

# University of Cincinnati

Date: 7/21/2021

I, Gifty A Arthur, hereby submit this original work as part of the requirements for the degree of Master of Science in Information Technology.

It is entitled:

**A Machine Learning Web Application for Predicting Neighborhood Safety in The City of Cincinnati**

Student's name: **Gifty A Arthur**

This work and its defense approved by:

Committee chair: M. Murat Ozer, Ph.D.

Committee member: Bilal Gonen, Ph.D.



40484

# **A Machine Learning Web Application for Predicting Neighborhood Safety in The City of Cincinnati**

A thesis submitted to the  
Graduate School  
of the University of Cincinnati  
in partial fulfillment of the  
requirements for the degree of

Master of Science

In the Department of School of Information Technology  
Of the College of Education, Criminal Justice, and Human Services

by

Gifty Ama Dua Arthur

B.A Luther College

May 2017

Committee Chair: Murat Ozer, Ph.D.

Committee Member: Bilal Gonen, Ph.D.

## ABSTRACT

Neighborhood safety is the seed or cell state of tackling a much bigger issue in large metropolitan cities like Cincinnati. The aim of this study is to present a novel approach for raising awareness of safety of a particular location at a specific time through a web application that can be easily assessed on mobile devices such as smart phones. This development serves as a contribution to new advanced geographic information systems today that can help tackle crime prediction problems in real-time, as criminal activities continue to evolve. The work presented explores various machine learning algorithms to determine how safe a neighborhood is by recommending a 'safety score' deduced from records of property crimes within the metropolitan area. To demonstrate the feasibility of this approach, the main focus is on real property crime data from the Cincinnati Police Department for 133,246 incidents of burglary or breaking entry, theft and unauthorized use from 2010 to 2021. The dataset was extracted live from the crime management portal where it is updated daily, and re-engineered to produce victim data to extract fine-grained information such as theft in a particular suburb and the victims involved. The proposed approach falls in line with addressing a much larger socio-economic issue since many previous research efforts have tackled the crime prediction problem by focusing on historical suspect data to determine repeat offenders. This research supports the hypothesis that victim data can be mined in combination with human behavioral activity through mobile devices such as smart phones to avoid repeat victimization. The end product is a responsive web-based application using *Streamlit*, that combines computed geocoded address with basic demographic information to predict the likelihood of an arrest at the location at a specific time, within the greater Cincinnati area. As the target class categories in the dataset are imbalanced, SMOTE oversampling method was used to solve the classification problem. The experimental results

after re-sampling the data helped Random Forest machine learning algorithm to outperform other algorithms with about 88% prediction accuracy. The main advantage in housing the machine learning project on a web server is that it provides a safety tool for users which can be easily accessed in their pocket. Machine learning algorithm offers a great predictive power for this kind of analysis and powerful web applications like *Streamlit* offer visual insights through the interactive interface to communicate relevant information with end-users. The application is deployed on the cloud through *Streamlit*'s cloud-sharing platform and made accessible on various platforms. In addition, this work provides a discussion on the implications of the findings and offers future research direction for data-driven crime analysis.

**Keywords:** crime prediction, real-time, Cincinnati, web application, Streamlit, smart phones



## ACKNOWLEDGMENTS

I would like to offer my sincere gratitude to my supervisor Murat Ozer, PhD. for supporting me throughout this process. I also want to thank my committee member Bilal Gonen, PhD. for taking keen interest in my work and for agreeing to help on this project.

A special thanks goes to Michael Zidar for his willingness probe my research with tough questions to narrow down my findings. A special thanks to Alfred Danquah and Seth Yeboah for their expertise on documentation and software development. Your contributions are of tremendous value and much appreciated.

I would also like to thank all those who shared their ideas and work on Medium and *Streamlit* open platforms, discussion forum and on YouTube.

# Table of Contents

ABSTRACT .....	ii
ACKNOWLEDGMENTS .....	ii
CHAPTER 1 .....	1
1. Introduction .....	1
1.1 Motivation and Background.....	1
1.2 Research Statement & Objectives.....	2
1.3 Approach .....	3
1.4 Scope and Limitation .....	4
1.5 Target group .....	5
1.6 Thesis Outline .....	5
CHAPTER 2 .....	6
2. Literature Review .....	6
CHAPTER 3 .....	12
3. Methodology .....	12
3.1 The Study Area .....	12
3.2 Preparing the Input Data .....	13
3.3 Introducing Streamlit .....	16
3.4 Machine Learning .....	17
3.5 Exploratory Data Analysis .....	18
3.6 Performance Metrics .....	30
3.7 Modeling .....	32
3.8 Model Selection .....	34

CHAPTER 4 .....	40
4.1 Evaluation.....	40
4.2 Prediction .....	41
4.2.1 Web Design - User Input Data .....	41
4.3 Results and Analysis .....	45
4.3.1 Probability of Arrest or No Arrest .....	45
4.3.2 Safety score, so what?.....	45
4.4 Recommendation.....	47
CHAPTER 5 .....	49
5.1 Conclusion.....	49
5.2 Future Work .....	51
REFERENCES .....	52



## Table of Figures

Figure 1: Map of Cincinnati's Neighborhoods <sup>[39]</sup> .....	12
Figure 2: SODA API call to crime dataset using Socrata .....	13
Figure 3: Snapshot of first five selected values for the crime data .....	18
Figure 4: Frequency of property crime types.....	20
Figure 5: Bar and line charts of crimes occurring in different years .....	21
Figure 6: Area graph of crime occurring in different months .....	22
Figure 7: Histogram of crimes occurring during different days of week .....	22
Figure 8: Crime occurring in different hours of the day (in 24 hours format).....	23
Figure 9: Map showing boundaries of reported crime coordinates .....	24
Figure 10: Value counts and histogram of victim gender .....	25
Figure 11: Value counts and histogram of victim race .....	26
Figure 12: Value counts and histogram of victim age .....	27
Figure 13: Snapshot of close code values for the crime data.....	28
Figure 14: Data frame showing Chi square score for best features .....	29
Figure 15: Heatmap showing feature correlation.....	30
Figure 16: Training and Test results after undersampling with TomekLinks .....	35
Figure 17: Confusion matrix of Random Forest algorithm for under sampled data.....	36
Figure 18: ROC AUC curve for Random Forest algorithm for under sampled data .....	36
Figure 19: Training and Test results after oversampling with SMOTE .....	37
Figure 20: Confusion matrix of Random Forest algorithm for over sampled data.....	38
Figure 21: ROC AUC curve of Random Forest algorithm for the over sampled data .....	38

Figure 22: Result of Random Forest Classifier on actual dataset using K-Fold cross validation.....	40
Figure 23: Segment with input parameters for collecting user information on Streamlit	41
Figure 24: Output of information entered by user in JSON format and on Pandas dataframe.....	42
Figure 25: Output of Safety Score and Feature Importance of best model trained on data .....	43
Figure 26: Screenshot of full User Prediction page on Streamlit web application .....	44

## **List of Tables**

Table 1: Attributes of the final crime dataset.....	19
Table 2: Frequency of Property Crime Category .....	20
Table 3: Result of best classifier (Random Forest) on balanced dataset .....	39

# CHAPTER 1

## 1. Introduction

### 1.1 Motivation and Background

Crime prevention is one of the keys to successful living in any society. Being able to assess how safe a location is within a city in at a specific time can serve as a powerful agent in the fight against crime and lead to a better feeling of security for the society with fewer law enforcement resources needed. A successful execution of this research can inform people on whether to move to a city or not, and where to avoid when traveling. One important way of achieving this community sense of security is by creating a system that can make extraction of crime information automatic (from street level, suburb level and city level) in order to engage in proactive ‘neighborhood crime watch’ <sup>[1]</sup> to reduce crime.

According to NeighborhoodScout Crime Risk Report, majority of the crimes that occur in Cincinnati are property crimes and the chances of being a victim is one in 23 <sup>[2]</sup>. Hence, if studies of criminal activities are centered around recognizing patterns relating to property crimes and making information available to users in real-time, opportunities for crime to occur will greatly reduce and thus tackle the overall crime rate in the area. The two main goals of this research, improved technology and accessibility, are in line with strategies to assist government’s efforts in making Cincinnati a smart city <sup>[3]</sup>. The insights gleaned can be valuable information for the allocation of resources to aid in policing strategies to prevent crime before it happens <sup>[4, 17]</sup>. It also serves to empower residents with the preventive tool they need to avoid crime traps <sup>[15]</sup>. The data source was from the Cincinnati’s Open Data host Socrata, which was retrieved in real-time from the web portal where it is updated daily by the Cincinnati Police Department <sup>[5]</sup>.

Based on the reasons above, the crime data was processed and analyzed to identify patterns and to predict future occurrences. Hosted on *Streamlit*’s server, end-users can have live

interaction with the model on their technological devices. The main idea of this data-crime analysis is to ensure that a sustainable community like Cincinnati is safe and perceived by its residents to be safe <sup>[21, 22]</sup>. The process serves as a way of enhancing policing strategies and empowering residents, thereby improving the quality of life for people <sup>[6]</sup>. Generating a safety score for a location gives an indication to the state of affairs within the city is based on the research idea that probability outcomes forecast likely probation violations <sup>[16]</sup> and predictive modeling can support decision-making and assist with the allocation of resources for prevention strategies and management of perceived risk in the society <sup>[7, 8]</sup>.

## **1.2 Research Statement & Objectives**

The use of historical crime data appeals to the school of thought that although crime could occur anywhere, crime activities and opportunities occur at places most familiar to criminals, providing an idea about their *Activity Space* <sup>[9]</sup>. Data from this familiar territory of crime activity can inform users on how to conduct themselves to mitigate risk. The conventional crime analysis used in this research will determine when areas will be most at risk of crime by the frequency of crimes in a given area by time and date. In this thesis, a machine learning technique will be compared among others to select the top ranking as the main engine for predicting location safety score from a user's input parameters for a specific location within a particular time combined with basic demographic data such as age, gender and race. Further, an approach for evaluating the quality of the model will be proposed. The problem relating to safety based on property crime is addressed in two distinct parts:

1. Exploratory data analysis to mine patterns in crime data.
  - Can patterns inferred from type of property crime offer insights into the crime nature, severity, location, duration and frequency?

- What are the seasonal changes in the crime rate and type?
  - Can possible areas of victimization based on geographical location be identified?
2. Build a prediction model that can accurately inform the likely occurrence of an arrest at a location, thus making it not safe.
- Does the database from Cincinnati Police Department have enough indicators to predict the likely occurrence of a property crime?
  - Which machine learning model is the most accurate in predicting property crime?
  - Can the model determine when areas will be at most risk of crime?
  - Can machine learning algorithm deduce profile of repeat victims?
  - Will the accuracy of the model inform the perception of safety for the people of Cincinnati and the government?

Both parts of the problem will make use of *Streamlit*'s powerful web framework for data preprocessing, visualization and analyzation of the spread of crime in Cincinnati. The prediction model will build on existing work and improve their results by examining different types of algorithms.

### **1.3 Approach**

This work focuses on using machine learning methods and algorithms in order to evaluate the outcome of a reported crime resulting in an arrest. First, data features will be filtered to mine victim information for property crimes. In the second attempt, the probability of an arrest occurring based on the frequency of reported incidents by crime victims, will serve as the metric for determining safety score.

The procedure for implementation will be tackled iteratively to answer the research questions above: The quality and relevance of data features for proposed model in a dataset, exploratory data analysis of relevant data features to mine patterns and for visual information, model training comparing six different machine learning algorithms to determine the best fit. The process and results will be reviewed, evaluated and compared to one another. Selected algorithm will be used as the model agent to predict location safety score based on user's input values. Visualization is by use of dot maps and will be based on spatial distribution of the crimes by cluster dots <sup>[10]</sup>.

#### **1.4 Scope and Limitation**

Due to time constraint and the newness of this proposed machine learning agent for the city's crime data, some limitations have been set to ensure that the work is finished in time.

- Dataset attributes are hand-selected and re-engineered to account for victim crime report only
- Target variable Close Code 'clsd' is encoded to indicate CLEARED BY ARREST – ADULT and CLEARED BY ARREST – JUVENILE as True, leaving the rest of the eleven unique values as False. The category encoding results in an imbalanced dataset which is addressed using the SMOTE oversampling technique.
- The algorithm will not cater to geocoded locations outside the geographical boundaries of the greater Cincinnati metropolitan area in Ohio, USA.
- The threshold for measuring safety will be compared against the overall property crime rate in Cincinnati based on 2021 data.

## **1.5 Target group**

This work is especially interesting for researchers in the area of perceived and actual risk assessment, large city neighborhood evaluation and predictive policing. The private interest group for this research are the police department, real estate agencies and international or tourist companies. Overall, this work is of interest to all kinds of customers <sup>[11]</sup> since variety of user data is captured within the model framework.

## **1.6 Thesis Outline**

The introduction chapter of this report is devoted to the motivation, purpose and problem formulation, scope and limitation of the research work given the time frame, and the intended target groups for the final product. The following Chapter 2 addresses the background and literature review for predictive crime analysis, employing knowledge discovery process, machine learning approaches, and web interactive visualization. After describing the methodology, the design, empirical data and respective results for the model will be shown in Chapter 3. Chapter 4 will examine the validity and reliability of presented results. It will introduce the web application and combine user's input data on the best model fit. Finally, Chapter 5 will summarize the work and offer possible future research and expansion this topic.



## CHAPTER 2

### 2. Literature Review

Until recently, crime data was collected and simply kept as records of crimes and to justify imprisonment. Not much concern was given to the data itself but only on the history and actions of a person. For a long time, crime was considered unique and random, and in some cases an act of opportunity and so crime forecasting made no sense since it is impossible to predict random events. This meant that there were no general patterns and or link between crime types since there was none. However, as time passed and collecting more data became a thing due to the advent of big data and storage mechanisms especially for crime locations, this problem became a thing of the past. Recent research shows that there are patterns within data and that crimes are not random and unique in every case <sup>[25]</sup>. With more data being recorded and produced each day, the need of massive computations to reduce this data to manageable sizes has also increased. Machine learning along with other computational machines have accelerated this automation in recent years.

