George Mason University
DAEN 690
Summer 2019

Knowledge Discovery for Crimes Against Children

Team KD-CAC

Robert Grillo, *Product Owner*
Waliyat Olayiwola, *Scrum Master*
Xianci Tang, *Developer*
Nikhitha Lingutla, *Developer*
Yao Zhang, *Developer*

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Abstract

This project seeks to discover knowledge about crimes against children through opportunistic data mining of the National Incident-Based Reporting System (NIBRS) crime data source. The main goal is to identify the characteristics of victims, offenders, and the situational threats in which children are especially vulnerable to particular categories of offenses. The team applies a breadth of approaches, with an emphasis on the use of interpretable data science techniques, including hierarchical clustering, penalized logistic regression, tree-augmented Bayesian networks, and gradient boosting with XGBoost. The resulting XGBoost models are interpreted using Shapley Additive Explanation (SHAP) and Local interpretable model-agnostic explanations (LIME).

Hierarchical clustering of features identified combinations of features sharing similarities across observations to highlight relevant scenarios within the data, such as the kidnapping of children aged 5 to 8 by a female parent aged 24 to 41 between the hours of 0600 to 1100 in the East-South-Central country division. The tree-augmented Bayesian networks were built in the IBM Watson Studio SPSS Modeler, which provides an interface for model deployment and testing. Offense characteristics can be entered to calculate the likelihood that a crime belongs to a particular offense category versus others. Interpretation of the XGBoost models revealed that teenagers are more vulnerable to sex offenses committed by 20 to 40-year-old offenders than younger children; however, this serves as a reminder that sex offenses against young children are known to be under reported. The geospatial visualizations of the project did not show a strong correlation between poverty rates and crime rates across U.S. counties.

There are a vast number of ways to expand upon this study. Recommendations include adding features to the data set, developing multi-class or multi-label models, applying time series analysis techniques to the data, creating subject matter expert informed Bayesian belief networks, and the analysis of other vulnerable population types, such as the elderly.

1 Introduction

1.1 Background and Rationale

Crimes committed against children, from birth through age 17, and issues related to child victimization are a problem of growing concern across the United States. The level of attention paid to such matters speaks volumes to the nature of the society in which they occur. In the current age of big data, it can be seen as a moral obligation to uncover as much information as possible about this topic using advanced analytical techniques on the most detailed crime data available. The *Knowledge Discovery for Crimes Against Children* project was designed and executed to explore solutions for this challenge.

1.2 Prior Research

According to David Finkelhor (1999), a researcher at the Crimes Against Children Research Center (CRCC), the different crimes committed against children consist of sexual assault, abduction, aggravated assault, child abuse in all its forms (physical, sexual, emotional), and child neglect. These issues can often leave child victims feeling alone and helpless. Despite substantial publicity about crime and children, the number of offenders is much more widely recognized than the disproportionate number of victims (Finkelhor, 1999).

Finkelhor and Shattuck (2012) studied characteristics of crimes against children using the National Incident-Based Reporting System (NIBRS) dataset through "an exploratory data analysis" and discovered that an estimated 1.3 million crimes against children victims came to the attention of police in the U.S. as a whole in 2008, comprising 9 percent of all crime victims. Out of these, 187,100 sex offenses against children were reported to the police, constituting 66 percent of sex crime victims of all ages, adult and children. According to a nationwide crime victimization survey, the sexual assault victimization rate for youths under 18 is 2.7 times (170%) higher than for adults, 51% of lifetime rapes occur prior to the age of 18, and 29% occur prior to the age of 12 (Kilpatrick et al., 1992).

Finkelhor and Shattuck go on to explain how crimes against children involve special investigatory and prosecutorial challenges due to how these crimes are reported to law enforcement officials. There are also challenges related to obtaining information from child victims, where approximately one quarter of them are under the age of 12. The intimate characteristics of the perpetrators introduce further complications, where 26 percent of whom are family members and 63 percent of whom are acquaintances (Finkelhor, & Shattuck, 2012). These factors help answer the question as to why such crimes remain underreported.

1.3 Project Objectives

This project attempts to discover data-driven knowledge about crimes against children through opportunistic data mining. The main goal is to identify the characteristics of victims, offenders, and the situational threats in which children are especially vulnerable to particular types of crimes. Predictive and unsupervised learning methods can help identify subgroups of children and incident characteristics for populations that are at risk, some of which may have not yet been recognized. Characteristics can

also be discovered using geospatial visualization techniques to explore the data collected for this project.

1.4 Problem Space and User Stories

The project team attempted to determine what knowledge can be discovered about crimes against children through the exploration and mining of the National Incident-Based Reporting System (NIBRS) crime dataset. These attempts at knowledge discovery emphasized the use of interpretable data science methods. This project exclusively focused on crimes against children (those younger than 18). Additionally, crimes involving property or against society were ignored.

The team used the Agile project management methodology to execute this study with a series of five Sprints over a 12-week period. Further information about the Agile management approach and the project plan that was executed can be found in Appendix B. This appendix also contains Information pertaining to the risks identified during the early phases of the project and their respective mitigation strategies.

1.4.1 Primary User Story:

Through training modules and intelligence bulletins, project findings can ultimately influence subtle decisions in the minds of law enforcement officers, social workers, school teachers/administrators, parents, childcare providers, and healthcare providers as they find themselves in the presence of children at risk of victimization. Law enforcement agencies across the world have resources specifically dedicated to crimes against children. These investigators would have particular interest in a project of this nature. In addition, the knowledge discovered from this project can potentially inspire researchers in the social sciences to investigate previously unrecognized threats to specific subpopulations of children.

With the goals established within the above primary user story in mind, the project team formulated the following research questions:

1) What characteristics of crimes against children are most associated with each type of offense category?
2) How can NIBRS data be used to predict the likelihood of a victimized child falling into a particular offense category given various known circumstances?
3) To what extent are state/county-level socioeconomic indicators associated with crimes against children?

1.4.2 Solution Space

The data used in the project includes four years of crime data from 2013 to 2016. The definition of rape changed in NIBRS starting in 2013, thus making it difficult to integrate with data from previous years (U.S. FBI 2012). 2017 data is not expected to be publicly available for NIBRS in its complete format until September 2019. Given the evolving nature of agency participation in the NIBRS program, the

project did not include in-depth temporal trend analysis beyond the application of simple forecasting methods on monthly crime rates.

1.4.3 Project Assumptions

The project team made a series of assumptions while defining the problem. It was assumed that NIBRS is a reliable source of crime data (i.e., the incomplete and inconsistent agency participation in NIBRS would not overly influence findings and missing data would not limit analyses). For modeling purposes, it was assumed the data contains a sufficient number of characteristics to represent the crimes that took place and methods exist to provide a useful level of model interpretation. Regarding project planning and execution, it was assumed the initial project plan could be completed within a 12-week period.

1.5 Product Vision and Sample Scenarios

The project team envisioned the study results being used by law enforcement officers, social workers, school teachers/administrators, parents, childcare providers, healthcare providers, and researchers in the social sciences to better understand the characteristics that distinguish among categories crimes against children. The use of statistical learning techniques on NIBRS crime incident data is a new way of researching the topic and has the potential to reveal new information about the characteristics of crimes against children for stakeholders to consider as they execute their respective missions. Unlike the traditional methods of testing hypotheses proposed by human intuition, the predictive and unsupervised learning models developed during this project could reveal patterns in the data that may only be recognized through these analytical techniques. Prior to initiating the project, it was also recognized that informative patterns may not exist in the data or are obscured by the incompleteness of the information captured in the NIBRS data source.

Three sample scenarios are outlined in the following subsections to demonstrate the value that can be delivered for the potential users of the knowledge discovered from this project. The three scenarios mirror the three research questions outlined in Section 1.4.1.

1.5.1 Scenario #1

A law enforcement agency needs to produce a training module for its officers to raise awareness of the characteristics of crimes against children that are most associated with particular types of offense categories. The literature they find is largely anecdotal. The unsupervised and supervised methods executed during this project can reveal new, data-driven information to improve a training module of such a nature.

1.5.2 Scenario #2

A criminal investigator is provided a limited set of details related to a crime that took place, or is believed to have taken place against a child, including information associated with various suspects and possible locations where the crime may have taken place. The investigator can use the predictive models developed in this project to run “what-if” scenarios and estimate the likelihood of different offense categories given the inputs provided for the scenario.

1.5.3 Scenario #3

A parent is considering moving to another state or county in the United States and is concerned about his or her children being a victim of a particular type of crime. The geospatial analysis produced in this project can inform the parent of the crime type rates and how the socio-economic conditions of the area might influence the propensity of various types of crimes against children.

1.6 Definition of Terms:

See Appendix C.

2 Data Acquisition

2.1 NIBRS Overview:

NIBRS is one of the largest and most comprehensive publicly available crime datasets. The Federal Bureau of Investigation (FBI) uses NIBRS to collect and maintain vast amounts of data from participating agencies to present a nationwide view of crime (NACJD 2019). A well-defined codebook allows for a common reporting language to be used across different criminal statutes, making it a unique source of uniform crime reporting for a project of this nature (ASUCRP 2019).

Data in NIBRS is organized in complex ways to reflect the many different aspects of criminal incidents. There are over 2 GB of data per year and over 5.5 million offenses per year in the timeframe under study in this project. The data is incident-focused with up to 10 offenses per incident for crimes against persons, property, and society. Data fields include 46 location types, victim characteristics, offender characteristics, and variables reflecting the situational aspects of each crime.

NIBRS is composed of 11 flat files for each master file year. The project dataset was created from five of these flat files, representing the Batch Head, Administrative, Offense, Victim, and Offender segments of NIBRS. Leveraging the relational design of the NIBRS datasets, the project team merged the applicable files into a single dataset for the purposes of this study.

2.2 Description of Used NIBRS Fields:

The following descriptions are provided for the NIBRS fields used to generate the base project dataset. Descriptive text written in italics is directly quoted from the NIBRS codebook (U.S. FBI 2016). The first six fields are dichotomous target variables, four of which are used as dependent variables throughout this project. Fields 7 through 58 were considered predictors while entering the model development phase of the project. Fields 59 through 81, which can be found in Appendix D, were used for generating the geospatial visualization dataset, or for other referencing purposes.

1) **TARGET.Assault** (Type: numeric) – This binary target variable distinguishes between offenses that are categorized as assaults versus those that are not assaults.
2) **TARGET.Homicide_Nonnegligent** (Type: numeric) – This binary target variable distinguishes between offenses that are nonnegligent homicides versus those that are not nonnegligent homicides.
3) **TARGET.Negligent_Manslaughter** (Type: numeric) – This binary target variable distinguishes between offenses that are negligent manslaughters versus those that are not negligent manslaughters.
4) **TARGET.Kidnaping_Abduction** (Type: numeric) – This binary target variable distinguishes between offenses that are kidnappings/abductions versus those that are not kidnappings/abductions.
5) **TARGET.Human_Trafficking** (Type: numeric) – This binary target variable distinguishes between offenses that are human trafficking versus those that are not human trafficking.
6) **TARGET.Sex_Offense** (Type: numeric) – This binary target variable distinguishes between offenses that are sex offenses versus those that are not sex offenses.

7) **V4018.AGE.OF.VICTIM** (Type: numeric) – This is the reported age of the victim.
8) **V4019.SEX.OF.VICTIM** (Type: string) – This is the reported sex of the victim.
9) **V4020.RACE.OF.VICTIM** (Type: string) – This is the reported race of the victim.
10) **V4021.ETHNICITY.OF.VICTIM** (Type: string) – This is the reported ethnicity of the victim.
11) **V4022.RESIDENT.STATUS.OF.VICTIM** (Type: string) – *The test for determining if the victim is a resident is whether the victim maintains his/her permanent home for legal purposes in the locality (i.e., town, city, or community) where the crime occurred.*
12) **OFFENDER.RELATIONSHIP** (Type: string) – *The relationship(s) of the victim to each offender (up to 10 offenders involved in the incident) are present. If there are more than 10 offenders, the 10 closest in relationship will be present.* The Victim segment contains 10 victim-offender relationship fields for the 10 possible offenders.
13) **V5007.AGE.OF.OFFENDER** (Type: numeric) – This is the reported age of the offender.
14) **V5008.SEX.OF.OFFENDER** (Type: string) – This is the reported sex of the offender.
15) **V5009.RACE.OF.OFFENDER** (Type: string) – This is the reported race of the offender.
16) **V5011.ETHNICITY.OF.OFFENDER** (Type: string) – This is the reported ethnicity of the offender.
17) **V2007.OFFENSE.ATTEMPTED.COMPLETED** (Type: string) – *This indicates if the offense was completed or merely attempted.*
18) **V2011.LOCATION.TYPE** (Type: string) – *This is where this offense occurred. Each offense will contain the location type in a multi-offense incident.*
19) **LOCATION.CATEGORY** (Type: string) – NIBRS has 46 location types. The project team used human judgement to consolidate the 46 location into 18 location categories.
20) **CIRCUMSTANCES.ARGUMENT** (Type: numeric) – This binary variable is derived from a set of two fields that describe the circumstances of either an aggravated assault or homicide offense. This particular binary variable identifies whether or not an argument took place during the incident.
21) **CIRCUMSTANCES.DRUG.DEALING** (Type: numeric) – This binary variable is derived from a set of two fields that describe the circumstances of either an aggravated assault or homicide offense. This particular binary variable identifies whether or not drug dealing took place during the incident.
22) **CIRCUMSTANCES.GANG** (Type: numeric) – This binary variable is derived from a set of two fields that describe the circumstances of either an aggravated assault or homicide offense. This particular binary variable identifies whether or not the incident was gang related.
23) **CIRCUMSTANCES.LOVER.QUARREL** (Type: numeric) – This binary variable is derived from a set of two fields that describe the circumstances of either an aggravated assault or homicide offense. This particular binary variable identifies whether or not a lover's quarrel was a factor in the incident.
24) **CIRCUMSTANCES.OTHER.FELONY** (Type: numeric) – This binary variable is derived from a set of two fields that describe the circumstances of either an aggravated assault or homicide offense. This particular binary variable identifies whether or not a felony of another type took place during the incident.
25) **CIRCUMSTANCES.OTHER.CIRC** (Type: numeric) – This binary variable is derived from a set of two fields that describe the circumstances of either an aggravated assault or homicide offense. This particular binary variable identifies whether or not other circumstances took place during the incident.

26) **CIRCUMSTANCES.NEGLIGENT.WEAPON** (Type: numeric) – This binary variable is derived from a set of two fields that describe the circumstances of either an aggravated assault or homicide offense. This particular binary variable identifies whether or not negligence with a weapon was a factor in the incident.
27) **CIRCUMSTANCES.NEGLIGENT.OTHER.KILLING** (Type: numeric) – This binary variable is derived from a set of two fields that describe the circumstances of either an aggravated assault or homicide offense. This particular binary variable identifies whether or not negligence resulting in a killing without a weapon was a factor in the incident.
28) **INJURY.TYPE.MINOR** (Type: numeric) – This binary variable is derived from a set of five fields that describe the victim's injuries resulting from the incident. This particular binary variable identifies whether or not the victim had minor injuries.
29) **INJURY.TYPE.BROKEN.BONES** (Type: numeric) – This binary variable is derived from a set of five fields that describe the victim's injuries resulting from the incident. This particular binary variable identifies whether or not the victim had broken bones.
30) **INJURY.TYPE.OTHER.MAJOR** (Type: numeric) – This binary variable is derived from a set of five fields that describe the victim's injuries resulting from the incident. This particular binary variable identifies whether or not the victim had other major injuries.
31) **INJURY.TYPE.INTERNAL** (Type: numeric) – This binary variable is derived from a set of five fields that describe the victim's injuries resulting from the incident. This particular binary variable identifies whether or not the victim had internal injuries.
32) **INJURY.TYPE.LOSS.TEETH** (Type: numeric) – This binary variable is derived from a set of five fields that describe the victim's injuries resulting from the incident. This particular binary variable identifies whether or not the victim lost teeth.
33) **INJURY.TYPE.LACERATION** (Type: numeric) – This binary variable is derived from a set of five fields that describe the victim's injuries resulting from the incident. This particular binary variable identifies whether or not the victim had laceration injuries.
34) **INJURY.TYPE.UNCONSCIOUSNESS** (Type: numeric) – This binary variable is derived from a set of five fields that describe the victim's injuries resulting from the incident. This particular binary variable identifies whether or not the victim experienced unconsciousness during the incident.
35) **OFFENDER.SUSPECTED.USING.ALCOHOL** (Type: numeric) – This binary variable is derived from a set of three fields that show if any offender was suspected of using drugs or consuming alcohol during or shortly before this offense, or if computer equipment was used to perpetrate this offense. This particular binary variable represents the use of alcohol.
36) **OFFENDER.SUSPECTED.USING.COMPUTER.EQ** (Type: numeric) – This binary variable is derived from a set of three fields that show if any offender was suspected of using drugs or consuming alcohol during or shortly before this offense, or if computer equipment was used to perpetrate this offense. This particular binary variable represents the use of computer equipment.
37) **OFFENDER.SUSPECTED.USING.DRUGS** (Type: numeric) – This binary variable is derived from a set of three fields that show if any offender was suspected of using drugs or consuming alcohol during or shortly before this offense, or if computer equipment was used to perpetrate this offense. This particular binary variable represents the use of drugs.
38) **WEAPON.FORCE.GUN** (Type: numeric) – This binary variable is derived from a set of three fields that show the type(s) of weapon/force involved with an offense. This particular binary variable refers to the use of a gun.

39) **WEAPON.FORCE.CUTTING** (Type: numeric) – This binary variable is derived from a set of three fields that show the type(s) of weapon/force involved with an offense. This particular binary variable refers to the use of a cutting instrument.
40) **WEAPON.FORCE.BLUNT** (Type: numeric) – This binary variable is derived from a set of three fields that show the type(s) of weapon/force involved with an offense. This particular binary variable refers to the use of a blunt instrument.
41) **WEAPON.FORCE.VEHICLE** (Type: numeric) – This binary variable is derived from a set of three fields that show the type(s) of weapon/force involved with an offense. This particular binary variable refers to the use of a vehicle.
42) **WEAPON.FORCE.HANDS.FEET.ETC** (Type: numeric) – This binary variable is derived from a set of three fields that show the type(s) of weapon/force involved with an offense. This particular binary variable refers to the use of body parts as weapons.
43) **WEAPON.FORCE.POISON** (Type: numeric) – This binary variable is derived from a set of three fields that show the type(s) of weapon/force involved with an offense. This particular binary variable refers to the use of poison as a weapon.
44) **WEAPON.FORCE.FIRE.EXP** (Type: numeric) – This binary variable is derived from a set of three fields that show the type(s) of weapon/force involved with an offense. This particular binary variable refers to the use of fire or explosives as weapons.
45) **WEAPON.FORCE.DRUGS** (Type: numeric) – This binary variable is derived from a set of three fields that show the type(s) of weapon/force involved with an offense. This particular binary variable refers to the use of drugs as a weapon.
46) **WEAPON.FORCE.ASPHYXIATION** (Type: numeric) – This binary variable is derived from a set of three fields that show the type(s) of weapon/force involved with an offense. This particular binary variable refers to the use of asphyxiation during the offense.
47) **WEAPON.FORCE.OTHER** (Type: numeric) – This binary variable is derived from a set of three fields that show the type(s) of weapon/force involved with an offense. This particular binary variable refers to the use of other weapons or force.
48) **WEAPON.FORCE.NONE** (Type: numeric) – This binary variable is derived from a set of three fields that show the type(s) of weapon/force involved with an offense. This particular binary variable refers to the use of no weapons or force during the offense.
49) **HATE.CRIME** (Type: numeric) – *This indicates any bias motivation in the commission of this offense.* There are various codes for the different possible biases; however, these were reduced to a binary variable representing the existence of a hate crime or not.
50) **V1007.INCIDENT.DATE.HOUR** (Type: string) – *This is the military time hour that the crime occurred. If the crime occurred at exactly midnight, the Incident Date will reflect as if the time was one minute past midnight. In other words, midnight is considered the beginning of the day.*
51) **V1008.TOTAL.OFFENSE.SEGMENTS** (Type: numeric) – *This is the number of offense segments (up to 10) in the Group A Incident Report.*
52) **V1009.TOTAL.VICTIM.SEGMENTS** (Type: numeric) – *This is the number of Victim Segments (up to 999) in the Group A Incident Report.*
53) **V1010.TOTAL.OFFENDER.SEGMENTS** (Type: numeric) – *This is the number of Offender Segments (up to 99) in the Group A Incident Report.*
54) **BH010.COUNTRY.DIVISION** (Type: string) – *Geographic division in which the state is located.*

55) **BH011.COUNTRY.REGION** (Type: string) – *Geographic Region in which the Country Division is located.*
56) **COLLEGE.UNIVERSITY** (Type: numeric) – Distinguishes incidents that occurred at colleges and universities.
57) **BH013.CORE.CITY** (Type: numeric) – *This designates if the agency is the central city of a Metropolitan Statistical Area (MSA).*
58) **AGENCY.POPULATION.TOTAL** (Type: numeric) – This variable is the sum of the (up to) three agency population coverage fields (BH019, BH023, BH027), thus providing the total population covered by the agency.

2.3 NIBRS Data Context:

The Uniform Crime Reporting (UCR) Program is a nationwide, cooperative, statistical effort of more than 17,000 city, county, and state law enforcement agencies voluntarily reporting data on crimes brought to their attention (ASUCRP 2019). The UCR consist of four data collections:

- The National-Based Reporting System (NIBRS)
- The Summary Reporting System (SRS)
- The Law Enforcement Officers Killed and Assaulted (LEOKA)
- The Hate Crime Statistics Program

NIBRS appears to be given top priority by the FBI.

This project focuses on the NIBRS data collection because of the level of detail it offers for each particular offense within distinct crime incidents. An additional benefit of NIBRS is that it largely contains the information from the other three UCR data collections, except at a more granular level. NIBRS data is collected on a single crime occurrence; that is, data is collected on each single incident and arrest with up to 10 offenses per incident. Each offense is reported to the law enforcement agency with various facts about the crime. The data collected includes injuries received, victim, offender and arrestee age, sex, race, and ethnicity, etc. Specialized crime data can be generated from this data such as crime against children, crimes against women, domestic violence, and crimes against elderly.

2.4 NIBRS Data Conditioning

The fields created for the project from the NIBRS dataset are taken from five of the NIBRS source files. The five files are listed and described as follows:

1) **Batch Header** – *(4 MB, 60 cols, 22,591 rows in 2016)* – This segment provides descriptive data related to the individual police agencies, which are coded by an Originating Agency Identifier (ORI) (U.S. FBI 2016). Each ORI only appears in the batch header segment once. Besides the ORI characteristics, a key component of this file is the Federal Information Processing Standard (FIPS) county codes associated with the reporting agencies. This provides a linkage to other county level data sources.
2) **Administrative** – *(319 MB, 16 cols, 5,364,179 rows in 2016)* – This segment provides a series of information that is only applicable at the incident level (U.S. FBI 2016). Each incident therefore only has one record in the administrative segment.

3) **Offense** – *(485 GB, 25 cols, 6,063,523 rows in 2016)* – Up to ten offenses can be associated with an incident in NIBRS (U.S. FBI 2016). Each offense is listed as a separate record in this segment. Each offense is intended to represent a distinct crime.
4) **Victim** – *(819 GB, 54 cols, 6,034,735 rows in 2016)* – This segment describes the victims involved in each incident. Up to 999 victims can be associated with an incident in NIBRS (U.S. FBI 2016). A separate set of information is captured for each victim involved in the incident.
5) **Offender** – *(301 MB, 11 cols, 6,105,830 rows in 2016)* – This segment describes the offenders involved in each incident regardless of whether or not they have been identified or an arrest was made. Up to 99 offenders can be associated with an incident in NIBRS (U.S. FBI 2016). It is intended to capture any offender information known to law enforcement concerning the offenders. Naturally, there are instances where there is no information about the offender(s) involved.

The construction of the main project dataset was accomplished through R code that leveraged the relational structure of the NIBRS flat files. The R code takes in the raw data and does all required manipulations to produce the base modeling data set used for this project. This approach ensures any perceived data errors are completely traceable back to the raw data.

The data set construction process started with the Victim files for each NIBRS reporting year. These were reduced down to victims under the age of 18, since those are the only incidents within the scope of this project. The victim files have 10 columns that can be populated with up to 10 different offenses that took place during the incident. This project was an offense-centered analysis, so these 10 offense columns were gathered into rows of victim-offense associations for each Victim data frame. The reshaped data frames were then filtered down to only the offenses of interest in this project. The Victim files also have up to 10 columns available for up to 10 involved offenders. These offender columns were also gathered into rows with the victim-offense associations to further reshape the Victim data frames so they each have one row for each possible victim-offense-offender relationship. Figure 1 provides an example of how this process unfolded for one incident from the NIBRS data source. Columns of data recording the total number of victims, offenses, and offenders involved in the overarching incident preserve some of the inherent relational information for each row of the resulting data set.

With the Victim data frames organized by victim-offense-offender associations, it was then possible to link fields from the other files to this restructured Victim data frame. This was accomplished with a series of left joins, starting with the Offense file, followed by the Offender file, then the Batch Head file, and finally the Administrative file. After executing this for each year, the resulting rows from each master file year were bound together into one data frame.

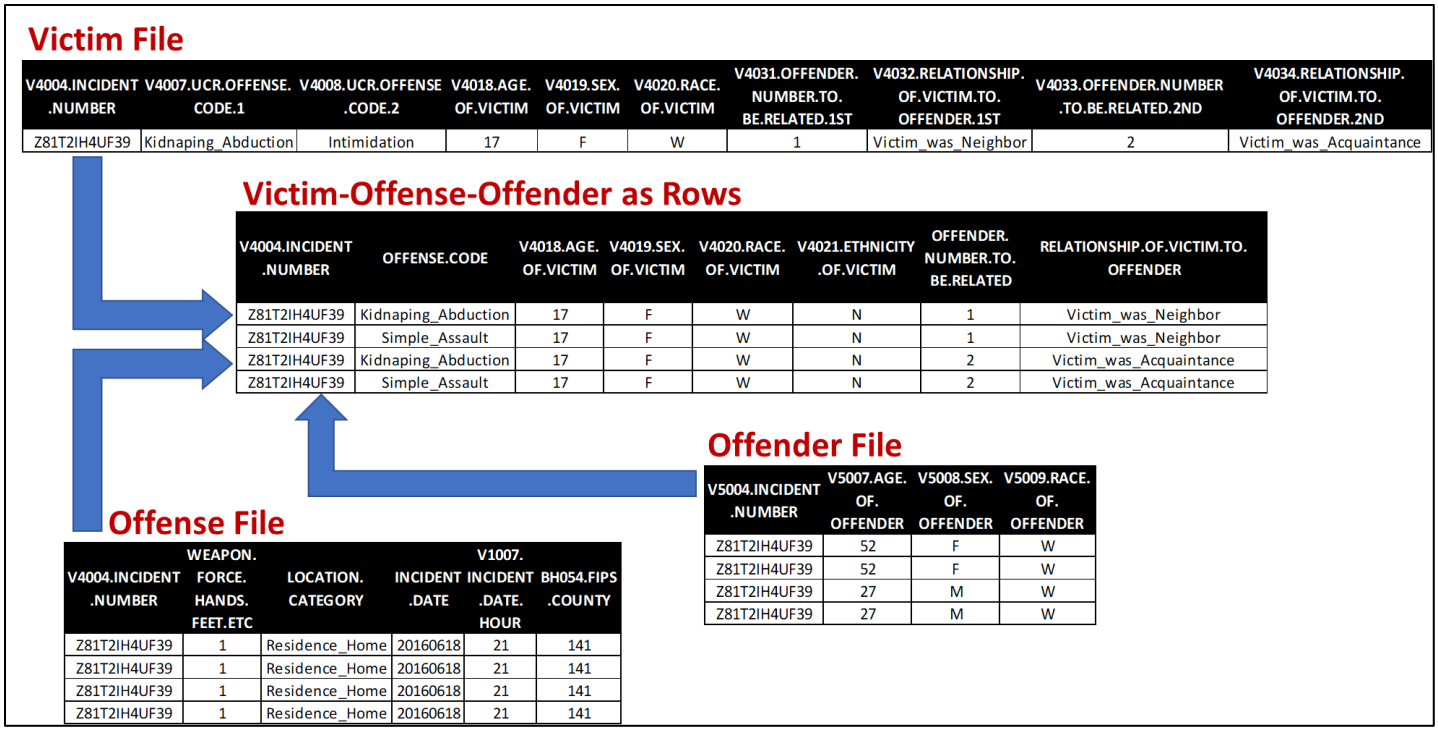


Figure 1: Example of Data Linkages Required for the Project Data Set Construction

Some columns in the dataset at this point still represented multiple possible entries for the same type of information. For example, there were three columns for up to three different entries for the weapons/force involved. This format was not particularly useful for analysis purposes. To fix this issue for applicable fields, a series of binary variables were created for each possible relevant descriptor. For example, variables were created for items like gun, cutting instrument, and asphyxiation. The three weapons/force columns were checked at each row and if one of them matched the gun category, for example, the gun binary variable would get flagged with a 1, otherwise a 0. This approach was used to address categorical data captured across multiple columns for weapons/force used, injury types, circumstances, hate crimes, and offender suspected of using substances/equipment.

