Crime Watch: Final Report

CEO

Director of Intelligence, FBI

Donald Wedding

CrimeWatch Consultants:

Damon Panahi

Jeremy Melville

Jin Choi

Setu Madhavi Namburu

Sheila Cloudcroft

Key Findings and Takeaways

SURVEY OPINIONS

Survey Data is based on opinions. Victims respond only if they feel safe, so this survey sample may not be representative of the real population experience.

ETHNICITY AND RACE

Ethnicity and race is a key factor when reporting to police.

THEFT

Largest proportion of crime types in this survey sample was theft.

NCVS DATA

Rich data available from raw data and other sources for deeper and more accurate insights as next step.

VICTIM SERVICES

In the survey sample when Victim Services are involved there is a much higher likelihood of reporting to police.

RAPE/SEXUAL ASSAULT

Underreporting of rape and sexual assault a big problem, especially for those ages between 12 and 14.

DEMOGRAPHICS

Never married, between ages of 12-24, and non-Hispanic are less likely to report crimes to police in this survey sample

TABLE OF CONTENTS

Overview

The Director of Intelligence at the Federal Bureau of Investigations (FBI) contracted with CrimeWatch to help them understand the frequency of and patterns behind reported and unreported incidents of crime. It is of concern to all law enforcement agencies that any crimes in the United States go unreported. The FBI also seeks to understand key drivers behind reporting and crime so the agency can develop more effective programs and campaigns to fight crime as well as to direct resources towards appropriate law enforcement and victim support agencies and programs. It is anticipated that law enforcement will improve the American people's confidence in their safety and security through achievement of these goals.

CrimeWatch is an experienced analytics consulting firm that has worked previously with the FBI, Bureau of Justice Statistics (BJS), and Criminal Justice Information divisions delivering solid, actionable intelligence via data science/AI tools. CrimeWatch funding is limited to a pilot project to achieve visibility into those issues and uses the Bureau of Justice Statistics' RESTful API National Crime Victimization Survey (NCVS) data.

Business Case

The Federal Bureau of Investigation's (FBI) criminal justice information division publishes different types of crime reports annually . The message from the Director of Intelligence indicates:

> *"As a nation, we continue to face an evolving crime landscape. To stay ahead of threats, and keep people safe, we need a clear and complete picture of what's going on in our communities. We need greater transparency and accountability in policing. The National Incident-Based Reporting System (NIBRS) will help make this happen."*[1]

According to BJS, the total number of estimated personal and household victimizations has fallen each year since 1992, except for 1993, despite increases in the U.S. population[2]. Law enforcements' ultimate goal is to reduce criminal activity, the fear of crime, and victimization in specific places and neighborhoods. To achieve this people must feel safe about reporting crime to police. The Director of FBI would like to understand if predictive analytics can reveal patterns and clusters regarding unreported crimes that ordinary evaluation of report data has not revealed.

A small data science team within CrimeWatch will conduct a pilot study using NCVS data and identify patterns and key drivers behind reported and unreported crimes. The team will build predictive models and interactive dashboards with mobile applications to identify regions and sub-groups where crimes are not reported so targeted campaigns or safety measures can be launched by the FBI.

[1] Department of Justice, Federal Bureau of Investigation, 2018 Crime in the United States, Message from the Director, Retrieved from: https://ucr.fbi.gov/crime-in-the-u.s/2018/crime-in-the-u.s.-2018/topic-pages/message-from-the-director on October 24, 2020.
[2] Office of Justice Programs, U.S. Department of Justice. "Office for Victims of Crime (OVC) Archive." *Chapter 1 - The Scope of Violent Crime and Victimization: Statistical Overview*, www.ncjrs.gov/ovc_archives/nvaa/supp/a-ch1.htm.

Project Objectives

The overall project goal is to help the Federal Bureau of Investigation (FBI) make informed decisions so they can: (1) effectively develop programs and campaigns to fight crime; (2) more accurately and precisely direct resources towards law enforcement and victim support agencies; and (3) improve the American people's confidence in their safety and security. The decision to use the Bureau of Justice Statistics (BJS) National Crime Victimization Survey (NCVS) dataset to establish a crime incident reporting baseline and for modeling was made after several discussions with the FBI team. The following objectives were developed:

Objective	Technical Deliverable	Business Success Criteria
1. Understand differences between crime types and key insights from the data	An Interactive descriptive and predictive insights dashboard to the FBI	Training available to FBI Agents on new tools through demo videos & Feedback collected
2. Understand the context of reported and under-reported crimes to police.	Exploratory data analysis & reporting of key findings. Reporting of nuances from raw data processing	Set of recommendations provided to the FBI that result in actions
3. Understand the key drivers causing the under-reporting of crimes to police.	Build explainable predictive models to uncover key drivers & actionable insights	Models approved by the FBI, and a list of programs and campaigns prioritized by the FBI
4. Provide a likelihood of reporting to police for each identified crime.	A mobile application for FBI Agents in the field to view and act on the recommendations	Funding secured to productionize models & pipelines from pilot. Tracking mechanism established to monitor against long-term ~5-10% improvement in targeted goals

Figure 1: Objectives, Technical Deliverables, Business Success Criteria

Scope Clarification & Limitations

After 4 weeks of wrangling, the project was assessed at risk for delivery on time due to size, complexity, and vagueness of joins across tables to construct single incidents for analysis using the NCVS raw crime dataset. Data collection surveys and storage variables underwent a myriad of changes over the years which made interpreting which variables to use in what context impossible. Subject matter expertise is necessary to process the data. Upon consulting the FBI team we were provided another dataset that is a pre-processed version of the raw dataset from BJS . It has fewer yet more logically organized factors than the original dataset and is heavily used by the BJS in their NCVS victimization analysis tool. Preparation of the raw data was terminated as a parallel effort. A brief of raw data issues and preliminary work are provided in Appendix E.

Project goals were refined to clarify that models will focus on whether victims report crime to police. This distinction was made because the NCVS definition of reporting includes notification to other authorities such as a school principal, parish priest, or human resources at a job.

Limitations

The source for this data are surveys. Those reporting an incident to the interviewer and/or to the police may be limited to those who feel safe enough to admit the crime was committed. This may be affected by such things as (a) a woman or young child being in the presence of family members who do not know full details of the crime, or worse, who may be their perpetrators, (b) reluctance to report the crime against a familiar such as a boss or spouse, (c) living in a community where reports to the police are serious breaches of territory etiquette (where fear of and/or direct experience with reprisal curbs admission), or (d) by cultural norms which project shame onto victims for crimes like rape or other sexual assault. Other reasons exist but for brevity are not listed. Inference to the general population is therefore affected to an unknown magnitude with respect to assessment of the frequency of crimes.

Certain demographic data elements such as age were transformed from elemental integer values into buckets which we suspect may confound results by mixing disparate probabilities for victimization and likelihood of reporting. Household size was aggregated for 1, 2-3, 4-5, then 6+ members. This recategorization made direct use of this information joined with family income and US Census poverty definitions unavailable to determine whether poverty impacted crime reporting to police.

The new dataset removes other rich, available data such as:

- victim feedback on reasons for reporting/not reporting to police
- cost of crime considerations like duration of hospitalization, cost of victimization tied to loss of job or work hours, or other financial considerations (lawsuits, remediation after identity theft, etc)
- size and characterization of dwellings such as gated community, private entrances, deterrent systems such as neighborhood watches or alarm systems

Information containing specifics around locale (address, city, state) were not architecturally presented in the dataset although gathered in the survey. The NCVS raw data has a variable called scrambled control number which contains this information so the capability to model with this exists. In the present design the level available for geographical classification distinction is region (NorthEast, MidWest, South, and West), population density, and whether the locale is or is not part of or external to a metropolitan statistical area (MSA). This is a limitation because crime patterns in two areas of population density less than 10,000 within an MSA in the same region can have vastly different crime rates and potentials for reporting when interacting with other model variables. As an example, under the current constraints crime rates from the most violent crime area in Detroit, MI, Forest Park, is mixed with other similar sized areas in the region such as Oakland Park, MI, which has a very low crime rate. We suspect if this data were provided the predictive capability of the models and emergence of clusters would show improvement.

Analytics Approach

The process followed during this project reflects the CRISP-DM process as shown in Figure 2. This is an iterative process. The CrimeWatch team has completed the Business Exploration task and moved into Advanced Prototyping. This project is a pilot to show the value of this analysis. The next step would be to move this to production and use study insights in fighting crime. The production refinement steps are out of scope of this project. During the 6 month pilot period we will collect feedback from users and refresh dashboards and mobile applications as needed.

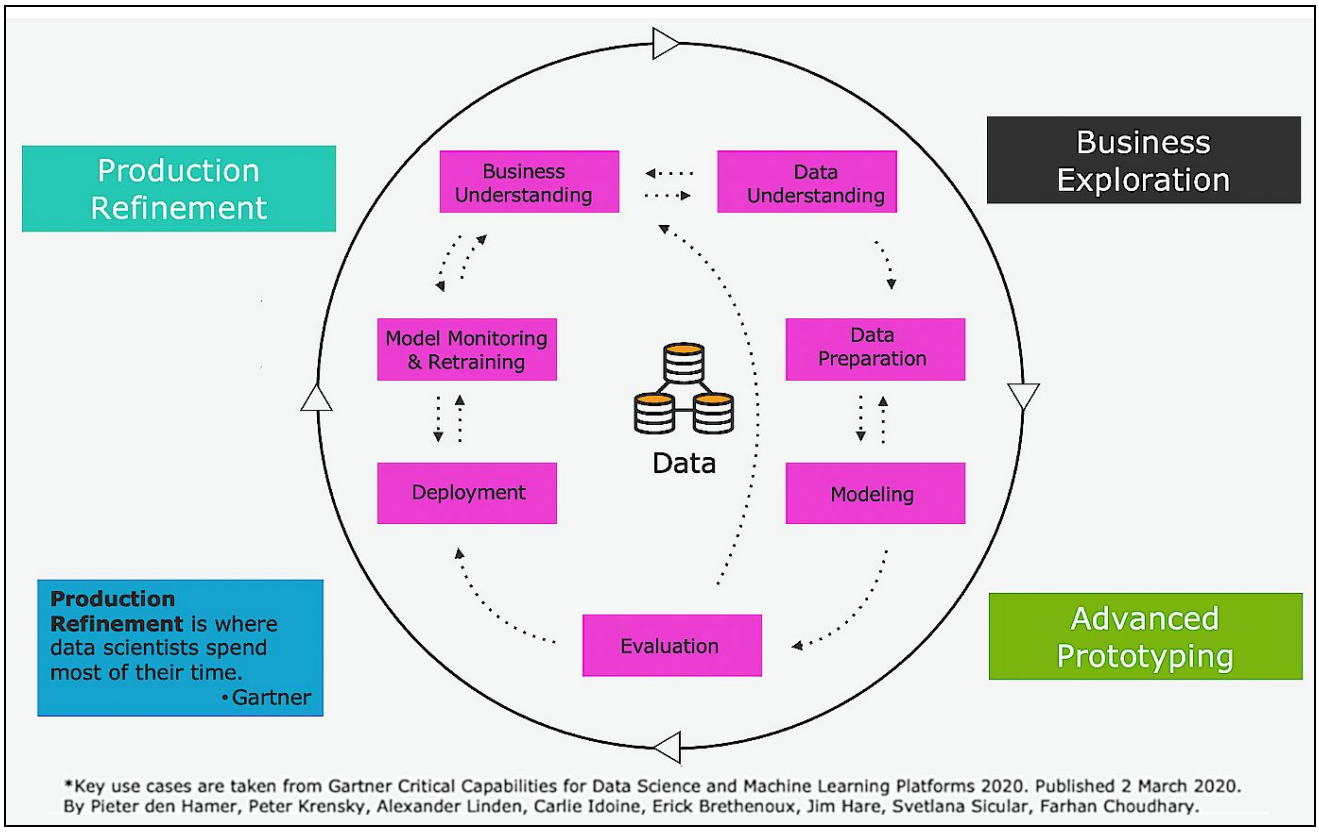


Figure 2: CRISP-DM Process

Data Understanding and Preparation Brief

As stated in the clarification of scope and limitations section, we pivoted to a new reduced, pre-processed dataset created and maintained for the NCVS victimization analysis tool. The new dataset does not have a unique identifier so it is impossible to relate the reduced record set with other information in the raw dataset.

In the new dataset there are two main categories of data: Household and Personal. Each has two csv files - one for population and one for victimization incidents. Our study focused on the Personal and Household

victimization files only as our mission was to identify patterns/clusters for under reporting of crime to the police. Appendix A and B provide data descriptions and more information for understanding the data .

Weighting is available in this study on both the personal and household level to adjust the sample to be representative of the population. Since the primary goal of the project is to understand the patterns in underreported crime incidents, we used unweighted data throughout this project.

The Justice Department reports household and personal victimization crime has declined over the years. A reverse trend is observed around 2012 [https://www.bjs.gov/content/pub/pdf/cv12.pdf]. Online sources note the many changes and sampling adjustments NCVS has undergone over the years. To ensure homogenous data was being used, we conducted statistical significance tests. These tests did not yield consistent results, which is attributed to underlying data complexity. The last 6 years worth of data (2013-2019) was chosen as it is homogenous within both the datasets and is considered of sufficient span and relevance to meet project goals.

The CrimeWatch team performed data cleaning, data transformation, data modeling, and visualization in order to arrive at final recommendations.

All variables in these processed Bureau of Justice Statistics datasets are categorical, with the exception of the year and weight variables. This is quite common in survey research. These datasets, on the whole, are clean. Missing values were re-coded into a non-response category since imputing to address non-response when so many factors influence it isn't appropriate. Variables that are highly correlated with each other remain in the dataset but a correlation analysis was used to determine the best variables for modeling. All cases that were not coded as 'yes' or 'no' in the notify police variable were removed. No duplicate records have been removed as these are seen as valid and indicate multiple events of the same type of incident within the same time period. For detailed information on the processing of the raw survey data refer to Appendix E.

Predictive Model Selection

After initial exploratory research was conducted using different supervised and unsupervised machine learning approaches including Factor and Cluster Analysis (see discussion in Appendices C and D), decision tree algorithms were selected to explore key underlying reasons for reporting or not reporting different crimes by category due to their high predictive capability and ease of interpretation. Logistic regression models were also explored as results provide parameter coefficients for the influential variables.

Decision tree

Decision tree is one of the most popular algorithms in supervised machine learning domain for understanding, interpreting, and predicting responses in a dataset. Depending upon continuity of the target output, decision trees can be used both for classification, where the target variable can take a discrete set of values, and regression problems, where the target variable is continuous. Decision trees follow a very similar pattern to human thinking and that is why they are usually used: they can help even nonexperts see the logic for interpretation of data. The algorithms consist of the following steps:

1. Create a node: Select the variable (attribute) which yields the best separation of the data space.
2. Define a decision rule: Ask a question.
3. Create branches: Follow the answer paths.

4. Repeat the above loop for each branch until you arrive to an outcome (leaf of the tree)

As seen, the two big questions are: 1) what is the best attribute that gives the most information gain for splitting the data space, and 2) what is the best question/condition for splitting the data.

Different decision tree algorithms use different metrics when picking the best attribute at each node. ID3 (Iterative Dichotomiser 3) and C4.5 algorithms are among the early algorithms that were developed based on this concept in 1986 and 1993 by Ross Quinlan. ID3 was designed for classification of categorical targets only while C4.5, which was the successor of ID3, removed this limitation and expanded the algorithm to continuous variables. C4.5 was later expanded (C5.0 algorithm) to support numerical target variables for regression analysis.

CART is a more recent algorithm in this area for building binary decision trees for regression and classification problems. This method, which employs a newer statistical metric called "Gini Index" to create the nodes (decision points), is the foundation of many other important algorithms (e.g. Random Forest and Gradient Boosted decision trees), which are widely used in different machine learning platforms.

Ensemble Methods

Despite the many advantages of decision trees there are some concerns that they may overfit the data and yield poor generalization performance. Other techniques developed address overfitting and other common errors like bias and variance in decision tree algorithms.

Random Forest and Gradient Boosting are two effective ensemble methods that combine multiple learners to create a more accurate model and address those issues.

- Random Forest is a collection of decision trees which differ from each other. The idea behind random forests is that if there are many trees, then overfitting can be reduced by averaging the result of each tree resulting in a more robust outcome. In other words, it uses wisdom of the crowd to deliver a more reliable model. These trees are combined employing a technique called "Bagging" in which each tree is treated as a parallel evaluator.
- Gradient Boosting (GBDT) works by building and trimming trees in a serial manner. It employs a method named "boosting" for connecting many different smaller trees sequentially. Algorithms build the model in stages by evaluating residuals (errors) of each tree and trying to fix them in the next tree.

Findings

Data Insights and Differences Between Crime Types

Thorough exploratory analysis was conducted using simple visualizations by slicing and dicing the data using various dimensions & studying correlations between the variables. The data description, Exploratory Data

Analysis (EDA), and initial findings are documented in Appendix B. The CrimeWatch dashboard and mobile application also offer opportunities for users to interact with and explore the data for insight.

Personal Crime

An initial question asked in the analysis was, "do demographics affect crime reporting?" Analysis of different incidents in the personal crime category reveal many important demographic nuggets about different types of crime.

Reporting Rate by Demographic: Gender. Figure 3 below shows that for the years 2013:2019 in this survey sample:

- rates of Personal Crime reporting differ by victim gender; consistently about 5% of all crimes are less reported by males than by females
- the lowest rate of reporting to police for the five Personal Crime categories is in the serious crime category Rape/sexual Assault; both sexes report less than 30% of Rape/sexual Assaults; alarmingly, over 70% of both sexes do not
- higher reporting rates occur in the serious crime categories Aggravated Assault, and Robbery, however men still do not report 40% of incidents while 35% of women do not
- reporting is slightly improved but still bad for Simple Assault and Personal Theft where a "both-sex" average of 60% of the incidents are not reported to police
    - 10% more males than females do not report Personal Theft
    - 5% more do not report Simple Assault

Distribution of reported (Yes) and unreported (No) personal crimes per gender of the victims

Newoff	Gender	No	Yes
Aggravated assault	Female	36.56%	63.44%
	Male	41.76%	58.24%
Personal theft	Female	55.23%	44.77%
	Male	65.44%	34.56%
Rape/sexual assault	Female	70.57%	29.43%
	Male	72.73%	27.27%
Robbery	Female	35.56%	64.44%
	Male	42.22%	57.78%
Simple assault	Female	55.77%	44.23%
	Male	59.68%	40.32%

Figure 3: Personal Crime Reporting to Police Characterized by Gender

Reporting Rate by Demographic: Location (region, MSA). Figure 4 below shows that for the years 2013:2019 in this survey sample:

- the lowest reporting rate for all five types of Personal Crime by region and MSA status is in the category Rape/sexual Assault; over 70% are not reported
- rates of Personal Crime reporting to police with a couple exceptions follow similar trends by region and whether the crime occurred within or outside of an MSA

- similar to cut by Gender, Personal Theft and Simple Assault reporting by all types of MSA and Region is slightly improved but still bad given an average of 60% of the incidents are not reported to police; this means only 40% are reported

Distribution of reported (Yes) and unreported (No) personal crimes in different regions

Newoff	Region	No	Yes
Aggravated assault	Midwest	39.33%	60.67%
	Northeast	46.43%	53.57%
	South	37.66%	62.34%
	West	38.90%	61.10%
Personal theft	Midwest	60.71%	39.29%
	Northeast	65.67%	34.33%
	South	55.32%	44.68%
	West	58.73%	41.27%
Rape/sexual assault	Midwest	70.98%	29.02%
	Northeast	72.28%	27.72%
	South	69.04%	30.96%
	West	71.90%	28.10%
Robbery	Midwest	34.51%	65.49%
	Northeast	37.50%	62.50%
	South	34.55%	65.45%
	West	49.22%	50.78%
Simple assault	Midwest	57.70%	42.30%
	Northeast	59.50%	40.50%
	South	55.25%	44.75%
	West	59.15%	40.85%

Distribution of reported (Yes) and unreported (No) personal crimes in different metropolitan statistical areas (MSA)

Newoff	Msa	No	Yes
Aggravated assault	Not part of principal city within MSA	39.78%	60.22%
	Outside MSA	36.71%	63.29%
	Principal city within MSA	40.04%	59.96%
Personal theft	Not part of principal city within MSA	50.00%	50.00%
	Outside MSA	63.16%	36.84%
	Principal city within MSA	65.90%	34.10%
Rape/sexual assault	Not part of principal city within MSA	70.35%	29.65%
	Outside MSA	72.45%	27.55%
	Principal city within MSA	70.92%	29.08%
Robbery	Not part of principal city within MSA	37.60%	62.40%
	Outside MSA	41.58%	58.42%
	Principal city within MSA	40.03%	59.97%
Simple assault	Not part of principal city within MSA	56.60%	43.40%
	Outside MSA	53.72%	46.28%
	Principal city within MSA	60.19%	39.81%

Figure 4: Personal Crime Reporting to Police Characterized by Region and MSA

Reporting Rate by Demographic: Age, Gender. Figure 5 below shows the level of reporting to police in the five Personal Crime categories characterized by age of victim and gender. These are a few important take-aways:

- approximately 78% of the Rape/sexual Assault victims are females between ages of 18-64 years old
  - one in five are between the ages of 12 to 20
  - one in three are between 25-49 years old
- approximately 28% of the Aggravated Assault victims are between 35-49 years old
- most Personal Theft victims are individuals older than 25 years old
- all crimes except Rape/sexual Assault occur roughly equally between the sexes, and happen most often to 'mature adult' victims between the ages of 25-64 years old

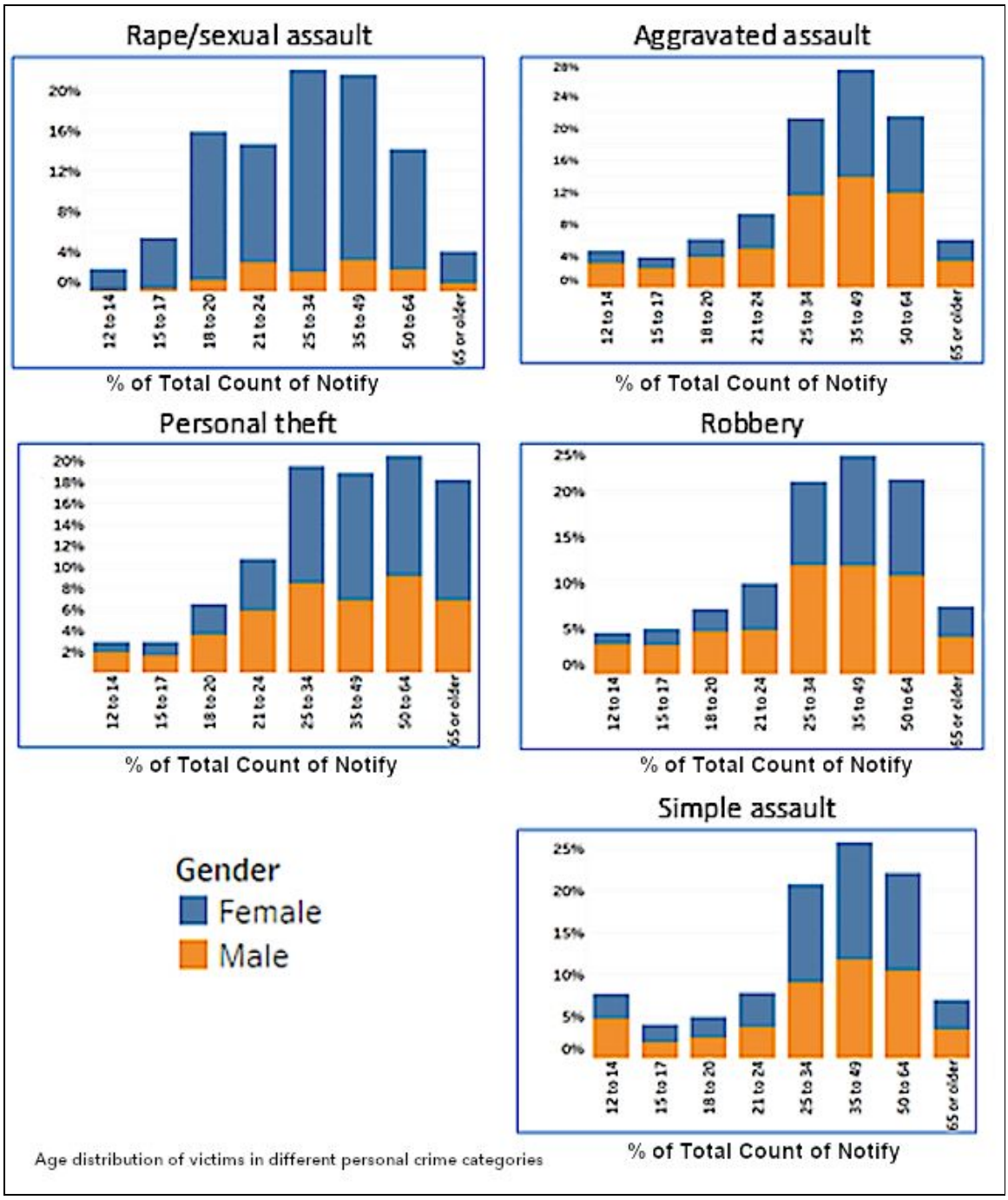


Figure 5: Personal Crime Category Reporting to Police by Age and Gender

Household Crime

Reporting Rate by Demographic: Location (region, MSA) Analysis of different incidents in the Household Crime category was also conducted. Figure 6 shows:

- the largest proportion of unreported Household Crimes (~70%) are in the "Theft" category regardless of region or MSA status
- the highest proportion of reported Household Crimes (~77%) are in the "Motor-vehicle Theft" category
- roughly half of all Burglary/trespassing crimes go unreported regardless of region or MSA status

Row Proportions: Distribution of reported (Yes) and unreported (No) property crimes in different regions

Newoff	Region	No	Yes
Burglary/trespassing	Midwest	49.94%	50.06%
	Northeast	49.24%	50.76%
	South	45.75%	54.25%
	West	54.93%	45.07%
Motor-vehicle theft	Midwest	20.44%	79.56%
	Northeast	21.21%	78.79%
	South	22.13%	77.87%
	West	23.76%	76.24%
Theft	Midwest	70.95%	29.05%
	Northeast	72.55%	27.45%
	South	69.58%	30.42%
	West	73.02%	26.98%

Row Proportions: Distribution of reported (Yes) and unreported (No) property crimes in different metropolitan statistical areas (MSA)

Newoff	Msa	No	Yes
Burglary/trespassing	Outside MSA	50.56%	49.44%
	Not part of principal city within ..	49.28%	50.72%
	Principal city within MSA	49.56%	50.44%
Motor-vehicle theft	Outside MSA	25.61%	74.39%
	Not part of principal city within ..	22.01%	77.99%
	Principal city within MSA	21.94%	78.06%
Theft	Outside MSA	69.04%	30.96%
	Not part of principal city within ..	69.59%	30.41%
	Principal city within MSA	73.72%	26.28%

Figure 6: Reported to Police by Property Crime Type by Region and MSA

Reporting Rate by Demographic: Gender and Location Figure 7 explores reporting to police by crime type and gender and by crime type and location. Trends observed in the survey sample are very similar to that noted for different regions.