Assessing the risk of a location using machine learning algorithms and data collection methods is not a new thing. Countless of work has been done to related to crime, with the new focus being able to predict when and where crime will happen. The study of the connection between crime and geography has been in the works for over nearly two centuries. In the early 19<sup>th</sup> century, Guerry (1833) identified a link between higher property crime rates and affluent locations <sup>[13]</sup>. He was the first to conduct comprehensive study relating to social demographics by analyzing data on crimes, suicides, literacy and other demographic statistics. On the side, Quetelet (1831, 1835) led his renowned work in Europe for the foundation of criminology and sociology with the study of “moral statistics”. The extensive works of Guerry and Quetelet are considered to be the foundation of modern social science <sup>[14]</sup>.

According to an analysis by Washington University Law Review on Policing Predictive Policing, identifying a future location of a criminal activity may be statistically possible by studying where and why past crime activities happened over time <sup>[18]</sup> although identifying the human “criminal” may be more challenging. The article offers three insights to the literature of predictive policing where the second discussion examines the rapid evolution from place-based property crimes to placed-based violent crimes and then to person-based crimes <sup>[19]</sup>. These factors prove how difficult the task is for a working solution to the prediction problem. However, over time the problem has been examined by new research and technologies through the introduction of Geographic Information Systems (GIS) to aid in data collection to establish baseline to make forecasting possible <sup>[25]</sup>.

In another work on predictive policing by Walter Perry et al., the study expands on four broad categories of predictive methods: methods for predicting crimes, methods for predicting offenders, methods for predicting perpetrators’ identities and methods for predicting victims of crimes. The fourth category which is relevant to this thesis, focuses on predictions about crime and its victims, touching on when and where it is most likely to occur, what is likely to cause it and who is likely to be its victim. The article explores both tactical (next incident) or strategic (long term) on past data to identify patterns. For each what, where, and when question about crime data, methods discussed include hotspot analysis, statistical regression, data mining and near-repeat for phase one, temporal and spatiotemporal methods for ambient populations for phase two and risk terrain analysis for geospatial factors for phase three. This thesis uses hotspot analysis to set the undertone for where a crime will occur and who is likely to be a victim <sup>[6]</sup>.

In a book entitled the Paradoxical reactions of property crime victims by Scott Beach and Martin Greenberg, the writers explore how far more attention has been given in previous work on reactions to crimes of violence such as rape, robbery and assault than on reactions to property

crimes such as burglary and theft <sup>[23]</sup>. They drew on research work done in greater Pittsburgh (United States) on victims of residential burglary, theft and non-victims during a 12-month period, prior to having two waves of phone interviews with them. The assessment was on victims' immediate emotions, beliefs and behavior and experience on their mental health. They found three conclusions that suggest that victims studied wanted control over future crime victimization and access to the necessary security precautions to enhance their perception of safety <sup>[24]</sup>. This thesis adds user demographics to its predictions to provide users the control highlighted from the research and improve their quality of life in Cincinnati neighborhoods.

The novel approach of using hotspot forecasting along with human behavioral activity on mobile devices is highlighted in Andrey Bogomolov et al.'s paper on Crime prediction from demographics and mobile data <sup>[28]</sup>. They propose that aggregated and anonymized human behavioral data derived from mobile network activity is the novel contribution that can tackle the crime prediction problem. The paper argues that since the number of mobile phones actively in use is approaching the 7 billion mark <sup>[29]</sup>, they become a valuable and persuasive source of human behavioral data. Mobile phones can be seen as sensors of aggregated human activity <sup>[30, 12]</sup> and in various cases have been used to monitor human mobility patterns and urban interactions <sup>[31, 32]</sup>. In this thesis human behavioral activity through interaction with the web app is used as a catalyst to increase awareness and responsiveness via mobile activity, together with open data related to crime events to predict crime hotspots in neighborhoods of Cincinnati metropolitan area.

Many works have preceded the idea behind this model. For instance, Naik et al. of MIT Media Lab built a dataset that had millions of online volunteers' perception of over 100,000 Google Street View images worldwide <sup>[8]</sup>. The crowd-sourced dataset was trained to develop a

model based on algorithms in computer vision for scenery, and in turn produce a predicted score of the perceived level of safety, wealth, liveliness or depression in the geography context.

Wang et al. in the paper Learning to Detect Patterns of Crime, proposed a pattern detection algorithm called Series Finder that is able to grow a pattern of data crimes from a ‘seed’ of a few crimes <sup>[33]</sup>. The algorithm was found to provide a more accurate prediction of crimes committed in the city of Cambridge that were missed by analysts. The algorithm focused more on determining repeat offender rather than the geographical “hotspot” locations. This machine learning data mining approach was set to complement the work of human analysts by use of social media.

Babak Jahromi in his project, Predicting Neighborhood Safety using Boosting Machine Algorithm <sup>[34]</sup>, predicts safety in the City of Chicago using the machine learning methodology to determine the likelihood of a crime resulting in an arrest at a specific location. The attributes of the dataset were related to crime date and time, and crime location. The end product was an application that was able to generate a safety score and map out all the crimes within 100 meters radius in the area location. Safety prediction in the city of Cincinnati adds on to this project idea by including demographic data to tailor the safety score to the likelihood of victimization at a specific time. Attributes like victim age and gender are part of the feature selection for this research model.

Angelov et al. assessed how different types crimes affect the sales price of a residential value in Pierce County, Washington USA. Combining data consisting of physical properties of a building and crime data for the county from July 2018 to July 2019, the team built 32 models and evaluated them with three machine learning algorithms – decision tree, artificial neural networks, and random forests. The models were evaluated based on prediction errors and were

concluded to have a high correlation in predicting the sales price of residential properties in Pierce County, Washington <sup>[35]</sup>.

Saltos and Cocea conducted an exploration of crime prediction using data mining on open data. Their paper explored models for predicting the frequency of several types of crimes by LSOA code (Lower Layer Super Output Areas - an administrative system of areas used by the UK Police) and the frequency of anti-social behavior crimes <sup>[35]</sup>. Conducting the data on over 600,000 records of crime data before preprocessing, three models, decision trees, instance-based learning and regression were evaluated on predictive performance and processing time. In the end decision tree algorithm was found to reliably predict crime frequency as well as anti-social behavior frequency as part of the measures to fight crime.

*Crime rate* is the statistic used to summarize the quantity and extent of criminal events. The appropriate denominator selected for this research is population data compiled from the data source compared with the overall crime rate for the city of Cincinnati. Studies have often restricted to using residential population data which Nick Malleson and Martin A. Andresen propose as ‘inadequate’ to describe the true ambient population, leading to misleadingly high or low crime rates <sup>[26]</sup>. Their research used ‘crowd-sourced’ data to measure the ambient population, which is the number of potential victims present at the time of the offense. The data source was messages generated on mobile devices (such as smart phones) and posted to the Twitter social media service as the ambient population. The goal was to identify spatio-temporal clusters of crimes that are significant even after taking account of the ambient population. The degree to which social media truly represented the population under study was considered as preliminary, leaving room for expansion and much scrutiny for future works. This thesis builds a model using the traditional data compiled from the historical crime data and marries it with temporarily dynamic mobile phone activity to determine the crime rate (probability of an arrest), which in

this case is translated as the safety score of the location. Residential population is the appropriate population at risk, given that property crimes usually occur at residential locations when the population is largely expected to not be at home. However, there is the need to account for ambient population estimates to assess risk at different times of the day <sup>[27]</sup> since some of these crimes, like theft, occur on the streets.

One problem with predictive crime modeling is being able to accurately measure the impact of various methods mentioned. Due to the successes of these methods over the years, or events like holidays and neighborhood watches or emigration of residents, crime rates continue to reduce across the nation. An analysis by Pew Research Center reveals that violent and property crime rates have plunged since 1990s regardless of data source <sup>[37]</sup>. This makes it hard to tell how successful crime predictive models are <sup>[36]</sup>, however this work seeks to complement existing works and the above-mentioned research efforts and contributions for increased responsiveness and expansion in other areas of crime data-analysis.

## CHAPTER 3

### 3. Methodology

The following chapter presents the parameter and methodology of the research. The first part introduces the study area and then moves to the preparation of the input data. Section 3.3 gives a short introduction to Streamlit web application and its embedded programming language Python, which is used for the prediction model. In section 3.4 Machine Learning and the concept of Supervised Learning are discussed. Section 3.5 and 3.6 are about data exploratory analysis and performance indicators. The last part is for data modeling and web design.

#### 3.1 The Study Area

The area studied is the City of Cincinnati, Ohio, USA. Cincinnati is about 202 years old and divided into five districts with 52 official neighborhoods (Figure 1). The total city area covers about 7.9.56 mi<sup>2</sup> and had a population estimate of about 303, 940 as of July 1, 2019 [38].

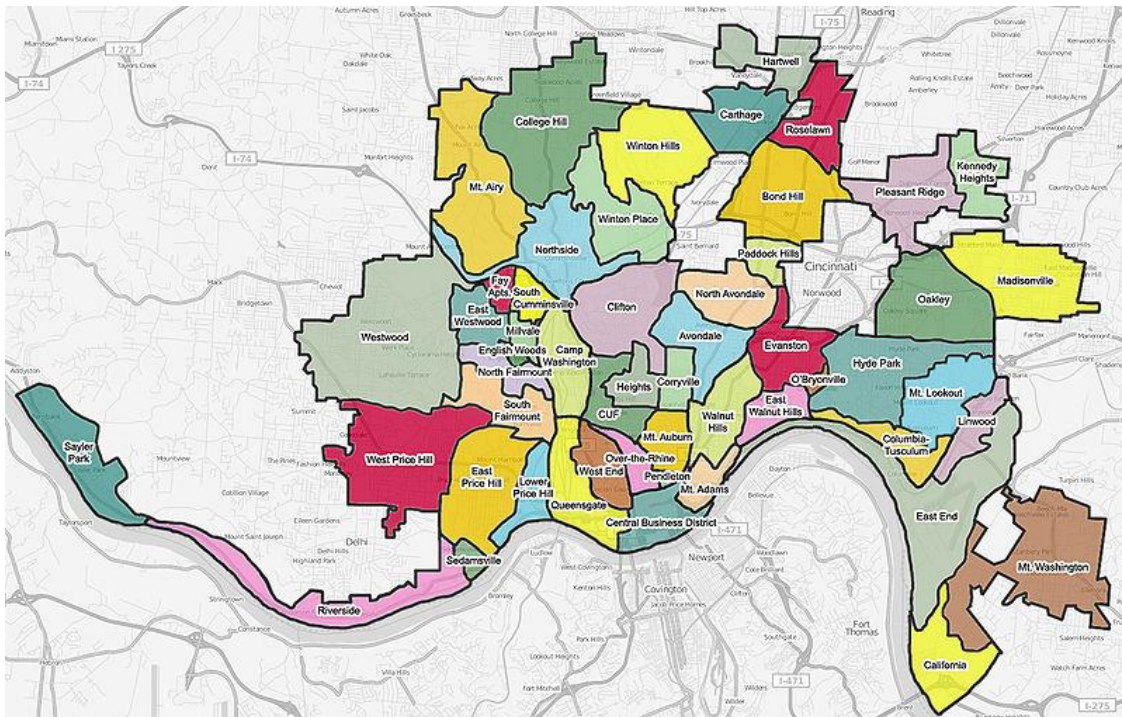


Figure 1: Map of Cincinnati's Neighborhoods [39]

### 3.2 Preparing the Input Data

The dataset used in this research project is real from City of Cincinnati crime dataset made available by the Cincinnati Police Department. The dataset can be found on the CincyInsights Open Data website<sup>[3]</sup> as part of the open data initiative. For this research, the dataset is extracted live from the web portal (Figure 2) where it is updated daily to reflect current reports. The Socrata Open Data API (SODA) provides programmatic access to the dataset including the ability to filter, query, and aggregate data.

```
# Load data
@st.cache(suppress_st_warning=True)
def load_data():
    client = Socrata("data.cincinnati-oh.gov", None)
    crimes = client.get("k59e-2pvf", limit=500000)
    crimes_na = pd.DataFrame.from_records(crimes)
    crimes_df = crimes_na.replace(r'^\s*$', np.nan, regex=True)
```

*Figure 2: SODA API call to crime dataset using Socrata*

The original dataset provides information on crime incidents that occurred in Cincinnati for the period of 1991-12-31 00:00:00-05:00 to 2021-04-27 03:14:00-04:00 with 428788 rows (at the time of writing) before preprocessing. Since the app makes an API call to the data source, the dataset and fields are updated daily in real-time. Below are the descriptions of each of the attributes:

- **InstanceId:** This is a text field. It refers to the incident id given to the crime reported. It is the analogous to the row number.
- **Incident\_No:** This is a text field. It refers to the incident number given to the crime reported. It is also similar to the row number.



