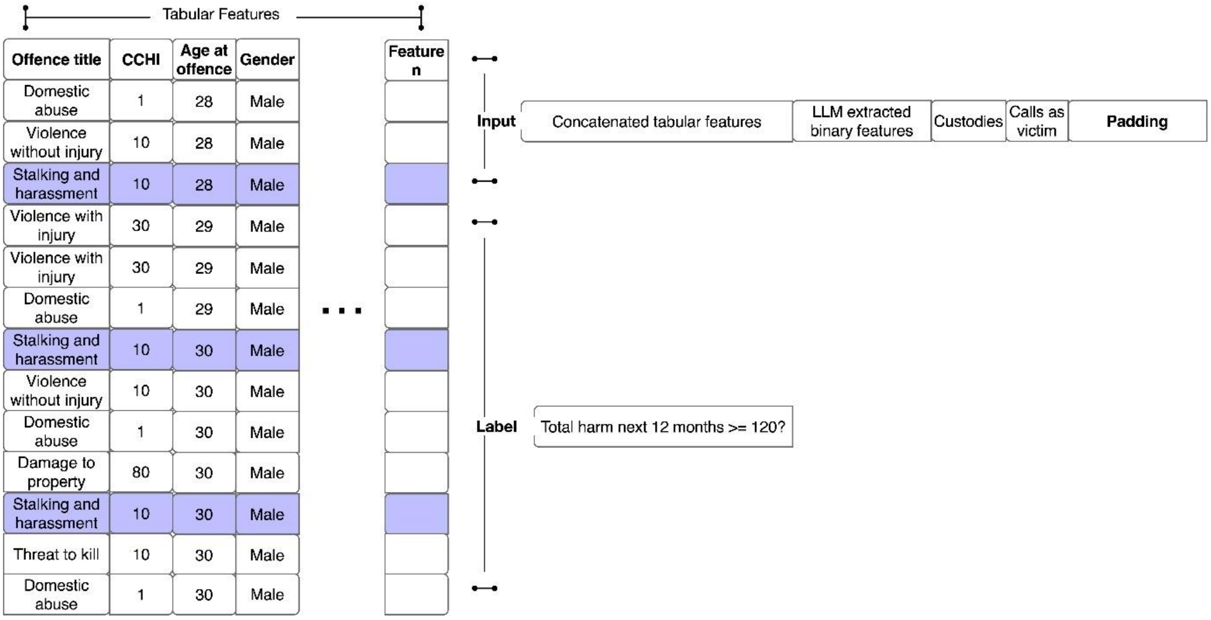


Figure 11. Ensemble model architecture with a shallow neural network meta-learner.

Since tree-based models do not inherently support sequential data, the data had to be reformatted accordingly. As per **Figure 12**, each tabular row, containing features such as harm score, gender, age, and so forth, up to the occurrence of stalking and harassment, was concatenated horizontally into a single extended feature vector. Additional features that remained constant across rows – such as the number of custodial incidents, calls as a victim, and binary features extracted via the LLM – were appended to the end of these concatenated rows. To ensure uniformity in length, each feature vector was padded to match the length of the longest vector found among them (the maximum unpadded length). Padding was done using a value of -9999, as opposed to zeros, to prevent zero values in binary features from being mistakenly treated as missing or ignorable features.


Tabular Features
...
Input
Concatenated tabular features
LLM extracted binary features
Custodies
Calls as victim
Padding
Label
Total harm next 12 months >= 120?

Offence title	CCHI	Age at offence	Gender	Feature n
Domestic abuse	1	28	Male	
Violence without injury	10	28	Male	
Stalking and harassment	10	28	Male	
Violence with injury	30	29	Male	
Violence with injury	30	29	Male	
Domestic abuse	1	29	Male	
Stalking and harassment	10	30	Male	
Violence without injury	10	30	Male	
Domestic abuse	1	30	Male	
Damage to property	80	30	Male	
Stalking and harassment	10	30	Male	
Threat to kill	10	30	Male	
Domestic abuse	1	30	Male	

Figure 12. Input data format for non-sequential modelling.

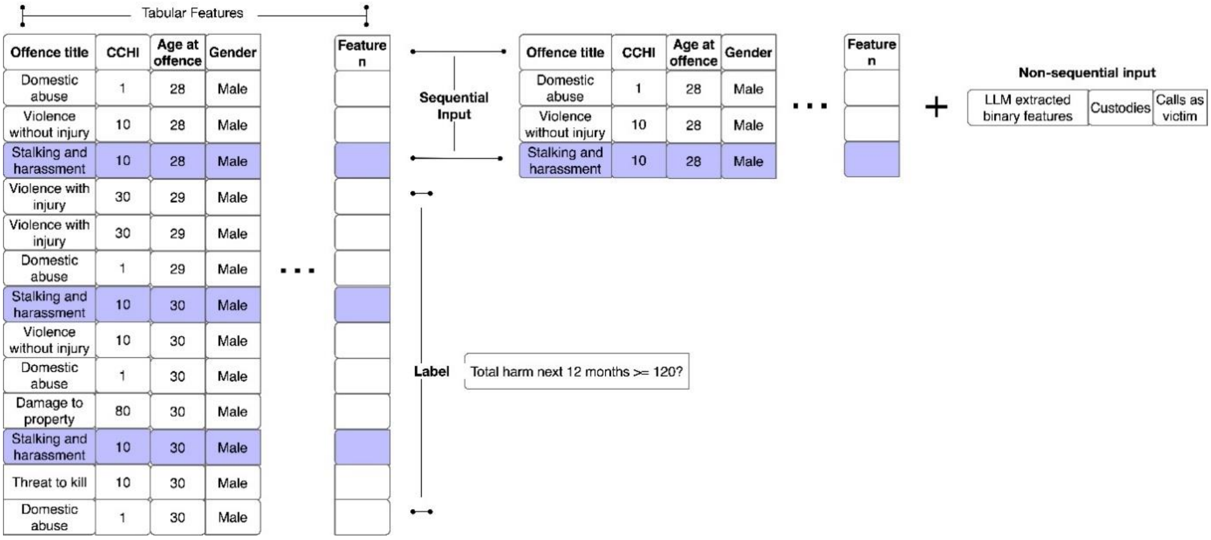
6.3.2 Neural network model with feature attention, multi-head attention, and cross-network integration

Two deep learning models were trained with distinct input formats. Both model architectures are described into detail in **Appendix C**.

1. **Model 1: Single Input Model** This model utilises the same input format as the XGBoost and ensemble models, as illustrated in **Figure 56**, **Appendix C**. The input is a concatenated feature vector that includes all the features associated with each incident, such as harm score, demographics, and any occurrences of offenses like stalking and harassment. These features are combined into a single input vector, allowing the model to process all information at once without distinguishing between sequential and non-sequential data.
2. **Model 2: Dual Input Model** The second model is designed to handle both sequential and non-sequential data separately, as illustrated in **Figure 57**, **Appendix C**. It takes in two inputs:

- **Sequential input**: A two-dimensional input comprising rows of features associated with each offense, including both prior incidents and the specific occurrence of stalking and harassment. This sequential structure allows the model to capture the temporal and contextual patterns within the sequence of offenses, leveraging the temporal relationships between events.
- **Non-sequential input**: A one-dimensional vector formed by concatenating non-sequential features such as the binary features extracted by the LLM, the number of custodial incidents, and the number of phone calls as a victim. This flat vector provides a summary of overall context that is not tied to any particular sequence of offenses.

The dual-input model required inputs to be formatted as illustrated in **Figure 13**. Unlike the single-input model, the tabular features – such as harm scores, derived metrics, gender, age at the time of offense, and binary crime indicators extracted from keywords – were not concatenated into a single long vector. Instead, they were retained in their original tabular format as a 2D array. The 1D input, on the other hand, was a vector created by concatenating features extracted from the LLM (large language model), including the number of custodial incidents and the number of calls recorded as a victim. This dual-format approach preserved the structure of sequential data in its native tabular form while capturing static contextual features in a flat vector.



Offence title	CCHI	Age at offence	Gender	Feature n
Domestic abuse	1	28	Male	
Violence without injury	10	28	Male	
Stalking and harassment	10	28	Male	
Violence with injury	30	29	Male	
Violence with injury	30	29	Male	
Domestic abuse	1	29	Male	
Stalking and harassment	10	30	Male	
Violence without injury	10	30	Male	
Domestic abuse	1	30	Male	
Damage to property	80	30	Male	
Stalking and harassment	10	30	Male	
Threat to kill	10	30	Male	
Domestic abuse	1	30	Male	

Offence title	CCHI	Age at offence	Gender	Feature n
Domestic abuse	1	28	Male	
Violence without injury	10	28	Male	
Stalking and harassment	10	28	Male	

LLM extracted binary features	Custodies	Calls as victim

Figure 13. Input data format for both sequential and non-sequential modelling.

6.3.3 Further considerations

When creating baseline models, the primary goal is to quickly assess whether a model can learn from the data and establish a basic starting point for performance. At this stage, using a single train-test split is often preferable to cross-validation because it is more efficient and less computationally demanding. Cross-validation requires training the model multiple times on different subsets of the data, which can be time-consuming. A single split allows for rapid evaluation and helps identify models that show promise before deeper optimization and more thorough performance assessment.

Additionally, **baseline models are part of an initial exploratory phase focused on understanding the feasibility of different model types, architectures, or feature sets. The goal is to gain an early indication of model performance, rather than achieving precise generalisation metrics.** However, **cross-validation has been incorporated as part of the final model selection process to provide a more robust estimate of performance and variability across data splits**. This ensures that resources are efficiently allocated to models that demonstrate strong potential and allows for confident final evaluation and tuning before deployment.

7 Model Exploration, Evaluation, and Performance Comparison

This section explores models specifically for offender data, with the goal of informing the final model selection approach for both offenders and victims. The insights gained from analysing harm escalation among offenders will help shape the overall strategy for predicting risk and identifying those at high risk within the context of stalking and harassment for both groups. An in-depth description of the chosen model hyperparameters for training the baseline models, the optimisation of these hyperparameters, and evaluation of performance metrics are provided in **Appendix E**.

7.1 XGBoost as a baseline model

Multiple XGBoost baseline models were trained using different combinations of features as outlined in **Table 24, Appendix E1**.

The model which utilises CCHI and derived metrics, male and female flags, age at offence, age < 30, number of calls as victim, number of custodies, day sin/cos transforms: achieves balance of high specificity (0.97) and reasonable recall for the positive class (0.31). Its ROC AUC of 0.73 indicates good discriminative ability. Notably, this model has the highest precision for the positive class (0.60) among the three, suggesting that when it predicts a high-risk individual, it's more likely to be correct. This feature is particularly valuable given the aim to minimise false positives and focus resources efficiently.

The model which utilises CCHI and derived metrics, male and female flags, age at offence, age < 30, number of calls as victim, number of custodies, binary features extracted from incident log keywords: shows a slight improvement in ROC AUC (0.75) and recall for the positive class (0.36). However, this comes at the cost of a small decrease in specificity (0.94) and precision for the positive class (0.51). The improved recall suggests this model is better at identifying potential offenders, capturing a larger proportion of true positives.

The model which utilises all available features: maintains the high ROC AUC (0.74) and recall (0.37) of the second model above, with the same specificity (0.94). Its precision for the positive class (0.50) is comparable to the second model, too. This model's performance suggests that combining all features provides a robust prediction, though the marginal improvement over Model B indicates potential redundancy or noise in some additional features.

These three models stand out from others in the initial evaluation due to their superior balance of key metrics. They consistently demonstrate high specificity (≥0.94), which aligns with minimising false positives (high specificity). Their ROC AUC scores (0.73-0.75) are the highest among all models, indicating better overall discriminative power. Moreover, they achieve this performance while maintaining reasonable recall for the positive class (0.31-0.37), capturing a significant portion of potential serious offenders.

When comparing these top models to others in the initial set, we observe that they strike an optimal balance between specificity and recall. Some other models achieved higher specificity but at the cost of much lower recall, potentially missing too many high-risk individuals. Conversely, models with higher recall often sacrificed too much in terms of specificity, which would lead to an inefficient allocation of monitoring resources.

In terms of recommending the best among these three models, each has its strengths:

1. The first model offers the highest precision for positive predictions, which could be important if the cost of intervention is very high.

2. The second model provides the best overall discriminative ability (highest ROC AUC) and a good balance of precision and recall.

3. The third model captures the most potential offenders (highest recall) while maintaining high specificity.

Given our objective to efficiently allocate limited monitoring resources while identifying as many potential serious offenders as possible, the **second model appears to offer the best overall performance**. Its higher ROC AUC suggests better discrimination between classes, and it provides a good balance between precision and recall for the positive class. The inclusion of binary features from incident log keywords likely contributes meaningful context that enhances prediction accuracy.

7.2 Ensemble of models with soft voting as a baseline

The models in the ensemble are as follows:

1. **XGBoost**:
   - XGBoost was configured with the same hyperparameters as described in **Table 23, Appendix E1**. It uses 200 estimators, a maximum depth of 20, a learning rate of 0.1, and subsamples 80% of the training data. The `scale_pos_weight` was set to the ratio of negative to positive samples, to handle class imbalance effectively. XGBoost's ability to efficiently handle large datasets and imbalanced data makes it a crucial part of the ensemble.
2. **Random Forest**:
   - The Random Forest model was also configured with the same hyperparameters: 200 trees (`n_estimators`) and a maximum depth of 20. The `class_weight='balanced'` parameter was used to ensure that the model gives more weight to the minority class, addressing the class imbalance issue.
3. **LightGBM**:
   - LightGBM was configured similarly to XGBoost, with a few key differences. While it uses the same number of trees (200) and depth (20), the learning rate was set to 0.05 (lower than XGBoost's 0.1), allowing LightGBM to train more gradually and potentially avoid overfitting. The `scale_pos_weight` was used in the same way as XGBoost, to handle class imbalance.
4. **SVM**:
   - The Support Vector Machine (SVM) model was configured with an **RBF kernel**, and the `class_weight='balanced'` parameter was used to handle the class imbalance. SVM was included in the ensemble to capture non-linear relationships in the data that might be missed by tree-based models. The `probability=True` setting ensures that the model outputs probability estimates, which are necessary for soft voting in the ensemble.

As per **Table 25, Appendix E1**, the ensemble approach has generally maintained or slightly improved the ROC AUC scores across most feature combinations. For instance, the best-performing ensemble model achieved a ROC AUC of 0.74, which is comparable to the top individual models. This suggests that the ensemble method has preserved the overall discriminative power of the predictions.

Areas of improvement over XGBoost baseline:

1. **Increased recall:** Many ensemble models show improved recall for the positive class. For example, the model using all features increased recall from 0.37 in the individual model to 0.55 in the ensemble. This significant improvement means the ensemble is better at identifying potential serious offenders, which aligns with our goal of capturing as many true positives as possible.
2. **Varied optimal thresholds:** The ensemble models show different optimal thresholds compared to their individual counterparts, with some increasing and others decreasing or remaining similar. This variation in thresholds suggests that the ensemble method is adjusting its decision boundary differently for each feature set, potentially finding new balance points between sensitivity and specificity.

Trade-offs and potential drawbacks:

1. **Decreased specificity:** The improved recall comes at the cost of lower specificity in most models. For instance, the “All features” model’s specificity dropped from 0.94 to 0.79. This trade-off means an increase in false positives, which could overload monitoring resources.
2. **Lower precision for positive class:** The ensemble models generally show lower precision for the positive class. This decrease means that when the model predicts a positive outcome, it is less likely to be correct compared to the individual models.

Among the ensemble models, three stand out as the best performers:

- Model trained on “CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, binary features extracted from incident log keywords”:
  - Highest ROC AUC (0.74), highest specificity (0.96), and highest precision for Class 1 (0.55)
  - Moderate recall for Class 1 (0.31)

- Model trained on “CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, day/month sin/cos transforms, binary features extracted from incident log keywords”:
  - Equally high ROC AUC (0.74), good specificity (0.91)
  - Better balance between precision (0.40) and recall (0.39) for Class 1

- Model trained on all features:
  - Slightly lower ROC AUC (0.73) and specificity (0.79)
  - Higher recall for Class 1 (0.55) but lower precision (0.29)

Given our dual objectives of minimising false positives while capturing as many true positives as possible, the second model (“CCHI and derived metrics, ... day/month sin/cos transforms, binary features extracted from incident log keywords”) appears to offer the best balance. It maintains high ROC AUC and specificity while providing a good trade-off between precision and recall for Class 1.

7.3 Ensemble of models with a meta-learner as a baseline

The ensemble of models was further extended by incorporating a meta-learner in a stacked architecture. The meta-learner consists of a shallow neural network featuring two hidden dense layers with 32 and 16 neurons, respectively, each utilising rectified linear unit (ReLU) activation functions. The output layer is a dense layer with a single neuron and a sigmoid activation function, designed to produce the final class probability.

According to **Table 26, Appendix E1**, the ensemble model with the meta-learner yielded the lowest performance among the tested models. The ROC AUC scores, ranging from 0.63 to 0.70 across different feature combinations, reflect moderate discriminative ability. Despite consistently high specificity scores, which indicate strong performance in correctly identifying negative instances, the model exhibits lower precision and recall for Class 1 (positive class). This suggests that it struggles to accurately identify and retrieve positive cases.

One particular observation is that the **optimal thresholds** for most feature combinations are low, ranging from **0.01 to 0.07**. This low threshold indicates that the model is highly sensitive to positive predictions, and only minimal probability is required for the model to classify an instance as positive. There are several potential reasons why these optimal thresholds are so low:

1. **Class imbalance:** If there is a significant imbalance between the positive and negative classes in the dataset, the model might favour predicting the negative class. In such cases, a low threshold might be necessary to detect a meaningful number of positive instances. This aligns with the low precision for the positive class (around 0.43 to 0.56), suggesting that the model is more inclined to predict the negative class unless a very low threshold is set.
2. **Trade-off between precision and recall:** The low threshold suggests that the model is optimised for **recall**, especially for the positive class, at the expense of precision. This is common in scenarios where missing positive instances (false negatives) is considered more costly than including some false positives. However, the recall for Class 1 is still relatively low (around 0.29 to 0.33), indicating that even with a low threshold, the model struggles to retrieve positive cases effectively.
3. **Model calibration:** It is also possible that the probabilities output by the base models are not well-calibrated. Ensembling methods, especially those combining different types of models (e.g., XGBoost, Random Forest, SVM), can sometimes result in poorly calibrated predictions. The neural network meta-learner might struggle to correct this, leading to lower probability outputs overall.
4. Similar results were observed when the meta-learner was XGBoost.

7.4 Review and analysis of baseline model outcomes

The dataset exhibits significant class imbalance, with 7,771 negative samples and only 1,218 positive samples. This imbalance presents a challenge for the model in effectively learning patterns related to the minority class (positive cases). During exploratory analysis, no clear linear relationships emerged between the features, further complicating the modelling process. Specifically, we observed inconsistent patterns in crime progression: similar behaviours prior to stalking and harassment led to different outcomes afterwards. For instance, offenders with a history of high-harm crimes unexpectedly showed low harm escalation following stalking and harassment, while those with a seemingly low-risk history escalated to high harm post-incident. Such inconsistencies make it difficult for the model to generalise and learn reliable relationships between pre- and post-stalking behaviours, which are important for predicting harm escalation.

Tree-based models like XGBoost and Random Forest are generally robust to class imbalance due to their inherent splitting mechanism, yet we still applied class weighting to mitigate any potential imbalance-related issues. Despite this, no specific subset of features provided a notable improvement in precision or recall for the positive class, which was confirmed during exploratory analysis. This suggests that the feature space itself may lack the necessary signal to differentiate between positive and negative outcomes.

A potential limitation arises from the restricted variability in the Cambridge Crime Harm Index (CCHI) scores, with only 30 unique values being observed across offences prior to stalking and harassment. This lack of variability in the harm metrics likely hinders the model's ability to leverage CCHI as a predictive feature, as the compressed range of scores may not capture the nuanced escalation in harm. Additionally, the limited discriminative features between offences committed pre- and post-stalking further increases the difficulty. The exploratory analysis revealed no consistent patterns in the feature set that could be reliably used to distinguish between low and high harm cases after stalking incidents.

One potential reason why the ensemble method (and the stacking approach with a meta-learner) failed to improve performance over a standalone XGBoost model is the inherent complexity of the underlying relationships between features. **The patterns governing harm escalation are non-linear and riddled with inconsistencies, which limits the model's ability to learn meaningful relationships**. While ensemble methods like XGBoost are designed to capture non-linear interactions and reduce bias, their effectiveness relies on the availability of coherent patterns in the data, which are sparse in this case.

The ensemble model did slightly improve recall for the positive class at the expense of precision, indicating that it became more sensitive to identifying positive cases but also more prone to false positives. This trade-off is likely due to the nature of the ensemble averaging probabilities across models. **When the individual base models are already struggling with identifying clear patterns, stacking their predictions leads to compounding errors rather than improvement.** The weak signals identified by the base learners, when combined by the meta-learner, resulted in further noise, explaining why the meta-learner approach performed the worst of all models (similar results were observed when the meta-learner was XGBoost).

The shallow neural network meta-learner likely compounded this issue because neural networks, while flexible, tend to require large amounts of data and clearer signal to learn effectively. With inconsistencies in the patterns and limited discriminative features, the neural network may have failed to extract meaningful insights from the base model outputs, leading to degraded performance.

Moreover, the class imbalance combined with inconsistent patterns likely caused the model to struggle in assigning the correct decision boundary, particularly for the positive class. While ensemble methods can sometimes help by leveraging multiple models to cover a broader range of decision boundaries, the lack of clear signals from the features likely limited the meta-learner’s ability to fine-tune this boundary effectively. As a result, the stacked model failed to outperform the base XGBoost model and, in fact, introduced further inaccuracies.

The limited number of unique CCHI scores (30) could also mean that the model is unable to differentiate between nuanced levels of harm. This lack of variability may lead to insufficient granularity in harm prediction, causing the model to generalise too broadly across different harm levels. As a result, the model may not be able to make accurate distinctions between cases of high and low harm escalation, especially in the presence of stalking and harassment. This lack of granularity likely diminishes the value of CCHI as a feature.

Lastly, while it is possible to exhaustively explore all combinations of hyperparameters and features, doing so would have been computationally expensive and time-consuming. Instead, the baseline models with different feature combinations were designed to provide a broad yet efficient exploration of both the feature and hyperparameter space. This approach allowed us to assess the potential for improvement in model performance without the need for exhaustive tuning. **Based on the results from these baselines, it became evident that further exploration of hyperparameters or additional feature combinations was unlikely to yield significant improvements.** The models did not show a strong response to various feature sets or hyperparameter changes, suggesting that **the limitations are more likely tied to the dataset itself – such as inconsistencies in crime patterns and feature interactions – rather than to the specific modelling or tuning choices**. Thus, a more exhaustive search was not undertaken as the explored space provided sufficient insight to conclude that additional complexity or tuning would likely have minimal to no impact on performance.

For most samples, the offending history prior to stalking and harassment is relatively short, providing insufficient detail to capture meaningful patterns of escalation or de-escalation. In contrast, the data used to determine harm escalation outcomes covers a much longer period – 12 months post-stalking and harassment – which often includes a significantly larger number of offences and more diverse behavioural patterns. This discrepancy creates **a fundamental challenge: we are attempting to predict highly complex post-stalking offending behaviour based on a sparse and limited set of pre-stalking offences**. **Most offenders in the dataset have only committed around three to five offences prior to stalking and harassment, whereas the number of offences committed after these events is substantially larger**.

7.5 Further observations

While similar patterns to those seen in the false positive and false negative samples are likely present among the true positives and true negatives, the difference lies in the subtle features that helped the model make the correct classifications in the true cases. These subtle features likely influenced whether a sample fell on one side of the decision boundary versus the other, differentiating true positives and negatives from the false classifications.

The fact that the reasons behind the false positives and false negatives are identifiable suggests that the model is not simply guessing. Instead, it is **attempting to make informed decisions, even in the absence of features known to be highly discriminative in the context of stalking and harassment**. **In many cases, a human evaluator would likely have reached the same conclusions as the model, particularly in the misclassified instances.**

Even when certain features are present – such as the number of custodies, criminal behaviour flags derived from incident log keywords, or LLM-extracted indicators like explicit threats, physical assault, or strangulation – these alone or in combination did not consistently provide the model with enough discriminative power. This was true even when harm scores and derived harm metrics suggested a potential escalation or de-escalation post-stalking and harassment.

A key issue is that the wide range of binary features (presence or absence of certain criminal behaviours) did not compensate for the lack of pre-stalking and harassment offenses or escalation. This pattern is observed in both false positives and false negatives. The presence, absence, or combination of these features, while helpful, did not always provide sufficient discriminative power for the model to accurately predict outcomes. These instances are also challenging for human evaluators, further highlighting the inherent difficulty in making accurate predictions based on these factors.

It is also difficult to pinpoint exactly why the model made certain errors during error analysis because of the complexity involved in how machine learning models, especially ensemble methods like XGBoost, make decisions. One key challenge is the large number of binary predictors representing over 40 crime-related behaviours. These binary features can interact in complex ways that are not immediately interpretable. During training, the model combines these features to create decision boundaries that may involve subtle or non-linear interactions between them. As a result, understanding the exact combination of predictors that influenced a classification decision can be difficult, particularly when the model uses hundreds of decision trees.

Moreover, XGBoost leverages multiple trees to arrive at a final prediction, meaning it’s aggregating results from many small, individual decisions. Each tree might use different combinations of features and thresholds, making it hard to trace back the exact path that led to an incorrect classification. Even though we can analyse feature importance or decision paths in some instances, the complexity of these interactions – along with the significant number of features – makes it challenging to directly attribute an error to any specific combination of predictors. This complexity is compounded in situations where subtle features, like minor crime behaviours, contribute to the model’s decision in ways that are not always immediately intuitive from a human perspective through error analysis.

7.6 Neural network model with feature attention, multi-head attention, and cross-network integration

A neural network model, enhanced with feature attention, multi-head attention, and cross-network layers, was trained using the same input format (concatenated features) as the XGBoost and ensemble baselines. The model’s hyperparameters are detailed in **Table 8**. All features were used.

As shown in **Figure 14**, the model starts to exhibit overfitting after approximately 140 epochs: there is no further improvement in validation accuracy, and the model seems to memorise the training data as indicated by the increasing difference between train and validation accuracies, negatively affecting its generalisation ability. To address this, the same model was retrained for only 140 epochs, and its performance metrics were compared with those of the initial model trained for 500 epochs.

Table 8. Parameters for the extended neural network model with feature attention, multi-head attention, and cross-network integration.

Feature importance layer	L1 regularisation strength of 0.01
Feature-wise attention layer	Sigmoid activation for calculating the attention weights
Three dense layers, each followed by a batch normalisation and dropout layer	• 256, 256, and 64 units respectively • Activation function: ReLU • Dropout rate of 0.3
Projection layer	32 units
Multi-head attention layer	4 heads, each with key embedding dimension of 8
Two dense layers, each followed by a batch normalisation and a dropout layer	• 128 and 64 units respectively • Activation function: ReLU • Dropout rate of 0.3
Classification output layer	1 unit with sigmoid activation function
Optimiser	Adam with learning rate of 0.001
Epochs	500
Batch size	32

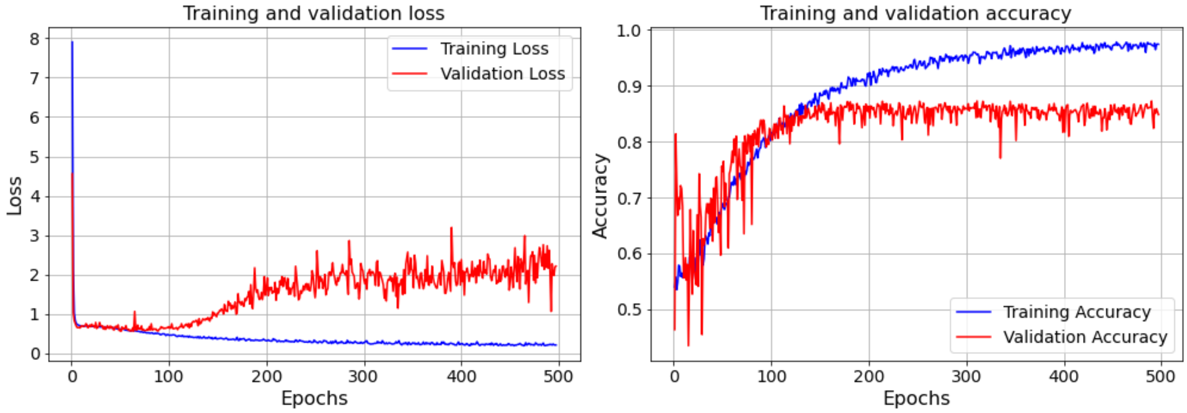


Figure 14. Train and validation loss and accuracy for the extended neural network model with feature attention, multi-head attention, and cross-network integration (500 epochs).

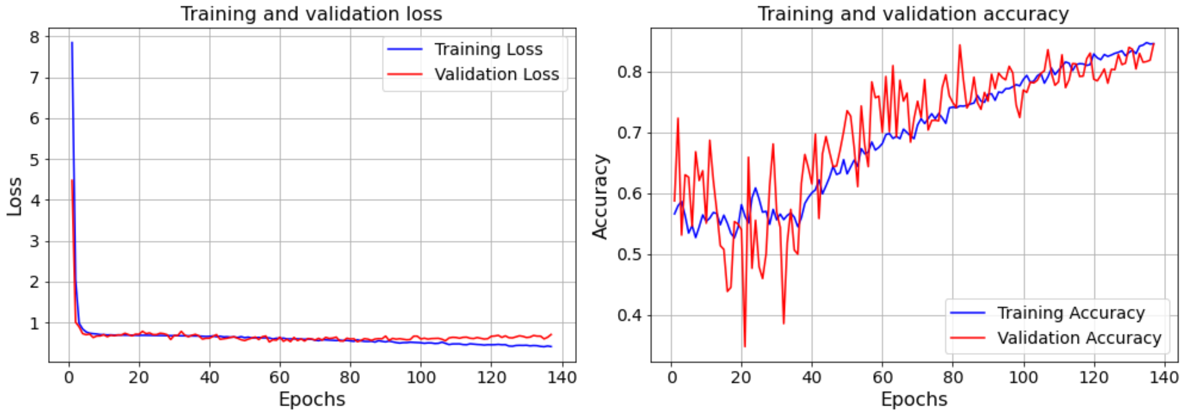


Figure 15. Train and validation loss and accuracy for the extended neural network model with feature attention, multi-head attention, and cross-network integration (140 epochs).

Figure 15 shows no obvious sign of overfitting, as the training and validation accuracies converge closely, with values of 0.85 and 0.84 after 140 epochs, respectively. However, the discrepancy observed earlier – where the training loss is 0.4 compared to a validation loss of 0.84 – may arise from several potential factors:

- The weighted binary cross-entropy amplifies the loss for the minority class (positive samples). If the validation set has a different class distribution than the training set, this could lead to a higher average loss without necessarily affecting accuracy as much.
- Accuracy is a threshold-based metric (typically at 0.5 for binary classification), while cross-entropy loss considers the confidence of predictions (via comparing probabilities). The model might be making correct predictions on the validation set but with less confidence, resulting in higher loss values without significantly impacting accuracy.
- Given the class frequencies in the training set (6204 vs 987), the imbalance is substantial. The weighting scheme might be working as intended, forcing the model to pay more attention to the minority class. This could lead to higher but more balanced loss values across classes.
- The weighting in the loss function (`pos_weight` of 3.64 for the minority class) directly scales the loss values. This scaling does not affect the optimisation process but does affect the absolute values of the loss. The difference in loss between train and validation might partly be an artifact of this scaling, especially if the class distributions differ between sets.
- The model might be well-calibrated for the training data but less so for the validation data. This means it could be making correct predictions (hence similar accuracies) but with probability estimates that are off, leading to higher loss.

Despite the disparity in loss values, the stability and similarity of accuracies between training and validation sets suggest that the model's predictive performance is consistent. This indicates that the higher validation loss might not be critically impacting the model's ability to make correct classifications, which is the primary goal. The similar accuracies between training and validation sets indicate that the model is generalising well (considering that data lacks strong discriminative features) in terms of its decision boundary. The absolute values of the loss are less important than their relative changes and the resulting model performance.

Table 9. Performance metrics for the neural network model with feature attention, multi-head attention, and cross-network integration (500 vs 140 epochs).

