Setu Madhavi Namburu

Sheila Cloudcroft

Team 52

Jeremy Melville

Jin Choi

Project management

Damon Panahi

Predictive Insights from Crime patterns

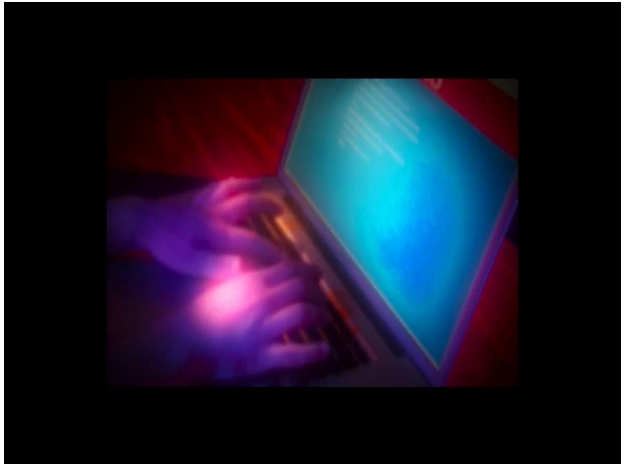
BJS *National Crime Victimization Survey*

Agenda: Approach and objectives, how we got here, and key findings and recommendations.

YouTube URL:

☆ https://youtu.be/_5Dj01Ny9tI

Project Summary



Customer: FBI, Law Enforcement

FBI Goal: Increase reporting of crime to police

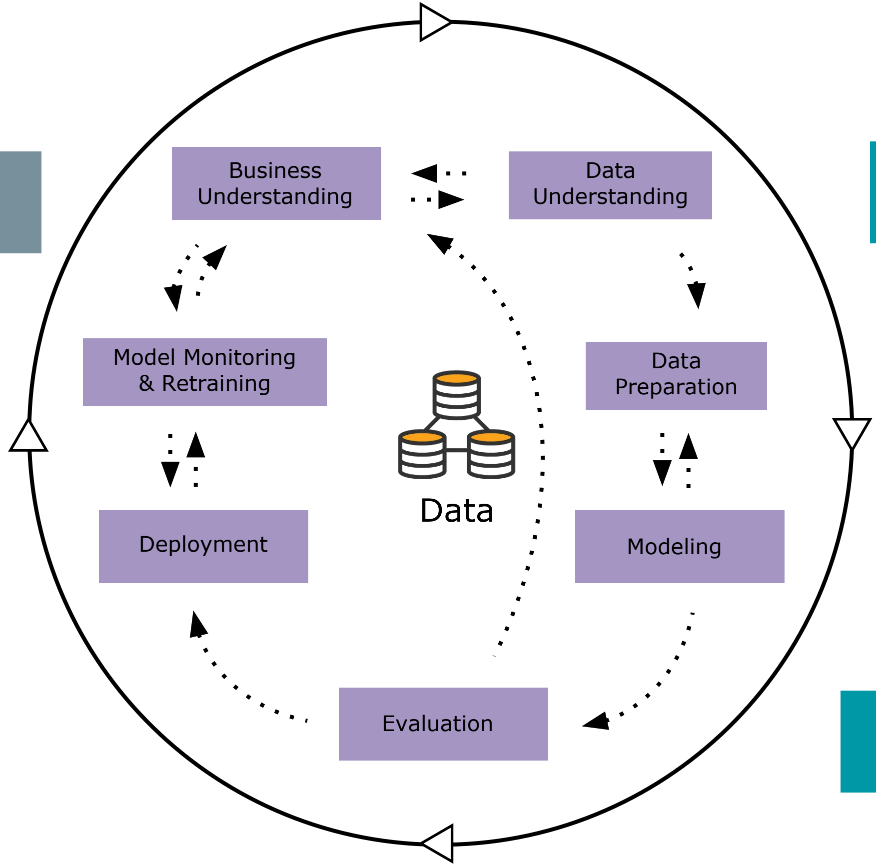
Our Goal: Identify influential factors

Scope: 10 week study

Video Courtesy of FBI NIBRS
https://www.fbi.gov/video-repository/nibrs-101.mp4

Our Approach: Iterative CRISP-DM

Production Refinement



Business Exploration

Nontraditional Data Science

Advanced Prototyping

Production Refinement is out of scope - POC pilot will run for 6 months

Project Journey

- Team 52 formed
- Explored potential project ideas

WEEK 1

WEEK 2

- Crime project selected by vote to use raw data

- Background research done
- Problem defined

WEEK 3

WEEK 4

- Preliminary data assessment done
- Project risk assessed

- Changed data to reduced, pre-processed data set
- One member pursuing raw data
- EDA started on reduced set

WEEK 5

WEEK 6

- EDA completed
- Modeling and Spotfire dashboard work begun

WEEK 7

- Modeling and dashboard continue, Android app begun
- Prelim model findings

- Android app functionality completed, modeling & dashboard work continue

WEEK 8

WEEK 9

- Models tuned, var selection
- Work on Final Project and Executive Summary

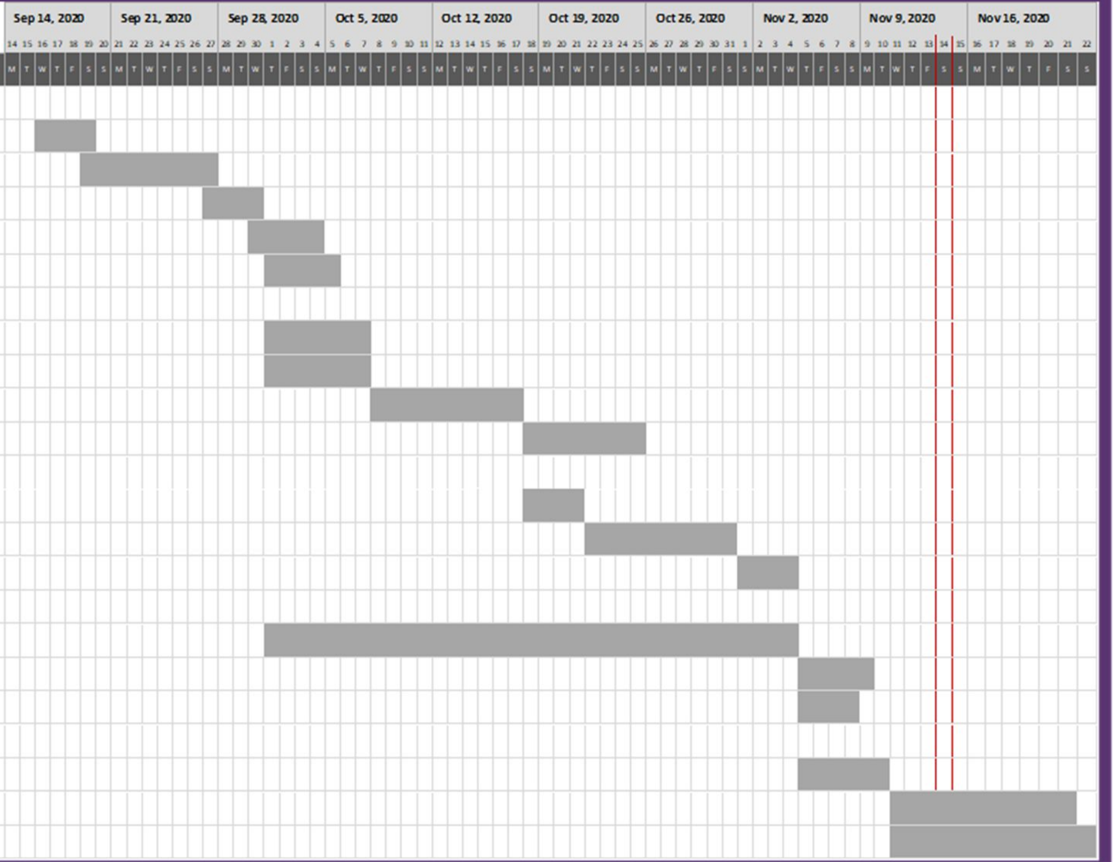
- Final Project Report submitted
- Executive Summary submitted

WEEK 10

Gantt Chart

Display Week: 1

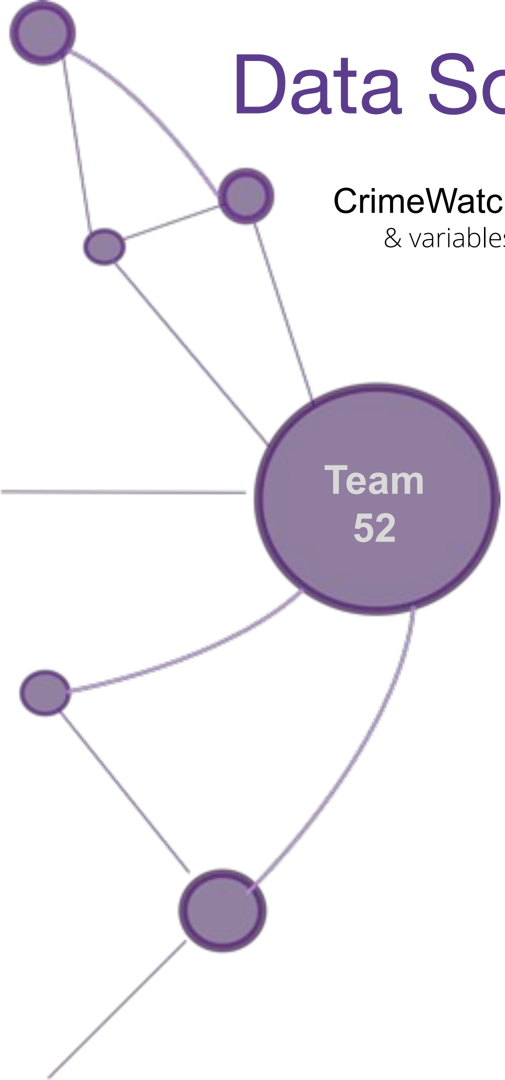
TASK	PROGRESS	START	END	Duration
Project Proposal				
Introduction and team setup	100%	9/16/20	9/19/20	3
Project selection	100%	9/19/20	**9/27/20**	8
Define the goal	100%	9/27/20	**9/30/20**	3
Project Proposal (report)	100%	9/30/20	10/4/20	4
Collect complimentary datasets	100%	10/1/20	10/5/20	4
EDA				
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	100%	11/5/20	11/9/20	4
App development	100%	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	100%	11/11/20	11/21/20	10
Final content organization	100%	11/11/20	**11/22/20**	11



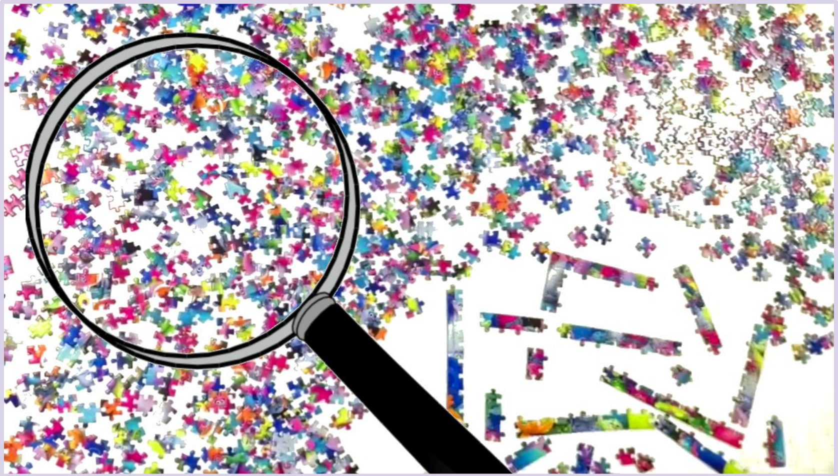
Objectives & Deliverables

	Objective	Technical Deliverable
1	**Understand differences between crime types and key insights from the data**	An Interactive descriptive and predictive insights **dashboard** to the FBI
2	**Understand the context of reported and under-reported crimes to police.**	Exploratory Data Analysis (EDA) & reporting of key findings. Reporting of nuances from raw data processing.
3	**Understand the key drivers causing the under-reporting of crimes to police.**	Build explainable predictive models to uncover key drivers & actionable insights
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