- approximately 82% of Theft crimes at schools are not reported to the police
- half of the Burglary/trespassing Household crimes which occur at or near a victim's home go unreported
- a very high percentage of Motor-vehicle theft in the Household category are reported to police except when it occurs at school - these crimes only show report less than one in five times
- there is little difference in the reporting rate of Household Crime based on gender

Row Proportions: Distribution of reported (Yes) and unreported (No) Property crimes per gender of the victims

Newoff	Hhgen	No	Yes
Burglary/trespassing	Female	49.49%	50.51%
	Male	49.80%	50.20%
Motor-vehicle theft	Female	24.10%	75.90%
	Male	20.57%	79.43%
Theft	Female	71.85%	28.15%
	Male	70.78%	29.22%

Row Proportions: Distribution of reported (Yes) and unreported (No) personal crimes at different locations

Newoff	Locationr	No	Yes
Burglary/trespassi..	At or near victim's home	49.63%	50.37%
Motor-vehicle theft	School	12.50%	87.50%
	At or near friend, neighbor, or relative's home	15.89%	84.11%
	Other location	12.82%	87.18%
	Commercial place, parking lot, or other public a..	14.95%	85.05%
	At or near victim's home	27.59%	72.41%
Theft	School	81.80%	18.20%
	At or near friend, neighbor, or relative's home	68.99%	31.01%
	Other location	70.36%	29.64%
	Commercial place, parking lot, or other public a..	64.29%	35.71%
	At or near victim's home	73.08%	26.92%

Figure 7: Reported to Police by Property Crime Type by Gender and Location

Reporting Rate by Demographic: Income Figure 8 explores reporting to police by crime type and income. Trends observed are:

- the rate of reporting/non-reporting of Household Crime incidents vary very little by income level.
- three out of four "Thefts" in all income groups are not reported to the police; the non-report rate increases with income
- half of all Burglary/trespassing incidents are unreported regardless of income, fewer are reported by individuals with income less than $24,999

Row Proportions: Distribution of reported (Yes) and unreported (No) Property crimes per the total income of the household head and all members of the household

Newoff	Hincome	No	Yes
Burglary/trespassing	Unknown	47.26%	52.74%
	Less than $7,500	58.94%	41.06%
	$7,500 to $14,999	57.18%	42.82%
	$15,000 to $24,999	53.70%	46.30%
	$25,000 to $34,999	49.73%	50.27%
	$35,000 to $49,999	48.77%	51.23%
	$50,000 to $74,999	46.68%	53.32%
	$75,000 or more	43.94%	56.06%
Motor-vehicle theft	Unknown	22.03%	77.97%
	Less than $7,500	29.11%	70.89%
	$7,500 to $14,999	27.86%	72.14%
	$15,000 to $24,999	18.65%	81.35%
	$25,000 to $34,999	20.09%	79.91%
	$35,000 to $49,999	25.00%	75.00%
	$50,000 to $74,999	21.62%	78.38%
	$75,000 or more	20.95%	79.05%
Theft	Unknown	69.77%	30.23%
	Less than $7,500	77.96%	22.04%
	$7,500 to $14,999	77.14%	22.86%
	$15,000 to $24,999	76.41%	23.59%
	$25,000 to $34,999	72.73%	27.27%
	$35,000 to $49,999	71.39%	28.61%
	$50,000 to $74,999	70.38%	29.62%
	$75,000 or more	67.74%	32.26%

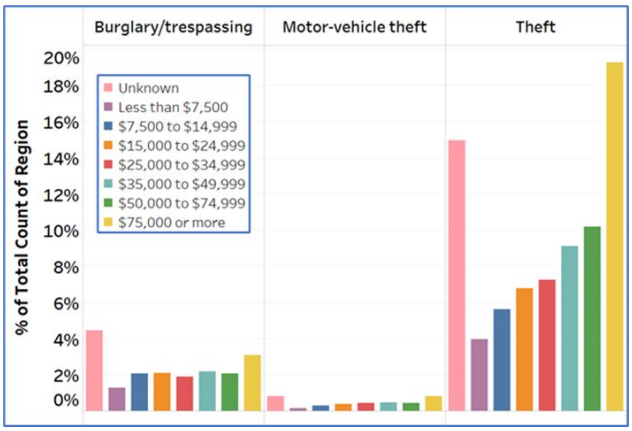


Figure 8: Reporting to Police by Income Level and Crime Category and by Region, Income, and Crime Category

Reporting Rate by Demographic: Population Figure 9 explores reporting to police by crime type and population size .Trends observed are:

- nearly three in four of the incidents in the Household Crime Theft category are not reported to the police regardless of population density
- more than three in four Household Crime Motor-vehicle Theft incidents in all population densities are reported
- non-reports to police in the Household Crime Theft category increase with population density, although for all densities three in four incidents of this type are left unreported
- Household Crime decreases with population size which could be due to increased tax-based financial resources for prevention such as law enforcement
- a much higher number of incidents go unreported in areas with less than 100,000 people

Distribution of reported (Yes) and unreported (No) Property crimes per populations of the place in which the housing unit is located

Newoff	Popsize	No	Yes
Burglary/trespassing	Not a place	48.82%	51.18%
	Under 100,000	50.37%	49.63%
	100,000 to 249,999	49.44%	50.56%
	250,000 to 499,999	48.00%	52.00%
	500,000 to 999,999	48.26%	51.74%
	1 million or more	51.21%	48.79%
Motor-vehicle theft	Not a place	22.81%	77.19%
	Under 100,000	22.53%	77.47%
	100,000 to 249,999	21.91%	78.09%
	250,000 to 499,999	22.16%	77.84%
	500,000 to 999,999	20.89%	79.11%
	1 million or more	23.08%	76.92%
Theft	Not a place	69.06%	30.94%
	Under 100,000	69.77%	30.23%
	100,000 to 249,999	73.25%	26.75%
	250,000 to 499,999	74.35%	25.65%
	500,000 to 999,999	75.27%	24.73%
	1 million or more	75.76%	24.24%

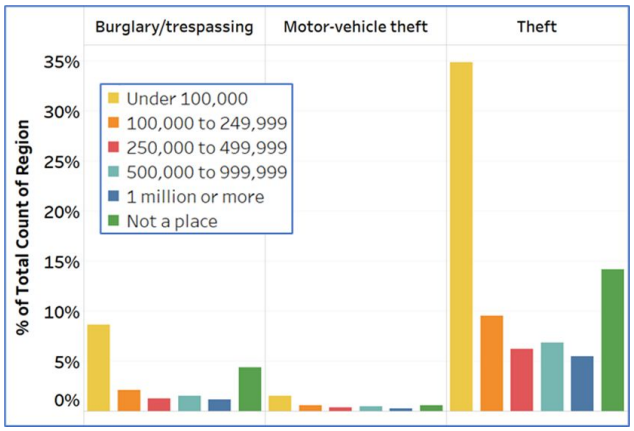


Figure 9: Reporting to Police by Population Size and Crime Category and by Region, Population Size and Crime Category

Context of Reporting and Not Reporting Crimes to Police

To understand if one can interpret any underlying patterns behind crime reporting we conducted multiple correspondence analysis, factor analysis, k-mean clustering, and visualized data using PCA and TSNE. Chi-square Pearson correlations, one-hot encoding and weight of evidence (WoE) representations were used to encode the categorical variables for preparing input data.

Figures 10 and 11 show association of survey variables with respect to reporting to police. Weight of Evidence (WOE) calculations were performed, using the notify categories yes and no, so their categories were separated with maximum information value and assigned a value/coefficient using a logistic regression model where a positive value indicates a high or greater likelihood of reporting to police and a negative correlation indicating a less likelihood of reporting to police. These WOE values were used in factor analysis.

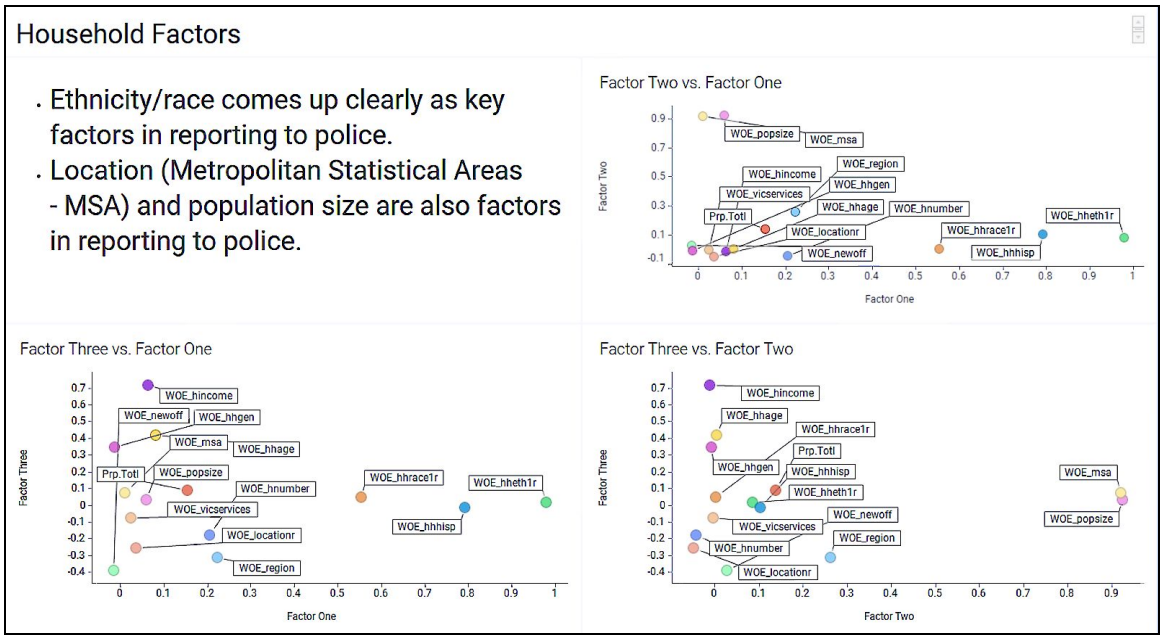


Figure 10: Key Household Factors in Reporting to Police Revealed by Factor Analysis

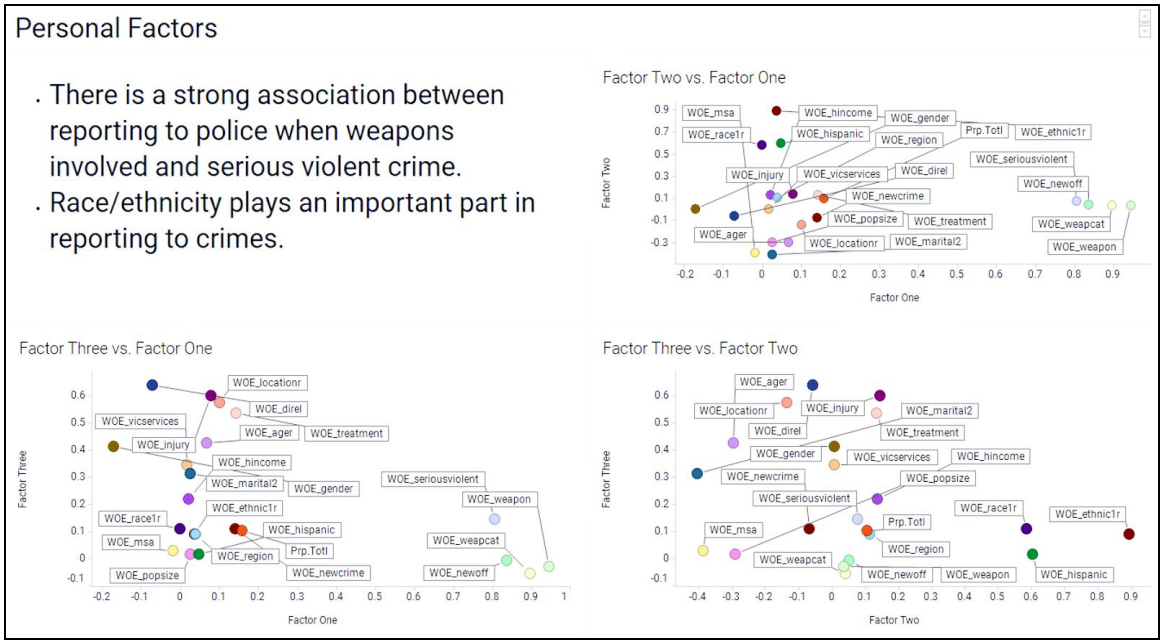


Figure 11: Key Personal Factors in Reporting to Police Revealed by Factor Analysis

While we did not arrive at any solid conclusions from these techniques, we observed some structural differences in factor loadings & TSNE cluster patterns between reported and unreported crimes in both the

datasets as shown in Figure 12 (green and orange dots started getting separated in particular subgroups which could be an indicative of response variation in different groups of people as observed in previous section).

The WoE & PCA processed data started showing some contrast in crime responses (yes or no) compared to original one-hot vectors. Applying k-means clustering on WoE & PCA processed data yielded AUC of more than 58% indicating some predictive power in the data (better than a no skill classifier) as shown in Figures 13 & 14. The detailed findings from factor and cluster analysis are presented in Appendix C Factor and Cluster Analysis.

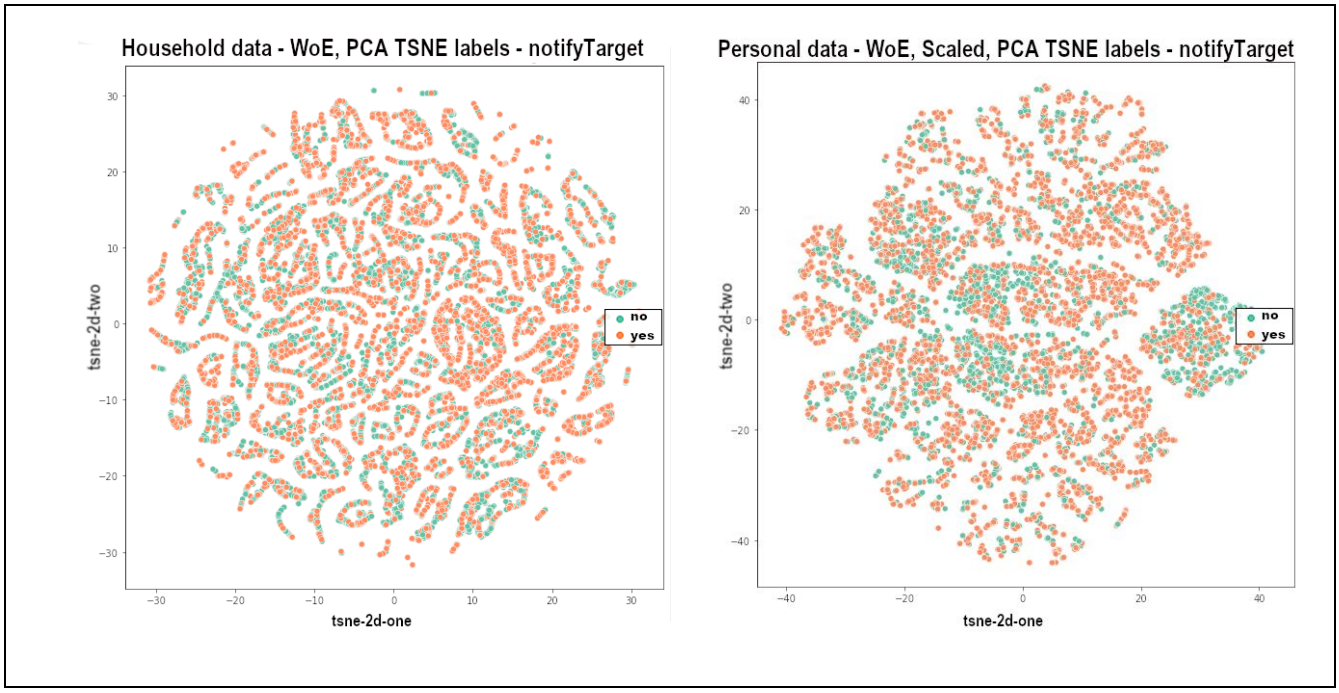


Figure 12: TSNE visualization of Household and Personal crime response patterns

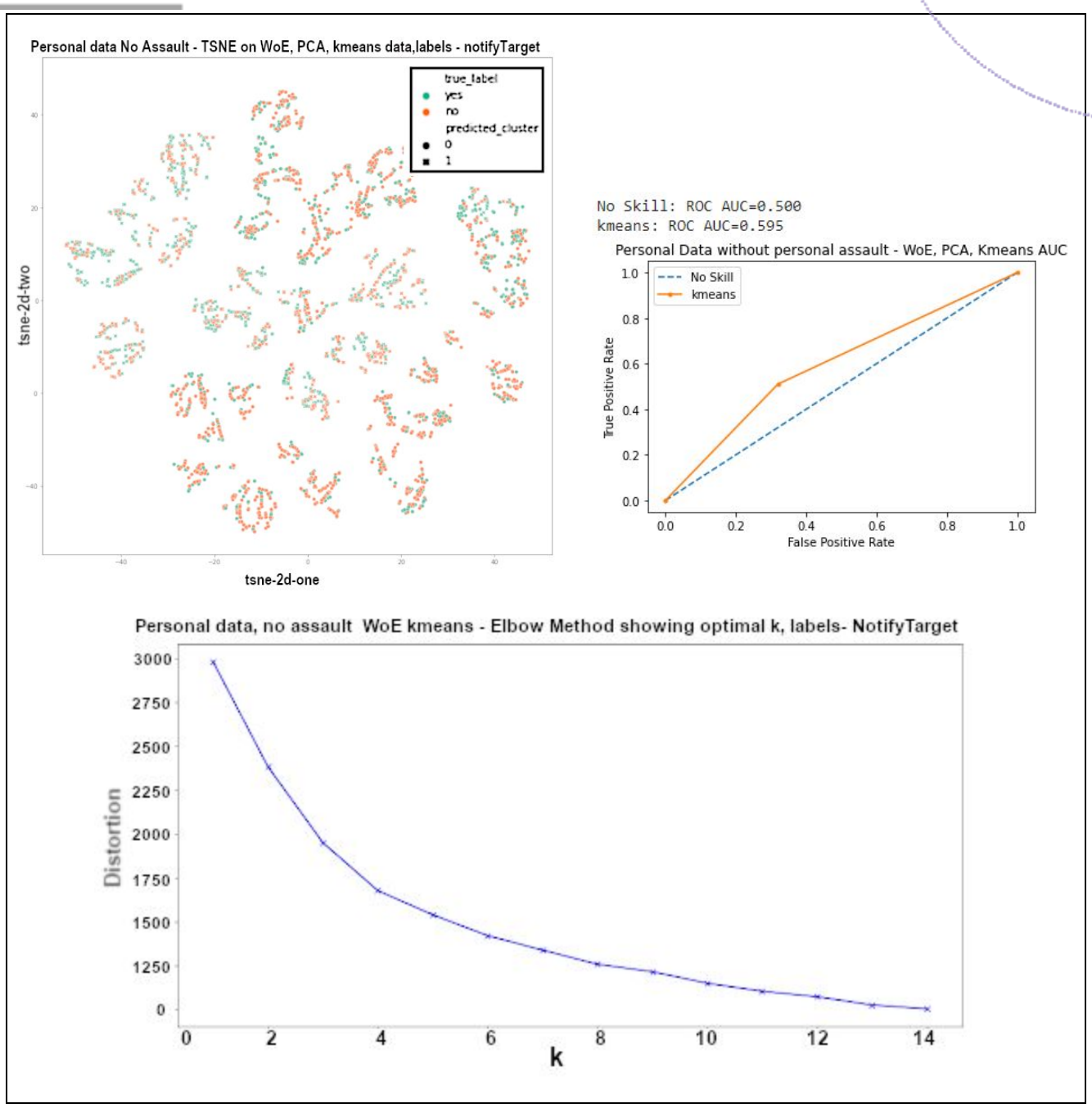


Figure 13: TSNE Visualization & K-means Clustering Results on Personal Data without Simple Assault Crimes

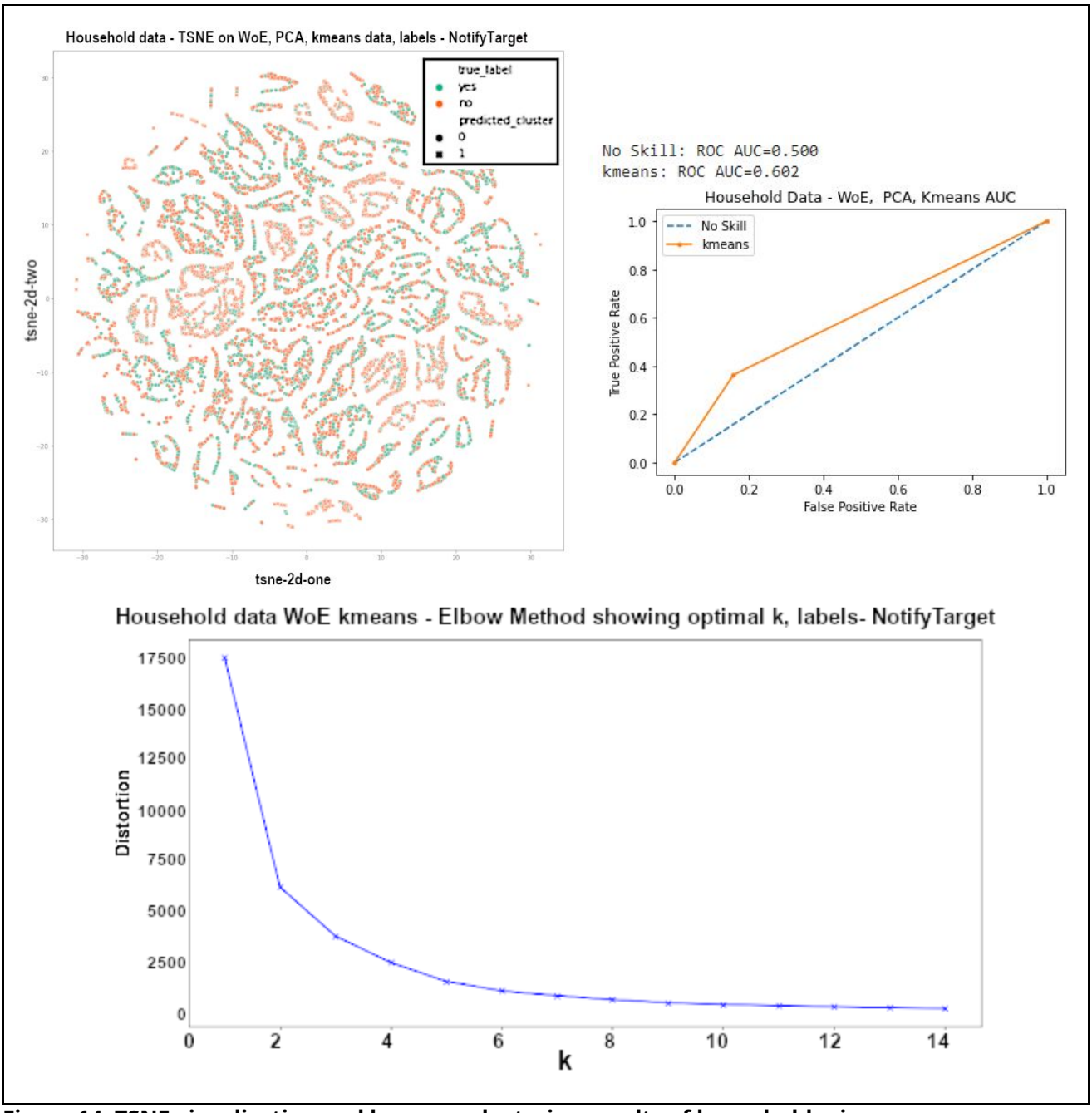


Figure 14: TSNE visualization and k-means clustering results of household crimes

Since there are no obvious interpretable groups from clustering techniques, we created additional advanced models to identify key drivers behind crime reporting patterns. We experimented with several regression and

tree based models by formulating it as a supervised classification problem with notify 'yes' or 'no' as a target variable.

Key Drivers of Not Reporting Crimes to Police

More than 10 predictive models were constructed including logistic regression and tree based models for each dataset to identify key drivers behind not reporting victimization to police. Notify was set as the target variable and the problem was formulated as a supervised classification problem.

The decision tree based models included the following methodologies:

- single tree algorithms (DecisionStump, HoeffdingTree, J48 and REPTree) using different metrics for picking the attributes
- ensemble models based on bagging (e.g. RandomForest, RandomTree)
- ensemble models based on boosting (AdaBoost with J48, REPTree and LME as the week learners)

While employing different variable selection criteria each model was tested for different selection of attributes as the predictor elements. This could potentially help to shed more light on the significance of different variables for predicting reporting/not reporting of crimes and reduce risk of overfitting by having less complexity in the models.

A 10-fold Cross Validation methodology was employed to reduce bias and variance and assure that the selected models have learned most of the patterns from the crime dataset correctly. The dataset was divided into 10 different subsets and the selected decision tree algorithm applied 10 times. Each time, one of these subsets act as the test set while the other 9 subsets act as the training set. In other words, every data point is placed in the validation and training sets 1 and 9 times, respectively, to improve accuracy of the predictions.

In the end the Random Forest algorithm turned out to be the front-runner in terms of consistent performance across both the datasets. It yielded the highest 'all crime' category score using 10-fold cross-validation with an accuracy of 69.1% for Personal and 68.5% for Household Crime reporting classification predictions as shown in Figures 15 & 16.

	All variables	Top 10 Attributes			Random set of attributes	
Personal crime		Pearson's correlation coefficient	Information Gain Based Feature Selection (Entropy)	Learner Based Feature Selection(through J48)	Set 1	Set 2
	1. newoff 2. locationr 3. weapon 4. weapcat 5. treatment 6. injury 7. vicservices 8. seriousviolent 9. marital2 10. ager 11. direl 12. hincome 13. msa 14. popsize 15. hispanic 16. region 17. newcrime 18. gender 19. race1r 20. ethnic1r	1. newoff 2. locationr 3. weapon 4. weapcat 5. treatment 6. injury 7. vicservices 8. seriousviolent 9. marital2 10. ager	1. locationr 2. newoff 3. treatment 4. weapcat 5. ager 6. weapon 7. vicservices 8. marital2 9. seriousviolent 10. injury	1. gender 2. ethnic1r 3. ager 4. Marital2 5. hincome 6. Popsize 7. region 8. newcrime 9. newoff 10. seriousviolent	1. newoff 2. locationr 3. weapon 4. weapcat	1. newoff 2. locationr

Mthodology	Decision tree algorithm	Accuracy	ROC (Avg)	Accuracy	ROC (Avg)	Accuracy	ROC (Avg)	Accuracy	ROC (Avg)	Accuracy	ROC (Avg)	Accuracy	ROC (Avg)
Single tree	DecisionStump	58.1	0.55	58.1	0.55	58.1	0.55	57.6	0.54				
	HoeffdingTree	63.0	0.66	63.9	0.67	63.9	0.67	59.6	0.63				
	J48 (ID3 = Iterative Dichotomiser 3)	64.9	0.67	63.9	0.67	63.9	0.67	61.8	0.63				
	LMT (Logistic Model Tree)	64.7	0.70	64.7	0.70	64.7	0.70	60.2	0.64				
	REPTree (CART)	62.8	0.66	63.6	0.67	63.6	0.67	61.0	0.64	62.2	0.65	62.3	0.65
Ensemble: Bagging algorithm	RandomForest	69.1	0.76	63.6	0.67	63.6	0.67	63.7	0.69	62.0	0.65	62.3	0.65
	RandomTree	62.9	0.64	63.3	0.66	63.3	0.66	61.3	0.65				
Ensemble : Boosting algorithm	AdaBoostM1 (with J48)	65.1	0.70										
	AdaBoostM1 (with REPTree)	64.6	0.69							62.2	0.65	60.7	0.64
	AdaBoostM1 (with LMT)	65.2	0.70										

Figure 15: Top predictors for non-reporting personal victimization crime to police by model accuracy.