Performance metric	Model trained for 500 epochs	Model trained for 140 epochs
Accuracy	0.85	0.84
ROC AUC score	0.68	0.72
Class 0 precision	0.9	0.91
Class 0 recall	0.93	0.91
Class 1 precision	0.43	0.43
Class 1 recall	0.32	0.41
Specificity	0.93	0.91
F1 score	0.37	0.42
Weighted precision	0.83	0.84
Weighted recall	0.84	0.84
Weighted F1 score	0.84	0.84

In the second model, recall for positive samples increased by 10% without any reduction in precision (**Table 9**). The confusion matrices displayed in **Figure 16** (**left:** model trained for 500 epochs; **right:** model trained for 140 epochs) indicate an increase of 23 true positives thus reducing the false negative by the same amount. When compared to the XGBoost and ensemble confusion matrices and performance metrics in **Tables 24**, **25**, and **26** in **Appendix E1**, the model trained for 140 epochs demonstrates a more balanced confusion matrix, suggesting minimal overfitting, as evidenced by the learning curves which align closely in terms of loss and accuracy. Importantly, this is achieved while maintaining a high specificity of 0.91. However, similar to all the previous models, this model is also likely to be biased toward the negative class, a reason that will be discussed at the end of this section.

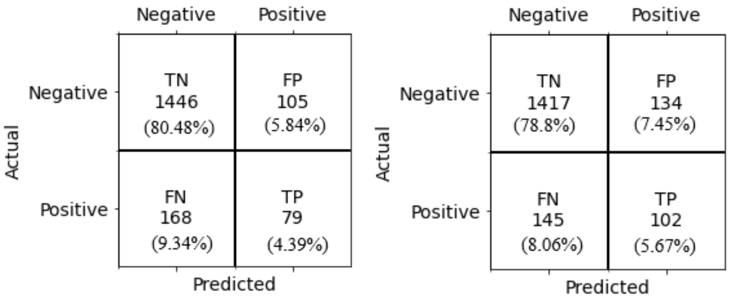


Figure 16. Confusion matrices for the extended neural network model with feature attention, multi-head attention, and cross-network integration (500 epochs vs 140 epochs).

The model is not overfitting in the traditional sense since both train and validation accuracies are similar. However, the model seems to be biased towards the majority class (negative samples). This is a common issue with imbalanced datasets.

The high accuracy (80%) is likely due to the model correctly classifying most of the majority class samples, which dominate the dataset and are also very simplistic in terms of features (e.g., lack in patterns). **To get a better picture of the model's performance, it is more useful to consider metrics that are less sensitive to class imbalance, such as F1 score, precision and recall for each class, and AUC-ROC.**

7.7 Neural network model with feature attention, multi-head attention, and cross-network integration, and sequence inputs

The model trained here shares the same hyperparameters and architecture as the previous model described in **Table 8**. The key distinction is that this model utilises two inputs: one sequential and one non-sequential. For the sequential input, two LSTM layers were employed, as depicted in **Figure 57, Appendix C** and discussed in **Subsection 5.3.2**.

In this model, the sequences are encoded by the LSTM layers and then concatenated with a one-dimensional array of features, including binary features extracted by the LLM, the number of custodial incidents, and the number of calls as a victim. However, this approach led to the model overfitting the training data more rapidly and ultimately performing worse than the non-sequential deep learning model, as shown in **Figure 17**. After 500 epochs, the training loss was significantly lower than the validation loss (0.09 vs. 2.7), with training and validation accuracies of 0.98 and 0.78, respectively. Overfitting became apparent early in the training process, specifically around the 100th epoch. Additionally, both the training and validation loss and accuracy curves exhibited high fluctuations across epochs, particularly in the validation data. This may indicate instability in the learning process or difficulty in capturing consistent temporal patterns, potentially due to the complexity of the sequential input.

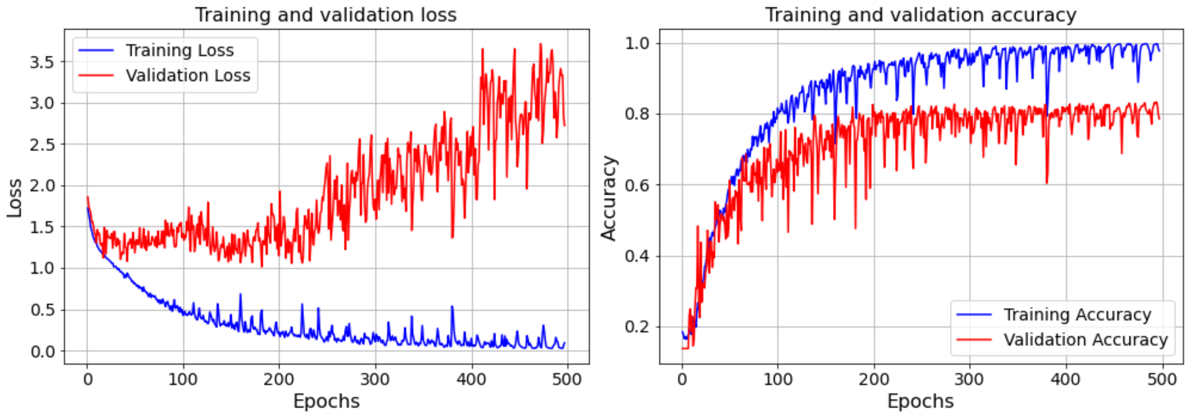


Figure 17. Train and validation loss and accuracy for the extended neural network model with sequence inputs, LSTM, feature attention, multi-head attention, and cross-network integration (500 epochs).

To further explore the model's behaviour, a second training session was conducted, limited to 100 epochs, to assess its key performance metrics (**Figure 18**). The shorter training duration made the overfitting more apparent, highlighting the model's challenges in generalising temporal interdependencies in the validation data. The training loss and accuracy curves were considerably smoother compared to the more erratic and fluctuating validation curves, suggesting that while the model fit the training data well, it struggled to maintain consistency on unseen data.

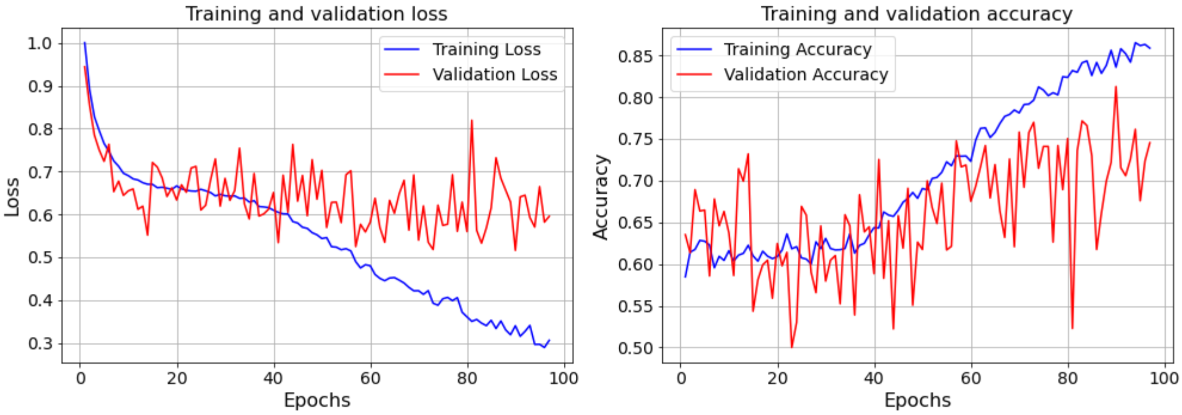


Figure 18. Train and validation loss and accuracy for the extended neural network model with sequence inputs, LSTM, feature attention, multi-head attention, and cross-network integration (100 epochs).

Limiting training to 100 epochs reduced overfitting and slightly improved performance metrics, including ROC AUC, F1 score, and precision and recall for both classes (**Table 10**). Nevertheless, the use of LSTM layers to learn temporal dependencies between sequences of offenses proved to be unstable. This instability might be due to a lack of variability in features from one time step to another. Manual inspection of a subsample of the training data revealed that the features across time steps were often repetitive; for example, similar offenses and features reoccurred across multiple time steps, making it difficult for the model to capture meaningful patterns or learn any new context from the sequences.

Table 10. Performance metrics for the neural network model with sequential inputs, LSTM, feature attention, multi-head attention, and cross-network integration (500 vs 100 epochs).

Performance metric	Model trained for 500 epochs	Model trained for 100 epochs
Accuracy	0.78	0.75
ROC AUC score	0.66	0.7
Class 0 precision	0.89	0.91
Class 0 recall	0.85	0.77
Class 1 precision	0.28	0.28
Class 1 recall	0.37	0.56
Specificity	0.85	0.89
F1 score	0.32	0.38
Weighted precision	0.81	0.83
Weighted recall	0.78	0.75
Weighted F1 score	0.8	0.78

As shown in the confusion matrices in **Figure 19**, limiting the training of the LSTM-based model to no more than 100 epochs (right subfigure) slightly improved generalisation on unseen data for the positive class. This resulted in an increase in true positives but also led to a higher number of false positives. Despite this improvement, the number of false positives remains considerably high, especially when compared to the model trained on non-sequential data (**Figure 16**, **Subsection 6.6**). The non-sequential model was able to maintain both true positives and false positives at more balanced and acceptable levels, indicating better overall performance in distinguishing between classes.

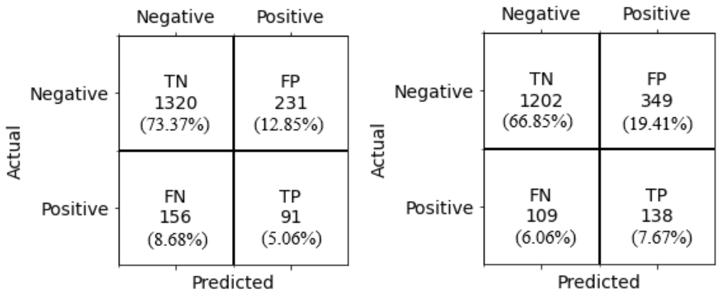


Figure 19. Confusion matrices for the extended neural network model with sequential input, LSTM, feature attention, multi-head attention, and cross-network integration (500 epochs vs 100 epochs).

7.8 Final considerations

This subsection discusses the challenges and limitations associated with modeling harm escalation. It examines key issues such as the sensitivity of metrics to individual events, limitations of momentum and decay calculations, and potential overgeneralisation in model predictions. The section also explores how similar historical incidents can lead to different outcomes, making predictive modeling difficult due to the presence of irregular patterns and unpredictable external factors. Furthermore, it addresses potential problems in the labelling process, which can misclassify outcomes based on rigid timeframes. Challenges related to balancing the dataset, such as the limitations of oversampling techniques and the impact of batch generation, are analysed, revealing how they can affect model performance. Finally, the section concludes that the primary issues lie in the data itself rather than the modeling approaches, with the inconsistencies and lack of clear patterns leading to difficulties in model generalisation and prediction accuracy.

7.8.1 Challenges with features derived from CCHI scores

High sensitivity to individual events: Each event in the CCHI sequence has a significant proportional influence on the calculated metrics due to the low total count. This high sensitivity means that a single atypical event can significantly skew results, potentially producing misleading patterns.

Issues with momentum calculation: Momentum is reliant on changes in the cumulative score over time. However, when working with a small sequence, meaningful trends or behavioural patterns are less likely to emerge. Consequently, the results might be influenced more by random variation rather than systematic changes, reducing the reliability of the momentum metric.

Limitations of decayed score calculation: The decay function emphasises recent scores while diminishing the influence of older ones. In shorter sequences, the distinction between old and recent events becomes blurred, which can undermine the purpose of decay and diminish its ability to provide additional insights.

Overgeneralisation of derived features: Features derived from short sequences may not generalise well to larger, more varied sequences, where meaningful patterns are easier to identify and quantify. As a result, models trained on these derived features might fail to generalise effectively to unseen data.

7.8.2 Contradictory outcomes in similar historical incidents

Differing outcomes despite similar histories: Cases such as domestic abuse (non-crime), domestic abuse, and stalking and harassment offenses may lead to vastly different outcomes for different individuals. For instance, these histories can lead to either an escalation to high-harm outcomes (e.g., grievous bodily harm, murder) or to low-harm outcomes (e.g., another stalking and harassment offense, domestic abuse without injury). This example applies to a range of crime history patterns not just to domestic abuse. Similar historical patterns can lead to different outcomes.

Dataset inspection and implications: Manual inspection of the dataset reveals numerous examples of irregular patterns, suggesting inconsistencies in how similar incidents lead to different outcomes. This unpredictability poses challenges in modeling and risk assessment.

Irregular and unexpected patterns: Several unexpected scenarios are evident:

- **Low pre-harm escalation followed by high harm:** Instances where individuals show very low harm before stalking and harassment, only to escalate to serious offenses in the subsequent 12 months.

- **High pre-harm escalation with no further serious offenses:** Cases where high harm prior to stalking and harassment might suggest a high risk for serious offenses, yet no escalation occurs within the next 12 months.

- **Influence of external factors**: Unmeasurable factors, such as changes in personal circumstances (e.g., finding a job, moving away, quitting harmful behaviour), may play a significant role in altering outcomes.

- **Chance encounters leading to high harm:** Some high-harm incidents arise from unpredictable events, like chance meetings between offender and victim in public. These situations often result in impulsive actions rather than premeditated behaviour, making prediction extremely difficult.

7.8.3 Difficulties with balancing the dataset

Limitations of traditional oversampling techniques: Traditional techniques like SMOTE [23] are not suitable due to a combination of binary and continuous features. While SMOTEN [24] and SMOTE-NC [25] are designed for mixed data types, they are not applicable for sequential data, such as a series of offenses per individual. Additionally, augmenting data by flipping binary features could alter the interpretation of criminal behaviours, which is undesirable.

Balanced batch generator and its implications: A balanced batch generator was employed to create batches with equal numbers of positive and negative samples, ensuring the model sees all negative samples per epoch while repeating positive samples. This approach led to training accuracy increasing disproportionately compared to validation accuracy, causing potential issues such as:

- **Irregular training and validation metrics:** Comparing all negative samples against a limited number of positive samples can result in unstable training and validation losses and accuracies.

- **Model performance concerns:** Overexposure to negative samples may bias the model, affecting its ability to generalise and potentially leading to overfitting or poor performance on unseen data.

7.8.4 Data, not modelling, as the core issue

Error analysis and model performance consistency: Extensive exploratory data analysis and baseline models trained indicate that misclassifications are largely consistent across models. This suggests that the core issue lies in the data itself, not the modeling approach. The inconsistent patterns observed in the dataset contribute to the model's difficulty in learning predictive relationships.

Challenges with inconsistent patterns:

- **Inconsistent harm escalation patterns:** High harm escalation before stalking and harassment does not necessarily continue afterward, and similarly, low harm escalation does not guarantee low harm in subsequent offenses.

- **Unexpected outcomes confuse the model:** While one might expect that serious offenses with high harm prior to stalking and harassment predict similar high-harm offenses in the following 12 months, such straightforward patterns are rare in the dataset. The presence of inconsistent escalation patterns leading to different outcomes can confuse the model.

- **Influence on model performance and regularisation**: The dominance of these unpredictable patterns, combined with the class imbalance, biases the model towards positive classes. This causes the model to predominantly predict obvious high-harm cases while struggling to identify less clear patterns.

Lack of necessity for further modelling exploration: Additional modeling efforts are unlikely to yield significant improvements. Current models have demonstrated no notable increase in precision and recall for the positive class, indicating that ensemble techniques or alternative modeling approaches would not lead to substantial gains. Minor metric improvements typically trade-off between precision and recall, failing to reach meaningful thresholds for enhancement.

Bias toward the negative class: The models exhibit a strong bias towards the negative class, as evidenced by high specificity values across all models. This remains true even with strong regularisation applied to deep learning models and class weighting during training, suggesting that the data lacks discriminative features necessary for the model to learn an effective decision boundary.

Label assignment based on a fixed timeframe: Labels were assigned based on the total harm occurring within 12 months following a stalking and harassment offense. However, if a serious offense occurred just after the 12-month period, the individual would be labeled as negative (i.e., not having committed a serious offense within the timeframe). This cutoff can misclassify cases and reduce the accuracy of the model. Nevertheless, this issue exists no matter what the chosen timeframe is.

Crime behaviours extracted from keywords and incident logs: Binary features indicating specific crime behaviours (e.g., property damage, breach of court order, substance abuse, blackmailing, coercive control, etc.) are recognised in the literature as predictive indicators of serious harm escalation. However, their utility is limited in the current context because they frequently appear in both high-harm cases and negative classes (where no serious harm occurs). This overlap means that, despite their theoretical association with serious harm escalation, the model struggles to learn a clear relationship between these features and high-risk outcomes. Since these behaviours are prevalent even in cases that do not lead to harm escalation, their predictive value is diluted, making it challenging for the model to differentiate between high-harm and low-harm instances based on these features alone.

8 Offender Model

This section focuses on the process of selecting the optimal model for offender prediction, along with the cross-validation and tuning strategies employed to enhance model performance. The selection process involved evaluating multiple models, including XGBoost and neural networks, trained on various feature sets. Cross-validation was used to ensure that the model's performance generalised well to unseen data, mitigating the risk of overfitting. Additionally, model tuning techniques, such as adjusting hyperparameters (e.g., learning rates, tree depth, and regularisation terms), were applied to improve key metrics such as precision, recall, and overall accuracy. The goal was to identify the model that offered the best balance between predictive performance and computational efficiency, ultimately leading to the selection of the XGBoost model trained on selected features without LLM-extracted data as the final model for offender prediction.

8.1 Model selection

The best-performing models in this analysis are the XGBoost model trained on all features and the XGBoost model trained on a subset of features, including CCHI scores and derived metrics, gender flags, age at offense, age under 30, number of custodial sentences, number of victim calls, and binary features extracted from incident log keywords. The neural network with feature attention, multi-head attention, and cross-network integration also performed well, but the XGBoost models showed comparable performance across key metrics.

Table 11. Performance metrics across selected baseline models.

Performance metric	**XGBoost baseline trained on all features**	**Neural network with feature attention, multi-head attention, and cross-network integration trained on all features**	**XGBoost baseline trained on CCHI scores and derived metrics, gender flags, age at offence, age lower than 30, number of custodies, number of calls as victim, and binary features extracted from incident log keywords**
ROC AUC score	0.74	0.72	0.75
Class 0 precision	0.91	0.91	0.9
Class 0 recall	0.94	0.91	0.94
Class 1 precision	0.5	0.43	0.51
Class 1 recall	0.37	0.41	0.36
Specificity	0.94	0.91	0.94
Weighted precision	0.85	0.84	0.85
Weighted recall	0.86	0.84	0.87
Weighted F1 score	0.86	0.84	0.86
TN | FP	1462 | 92	1417 | 134	1468 | 86
----------	-------------	--------------	-------------
FN | TP	153 | 81	145 | 102	156 | 88

As shown in **Table 11**, the XGBoost model trained on all features and the model trained without LLM-generated features demonstrated similar results in terms of ROC AUC score (0.74 vs. 0.75), Class 0 recall (0.94 for both), weighted precision (0.85 vs. 0.87), and weighted F1 score (0.86 for both). The key differentiating factor is that the inclusion of LLM-generated features provided only a marginal increase in performance at a much higher cost in terms of time and resources. Given that the recall for Class 1 (0.43 vs. 0.51) and other important metrics showed only slight variation between these models, the model without LLM-extracted features provides a more efficient choice.

Thus, the XGBoost model trained on the selected features without LLM-generated data is the optimal choice. It provides comparable predictive performance while being more resource-efficient, making it suitable as the final model for both offender and victim modeling tasks. This decision is further supported by the balanced performance across key metrics like specificity (0.94), precision, and recall, with minimal trade-offs compared to models using more complex feature sets.

8.2 Model cross-validation

For cross-validation, the dataset was initially split into two sets: a training set and a test set. Stratification was applied to ensure that the class distribution in both sets was representative of the overall dataset, particularly important given the class imbalance. The test set comprised 20% of the total data, resulting in 1,798 samples, while the remaining 80% (7,191 samples) was used for training. The training set was further split into 5 folds, following a 5-fold cross-validation procedure. This approach ensured that each fold was used as a validation set once while the remaining folds were used for training, leading to the training of 5 models. This method aimed to improve generalisation and ensure that the model's performance was not dependent on a single partition of the data.

Table 12. Offender model cross-validation metrics

Performance metric	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**
ROC AUC score	0.76	0.76	0.76	0.75	0.73
Class 0 precision	0.91	0.91	0.9	0.91	0.91
Class 0 recall	0.94	0.95	0.95	0.89	0.97
Class 1 precision	0.51	0.54	0.58	0.41	0.63
Class 1 recall	0.39	0.39	0.4	0.46	0.36
Specificity	0.94	0.95	0.95	0.89	0.97
Weighted precision	0.86	0.86	0.85	0.84	0.87
Weighted recall	0.87	0.88	0.87	0.83	0.88
Weighted F1 score	0.86	0.87	0.86	0.83	0.87
Optimal threshold	0.26	0.28	0.24	0.1	0.33
TN | FP ---------- FN | TP	1175 | 73 --------------- 116 | 75	1184 | 64 --------------- 115 | 75	1162 | 62 --------------- 128 | 86	1097 | 138 -------------- 109 | 94	1200 | 42 --------------- 125 | 71

The results of the 5-fold cross-validation for the offender model are shown in **Table 12.** The performance metrics across all folds demonstrate consistent results, indicating the robustness of the model. The ROC AUC score ranges from 0.73 to 0.76 across the folds, indicating moderate discriminative performance in separating positive and negative classes. While the scores suggest that the model is reasonably capable of distinguishing between harm escalation and non-escalation cases, there is still room for improvement, particularly in predicting the positive class.

Key Metrics:

- **Class 0 precision** remains stable at 0.91 across all folds, suggesting that the model is consistently effective in predicting the negative class (non-escalation cases).
- **Class 0 recall** is also high, ranging from 0.89 to 0.95, indicating that the model accurately captures the majority of the negative class across all folds.
- **Class 1 precision**, representing the model's ability to correctly predict high harm escalation cases, varies from 0.41 to 0.63 across folds. While this indicates some fluctuation in the model's ability to precisely predict high harm cases, the values are still reasonable given the complexity of the task.
- **Class 1 recall**, which is critical for identifying true positive cases of harm escalation, shows slightly lower values, ranging between 0.36 and 0.46. This suggests that the model may still miss some cases of escalation, but it performs adequately given the inherent imbalance in the dataset.

In terms of specificity, the model performs well across all folds, with values consistently around 0.94 to 0.95. However, this high specificity might be a reflection of the imbalanced nature of the dataset, where the negative class (non-escalation cases) is significantly larger than the positive class (escalation cases), as also noted in the previous section. As a result, the model may be highly effective in identifying non-escalation cases, but this does not necessarily mean that it performs equally well in predicting harm escalation. The relatively strong weighted metrics, including weighted precision, recall, and F1 score (all above 0.83), further indicate that the model is skewed toward the majority class. While these metrics suggest good overall performance, they may mask difficulties in detecting the positive (minority) class.

The optimal decision threshold, which balances precision and recall, varies slightly between the folds, ranging from 0.1 to 0.33. To finalise the model, the average of these thresholds (0.26, 0.28, 0.24, 0.1, and 0.33) was calculated and used as the final threshold. This averaged threshold of 0.24 ensures the best trade-off between identifying true positive cases (high-harm cases) and minimising false positives, providing a balanced approach to harm prediction.

8.3 Model tuning

For hyperparameter tuning, Optuna[8] was utilised. Optuna is a powerful and flexible framework designed to automate hyperparameter optimisation, employing an efficient search strategy known as Bayesian optimisation[9]. This method is well-suited for finding optimal

[8] GitHub - optuna/optuna: A hyperparameter optimization framework

[9] Bayesian Optimization simplified: Master advanced hyperparameter tuning for Machine Learning | by Hanish Paturi | Medium

hyperparameters in machine learning models by leveraging previous search results to guide future iterations more intelligently.

Bayesian optimisation is a sequential model-based optimisation technique that aims to find the best hyperparameters with fewer evaluations compared to traditional methods like grid search or random search. Unlike these methods, Bayesian optimisation builds a probabilistic model of the objective function (usually a Gaussian process[10]) and uses it to decide where to evaluate next, focusing on regions that are likely to yield better performance.

In each iteration, Bayesian optimisation balances exploration (trying hyperparameters in areas of the space that have not been well explored) with exploitation (focusing on areas where the model has already identified promising results). It iteratively refines its search space, learning from past trials to prioritise the most promising hyperparameter combinations, thereby reducing the number of required evaluations.

Advantages of Bayesian optimisation over grid and random search

1. **Efficiency:**

    - Grid search evaluates every possible combination of hyperparameters within a predefined grid, which can be extremely time-consuming, especially when the number of hyperparameters and their possible values is large. This method often requires an exhaustive search over a large space, leading to inefficiencies, as many configurations are evaluated that do not significantly improve performance.

    - Random search is somewhat more efficient than grid search because it samples hyperparameters randomly. However, it still does not leverage information from previous evaluations, meaning that some trials may not be relevant or provide significant improvements.

    - Bayesian optimisation, in contrast, guides the search by focusing on the most promising regions of the hyperparameter space. This significantly reduces the number of evaluations needed to find an optimal solution, making it much more computationally efficient than grid or random search, especially for large search spaces.

2. **Incorporation of prior knowledge:**

    - Bayesian optimisation builds a probabilistic model that uses prior knowledge (information from previous trials) to inform future hyperparameter suggestions. This iterative learning process helps to converge more quickly on the best hyperparameters by focusing on areas that are more likely to yield improvements, rather than randomly searching or exhaustively covering the entire space.

3. **Balance of exploration and exploitation:**

    - One of the key strengths of Bayesian optimisation is its ability to balance the trade-off between exploration (searching new, unexplored hyperparameter values) and exploitation (fine-tuning hyperparameters in promising areas). This balance helps

[10] Gaussian Processes, not quite for dummies (thegradient.pub)

> prevent the optimisation process from getting stuck in local optima and ensures a more thorough search for global optima, which grid and random search methods may miss.

The ranges used for hyperparameter optimisation in the objective function are outlined in **Table 13**. These ranges were tested across 1000 trials during the optimisation process using Optuna. For each trial, the model was fine-tuned by adjusting key parameters such as `n_estimators`, `max_depth`, `learning_rate`, and others, as part of the Bayesian optimisation approach. This process aimed to identify the optimal hyperparameter configuration to maximise model performance, as evaluated through cross-validation.

Table 13. Hyperparameter ranges for Bayesian optimisation trials.

Hyperparameter	Range value
n_estimators	[100, 500]
max_depth	[3, 20]
learning_rate	[0.01, 0.2]
subsample	[0.6, 1]
colsample_bytree	[0.6, 1]
gamma	[0, 0.4]
min_child_weight	[0.5, 4]
scale_pos_weight	[1, 5]

The hyperparameter ranges were selected based on the specific characteristics of this binary classification problem. The `max_depth` parameter ranges between 3-20 to balance between underfitting and overfitting, while learning rates from 0.01-0.2 allow for both fine and coarse gradient updates. Sampling parameters (`subsample` and `colsample_bytree`) range from 0.6-1.0 to test different levels of randomness for preventing overfitting. The `gamma` threshold (0-0.4) and `min_child_weight` (0.5-4) ranges help explore different tree pruning strategies. Given the class imbalance, `scale_pos_weight` is tested between 1-5 to address the skewed distribution. The `n_estimators` range of 100-500 ensures testing both simpler and more complex combinations while maintaining computational feasibility.

Table 14. Offender model optimal hyperparameters.

Hyperparameter	Optimised value
n_estimators	312
max_depth	11
learning_rate	0.0568
subsample	0.874
colsample_bytree	0.724
gamma	0.002
min_child_weight	0.6
scale_pos_weight	4.177

8.4 The final offender model

The final XGBoost offender model was trained using the optimised hyperparameters determined through Bayesian optimisation, as outlined in **Table 14**. The performance metrics for this model are presented in **Table 15**. The train and test splits remained consistent with those used during the hyperparameter optimisation process, ensuring comparability and consistency in model evaluation.

Table 15. The offender final model and its performance metrics.

Performance metric	Value
ROC AUC score	0.74
Class 0 precision	0.91
Class 0 recall	0.95
Class 1 precision	0.54
Class 1 recall	0.39
Specificity	0.95
Accuracy	0.87
Weighted precision	0.86
Weighted recall	0.87
Weighted F1 score	0.86
TN | FP ---------- FN | TP	1472 | 82 --------------- 148 | 96

The performance metrics of the final XGBoost offender model, presented in **Table 15**, reflect similar patterns observed during the model selection and error analysis stages. The ROC AUC score of 0.74 and the high specificity value of 0.95 indicate that the model remains heavily biased towards the negative class (non-escalation cases). This aligns with the previous discussions, where class imbalance posed a challenge. As noted earlier, traditional class balancing techniques have proven ineffective in addressing this issue due to the nature of the data.

The model's Class 0 precision (0.91) and recall (0.95) further demonstrate its strength in predicting non-escalation cases. However, the model's performance for predicting the positive class (escalation cases) remains less optimal, with a Class 1 precision of 0.54 and recall of 0.39. Despite this, the overall weighted metrics – precision, recall, and F1 score – are still reasonably high, indicating a balanced model performance.

Lastly, the final classification threshold was set to 0.24, as identified during cross-validation. This threshold was chosen to balance precision and recall for the positive class, though the model's inherent bias towards the negative class persists, a limitation that, as previously discussed, is challenging to address given the dataset's structure.

9 Victim Model

This section covers the victim model, which is designed to predict the probability of re-victimisation following a stalking and harassment incident. In this context, re-victimisation refers to serious offenses such as murder, sexual assault, physical assault causing grievous bodily harm, and other severe crimes discussed in previous sections. The data processing steps, the filtering conditions to retain victim records, and features used for this model are identical to those employed in the offender model, ensuring consistency in approach. Cross-validation was utilised to validate the model and assess its performance across different data splits. Hyperparameter tuning was conducted using Optuna to identify the optimal set of parameters for the model. Finally, the victim model was trained using these optimal hyperparameters to maximise predictive accuracy.

9.1 Model cross-validation

The victim model was trained on a dataset comprising 5,031 samples, with 4,375 representing the negative class and 656 representing the positive class. This reflects a severe class imbalance, similar to what was encountered in the offender modeling process. Such imbalance can present challenges in model training and performance, as the model may become biased toward the majority class. The test set, which represents 20% of the entire dataset, consists of 1,258 samples – 1,094 belonging to the negative class and 164 to the positive class.

The results of the 5-fold cross-validation for the victim model are presented in **Table 16**. Similar to the offender model, the victim model demonstrates consistent performance across all folds, indicating the stability of the model. The ROC AUC scores range from 0.68 to 0.75, suggesting moderate discriminative capability in differentiating between cases where victims are re-victimised and those where they are not. While the model shows a reasonable ability to distinguish between these classes, there remains room for improvement, particularly in predicting instances of high-harm re-victimisation.

Table 16. Victim model cross-validation metrics

Performance metric	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**
ROC AUC score	0.75	0.73	0.71	0.71	0.68
Class 0 precision	0.92	0.91	0.9	0.92	0.91
Class 0 recall	0.95	0.94	0.91	0.94	0.94
Class 1 precision	0.52	0.51	0.46	0.47	0.45
Class 1 recall	0.39	0.43	0.42	0.38	0.33
Specificity	0.95	0.94	0.91	0.94	0.94
Weighted precision	0.87	0.86	0.83	0.86	0.85
Weighted recall	0.88	0.87	0.84	0.87	0.86
Weighted F1 score	0.87	0.86	0.83	0.87	0.86
Optimal threshold	0.35	0.19	0.14	0.21	0.38
TN | FP ---------- FN | TP	835 | 46 --------------- **77** | 49	815 | 56 --------------- **77** | 58	778 | 76 --------------- 88 | 64	829 | 54 -------------- 76 | 47	828 | 51 --------------- 85 | 42

Key Metrics:

- **Class 0 precision** is stable across the folds, ranging between 0.90 and 0.92, which reflects the model's consistent accuracy in identifying non-victimisation cases.
- **Class 0 recall** is similarly high, ranging from 0.91 to 0.95, meaning that the model successfully captures most cases where victims are not re-victimised.
- **Class 1 precision**, which measures the model's ability to accurately predict re-victimisation cases, varies more significantly, from 0.45 to 0.52 across folds. Although this variation reflects some inconsistency in precisely identifying high-risk victims, the values are still within a reasonable range given the difficulty of the task.
- **Class 1 recall** values, which are critical for identifying true positive cases of re-victimisation, are somewhat lower, ranging from 0.33 to 0.43. This suggests that while the model can detect some instances of re-victimisation, it may still miss a portion of cases, likely due to the inherent class imbalance in the dataset.

The model performs well in terms of **specificity**, with values between 0.91 and 0.95, indicating its strong ability to accurately identify non-victimisation cases. However, as with the offender model, this high specificity may be partially driven by the class imbalance, where non-victimisation cases heavily outweigh re-victimisation cases. Consequently, while the model is effective at identifying the majority class, it struggles more with accurately predicting the minority class. The strong **weighted precision**, **recall**, and **F1 scores** (all above 0.83 across folds) suggest solid overall performance, though these metrics may obscure the model's difficulty in detecting the minority class.