Six target variables were created from the data set for the purposes of analysis. They represent the six offense categories that were determined by the project team by consolidating the 14 pertinent NIBRS offense codes into six more manageable groups. Each of the six binary variables distinguish the target offense from all of the others. A summary of the target variables is shown in Table 1. As discussed in later sections, the Negligent Manslaughter and Human Trafficking target variables were ultimately excluded from the analysis due to limited data availability.

NIBRS Code	NIBRS Offense Name	KD-CAC Offense Category
13A	Aggravated Assault	Assault
13B	Simple Assault	
13C	Intimidation	
11A	Rape	Sex Offense
11B	Sodomy	
11C	Sexual Assault With An Object	
11D	Fondling, Indecent Liberties, Child Molesting	
36A	Incest	
36B	Statutory Rape	
100	Kidnaping Abduction	Kidnaping/Abduction
09A	Murder Nonnegligent Manslaughter	Homicide
09B	Negligent Manslaughter	Negligent Manslaughter
64B	Human Trafficking Involuntary Servitude	Human Trafficking
64A	Human Trafficking Commercial Sex Acts	

Figure 2: Offense Categorization for Determining Target Variables

One remaining conditioning step involved the consolidation of some categorical variables into a smaller set of categories. This was particularly important for the offense location type variable, which had 48 possible categories. Besides the fact that this unwieldly number of categories contained some with very small counts, some algorithms under consideration for this project only allowed up to 32 categories. A team consolidation effort reduced the 48 location type categories down to 18.

Some basic data cleaning operations were required to prepare the main project dataset for analysis. For most categorical variable, it was logical to replace blanks or NA values with a "UNK" representing an "Unknown" category within the field. Binary variables without an "Unknown" category were converted to a 0/1 representation. Victim age variables needed to be converted from a categorical (most often it was an integer) to a continuous variable in order to represent the various possible ages of infants (less than one year of age).

Another critical data cleaning step involved the identification and removal of duplicate incidents. For unknown reasons, some master year files contained repeats of incidents from a previous year. If such incidents were identified as already occurring in a previous master year file, they were removed from the newer master year file. Overall there were a little over 7,000 rows of data removed for the purposes of this correction.

After all conditioning, the base project dataset was 372 MB with 1,137,211 rows of victim-offense-offender information and 81 columns (58 of which were initially considered predictors). Exploratory data analysis techniques were applied to this data set to assess quality and gain understanding prior to modeling. This assessment is described in the next section.

2.5 NIBRS Data Quality Assessment:

- **Completeness:** The highest percentages of missing data were found to be for offender age (9.3%), county code (6.3%), and agency population coverage (3.1%) fields. For models that cannot handle missing data, such as logistic regression, the `MICE` package was used in `R` with its default settings to fill in the missing information. Overall, the amount of missing information in the dataset is not believed to have been an overly influential issue during model development. When data was due to unavailability (i.e., unknown) it was not considered to represent a lack of completeness.
- **Uniqueness:** For unknown reasons the raw data across master file years contained some duplicated incidents. Roughly 7,000 rows of data were identified as duplicated records and removed from the dataset. The dataset otherwise appeared to satisfy the uniqueness criterion.
- **Accuracy:** The NIBRS set appears to be accurate as reported with the exception of the victim-offender relationship field. Each category in this field is defined in terms of "victim was…" There are instances such as "victim was grandparent" or "victim was step-parent" which do not make sense because all victims are under the age of 18 in this reduced data set. It is believed that some people entering the data may have misinterpreted the intent of this field and were entering it in terms of "offender was…" instead. This appears to have been an issue with less than 5% of the data in this field.
- **Atomicity:** The atomicity criteria does not apply to the datasets in this project.
- **Conformity:** After conducting exploratory data analysis, there does not appear to be any issues with data conformity.
- **Overall Quality:** Overall, there were very few concerns with the quality of the NIBRS data. The NIBRS dataset does, however, have some limitations. It does not provide a representative sample of the crime in the United States. The data only has information on reported incidents and address issues where some offenses are reported more than others. For example, human trafficking and sex offenses are believed to more likely to go unreported. Finally, given the wide range of possible context surrounding crimes, it is difficult to consistently capture all of the relevant details of such instances in a database.

2.6 Socio-Economic Data Context and Conditioning

The project team integrated several socio-economic datasets with the NIBRS data set in order to execute the geospatial component of this study. County population data of those under the age of 18 from 2013 to 2016 were collected from the National Institute of Health, National Cancer Institute (NIH, NCI 2019). Additionally, poverty level averages and unemployment rate data for 2013 to 2017 were collected from the U.S. Census Bureau and consolidated into a format suitable for integration with the NIBRS base project dataset (U.S. Census Bureau 2019). Whenever possible, crime rates were calculated by location using the population information that was compiled.

2.7 Socio-Economic Data Quality Assessment

County poverty rates were used to assess the relationship between crime rates and the economic status of the county. A drawback in this procedure was that the county names in the NIBRS data were sometimes unknown, as described in Section 2.5. Additionally, when compared to a comprehensive list of ORIs published by the United States Bureau of Justice Statistics, only approximately 40 percent of the agencies were found to be represented in the project dataset (USBJS 2012).

3 Analytics and Algorithms

3.1 Exploratory Data Analysis

Basic exploratory data analysis was conducted to assess the quality of the data and better understand it prior to developing models. Figure 3 shows the distribution of the four target variables selected for analysis in this project. Table 1 shows the counts of all six target variables initially considered at the beginning of this study. There are clearly severe imbalances across the classes. With Negligent Manslaughter and Human Trafficking each representing less than 0.004% of the rows in the base project dataset it was eventually decided to exclude these categories as target variables. To avoid the complications associated with attempting multinomial classification with severe class imbalances, the project team ultimately created separate models with binary target variables for each of the four remaining offense categories.

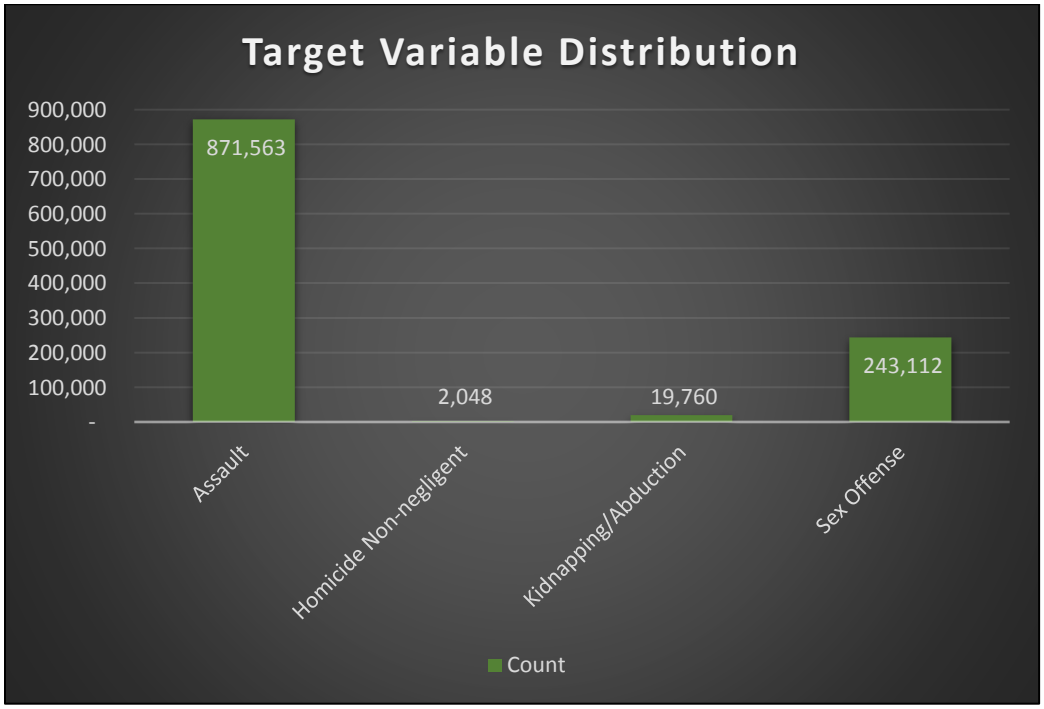


Figure 3: Target Variable Distribution (Bar chart)

Table 1: Target Variable Distribution (Table of Counts)

Target Variable	
Assault	871,536
Homicide Nonnegligent	2,048
Human Trafficking	398
Kidnaping and Abduction	19,760
Manslaughter	357
Sex Offense	243,112

The project team took a close look at the correlations among the variables in the NIBRS dataset. These are shown in Figures 4, 5, and 6 for display purposes. Examination of these correlations did not reveal any major concerns with regard to the modeling to be conducting in later stages of the project. Model-specific mitigation strategies were employed to address some of the minor correlations found among the predictors.

The correlation between continuous variables shown in Figure 4 was calculated using Pearson correlation, which indicates how two variables are linearly related. The calculation of correlations between continuous and categorical variables, as shown in Figure 5, is more complex and was generated using the *Eta* Correlation Ratio. The *Eta* Correlation Ratio is related to the root *R*-squared value of the one-way ANOVA procedures (MATLAB Correlation, 2013).

With regard to the categorical variable correlations shown in Figure 6, Cramer's *V* was used as the method of comparison. Cramer's *V* is more typically used as a post-test to determine strengths of association after *chi*-square has determined significance; however, it is often also used to represent the correlation of categorical variables (Zychlinski 2018). Cramer's *V* ranges between 0 and 1. When the value is close to 0 it shows little association between variables. When it is close to 1, it indicates a strong association.

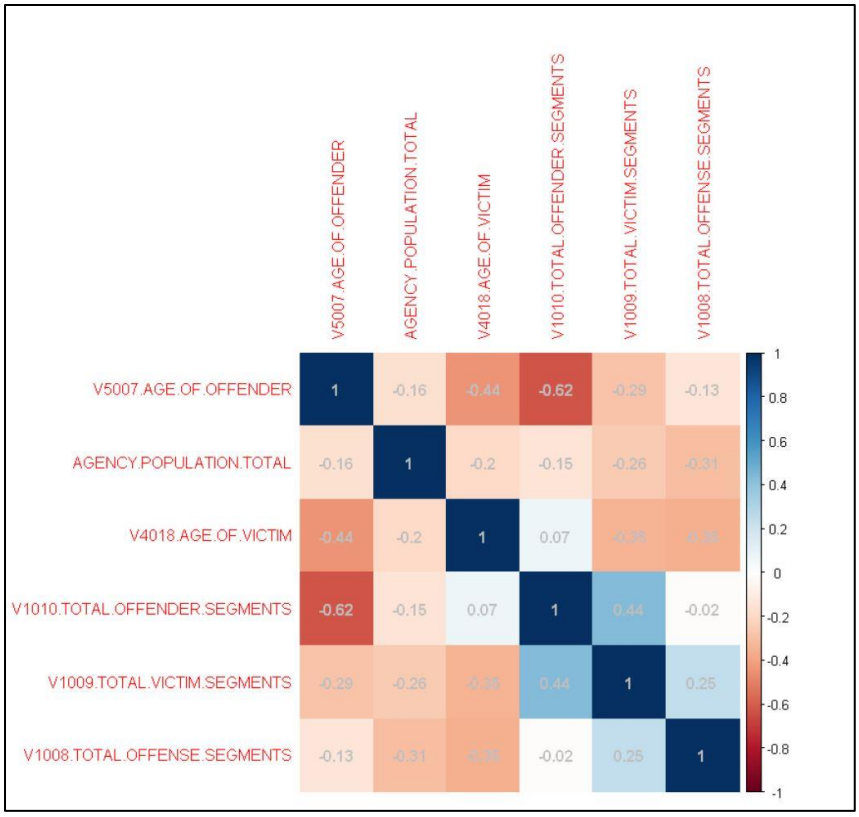


Figure 4: Continuous Variable Correlations

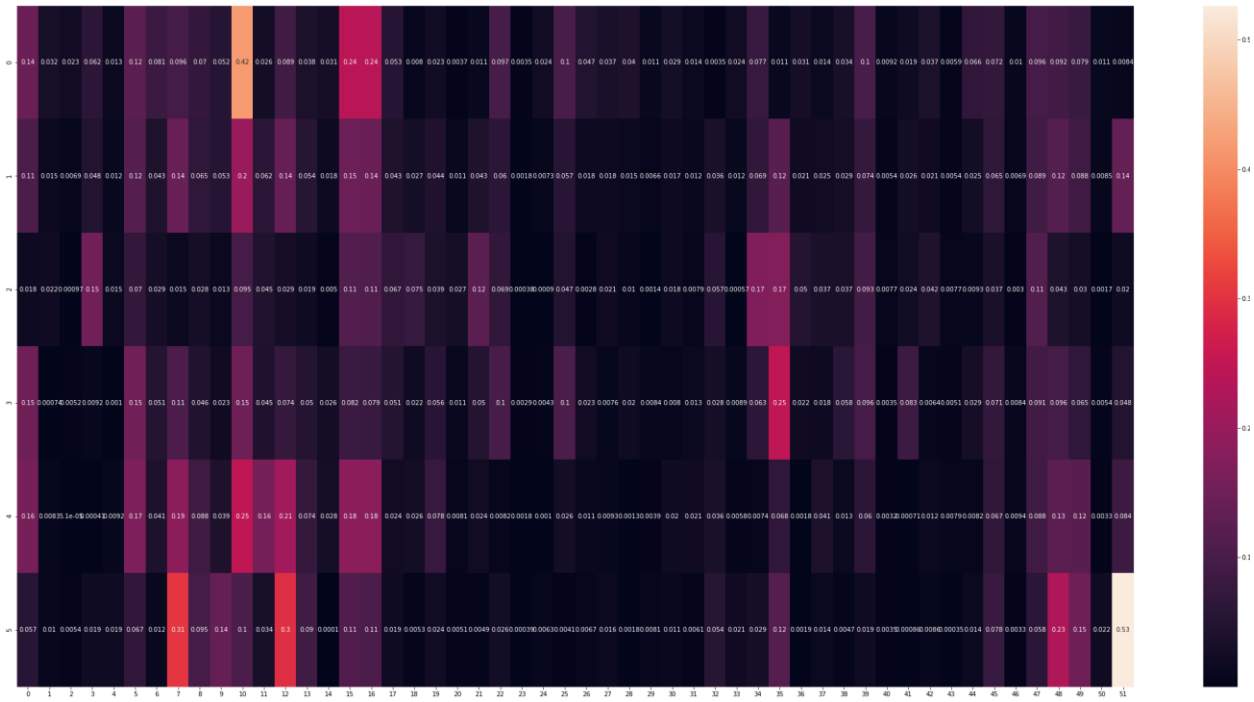


Figure 5: Continuous vs. Categorical Variable Correlations

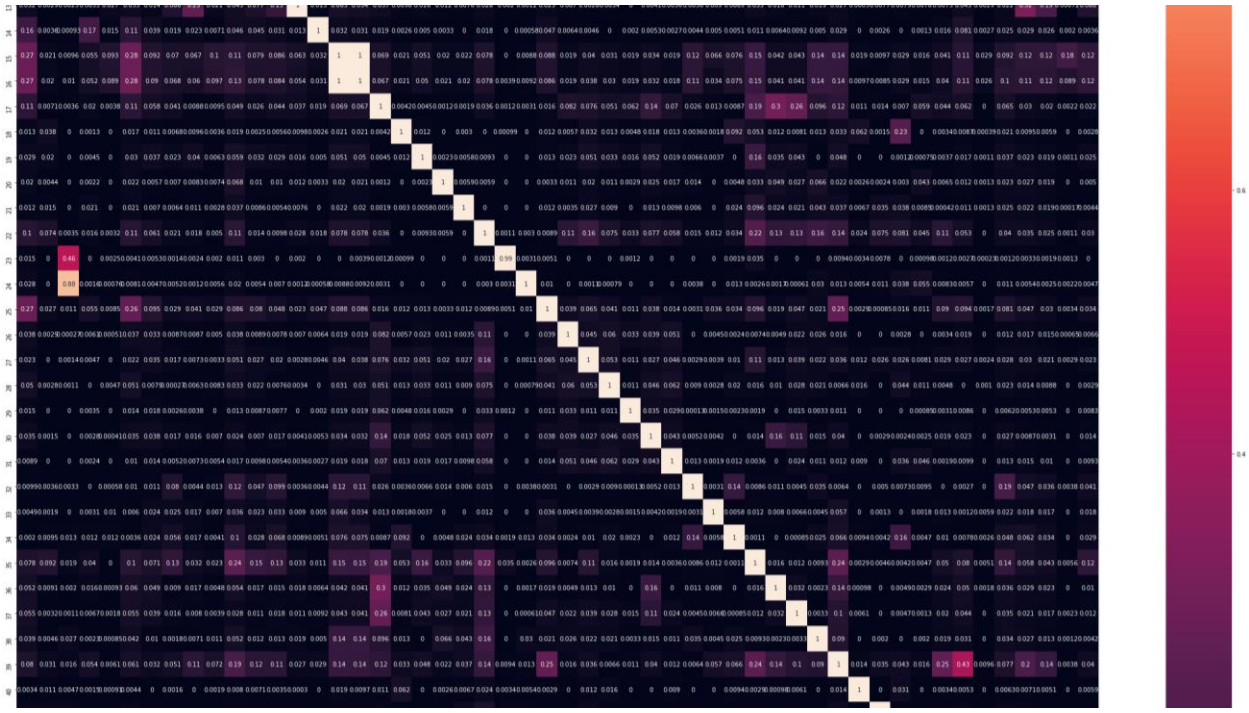


Figure 6: Categorical Variable Correlations

3.2 Analytics Model Development Process

The analytics and algorithm portion of this project started with an exploratory modeling phase where a wide variety of models were attempted with selected subsets of the data to see which were most promising for full implementation. Unsupervised learning techniques explored included K-Modes clustering, hierarchical clustering, and association rule learning. Classification techniques attempted in this project included penalized logistic regression, basic decision trees, random forest, gradient boosting with XGBoost, Bayesian networks, and neural networks.

The IBM Watson SPSS modeler was used for automated machine learning model exploration with a focus on Bayesian networks and neural networks. Python was used for developing the XGBoost models. R was used for all other modeling tasks.

As shown in Figure 7, this assortment of models was evaluated and down-selected to hierarchical clustering, penalized logistic regression, gradient boosting, and Bayesian networks for in-depth model development. Using these methods, models were developed in-depth and compared across the four selected target variables.

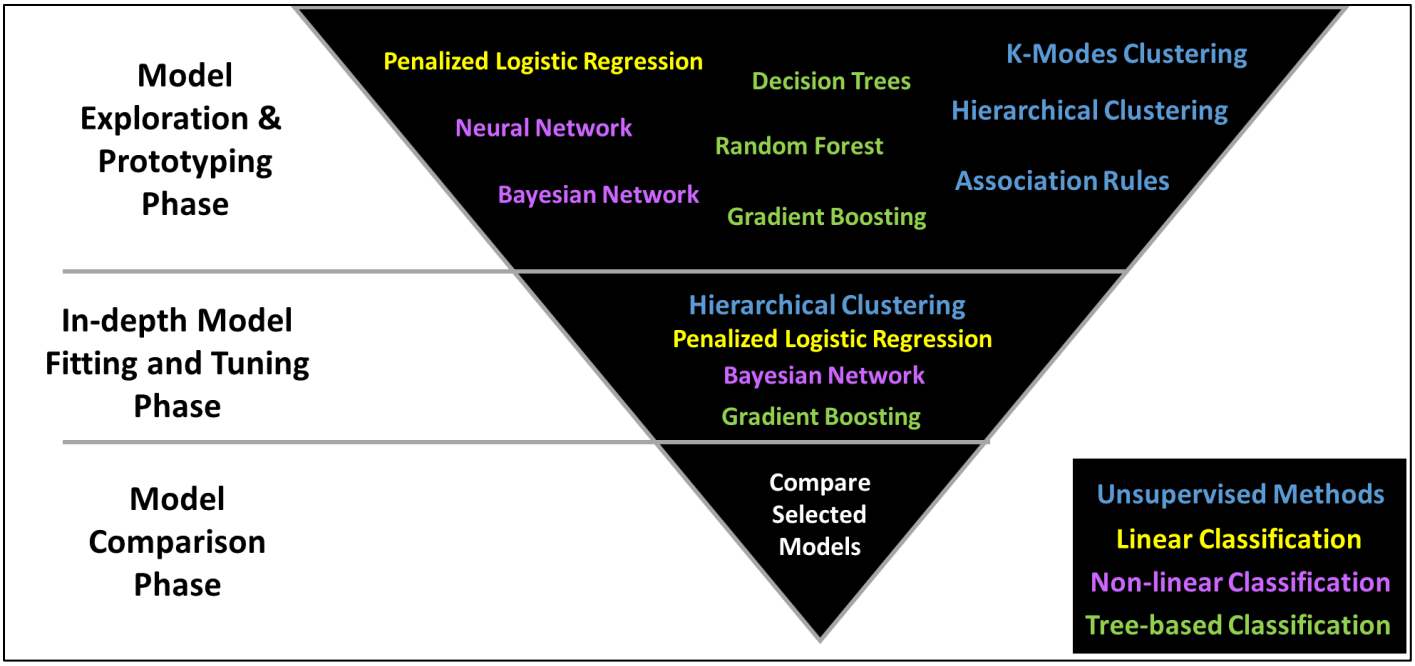


Figure 7: Predictive Model Development Process for the KD-CAC Project

3.3 Unsupervised Methods

3.3.1 K-Modes Clustering

K-modes clustering was the first unsupervised method attempted, but quickly abandoned. Given a predetermined number of clusters, the K-modes clustering algorithm attempts to assign each observation to one of the developing clusters based on how similar it is to the feature modes of the observations currently in each of the clusters (Kar 2017). The initial thought was that the resulting modes of these clusters would reveal the most important features across subgroups of the population of data. Predefining too few clusters yielded subgroups sharing too many common characteristics and ignoring infrequently occurring crime characteristics. Too many clusters yielded more diverse groups, some still sharing common characteristics, but it was too difficult to discern which clusters were most relevant. As a result, other methods were pursued.

3.3.2 Association Rule Learning

Association rule learning is useful for discovering interesting relationships that may reside in large data sets. An association rule is expressed in the form $X \rightarrow Y$, where X and Y are collections of dichotomous features called item sets (Pang-Ning, Steinbach, and Kumar 2015). X is called the antecedent and Y is referred to as the consequent. The consequent is typically one item that is implied by the collection of items in the antecedent. The most common example of an item set is $\{Milk, Diaper\} \rightarrow \{Beer\}$. This famous example describes how supermarkets discovered that fathers often went to purchase milk and diapers for a baby in the house, but also purchased beer to sooth themselves. Supermarkets and other businesses can mine the transactions in their databases to identify interesting relationships that can assist them with their marketing strategies.

Given the project dataset was predominantly made up of categorical variables and the project goal of identifying interesting relationships, the team felt association rule learning was an unsupervised method worth pursuing. The base project data set was modified into a transactional data set where the characteristics present in each row of data where consolidated into one field with each present item separated by commas in a string. In this state, the data could be used with the functions available in the `arules` and `arulesviz` R packages.

The `arules` package implements the *apriori* algorithm, which conducts frequent itemset mining and association rule learning on transactional data sets. It has settings to limit the rules identified to those having a minimum amount of support, confidence, lift, or other interest measure types. Support is the fraction of transactions that contain both X and Y, while confidence measures how often items in Y appear in transactions that contain X (Pang-Ning, Steinbach, and Kumar 2015). Lift is the confidence of $X \rightarrow Y$ divided by the support of Y. Lift provides a measure that can potentially highlight interesting rules that involve fairly rare items that are otherwise ignored due to the low support of the rule as a whole (Pang-Ning, Steinbach, and Kumar 2015).

After dedicating a good deal of project resources to the association rule approach, the team felt the results were not very useful. Even after pruning rules down to those that were statistically significant, removing rules that were redundant, and sorting rules by lift and other more complex interest measures, a long list of measure remained. Similar to the experience with K-modes clustering, the resulting rule lists had too many common characteristics and did not sufficiently represent infrequently occurring crime characteristics. Still, association rule lists were generated for each of the four offense categories under consideration in this project. The `arulesviz` package makes it possible to graph association rule sets to more easily understand the most interesting relationships. An example of such a graph will be show in Section 4 of this report.

3.3.3 Hierarchical Clustering

Agglomerative (bottom-up) hierarchical clustering was found to be the most useful approach for gaining an understanding of what characteristics are most associated with each other within each of the four offense categories. This approach provides an intuitive tree-based visualization called a dendrogram and has the added benefit of not needing to predefine a number of clusters. Most commonly, this technique is used to cluster observations on the basis of features, as typically done in K-means or K-modes clustering. It is, however, quite acceptable to cluster features on the basis of the observations to gain an understanding of which features share similarities across the observations (James, Witten, Hastie, & Tibshirani 2013). Obtaining this type of information was a key goal of this project.

Clustering features on the basis of the observations is accomplished by transposing the matrix of data, so that features are represented in rows and observations are represented in columns. Agglomerative hierarchical clustering is accomplished by starting at the bottom of a tree where each feature (in our case) is considered an individual cluster. At each step in the algorithm, a pair of clusters is merged until only one cluster is left (James, Witten, Hastie, & Tibshirani 2013). In order to execute this procedure, there is a need for a measure of distance between the features (i.e. their similarity) and a method for deciding which clusters to merge.

