NHH

Norwegian School of Economics
Bergen, Fall 2021

# Likelihood of Arrests for Violent Crime Incidents in America

*An exploratory study using logistic regression and random forest methods*

**Mayank Shukla**

**Supervisor: Dr. Evelina Gavrilova-Zoutman**

Master thesis, Economics and Business Administration

Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

This thesis was written as a part of the Master of Science in Economics and Business Administration at NHH. Please note that neither the institution nor the examiners are responsible – through the approval of this thesis – for the theories and methods used, or results and conclusions drawn in this work.

# Acknowledgements

Despite the challenges of a virtual learning environment, I am proud to present this thesis as my academic chapter comes to a close during a global pandemic. I would like to thank the staff and administration of the Norwegian School of Economics, and Ivey Business School for being flexible and making it possible to pursue a dual degree. A sincere thank you to Dr. Evelina Gavrilova-Zoutman and their patience while guiding my interest and research approach in this topic. This project would not be possible without my friends and colleagues for reviewing this body of work and motivating me to continue on. I would like to acknowledge my family for giving me encouragement to pursue my goals.

Norwegian School of Economics

Bergen, December 2021

Mayank Shukla

# Abstract

The use of policing algorithms to predict for arrest is rising in America. However, research indicates that these algorithms may be biased against certain populations. These false perceptions of who commits these crimes, and who is impacted by them is also skewed by the media. Hence, it is important to understand which demographic and situational characteristics of a violent crime incident impact the likelihood of arrest. In this thesis, I will predict for arrest in incidents of violent crime as reported in the National Incident-Based Reporting System 2014. The outcome of arrest was predicted using two types of classification methods, logistic regression and random forest. The models that were built for the aggregate of all violent crime, as well as the subsets of offense types had a good predictive power with an accuracy of greater than 50%. Additionally, adjusted models were built to address class imbalance and leveraged cross-validation methods. Using odds ratios from the logistic regression results, and the variable importance plots from the random forest - likelihood of arrest was ascertained. The results indicate that generally the likelihood of arrest increases under certain conditions. These conditions are: in incidents where the race of the offender is white, in incidents where the race of the victim is white, in incidents where the offender is a female (for aggravated assault instances), and in incidents where if the victim of a violent crime is a female. Generally, the likelihood of arrest decreases as the age of the offender increases, and the likelihood of arrest increases as the age of the victim increases. The likelihood of arrest decreases for incidents where the offender is armed with a deadly weapon, and where the offender and victim are strangers. Additionally, the likelihood of arrest increases for all violent crimes if the incident takes place at night time compared to day time, and in incidents where the offender is using substances. The results show that media perceptions, and predictive policing algorithms are skewed. These typically represent black individuals as more dangerous more likely to be incarcerated than white offenders. However, the results from this thesis show the converse relationship. Additionally, this thesis also shows that variables such as time of day, substance use, and the age of the victim and offender can be leveraged to make more powerful predictions on the likelihood of arrest.

***Keywords*** – arrests, FBI, NIBRS, classification models, logistic regression, random forest

# Contents

# List of Figures

# List of Tables

# 1 Introduction

Violent crime in America saw a peak in the 1990's and has been declining since then. However, polling by the Center from American Progress in 2017 showed that 88 percent of survey respondents regarded violent crime on the national level as either a “major problem” or an “immediate crisis" (Sun, 2018). Despite the downward trend in national violent crime rates, Americans still seem to perceive that crime is up (Gramlich, 2020). It can be challenging to draw out objective conclusions about violent crime incidents since they are often the target of public speculation and frenzy (Horton, 2008). These misconceptions about crime rates usually draw contentious debates within the country. The public often starts questioning the demographic characteristics like the offender's or other situational factors of the incident.

An example is that news reporting of violent crime over-represents black men in particular as perpetrators of crime and more threatening than white men (Sun, 2018). It is crucial to study the association of critical demographic and situational characteristics of the incident and their likelihood of arrest, rather than rely on biased notions and media presentations. My research objective is to evaluate which demographic and situational characteristics of a violent crime incident contribute to a higher likelihood of arrest.

I will explore this research area using the most comprehensive crime incident dataset in the United States called the National Incident-Based Reporting System (NIBRS) from 2014. The Federal Bureau of Investigation (FBI) maintains the NIBRS, and it can be used to provide statistically sound conclusions about violent crime characteristics. Setting this up as a classification problem, I will predict the outcome of arrest (1) and no arrest (0) for reported violent crime incidents using the characteristics of the crime, offenders, and victims as predictors. As with some research articles in this field, I focus on violent crime incidents with a single offender and a single victim. In these incidents, the victim can get some indication of the offender's demographic characteristics. I posit the question, what features of the reported incidents of violent crime lead to a greater likelihood of an arrest.

I will be applying two classification methods: logistic regression and random forest, to test which characteristics of the violent crime incident increased the likelihood of arrest. I comment on the descriptive statistics and how they compare to the literature. After a discussion on methodology, I present and compare the results of the base model (all violent crime) and subset models by offense types from the logistic regression and random forest classifiers.

I will report results for predicting arrest in all violent crimes and the subsets of violent crime offenses at the base case. Then the adjusted model will be implemented using concepts from class imbalance and cross-validation. I will use metrics such as sensitivity, specificity, accuracy, and area under the curve (AUC) to compare the adjusted and base case for each model. The results section will conclude with a discussion of which classification method was best suited for predicting each type of violent crime offense.

I will explain the implications of the classification models in the discussion section. Using the log odds from logistic regression and variable importance from random forests, I will compare the impact of predictive variables on the arrest outcome to what is observed in the literature review.

Finally, contributions from my work will shed light empirically on the impact of demographic and situational variables of violent crimes on the outcome of the arrest. This paper will describe which population groups or factors involved in the commission of the crime would lead to higher arrests than others. Lastly, there will be a discussion on the limitations of this study.

# 2 Background

## 2.1 Violent Crime - Definition and Trends

The FBI defines *violent crimes* as offenses that involve force or threat of force against a victim (FBI, 2019). This encompasses both crimes in which violence is the objective, such as murder, as well as crimes where violence is the means to an end, such as a robbery (FBI, 2019). In 2019, the FBI reported a total of 379 violent crimes per 100,000 people (Gramlich, 2020). Additionally, violent crimes may be committed with and without weapons (FBI, 2020). In the United States, violent crime can be composed of four offenses: i) murder and non negligent manslaughter, ii) rape, iii) robbery, and iv) aggravated assault (FBI, 2020). Among violent crimes, aggravated assault was the most common offense, followed by robbery, rape, and murder/non-negligent manslaughter (Gramlich, 2020).

Looking at historical trends, the peak of violent crimes was in 1990's where approximately 758.2 offenses took place nationally per 100,000 population. Fortunately, violent crimes have dropped by 49% between 1993 and 2019 (Gramlich, 2020). FBI Statistics from 2019 indicated that there was an estimate of 366.7 instances of violent crime per 100,000 inhabitants in the United States (FBI, 2019). As mentioned previously, despite the sharp downward trend in national violent crime rates, Americans still perceive crime is up (Pew Research, 2020).

Why public views on crime have grown more negative is unclear, though many point the blame to the 24 hour cycle news coverage and political rhetoric. (Baer, 2016). In news media, the saying goes, "if it bleeds, it leads". This inaccurate perception of the trends in crime has a disproportionate impact on communities of colour (Ghandnoosh, 2014). Reporting shows that the mainstream news exaggerate rates of black individuals offending, white victimization and depict black suspects in a less favorable light than whites (Ghandnoosh, 2014). Studies also indicate that Black and Hispanic individuals are more likely to be stopped by the police (Coviello and Persico, 2016), more likely to be incarcerated compared with White persons (Binswanger et al., 2012) and police interactions among racial minorities is more likely to result in arrest (Kochel et al., 2011).

## 2.2 Predictive Algorithms for Policing

There has been a focus on applying data mining methods in crime analysis in the last decade (Sun et al., 2014). This is due to the vast amount of criminal data that law enforcement agencies around the country are accumulating. The FBI's NIBRS data set collects variables pertinent in the predictive model building process to combat and prevent crime (Sun et al., 2014). In the United States, law enforcement agencies in California, Washington, South Carolina, Alabama, Arizona, Tennessee, New York, and Illinois have all incorporated predictive algorithms for crime detection and classification (Friend, 2013). The Santa Cruz Police Department in California leveraged verified crime data to predict future offenses in specific locations in one such use case. By strategically placing police paroles in these hot spot areas, they saw a reduction in robberies by 19 percent in 6 months after predictive algorithm was put into effect (Friend, 2013). While this is positive news, these algorithms' accuracy and predictive power have been questioned.

The RAND Corporation, a non-profit think tank, published a report in which no statistical evidence was found that crime was reduced when predictive policing was implemented (Perry, 2013). The limitation with these predictive policing algorithms is that they are only as good as the data fed in (Patel, 2015). Dissidents of predictive policing also liken the practice to racial profiling. Take, for instance the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), an algorithm and decision support tool used across the country. The tool assigns risk assessment scores to offenders based on their characteristics (race, sex, age) and offense type. These scores determine factors like bail amount and criminal sentencing when the offender is charged and found guilty. Propublica has found that using COMPAS, black individuals are almost twice as likely as white individuals to be labeled a higher risk but not re-offend in the future. While the algorithm assigned white individuals with lower risk scores, they were more likely than black individuals to commit other crimes. There are many more examples of racial bias in predictive models for crime but they fall outside the scope of this paper. Due to the contentious debate they invite and inaccurate predictions, early adaptors of predictive policing tools such as the Santa Cruz PD (mentioned earlier) have announced recently that they would no longer be using them (Heaven, 2021)

# 3 Literature Review

This section will introduce relevant literature about characteristics such as race, sex and age of victims and offenders, and arrest probability. The literature presented here would serve as the basis for some of the variables used to predict the likelihood of arrest for violent crime incidents.

## 3.1 Race and Probability of Arrest

The first seminal large-scale quantitative paper on race and probability of arrests for violent crime was conducted by Michael Hindelang (1978). Hindelang compared race-specific arrest data from the Uniform Crime Records (UCR) against offender data from the National Crime Victimization Study (NCVS) (Hindelang, 1978). The UCR was the best available source at the time to collect information about race, sex and age of offenders. Whereas the NCVS reported data on the race of the offender as determined by the victim of the crime. Hindelang's goal was to observe if there were substantial differences in the two datasets on the race of the offender for specific violent crimes. The study showed that black individuals were overrepresented by about 10 percentage points in the UCR arrest data for crimes of rape, robbery, and assault (Hindelang, 1978). While this might suggest that black offenders were likely profiled and meant they were arrested at a higher rate, the author concluded this was not the case. In fact, Hindelang stated that the disparity is due to crimes involving black offenders were less apt to be reported to police than crimes involving white offenders. Fundamentally, while Hindelang's work has been an inspiration for researchers in the area, one major criticism of Hindelang's work is whether or not the UCR and NCVS are measuring the same outcome. This is due to a considerable amount of studies that show there were sizable differences in relative crime levels reported in the UCR and NCVS (Booth, Johnson & Choldin 1977; O'Brien 1983; O'Brien, Shichor & Decker 1980). Another criticism is that both UCR and NCVS are aggregate-level data. One cannot conclude that black individuals were more likely than white to be arrested for similar violent crime infractions.

Building on Hindelang's work, D'Alessio and Stolzenberg (2003) use the NIBRS 1999 dataset to determine the impact of race on the probability of arrest for 335,619 incidents of violent crime across the country. The study looked at four types of violent offenses-rape, robbery, aggravated, and simple assault; in these crimes, the victim is confronted by the offender (D'Alessio and Stolzenberg, 2003).. Hence, the victim can get some indication of the offender's demographic characteristics. Using multivariate logistic regression, the authors determined that the odds of arrest for white offenders is approximately 22% higher for robbery, 13% higher for aggravated assault, and 9% higher for simple assault than they are for black offenders. Lastly, the race of the offender played no role in the probability of arrest for the crime of rape. The advantage of the NIBRS data over Hindelang's study is that a reported crime incident can be linked to the subsequent arrest; which is not possible with the UCR or the NCVS. Additionally, the study went beyond the Hindelang study. It incorporated the victim/offender relationship, time and place of occurrence, weapon use, and victim injury in the model to calculate the probability of arrest.

## 3.2 Sex and Probability of Arrest

Stolzenberg and D'Alessio (2004) extend their research and look into the relationship between sex differences in the likelihood of arrest. Their literature finds that women are arrested at a much lower rate than men. While women account for almost 51% of the national population, they represent only about 12 percent of the arrests for violent crimes in 2000 Stolzenberg and D'Alessio (2004). The authors note, however, that most previous studies rely on observational data and hence is unreliable Stolzenberg and D'Alessio (2004). Using a similar approach from their previous research on race and probability of arrest, they analyzed the impact of an offender's sex for 555,752 incidents of kidnapping, sexual assault, aggravated assault, simple assault, and intimidation in 19 states using the NIBRS 2000 dataset (Stolzenberg and D'Alessio, 2004). They explain that they limited their observations to types of crime where the one offender, comes into contact with one victim. This would make identifying important demographic variables of the offender by the victim possible. Using logistic regression modeling, they showed that the probability of arrest for females was 28% lower for kidnapping, 48% lower for fondling, 9% lower for simple assault, and 27% lower for intimidation than for males. The authors undertook a further supplementary investigation to determine if the likelihood of arrest for Black females was higher than that of White females. They specifically looked at the interaction of gender and race on the likelihood of arrest and found race conditioned the relationship between gender and the likelihood of arrest for simple assault and aggravated assault (Stolzenberg and D'Alessio, 2004).

## 3.3 Characteristics of Age and Violent Crimes

Literature examining the relationship of age and likelihood of arrests for violent crimes is hard to come by. However, there was a body of work looking at the characteristics of age and in relation to violent crimes that should be highlighted.