- **Date\_Reported:** This is a Date-Time field. It specifies the exact date the crime was reported.
- **Date\_From:** This is a Date-Time field. It specifies the exact date the crime investigation was started.
- **Date\_To:** This is a Date-Time field. It specifies the exact date the crime investigation ended.
- **CLSD:** This is a text field. Specifies the resolution of the incident. Originally, there are 13 distinct values (such as *F--CLEARED BY ARREST - ADULT*, *K--UNFOUNDED*, *J--CLOSED*, *G—CLEARED BY ARREST – JUVENILE*, etc.)
- **UCR:** This is a numeric field. It gives the crime category code reported incident (such as *551.*, *810.*, *600.*, *301.*, *1493.*, *552.*, *201.*, *303.*, *401.*, etc.)
- **DST:** This is a text field. There are about 7 distinct values in the dataset (such as *'2'*, *'3'*, *'5'*, *'4'*, *'1'*, *'CENTRAL BUSINESS'*, *'OTHER'*, etc.)
- **BEAT:** This is a text field. There are about 14 distinct values in the dataset (such as *1.0*, *2.0*, *6.0*, *4.0*, *nan*, *3.0*, *5.0*, *'3'*, *'2'*, *'1'*, *'5'*, *'6'*, *'4'*, etc.)
- **Offense:** This is a text field. There are over 100 distinct values in the dataset (such as *'DISCHARGE FIREARM ON/NEAR PROHIBITED PREMISES'*, etc.)
- **Location:** This is a text field. It gives the crime location. Originally, there are over 100 distinct values in the dataset (such as *'02-MULTI FAMILY'*, *'47-STREET'*, *'01-SINGLE FAMILY HOME'*, *'26-BAR'*, *43-YARD*, *48-PARKING LOT*, *01-SINGLE FAMILY 2 STORY*, etc.)
- **Theft\_Code:** This is a text field. It gives the theft code of the crime category. Originally, there are over 20 attributes (such as *23D-THEFT FROM BUILDING*, *23-C SHOPLIFTING*, *24I-THEFT OF LICENSE PLATE*, etc.)
- **Floor:** This is a text field. It gives detail about crime scene inside a building. Originally, there are over 10 attributes (such as *1-BASEMENT*, *2-FIRST FLOOR*, *4 – OTHER*, etc.)
- **Side:** This is a text field. This gives detail about part of location where crime scene occurred. Originally, there are over 12 distinct values in the dataset (such as *5 – OTHER*, *3-SIDE*, *4-ROOF*, *2-SIDE*, *3-REAR*, etc.)
- **Opening:** This is a text field. Originally, there are about 13 distinct values (such as *'1 - DOOR'*, *nan*, *'2-WINDOW'*, *'1-DOOR'*, *'5-OTHER'*, *'3-GARAGE'*, etc.)

- **Hate\_Bias:** This is a text field. Originally, there are about 30 distinct values in the dataset (such as *13--ANTI AMERICAN INDIAN/ALASKAN NATIVE*, *89--GANG RELATED*, *88--DOMESTIC VIOLENCE*, *31--ANTI-ARAB*, *etc.*)
- **DayOfWeek:** This is a text field. Day of the week that crime occurred. It takes on one of the values from: *Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday*.
- **RPT\_Area:** This is a text field. It reports the RPT location area where crime is recorded.
- **CPD\_Neighborhood:** This is a text field. Cincinnati Police Department Neighborhood.
- **SNA\_Neighborhood:** This is a text field. Statistical Neighborhood Approximations.
- **Weapons:** This is a text field. Specifies the types of weapons used at the scene of the crime. Originally, there are over 60 distinct values (such as *12 – AUTOMATIC HANDGUN*, *14 – SHOTGUN*, *70 – DRUGS/NARCOTICS/SLEEPING PILLS*, *etc.*)
- **Date\_Of\_Clearance:** This is a Date-Time field. It gives the date that incident was cleared on record.
- **Hour\_From:** This is a date-time field. It gives the hour that the incident investigation started.
- **Hour\_To:** This is a date-time field. It gives the hour that the investigation ended.
- **Address\_X:** This is a text field. It gives the street address of the crime.
- **Longitude\_X:** This is a geographic field. It gives the longitudinal coordinate of the crime.
- **Latitude\_X:** This is a geographic field. It gives the latitudinal coordinate of the crime
- **Victim\_Age:** This is a text field. It tells the age of the victim at crime scene.
- **Victim\_Race:** This is a text field. It gives the race of the victim at crime scene.
- **Victim\_Ethnicity:** This is a text field. It gives the ethnicity of the victim. Originally, there are over 17 distinct values (such as *AFRICAN AMERICAN*, *CHINESE*, *EUROPEAN*, *HISPANIC ORIGIN*, *PAKASTANI*, *etc.*)
- **Victim\_Gender:** This is a text field. It gives the gender of the victim.
- **Suspect\_Age:** This is a text field. It gives the age of the suspect caught at the crime scene.
- **Suspect\_Race:** This is a text field. It gives the race of the suspect caught at the crime scene. Originally, there are over 10 distinct values in the dataset (such as *WHITE*, *BLACK*, *HISPANIC*, *AMERICAN INDIAN/ALA*, *etc.*)

- **Suspect\_Ethnicity:** This is a text field. It gives the ethnicity of the suspect caught at the crime scene. Originally, there are over 17 distinct values (such as *AMERICAN INDIAN*, *AFRICAN AMERICAN*, *CHINESE*, *EUROPEAN*, *HISPANIC ORIGIN*, *etc.*)
- **Suspect\_Gender:** This is a text field. It gives the gender of the suspect caught at the crime scene.
- **TotalNumberVictims:** This is an int field. It tells the total number of victims reported at crime scene.
- **Total\_Suspects:** This is an int field. Total number of suspects reported at crime scene.
- **UCR\_Group:** This is a text field. It gives the crime category of the reported crime. Originally, there are 8 distinct values (such as *BURGLARY/BREAKING ENTERING*, *PART 2 MINOR*, *THEFT*, *ROBBERY*, *RAPE*, *UNAUTHORIZED USE*, *etc.*)
- **Community\_Council\_Neighborhood:** This is a text field. This tells the community neighborhood where crime is recorded.
- **Zip:** This is a numerical field. It gives the zip code of the location of crime scene.

### 3.3 Introducing Streamlit

Streamlit<sup>[41]</sup> is the front-end tool for hosting the machine learning code. It is a Python-based web application that has numerous widgets to create user-friendly experience. By far, Streamlit is the fastest way to build data apps and share them. With the application there is no need to learn Frontend, Backend, Django, Flask, Heroku or another web framework. Basically, it reduces the entire model deployment significantly by seamlessly integrating the frontend and backend of a working model to create beautiful, performant apps in a few hours. This is all built in Python and is entirely free. Streamlit was founded by Adrien Treuille in 2018 along with two other Google X engineers and in the short amount of time since its inception, Streamlit is fast growing as the preferred interactive tool for machine learning engineers. Streamlit also has a web hosting platform for sharing completed work project<sup>[40]</sup>. To install Streamlit, open terminal or command prompt and type -> \$ pip install streamlit.

Python is a free and applicable programming language invented by Guido van Rossum in the 1990s. The language contains vast standard library, with new modules and resources that allows it to be widely applicable. The Python interactive interpreter allows real-time code development due to its expressive syntax and availability <sup>[20]</sup>. For this thesis, the entire project was built using PyCharm as the IDE for the Python code and pushed to GitHub <sup>[42]</sup>.

### **3.4 Machine Learning**

Machine Learning (ML) is a branch of artificial intelligence that employs statistical methods to teach computers to learn from past experiences. It is a science that teaches computers to make decisions without human intervention. Machine learning along with other computational machines have accelerated automotive analysis. Machine learning algorithms can be classified under supervised, unsupervised and reinforcement learning. This study employs supervised learning because of the model design's required target output and select input features. Under supervised learning, classification predicts a discrete class label whereas regression predicts a continuous quantity. This project study attempts to predict the likelihood of an arrest at a particular location, hence the use of classification.

Supervised Learning (SL) is a branch of machine learning task that predicts outputs from a set of inputs. The machine is fed with expected outcomes (target variable), so it knows what to look out for. Separating training and testing data helps supervised learning to avoid overfitting. Supervised learning algorithm can be used to solve regression and classification problems. For this dataset, the goal is to predict the category of the close code of the reported crime incident. This groups the crime analysis problem under classification. Based on the defined goals and objectives centered around supervised learning the following machine learning algorithms were


used: k-Nearest Neighbor, Support Vector Machine, Decision Tree, Logistic Regression, Naïve Bayes and Random Forest.

### 3.5 Exploratory Data Analysis

#### 3.5.1 Feature selection

To make the data relevant for the study topic of this thesis, the data is filtered, cleaned (removed missing values) and reengineered (encoded) to extract victim crime information. In essence the final dataset derived tells the story about a who had the crime occurrence, when it happened, where it happened and what was the outcome. New features were also extracted from the original attributes to provide useful information for the model. The total number of features after completing preprocessing was eleven and this was based on human assumption that some features are more useful than others. Table 1 describes the final attributes and Figure 3 shows the attributes as displayed on the web application.

#### Information about dataset



Snapshot of Data

	0	1	2	3	4
year	2017	2019	2013	2017	2014
month	2	7	8	3	10
day	26	15	17	28	26
hour	9	13	23	8	3
dayofweek	4	5	5	0	5
victim_age	4	5	3	3	4
victim_race	0	0	0	0	1
victim_gender	1	1	0	0	0
lon	-84.5592	-84.5674	-84.5151	-84.5207	-84.6095
lat	39.1986	39.1033	39.1215	39.1480	39.1433
clsd	0	0	0	0	0

Shape of dataset (133247, 11)

Figure 3: Snapshot of first five selected values for the crime data

year	Year from timestamp when crime occurred
month	Month from timestamp when crime occurred
day	Day from timestamp when crime occurred
hour	Hour from timestamp when crime occurred
minute	Minute from timestamp when crime occurred
dayofweek	Day of week
victim_age	Age of victim at the location of the crime
victim_race	Race of victim at the location of the crime
victim_gender	Gender of victim at the location of the crime
lon	Signifies the longitude of the location coordinates of the crime
lat	Signifies the latitude of the location coordinates of the crime
ucr_group	Type of property crime
clsd	How the crime was resolved. This is the target label for the data.  There are two types of resolutions: Arrest or No arrest.

*Table 1: Attributes of the final crime dataset*

### **3.5.1 Crime Category or UCR Group**

Three main subcategories of crimes were analyzed as they relate to property crimes in the Cincinnati Crime Dataset. These are charge codes related Part I property crimes. These crime types are considered as classes and having three distinct classes makes it a multi-class problem. Table 2 shows the crime labels and their frequencies for incidents reported between 2010 to 2021. Figure 4 gives a visual representation for the table data. As seen, reported theft cases are significantly higher than burglary/breaking entry and unauthorized use.

Crime Category	Frequency
THEFT	145121
BURGLARY/BREAKING ENTERING	56604
UNAUTHORIZED USE	2938

Table 2: Frequency of Property Crime Category

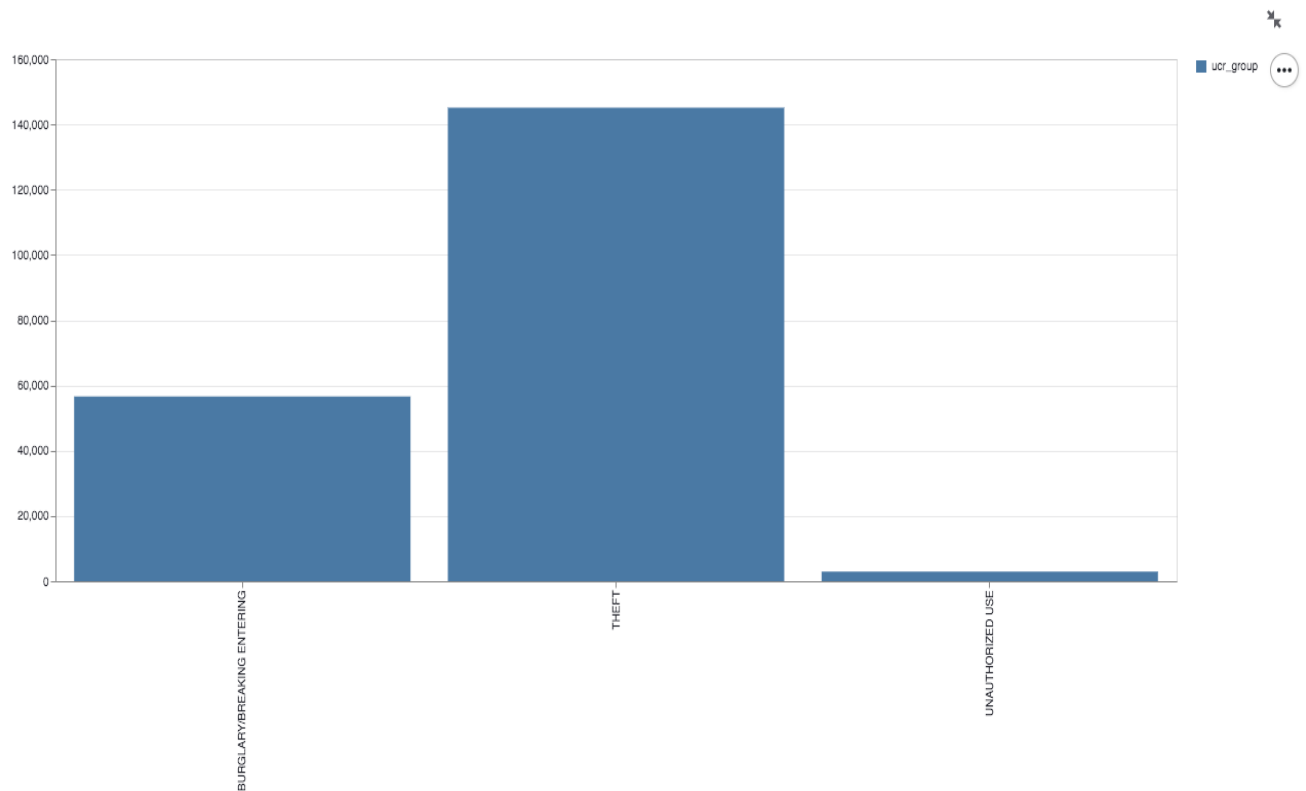


Figure 4: Frequency of property crime types

### 3.5.2 Time of Crime Incident

Reviewing visualizations of crime charts for the 'date\_reported' attribute of the original dataset, information can be gleaned about the seasonal changes of crimes for specific times throughout

the day. From the datetime stamp, five main features were extracted - year, month, date, hour and minute.