The optimal decision threshold for this model also varies between folds, ranging from 0.14 to 0.38. To ensure a balanced trade-off between precision and recall, the average of these thresholds was calculated, resulting in a final threshold of 0.23. This value helps the model maintain the best balance between correctly identifying re-victimisation cases while minimising false positives, allowing for more reliable harm prediction for victims.

9.2 Model tuning

The hyperparameter ranges used for optimisation in the objective function are identical to those employed in the offender model optimisation, as outlined in **Table 13**. These ranges were evaluated over 1,000 trials during the optimisation process using Optuna. The optimal parameters are shown in **Table 17**.

Table 17. Victim model optimal hyperparameters.

Hyperparameter	Optimised value
n_estimators	211
max_depth	10
learning_rate	0.1583
subsample	0.8216
colsample_bytree	0.6798
gamma	0.321
min_child_weight	1.6674
scale_pos_weight	2.3394

9.3 The victim final model

The final XGBoost victim model was trained using the optimised hyperparameters determined through Bayesian optimisation, as outlined in **Table 17**. The performance metrics for this model are presented in **Table 18**. The train and test are the same as those using during the hyperparameter optimisation process.

Table 18. The victim final model and its performance metrics.

Performance metric	Value
ROC AUC score	0.71
Class 0 precision	0.9
Class 0 recall	0.93
Class 1 precision	0.42
Class 1 recall	0.34
Specificity	0.93
Accuracy	0.85
Weighted precision	0.84
Weighted recall	0.85
Weighted F1 score	0.85
TN | FP	1017 | 77
----------	---------------
FN | TP	108 | 56

The **ROC AUC score** of 0.71 indicates moderate discriminative ability, which aligns with the performance seen in the offender model, suggesting a similar level of effectiveness in distinguishing between serious crime re-victimisation and non-re-victimisation cases following a stalking and harassment incident.

- **Class 0 precision** (0.9) and **Class 0 recall** (0.93) show that the model performs well in predicting non-re-victimisation cases, as it accurately identifies the majority of the negative class.
- However, the model struggles more with identifying re-victimisation cases, as evidenced by the **Class 1 precision** of 0.42 and **Class 1 recall** of 0.34. This indicates that while the model captures some true positives in re-victimisation, it misses a significant number of these cases, likely due to the severe class imbalance present in the dataset.

The **specificity** (0.93) further confirms the model's ability to effectively identify negative cases, though this high specificity, similar to the offender model, may be influenced by the class imbalance. The **accuracy**, **weighted precision**, **weighted recall**, and **weighted F1 score** all sit around 0.85, which reflects overall balanced performance. However, these metrics might mask the challenges the model faces in accurately predicting the positive class due to the imbalanced nature of the data. The confusion matrix also highlights these difficulties, with a larger number of false negatives (108) compared to true positives (56), which could limit the model's effectiveness.

10 Time-to-Event Modelling

This section introduces the application of the Random Forest Survival Model to predict when nominals are likely to escalate to a serious offense following a stalking and harassment incident, or when a victim is likely to be re-victimised with a high-harm crime. The model focuses on a 12-month time frame from the initial stalking and harassment event, providing insights into the timing of potential escalations. Additionally, the model can be used as a prioritisation tool, ranking individuals based on their escalation risk to help prioritise interventions.

A key concept in survival analysis is *censored data*. Censored data occurs when the event of interest (escalation or re-victimisation) has not occurred by the end of the observation period. For instance, if a nominal has not escalated to a serious offense within the 12-month time frame, this individual's data is considered right-censored because we know they have not escalated within the study period, but we do not know if or when the event might occur in the future. This is crucial for survival models, as it allows the model to account for individuals who have not yet experienced the event, rather than discarding or misinterpreting their data.

In survival analysis, the Random Forest Survival Model produces an estimate known as the *survival function*, which gives the probability that the event has not occurred by a certain time. Higher survival function values indicate a lower likelihood of the event occurring within the time frame under consideration. This is useful for prioritisation, as individuals with Higher probability of suffering an incident can be flagged as higher risk, allowing for timely interventions before escalation or re-victimisation occurs. By leveraging censored data and providing time-sensitive risk predictions, this model supports more informed decision-making in managing and prioritising high-risk cases.

The **concordance index (C-index)**[11] is a performance metric used in survival analysis to evaluate the predictive accuracy of time-to-event models. It measures the model's ability to correctly rank pairs of observations based on their risk of experiencing an event. Specifically, the C-index calculates the proportion of all usable (non-tied) pairs in which the model correctly predicts that the individual with the shorter survival time has a higher risk score than the one with a longer survival time.

A **usable pair** is one in which one nominal has experienced the event (e.g., escalation to high-harm crime) at an earlier time than the other nominal. A pair of nominals is considered "usable" if one nominal experienced the event (e.g., a high-harm crime) sooner than the other. This allows the model to show whether it correctly predicted which nominal would have a shorter time to the event. For instance:

- Nominal A accumulated high-harm at 180 days;
- Nominal B accumulated high-harm at 300 days.

These are usable pairs because their event times can be compared. For the C-index to use a pair, the nominals must have distinct event times. If both nominals have the same event time (e.g., both accumulated high-harm at 180 days) or if both were censored at the same time (did not accumulate high-harm within 365 days), they form a **tied pair**. Tied pairs are not use

[11] Evaluating Survival Models – scikit-survival 0.23.0

because there is no meaningful way to rank one nominal above the other. Therefore, these pairs do not contribute to the C-index calculation.

A C-index value ranges from 0.5 to 1.0, where:

- **0.5** indicates no predictive ability (equivalent to random chance).
- **1.0** indicates perfect prediction accuracy.

In the context of the Random Forest Survival model, a higher C-index indicates that the model is more effective at distinguishing between individuals who experience a high-harm event sooner and those who experience it later (or not at all), making it a valuable metric for assessing how well the model captures risk rankings over time.

10.1 Offender survival model

In the context of the Random Forest Survival (RFS) model, the censoring point was set to 365 days. In addition to the features used to train the previous models, the RFS model requires two additional values: a boolean flag indicating whether a high-harm crime occurred within the 365-day period, and the number of days at which this high harm level was reached, if applicable. Although these values play a critical role in the survival analysis, they may be better referred to as *target variables* or *outcome indicators*, as they serve to guide the model in estimating time-to-event outcomes.

The steps for calculating these outcome indicators are as follows:

1. For each nominal's offending history prior to the stalking and harassment incident, calculate the cumulative CCHI score after the stalking and harassment incident.
2. If the cumulative CCHI score reaches or exceeds 120, record the number of days that have elapsed from the incident to this point.
3. Store these values in a tuple of the form (status, time), where *status* is a boolean indicating whether the harm level reached 120 or more, and *time* represents the number of days it took to reach that harm threshold.

These `(status, time)` values allow the RFS model to predict not only the probability of high-harm escalation but also the estimated time frame within which this escalation is likely to occur.

Cross-validation was employed to establish a baseline C-index for the RSF model, with values from each of the five folds as follows: 0.604, 0.626, 0.662, 0.661, and 0.634. These scores indicate consistent performance across the folds, providing a reliable baseline for assessing model improvements. Subsequently, hyperparameter optimisation was conducted to explore whether these baseline results could be enhanced. The ranges for the model's hyperparameters are detailed in **Table 19**. Given the time-intensive nature of RSF optimisation – significantly more demanding than XGBoost optimisation – the optimisation process was limited to 100 trials (which took 18 hours to complete). The feature set used remained consistent with those applied in previous models, ensuring comparability across modeling approaches.

Table 19. Hyperparameter ranges for Bayesian optimisation trials for the offender time-to-event model.

Hyperparameter	**Range value**
n_estimators	[100, 500]
max_depth	[3, 20]
min_sample_split	[2, 20]
min_sample_leaf	[1, 20]
max_features	[0.1, 1]

The optimal hyperparameters identified using Optuna are presented in **Table 20**. Post-optimisation, the model achieved a C-index of 0.6831, representing an improvement over all baseline C-index values obtained with the initial, untuned hyperparameters. This increase indicates that the hyperparameter tuning process effectively enhanced the model's ability to rank individuals based on the risk of the event occurring sooner. The final model trained using the optimised hyperparameters yielded a C-index score of 0.6771 on the test set (N=1,797).

Table 20. Optimal hyperparameter values for the offender time-to-event model after 100 trials.

Hyperparameter	**Range value**
n_estimators	498
max_depth	20
min_sample_split	4
min_sample_leaf	4
max_features	0.7955
Best C-index	0.6832

Figures 20 and **21** illustrate the survival curves for all nominals who either escalated or did not escalate to high harm within 12 months following a stalking and harassment incident. In these plots, lower probability values indicate a higher likelihood that a nominal will commit a high-risk offense against their victim within the 12-month period. A red dashed line marks a threshold probability of 0.5; any survival probability below this line suggests that the nominal is likely to commit a high-risk crime on or before the day indicated on the x-axis.

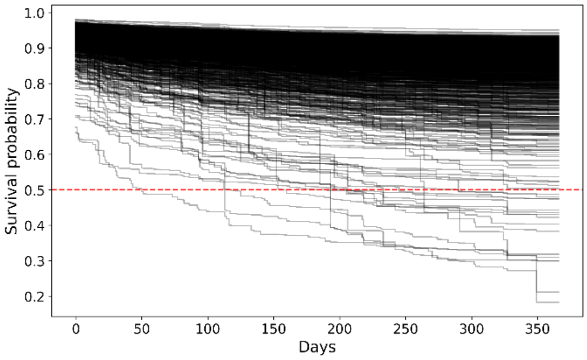


Figure 20. Survival curves for nominals who did not escalate to high risk (total CCHI < 120, N=1,505) in the next 12 months following the occurrence of a stalking and harassment offence.

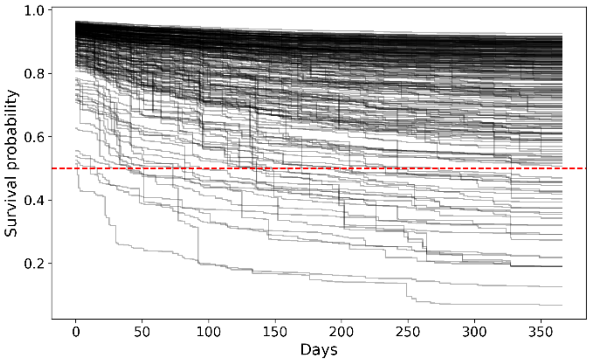


Figure 21. Survival curves for nominals who escalated to high risk (total CCHI >= 120, N=293) in the next 12 months following the occurrence of a stalking and harassment offence.

The model successfully flagged a substantial portion of high-risk nominals as likely to escalate within the next 12 months, while also accurately filtering out many who did not commit high-risk offenses within this period. This demonstrates the model's capacity to distinguish between high- and low-risk individuals effectively.

10.2 Victim survival model

An initial baseline victim survival model was trained using cross-validation with the following hyperparameters: `n_estimators=100`, `max_depth=10`, `min_samples_split=2`, and `min_samples_leaf=2`. This baseline model produced C-index values of 0.63, 0.653, 0.631, 0.629, and 0.631 across the five folds, showing consistent but moderate performance. To improve upon this, Optuna was employed to identify optimal hyperparameters, using the same parameter ranges outlined in **Table 20**. After 100 optimisation trials, the optimal hyperparameter values were determined and are listed in **Table 21**. The final model trained using the optimised hyperparameters yielded a C-index score of 0.6443 on the test set (N=1,258).

Table 21. Optimal hyperparameter values for the victim time-to-event model after 100 trials.

Hyperparameter	Range value
n_estimators	160
max_depth	19
min_sample_split	8
min_sample_leaf	10
max_features	0.9579
Best C-index	0.6717

Similar to the offender time-to-event model, the victim time-to-event model effectively filtered out the majority of individuals unlikely to experience high-risk escalation following a stalking and harassment incident, as seen in **Figures 22** and **23**. Additionally, the model successfully identified a modest proportion of victims at risk of becoming subject to a high-harm offense within the next 12 months. This demonstrates the model's capacity to differentiate between lower- and higher-risk individuals, providing valuable insights for targeted prevention efforts.

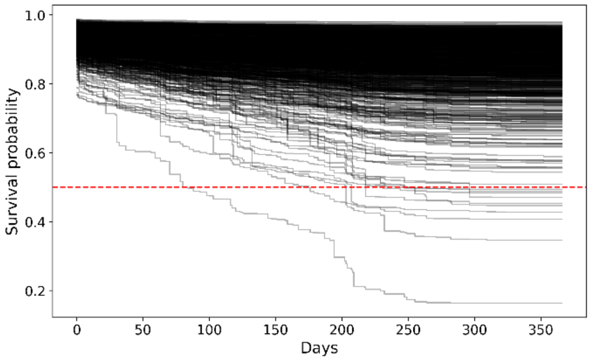


Figure 22. Survival curves for nominals who were not victims of escalation to high risk (total CCHI < 120, N=1,069) in the next 12 months following the occurrence of a stalking and harassment offence.

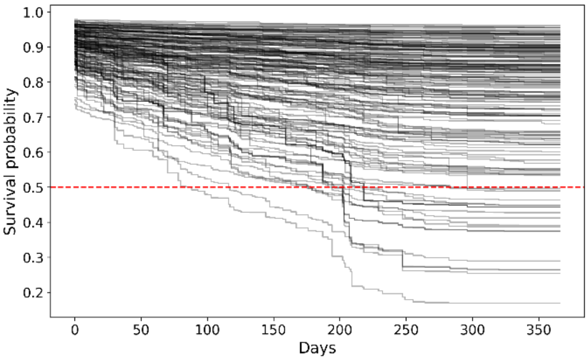


Figure 23. Survival curves for nominals who were victims of escalation to high risk (total CCHI >= 120, N=189) in the next 12 months following the occurrence of a stalking and harassment offence

11 Stalking Tool in Use

The four models shown in **Figure 24** will be deployed as endpoints in the cloud environment, enabling the Stalking and Triage Clinic to submit cases based on their specific needs. These models support both independent and combined decision-making, with potential use cases as follows:

- **Dual Assessment for offenders and victims**: A single offender can be evaluated using both models to obtain a probability of committing a high-risk crime within the next 12 months and an estimated time frame, indicating the number of days until such a crime is likely. The same assessment can be applied to a victim, particularly when there is a known connection to the offender.
- **Complementary insights for case officers**: The combined output of these models provides multiple perspectives to enhance decision-making. For example, if the offender classification model yields a probability below 0.5 (suggesting low risk within the next 12 months) but the time-to-event model flags the offender as high-risk within a certain timeframe, the models collectively address each other's limitations. This layered insight can be valuable, especially considering prior challenges in accurately identifying high-risk individuals.
- **Victim-based risk identification**: Even if the offender models do not flag an individual, known victims of the offender can be evaluated using the victim models. If any victim is flagged as high-risk, the associated offender could also be flagged as potentially high-risk based on the victim's vulnerability.
- **Prioritisation of cases by risk level**: All models enable case officers to rank offenders or victims by the immediacy of risk or probability of a high-harm offence occurring in the next 12 months, or both, allowing for a targeted and prioritised approach to intervention.

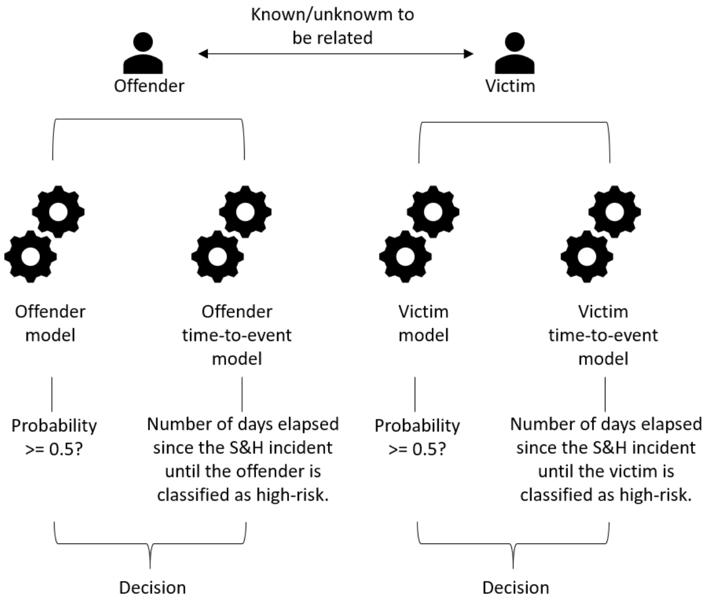

Known/unknowm to be related
Offender
Victim
Offender model
Offender time-to-event model
Victim model
Victim time-to-event model
Probability >= 0.5?
Number of days elapsed since the S&H incident until the offender is classified as high-risk.
Probability >= 0.5?
Number of days elapsed since the S&H incident until the victim is classified as high-risk.
Decision
Decision

Figure 24. Example of model usage.

11.1 Impact of changes in Home Office crime recording rules

The most recent update to the Home Office Counting Rules (HOCR) affecting stalking and harassment offences occurred in **April 2020**[12]. This update mandated that, in any case involving a "course of conduct" between a victim and their former partner, the offence should be recorded as **stalking by default** – unless it was clearly established that only harassment (without elements specific to stalking) was legally applicable. This change aimed to provide a more cautious and victim-centred approach in cases where stalking and harassment overlap.

In **May 2023**, another update to the HOCR introduced a refinement for offences that involve stalking and harassment. This update removed the requirement to record **two separate crimes** when both stalking and harassment elements are present in the same case. Instead, it streamlined the process to require recording just one offence to better align with legal handling, while still capturing the nature of the conduct comprehensively.

Importantly, this May 2023 update **does not override or alter the April 2020 rule**. The 2020 rule still stands: if there's a course of conduct between a victim and a former partner, it should be recorded as stalking unless it is definitively only harassment. The 2023 update simply makes recording these offences administratively simpler but leaves the victim-protective intent of the 2020 rule intact.

As shown in the report's first figure, the recorded number of stalking and harassment offences in the West Midlands has risen sharply. This increase aligns with recent changes to the HOCR, which have broadened the criteria for recording these offences, meaning more offenders and victims are now categorised under stalking and harassment. Since it has been over four years since the last update that directly affected the classification of offenders and victims, such reclassifications are relatively infrequent. Although it's challenging to account for these changes directly within the classification models, they do not impact the model's deployment or use by the Stalking Triage Clinic. These models can be retrained quickly and can be updated frequently as new data is collected. Therefore, any effects of the counting rule changes on model performance or applicability are minimal or non-existent.

11.2 Offender outliers and implications for the overall model performance

The exploratory data analysis highlighted that the dataset contains a limited number of outlier samples, where "outliers" represent cases with exceptionally high or extreme total harm values following stalking and harassment. This imbalance is further complicated by inconsistencies in outcomes for these outliers (and a significant proportion of all other samples). Specifically, some offenders with similar pre-stalking and harassment offending patterns show vastly different outcomes, while others with differing pre-stalking behaviours end up with similar post-stalking harm levels. This variability, as previously noted, poses challenges for the model's ability to effectively optimise its decision boundary, as it creates contradictory patterns in behaviour and outcomes.

[12] count-violence-apr-2018 (1).pdf

In the model error analysis, it was also observed that many of the misclassified samples are cases where human evaluators might also struggle to make accurate classifications. For example, if an offender shows clear signs of harm escalation prior to stalking and harassment, one might expect them to be classified as high-risk and flagged for close monitoring due to the potential for further harm escalation. However, in many cases, no further escalation occurs. Similarly, there are instances where offenders with minimal prior escalation unexpectedly engage in high-harm behaviour post-stalking and harassment.

Such cases lead to situations where both modeling approaches and human evaluators might incorrectly assess offenders as low- or high-risk post-stalking and harassment, based on (misleading/unpredictable) patterns that do not align with expected outcomes.

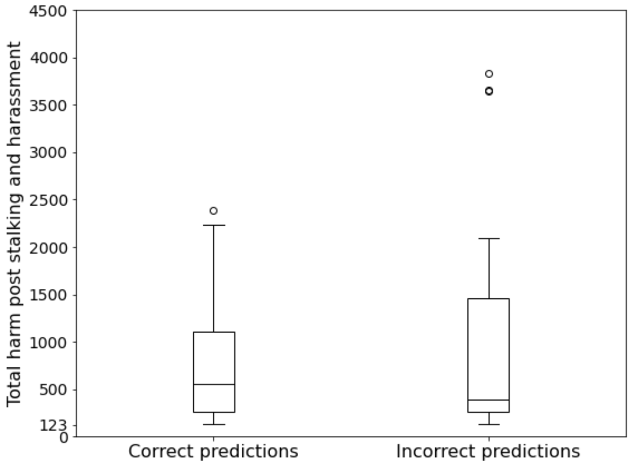


Figure 25. Total harm post-stalking and harassment for both flagged and missed nominals from the test set.

As seen in **Figure 25**, the range of total harm scores post-stalking and harassment for correctly flagged high-risk nominals falls between 130 and 2400, with more than 50% of these cases scoring above 550. This suggests that the model is effective in capturing higher-harm offenders within this group. In contrast, the total harm scores for incorrectly flagged high-risk nominals range from 130 to 3800, with more than 50% of these cases starting from a harm score of 386. **This lower median threshold in the incorrect predictions indicates that many of these cases involve less severe harm, making the model's correct identification of high-risk nominals comparatively stronger**. As previously discussed in the model error analysis, it is important to note that many of these incorrect predictions involve cases where harm escalation is ambiguous or misleading – cases likely to be challenging for a human evaluator to classify accurately as well.

By proactively addressing and mitigating risks, the envisaged use of the model contributes to preventing incidents that could result in substantial burdens on public services and most importantly harm created within the West Midlands. Therefore, even with its limitations, the model could play an important role in enhancing public safety and reducing harm.

12 Unsuccessful Approaches and Further Recommendations

This section discusses several approaches that were initially applied to different parts of the dataset (e.g., numerical data and free text) to explore the extent to which patterns could be identified and learned. These methods were tested during the initial investigation phase to assess their predictive potential. Following this, the section addresses a number of analytical challenges that emerged, highlighting issues that are likely to impact future efforts in modeling events such as the occurrence of serious crimes following stalking and harassment.

12.1 GPT-2 model

A GPT-2[13] model was trained from scratch to determine whether the occurrence of crimes in a sequence could be predicted or if there were any discernible patterns. GPT-2, being an autoregressive model, is well-suited for sequential tasks, making it a logical choice for understanding if future crimes could be predicted based on the history of prior offenses. For example, given the crime sequence `[domestic violence, domestic violence, physical assault, domestic violence, stalking and harassment, physical assault, damage to property]`, the model was provided with all the preceding crimes to learn sequential patterns and predict that the next crime would be damage to property.

However, the model's performance did not surpass that of a random predictor, suggesting that there were no specific patterns in the order or types of crimes that lead to stalking and harassment. Additionally, the model failed to accurately predict the type of crime likely to occur after stalking and harassment. Instead, it disproportionately predicted domestic non-crime offenses, the most common type of crime in the dataset. This reliance on the most frequent crime type likely hindered the autoregressive model's ability to predict less common crimes, leading to its suboptimal performance.

12.2 DistilBert Masked Language Model

A DistilBERT[14] model was tested using masked language modelling (MLM) as a natural language processing and classification approach. Each row in the dataset was transformed into artificial sentences that were then compiled into paragraphs, one per nominal, to describe their past offenses and future criminal behaviour following a stalking and harassment incident. For example, a generated paragraph might read:

```
"Prior to the reported incident, the {gender} offender committed the following offences: {prior offences}. These offences have a cumulative Cambridge Crime Harm Index Score (CCHI) of {total_prior_cchi}. The {gender} offender then committed the following offences in the next 12 months: {future_offences}. These offences have a cumulative Cambridge Crime Harm Index Score (CCHI) of {total_cchi}."
```

The DistilBERT model was first fine-tuned for domain adaptation and later had a classification head added and trained to predict whether a serious offense would occur within the 12 months following the stalking and harassment event. Despite this approach, the model's performance

[13] OpenAI GPT2 (huggingface.co)

[14] DistilBERT (huggingface.co)

did not exceed that of a random classifier, indicating that this method, too, failed to capture predictive patterns related to high-harm offenses.

12.3 Retrieval Augmented Generation with LLM evaluation

The Retrieval Augmented Generation (RAG)[15] model was also tested using several large language models (LLMs) to enhance prediction. The approach began by utilising a pre-trained embedding model called BGE-M3[16] [17] to embed incident logs. BGE-M3, a model designed for semantic understanding, was used to create embeddings of the crime logs. These embeddings were then stored in a vector store, such as the Llama Index[18] or Hierarchical Navigable Small World (HNSW)[19], allowing for fast retrieval of nearest neighbours using similarity metrics like cosine similarity.

For each nominal, all prior incident logs and crime details preceding the stalking and harassment event were gathered. The closest historical incidents, both semantically and syntactically, were retrieved from the vector store. The LLM was then provided with a specially designed prompt that instructed it to assess the likelihood of the individual committing a serious offense in the next 12 months, based on the criminal histories of similar offenders.

While the model produced assessments and justifications that were aligned with manual evaluations and met our expectations as Data Scientists, the outcomes – like those of the previous models – did not match what occurred within the subsequent 12 months. **This may suggest that even with enhanced contextual understanding and comparison with similar offenders, predicting high-risk individuals remains a challenge, possibly implying that even informed human judgment often cannot accurately identify high-risk offenders**.

12.4 Autonecoders

The features used to train both the offender and victim models, as described in their respective sections, were also employed to train an autoencoder. One version of the autoencoder included attention mechanism layers, while another incorporated two optimisation functions – one to minimise the loss for the Cambridge Crime Harm Index (CCHI) score, and the other for binary classification loss. This dual-objective setup allowed the autoencoder to force the bottleneck layer to compress the features into a smaller representation influenced by both the class label and the associated CCHI score.

Once the autoencoder was trained, the decoder was discarded, and the encoder was used to generate embeddings that represented the offending history of each nominal. These embeddings were then utilised in a binary classifier. Despite the autoencoder effectively learning meaningful embeddings, as evidenced by the low mean squared error (MSE) and mean

[15] What is RAG? - Retrieval-Augmented Generation AI Explained - AWS (amazon.com)

[16] BAAI/bge-m3 · Hugging Face

[17] [2402.03216] BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation (arxiv.org)

[18] Vector Store Index - LlamaIndex

[19] Hierarchical Navigable Small Worlds (HNSW) | Pinecone

absolute error (MAE), the binary classifier trained on these embeddings underperformed compared to the baseline models discussed in previous sections.

12.5 Further recommendations

As already discussed in detail in **Subsection 6.8**, the sensitivity of the Cambridge Crime Harm Index (CCHI) to individual events, momentum calculations, and decayed score limitations present significant challenges when working with short crime sequences. A single atypical event can disproportionately skew results, potentially creating misleading patterns. Momentum, which relies on changes in cumulative scores over time, may be influenced more by random variations than by systematic trends when sequences are short. Similarly, the decay function, which emphasises recent events, becomes less effective as the distinction between old and recent events blurs in smaller datasets. Overgeneralisation of derived features is also problematic, as features extracted from short sequences may not translate well to larger datasets, leading to models that fail to generalise effectively to unseen data.

Additionally, the dataset shows irregular patterns, such as similar histories leading to vastly different outcomes – some resulting in high-harm offenses and others in low-harm offenses. External factors like personal circumstances (e.g., job changes, moving away) about which WMP have no data may further complicate predictions. These inconsistencies make it difficult for models to detect reliable patterns, and traditional oversampling techniques such as SMOTE, SOMEN, and SMOTE-NC are not suitable for this kind of sequential data. Balanced batch generation introduced its own set of issues, such as irregular training and validation metrics and model overfitting. Error analysis indicates that the core problem lies in the data itself, as models tend to misclassify the similar cases. The models are also biased toward the negative class due to the class imbalance and lack of strong distinguishing features, and further modelling exploration is unlikely to yield significant improvements. Additionally, binary features extracted from incident logs, while theoretically linked to high-harm escalation, frequently appear in negative cases as well, making it difficult for the model to learn a meaningful association with high-risk outcomes.

The problem statement may need re-framing to address broader criminal behaviours beyond just stalking and harassment. Currently, the focus on stalking and harassment suggests that crimes occurring before and after these incidents are closely linked, which might not always be the case. This narrow focus excludes many relevant criminal behaviours from the dataset, potentially limiting the model’s ability to learn from patterns that may generalise to stalking and harassment-related offenses. Expanding the scope to include other crime behaviours could provide more comprehensive insights, but this introduces a challenge: determining when to cut off the data for training versus when to generate labels.

Currently, for each nominal, all offenses up to and including stalking and harassment are used to train the model, while offenses occurring afterward are used to generate binary labels, which indicate whether the cumulative Cambridge Crime Harm Index (CCHI) score exceeds 120 within the next 12 months. One alternative is to base the cut-off on when the first serious offense (CCHI $\geq$ 120) occurs. However, this presents its own issues, as those who never committed a serious offense within the 12-month window would lack a clear cut-off point, making labelling impossible for these cases.

13 Appendix

A Exploratory Data Analysis

A.1 Total harm pre- and post-stalking and harassment

This subsection examines the relationship between harm levels before and after stalking and harassment incidents, analysing the severity of harm from several perspectives.

One angle centres on cases with minor changes in harm levels (a difference of 30 or less), where overall harm scores remain below 180. This reflects less severe offenses both before and after the incidents. The 30-point difference was chosen because it aligns with a penalty of roughly 30 days in prison, capturing minor but still notable offenses. This threshold serves as a practical benchmark for identifying these cases.

Another perspective considers cases with similar minor harm differences (30 or less), yet where both pre- and post-stalking harm levels exceed 180. These cases represent consistently high levels of harm, indicating severe offenses that remain above a critical threshold throughout.

The analysis then shifts to instances where harm escalates significantly after stalking, moving from below 180 to above the threshold. This rise signals a substantial increase in offense severity following the stalking and harassment incidents.

Lastly, the discussion addresses cases with a marked reduction in harm after stalking, where harm levels decrease from above 180 to below the threshold. This suggests a notable decrease in offense severity following the stalking and harassment incidents.

A.1.1 Difference between the total harm up to and after stalking and harassment is 30 or less, with both the pre- and post-stalking and harassment harm levels remaining below 180

The discussion below focuses on instances where the difference between the total harm up to and after stalking and harassment is 30 or less, with both the pre- and post-stalking and harassment harm levels remaining below 180. This category represents cases with relatively low harm changes and overall harm scores below the critical threshold, indicating less severe offences before and after the stalking and harassment incidents.

There are 4298 offending sequences where the difference between the total harm up to and after stalking and harassment is 30, and the total harm up to and after stalking and harassment

is less than 180. Notably, 31.43% of these sequences involve individuals with only one offence prior to the stalking and harassment incident, highlighting a significant portion of offenders who have already engaged in multiple offences before escalating to stalking and harassment.

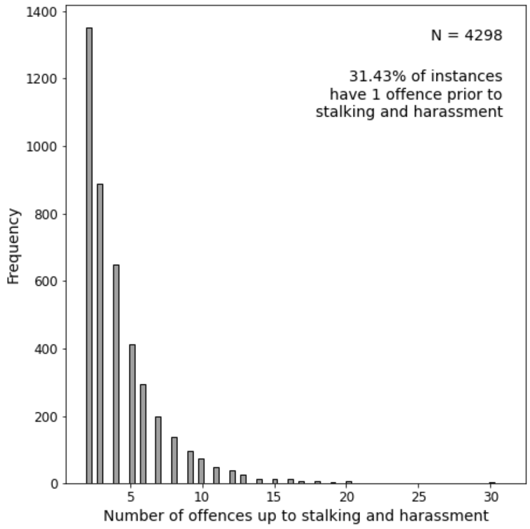


Figure 26. Number of offences leading up to stalking and harassment where the difference between the total harm up to and after stalking and harassment is lower/equal than 30, and the total harm up to and after stalking and harassment is less than 180.

This low frequency of prior offences poses a challenge in precisely quantifying harm escalation, as limited offence histories provide insufficient data to clearly trace patterns of increasing harm. For the predictive model, this raises concerns about accurately identifying individuals who may commit high-harm crimes (harm > 180) after a stalking and harassment offence, given the difficulty in measuring harm escalation in those with few previous offences. It underscores the importance of focusing not only on the number of prior offences but also on understanding the nature and context of these offences to better assess the risk of future high-harm crimes.

Further, it highlights the need for the model to be sensitive to subtle patterns of escalation, even among offenders with low offence counts, to improve its predictive accuracy in identifying high-risk individuals following stalking and harassment incidents.

A.1.2 Difference between the total harm before and after stalking and harassment is 30 or less, with both the pre- and post-stalking and harassment harm levels exceeding 180

The analysis below concentrates on instances where the difference between the total harm before and after stalking and harassment is 30 or less, with both the pre- and post-stalking and harassment harm levels exceeding 180. This category represents cases with minimal harm changes but with overall harm scores above the critical threshold, highlighting more serious offences that maintain high harm levels before and after the stalking and harassment incidents.