Household crime	All variables	Top 10 Attributes			Random set of attributes	
		Pearson's correlation coefficient	Information Gain Based Feature Selection (Entropy)	Learner Based Feature Selection(through J48)	Set 1	Set 2
	1. newoff 2. vicservices 3. locationr 4. hincome 5. region 6. msa 7. popsize 8. hhage 9. hnumber 10. hheth1r 11. hhrace1r 12. hhhisp 13. hhgen 14. newcrime	1. newoff 2. vicservices 3. msa 4. region 5. hincome 6. popsize 7. locationr 8. hnumber	1. newoff 2. vicservices 3. locationr 4. hincome 5. region 6. msa 7. popsize 8. hhage	1. hincome 2. hheth1r 3. hnumber 4. popsize 5. region 6. newoff 7. vicservices 8. locationr	1. newoff 2. vicservices 3. locationr 4. hincome	1. newoff 2. vicservices

Mthodology	Decision tree algorithm	Accuracy	ROC (Avg)	Accuracy	ROC (Avg)	Accuracy	ROC (Avg)	Accuracy	ROC (Avg)	Accuracy	ROC (Avg)	Accuracy	ROC (Avg)
Single tree	DecisionStump	67.6	0.59	67.6	0.60	67.6	0.60	67.6	0.60				
	HoeffdingTree	68.4	0.64	68.3	0.65	68.3	0.64	68.2	0.64				
	J48 (ID3 = Iterative Dichotomiser 3)	69.1	0.59	68.8	0.61	68.6	0.61	69.2	0.61				
	LMT (Logistic Model Tree)	68.7	0.65	68.5	0.65	68.3	0.66	68.8	0.65				
	REPTree (CART)	68.0	0.64	68.5	0.65	68.2	0.65	68.4	0.64	68.2	0.65	67.8	0.61
Ensemble: Bagging algorithm	RandomForest	66.7	0.65	67.0	0.64	66.7	0.63	66.9	0.63	68.5	0.65	67.8	0.61
	RandomTree	65.4	0.61	66.8	0.63	66.7	0.62	66.8	0.62				
Ensemble : Boosting algorithm	AdaBoostM1 (with J48)	65.7	0.64										
	AdaBoostM1 (with REPTree)	65.7	0.64							68.4	0.65	67.8	0.61
	AdaBoostM1 (with LMT)												

Figure 16: Top predictors for non-reporting personal victimization crime to police by model accuracy.

Logistic Regression, Decision Trees, XGBoost models and Multiple Variable Selection methods for Random Forest were also constructed/experimented with to identify key drivers and improve model performance in this analysis and in parallel analysis performed by other analysts as detailed in Appendices C and D.

Crime Type as Primary Driver

The models indicated crime type as a primary driver behind 'reporting to police' choice. For example, the Personal Crime of Rape/sexual Assault has a high likelihood of not being reported whereas Motor-vehicle theft has a high likelihood of being reporting compared to other Household Crime types. Location of the crime and utilization of victimization services are among the other key drivers behind reporting or non-reporting based on XGBoost algorithms as shown by SHAP values in Figures 18 & 21.

Personal Victimization

The parameter estimates for the Personal Victimization Logistic Regression tells an interesting story (detailed in Appendix D). Focusing on the situations of serious crime Rape/sexual Assault where the victim did not notify police (negative coefficient values), the largest predictors of this crime are 'no victim services', and when

the victims were between the ages of 12 and 14, as shown in Figure 17. Note: negative values indicate stronger likelihood of not reporting the crime to police.

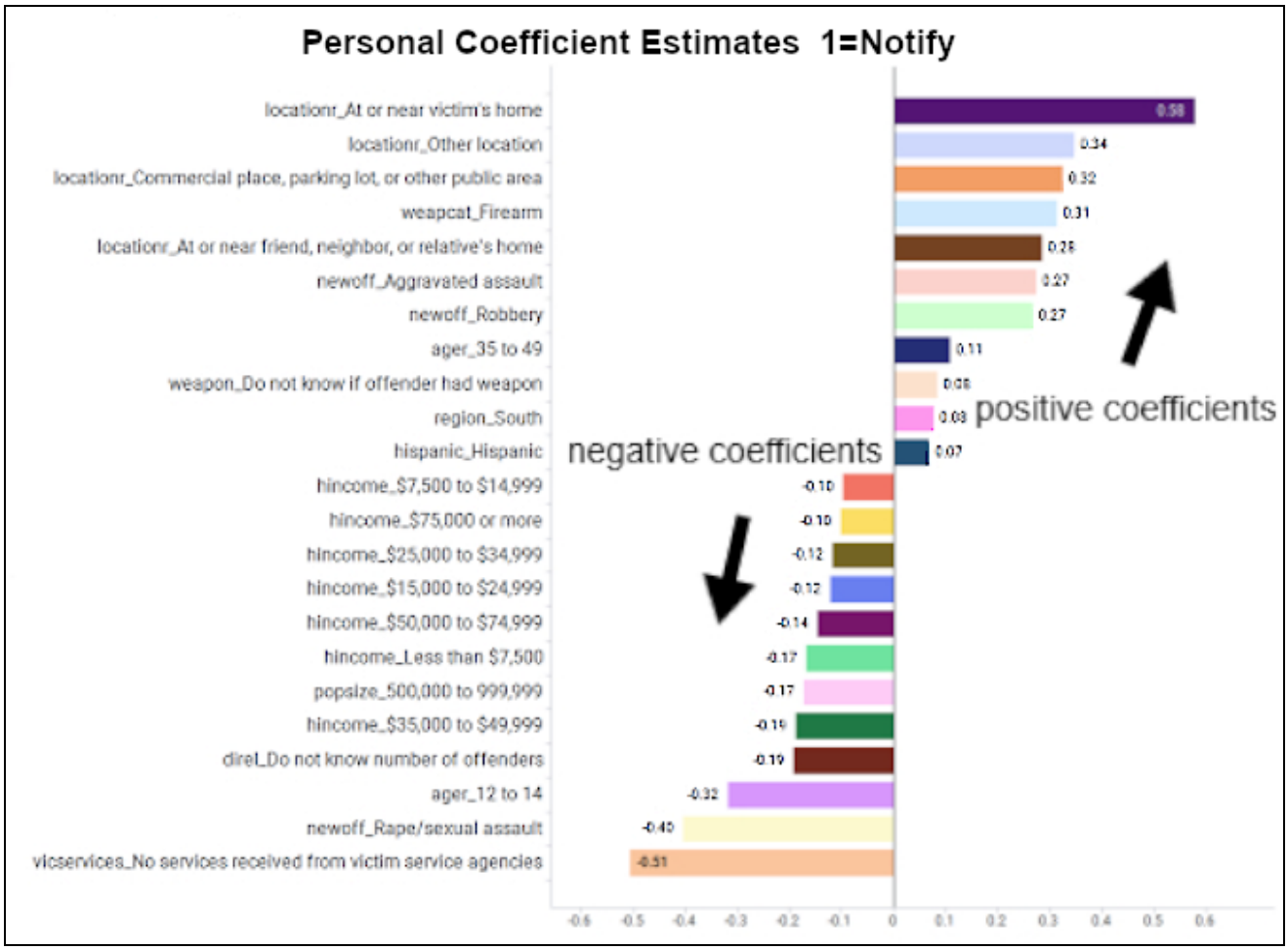


Figure 17: Personal Victimization Coefficient Estimates for Logistic Regression

The Personal Crime victimization Trees model using the Chi-square Automatic Interaction Detection (CHAID) method shown in Figure 18 also revealed less likelihood that the crime will be reported to police if it was committed away from the victim's home. However, in circumstances when crimes were experienced away from home and a firearm was involved, the crime is more likely to be reported to police.

This modeling technique also showed crimes committed without a firearm against victims away from home wherein the victim didn't receive services from victim services agencies have a high likelihood of not being reported, especially for victims between the ages of 12 and 14 and where the Personal Crime of Rape/sexual Assault was involved.

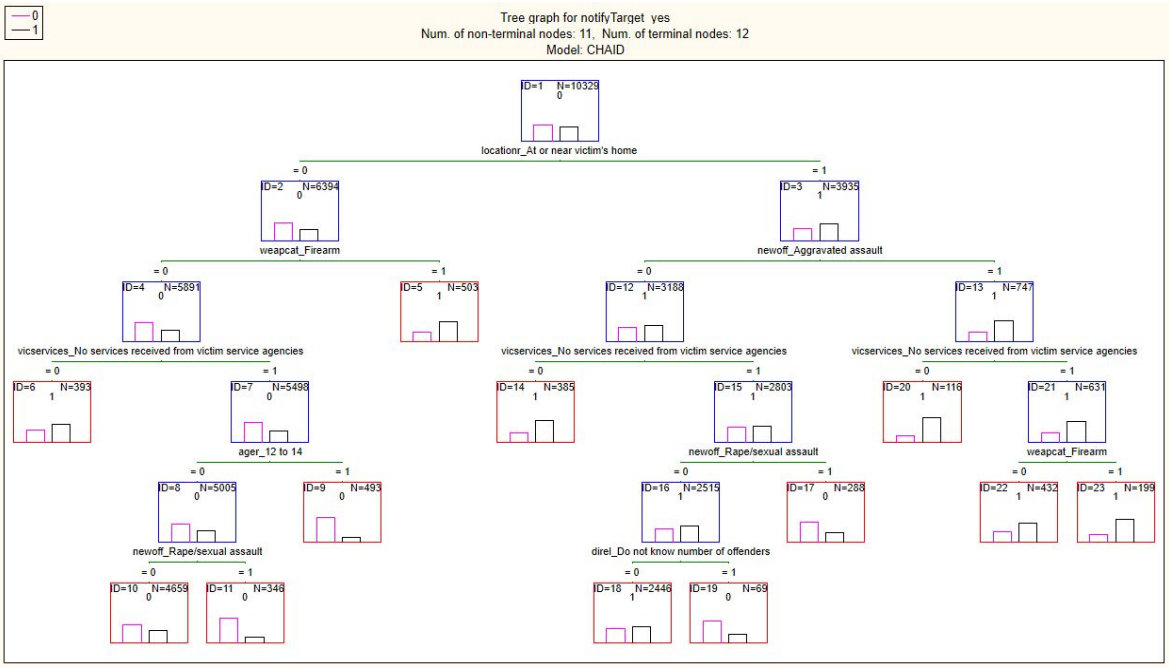


Figure 18: Personal Victimization CHAID (Chi-square Automatic Interaction Detector) model tree output where 1 = notify and 0 = didn't notify police.

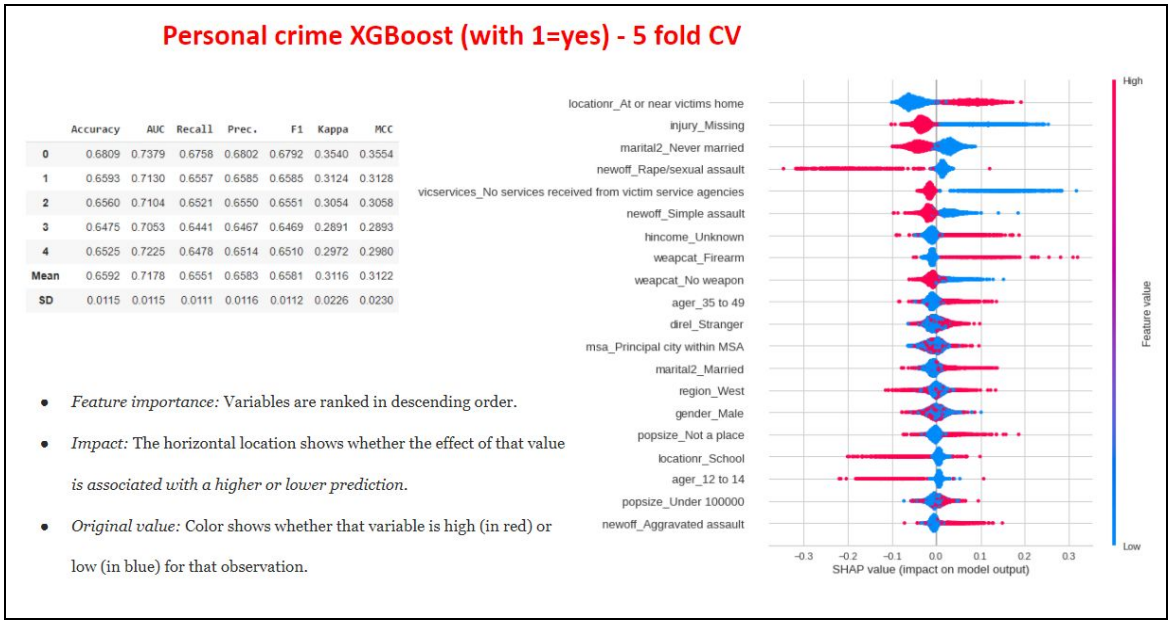


Figure 19: Personal Crime XGBoost model with SHAP Value for Key Variables Impacting Reporting to Police

Household Victimization

The parameter estimates for Logistic Regression regarding Household Crime victimization also tell an interesting story (see details in Appendix D). Focusing on the situations where the victim did not notify police (negative coefficient values), the largest predictors include crimes of theft and where there were no services received from victim service agencies. When crimes were committed at school, there was less likelihood a victim will report the crime to police, as shown in Figure 20. Note: negative values indicate stronger likelihood of not reporting the crime to police.

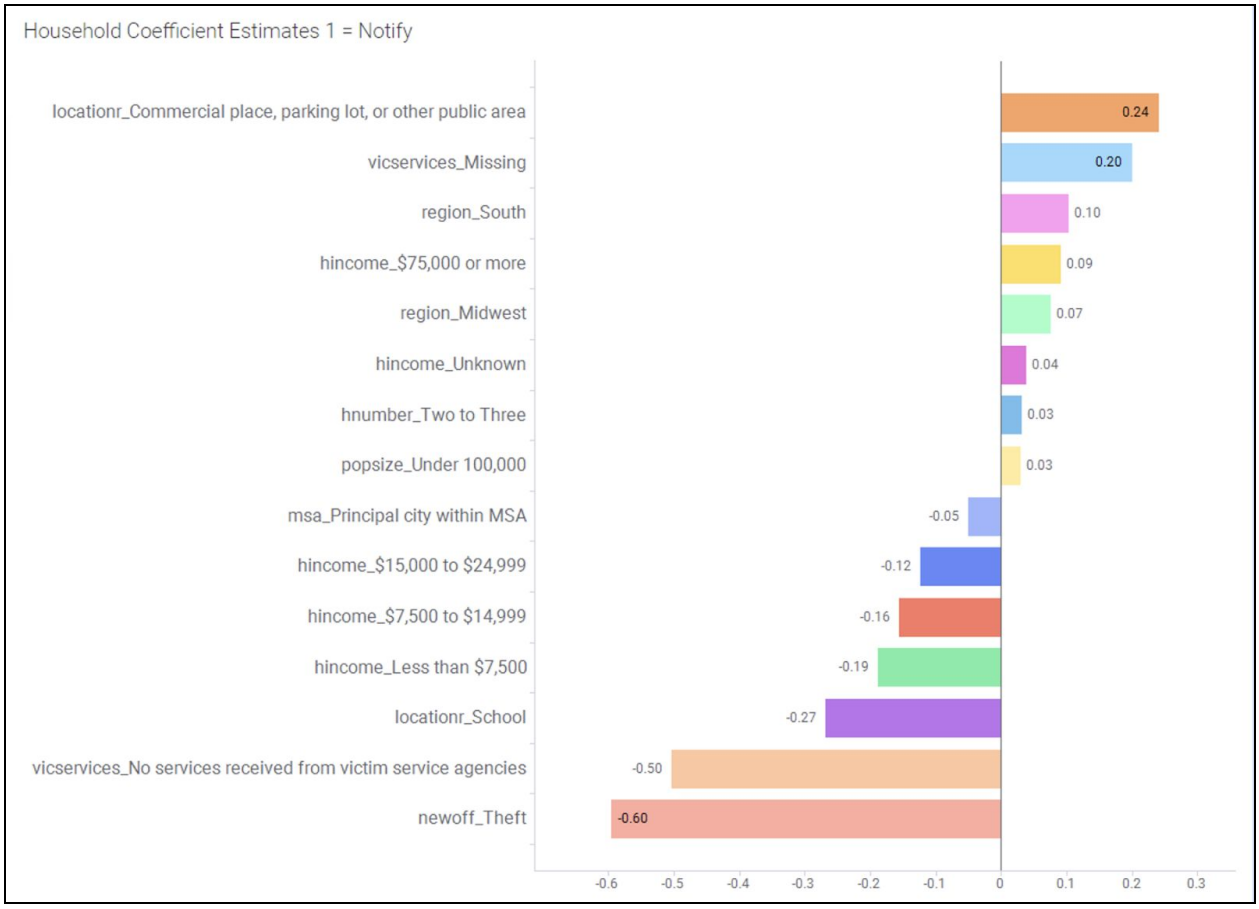


Figure 20: Household Victimization Coefficient Estimates for Logistic Regression

The Household Crime victimization (C&RT) Trees model is rather simple yet provides about 88% accuracy in predicting those who are likely to report the crime to police. When Household Theft is involved there is a strong likelihood that this crime will not be reported to police. There is a slightly greater likelihood the crime will not be reported to the police if the Household Theft crime was not committed at a commercial place, parking lot, or other public area, as shown in Figure 21. Note: Branches show decisions where 1 = notify and 0 = did not notify police.

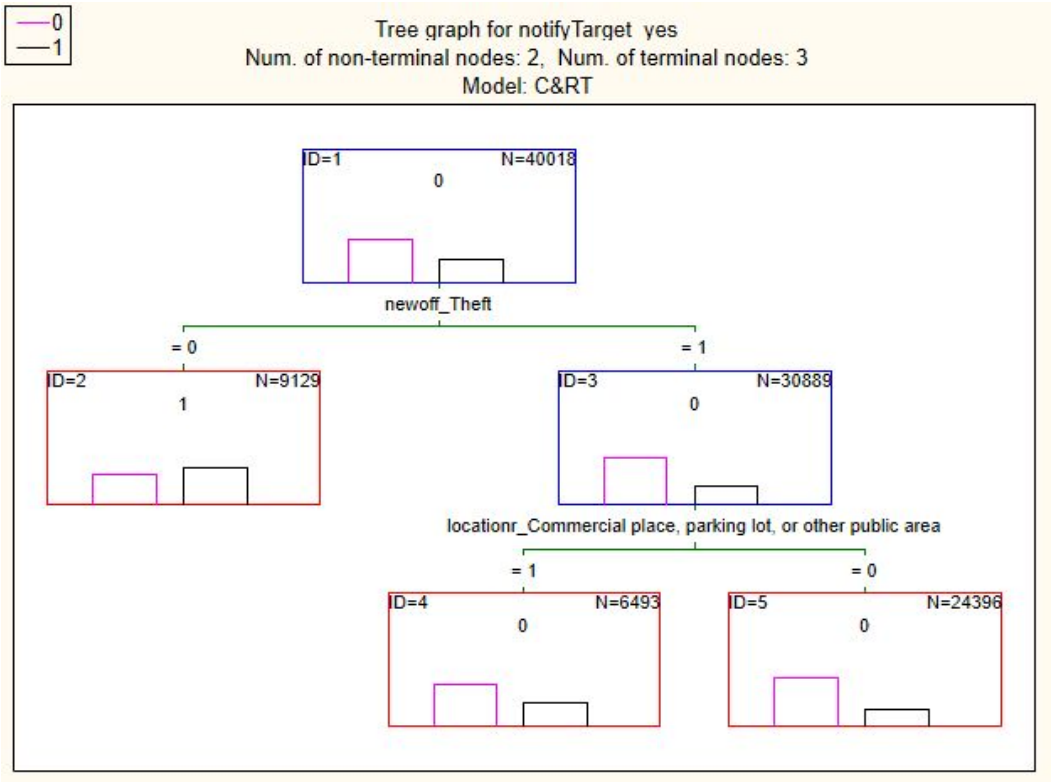


Figure 21: Household Victimization C&RT (Classification and Regression Tree) model tree output

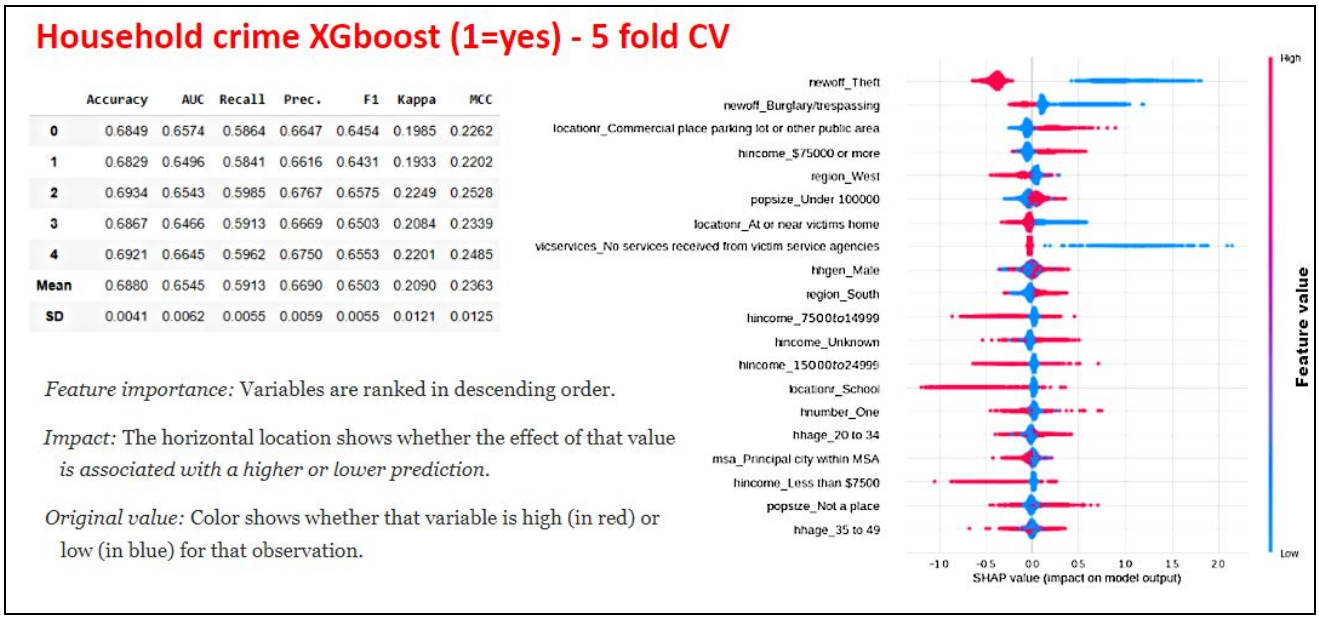


Figure 22: XGBoost & SHAP Value on household crimes

Likelihood of Reporting to Police By Crime Type

Individual models by crime type within both the Personal and Household categories were constructed to identify the likelihood of not reporting to police. Each model identifies the key drivers behind the crime type. Random Forest algorithms constructed using Correlation-based Variable Selection consistently performed well in predicting reporting to police so this technique was used. Model results are shown in Figures 23-24 for Household and Personal Crime by crime type.

One insight gleaned from the Household and Personal Crime analysis is that the most severe type of crimes, Motor-vehicle Theft & Rape/sexual Assault yielded the highest prediction cross-validation accuracies of 78.2% and 72.3% respectively using Random Forest classification models. This is an indicator that relatively high informative variables are in the data for interpreting these crimes.

Victims may have had mixed response choices behind other less severe type crimes in reporting which could have caused poor classification accuracy. Nuances in raw data could also explain some of the short-comings of these models. Full details on these limitations are documented in this document in section Scope Clarification and Limitations.

Random Forest algorithm	Training model including all variables (Full Model)		Top Attributes selected by Correlation-based Feature (CfsSubsetEval) methodology: This method evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.			Evaluation of the Full Model on the TEST dataset	
			Training model including only the most important variables				
Crime category	**Accuracy**	**ROC**	**Most informative attributes**	**Accuracy**	**ROC**	**Accuracy**	**ROC**
Theft	68.2%	0.60	1. vicservices 2. locationr	71.4	0.55	68.3%	0.61
Motor-vehicle theft	79.0%	0.60	1. vicservices 2. locationr	78.2	0.56	78.1%	0.66
Burglary	59.0%	0.63	1. hincome 2. vicservices	54.4	0.55	59.3%	0.64

Figure 23: Random Forest Model Classification Accuracies for Household Crime by Crime Type

Random Forest algorithm	Training model including all variables (Full Model)		Top Attributes selected by Correlation-based Feature (CfsSubsetEval) methodology: This method evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.			Evaluation of the Full Model on the TEST dataset	
			Training model including only the most important variables				
Crime category	**Accuracy**	**ROC**	**Most informative attributes**	**Accuracy**	**ROC**	**Accuracy**	**ROC**
Theft	62.6%	0.66	1. ager 2. marital2 3. hincome 4. msa 5. direl	64.3%	0.67	57.6%	0.68
Simple Assault	68.3%	0.74	1. ager 2. treatment 3. vicservices 4. locationr	63.3%	0.65	68.5%	0.75
Robbery	67.1%	0.70	1. race1r 2. ager 3. region 4. weapcat 5. treatment 6. vicservices 7. locationr	63.7%	0.67	67.3%	0.65
Rape/Sexual Assault	77.1%	0.78	1. weapcat 2. Treatment 3. vicservices	72.3%	0.67	77.9%	0.78
Aggravated Assault	68.6%	0.72	1. ager 2. weapcat 3. treatment 4. vicservices 5. locationr	0.62%	0.63	66.6%	0.71

Figure 24: Random Forest Model Classification Accuracies for Personal Crime by Crime Type

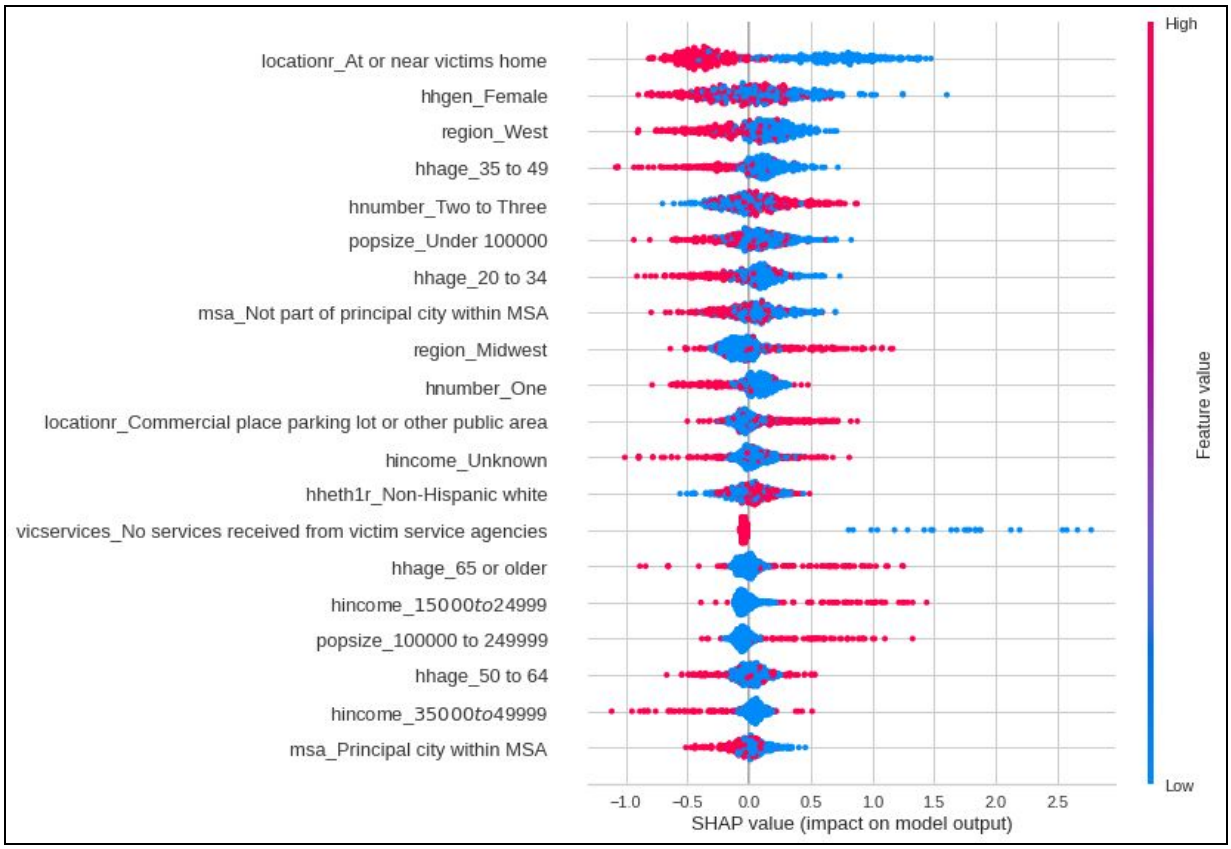


Figure 25: Interpretable SHAP values from Motor Vehicle Theft XGBoost model

Ensemble XGBoost models were also built for Motor-vehicle Theft & Rape/sexual Assault crime types to identify the key variables/patterns behind reporting choices made by people in the survey. SHAP values are plotted for model interpretability in Figure 25 and 26 respectively. Key takeaways are:

- The Motor-vehicle Theft model indicates that reporting to police has a high likelihood of occurring if the crime took place away from home (the blue bar on right indicates less of this, more of yes in reporting).
- The Rape/sexual Assault model indicates that if victim services are received from victim agencies there is a high likelihood of reporting (the red bar on right side indicates, more of this, more of yes in reporting).
- Ages 21-24, never married, and those victims of non-Hispanic ethnicity show less likelihood of reporting (red bar on to the left). These are very interesting findings as it can represent psychological differences in different groups of people.