Data Sources



CrimeWatch Project at risk raw data issues, how to construct single crime incidents unclear, significant changes in survey & variables, which variables to use in what context? multiple variables for person, Subject Matter Expertise needed



Data Sources

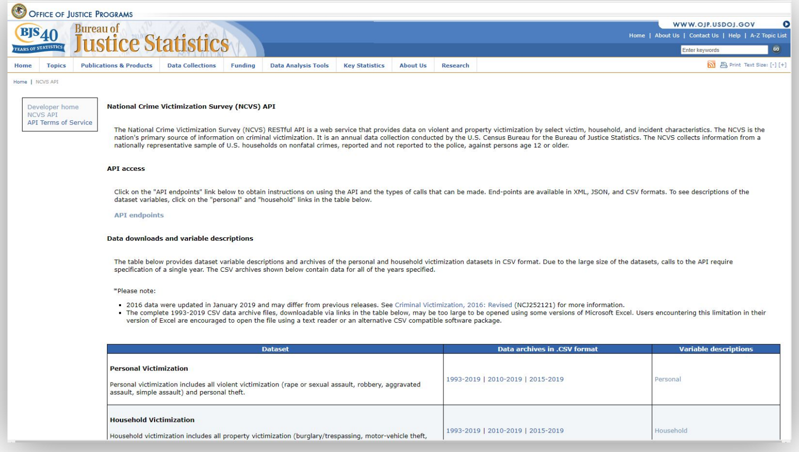
Team 52

CrimeWatch Project at risk raw data issues, how to construct single crime incidents unclear, significant changes in survey & variables, which variables to use in what context? multiple variables for person, Subject Matter Expertise needed

(FBI) New Data: NCVS RESTful API 2010-2019, subset 2013-2019 of ICPSR 37689 V1 (US Dept of Commerce. Census Bureau)

File Dimensions: Household: 69,400 records in 17 variables; Person: 17,515 records in 23 variables; Unweighted data used - pattern identification in reporting to police (unused: population files)

Universe: age 12+ living in US & District of Columbia - but not crews of vessels, institutions (prisoners, nursing homes), members of the armed forces


Office of Justice Programs

BJS 40 Years of Statistics

Bureau of Justice Statistics

WWW.OJP.USDOJ.GOV

Home | About Us | Contact Us | Help | A-Z Topic List

Enter keywords GO

Home | Topics | Publications & Products | Data Collections | Funding | Data Analysis Tools | Key Statistics | About Us | Research

Print Text Size: [-] [+]

Home | NCVS API

Developer home
NCVS API
API Terms of Service

National Crime Victimization Survey (NCVS) API

The National Crime Victimization Survey (NCVS) RESTful API is a web service that provides data on violent and property victimization by select victim, household, and incident characteristics. The NCVS is the nation's primary source of information on criminal victimization. It is an annual data collection conducted by the U.S. Census Bureau for the Bureau of Justice Statistics. The NCVS collects information from a nationally representative sample of U.S. households on nonfatal crimes, reported and not reported to the police, against persons age 12 or older.

API access

Click on the "API endpoints" link below to obtain instructions on using the API and the types of calls that can be made. End-points are available in XML, JSON, and CSV formats. To see descriptions of the dataset variables, click on the "personal" and "household" links in the table below.

API endpoints

Data downloads and variable descriptions

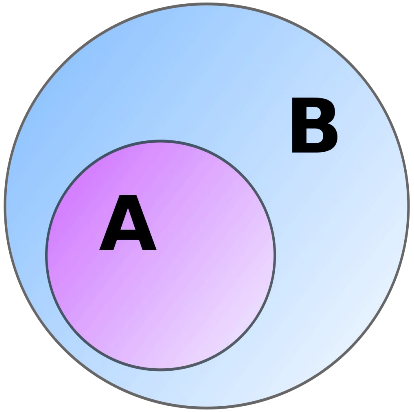
The table below provides dataset variable descriptions and archives of the personal and household victimization datasets in CSV format. Due to the large size of the datasets, calls to the API require specification of a single year. The CSV archives shown below contain data for all of the years specified.

*Please note:

- 2016 data were updated in January 2019 and may differ from previous releases. See Criminal Victimization, 2016: Revised (NCJ252121) for more information.
- The complete 1993-2019 CSV data archive files, downloadable via links in the table below, may be too large to be opened using some versions of Microsoft Excel. Users encountering this limitation in their version of Excel are encouraged to open the file using a text reader or an alternative CSV compatible software package.

Dataset	Data archives in .CSV format	Variable descriptions
Personal Victimization Personal victimization includes all violent victimization (rape or sexual assault, robbery, aggravated assault, simple assault) and personal theft.	1993-2019 | 2010-2019 | 2015-2019	Personal
Household Victimization Household victimization includes all property victimization (burglary/trespassing, motor-vehicle theft,	1993-2019 | 2010-2019 | 2015-2019	Household

Team 52



Scope Clarification & Limitations

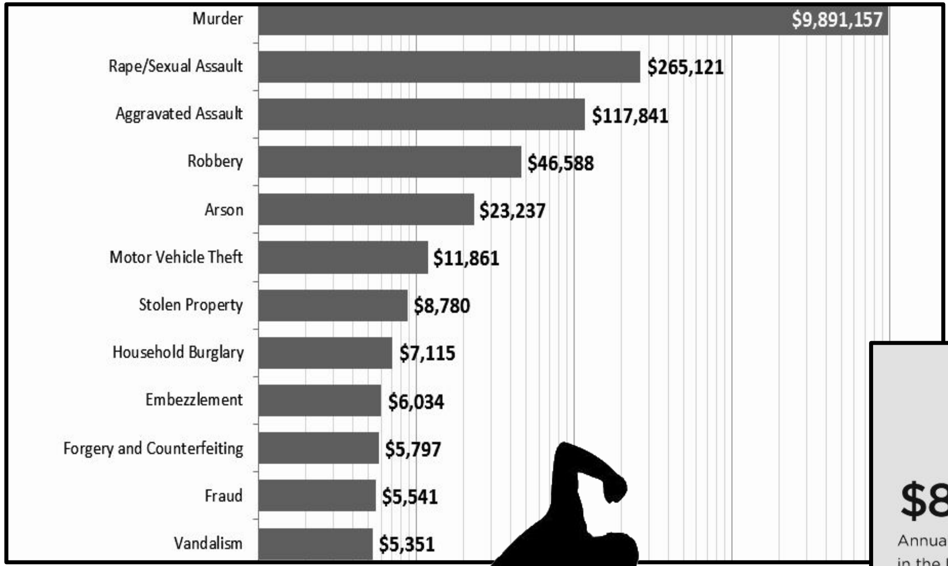
Reduction of Scope 2013-2019, difference of means tests, homogeneous with larger dataset, sufficient for goals

Whether victims report crime to police not other authorities

Data Sources

Team 52

New set more organized, condensed factors, less rich information (lost: **cost of victimization**, security, specific reasons no report); collapses variables - confounds results, - age, household size and poverty; location (city/state)



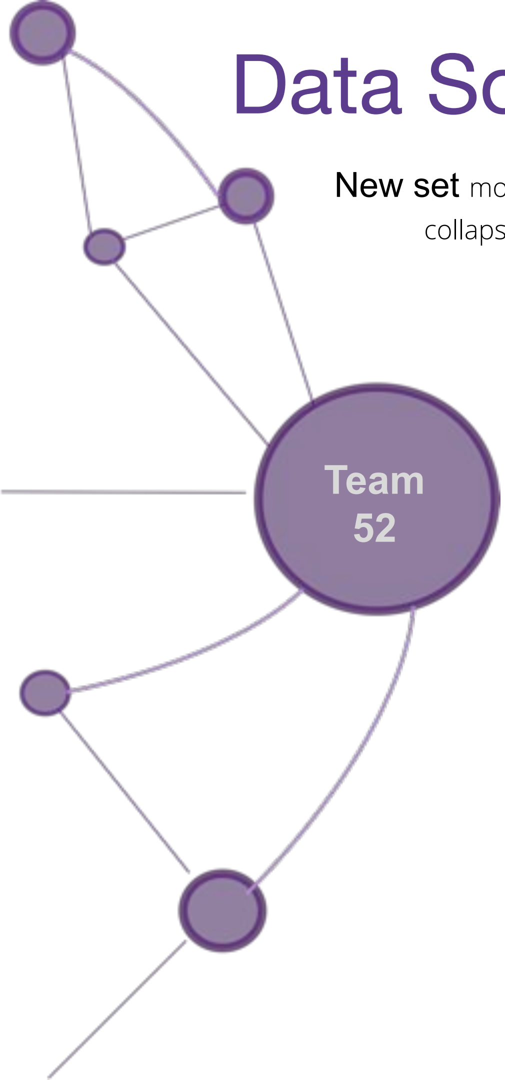
Data Sources

Team
52

New set more organized, condensed factors, less rich information (lost: cost of victimization, **security, specific reasons no report**); collapses variables - confounds results, - age, household size and poverty; location (city/state)

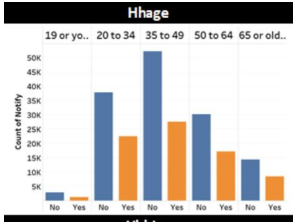
Why Didn't They Report?

- Reported to Different Official
- Offender is a Police Officer
- Not important to Police
- Advised Not to Report
- Not Clear a Crime
- Child Offender
- Lack of Proof
- Personal Matter
- To Protect Offender
- Police Biased
- Fear of Reprisal
- Too Inconvenient
- Other Reason

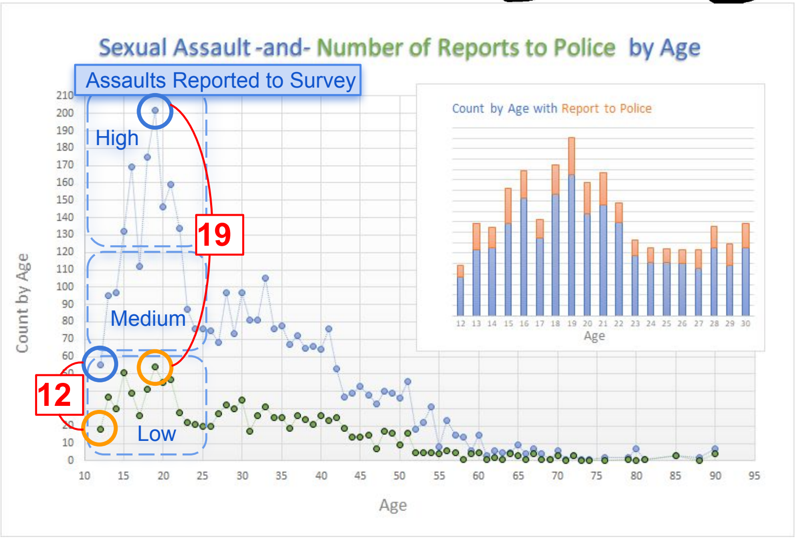


Data Sources

New set more organized, condensed factors, less rich information (lost: cost of victimization, security, specific reasons no report); collapses variables - confounds results, - **age**, household size and poverty; location (city/state)