There are 40 offending sequences where the difference between the total harm before and after stalking and harassment is 30 or less, and the total harm before and after stalking and harassment is greater than 180. These are offenders who consistently operate above the harm threshold and pose significant risks.

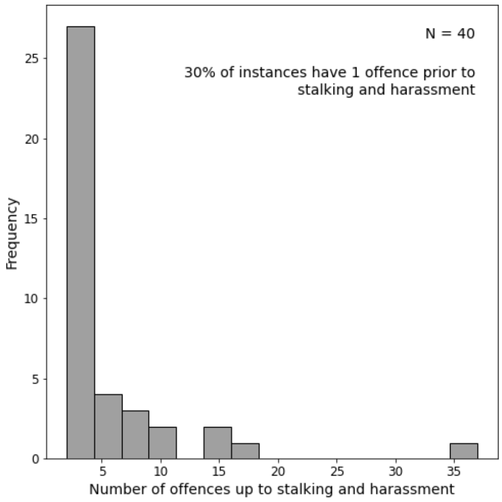


Figure 27. Number of offences leading up to stalking and harassment where the difference between the total harm up to and after stalking and harassment lower/equal than 30, and the total harm up to and after stalking and harassment is higher than 180.

The histogram in **Figure 27** shows that most offenders in this category have few offences, with a significant spike at around two prior offences. Despite the low count, these offences contribute to consistently high harm scores, indicating that harm escalation may not always be marked by an increase in offence frequency but rather by the severity of individual offences.

For the predictive model, this suggests a need to factor in not just the number of prior offences but also the nature and severity of those offences. The challenge lies in accurately predicting future high-harm offences (>=180) based on limited but impactful offence histories. Therefore, the model should be sensitive to cases where minimal changes in offence patterns still pose a significant risk, highlighting the importance of detecting persistent high harm even in the absence of clear harm escalation.

A.1.3 Difference between the total harm before and after stalking and harassment is greater than 30, with the pre-stalking and harassment harm level below 180 and the post-stalking and harassment harm level exceeding 180

The subsequent discussion addresses instances where the difference between the total harm before and after stalking and harassment is greater than 30, with the pre-stalking and harassment harm level below 180 and the post-stalking and harassment harm level exceeding 180. This category captures cases where harm levels significantly increase after the stalking and harassment offence, crossing the critical threshold of 180, indicating a substantial escalation in the severity of offences following the initial stalking and harassment incidents.

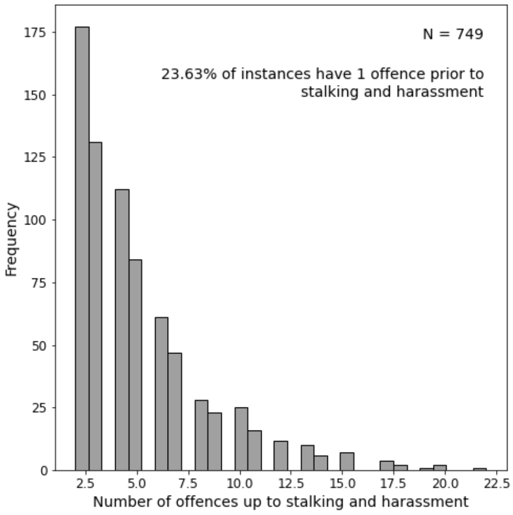


Figure 28. Number of offences leading up to stalking and harassment where the difference between the total harm up to and after stalking and harassment is higher than 30, and the total harm up to and after stalking and harassment is higher than 180.

There are 749 offending sequences where the difference between the total harm before and after stalking and harassment is greater than 30, with the total harm before stalking and harassment less than 180 and total harm after stalking and harassment greater than 180. The histogram in **Figure 28** shows that 23.63% of these instances involve offenders with just one offence prior to stalking and harassment, illustrating that even minimal prior offending can lead to significant harm escalation.

The histogram also highlights that while a significant portion of offenders had only one offence prior to the stalking and harassment incident, the escalation in harm is pronounced, pushing the overall harm score above the critical threshold of 180. This suggests that even offenders with a low count of prior offences can still pose a high risk of escalating to severe harm levels after stalking and harassment.

For the predictive model, this emphasises the importance of recognising harm escalation patterns even in offenders with minimal prior offending. The model should be designed to detect not just the quantity of prior offences but also how these offences contribute to potential high-harm outcomes.

A.1.4 Difference between the total harm before and after stalking and harassment is greater than 30, with the pre-stalking and harassment harm level exceeding 180 and the post-stalking and harassment harm level falling below 180

The following discussion centers on instances where the difference between the total harm before and after stalking and harassment is greater than 30, with the pre-stalking and harassment harm level exceeding 180 and the post-stalking and harassment harm level falling below 180. These are cases where there is a significant reduction in harm following the stalking and harassment offence, moving from above the critical threshold to below it, indicating a notable decrease in the severity of offences after the stalking and harassment incidents.

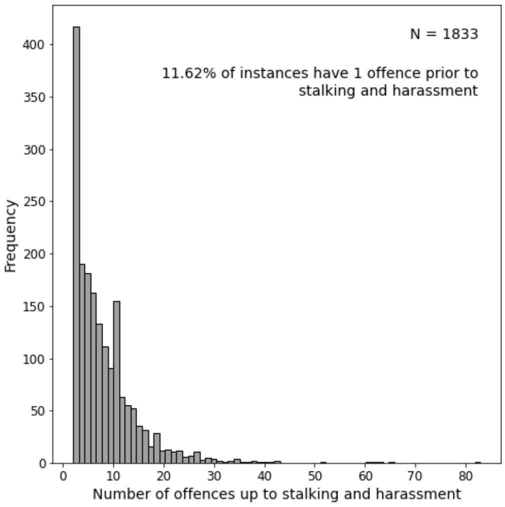


Figure 29. Number of offences leading up to stalking and harassment where the difference between the total harm up to and after stalking and harassment is higher than 30, and the total harm up to and after stalking and harassment is lower than 180.

There are 1,833 offending sequences where the difference between the total harm before and after stalking and harassment is greater than 30, with the total harm before stalking and harassment greater than 180 and the total harm after stalking and harassment less than 180. The histogram in **Figure 29** reveals that 11.62% of these instances involve offenders with only one offence prior to stalking and harassment, suggesting that harm reduction can occur even among those with a minimal offence history.

The histogram also confirms that this group has the highest number of offences prior to stalking and harassment compared to other cases, indicating that offenders with numerous prior offences are more prevalent here. This pattern suggests that even with a substantial offence history, offenders can still exhibit a considerable reduction in harm, complicating the model's task of predicting future high-harm offences based solely on past behaviour.

For the predictive model, this highlights the complexity of assessing future risk, as initial high-harm behaviour does not necessarily predict continued escalation. The model needs to account for scenarios where harm decreases significantly, emphasizing the importance of understanding not just the offence count but the broader context and trajectory of an individual's offending patterns.

Furthermore, these instances could pose the greatest challenge for the predictive model to learn, as it involves offenders who transition from high harm before stalking and harassment to significantly lower harm levels afterward (the next lower score below 180 is 120). The model must differentiate between those likely to continue on a path of harm reduction and those who might not, which is complex given the high number of instances (1,833) that fall into this category.

A.2 Relationship between harm and other predictor features

A total of 8,989 crime sequences across all nominals involve stalking and harassment that was both preceded and followed by at least one other offence committed by a nominal who played the role of an offender in the crime. If a nominal committed two or more stalking and harassment offences, each offence is treated as a separate instance, resulting in multiple entries for the same individual, such as nominalA_1, nominalA_2, and so on (as described in Subsection 3.2).

A.2.1 Distribution of harm scores

Figure 30 presents a comparative analysis of harm scores across three distinct contexts: **(1)** all harm scores, **(2)** total harm scores up to and including the first stalking and harassment offence, and **(3)** total harm scores after the first stalking and harassment offence. In the first boxplot, representing all harm scores (N=74,099), the mean harm score is 45.82 with a median of 2, suggesting a heavily right-skewed distribution where the majority of harm scores are low, with a significant proportion (46%) falling between the first quartile (Q1 = 1) and the median, and a notable 31.8% above the third quartile (Q3 = 10).

The second boxplot, which illustrates total harm scores per nominal up to and including the first stalking and harassment offence (N=8,989), shows a much higher mean harm score of 258.78 and a median of 31. This substantial increase in mean and median harm scores compared to all harm scores indicates that the offences preceding and including stalking and harassment are generally associated with higher harm levels. The interquartile spread is also broader (Q1 = 13, Q3 = 184), highlighting a greater variability in harm, with 25% of scores exceeding the third quartile threshold.

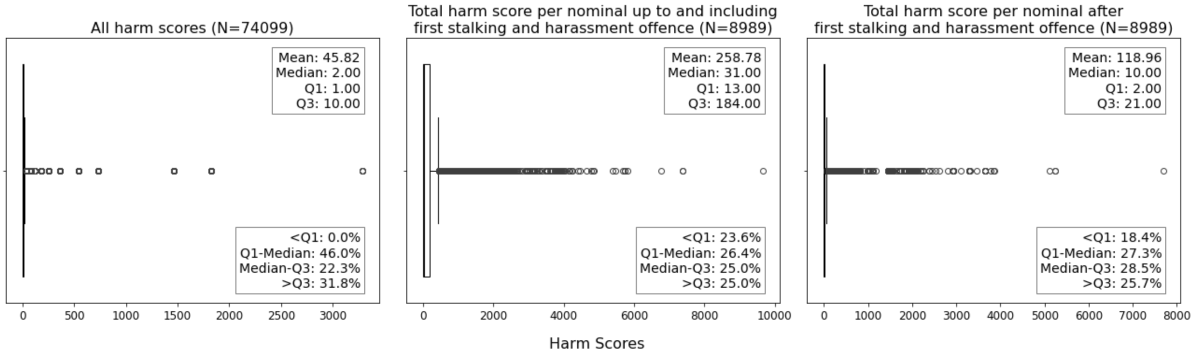


Figure 30. Distribution of harm scores across all crimes, crimes up to stalking and harassment, and crimes after stalking and harassment.

The third boxplot, which depicts total harm scores per nominal after the first stalking and harassment offence (N=8,989), reveals a notable reduction in harm scores, with a mean of 118.96 and a median of 10. This decline suggests a potential de-escalation in offending severity post-stalking and harassment. The interquartile range narrows (Q1 = 2, Q3 = 21), and the distribution indicates that a smaller proportion of harm scores are above the third quartile (25.7%), compared to the pre-stalking and harassment phase. These observations suggest that while stalking and harassment offences are associated with an escalation in harm scores, there is often a subsequent reduction in offending severity following these incidents, which may reflect the impact of interventions or a natural de-escalation in criminal behaviour. This pattern highlights the complexity of predicting future high-harm offences based on past behaviour and underscores the importance of understanding the dynamics of harm escalation and reduction in offenders with stalking and harassment offences.

A.2.2 Distribution of log transformed harm scores

Figure 31 presents log-transformed harm scores for three distinct contexts: all harm scores, total harm scores up to and including the first stalking and harassment offence, and total harm scores after the first stalking and harassment offence. This transformation is applied to the same data as the previous figure, offering a clearer view of the distribution by mitigating the effect of extreme values.

In the first boxplot, representing all harm scores (N=74,099), the log transformation reveals a more compact distribution with a median of 1 (log-transformed value of 2) and a third quartile (Q3) of 2.4 (log-transformed value of 10). Nearly half of the values fall between Q1 (0.69) and the median, indicating a concentration of lower harm scores and a substantial portion (31.8%) exceeding the third quartile.

The second boxplot, showing total harm scores up to and including the first stalking and harassment offence (N=8,989), demonstrates a broader distribution post-log transformation, with a median of 3 (log-transformed value of 19) and Q3 of 5.22 (log-transformed value of 184). This spread underscores that offences associated with stalking and harassment exhibit greater variability in harm scores, with a balanced distribution across the quartiles and 25% of cases surpassing the third quartile.

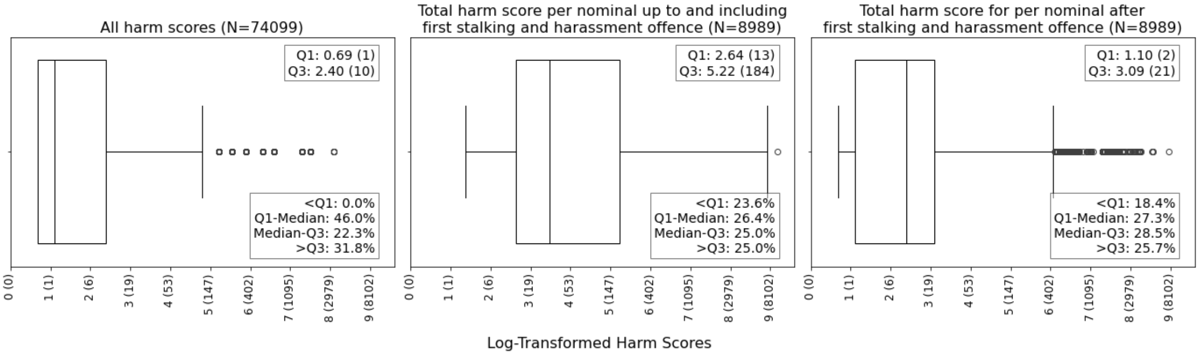


Figure 31. Distribution of log transformed harm scores across all crimes, crimes up to stalking and harassment, and crimes after stalking and harassment

The third boxplot illustrates total harm scores after the first stalking and harassment offence (N=8,989), highlighting a reduction in harm scores with a median of 1.1 (log-transformed value of 2) and Q3 of 3.09 (log-transformed value of 21). The transformation emphasises that post-offence harm scores tend to concentrate in the lower range, suggesting a notable decline in offending severity. Additionally, 25.7% of harm scores are above the third quartile, indicating that a quarter of the cases still maintain relatively higher harm levels.

Overall, the log-transformed plots provide a clearer visualisation of harm distribution across the different stages, confirming patterns observed in the raw data: stalking and harassment are often preceded by higher variability and elevated harm scores, with a tendency towards harm reduction after these offences.

A.2.3 Relationship between harm up to and after stalking and harassment

The left scatter plot depicts the raw total harm scores, while the right plot shows the log-transformed values, providing a clearer view of the distribution by compressing the extreme values and making patterns more discernible.

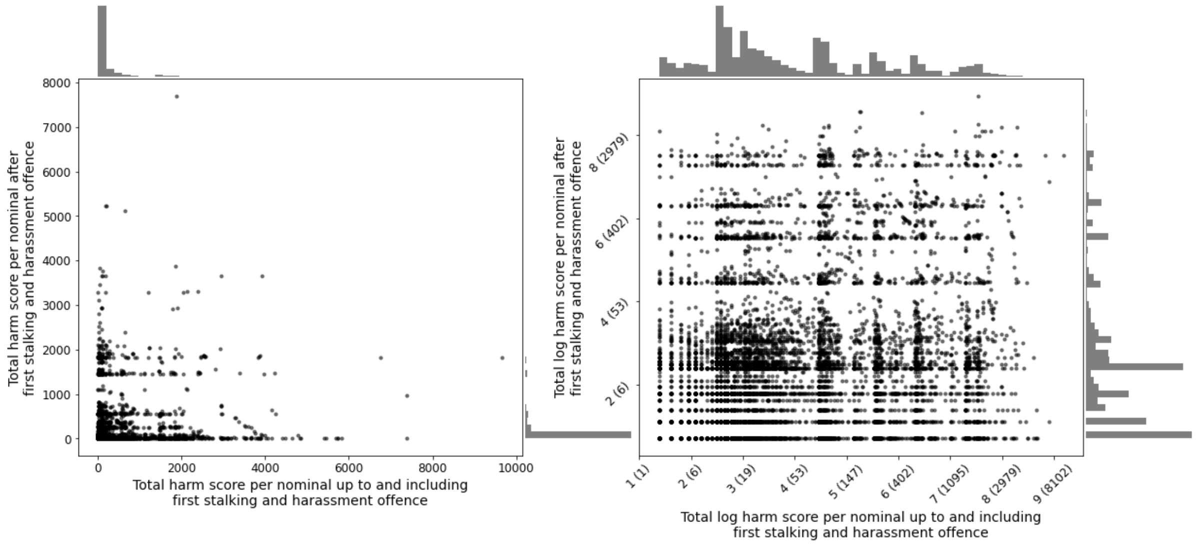


Figure 32. Relationship between harm up to and after stalking and harassment.

From the scatter plots, it is evident that there is no clear linear relationship between harm scores before and after the stalking and harassment offence. The data points are widely dispersed with no discernible trend or line of best fit, indicating that harm levels post-stalking and harassment do not consistently increase or decrease relative to pre-stalking and harassment harm levels in a predictable manner. The spread of data points, particularly in the log-transformed scatter plot, shows a clustering of lower harm scores along the axes, with fewer instances of high harm escalation or de-escalation.

In the right scatter plot, which uses log-transformed scores, the distribution appears somewhat more structured, yet still lacks a clear linear or nonlinear pattern. The log transformation makes the concentrations of harm scores near lower values more visible, highlighting the density of cases with low post-offence harm despite varying pre-offence harm levels.

Figure 33 shows the total harm scores up to and after the first stalking and harassment occurrence, sorted by the pre-stalking and harassment scores and plotted against each other. The grey line represents the total harm score up to the first stalking and harassment offence, while the black line shows the total harm score after the first occurrence. The Pearson correlation coefficient is 0.13 and the Spearman correlation coefficient is 0.15, both with p-values of 0.0, indicating statistically significant but very weak correlations (note that statistical significance is not difficult to achieve with large datasets).

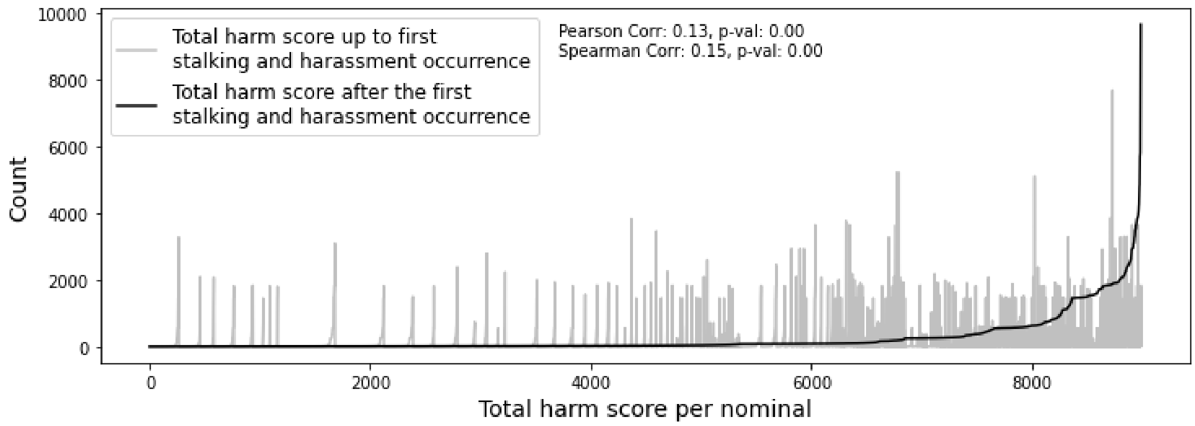


Figure 33. Relationship between harm up to and after stalking and harassment plotted by the sorted pre-stalking and harassment harm scores in ascending order.

From the plot, it is apparent that the post-stalking and harassment harm scores do not consistently follow the pre-offence scores. The grey line, which represents the pre-offence scores, rises steadily but exhibits significant fluctuations and does not align closely with the post-stalking and harassment scores (black line). The variability between the lines suggests that the harm scores after stalking and harassment do not systematically increase or decrease in a manner directly proportional to the harm scores before the offence.

A.2.4 Probabilities and conditional probabilities of pre-harm resulting in post-harm within a threshold

Figure 34 presents the probabilities and conditional probabilities related to the transition from pre-stalking and harassment total harm to post-stalking and harassment total harm, analysed against various harm thresholds. The left plot displays the conditional probabilities, where two specific conditions are evaluated: **(1)** the probability that the total post-stalking and harassment harm is less than or equal to the threshold given that the total pre-stalking and harassment harm is less than or equal to the total post-stalking and harassment harm, and **(2)** the probability that the total post-stalking and harassment harm is less than or equal to the threshold given that the total pre-stalking and harassment harm is greater than the total post-stalking and harassment harm. The right plot shows the overall probability that the total post-harm is less than or equal to the threshold.

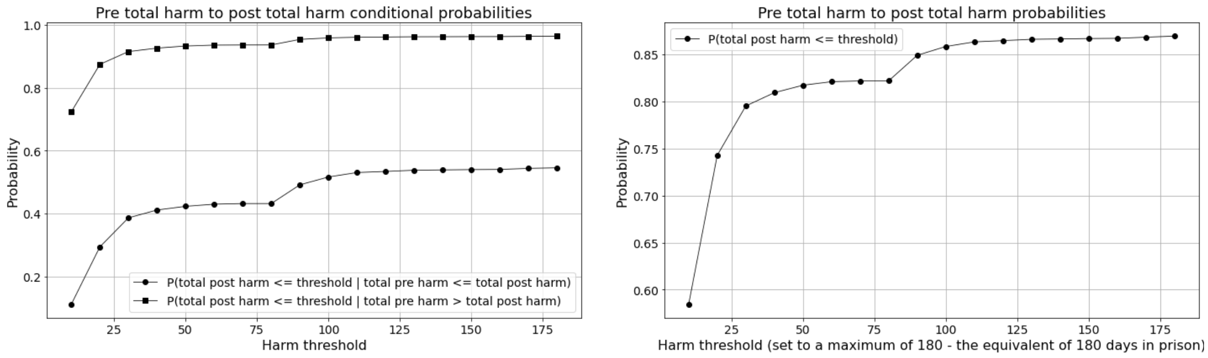


Figure 34. Probabilities and conditional probabilities of pre-harm resulting in post-harm within a threshold.

The conditional probabilities (left plot) indicate distinct patterns based on the relationship between pre- and post-harm scores. For cases where pre-harm is less than or equal to post-harm, the probability that post-harm remains below or equal to the threshold starts high and gradually approaches 1 as the threshold increases. This suggests a strong tendency for offenders whose harm remains stable or increases slightly to still have post-harm scores below the threshold, especially at higher thresholds.

In contrast, for cases where pre-harm is greater than post-harm, the conditional probability remains relatively lower, showing a modest increase with rising thresholds but plateauing around 0.5. This reflects that when harm decreases after the stalking and harassment offence, there is a more balanced chance of post-harm staying under or going above the threshold, indicating less predictability.

The overall probability that the total post-harm is less than or equal to the threshold shows a generally upward trend as the harm threshold increases, with probabilities nearing 1 as the threshold approaches the upper limits of 180 (equivalent to 180 days in prison). This consistent rise suggests that, broadly, the likelihood of post-harm scores falling within the threshold grows as the threshold itself becomes less restrictive.

A.3 Gender, age, offending, and harm

This section examines the distribution of crimes by gender and age, followed by a comparison of total harm scores across different genders and age groups.

A.3.1 Distribution of crimes per gender and age

Figure 35 illustrates the distribution of crimes based on gender and age at the time of offence across three subfigures: **(1)** the number of crimes per gender, **(2)** the overall distribution of crimes by age, and **(3)** the distribution of crimes by age and gender.

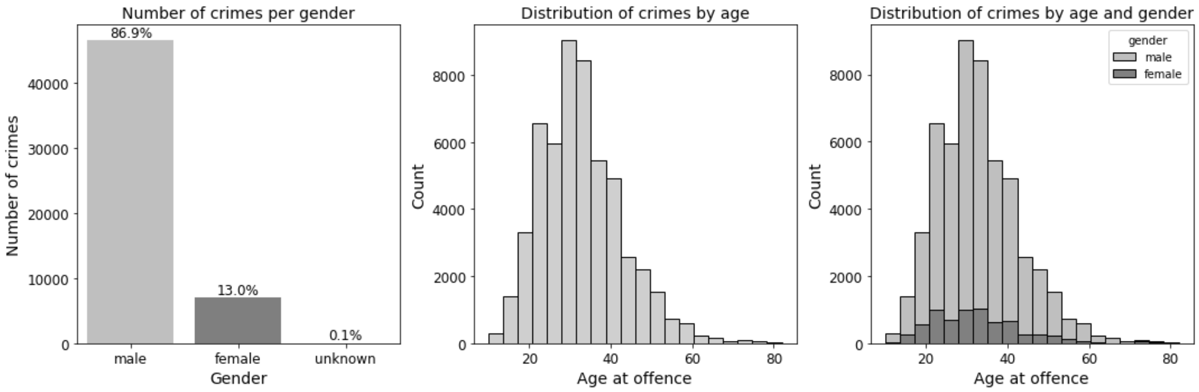


Figure 35. Distribution of crimes per gender and age.

The left subfigure shows a bar chart of the number of crimes committed by gender. The overwhelming majority of crimes were committed by males, accounting for 86.9% of all offences. Females are responsible for 13% of the offences, while a negligible 0.1% of crimes are attributed to individuals with an unknown gender. This significant disparity suggests a strong gender skew, with males being the predominant offenders.

The middle subfigure displays the overall age distribution of crimes, revealing that most offences are committed by individuals in their early 20s to late 40s, with a peak around the late 20s to early 30s. The histogram shows a bell-shaped distribution with a gradual decline in crime rates as age increases beyond 40.

The right subfigure breaks down this age distribution by gender, highlighting that both male and female offenders generally follow a similar age distribution pattern. However, the distribution shows that male offenders are consistently more numerous across all age groups, reflecting the overall higher number of crimes committed by males. The peak age for male offenders also aligns closely with the overall peak, while female offenders are comparatively fewer across all age brackets.

A.3.2 Total harm per gender up to and after first stalking and harassment offence

Figure 36 compares total harm scores by gender, both up to and after the first stalking and harassment offence, highlighting differences in harm patterns between males and females over these periods. For male offenders, the distribution of harm scores is broad, with numerous extreme outliers, especially before the first S&H offence. These outliers suggest that while most male offenders have relatively low harm scores, a subset contributes significantly higher harm, resulting in a wide spread. After the first stalking and harassment offence, male harm scores tend to decrease, although outliers remain present, indicating ongoing variability in offending severity.

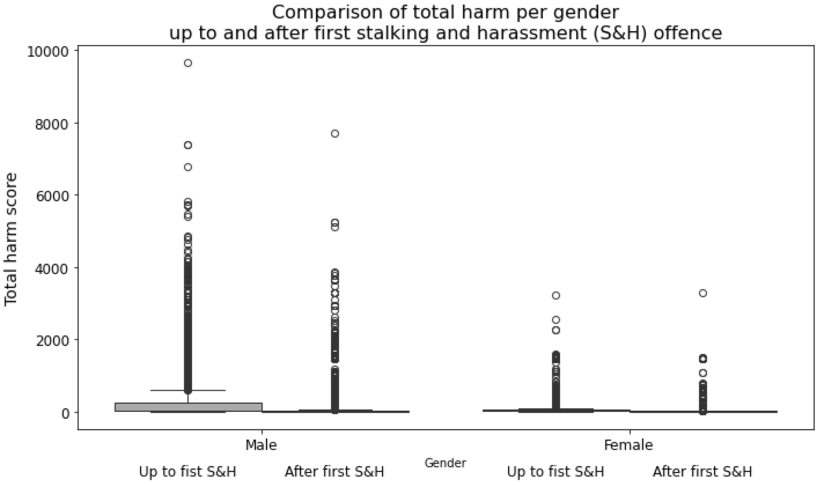


Figure 36. Total harm per gender up to and after first stalking and harassment offence

In contrast, female offenders exhibit lower overall harm scores, with fewer extreme outliers compared to their male counterparts. The harm scores for females are more concentrated at the lower end of the scale both before and after a stalking and harassment offence, indicating that females generally contribute less to total harm. The harm distribution among females remains relatively stable across the two time frames, with no significant increase in harm severity observed post-offence.

Comparing across genders and time frames, the data suggests a potential de-escalation in offending severity following the first stalking and harassment offence, particularly among males who show a marked reduction in harm scores post-offence. This pattern is less pronounced among females, whose harm scores are consistently lower and more stable.

A.3.3 Relationship between age and total harm after first stalking and harassment occurrence

Figure 37 explores the relationship between age at the time of offence and the total harm scores after the first stalking and harassment occurrence. The scatter plot visualises this relationship with data points colored by intensity, ranging from blue to red, across a wide age spectrum from approximately 10 to 80 years. Notably, both the Pearson correlation coefficient (-0.01, p-value = 0.12) and the Spearman correlation coefficient (-0.01, p-value = 0.13) indicate a very weak, statistically insignificant negative correlation between age and post-offence total harm scores.

The scatter plot shows a broad distribution of harm scores across all age groups, with no clear trend indicating a strong relationship between age and the severity of harm post-offence. High harm scores are observed across a wide range of ages, particularly clustering in younger offenders from their teens to their 30s. However, the plot does not show a consistent pattern of increasing or decreasing harm scores with age, as high and low harm scores are scattered relatively evenly across the entire age range.

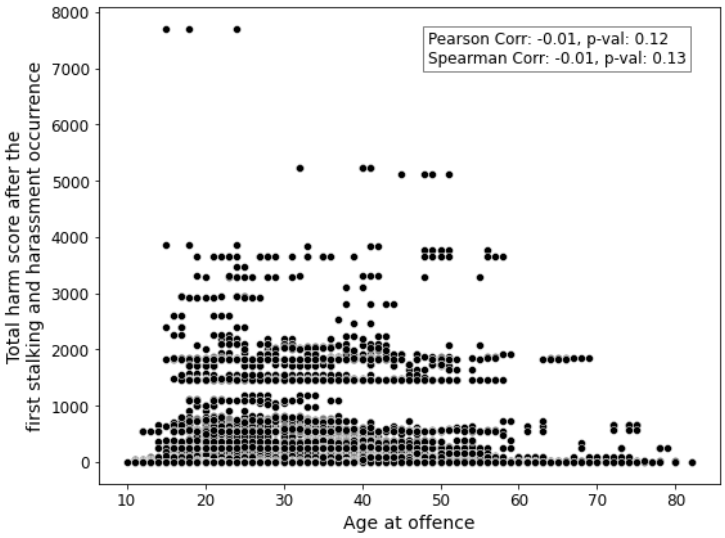


Figure 37. Relationship between age and total harm after first stalking and harassment occurrence.

For younger offenders (under 30 years), there is a noticeable concentration of high harm scores, with some instances reaching values above 7,000. As age increases, the density of high harm scores decreases slightly, with harm levels appearing more stable and less extreme, especially beyond 50 years of age. Nonetheless, even older age groups display variability in harm scores, albeit with a lower overall intensity and fewer extreme outliers.

A.4 Role in crime

A.4.1 Distribution of total harm scores after first stalking and harassment offence by crime role

Figure 38 explores the distribution of total harm scores after the first stalking and harassment offence, comparing nominals who were both offenders and victims to those who were solely offenders. The analysis is presented across three subfigures: the first subfigure compares the number of nominals by their role in crime, while the second and third subfigures display the harm score distributions for each group.

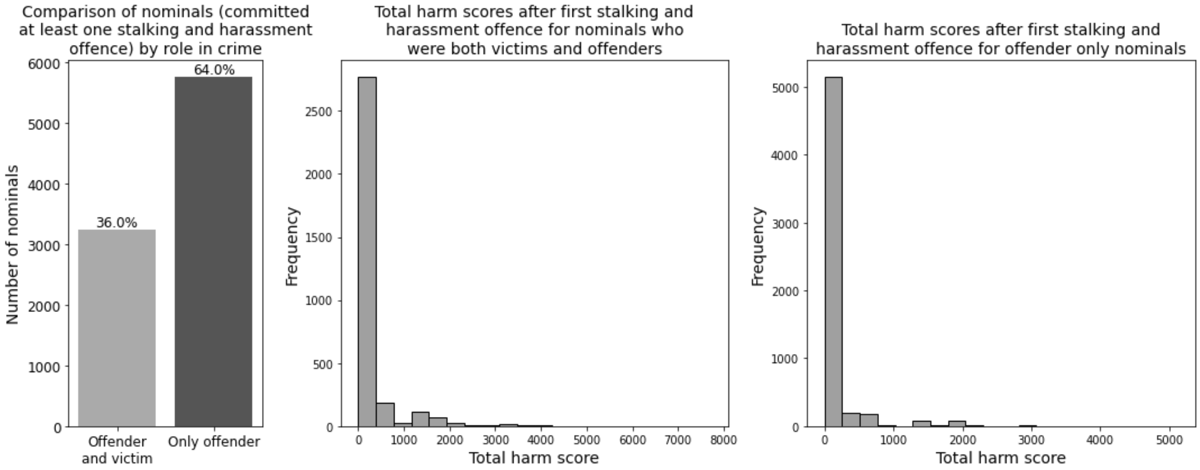


Figure 38. Distribution of total harm scores after first stalking and harassment offence by crime role.