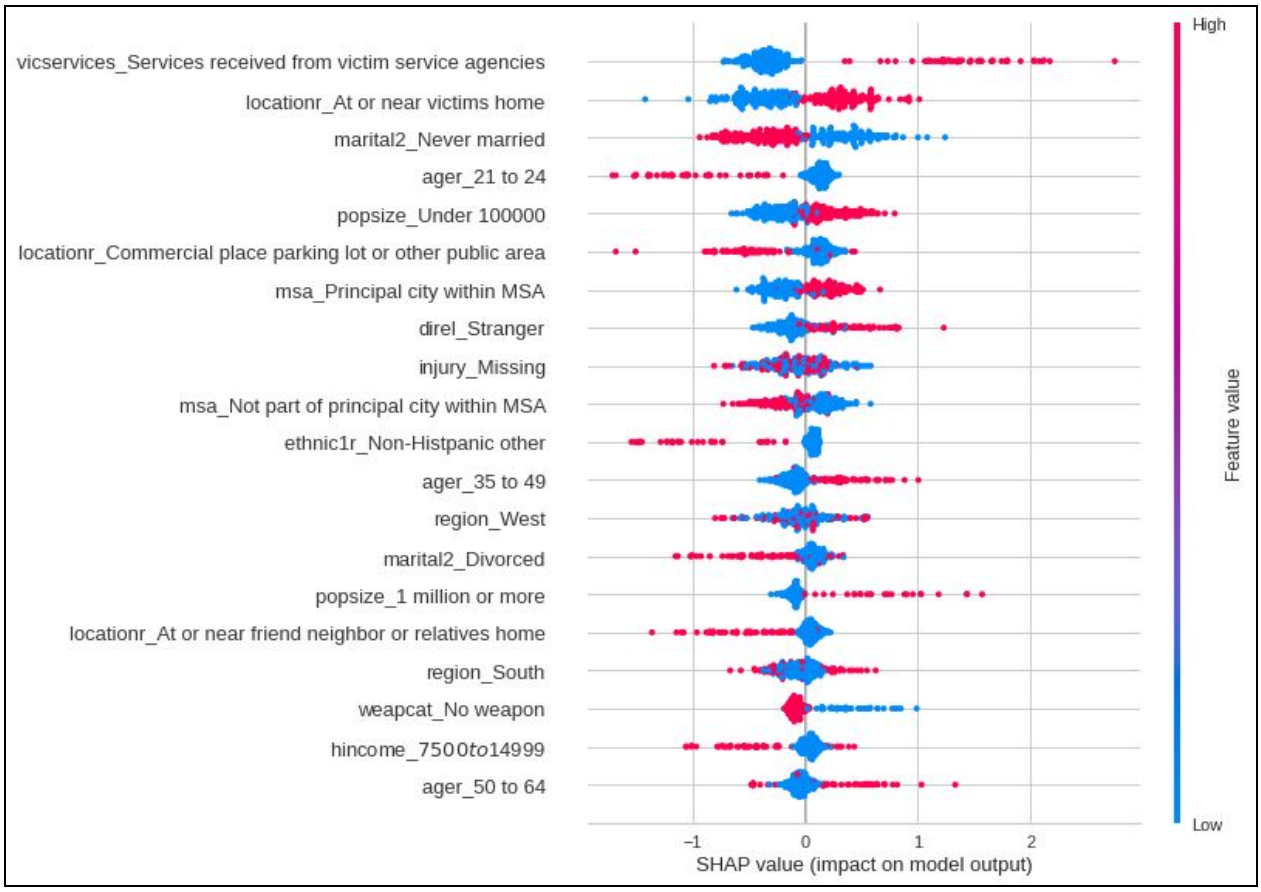


Figure 26: Interpretable SHAP values from Rape/Sexual Assault XGBoost model

The likelihood of reporting are also displayed as key metrics on the Android mobile application as shown following, in the document section Dashboard and Mobile Application.

Focusing on Household Crime type Theft generally, probabilities are over 70% that the crime will not be reported regardless of whether it occurs at or not at a school. Theft victimization at school 82% of the time will not be reported to police. Reporting Theft to the police is 23% less likely when services were not received from victim services agencies, as shown in Figure 27.

Figure 27: Average Probabilities of Not Reporting Household Theft

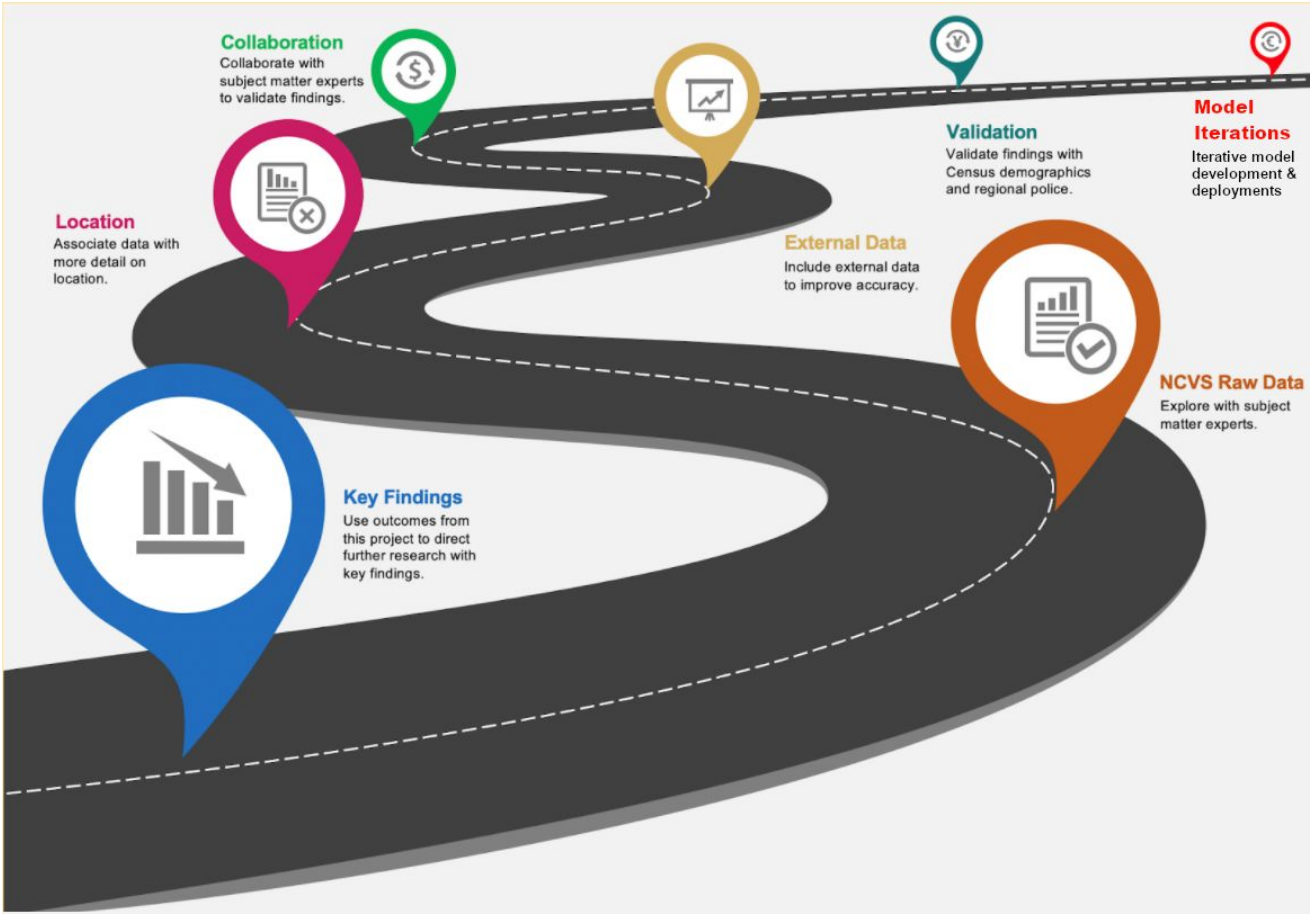
Focusing on personal rape/sexual assault and personal assault, rape and sexual assault much more likely to not be reported to police, especially when no services were received from victim services agencies; other types of assault less so, as shown in Figure 28.

Figure 28: Average Probabilities of Not Reporting Personal Rape/Sexual Assault and Assault.

Conclusions & Recommendations

The top model results using Ensemble Method Random Forest for predicting "will victims report a crime to the police " for all types of Personal Crime had an accuracy of 62.3% and revealed Type of Crime (newoff) and Location (locationr) are most influential in the decision; for all types of Household Crime accuracy using same technique was 67.8% and Type of Crime (newoff) and Victim Services Received (vicservices) were the most influential. For the serious Personal Crime type Rape/sexual Assault, the model had an accuracy of 72.3% and the top influenctial variables influencing reporting are Weapon Category (weapcat), Medical Services for Physical Injuries (treatment), and Victim Services Received (vicservices). For Motor-vehicle Theft Random Forest predictive accuracy was 78.2% using Victim Services Received (vicservices) and Location (locationr). Adding the full model variables in all cases produced negligible additional benefit.

The Road to Reduced Crime



The stated four objectives are successfully met by this pilot project using available data from the FBI.

Understand differences between crime types and key insights from the data
Understand the context of reported and under-reported crimes to police.
Understand the key drivers causing the under-reporting of crimes to police.
Provide a likelihood of reporting metric to police for each identified crime.

- This project also created clarity on modeling & technical solution strategy around crime patterns & reporting to police which is a major leap into predictive analytics for the FBI.
- Some of the models we developed showed better classification accuracy than that are available in the literature (for example, rape/sexual assault & motor vehicle theft models).
- The FBI should leverage this for establishing baselines to track long term goals & continue to invest into predictive technologies & automate insights generation using AutoML.
- Resulting insights from severe crime type models including rape/sexual assault and motor vehicle theft should prompt further actions from the FBI
    - The groups identified as not reporting rape/sexual assault should be given special attention & care.
    - Mixed set of responses in other crime types (as acknowledged by poor classification performance) needs further attention into re-designing survey questions & sampling, for example creating custom questionnaires for different demographic groups and crime types.

Future recommendations to further improve the models:

- Future iterations of this project should include a budget to collaborate with subject matter experts so the wider raw data can be processed.
- More external linkages and restructuring of variables contained in the raw dataset should be explored to create additional variables for insight into reporting/non-reporting to police.
- The NCVS data lacks sufficient information for predicting report/non-report to police across all crime types. Some research indicates the NCVS raw data is defective due to complexity and what the data does not contain. Reported crime data should be obtained from regional police stations, as well as Census demographics by area to compare how accurate these surveys are in reporting actual crime rates and reporting. For example, if a given region shows crime reporting is relatively high from our data, is the crime rate also high from police data in those regions?
- The variable "Scrambled Control Number" contains information which may help improve model prediction capability. It contains the household address, city, and state. At present the only way crime can be explored geographically is through region and population density/type with respect to Metropolitan Statistical Areas. Crime patterns in an area of population density less than 10,000 within two MSA's in the region "West" can have vastly different crime rates and potentials based on particulars of city and demographics. For example, crime rates from Forest Park, one of the most violent areas in Detroit, Michigan, under the present constructs are mixed with those of Oakland Park, MI, one of the safest but similar sized cities in Michigan, the same region.
- Finally, partnership is suggested for designing surveys and data collection for analytics purposes.

Dashboard and Mobile Application

An Android Studio mobile app is being developed which allows the customer to view data and analytics related to the CrimeWatch project . With it the FBI will be able to view details about the data going into the study as well as the major contributors to why individuals do not report crimes against both themselves and their households.

Figure 29: Login Screen to the CrimeWatch app interface.

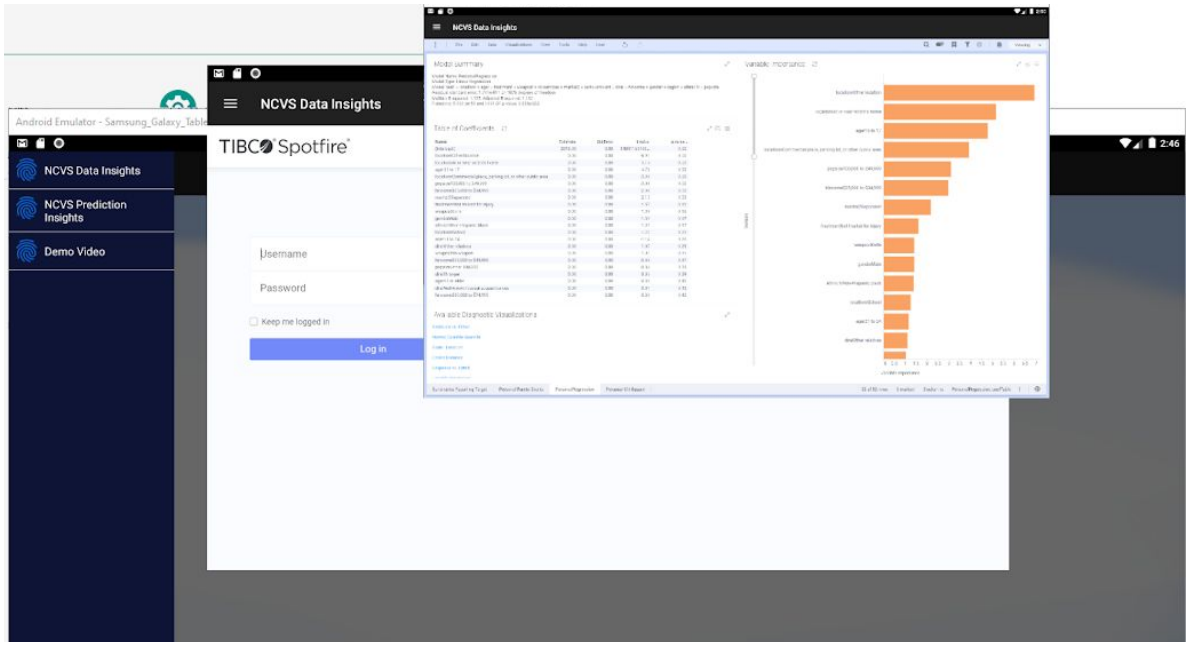
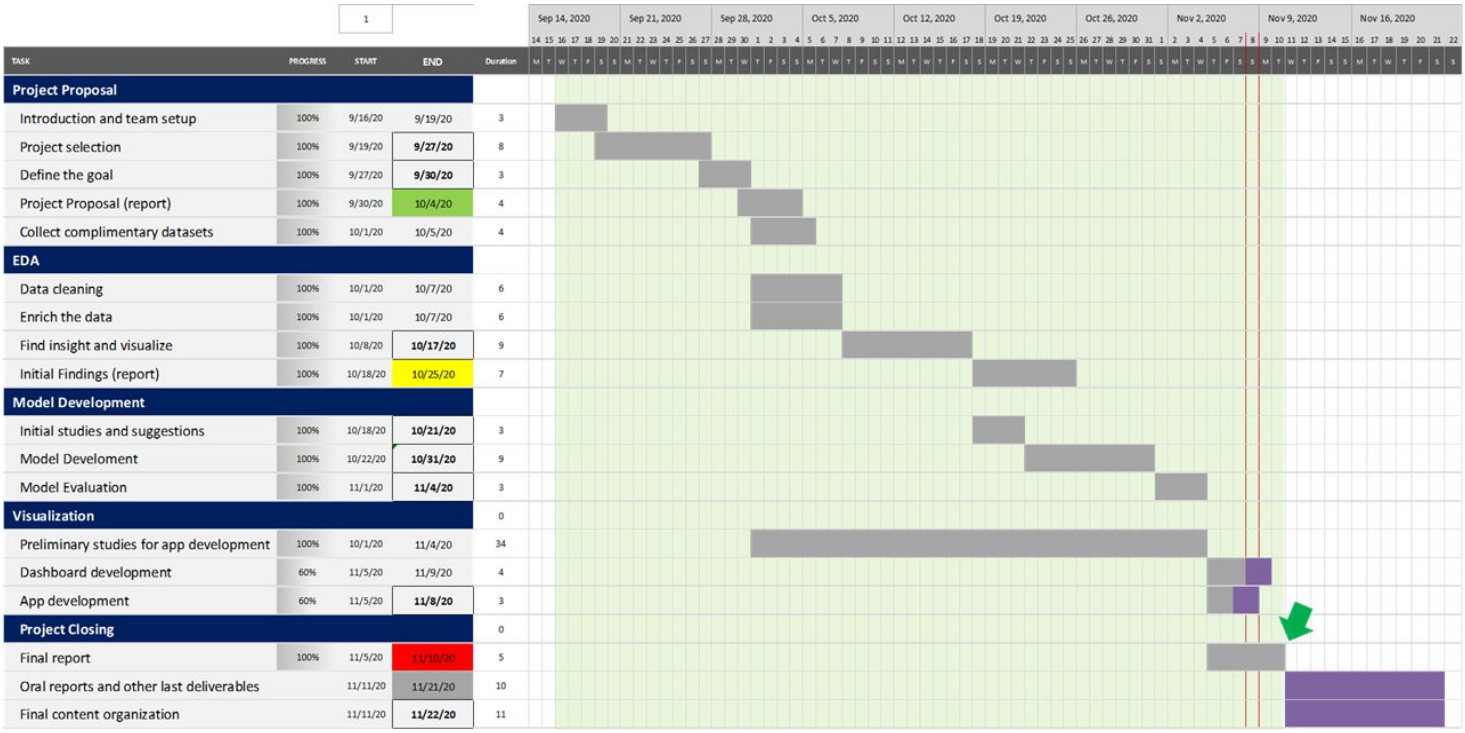


Figure 30: App Access to the CrimeWatch Data Dashboard.

Figure 31: From Mobile App: Samples of the CrimeWatch Dashboard

Project Management Summary

The project Gantt chart is presented below in Figure 26.



TASK	PROGRESS	START	END	Duration
Project Proposal				
Introduction and team setup	100%	9/16/20	9/19/20	3
Project selection	100%	9/19/20	**9/27/20**	8
Define the goal	100%	9/27/20	**9/30/20**	3
Project Proposal (report)	100%	9/30/20	10/4/20	4
Collect complimentary datasets	100%	10/1/20	10/5/20	4
EDA				
Data cleaning	100%	10/1/20	10/7/20	6
Enrich the data	100%	10/1/20	10/7/20	6
Find insight and visualize	100%	10/8/20	**10/17/20**	9
Initial Findings (report)	100%	10/18/20	10/25/20	7
Model Development				
Initial studies and suggestions	100%	10/18/20	**10/21/20**	3
Model Develoment	100%	10/22/20	**10/31/20**	9
Model Evaluation	100%	11/1/20	**11/4/20**	3
Visualization				0
Preliminary studies for app development	100%	10/1/20	11/4/20	34
Dashboard development	60%	11/5/20	11/9/20	4
App development	60%	11/5/20	**11/8/20**	3
Project Closing				0
Final report	100%	11/5/20	11/10/20	5
Oral reports and other last deliverables		11/11/20	11/21/20	10
Final content organization		11/11/20	**11/22/20**	11

Figure 32: Current Project Gantt Chart

As seen, the project is progressing according to the initial plan.

EDA, and Model development stages are completed successfully. Exploratory studies for development of the dashboard and mobile app are also done and final versions of them are under construction as a part of the final deliverables to FBI.

Approval

We approve the project as described above, and acknowledge the team completed this project successfully.

Name and Title: _____ Date: _____ Signature: _____

References

1. The Regents of the University of Michigan (2020). *National Crime Victimization Survey (NCVS) Series.* Bureau of Justice Statistics. https://www.icpsr.umich.edu/web/NACJD/series/95

2. Bureau of Justice Statistics (2020). *NCVS Victimization Analysis Tool (NVAT)*. Office of Justice Programs. https://www.bjs.gov/index.cfm?ty=nvat

3. Bachman, R. 1993. Predicting the Reporting of Rape Victimizations: Have Rape Reforms Made a Difference? Criminal Justice and Behavior, 20(3), 254–270. https://doi.org/10.1177/0093854893020003003

4. Baumer, E. P., Lauritsen, J. L. 2010. Reporting crime to the police, 1973-2005: multivariate analysis of long term trends in the national crime survey (ncs) and national crime victimization survey (ncvs). Criminology, 48(1), 131-186.

5. Baumer, Eric P., Richard B. Felson, and Steven F. Messner. 2003. Changes in police notification for rape, 1973-2005. *Criminology* 41:841-72.

6. Clay-Warner, Jody, and Callie Harbin Burt. 2005. Rape reporting after reforms: Have times really changed? *Violence Against Women* 11:150-76.

7. Felson, Richard B., and Paul-Philippe Pare. 2005. The reporting of domestic violence and sexual assaults by nonstrangers to the police. *Journal of Marriage and the Family* 67:597-610.

8. Felson, Richard B., and Paul-Philippe Pare. 2005. The reporting of domestic violence and sexual assaults by nonstrangers to the police. *Journal of Marriage and the Family* 67:597-610.

9. Gartner, Rosemary, and Ross Macmillan. 1995. The effect of victim- offender relationship on reporting crimes of violence against women. *Canadian Journal of Criminology* 37:393-429.

10. Gutierrez, Juliette and Leroy, Gondy, 2007. Predicting Crime Reporting with Decision Trees and the National Crime VictimizationSurvey. AMCIS 2007 Proceedings.Paper, 185.http://aisel.aisnet.org/amcis2007/185

11. Hart, Timothy C., and Callie Rennison. 2003. *Reporting Crime to the Police, 1992-2000.* Bureau of Justice Statistics Special Report. Washington, DC: U.S. Department of Justice.

12. Zychlinski, S. (2018 February 23). *The Search for Categorical Correlation*. Towards Data Science. https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9

Project Team

Damon Panahi. Project role: Project Manager

Damon works as a data scientist and Project Manager on the CrimeWatch team. He has more than 12 years experience in engineering and academic environments. Damon has led several product development projects for manufacturing companies like ArcelorMittal where he is currently working as a senior research engineer in the R & D department.

Sheila Cloudcroft. Project role: Data preparation and authentication, model development, Android app development, and content organization.

Sheila works as a senior data scientist engineer on the CrimeWatch team. She brings 23 years of experience working with data and modeling using high end statistical analysis to projects. She helps ensure data integrity in data inputs, and model accuracy in proposed solutions. She has advanced skills in predictive analytics, data science, machine learning, Android programming, and nearly 4 decades of experience in software development.

Setu Madhavi Namburu. Project role: EDA, model development, reporting and storytelling.

Setu works as a data scientist and story teller on the CrimeWatch team. Her primary responsibility is to keep the technical content flow organized while contributing to overall project execution. She has more than 12 years of experience in analyzing, modeling, deriving data-driven insights and helping business leaders with strategic decision making in various business areas from research to warranty/service to telematics to supply chain demand analytics in automotive industry.

Jeremy Melville. Project role: Reporting and storytelling, data visualization and dashboard development.

Jeremy works as a Data Scientist on the CrimeWatch team. He helps manage the reporting content and contributes to the overall analytics. Jeremy has almost 25 years of experience working in the field of data science and artificial intelligence. For the first 15 years of his career he worked as a consultant at SPSS, now an IBM company, before starting his own consulting company. More recently he has worked at TIBCO and CrimeWatch.

Jin Choi. Project Role: Data preparation, content organization, EDA, model development, data visualization and dashboard development.

Jin works as a data scientist on the CrimeWatch Team. Jin is a data scientist engineer with 8 years of experience working in the field of data science, analytics modeling, and data visualization. He has worked at the Wells Fargo bank as an Analytic consultant handling consumer lending data with SQL, SSIS, Teradata, SAS, and Tableau.

Appendix A

BOJ Statistics National Crime Victimization Survey (NCVS) RESTful API Datasets

Personal Victimization Dataset

Variable_Name	Variable_ID	No_Distinct_Values	Sample_Values
Age	ager	8	['12 to 14' '15 to 17' '18 to 20' '21 to 24' '25 to 34' '35 to 49' '50 to 64' '65 or older']
Victim-offender relationship	direl	6	['Do not know number of offenders' 'Do not know relationship' 'Intimates' 'Other relatives' 'Stranger' 'Well-known/casual acquaintances']
Race/Hispanic origin	ethnic1r	4	['Hispanic' 'Non-Hispanic black' 'Non-Hispanic white' 'Non-Histpanic other']
Sex	gender	2	['Female' 'Male']
Household income	hincome	8	['$15,000 to $24,999' '$25,000 to $34,999' '$35,000 to $49,999' '$50,000 to $74,999' '$7,500 to $14,999' '$75,000 or more' 'Less than $7,500' 'Unknown']
Hispanic origin	hispanic	3	['Hispanic' 'Non-Hispanic' 'Unknown']
Injury	injury	2	['Missing' 'Not injured']
Location of incident	locationr	5	["At or near friend, neighbor, or relative's home" "At or near victim's home" 'Commercial place, parking lot, or other public area' 'Other location' 'School']

Marital status	marital2	6	['Divorced' 'Married' 'Never married' 'Separated' 'Unknown' 'Widowed']
Location of residence	msa	3	['Not part of principal city within MSA' 'Outside MSA' 'Principal city within MSA']
Aggregate type of crime	newcrime	2	['Personal theft/larceny' 'Violent victimization']
Type of crime	newoff	5	['Aggravated assault' 'Personal theft' 'Rape/sexual assault' 'Robbery' 'Simple assault']
Reporting to the police	notify	4	['8' 'Do not know' 'No, did not report to the police' 'Yes, reported to the police']
notifyTarget	notifyTarget	2	['no' 'yes']
Population size	popsize	6	['1 million or more' '100,000 to 249,999' '250,000 to 499,999' '500,000 to 999,999' 'Not a place' 'Under 100,000']
Race	race1r	3	['Black' 'Other' 'White']
Region	region	5	['Midwest' 'Northeast' 'South' 'Unknown' 'West']
Violent crime excluding simple assault	seriousviolent	3	['Personal theft' 'Simple assault' 'Violent crime excluding simple assault']
Medical treatment for physical injuries	treatment	3	['Missing' 'Not injured' 'Not treated for injury']

Victim services	vicservices	3	['Missing' 'No services received from victim service agencies' 'Services received from victim service agencies']
Weapon category	weapcat	6	['Do not know if offender had weapon' 'Firearm' 'Knife' 'No weapon' 'Other type weapon' 'Type weapon unknown']
Presence of weapon	weapon	3	['Do not know if offender had weapon' 'No, offender did not have weapon' 'Yes, offender had weapon']
Weight	weight	50146	[5.98210600e+01 6.02814300e+01 6.09631300e+01 ... 1.01508901e+05 1.34254865e+05 1.53486648e+05]
Year	year	27	[1993. 1994. 1995. 1996. 1997. 1998. 1999. 2000. 2001. 2002. 2003. 2004. 2005. 2006. 2007. 2008. 2009. 2010. 2011. 2012. 2013. 2014. 2015. 2016. 2017. 2018. 2019.]