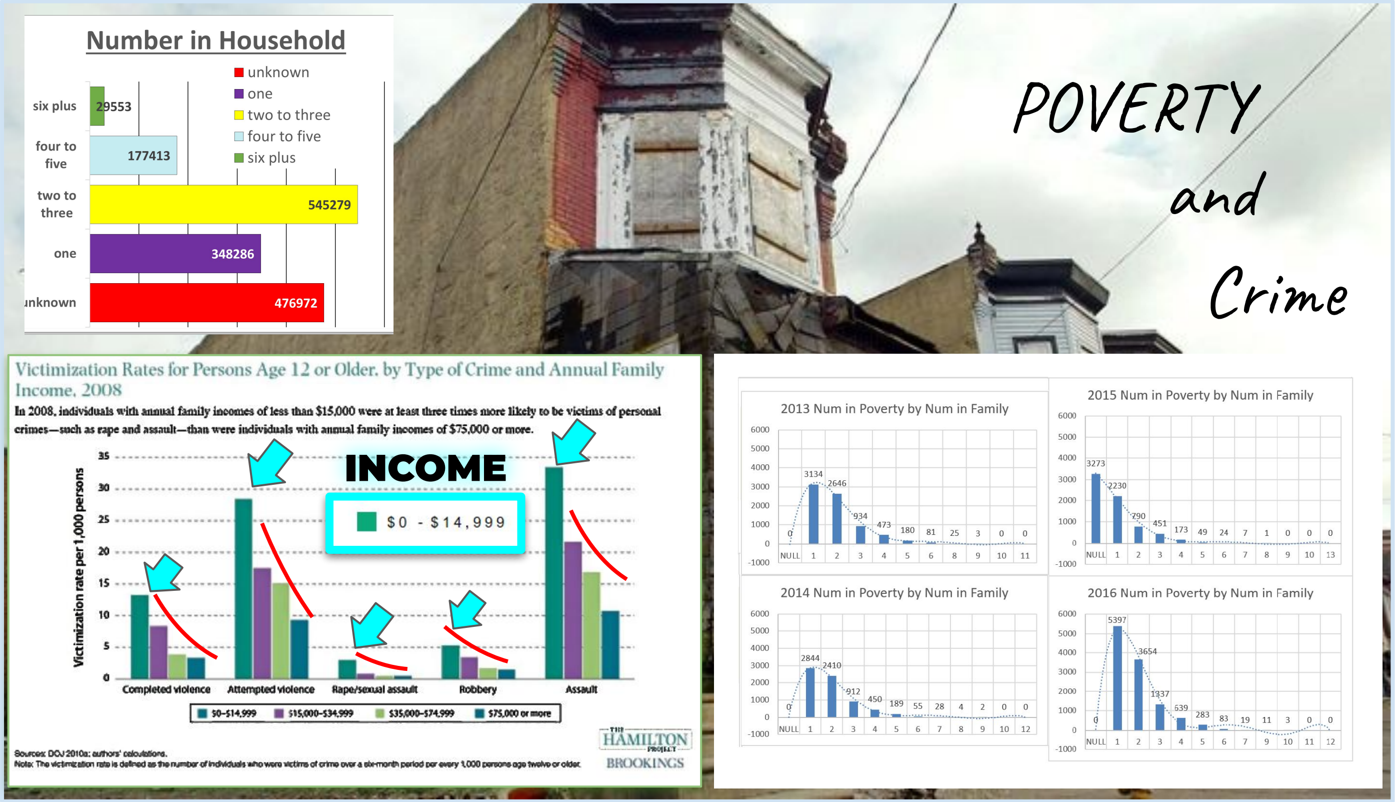
I can't tell anyone



Data Sources

Team 52

New set more organized, condensed factors, less rich information (lost: cost of victimization, security, specific reasons no report); collapses variables - confounds results, - age, **household size and poverty**; location (city/state)



Data Sources

New set more organized, condensed factors, less rich information (lost: cost of victimization, security, specific reasons no report); collapses variables - confounds results, - age, household size and poverty; **location (city/state)**

Forest Park, Detroit MI

Oakland Park Township, MI

Same region, population size, MSA classification
vastly different crime rates, demographics

Violent crime

The top 100 Safest Cities in America have a violent crime rate **12.33 times lower** than the national average.

Top 100 average
0.3 per 1,000 people

Compared to the national average
3.7 per 1,000 people

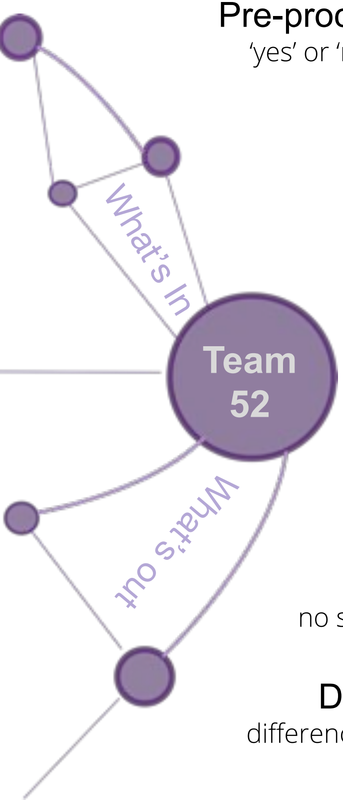
9

Data Characterization/Preparation

New set two main categories of data: Household and Personal. All variables are categorical, with the exception of the year and weight variables.

Pre-processed datasets, already clean, added new 'yes' or 'no' variable notifyTarget (police), eda: all data considered; modeling: all variables not 'yes' or 'no' removed

Dataset	Data archives in .CSV format	Variable descriptions
Personal Victimization Personal victimization includes all violent victimization (rape or sexual assault, robbery, aggravated assault, simple assault) and personal theft.	1993-2019 | 2010-2019 | 2015-2019	Personal
Household Victimization Household victimization includes all property victimization (burglary/trespassing, motor-vehicle theft, and theft).	1993-2019 | 2010-2019 | 2015-2019	Household



Missingness missing values did not exist, those where category meant NA or Out of Universe were removed, imputing when so many factors influence it isn't appropriate.

Correlation highly correlated variables remain in the dataset, during modeling determine best variables to include, no single variable with high correlation to notifyTarget (police).

Duplicate rows found what appear to be 5.6% duplicate rows, checking raw data: similar incidents multiple times with differences in variables removed from new dataset - none removed: valid, indicate 'like' multiple events in the same time period.

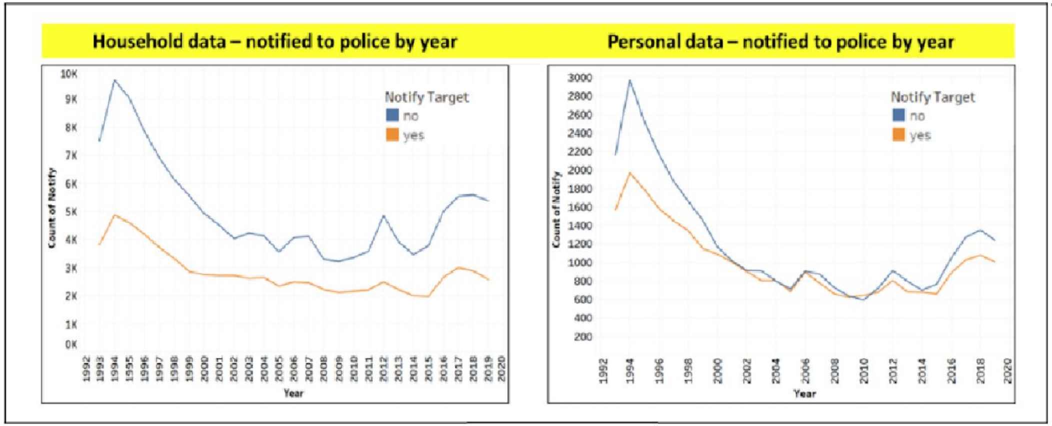
Exploratory Data Analysis: Interface

- Data Analytics Tools
- EDA preparation/ Data Profiling

Team 52

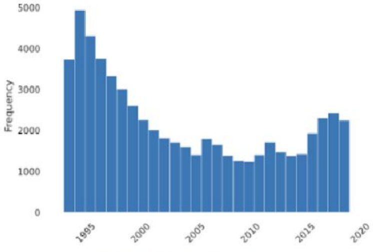
year	weight	gender	race1R	hispanic	ethnic1R	ager	marital2	hincome
2019	3313.01957	2	1	2	1	3	1	88
2019	3313.01957	2	1	2	1	3	1	88
2019	3313.01957	2	1	2	1	3	1	88
2019	1221.08481	2	1	2	1	7	4	88
2019	1221.08481	2	1	2	1	7	4	88

popsize	region	msa	direl	notify	weapon	weapcat	newcrime	newoff
1	2	3	4	2	3	5	1	4
1	2	3	4	2	2	0	1	4
1	2	3	4	2	3	5	1	4
5	4	1	4	1	2	0	1	4
5	4	1	4	1	2	0	1	1



Dataset statistics

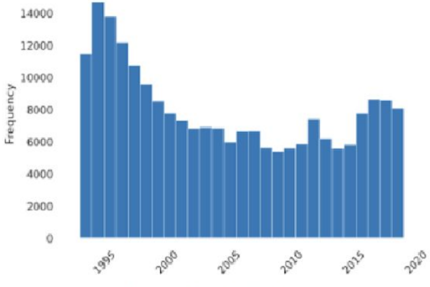
Number of variables	24
Number of observations	60034
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	3379
Duplicate rows (%)	5.6%
Total size in memory	11.0 MiB
Average record size in memory	192.0 B



Histogram with fixed size bins (bins=27)

Dataset statistics

Number of variables	18
Number of observations	216269
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	19467
Duplicate rows (%)	9.0%
Total size in memory	29.7 MiB
Average record size in memory	144.0 B



Histogram with fixed size bins (bins=27)

Exploratory Data Analysis: Interface

- Personal Crime Variable.

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%

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%

Team 52

Exploratory Data Analysis: Interface

- Household Crime Variable.

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%

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%

Team 52

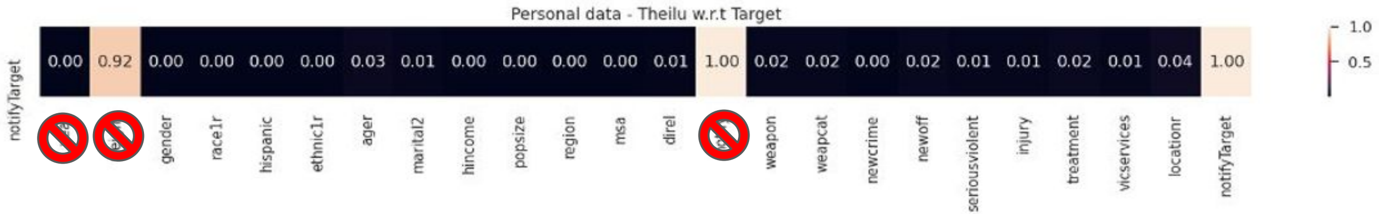
Modeling for pattern identification & Interpretation