The left subfigure shows a bar chart comparing the number of nominals who committed at least one stalking and harassment offence based on their role in crime. A significant proportion, 64%, were only offenders, while 36% of nominals were involved as both offenders and victims. This distribution indicates that a majority of individuals involved in stalking and harassment offences do not have a dual role as victims, highlighting a distinct separation in offender and victim roles within the data.

The middle subfigure displays the distribution of total harm scores after the first stalking and harassment offence for nominals who were both offenders and victims. The distribution is heavily right skewed with the majority of harm scores clustered below 1,000. A small number of cases exhibit much higher harm scores, extending up to 8,000, although these are outliers within the broader trend of relatively low post- stalking and harassment harm among this group.

The right subfigure is the total harm scores for nominals who were solely offenders. Similar to the middle subfigure, the distribution is also heavily right skewed with the majority concentrated below 1,000. However, the range of harm scores is slightly narrower, and the highest values do not extend as far as those seen among nominals who were both offenders and victims. This suggests that while both groups predominantly exhibit low harm scores post-stalking and harassment, those with dual roles occasionally reach higher levels of harm.

A.4.2 Relationship between total harm score after the first stalking and harassment offence and role in crime

Figure 39 explores the relationship between total harm scores after the first occurrence of stalking and harassment and the nominal's role in the crime, comparing individuals who were both offenders and victims to those who were solely offenders. The scatter plot uses color intensity to represent the magnitude of the harm scores, with redder colors indicating higher harm levels.

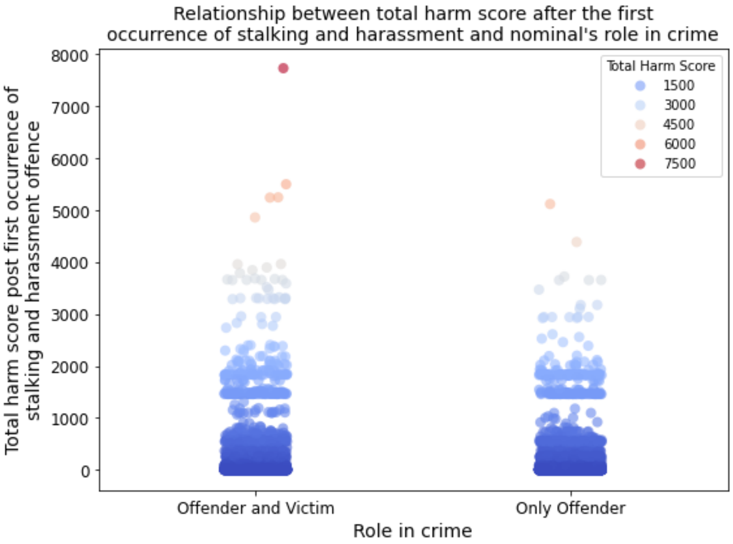


Figure 39. Relationship between total harm score after the first stalking and harassment offence and role in crime.

For individuals who were both offenders and victims, the harm scores are mostly concentrated at the lower end, similar to the offender-only group. However, there is a noticeable spread of harm scores extending into higher ranges, including some extreme values that exceed 7,000. This indicates that while most dual-role individuals have relatively low harm scores, there are instances of significantly higher harm, suggesting a subset of individuals with both offender and victim roles who contribute disproportionately to total harm.

In the group of individuals who were only offenders, the harm scores also predominantly cluster at the lower end, with most scores below 1,000. The distribution is slightly narrower compared to the offender and victim group, with fewer high harm outliers. This suggests that offenders without a victim role tend to exhibit more uniform, lower harm scores after their first stalking and harassment offence.

A.5 Victim - offender/suspect relationship

This subsection examines the extent to which the nature of the victim-offender relationship is linked to the escalation in harm following stalking and harassment offences.

A.5.1 Prevalence of offender-victim relationships in stalking and harassment offences

Figure 40 displays the prevalence of various offender-victim relationships in stalking and harassment offences, highlighting the different types of relationships between offenders and their victims. The bar chart categorises these relationships into six groups: Partner, Ex-Partner, Family Member, Professional, Other/Unknown, and Neighbour/Friend/ Acquaintance, with each bar representing the count and percentage of offences within each category.

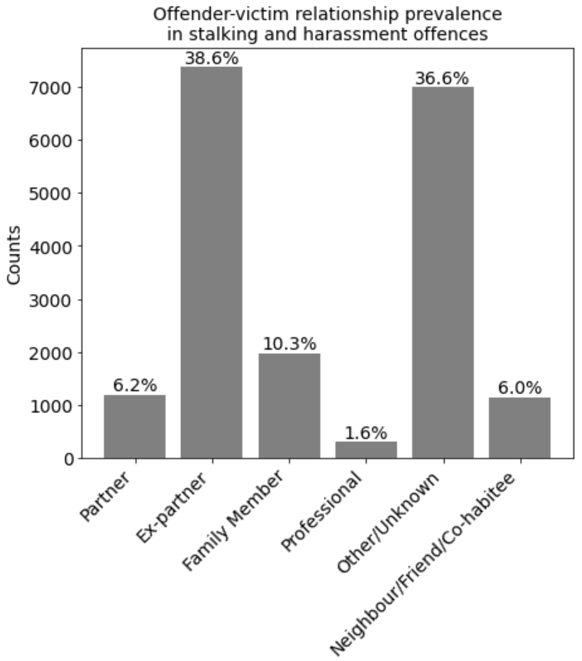


Figure 40. Prevalence of offender-victim relationships in stalking and harassment offences.

The most prevalent offender-victim relationship in stalking and harassment offences is with an ex-partner, accounting for 38.6% of the cases. This suggests that a significant portion of stalking and harassment incidents are linked to past intimate relationships, where tensions and unresolved conflicts may drive offending behaviour.

The second most common category is "Other/Unknown", comprising 36.6% of the offences. This broad category likely includes various relationships that do not fit neatly into specific classifications or where the relationship type was not clearly identified (see **Table 4**).

Family members and neighbours, friends, or acquaintances account for 10.3% and 6% of the cases, respectively. These figures indicate that while these relationships are less prevalent than those involving ex-partners, they still represent a notable proportion of stalking and harassment offences, suggesting that personal and social connections are also common contexts for such crimes.

Current partners represent 6.2% of the cases, while professional relationships are the least common, making up only 1.6% of the total. The lower prevalence in these categories might reflect differing dynamics in these relationships, such as ongoing partnerships or professional boundaries that potentially reduce the likelihood of stalking and harassment.

A.5.2 Distribution of total harm pre-stalking and harassment by relationship type

Figure 41 examines the distribution of total harm scores by offender-victim relationship type up to the first stalking and harassment offence, along with the percentage of nominals with high total harm scores (≥180) categorised by relationship type. The relationship type reflects the status between the offender and the victim at the time of the stalking and harassment offence.

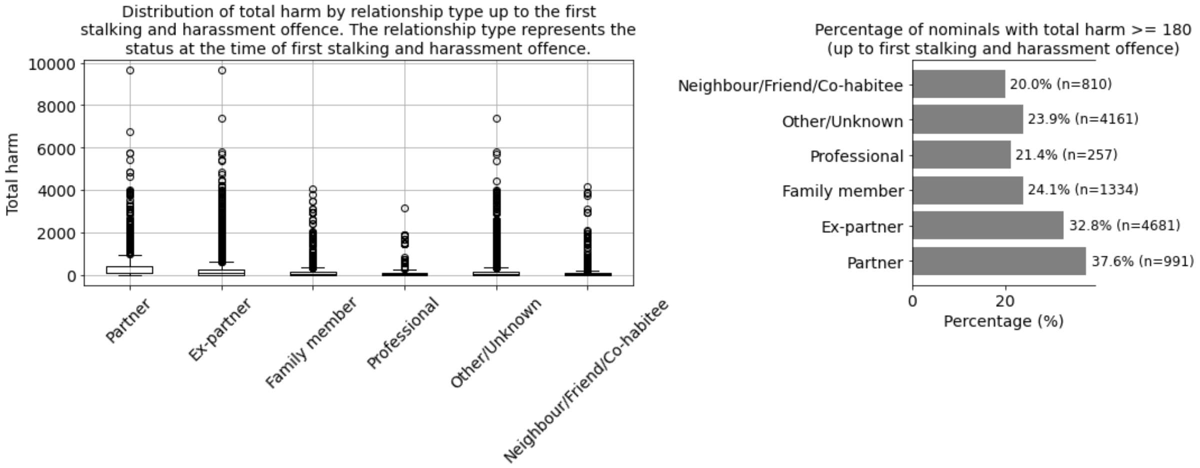


Figure 41. Distribution of total harm pre-stalking and harassment by relationship type.

The boxplots in the left subfigure show the spread and variability of total harm scores across different relationship types, including Partner, Ex-Partner, Family Member, Professional, Other/Unknown, and Neighbour/Friend/Co-habitee. Across all categories, the distributions are heavily skewed with median values consistently low. However, there are numerous high outliers across each relationship type, indicating that a small number of cases within each category contribute disproportionately high harm scores.

- **Partner and Ex-Partner:** Both categories show a wide range of harm scores with many extreme outliers, suggesting high variability in harm levels. Notably, ex-partners have a broader spread and more high-harm outliers compared to current partners.

- **Family Member and Neighbour/Friend/Co-habitee:** These categories also exhibit variability in harm scores but with fewer high outliers compared to partner-related categories.

- **Professional:** This category shows the least variability and fewest high-harm outliers, indicating that offences involving professional relationships generally have lower harm levels.

- **Other/Unknown:** This group shows a similar spread to other categories but with a notable number of high outliers, reflecting a range of unknown or less common relationships with varying harm impacts.

The bar chart on the right illustrates the percentage of nominals with total harm scores of 180 or greater, categorised by relationship type. The data reveals distinct differences in the likelihood of high harm scores based on relationship type:

- **Partner (37.6%) and Ex-Partner (32.4%):** These groups have the highest percentages of high-harm scores, underscoring the significant risk associated with intimate relationships, particularly when involving former partners.
- **Family Member (24.1%) and Other/Unknown (23.9%):** These categories show moderate levels of high-harm cases, reflecting that familial and various other relationships also contribute to elevated harm, though to a lesser extent than partner-related categories.
- **Professional (21.4%) and Neighbour/Friend/Co-habitee (20%):** These groups have the lowest percentages of high-harm scores, suggesting that offences involving these relationships are generally less severe in terms of harm.

A.5.3 Distribution of total harm post-stalking and harassment by relationship type

Figure 42 examines the distribution of total harm scores by offender-victim relationship type after the first stalking and harassment offence, as well as the percentage of nominals with high total harm scores (≥180) in each relationship category. The relationship type reflects the offender-victim status at the time of the first stalking and harassment offence.

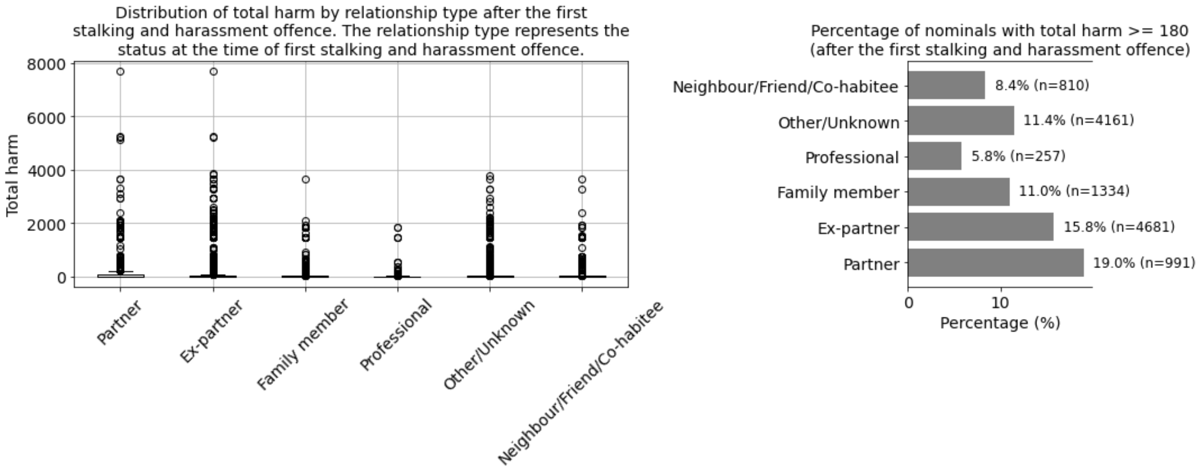


Figure 42. Distribution of total harm post-stalking and harassment by relationship type.

The boxplots in the left subfigure display the spread of total harm scores across different relationship types, including Partner, Ex-Partner, Family Member, Professional, Other/Unknown, and Neighbour/Friend/Co-habitee. Across all categories, the distributions are heavily right skewed with most values clustered near the bottom of the scale. However, each category also includes high outliers, indicating that although most cases have low harm levels post-offence, there are instances of significant harm in all relationship types:

- **Partner and Ex-Partner:** These categories continue to show wide variability in harm scores with numerous high outliers, reflecting that intimate partner-related offences often involve substantial harm, even after the stalking and harassment offence.
- **Family Member and Neighbour/Friend/Co-habitee:** These relationships also display variability but generally have fewer extreme outliers compared to partner-related categories, indicating a somewhat more stable pattern of lower harm levels.
- **Professional:** This group has the least variability and the fewest high outliers, suggesting that stalking and harassment involving professional relationships are typically less severe in terms of harm.
- **Other/Unknown:** This category maintains a range of harm scores with notable high outliers, representing various less common or undefined relationship types that still contribute to overall harm variability.

The bar chart on the right illustrates the percentage of nominals with total harm scores of 180 or greater after the first stalking and harassment offence, categorised by relationship type:

- **Partner (19%) and Ex-Partner (15.8%):** These categories have the highest percentages of high-harm cases, reinforcing the elevated risk associated with intimate relationships, particularly in post-offence scenarios.
- **Family Member (11%) and Other/Unknown (11.4%):** These categories exhibit moderate levels of high-harm instances, showing that family and various other relationships contribute meaningfully to harm but are less severe than partner-related cases.
- **Neighbour/Friend/Co-habitee (8.4%) and Professional (5.8%):** These groups have the lowest percentages of high-harm scores, indicating that offences within these relationship types generally result in lower harm levels post-stalking and harassment.

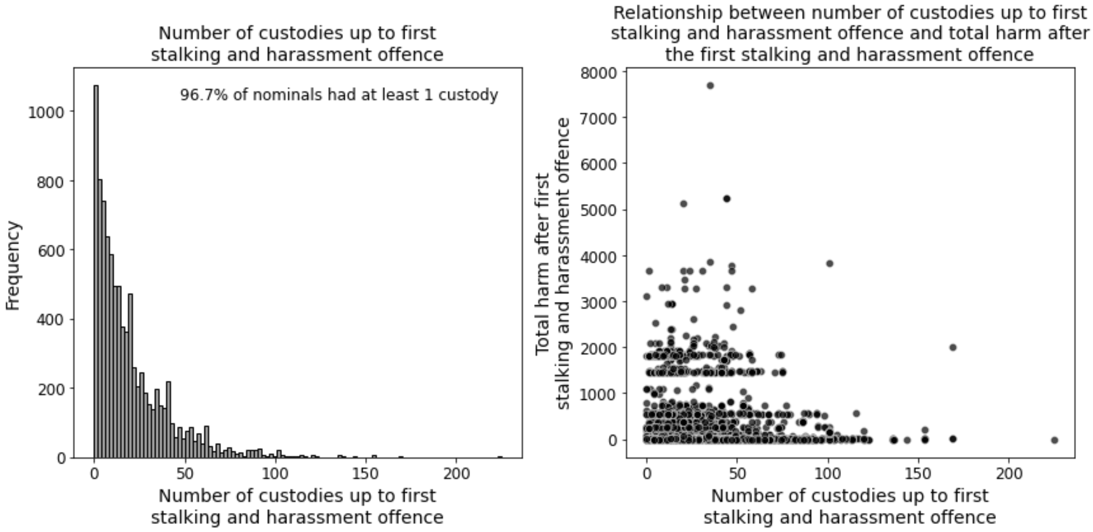


Figure 43. Offenders who were in custody up to the first stalking and harassment offence.

Figure 43 explores the relationship between the number of custodial sentences nominals received up to their first stalking and harassment offence and the total harm scores following that offence. The left subfigure shows a histogram where the data is skewed towards fewer custodial sentences, with 96.7% of nominals having at least one custody. Most nominals have fewer than 10 custodial instances, with a rapid decline in frequency as the number increases. The right subfigure presents a scatter plot revealing no clear trend between the number of custodial sentences and subsequent harm scores, which mostly cluster below 2,000. Even for nominals with many custodial sentences, harm levels remain inconsistent, suggesting that prior custodial frequency is not a strong predictor of harm severity post-offence.

A.6 Calls made to the police as a victim

Figure 44 examines the number of times nominals were flagged as victims in calls to the police before their first stalking and harassment offence and their impact on subsequent harm scores. The left subfigure shows a highly skewed distribution where 9.3% of nominals had at least one such call, but most had none or very few instances. The scatter plot in the right subfigure indicates no clear relationship between victim-related calls and harm scores (resulting from nominals' behaviour), which generally stay below 2,000. High harm scores appear both among those with no calls and those with a few, suggesting that being identified as a victim does not strongly influence harm levels after the offence.

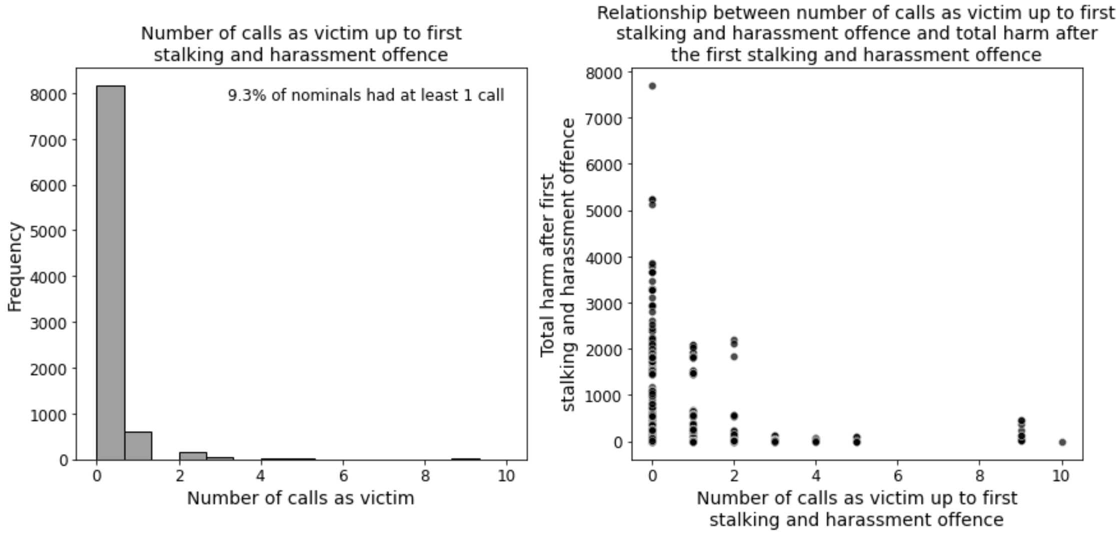


Figure 44. Calls made by offender nominals when also recorded as victim.

A.7 Number of crimes up to the first stalking and harassment offence

Figure 45 investigates the number of previous offences nominals had before their first stalking and harassment offence and their correlation with subsequent harm scores. The left subfigure displays a histogram showing a skewed distribution with 37% having at least five prior offences and 12.9% having at least ten. The scatter plot in the right subfigure shows no clear relationship between the number of previous offences and harm scores, which mostly cluster below 2,000. While high harm scores are observed across varying offence counts, there is no consistent increase in harm with a greater number of prior offences, indicating that the number of previous crimes is not a strong predictor of harm severity post-offence.

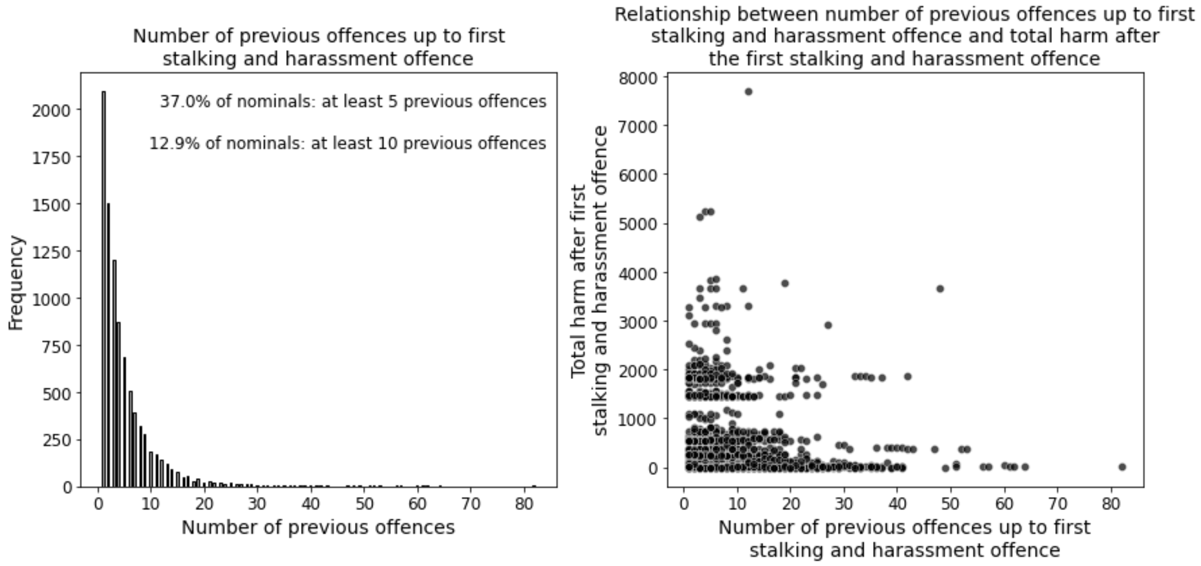


Figure 45. Crimes up to the first stalking and harassment offence.

A.8 Relationship between crime type indicators and harm post-stalking and harassment

This subsection examines the relationship between the types of crimes identified in pre-stalking and harassment offences, using keyword indicators as outlined in **Table 5**, and the total harm committed after the stalking and harassment offence.

A.8.1 Physical violence, physical assault, and post total harm

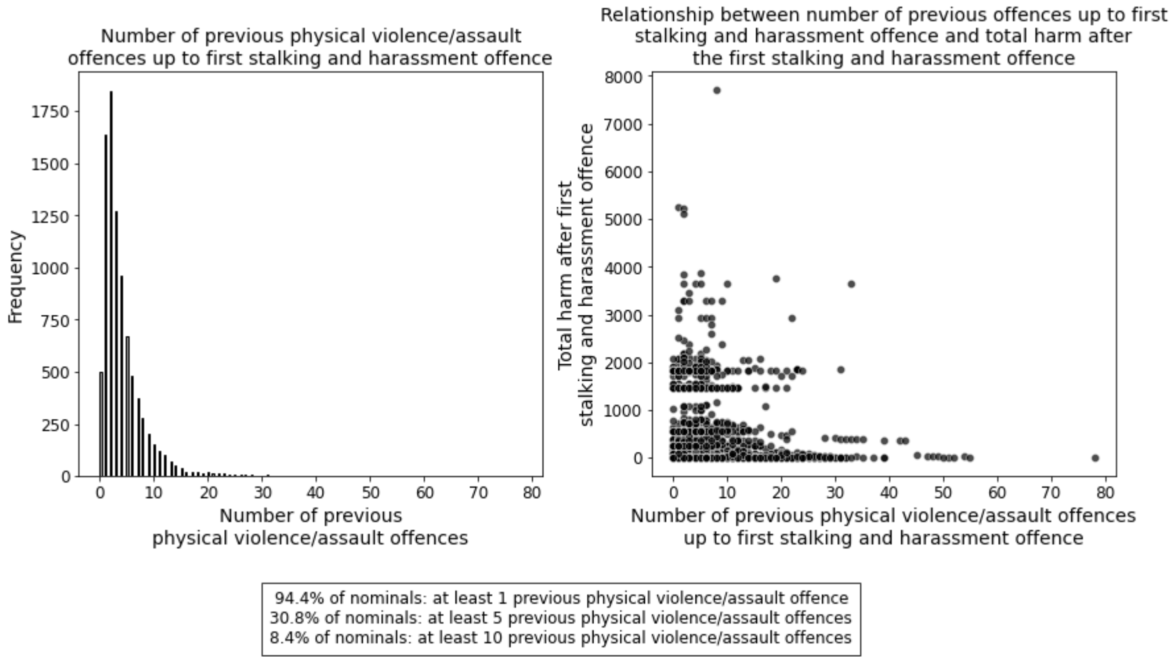
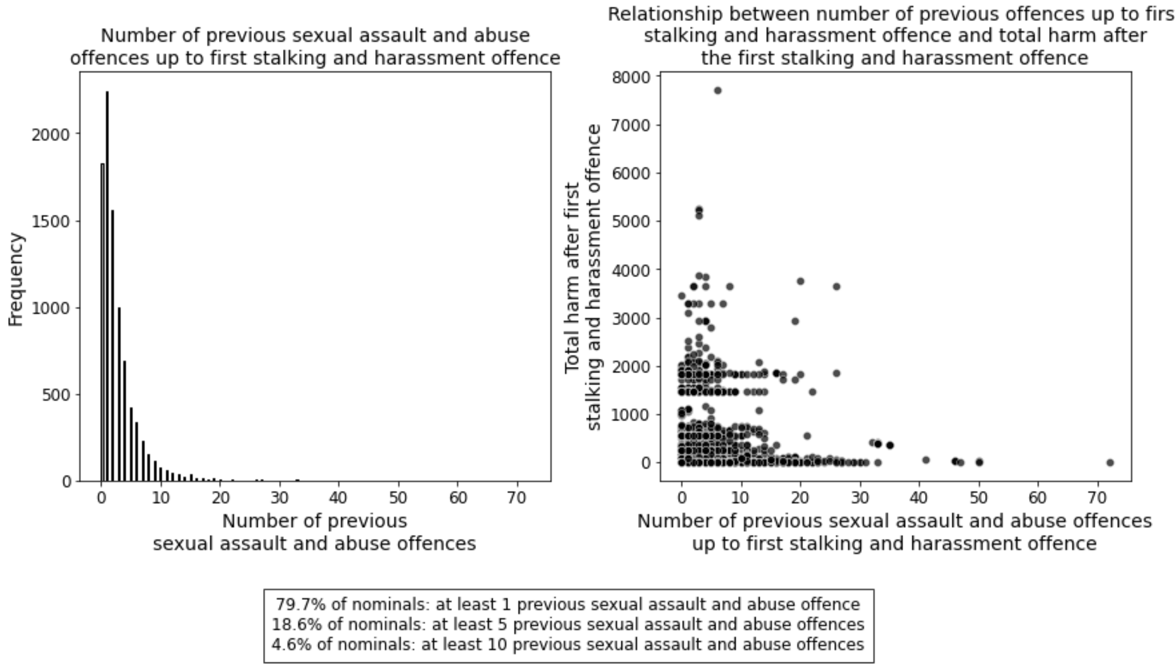


Figure 46. Relationships between physical violence, physical assault and post stalking and harassment harm.

Figure 46 examines the relationship between the number of prior physical violence or assault offences before the first stalking and harassment offence and the subsequent harm scores. The left subfigure shows a histogram with a skewed distribution; 94.4% of nominals had at least one prior offence, but only 8.4% had ten or more, indicating that accumulating numerous offences is uncommon. The scatter plot in the right subfigure reveals no clear trend between the number of prior physical violence offences and post-offence harm scores, which generally cluster below 2,000, suggesting that prior physical violence does not strongly predict harm severity.

A.8.2 Sexual assault, sexual abuse, and post total harm

Figure 47 explores the link between the number of previous sexual assault and abuse offences and harm scores following the first stalking and harassment offence. The left subfigure displays a histogram showing that while 79.7% of nominals had at least one such offence, only 4.6% had ten or more, reflecting the rarity of high offence accumulation. The scatter plot in the right subfigure indicates a broad distribution of harm scores, with most remaining below 2,000 regardless of the number of prior sexual offences. The absence of a consistent pattern suggests that the frequency of past sexual offences does not reliably predict post-offence harm severity.

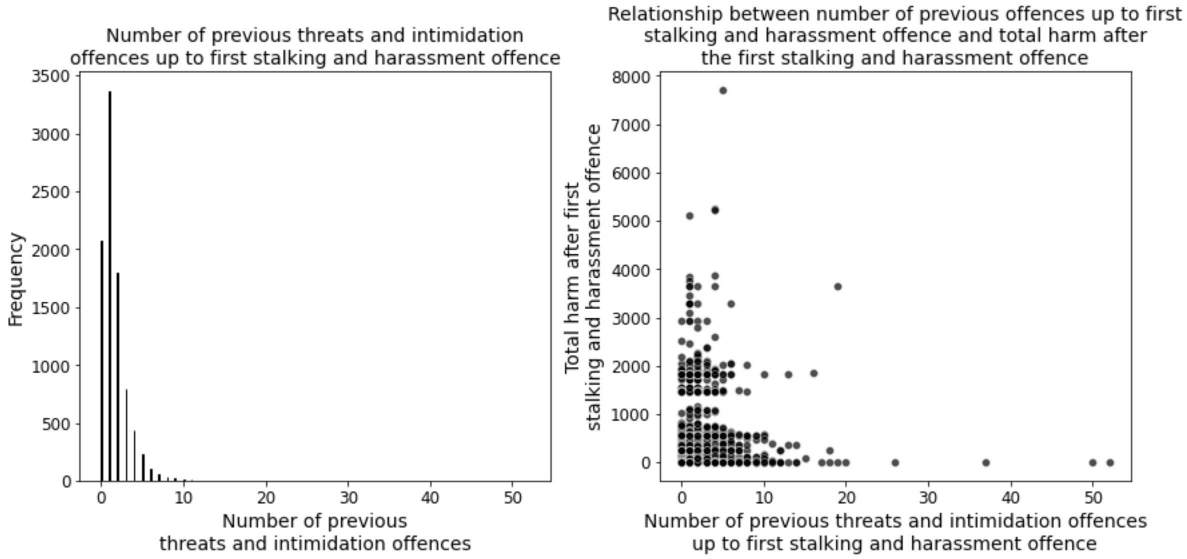


79.7% of nominals: at least 1 previous sexual assault and abuse offence
18.6% of nominals: at least 5 previous sexual assault and abuse offences
4.6% of nominals: at least 10 previous sexual assault and abuse offences

Figure 47. Relationship between sexual assault, sexual abuse, and post stalking and harassment harm.

A.8.3 Threats, intimidation, and post total harm

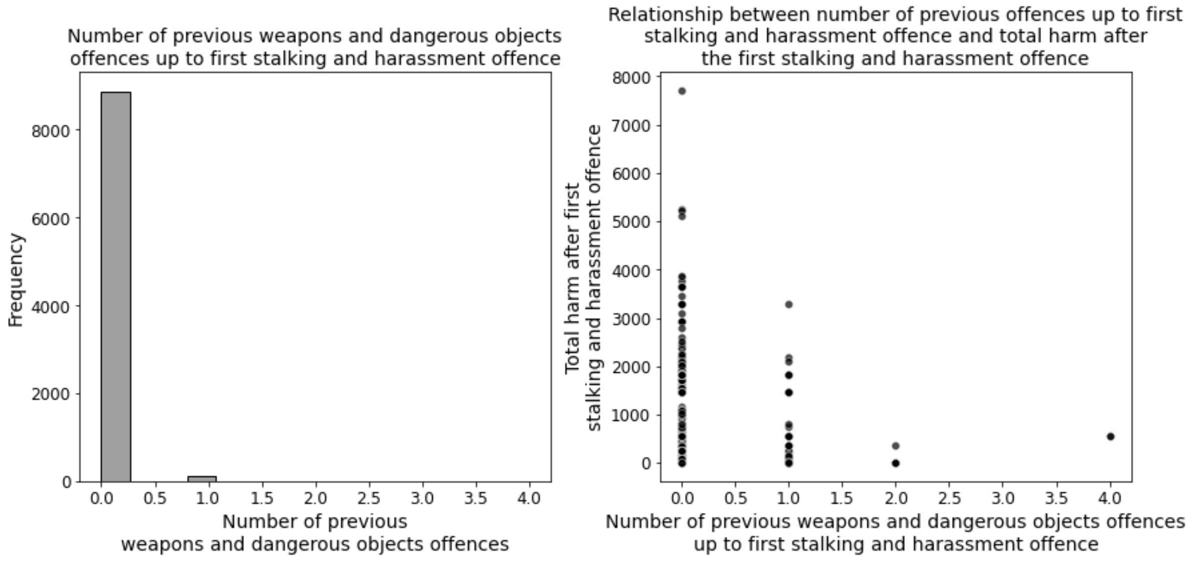
Figure 48 investigates the relationship between the number of prior threats and intimidation offences and subsequent harm scores after the first stalking and harassment offence. The left panel shows a skewed distribution, with 76.9% of nominals having at least one offence, but only 0.6% reaching ten or more, indicating that extensive histories of threats are rare. The scatter plot in the right subfigure reveals a wide dispersion of harm scores, mostly clustering below 2,000, with no clear relationship to the number of prior offences. This suggests that the count of past threats and intimidation offences is not a strong predictor of harm severity after the offence.