Table Apx-A1. Personal Victimization Data Description: BoJ Statistics Dataset

Household Victimization Dataset

Variable_Name	Variable_ID	No_Distinct_Values	Sample_Values
Head of household age	hhage	5	['19 or younger' '20 to 34' '35 to 49' '50 to 64' '65 or older']
Race/Hispanic origin of head of household	hheth1r	4	['Hispanic' 'Non-Hispanic black' 'Non-Hispanic other' 'Non-Hispanic white']

Sex of head of household	hhgen	2	['Female' 'Male']
Hispanic origin of head of household	hhhisp	3	['88' 'Hispanic' 'Non-Hispanic']
Race of head of household	hhrace1r	3	['Black' 'Other' 'White']
Household income	hincome	8	['$15,000 to $24,999' '$25,000 to $34,999' '$35,000 to $49,999' '$50,000 to $74,999' '$7,500 to $14,999' '$75,000 or more' 'Less than $7,500' 'Unknown']
Household size	hnumber	4	['Four to Five' 'One' 'Six or more' 'Two to Three']
Location of incident	locationr	5	["At or near friend, neighbor, or relative's home" "At or near victim's home" 'Commercial place, parking lot, or other public area' 'Other location' 'School']
Location of residence	msa	3	['Not part of principal city within MSA' 'Outside MSA' 'Principal city within MSA']
Aggregate type of crime	newcrime	1	['Property Victimization']
Type of crime	newoff	3	['Burglary/trespassing' 'Motor-vehicle theft' 'Theft']

Reporting to the police	notify	4	['8' 'Do not know' 'No, did not report to the police' 'Yes, reported to the police']
notifyTarget	notifyTarget	2	['no' 'yes']
Population size	popsize	6	['1 million or more' '100,000 to 249,999' '250,000 to 499,999' '500,000 to 999,999' 'Not a place' 'Under 100,000']
Region	region	5	['Midwest' 'Missing' 'Northeast' 'South' 'West']
Victim services	vicservices	3	['Missing' 'No services received from victim service agencies' 'Services received from victim service agencies']
Weight	weight	161328	[3.37986500e+01 5.39848400e+01 5.71586100e+01 ... 6.66650982e+04 7.10948084e+04 7.97670744e+04]
Year	year	27	[1993. 1994. 1995. 1996. 1997. 1998. 1999. 2000. 2001. 2002. 2003. 2004. 2005. 2006. 2007. 2008. 2009. 2010. 2011. 2012. 2013. 2014. 2015. 2016. 2017. 2018. 2019.]

Table Apx-A2. Household Victimization Data Description: BoJ Statistics Dataset

Appendix B

Data Understanding

This project uses survey data collected from the National Crime Victimization Survey (NCVS). The NCVS was designed with four primary objectives: (1) to develop detailed information about the victims and consequences of crime, (2) to estimate the number and types of crimes not reported to the police, (3) to provide uniform measures of selected types of crimes, and (4) to permit comparisons over time and types of areas.

Data Source

The National Crime Victimization Survey (NCVS) has been collecting data on personal and household victimization through an ongoing survey of a nationally-representative sample of residential addresses since 1973. Online sources note that the NCVS has undergone many changes and sampling adjustments over the years. To ensure homogenous data was being used, we conducted statistical significance tests. These tests did not yield consistent results, which is attributed to underlying data complexity. The last 6 years worth of data (2013-2019) was chosen as it is homogenous within both the datasets and is considered of sufficient span and relevance to meet project goals.

From the product documentation:

- Study Sample: NCVS RESTful API 2010-2019, truncated to 2013-2019, a pre-processed subset of ICPSR 37689 V1, NCVS [United States] 1992-2019, Concatenated File United States. Bureau of Justice Statistics released 2020-09-21
- **Data Collector:** United States Department of Commerce. Bureau of the Census
- **Mode of Collection:** telephone interviews, computer-aided telephone interviews, face-to-face interviews
- **Universe:** Individuals 12 years of age and older living in households and group quarters within the United States and the District of Columbia. Excluded are persons who are crews of vessels, in institutions (e.g., prisons and nursing homes) or members of the armed forces living in military barracks. Interviews are translated for non-English speaking respondents.
- **Time Method:** Rotating panel survey. Once in sample, respondents are interviewed every six months for a total of seven interviews over a three-year period. After the seventh interview the household leaves the panel and a new household is rotated in.
- **Frequency of Data Collection:** Semi-annually
- File Dimensions: Household No. of Cases: 69,400 records in 17 variables ; Person: 17515 records in 23 variables.
- **Year Reported:** Victimizations are counted in the year the interview is conducted, regardless of the year when the crime incident occurred.

Data Overview

As stated in document section "Clarification of Scope and Limitations", a new reduced, pre-processed dataset created and maintained for NCVS victimization analysis tool was pivoted to for analysis purposes.

The new dataset does not have a unique identifier so it is impossible to relate the reduced records with other information in the raw dataset. The truncated work with raw data is summarized in Appendix C.

In the new dataset there are two main categories of data. Each has two csv files - one for population and one for victimization incidents. Our study focused on personal and household victimization files only as our mission was to identify patterns/clusters for under reporting of crime to the police. Unweighted data was used to identify crime patterns from reported incidents.

Each household/personal victimization and general/household population record is recorded at a year level and has a weight assigned to each record to estimate national victimization levels. These weights are shown below in Figures Apx-B2 & Apx-B3.

Sampling Procedure

Each month the U.S. Census Bureau randomly selects about 160,000 unique persons in about 95,000 households for the NCVS using a "rotating panel" sample design. They conduct approximately 240,000 interviews on criminal victimization of all age-eligible individuals who become part of the panel.

Weighting

Components of the NCVS weights	Household-level estimates			Person-level estimates		
	Household	Victimization	Incident	Person	Victimization	Incident
Base weight	×	×	×	×	×	×
GQ subsampling adjustment	×	×	×	×	×	×
Household nonresponse	×	×	×	×	×	×
Within-household nonresponse				×	×	×
Ratio adjustment	×	×	×	×	×	×
Bounding adjustment		×	×		×	×
TIS adjustment		×	×		×	×
Series crime adjustment		x	x		x	x
Multiple victim adjustment						×

Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2016.

Weighting is available in this study on both the personal and household level to adjust the sample to be more representative of the population. These weights will be used as often as possible to reduce bias due to undercoverage of various portions of the population. The weighting is calculated counting for estimates of crime experienced by the U.S. population age 12 or older and then several adjustments are made to account for sampling nonresponse bias and differences in distributions across demographics. Figure Apx-B1 describes the adjustments made on the household and person-level estimates. Weights found in the two main datasets contain these target weights for each level.

Figure Apx-B1: Treatments and adjustments made to weighting to better reflect the true population

Weight	Personal Victimization
Value	**Description**
Population	This weight is attached to the person population file and is used to calculate an estimate of persons covered by the NCVS. In a calculation of person victimization rate, the weight is used to determine the denominator.
Victimization	The weight used to calculate an estimate of victimizations. In a calculation of victimization rate, they are used to determine the numerator. This weight also accounts for high-frequency repeat victimizations, or series victimizations, which are six or more similar but separate victimizations that occur with such frequency that the victim is unable to recall each individual event or describe events in detail. BJS has decided to count series victimizations using the victim's estimate of the number of times the victimizations occurred during the prior 6 months, capping the number within each series at a maximum of 10 victimizations. Including series victimizations in national estimates can substantially increase the number and rate of violent victimization. However, trends in violence are generally similar regardless of whether series victimizations are included.

Figure Apx-B2: Personal Victimization Weight Description.[2]

Weight	Household Victimization
Value	**Description**
Household	The weight is attached to the household population file and is used to calculate an estimate of households covered by the NCVS. In a calculation of household victimization rate, the weight is used to determine the denominator.
Victimization	The weight used to calculate an estimate of victimizations. In a calculation of victimization rate, they are used to determine the numerator. This weight also accounts for high-frequency repeat victimizations, or series victimizations, which are six or more similar but separate victimizations that occur with such frequency that the victim is unable to recall each individual event or describe events in detail. BJS has decided to count series victimizations using the victim's estimate of the number of times the victimizations occurred during the prior 6 months, capping the number within each series at a maximum of 10 victimizations. Including series victimizations in national estimates can substantially increase the number and rate of violent victimization. However, trends in violence are generally similar regardless of whether series victimizations are included.

Figure Apx-B3: Household Victimization Weight Description.[3]

[2] Bureau of Justice Statistics, Office of Justice Programs, National Crime Victimization Survey (NCVS) API, retrieved from: https://www.bjs.gov/developer/ncvs/personalFields.cfm.

[3] Bureau of Justice Statistics, Office of Justice Programs, National Crime Victimization Survey (NCVS) API, retrieved from: https://www.bjs.gov/developer/ncvs/householdFields.cfm.

Variable Selection Methods

Analysis of all the existing variables in the NCVS through regression (or other similar approaches) is not possible. In fact, these type of evaluations through descriptive statistics require some domain knowledge to manually select some variables for more detailed understanding (e.g., Baumer, Felson, and Messner, 2003; Clay-Warner and Burt, 2005; Felson and Pare, 2005; Gartner and Macmillan, 1995).

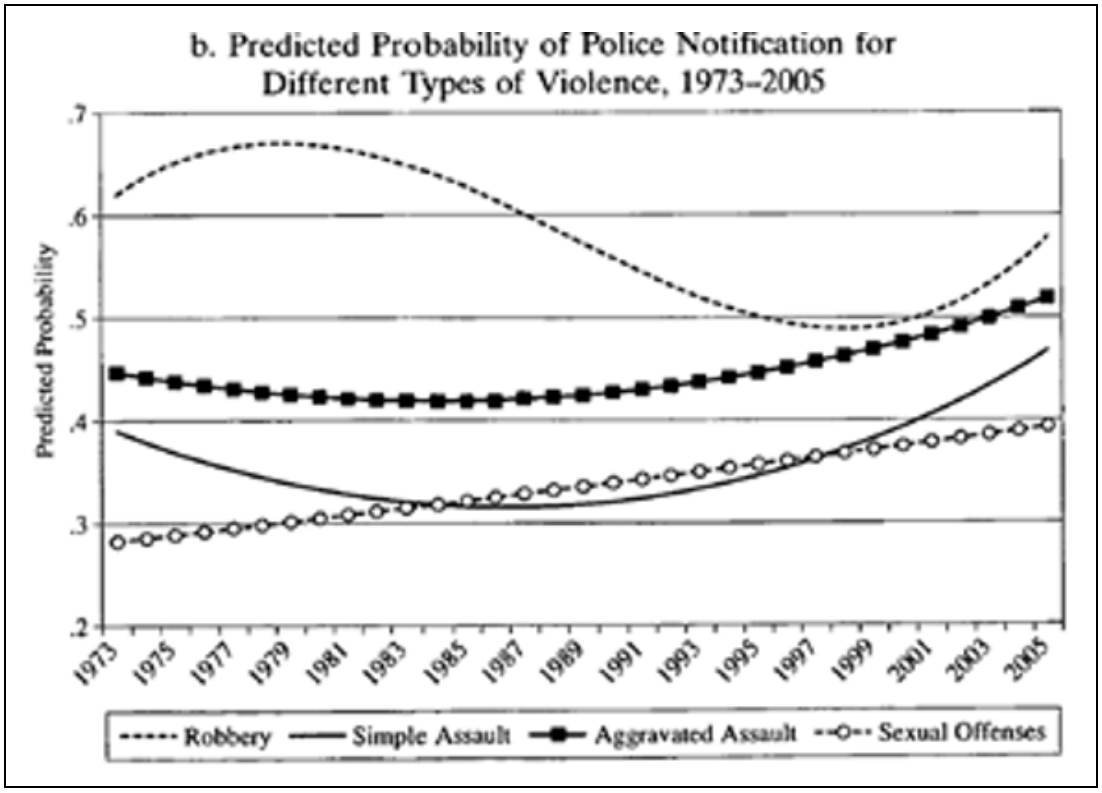


Figure Apx-B4: Trends for Probability of Police Notifications for Some of the Investigated Crime Types.

As a result, outcomes of these approaches yield an incomplete picture of the problem. In other words, there still might be many more variables that influence the decision making process of a victim (or third-party people) to report a crime to the police.

That is why selective processing approaches using filters and wrappers, have been explored by different researchers as an alternative way for analyzing the existing variables in the crime data sets.

Juliette et. al. (Juliette et. al. 2007) compared results of different decision trees including trees built based on manually selected variables through domain knowledge, and those built by automatic selection of variables through Chi-squared filter, Cramer's V Coefficient filter, and a forward selection wrapper. Results showed that automatic selection of variables may show unexpected variables important for understanding reporting of a crime.

Figure Apx-B5 shows the most important variables based on the employed selection methods.

As seen, each method discovers a different set of variables. These types of observations can potentially identify new areas for further research, allocation of resources and direction for law enforcement.

Forward selection	Cramer's V	Chi-squared
Description	**Description**	**Description**
Type of Crime Code	Help From Victim's Agencies?	Help From Victim's Agencies?
Activity at Time of Incident	Was Victim Agency Government or Private?	Was Victim Agency Government or Private?
Single Offender: How Did Respondent Know Offender?	Incident Occur at Work Site?	Incident Occur at Work Site?
Check B: Attack, Threat, Theft	Job Located in City/Suburb/Rural Area?	Job Located in City/Suburb/Rural Area?
Which Best Describes Your Job?	Is the Business Incorporated?	Is the Business Incorporated?
How Many Times Incident Occurred?	Type of Industry at Time of Incident?	Type of Industry at Time of Incident?
Total Number Days Lost?	Anything Damaged?	Usually Work Days or Nights?
Help From Victims Agencies?	Usually Work Days or Nights?	Type of Crime Code
Age (Allocated)	Type of Crime Code	Where Did Incident Happen?
Anything Damaged?	Where Did Incident Happen?	Total Amount of Medical Expenses?
Covered By Medical Insurance	Total Amount of Medical Expenses?	Any Others Harmed or Robbed?
Number of Household Members Harmed/Robbed	How Other's Action Helped?	Number of Others Harmed or Robbed?
Stolen, Attack, Threat: Offender Known	How Attacked?	Activity at Time of Incident?
Thought Crime But Didn't Call Police	Residue: How Other's Helped	Medical Care: Emergency Room
Something Taken? (Allocated)	Any Others Harmed or Robbed?	Is the Business Incorporated?
Number of Others Harmed or Robbed	Number of Others Hurt or Robbed?	One or More Than One Offender?
	Residual: Medical Care	Medical Care: Home, Neighbor's, Friends
	How Other's Actions Worsened Situation	Current Job?
	Activity at Time of Incident	Attempt/Threat: Weapon Present?
	How Other's Action Hurt	How Offender Threatened or Tried to Attack

Figure Apx-B5: Most Important Variables Based on the Employed Selection Methods.

Accuracy of the predictions for different variable selection methods are also summarized in Figure Apx-B6 below.

Method	Leaves	Size	Overall Accuracy	Root Node	Description	Accuracy YES	Accuracy NO	Accuracy MAYBE
Chi-squared	1,075	2,149	66%	V4130	Medical Care: Home, Neighbor's, Friends	7,748/12,838 = 66%	10,473/14,427 = 73%	12/350 = 3%
Cramer's V	1,003	2,005	66%	V4136	Residue: Medical Care Site	7,801/12,838 = 61%	10,400/14,427 = 72%	9/350 = 3%
Forward Selection	678	1,355	69%	V4498	Total Number Days Lost	8,242/12,838 = 64%	10,755/14,427 = 75%	15/350 = 4%
Domain Knowledge	126	251	64%	V4529	Type of Crime Code	7,948/12,838 = 62%	9,860/14,427 = 68%	0/350 = 0%

Figure Apx-B6: Accuracy for Different Variable Selection Methods.

As seen, selection of variables based on domain knowledge gives the lowest accuracy (~64%) while the "Forward Selection" approach gives the highest accuracy (~69%) among these methods.

In general, looking at these studies, it seems that employing these types of automatic variable selection methods can be an interesting approach which is worth exploring further. Many substantial social, legal, technological and cultural shifts that have occurred in the United States during the past 5-10 years make

expansion of these studies to datasets constrained to this time period another interesting focus for future research.

Exploratory Data Analysis (EDA)

Data Analytics Tools

We used Python, R, Tableau, Power BI, TIBCO Statistica, Minitab, and Spotfire to explore and visualize this data. The findings for household victimization and personal victimization datasets are reported separately in below sections.

Thorough EDA is conducted by visualizing counts over the years and slicing the data by different variables and their categories. The counts shown in Figure Apx-B7 indicate sample population. Although it appears there may be a decline overall, to get true population representation the counts need to be multiplied by weights for each category of data. One insight from these plots is that the personal incidents dataset seems to be fairly balanced in terms of reporting to police as compared to household incidents.

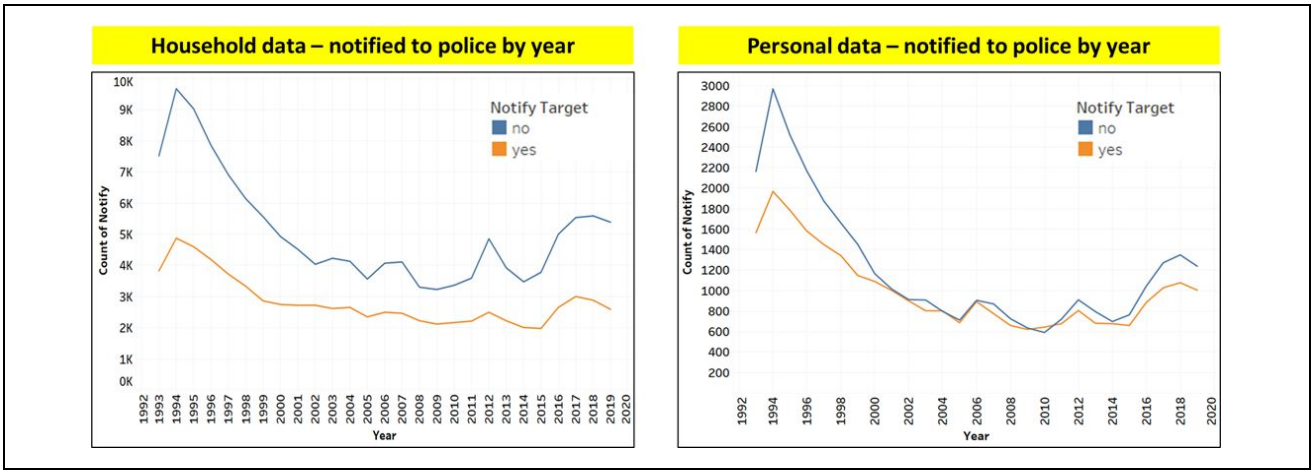


Figure Apx-B7: Count Plots of Household Victimization and Personal Victimization Notified to Police.

Data Profiling

Except the year and weight columns, all other variables are categorical. The csv's contain categorical variables in numerical value form (eg. (2) Female is '2') with definitions provided on the API website. The attribute values were replaced by their labels using the ETL tool in Spotfire to make the data more readable and facilitate interpretation while analyzing it (see Appendix A).

Personal Victimization

Personal victimization includes all violent victimization (rape or sexual assault, robbery, aggravated assault, simple assault) and personal theft. The dataset for it has 23 variables. We added an additional variable called notifyTarget based on the variable 'notify' to get quick gist of the data. A list of variables, their descriptions and sample values are provided in Appendix A.

Personal victimization data has 60034 observations. A preview of this data is shown below in Figure Apx-B8.

year	weight	gender	race1R	hispanic	ethnic1R	ager	marital2	hincome	popsize	region	msa	direl	notify	weapon	weapcat	newcrime	newoff
2019	3313.01957	2	1	2	1	3	1	88	1	2	3	4	2	3	5	1	4
2019	3313.01957	2	1	2	1	3	1	88	1	2	3	4	2	2	0	1	4
2019	3313.01957	2	1	2	1	3	1	88	1	2	3	4	2	3	5	1	4
2019	1221.06481	2	1	2	1	7	4	88	5	4	1	4	1	2	0	1	4
2019	1221.06481	2	1	2	1	7	4	88	5	4	1	4	1	2	0	1	1

seriousviolent	injury	treatment	vicservices	locationr
2	0	0.0	2.0	3
2	0	0.0	2.0	3
2	0	0.0	2.0	3
2	0	0.0	2.0	3
1	0	0.0	2.0	3

Figure Apx-B8: Personal Victimization Data Preview

Figure Apx-B9 provides a summary of the data and the frequency of observations by year. While the report shows no missing cells, there are other ways in which missing variables are coded (e.g., unknown, not available etc.). We will treat missingness as another coded category rather than a missing value for impute.

The Personal Victimization data shows 5.6% duplicate rows but we understood from checking the raw data that the same individuals reported very similar incidents that happened multiple times with differences in other attributes unavailable in the new dataset. An example of this is shown in Appendix B. We did not remove any duplicate rows as it can lead to underrepresented reported incident counts.

Dataset statistics

Number of variables	24
Number of observations	60034
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	3379
Duplicate rows (%)	5.6%
Total size in memory	11.0 MiB
Average record size in memory	192.0 B

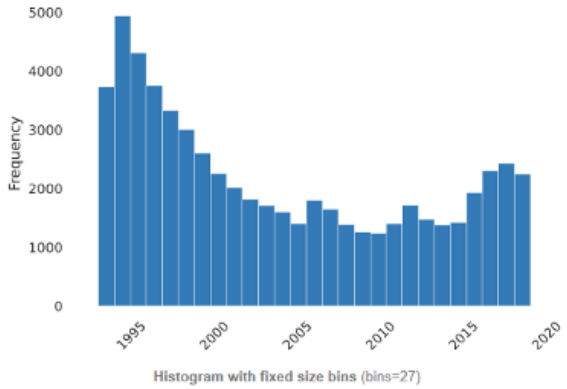


Figure Apx-B9: Summary of Personal Victimization data and number of observations over years

Household Victimization

Household victimization includes all property victimization (burglary/trespassing, motor-vehicle theft, and theft). Household victimization data has 17 variables of which non-incident related variables are also available in the personal population data. We added an additional variable called notifyTarget based on the variable 'notify' to get quick gist of the data. The list of variables, their descriptions and sample values are presented in Appendix A.

Household victimization data has 216269 observations. A preview of this data is shown in Figure Apx-B10.

year	weight	msa	hincome	hhage	hhgen	hhhisp	hhrace1r	hheth1r	hnumber	popsize	region	notify	newcrime	newoff	vicservices	locationr
2019	432.34504	1	7	5	1	2	1	1	2	4	4	1	3	8	2.0	1
2019	1733.58364	1	7	3	1	2	1	1	2	2	2	2	3	8	2.0	1
2019	449.44243	1	7	3	2	2	1	1	3	1	4	1	3	8	1.0	1
2019	495.34413	1	5	2	2	2	1	1	3	2	4	2	3	8	2.0	3
2019	476.64194	1	7	4	2	2	1	1	3	4	4	1	3	8	2.0	1

Figure Apx-B10: Household Victimization data preview.

Figure Apx-B11 provides a summary of the data and the frequency of observations by year. Further work on Household profiles provides detailed statistics about each individual variable in the data.

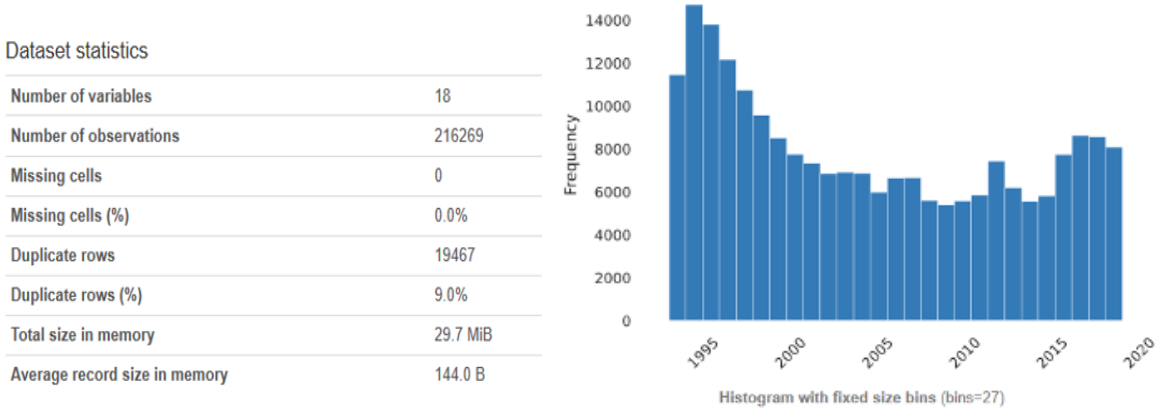


Figure Apx-B11: Summary of Household Victimization data and number of observations over the years

The Household Victimization data shows 9% duplicate rows but we understood from checking the raw data that the same household/individuals reported very similar incidents that happened multiple times with differences in other attributes unavailable in the new dataset. An example of this is shown in Appendix B. We did not remove any duplicate rows as it can lead to underrepresented reported incident counts.

As can be seen in both household and personal victimization counts against years, there is an apparent decline in the number of incidents over the years. This likely shows efficacy in measures implemented by FBI and perhaps the influence of cultural and technological changes.

Note: DOJ NCVS documents indicate a significant change in survey methodology occurred in 2016. Although outside the scope of this project, further study of the variation in reported incidents to understand the drivers behind unreported crime to police (perhaps two segments: 2003 to 2015 where the counts looked similar; and 2016 onwards) might be useful to determine if there are similar clusters over the years.

Count of reported and unreported personal and household victimization incidents by different categorical variables are shown in Figure Apx-B12 and Apx-B13 respectively. While there are imbalances in the counts of reported vs unreported in each of the sub-categories, there is not a single variable or two that can help us understand non-report incidents without employing advanced techniques. This is further verified by visualizing personal victimization data using combinations of variables.