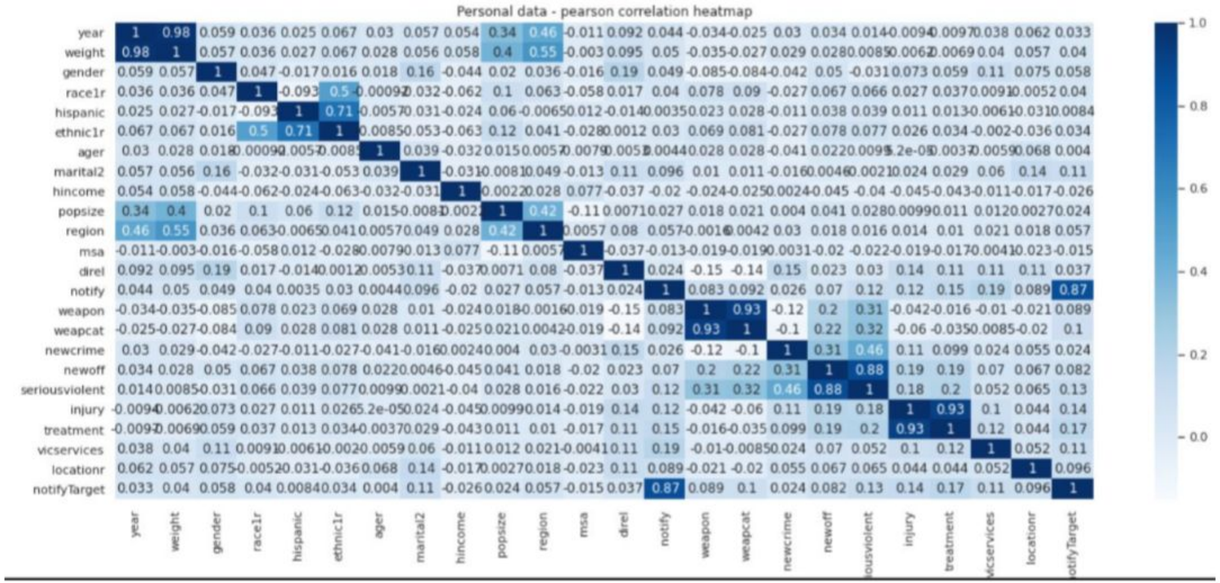
Crime Watch

- Correlation analysis
  - Cramer’s V (Symmetric metric)
  - Theil’s U (Asymmetric metric)
- Variable Encodings
  - Cramer’s V correlation matrix
  - One-hot-encoding
  - Weight of Evidence (WoE)
- Multiple Correspondence analysis(MCA) & Factor Analysis
- T-NSE visualization, PCA & K-means Clustering
- Interpretable models - Experiments
  - Logistic Regression
  - Decision trees
  - XGBoost

Personal Victimizations

EDA - Correlation Analysis



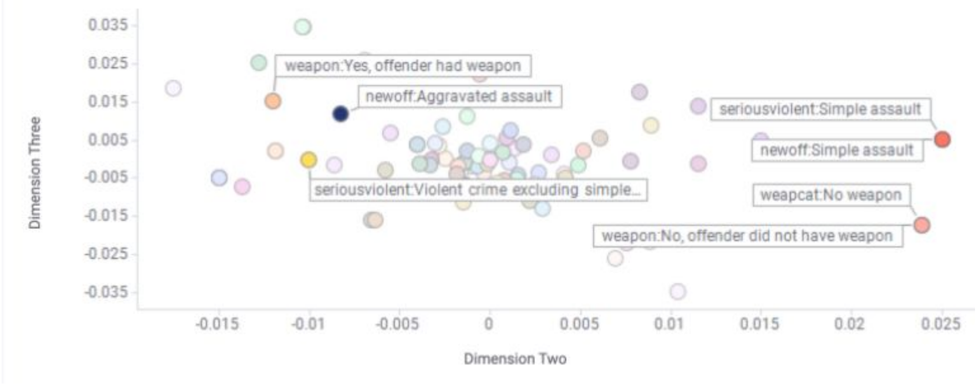


MCA on WoE Encodings

Crime Watch

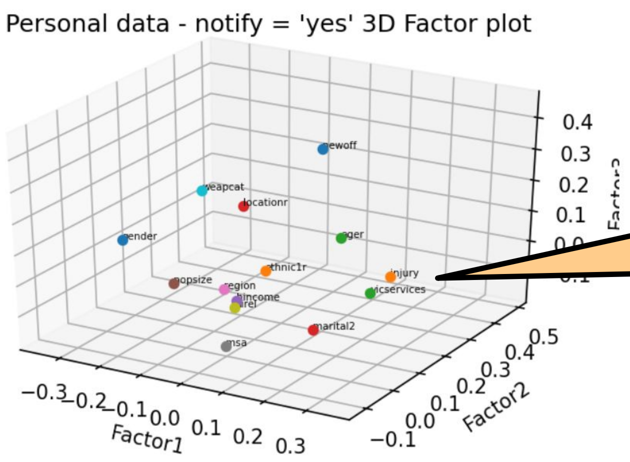
Personal Multiple Correspondence Analysis

- Associations between simple assault and no weapons.
- Serious violent crime, aggravated assault, and the presence of a weapon often occurred together.

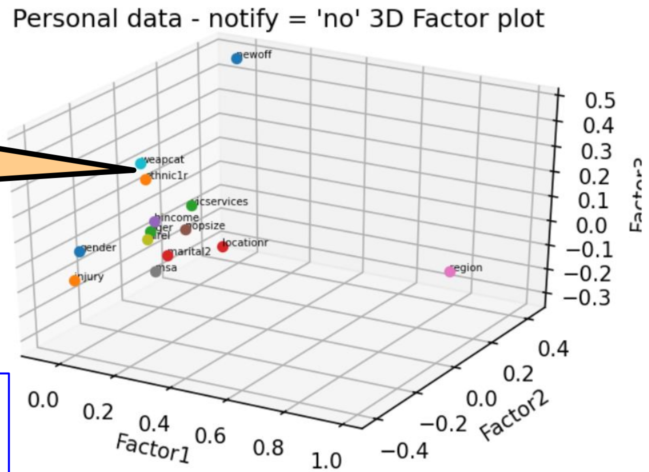


Crime Watch

Factor Analysis on Correlation Matrix



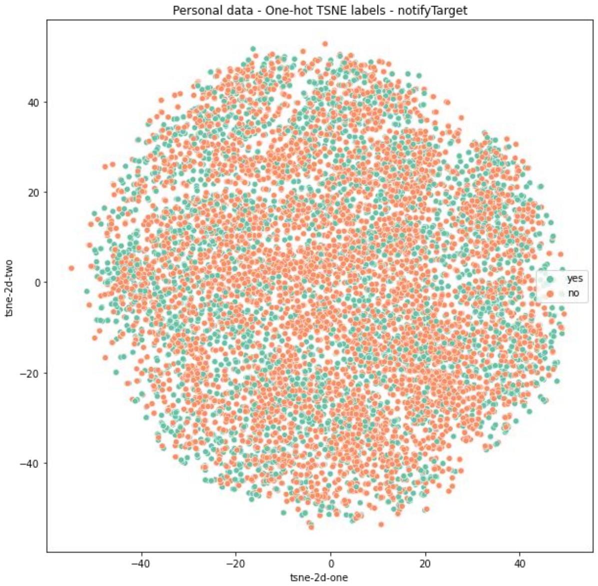
It seems like, it is referring to some psychological factors behind reporting (if married, crime resulted in injuries & victim services received)



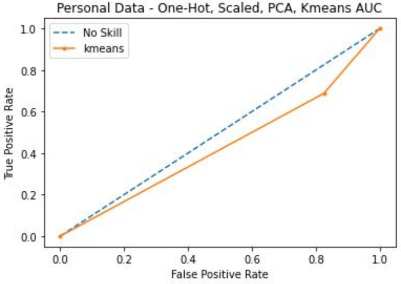
May be people chose not to report where certain demographics and types of crimes are involved

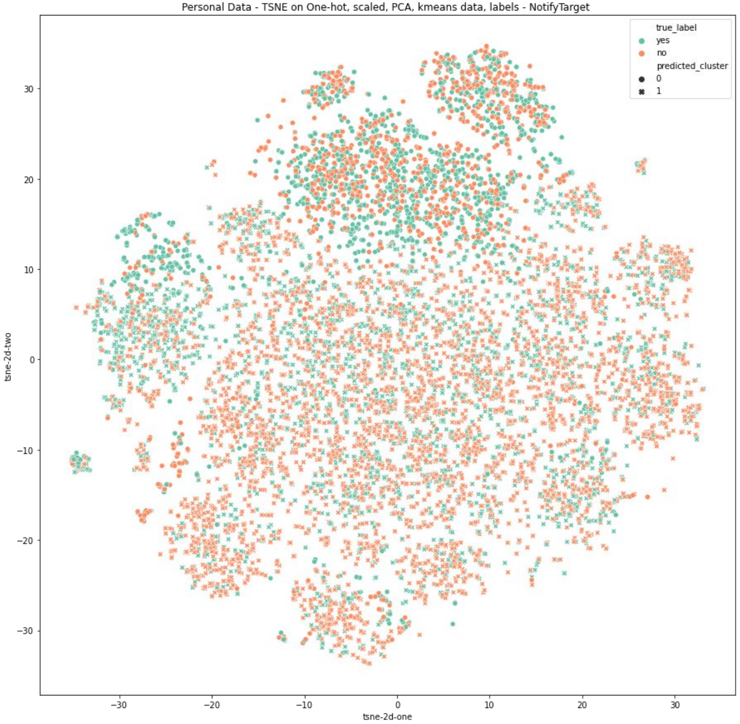
Interpretations can be very subjective but underlying structural differences observed

One-hot T-SNE Visualization & Clustering



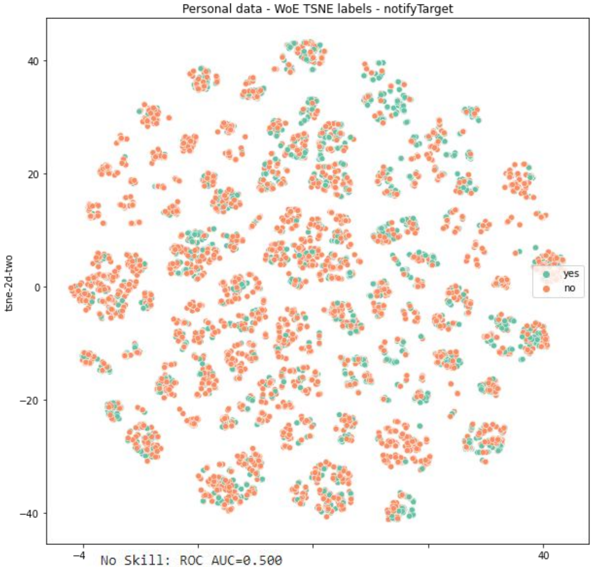
No Skill: ROC AUC=0.500
kmeans: ROC AUC=0.431



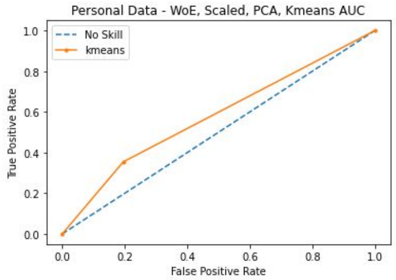


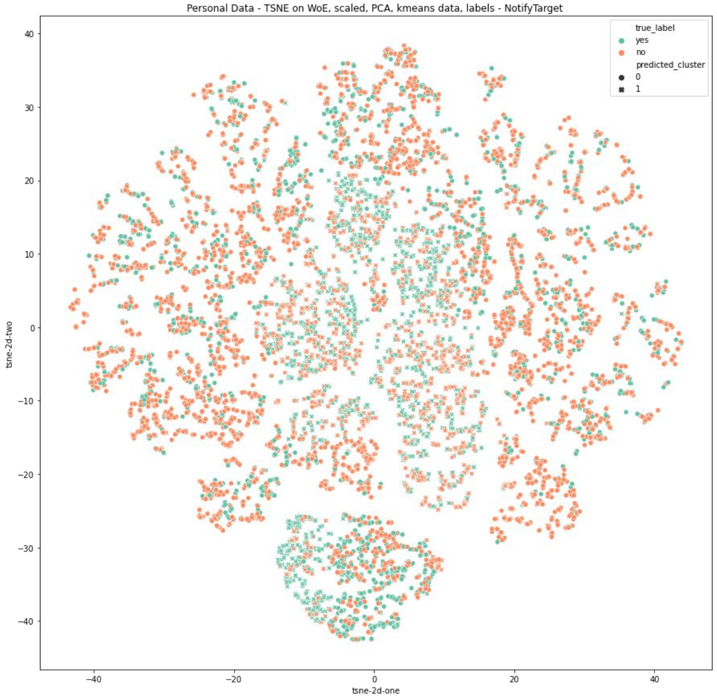
Even though some subgroups in reporting are observed, clustering performance is poor