The paper was presented by the Bureau of Justice Statistics in 1997, looking into age patterns of victims of serious violent crime (Perkins, 1997). Violent crime victimization rate increases during the teenage years, and peaks at the age of 20, then steadily reduces as an individual gets older. The paper finds that victims between ages 12 and 24, which represent a fourth of the national population account for almost half of all serious violent crime (Perkins, 1997).This pattern was observed by the authors, across all race, sex and ethnic groupings, with some exception. Between the years, 1992 to 1994, about 1 in every 2 persons who reported an aggravated assault was younger than 25 (Perkins, 1997). Looking at the race grouping for aggravated assault, Black and Hispanics individuals which represent 20% of the general population were about 28% of aggravated assault victims. In instances of robbery, it was observed that half of all robbery victims were age 26 or younger (Perkins, 1997). Additionally, Black and Hispanic individuals under the age of 22 had robbery rates about twice as those for white individuals. Lastly, for instances of rape or sexual assault, a little more than 1/5 of all victims were aged 18 to 21 with the average age of victims being 27.

# 4 Data

In the data section, I will be covering briefly the history of crime data collection in the United States and introduce the NIBRS dataset. The 2014 NIBRS dataset was provided by my supervisor, Dr. Evelina Gavrilova-Zoutman. Next, I will explain how I linked the various segments of NIBRS to build the violent crime incidents dataset. Lastly, I will describe how the variables were selected, and manipulated as input for the classification model. This section will conclude on exploratory data analysis, looking at proportions that were derived from these variables.

## 4.1 Introduction to NIBRS

Before the 1930s, law enforcement agencies had been individually collecting summary counts of crime data (Maltz, 1999). The methods of collecting this data had varied from one agency to another which did not allow for comparisons or aggregation of crime statistics at state or national level. This was problematic, because even back then newspapers were manufacturing supposed “crime waves” out of thin air (Maltz, 1999). To address this issue, the International Association of Chiefs of Police (IACP) and 400 cities from 43 states representing 20 million people, began participating in the Uniform Crime Record (UCR) reporting system. The UCR fulfilled this need by providing useful statistics such as compiled counts of offenses, clearances, and arrests (Maltz, 1999).

By 1982, data collection became outdated and there were limitations with the UCR such as the types of crimes it was able to collect (Strom and Smith, 2017). Hence, to bring crime reporting into the "21st century", the FBI working with the Bureau of Justice Statistics presented the National Incident-Based Reporting System (NIBRS) to compile aggregate level crime data and statistics across the United States (Strom and Smith, 2017). Unlike the UCR system that collects data on only eight types of crimes, NIBRS collects 24 crime categories made up of 52 specific crimes called Group A offenses. The NIBRS dataset is unique as it captures details on crime incidents reported to law enforcement agencies participating in the program. The details are broken up over six types of data segments: administrative, offense, victim, property, offender, and arrestee (Strom and Smith, 2017). The advantage of working with this dataset is that there is a standard set of definitions for criminal offenses across jurisdictions. The NIBRS also provides information about the demographics of victims and offender, and collects details about the circumstances of each incident (Strom and Smith, 2017). Another significant advantage of the NIBRS is that the data is collected annually, meaning that it is a consistent source for studying crime trends over time. One disadvantage of the NIBRS is that it is not representative of all crime that takes place in America. This is primarily due to the fact that reporting to the FBI by local agencies is voluntary. In Figure A0.1, the states highlighted in red were represented in the NIBRS 2014 (FBI, 2014). This does not mean that all local agencies within the state participated by sending data to the FBI. In fact, of the 18,489 agencies in the country only about 6520 agencies sent in their data to the NIBRS 2014. The NIBRS 2014 is only representative of 93,330,000 individuals, where the population of America exceeds 318 million individuals (FBI, 2014).

### 4.1.1 Data Requirement & Linking Segments

First I need to discuss what the data requirements are in order to get our dataset for analysis. The first requirement is to gather incidents where there is one victim and one offender. I chose to follow a similar approach as (D'Alessio and Stolzenberg, 2003). The research question is interested in crimes where the victim and offender come into direct contact. I will exclude murder from our criteria for violent crime since we are interested in identification which may be collected by the victim when the incident is reported.

In the NIBRS 2014, there are 6,520 ORIs reporting across 38 states which cover a total population of 93,330,000 individuals or approximately 30% of the population in the United States. (Image of MAP). After looking at the data segments, I selected six segments: batch, administrative, offense, victim, offender, and arrestee. While these segments report various levels of detailed information, there are a three fields that are of interest for linking. These fields are Originating Agency Identifier (ORI), Incident Number, and the Incident Date. The ORI is a unique nine character identifier assigned to a agency law enforcement agency. Each Incident Number represents a unique crime instance under that Originating Agency Identifier. By merging various segments on these fields we are able to string together data for a crime incident from various NIBRS segments.

First I filtered for incidents where there was one offender, one victim, one offense and one arrest. It is a necessary data filtering process as an incident with multiple offenders and victims would make it difficult to estimate the probability of arrest based on demographic details. An incident with multiple offenses would make it difficult to pinpoint probability of arrest to a particular violent crime. Arrest was determined based on the presence of a record with the incident number and matching ORI in the arrestee segment. Next, it is necessary for me to select the offense types that fall into violent crime. Based on literature, I selected 4 categories of violent crime to predict arrest upon based on the data requirements described previously. These categories are: aggravated assault, forcible sexual offense, robbery and simple assault. After these considerations are made, the final dataset for analysis contains 745,382 incidents of violent crimes that were reported to police. This includes a total of 119,952 aggravated assaults, 63,509 forcible sexual offenses, 38,543 robberies and 523,378 simple assaults.

## 4.2 Variable Selection

In this subsection, I explain how I constructed dummy variables for all predictor variables and the target variable.

### 4.2.1 Predictors

The main demographic predictors the study is interested in is race, sex and age of the victims and offenders. Race falls into four categories: Black, White, American Indian/Native or Asian/Pacific Islander in the NIBRS. The dummy variables I created to satisfy a binary outcome are "Offender Black" and "Victim Black", where 1 represents a black offender or victim and 0 represents a non-black victim or offender. Sex was captured in a similar method; the NIBRS collects information on the sex of the offender and victim which fall into two categories: male or female. I created two dummy variables called "Victim Male" and "Offender Male" to discern if the victim or offender is a male (1) or female (0) Lastly, age of the offender and victim is also collected as a numeric variable. I cleaned the age variable by excluding those aged 1 year or less and those aged 98 or above.

Other predictors of interest are whether or not the offender was a stranger, if the offender had a deadly weapon and if the violent crime incident took place in the daytime or nighttime. I will refer to literature in building out these dummy variables as it is important to transform them into binary variables. Victim-offender relationship is captured in the NIBRS and falls into 3 categories: family, acquaintance or unknown/stranger. Yang and Olafsson (2011), suggest using a binary victim–offender relationship such as stranger vs. known. In their work, which is set up as a classification problem similar to mine, they use predictors to try to predict for the victim-offender. They argue that it would be hard to attempt a classification model when their target variable is divided into three categories. Their solution was to use "Stranger vs Non-Stranger". Hence, for my paper, I constructed a binary dummy factor "Relationship" which breaks down into "Stranger", if the victim does not know the offender or "Known", where the victim knows the offender.

Definitions of what qualifies as a deadly weapon used by the offender vary from state to state, which makes it difficult to transform this variable as a dummy variable. The (School, 2021), defines a deadly weapon as "an object, instrument, substance, or device which is intended to be used in a way that is likely to cause death, or with which death can be easily and readily produced", which casts a wide net as to the what fits into my criteria. To make things simple, I define a deadly weapon as any firearm (regardless of type), knifes or cutting instruments, blunt objects, motor vehicles, explosives or fire. The other personal weapons that are used by an offender, like hands, feet, teeth are not captured as deadly weapons in my definition. I coded this dummy variable as "Deadly Weapon", where 1 indicates that the offender was armed with a deadly weapon and 0 in instances where there was no weapon or no deadly weapon involved.

Another variable of interest that could help in the classification of arrest, given testimony from the victim about the offenders characteristics is time of day. If the violent crime incident takes place in the daytime, there could be important details captured versus if the incident takes place at night. The NIBRS collects date-time information for when the incident took place based on the ORI. Looking at literature, Nix et al. (2019), study the danger to a police officer responding to a domestic incident compared to a non-domestic incident while controlling for other potentially important variables. They leverage the date-time stamp on the incident from the NIBRS to control for Day time / Night time. Using their approach, I will control for time of day with a dummy variable called "Time Of Day", where the evening and night hours are 6PM to 5:59 AM, and daytime as 6 AM to 5:59 PM.

### 4.2.2 Target Variable

The target variable is the presence of an arrest in the arrest segment that matches the ORI and incident number from other linked segments. I built a dummy variable labelled "ArrestYN", that captures arrests as 1 and no arrests as 0.

## 4.3 Exploratory Data Analysis

To gain initial insights into the violent crime dataset consider the relevant summary statistics in Table 4.1. It is observed that in a majority of instances of forcible sexual assault (19.1%) and robbery (18.2%), an offender is arrested. For aggravated (54.0%) and simple assault (53.9%), the rate of arrest of an offender is higher. In the race groupings of offenders, it is observed that black offenders are the minority in instances of aggravated assault (35.4%), forcible sexual offense (20.7%), robbery (44.4%) and simple assault (32.4%). Black victims are also the minority in incidents of aggravated assault (31.3%), forcible sexual offense (16.1%), robbery (25.1%) and simple assault (25.9%). Male offenders represent the majority in all violent crime incidents especially forcible sexual assault (87.0%). Whereas the majority of victims are female in all types of violent crime, except for robbery where male victims represent 45.1% of the incidents.

When looking at substance use among offenders, it is observed that in a majority of instances, the offender is not under the influence of substances. Offenders who commit aggravated assault (15.1%), or simple assault (15.0%) were more likely to be using substances than offenders who commit forcible sexual offenses (9.3%) or robbery (3.7%). The use of a deadly weapon as defined by my criteria was seen in the majority of incidents of aggravated assault (57.9%). Deadly weapons were also more likely to be seen in instances of robbery (46.1%), however in instances of forcible sexual offenses (0.1%) and simple assault (0.0%), deadly weapons were rarely observed. This means that my criteria in defining what a deadly weapon is was logical, offenders who commit assault without a deadly weapon are rarely ever charged with aggravated assault. A notable observation when looking at the mean age of victims shows that victims are most likely to be younger in instances of forcible sexual offenses (19 years old), as compared to other violent crimes. The mean age of offenders show no notable findings. The time of the incident gives me a rough clue if the incident took place in the daytime or night time. In the majority of instances of violent crime, the victim reports that the incident took place during the day-time. However as compared to other types of offenses, robbery (38.7%) was most likely to take place during the night-time. The victim also reported in a majority of incidents of aggravated assault (75.3%) , forcible sexual offense (77.7%), and simple assault (84.5%), that the offender was not a stranger. For cases of robbery however, the responses were evenly distributed in that the offender either claims that the offender was a stranger (37.9%), a known person (37.5%), or is unable to determine one way or another (24.6%).

**Table 4.1:** Percentage Distributions of Characteristics of Crimes, Offenders and Victims by Type of Violent Crime, NIBRS 2014

| | Aggravated Assault | Sexual Offense | Robbery | Simple Assault |
|---|---|---|---|---|
| | **N = 119,952** | **N = 63,509** | **N = 38,543** | **N = 523,378** |
| Offender Arrested | | | | |
| 0 = No | 54 | 80.9 | 81.1 | 53.9 |
| 1 = Yes | 46 | 19.1 | 18.2 | 46.1 |
| Offender Black | | | | |
| 0 = No | 52 | 64.8 | 32 | 61.2 |
| 1 = Yes | 35.4 | 20.7 | 44.4 | 32.4 |
| Missing Data | 12.6 | 14.5 | 23.5 | 6.3 |
| Victim Black | | | | |
| 0 = No | 55.9 | 73.7 | 43.8 | 65.5 |
| 1 = Yes | 31.3 | 16.1 | 25.1 | 25.9 |
| Missing Data | 12.8 | 10.2 | 31.1 | 8.6 |
| Offender Male | | | | |
| 0 = No | 20.2 | 0.4 | 5.4 | 26.3 |
| 1 = Yes | 69.3 | 87 | 65.7 | 69.3 |
| Missing Data | 10.5 | 8.8 | 31.1 | 4.4 |
| Victim Male | | | | |
| 0 = No | 40.5 | 80.1 | 23.8 | 58.7 |
| 1 = Yes | 48.6 | 13.3 | 45.1 | 34.7 |
| Missing Data | 10.9 | 5.7 | 31.1 | 6.6 |
| Offender Stranger | | | | |
| 0 = No | 75.3 | 77.7 | 37.5 | 84.5 |
| 1 = Yes | 18.8 | 16 | 37.9 | 12.4 |
| Missing Data | 5.9 | 6.3 | 24.6 | 3.1 |
| Offender Substance Use | | | | |
| 0 = No | 84.9 | 90.4 | 96.3 | 85 |
| 1 = Yes | 15.1 | 9.3 | 3.7 | 15 |
| Missing Data | 0 | 0.3 | 0 | 0 |
| Deadly Weapon | | | | |
| 0 = No | 38.7 | 94.8 | 46.3 | 97.8 |
| 1 = Yes | 57.9 | 0.1 | 46.1 | 0.0 |
| Missing Data | 3.4 | 5.7 | 7.6 | 2.2 |
| Offender's Age | | | | |
| Mean years | 34 | 31 | 29 | 33 |
| Missing Data | 14 | 14 | 42 | 6 |
| Victim's Age | | | | |
| Mean years | 34 | 19 | 35 | 33 |
| Missing Data | 12 | 6 | 31 | 6 |
| Time of Day | | | | |
| Day-time | 62.1 | 71.7 | 59.8 | 62.9 |
| Night-time | 35.2 | 21.9 | 38.7 | 34.9 |
| Missing Data | 2.7 | 6.4 | 1.6 | 2.2 |
| Relationship | | | | |
| Known | 75.3 | 77.8 | 37.5 | 84.5 |
| Stranger | 18.8 | 16.0 | 37.9 | 12.4 |
| Missing Data | 5.9 | 6.3 | 24.6 | 3.1 |

# 5 Methodology

The methodology section will be broken up into two parts. First, I describe the two classification algorithms, logistic regression and random forest. Then, I explain why they are ideal to use for this research question. The second part of this section focuses on the methods that will be used to evaluate and validate the output of the models.