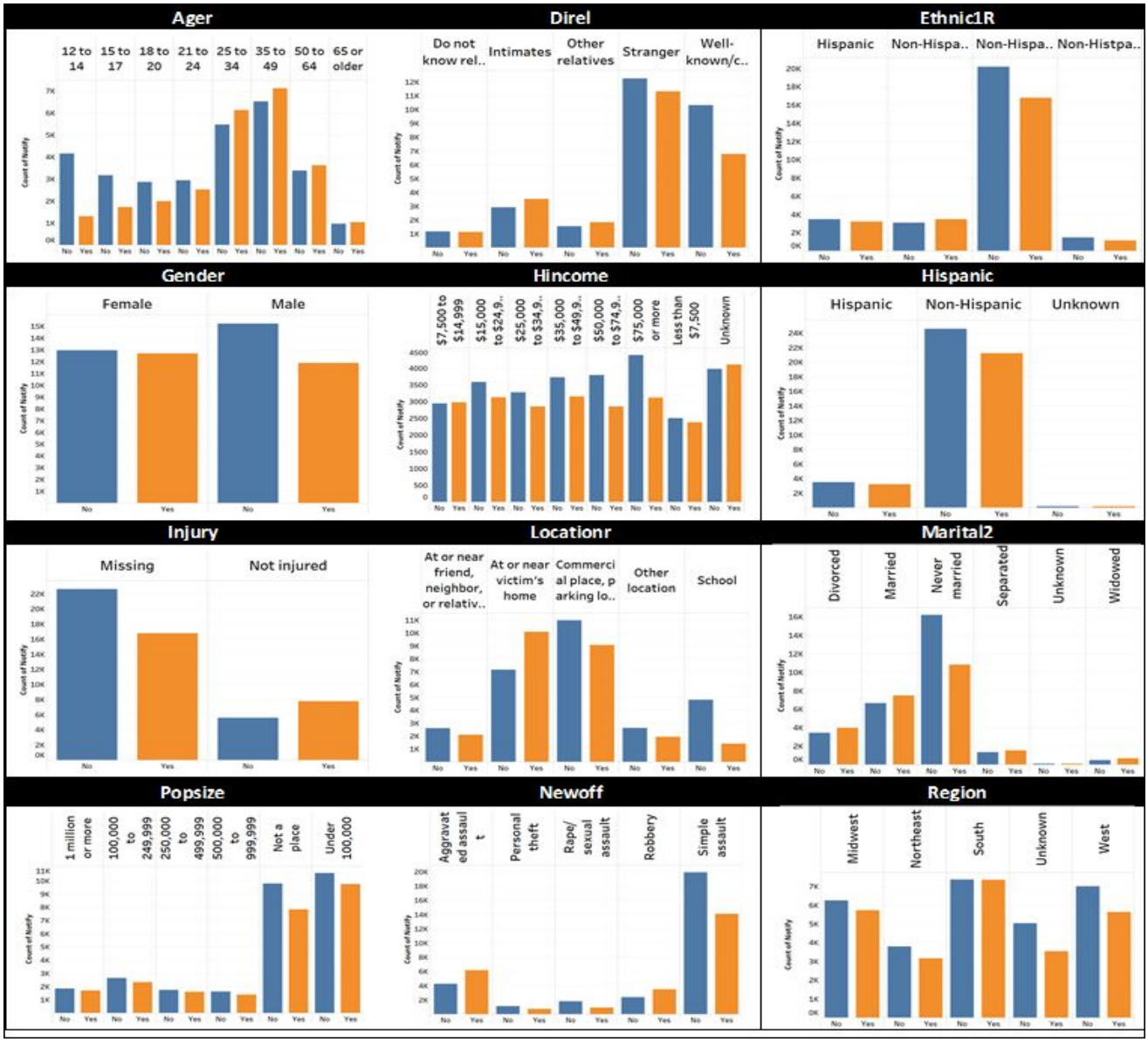


Figure Apx-B12: Counts of Reported and Unreported Personal Victimizations by Categorical Variables.

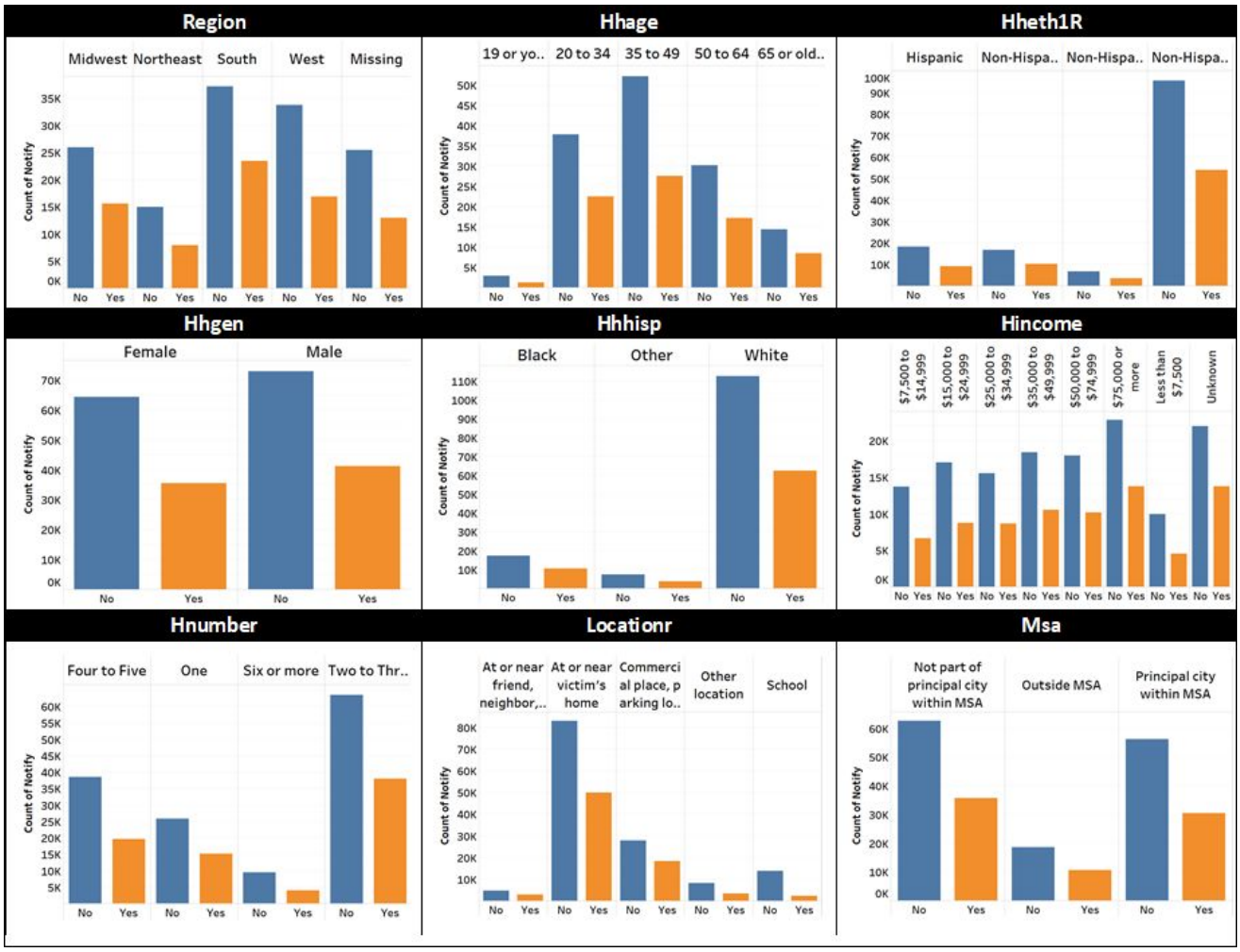


Figure Apx-B13: Counts of Reported and Unreported Household Victimizations by Categorical Variables.

During the modeling phase, clustering and classification experiments can be conducted using all or few variables to understand their significance. Some variables grouped sub-levels captured by other variables into higher levels (e.g., weapon, weaponcat) - this may help achieve better classification accuracy.

Figure Apx-B14 shows trends of total number of personal victimizations not reported per age group of the victim.

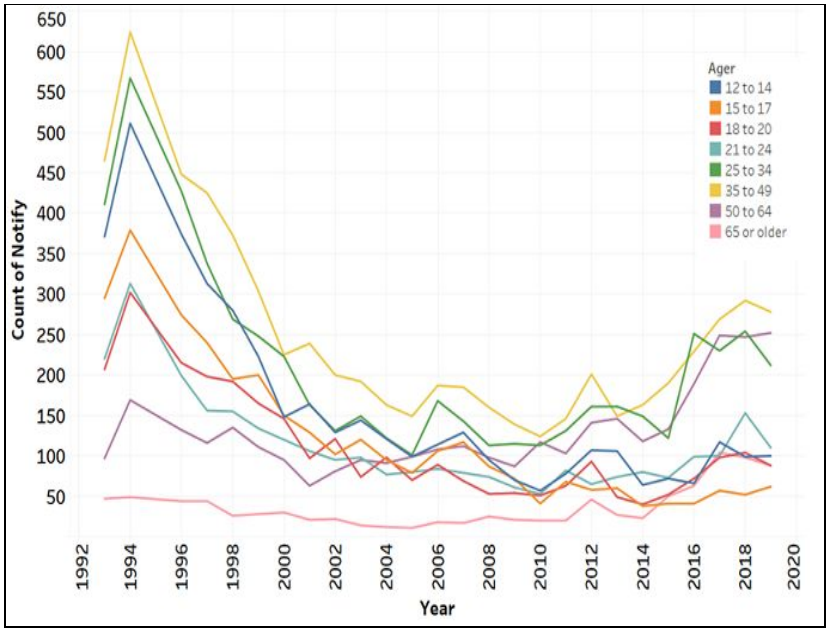


Figure Apx-B14: Trends of Total Number of Personal Incidents that are not Reported per Age Group.

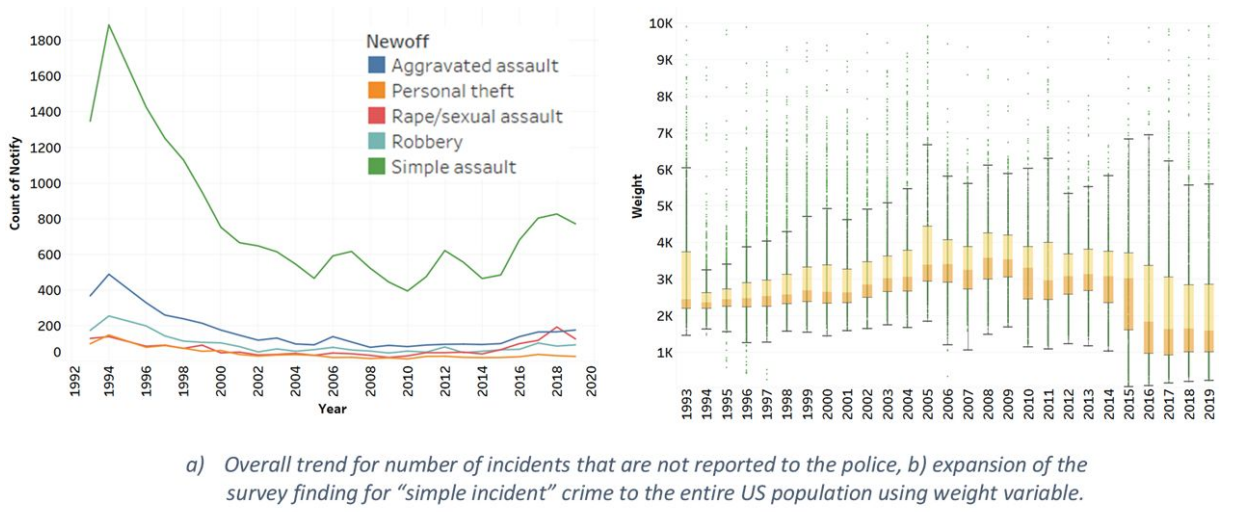
Figure Apx-B15 shows counts of Personal Crime victimizations reported vs unreported by region and type of crime.

Newoff	Region	8	Do not know	No, did not report to..	Yes, reported to the ..
Aggravated assault	Midwest	1	26	850	1,314
	Northeast	2	20	468	661
	South	8	38	1,127	1,850
	Unknown	2	32	1,194	1,388
	West	3	49	987	1,415
Personal theft	Midwest	1	2	198	169
	Northeast		6	227	136
	South		5	239	189
	Unknown		3	341	159
	West	1	3	202	124
Rape/sexual assault	Midwest	10	3	428	221
	Northeast	1		217	90
	South	4	7	466	281
	Unknown		1	373	147
	West	3	5	438	219
Robbery	Midwest	4	4	417	763
	Northeast		9	310	457
	South	3	12	609	1,050
	Unknown		10	602	751
	West	1	15	610	724
Simple assault	Midwest	25	118	4,404	3,291
	Northeast	16	60	2,624	1,849
	South	29	120	4,988	4,028
	Unknown	2	112	4,772	2,736
	West	20	125	4,821	3,189

Figure Apx-B15: Counts of Personal Victimizations Reported vs Unreported by Region and Type of Crime.

The decline in number of incidents can be further verified by observing year to year analysis of the Personal Crime victimization types as shown in Figure Apx-B15. It also shows very interesting trends for reported and unreported incidents to police. For example, graph (a) in Figure Apx-B16 below, one may see that the number of "unreported" incidents in different crime categories (specially for "simple assault" crimes) has decreased

between 1993 to 2015. However, extrapolation of the survey results to the entire U.S. population using the "weight" variable in the dataset (graph (b)), shows a rather different story suggesting an overall increase in the number of incidents that were not reported to the police. Interestingly, these trends have been reversed after 2015 when probably better policing and data collection resulted in an improvement in reporting of different crimes (especially "simple assault" crimes) to the police. Since our goal of the project is to understand the patterns in crime incidents, we will continue to use unweighted data throughout this project.



a) Overall trend for number of incidents that are not reported to the police, b) expansion of the survey finding for "simple incident" crime to the entire US population using weight variable.

Figure Apx-B16: Year to Year Analysis of the Personal Victimization Types.

Potential Duplication of Records in Data

There was a question of whether the new dataset contained duplicate records. Using the raw dataset it was shown that incidents which appeared to be duplicates in the new dataset yet were actually multiple similar incidents which differed in other variables not included in the new dataset. Figure Apx-B17 reveals the existence of such duplicates.

YEARQ	2016.3	2016.3	2016.3	2016.3	2016.3	
IDHH	30120368253917675634851 36	30120368253917675634851 36	30120368253917675634851 36	30120368253917675634851 36	30120368253917675634851 36	
IDPER	30120368253917675634851 3601	30120368253917675634851 3601	30120368253917675634851 3601	30120368253917675634851 3601	30120368253917675634851 3601	
V4002	35110	35110	35110	35110	35110	
V4003	(163) 2016 3rd quarter	(163) 2016 3rd quarter	(163) 2016 3rd quarter	(163) 2016 3rd quarter	(163) 2016 3rd quarter	
V4004	30	30	30	30	30	
V4005	12036825391767563485	12036825391767563485	12036825391767563485	12036825391767563485	12036825391767563485	
V4006	1	1	1	1	1	
V4008	36	36	36	36	36	
V4009	1	1	1	1	1	
V4010	1	1	1	1	1	
V4011	(36) 36:Indiv scrn quest	(36) 36:Indiv scrn quest	(39) 39:Hhld scrn quest	(39) 39:Hhld scrn quest	(98) Residue	Screen Q which prompted report
V4012	1	2	1	2	1	Incident number
V4014	(04) April	(04) April	(05) May	(04) April	(06) June	Month incident occurred
V4015	2016	2016	2016	2016	2016	
V4016	2	2	3	2	1	How many times incident occur last 6 months
V4017	(1) 1-5 incidents	(1) 1-5 incidents	(1) 1-5 incidents	(1) 1-5 incidents	(1) 1-5 incidents	Check Item B: How many incidents?
V4021B	(07) Aft 12pm-6am	(07) Aft 12pm-6am	(07) Aft 12pm-6am	(07) Aft 12pm-6am	(07) Aft 12pm-6am	
V4022	(3) Same city etc	(3) Same city etc	(3) Same city etc	(3) Same city etc	(3) Same city etc	
V4023B	(2) No	(2) No	(2) No	(2) No	(2) No	
V4024	(16) Park-noncomm	(05) N/hme-own yrd	(02) R/hme-det bldg	(02) R/hme-det bldg	(05) N/hme-own yrd	Where did the incident happen?
V4025			(2) No	(2) No		Did offender have a right to be there?
V4026			(1) Yes	(1) Yes		Did offender get inside?
V4028			(2) No	(2) No		Evidence of forcible entry
V4040			(98) Residue	(98) Residue		How did offender get in?
V4041C	(1) Open to the pub					Was area open to public?
V4042	(1) Indoors	(2) Outdoors			(2) Outdoors	Incident occurred inside, outside, or both?

Figure Apx-B17: Duplicate Data Verified As Distinct Records Using Full Dataset

Figures Apx-B18 and Apx-B19 show transformation in Household Income between datasets.

Household income (hincome)

The total income of the household head and all members of the household for the 12 months preceding the interview. Includes wages, salaries, net income from businesses or farms, pensions, interest, dividends, rent, and any other form of monetary income.

Value	Description
1	Less than $7,500
2	$7,500 to $14,999
3	$15,000 to $24,999
4	$25,000 to $34,999
5	$35,000 to $49,999
6	$50,000 to $74,999
7	$75,000 or more
88	Unknown

Figure Apx-B18 New Dataset Household Income

V2026 - HOUSEHOLD INCOME

Location: 42-43 (width: 2; decimal: 0)

Variable Type: numeric

Question:

Household income

Text:

Source code: 214

Notes: Variables V2013 Through V2026 refer to transcription items from control card.

Value	*Label*
01	Less than $5,000
02	$5,000 to $7,499
03	$7,500 to $9,999
04	$10,000 to $12,499
05	$12,500 to $14,999
06	$15,000 to $17,499
07	$17,500 to $19,999
08	$20,000 to $24,999
09	$25,000 to $29,999
10	$30,000 to $34,999
11	$35,000 to $39,999
12	$40,000 to $49,999
13	$50,000 to $74,999
14	$75,000 and over
98	Residue
99 (M)	Out of universe

Figure Apx-B19: Raw Data Household Income Categories

As seen to the left (V2026 - Household Income), raw data has more categories compared to new dataset, likely restructured based on NCVS reporting needs.

Categorical Correlation

From the exploratory analysis plots, we concluded that simple reports and plots are not able to provide insights into underreporting of the crimes. Correlation analysis is widely used to understand if there is a relationship between explanatory variables and a target variable (in our case, reported to police - yes or no). Since all the variables are categorical variables, regular correlations metric would not make much sense here. An alternative is to convert the categorical variables into one-hot-encoded vectors and use Pearson correlation metric. But that would be too messy as there are multiple levels in each categorical variable. Cramer's V is based on a nominal variation of Pearson's Chi-Square Test. This metric is used for assessing correlations between categorical variables. Figures Apx-B20 and Apx-B21 show categorical correlations between all the variables in both personal and household victimization datasets. As can be seen, there is no variable with high correlation with our target (notify/notifyTarget).

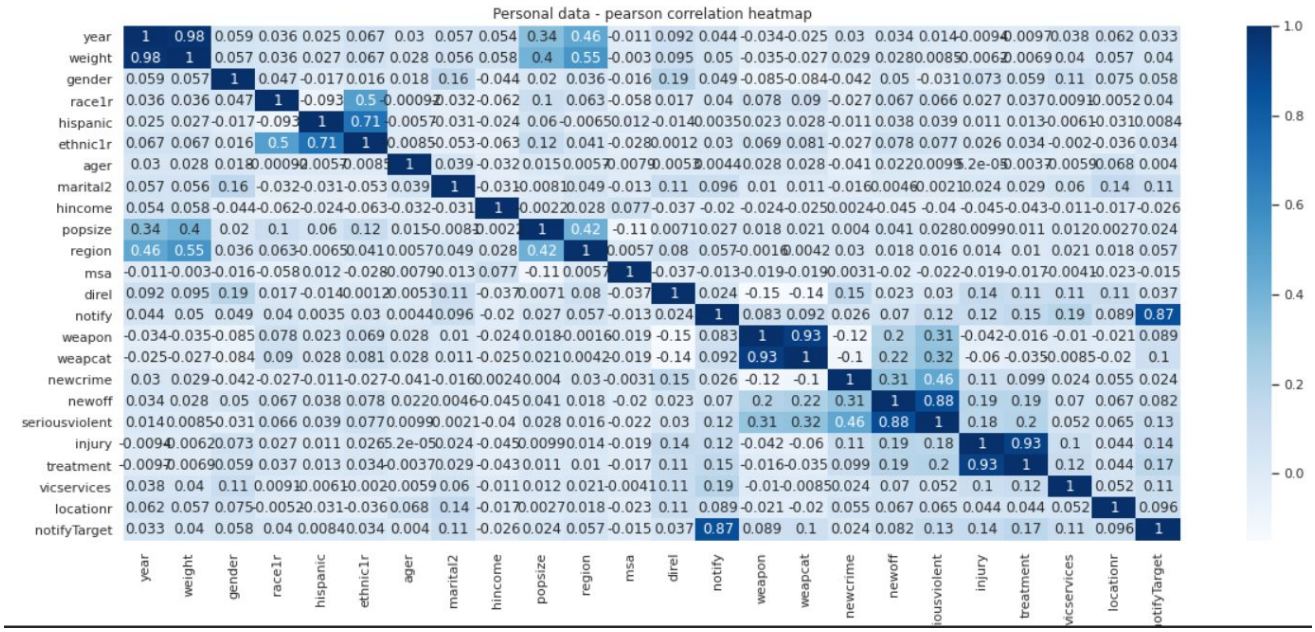


Figure Apx-B20: Cramer's V Correlation Heatmap of Personal Victimization Variables

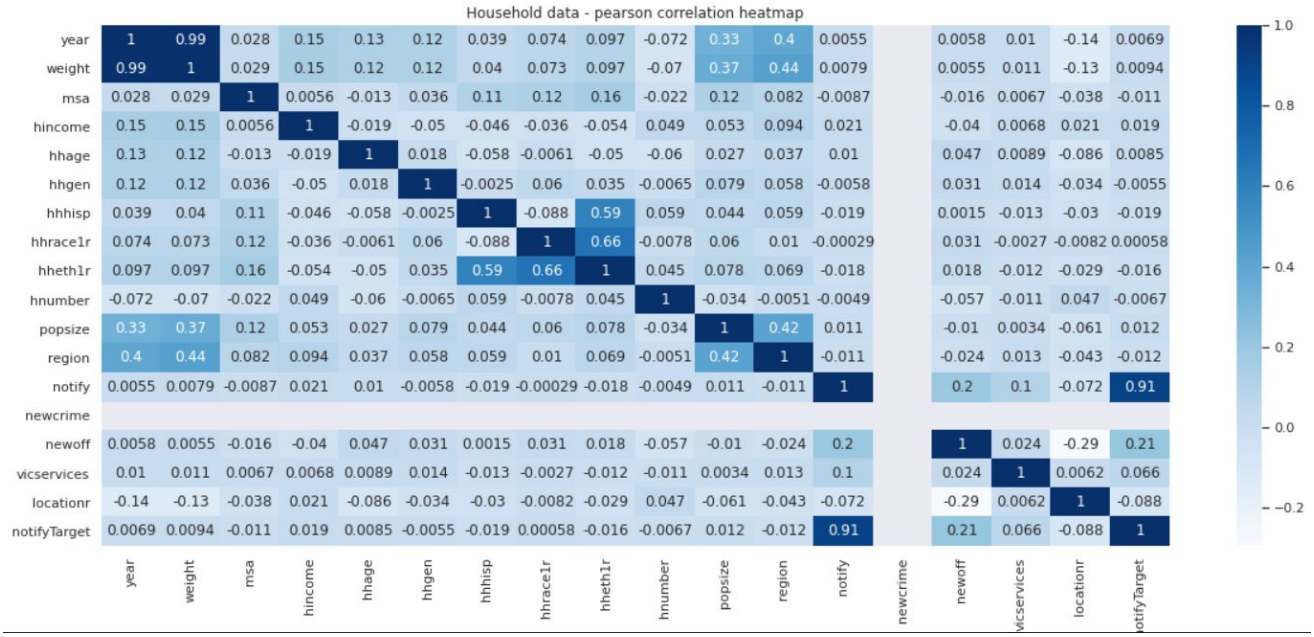
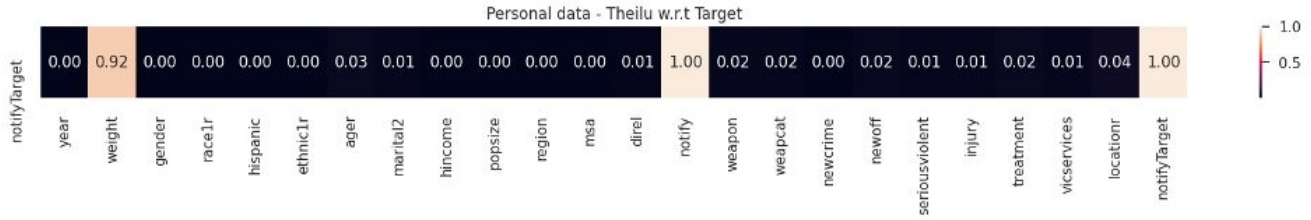


Figure Apx-B21: Cramer's V Correlation Heatmap of Household Victimization Variables

One drawback with Cramer's V metric is that it is symmetric and can misrepresent correlations with the type of variables we have (e.g., weapon, weaponcat). Theil's U is an asymmetric measure which can prevent the loss of information due to symmetric metric like Cramer's V.

Theil's U [1], also referred to as the Uncertainty Coefficient, is based on the conditional entropy between x and y — or in human language, given the value of x, how many possible states does y have, and how often do they occur. Just like Cramer's V, the output value is on the range of [0,1], with the same interpretations as before — but unlike Cramer's V, it is asymmetric, meaning U(x,y)≠U(y,x) (while V(x,y)=V(y,x), where V is Cramer's V). Using Theil's U will let us find out that knowing y means we know x, but not vice-versa.



As can be seen in Figure Apx-B22, none of the variables have high Theil's U coefficient with respect to target variable (notify is re-coded into notifyTarget) further confirming our EDA results. The weight is a continuous variable, so that can be ignored here.

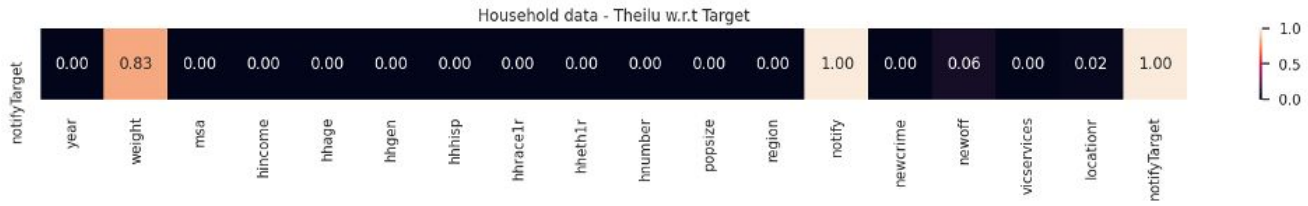


Figure Apx-B22: Theil's U Coefficients of Personal and Household Victimization Variables with Respect to Target.

In order to clean or remove redundant variables, we further need to study the relationship between explanatory variables as correlated X's can result in model instability. Since our variables are all categorical, Theil's U metric is run between pairs of variables and the combinations with value >0.1 are plotted in Figures Apx-B23 and Apx-B24, respectively, for both the datasets.

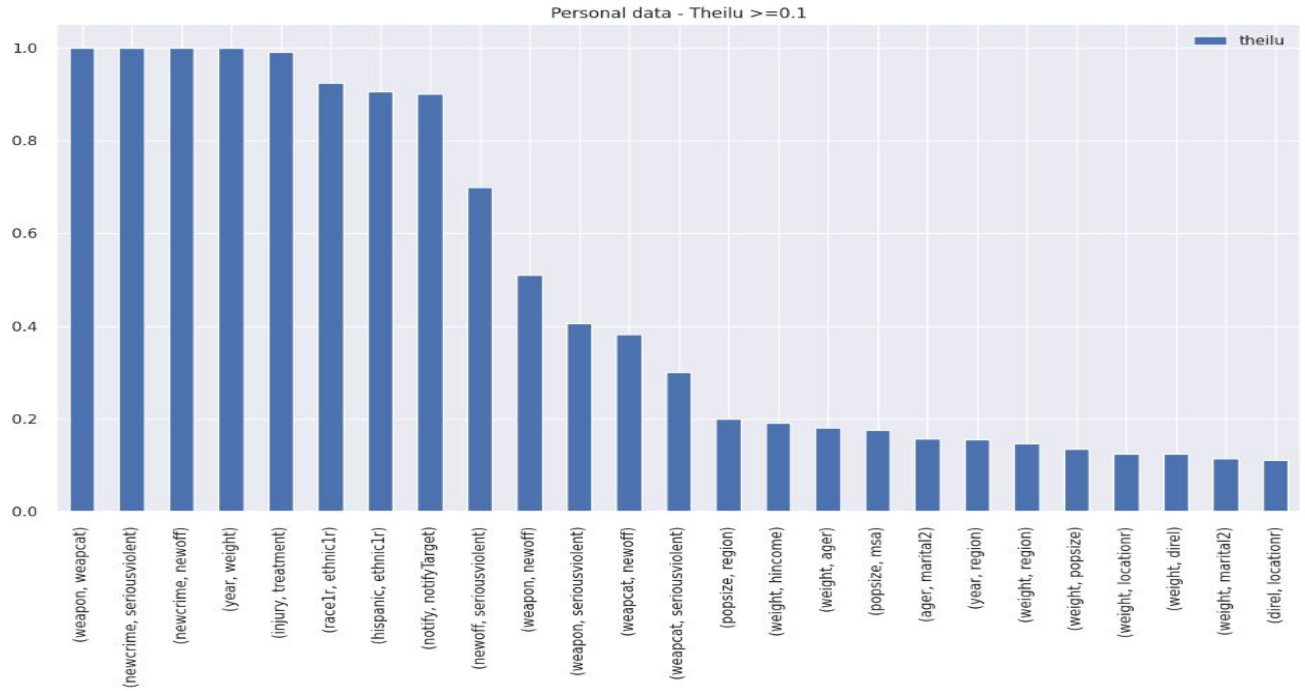


Figure Apx-B23: Theil's U Between pairs of Personal Victimization Variables.