### 3.5.2.1 Year

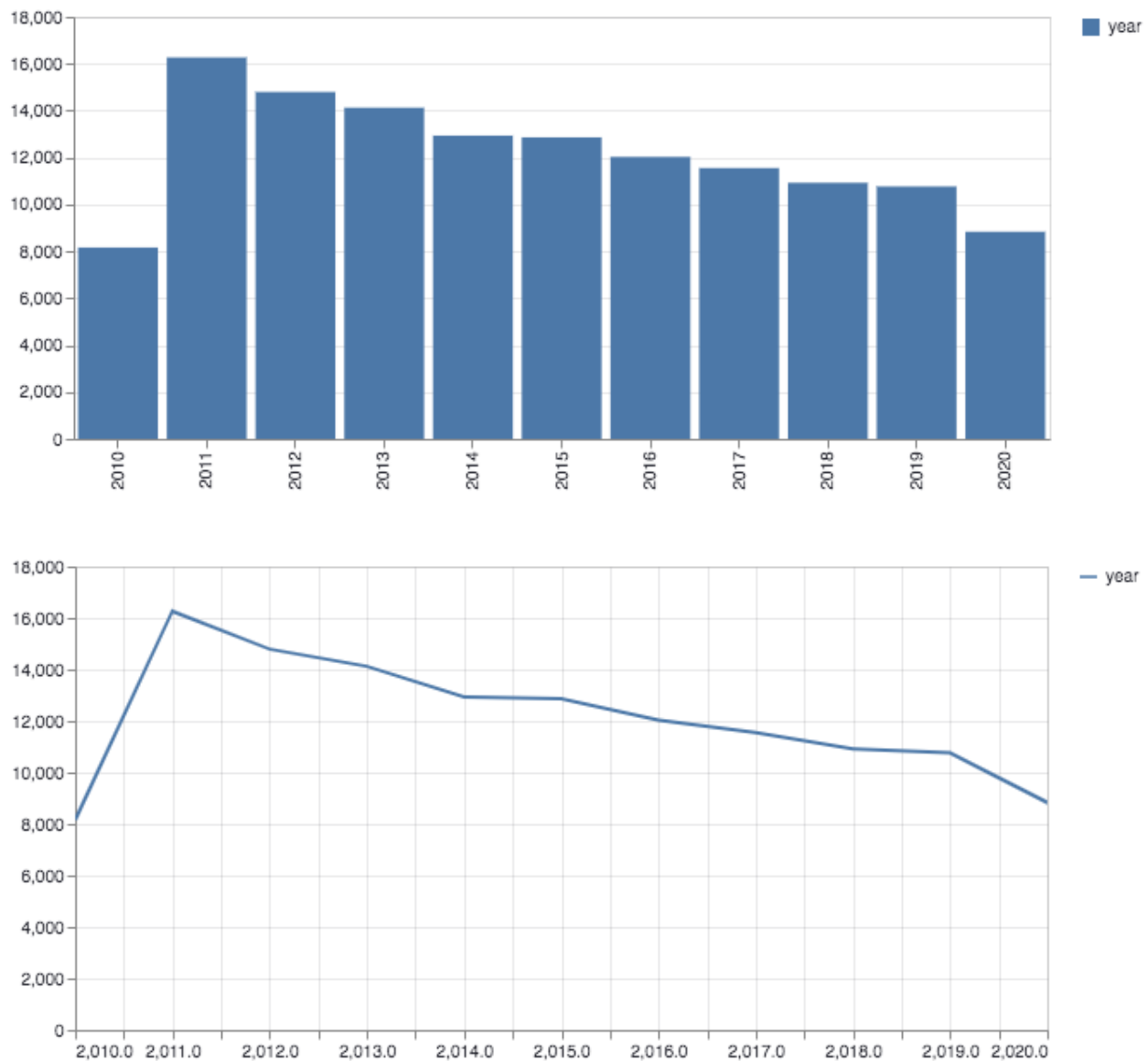


Figure 5: Bar and line charts of crimes occurring in different years

Looking at the plot of criminal activities over the course of 10 years, it is apparent that property crime rates are on the decline, beginning from the year 2012 as shown in Figure 5.



### 3.5.2.2 Month

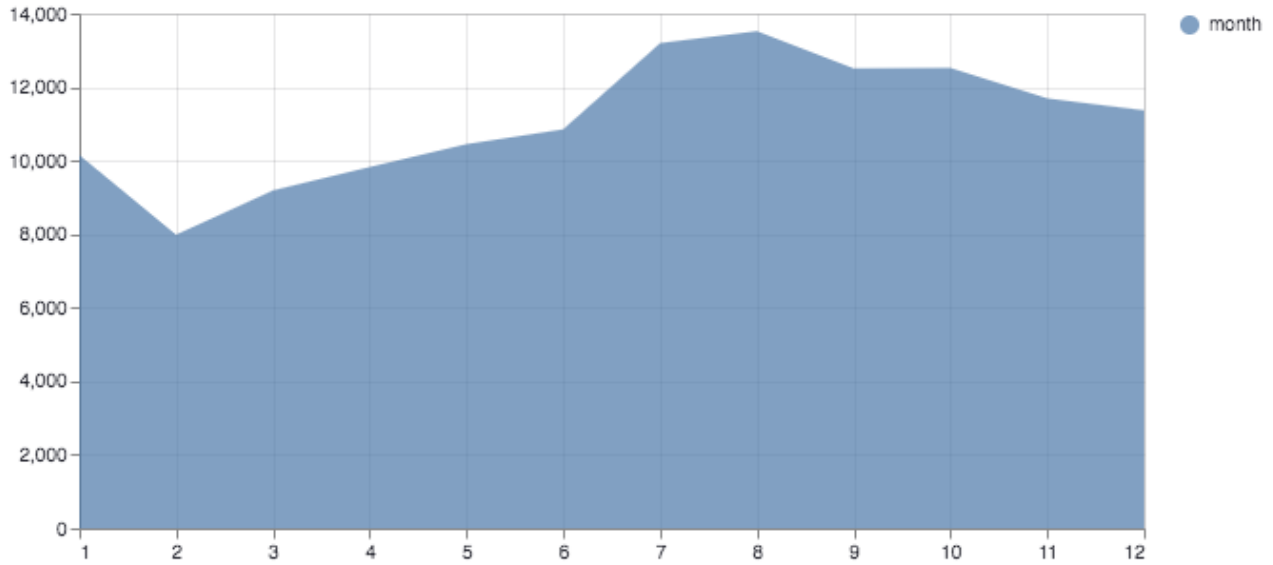


Figure 6: Area graph of crime occurring in different months

From the plot of crimes occurring in different months throughout the year as shown in Figure 6 above, it is apparent that most criminal activities occur during the Summer (June to August) and the least criminal activity occur in February during the season of Winter.

### 3.5.2.3 Day of Week

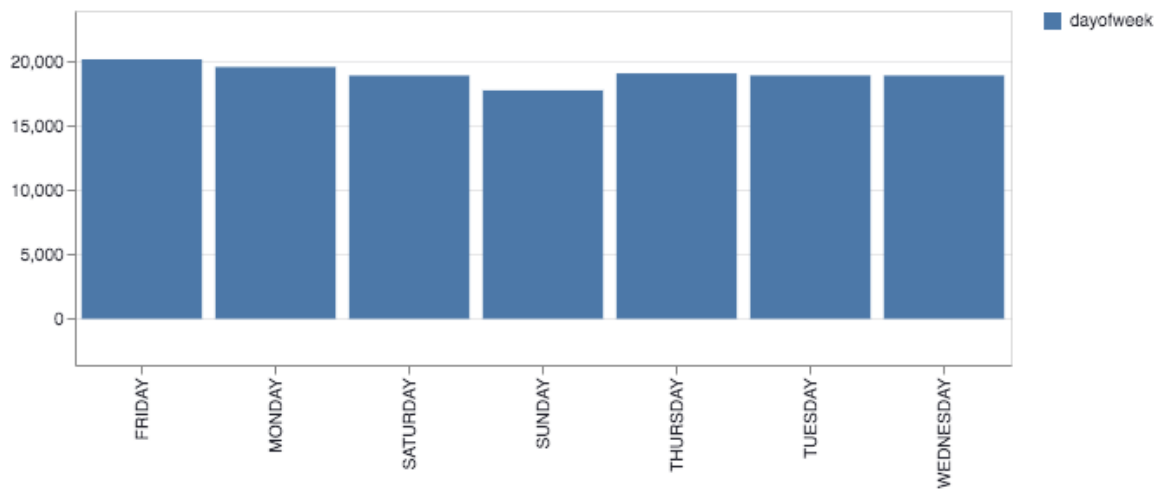


Figure 7: Histogram of crimes occurring during different days of week

From Figure 7 above, most crimes occur on Friday, and the least crimes occur on Sunday. Crimes dip on Sundays and pick up again sharply on Mondays. The rates gradually increase through the weekdays until they reach their peak on Friday.

#### 3.5.2.4 Hour

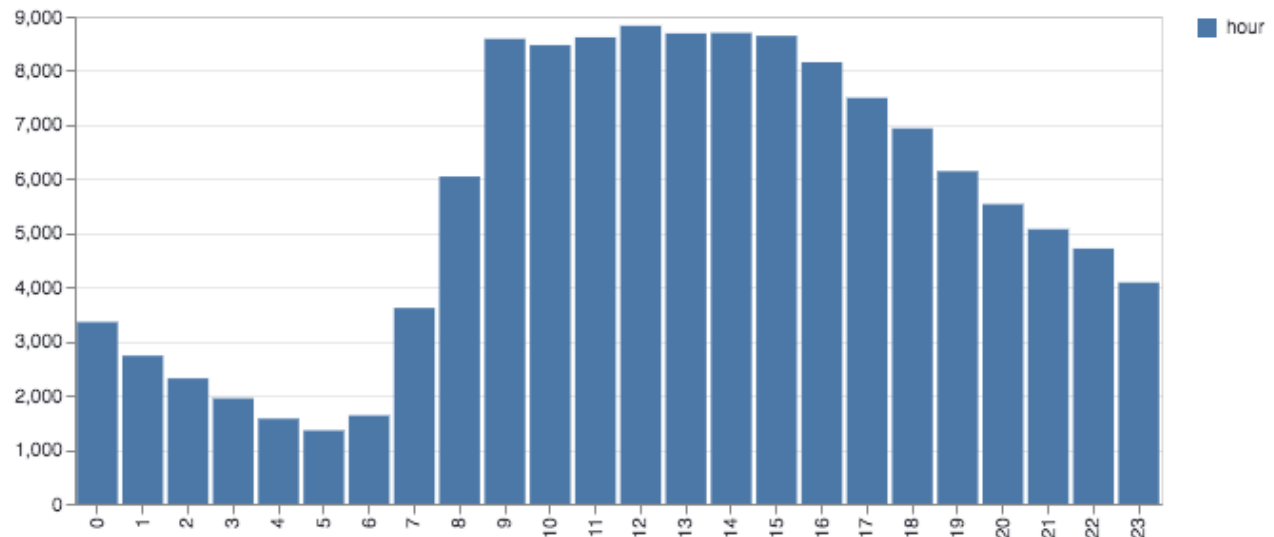


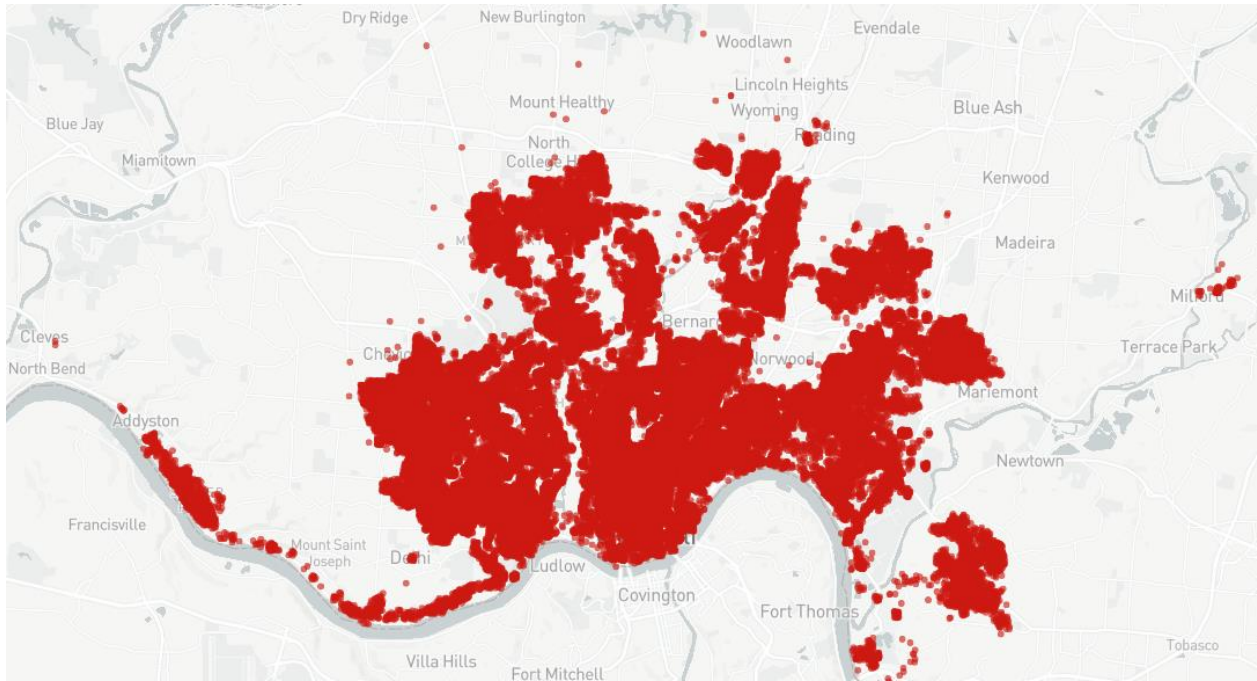
Figure 8: Crime occurring in different hours of the day (in 24 hours format)

From Figure 8 above, most crimes occur during late mornings to afternoons. Looking at the chart, there is an upsurge of criminal activities from 9AM and 10AM. The reports reach their peak at 12 noon and gradually decline throughout the evening. They are at their lowest after midnight through to the early hours of the morning.