WoE T-SNE Visualization & Clustering



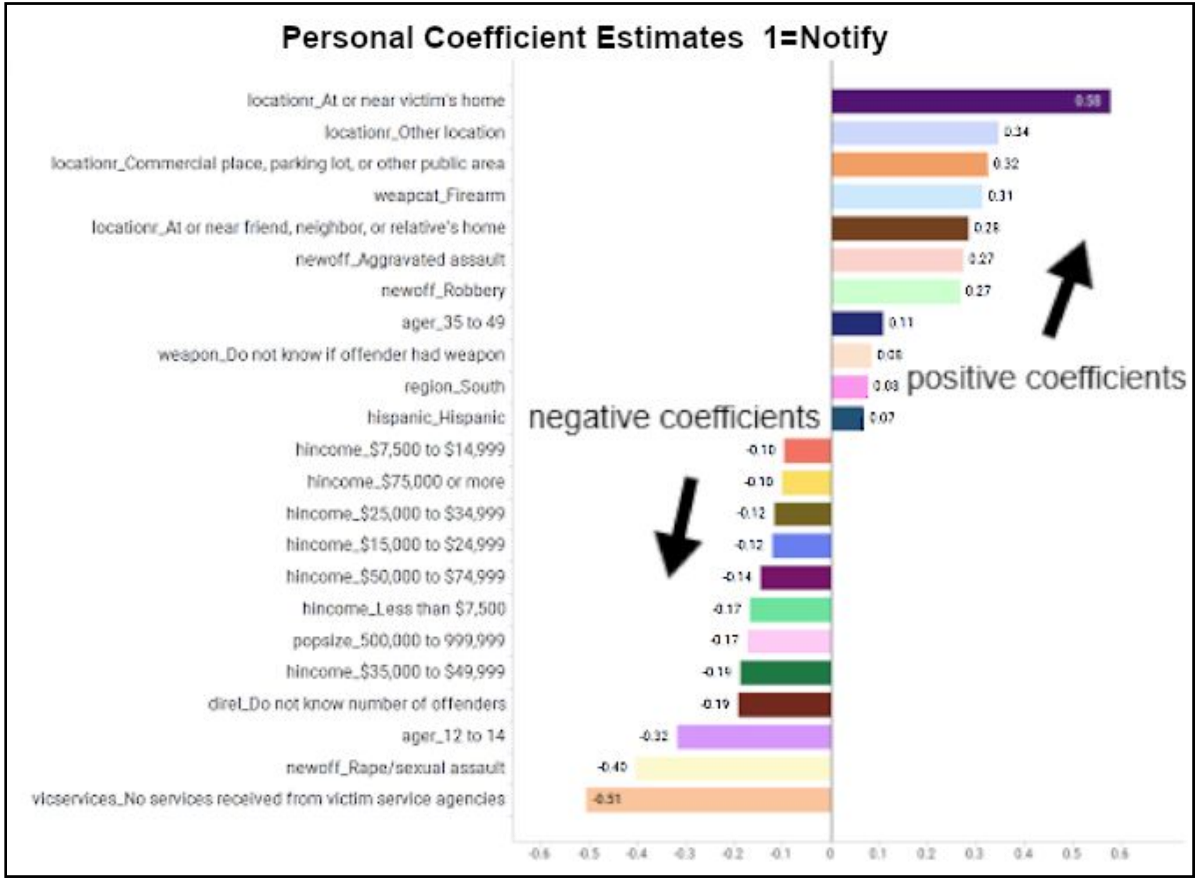

```
No Skill: ROC AUC=0.500
kmeans: ROC AUC=0.580
```





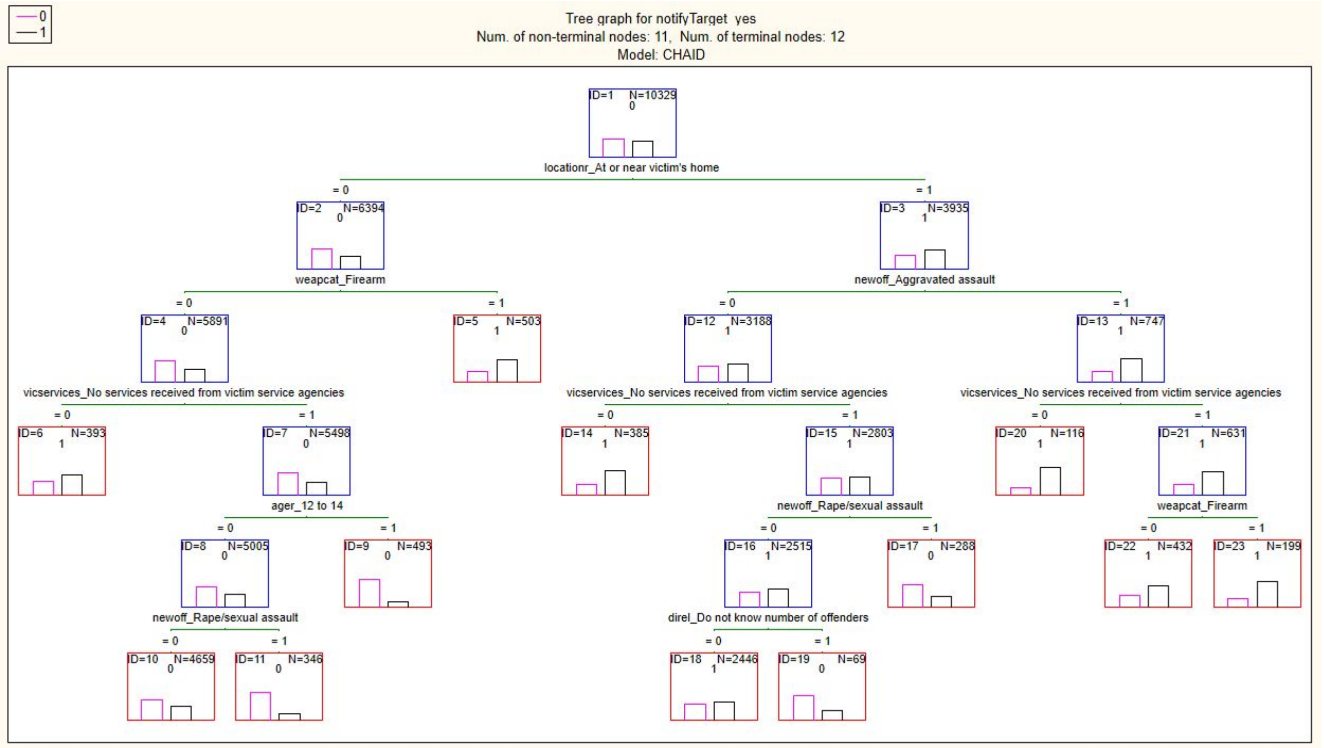
Better performance than one-hot representation. The data has some predictive potential for crime reporting classification

Interpretable Models - Logistic Regression



Decision Tree Results

Crime Watch

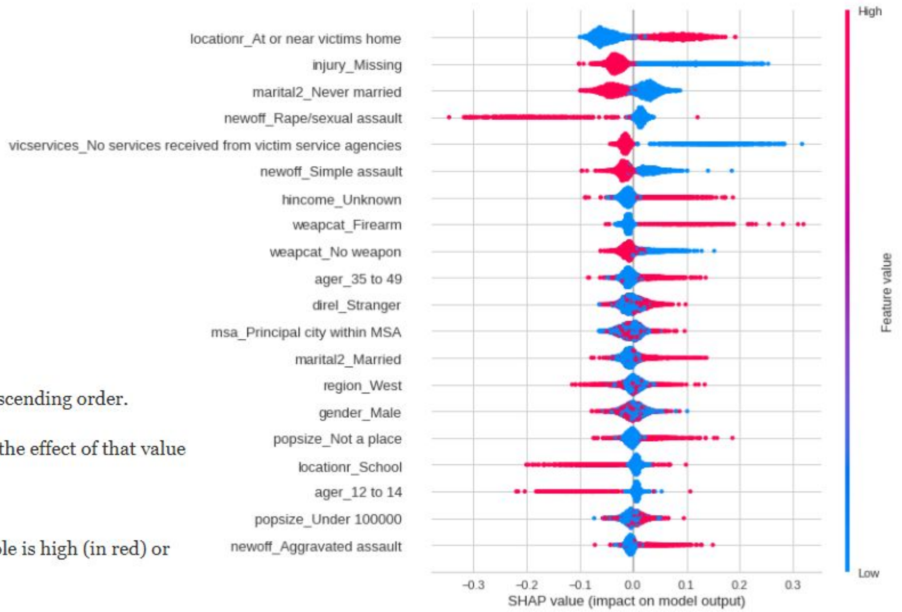


Variable Impact - XGBoost SHAP Values

Personal crime XGBoost (with 1=yes) - 5 fold CV

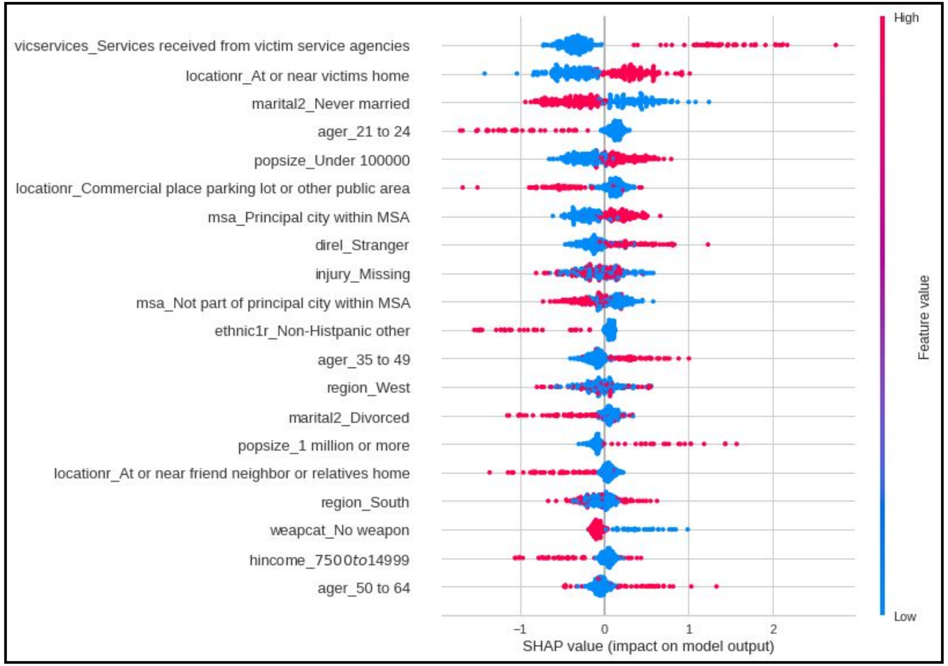
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.6809	0.7379	0.6758	0.6802	0.6792	0.3540	0.3554
1	0.6593	0.7130	0.6557	0.6585	0.6585	0.3124	0.3128
2	0.6560	0.7104	0.6521	0.6550	0.6551	0.3054	0.3058
3	0.6475	0.7053	0.6441	0.6467	0.6469	0.2891	0.2893
4	0.6525	0.7225	0.6478	0.6514	0.6510	0.2972	0.2980
Mean	0.6592	0.7178	0.6551	0.6583	0.6581	0.3116	0.3122
SD	0.0115	0.0115	0.0111	0.0116	0.0112	0.0226	0.0230

- *Feature importance:* Variables are ranked in descending order.
- *Impact:* The horizontal location shows whether the effect of that value *is associated with a higher or lower prediction.*
- *Original value:* Color shows whether that variable is high (in red) or low (in blue) for that observation.