## 5.1 Classification Algorithms

The exercise of predicting qualitative responses is known as classification (James, Witten, Hastie & Tibshirani, 2013). The research question of interest is predicting whether an violent crime incident would lead to an arrest. Additionally, our response variable *Arrest* (Y) is a dichotomous variable (5.1) , where it falls into one of the two categories: *no arrest (0)* or *arrest (1)*.

$$Y = \begin{cases} 0 & \text{no arrest} \\ 1 & \text{arrest} \end{cases} \tag{5.1}$$

### 5.1.1 Logistic Regression

Logistic regression is the traditional classification approach based on the maximum likelihood method. Maximum likelihood is a general approach that is used to fit many non-linear models (James et al., 2013). Using the logistic function (Equation: 5.2), we can model probability p(X) and have the output fall between 0 and 1 for all values of X.

$$p(X) = \frac{e^{\beta_0+\beta_1X_1+...\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+...+\beta_pX_p}} \tag{5.2}$$

The logistic function always produce an S-shaped curve (Figure A0.2, and so regardless of the value of X, we will obtain a sensible prediction. Linear regression would not an appropriate method for this scenario; in a linear regression model approach, we might estimate probabilities of arrest in violent crime lie outside the [0,1] interval (Figure A0.2. In the logistic function, p(X) represents the probability that a violent crime incident would lead to no arrest or arrest James et al. (2013). For example, one can predict that any violent crime incident with a p(X) $>$ 0.5 has resulted in an arrest. By performing a logistic transformation of probability we get the following equation.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{5.3}$$

The left hand side of the equation represents the *logit* or *log-odds*. This is the log of the odds of P(Y $=$ 1|X) versus P(Y $=$ 0|X). On the right hand side, it is observed that the logit of the logistic regression model is *linear* in X (James et al., 2013). In the logistic function (5.2), the values of $\beta_p$ are unknown. One way to estimate these values based on the training set of the NIBRS data available is to use the maximum likelihood method (5.4). James et al. (2013), explains that the idea is to seek estimates for values of $\beta_p$ such that the predicted probability $p(X)$ of arrest in violent crime incidents is as close to the actual outcome of the incident as possible. Simply put, we try to find values for $\beta_p$ gives a number close to 1 for incidents that have lead to an arrest, and 0 for incidents that lead to no arrests.

$$\ell\left(\beta_0, \beta_1\right) = \prod_{i: y_i = 1} p\left(x_i\right) \prod_{i^{\prime}: y_{i^{\prime}} = 0}\left(1 - p\left(x_{i^{\prime}}\right)\right) \tag{5.4}$$

The estimates of $\beta_p$ are chosen to maximize this likelihood function. These estimates are reported as coefficients in the logistic regression output along with a level of significance at alpha $=$ 0.05. These sign attached to these coefficients give an indication of the relationship of the variable to the outcome of interest (Alber, 2021). For each predictive variable, negative coefficients imply reduced likelihood of arrests, and positive coefficients imply increased likelihood of arrest for that variable. By transforming the coefficients with exponentiation, the odds ratios can be obtained which are more intuitive than the coefficients (Alber, 2021). The odds ratio provides the magnitude of the outcome of interest taking place. For example, if the odds ratio was 0.72, then the odds of arrest would be reduced by 28 percent for that variable, after controlling for all other variables.

I will be using the *glm* and *caret* packages from CRAN in order to build the logistic regression models (Friedman et al., 2010) (Kuhn, 2008).

### 5.1.2 Random Forest

Random forest is a machine learning technique that uses ensemble learning for regression and classification problems (James et al., 2013). Before discussing random forests and how they can be implemented for the arrest classification question, it is important to explain decision trees and bootstrapping.

Decision trees are the building blocks for a random forest. Decision trees use the greedy approach and are built top down, meaning that, it uses information about predictors (branches) to help inform us about the target variable (leaves) (James et al., 2013). At each split, two new nodes are created on the decision tree. This process is known as *recursive binary splitting* (James et al., 2013). Decision trees are advantageous because they mimic "human decision-making", but suffer from high variance which leads to low accuracy (James et al., 2013).

One way to account for this variance is to bootstrap. This can be done by taking repeated samples of the training data, and training our decision trees on each of the bootstrapped training samples. The goal would be to then aggregate multiple decision trees and combine them to give us a single averaged prediction. One issue with bootstrapping is that it might run into the problem of collinearity. This is because decision trees use the most significant variable to decide a split at the very top of the tree. When averaging these correlated decision trees there may not be a large reduction in variance, which also will lead to a low accuracy of the model (James et al., 2013).

This is where the random forest method comes handy, as it is able to decorrelate the process of bootstrapping multiple decision trees. Instead of using the most important feature at the top of the tree when splitting a node, this approach uses a random subsample of $m$ predictors from the full set of $p$ predictors. With each split, a new sample of $m$ predictors is considered. James et al. (2013) explains that the way to calculate $m$ is by using this equation: $m = \sqrt{p}$. That is, the number of predictors considered at each split is equal to the square root of the number of total predictors. If we had 10 predictors, 3 are considered at each split ( $\sqrt{10} \approx 3$). This results in the average of the resulting trees has less variability and hence is more reliable (James et al., 2013).

The next concept to introduce in evaluating random forest models is Gini Impurity.

$$G = \sum_{k=1}^{K} \hat{p}_{mk} \left(1 - \hat{p}_{mk}\right) \tag{5.5}$$

where $\hat{p}_{mk}$ is the proportion of observations in class $k$ in node $m$. The Gini formula (5.5), indicates that if all $\hat{p}_{mk}$'s are part of one class, the Gini index shifts towards zero (James et al., 2013). Ideally, if the values of Gini are small, it means that the observations in a node predominantly fall into a single category, meaning it is a pure node. In order to measure the variable importance for predicting the target variable of *arrest*, one can inspect the mean decrease in Gini index averaged over all trees. The higher the mean decrease in Gini over all trees, implies that that variable is of a higher importance (James et al., 2013).

I will be implementing the random forest classifier using the *randomForest* package and *caret* package, available on CRAN (Liaw and Wiener, 2002) (Kuhn, 2008). Within these packages, I am able to pre-select number for the sub sample of $m$ predictors, as well as specify the metric used for evaluation, gini index.

## 5.2 Model Evaluation and Validation

In this section, I will discuss the metrics I will be using to check the performance of the data. The latter half of this section focuses on how I deal with class imbalance and perform cross validation for both logistic regression and random forest methods.

### 5.2.1 Performance Metrics

**Confusion Matrix**

A confusion matrix captures the classification performance of a classification algorithm with respect to some test data (Ting, 2010). For this paper, this matrix of two dimensions shows the true class of a violent incident on one side; and on the other side, it shows the class that the classifier has determined.

| | | Actual | |
|---|---|---|---|
| | | Arrest | No Arrest |
| Predicted | Arrest | TP | FP |
| | No Arrest | FN | TN |

(5.6)

Figure 5.6 shows the template of the confusion matrix for arrest classification. Since *Arrests* has a two-class outcome (arrest vs. no arrest), the True Positive (TP), and True Negative (TN) values refer to correct predictions (Ting, 2010). These would be populated if the classifier correctly predicts a violent crime incident that leads to an arrest (TP), or correctly predicts an incident lead to no arrest (TN). On the contrary, the False Positive (FP) and False Negative (FN) values refer to incorrect predictions (Ting, 2010). These categories would be populated when the classifier incorrectly predicts a violent crime incident led to an arrest, when it did not (FP), or incorrectly predicts an incident led to no arrest (FN), when there was an arrest.

### True Rates, False Rates and Accuracy

True and false rates can be calculated using the confusion matrix. True positive rate (TPR) is also known as *sensitivity* or recall. It refers to how often the classifier is able to predict arrests in all violent crime incidents which have lead to an arrest. True negative rate (TNR) is also known as *specificity*. It refers to how often the classifier is able to predict no arrests in all violent crimes incidents which have lead to no arrest. The TPR and TNR can be calculated using the following

$$TPR = \frac{TP}{TP+FN} \quad , \quad TNR = \frac{TN}{TN+FP} \tag{5.7}$$

The accuracy of the classification algorithm can be calculated by taking the total number of correct predictions (both arrests and no arrests) divided by the total number of violent crime incidents. This can be expressed as:

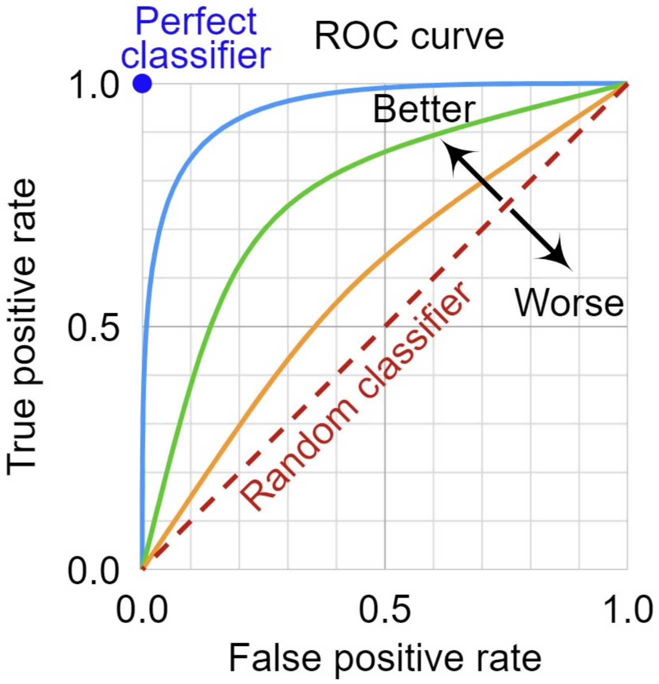
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5.8}$$

For any classification problem, we want to have a highly sensitive and highly specific model; however, there's always a trade off between specificity and sensitivity (Chu, 1999). The relationship between specificity and sensitivity depends on the cut-off value we use to define an incident from an arrest or no arrest. Suppose that the cut off value is 0.5, if classifier scores a violent crime incident >0.5, the prediction is that an arrest took place. If the classifier scores a violent crime incident as <0.5, then the prediction is that an arrest did not take place (Chu, 1999). If the cut-off value is decreased, the sensitivity would increase and specificity would decrease. That is, the number of false negatives; or the number of incidents predicted incorrectly as no arrest, would increase. Increasing the cut off would have the opposite effect, and specificity of the test would increase and the sensitivity would decrease. That is, the number of incidents predicted incorrectly as resulting in arrest, would increase (Chu, 1999).

### ROC Curve and AUC

The ROC (Receiver Operating Characteristic) curve 5.1 gives us a graphical representation of the performance of a binary classification algorithm. The plot is created by plotting the true positive rate against the false positive rate. Recall that the true positive case in my case is the proportion of violent crime incidents that were correctly predicted to lead to an arrest out off all arrests (TP/(TP+ FN)), while the false positive rate is the proportion of violent crime incidents that are were incorrectly predicted to lead to an arrest out of all no arrest incidents (FP/TN + FP)). ROC curves shows the trade-off that is made between sensitivity and specificity. The comparison of ROC curves from two classification models allows us to select the more superior model. Since the classifiers I am using are probabilistic the output I have is a probability of arrest vs no-arrest which can be plotted as a curve based on what my cut-off value is. James et al. (2013) notes that a classifier with a threshold of 50% would give us the highest overall accuracy. A classifier with a threshold of 50% essentially has no predictive value since it is a random guess. in 5.1 the diagonal red dotted line represents the performance of a random guess. The further up and to the left the ROC curve lies from the diagonal line the better the performance is.

**Figure 5.1:** Example of a ROC Curve



The AUC (Area Under the Curve) is directly linked to the ROC curve. It is useful in the measurement of the overall performance of the classifier, summarized over all threshold values James et al. (2013). The AUC calculates the entire two-dimensional area underneath the entire ROC curve. The higher the AUC value, the better the classifier is in predicting arrests and no arrests.

### 5.2.2 Class Imbalance

James et al. (2013) notes that a classifier with a threshold of 50% would give us the highest overall accuracy. This would require the probability of arrest for a violent crime to be at, or more than 50% to be classified as arrest, otherwise it would be classified as non-arrest. In order for this to happen, the dataset has to be somewhat balanced for the classifier to determine arrest probability. A common issue in classification problems is that the *event* that we are measuring could be a rare event. For example, if I had a data set with far fewer arrests than non-arrests, there would be a class distribution that is skewed towards non-arrests. In these situations, we might end up model with a high accuracy, but the specificity would be poor. The classification model would predict all incidents as "non-arrests", and it would be useless.

**Table 5.1:** Types of Violent Crime and Outcomes (Omitting rows with NAs)

| | No-Arrest (0) | Arrest (1) |
|---|---|---|
| All Violent Crime | 311044 | 271279 |
| Aggravated Assault | 40045 | 46517 |
| Sexual Offence | 35501 | 9890 |
| Robbery | 9997 | 3426 |
| Simple Assault | 229371 | 213569 |

Looking at the counts of arrests vs non-arrests for all violent crime and by subsets of offense types (Table 5.1, there are a few considerations to make in the analysis. Class imbalance does not seem to be an issue when reviewing counts for arrests in *All Violent Crime* incidents (47%), *Aggravated Assault* (54%) incidents, and *Simple Assault* incidents (48%). However, class imbalance is very skewed against arrests in cases of *Sexual Offense* (22%), and *Robbery* (25%).

One way to address this issue would be to run an adjusted case for each model in addition to running the classification exercise with the base case (no treatment of class bias). In the adjusted sample, the objective is to draw an equal proportion of arrests and non-arrests for the trainingData set (Prabhakaran, 2016). This method is known as down-sampling, as it is taking a lower number of non-arrest counts (which is the majority), and matching it to the same number of arrests, which is the minority class (Prabhakaran, 2016). The trainingData with an equal proportion, would be best suited to create a model that would predict for both arrests and non-arrests. The remaining sample of arrests and non-arrests, which is not included in the trainingData, would be used for the testData (Prabhakaran, 2016).