#### 3.5.3 Location of Crime Incident

The Cincinnati crime dataset already contained location coordinates for each address recorded for crime scene. Thus, there was no computational geocoding for the exploratory analysis of the

crime map. 'Latitude' (X) and 'Longitude' (X) had 133247 unique values. These values were plotted using Streamlit's API to give a visual representation of the distribution of crimes.



*Figure 9: Map showing boundaries of reported crime coordinates*

As seen above in Figure 9, crimes are densely concentrated in neighborhoods on the border of the basin separating Ohio and Kentucky and sparsely distributed further from the border. These dense areas are downtown areas where there are a lot of movement and activity. The City of Cincinnati, often referred to as the Queen City, has nearly 300,000 Cincinnatians living in the downtown core and over 2.2 million residents in the surrounding metropolitan area. This makes Cincinnati the third largest city in the state of Ohio, and the 65<sup>th</sup> most populous city in the United States [43].

With the influx of citizens and about 2 million visitors each year, there could be a statistical correlation between the population density and crime activity. Since 2008, the Cincinnati Police Department (CPD) has expanded its citywide surveillance with the installation

of over 200 cameras through the Omnicast video surveillance system. The Omnicast allows the security agencies to see what is happening in real-time and respond efficiently. This project through its final product and continuous expansion can also help Cincinnatians and its visitors to predict and respond in real-time without the use of cameras, complementing police efforts.

### 3.5.4 Victim Gender

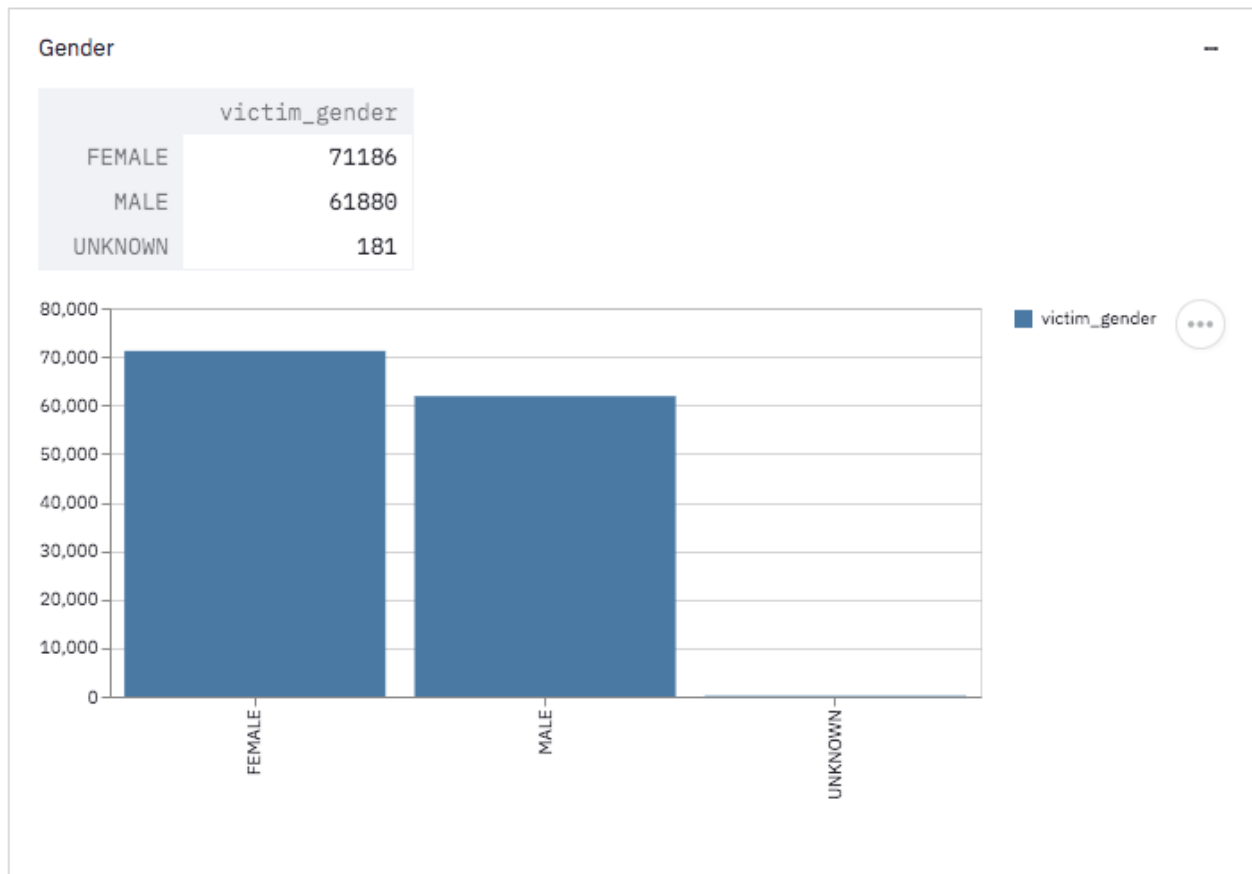


Figure 10: Value counts and histogram of victim gender

The dataset for victim gender as shown in Figure 10 above, gives the distribution of female versus male occurrences. Both genders have a high count of incidents over the past 10 years, with females being the most profiled.

### 3.5.5 Victim Race

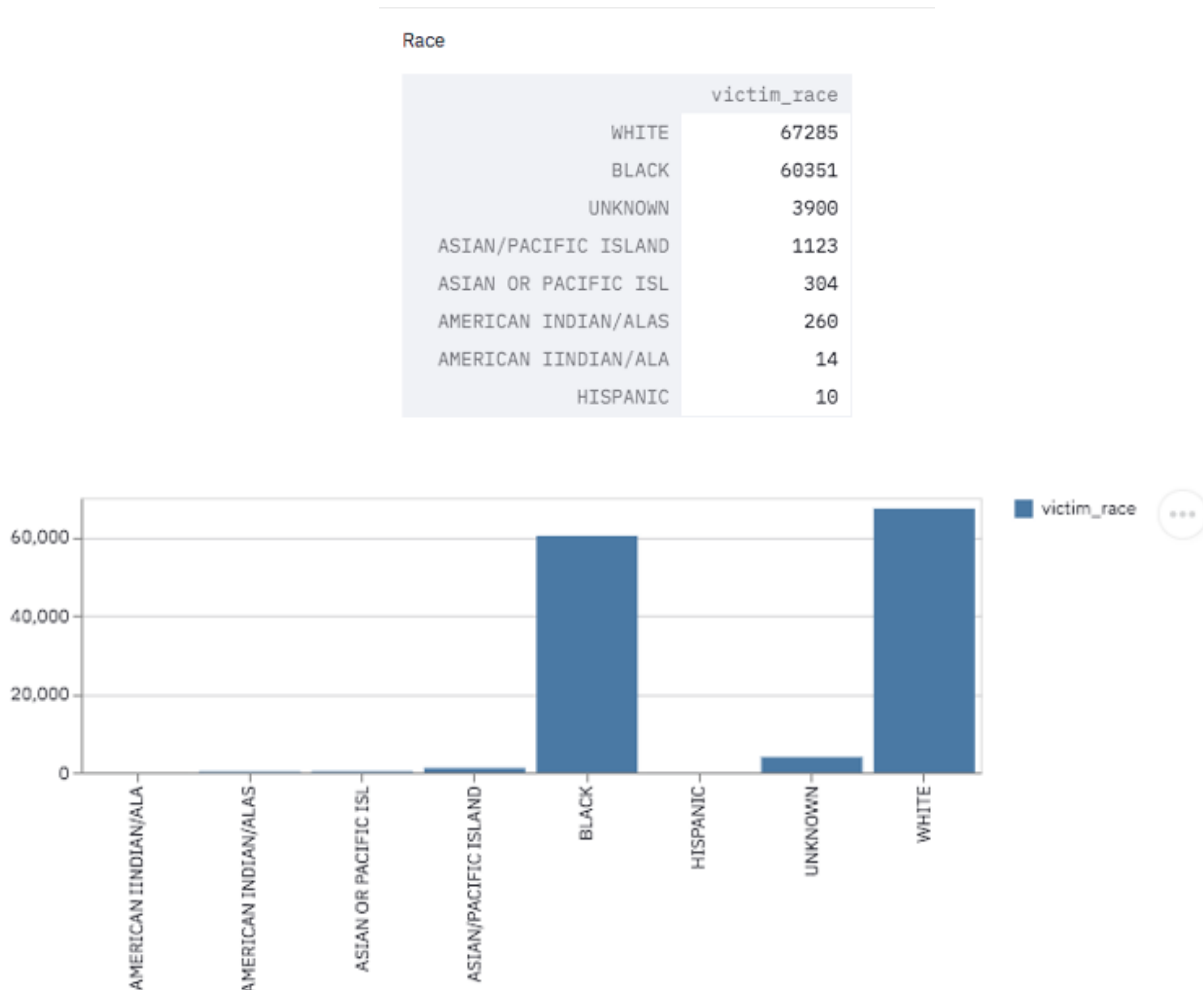


Figure 11: Value counts and histogram of victim race

The dataset for victim race in Figure 11 above, also shows the race distribution count of incidents for 8 distinct races over the 10-year period. Whites have the most occurrences followed by blacks. During model encoding, some of these are grouped to aid with encoding. Asians or American Indians are grouped under the White category to provide useful information for the model.

### 3.5.6 Victim Age

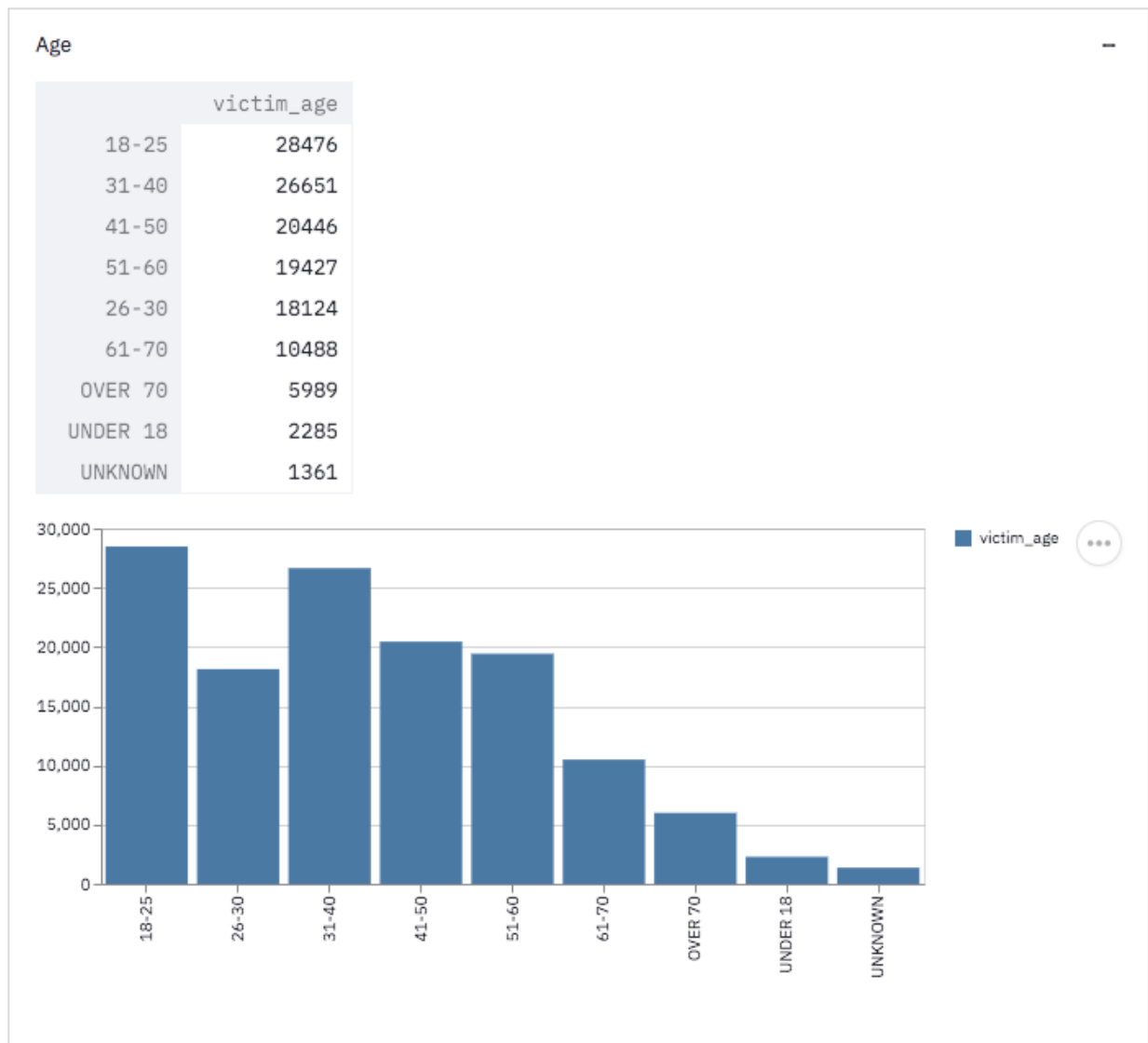


Figure 12: Value counts and histogram of victim age

Figure 12 above shows the age distribution for the dataset recorded in age groups. From the value counts and bar charts above most of the victims recorded are from the young age group of 18-25 followed by 31-40. This distribution suggests that older gens are much safer when it comes to property crime risks.