The Jaccard similarity measure was chosen as the distance method for use in the agglomerative hierarchical clustering algorithm. It provides a measure of similarity between binary vectors. The base data set had already been converted to a binary data set for the purposes of association rule learning. The age of victim, age of offender, and hour-of-day features needed to be discretized into groups in order to construct this purely binary data set. In this state, calculating the Jaccard similarity coefficients was relatively straightforward with the following quantities for f_{pq} (Pang-Ning, Steinbach, and Kumar 2015):

$$
\begin{aligned}
f_{01} &= number\ of\ attributes\ where\ p\ was\ 0\ and\ q\ was\ 1 \\
f_{10} &= number\ of\ attributes\ where\ p\ was\ 1\ and\ q\ was\ 0 \\
f_{00} &= number\ of\ attributes\ where\ p\ was\ 0\ and\ q\ was\ 0 \\
f_{11} &= number\ of\ attributes\ where\ p\ was\ 1\ and\ q\ was\ 1
\end{aligned}
$$

The Jaccard similarity coefficient is then calculated as follows:

$$Jaccard\ Similarity\ Coeffeicient = \frac{Number\ of\ 11\ matches}{Number\ of\ nonzero\ attributes} = \frac{f_{11}}{(f_{01} + f_{10} + f_{11})}$$

An alternative to the Jaccard similarity coefficient could have been the simple matching coefficient method, which is calculated as follows:

$$Simple\ Matching\ Coefficient = \frac{Number\ of\ matches}{Number\ of\ attributes} = \frac{f_{11} + f_{00}}{(f_{01} + f_{10} + f_{11} + f_{00})}$$

The Jaccard similarity coefficient was believed to be superior to the simple matching coefficient for the purposes of this project due to the nature of the dataset. The project dataset is sparse in its binary state. Because of this, many features would have been found to be similar on the basis of many characteristics being not present in the offenses and would therefore drown out the similar characteristics that were present in offenses.

With a matrix of distance measures calculated, the next step was to choose a clustering selection method. It may seem like the best approach would be to combine the most similar clusters at each step in the tree, but this often does not yield the best results. Ward's clustering method is one of the more popular cluster selection approaches currently available and was found to be the most effective for the data in this project. Ward's method minimizes the total within-cluster variance. At each step, the algorithm finds the pair of clusters that leads to the minimum increase in the total within-cluster variance after making the merger. Ward's method is known to be less susceptible to noise and outliers, while also producing dense globular cluster, which was desirable for the purposes of this project (James, Witten, Hastie, & Tibshirani 2013).

The agglomerative hierarchical clustering approach proved to be useful and informative. It was the unsupervised method selected for in-depth development during the analytics project sprint. The

visualizations ultimately informed other aspects of this project, such as where to focus attention while interpreting the Bayesian network model outputs. The resulting visualizations are shown in Section 4.

3.4 Data Balancing for Supervised Methods

As described in Section 3.1, this project involved unbalanced target variables. For such cases, there are three ways to balance the training set in order to avoid misleading model fitting and evaluation issues (Fawcett 2016):

1) **Oversample the minority class:** This approach involved replicating observations from the minority class to balance the data.
2) **Undersample the majority class:** This approach involves reducing the number of observations from majority class to balance the data.
3) **Synthesize new minority classes:** This approach involves creating new minority examples by interpolating between existing ones.

The first two approaches are demonstrated in Figure 8. These two methods have important drawbacks. Oversampling results in more data, but replicating data makes variables appear to have lower variance than they do in reality. Undersampling randomly downsamples the majority class by throwing away data. If the dataset is not sufficiently large, this will cause information loss (Fawcett 2016).

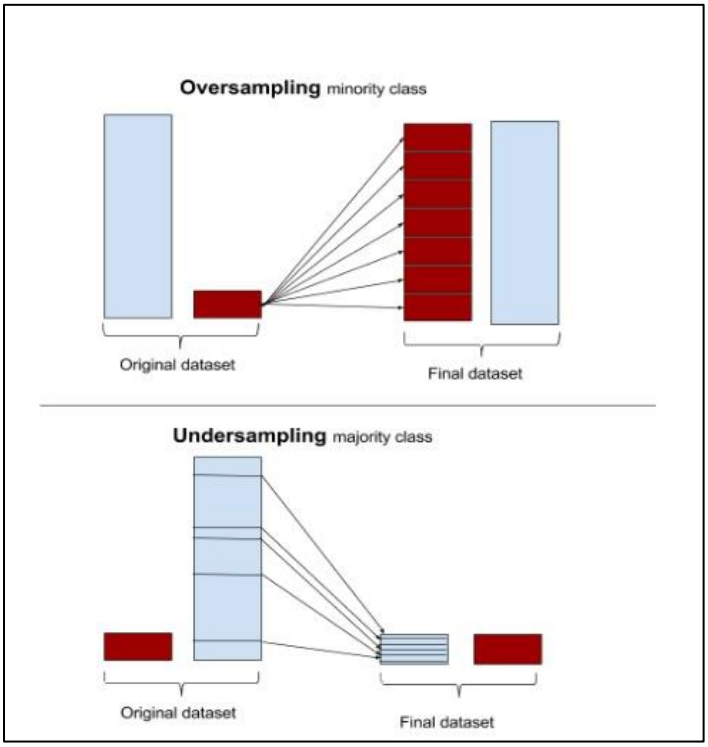


Figure 8: Demonstration of Oversampling and Undersampling Methods for Dealing with Class Imbalance (Fawcett 2016)

The Synthetic Minority Oversampling Technique (SMOTE) is a popular method for synthesizing new minority class data. SMOTE creates artificial data based on feature space similarities from minority samples (Torgo 2011). The imbalance between Sex Offense and other crimes is large, but quite

appropriate for use in the SMOTE method. In the cases of Kidnapping/Abduction and Homicide the imbalanced can be considered severe, but perhaps manageable. SMOTE data sets were produced for these four offense categories and used for model development. For Human Trafficking and Negligent Manslaughter, each of which represented 0.004% of the total data set. The team felt the imbalances were too extreme for the SMOTE procedure to be useful. Still, SMOTE data sets were developed for these two offense categories. When used during prototype modeling, the results did not appear to be trustworthy.

SMOTE is effective with high-dimensional data, so it was believed to be an ideal solution for this particular project (Torgo 2011). The algorithm first it finds the k-nearest neighbors in the minority class for each of the samples in the class. It then calculates a line between the neighbors and generates random points along the lines. These steps are further described and demonstrated in Figures 9 (Fawcett 2016).

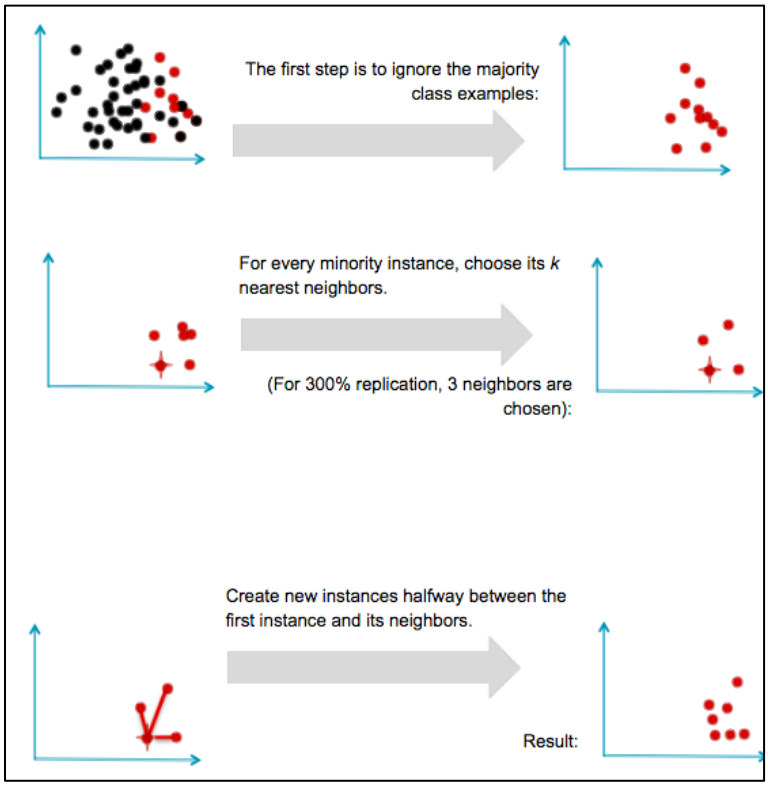


Figure 9: SMOTE Process Description (Fawcett 2016)

SMOTE has three key parameter settings in its R implementation (definitions taken from Torgo 2011):

- **Perc.over:** *A number that drives the decision of how many extra cases from the minority class are generated*
- **Perc.under:** *A number that drives the decision of how many extra cases from the majority classes are selected for each case generated from the minority class*
- **k:** *A number indicating the number of nearest neighbors that are used to generate the new examples of the minority class*

After defining the following three parameters, the new number of observations in the SMOTE generated data set can be determined using the equations shown in Table 2, where:

$$xx = perc.over$$

$$yy = perc.under$$

$$n = original\ number\ of\ observations\ for\ the\ minority\ class$$

Table 2: SMOTE Parameter Settings

Class	New Number of Observations
Minority	$(1 + xx/100) * n$
Majority	$yy/100 * (xx/100) * n$

Through trial and error, it was determined that sensible settings for the SMOTE function for the Kidnapping/Abduction, Assault, and Sex Offense data sets were $k = 10$, $perc.over = 100$, and $perc.under = 200$. To address the exceptionally severe imbalance related to the homicide data, the most sensible SMOTE settings were found to be $k = 10$, $perc.over = 5000$, and $perc.under = 200$. This process provided the team with four balanced data sets (with approximately 50% - 50% distributions), one for each of the four selected target variables. The use of the ARGO cluster was essential for executing the SMOTE procedure due to the extensive computation requirements.

3.5 Supervised Methods

The performance of supervised models will be discussed in terms of their prediction accuracy and Receiver Operating Characteristics (ROC) Area Under the Curve (AUC). AUC measures how true positive rate (recall) and false positive rate trade off. The AUC is typically preferred over accuracy for binary classification, especially when dealing with an unbalanced target variable where algorithms tend to overfit to a single class. This project certainly has class imbalance issues to navigate; however, as discussed in Section 3.4, SMOTE was used to address this issue. Accuracy therefore should still be a relatively reliable performance measure in this project alongside the AUC.

All prototyping was executed using subset of the Sex Offense data set to compare performance consistently and efficiently. Penalized logistic regression, gradient boosting with `XGBoost`, and Bayesian network were ultimately selected for in-depth model development. A brief overview of the prototype models that were developed will be provided first.

3.5.1 Prototyped Models

3.5.1.1 Decision Trees

The Classification and Regression Trees (CART) decision tree modeling approach was implemented on the Sex Offense data set using the `rpart` package in R, but quickly abandoned due to the poor performance results in the prototype models. The initial thought was that the resulting trees

could provide meaningful insight into the relationships among the variables; however, it was also evident that the resulting trees were sensitive to change. Slight changes in parameter settings or the random split of training and test data seemed to result in trees with completely different variables and relationships.

3.5.1.2 Random Forest

Random Forest models work by growing many classification trees. To classify a new object from an input vector, it runs the input vector down each of the trees in the forest. Each tree gives a classification, so in a sense, each tree "votes" for a class. The forest chooses the classification having the most votes over all the trees in the forest. The proportion of votes in each class across the ensemble is the predicted probability vector.

The tuning parameters of the highest importance are: mtry, number of trees, and minimal node size. The mtry value is the number of randomly selected predictors to choose at each split, while the number of trees is how many trees the random forest model needs to create. The minimal node size represents the minimum size of the final node of any tree in the forest. The Random Forest model performed quite well during prototyping, but was ultimately not chosen for in-depth development because it was computationally burdensome compared to the `XGBoost` models, which performed equally well.

3.5.1.3 Neural Network

Neural Network and Bayesian Network prototype models were built using IBM Watson - SPSS Modeler. While the Neural Network model was never developed beyond the prototyping phase, the experience with building such a model in IBM Watson - SPSS Modeler is discussed in this subsection. The following parameters were set for the Sex Offense prototype model:

- **Confidence** – Based on probability of the predicted value
- **Overfit Prevention** – Set to 30%
- **Combining rule for categorical target** – Voting
- **Combining rule for continuous target** – Mean
- **Number of components used for boosting and bagging** – 10

To fit a model, the software used Multilayer Perceptron (MLP) as the model building method, with two neurons in the output layer and SoftMax as the activation function in the output layer. The model formulation is displayed graphically in Figure 11. The resulting model evaluation measures are shown in Table 3.

Table 3: NN Model Evaluation Metrics

Accuracy	0.869
True Positive Rate	0.869
False Positive Rate	0.131
Precision	0.869
Recall	0.869
F_1 Measure	0.869

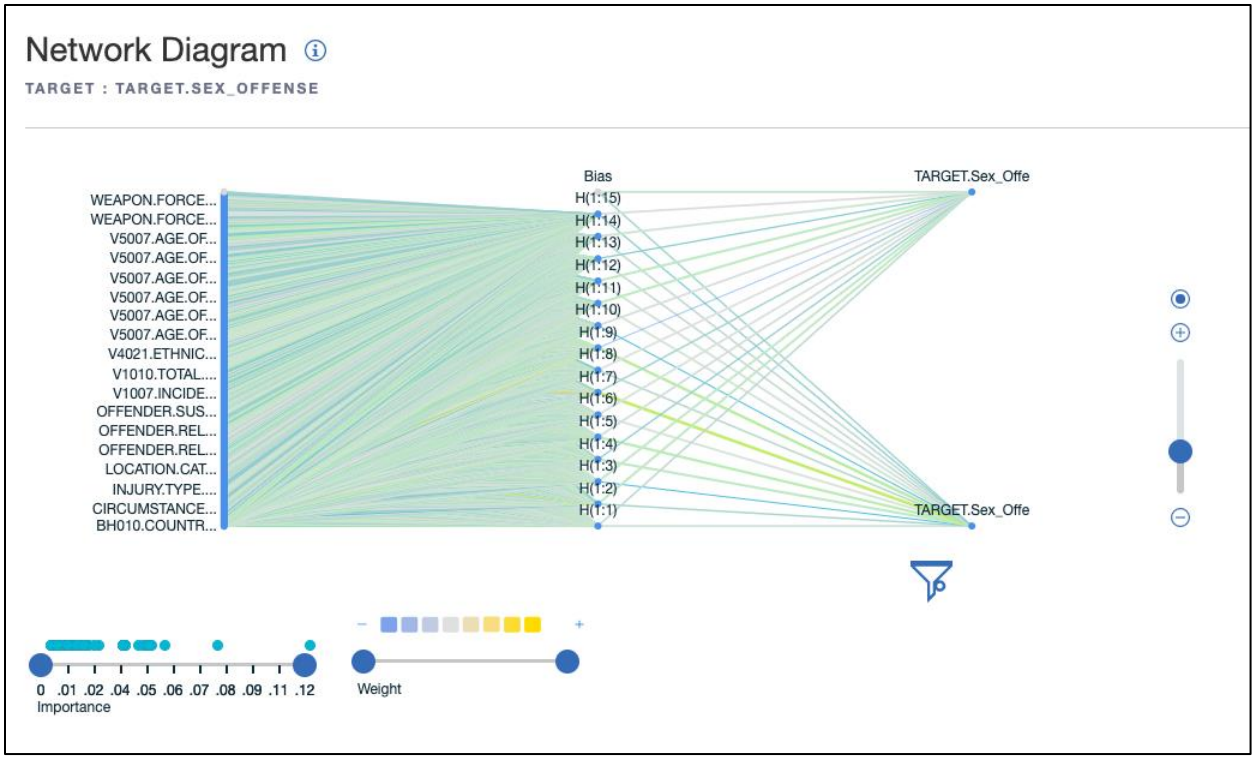


Figure 10: Neural Network Graph

By looking at the network graph in Figure 10, the two neurons (output node) on the right represent the two binary sex offense values and the final output of the neural network, the left neurons (input node) are the raw data variables. The weights represent the relative importance of predictors and their associations with the outcome variable. The relative importance of each predictor can be adjusted in the graph to see how important a predictor variable is, for example, the most important predictor in the neural network above is location category followed by the type of weapon used, others include age of victim, circumstances, and sex of victim. The presence of bias in the middle of the network graph shows that the model is flexible. It also shows the bias of each neuron, which helps in controlling the value. The output is computed by multiplying the input by the weight and passing the result through a SoftMax activation function. The Neural Network approach ultimately lacked in performance and interpretability.

3.5.2 Penalized Logistic Regression

Penalized Logistic Regression was believed to be an ideal baseline model for the purpose of this project because the high interpretability inherent in the resulting coefficients and, as an added benefit, it can be tuned relatively quickly.

A popular method for executing penalized logistic regression is referred to as an Elastic Net, which offers a combination of the advantages provided by Ridge Regression and Lasso Regression. In their pure forms, Ridge Regression aims to balance coefficients that are correlated, and Lasso Regression aims to reduce one or more of the predictor coefficients that are strongly correlated.

Elastic net logistic regression attempts to minimize the loss function below when it's transformed into its logistic regression form. The $\hat{\beta}_j^2$ term represents the penalty derived through Ridge

Regression while the $|\hat{\beta}_j|$ term represents the penalty derived through Lasso Regression. The α term provides a tuning parameter to weight the amount of penalty derived between the $\hat{\beta}_j$ terms. The λ term provides a tuning parameter to adjust the overall combined penalty applied to the coefficients in the model.

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j|\right)$$

Tuning the α and λ parameters for the Elastic Net models was accomplished with a course grid search. α was allowed to take on the values of 0, 0.25, 0.5, 0.75, and 1. λ was sequenced across four values starting at 0.00001 and ending at 0.1. The resulting tuning parameters and performance metrics for each of the four models can be seen in Table 4. The algorithm clearly favored low λ values in all of the models, which was not surprising given there was very little correlation found among the predictors. The α values varied more among the models, where the Assault and Kidnapping/Abduction models had α = 1 making them purely Lasso regression models. Visualization and interpretation of these four Elastic Net models will be described in Section 4.

Table 4: Tuning Parameters and Performance Results for Final Penalized Logistic Regression Models

Model	Alpha	Lambda	Accuracy	AUC	Sensitivity	Specificity
Assault	1.00	0.0001	0.843	0.921	0.849	0.838
Sex Offense	0.50	0.00001	0.864	0.938	0.854	0.875
Homicide	0.75	0.00001	0.917	0.971	0.905	0.928
Kidnap/Abduction	1.00	0.0001	0.859	0.926	0.869	0.849

3.5.3 Gradient Boosting

The goal of Random Forest is to build complex trees that are independent of one another in order to reduce correlation between them. Gradient Boosting models differ in that they build successive trees that learn from the previous one. The models are then trained on the residuals from the previous iteration, rather than the dependent variables. This allows the models to find patterns in the incorrectly classified data and learn from it. Since the models are added sequentially, a new weak base learner is trained at each iteration with respect to the error of the whole ensemble learned so far.

`XGBoost` is an algorithm that implements gradient boosting. It is written in C++, with the focus on being computationally efficient. The Random Forest and XGBoost prototype models had comparable performance; however, `XGBoost` was ultimately selected because of its superior processing speed and the belief that its performance could be greatly improved if attention was given to manipulating its many tuning parameters. The use of the ARGO cluster was essential for building the `XGBoost` models due to the extensive computation requirements for tuning the parameters.

Model tuning for the `XGBoost` models was accomplished through the hyperparameter tuning method, using the `xgb.cv` function in `Python`. This method attempts to maximize the AUC value of

the model under a given set of constraints for selected parameters. After some preliminary experimentation, the project team selected the following for tuning parameters (He, et al. 2019):

- **`max_depth`:** This is the maximum tree depth for base learners
- **`learning_rate "eta"`:** This is the boosting learning rate
- **`n_estimators`:** This is the number of boosted trees to fit
- **`min_child_weight`:** This is the minimum sum of instance weight (hessian) needed in a child
- **subsample:** This is the subsample ratio of the training instance
- **`colsample_bytree`:** This is the subsample ratio of columns when constructing each tree

A comparison of the performance for all four `XGBoost` models is shown in Figure 11. The homicide model had the best performance. The other three models also performed well and were quite comparable to each other. The exceptional performance of these `XGBoost` models provide the strongest support for the second research question of this project; however, some advanced model interpretation techniques were applied to these models in order to assist in addressing the first research question as well. These techniques and the accompanying findings will be demonstrated and discussed in Section 4.

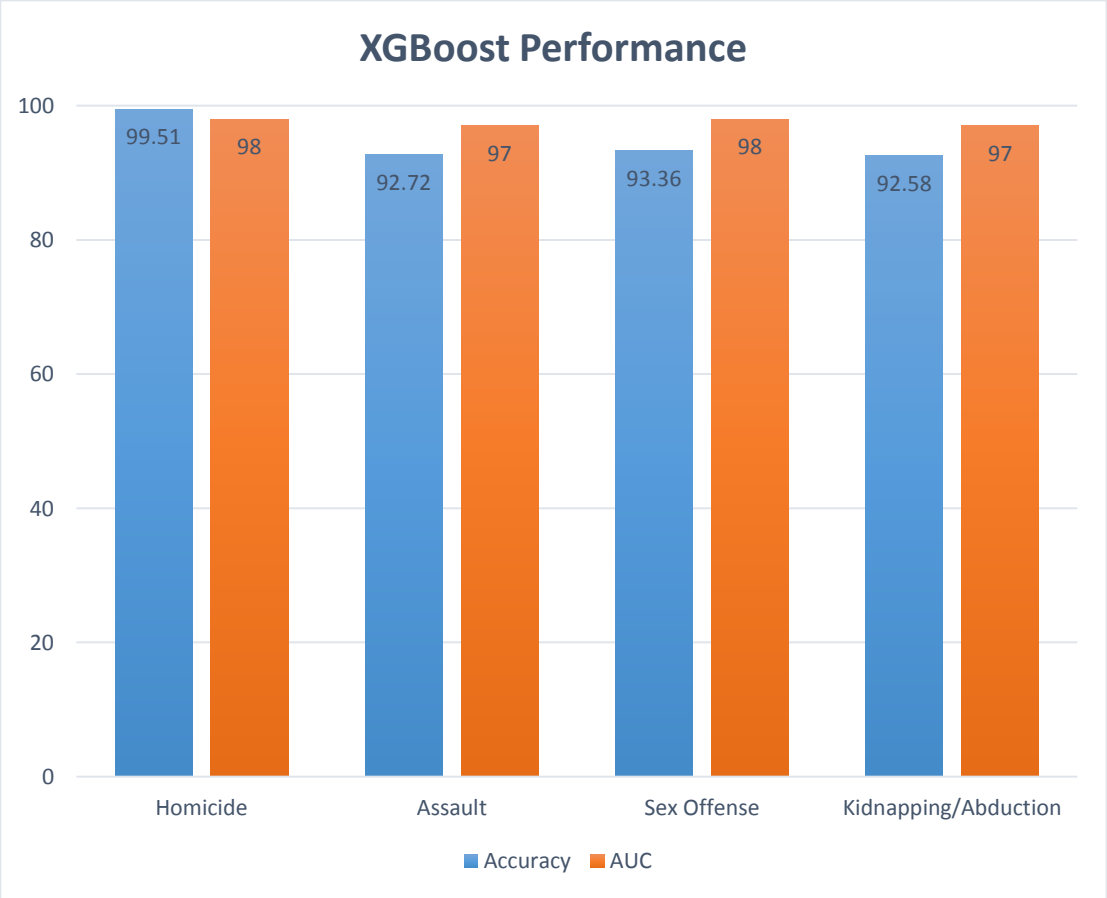


Figure 11: Final XGBoost Model Results

3.5.4 Bayesian Network

A Bayesian network consists of a directed acyclic graph whose nodes represent random variables and links express dependencies between nodes (Park, Chang, & Nam, 2018). Assuming random variables $V_i \in V(1 \le i \le n)$, the Bayesian network is represented by a directed acyclic graph $G = (V, A, P)$ with links between $A \subseteq V \times V$ and *P, a joint probability distribution.* More specifically, *P is a* joint probability distribution over V (Park et al., 2018). It is a directed acyclic graph in which nodes are represented by random variables while the edges indicate the influence of one node on another (Koller, Friedman, Getoor, & Tasker, 2007).

When training the Bayesian network, the goal is to find optimal Bayesian structure through a process of learning the parameters by estimating the parameter set of P that best represents a given dataset with a set of labeled instances (Witten, Frank, & Hall, 2016). The learning process measures and compares the quality of Bayesian networks to evaluate how well the represented distribution explains the given dataset. The maximum log-likelihood was used for measuring the quality of the learning process in the Bayesian networks in this project.

3.5.4.1 Tree Augmented Naïve Bayes

The project team fitted Tree Augmented Naïve Bayes (TAN) models on the dataset to get Conditional Probability Tables (CPT) from the Bayesian Network classifiers. Learning a Bayesian network can be very complex because, for a large dataset, there can be many random variables. Each random variable can take on many values. Also, a single random variable can have many parents and finding the conditional probability of those parents can increase the complexity (Padmanaban, 2014). A general Bayesian network may not have edges from the class node to all the variables, which can sometimes lead to lower classification accuracy. According to (Padmanaban, 2014), for better performance, a Bayesian network that encodes the structure of the Naïve Bayes model and also captures the correlations between the variables is needed. The solution to that is to use a TAN, which reduces the computational complexity, maintains the structure by adding edges between the variables, and captures the correlations between the attributes. In addition, each variable can be connected to another variable in the network (Padmanaban, 2014).

According to (Chow, & Liu, 1968), the key feature of a TAN is the structure. To construct the tree, the correlation between each pair of variables is measured and it adds edges only between the most highly correlated variables. The TAN algorithm is shown below (Padmanaban, 2014).

TAN Algorithm; The TAN algorithm is $O(n^2 \log n)$, where *n* is the number of graph vertices:

1) Compute the mutual information between each pair of attributes.
2) Build a complete undirected graph in which the vertices are the attributes (the *n* variables). The edges are weighted according to their pairwise mutual information.
3) Build a maximum weighted spanning tree.
4) Transform the resulting undirected graph to a directed graph by selecting the class variable as the root node and setting the direction of all edges outward from it.
5) Construct a TAN model by adding an arc from the class variable to all other variables.

As a result, the TAN allows each predictor to depend on another predictor in addition to the target variable. After constructing the trees, the conditional probability of each of the attributes conditioned on its parent and class label is calculated and stored. The entire framework of the Bayesian network is shown in Figure 12. The analysis node provides the performance of the model and the TARGET nodes provide the model output which contains the feature importance for each of the model, the acyclic graph, and the conditional probability table. A total of 57 fields were used at first which resulted in an unreadable acyclic graph, like that which is shown in Figure 13. Graphs with fewer variables that are more readable will be shown in Section 4.