### 5.2.3 K-fold Cross-Validation

Typically, classification problems leverage the *validation set approach.* The process would involve randomly splitting the dataset into two parts - the *training set* and *testing set* (James et al., 2013). This method will be used in this paper to run all base cases for each model in logistic regression and random forest. However, there are two drawbacks in using this approach. James et al. (2013), states that the validation set approach would, firstly result in a highly variable test error rate depending on which records are assigned randomly to the training set and testing set. Secondly, the validation set approach only uses the training set to train the model, due to the lower sample size the model is likely to overestimate the test error rate for the model fit on the entire data set (James et al., 2013). The way to address these two issues is to leverage the k-fold cross-validation technique.

The k-fold cross-validation method is a robust way in which to estimate the accuracy of a model (James et al., 2013). The process would involve randomly splitting the training set into k-folds, i.e, if k =5, there would be 5 k-folds. Then one of the folds is excluded, and the model is trained on the remaining 4 k-subsets. Next, the model is tested on the subset that was excluded, and the prediction error is recorded. This process is repeated until each of the subsets has had a chance to be the test set. The prediction errors are then averaged, to give us the performance metrics for the cross-validation.

k-fold CV with lower values of$k$ result in lower variance but higher bias, while higher values of k leads to lower bias but higher variance (James et al., 2013). So the question remains, how many folds is ideal for the research question at hand? James et al. (2013), advises that classification with k-fold CV should be performed using k = 5, or k = 10. The most advantage is computational and since the NIBRS dataset is quite large, it would be advisable to use k=5.

Continuing the discussion from the previous section of class imbalance; in this paper, the class imbalance method and k-fold CV are used in tandem in order to create an adjusted case. The combination of these two methods would result in a better performance in predicting for arrest in all violent crime, and the subsets by each offense type (aggravated assault, sexual offense, robbery and simple assault).

# 6 Results

This section will briefly explain how models are set up to predict the likelihood of arrest. For each type of classification method, there will be five models - one for each of the four violent crime offense types (aggravated assault, simple assault, sexual offense & robbery), and one for modeling all violent crime. The unadjusted base case is implemented within each model to predict the likelihood of arrest. Then using concepts of class imbalance and cross-validation, the adjusted model is implemented. The key performance metrics are discussed for the base and adjusted cases within each of the four offense types and overall violent crime based on the outputs. Then, the predictors from the model are evaluated for the importance they carry in predicting arrests.

## 6.1 Logistic Regression

The logistic regression method is appropriate in utilizing categorical and continuous independent variables to analyze a dichotomous target variable such as arrest: *no arrest (0)* or *arrest (1)*. The outputs are leveraged to summarise the probability of arrest based on each characteristic of the violent crime incident, after controlling for all other predictive variables in the model.

Table A0.1 displays the coefficients from the five logistic regression models. The coefficients show the direction and magnitude of that particular variable's contribution to the likelihood of arrest. The significance level is also recorded with *** indicating significance at <0.01 and ** indicating significance at <0.05. More importantly, In Table A0.2, the calculated odds ratios from can be used to find the odds of arrest for each variable while controlling for all other predictive variables.

### 6.1.1 All Violent Crime

**Table 6.1:** Performance Metrics for Logistic Regression - All Violent Crime

| | Sensitivity | Specificity | Accuracy | AUC |
|---|---|---|---|---|
| Base Model | 0.5769 | 0.6058 | 0.5944 | 0.6331 |
| CV and Adjusted | 0.6644 | 0.5222 | 0.5742 | 0.6327 |

In Table 6.1, the base logistic model predicting for arrests in all violent crimes yields an accuracy of 59.44%. The sensitivity metric of the base model shows that it can predict 57.69% of the actual arrests, and specificity shows that can predict 60.58% of the non-arrests for all violent crime incidents. The cross-validated adjusted model has a higher sensitivity metric of 66.44% but a lower specificity metric of 52.22% and accuracy of 57.42%. Figure 5.1 shows that for both the base and adjusted model, the AUC are similar - 0.6331 and 0.6327, respectively. When looking at the adjusted model for all violent crime incidents, all ten predictive variables were significant ($p<0.05$).

Odds ratios are reported based on the adjusted model. The odds ratios show that with every incremental increase of 1 in the age of the offender, the odds of arrest decrease by 0.5 percent. This finding indicates that younger offenders are more likely to be arrested for overall violent crime incidents than older offenders. Conversely, with every incremental increase of 1 in the age of the victim, the odds of arrest increase by 0.7 percent, indicating that incidents with older victims are more likely to result in an arrest than incidents with younger victims. In terms of the impact of race, incidents with a white offender lead to increased odds of arrest for all violent crime by 20 percent compared to black offenders, net of other predictive variables. In instances with a white victim, the odds of arrest for all violent crimes increased by 35 percent, compared to black victims. For incidents with a male offender - the odds of arrest decreased slightly by 2 percent compared to females, and for incidents with a male victim - the odds of arrest decreased by 9 percent than for females.

Turning to the other characteristics of the offense and individuals involved, it is reported that if the offender was a stranger to the victim, the odds of arrest decreased by 23 percent than if it was a known person. If the offender was reported to be using substances, the odds of arrest increased by 84 percent than if the offender was not using substances. Interestingly, the odds of arrest decreased by 20 percent if the offender was armed with a deadly weapon. For violent crime incidents that took place at night, the odds of arrest increased by 9 percent, compared to incidents during the day. Lastly, taking aggravated assault as the reference, the odds of being arrested for sexual offenses was decreased by about 79 percent, decreased by 65 percent for robbery, and decreased by 32 percent for simple assault.

### 6.1.2 Subset by Offense Type

**Aggravated Assault**

**Table 6.2:** Performance Metrics for Logistic Regression - Aggravated Assault

| | Sensitivity | Specificity | Accuracy | AUC |
|---|---|---|---|---|
| Base Model | 0.6961 | 0.4579 | 0.5580 | 0.6179 |
| CV and Adjusted | 0.8258 | 0.3031 | 0.5870 | 0.6206 |

In Table 6.2, the base logistic model predicting arrests in aggravated assault incidents yields an accuracy of 55.80%. The sensitivity metric of the base model shows that it can predict 69.61% of the actual arrests, and specificity shows that it can predict 45.79% of the non-arrests for aggravated assault incidents. The cross-validated adjusted model has a higher sensitivity metric of 82.58% but a lower specificity metric of 30.31%. The accuracy, however, improved over the base model to 58.70% and the AUC from Figure A0.4 show that both the base and adjusted model are similar, 0.6179 and 0.6206. The coefficient table for the adjusted aggravated assault model shows that 9 out of 10 variables were significant. The one variable that was not significant was the age of the offender.

Odds ratios from the adjusted model indicate that with an incremental increase of 1 in the age of the victim, the odds of arrest increased by 0.8 percent. For the race-specific variables, it is reported that the odds of arrest for an aggravated assault involving a white offender increased by 19 percent compared to black offenders. Whereas the odds of arrest involving a white victim result in a 40 percent increase compared to black victims. Sex-specific predictors indicate that odds of arrest for male offenders were lowered by 13 percent, and for male victims, it was lowered by 14 percent than for female offenders and victims. Other predictive variables indicate that the odds of arrest was lowered by 43 percent if the offender was a stranger to the victim, and increased by 75 percent if the offender was using substances. The use of a deadly weapon in the incident reduced the odds of arrest by 21 percent compared to no deadly weapon use and odds of arrest increased by 12 percent if the aggravated assault took place in the nighttime compared to daytime.

**Sexual Offense**

**Table 6.3:** Performance Metrics for Logistic Regression - Sexual Offense

| | Sensitivity | Specificity | Accuracy | AUC |
|---|---|---|---|---|
| CV and Adjusted | 0.9401 | 0.0713 | 0.4091 | 0.5438 |

The base logistic regression classifier for sexual offenses failed to provide predictions for both arrests and non-arrests. Instead, it classified all sexual offense incidents as non-arrests. This finding is due to a significant class imbalance in the dataset, where only 21 percent of incidents lead to arrests.Hence, excluding the base model from the logistic regression analysis is the best option. On the other hand, the adjusted model that was treated for class imbalance and executed using cross-validation was able to provide some predictive value; albeit poor. In Table 6.3, the adjusted logistic regression model predicting arrests in sexual offense incidents yields an accuracy of 41%. While the sensitivity is high - 94.01% and the model can predict actual arrests correctly, the model has a poor specificity of only 7.13%. Given that the model is worse than random guessing in predicting for non-arrests, it makes little sense to examine the coefficients and log-odds extracted from this model.

**Robbery**

**Table 6.4:** Performance Metrics for Logistic Regression - Robbery

| | Sensitivity | Specificity | Accuracy | AUC |
|---|---|---|---|---|
| Base Model | 0.0029 | 0.9955 | 0.7425 | 0.6277 |
| CV and Adjusted | 0.5277 | 0.6376 | 0.6281 | 0.6147 |

The base model logistic regression classifier performed poorly compared to the cross-validated adjusted model for robbery. From Table 6.4, the base model yields an accuracy of 74.25% and specificity of 99.55%. This finding means that the base model can predict non-arrests almost perfectly. However, the model performs poorly on sensitivity as it can correctly predict arrests in 0.29% of all robbery incidents. On the other hand, the adjusted model performs better as the sensitivity and specificity are both >50%. It can predict 52.77% of the actual arrests and predict 63.76% of the non-arrests for robbery incidents. The AUC for the base model was 0.6277, and for the adjusted model, it was 0.6147, which is pretty similar. From the coefficients table A0.1 six of the ten predictors were significant ($p<0.05$) in the adjusted robbery logistic regression model. These variables are: Age of offender, Offender Not Black, Victim Not Black, Offender Stranger, Offender Substance Use, Deadly Weapon.

Odds ratios extracted from Table A0.2 for the adjusted robbery model show that with each incremental increase in age of the offender, the odds of arrest increase by 0.6 percent. The odds of arrest for robbery are also increased by 43 percent for incidents with white offenders and by 33 percent for incidents with white victims. Instances in which offenders are strangers lead to a decreased odds of arrest by 44 percent. If the offender was using substances, the odds of arrest increased by 73 percent, and if there was a deadly weapon involved, the odds of arrest decreased by 15 percent. Lastly, the time of day was significant in the base model but is not in the adjusted model.

### Simple Assault

**Table 6.5:** Performance Metrics for Logistic Regression - Simple Assault

| | Sensitivity | Specificity | Accuracy | AUC |
|---|---|---|---|---|
| Base Model | 0.4456 | 0.6896 | 0.5719 | 0.6001 |
| CV and Adjusted | 0.5827 | 0.5669 | 0.5736 | 0.6016 |

From Table 6.5, the base model yields an accuracy of 57.19%, sensitivity of 44.56%, and specificity of 68.96%. This finding means that the base model can predict non-arrests higher than actual arrests. The base model's specificity is worse than pure guessing and needs to be adjusted. On the other hand, the adjusted model performs better as the sensitivity and specificity are both >50%. It can predict 58.27% of the actual arrests and predict 56.69% of the non-arrests for simple assault incidents. The accuracy of the adjusted model is also slightly better than the base model with 57.36%. The AUC for the adjusted model was 0.6016 and for the base model, it was 0.6001; which is also slightly better. From the coefficients table A0.1, eight of the nine predictors were significant ($p<0.05$) in the adjusted robbery logistic regression model. The variable Deadly Weapon was removed from the analysis as there were no instances of deadly weapon use in simple assault incidents. The significant variables are Age of Offender, Age of Victim, Offender Not Black, Victim Not Black, Victim Male, Offender Stranger, Offender Substance Use, Time of Day - Night time.

Odds ratios extracted from Table A0.2 for the adjusted simple assault model shows that with each incremental increase in age of the offender, the odds of arrest decrease by 0.6 percent, and with each incremental increase in age of victim, the odds of arrest increase by 0.9%. The odds of arrest for simple assault are also increased by 21 percent for incidents with white offenders and by 34 percent for incidents with white victims. In simple assault incidents where the victim is a male, the odds of arrest decrease by 10%. Instances in which offenders is a stranger lead to a decreased odds of arrest by 19 percent. If the offender was using substances, the odds of arrest increased by 92 percent. Lastly, the time of day was significant in the adjusted model leading to a 9% increase in odds of arrest for night-time incidents.

## 6.2 Random Forest

The random forest classification algorithm employs decision trees and bootstrapping methods to predict if a violent crime incident would lead to an arrest or no arrest. Unlike the logistic regression modeling and its coefficients, it is not possible to track the significance or the magnitude of the odds of arrest. Instead, the random forest outputs the variable importance plot. The importance of a variable to the arrest outcome is determined by the mean decrease in the Gini measure, averaged across all trees. The ranking of each variable's mean decrease in Gini would indicate as to which of the variables are most important for the likelihood of arrest that particular violent crime.