As discussed earlier, weapon and weaponcat show very high Theil's U values indicating they are redundant to each other. We will pick only one of the variables in each of these pairs with Theil's U> 0.5 (the number of

categories in each variable can play a role in model performance so we will experiment with different combinations of variables).

The take away from correlation analysis is that there is no single variable with high correlation to target and there are redundant categorical variables which may impact performance of modeling techniques.

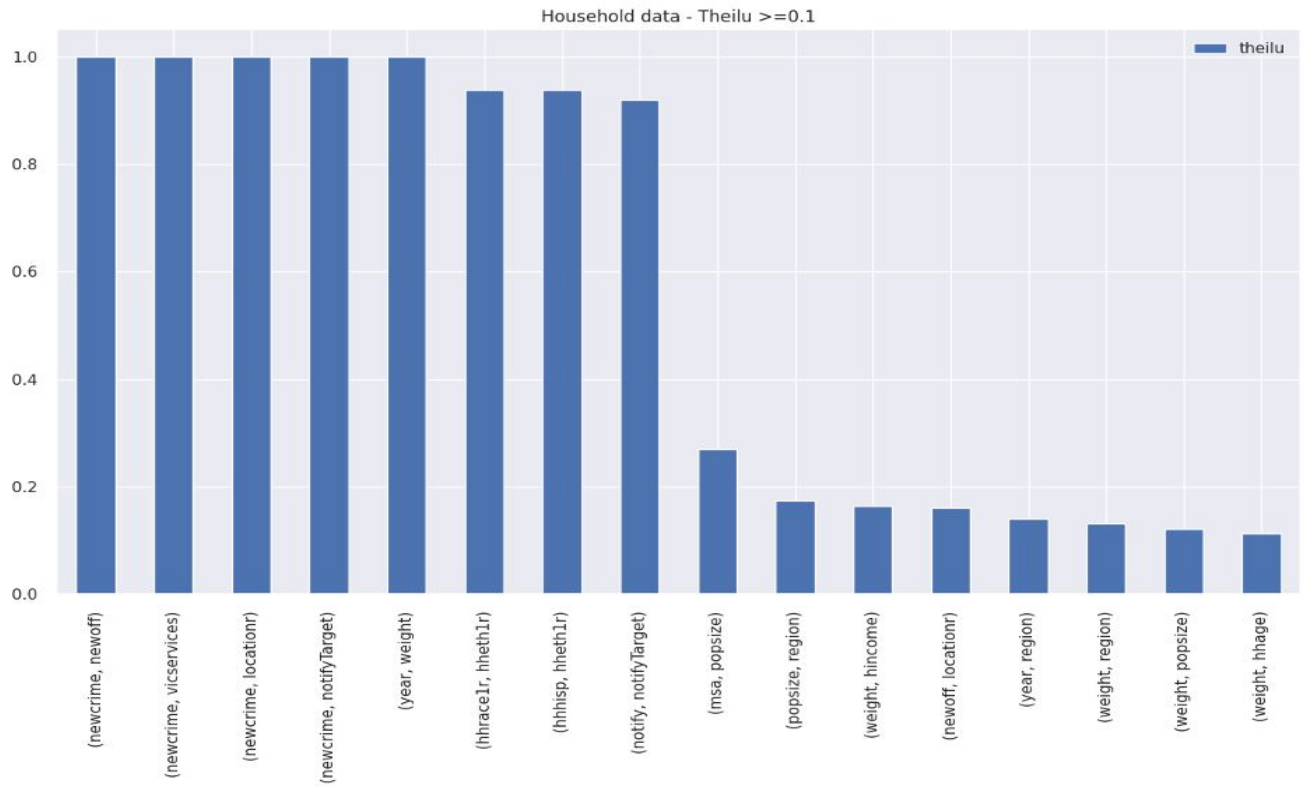


Figure Apx-B24: Theil's U Between pairs of Household Victimization Variables.

Literature Review

Looking at some of the available literature data and studies on the similar datasets from NCVS shows valuable observations by different scientists.

For example, we can see that financial loss due to different types of crimes across the U.S. in 2004 is estimated to be close to $15.85 billion. Interestingly, around 57.5% of these crimes were not even reported to the police department (Juliette et. al. 2007). That is why identifying the most important variables for not reporting a crime has been the subject of many studies in the past.

In most of these studies researchers employ regression techniques and frequency distributions to analyze their datasets to find the underlying parameters related to reporting a crime.

Most of these cases are limited to specific crime types. For example, in a study with focus on rape victimization, Bachman R. (Bachman R 1993) showed that use of physical force during rape and necessity of medical attention for injuries among the parameters that increases likelihood of a crime to be reported significantly.

Much smaller number of studies have explored the overall trends and underlying reasons for reporting all crimes. For example, Felson et. al (Felson et. al. 2002) used regression analysis and identified three major variables that increase likelihood of crime reporting. That includes the relationship of the offender to the victim (if committed by a stranger), age of the victim (above 65 years old), and prior experiences of victims with victimization and reporting, are associated positively with likelihood of a crime to be reported to the police.

In another similar study by Hart et.al., victim age, the presence of a weapon, and injury to the victim were identified as the major variables (Hart and Rennison, 2003).

Baumer et.al. conducted an overall study in 2010 and “compared changes in the likelihood of crime reporting to the police for several different types of crime (robberies, aggravated assaults, simple assaults, rapes and sexual assaults, burglaries, motor vehicle thefts, larceny, stranger and nonstranger violence, violence experienced by men and women, and crime experienced by members of different racial ethnic subgroups) between 1973 and 2005”.

Appendix C

Factor Analysis & Clustering

Factor Analysis

Factor analysis is used to explore factors (unobserved) that expresses itself through its relationship with other observed variables. It allows us to investigate concepts that are not easily measured directly by collapsing a large number of variables into few interpretable underlying factors. Each factor captures certain amount of the overall variance in the observed variables. We can interpret the constructs (e.g., socioeconomic conditions, mental safety/attitudes that are unmeasurable) behind the measured variables by extracting and visualizing the factors. Additionally, it offers axis rotation methods to allow subjective interpretations of the factors (so there will be some trade-off between fitting the data and interpretability).

We used two types of encodings of the categorical data (Chi-square based correlations, WoE vectors) as input data for Factor Analysis. Different estimation and rotation methods are experimented to analyze the factors. Factor analysis is mainly driven by how the analyst wants to interpret the underlying factors by rotating the axis. While we did not have sufficient subject matter expertise required to interpret the underlying factors, we did notice structural differences in factor loadings in reported vs unreported crime data. This could be an indicative of different underlying factors driving under reporting of crime to police.

Chi-square Correlation Matrix as Input Data

We used the Maximum-likelihood method for fitting the data and Verimax rotation to visualize and interpret the data. We kept variables with low granularity and removed others which seem to have high correlations (e.g., weapcat and weapon convey similar information, so we only kept weapcat due to more levels).

Interpretation from Personal Data

Looking at overall data loadings in Figure Apx-C1, Factor 1 may be referring to perpetrator, Factors 2-4 refer to demographics and victims' characteristics. Looking at differences in reporting vs not reporting factor loadings, it seems like, Factor1 & 2 in reported is referring to some psychological factors (if married and well to do and crime resulted in injuries in populated areas). Factor 1 and 3 in not reported seem to show some pattern about demographics & perpetrators (may be people chose not to report where certain demographics and types of crimes are involved).

Factor Loadings

All data:

	Factor 1	Factor 2	Factor 3	Factor 4
gender	-0.071196	-0.307071	0.073271	-0.084161
ethnic1r	0.017355	-0.003851	0.093234	-0.095131
ager	0.023271	-0.061383	0.627818	0.171234
marital2	0.033470	0.284626	-0.048226	0.053788
hincome	0.017707	0.031266	0.051575	-0.043086
popsize	0.007933	-0.065819	-0.002988	0.024316
region	0.020549	-0.002670	-0.023053	0.070236
msa	-0.060379	0.028348	0.038189	0.232023
direl	0.071786	0.056614	0.033763	0.384324
weapcat	0.091751	0.030013	0.014438	0.053292
newoff	0.973612	0.131610	0.037557	-0.168368
injury	0.077754	0.395596	0.094425	-0.034811
vicservices	0.034472	0.251792	0.027036	0.030524
locationr	-0.059493	-0.388975	0.326666	-0.207105

notify = no

	Factor 1	Factor 2	Factor 3	Factor 4
gender	0.026158	-0.388916	0.062612	0.133689
ethnic1r	-0.055460	0.021432	0.147601	0.080137
ager	-0.047975	0.027049	-0.062587	0.367921
marital2	-0.010647	0.057585	-0.170895	-0.484729
hincome	-0.001964	-0.011237	0.006272	0.012869
popsize	0.125276	-0.027517	0.006670	-0.027825
region	0.989307	0.071947	-0.022810	0.102833
msa	-0.051464	0.051290	-0.243086	-0.037470
direl	0.008797	-0.058611	-0.044454	0.225213
weapcat	0.001921	-0.069052	0.265691	-0.099382
newoff	-0.027285	0.437605	0.469337	-0.040915
injury	0.019932	-0.410271	-0.044389	-0.012889
vicservices	-0.006877	0.175640	-0.015986	-0.030838
locationr	-0.062819	0.401263	-0.305103	0.083531

notify = yes

	Factor 1	Factor 2	Factor 3	Factor 4
gender	-0.348699	0.128994	0.016207	-0.007748
ethnic1r	0.031779	0.094007	0.032410	-0.056363
ager	0.066003	0.324682	0.028626	0.249781
marital2	0.285333	-0.118255	0.024185	-0.003067
hincome	0.024669	-0.000117	-0.017975	0.001721
popsize	-0.101362	-0.037839	0.025159	0.009802
region	-0.005750	0.001846	0.010882	0.029835
msa	0.009180	-0.015739	-0.165243	0.277741
direl	0.095748	-0.110097	0.037242	0.323443
weapcat	-0.037553	-0.023905	0.327758	0.113647
newoff	0.153668	0.112365	0.438716	-0.092075
injury	0.349694	0.061188	0.114460	-0.005224
vicservices	0.318323	0.037158	0.064451	0.048293
locationr	-0.288671	0.486747	-0.032357	-0.137401

Figure Apx-C1: Overall Factor Loadings for Personal Crimes

Interpretation from Household Data

In the overall analysis, Figure Apx-C2 seems to indicate that Factor 1 & 2 may be referring to regional crime types and Factors 3 & 4 to victim demographics and characteristics. Reporting to police seems to surface with ethnicity in certain localities (Factors 1 & 2).

Factor Loadings

	Factor 1	Factor 2	Factor 3	Factor 4
msa	0.076802	-0.056486	0.185842	-0.014701
hincome	-0.013758	-0.004396	0.022910	0.205628
hhage	-0.020256	0.054901	0.225938	0.008279
hhgen	-0.012179	0.016068	-0.031031	-0.035595
hheth1r	-0.030485	-0.015738	0.000699	-0.184111
hnumber	-0.001367	-0.098689	-0.134636	0.271694
popsize	0.027053	-0.005428	-0.025370	0.050707
region	0.981452	0.001494	0.145680	0.102600
newoff	-0.007771	0.604383	-0.192996	-0.081846
vicservices	-0.002166	0.006580	0.020402	0.042622
locationr	-0.006867	0.406888	0.127443	0.027255

notify = yes

	Factor 1	Factor 2	Factor 3	Factor 4
msa	0.994672	0.047134	-0.051720	0.027021
hincome	-0.024803	0.064616	0.081268	-0.156212
hhage	0.032422	-0.006714	-0.035223	0.315622
hhgen	-0.004122	-0.027193	-0.033967	0.018551
hheth1r	-0.033211	0.994252	0.024564	0.068969
hnumber	-0.001286	-0.039456	-0.022244	-0.100091
popsize	0.194433	-0.133543	0.008532	-0.008303
region	0.042633	-0.091533	-0.037556	-0.037167
newoff	0.035107	-0.035218	-0.498465	0.061038
vicservices	-0.014899	0.010793	-0.008690	0.040587
locationr	0.000915	0.022735	0.607849	-0.002574

notify = no

	Factor 1	Factor 2	Factor 3	Factor 4
msa	0.103796	-0.071936	0.191029	0.001931
hincome	-0.016991	-0.017539	0.036567	0.208046
hhage	-0.010647	0.046889	0.239928	0.009929
hhgen	-0.016169	0.002895	-0.012176	-0.035410
hheth1r	-0.040097	-0.023766	0.016256	-0.188803
hnumber	-0.029431	-0.104066	-0.118472	0.271982
popsize	0.032624	-0.005368	-0.050952	0.063262
region	0.803324	0.009683	0.056659	0.124038
newoff	-0.018361	0.519210	-0.144189	-0.076904
vicservices	0.002835	0.027679	0.033813	0.044313
locationr	0.002713	0.427221	0.211695	0.018385

Figure Apx-C2: Overall Factor Loadings for Household Crimes

WoE vectors as Input Data

Multiple Correspondence Analysis is another method from the family of Factor Analysis techniques. Experiments were conducted with both MCA & FA and the results are as shown in Figures Apx-C3 and Apx-C4. PCA is used as an estimating procedure and Verimax Rotation is used to interpret the variables.

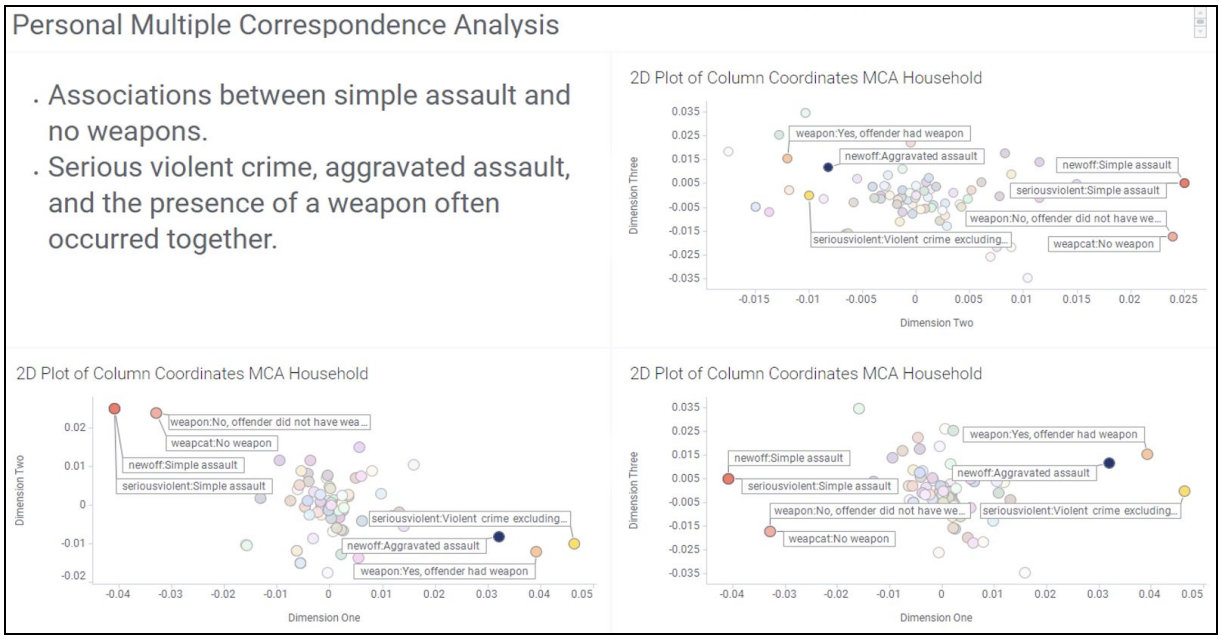


Figure Apx-C3: Multiple Correspondence Analysis Plots for Personal Crimes

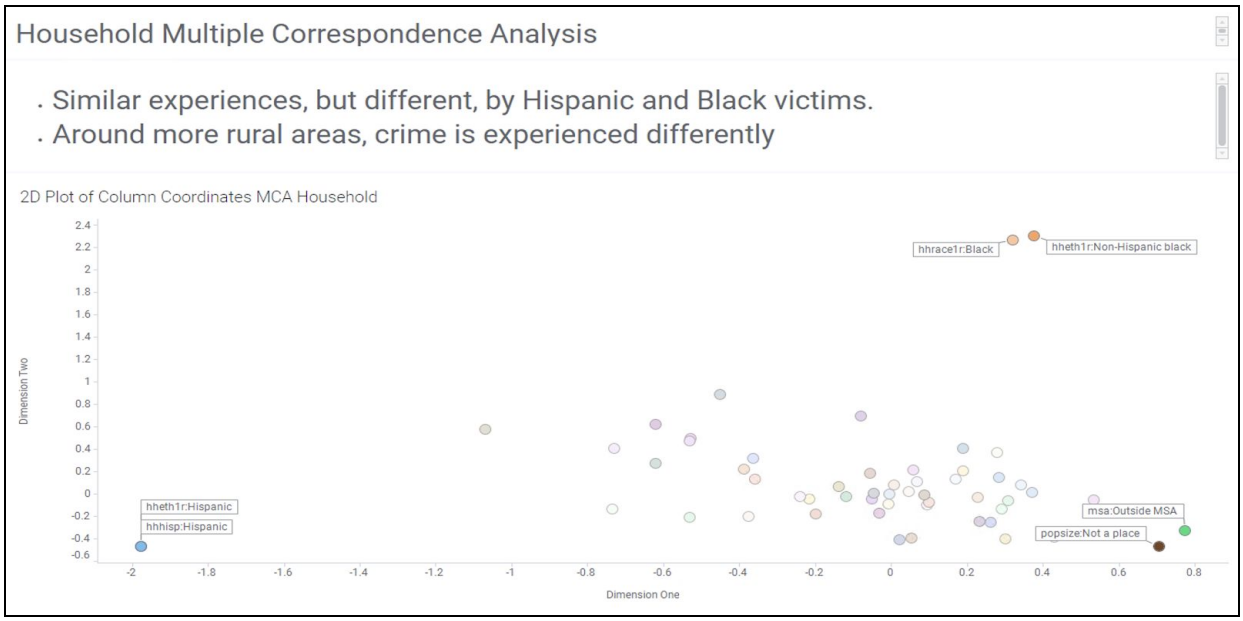


Figure Apx-C4: Multiple Correspondence Analysis Plots for Household Crimes

Figures Apx-C5 and Apx-C6 show and discuss the resulting Personal Crime and Household Crime factors.

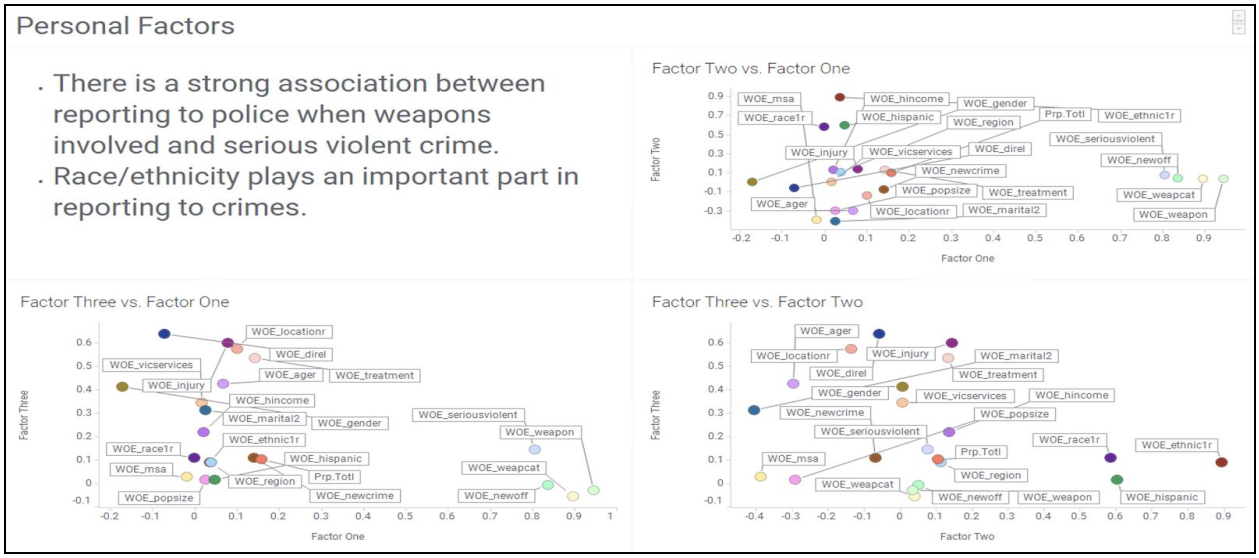


Figure Apx-C5: Factor Analysis Plots for Personal Crimes

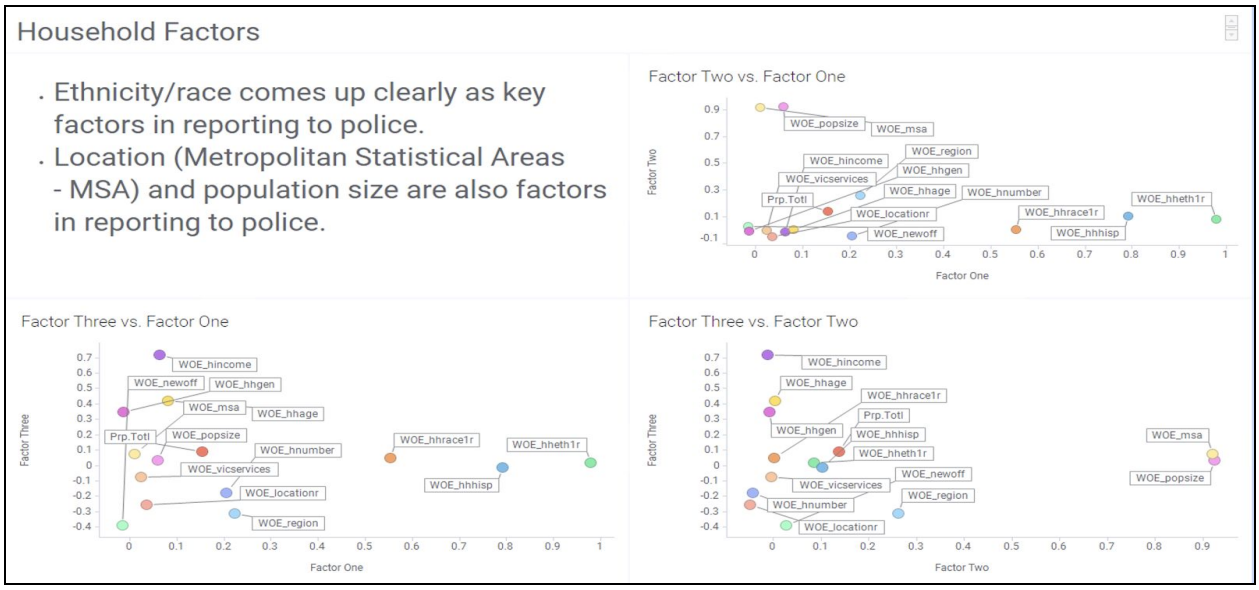


Figure Apx-C6: Factor Analysis Plots for Household Crimes

Clustering

Clustering is a useful exploratory analysis tool to project high-dimensional data into lower dimensions and investigate natural patterns in the data. PCA and TSNE are used to project data into lower dimensions and k-means clustering is used to classify the data into natural clusters.

We tried to interpret these clusters by relating them to our target variable and measuring the classification accuracy. The patterns look very noisy with one-hot vector representations of the categorical variables while WoE representations provided AUC of 0.58 indicating the data has some predictive potential, better than a random classifier. Without a dominating crime type, simple assault in the data, other crime types showed some separating clusters in TSNE visualization indicating there might be different sub-groups in the data behind responses.

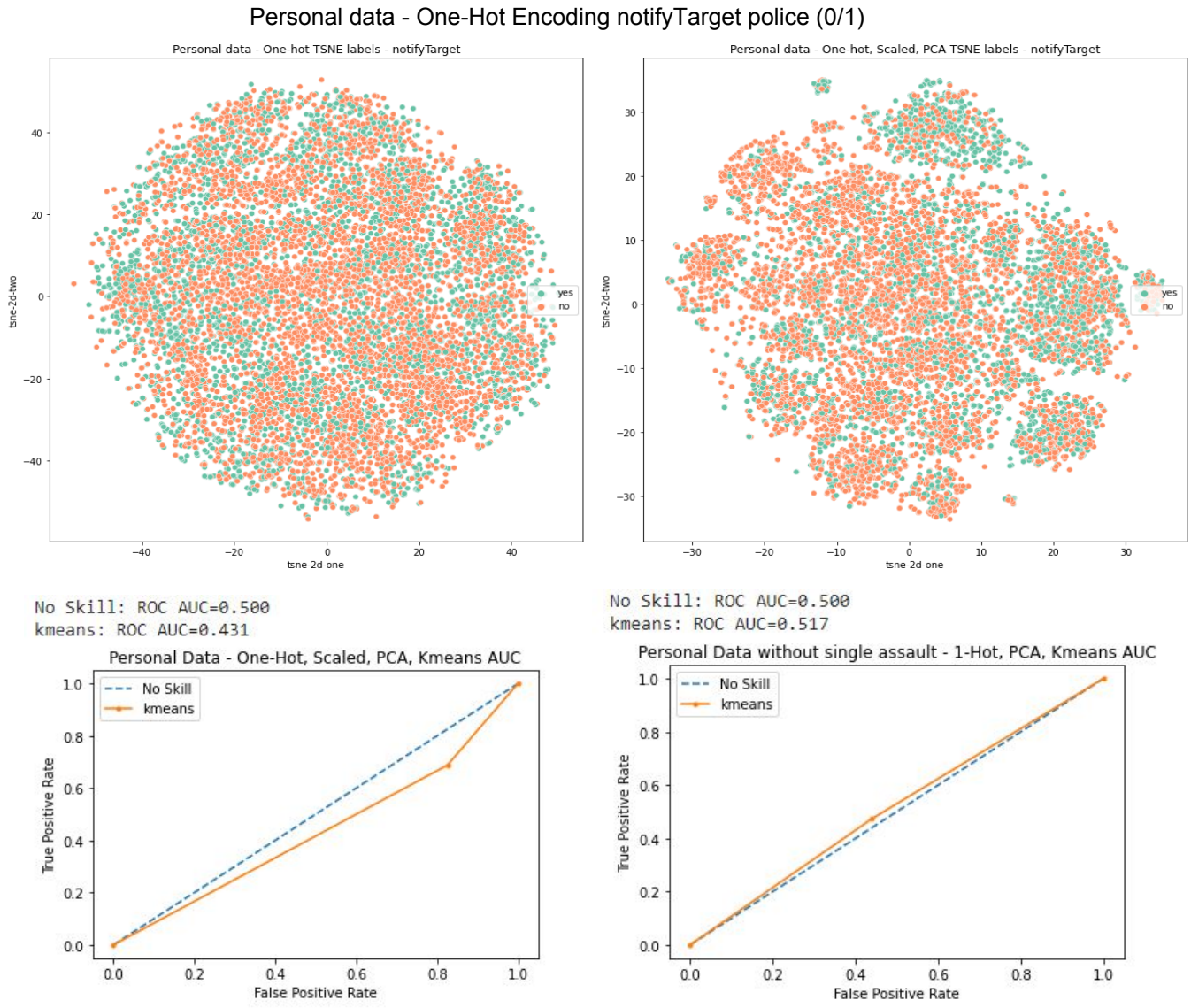


Figure Apx-C7: K-means Clustering with AUC for Personal Crimes using One-Hot Encoding

Appendix D

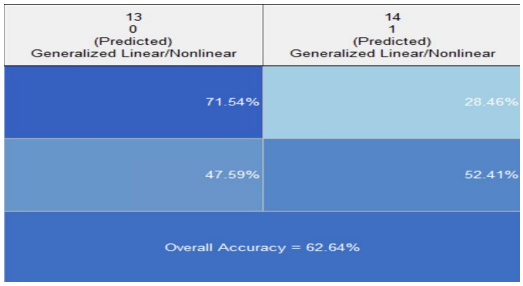
Classification Modeling

Personal Victimization Logistic Regression

The Logistic Regression model for Personal Victimization showed a total of around 63% overall accuracy and 72% accuracy for cases where people do not notify police.