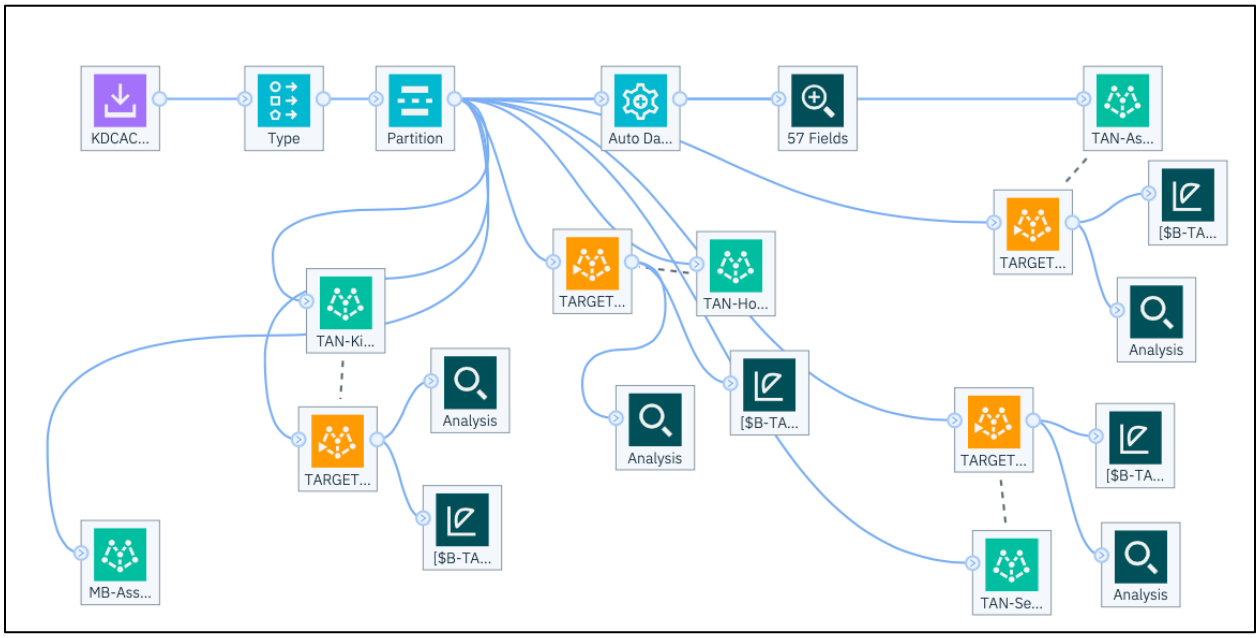


Figure 12: The purpose framework of TAN structure of Bayesian Network

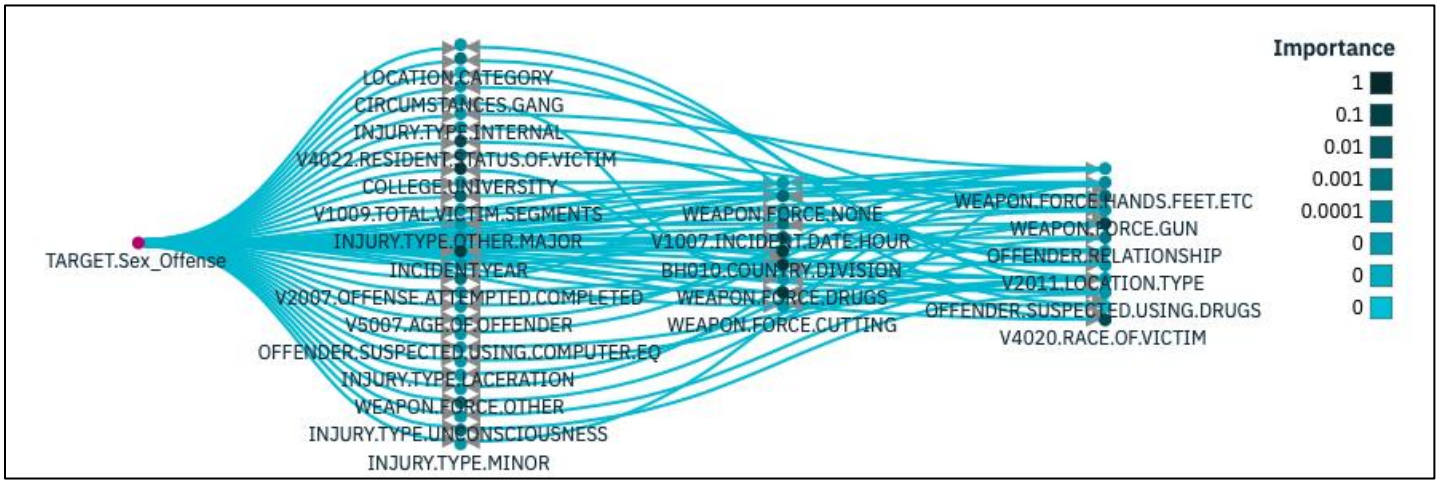


Figure 13: Full Network Graph for Sex Offense

The final comparison of the performance on all four is shown in Figure 14. The performance of the Bayesian network was evaluated based on the AUC and accuracy. The AUC's of the four models were fairly different (0.945 with 0.868 accuracy for Homicide, 0.880 with 0.815 accuracy for Assault, 0.901 with 0.853 accuracy for Sex Offense, and 0.864 with 0.845 accuracy for Kidnapping/Abduction).

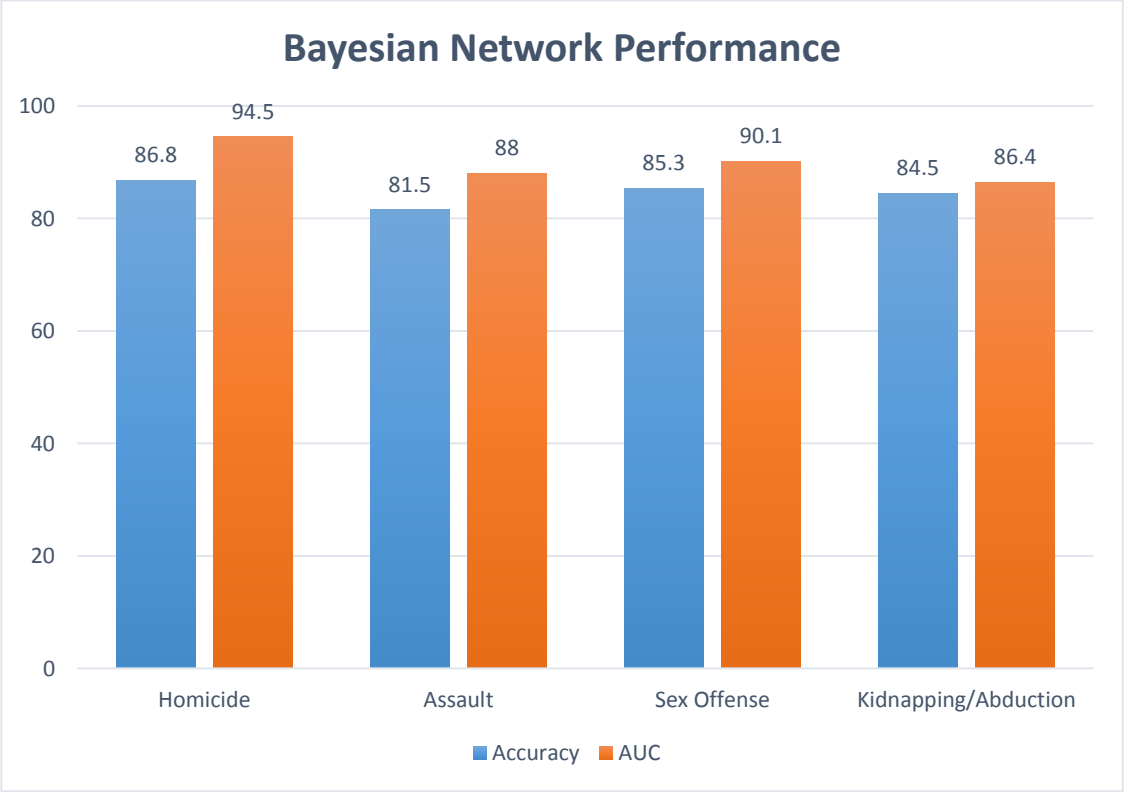


Figure 14: Bayesian Network Performance

3.5.5 Final Predictive Model Comparison

As mentioned above, three different predictive modeling approaches were selected for in-depth development. XGBoost outperformed the other approaches with AUC values of either 98 or 97 on all of the four target variables (Homicide, Assault, Sex Offense, Kidnapping/Abduction). The accuracy and AUC values of penalized logistic regression and Bayesian network were a good deal lower; however, their performance was believed to be acceptable given the superior interpretability benefits they offer. The final predictive model comparison of the three approaches is shown in Figures 15 and 16.

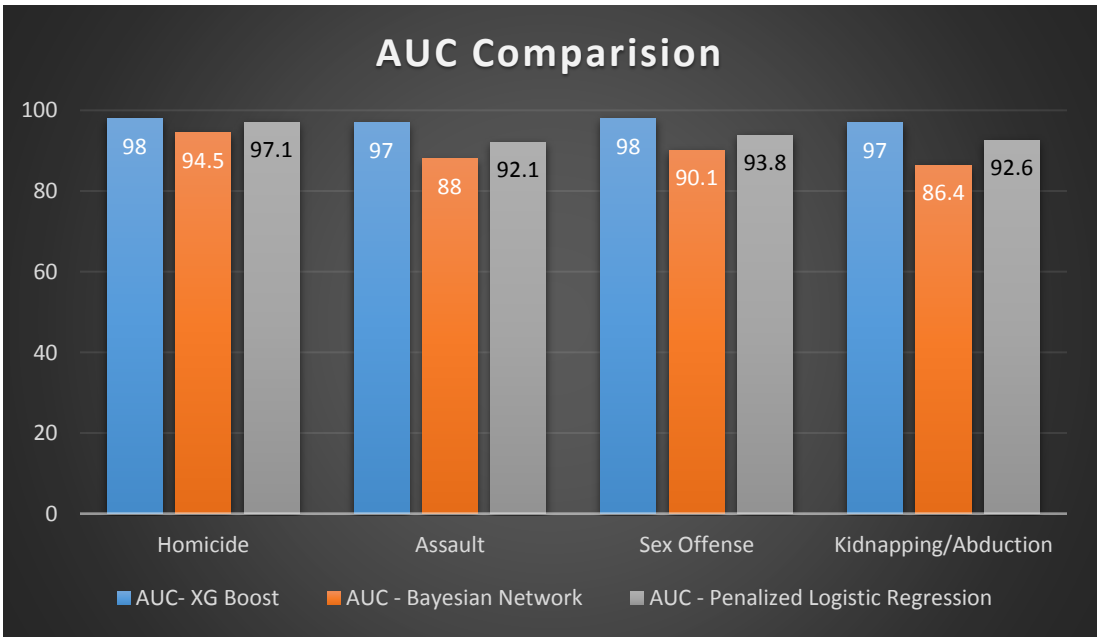


Figure 15: AUC Comparison of all predictive models

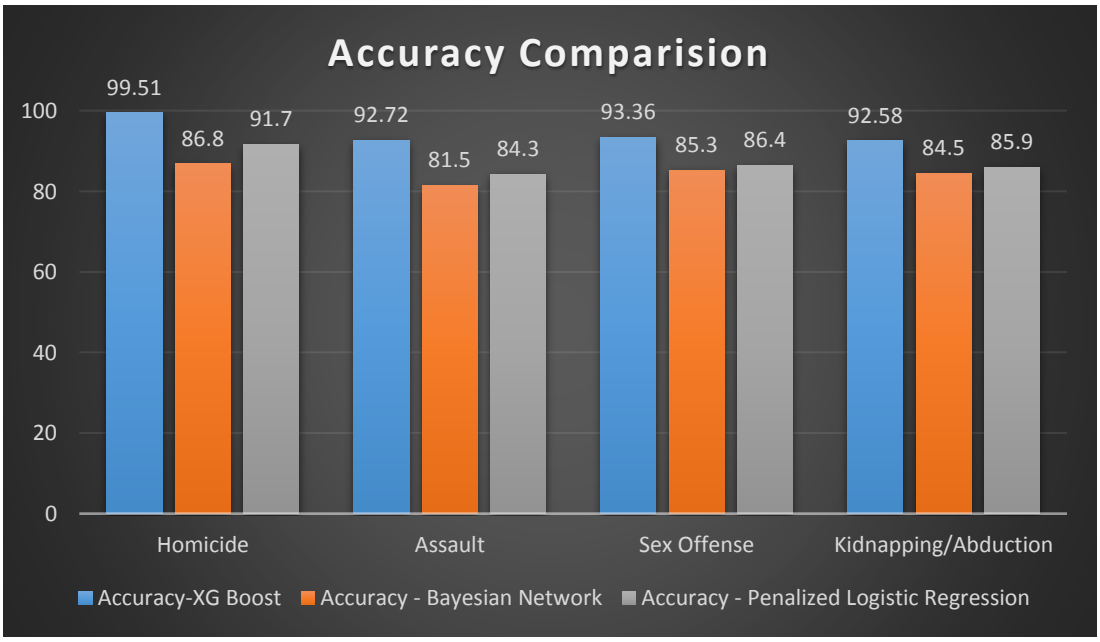


Figure 16: Accuracy Comparison of all predictive models

4 Visualization

Visual knowledge discovery was attempted using geospatial exploration in Tableau, cluster diagrams, and interpretable machine learning visualizations, such as simple dependence plots, local interpretation plots, and conditional probability tables.

4.1 Geospatial Visualization

After some experimentation, it was clear that it made the most sense to visualize the project data geospatially at the county level. The data does not contain complete participation from within states and thus analysis at the state level was abandoned to avoid misinterpretation of states crime rates. The NIBRS project dataset provided a city for each crime, but the population of cities as they were captured by NIBRS was not readily available. For example, there were incidents that were reported by universities (e.g. George Mason University) and other incidents reported by the city in which the university is located (e.g. Fairfax). The complex deconfliction of such situations was outside the scope of this 12-week project, therefore analysis at the city level was also abandoned.

4.1.1 Tableau Exploratory Worksheets

Figure 17 shows a map of the participating counties within the NIBRS base project data set. With the count of the crimes in each county and the population of each county, it was possible to calculate county crime rates. The crime rates discussed in this section represent the number of crimes reported to law enforcement agencies in each county for every 100,000 persons within the county population. This assumes full participation from the counties in the data. County FIPS codes were used to integrate the NIBRS project data with the county populations and poverty rates. Microsoft Excel was used for the data integration tasks related to the geospatial visualization portion of the project. Figure 18 shows participating county population densities. Here, darker blue coloring represents higher population density.

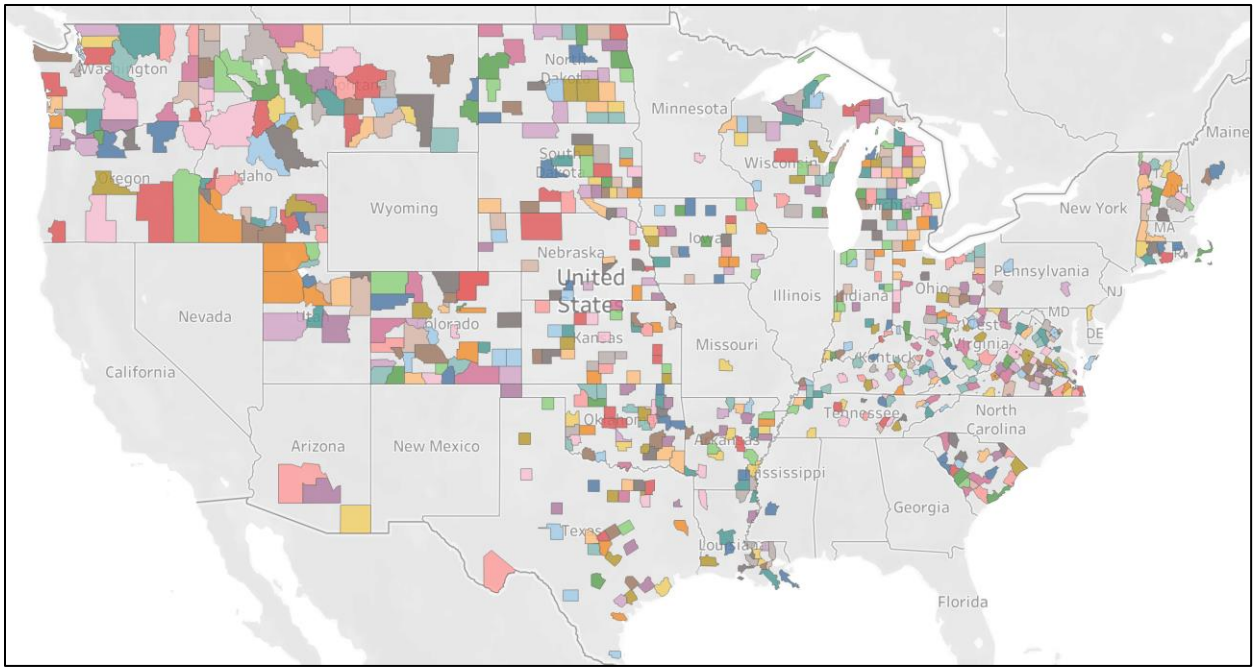


Figure 17: Map of Participating Counties in the NIBRS Project Data Set

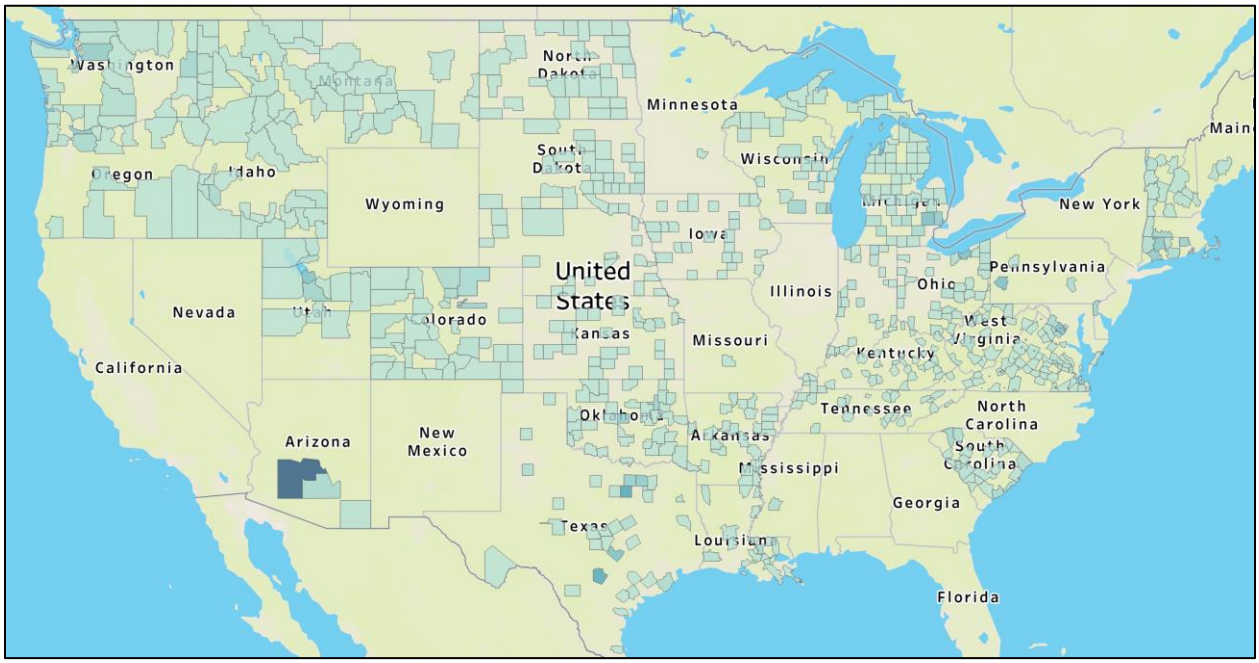


Figure 18: Map of County Population Density - Darker Color Indicates Higher Population

A key goal of the project was to investigate the impact of social and economic factors on crime rates. To do this, the team integrated data with poverty rates and unemployment rates acquired from the U.S. Census Bureau. Figure 19 shows a map of participating counties from project dataset with a simultaneous display of the county poverty rate and county crime rate. The size of the circles

represents crime rate, where the bigger the size implies a higher crime rate. Color represents poverty rate, where darker red color implies the highest poverty rates. Interestingly, it was observed that the poverty rates of the counties appear to have little correlation with crime rates across the counties.

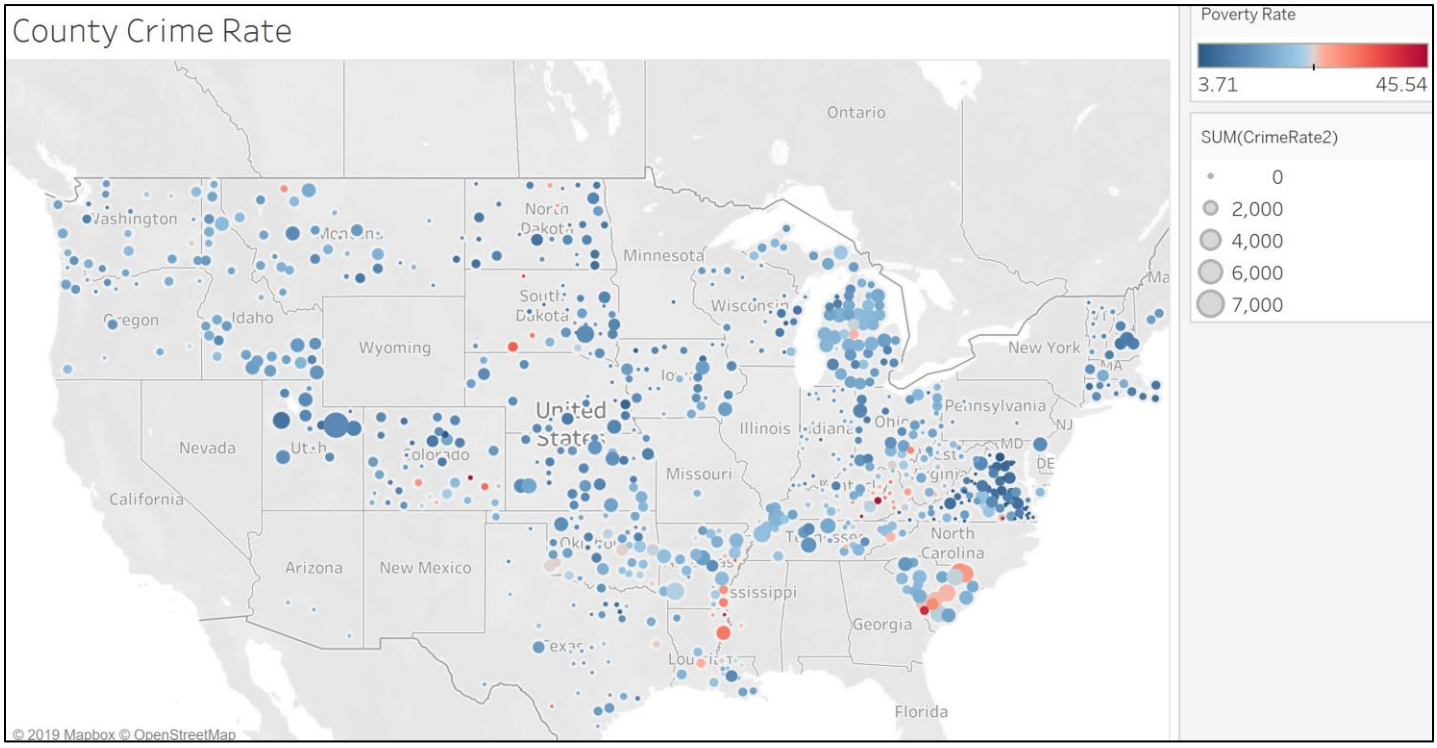


Figure 19: Map of Crime Rates and Poverty Rates for Participating Counties in the Project Data Set

The Tableau exploratory worksheets produced for this project allowed for the inspection of correlation between poverty rates, unemployment rates, and the crime rates for particular offense categories. Filters such as location types, race of victim, sex of victim, resident status of victims, and year of the incidents, allowed for even further insight. Filtering the data reveals that so called "safe places" are as safe as one would hope. Location types like "day care," "parks or play grounds," and "elementary schools" are supposed to be the safe places for kids; however, these were found to have relatively high rates of the Kidnapping/Abduction and Sex Offense crime categories while exploring the data across counties.

Since the project dataset only included data for four years (2013 to 2016), viewing the data by year provided limited value. Instead, the team explored crime rates for each offense category by month over the four-year period. Figure 20 shows an example of such a graph with a time series trend line for Duchesne County, Utah.

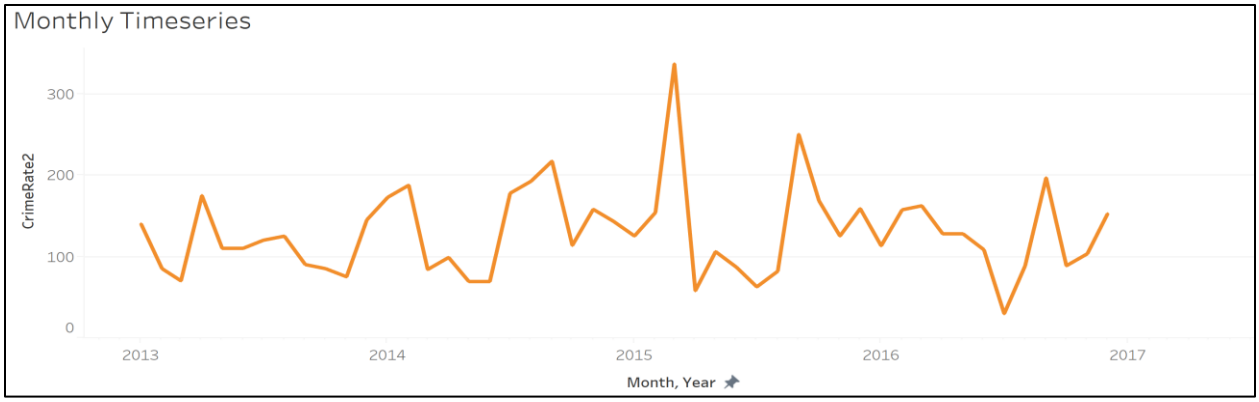


Figure 20: Monthly Crime Rates for Duchesne County from January 2013 to December 2016

4.2 Tableau Dashboards

A series of dashboards were created to explore the project data more efficiently. Figure 21 shows a dashboard for crime rate distributions over offense types, victim-offender relationships, and crime location type categories. This particular example shows how the location types "Residence/Home" and "Elementary School" have higher offense rates displayed for the victim-offender relationship "Acquaintance" in Fairfax County, Virginia or the year 2014. The type of offense which took place most was "Assault."

Figure 22 shows a dashboard that visualizes the crime rate across victim race and the age of the victim. As can be observed, crime rates increased as the age increased in this example for New Castle County, Delaware. Most of the victims belong to the Black or African American race, which can likely be attributed to the underlying race distribution of this particular county.

Figure 23 shows a dashboard that has all the county crime rates from 2013 to 2016 spread over United States with distribution across the race of victim. Circle sizes represent the crime rate and the darker color red circles represent the highest poverty rates. In the example shown, Duchesne County in Utah has the White race as the predominant race of victims, which can likely also be attributed to the underlying race distribution of this particular county. The time series plot in this dashboard shows a second trendline for the poverty rate of the county to examine if there is a relationship between poverty and crime rates in this county over time.