### 6.2.1 All Violent Crime

**Table 6.6:** Performance Metrics for Random Forest - All Violent Crime

| | Sensitivity | Specificity | Accuracy | AUC |
|---|---|---|---|---|
| Base Model | 0.4577 | 0.5486 | 0.5001 | 0.6128 |
| CV and Adjusted | 0.6795 | 0.5402 | 0.5962 | 0.6593 |

**Figure 6.1:** Variable Importance - All Violent Crimes

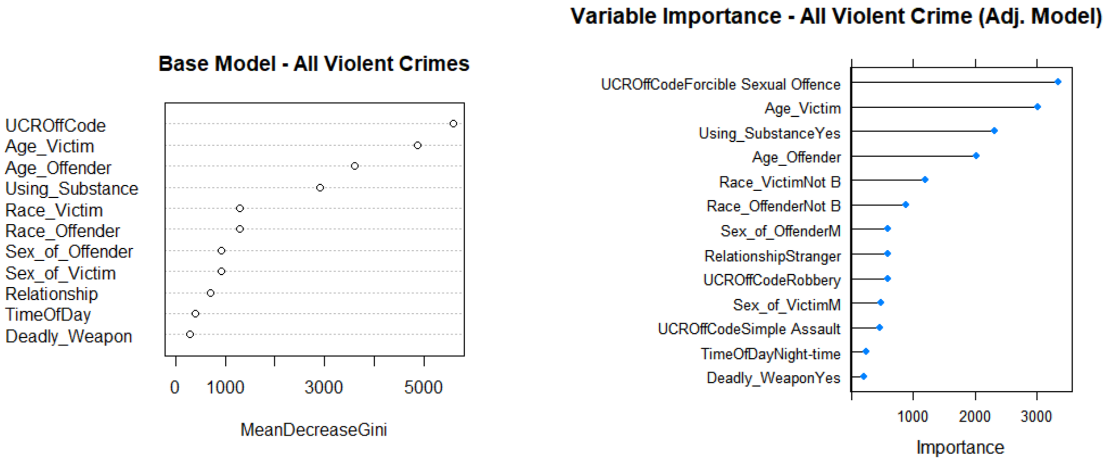


Table 6.6 displays the random forest performance metrics for the base and adjusted models predicting arrest in all violent crimes. The base model had a poor sensitivity score of lower than 50% (45.77%). The sensitivity of the base model was 45.86%, and the accuracy was 50.01%. The base model performed just slightly better than random guessing. The adjusted model, which adjusted for class imbalance and applied cross-validation methods, performed better. Compared to the base model, the adjusted model yielded a sensitivity of 67.95%, specificity of 54.02% and an accuracy of 59.62%. The AUCs obtained from the ROC plot (Figure A0.8) show that the adjusted model (0.6593) was higher than the base model (0.6128).

Figure 6.1 shows the variable importance for all violent crime incidents in the base and adjusted models. The variables are sorted according to the highest mean decrease in Gini averaged across all trees. The variable importance for the adjusted model predicting for arrests in all violent crime indicates that the UCR offense code (Sexual Offense) had the highest importance, followed by the Age of the Victim and Use of Substances (Yes) by the offender. On the lower end of importance are the UCR offense code for simple assault, Time of Day (Night-time) and the offenders use of a Deadly Weapon.

## 6.2.2 Subset by Offense Type

### Aggravated Assault

**Table 6.7:** Performance Metrics for Random Forest - Aggravated Assault

| | Sensitivity | Specificity | Accuracy | AUC |
|---|---|---|---|---|
| Base Model | 0.7269 | 0.4653 | 0.6230 | 0.5961 |
| CV and Adjusted | 0.5990 | 0.6013 | 0.6004 | 0.6001 |

**Figure 6.2:** Variable Importance - Aggravated Assault

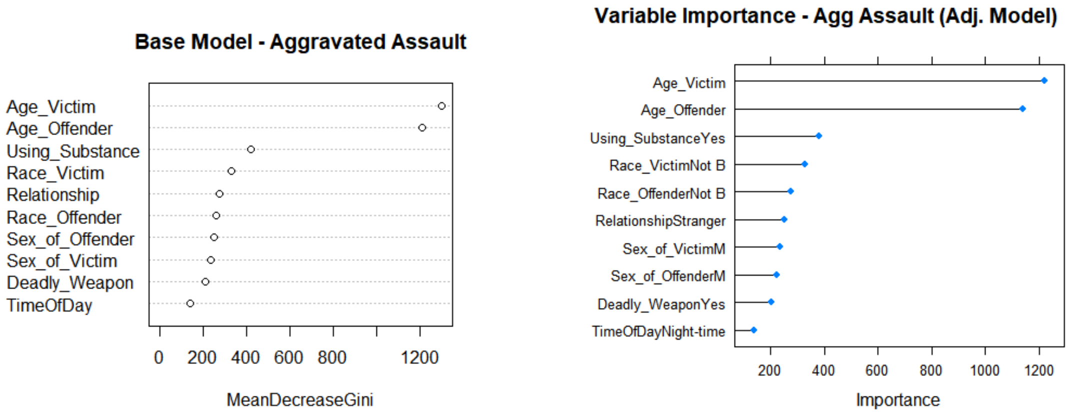


Table 6.7 displays the random forest performance metrics for the base and adjusted models predicting arrest in aggravated assault incidents. The base model has a relatively high sensitivity score of 72.69% while the specificity was poor at 46.53% (less than 50%). The base model yielded an accuracy of 62.30%. In comparison, the adjusted model yielded a lower sensitivity of 59.90%, but higher specificity of 60.31% and an accuracy of 60.04%. The AUCs obtained from the ROC plot (Figure A0.9), show that the adjusted model (0.6001) was slightly higher than the base model (0.5961).

Figure 6.2, shows the variable importance for aggravated assault incidents in the base and adjusted models. The variable importance for the adjusted model predicting for arrests in aggravated assault instances indicates that the Age of the Victim and Age of the Offender had the highest importance , followed by Substance Use by the offender and Race of the Victim (Not Black). On the lower end of importance are the Sex of the Offender (Male), use of Deadly Weapon (Yes) by the offender, and Time of Day (Night-time).

### Sexual Offense

**Table 6.8:** Performance Metrics for Random Forest - Sexual Offense

| | Sensitivity | Specificity | Accuracy | AUC |
|---|---|---|---|---|
| Not Adjusted | 0.0025 | 0.9997 | 0.7760 | 0.5011 |
| CV and Adjusted | 0.6021 | 0.5590 | 0.6002 | 0.6001 |

**Figure 6.3:** Variable Importance - Sexual Offense

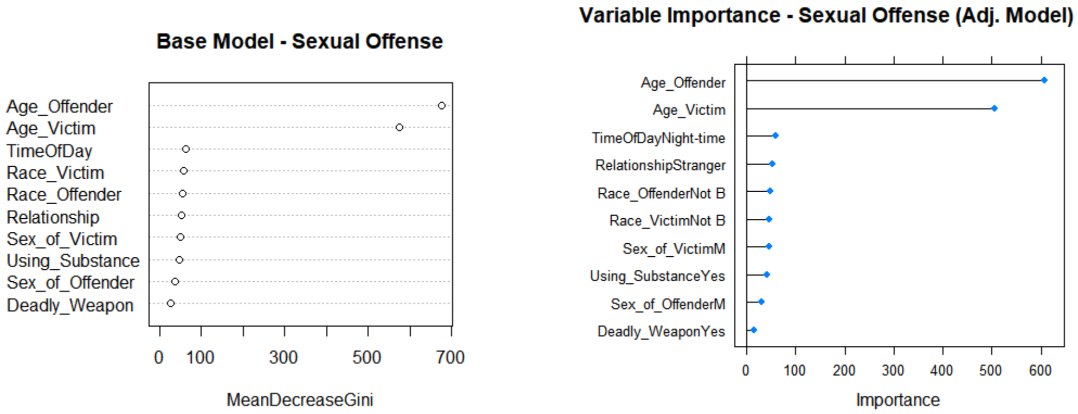


Table 6.8 displays the random forest performance metrics for the base and adjusted models predicting for arrest in sexual offense incidents. The base model has a very low sensitivity score of 0.25% while the specificity was almost perfect at 99.9%. The base model yielded an accuracy of 77.60%. In comparison, the adjusted model yielded a better sensitivity of 60.21%, and a lower specificity, (but still acceptable) of 55.90% and an accuracy of 60.02%. The AUCs obtained from the ROC plot (Figure A0.10), show that the adjusted model (0.6001) was higher than the base model (0.5011).

Figure 6.3, shows the variable importance for sexual offense incidents in the base and adjusted models. The variable importance for the adjusted model predicting for arrests in sexual offense instances indicates that the Age of the Offender and Age of the Victim had the highest importance , followed by Time of Day (Night time) and the Relationship (Stranger). On the lower end of importance are Use of Substances by Offender (Yes), Sex of the Offender (Male), use of Deadly Weapon (Yes) by the offender.

### Robbery

**Table 6.9:** Performance Metrics for Random Forest - Robbery

| | Sensitivity | Specificity | Accuracy | AUC |
|---|---|---|---|---|
| Not Adjusted | 0.0324 | 0.9872 | 0.7329 | 0.510 |
| CV and Adjusted | 0.6886 | 0.4957 | 0.6719 | 0.658 |

**Figure 6.4:** Variable Importance - Robbery

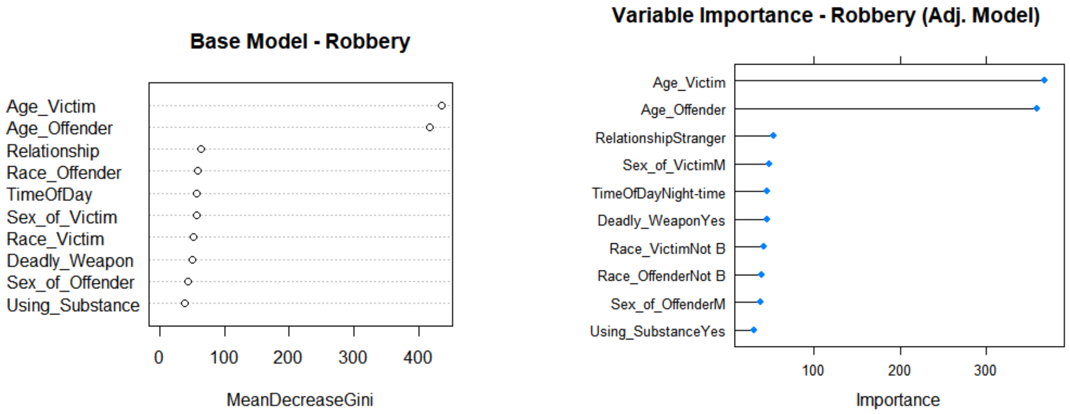


Table 6.9 displays the random forest performance metrics for the base and adjusted models predicting for arrest in robbery incidents. The base model has a very low sensitivity score of 3.24% while the specificity was almost perfect at 98.72%. The base model yielded an accuracy of 73.29%. In comparison, the adjusted model yielded a better sensitivity of 68.86%, but a poor specificity (lower than 50%) of 49.57%. The accuracy of the adjusted model was 67.19%. The AUCs obtained from the ROC plot (Figure A0.11), show that the adjusted model (0.658) performed better than the base model (0.510).

Figure 6.4, shows the variable importance for robbery incidents in the base and adjusted model. The variable importance for the adjusted model predicting for arrests in robbery instances indicates that the Age of the Victim and Age of the Offender had the highest importance, followed by the Relationship (Stranger), and Sex of the Victim (Male). On the lower end of importance are the Race of Offender (Not Black), Sex of the Offender (Male), use of substances (Yes) by the offender.

### Simple Assault

**Table 6.10:** Performance Metrics for Random Forest - Simple Assault

| | Sensitivity | Specificity | Accuracy | AUC |
|---|---|---|---|---|
| Not Adjusted | 0.5681 | 0.6285 | 0.5994 | 0.5983 |
| CV and Adjusted | 0.6108 | 0.5907 | 0.5997 | 0.6008 |

**Figure 6.5:** Variable Importance - Simple Assault

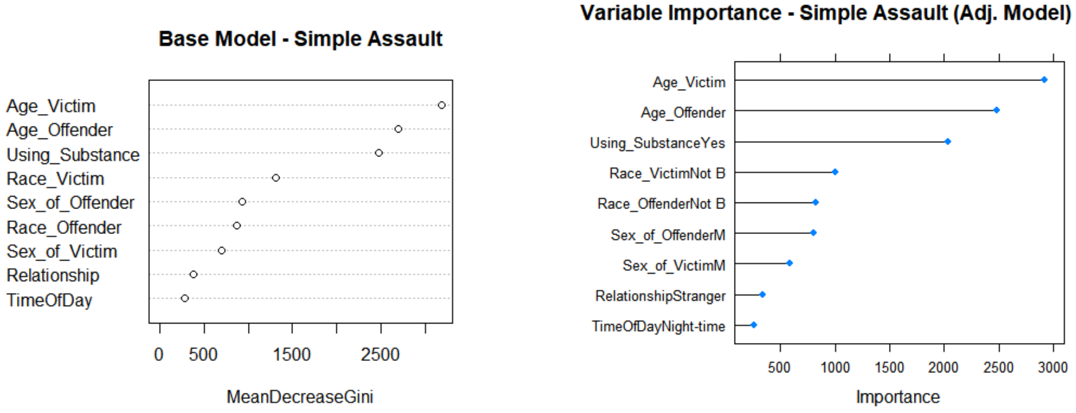


Table 6.10 displays the random forest performance metrics for the base and adjusted models predicting arrest in simple assault incidents. The base model has a sensitivity score of 56.81%, while the specificity is 62.85%. The base model yielded an accuracy of 59.94%. The adjusted model yielded a slightly better sensitivity of 61.08%, but a slightly worse specificity of 59.07%. The accuracy of the adjusted model was 59.97%, which is almost the same as the base model. The AUCs obtained from the ROC plot (Figure A0.12), show that the adjusted model (0.6008) performed slightly better than the base model (0.5983).

Figure 6.5, shows the variable importance for robbery incidents in the base and adjusted model. The variable importance for the adjusted model predicting for arrests in robbery instances indicates that the Age of the Victim and Age of the Offender had the highest importance, followed by the use of substances (Yes) by the offender, and Race of the Victim (Not Black). On the lower end of importance is the Sex of the Victim (Male), Relationship (Stranger), and Time of Day (Night-time).