### 3.5.7 Close Code or Outcome

The close code or 'clsd' attribute of the crime dataset is designated as the target variable of the crime dataset. The attribute tells the resolution code for the reported crime incident. Originally, there were 13 distinct resolution codes in the dataset. These were grouped to set all instances of arrest (whether adult or juvenile) to a True instance (ARREST == 1) and all other instances to a False instance (NO ARREST == 0). Below is the distribution of values after processing the close code category:

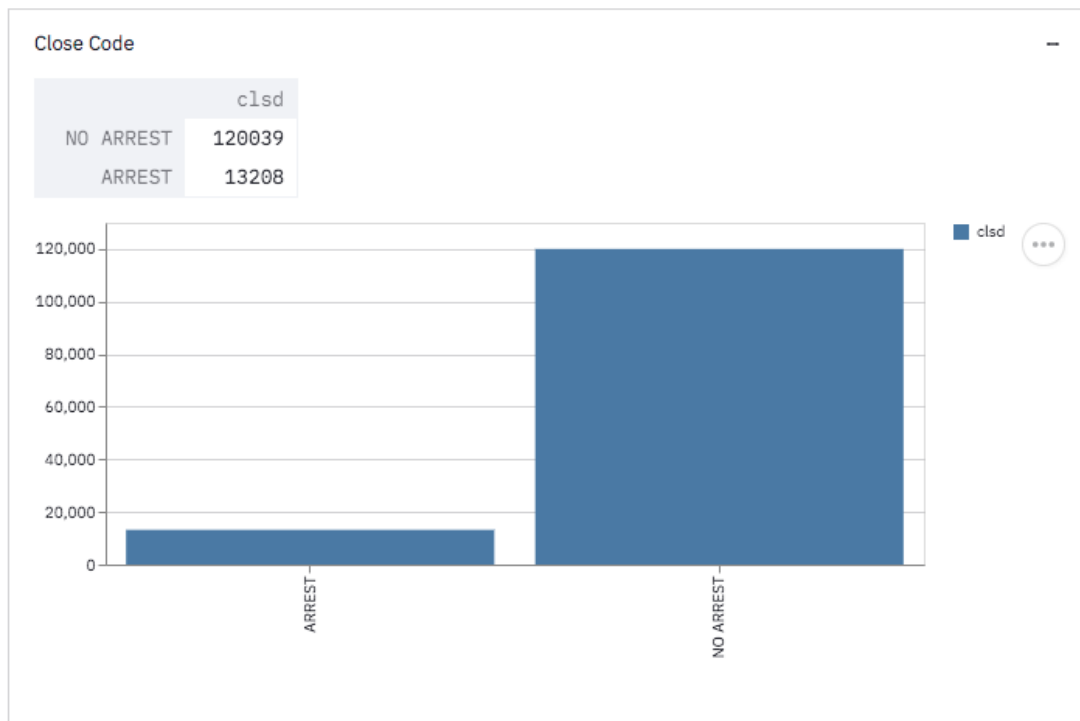


Figure 13: Snapshot of close code values for the crime data

As seen in Figure 13 above, instances of no arrest are significantly higher than instances of an arrest. This results in an imbalanced dataset for classification modeling. This issue will be addressed in another chapter of this thesis, where re-sampling methods were adopted to solve the classification problem.

Still on the distribution of the close code category, according to US Department of Justice Federal Bureau of Investigation, an offense is cleared by arrest, or solved for crime reporting purposes when at least one person has been arrested, charged with commissions of the offense and turned over to the court for prosecution <sup>[44]</sup>. There are many other instances where law enforcement agency cleared the incident by exceptional means that are not used as positive close code categories for this analysis. These instances are crimes that did occur and there were sufficient probable causes to support arrest or charge the offender, however circumstances beyond the law enforcement's control allowed the arrest. Exceptional clearances may include situations like death of the suspect, or decline of prosecution by victim, etc. As a point of reference for expansion on this topic, the inclusion of exceptional clearances to the positive class category of the target code can be deliberated.

### 3.5.7 Heatmap of Features

This section aims to show the correlation between each of the selected features to determine which of the features is most impactful for the outcome of a reported crime incident. Figure 14 below shows the output for which of the independent variables have the greatest effect on the output of the dependent variable. The variables are ranked in order of relative importance.

	Specs	Score
5	victim_age	83.3880
7	victim_gender	55.5563
1	month	24.2253
3	hour	7.4447
6	victim_race	4.1334
4	dayofweek	3.3781
2	day	0.6789
0	year	0.0004

Figure 14: Data frame showing Chi square score for best features



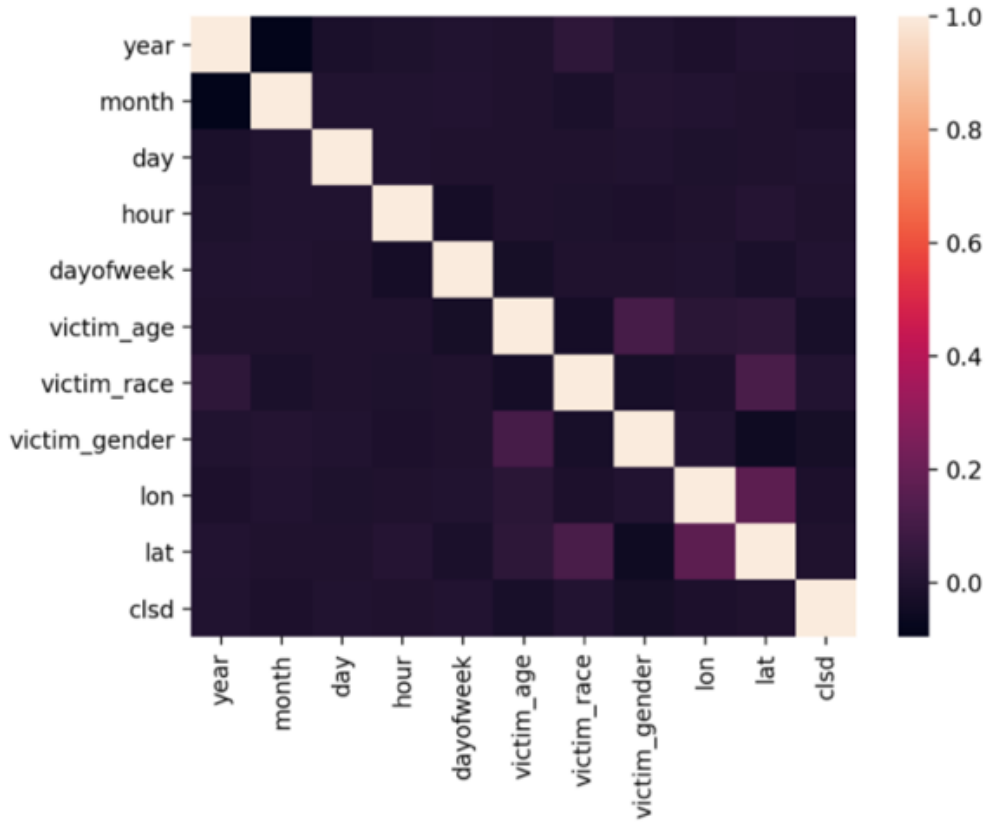


Figure 15: Heatmap showing feature correlation

Figure 15 showing the heatmap above, reveals a strong correlation between `victim_age` and `victim_gender`. This is explained in Figure 14 which gives the scores for best features. The selection of most relevant feature for determining close code of each incident is calculated using Chi squared algorithm and best features package from Python's *sklearn* module.

### 3.6 Performance Metrics

In order to propose a good fit for the input data, performance metrics is used to evaluate how well an algorithm is performing. Six metrics were used to compare and measure performance of different classification models on the crime dataset. The metrics under consideration include accuracy, precision score, recall score, F-score, confusion matrix and log-loss <sup>[47]</sup>.

### 3.6.1 Accuracy

Accuracy measures how many predictions matched exactly with the actual or true value of the testing dataset and returns the percentage of the correct results. This is also known as the benchmark. In other words, the value of an indicator is compared to a reference value, strengthening its statement. One may assume that high accuracy results in best model, however that depends on whether the dataset is symmetric and false positive and false negatives are almost the same. Hence the need to include other performance indicators in evaluation.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + False\ Negative + True\ Negative}$$

### 3.6.2 Precision Score

Precision score or positive prediction value is the relative amount of correctly predicted positive observations to the total predicted observations. Precision score answers the question: Of all reported crime instances that labeled as arrested, how many actually resulted in an arrest?

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

### 3.6.3 Recall Score

The recall score also called the sensitivity value is defined as the ratio of correctly predicted observations among all true instances. For example, of all the crime instances that truly resulted in an arrest, how many were labeled?

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

### 3.6.4 F-score

F-score of F1 score is the weighted average or mean between Precision and Recall. F1 does take into account False Positives and False Negatives into account, hence serves as a better indicator for uneven class distribution.

$$F1\ Score = 2 * (Recall * Precision) / (Recall + Precision)$$

### 3.6.5 Confusion Matrix

Confusion matrix also known as the contingency table distinguishes between true positive, false positive, true negative and false negative predictions. The only disadvantage being that confusion matrix requires human interpretation.

### 3.6.6 Log-Loss

Log loss is used to measure performance of classifier by penalizing false positives. This means that the smaller the log loss value, the more accurate the classifier. This is because there is a low uncertainty or entropy for the model. Log-loss will be used for the binary classification of this project.

## 3.7 Modeling

### 3.7.1 Data Preprocessing

*Streamlit* inherits properties from Python's library Scikit-learn (sklearn) for data processing.

Duplicates or NA values were dropped and datetime stamps were also parsed to extract date and time features. Out of the whole dataset containing 40 attributes, only about 10% of the data values were numeric. In order to use this dataset in the machine learning models, the text features

were converted to numeric values. The dataset was also split on an 80 by 20 ratio for training and testing dataset to deal with overfitting issues.

### 3.7.1.1 Data Encode

Machine learning requires all input and output variables to be numeric. This means that if the data contains categorical data, it must be encoded to numbers before it can fit and evaluate a model. Attributes with string data from the filtered dataset were “victim\_age”, “victim\_race”, “victim\_gender”, “clsd” columns. These were manually encoded to convert string data to numeric data by assigning each unique item an integer value. Datetime attribute although string, was also converted to into datetime object to deduce five different attributes: “year”, “month”, “day”, “hour” and “minute”. The technique adopted for this project is integer encoding, where a unique label is mapped to an integer. The labels were categorized as follows:

```
gender_dict = {'FEMALE': 0, 'F - FEMALE': 0, 'MALE': 1, 'M - MALE': 1, 'UNKNOWN': 2, 'NON-PERSON (BUSINESS)': 2}

age_dict = {'UNDER 18': 0, 'JUVENILE (UNDER 18)': 0, '18-25': 1, '26-30': 2, '31-40': 3, '41-50': 4, '51-60': 5, '61-70': 6, 'OVER 70': 7,
'UNKNOWN': 8, 'ADULT (18+)': 8, '00': 8}

race_dict = {'WHITE': 0, 'BLACK': 1, 'ASIAN/PACIFIC ISLAND': 2, 'AMERICAN INDIAN/ALAS': 3, 'UNKNOWN': 4,
'ASIAN OR PACIFIC ISL': 5, 'AMERICAN IINDIAN/ALA': 6, 'HISPANIC': 7}

target_dict = {'D--VICTIM REFUSED TO COOPERATE': 0, 'Z--EARLY CLOSED': 0, 'J--CLOSED': 0, 'H--WARRANT ISSUED': 0,
'F--CLEARED BY ARREST - ADULT': 1, 'K--UNFOUNDED': 0, 'G--CLEARED BY ARREST - JUVENILE': 1,
'I--INVESTIGATION PENDING': 0, 'B--PROSECUTION DECLINED': 0, 'E--JUVENILE/NO CUSTODY': 0,
'A--DEATH OF OFFENDER': 0, 'U--UNKNOWN': 0, 'C--EXTRADITION DENIED': 0}

week_dict = {'SUNDAY': 6, 'MONDAY': 0, 'TUESDAY': 1, 'WEDNESDAY': 2, 'THURSDAY': 3, 'FRIDAY': 4, 'SATURDAY': 5}
```

### 3.7.1.2 Training and Testing Dataset

The goal of splitting dataset into two portions: training and testing dataset, is to avoid overfitting and get more realistic accuracy. Training dataset contains all of the features in the dataset including the target label. Testing dataset contains features which the machine learning uses to

predict the target label. Scikit-learn has a package called `model_selection` which contains `test_train_split` that splits the dataset to the training and testing dataset. The test dataset size is set to 20% of the original dataset. This is used for conducting the experiment.

### **3.8 Model Selection**

Six machine learning algorithms, specialized for classification problems were selected, trained and evaluated. The goal was to obtain the best performing algorithm for the crime dataset to make prediction on user's input data. The models were evaluated on the basis of score accuracy, precision, recall, F-1 score and log loss, all imported from metric in the Python sklearn module.

#### **3.8.1 Imbalanced Dataset**

As discussed earlier in this chapter, the dataset for this project study is imbalanced since the classification categories are not approximately equal. Out of 133,247 instances reported as property crime incidents, only 13,208 resulted in an arrest. To deal with the imbalanced dataset problem, resampling methods (undersampling and oversampling) were used.