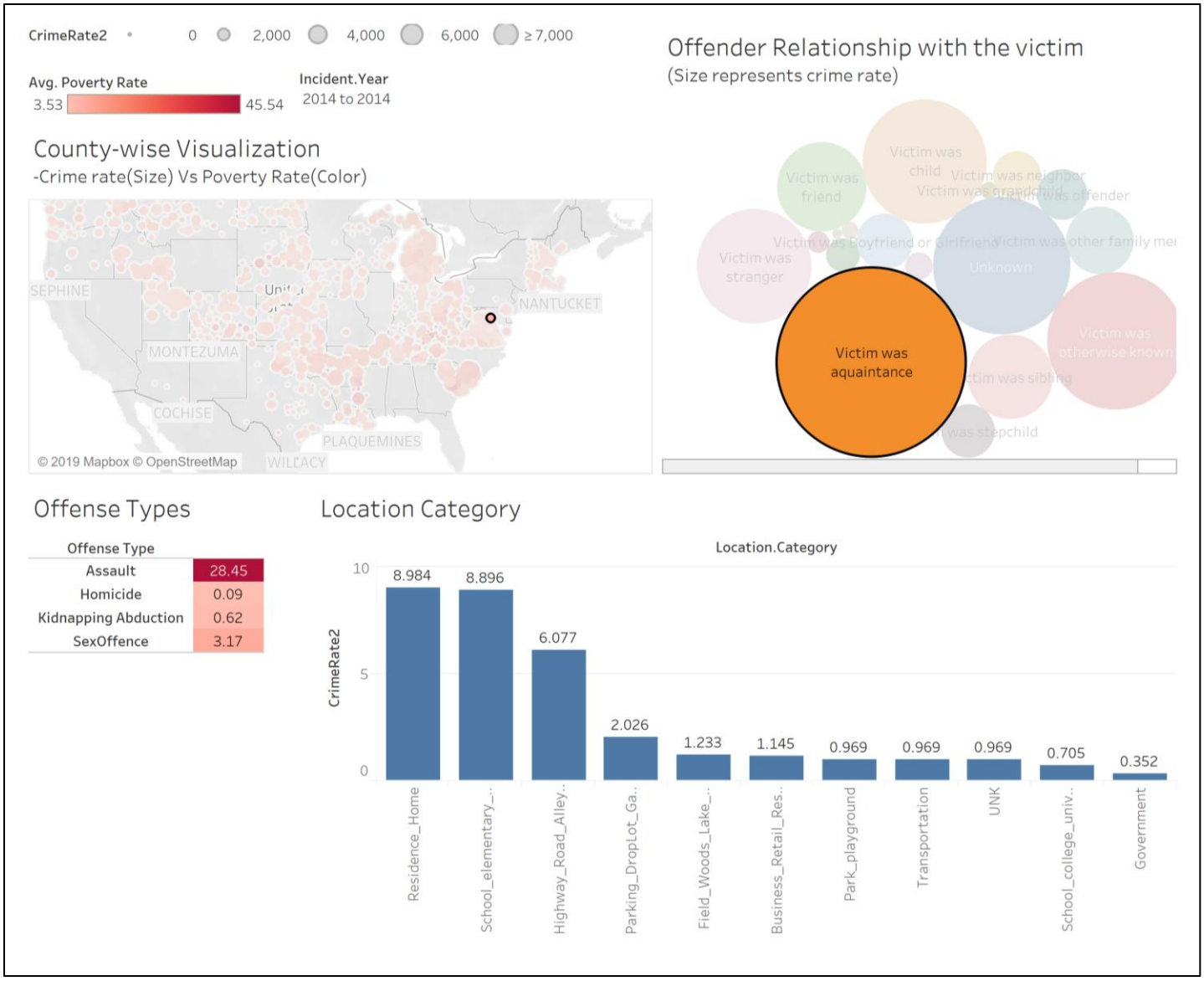


Figure 21: Dashboard for Crime Rate Distributions Over Offense Types, Relationships, and Location Categories

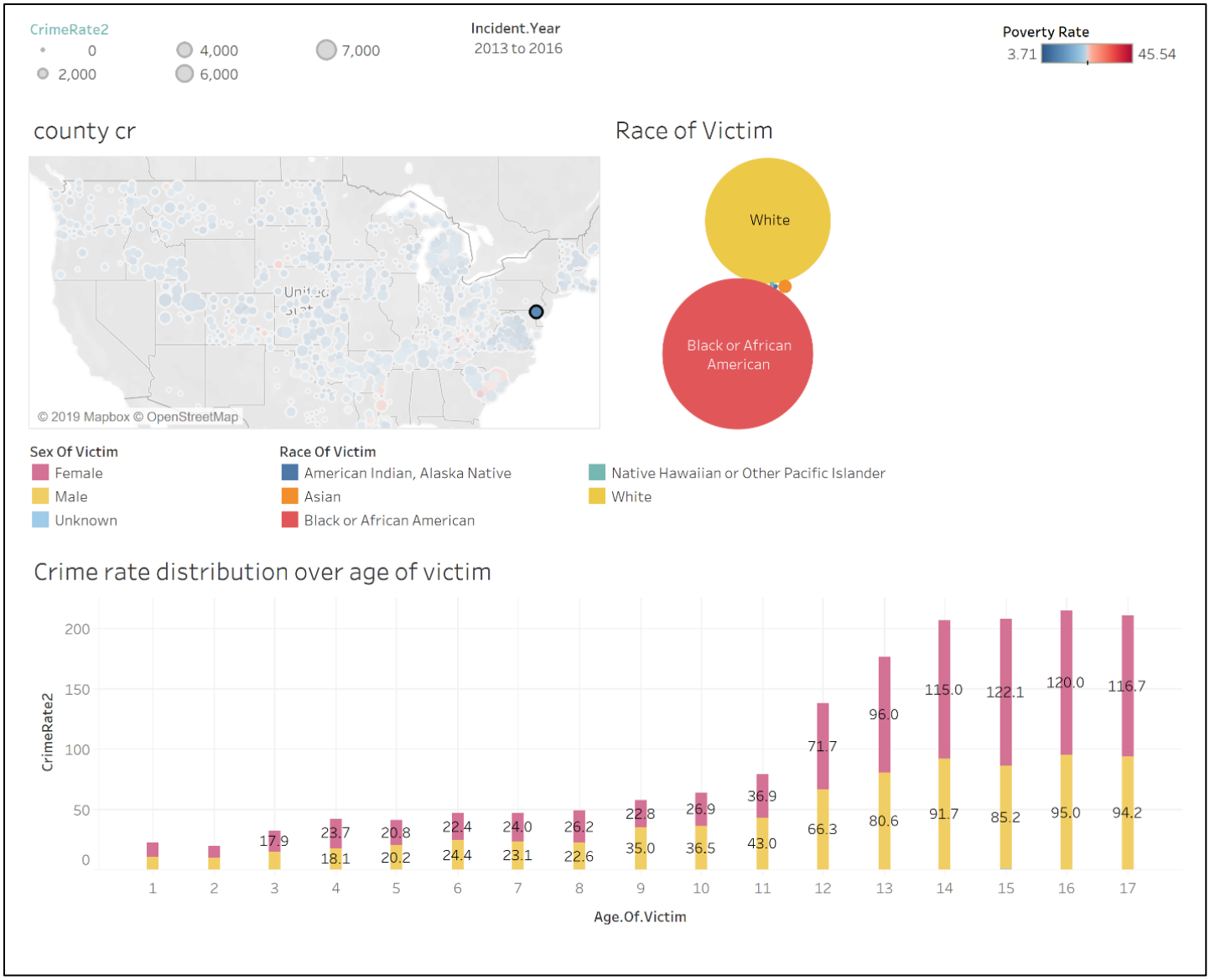


Figure 22: Dashboard for Crime Rate Across Victim Race and the Age of Victim

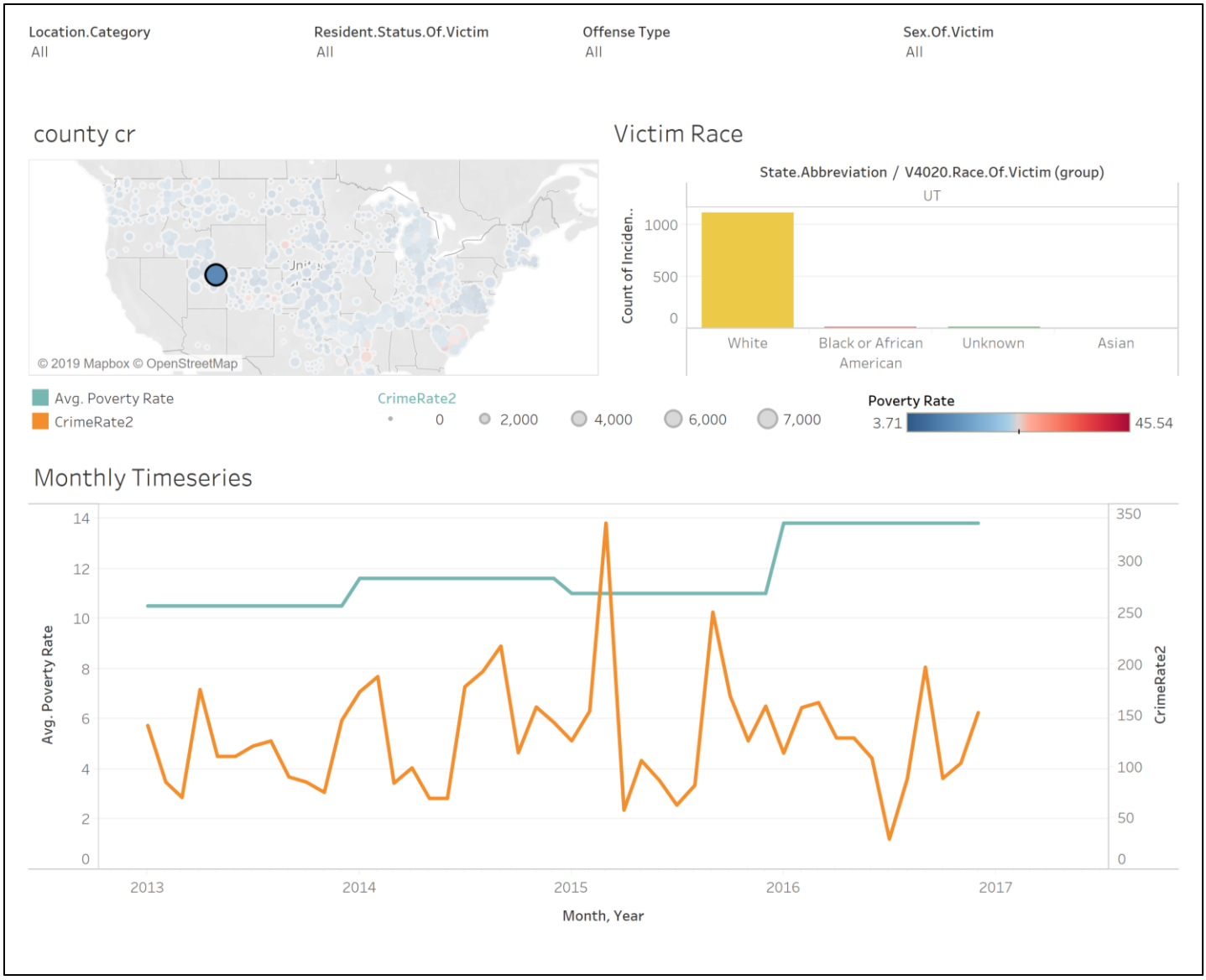


Figure 23: Dashboard of County Rates from 2013 to 2016 with Distributions Across Race of Victim

Figure 24 shows a dashboard visualizing the crime rate distribution from January 2013 to December 2016 along with a forecast for January 2017 to December 2018. It also shows a crime rate distribution across race of the victim with offense category distributions for Providence County in Connecticut. The forecasts were estimated using the exponential smoothing features made available in Tableau. The trend line shown in Figure 24 is a polynomial trend model of degree 8. It is computed for both Unemployment Rate and Crime Rate. The model formula used is defined as follows:

$$Rate * (Attribute\text{^}8 + Attribute\text{^}7 + Attribute\text{^}6 + Attribute\text{^}5 + Attribute\text{^}4 + Attribute\text{^}3 + Attribute\text{^}2 + Attribute + Intercept)$$

The smoothing coefficients used in Tableau were Alpha, Beta, Gamma. Alpha is the level smoothing coefficient, Beta the trend smoothing coefficient, and Gamma the seasonal smoothing coefficient (Tableau 2019).

Figure 25 shows a similar dashboard, but this time between Crime Rate and Poverty Rate with forecasting. Figure 26 shows side-by-side maps comparing the association of crime rate with poverty rate and unemployment rate, offering a different perspective of the collection of variables. Forecasting of crime was not a goal established during project planning. The concept was introduced very late during the execution of the project; therefore, the methods used were fairly rudimentary and left to those readily available in Tableau. Future work is encouraged in this area.

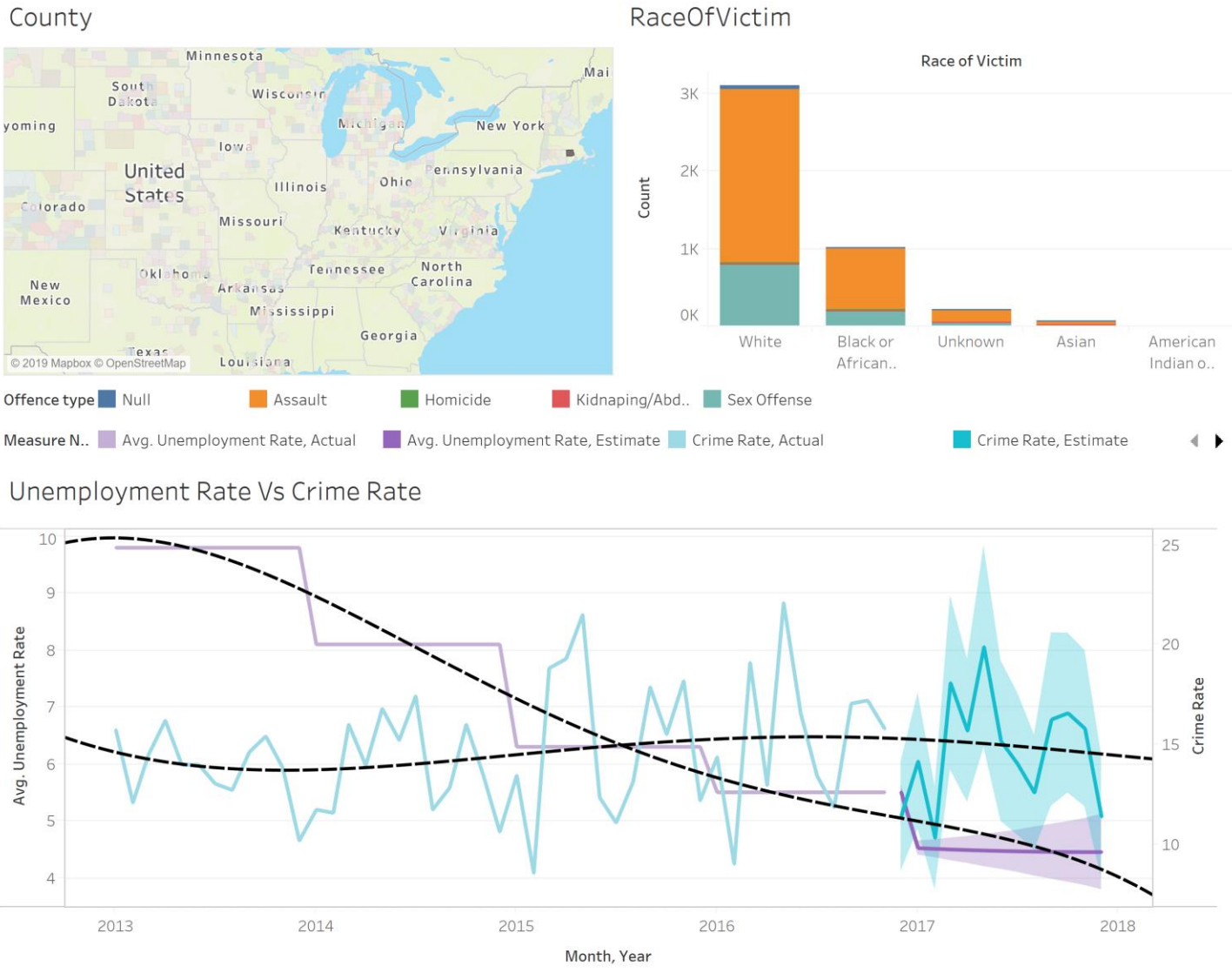


Figure 24: Forecasting Unemployment Rate vs. Crime Rate

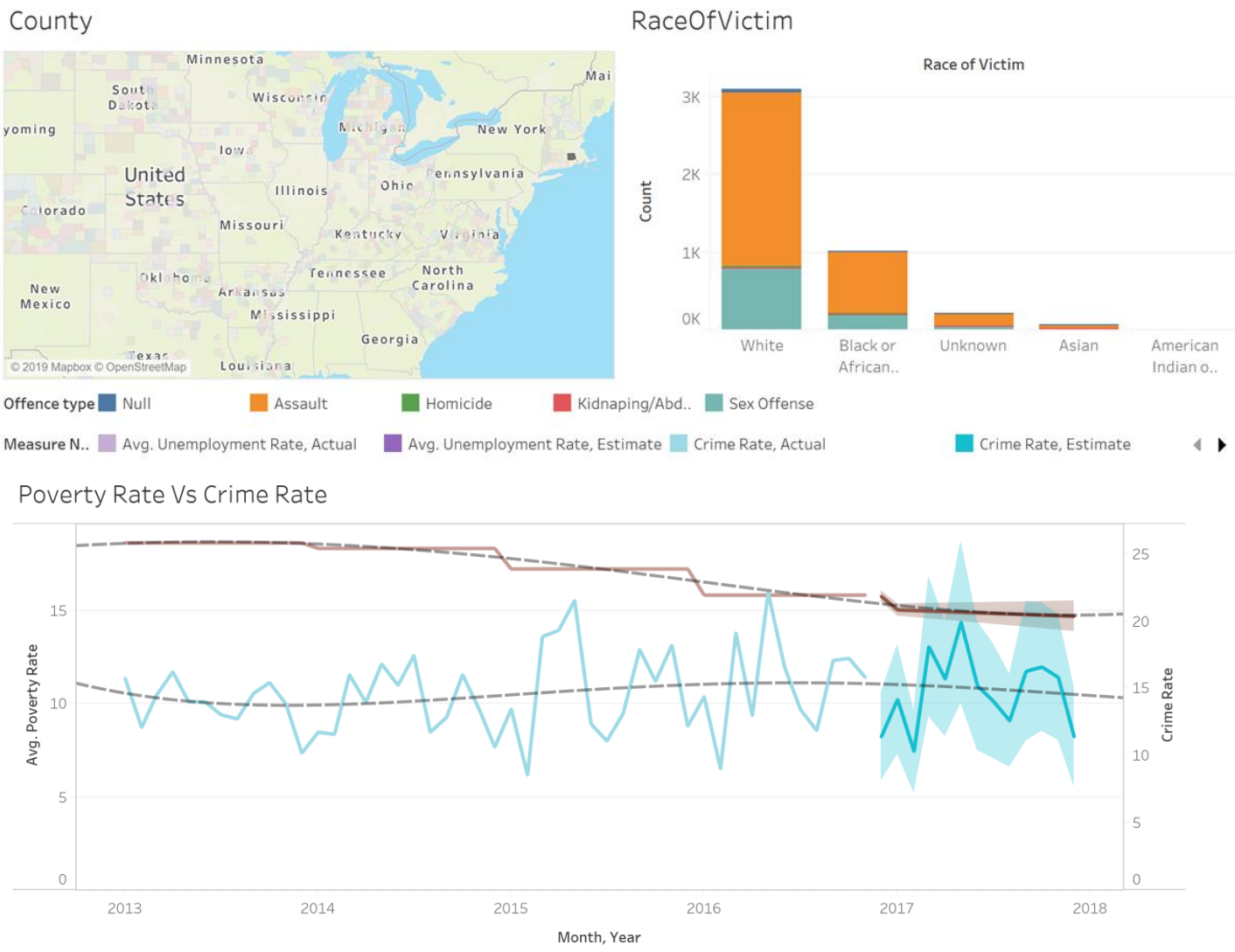


Figure 25: Forecasting Poverty Rate vs. Crime Rate

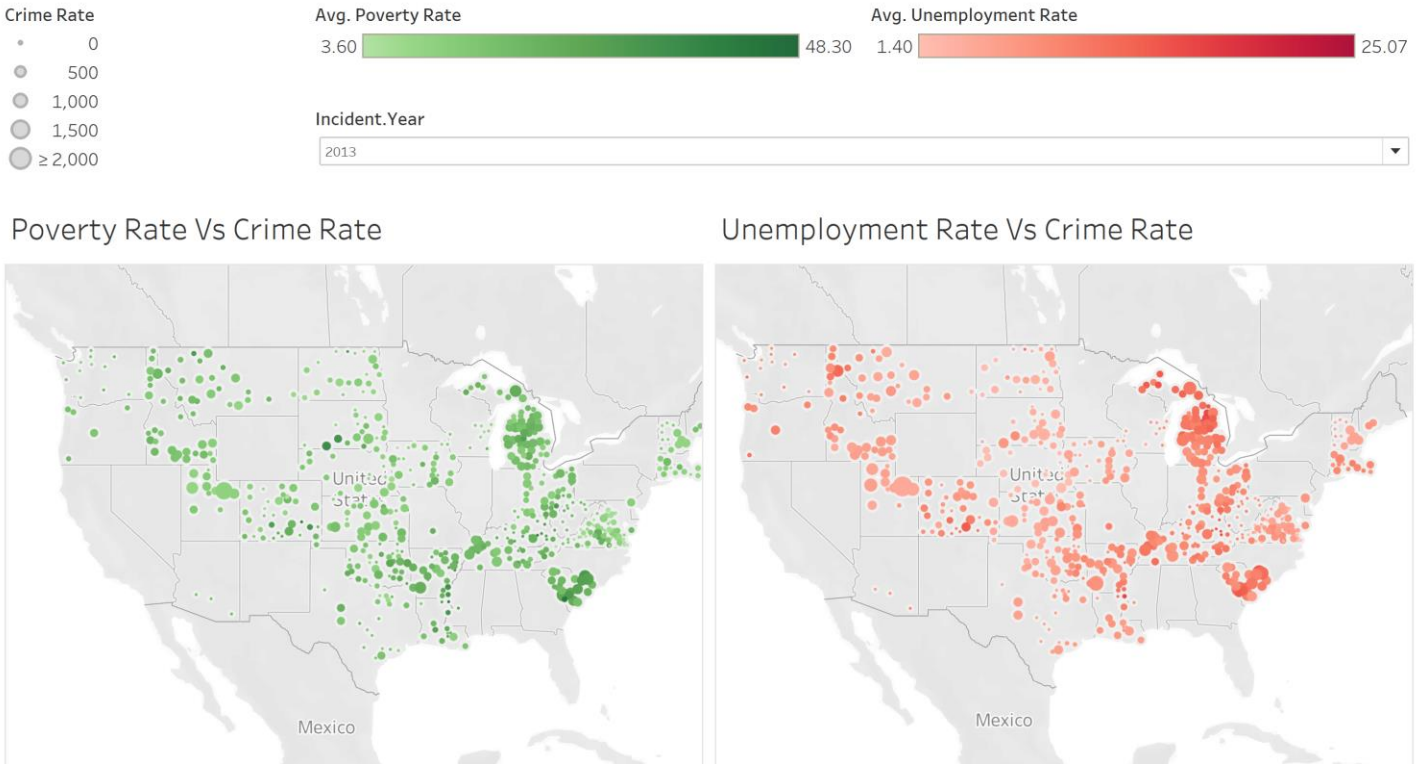


Figure 26: Side-by-side Comparison of Crime Rates with Poverty Rate and Unemployment Rate

4.3 Unsupervised Learning Visualization

4.3.1 Association Rule Learning Visualization

As described in Section 3.3.2, the association rule learning approach was found to have little value with regard to identifying definitive sets of characteristics associated with each of the four offense categories under consideration. A good deal of time was invested in this approach and perhaps dedicating even more resources to the approach could have led to getting more utility out of the association rules. For example, loading the resulting association rules into software designed for network modeling, such as *Gephi*, may have made the most interesting rules more evident.

The project team only got as far as producing visualizations such as the one shown in Figure 27, which was created using the Kidnapping/Abduction dataset. Here the top five rules when sorted by lift with the 15 to 17-year old age group as the consequent are graphed, with each circle representing a rule composed of the variables pointing to it. The size of the circle represents the amount of support the rule has. The darkness of the circle represents the amount of confidence in the rule. Such visualizations are much more useful than reading a list of rules; however, they tend to become overly cluttered with more than 5 to 10 rules in one graph using a fixed plot.

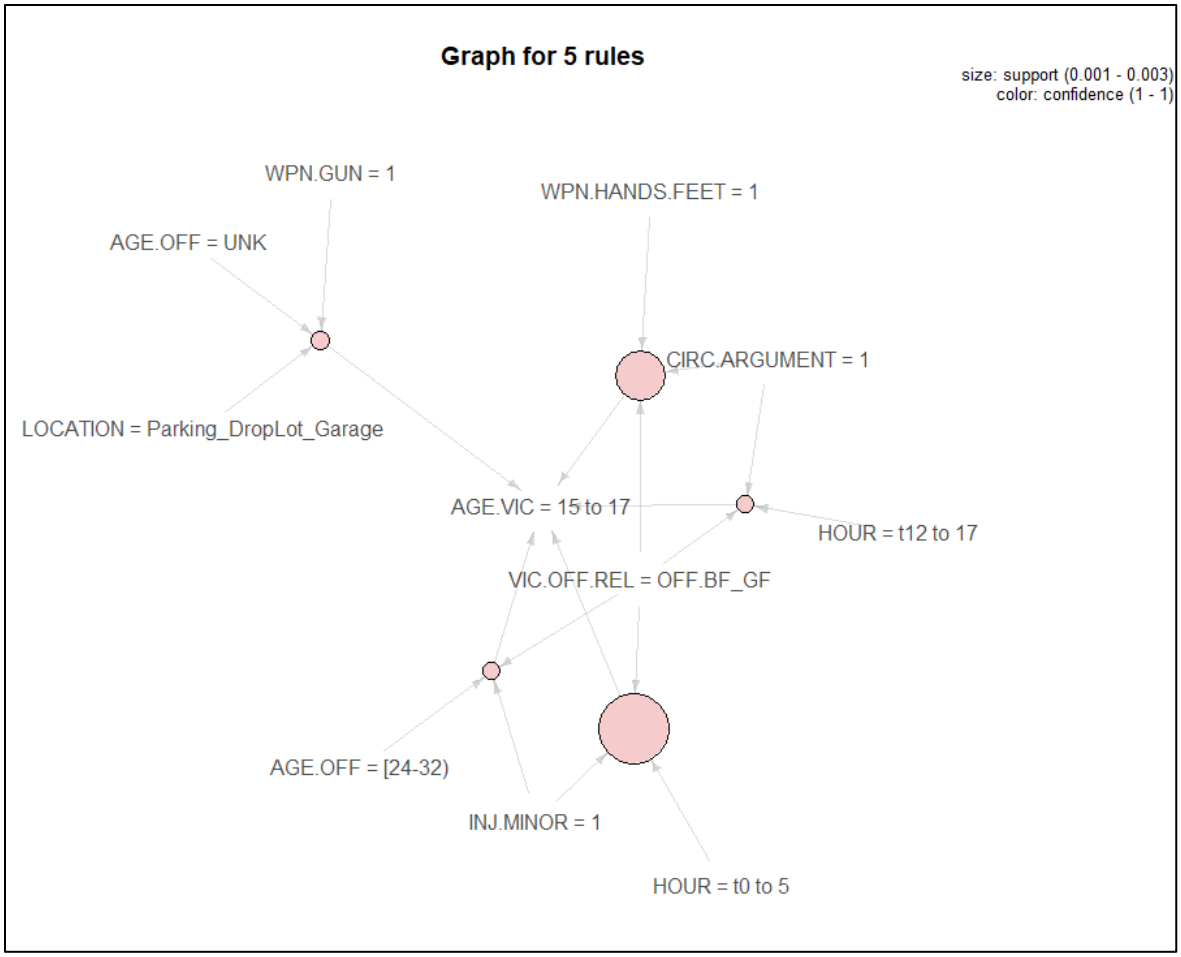


Figure 27: Top 5 Association Rules for the 15 to 17-Year-Old Age Group in Kidnapping/Abduction

4.3.2 Hierarchical Clustering Visualization

Agglomerative hierarchical clustering was found to produce visualizations that were highly informative with regard to the first research question of this project. Hierarchical clustering is typically represented in the form of a dendrogram, like the one shown in Figure 28. Here, five major clusters are distinguished by color. The height (in this case length due to the horizontal orientation) represents the dissimilarity among the variables. This is a useful view for determining the characteristics that are associated with each other for this particular offense (Kidnapping/Abduction); however, it can be difficult to identify the relevant subclusters. In addition, variables directly above and below one another may in fact be separated by a good deal of distance in the plot.