## 6.3 Comparison of Classification Methods

Table 6.11, is a summary of the adjusted performance metrics from the models. Overall, the random forest classifier performs better than the logistic regression classifier for aggravated assault, sexual offenses, simple assault, and overall grouping of all violent crimes. The adjusted random forest metrics are better than the adjusted logistic regression metrics by observing the sensitivity, specificity, and accuracy for these subsets. However,

**Table 6.11:** Comparing Performance Metrics from Adjusted Models for All Violent Crime and Offense Types

| | | Sensitivity | Specificity | Accuracy | AUC |
|---|---|---|---|---|---|
| All Violent Crime | Logistic Regression | 0.6644 | 0.5222 | 0.5742 | 0.6331 |
| | Random Forest | 0.6795 | 0.5402 | 0.5962 | 0.6593 |
| Aggravated Assault | Logistic Regression | 0.8258 | 0.3031 | 0.5870 | 0.6206 |
| | Random Forest | 0.5990 | 0.6013 | 0.6004 | 0.6001 |
| Sexual Offense | Logistic Regression | 0.9401 | 0.0713 | 0.4091 | 0.5438 |
| | Random Forest | 0.6021 | 0.5590 | 0.6002 | 0.6001 |
| Robbery | Logistic Regression | 0.5277 | 0.6376 | 0.6281 | 0.6147 |
| | Random Forest | 0.6886 | 0.4957 | 0.6719 | 0.6580 |
| Simple Assault | Logistic Regression | 0.5827 | 0.5669 | 0.5736 | 0.6016 |
| | Random Forest | 0.6108 | 0.5907 | 0.5997 | 0.6008 |

the adjusted logistic regression classifier performs better for the crime of robbery, although it has a lower Accuracy and AUC than the adjusted robbery random forest model.

# 7 Discussion

Discovering which of the characteristics of the offender, victim, and situational factors impact the likelihood of arrest could put to rest biased and preconceived notions about who commits violent crime and who is affected by it. In this thesis, I leveraged two classification methods to study the impact of these characteristics on arrest in all violent crime incidents, as well as, by subsets of offense type. In this section, I will be providing an answer to the research question, and comparing the findings to the literature review. Lastly, there will be a discussion on the limitations of this study, the data, and this type of analysis.

## 7.1 Likelihood of Arrest for Violent Crimes

The classification models were built to predict for arrests to answer the research question of which which demographic and situational characteristics of a violent crime incident contribute to a higher likelihood of arrest. Looking at the overall instances of violent crimes, the AUC and accuracy of the adjusted random forest model (0.6593, 59.62%) are slightly better than the AUC and accuracy of the adjusted logistic regression model (0.6331, 57.42%). Additionally, the sensitivity and specificity metrics for the random forest model (0.6795, 0.5402) are slightly better than the sensitivity and specificity metrics from the adjusted logistic regression model (0.6644, 0.5222). These results prove that both these classifiers have a certain degree of predictive power.

In the adjusted logistic regression model, if the offender's race was white rather than black, the odds of arrest increase for aggravated assault, simple assault, robbery and overall all violent crime. This result was statistically significant. However, this association was not significant for sexual offenses. Similarly, the D'Alessio and Stolzenberg (2003) study using logistic regression found that the odds of arrest for white offenders was statistically higher than black offenders for aggravated assault, simple assault, robbery. The authors also conclude that similar to this thesis, the association was not statistically significant for the crime of rape (a type of sexual offense). The results from this thesis and the D'Alessio and Stolzenberg (2003) study are contrary to the reporting of mainstream news which exaggerate rates of black individuals offending (Ghandnoosh, 2014). Empirically, it can be concluded that mainstream news depict black offenders in a negative light while the research says that white offenders are actually more likely to be arrested for committing violent crimes. If the victim's race was white rather than black, the odds of arrest increase significantly for the aggregate of all violent crime, in addition to aggravated assault, robbery and simple assault. This association was found to be not significant for sexual offenses. The D'Alessio and Stolzenberg (2003) study, found similar results where the odds of arrest for incidents involving white victims were higher for aggravated assault, simple assault, and robbery. The difference however was that in the D'Alessio and Stolzenberg (2003) study only simple assault was found to be statistically significant; aggravated assault and robbery incidents were not significant.

In the adjusted logistic regression model, if the offender's sex was male rather than female, the odds of arrest decrease for aggravated assault and overall all violent crime. This result was statistically significant. However, this association was not significant for robbery and simple assault. Similarly, the Stolzenberg and D'Alessio (2004) study using logistic regression found that the odds of arrest for female offenders was statistically higher than male offenders for aggravated assault. The authors also found that female offenders had a significantly lower odds of arrest for simple assault, and the association was not significant for robbery incidents. If the victim's sex was male rather than female, the odds of arrest decrease significantly for the aggregate of all violent crime, in addition to aggravated assault, robbery and simple assault. Furthermore, the Stolzenberg and D'Alessio (2004) study found similar results where the odds of arrest for incidents involving male victims were statistically higher for aggravated assault and simple assault but not statistically higher for incidents of robbery.

The age of the offenders and victim also played a role in the likelihood of arrest. The variable importance plots obtained from the adjusted random forest models indicate that the age of the offender and victim are both one of the most important variables in determining the outcome of arrest. Using the adjusted logistic regression model, the likelihood of arrest decreases as the age of the offender increases for the aggregate of all violent crimes, as well as simple assault. This observation was statistically significant. Meanwhile the likelihood of arrest in robbery incidents increases as the age of the offender increases. Additionally, as the age of the victim increases, so does the likelihood of arrest.

This can be explained by the fact that younger victims of violent crimes might be less likely to report the incident to law enforcement.

The likelihood of arrest was decreased significantly for all violent crime, and three of the four subsets of violent crime (aggravated assault, sexual offenses, and robbery) if the offender was using of a deadly weapon. It should be noted that no instance of simple assault involved the use of a deadly weapon. This result was matched in the D'Alessio and Stolzenberg (2003) study, where the authors also found that the use of a deadly weapon actually decreased significantly the odds of arrest. One potential reasoning for this could be that the recollection of the offender by the victim was not reliable, as the victim was preoccupied by the potential danger posed by the deadly weapon. If the reported characteristics of the offender by the victim is incorrect due to the presence of a deadly weapon, it would most likely lead to the outcome of no arrest. Relationship between the offender and victim impacted the likelihood of arrests for all violent crimes, and in the the subset of offense types: aggravated assault, simple assault and robbery in the adjusted logistic regression models. There was a significant negative association for odds of arrest if the offender was a stranger to the victim. This finding would not be surprising as the victim would have a lower recall for an offending stranger's demographic characteristics, than if the offending person was known to the victim. These findings also agree with the D'Alessio and Stolzenberg (2003) study where the likelihood of arrest was significantly reduced if the offending person was a stranger for incidents of aggravated assault, simple assault, and robbery.

Examining the impact of Offender Substance Use, the adjusted logistic regression model in this thesis shows that the odds of arrests increase significantly for the aggregate of all violent crime, in addition to the the following types of offenses: aggravated assault, robbery and simple assault. This finding is also consistent with the work of D'Alessio and Stolzenberg (2003) and Stolzenberg and D'Alessio (2004).

A unique contribution that this thesis provides is the odds of arrest within each subset of violent crime. The adjusted logistic regression model for all violent crime shows that the odds of being arrested for sexual offenses was decreased by about 79 percent, in comparison to the reference (aggravated assault). This finding is backed up by the results of the adjusted random forest model for all violent crime. It is observed that the UCR

Offense code for sexual offenses was the most important variable in determining arrest via the mean decrease in Gini measure.

Another unique contribution that this thesis makes is the addition of one other predictive variable in predicting for arrests: Time of Day. The time of day (night-time) was significant in increasing the odds of arrest for the aggregate of all violent crimes, in addition to the following subsets of offenses: aggravated assault and simple assault. In the adjusted random forest model for aggravated assault, Time of day was the third most important variable whereas for the adjusted random forest model for simple assault, Time of Day was the least important variable in determining arrest via the mean decrease in Gini measure.

## 7.2 Limitations

It is important to point out that the NIBRS data is not the exact representation of all crime that takes place in the United States. This is because, only half of the violent crimes in the US are actually reported to authorities (Gramlich, 2020), and the data is subject to reporting bias as it is collected through the *voluntary* reporting of local-level agencies (Pepper and Petrie, 2003). This means that my analysis is looking at a sliver of crime that is reported to the local-level reporting agencies which as to then submitted to the FBI for reporting (FBI, 2014). In the Data section, I explain that a significant disadvantage of the data set is that of the 18,489 local-level agencies in the country only about 6520 agencies sent in their data to the NIBRS 2014. The NIBRS 2014 is only representative of 93,330,000 individuals, where the population of America exceeds 318 million individuals (FBI, 2014).

Additionally, predictions can be fundamentally biased as the research design leverages data from offenders to predict the arrests and not from conviction data. This means that although an arrest is made for a violent crime incident, there is no way to link it to a conviction or an acquittal. We do not have the data to inherently prove that one gender or race commits more crime, only that these incidents might be a) reported more frequently, or b) are arrested more often compared to other groups.

Lastly, the two classification methods used report the impact of the predictive variables on the outcome of arrest differently. The logistic regression model is able to output coefficients and odds ratios to quantify the likelihood of arrest. On the other hand, the random forest model is only able to rank the importance of all predictive variables for each of the individual models.

# 8 Conclusion

The main objective of the thesis was to investigate demographic and situational characteristics of violent crime incidents for their impact on likelihood of arrest in America. The focus was to develop empirical evidence instead of relying on biases and preconceived notions, or skewed policing algorithms about who commits crime, and who is impacted by it.

The models that were created leveraged logistic regression and random forest methods and they provided some predictive value in classification of arrests. The models for the aggregate of all violent crime, as well as the subsets of offense types had a good predictive power with an accuracy of greater than 50%. Additionally, adjusted models were built to address class imbalance and leveraged cross-validation methods. Using odds ratios from the logistic regression results, and the variable importance plots from the random forest - likelihood of arrest was ascertained.

The results indicate that generally the likelihood of arrest increases under certain conditions. These conditions are: in incidents where the race of the offender is white, in incidents where the race of the victim is white, in incidents where the offender is a female (for aggravated assault instances), and in incidents where if the victim of a violent crime is a female. Generally, the likelihood of arrest decreases as the age of the offender increases, and the likelihood of arrest increases as the age of the victim increases. The likelihood of arrest decreases for incidents where the offender is armed with a deadly weapon, and where the offender and victim are strangers. Additionally, the likelihood of arrest increases for all violent crimes if the incident takes place at night time compared to day time, and in incidents where the offender is using substances.

Some of the limitations of the study were that the NIBRS dataset may not be representative of all violent crime that takes place in America. This is due to the low proportion of local-level reporting agencies reporting their data to the FBI. One unique contribution of this paper is that it leverages Time of Day as one of the predictive variable, which had not been used in the literature that was reviewed for this paper.

# References

Alber, S. (2021). Logistic regression. University of California Davis: Course Material.

Baer, D. (2016). The Psychology of Why Americans Are Afraid of Historically Low Crime Levels. https://www.thecut.com/2016/07/psychology-why-americans-afraid-low-crime-levels.html [Accessed: 2021-10-31].

Binswanger, I. A., Redmond, N., Steiner, J. F., and Hicks, L. S. (2012). Health disparities and the criminal justice system: an agenda for further research and action. *Journal of Urban Health*, 89(1):98–107.

Chu, K. (1999). An introduction to sensitivity, specificity, predictive values and likelihood ratios. *Emergency Medicine*, 11(3):175–181.

Coviello, D. and Persico, N. (2016). An economic analysis of black-white disparities in nypd's stop and frisk program (working paper no. 18803).

D'Alessio, S. and Stolzenberg, L. (2003). Race and the probability of arrest. *Social forces*, 81(4):1381–1397.

FBI (2014). Participation by state, 2014.

FBI (2019). 2019 Violent Crime in the United States. https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/violent-crime [Accessed: 2021-10-31].

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Friend, Z. (2013). Predictive policing: Using technology to reduce crime. *FBI Law Enforcement Bulletin*, 82(4):1–4.

Ghandnoosh, N. (2014). Race and Punishment: Racial Perceptions of Crime and Support for Punitive Policies. https://www.sentencingproject.org/publications/race-and-punishment-racial-perceptions-of-crime-and-support-for-punitive-policies [Accessed: 2021-10-31].

Gramlich, J. (2020). What the data says (and doesn't say) about crime in the United States. https://www.pewresearch.org/fact-tank/2020/11/20/facts-about-crime-in-the-u-s/ [Accessed: 2021-10-31].

Heaven, W. D. (2021). Predictive policing is still racist—whatever data it uses.

Hindelang, M. J. (1978). Race and involvement in common law personal crimes. *American Sociological Review*, 43(1):93–109.

Horton, A. (2008). Violent crimes and racial profiling, what the evidence suggests. *Journal of Human Behavior in the Social Environment*, 6(4):87–106.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Kochel, T. R., Wilson, D. B., and Mastrofski, S. D. (2011). Effect of suspect race on officers'arrest decisions. *Criminology*, 49(2):473–512.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, 28(5):1–26.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Maltz, M. D. (1999). *Bridging gaps in police crime data.* DIANE Publishing.

Nix, J., Richards, T. N., Pinchevsky, G. M., and Wright, E. M. (2019). Are domestic incidents really more dangerous to police? findings from the 2016 national incident based reporting system. *Justice Quarterly*, pages 1–23.

Patel, F. (2015). Be cautious about data-driven policing. *The New York Times. Retrieved from https://www. nytimes. com/roomfordebate/2015/11/18/can-predictivepolicing-be-ethical-and-effective/be-cautious-about-data-driven-policing.*

Pepper, J. and Petrie, C. (2003). *Measurement problems in criminal justice research: Workshop summary.* National Academies Press.

Perkins, C. A. (1997). *Age patterns of victims of serious violent crime.* US Department of Justice, Office of Justice Programs, Bureau of Justice . . . .

Perry, W. L. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations.* Rand Corporation.

Prabhakaran, S. (2016). Logistic Regression. http://r-statistics.co/Logistic-Regression-With-R.html. Accessed: 2021-12-10.

School, C. L. (2021). Deadly Weapon. https://www.law.cornell.edu/wex/deadly_weapon [Accessed: 2021-10-31].

Stolzenberg, L. and D'Alessio, S. J. (2004). Race and the probability of arrest. *Journal of Criminal Justice*, 32(5):443–454.

Strom, K. J. and Smith, E. L. (2017). The future of crime data: The case for the national incident-based reporting system (nibrs) as a primary data source for policy evaluation and crime analysis. *Criminology & Public Policy*, 16(4):1027–1048.