Rape/Sexual Assault Model Variable Impact & Likelihoods

Crime Watch



No services from victim services agencies

0.72

Avg(Probability of not reporting rape and sexual assault)

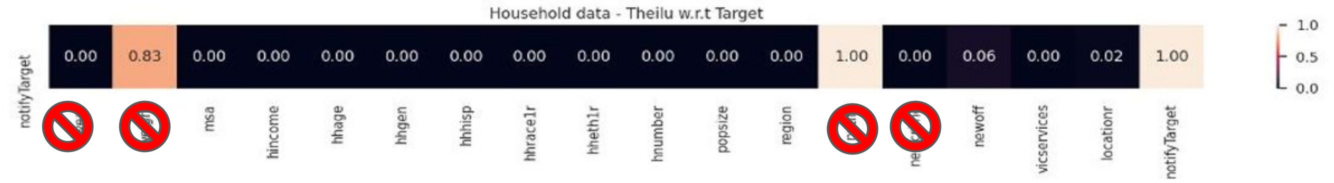
Services from victim services agencies

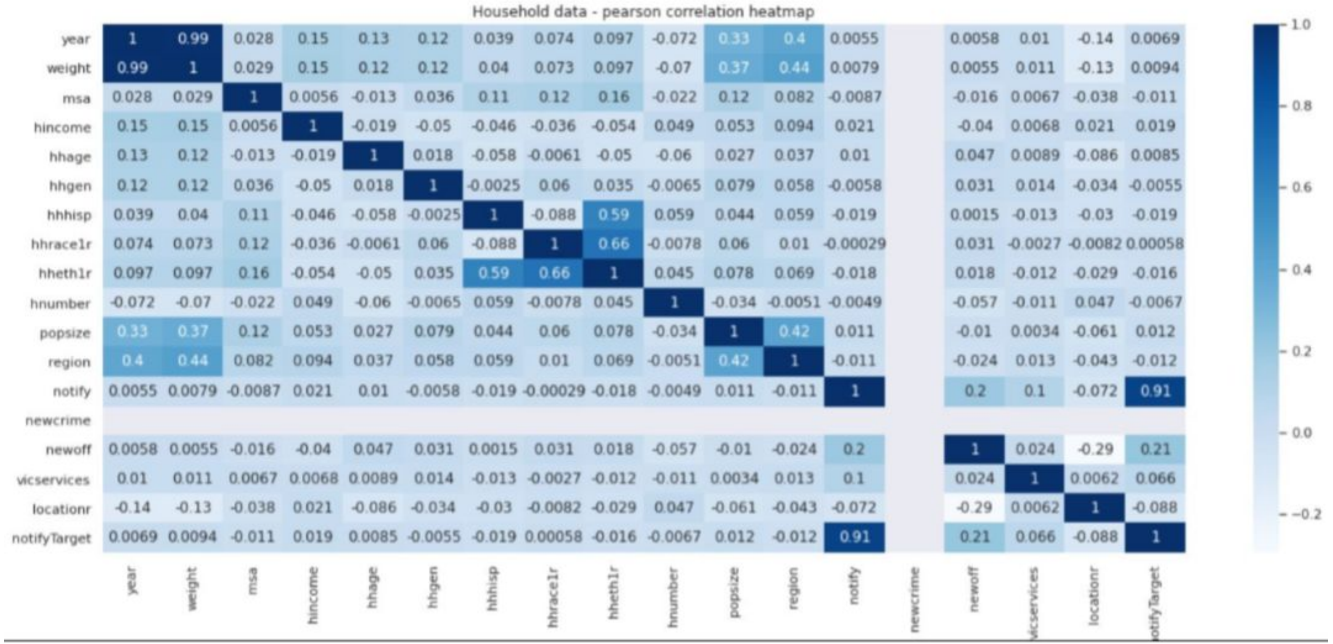
0.58

Avg(Probability of not reporting rape and sexual assault)

Household Victimizations

EDA - Correlation Analysis



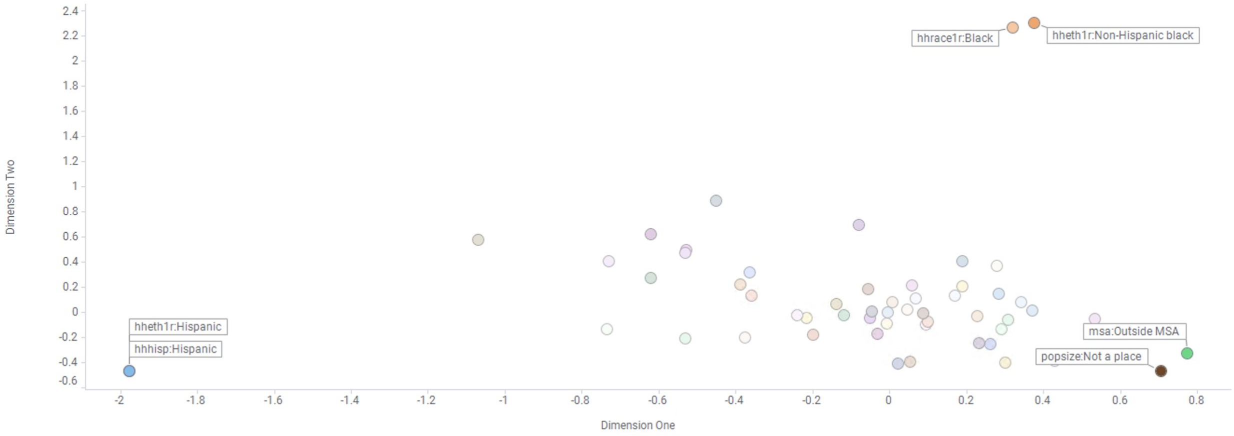


MCA on WoE Encodings

Household Multiple Correspondence Analysis

- Similar experiences, but different, by Hispanic and Black victims.
- Around more rural areas, crime is experienced differently

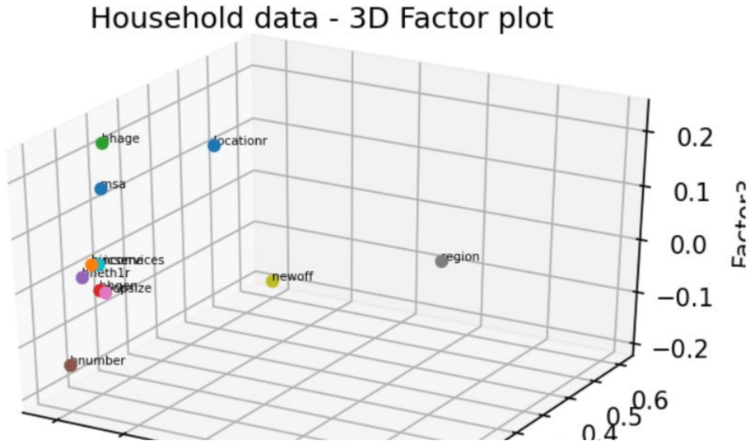
2D Plot of Column Coordinates MCA Household

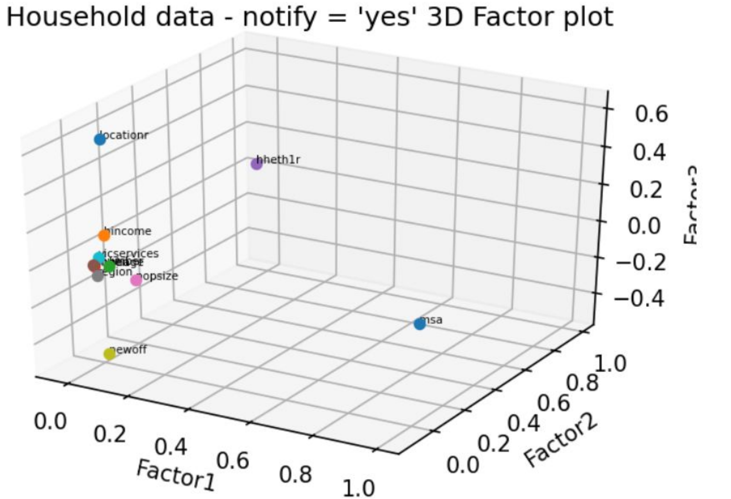


Factor Analysis on Correlation Matrix

In the overall analysis, Factor 1 & 2 may be referring to regional crime types and Factor 3&4 refer to victim's demographics and characteristics. Reporting to police seem to surface ethnicity in certain regions

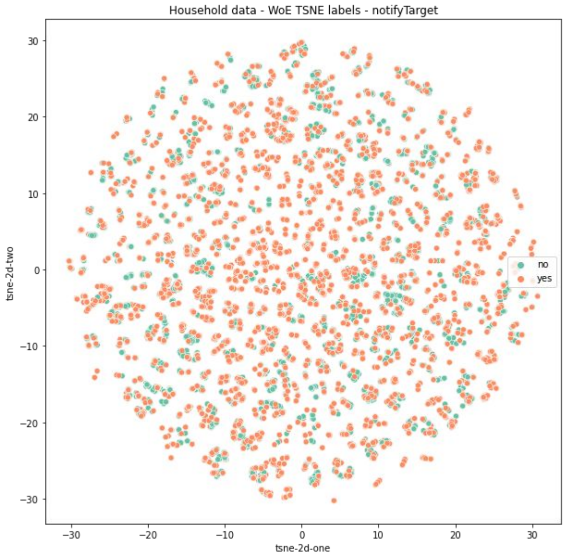
Interpretations can be very subjective but underlying structural differences observed