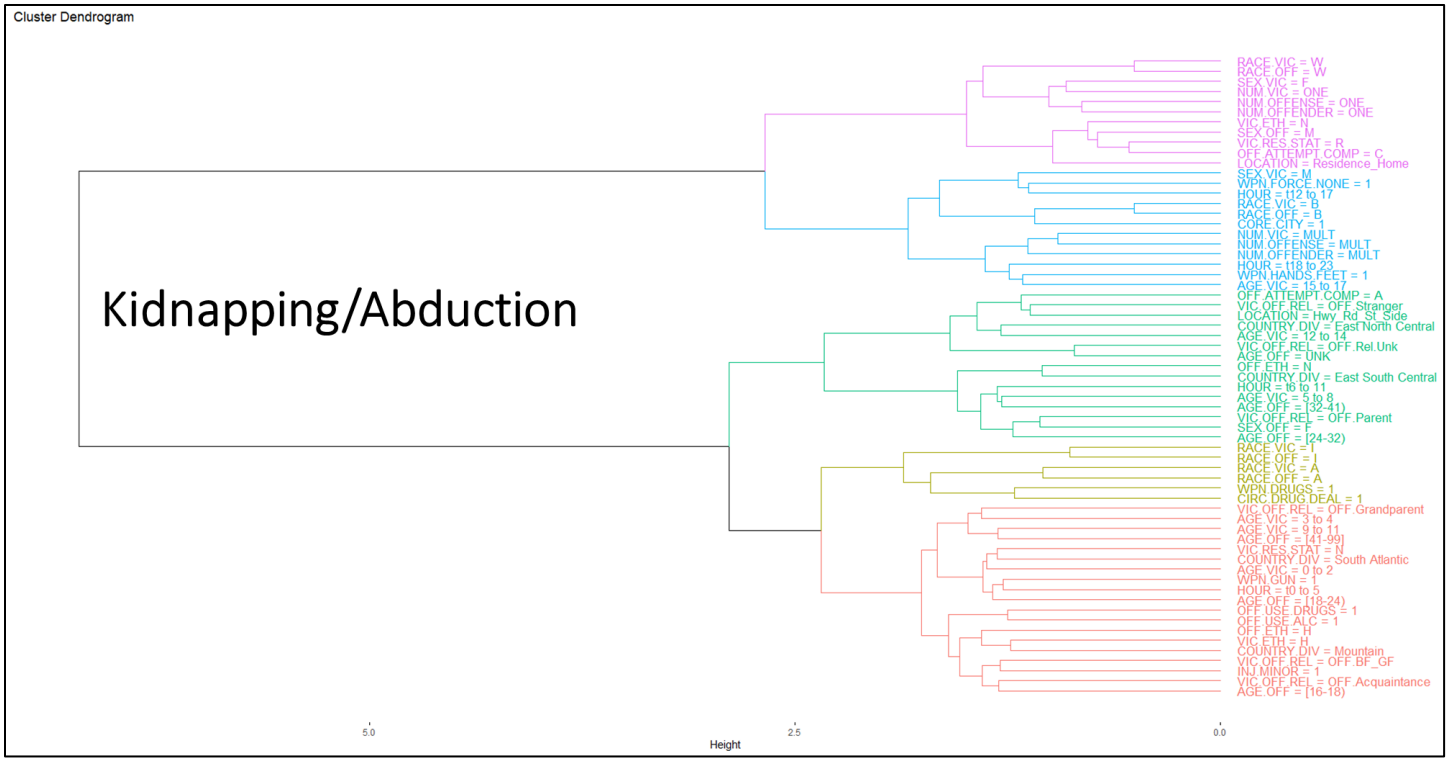


Figure 28: Dendrogram for the Kidnapping/Abduction Offense Category

Another useful layout for the hierarchical clustering approach is shown in Figure 29. This is referred to as a phylogenetic tree diagram, which is more typically used to represent the evolutionary relationships among species. It represents the same information as the dendrogram in that the distances, when following the lines, represents the extent of the dissimilarity between the variables. Subclusters become more evident when viewed this way and it becomes easier in some ways to understand the proximity among the variables.

The clusters of variables resulting from this procedure provide patterns of associations that can point law enforcement, caretakers of children, and social science researchers towards combinations of characteristics of interest. The patterns of variables in the clustering diagrams can be seen more clearly in the project presentation slides.

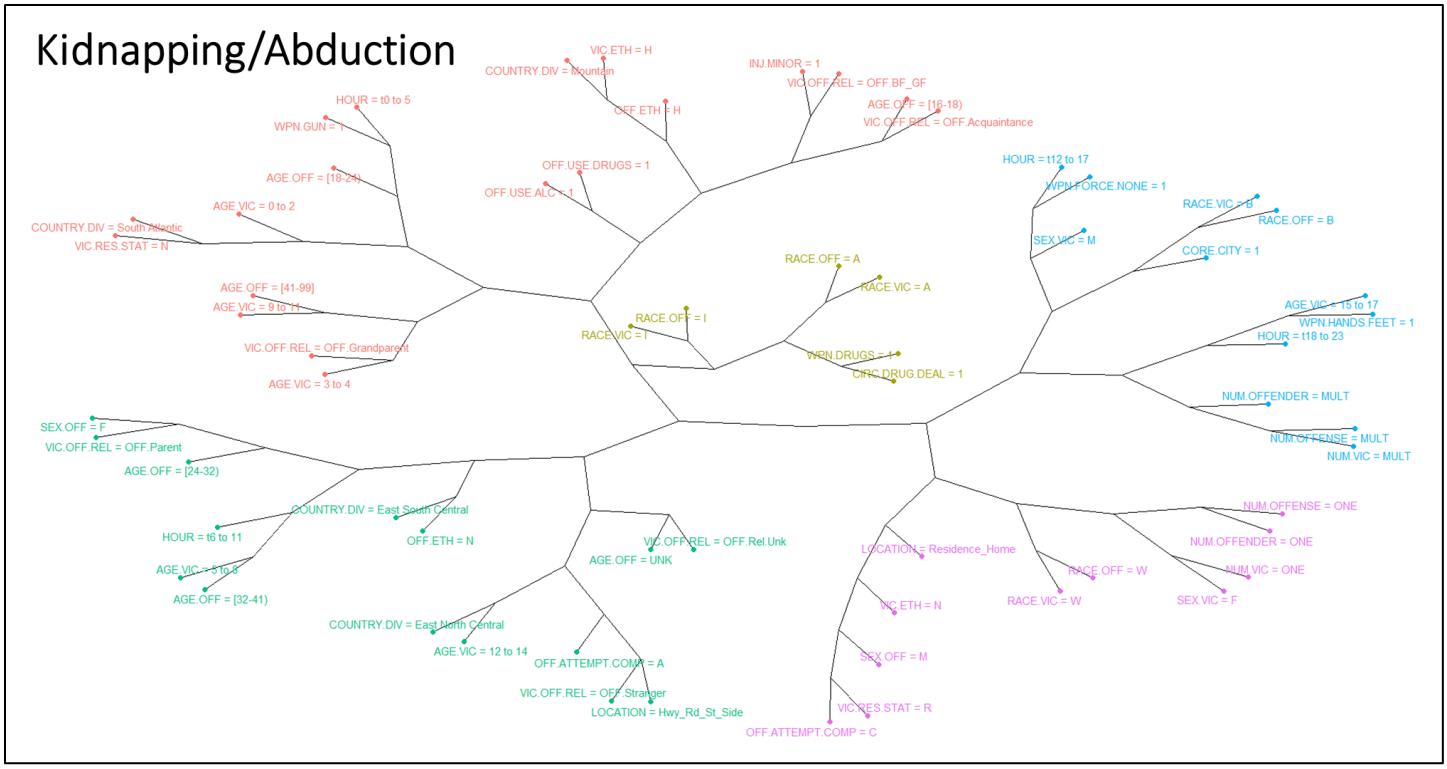


Figure 29: Phylogenetic Tree Diagram for Kidnapping/Abduction

In a sense, each cluster in the dendrograms and phylogenetic trees tells a story. The team reviewed the patterns in the cluster diagrams to produce the following summaries for each of the four offense categories. This is type of information the team was hoping to get out of the association rule approach.

Assault:

1) Victim aged 15 to 17; Victim is Female; Offender is Female; Time is 1200 to 2300; Minor Injuries; Multiple victims, offenders, and offenses; In a city environment within the East-North Central country division
2) Victim is male; Offender is male; Offense occurs in a residence/home; Using hands/feet as weapons; One victim, one offender, one offense
3) Victim aged 12 to 14; Offender age is 13 to 18; Time is 0600 to 1100; In schools within the South-Atlantic country division
4) Victim aged 0 to 2 or 9 to 11; Offender aged 1 to 13 or 24 to 32; Offender is parent or of unknown relationship; In the East-South-Central country division
5) Offender age 18 to 24 and 41 or older; Multiple offenses; Offense take place along roads/highways; In the West-South-Central country division

Sex Offense:

1) Victim is male aged 5 to 11; Offender is parent or of unknown relationship aged 24 or older; Time is midnight to 1100; Multiple victims, offenders, and offenses; In the East-South-Central and South-Atlantic country division
2) Victim aged 12 to 17; Offender is an acquaintance aged 18 to 24; Time is 1200 to 2300; No weapon involved; In a city environment within the East-North-Central country division
3) Male offender; Female victim; Hands/feet as weapon; Offense in residence/home; Offense completed
4) Offender is sibling aged 1 to 13; Offender female grandparent; Offender is boyfriend/girlfriend; Alcohol and/or drugs involved; West-South-Central country division

Kidnapping/Abduction:

1) Victims aged 0 to 2; Offenders aged 18 to 24; Time is midnight to 0500; Using gun as a weapon; In the South-Atlantic country division
2) Victims aged 3 to 4 and 9 to 11; Offender is Grandparent
3) Victim aged 5 to 8; Offender is female parent aged 24 to 41; Time is 0600 to 1100; In the East-South-Central country division
4) Victims aged 12 to 14; Attempted offenses along roads/highways; Offender a stranger or unknown; In the East-North-Central country division
5) Offenders Aged 16 to 18; Victim and Offender are Boyfriend/Girlfriend; Offenders using drugs and/or alcohol; In the Mountain country division
6) Male victims; Time is 1200 to 2300; Using hands/feet as weapons; Multiple victims, offenders, and offenses; In a city environment
7) Female victim; Male offender; Offense occurs in a residence/home; One victim, one offender, one offense

Homicide:

1) Victim is male; Offender is male; Offense occurs in a residence/home; One victim, one offender, one offense
2) Drug dealing involved; Drugs as a weapon; Offender using drugs Lacerations, internal, and other major injuries; In the Mountain and West-South-Central country divisions
3) Female victim aged 0 to 2; Female offender aged 24 to 32; Time is 1200 to 1700; Hands/Feet as weapons; In the East-North-Central country division
4) Victim is aged 15 to 17; Offender is aged 18 to 24; Time is 1800 to 2300; Gun used as a weapon; Multiple victims, offenders, and offenses; In a city environment
5) Victims age is 3 to 4; Offender age is 16 to 18 or 32 to 41; Offender relationship is an acquaintance or unknown Time is midnight to 1100 hours along roads/highways; Circumstances involve an argument In the East-South-Central and South-Atlantic country division

4.4 Predictive Modeling Visualizations

4.4.1 Penalized Logistic Regression Visualization and Interpretation

The variable importance plots resulting from the four final Elastic Net models can be examined in Figure 30. For logistic regression models, variable importance is represented by normalizing the absolute values of the resulting model coefficients to 100 for each feature in the model. Variable importance plots are useful for understanding which features are most predictive of each crime; however, it is difficult to compare which features have the strongest positive and negative influence on each offense category. To get at this, the team departed from the traditionally used variable importance plots and instead made a heat map of the coefficient values resulting from each model.

The heat map displayed in Figure 31 shows to what extent each of the features is predictive (positively or negatively) within each of the offense categories. The shade of the colors indicates the relative value of the coefficients within its respective model. When viewing the model outputs in this way, some interesting features stand out in each of the models. For example, a lack of weapon/force use has a positive influence on predicting Sex Offense and Kidnapping/Abduction, but a negative influence on predicting Assault and Homicide. The sex of offender feature is male feature shows a positive predictive influence for Sex Offense and Homicide, but negative for Assault and Kidnapping/Abduction, implying these latter two are more predictively associated with females. As a final example, the "location of field, woods, lake, and beach" has a strong predictive influence for homicide when compared to the other offense categories. A summary of more findings that were identified in this approach is provided in Section 6. It is true that much of this could be discovered through descriptive statistics, but this approach allows us to see the influence of a feature while in the context of all other features.

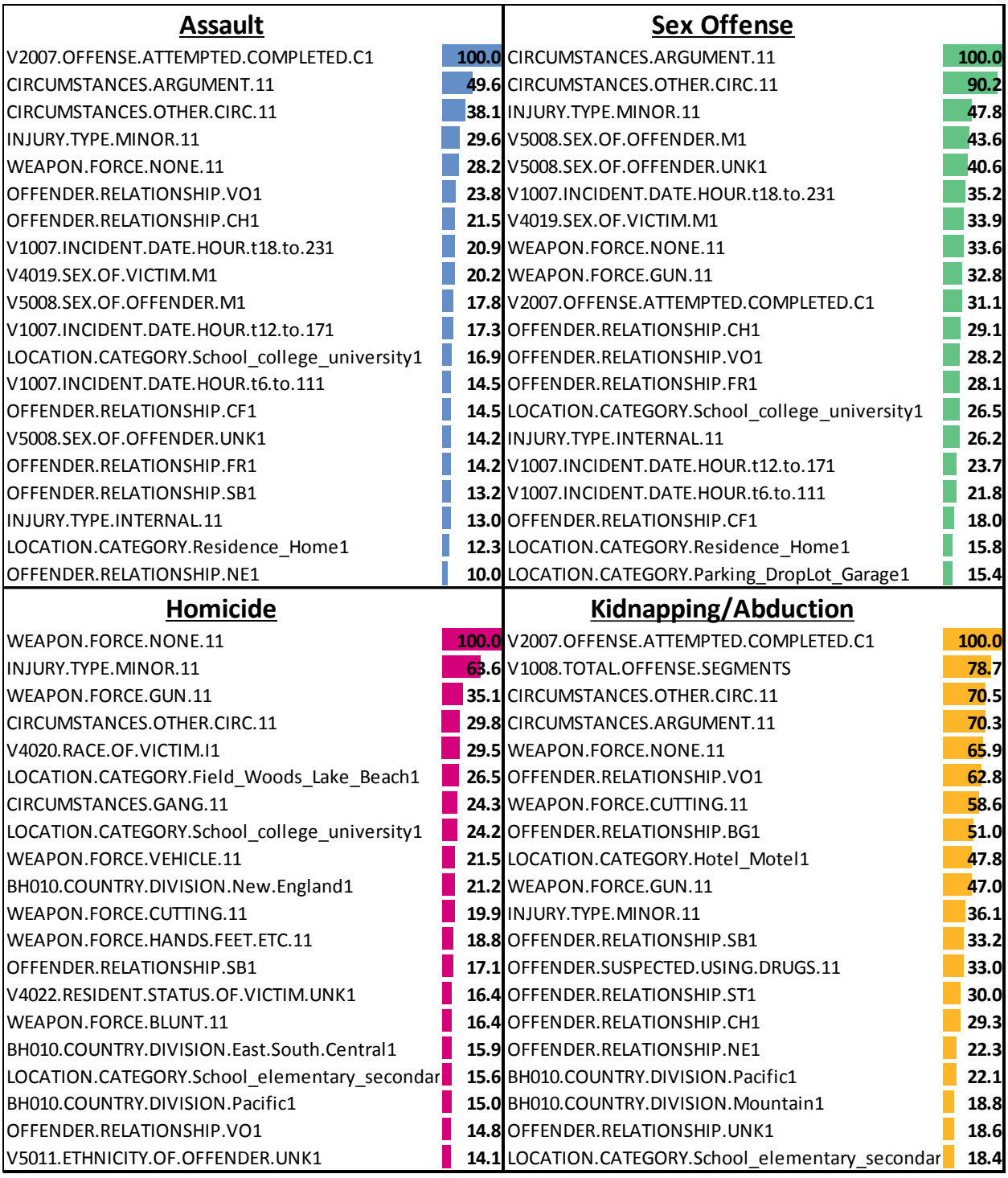


Figure 30: Variable Importance Across the Four Final Penalized Regression Models

Feature	Assault	Sex Off	Kid/Abd	Homicide
V2007.OFFENSE.ATTEMPTED.COMPLETED.C1	6.9	-1.5	-3.2	0.0
WEAPON.FORCE.NONE.11	-1.9	1.6	2.1	-7.7
CIRCUMSTANCES.ARGUMENT.11	3.4	-4.8	-2.2	0.5
CIRCUMSTANCES.OTHER.CIRC.11	2.6	-4.4	-2.2	2.3
INJURY.TYPE.MINOR.11	2.0	-2.3	-1.1	-4.9
WEAPON.FORCE.GUN.11	-0.2	-1.6	1.5	2.7
OFFENDER.RELATIONSHIP.VO1	1.6	-1.4	-2.0	-1.1
V5008.SEX.OF.OFFENDER.M1	-1.2	2.1	-0.4	1.1
V1007.INCIDENT.DATE.HOUR.t18.to.231	1.4	-1.7	0.0	0.0
LOCATION.CATEGORY.School_college_university1	1.2	-1.3	-0.2	-1.9
V4019.SEX.OF.VICTIM.M1	1.4	-1.6	0.0	0.1
V5008.SEX.OF.OFFENDER.UNK1	-1.0	2.0	-0.3	0.5
OFFENDER.RELATIONSHIP.CH1	1.5	-1.4	0.9	0.7
V1008.TOTAL.OFFENSE.SEGMENTS	-0.3	-0.3	2.5	0.9
OFFENDER.RELATIONSHIP.BG1	0.3	-0.4	1.6	-1.0
LOCATION.CATEGORY.Field_Woods_Lake_Beach1	-0.4	0.5	0.0	2.1
WEAPON.FORCE.CUTTING.11	-0.1	-0.6	1.8	1.5
OFFENDER.RELATIONSHIP.SB1	0.9	-0.7	-1.0	1.3
V1007.INCIDENT.DATE.HOUR.t12.to.171	1.2	-1.1	0.1	-0.2
OFFENDER.RELATIONSHIP.FR1	-1.0	1.4	0.0	-0.3
V4020.RACE.OF.VICTIM.I1	0.0	0.0	0.0	2.3
INJURY.TYPE.INTERNAL.11	-0.9	1.3	0.0	0.0
WEAPON.FORCE.HANDS.FEET.ETC.11	-0.7	0.6	0.4	-1.5
V1007.INCIDENT.DATE.HOUR.t6.to.111	1.0	-1.1	0.5	0.5
BH010.COUNTRY.DIVISION.East.South.Central1	-0.3	0.7	0.3	-1.2
CIRCUMSTANCES.GANG.11	0.0	0.0	0.0	1.9
OFFENDER.RELATIONSHIP.CF1	1.0	-0.9	-0.1	0.7
BH010.COUNTRY.DIVISION.Pacific1	-0.4	0.2	0.7	-1.2
V4022.RESIDENT.STATUS.OF.VICTIM.UNK1	0.4	-0.6	-0.2	1.3
BH010.COUNTRY.DIVISION.New.England1	0.1	0.0	-0.2	-1.6
LOCATION.CATEGORY.School_elementary_secondary1	0.5	-0.2	-0.6	-1.2
WEAPON.FORCE.VEHICLE.11	0.0	0.0	0.0	1.7
OFFENDER.RELATIONSHIP.ST1	-0.3	-0.7	0.9	0.4
LOCATION.CATEGORY.Residence_Home1	-0.8	0.8	0.0	0.6
LOCATION.CATEGORY.Hotel_Motel1	0.0	0.0	1.5	0.0

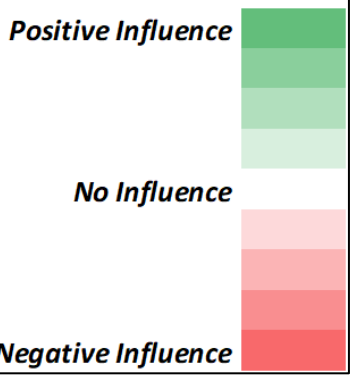


Figure 31: Variable Influence Comparison Across Elastic Net Models

4.4.2 XGBoost Visualization and Interpretation

4.4.2.1 Shapley Additive Explanation (SHAP)

One way to attempt to interpret a XGBoost model is to examine the tree leaf values for each classification instance. A sample will terminate in one leaf for each tree. The logit for a sample is the sum of the values from all sample leaves. Gradient boosted ensembles sum over the predictions of all trees, so the logit can then be used to compute the predicted probability of class membership. Because of this ensemble method, visually examining the relationships of an individual tree provides little insight into how the model is making decisions. Other methods were instead used to explore the behaviors of the XGBoost models.

Feature importance for XGBoost is typically calculated based on an information gain method. Trees are constructed greedily, so features near the root of the tree tend to be more important than features split near leaves, yet the gain method is often biased to attribute more importance to lower

splits (Lundberg 2012). As an alternative for generating feature importance rankings using the gain method, the team used the Shapley Additive Explanation (SHAP) approach. While it can be computationally expensive, SHAP is both consistent and accurate because it attempts to fairly distribute the difference between the prediction and the average prediction among the feature values of each instance (Molnar 2019).

The summary plot in Figure 32 shows the SHAP values for every instance from the training data set for the Sex Offense category. The *x* position of the dot is the impact of that feature on the model's prediction for the row, and the color of the dot represents the value of that feature for the row. For example, for the CIRCUMSTANCE.ARGUMENT feature, the red values indicate rows where there was an argument during the crime. These points are towards the far left of the plot, indicating a strong negative influence on the prediction of a sex offense. This feature SHAP value chart can be simplified by taking the mean SHAP value for each feature. In Figure 33, the *x*-axis is the average magnitude change in model output when a feature is hidden from the model (for this model the output has log-odds units) (Lundberg, n.d.).

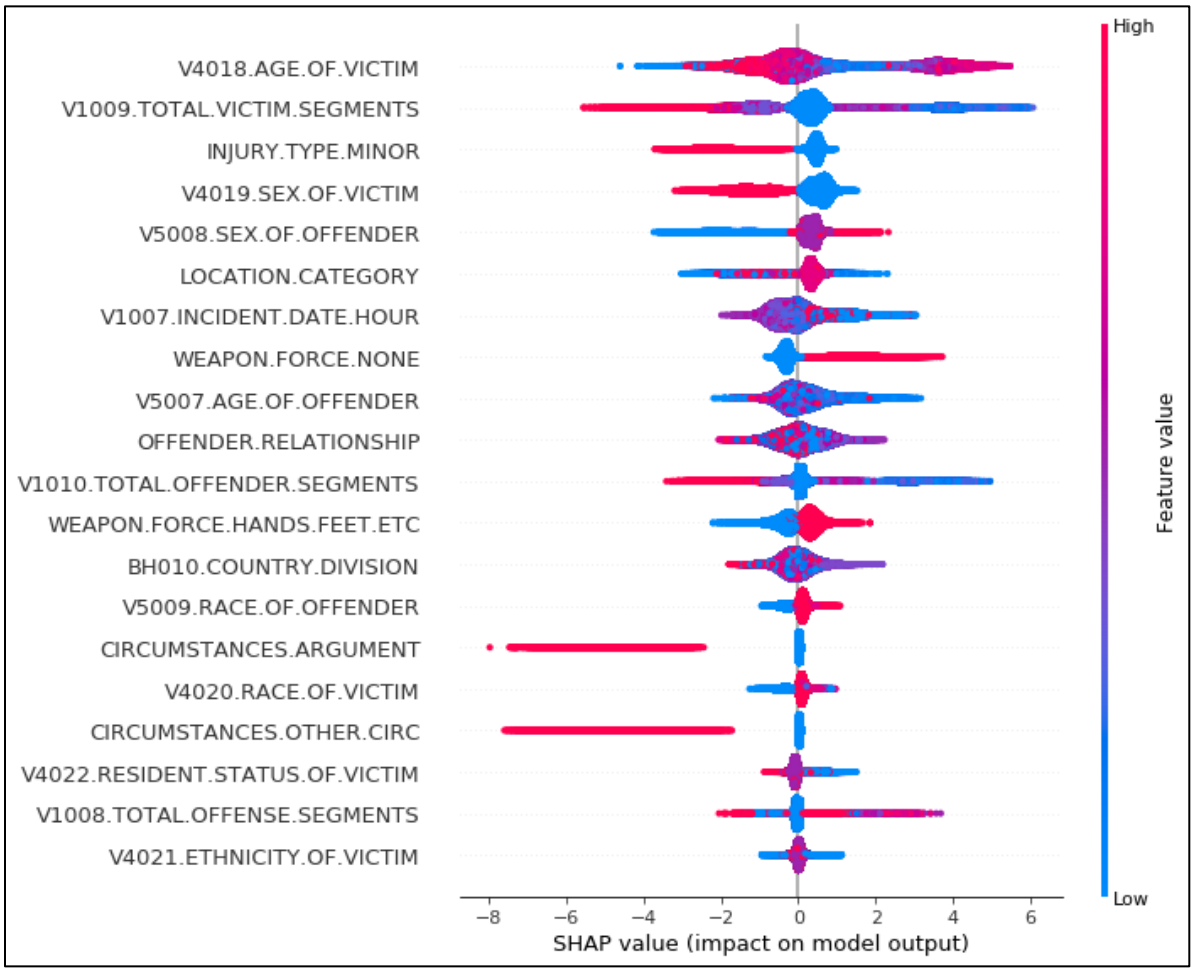


Figure 32: SHAP Feature Value Chart for Sex Offense

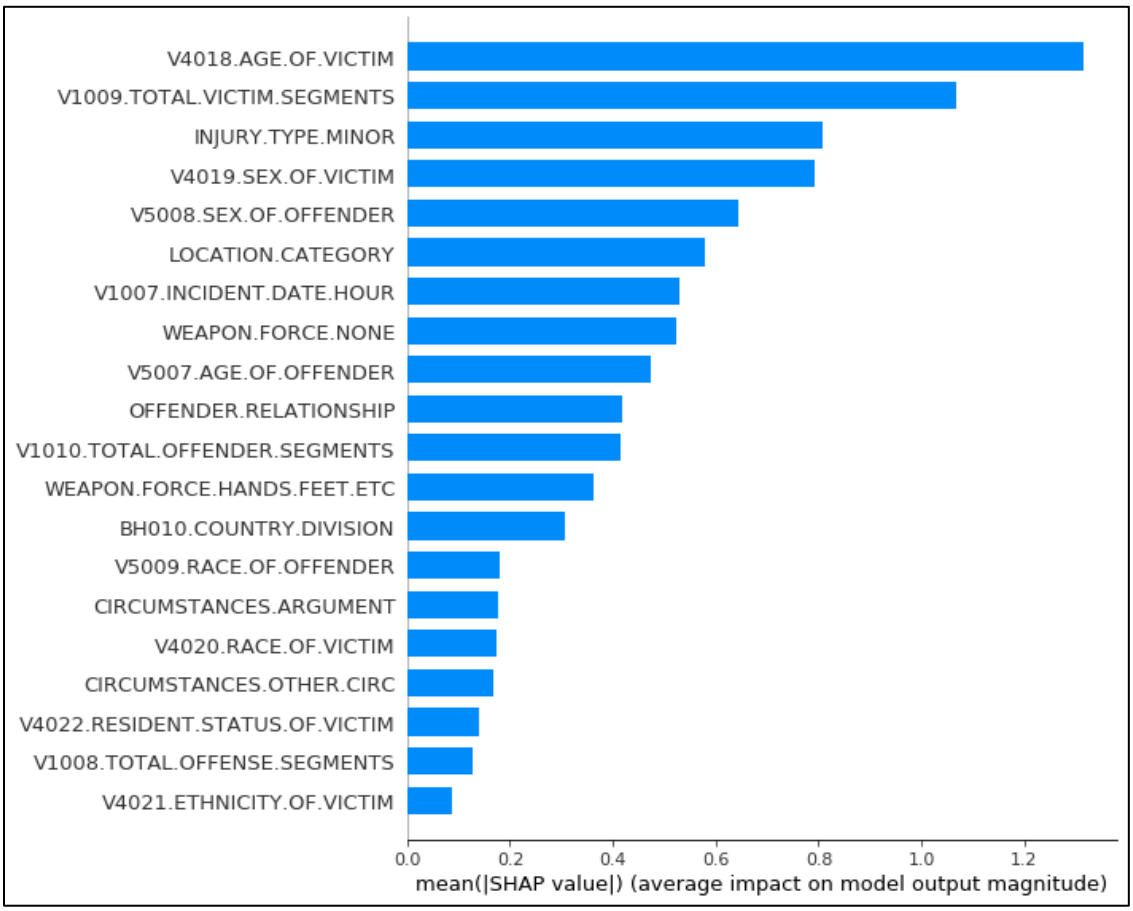


Figure 33: Mean SHAP Variable Importance Plot for Sex Offense

SHAP values can be used to create simple dependence plots. A simple dependence plot is a scatter plot that shows the effect that a single feature has on the predictions made by the model (Molnar 2019). Two examples of such plots are shown in Figures 34 and 35. Each dot in these figures represents a single prediction (a row) from the dataset. The *x*-axis is the value of the feature taken from the training data set. The *y*-axis is the SHAP value for that feature, which represents how much knowing that feature's value changes the output of the model for that sample's prediction (Molnar 2019).