Sun, C.-c., Yao, C., Li, X., and Lee, K. (2014). Detecting crime types using classification algorithms. *J. Digit. Inf. Manag.*, 12(5):321–327.

Sun, E. (2018). The Dangerous Racialization of Crime in U.S. News Media. https://www.americanprogress.org/issues/criminal-justice/news/2018/08/29/455313/dangerous-racialization-crime-u-s-news-media/ [Accessed: 2021-10-31].

Ting, K. M. (2010). *Confusion Matrix*, pages 209–209. Springer US, Boston, MA.

Yang, R. and Olafsson, S. (2011). Classification for predicting offender affiliation with murder victims. *Expert Systems with Applications*, 38(11):13518–13526.

# Appendix

**Figure A0.1:** Participation by State, NIBRS 2014

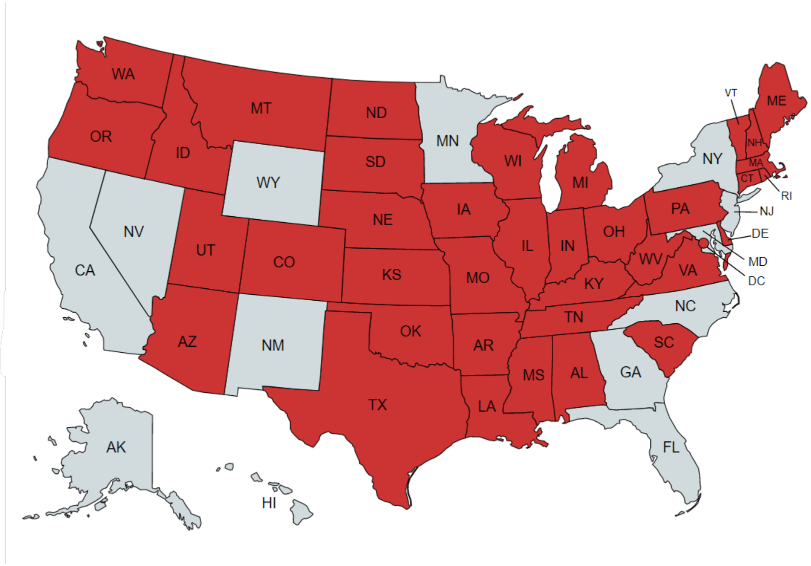


**Figure A0.2:** Linear vs. Logistic Regression

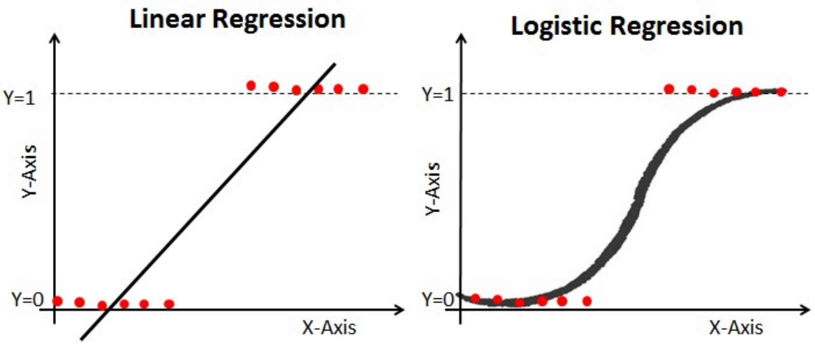


**Figure A0.3:** ROC Curves for Logistic Regression - Base and Adjusted Models (All Violent Crime)

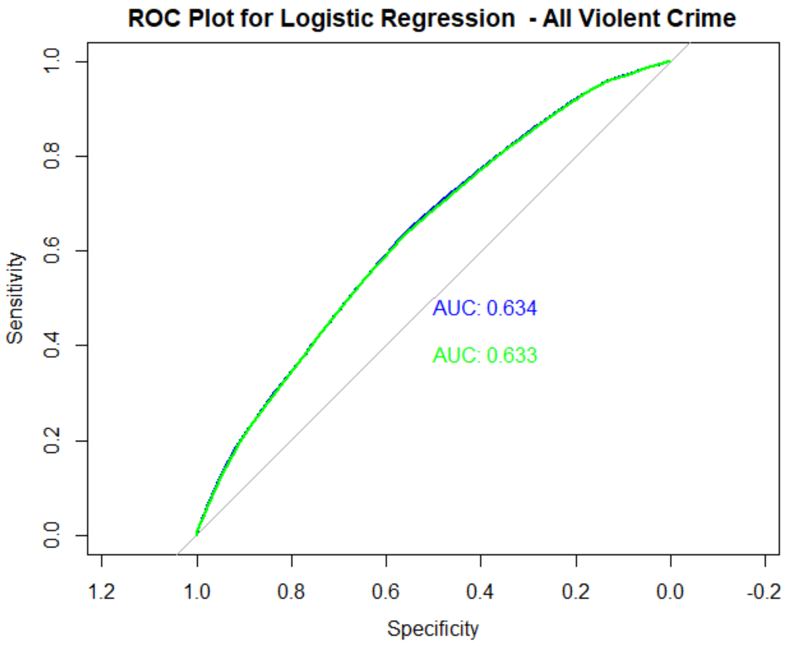


**Figure A0.4:** ROC Curves for Logistic Regression - Base and Adjusted Models (Aggravated Assault)

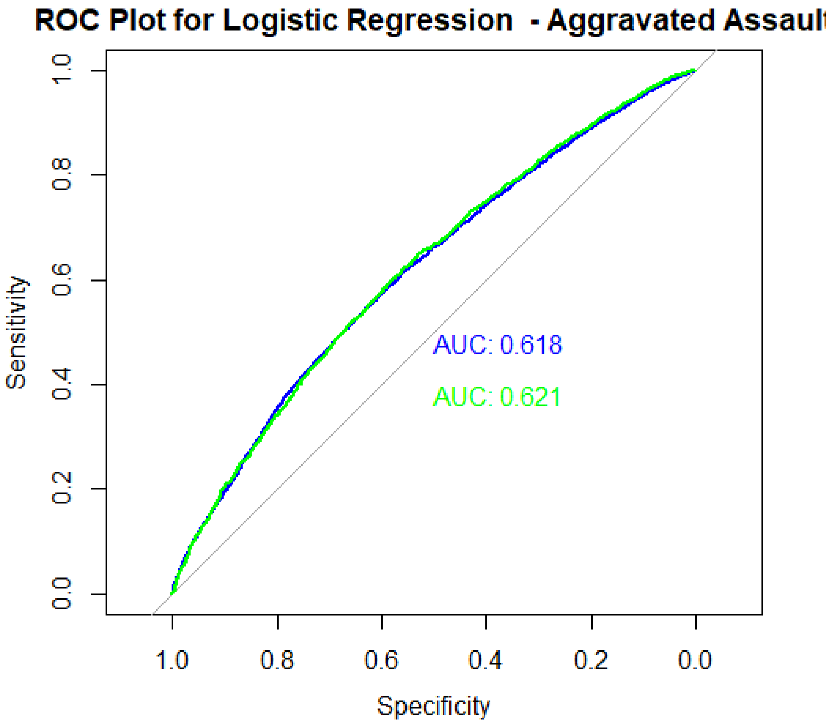


**Figure A0.5:** ROC Curve for Logistic Regression - Adjusted Model (Sexual Offense)

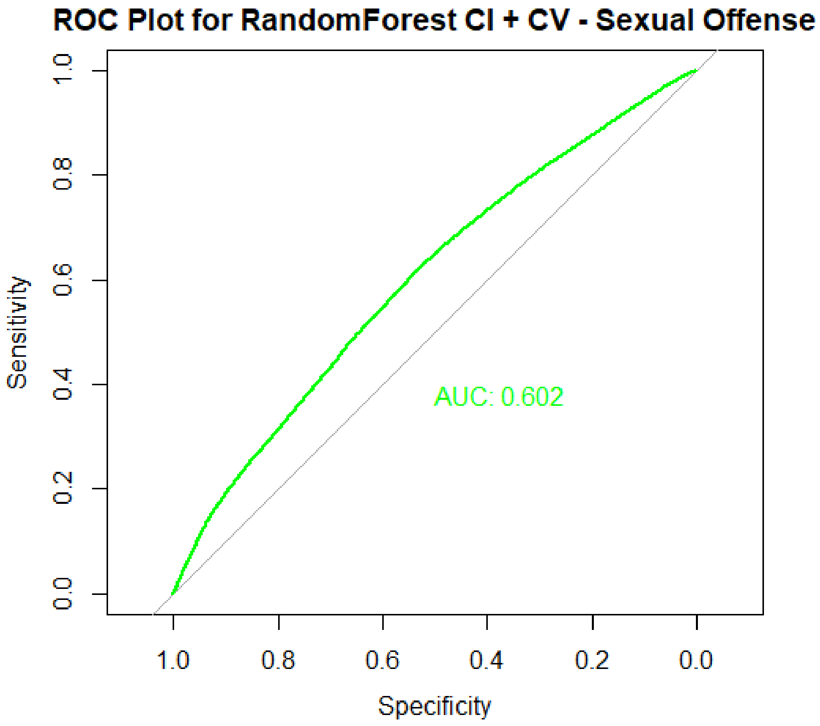


**Figure A0.6:** Comparing ROC Curves for Logistic Regression - Base and Adjusted Models (Robbery)

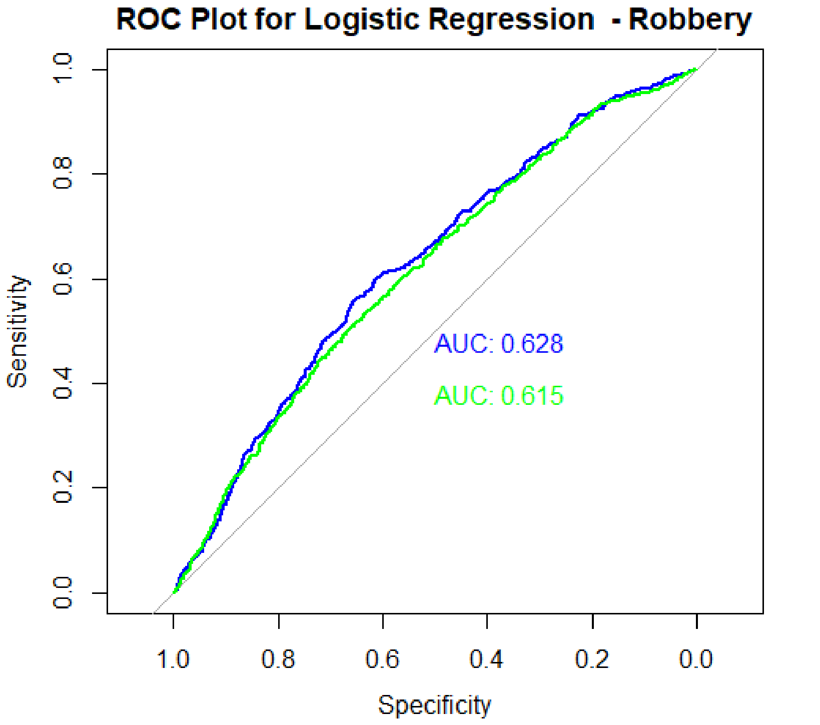


**Figure A0.7:** Comparing ROC Curves for Logistic Regression - Base and Adjusted Models (Simple Assault)

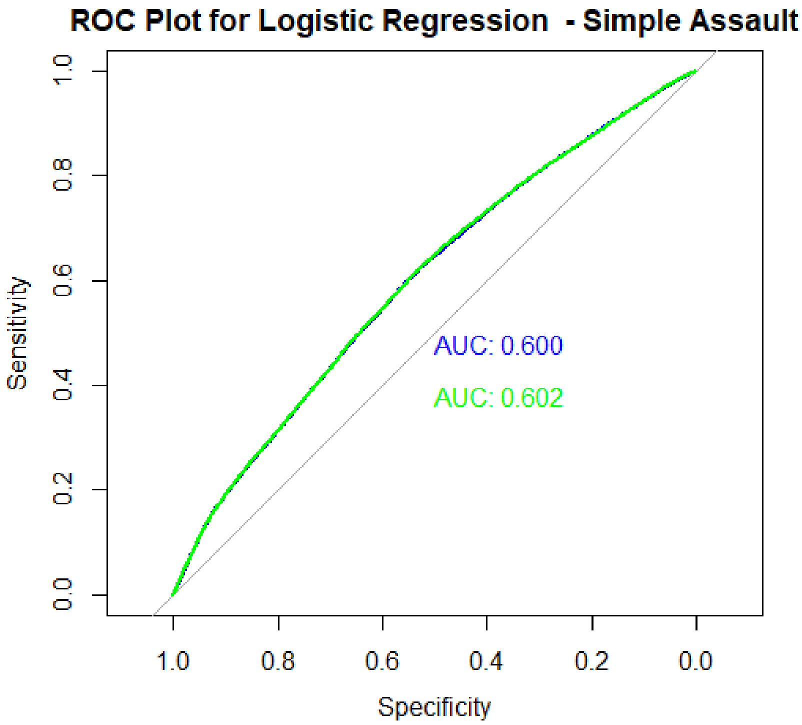


**Figure A0.8:** ROC Curve for Random Forest - Base and Adjusted Models (All Violent Crime)

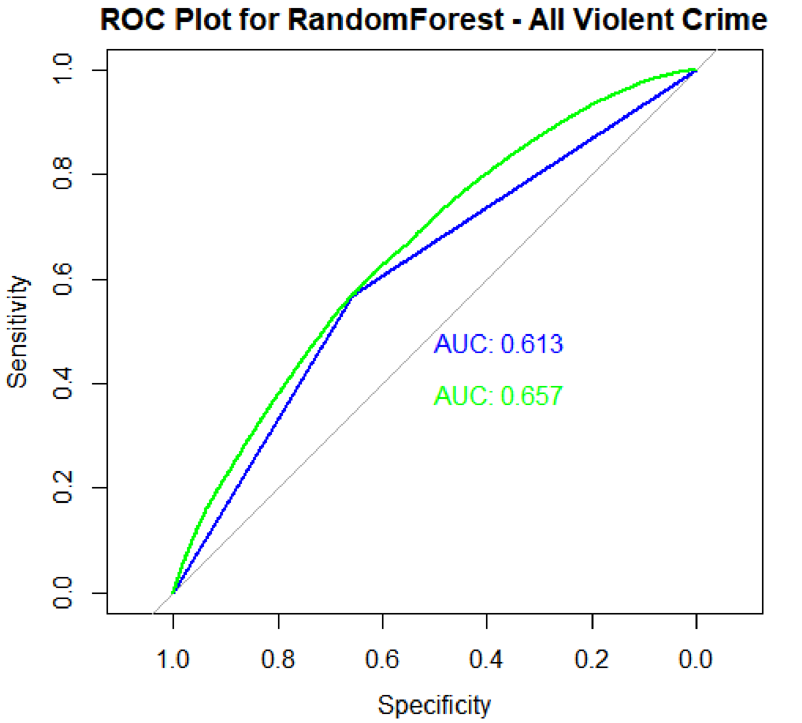


**Figure A0.9:** ROC Curve for Random Forest - Base and Adjusted Models (Aggravated Assault)

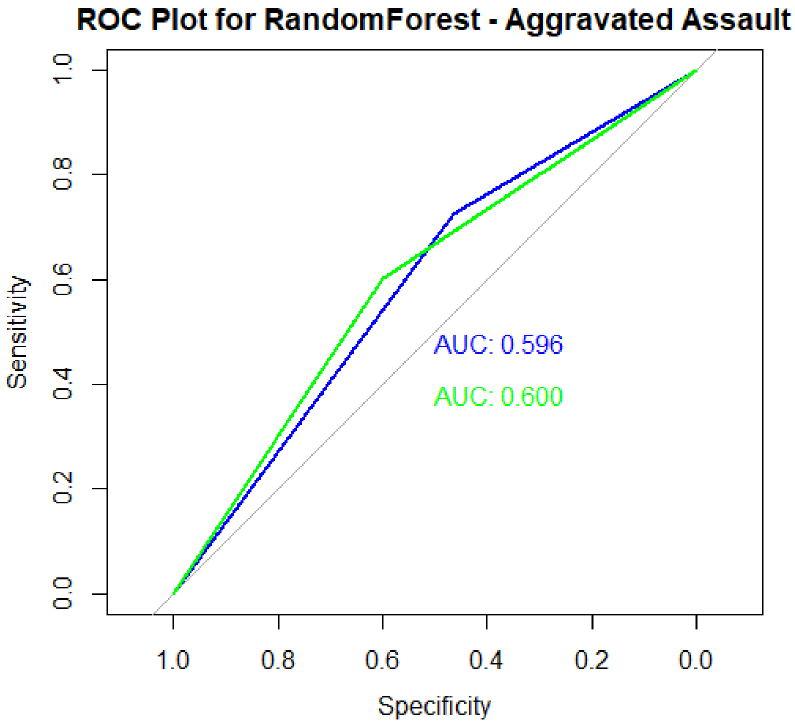


**Figure A0.10:** ROC Curve for Random Forest - Base and Adjusted Models (Sexual Offense)

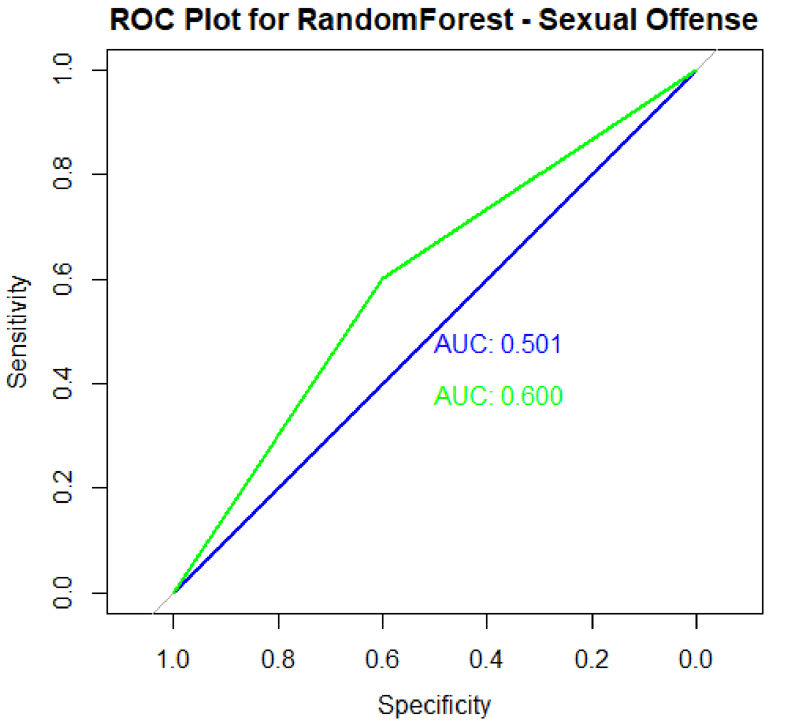


**Figure A0.11:** ROC Curve for Random Forest - Base and Adjusted Models (Robbery)

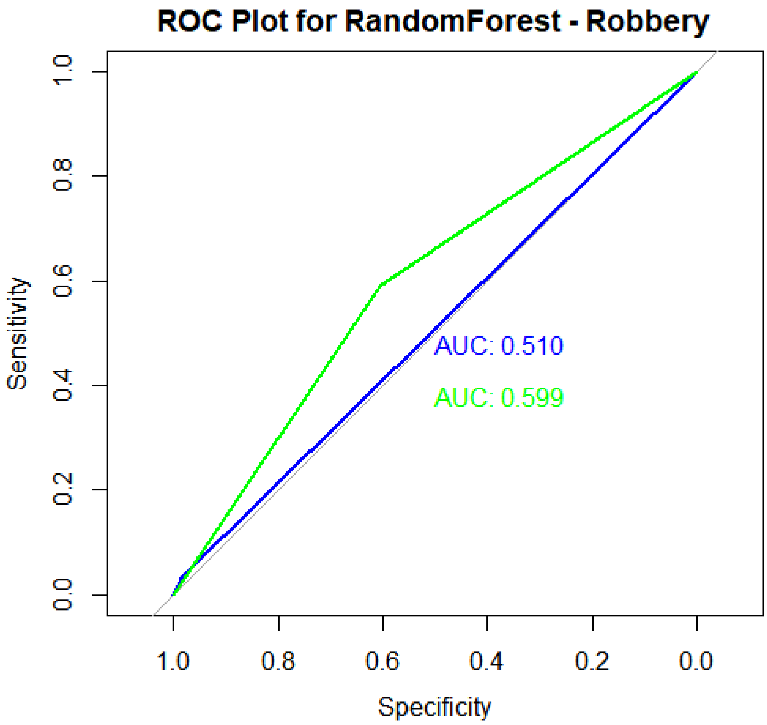


**Figure A0.12:** ROC Curve for Random Forest - Base and Adjusted Models (Simple Assault)

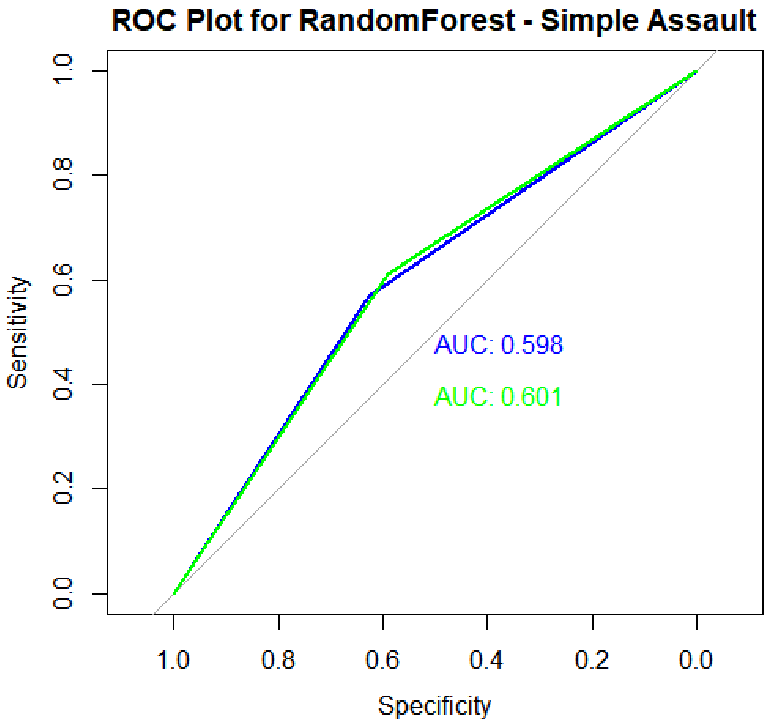


**Table A0.1:** Logistic Regression Coefficients predicting the probability of arrest by type of violent crimes, 2014

| | **Model 1** | | **Model 2** | | **Model 3** | | **Model 4** | | **Model 5** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **All Violent Crime** | | **Aggravated Assault** | | **Sexual Offense** | | **Robbery** | | **Simple Assault** | |
| Variables | Base | Adj. | Base | Adj. | Base | Adj. | Base | Adj. | Base | Adj. |
| Age of Offender | -0.004 *** | -0.005 *** | -0.000 *** | 0.000 | - | 0.004 *** | 0.006 ** | 0.006 * | -0.007 *** | -0.007 *** |
| Age of Victim | 0.007 *** | 0.007 *** | 0.008 *** | 0.008 *** | - | -0.008 *** | -0.000 | -0.000 | 0.008 *** | 0.008 *** |
| Offender Not Black | 0.177 *** | 0.183 *** | 0.177 *** | 0.173 *** | - | 0.047 | 0.311 *** | 0.360 *** | 0.189 *** | 0.192 *** |