The following binary predictors were used in the logistic regression and CHAID models: hispanic_Hispanic, newoff_Rape/sexual assault, vicservices_No services received from victim service agencies, locationr_At or near victim's home, weapcat_Firearm, locationr_Commercial place, parking lot, or other public area, newoff_Robbery, locationr_Other location, locationr_At or near friend, neighbor, or relative's home, hincome_$35,000 to $49,999, newoff_Aggravated assault, hincome_$50,000 to $74,999, popsize_500,000 to 999,999, hincome_$75,000 or more, ager_12 to 14, hincome_Less than $7,500, hincome_$15,000 to $24,999, region_South, weapon_Do not know if offender had weapon, hincome_$25,000 to $34,999, direl_Do not know number of offenders, ager_35 to 49, and hincome_$7,500 to $14,999.

Misclassification Matrix Testing



Misclassification Matrix Training

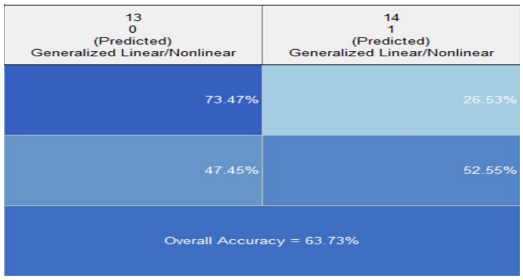
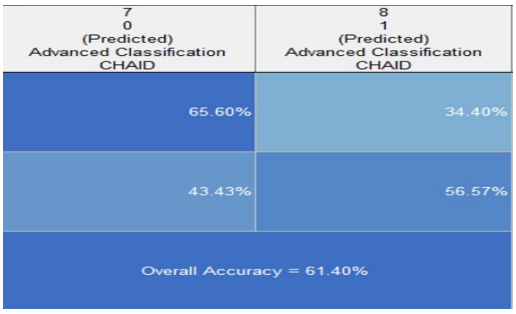


Figure Apx-D1: Personal Victimization Misclassification Matrices for Logistic Regression Training and Testing Datasets

Personal Victimization CHAID (Chi-square Automatic Interaction Detector)

The CHAID model for Personal Victimization showed a total of around 61% overall accuracy and 66% accuracy for cases where people do not notify police.

Misclassification Matrix Testing



Misclassification Matrix Training

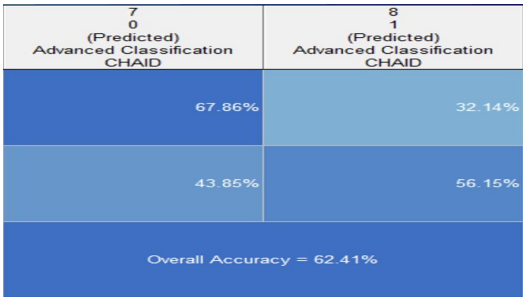


Figure Apx-D2: Personal Victimization Misclassification Matrices for CHAID Training and Testing Datasets

Household Crime Victimization Logistic Regression

The Logistic Regression model for Household Victimization showed a total of around 68% overall accuracy and 87% accuracy for cases where people do not notify police.

The following predictors for Household Crime victimization shown in Apx-D3 were used in the logistic regression and C&RT models:

hincome_Less than $7,500
hincome_Unknown
hincome_$15,000 to $24,999
hincome_$7,500 to $14,999
hincome_$75,000 or more

msa_Principal city within MSA

hhage_20 to 34

popsize_Under 100,000

newoff_Theft

vicservices_No services received from victim service agencies
vicservices_Missing

locationr_At or near victim's home
locationr_Commercial place, parking lot, or other public area
locationr_At or near friend, neighbor, or relative's home
locationr_School

hnumber_Two to Three

region_Southregion_Midwest

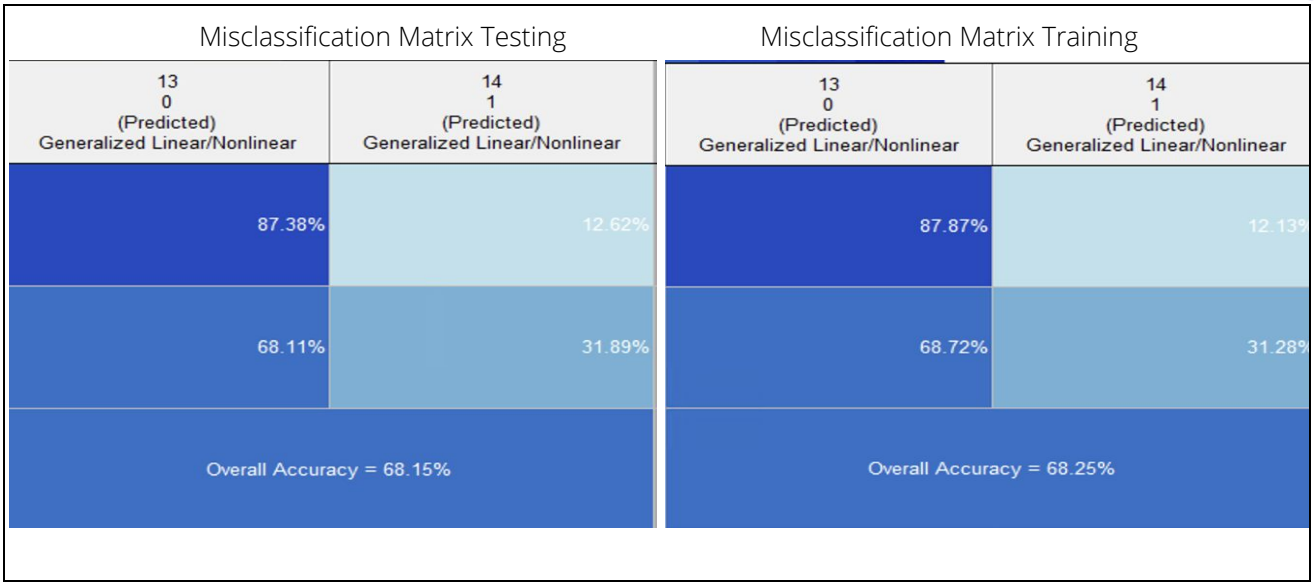


Figure Apx-D3: Household Victimization Misclassification Matrices for Logistic Regression Training and Testing Datasets

Household Victimization C&RT (Classification & Regression Tree)

The C&RT model for Household Victimization showed a total of around 68% overall accuracy and 84% accuracy for cases where people do not notify police.

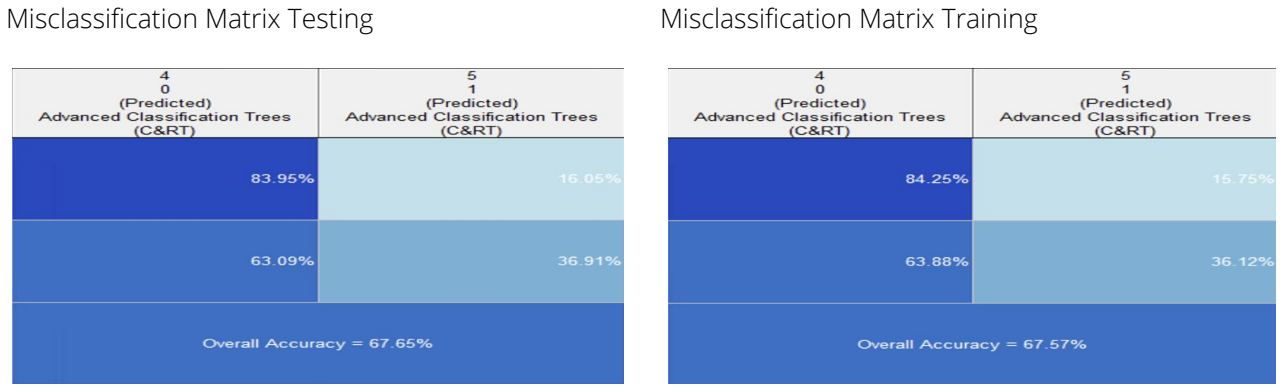


Figure Apx-D2: Household Victimization Misclassification Matrices for C&RT Training and Testing Datasets

Appendix E

Raw Data

The National Crime Victimization Survey for the period of 1992-2016 contains approximately 7.4 million records regarding crime gathered from households sampled randomly so as to be representative of households across the nation. The information is placed in three tables, *household, person, and incident*, in 1,371 variables. Interviews with the households in the survey are typically in person for the first interview then by phone once every six months for a period of 3 years. The interview includes every member of the sampled household when available.

Each record in Table 1 household and Table 2 person contains information on the randomly selected household occupants. The data includes housing type, setting (urban, rural, population of locale, if college, group home, etc), income, household composition, each members' educational attainment, occupation, race, sex, disability, and relationship status . There is one record for each household member aged 12 or older in person so the relationship of household to person is one to many. Information on those under age 12 is also recorded as "Yes/No" there are members under that age. Some households are a group of single people in group settings (eg in a convent or halfway house), some are single couples, some are single individuals, some have children or live with parents, some are single parents, some are multi-generational families, and every mix of the aforementioned.

Table 3 Incident contains information on the victim including race, sex, age, marital status, on the two primary household members related to the victim, on the perpetrator or group of them who committed the crimes such as sex, race, age, and familiarity of the victim with the perpetrator(s), on the crime and details of it (such as month and year of the crime, the object and type of crime eg. vandalism, whether it was thwarted or successful, whether it involved weapons, injuries sustained), what losses were experienced via job or financial losses, whether medical treatment was involved, on the setting of the victimization, on inhibitors to crime such as neighborhood watches, whether the victim was targeted in a hate crime, and on what the victim believes are the key factors in the crime, including if the crime was a single event or a series of similar events. Each record also indicates if they did or did not report the crime to authorities or to the police, and in some cases why they did or did not make the report.

Assessment of Raw Data

The database contains superfluous information. Some relate only to the survey, such as the best way to contact participants and whether an interview was successfully conducted or not. There are 19 exactly duplicated variables from household in person and 349 exactly duplicated variables from household and person in incident. There are 7 variables exactly replicated as 40xx from 30xx and other 30xx from 20xx. It is noted that only with respect to Table3 Incident, without some (not all) of this replication there is no single way to tie incidents uniquely to the data in Table1 Household or Table2 Person. No primary key exists for distinct Incident data. Over 15 variables exist delineating a “line number” of some sort for person which also makes join of data for unique incidents impossible without a subject matter experts’ assistance.

The organization of the data tables is very happenstance, which when added to the sheer volume of information makes working with the data unwieldy. There are instances when data is in dummy variables

(Yes/No) when they could be factors, and instances when factors exist when information in them should be reorganized differently so the levels make more sense.

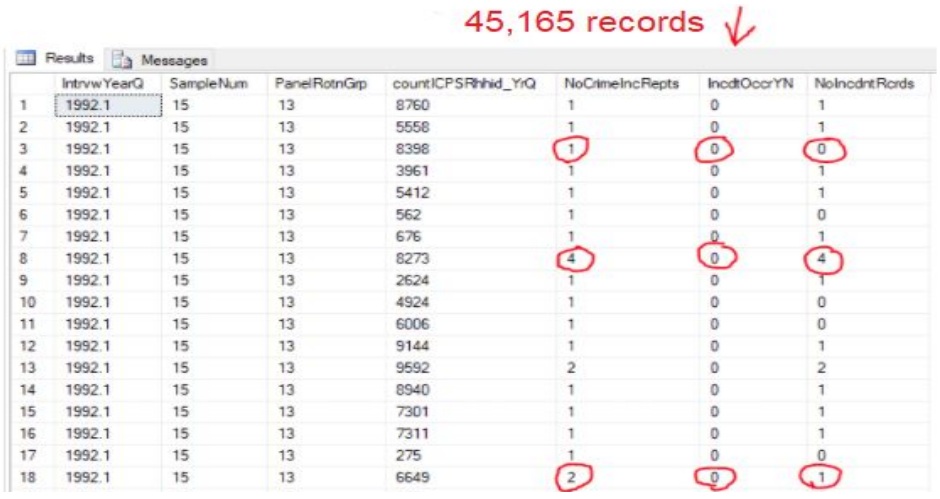
Data is missing at random based on the response of participants, but also not completely at random when it was only gathered for five years, or two years, or the first 20 years then stopped. Nearly every variable is highly correlated to others. The magnitude of studying such correlation was extremely difficult.

It has been very difficult for the team to work with the data in raw format. It has to be highly transformed to make modeling possible. That work is still proceeding and is being reported here. The team has incorporated a filler set with reduction of variables for modeling. The two sets will be used for results.

One of the most relevant variables to understanding national crime incidents would be the Control Number which contains the address, city, and state of the household. It is not included in the survey data except as "scrambled control number," which makes matching to an exact locale impossible. One can only understand region (eg. Northeast, Midwest, South, or West), and size of the location of the crime within that region (eg. a town of 1700 individuals, or a city of >5Million people).

VFLAG, Incident Reports, and Incident Records

VFLAG1. VFLAG (IncdtOccrYN) is designed to enable rapid pull of the set of data where incidents have occurred. Going through the data, cases exist where VFLAG does not positively affirm an incident occurred and was recorded when incident records exist in the data set. There are 45,165 records where there is no VFLAG present with either a Crime Incident Report (NoCrimeIncRepts) > 0 or Number of Incident Records (NoIncdntRcrds) > 0. Figure Apx-E1 shows a snapshot of cases where this is true.



	IntrvwYearQ	SampleNum	PanelRotnGrp	countICPSRhhid_YrQ	NoCrimeIncRepts	IncdtOccrYN	NoIncdntRcrds
1	1992.1	15	13	8760	1	0	1
2	1992.1	15	13	5558	1	0	1
3	1992.1	15	13	8398	1	0	0
4	1992.1	15	13	3961	1	0	1
5	1992.1	15	13	5412	1	0	1
6	1992.1	15	13	562	1	0	0
7	1992.1	15	13	676	1	0	1
8	1992.1	15	13	8273	4	0	4
9	1992.1	15	13	2624	1	0	1
10	1992.1	15	13	4924	1	0	0
11	1992.1	15	13	6006	1	0	0
12	1992.1	15	13	9144	1	0	1
13	1992.1	15	13	9592	2	0	2
14	1992.1	15	13	8940	1	0	1
15	1992.1	15	13	7301	1	0	1
16	1992.1	15	13	7311	1	0	1
17	1992.1	15	13	275	1	0	0
18	1992.1	15	13	6649	2	0	1

Figure Apx-E1: Missing VFLAG Indicators When Incidents Reside in Database

In the next set of records for ICPSRhhid_YRQ '9503', Figure Apx-E2 shows 2 incidents reports were filed according to the survey respondent, while only one incident record was created. The three rows represent three people tied to the incident.

IntrvwYearQ	SampleNum	PanelRotnGrp	ICPSRhhid_YrQ	FrstTmInSample	NoCrimeIncRepts	IncdtOccrYN	NoIncdntRcrds	S
1992.1	15	13	7956	1	0	0	0	3
1992.1	15	35	9392	1	1	1	1	6
1992.1	15	34	8154	1	0	0	0	5
1992.1	15	34	8154	1	0	0	0	5
1992.1	15	34	9505	1	3	0	3	6
1992.1	15	34	9505	1	3	0	3	6
1992.1	15	34	9505	1	3	0	3	6
1992.1	15	34	9503	1	2	1	1	6
1992.1	15	34	9503	1	2	1	1	6
1992.1	15	34	9503	1	2	1	1	6
1992.1	15	34	8152	1	1	1	1	5

Figure Apx-E2: Differing Crime Incident Reports vs Crime Incident Records

The interpretation is that NoCrimeIncidentReports may be the number filed with or dialed into police while NoIncidentRecords reflects what got recorded in the database. That number may be less because inclusion in the study filters complaints out by preset criteria such as victim age 12 yrs or older. Alternatively, it may be less than the number of reports when multiple things happened on one day and the Incident Record aggregated more than one crime in one incident record.

If VFLAG is used as designed to pull the set of data where incidents have occurred it will miss some 'n' number of incidents. For the full dataset analysis, VFLAG was updated into VFLAG1 to properly capture all incidents. Hopefully the NCVS staff use a "where exists" on incident records (raw field V2113) or also update the VFLAG field else reports using the datasets will include less incidents then actually in the database.

Transformations, Restructuring, and Use of Outside Information

Type of Crime. To determine whether a *type of crime* influences a person to report to police, in the old dataset 7 variables from Table1 Household which originally had 5-7 levels were combined into a single new variable "TypeOfCrime". Transformation occurred when we took multiple factors and combined them, evaluating them with new criteria.

```
-- TYPE OF CRIME
,CASE
    WHEN  A.V2076 = 1 THEN 1 --breakin or attempt (1) Yes (2) No (3) Refused (8) Residue (9) OOU
    WHEN  A.V2079 = 1 THEN 2 --motor vehicle theft (1) Yes (2) No (3) Refused (8) Residue (9) OOU
    WHEN A.V2080A = 1 THEN 3 --used Credit Card w-out permission (1) Yes (2) No (3)(8)(-2)(-1)(9) NULL Don't Know
    WHEN A.V2080B = 1 THEN 4 --used other Accts w-out permission
    WHEN A.V2080C = 1 THEN 5 --used Personal Info for Theft/Fraud
    WHEN  A.V2081 = 1 THEN 6 --vandalism against Household
    WHEN  A.V2104 = 1 THEN 7 --attack, threat, theft during vandalism ,A.V2104
    ELSE NULL
END AS TypeOfCrime
```

Figure Apx-E3: Type of Crime Transformation / Consolidation

In the reduced dataset this was explored through the variable 'newoff' in models.

Poverty. Initially for the full dataset, a new variable was created to determine if *poverty* influences whether crimes are reported to police. Data on poverty thresholds was obtained from the US Census for the period of 1959 to 2019 and truncated to the study years, 1992-2019. Figure Apx-E4 shows the Census data color-coded with NCVS study reported household income bands, by year, by family size for the raw dataset, and Apx-E5 for the reduced dataset. The top of the color band based on family size, 1-9+ members is a

poverty threshold for that reported income. There are 189,908 times wherein this is the case. When taking this methodology into the reduced dataset, rebucketized income in fewer categories on the lower end plus combination of households into 2-3, 4-5, and 6+ lost insights available in the full dataset. An estimated weighted year where poverty existed would have had to be used.

Table 1. Weighted Average Poverty Thresholds for Families of Specified Size: 1959 to 2019
(Population in thousands. Population as of March of the following year)

Year	Unrelated individuals		Families of 3 people or more						
	1 person	2 people	3 people	4 people	5 people	6 people	7 people	8 people	9 people or more
2019	13,011	16,521	20,335	26,172	31,021	35,129	40,016	44,461	52,875
2018	12,784	16,247	19,985	25,701	30,454	34,533	39,194	43,602	51,393
2017	12,485	15,880	19,515	25,093	29,716	33,610	38,170	42,642	50,723
2016	12,228	15,569	19,105	24,563	29,111	32,928	37,458	41,781	49,721
2015	12,082	15,391	18,871	24,257	28,741	32,542	36,998	41,029	49,177
2014	12,071	15,379	18,850	24,230	28,695	32,473	36,927	40,968	49,021
2013	11,880	15,139	18,554	23,844	28,234	31,887	36,239	39,930	48,343
2012	11,720	14,937	18,284	23,492	27,827	31,471	35,743	39,688	47,297
2011	11,484	14,657	17,916	23,021	27,251	30,847	35,085	39,064	46,572
2010	11,137	14,216	17,373	22,315	26,442	29,904	34,019	37,953	45,224
2009	10,956	13,991	17,098	21,954	25,991	29,405	33,372	37,252	44,366
2008	10,991	14,051	17,163	22,025	26,049	29,456	33,529	37,220	44,346
2007	10,590	13,540	16,530	21,203	25,080	28,323	32,233	35,816	42,739
2006	10,294	13,167	16,079	20,614	24,382	27,560	31,205	34,774	41,499
2005	9,973	12,755	15,577	19,971	23,613	26,683	30,249	33,610	40,288
2004	9,646	12,335	15,066	19,307	22,830	25,787	29,233	32,641	39,062
2003	9,393	12,015	14,680	18,810	22,245	25,122	28,544	31,589	37,656
2002	9,183	11,756	14,348	18,392	21,744	24,576	28,001	30,907	37,062
2001	9,039	11,569	14,128	18,104	21,405	24,195	27,517	30,627	36,286
2000	8,791	11,235	13,740	17,604	20,815	23,533	26,750	29,701	35,150
1999	8,499	10,864	13,289	17,030	20,128	22,730	25,918	28,970	34,436
1998	8,316	10,634	13,003	16,660	19,680	22,228	25,257	28,166	33,339
1997	8,183	10,473	12,802	16,400	19,380	21,886	24,802	27,593	32,566
1996	7,995	10,233	12,516	16,036	18,952	21,389	24,268	27,091	31,971
1995	7,763	9,933	12,158	15,569	18,408	20,804	23,552	26,237	31,280
1994	7,547	9,661	11,821	15,141	17,900	20,235	22,923	25,427	30,300
1993	7,363	9,414	11,522	14,763	17,449	19,718	22,383	24,838	29,529
1992	7,143	9,137	11,186	14,335	16,952	19,137	21,594	24,053	28,745

For information on confidentiality protection, sampling error, nonsampling error, and definitions, see <https://www2.census.gov/programs-surveys/cps/techdocs/cpsmar20.pdf>)

The way you read this is:

value 1 is always below the poverty levels from 1992-2019.
value 2 is below the poverty level for a single person beginning in 1994, and is always below the poverty level for 2+ households.
value 3 becomes below the poverty level in 2006 for 1 person, and in 1996 for 2. It is always below poverty level for families 3+.

...and so on.

https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-people.html

Value	Label
1	Less than $5,000
2	$5,000 to $7,499
3	$7,500 to $9,999
4	$10,000 to $12,499
5	$12,500 to $14,999
6	$15,000 to $17,499
7	$17,500 to $19,999
8	$20,000 to $24,999
9	$25,000 to $29,999
10	$30,000 to $34,999
11	$35,000 to $39,999
12	$40,000 to $49,999
13	$50,000 to $74,999
14	$75,000 and over
98	Residue
99 (M)	Out of universe

Figure Apx-E4: US Census Weighted Avg Poverty Thresholds By Year,Family Size

Table 1. Weighted Average Poverty Thresholds for Families of Specified Size: 1959 to 2019
(Population in thousands. Population as of March of the following year)

Year	Unrelated individuals	Families of 3 people or more	
	1 person	2-3 people	4-5 people	6+ people
2019	13011	16521	20335	26172
2018	12784	16247	19985	25701
2017	12485	15880	19515	25093
2016	12228	15569	19105	24563
2015	12082	15391	18871	24257
2014	12071	15379	18850	24230
2013	11880	15139	18554	23844
2012	11720	14937	18284	23492
2011	11484	14657	17916	23021
2010	11137	14216	17373	22315
2009	10956	13991	17098	21954
2008	10991	14051	17163	22025
2007	10590	13540	16530	21203
2006	10294	13167	16079	20614
2005	9973	12755	15577	19971
2004	9646	12335	15066	19307
2003	9393	12015	14680	18810
2002	9183	11756	14348	18392
2001	9039	11569	14128	18104
2000	8791	11235	13740	17604
1999	8499	10864	13289	17030
1998	8316	10634	13003	16660
1997	8183	10473	12802	16400
1996	7995	10233	12516	16036
1995	7763	9933	12158	15569
1994	7547	9661	11821	15141
1993	7363	9414	11522	14763
1992	7143	9137	11186	14335
	1994	2009	2012	1999
				2010

		bucket
2	2012	2010
3	2003	
4	1994	1991
5	1983	
2	2012	2010
3	2003	
4	1994	1991
5	1984	
4	2016	2013
5	2006	

modified for 1992-2019

(for reduced dataset)

Figure Apx-E5: US Census Weighted Avg Poverty Thresholds by Year, Family Size

Series Crime. Whether a *series of crimes* influences reporting to police is also captured. The present series variable shown in Figure Apx-C5 does not include all TypeOfCrime categories, so a new variable paralleling that was created. The NCVS only considers a string of crimes a series when the number of similar victimizations exceeds 6. When a series of crimes exist, the incident weight used to extrapolate their frequency to the US changes.

V4526 - CHECK ITEM V1: SERIES CRIME DESCRIPTION

Location: 1623-1624 (width: 2; decimal: 0)

Variable Type: numeric

Text:
Source code: 895

Notes: Prior to quarter 1, 1999 this variable was Check Item V.

V4017 - CHECK ITEM B: HOW MANY INCIDENTS

Value	*Label*
1	1-5 incidents (not a "series")
2	6 or more incidents
8	50
9 (M)	Out of universe

Value	*Label*
	Contact crimes
01	Completed or threatened violence in the course of the victim's job (police officer, security guard, psychiatric social worker, etc.)
02	Completed or threatened violence between spouses, other relatives, friends, neighbors, etc.
03	Completed or threatened violence at school or on school property
04	Other contact crimes (other violence, pocket picking, purse snatching, etc.)
	Noncontact crimes
05	Theft or attempted theft of motor vehicle
06	Theft or attempted theft of motor vehicle parts (tire, hubcap, battery, attached tape deck, etc.)
07	Theft or attempted theft of contents of motor vehicle, including unattached parts
08	Theft or attempted theft at school or on school property
09	Illegal entry of, or attempt to enter, victim's home, other building on property, second home, hotel, motel
10	Theft or attempted theft from victim's home or vicinity by person(s) known to victim (roommate, babysitter, etc.)
11	Theft or attempted theft from victim's home or vicinity by person(s) unknown to victim
12	Other theft or attempted theft (at work, while shopping, etc.)
98	Residue
99 (M)	Out of universe

Figure Apx-E6: Parallel New Variable Indicating Series for Type of Crime

In the reduced dataset this information was not available so was not used.

Hatecrime. Whether incidents being *hatecrimes* influences reporting to police is captured. Information from 7 variables in Table 1 Household and 16 variables in Table 3 incident were combined so only when evidence exists that a crime is targeted as a hate crime will the new variable indicate it affirmatively. In the reduced dataset this information was not available so was not used.