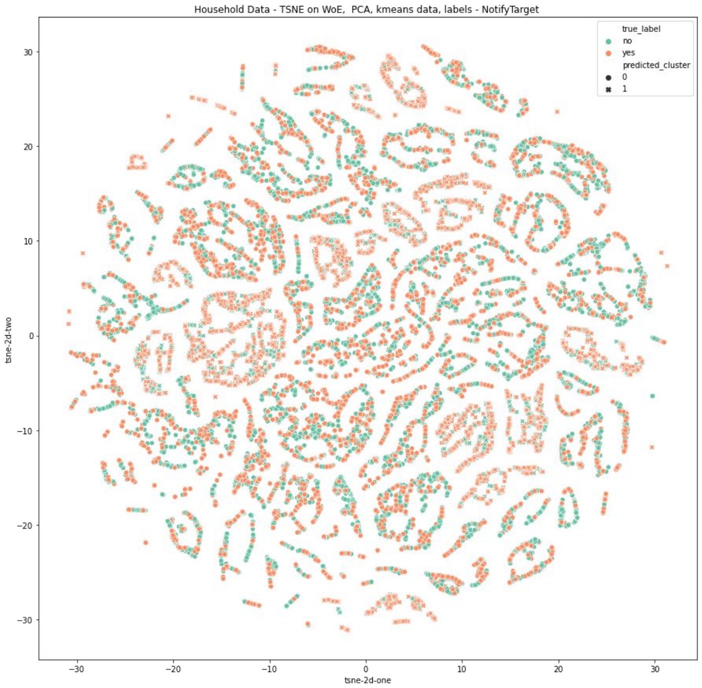


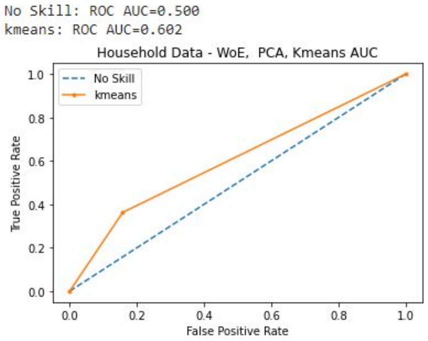


WoE T-SNE Visualization & Clustering

Team
52

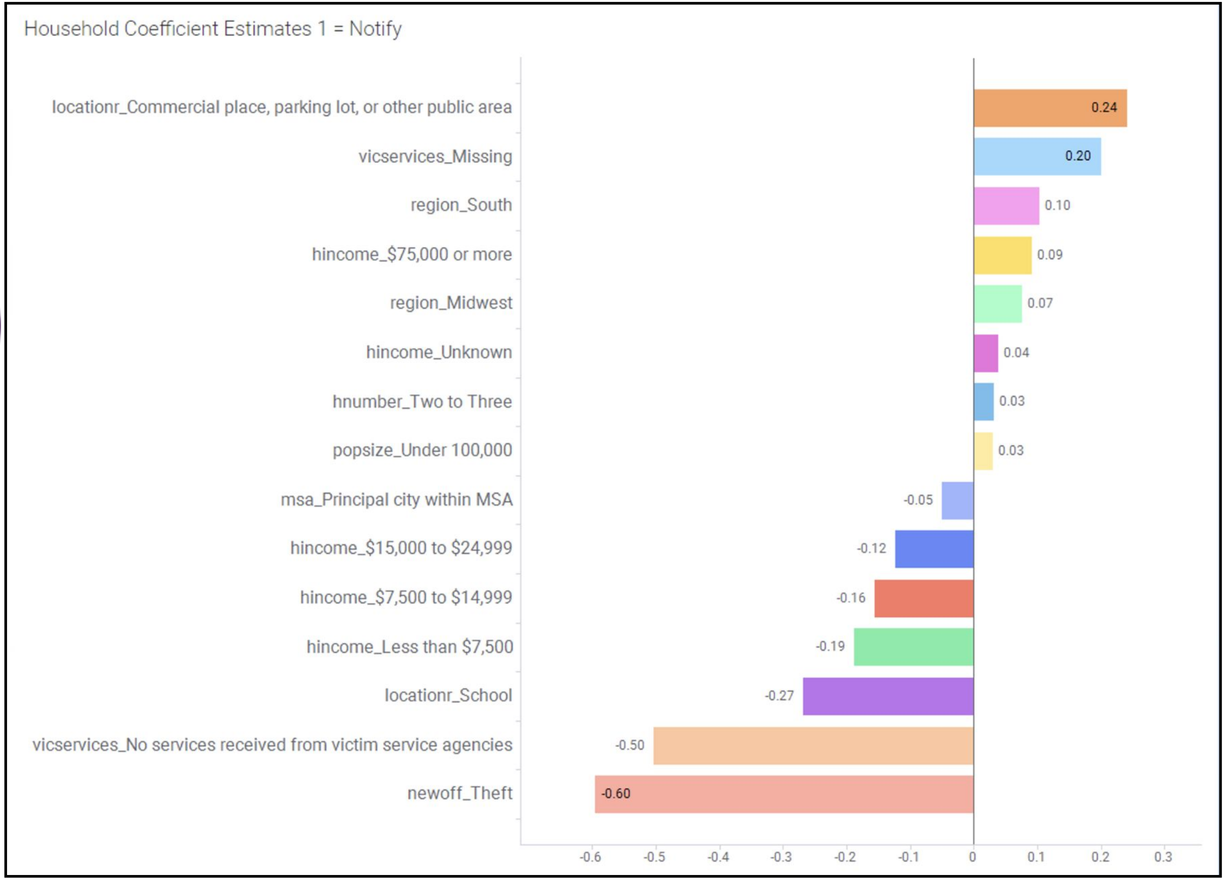




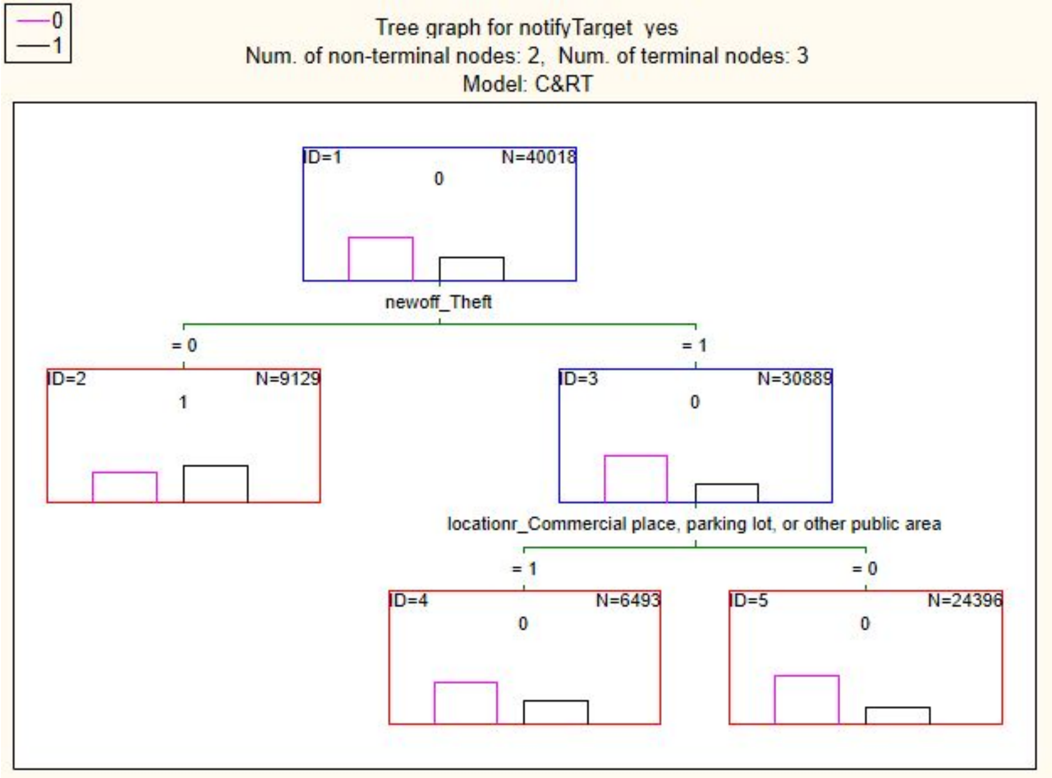


The data has some predictive potential for crime reporting classification

Interpretable Models - Logistic Regression



Decision Tree Results



Variable Impact - XGBoost SHAP Values

Crime Watch

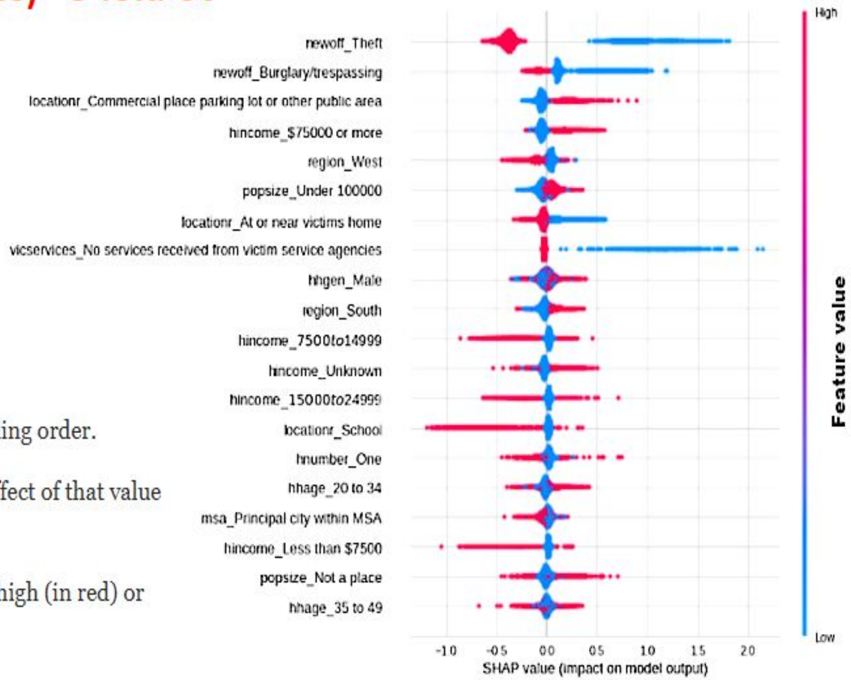
Household crime XGboost (1=yes) - 5 fold CV

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.6849	0.6574	0.5864	0.6647	0.6454	0.1985	0.2262
1	0.6829	0.6496	0.5841	0.6616	0.6431	0.1933	0.2202
2	0.6934	0.6543	0.5985	0.6767	0.6575	0.2249	0.2528
3	0.6867	0.6466	0.5913	0.6669	0.6503	0.2084	0.2339
4	0.6921	0.6645	0.5962	0.6750	0.6553	0.2201	0.2485
Mean	0.6880	0.6545	0.5913	0.6690	0.6503	0.2090	0.2363
SD	0.0041	0.0062	0.0055	0.0059	0.0055	0.0121	0.0125

Feature importance: Variables are ranked in descending order.

Impact: The horizontal location shows whether the effect of that value *is associated with a higher or lower prediction.*

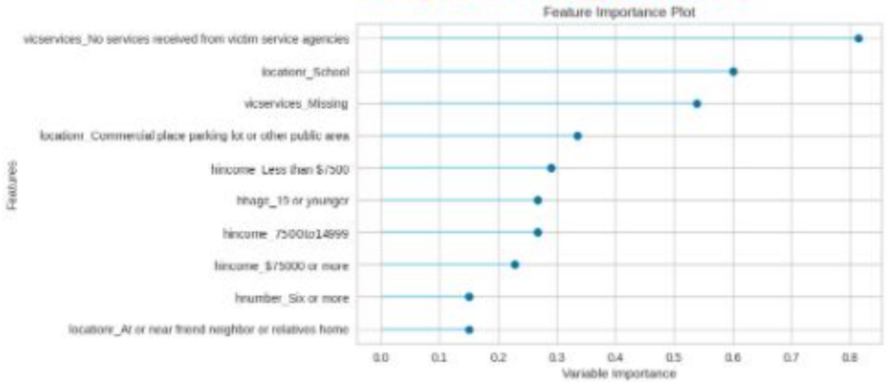
Original value: Color shows whether that variable is high (in red) or low (in blue) for that observation.



Theft Models - Variable Importance & Likelihoods

Crime Watch

Logistic Regression



At School	Not at School
0.82 Avg(Probability of not reporting theft)	**0.71** Avg(Probability of not reporting theft)

No services from victim services agencies	Services from victim services agencies
0.72 Avg(Probability of not reporting theft)	**0.49** Avg(Probability of not reporting theft)

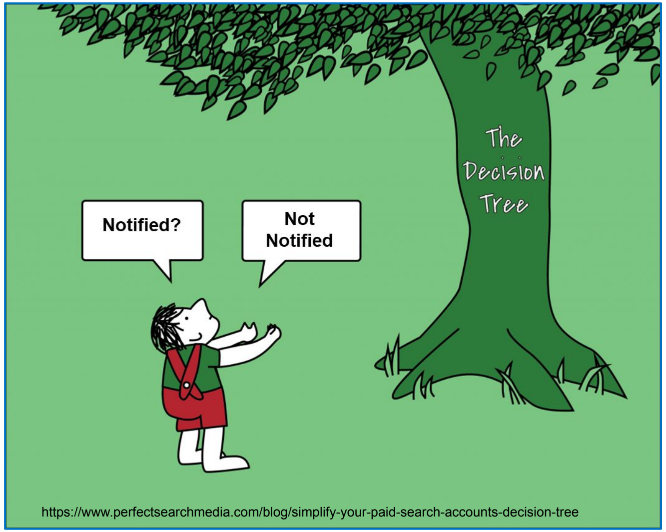
Key Drivers of Not Reporting to Police

More than 10 predictive models were constructed 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 decisionTree-based models were constructed using the following methodologies:

- **Single Tree** algorithms: such as DecisionStump, HoeffdingTree, J48 and REPTree (CART), with different metrics for picking the attributes.
- **Ensemble models:**
  - based on bagging methods: such as RandomForest, RandomTree
  - based on boosting methods: such as AdaBoost with J48, CART and LME as the week learners



In the end the Random Forest algorithm turned out to be the front-runner in terms of consistent performance across both the datasets.