| Victim Not Black | 0.299 *** | 0.302 *** | 0.344 *** | 0.335 *** | - | 0.036 | 0.406 *** | 0.286 *** | 0.299 *** | 0.294 *** |
| Offender Male | -0.024 *** | -0.020 *** | -0.155 *** | -0.136 *** | - | 0.543 *** | -0.094 | -0.020 | -0.008 | -0.001 |
| Victim Male | -0.096 *** | -0.091 *** | -0.151 *** | -0.147 *** | - | -0.111 * | 0.001 | 0.100 | -0.090 *** | -0.094 *** |
| Offender Stranger (Y) | -0.257 *** | -0.252 *** | -0.389 *** | -0.407 *** | - | 0.013 | -0.403 *** | -0.414 *** | -0.217 *** | -0.207 *** |
| Offender Substance Use (Y) | 0.615 *** | 0.611 *** | 0.561 *** | 0.556 *** | - | 0.126 * | 0.410 *** | 0.547 *** | 0.656 *** | 0.655 *** |
| Deadly Weapon (Y) | -0.229 *** | -0.213 *** | -0.239 *** | -0.240 *** | - | 0.414 ** | -0.142 ** | -0.159 ** | - | - |
| Time of Day - Night time | 0.088 *** | 0.085 *** | 0.112 *** | 0.112*** | - | -0.002 | -0.110 * | -0.083 | 0.089 *** | 0.090 *** |
| UCR Code Sexual Offense | -1.540 *** | -1.535 *** | - | - | - | - | - | - | - | - |
| UCR Code Robbery | -1.054 *** | -1.054 *** | - | - | - | - | - | - | - | - |
| UCR Code Simple Assault | -0.395 *** | -0.382 *** | - | - | - | - | - | - | - | - |
| Intercept | -0.134 | -0.022 | -0.183 | -0.354 | - | -0.576 | -1.193 | -0.185 | -0.524 | -0.463 |

**Table A0.2:** Log-odds predicting the probability of arrest by type of violent crimes, 2014

| | **Model 1** | | **Model 2** | | **Model 3** | | **Model 4** | | **Model 5** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **All Violent Crime** | | **Aggravated Assault** | | **Sexual Offenses** | | **Robbery** | | **Simple Assault** | |
| *Variables* | *Base* | *Adj* | *Base* | *Adj* | *Base* | *Adj* | *Base* | *Adj* | *Base* | *Adj* |
| Age of Offender | 0.999 | 0.995 | 0.999 | 1.000 | - | 1.004 | 1.006 | 1.006 | 0.992 | 0.993 |
| Age of Victim | 1.008 | 1.007 | 1.008 | 1.008 | - | 0.991 | 0.999 | 0.999 | 1.009 | 1.009 |
| Offender Not Black | 1.194 | 1.201 | 1.193 | 1.188 | - | 1.048 | 1.365 | 1.433 | 1.208 | 1.212 |
| Victim Not Black | 1.349 | 1.353 | 1.410 | 1.398 | - | 1.037 | 1.502 | 1.330 | 1.349 | 1.342 |
| Offender Male | 0.976 | 0.980 | 0.856 | 0.872 | - | 1.721 | 0.910 | 0.979 | 0.991 | 0.998 |
| Victim Male | 0.908 | 0.912 | 0.859 | 0.863 | - | 0.894 | 1.001 | 1.106 | 0.914 | 0.909 |
| Offender Stranger (Y) | 0.773 | 0.777 | 0.677 | 0.665 | - | 1.013 | 0.668 | 0.661 | 0.805 | 0.813 |
| Offender Substance Use (Y) | 1.849 | 1.843 | 1.751 | 1.749 | - | 1.135 | 1.507 | 1.729 | 1.928 | 1.926 |
| Deadly Weapon (Y) | 0.794 | 0.807 | 0.787 | 0.786 | - | 1.513 | 0.867 | 0.853 | - | - |
| Time of Day - Night time | 1.092 | 1.088 | 1.118 | 1.119 | - | 0.997 | 0.895 | 0.919 | 1.09 | 1.095 |
| UCR Code Sexual Offense | 0.214 | 0.215 | - | - | - | - | - | - | - | - |
| UCR Code Robbery | 0.348 | 0.348 | - | - | - | - | - | - | - | - |
| UCR Code Simple Assault | 0.673 | 0.682 | - | - | - | - | - | - | - | - |