For Figures 34 and 35, the units are log-odds of being a homicide offense versus the other three offenses. The color corresponds to a second feature that may have an interaction effect with the feature we are plotting (by default this second feature is chosen automatically). If an interaction effect is present between this other feature and the feature being plotting, it will show up as a distinct vertical pattern with separate coloring (Lundberg, n.d.). For Figure 34, not being 'other circumstances' with gun as a weapon is more likely to be a homicide offense. This suggests an interaction effect between them. For Figure 35, the use of a gun is represented by red color and having no gun is represented by blue dots. Given there are many more blue dots towards the top of the chart, it can be concluded that the not using a gun feature value has a stronger positive influence on the prediction of homicide than the using a gun feature value.

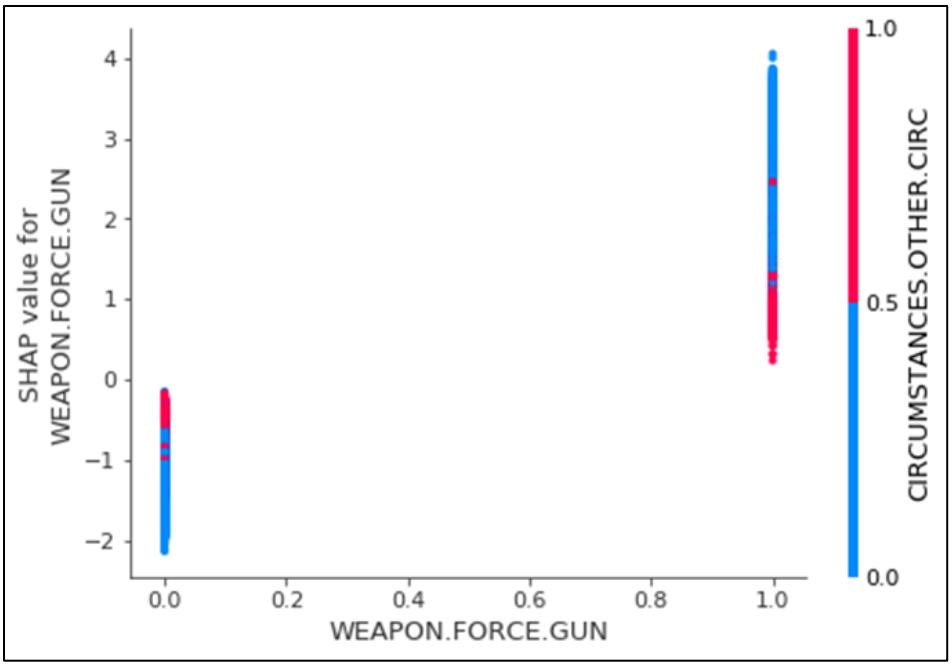


Figure 34: Simple Dependence Plot for Use of a Gun

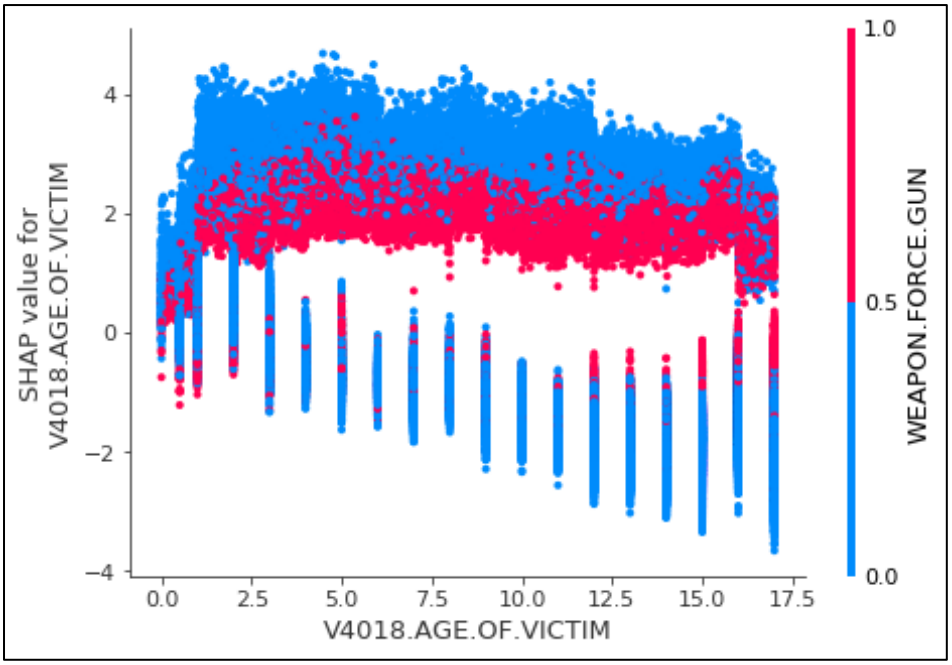


Figure 35: Simple Dependence Plot for Age of Victim and Use of a Gun

Other findings of interests that were derived from XGBoost single dependence plots are described below:

Assault:

- Females tend to assault other females, while males tend to assault other males.
- Assaults in the East-South-Central country tend to involve weapons more so than in other divisions of the country.
- Assault offenses in cities are more likely to involve weapons than for other offense categories.

Sex Offense:

- Female tends to commit sex offenses against males, while males tends to commit sex offenses against female.
- A sex offense is more likely to be committed during late night and early morning hours compared to the other offense categories.
- Teenagers are more vulnerable to sex offenses committed by 20 to 40-year-old offenders than younger children.

Kidnapping/Abduction:

- Kidnapping/Abduction is less likely to involve a weapon or other use of force.
- Older aged females are more likely to commit kidnapping than young females. Younger aged and middle-aged males are more likely to commit kidnapping than older aged males.
- Teenage girls are more vulnerable to kidnapping than younger girls, while younger boys are more vulnerable to kidnapping than teenage boys.

Homicide:

- The use of a gun is more likely in homicides when the offender is between 20 and 40 years of age.
- The "Field, Woods, Lake, and Beach" location type has the highest positive influence on the prediction of homicide.
- The "Sibling" victim-offender relationship type stands out as the highest positive influence on the prediction of homicide.

4.4.2.2 Local Interpretable Model-Agnostic Explanations (LIME)

Local interpretable model-agnostic explanations (LIME) provides methods to explain why an individual prediction was made for a given observation (Sharma 2018). With this method we select one observation from the dataset and examine how it came to make the prediction. Figure 36 shows the results of the LIME procedure for Sex Offense observation number 144. The scale length shows the impact strength. The left most part gives the prediction probabilities for class 0 (not sex offense) and class 1 (sex offense). The middle part specifies a range of feature values that are causing that feature to have its influence. It ranks the most important features from top to bottom. Float point number on the horizontal bars represent the relative importance of these features. Attributes having orange color

support class 1 and those with blue color support class 0. $CIRCUMSTANCES.OTHER.CIRC = 0$ means when this feature's value satisfies this criterion, it supports class 1. The right most part follows the same color coding as the bars in the middle. It contains the actual values of the variable in this observation. The features that contribute most to the result may be different from the overall model.

Figure 37 shows the LIME results for Assault observation number 144, which in this case resulted in a very high prediction probability for the Assault offense category, as opposed to Figure 36 where this observation was highly unlikely to be a Sex Offense.

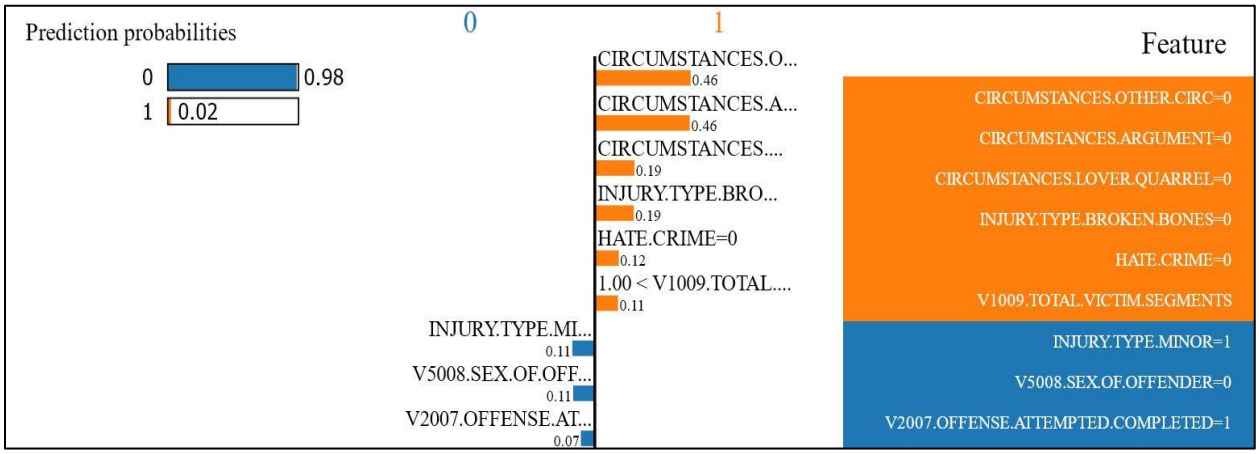


Figure 36: LIME for Sex Offense, Observation 144

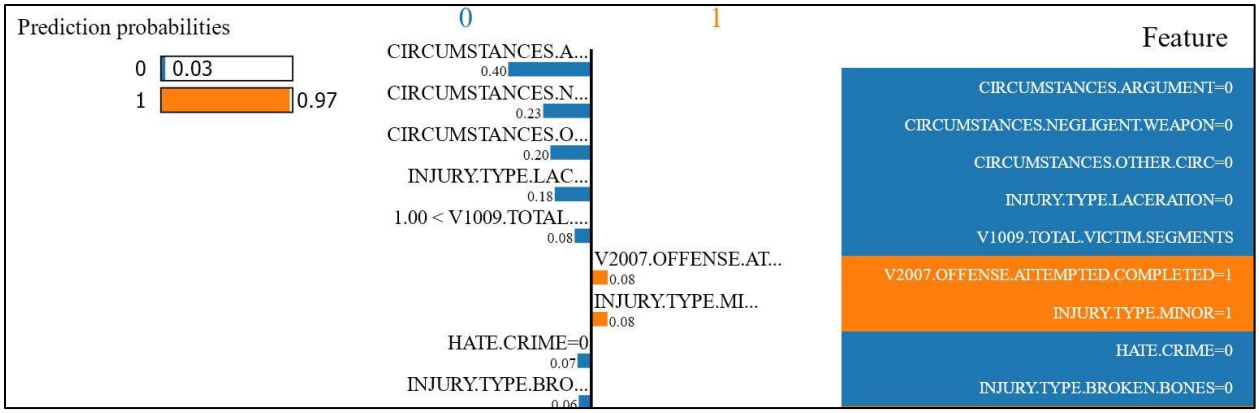


Figure 37: LIME for Assault, Observation 144

4.4.3 Bayesian Network Visualization and Interpretation

Using the dataset, we trained four tree-augmented Bayesian network classifiers; (1) Tree-augmented Bayesian network for the Assault target variable, (2) Tree-augmented Bayesian network for the Homicide target variable, (3) Tree-augmented Bayesian network for the Kidnapping/Abduction target variable, and (4) Tree-augmented Bayesian network for the Sex Offense target variable. The goal is to explore causes and effects based on the conditional probability between nodes.

The Bayesian Networks developed from the dataset are presented in the following sections for each of the target variables. The results and performance for each of the target variables are also explained in each section. A total of 57 fields were shown at first in each model which resulted in an unreadable acyclic graph. Some features were dropped from the visualizations to produce more readable networks and assist with the interpretations of predictions.

4.4.3.1 Target Assault

It can be observed in Figure 38 that the leftmost node (TARGET Assault) has no parent, hence, it is the root node of the tree and all other nodes are pointing away from the root node. This is a typical structure of a TAN model, and as mentioned earlier, the TAN structure allows each predictor to depend on another predictor in addition to the target variable. This can, in some cases, improves the classification accuracy. The model also ranks the variables by importance. The most effective variable for predicting assault was Incident Date Hour. Other important variables include sex of offender, sex of victim, offender relationship, race of offender, and location category. In the acyclic graph, the importance is depicted by darker color. It can be seen on the lower part of the graph in the Figure 38 (middle-lower part), that attribute Sex of Victim has two-parent, TARGET Assault and Sex of Offender. This implies that Sex of Offender and Sex of Victim are among the highly correlated predictors in the dataset for this target variable.

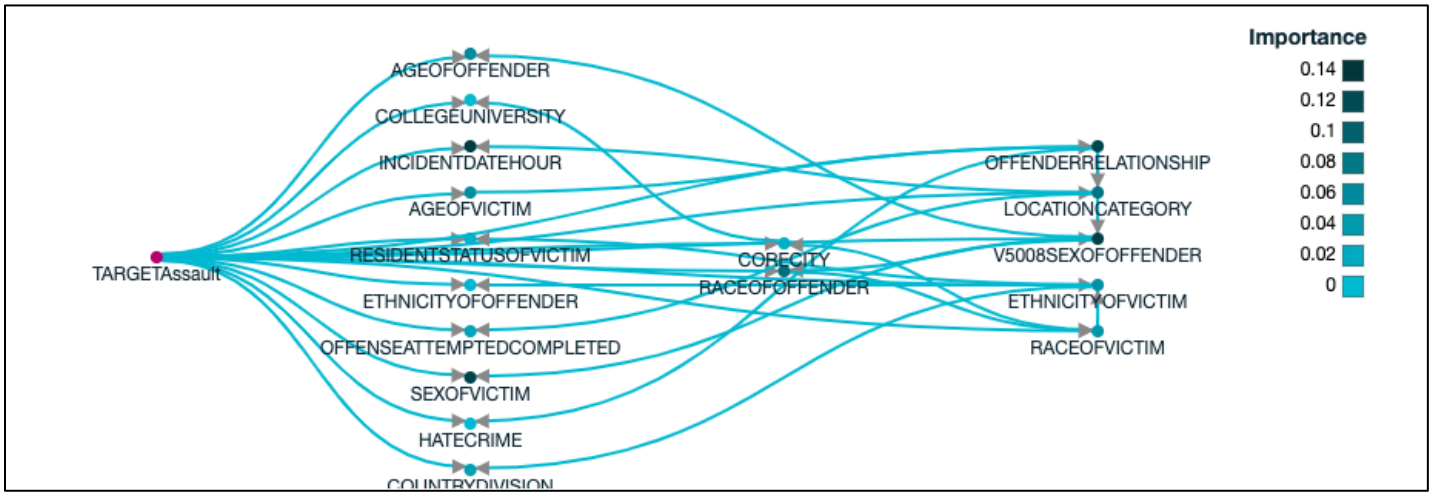


Figure 38: Bayesian Network Graph for Assault

The graph is only half of the story. There is a conditional probability table (CPT) for each node. This information defines the probability distribution used to predict the class probability for any given instance. Table 5 shows the CPT for Sex of Victim given Sex of Offender and the target variable. The UNK value represents the unknown in the dataset (i.e. the sex was unknown). From these results, the belief is that the conditional likelihood of a female assaulting a male is 0.279 and the conditional likelihood of a male assaulting a female is 0.336. Females are more likely to assault other females compared to the male gender. As a reminder, these probabilities describe the likelihood of Assault versus the other three offense categories, and do not account for the fact that a crime may not occur at all.

Table 5: Conditional Probability Table for Sex of Victim - Assault

TARGET Assault	SEX OF OFFENDER	Sex of victim		
		F	M	UNK
1	F	0.717	0.279	0.004
1	M	0.336	0.660	0.004
1	UNK	0.421	0.572	0.007

4.4.3.2 Target Sex Offense

The root node in Figure 39 is the TARGET Sex Offense. Unlike the Assault model, the most effective variable for predicting Sex Offense is Sex of Victim. The other important variables are similar to that of Assault and include Incident Date Hour, Location Category, Sex of Offender, and Country Division.

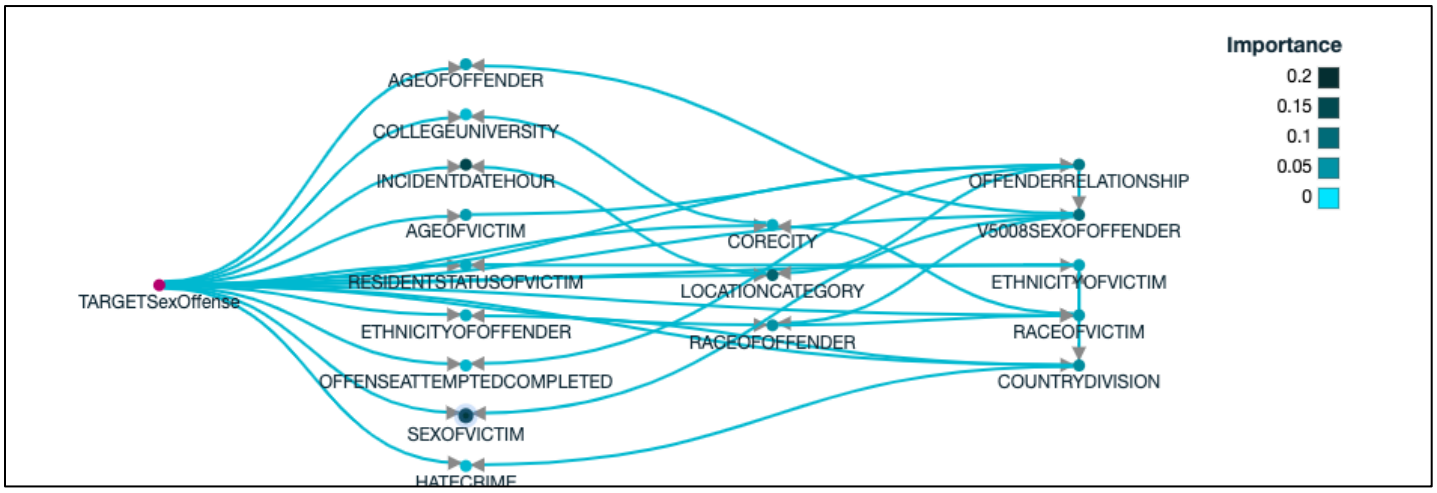


Figure 39: Bayesian Network Graph for Sex Offense

Table 6 shows the conditional probability table for Sex of Victim given Sex of Offender. Here the belief is that the conditional likelihood of a female committing a sex offense (instead of one of the other offense categories) against another female is 0.607 and 0.818 is the conditional likelihood of a male committing sexual offense against a female. This makes sense since most of the sexual offenses committed are against female. Males are less likely to commit a sex offense against another male under these conditions.

Table 6: Conditional Probability Table of Sex of Victim for Sex Offense

TARGET Sex Offense	SEX OF OFFENDER	Sex of Victim		
		F	M	UNK
1	F	0.607	0.391	0.002
1	M	0.818	0.181	0.001
1	UNK	0.838	0.16	0.003

4.4.3.3 Target Homicide

The root node in Figure 40 is the TARGET Homicide. Unlike the other two previously explained target variables, the most effective variable for predicting Homicide is Age of Victim. The other important variables are Age of Offender, Country Division, Incident Date Hour, Location Category, Offender Relationship.

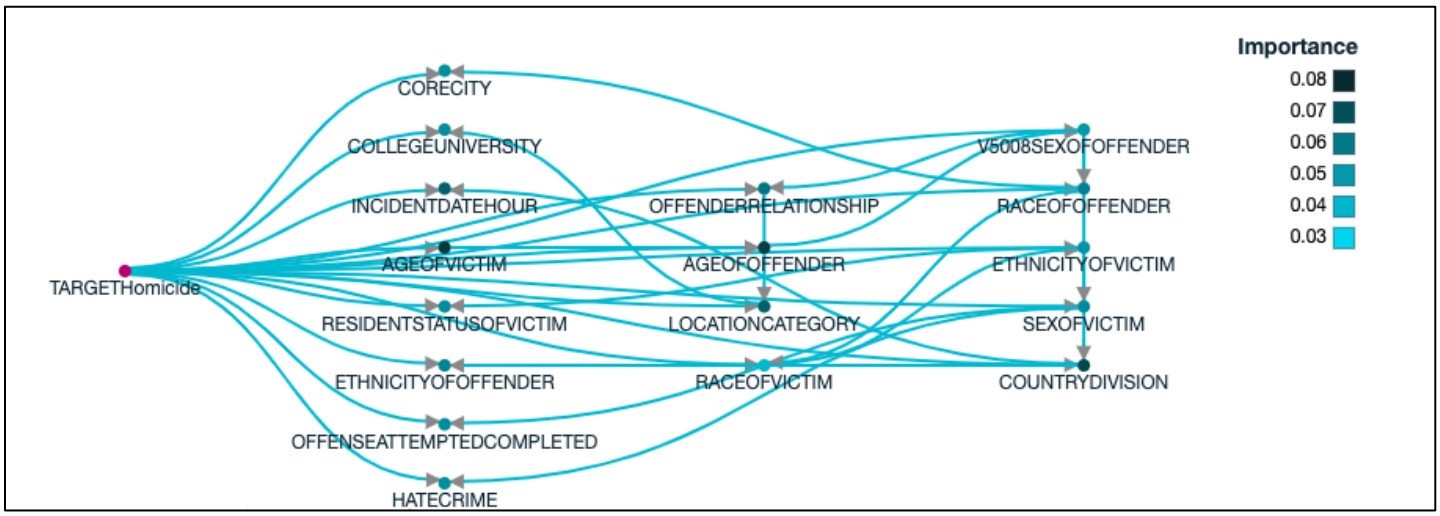


Figure 40: Bayesian Network Graph for Homicide

Table 7 shows the conditional probability table for Age of Victim. This particular node does not depend on any other predictors. It is also the most important feature. The age in the dataset includes children less than a year old. From the results, the belief is that the conditional likelihood of a child over the age of 13.6 becoming a victim of homicide, rather than one of the other offense categories, is 0.303, which is more compared to the other age groups. For example, children between the age of 10.2 to 13.6 are less likely to be a victim of homicide verses the other offense categories.

Table 7: Conditional Probability Table of Age of Victim for Homicide

TARGET Homicide	<= 3.4	3.4 ~ 6.8	6.8 ~ 10.2	10.2 ~ 13.6	> 13.6
1	0.149	0.192	0.177	0.18	0.303

4.4.3.4 Target Kidnapping and Abduction

The graph in Figure 41 represents the network for Kidnapping/Abduction. The root node in the graph is TARGET Kidnapping/Abduction and the most effective variable is the Age of Offender. The other important variables are Offender Relationship, Incident Date Hour, Age of Victim, Country Division, and Location Category.

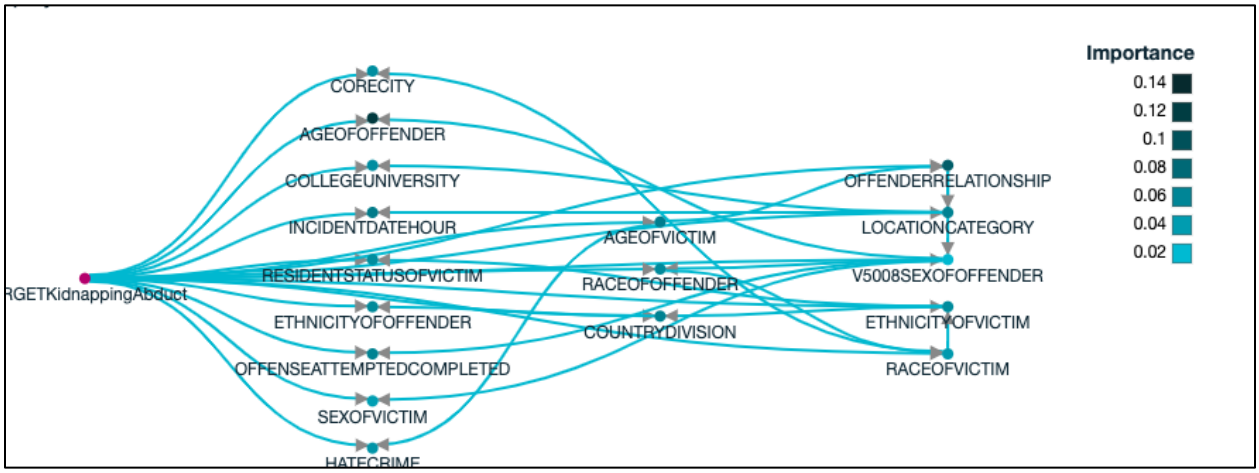


Figure 41: Bayesian Network Graph – Kidnapping/Abduction

Table 8 shows the conditional probability table for Location Category given the offender relationship. In the table, residence/home clearly has the highest conditional probabilities. Some offender relationships that stand out as having high likelihood are Babysitter, Siblings, Step-Parents, and Grand Parents. Also, we can see the conditional likelihood of Victim is Offender is 0.821, indicating that a victim kidnapped themselves. It is believed that this implies a victim ran away from home; however, this could not be confirmed.