##### **3.8.1.1 Undersampling dataset**

The goal of undersampling technique is to remove instances from the training dataset that are in the majority class in order to balance the class distribution and then use it to fit the machine learning model. *TomekLinks* from sklearn module is used to achieve this. The python code for *TomekLinks* is given below with the results for both training and test dataset.

```
from imblearn.under_sampling import TomekLinks
```

```
tomeklinks = TomekLinks()
```

```
X_tl, y_tl = tomeklinks.fit_resample(X, y)
```

#### 1.4. Training Set

Training

	Model	Accuracy	Precision	Recall	F1 score	Log Loss
5	Random Forest	99.9961	1	0.9996	0.9998	0.0013
0	K-Nearest Neighbors	91.0835	0.7083	0.2137	0.3284	3.0797
1	Decision Tree	89.8013	0	0	0	3.5225
2	Support Vector	89.8013	0	0	0	3.5225
3	Logistic Regression	89.8013	0	0	0	3.5225
4	Naive Bayes	89.8013	0	0	0	3.5225

#### 1.5. Test Set

Testing

	Model	Accuracy	Precision	Recall	F1 score	Log Loss
5	Random Forest	91.2493	0.9954	0.1600	0.2757	0.0013
1	Decision Tree	89.5913	0	0	0	3.5225
2	Support Vector	89.5913	0	0	0	3.5225
3	Logistic Regression	89.5913	0	0	0	3.5225
4	Naive Bayes	89.5913	0	0	0	3.5225
0	K-Nearest Neighbors	89.4441	0.4714	0.1165	0.1868	3.0797

Figure 16: Training and Test results after undersampling with TomekLinks

The undersampling method proved good for the training data with 99.99% accuracy for the Random Forest classifier. The test data also gave about 92% accuracy but a low recall score. Below are the confusion matrix and ROC AUC Curve for the under sampled dataset.

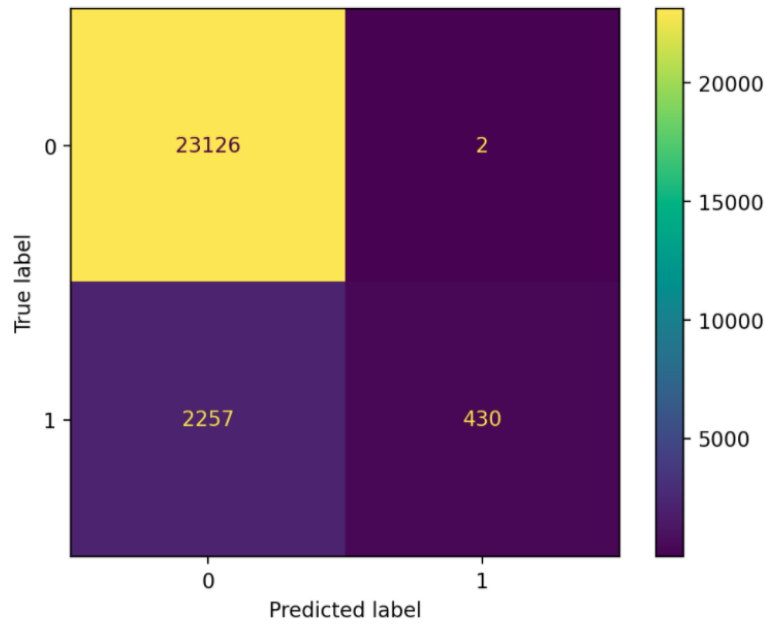


Figure 17: Confusion matrix of Random Forest algorithm for under sampled data

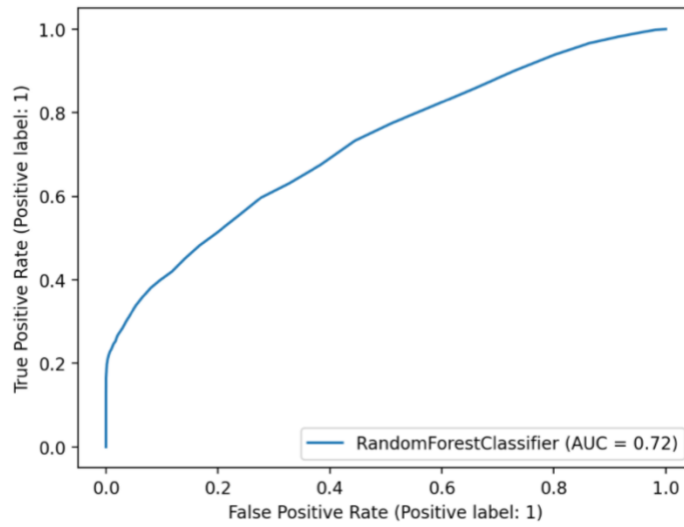


Figure 18: ROC AUC curve for Random Forest algorithm for under sampled data

Figure 17 for the confusion matrix shows that most of the classifications in this dataset using undersampling method were correct. A few incidents were misclassified, and the ROC AUC curve (Figure 18) gave a score of 0.73. This score was compared with next resampling technique to determine the best outcome.

### 3.7.1.3.1 Oversampling dataset

The goal of oversampling technique in solving imbalanced dataset problem is to oversample the minority class in order to balance the class distribution and then use it to fit the machine learning model. Synthetic Minority Oversampling Technique or SMOTE is the method used to generate new examples that are synthesized from the existing minority class. Below is the Python code for the SMOTE oversampling technique and the results as displayed in the web application.

```
from imblearn.over_sampling import SMOTE

smt = SMOTE()

X_train_res, y_train_res = smt.fit_resample(X, y)
```

## 1.4. Training Set

### Training

	Model	Accuracy	Precision	Recall	F1 score	Log Loss
5	Random Forest	100	1	1	1	0.0000
0	K-Nearest Neighbors	89.8379	0.8346	0.9934	0.9071	3.5100
3	Logistic Regression	62.4110	0.6117	0.6786	0.6434	12.9830
4	Naive Bayes	61.3113	0.6362	0.5273	0.5766	13.3627
1	Decision Tree	60.5438	0.5827	0.7409	0.6524	13.6279
2	Support Vector	50.0307	0	0	0	17.2588

## 1.5. Test Set

### Testing

	Model	Accuracy	Precision	Recall	F1 score	Log Loss
5	Random Forest	88.3242	0.8791	0.8894	0.8842	0.0000
0	K-Nearest Neighbors	85.2626	0.7793	0.9849	0.8701	3.5100
3	Logistic Regression	62.4734	0.6134	0.6798	0.6449	12.9830
4	Naive Bayes	61.5529	0.6409	0.5298	0.5801	13.3627
1	Decision Tree	60.6448	0.5848	0.7405	0.6535	13.6279
2	Support Vector	49.8771	0	0	0	17.2588

Figure 19: Training and Test results after oversampling with SMOTE



The oversampling method proved good for the training data with a 100% accuracy, and 88% on test data. The recall score for the best classifier (Random Forest) was 0.88 and log loss was 0.00 making SMOTE the preferred resampling method for advancing the research project. The confusion matrix and ROC AUC curve for Random Forest classifier on the oversampled data were used to support this statement.

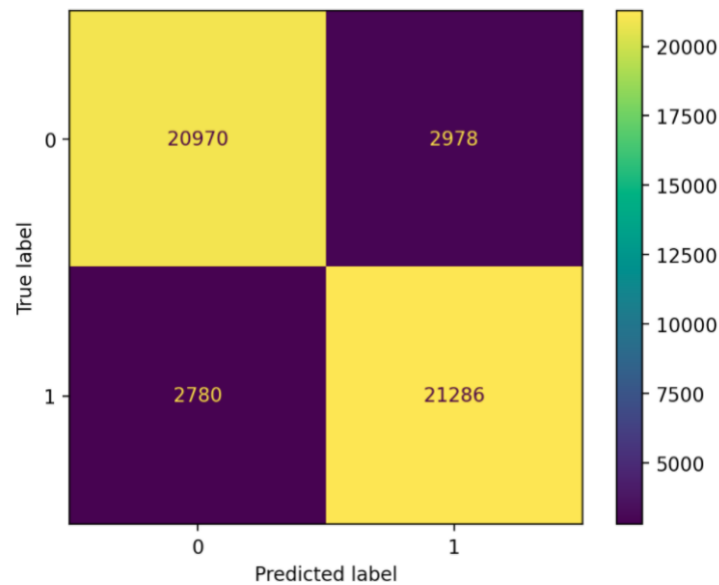


Figure 20: Confusion matrix of Random Forest algorithm for over sampled data

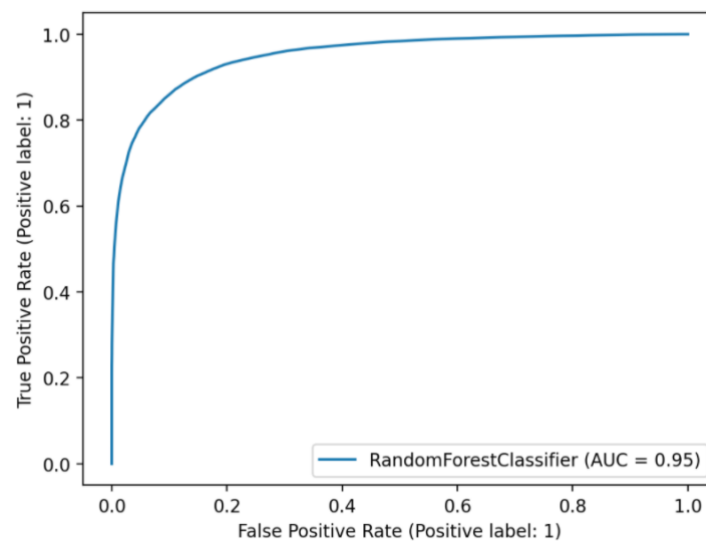


Figure 21: ROC AUC curve of Random Forest algorithm for the over sampled data

The confusion matrix shows that most of the classifications in this dataset using oversampling method were correct, as visible from the diagonal line in Figure 20 above. A few incidents were misclassified but overall, the ROC AUC curve gave a much higher score of 0.95 compared to the undersampling sample from previous analysis. Thus, SMOTE oversampling technique was used to solve the imbalanced classification problem of the dataset.

### 3.7.1 Model Selection

#### 3.7.1.1 Best Model - Random Forest Classifier

Among all the algorithms evaluated, Random Forest had the best results in performance metrics and in all instances of undersampling and oversampling techniques. Table 3 below gives a summary of the results of the re-sampling techniques.

Sampling	Method	Best Classifier	Accuracy	Log loss	Recall
Oversampling	SMOTE	Random Forest	88.47	0.0000	0.88
Undersampling	TomekLinks	Random Forest	91.50	0.0017	0.18

*Table 3: Result of best classifier (Random Forest) on balanced dataset*

Based on the results from Table 3 above, undersampling method using TomekLinks outperformed the SMOTE oversampling method on accuracy. However, further evaluation of the re-sampling methods on the dataset using F-1 and recall scores, confusion matrix and ROC AUC curve determined that SMOTE was the best resampling technique. The oversampling method also produced zero log loss compared to the undersampling technique.

## CHAPTER 4

### 4.1 Evaluation

The Random Forest algorithm ranked highest in all instances of performance matrices both on training and test data. An explanation can be because of the strong non-linear relational data processing ability and high prediction accuracy Random Forest has in many fields <sup>[45]</sup>. To validate the results, the algorithm was tested on the actual dataset using the K-Fold validation on 10 random data splits. KFold ensures that every observation from the original dataset has the chance of appearing in the training and test set. The Python code and the results of the validation are given below:

```
rf_cv = RandomForestClassifier()
cv_scores = cross_val_score(rf_cv, X, y, cv=10)
st.write(f'K-Fold = 10')
st.write(f"cv_score:", cv_scores.mean())
st.write(model_report)
```

#### 1.6. Model Choice: Random Forest

K-Fold = 10

cv\_score: 0.9184822152242029

precision recall f1-score support

0	0.89	0.88	0.88	23843
1	0.88	0.89	0.88	24173
accuracy			0.88	48016

macro avg 0.88 0.88 0.88 48016 weighted avg 0.88 0.88 0.88 48016

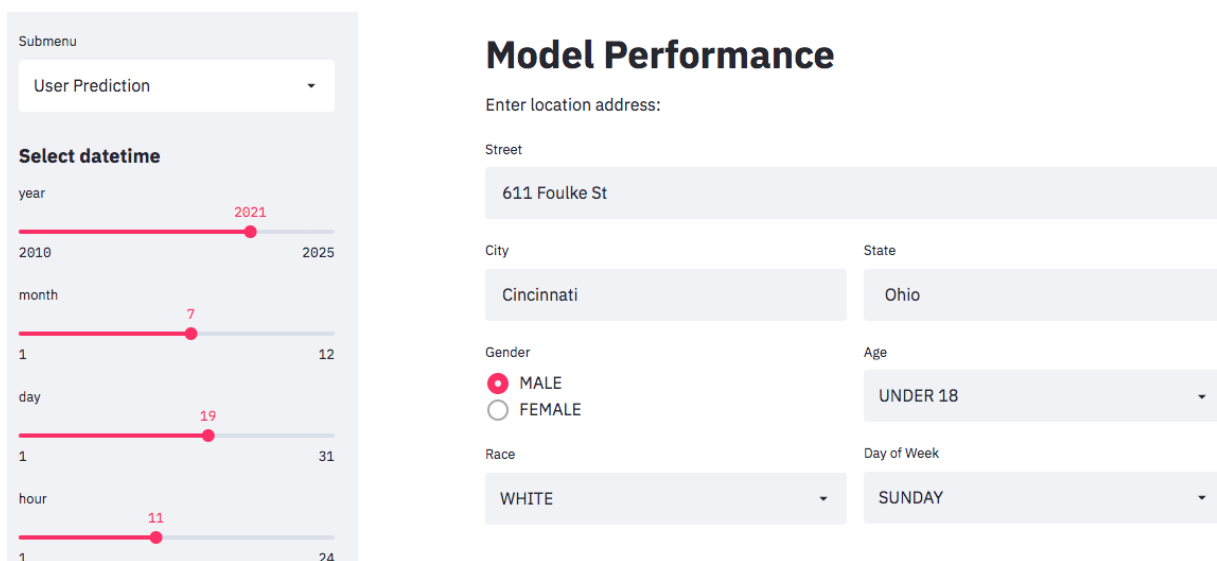
Figure 22: Result of Random Forest Classifier on actual dataset using K-Fold cross validation

The classification report for the K-Fold cross validation for the actual data for 10-fold valuation of the dataset, gave a mean cross validation `cv_score` of about 0.92 which is higher than the train test split score of the test dataset. This means that the Random Forest algorithm is consistent and can be used in production to lead to similar performance.