76.9% of nominals: at least 1 previous threats and intimidation offence
5.7% of nominals: at least 5 previous threats and intimidation offences
0.6% of nominals: at least 10 previous threats and intimidation offences

Figure 48. Relationship between threats, intimidation, and post stalking and harassment harm.

A.8.4 Weapons, dangerous objects, and post total harm



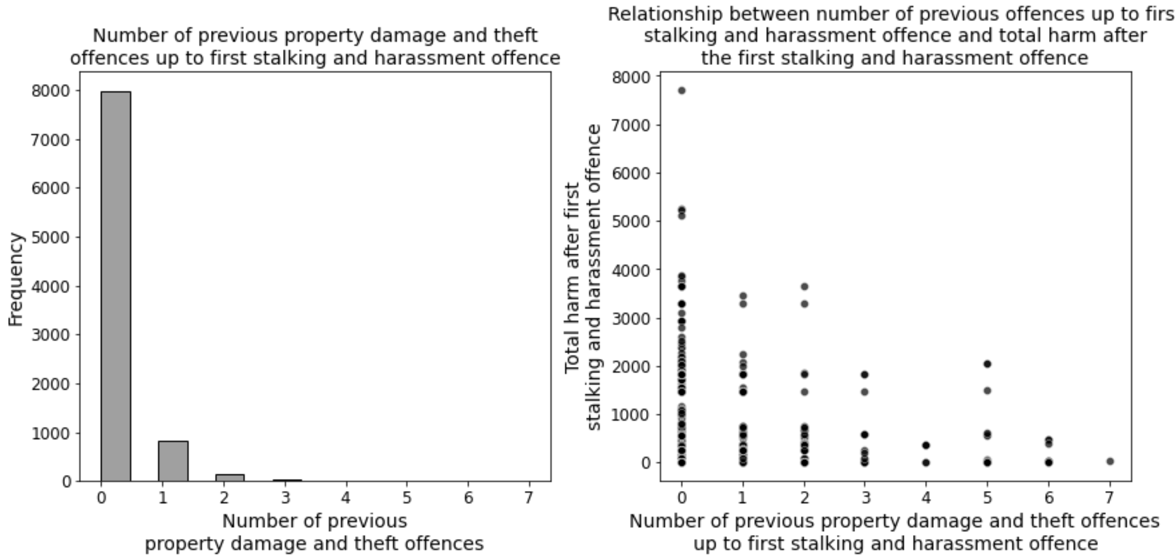
1.4% of nominals: at least 1 previous weapons and dangerous objects offence
0.0% of nominals: at least 5 previous weapons and dangerous objects offences
0.0% of nominals: at least 10 previous weapons and dangerous objects offences

Figure 49. Relationship between weapons, dangerous objects, and post stalking and harassment harm

Figure 49 explores the relationship between prior weapons and dangerous objects offences and harm scores after the first stalking and harassment offence. The left panel shows a histogram indicating that such offences are extremely rare, with only 1.4% of nominals having at least one offence and none accumulating five or more. The scatter plot in the right subfigure shows that harm scores mostly cluster below 2,000, with no clear pattern linking the frequency of these offences to increased harm severity, highlighting the infrequency and limited predictive value of these offences in harm outcomes.

A.8.5 Property damage, theft, and post total harm

Figure 50 examines the relationship between previous property damage and theft offences and harm scores after the first stalking and harassment offence. The left subfigure shows a skewed distribution with 11.4% of nominals having at least one such offence, but only 0.2% having five or more, highlighting the rarity of frequent offences. The scatter plot in the right subfigure reveals no clear correlation between the number of prior property damage and theft offences and harm scores, which largely remain below 2,000, suggesting these offences are not strong predictors of increased harm severity.

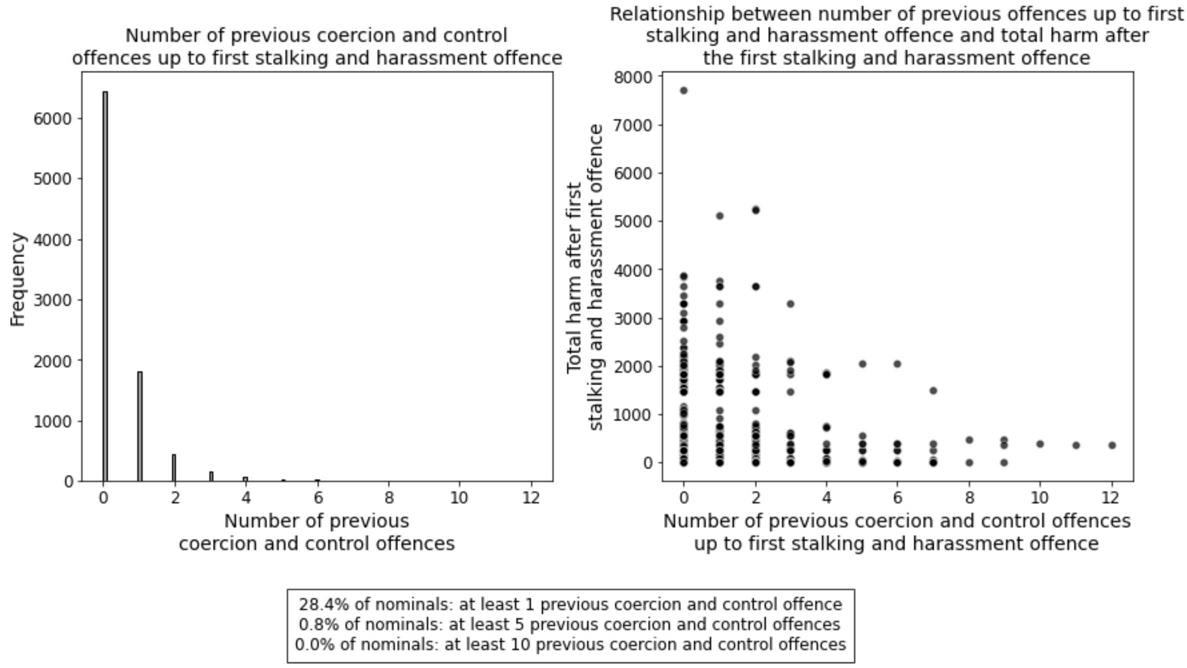


11.4% of nominals: at least 1 previous property damage and theft offence
0.2% of nominals: at least 5 previous property damage and theft offences
0.0% of nominals: at least 10 previous property damage and theft offences

Figure 50. Relationship between property damage, theft, and post stalking and harassment harm

A.8.6 Coercion, control, and post total harm

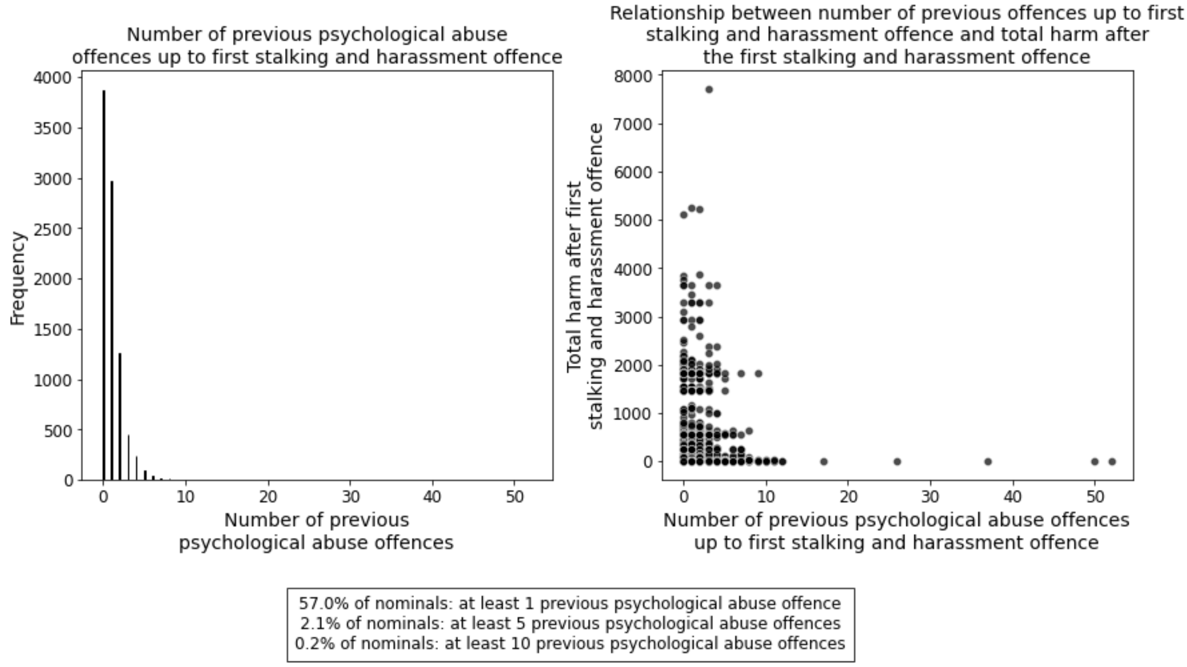
Figure 51 investigates the relationship between prior coercion and control offences and harm scores following the first stalking and harassment offence. The left subfigure presents a histogram showing that 28.4% of nominals had at least one offence, but accumulating five or more is rare, with only 0.8% reaching this level. The right scatter plot in the right subfigure displays no clear trend between the number of coercion and control offences and harm scores, which mostly cluster below 2,000. This suggests that these offences do not consistently predict harm severity, pointing to the influence of other factors beyond offence count.



28.4% of nominals: at least 1 previous coercion and control offence
0.8% of nominals: at least 5 previous coercion and control offences
0.0% of nominals: at least 10 previous coercion and control offences

Figure 51. Relationship between coercion, control, and post stalking and harassment harm.

A.8.7 Psychological abuse and post total harm



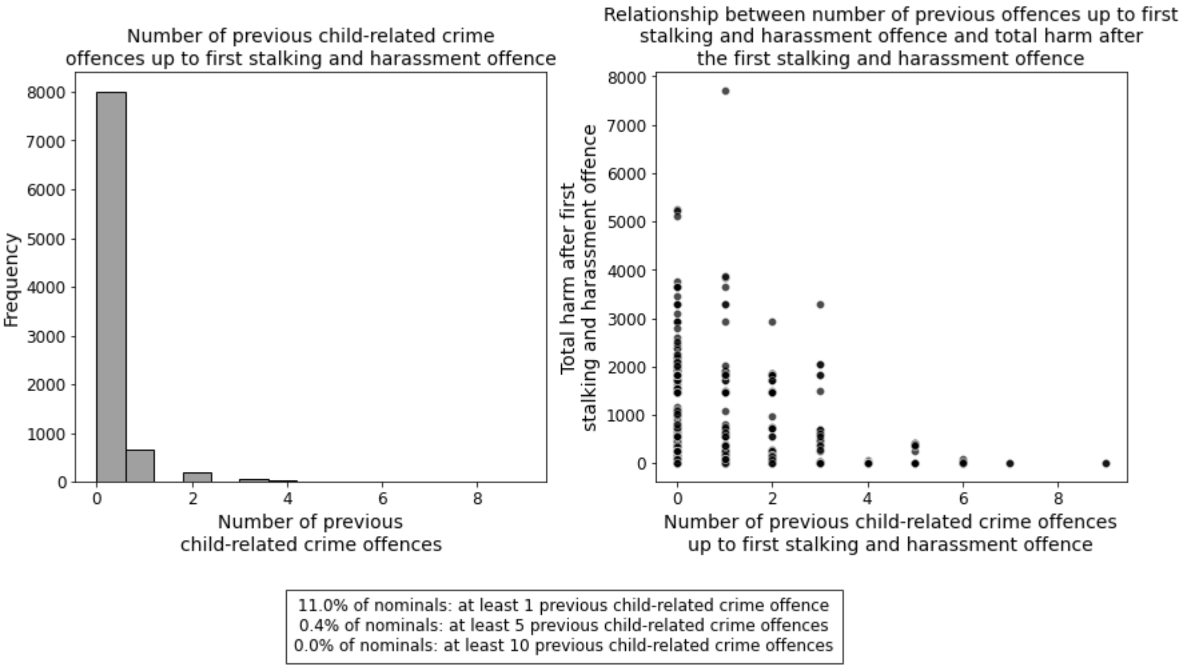
57.0% of nominals: at least 1 previous psychological abuse offence
2.1% of nominals: at least 5 previous psychological abuse offences
0.2% of nominals: at least 10 previous psychological abuse offences

Figure 52. Relationship between psychological abuse and post stalking and harassment harm.

Figure 52 explores the relationship between prior psychological abuse offences and harm scores following the first stalking and harassment offence. The left subfigure shows a histogram with a skewed distribution: 57% of nominals had at least one offence, but only 2.1% had five or more, and 0.2% reached ten, indicating that while psychological abuse is somewhat common, frequent occurrences are rare. The scatter plot in the right subfigure shows no clear correlation between the number of psychological abuse offences and harm severity, with most scores clustering below 2,000. High harm scores are scattered across varying offence levels, suggesting that other factors beyond offence count contribute to harm outcomes.

A.8.8 Child-related crimes and post total harm

Figure 53 examines the link between previous child-related crime offences and harm scores after the first stalking and harassment offence. The left subfigure presents a histogram revealing that 11% of nominals had at least one child-related offence, but only 0.4% had five or more, and none reached ten, highlighting the infrequency of these offences. The scatter plot in the right subfigure shows a sparse distribution of harm scores, mostly below 2,000, with no clear trend between offence frequency and harm severity. The scattered instances of higher harm scores suggest that factors other than the number of child-related offences likely influence harm outcomes.



11.0% of nominals: at least 1 previous child-related crime offence
0.4% of nominals: at least 5 previous child-related crime offences
0.0% of nominals: at least 10 previous child-related crime offences

Figure 53. Relationship between child-related crimes and post stalking and harassment harm.

A.8.9 Life endangering crimes and post total harm

Figure 54 examines the relationship between previous life-endangering offences and harm scores following the first stalking and harassment offence. The left subfigure shows a histogram where 9.2% of nominals had at least one life-endangering offence, but only 0.1% reached five or more, indicating the rarity of such offences. The scatter plot in the right subfigure reveals no clear pattern between the number of life-endangering offences and harm scores, which mostly cluster below 2,000. The scattered high harm scores across varying offence levels suggest that factors other than offence frequency likely play a significant role in determining harm severity.

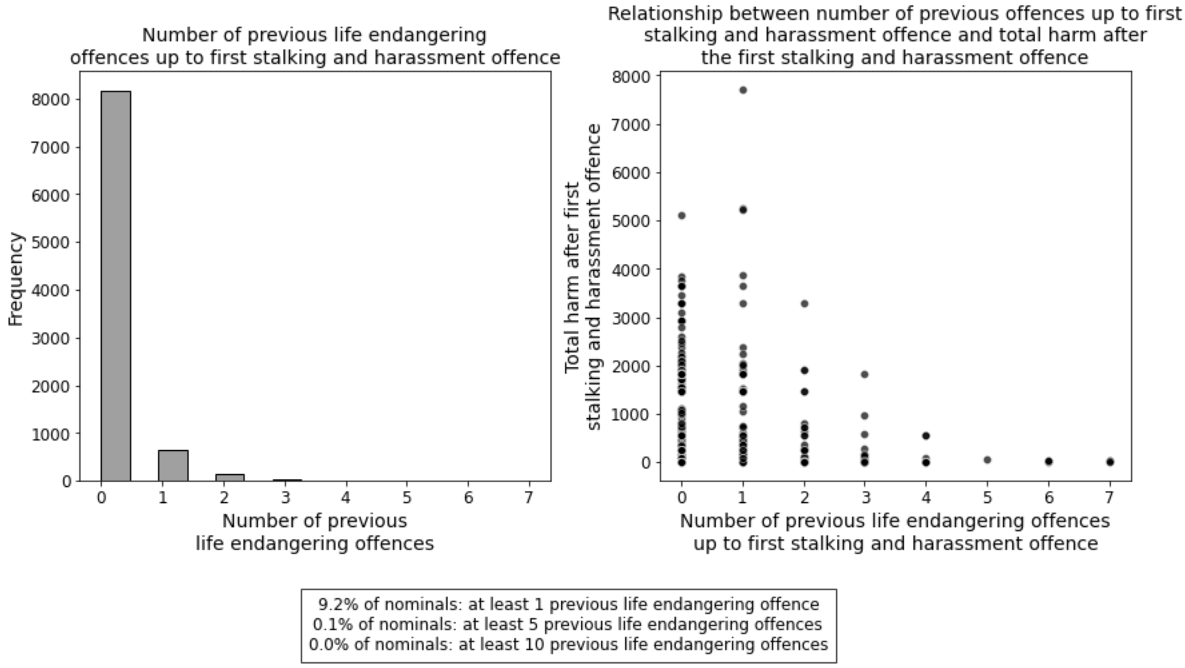
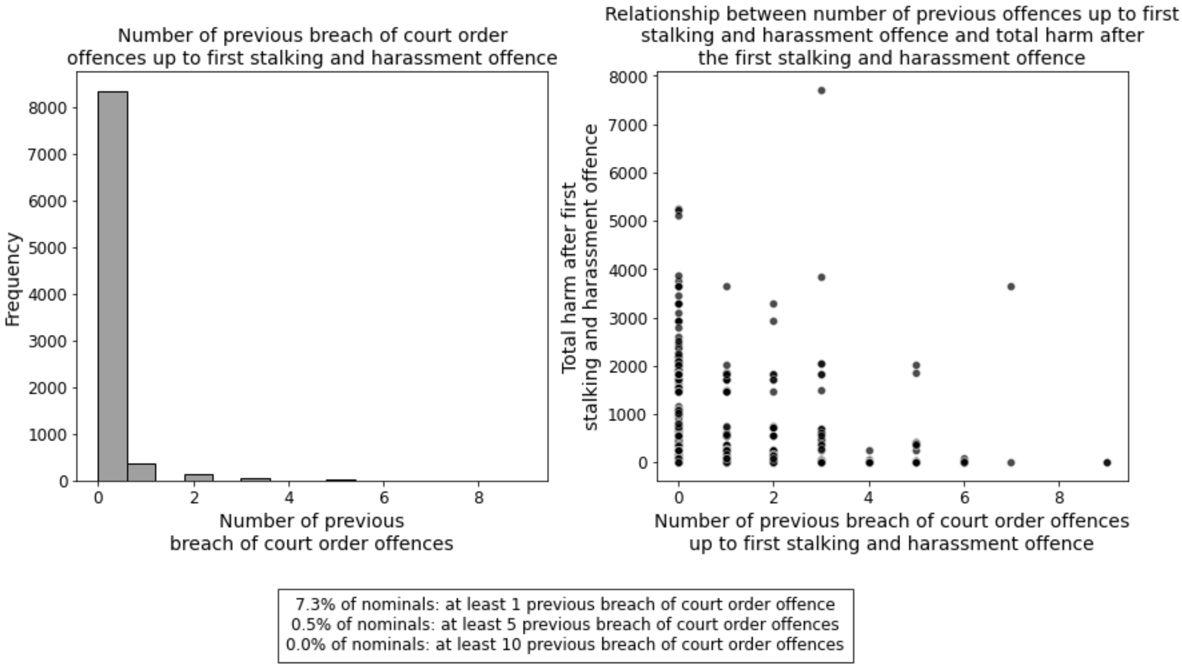


Figure 54. Relationship between coercion, control, and post stalking and harassment harm.

A.8.10 Breach of legal orders and post total harm

Figure 55 explores the relationship between previous breaches of court orders and harm scores after the first stalking and harassment offence. The left subfigure displays a histogram indicating that 7.3% of nominals had at least one breach offence, but only 0.5% accumulated five or more, underscoring their infrequency. The scatter plot in the right subfigure shows a broad distribution of harm scores with no clear trend, as most scores cluster below 2,000 regardless of breach frequency. This suggests that the number of breach offences does not strongly predict harm outcomes, pointing to other influential factors beyond the count of breach offences.



Number of previous breach of court order offences up to first stalking and harassment offence

Relationship between number of previous offences up to first stalking and harassment offence and total harm after the first stalking and harassment offence

7.3% of nominals: at least 1 previous breach of court order offence
0.5% of nominals: at least 5 previous breach of court order offences
0.0% of nominals: at least 10 previous breach of court order offences

Figure 55. Relationship between breach of legal orders and post stalking and harassment harm.

B Harm Scores and Related Offences

The number of unique harm scores, after filtering the data for exploratory data analysis as explained in the previous section, is 30. These scores are listed in **Table 22**.

Table 22. Unique CCHI scores pre and post-stalking and harassment.

1	2	3	4	5	7	10	19	42	84
120	182	182.5	252	270	357	365	547.5	730	912.5
1095	1460	1825	2190	2555	2920	3285	3650	4380	5475

CCHI score of 1. The crime categories largely refer to a range of incidents and investigations that include both criminal and non-criminal elements. They often involve property damage, such as vandalism or attempted damage to buildings, vehicles, or residences, usually valued under £5,000. Additionally, there are numerous cases categorised as non-crimes that still require police intervention, including domestic and child abuse incidents, anti-social behaviour, harassment, and safeguarding disclosures like Clare's Law. The focus is also on preventive measures, referrals, and protective actions, particularly concerning vulnerable individuals, abuse cases, and various forms of exploitation. Overall, the pattern suggests a broad engagement with incidents that may not always be prosecutable as crimes but still necessitate police involvement for public safety, protection, and community support.

CCHI score of 2. The crime categories predominantly involve offences related to theft, property damage, and personal assaults, often with specific aggravating factors such as targeting public service workers or being racially or religiously motivated. These include various forms of theft, ranging from stealing from vehicles and dwellings to unauthorised taking of items like bicycles or mail. Property-related crimes frequently involve damage or threats of damage to residences, vehicles, and businesses, often with significant financial thresholds. Assaults include those on constables and attempts to resist arrest, highlighting a pattern of violence against authorities. Additionally, there are elements of harassment and communication offences intended to cause distress, as well as violations involving control over animals and postal services.

CCHI score of 3. The crime categories primarily involve actions that disrupt personal safety and privacy or interfere with property. They include offences like causing intentional harassment, alarm, or distress, which directly impact individuals' sense of security. Property-related offences, such as interference with or tampering with motor vehicles, reflect actions that compromise the security of personal belongings, specifically vehicles. Additionally, the unauthorised handling of personal data indicates violations of privacy and control over personal information. Overall, these categories focus on offences that cause emotional distress, disrupt privacy, or involve unauthorised access or damage to property.

CCHI score of 4. The only crime associated with this score is bigamy. It is the act of marrying someone while already being legally married to another person.

CCHI score of 5. The crime categories primarily encompass offences involving threats, harassment, breaches of protective orders, and possession of dangerous items, often with aggravating factors like violence or public indecency. They include actions that intimidate or cause distress, such as disclosing private sexual images without consent, harassment that puts people in fear of violence, and breaches of restraining or non-molestation orders. There are also offences related to the possession and threatening use of bladed articles or offensive weapons, both in public and private settings. Additionally, some categories involve property-related offences like unauthorised taking of vehicles and arson that does not endanger life.

CCHI score of 7. The only crime associated with this score is unlawful eviction of an occupier. It refers to the illegal removal or exclusion of a tenant or occupant from a property by a landlord or another party without following the proper legal procedures.

CCHI score of 10. The crime categories cover a broad spectrum of serious offences primarily related to personal safety, property, and public order, with a strong emphasis on violent, sexual, and privacy-related crimes. They include violent acts such as assaults, affray, and threats to kill, as well as property crimes like aggravated vehicle taking and burglary. There are significant sexual offences, including those involving children, such as inciting or engaging in sexual activities and the distribution or threat to share intimate images without consent. The list also includes various breaches of protective orders related to sexual risks, stalking, and anti-social behaviour. Additionally, crimes that exploit or harm vulnerable individuals, such as child cruelty, endangering road users, and encouraging self-harm, are highlighted.

CCHI score of 19. The crime categories primarily focus on offences that involve personal safety, sexual misconduct, and violations of privacy. Many of the listed offences are attempts or actual cases of sexual assault or coercion, often targeting individuals without consent and involving non-penetrative actions. Additionally, there are crimes involving residential burglary, highlighting risks to personal property and security within homes. Some categories address specific duties of care, such as breaches by care providers or workers, which result in neglect or ill-treatment of vulnerable individuals. The list also includes racially or religiously aggravated assaults, reflecting crimes that have discriminatory motivations. Child protection is another significant theme, with offences related to child abduction and possession of indecent images.

CCHI score of 42. The crime categories focus on offences that involve intimidation, threats, or harm directed at individuals involved in the judicial process, such as witnesses, jurors, or those assisting in investigations. These actions are intended to obstruct justice by creating fear or coercing individuals into altering their testimony, not participating, or providing false information.

CCHI score of 84. The crime categories revolve around offences that involve controlling, coercive, or threatening behaviours in close relationships. They include attempts or actual instances of controlling or coercive behaviour within intimate or family relationships, where one party seeks to dominate or manipulate the other, often through emotional or psychological abuse. Harassment and stalking offences involve persistent and intrusive actions that instill fear of violence in the victim, significantly impacting their sense of safety and mental well-being. Additionally, breaching a marriage protection order indicates a failure to adhere to legal measures put in place to protect individuals from harm within the marital context.

CCHI score of 120. The only crime associated with this harm score is “attempting to pervert the course of public justice”. It refers to actions intended to obstruct, interfere with, or undermine the legal process and the administration of justice. This can include acts like providing false information to the authorities, intimidating or bribing witnesses, destroying evidence, or any other behaviour designed to mislead law enforcement, court proceedings, or investigations.

CCHI score of 180. The crime categories primarily involve serious offences related to exploitation, coercion, and violence against individuals, often targeting vulnerable populations. They include crimes like arranging or facilitating travel for exploitation purposes, which are typically linked to human trafficking and modern slavery. Offences involving holding a person in slavery, servitude, or requiring forced labor reflect severe breaches of human rights and personal freedom. There are also violent crimes against specific protected groups, such as assaulting emergency workers and police officers, as well as sexually exploiting individuals with mental disorders or minors under 13 without penetration. Additionally, these categories include threatening behaviours with weapons, highlighting risks to public safety.

CCHI score of 252. The crime categories focus on offences that cause significant physical and psychological harm, often with aggravating factors such as racial or religious motivations. They include stalking that escalates to causing serious alarm or distress, highlighting persistent, invasive behaviours that deeply impact the victim's mental and emotional well-being. The offences related to non-fatal strangulation and suffocation, including attempts and those aggravated by racial or religious bias, are particularly severe as they involve direct physical violence that poses a significant risk to life and safety.

CCHI score of 270. The only offence associated with this score is "causing or allowing a child or vulnerable adult to suffer serious physical harm". It involves actions or omissions that result in significant injury or danger to the well-being of a child or vulnerable adult under one's care or responsibility.

CCHI score of 357. The only offence associated with this score is "racially or religiously aggravated wounding or grievous bodily harm". It refers to severe physical assaults where the perpetrator's actions are motivated by hostility or prejudice based on the victim's race, ethnicity, or religion.

CCHI score of 365. The crime categories encompass a range of serious offences that involve violence, exploitation, and endangerment of life, often targeting vulnerable individuals or valuable properties. They include crimes like arranging or facilitating the sexual exploitation of children aged 13 to 17, highlighting severe breaches of trust and significant harm to minors. There are also violent property-related offences, such as arson and criminal damage that endanger life, as well as assaults and robberies, both personal and business-related, which involve a direct threat or use of force. Crimes like blackmail, controlling prostitution for gain, and various forms of distraction burglary involve deceit and manipulation, often exploiting the victim's trust or vulnerability. Additionally, there are causing serious injury by dangerous driving and offences involving sexual penetration of minors by adults or other minors.

CCHI score of 547. The crime categories involve serious offences related to physical harm, the misuse of personal images, drug production, and threats involving weapons. They include various forms of assault, such as administering poison with intent to injure or annoy, malicious wounding, and inflicting grievous bodily harm (GBH), which reflect deliberate attempts to cause significant injury. Several offences also focus on attempts or actual acts of disclosing private sexual photographs or films without consent, intended to cause distress, highlighting violations of privacy and personal dignity. There are also crimes involving threats with blades or sharply pointed articles on school premises, emphasising risks to public safety, especially in vulnerable environments like schools. Additionally, the production of controlled drugs like cannabis and the creation or distribution of indecent images of children underscore illegal activities that exploit or harm individuals, particularly minors.

CCHI score of 730. The crime categories include serious offences primarily involving violent and sexual crimes, many of which target vulnerable individuals, such as children and victims of aggravated burglaries. Some examples are administering a substance with intent to commit a sexual crime, causing or inciting sexual activity without consent, and sexual assault, with or without penetration, often involving minors or those unable to consent. Aggravated burglaries, which involve the use of weapons or threats during break-ins at residential properties, highlight a significant threat to personal safety and property. There are also severe crimes like manslaughter and causing or allowing the death of a child or vulnerable person, which indicate extreme neglect or violence resulting in death. Breaches of legal protections, such as Sexual Harm Prevention Orders, reflect ongoing risks posed by offenders failing to comply with restrictions meant to safeguard the public.

CCHI score of 912. The only offence with this harm score is "assault police - assault with injury - s.20 - malicious wounding: wounding or inflicting grievous bodily harm" involves a serious assault on a police officer that results in significant injury, such as wounding or grievous bodily harm (GBH).

CCHI score of 1095. The only offence with this score is "causing death by dangerous driving". It refers to the act of operating a vehicle in a manner that falls significantly below the expected standard of a competent and careful driver, resulting in a fatality.

CCHI score of 1460. The crime categories include various forms of kidnapping and false imprisonment, which involve the unlawful taking or restraining of a person, highlighting serious violations of personal liberty and safety. Assaults with intent to cause serious harm, including those involving bodily injury by explosion or wounding with intent to cause grievous bodily harm, represent extreme acts of violence with the deliberate aim to inflict severe physical injury. Other crimes are related to child abduction and sexual assault on minors, especially those involving penetration, and procuring illegal abortion through the administration of drugs.

CCHI score of 1825. The crime categories involve serious offences characterised by extreme violence, the use or possession of weapons, and sexual assault, often with the intent to cause severe harm or fear. They include attempted assaults on emergency workers with intent to cause grievous bodily harm, highlighting the dangers faced by frontline personnel. The offences of attempted rape and rape of individuals aged 16 or over reflect severe sexual violence that causes profound physical and psychological trauma to victims. Other crimes involve the possession of firearms, shotguns, or air weapons, whether real or imitation, with the intent to cause fear, violence, or endanger life, and causing death by aggravated vehicle taking.

CCHI score of 2190: The crime categories involve severe offences that pose significant threats to life, freedom, and personal autonomy, often targeting vulnerable individuals such as children. They include acts of endangering life through attempts to choke or otherwise incapacitate someone in order to commit another serious crime. Crimes related to kidnapping, particularly those linked to forced marriage, involve the unlawful restraint of a person to compel them into a marriage against their will. Other offences are related to child sexual exploitation, such as inciting a male child under 13 to engage in penetrative sexual activity and conducting actions to coerce a child into marriage.

CCHI score of 2555. The crime categories involve severe sexual offences that primarily target minors and vulnerable individuals. They include various forms of rape, specifically detailing offences against females and males in the 13-15 age range, as well as rape of females 16 and over by multiple undefined offenders. The categories also encompass incestuous sexual activity, particularly highlighting penetrative acts committed by family members 18 or over against male children aged 13-17.

CCHI score of 2920. The crime categories involve extremely serious sexual offences targeting highly vulnerable individuals, including young children and those with mental disorders. They encompass attempted and completed rape of female children under 13 by male perpetrators, as well as rape of male children under 13 by males, highlighting the severe exploitation of very young victims. The categories also include offences against individuals with mental disorders that impair their ability to make choices, specifically involving sexual activity with penetration and actions that cause or incite such activity.