Table 8: Conditional Probability Table for Location Category

TARGET Kidnapping / Abduction	OFFENDER RELATIONSHIP	Location Category															
		Business_ Retail_ Restaurant	Construction Site_ Abandoned	Daycare_ Facility	Entertainment	Field_Woods_Lake_ Beach	Government	Healthcare	Highway_Road_Alley _Street_Sidewalk	Hotel_Motel	Park_playground	Parking_DropLot_Gar age	Religious_Building_ Shelter	Residence_Home	School_college_ university	School_elementary_ secondary	Transportation
1	Acquaintance	0.018	0.003	0	0	0.02	0.005	0.001	0.118	0.034	0.011	0.035	0	0.612	0.029	0.049	0.004
1	Babysitter	0	0	0.049	0	0	0	0	0.024	0.024	0	0	0	0.902	0	0	0
1	Boyfriend_Girlfriend	0.017	0	0	0.001	0.004	0.006	0.003	0.131	0.024	0.004	0.037	0	0.688	0.009	0.032	0.004
1	Child_of_Boyfriend_ Girlfriend	0.013	0	0	0	0.013	0.007	0	0.079	0.062	0.007	0.016	0	0.744	0	0.007	0
1	Child	0.021	0.001	0.002	0.001	0.003	0.016	0.001	0.108	0.019	0.003	0.019	0	0.729	0.007	0.023	0.005
1	Common law spouse	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1	Employee	0.222	0	0	0	0	0	0	0	0	0.222	0.111	0	0.444	0	0	0
1	Employer	0.25	0	0	0	0	0	0	0.125	0	0	0	0	0.375	0.125	0	0
1	Friend	0.023	0.006	0	0.003	0.012	0.009	0.006	0.137	0.038	0.015	0.026	0	0.596	0.047	0.049	0
1	Grand Child	0.017	0	0.004	0	0.007	0.007	0	0.037	0.041	0	0.011	0	0.794	0.007	0.021	0.004
1	Grand Parent	0	0	0	0	0	0	0	0	0	0	0	0	0.75	0	0.25	0
1	Homosexual Relationship	0	0	0	0	0	0.125	0	0	0	0	0	0	0.75	0	0.125	0
1	In Law	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1	Neighbor	0	0.017	0	0	0.017	0	0	0.168	0.034	0	0.042	0	0.639	0.008	0	0
1	Other family member	0.016	0	0.003	0.001	0.005	0.011	0.002	0.093	0.025	0.002	0.026	0	0.765	0.006	0.01	0.001
1	Otherwise unknown	0.013	0.001	0	0	0.006	0.013	0.002	0.128	0.177	0.008	0.025	0	0.509	0.021	0.037	0.003
1	Parent	0	0	0	0	0	0	0	0	0	0	0.25	0	0.5	0	0	0
1	Sibling	0	0	0	0	0	0.019	0	0.056	0.006	0	0	0	0.87	0	0.012	0
1	Stepchild	0.016	0	0	0	0	0	0	0.06	0.033	0	0.016	0	0.832	0.005	0.011	0
1	Spouse	0	0	0	0	0	0	0	0	0.182	0	0	0	0.818	0	0	0
1	Step parent	0.133	0	0	0	0	0	0	0	0.067	0	0	0	0.733	0	0	0
1	Step sibling	0.167	0	0	0	0.083	0	0	0.083	0	0	0	0	0.5	0	0.083	0
1	Stranger	0.045	0	0	0.005	0.016	0.005	0.001	0.379	0.029	0.025	0.067	0	0.322	0.007	0.017	0.021
1	Offender	0	0	0	0	0	0	0	0.036	0.036	0.036	0	0	0.821	0	0.036	0
1	Ex-Spouse	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

4.4.3.5 Deploying and Testing the Bayesian Model

A web service deployment was created from the models using IBM Watson, which returns predictions given various input characteristics. More specifically, the interface returns the likelihood of a crime being classified as one of the offense categories given various conditions. For testing, the project team maintained the same conditions for all four target variables in the Bayesian network model to see the difference in the various likelihoods that they returned. The conditions that were specified are listed below:

{

"Fields":["Age of Victim", "Sex of Victim", "Race of Victim", "Ethnicity of Victim", "Resident Status of Victim", "Offender Relationship", "Age of Offender", "Sex of Offender", "Race of Offender", "Ethnicity of Offender", "Location Category", "Country Division"],"values":11, Female, Black, Non-Hispanic or Latino, Resident, Offender is a stranger, 14, Female, Black, Non-Hispanic or Latino, School, South Atlantic"]

}

With the question, "Given a child is believed to be at risk for a crime with certain known conditions, what is the likelihood that the crime will be a Sex Offense, Assault, Kidnapping/Abduction, or Homicide?" Using the listed conditions, the tests on all of the models and their respective results are displayed in Table 9.

Table 9: Deployed Model Test

Offense Type	Likelihood
Likelihood that the crime will be a **Sex Offense**	0.023
Likelihood that the crime will be an **Assault**	0.957
Likelihood that the crime will be a **Homicide**	0.004
Likelihood that the crime will be a **Kidnapping/Abduction**	0.036

In Table 10, we can see that the offense with the highest likelihood is an Assault, with a value close to one (0.957). This makes some sense given the age of the offender and age of the victim is less than 15, and the location category condition given is school. According to Stop Sexual Assault in Schools, an organization created to address sexual harassment/assault and k through 12 students' rights, sexual assault and harassment occurs often among school age children (SSAIS, 2015). A particular value Sex of Offender may enhance the prospect of a particular value Sex of Victim, but it certainly doesn't cause it.

4.4.3.6 Bayesian Modeling Summary

Interpreting machine learning models are very important in every field. Users need to understand the reasons behind predictions. The Bayesian network in this project presents results and performance by providing post crime outcomes/predictions based on the dataset extracted from NIBRS. The Bayesian network models deployed on IBM Watson were tested with different random variables which gave the probability of a victimized child falling into an offense category. Also, we were able to

see the characteristics of crimes against children in four of the offense categories in terms of conditional probabilities. All of this directly addresses the second research question established for this project: "How can crime data be used to predict the likelihood of a victimized child falling into a particular offense category given various known circumstances?"

5 Findings

While the following summary of findings is likely only scratching the surface of what the developed capabilities can reveal from the data, the team did ultimately discover a good deal of information related to crimes against children. Interestingly, the geospatial visualizations did not show poverty to be closely associated with rates of crimes against children. Further research to validate this observation is encouraged. The geospatial visualizations also showed that residences, daycares, and schools are unfortunately among the least safe locations for the Sex Offense and Kidnapping/Abduction offense categories.

With regard to the findings from the unsupervised and supervised modeling methods, the vast majority of conclusions appeared to be consistent across models. For the Assault offense category, the most predictive characteristics found through logistic regression modeling were: offense completed; involves an argument; minor injuries; use of a weapon; college/university location; offender is sibling or victim is offender; and other circumstances. The fact that "other circumstances" was found to be predictive seems to imply that NIBRS needs to add more circumstance codes to better capture the complexities of criminal incidents. Using the hierarchical clustering model for assaults, it was found that females tend to assault other females, causing minor injuries and males have a tendency to assault other males inside the residence/home, using hands & feet as weapons. These conclusions were also noticed in the predictive modeling results.

For the Sex Offense category, the most predictive characteristics found through logistic regression modeling were: known circumstances; does not involve an argument; result in serious injuries (often internal); sex of offender is male or unknown; does not involve weapons; offender is a friend; occurs between 0000 and 0600 hours; and takes place in a residence/home. In addition, using the Bayesian network, it was found that children above age 14 are more likely to be a victim of sex offense compared to those in age groups younger than 14, which is not surprising due to sex offenses being under-reported by those younger in age.

For the Kidnapping/Abduction offense category, the most predictive characteristics found through logistic regression modeling were: serious injuries; location either in hotel/motel or unknown; offender is boyfriend/girlfriend, a stranger, or unknown; involves a gun or cutting weapon; involves multiple offenses; and offense is attempted rather than completed. The Bayesian network showed us that the most common location is the victim's residence/home and the likelihood is higher for certain victim-offender relationships. In the XGBoost model, it was observed that older aged females are more likely to commit kidnapping than young females, while younger and middle-aged males are more likely to kidnap than older aged males. The XGBoost model also revealed that teenage girls are more vulnerable to kidnapping than younger girls, while younger boys are more vulnerable to kidnapping than teenage boys.

For the Homicide offense category, the most predictive characteristics found through logistic regression modeling were: use of a weapon (gun, knife, blunt, vehicle); sibling or unknown relationship;

major injuries; gang-related circumstances; and the field/lake/woods/beach location. Using the SHAP interpretation method on the XGBoost model, it was found that the likelihood of offender age and use of a gun jumps at age 20 and stays high through age 40.

6 Summary

This project set an ambitious goal to discover knowledge about crimes against children through opportunistic data mining of the NIBRS crime data source. The team applied a breadth of approaches, with an emphasis on the use of interpretable data science techniques to identify the characteristics of victims, offenders, and the situational threats in which children are especially vulnerable to particular categories of offenses. The geospatial visualizations developed for this project provide a unique way to explore the vast amount of crime data available in NIBRS. The other analytical methods used included hierarchical clustering, penalized logistic regression, tree-augmented Bayesian networks, and gradient boosting with XGBoost.

Hierarchical clustering of features identified combinations of features sharing similarities across observations to highlight relevant scenarios within the data, such as the assault of males by other males in a residence/home using hands and feet as weapons. This was a somewhat expected finding, which lends some credibility to the other scenarios derived from the hierarchical clustering approach.

The penalized logistic regression models developed for each offense category provided the most accessible level of interpretation. By laying out the resulting coefficients of the models side-by-side in a heat map, the team was able to quickly see the features that had the most positive and negative predictive influence on each of the offense categories. This approach combined with the findings from hierarchical clustering gave the team ideas for where to look for features of interest in the more complex models developed for this project.

The developed XGBoost models were interpreted using SHAP and LIME to gain an understanding of the behaviors of the underlying models with regard to each individual feature. These relatively new techniques allowed the team to derive interesting findings from a fairly advanced machine learning technique that is not often recognized as being interpretable

Constructing tree-augmented Bayesian networks in the IBM Watson Studio SPSS Modeler allowed for the creation of an interface for model deployment and testing where offense characteristics could be entered to calculate the likelihood that a crime belongs to a particular offense category versus others. This approach was not only valuable for the purposes of this project, but could be expanded upon to create a mobile application that would allow law enforcement officers and others entrusted with the care of children to enter known information and obtain a real-time assessment of the types of offenses a particular child is most at risk for victimization.

7 Future Work

This project was conducted using a very large and complex dataset from NIBRS. Although, some features were generated, we believe more features like race difference, age difference between offender and victims can be generated in order to leverage more crime characteristics. In our project, none of the variables in the socio-economic dataset were included as predictors. Adding the socioeconomic data as predictors might also provide a more robust prediction. Geospatial-specific features for predictive analytics is another area that can be explored.

The NIBRS dataset is not limited to the type of crimes that we included in this project, those against children (below the age of 18). The applied techniques can be expanded to crimes against adults, or focusing on a certain gender, the elderly, or hate crimes against specific races to get the characteristics of crimes against these categories of individuals.

For the association rule approach, a network visualization tool (e.g. Gephi) for exploring the association rules could help identify interesting relationships among the variables. Other supervised methods can be attempted to seek better predictive performance. Mutli-class or multi-label modeling could yield more findings. While more data becomes available, more in-depth time series analysis techniques with proper data normalization can also be used to explore and better understand how crimes reported to NIBRS changes overtime. With regard to the Bayesian network models, subject matter expert informed Bayesian belief networks can be explored to see if they perform better and provide more utility than TAN models.

Overall, the team is confident that they addressed the three research questions established for this project. The tools developed for this project allow for the exploration of NIBRS data in truly novel ways. The only remaining recommendation for future work is for others to leverage the tools developed during this project to continue exploring the data and discover knowledge about crimes against children. It is likely that, due to the short timeframe established for this project, there are still many more interesting findings to be discovered in the dataset using these newly developed tools.

Appendix A

Code references

Code files and other files used for analyses can be found on the Blackboard project file exchange.

Appendix B

Project Planning and Risk Section

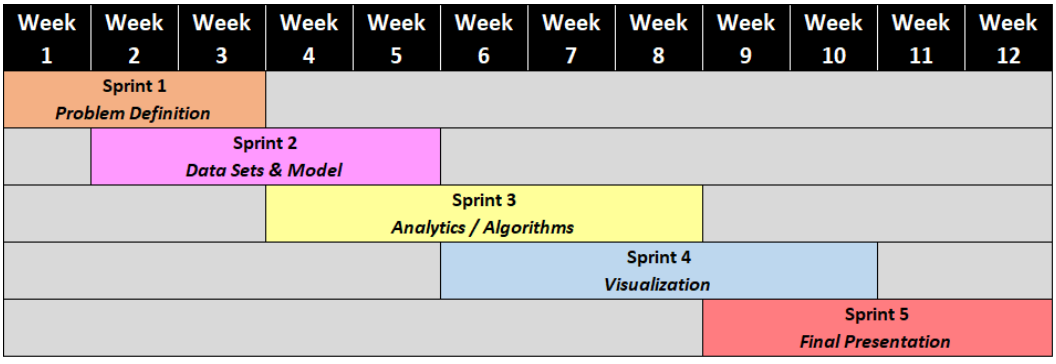
The team collaborated to identify the following risks and mitigation strategies:

Analytic methods fail to yield useful findings: There is believed to be a medium range probability of producing findings that are not useful to stakeholders, which would have a high impact on the project final results. To mitigate this, the team will try many modeling approaches within the time and resource constraints of the project.

Failure to integrate meaningful data: There is a possibility that technical implementation challenges will arise while integrating data from disparate and potentially incompatible sources. There also may be challenges with finding other useful data sources at the state/county-level. This is expected to have a low probability of occurring and a medium range impact on further steps of analysis. To overcome this issue the team is planning to actively pursue other data sources and engineer the features where possible.

Failure to succinctly define veracity: Veracity refers to the biases, noise, and abnormality in data. Is the data that is being mined meaningful to the problem being analyzed? There is believed to be a medium probability of this being an issue in the project with a medium impact on results. To overcome this, the team will invest a good deal of resource in trying to understand the limitations of the NIBRS data source.

Failure to meet deadlines with all deliverables: There is a medium probability of not meeting the deadlines due to the need to create a data set and the wide range of methods that will be applied. This would have a medium impact on the deliverables. To mitigate this, the team will follow the "Parallel Sprints" plan shown below. Under this plan, the team will start the sprints early and run them in parallel with the previous sprints.



Data exceeds processing capabilities: Since the project involves a large amount of data, there is a low probability that processing units will face issues. Since the issues, if any, are expected to be manageable, this is believed to have a low impact on the project deliverables. To mitigate this risk, if necessary, the analytical methods can be applied to a subset of the data.

The team collaborated to evaluate the complexity level of each sprint designed for this project. The resulting team assessment is shown in the figure below. The team believes they have a strong problem definition and project plan. The construction of a project-specific dataset and the integration of various data sources leads to a level of complexity for Sprint 2 that is likely atypical for a DAEN-690 project. While the complexity of the analytics/algorithms that have been proposed are generally well established, the breadth of approaches that will be explored will provide challenges for the team. The team plans to produce complex dynamic visualizations, including geospatial components during Sprint 4. Sprint 5 will present the challenge of capturing the results from a wide range of methods and hopefully a series of impactful findings.

Sprint	Base	Sought	Ultimate
Problem Definition and Project Plans	• Problem (decision) defined	• Project plan developed	• Complexity understood • Risks understood and mitigation planned
Data Sets	• Data sets represented by conventional tools	• At scale data sets	• Multiple data sets integrated at scale
Analytics/ Algorithms	• Algorithms supporting convention analysis	• Algorithms running at scale demonstrating 'big data' techniques	• Algorithms requiring high performance computing resources
Visualizations	• Conventional graphic presentations	• Visualization presenting at scale results	• Visualization dynamically presenting at scale results
Presentations	• Scope includes all value chain components	• Some 'big data' approaches used	• Multiple 'big data' approaches running at scale

Appendix C

Lexicons

Child: For the purposes of this project, a child is defined as a person who is not an adult and who has not attained the age of 18 years.

Crimes Against Children: Crimes against children include physical and emotional abuse; neglect; and exploitation, such as through child pornography or sex trafficking of minors.

Incident: An incident refers to a particular happening, sometimes criminal but always noteworthy.

Offender: An offender is a criminal, someone who breaks the law.

Offense: A breach of a law or rule; an illegal act.

Sex-Offender: A sex offender is a person who's been convicted of certain sex crimes, such as sexual assault or sexual conduct with a minor.

Victim: A victim is a person directly and proximately harmed as a result of the commission of an offense

Appendix D

Additional Field Descriptions for Base Project Data Set

59) **INCIDENT.ID** (Type: string) – This is a unique incident ID created by concatenating V4003 and V4004.
60) **V4003.ORIGINATING.AGENCY.IDENTIFIER.ORI** (Type: string) – *This identifies the agency in which any reported crime data occurred.*
61) **V4004.INCIDENT.NUMBER** (Type: string) – *This is a unique case number assigned to the incident by the agency; it links all segments together for the incident. No other incident submitted by the ORI has this same number.*
62) **V4005.INCIDENT.DATE** (Type: numeric) – *This is the same date value from the Administrative Segment (Format: YYYYMMDD).*
63) **V4002.NUMERIC.STATE.CODE** (Type: numeric) – *This is a two-digit value assigned to the state.*
64) **BH007.CITY.NAME** (Type: string) – *This is the city in which the ORI is associated.*
65) **BH008.STATE.ABBREVIATION** (Type: string) – *This is the state abbreviation. See Numeric State Code for the corresponding numeric state codes as well as the alphabetic codes.*
66) **V4007.UCR.OFFENSE.CODE** (Type: string) – *Each offense committed will always have one or more victims. Multiple offenses that affected this victim will be shown here. Every offense will have a victim or victims, but every victim may not be linked to each offense within an incident.*
67) **V4006.VICTIM.SEQUENCE.NUMBER** (Type: numeric) – *This uniquely identifies a victim and separates one victim from another within a multi-victim Group A Incident Report.*
68) **OFFENDER.SEQ** (Type: numeric) – This field provides a sequence number that links the offenders involved in an incident from the Victim segment to the offender information contained in the Offender segment. The Victim segment contains 10 offender sequence number fields for the 10 possible offenders.
69) **BH019.CURRENT.POPULATION.1** (Type: string) – *This is the population for the agency or the population of the portion of the agency which is located in the county.* (1 of 3)
70) **BH020.UCR.COUNTY.CODE.1** (Type: string) – *This is the UCR county code for the agency or the UCR county code of the portion of the agency which is located in the county.* (1 of 3)
71) **BH023.CURRENT.POPULATION.2** (Type: string) – *This is the population for the agency or the population of the portion of the agency which is located in the county.* (2 of 3)
72) **BH024.UCR.COUNTY.CODE.2** (Type: string) – *This is the UCR county code for the agency or the UCR county code of the portion of the agency which is located in the county.* (2 of 3)
73) **BH027.CURRENT.POPULATION.3** (Type: string) – *This is the population for the agency or the population of the portion of the agency which is located in the county.* (3 of 3)
74) **BH028.UCR.COUNTY.CODE.3** (Type: string) – *This is the UCR county code for the agency or the UCR county code of the portion of the agency which is located in the county.* (3 of 3)
75) **BH040.NUMBER.OF.MONTHS.REPORTED** (Type: numeric) – This is used by UCR to indicate if the agency has submitted data for only a portion of the months during the reporting year.
76) **BH041.MASTER.FILE.YEAR** (Type: numeric) – *This is the year of the NIBRS master file requested.*
77) **BH054.FIPS.COUNTY.1** (Type: string) – *Each relative position of the UCR county code has a corresponding FIPS county code equivalent in the same relative position.* (1 of 3)

78) **BH055.FIPS.COUNTY.2** (Type: string) – *Each relative position of the UCR county code has a corresponding FIPS county code equivalent in the same relative position.* (1 of 3)
79) **BH056.FIPS.COUNTY.3** (Type: string) – *Each relative position of the UCR county code has a corresponding FIPS county code equivalent in the same relative position.* (1 of 3)
80) **INCIDENT.YEAR** (Type: numeric) – This is the year that the incident took place
81) **YEAR.DIFF.FLAG** (Type: numeric) – This binary variable flags a row of data if the reported incident year is different from the master file year in which it is contained.

Appendix E

References

Association of State Uniform Crime Reporting Programs (ASUCRP) "National Incident Based Reporting System (NIBRS) – Association of State Uniform Crime Reporting Programs." n.d. Accessed June 8, 2019. http://www.asucrp.net/uniform-crime-reporting-program/national-incident-based-reporting-system-nibrs/.

Association of State Uniform Crime Reporting Programs (ASUCRP). Uniform Crime Reporting Program – Association of State Uniform Crime Reporting Programs. (2019). Retrieved July 20, 2019, from http://www.asucrp.net/uniform-crime-reporting-program/

Chow, C.K. & C.N. Liu (1968). Approximating discrete probability distributions with dependence trees. IEEE Trans. on Info. Theory, 14, 462- 467.

David Finkelhor, and Anne Shattuck. "Characteristics of Crimes against Juveniles," 2012. https://www.academia.edu/9728770/Characteristics_of_crimes_against_juveniles.

David Finkelhor. "About the Crimes Against Children Research Center (CCRC)." Research. 1999. http://www.unh.edu/ccrc/about-ccrc.html.

Fawcett, Tom. "Learning from Imbalanced Classes. Silicon Valley Data Science." (2016). Retrieved July 8, 2019, from https://www.svds.com/learning-imbalanced-classes/

Finkelhor, D., & Ormrod, R. (2001). Crimes Against Children by Babysitters. PsycEXTRA Dataset. doi:10.1037/e317992004-001

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. Machine learning, 29(2), 131-163.

He, Tong et al. (2019). "Extreme Gradient Boosting." R package. Version 0.90.0.2.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (Eds.). (2013). *An introduction to statistical learning: With applications in R*. New York: Springer.

Kar, Abhishek. "Using K-modes for clustering categorical data." All About Analytics. (2017). Retrieved June 20, 2019, from https://analyticsdefined.com/using-k-modes-clustering-categorical-data/

Kilpatrick, D.C., Edmunds, C., Seymour, A. 1992. "Rape in America: A Report to the Nation" from "The National Women's Study" sponsored by the National Institute of Drug Abuse, National Victim's Center and National Crime Victims Research and Treatment Center at the Medical University of South Carolina. Washington, DC.

Koller, D., & Friedman, N. (2009). Probabilistic graphical models: Principals and techniques MIT Press. ISBN 978-0262013192

Koller, D., Friedman, N., Getoor, L., & Taskar, B. (2007). Graphical models in a nutshell. In L. Getoor, & B. Taskar (Eds.), An introduction to statistical relational learning MIT Press.

Lundberg, Scott. "Documentation by example for shap.dependence_plot." (n.d.). Retrieved July 10, 2019, from https://slundberg.github.io/shap/notebooks/plots/dependence_plot.html

Lundberg, Scott. "Interpretable Machine Learning with XGBoost." Towards Data Science. (2018). Retrieved July 10, 2019, from https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27

Matlab correlation coefficient for non-dichotomous nominal variable and ordinal or numeric variables. Cross Validated. (2013). Retrieved July 3, 2019, from https://stats.stackexchange.com/questions/73065/correlation-coefficient-for-non-dichotomous-nominal-variable-and-ordinal-or-nume

Molnar, C. (2019). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Creative Commons License.

National Archive of Criminal Justice Data (NACJD) | Uniform Crime Reporting Program Resource Guide. (2019). Retrieved May 26, 2019, from https://www.icpsr.umich.edu/icpsrweb/content/NACJD/guides/ucr.html

National Incident-Based Reporting System Data. (n.d.). Retrieved May 26, 2019, from https://www.icpsr.umich.edu/icpsrweb/NACJD/series/128

National Incident-Based Reporting System Resource Guide. (n.d.). Retrieved May 26, 2019, from https://www.icpsr.umich.edu/icpsrweb/NACJD/NIBRS/index.jsp

National Institute of Health, National Cancer Institute. "U.S. Population Data, SEER Population Data." (2019). Retrieved June 16, 2019, from https://seer.cancer.gov/popdata/download.html

Padmanaban, H. (2014). "Comparative Analysis of Naive Bayes and Tree Augmented Naive Bayes Models."

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. (2005). Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Park, E., Chang, H., & Nam, H. S. (2018). A Bayesian Network Model for Predicting Post-stroke Outcomes With Available Risk Factors. Frontiers in Neurology, 9. https://doi.org/10.3389/fneur.2018.00699

Sharma, Abhishek. "Decrypting your Machine Learning model using LIME." Towards Data Science. (2018). Retrieved July 28, 2019, from https://towardsdatascience.com/decrypting-your-machine-learning-model-using-lime-5adc035109b5

Stop Sexual Assault in Schools (SSAIS). (June 10, 2015). "Crisis In Our Schools." Retrieved July 26, 2019, from Stop Sexual Assault in Schools website: https://stopsexualassaultinschools.org/crisis-in-our-schools/

Tableau. Describe Forecast Dialog Box. (2019). Retrieved August 4, 2019, from https://help.tableau.com/v2019.2/pro/desktop/en-us/forecast_describe.htm

Torgo, L. (2011). *Data mining with R: Learning with case studies*. Boca Raton: Chapman & Hall/CRC.

United States Census Bureau. "County Poverty Rate for the United States: 2013-2017." (2018). Retrieved August 3, 2019, from https://www.census.gov/library/visualizations/2018/comm/acs-5yr-poverty-all-counties.html

United States Federal Bureau of Investigation. (2012). UCR Program Changes Definition of Rape. (2012). Retrieved July 20, 2019, from https://www.fbi.gov/services/cjis/cjis-link/ucr-program-changes-definition-of-rape

United States Federal Bureau of Investigation. (2016). "Codebook for Uniform Crime Reporting Program" Data: National Incident-Based ICPSR 37065 Reporting System, [United States], 2016. Inter-university Consortium for Political and Social Research.

Witten I.H., Frank E, Hall MA, Pal CJ. *"*Data Mining: Practical machine learning tools and techniques." Morgan Kaufmann; *2016 Oct 1.* [online] Available at: http://www.sciepub.com/reference/207638

Zychlinski, Shaked "The Search for Categorical Correlation." Towards Data Science. (2018). Retrieved June 18, 2019, from https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9