Personal Crimes

The models indicated crime type as a primary driver behind 'reporting to police' choice.

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

Methodology	Decision tree algorithm	All variables Accuracy	All variables ROC	Pearson's Accuracy	Pearson's ROC	Information Gain Accuracy	Information Gain ROC	Learner Based Accuracy	Learner Based ROC	Set 1 Accuracy	Set 1 ROC	Set 2 Accuracy	Set 2 ROC
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										

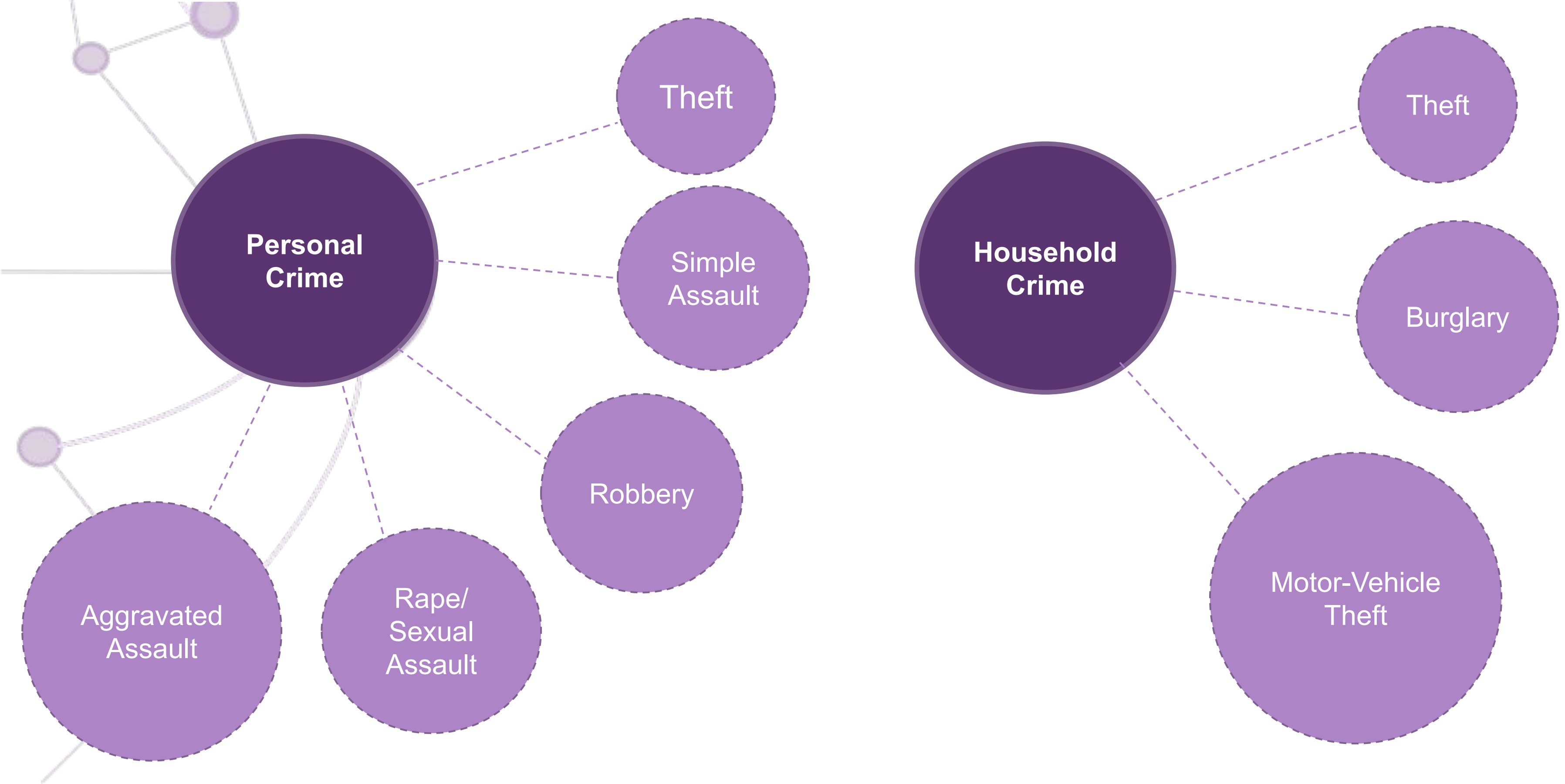
Household Crimes

The models indicated crime type as a primary driver behind 'reporting to police' choice.

		All variables		Top 8 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	
Methodology	**Decision tree algorithm**	**Accuracy**	**ROC**	**Accuracy**	**ROC**	**Accuracy**	**ROC**	**Accuracy**	**ROC**	**Accuracy**	**ROC**	**Accuracy**	**ROC**
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)												

Likelihood of Reporting to Police

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.



Likelihood of Reporting to Police

Personal Crime

	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.		
			Training model including only the most important variables		
	Accuracy	ROC	Most informative attributes	Accuracy	ROC
Theft	62.6%	0.66	1. ager 2. marital2 3. hincome 4. msa 5. direl	64.3%	0.67
Simple Assault	68.3%	0.74	1. ager 2. treatment 3. vicservices 4. locationr	63.3%	0.65
Robbery	67.1%	0.70	1. race1r 2. ager 3. region 4. weapcat 5. treatment 6. vicservices 7. locationr	63.7%	0.67
Rape/Sexual Assault	77.1%	0.78	1. weapcat 2. Treatment 3. vicservices	72.3%	0.67
Aggravated Assault	68.6%	0.72	1. ager 2. weapcat 3. treatment 4. vicservices 5. locationr	62.0%	0.63

Likelihood of Reporting to Police

Household Crime

	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.		
			Training model including only the most important variables		
	Accuracy	**ROC**	**Most informative attributes**	**Accuracy**	**ROC**
Theft	68.2%	0.60	1. vicservices 2. locationr	71.4	0.55
Motor-vehicle theft	79.0%	0.60	1. vicservices 2. locationr	78.2	0.56
Burglary	59.0%	0.63	1. hincome 2. vicservices	54.4	0.55

Key Findings

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.

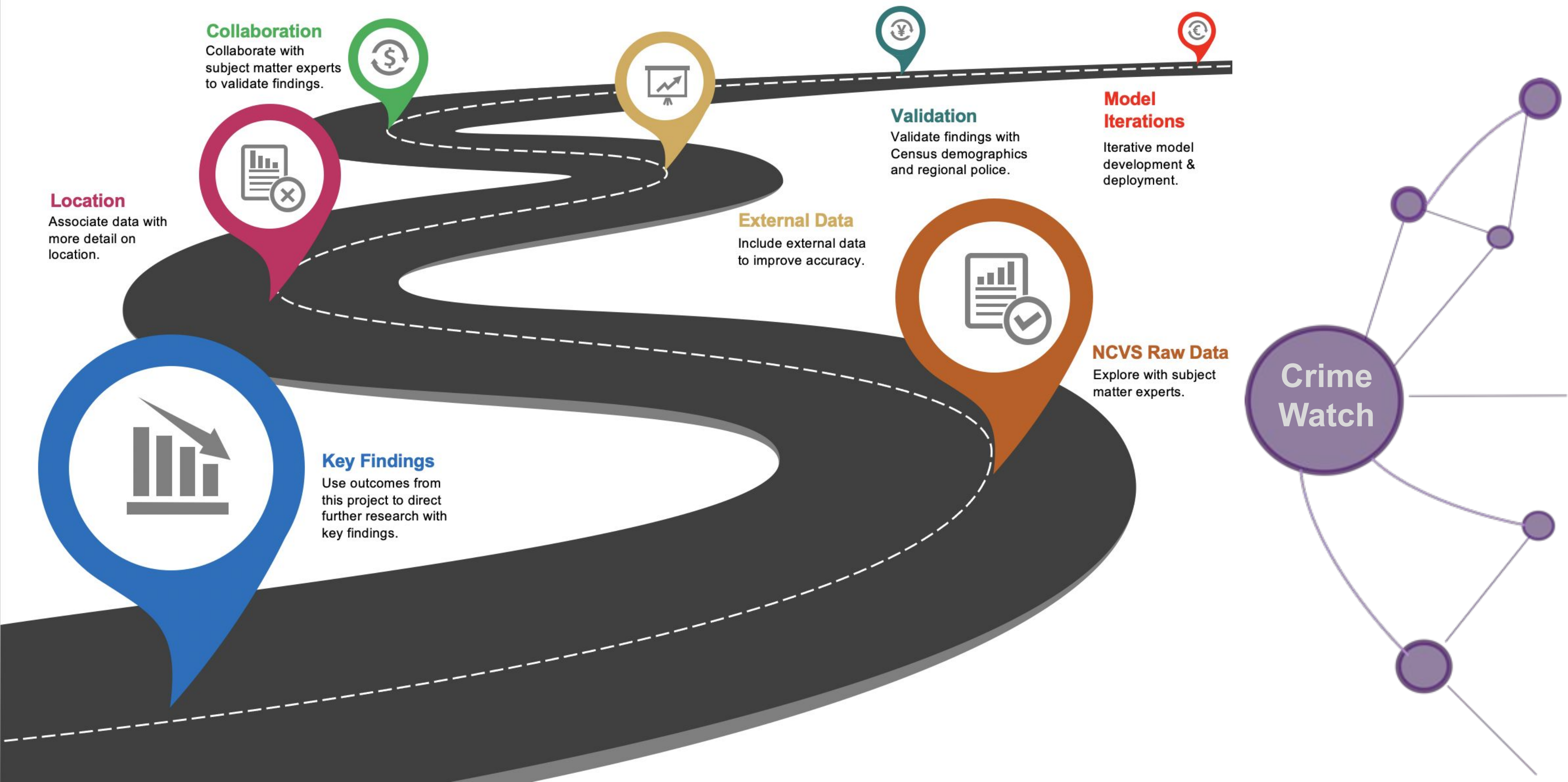
RAPE/SEXUAL ASSAULT

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

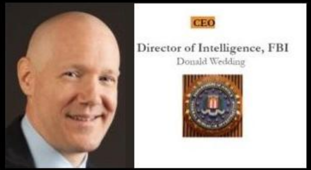
DEMOGRAPHICS

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

Recommendations



Thank you



Delivering solid, actionable intelligence via data science / AI tools

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.

"Your perspective is very positive, and class video's are really funny which makes the class engaging ~ they make the hard work and crazy timeline bearable."
-Sheila

"Thanks for your words of wisdom and emphasis on practical applications of data science." -Jeremy

"I sincerely appreciate your useful posts during this course. I specially befitted from your words of wisdom." -Damon

"To me words without wisdom and work without wit feels very boring, you provided both of them in plenty, can't ask for more :-)." -Setu

I really appreciate your hard work for preparing us these great content. -Jin