CCHI score of 3285. The crime categories involve severe offences that pose direct and deliberate threats to human life. They include attempted murder, which represents a failed but intentional effort to end someone's life, demonstrating a clear intent to kill even though the act was not completed. The category also covers the act of endangering life through the administration of poison, specifically with the intent to put someone's life at risk.

CCHI score of 3650. This only crime with this score is "causing a person with a mental disorder to engage in sexual activity by inducement, threat or deception – penetration". It involves a serious offence targeting individuals with mental disorders, exploiting their vulnerability through manipulative and coercive means.

CCHI score of 4380. The only crime with this score is "intentional destruction of a viable unborn child". It involves the deliberate termination of a viable unborn child's life.

CCHI score of 5475. The crime categories involve severe offences centered around the premeditated taking of human life or the planning thereof. They include conspiracy to murder, which encompasses both agreeing to commit murder and actively soliciting others to carry out a murder.

Based on the crimes listed above, those with a **harm score of 180 (days) or higher** represent the most serious types of offences. This score threshold was also validated by a subject matter expert.

C Architectures for Neural Network Model with Feature Attention, Multi-Head Attention, and Cross-Network Integration

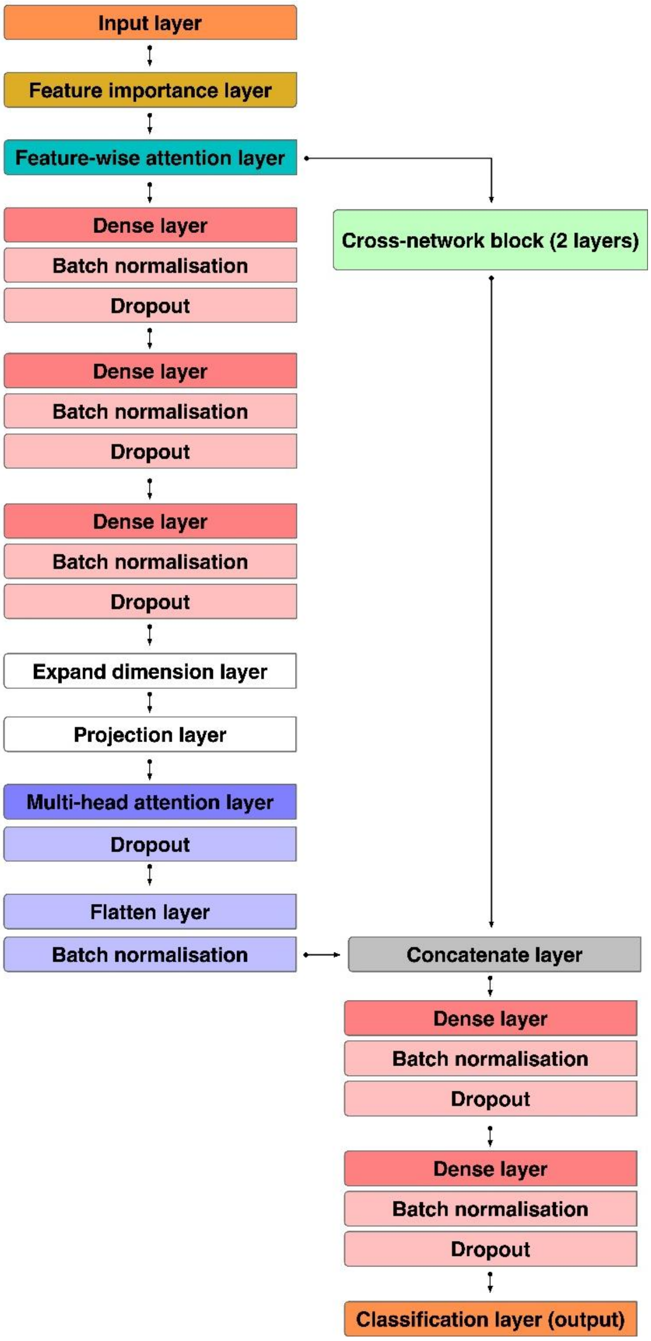


Figure 56. Model architecture for non-sequential inputs.

Both models, with the single and dual inputs, share the following architecture components (shown in **Figure 56**). The model begins with an input layer that takes in a vector of features. The input shape is determined by the number of features in the dataset.

Feature importance layer

This is a custom layer that applies learnable importance weights to each input feature. By identifying and scaling important features, this layer enhances model interpretability and improves its ability to learn from the most relevant data.

Initialisation and regularisation:

- The layer initialises an importance weight for each feature with a value of 1.0, allowing the model to start without bias toward any specific feature. These weights are trainable, meaning they are updated during backpropagation as the model learns.
- To prevent overfitting and encourage sparsity in the weights, L1 regularisation is applied with a strength of 0.01. This regularisation drives some weights to zero, effectively removing the influence of less important features.

Calculation of feature importance:

- During the forward pass, each input feature is multiplied by its corresponding importance weight. This scaling means that features with higher learned importance weights have a larger effect on the model's predictions, while those with weights closer to zero have less impact.
- The multiplication operation ensures that the feature contributions are re-weighted dynamically throughout training, allowing the model to adaptively focus on the most relevant features based on the task at hand.

Feature-wise attention layer

The `FeatureWiseAttention` layer is a custom neural network layer that applies attention weights to each feature in the input. Unlike the `FeatureImportanceLayer`, which assigns static weights to features, the attention mechanism is dynamic and can adjust focus based on the context provided by the input.

Attention weights initialisation and learning

The layer uses a dense neural network to compute attention weights for each feature in the input vector:

- **Weight calculation:** A dense layer with a sigmoid activation function is applied to the input features, generating attention scores ranging between 0 and 1. These scores represent how much "attention" or focus the model should place on each feature.
- **Kernel initialisation**: The kernel of the dense layer is initialised using the `glorot_uniform`[20] initialiser, ensuring weights are balanced initially.

The attention weights are trainable, meaning they are continuously adjusted during the backpropagation process to improve the model's performance on the task.

[20] Weight Initialization Schemes - Xavier (Glorot) and He | Mustafa Murat ARAT (mmuratarat.github.io)

Dynamic calculation of feature attention

For each feature, an attention score is calculated. The attention score indicates the relative importance of the feature within the context of the input sample. The input features are then multiplied by their corresponding attention scores. This allows the layer to scale features dynamically based on their calculated importance for each input sample.

Unlike static feature importance, the attention scores are context-dependent, which means they can vary based on the content of the entire input vector. This allows the model to emphasise different features under different conditions.

Advantages of the feature-wise attention layer

- **Dynamic focus:** The attention mechanism allows the model to adaptively focus on different features for different samples. This dynamic reweighting is especially beneficial for complex datasets where the significance of features varies based on their context or combinations with other features.

- **Improved performance and interpretability:** By calculating attention scores, the model can not only improve performance by focusing on the most relevant features at each step but also provide interpretability, revealing which features were given more focus for a particular prediction.

Cross Network Layer

The `CrossNetwork`[21] layer is designed to learn feature interactions by performing multiplicative combinations of features over multiple layers. This enables the model to capture complex relationships between features in an efficient manner, unlike static feature transformations (e.g., polynomial expansions).

Initialisation and structure

The `CrossNetwork` consists of multiple layers (in this case, 2 layers), each learning to cross features from the input data iteratively. Each layer has its own set of cross weights and cross biases. The cross weights are initialised using the `glorot_uniform` initialiser, which balances the weights, and the biases are initialised as zeros. The weights and biases are learnable parameters, meaning they are updated during training via backpropagation.

Calculation of feature crossings

The `CrossNetwork` learns interactions between features by iteratively performing cross-product transformations:

- **First layer**:
  - The layer starts by taking the original input vector x_0 and an initial transformed vector x, which is set to x_0.

[21] Deep & Cross Network (DCN) | TensorFlow Recommenders

- For each feature in x, an outer product is computed with x_0, effectively creating cross-terms between features. This means that each feature interacts multiplicatively with every other feature in the original input.
- The resulting cross-term is then combined with the learned cross weights and biases, and added to the original feature values.

- **Subsequent layers**:
  - Each layer in the `CrossNetwork` takes the output of the previous layer and performs a similar cross-product transformation with x_0. This iterative process allows the model to learn increasingly complex interactions over the layers.

Unlike traditional polynomial feature expansion (e.g., squaring or multiplying features manually to create higher-order terms), which produces a fixed set of combinations, the `CrossNetwork` learns only the useful feature interactions dynamically during training. Since the weights are trainable, the layer can adapt to the data, creating only the necessary feature combinations. This improves model efficiency and avoids the explosion of feature combinations that often occurs with polynomial feature expansions.

Dense layers with batch normalisation and dropout

Following the cross-feature interaction, the model includes a series of dense layers to further learn complex non-linear relationships. The combination of dense layers, batch normalisation, and dropout helps the model learn robust feature representations while maintaining generalisation to new data.

- **Dense layers:** Three dense layers with 256, 256, and 64 neurons, respectively, using ReLU activation.
- **Batch normalisation[22]:** Applied after each dense layer to stabilise training and reduce internal covariate shift.
- **Dropout[23]:** Each dense layer is followed by a dropout with a rate of 0.3 to prevent overfitting.

Expand Dimensions layer and higher embedding dimension

To enable self-attention across features, an `ExpandDimsLayer` reshapes the output of the dense layers:

- **Purpose:** Reshapes the dense layer output to introduce an additional dimension, preparing it for the multi-head attention mechanism.
- **Projection to embedding:** A dense layer projects the reshaped input to a higher embedding dimension of 32, allowing for a richer representation of features.

[22] *BatchNormalization layer (keras.io)*
[23] *Dropout layer (keras.io)*

Multi-head attention layer

The multi-head attention layer[24] [25] is a mechanism designed to learn contextual relationships between features by attending to different parts of the input data simultaneously. This layer applies multiple "attention heads" to the input, each learning unique patterns and interactions between features, which enhances the model's ability to capture complex dependencies.

Structure and initialisation

- The layer uses multiple attention mechanisms, or "heads", to process the input data. Each head learns to focus on different aspects or combinations of features.
- The use of several heads allows the model to explore multiple perspectives on the data, providing a more nuanced and comprehensive representation of feature interactions.
- In the model used, the attention is distributed across 4 heads. A dropout rate of 0.3 is applied to the attention outputs to ensure regularisation.
- Each attention head has its own set of weights that are initialised using strategies like `glorot_uniform`. These weights are trainable and are updated during backpropagation to learn the optimal attention patterns over the course of training.

Calculation of feature attention

- Each head processes the input in parallel, computing attention scores that reflect the importance of one feature relative to others in the context of the input data.
- These scores are then used to create weighted combinations of the input features, allowing the model to focus on different relationships and patterns in the data.
- Once each head has computed its respective attention output, all the outputs are concatenated together and then projected into a final output through a dense layer. This merging of multiple attention heads provides a more robust and detailed representation of the input features.

Flattening and concatenation of outputs

- **Flattening:** The attention output is flattened to form a one-dimensional vector, enabling compatibility with subsequent dense layers.
- **Batch normalisation:** Applied to the flattened attention output for stable training.
- **Concatenation:** The flattened attention output is concatenated with the cross-network output, combining both cross-feature interactions and attention-enhanced feature representations.

[24] *MultiHeadAttention layer (keras.io)*

[25] *Tutorial 6: Transformers and Multi-Head Attention – UvA DL Notebooks v1.2 documentation (uvadlc-notebooks.readthedocs.io)*

Final dense layers and output

The concatenated output is fed into additional dense layers to refine the learned representations before generating the final classification:

- **Dense layers:** Two dense layers with 128 and 64 neurons, using ReLU activation for non-linear transformations.
- **Batch normalisation and dropout:** Applied to each layer to maintain stable training and reduce overfitting.
- **Output layer:** A final dense layer with 1 neuron and sigmoid activation produces the binary classification output.

The second model, shown in **Figure 57**, is an extension of the original architecture, specifically designed to handle both sequential and non-sequential data through the use of Long Short-Term Memory (LSTM) layers and two separate input channels. This adaptation enables the model to better capture temporal patterns and contextual relationships in the data, particularly where sequences of events are important for predictions.

The adapted model takes in two types of input, handled by separate input layers:

1. **Sequential input (2D)**: This input captures sequential data, represented as a 2D array with varying lengths of sub-arrays (each sub-array represents a series of features over time, such as sequences of events or transactions). The sequential input allows the model to process ordered, context-dependent data, such as the progression of offences or changes over time.
2. **Non-Sequential Input (1D)**: The second input layer takes in flat, non-sequential features as a 1D array. These features are not associated with any temporal sequence and include static characteristics or aggregated information (e.g., number of custodies, number of calls as victim, and binary features extracted with the LLM).

LSTM Layers for sequence processing

The model introduces LSTM layers to handle the sequential data:

- **Masking for variable-length sequences**:
    - A `Masking` layer is applied to handle variable-length sequences within the 2D input. It uses a specified mask value (-9999) to identify and ignore padding within sequences, ensuring the model focuses only on the meaningful portions of the input.
- **LSTM layers**:
    - The sequential data passes through two LSTM layers:
        - **First LSTM layer**: The first LSTM layer has 128 units and is configured to return sequences (`return_sequences=True`). This means the output from this layer contains a sequence of vectors, allowing the model to maintain temporal context across all time steps.
        - **Second LSTM layer**: The second LSTM layer, also with 128 units, processes the output of the first LSTM but is configured to return only

the final output vector (`return_sequences=False`). This reduces the sequence to a single vector representation, capturing the most significant information across the sequence.

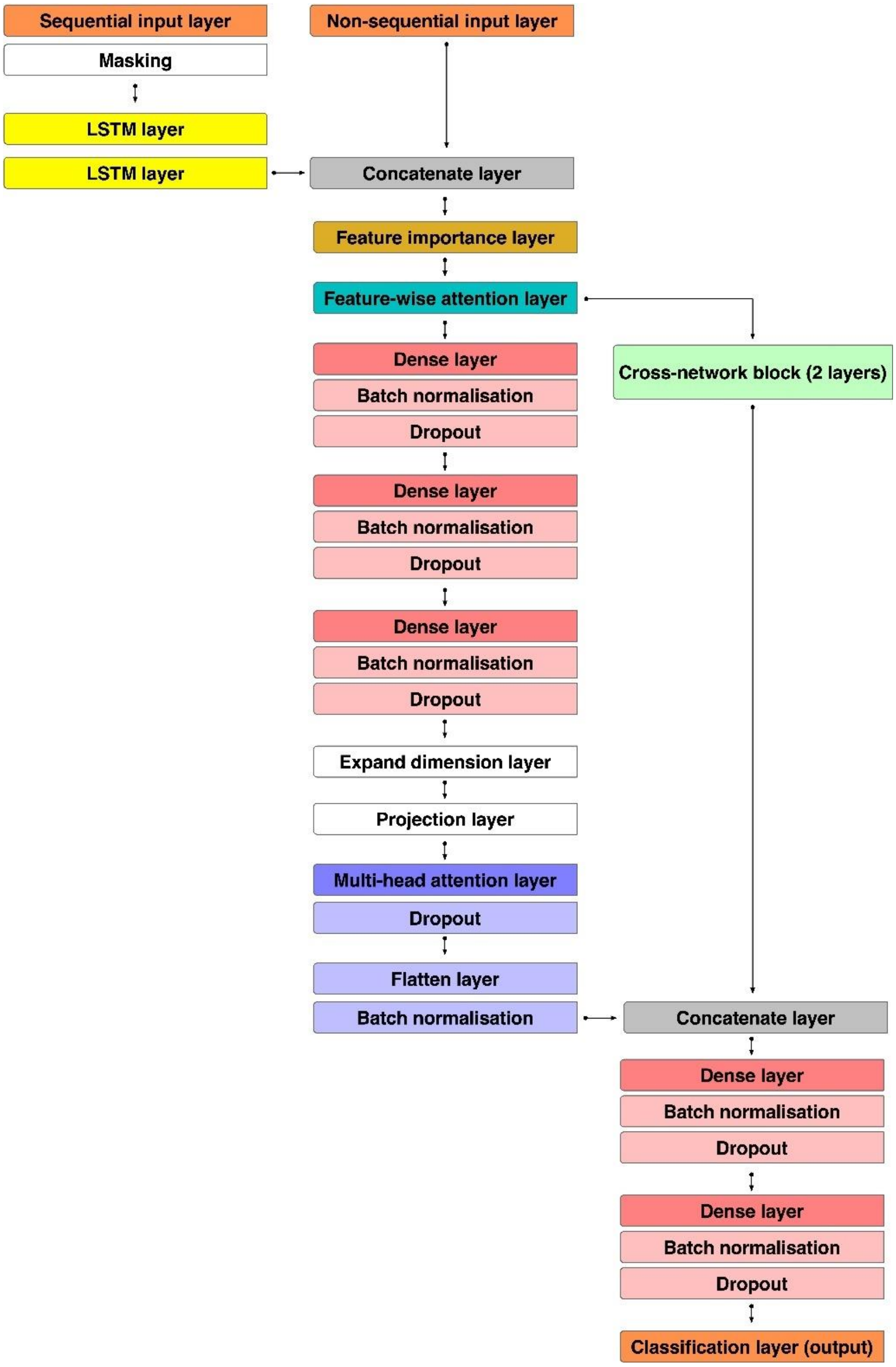


Figure 57. Model architecture for both sequential and non-sequential inputs.

Both the dual and single-input models were trained using the Adam[26] optimiser with a learning rate of 0.001, and the weighted binary cross-entropy[27] was used as the loss metric, with the weight for the positive class being 3.64.

[26] *Adam (keras.io)*

[27] *Use weighted loss function to solve imbalanced data classification problems | by hengtao tantai | Medium*

D Dealing with class imbalance

The dataset is highly imbalanced, containing 7,771 samples from the negative class and only 1,218 samples from the positive class.

For the XGBoost and other ensemble models, this class imbalance was handled using the scale_pos_weight argument. This parameter adjusts the balance of positive and negative classes during training by weighting the positive class more heavily. The value of scale_pos_weight is calculated as the ratio of negative to positive samples, which ensures that the model places more emphasis on correctly predicting the minority class.

For the deep learning models, class imbalance was addressed by using class weights during training. Class weights were calculated to ensure that the minority class receives a higher penalty during model training. Specifically, the class frequencies were first computed to understand the distribution of the classes. Using these frequencies, manual class weights were derived by taking the ratio of the total number of samples to the product of the number of classes and individual class counts. This approach effectively increases the contribution of the minority class to the loss function, making the model more sensitive to its predictions. Additionally, the compute_class_weight function from scikit-learn Python library was used to confirm these weights, ensuring they were balanced appropriately.

Weighted binary cross-entropy is a loss function designed to handle class imbalance by assigning different weights to the classes during training. In standard binary cross-entropy, each sample contributes equally to the loss, which can lead to a bias toward the majority class in imbalanced datasets. The weighted version modifies this by introducing a weight factor for each class, giving higher weight to the minority class and lower weight to the majority class. Specifically, the weight vector is calculated based on the class distribution, and the loss function scales the standard cross-entropy loss for each sample according to its class weight. This ensures that errors made on minority class samples have a greater impact on the loss, guiding the model to focus more on correctly classifying these underrepresented samples.

Alongside the use of class weights, other techniques were explored to address the imbalance. A focal loss function was implemented to focus learning on harder-to-classify samples, and a balanced batch generator was designed to provide an equal representation of both classes in each batch. However, these methods resulted in lower performance compared to using class weights, leading to their exclusion in the final model.

E Appendix Model hyperparameter selection, optimisation, and evaluation of performance metrics

E.1 XGBoost as a baseline model

The dataset is split into training and testing sets using an 80-20 split, meaning 80% of the data is allocated for training the model, while 20% is reserved for validating the model's performance. The split is done with stratification to ensure that the proportion of positive and negative classes is maintained in both the training and validation sets, which is particularly important in datasets with class imbalance. A fixed random seed (`random_state`) is used to make the splitting process reproducible, ensuring consistent results across multiple runs.

Table 23. XGBoost model hyperparameters.

Hyperparameter	Description	Value
n_estimators	The number of boosting rounds or trees to be built in the model. More trees can improve accuracy but may increase overfitting or training time.	200
max_depth	The maximum depth of a tree. Higher values increase the model's capacity to capture complex patterns, but also risk overfitting.	20
learning_rate	The step size or shrinkage factor used in updating the weights after each boosting step. Smaller values make learning slower but may improve performance.	0.1
subsample	The fraction of the training data that is randomly sampled for training each tree. Helps prevent overfitting.	0.8
colsample	The fraction of features (columns) to be randomly sampled for training each tree. Helps with overfitting by reducing correlation between trees.	0.8
gamma	The minimum loss reduction required to make a further partition on a leaf node of the tree. It controls the complexity of the model, with larger values making the algorithm more conservative.	0.1
min_child	The minimum sum of instance weights (or sample count) needed in a child. Higher values prevent the model from learning overly specific patterns to small data subsets.	0.8
scale_pos_weight	Balances the weight of positive and negative classes to address class imbalance. It is calculated as the ratio:	$\frac{Num. of\ negative\ samples}{Num.\ of\ positive\ samples}$
objective	The learning task or objective function. binary:logistic is used for binary classification with logistic regression applied at the output.	binary:logistic
eval_metric	The evaluation metric used to assess the model's performance. auc refers to the Area Under the ROC Curve, which measures the model's ability to distinguish between positive and negative classes.	auc
random_state	A seed value to ensure reproducibility by controlling the randomness in data splitting, sampling, and model training.	42

The dataset is split into training and testingsets using an 80-20 split, meaning 80% of the data is allocated for training the model, while 20% is reserved for validating the model's performance. The split is done with stratification to ensure that the proportion of positive and negative classes is maintained in both the training and validation sets, which is particularly important in datasets with class imbalance. A fixed random seed (`random_state`) is used to make the splitting process reproducible, ensuring consistent results across multiple runs.

The hyperparameters chosen for the XGBoost model, as **outlined in Table 23**, reflect values that are commonly effective and have been shown to work well in various machine learning tasks, even without extensive hyperparameter tuning:

- **n_estimators (200)**: The number of trees in the model is set to 200, which is a reasonable value for achieving good performance without excessively long training times. It strikes a balance between underfitting and overfitting.

- **max_depth (20)**: A higher maximum tree depth allows the model to capture more complex patterns in the data. While 20 is on the high side and could risk overfitting, it is used here to aim that no important patterns are missed.

- **learning_rate (0.1)**: A learning rate of 0.1 is a common default in gradient boosting models. It allows the model to learn gradually, reducing the risk of overshooting optimal solutions while balancing training time and performance.

- **subsample (0.8) and colsample_bytree (0.8)**: Both `subsample` and `colsample` settings are set to 80%, which helps in preventing overfitting by randomly selecting a fraction of the training data and features to build each tree. This makes the model more generalisable.

- **gamma (0.1)**: The gamma value controls the minimum reduction in loss required to make a further split in a tree. A small value like 0.1 ensures that splits are made only when they lead to significant improvements in the model's performance, adding a layer of regularisation.

- **min_child_weight (0.8)**: This parameter prevents the model from learning patterns that are specific to small subsets of data by controlling the minimum sum of instance weights in a leaf. It helps in reducing overfitting, particularly in imbalanced datasets.

- **scale_pos_weight**: This parameter is used to address class imbalance by adjusting the weight of the minority class (positive class in this case). The value is calculated as the ratio of the number of negative to positive samples, ensuring the model pays more attention to the minority class.

As per the results in **Table 24**, three models emerged as the top performers:

- The model incorporating CCHI and derived metrics, male and female flags, age at offence, age < 30[28], number of custodies, number of calls as victim, day sin/cos transforms;

- The model incorporating CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, binary features extracted from incident log keywords;

- The model utilising all available features.

[28] This, coupled with the age at offence will introduce collinearity in the dataset, but this does not affect the method used as it uses recursive partitioning.

Table 24. Classification results for the XGBoost baseline.

Features	ROC AUC	Specificity	Precision Class 0	Recall Class 0	Precision Class 1	Recall Class 1	Weighted Precision	Weighted Recall	Weighted F1 Score	Optimal Threshold	TN FP FN TP
CCHI and derived metrics	0.72	0.8	0.92	0.8	0.29	0.54	0.83	0.76	0.79	0.08	1238 | 316 113 | 131
CCHI and derived metrics, male and female flags	0.73	0.81	0.91	0.81	0.3	0.5	0.83	0.77	0.79	0.1	1265 | 289 121 | 123
CCHI and derived metrics, male and female flags, age at offence, age < 30	0.73	0.96	0.9	0.96	0.53	0.3	0.85	0.87	0.85	0.42	1490 | 64 171 | 73
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies	0.73	0.91	0.91	0.91	0.42	0.42	0.84	0.84	0.84	0.2	1411 | 143 142 | 102
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim	0.73	0.92	0.91	0.92	0.44	0.39	0.84	0.85	0.85	0.22	1431 | 123 149 | 95
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, day sin/cos transforms	0.73	0.97	0.9	0.97	0.6	0.31	0.86	0.88	0.86	0.37	1505 | 49 169 | 75
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, day/month sin/cos transforms	0.73	0.86	0.91	0.86	0.35	0.48	0.84	0.81	0.82	0.11	1331 | 223 126 | 118
Female and male flags, age < 30	0.51	0.07	0.93	0.07	0.14	0.97	0.82	0.19	0.15	0.01	110 | 1444 8 | 236
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, binary features extracted from incident log keywords	0.75	0.94	0.9	0.94	0.51	0.36	0.85	0.87	0.86	0.29	1468 | 86 156 | 88
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, day/month sin/cos transforms, binary features extracted from incident log keywords	0.73	0.93	0.91	0.93	0.47	0.4	0.85	0.86	0.85	0.21	1442 | 112 146 | 98
Binary features extracted from incident log keywords	0.69	0.76	0.91	0.76	0.25	0.51	0.82	0.73	0.76	0.12	1183 | 371 119 | 125
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, binary features extracted from incident logs with the LLM	0.74	0.87	0.91	0.87	0.35	0.44	0.83	0.81	0.82	0.11	1355 | 199 136 | 108
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, day/month sin/cos transforms, binary features extracted from incident logs with the LLM	0.72	0.95	0.9	0.95	0.5	0.3	0.84	0.86	0.85	0.29	1481 | 73 171 | 73
Binary features extracted from incident logs with the LLM	0.61	0.49	0.91	0.49	0.17	0.68	0.81	0.52	0.59	0.07	762 | 792 77 | 167
Binary features extracted from incident logs with the LLM, male and female flags, age < 30	0.56	0.26	0.9	0.26	0.15	0.81	0.8	0.33	0.38	0.01	404 | 1150 46 | 198
All features	0.74	0.94	0.91	0.94	0.5	0.37	0.85	0.86	0.86	0.23	1462 | 92 153 | 91

Table 25. Classification results for the baseline ensemble with soft voting.

Features	ROC AUC	Specificity	Precision Class 0	Recall Class 0	Precision Class 1	Recall Class 1	Weighted Precision	Weighted Recall	Weighted F1 Score	Optimal Threshold	TN FP FN TP
CCHI and derived metrics	0.71	0.83	0.91	0.83	0.31	0.49	0.83	0.79	0.8	0.26	1295 | 259 125 | 119
CCHI and derived metrics, male and female flags	0.7	0.84	0.91	0.84	0.3	0.44	0.82	0.79	0.8	0.27	1304 | 250 136 | 108
CCHI and derived metrics, male and female flags, age at offence, age < 30	0.71	0.81	0.91	0.81	0.29	0.48	0.82	0.77	0.79	0.25	1266 | 288 127 | 117
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies	0.72	0.94	0.9	0.94	0.47	0.36	0.84	0.86	0.85	0.33	1456 | 98 157 | 87
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim	0.72	0.92	0.9	0.92	0.42	0.39	0.84	0.85	0.84	0.31	1426 | 128 150 | 94
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, day sin/cos transforms	0.72	0.94	0.9	0.94	0.48	0.36	0.85	0.86	0.85	0.32	1461 | 93 157 | 87
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, day/month sin/cos transforms	0.73	0.92	0.91	0.92	0.43	0.4	0.84	0.85	0.84	0.3	1423 | 131 147 | 97
Female and male flags, age < 30	0.52	0.06	0.95	0.06	0.14	0.98	0.84	0.18	0.13	0.11	93 | 1461 5 | 239
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, binary features extracted from incident log keywords	0.74	0.96	0.9	0.96	0.55	0.31	0.85	0.87	0.86	0.38	1492 | 62 168 | 76
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, day/month sin/cos transforms, binary features extracted from incident log keywords	0.74	0.91	0.9	0.91	0.4	0.39	0.84	0.84	0.84	0.29	1407 | 147 148 | 96
Binary features extracted from incident log keywords	0.67	0.69	0.91	0.69	0.23	0.57	0.82	0.68	0.73	0.27	1080 | 474 105 | 139
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, binary features extracted from incident logs with the LLM	0.72	0.84	0.91	0.84	0.31	0.47	0.83	0.79	0.81	0.24	1305 | 249 130 | 114
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, day/month sin/cos transforms, binary features extracted from incident logs with the LLM	0.72	0.89	0.9	0.89	0.37	0.39	0.83	0.83	0.83	0.27	1390 | 164 148 | 96
Binary features extracted from incident logs with the LLM	0.63	0.66	0.9	0.66	0.2	0.54	0.81	0.64	0.7	0.27	1025 | 529 112 | 132
Binary features extracted from incident logs with the LLM, male and female flags, age < 30	0.6	0.56	0.9	0.56	0.18	0.59	0.8	0.57	0.63	0.21	873 | 681 99 | 145
All features	0.73	0.79	0.92	0.79	0.29	0.55	0.83	0.75	0.78	0.22	1223 | 331 111 | 133

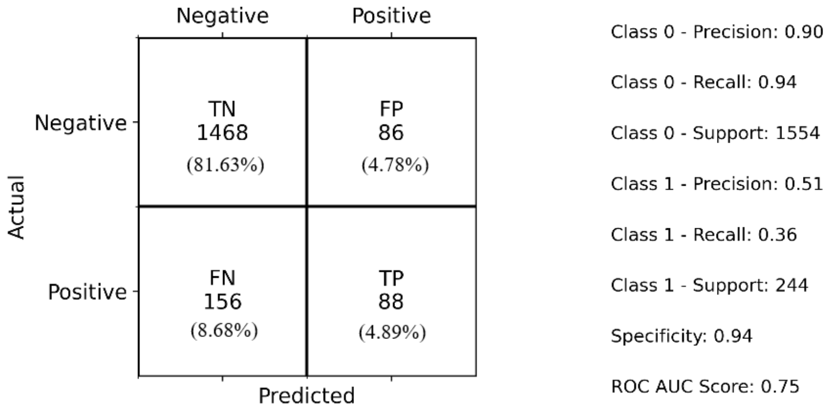
Table 26. Classification results for the baseline ensemble with a meta-learner in a stacked architecture

Features	ROC AUC	Specificity	Precision Class 0	Recall Class 0	Precision Class 1	Recall Class 1	Weighted Precision	Weighted Recall	Weighted F1 Score	Optimal Threshold	TN FP FN TP
CCHI and derived metrics	0.7	0.94	0.89	0.94	0.43	0.29	0.83	0.85	0.84	0.01	1461 | 93 174 | 70
CCHI and derived metrics, male and female flags	0.65	0.93	0.9	0.93	0.42	0.32	0.83	0.85	0.84	0.01	1449 | 105 167 | 77
CCHI and derived metrics, male and female flags, age at offence, age < 30	0.63	0.94	0.9	0.94	0.44	0.32	0.83	0.85	0.84	0.01	1454 | 100 167 | 77
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies	0.66	0.96	0.9	0.96	0.56	0.32	0.85	0.87	0.86	0.05	1493 | 61 167 | 77
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim	0.7	0.96	0.9	0.96	0.56	0.33	0.85	0.87	0.86	0.07	1490 | 64 164 | 80
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, day sin/cos transforms	0.67	0.96	0.9	0.96	0.54	0.3	0.85	0.87	0.85	0.01	1491 | 63 170 | 74
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, day/month sin/cos transforms	0.65	0.96	0.9	0.96	0.55	0.32	0.85	0.87	0.86	0.02	1491 | 63 166 | 78
Female and male flags, age < 30	0.49	0	0	0	0.14	1	0.02	0.14	0.03	0.0	0 | 1554 0 | 244
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, binary features extracted from incident log keywords	0.67	0.94	0.9	0.94	0.48	0.36	0.85	0.86	0.85	0.01	1458 | 96 156 | 88
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, day/month sin/cos transforms, binary features extracted from incident log keywords	0.63	0.97	0.9	0.97	0.6	0.31	0.86	0.88	0.86	0.1	1504 | 50 169 | 75
Binary features extracted from incident log keywords	0.57	0.92	0.89	0.92	0.34	0.27	0.81	0.83	0.82	0.01	1423 | 131 178 | 66
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, binary features extracted from incident logs with the LLM	0.64	0.95	0.9	0.95	0.51	0.32	0.85	0.87	0.85	0.01	1480 | 74 167 | 77
CCHI and derived metrics, male and female flags, age at offence, age < 30, number of custodies, number of calls as victim, day/month sin/cos transforms, binary features extracted from incident logs with the LLM	0.65	0.95	0.9	0.95	0.51	0.31	0.84	0.87	0.85	0.01	1481 | 73 169 | 75
Binary features extracted from incident logs with the LLM	0.58	0.77	0.88	0.77	0.2	0.36	0.79	0.71	0.74	0.04	1191 | 363 156 | 88
Binary features extracted from incident logs with the LLM, male and female flags, age < 30	0.5	0	0	0	0.14	1	0.02	0.14	0.03	0.00	0 | 1554 0 | 244
All features	0.71	0.96	0.9	0.96	0.56	0.3	0.85	0.87	0.86	0.02	1496 | 58 170 | 74

E.2 XGBoost baseline error analysis

The confusion matrix in **Figure 58** presents the performance of the XGBoost baseline classifier trained to predict whether nominals are likely to commit a serious crime in the 12 months following instances of stalking and harassment. The test dataset consists of 1,544 nominals who did not commit a serious crime within 12 months following incidents of stalking and harassment, and 244 nominals who did. The confusion matrix provides a breakdown of the model's true and false predictions as follows:

1. **True Negatives (TN)**: These are cases where the model correctly predicted that a nominal would not go on to commit a serious crime in the next 12 months. In the matrix, there are **1,468 true negatives**, indicating that the model was quite effective at identifying those who were not likely to commit a serious crime.
2. **False Positives (FP)**: These represent instances where the model incorrectly predicted that a nominal would commit a serious crime when, in reality, they did not. The matrix shows **86 false positives**, meaning the model incorrectly flagged these individuals as high risk when they were not.
3. **False Negatives (FN)**: These are cases where the model failed to identify nominals who went on to commit a serious crime, incorrectly predicting that they would not. The matrix indicates **156 false negatives**, which highlights the model's failure to catch individuals who ultimately posed a high risk.
4. **True Positives (TP)**: These are instances where the model correctly predicted that a nominal would commit a serious crime. The matrix displays **88 true positives**, showing the cases where the model successfully identified high-risk individuals.
5. **Specificity**: Indicates that the model either overfit to the negative class or the balancing approach was not as effective as expected.



	Negative	Positive
Negative	TN 1468 (81.63%)	FP 86 (4.78%)
Positive	FN 156 (8.68%)	TP 88 (4.89%)

Class 0 - Precision: 0.90

Class 0 - Recall: 0.94

Class 0 - Support: 1554

Class 1 - Precision: 0.51

Class 1 - Recall: 0.36

Class 1 - Support: 244

Specificity: 0.94

ROC AUC Score: 0.75

Figure 58. Confusion matrix for the XGBoost baseline model

E.2.1 False positives

The first subplot in **Figure 59** illustrates the sorted pre-stalking and harassment total harm scores compared to the post-stalking and harassment total harm scores. The second subplot displays the cumulative percentage of nominals based on their offending history, up to and including stalking and harassment incidents, with specific counts of offenses. Both plots highlight a clear trend: more than half of the false positives demonstrate a significant escalation in harm prior to the stalking and harassment incident. In the second subplot, it is also evident that over 50% of the nominals had committed at least five offenses prior to the stalking and harassment incident, offering sufficient context for the model to learn from.

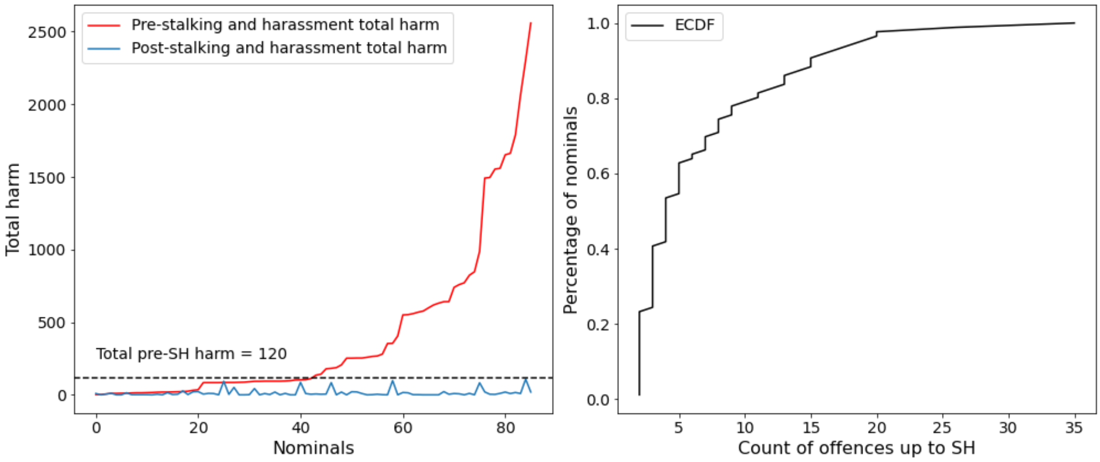


Figure 59. Total harm pre- and post-stalking and harassment and count of offences pre-stalking and harassment for false positives for the XGBoost mode.

The escalation in harm is highly elevated for a large subset of nominals, which explains why the model might have classified them as false positives. However, it is less clear for those nominals whose harm escalation was more moderate. **Figure 60** further shows that many nominals exhibited multiple crime behaviours identified through both keyword-based and LLM-extracted indicators from incident logs. For nominals with pre-stalking and harassment harm scores lower than 120, the number of custodies is notably frequent – nearly all had at least one custodial offense, and 60% of them had at least 10 custodies.

Despite this, it remains challenging to determine which specific feature combinations led to these false positives, as no clear linear relationships emerged from the exploratory data analysis. This makes it difficult to understand why the model misclassified these cases as false positives, given that both the harm escalation and the presence of crime indicators would have suggested a higher risk. **These false positives, however, would likely have been classified similarly by a human evaluator, given the clear escalation in harm leading up to the stalking and harassment offense.**

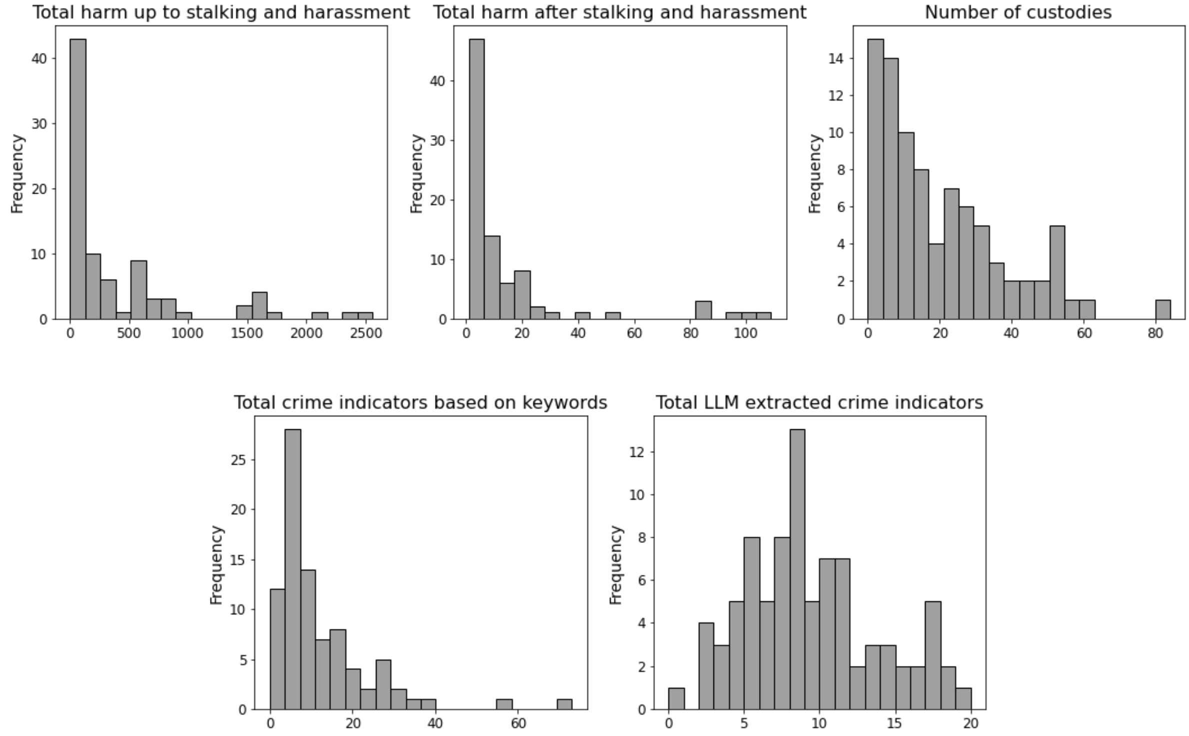


Figure 60. Distribution of key features associated with false positives for the XGBoost model.

E.2.2 False negatives

For the false negatives, **Figure 61** shows that most nominals showed little to no escalation in harm before the stalking and harassment incident, whereas the harm escalated significantly afterward, with frequent peaks reaching extreme levels. The small number of offenses or very low-harm offenses (e.g., non-crime domestic abuse, which carries a harm score of 1) may not have provided enough indication of a continued escalation post-stalking and harassment. For instance, while 55% of nominals with pre-stalking and harassment harm scores below 120 had 10 or more custodial events, these factors did not appear to be discriminative enough for the model to correctly classify them as true positives.

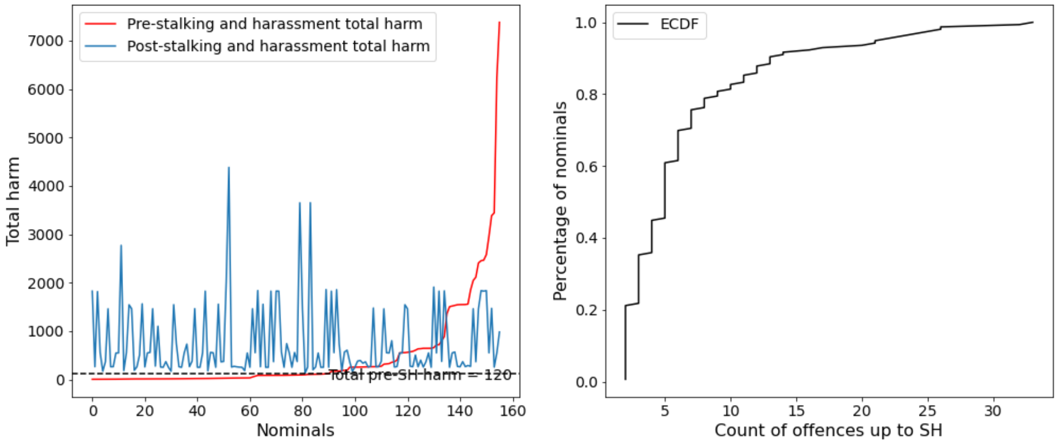


Figure 61. Total harm pre- and post-stalking and harassment and count of offences pre-stalking and harassment for false negatives for the XGBoost model.

As shown in **Figure 62**, the distribution of the number of custodies and crime indicators – whether based on keywords or extracted by the LLM from incident logs – does not differ significantly from those associated with false positive instances. The key issue seems to be the minimal or nonexistent escalation in harm prior to the stalking and harassment incident, which likely reduced the model's ability to discriminate effectively, even in the presence of stronger indicators like criminal behaviour flags and the number of custodies. Despite the presence of these features, many of these instances would likely have been similarly assessed as low risk by a human evaluator, given the limited offending history before the stalking and harassment offense.

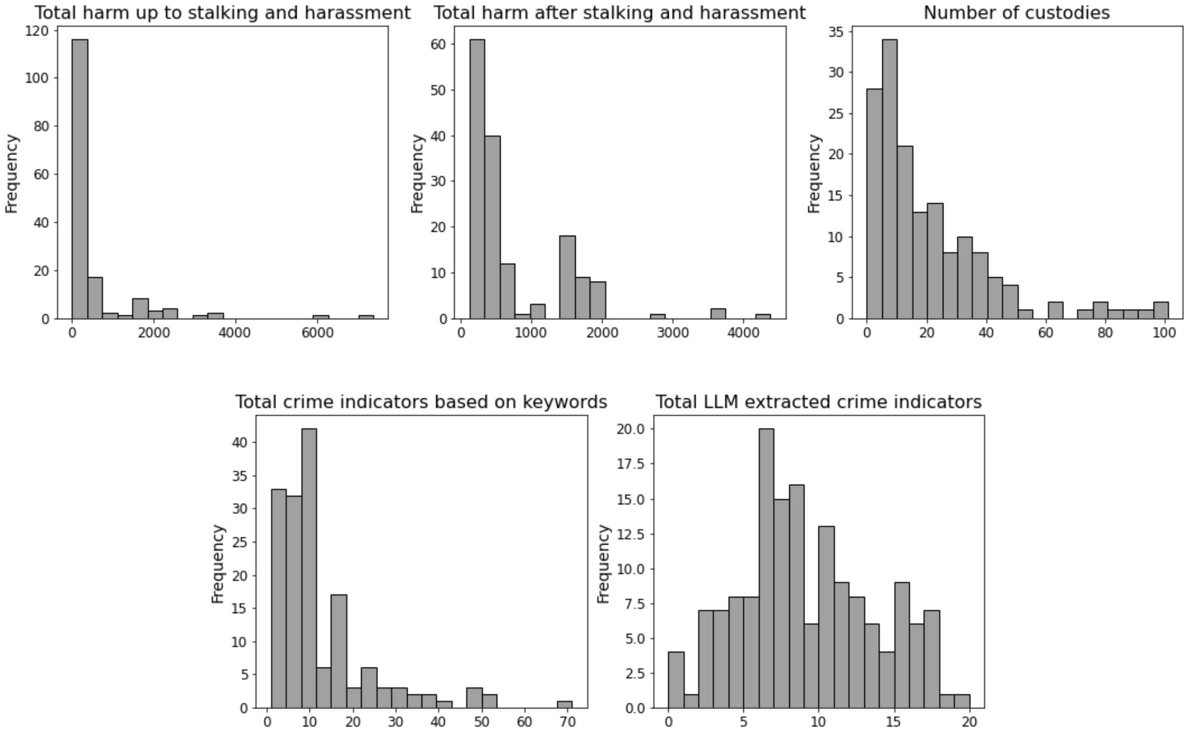


Figure 62. Distribution of key features associated with false negatives for the XGBoost model.

E.2.3 Error analysis for false positives

The first subplot in **Figure 63** shows the sorted total harm scores before stalking and harassment incidents compared to those after. The second subplot shows the cumulative percentage of individuals based on their offense history, including stalking and harassment incidents, with a breakdown of the number of offenses. Both plots reveal a noticeable trend: approximately half of the false positives exhibit a significant increase in harm before the stalking and harassment incident. Additionally, the second subplot indicates that more than 50% of individuals had committed at least five offenses prior to the incident, providing ample context for the model to learn from.

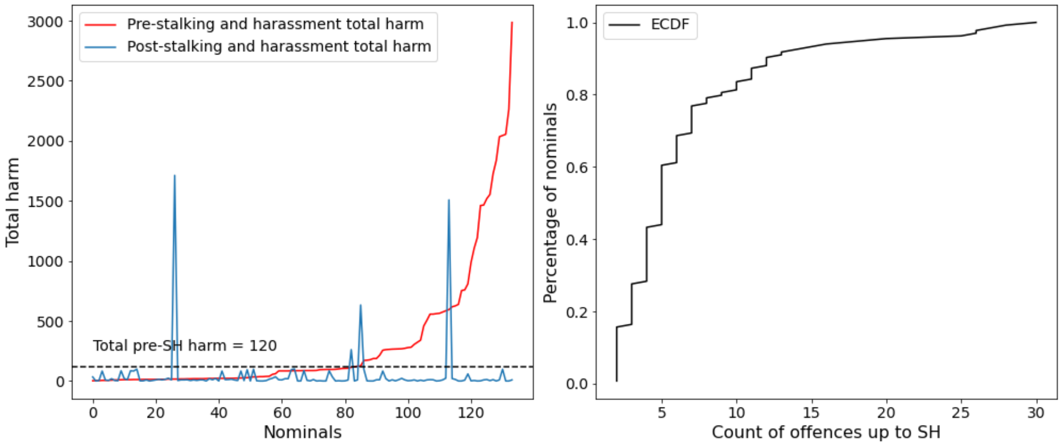


Figure 63. Total harm pre- and post-stalking and harassment and count of offences pre-stalking and harassment for false positives for the deep learning model.

The harm escalation is significantly higher for a large subset of nominals, which may explain why the model classified them as false positives. However, as seen in the error analysis in **Subsection 6.5.1**, the reasoning is less clear for those nominals whose harm escalation was more moderate. Additionally, there are two distinct spikes in harm after stalking and harassment (represented in blue above the red curve), which the model likely could not have classified as true positives. This is due to the fact that the pre-stalking and harassment harm levels were minimal (10 or under), while the post-stalking and harassment harm drastically increased (exceeding 600). These spikes represent unexpected or outlier events that would be difficult to predict, even for humans.

Figure 64 further demonstrates that many nominals displayed multiple crime behaviours, as identified through both keyword-based and LLM-extracted indicators from incident logs. Among nominals with pre-stalking and harassment harm scores below 120, custodial offenses were particularly prevalent – nearly all had at least one instance of custody, and 76% had been in custody at least 10 times.

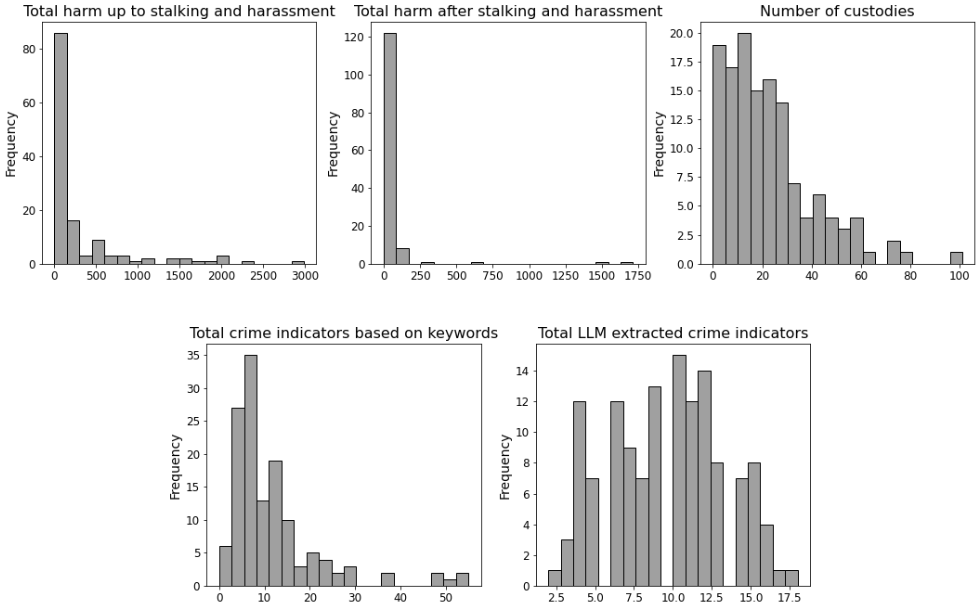


Figure 64. Distribution of key features associated with false positives for the deep learning model.

Similar to the previous error analysis of the XGBoost model, harm escalation observed before stalking and harassment does not appear to continue afterward. Almost all total harm scores remain very low, with only a few large outliers. It is evident that there are escalation patterns prior to stalking and harassment that are similar but lead to different outcomes, creating confusion for the model in predicting the correct classification. **Figure 63** illustrates this issue, showing instances where the model predicted low-risk harm before stalking and harassment as positives, despite there being minimal prior escalation that did not indicate further harm. This pattern was also seen in the XGBoost model, suggesting that low harm before stalking and harassment can result in both high and low harm outcomes afterward.

Furthermore, the binary features derived from keywords and processed by the LLM from incident logs did not exhibit significant discriminative power, making it challenging for the model to learn the distinction between inputs leading to different outcomes. This issue is similar to training an image classification model on pictures of dogs but inconsistently labelling them as “dog” or “cat”, leaving the model unable to learn any concrete patterns. The error analyses support this observation, confirming that the model struggles to differentiate cases, a finding that is also consistent with manual data inspection.

E.2.4 Error analysis for false negatives

The analysis of false negatives in **Figure 65** reveals a distinct pattern of harm escalation. Before the stalking and harassment incidents, most nominals demonstrated little to no increase in harm, but after these incidents, there was a significant escalation with frequent and extreme peaks. This mirrors the findings from the error analysis of the XGBoost model, suggesting that the model faced challenges in predicting certain cases accurately.

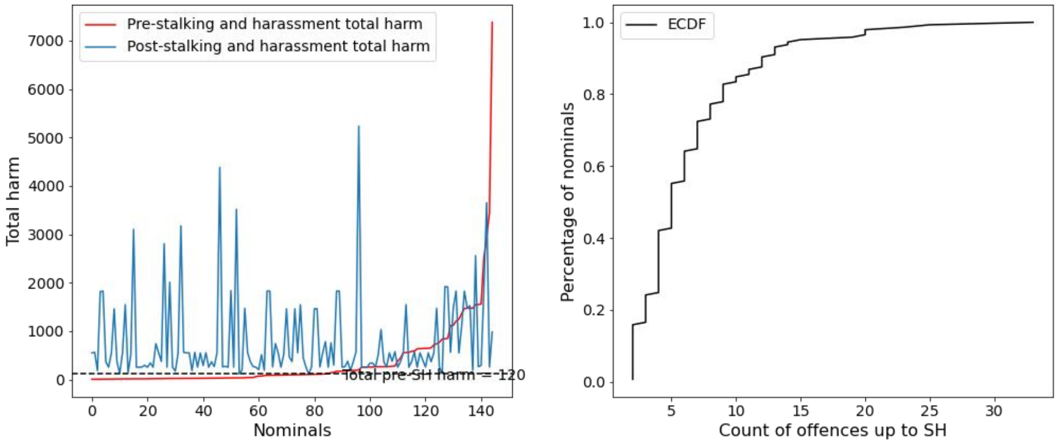


Figure 65. Total harm pre- and post-stalking and harassment and count of offences pre-stalking and harassment for false negatives for the deep learning model.

One potential issue is that a small number of offenses or offenses with very low harm (like non-crime domestic abuse with a harm score of 1) did not serve as strong indicators of a trend toward higher harm following stalking and harassment. Consequently, these low-level offenses may have contributed to the model’s inability to recognise the potential for future harm and correctly classify these cases as true positives.

An example highlighting this is the group of nominals whose pre-stalking and harassment harm scores were below 120. Despite the fact that 64% of these individuals experienced 10 or more custodial events, these occurrences did not provide enough discriminative power to aid the model in making accurate predictions. The complexity of harm escalation patterns and their variability across cases makes it challenging for the model to distinguish which individuals are likely to experience continued or heightened harm post-stalking and harassment.

Figure 66 reveals that the distribution of key features, such as the number of custodies and crime indicators (whether identified through keywords or extracted by an LLM from incident logs), shows little difference between the false positives and other instances, which is similar to the pattern seen in the false negatives from the XGBoost model. A significant issue appears to be the lack of escalation in harm before the stalking and harassment incident, limiting the model's ability to effectively discriminate these cases, even when stronger indicators like criminal behaviour flags and the number of custodies are present.

This minimal or absent harm escalation prior to the incident likely makes it difficult for the model to distinguish between low- and high-risk cases. Consequently, despite the presence of features typically associated with risk, these instances would likely have been similarly assessed as low risk by a human evaluator, due to the limited offending history before the stalking and harassment offense. This highlights the challenge in predicting escalation in cases where the pre-incident behaviour does not signal significant risk.

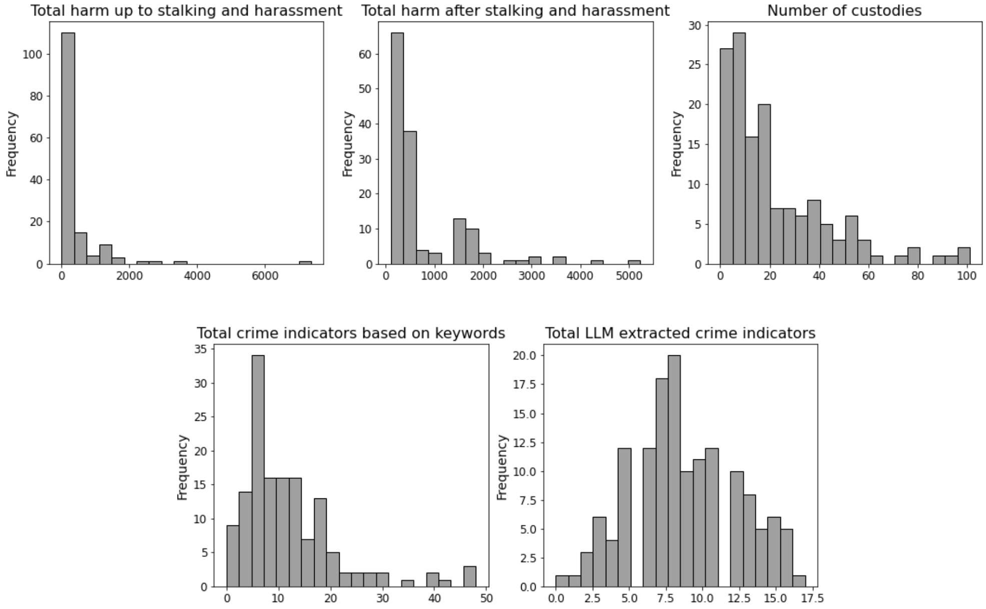


Figure 66. Distribution of key features associated with false negatives for the deep learning model.

F LLM Feature Extraction Template

```
## 1. Relationship with Victim
- **Current relationship type**:
  - *Partner* (Yes/No)
  - *Ex-partner* (Yes/No)
  - *Family member* (Yes/No)
  - *Friend/Acquaintance* (Yes/No)
  - *Other* (Yes/No)
- **Shared child custody present** (Yes/No)

## 2. Threats
- **Explicit threats** (Yes/No)

## 3. Physical Violence
- **Physical assault or beating** (Yes/No)
- **Sexual assault** (Yes/No)
- **Strangulation/Choking** (Yes/No)
- **Use of weapons** (Yes/No)
- **Serious bodily harm (GBH/ABH)** (Yes/No)
- **Evidence of persistent violent pattern** (Yes/No)

## 4. Coercive and Controlling Behaviour
- **Coercive control** (Yes/No)
- **Retaliation for leaving** (Yes/No)
- **Evidence of wanting to regain control** (Yes/No)
- **Spatially confining or restraining victim** (Yes/No)
- **Evidence of financial, social, or physical control** (Yes/No)

## 5. Stalking and Surveillance
- **Following and surveillance behaviours** (Yes/No)
- **Use of tracking devices** (Yes/No)
- **Unwanted contact and communications** (Yes/No)
- **Unwanted intrusions** (Yes/No)

## 6. Property Damage
- **Vandalism or arson** (Yes/No)

## 7. Psychological Abuse
- **Victim intimidation** (Yes/No)
- **Spreading false rumors** (Yes/No)
- **Blackmailing** (Yes/No)
- **Impersonating victim** (Yes/No)
- **Evidence of fixation/obsession** (Yes/No)
- **Public confrontation/arguments** (Yes/No)

## 8. Legal and Procedural
- **Court order issued** (Yes/No)
- **Breach of legal orders** (Yes/No)

## 9. Opportunistic Factors
- **Accessibility to the victim** (Yes/No)
- **Gathering personal information** (Yes/No)
- **Increased frequency/intensity of pursuit** (Yes/No)
- **Standing/littering around victim's home/school/work** (Yes/No)

## 10. Other Risk Factors
- **Under age 30** (Yes/No)
- **Substance or alcohol abuse** (Yes/No)
- **Documented mental health issues** (Yes/No)
- **Evidence of social instability** (Yes/No)
- **Victim fear** (Yes/No)
- **History of domestic violence** (Yes/No)
- **Presence of step-child** (Yes/No)
- **Victim's pregnancy** (Yes/No)
- **Separation from the perpetrator** (Yes/No)
```

14 References

Bendlin, M., & Sheridan, L. (2019). Nonfatal strangulation in a sample of domestically violent stalkers: The importance of recognizing coercively controlling behaviors. *Criminal justice and behavior, 46*(11), 1528-1541.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*(16), 321-357.

Eke, A., Hilton, N., Meloy, J., Mohandie, K., & Williams, J. (2011). Predictors of recidivism by stalkers: A nine-year follow-up of police contacts. *Behavioral Sciences & the Law, 29*(2), 271-283.

Ferguson, C., & Freya, M. (2023). Continuing coercive control after intimate partner femicide: The role of detection avoidance and concealment. *Feminist Criminology, 18*(4), 353-375.

Garcia-Vergara, E., Almeda, N., Martin Rios, B., Becerra-Alonso, D., & Fernandez-Navarro, F. (2022). A comprehensive analysis of factors associated with intimate partner femicide: a systematic review. *International journal of environmental research and public health, 19*(12), 7336.

James, D., & Farnham, F. (2003). Stalking and serious violence. *Journal of the American Academy of Psychiatry and the Law Online, 31*(4), 432-439.

Johnson, E., & Thompson, C. (2016). Factors associated with stalking persistence. *Psychology, Crime & Law, 22*(9), 879-902.

Johnson, H., Li, E., Paul, M., & Richard, W. (2019). Intimate femicide: The role of coercive control. *Feminist Criminology, 14*(1), 3-23.

Katz, E., Nikupeteri, A., & Laitinen, M. (2020). When coercive control continues to harm children: Post-separation fathering, stalking and domestic violence. *Child abuse review, 29*(4), 310-324.

McEwan, T., Daffern, M., MacKenzie, R., & Ogloff, J. (2017). Risk factors for stalking violence, persistence, and recurrence. *The Journal of Forensic Psychiatry & Psychology, 28*(1), 38-56.

McEwan, T., Mullen, P., & MacKenzie, R. (2009). "A study of the predictors of persistence in stalking situations. *Law and human behavior*(33), 149-158.

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

Meloy, J. (1997). The clinical risk management of stalking:"Someone is watching over me…". *American journal of psychotherapy, 51.2*, 174-184.

Meloy, J. (1999). Stalking: An old behavior, a new crime. *Psychiatric Clinics of North America 22*(1), 85-99.

Messing, J., Patch, M., Sullivan Wilson, J., Kelen, G., & Campbell, J. (2018). Differentiating among attempted, completed, and multiple nonfatal strangulation in women experiencing intimate partner violence. *Women's health issues, 28*(1), 104-111.

Mullen, P., Mackenzie, R., Ogloff, J., Pathé, M., McEwan, T., & Rosemary, P. (2006). Assessing and managing the risks in the stalking situation. *Journal of the American Academy of Psychiatry and the Law Online, 4*(39), 439-540.

Nobles, M., Cramer, R., Zottola, S., Desmarais, S., Gemberling, T., Holley, S., & Wright, S. (2018). Prevalence rates, reporting, and psychosocial correlates of stalking victimization: results from a three-sample cross-sectional study. *Social Psychiatry and Psychiatric Epidemiology, 53*, 1253-1263.

Protection from Harassment Act 1997. (n.d.). Retrieved from https://www.legislation.gov.uk/ukpga/1997/40/contents

RANDALL, K., & Cook, A. (2014). Intimate partner violence, stalking, and femicide. In *International handbook of threat assessment* (pp. 178-194).

Ratih, I., Retnaningsih, S., Islahulhaq, I., & Dewi, V. (2022). Synthetic minority over-sampling technique nominal continous logistic regression for imbalanced data. *AIP Conference Proceedings, 2668*(1).

Ratih, I., Retnaningsih, S., Islahulhaq, I., & Dewi, V. (2022). Synthetic minority over-sampling technique nominal continous logistic regression for imbalanced data. *AIP Conference Proceedings, 2668*(1).

Resnick, P. (2007). Stalking: Psychiatric perspectives and practical approaches. In *Stalking risk assessment.* (pp. 61-84). Oxford University Press.

Sheed, A., Cleo, B., & Troy E., M. (2024). The Relationship Between Stalking, Homicide, and Coercive Control in an Australian Population. *Homicide Studies*.

Sheridan, L., & Roberts, K. (2011). Key questions to consider in stalking cases. *Behavioral sciences & the law, 29*(2), 255-270.

Storey, J., Afroditi, P., & Williams, C. (2023). The impact of stalking and its predictors: characterizing the needs of stalking victims. *Journal of interpersonal violence, 38*(21-22), 11569-11594.

T. K, L., & Walker, R. (2010). Toward a deeper understanding of the harms caused by partner stalking. *Violence & Victims, 25*(4).

Tyson, D. (2020). Coercive control and intimate partner homicide. *Criminalising coercive control: Family violence and the criminal law*, 73-90.