## 4.2 Prediction

### 4.2.1 Web Design - User Input Data

The goal of this study is to predict the safety of a specific location within a particular time based on a user's demographic information. *Streamlit* web application has many widgets that enable best user interface and interaction. It has text fields, radio buttons, sliders and dropdowns to assist users with populating fields with the right information. The data obtained from user is encoded and stored into a dataframe and then used on the saved Random Forest model classifier to predict safety score. Figure 23 below shows the design of the input parameters for collecting user data.



The image displays a Streamlit web application interface. On the left, a sidebar contains a 'Submenu' with 'User Prediction' selected. Below it, a 'Select datetime' section features four sliders: 'year' (set to 2021), 'month' (set to 7), 'day' (set to 19), and 'hour' (set to 11). The main content area is titled 'Model Performance' and includes a text input for 'Enter location address:' (611 Foulke St). Below this are several input fields: 'City' (Cincinnati), 'State' (Ohio), 'Gender' (MALE, selected with a radio button), 'Age' (UNDER 18, selected from a dropdown), 'Race' (WHITE, selected from a dropdown), and 'Day of Week' (SUNDAY, selected from a dropdown).

Figure 23: Segment with input parameters for collecting user information on Streamlit

To predict the probability of no crime, the user needs to provide four related features of the crime besides the date. The required features are location address, gender, age and race. The datetime feature is set by default to give the real-time date and hours during the course of the simulation. This feature can also be reconfigured to determine past and predict future data. For Cincinnati, the location has to be in one of the 52 neighborhoods (See Figure 1).

### User input data:

```
{
  "address" : "611 Foulke St, Cincinnati, Ohio, USA"
  "gender" : "MALE"
  "race" : "WHITE"
  "age" : "UNDER 18"
  "week" : "SUNDAY"
  "year" : 2021
  "month" : 7
  "day" : 19
  "hour" : 11
  "longitude" : -84.52697424257504
  "latitude" : 39.1363811
}
```

### Encoded data

	year	month	day	hour	week	age	race	gender	lat	lon
0	2021	7	19	11	6	0	0	1	39.1364	-84.5270

Figure 24: Output of information entered by user in JSON format and on Pandas dataframe

User Prediction segment under Model Performance (See Figure 24) gives the printout of the information collected in JSON format. The data is reengineered to display in the dataframe entitled 'Encoded data'.

## Prediction

Your safety score is 75.0%

## Feature Importance

	0
year	0.0905
month	0.0928
day	0.1308
hour	0.1126
week	0.0706
age	0.0667
race	0.0266
gender	0.0478
lat	0.1874
lon	0.1740

*Figure 25: Output of Safety Score and Feature Importance of best model trained on data*

The 'Prediction' segment of the web page gives the model output (safety score) and the feature of importance of Random Forest (the best trained model) on the data features. Every result for 'Prediction' is a number between 0 and 1 that predicts the likelihood of no arrest (thus safety) for the location at the time, given the user's demographics. The output is color-coded to show approval or disapproval of the user's presence at the location, using Cincinnati crime rate as threshold. Explanation of the threshold for safety result is expanded in the next section.

The 'Feature Importance' segment of the web page gives the scores of the importance of each of the trained data to the model. The higher the score, the more relevant the feature is toward the target variable (Close Code or clsd). Feature importance is an inbuilt class that is part of Tree Based Classifiers like Random Forest. Figure 26 below shows the full webpage as displayed in the Streamlit web application.

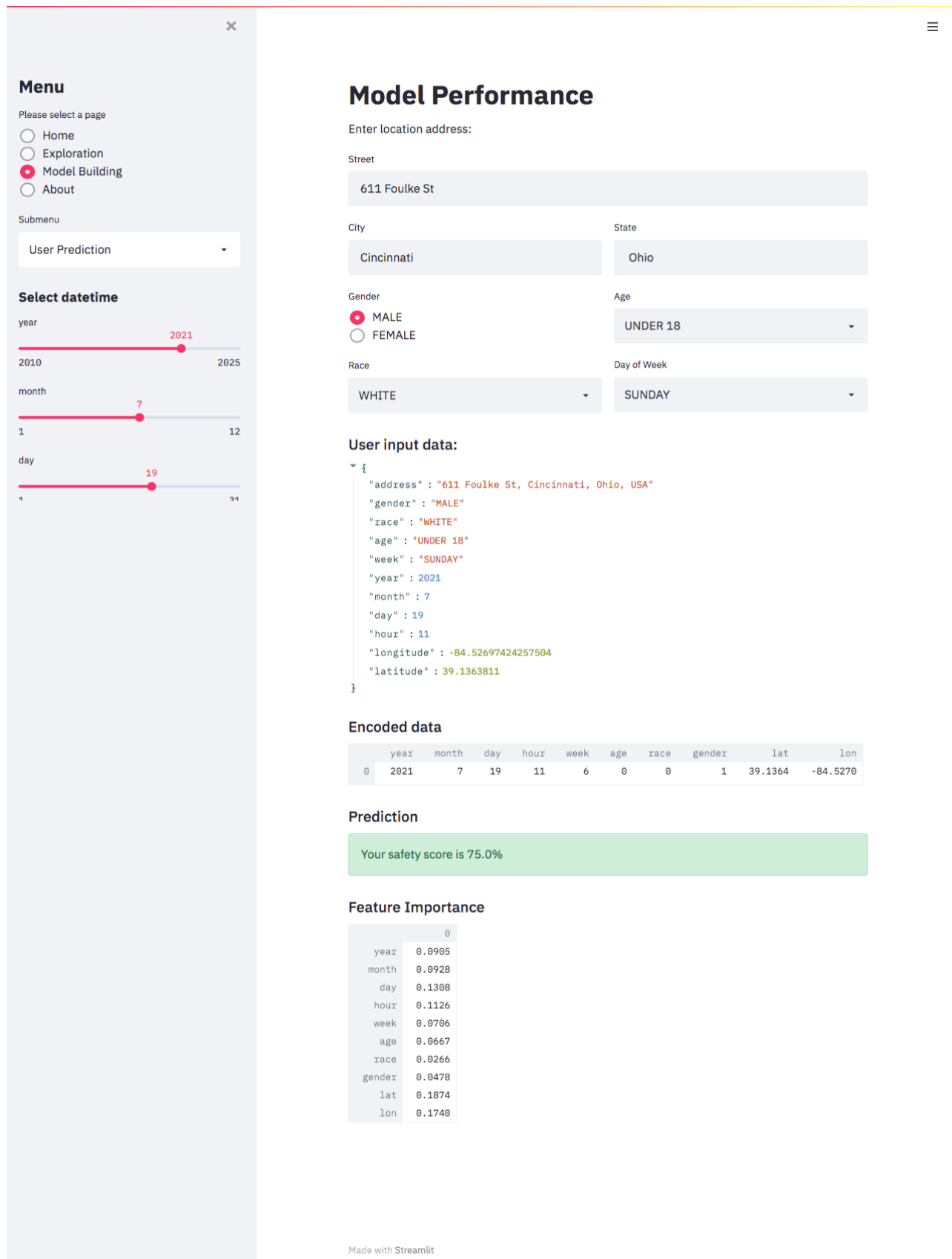


Figure 26: Screenshot of full User Prediction page on Streamlit web application

## 4.3 Results and Analysis

### 4.3.1 Probability of Arrest or No Arrest

A hard classification is the process of taking an observation and predicting which class it belongs to. In contrast a soft classification indicates the level of confidence that the model has in its prediction <sup>[46]</sup>. For this project, the safety score for the user's sample data employs soft classification, in that it measures the probability of a positive classifier (1 == ARREST) at the geocoded location. The higher the score means that the model is more confident that the observation belongs to the positive class. For this study, this is translated to mean that the higher the probability or chance of an arrest at the location, the less safe it is for the user. The metric used is obtained from Scikit-learn's `predict_proba` method, where the safety score is calculated mathematically as one minus the probability of getting arrested at the location or simply as the probability of no arrest at the geocoded location.

$$\text{Safety score} = 1 - \text{model.predict\_proba}(\text{user\_sample})$$

### 4.3.2 Safety score, so what?

The threshold for determining how safe a location is from the model safety score is measured against the overall property crime rate in the city of Cincinnati. According to crime stats on BestPlaces, property crime rate in Cincinnati is 74.6 whereas in the U.S it is 35.4 <sup>[48]</sup>. Using City of Cincinnati crime rate as the benchmark, the model was translated such that if the safety score is greater than 74.6, the location will be judged very safe and color coded to match the perceived risk. In Python, the code is given as:



```
if safety_score > .746:  
    st.success(f"Your safety score is {safety_score*100}%")  
else:  
    st.warning(f"Your safety score is {safety_score*100}%")
```

From Figure 25 above, the prediction output was color coded in green because the 75.0% chance of safety, outputted as safety score, is greater than the national property crime rate in Cincinnati, which is 74.6. The goal of adding this benchmark is to confirm or inform user's perception of safety which cannot be easily interpreted from the numeric score given.

Safety score of a location can have clear practical implications by informing residents, visitors and commuters on where and when to avoid certain locations, and how to react proactively to a predicted crime activity. The ability to predict the safety of a geographic location can give information about explanatory variables that can explain the underlying causes of these crimes (See Figure 25), hence enable users to intervene or have control over the perceived risk.

#### **4.2.2.2 Feature Importance**

The results in Figure 25 under 'Feature Importance', rank the importance of crime indicators to predict crime. Feature or variable importance is one of the ways to see how the model is fitted for the model. The result unveils that (by order of ranking) longitude, latitude, day of the week and hour were the most important variables for determining the likelihood of crime activity on the user, at the time of simulation. Nonetheless, after much testing, user age followed by gender, always ranked top for user demographic correlation to crime prediction.

## 4.4 Recommendation

The data was extracted automatically and live from the Cincinnati crime portal. The live and automatic extraction is essential to maintaining actual data and quality. The web application is also designed to collect data that give little room for user to insert wrong or inconsistent values. Unlike other crime data that only provide information on crime location, the crime dataset used for this analysis has information on crime occurrence date and time which are essential for the goals of this research. Since data remains the basis of all analysis, all stakeholders of the application must ensure that the integrity of the criminal data is maintained.

The close code variable had only two data values marked for the positive class (ARREST==1) in the original dataset. This accounts for only a small portion of property crime activities that did occur but did not result in an arrest. These crimes recorded as ‘cleared by exceptional means’ were contained in the original dataset but were not used for analysis due to criminal justice administrative reasons. Exceptional clearances with close codes ‘D—VICTIM REFUSED TO COOPERATE’, or ‘A—DEATH OF OFFENDER’, can be added to the positive class, to provide more data quantity for analysis and reduce the synthetic data generated from SMOTE oversampling method.

For this research, only property crimes were used, but many researches have been formulated for violent crimes as well. This machine learning agent could be designed to incorporate both property and violent crimes to give a broader perspective on safety to the user since it contains delicate crime codes like rape and aggravated assault. In addition, more information can be gleaned from the data such as crime weaponry, crime hotspots, and preferred neighborhoods. Even though this project focused on analyzing crimes in the city of Cincinnati. The research can be extended to other areas of the globe provided the necessary data is available for the area.

Also, a season-based predictive model can be developed to create crime clusters for a given timeframe. A sensitivity designed model can also be used to determine when the next crime will take place. As a word of advice, this proposed machine learning web application using Streamlit, does not take away user's discretion since it does not take full control of the decision making of the user and of the police. If harnessed properly, it can offer real-time valuable information about which the user can take advantage of to better protect themselves.

## CHAPTER 5

### 5.1 Conclusion

This thesis proposes that machine learning agents can move beyond the basic indicators of criminal activities of an area to classify the safety of a location at a specific time when given an address, datetime and user (potential victim) demographic information. In summary, the use of Geographical Information Systems (GIS) in the form of a Python crime predictive web application using Streamlit, combined with human behavioral activity on mobile devices can help increase awareness of crime activity and enhance predictive policing strategies in the City of Cincinnati. This research used a 10-year historical record of crime data from the Cincinnati Police Department to formulate a theory for location safety to help users protect themselves as they commute. Specifically, this study described a methodology to automatically extract crime information from the City of Cincinnati crime portal, model the data and then predict with 88% accuracy whether a user will experience property crime activity at a specific geographical address in the Cincinnati neighborhood and at a specific time when given demographic data such as age, gender and race.

The exploratory data analysis provided more visual information about crime types, the seasonality, victim information and explores the relationship between the variables and the target class. Victim age, was the most probable cause of property crime across various simulations, followed by gender and then race.

In this research approach, six machine learning algorithms were evaluated on property crime dataset and Random Forest ranked highest with the most accurate result. The experiments showed that the imbalanced dataset obtained during preprocessing benefited from using SMOTE oversampling method. The experiment was also tested using 10-fold cross-validation on the actual dataset and it gave a mean score of 92% which was higher than the test result. The safety

score was then given theoretically as the “confidence” that the model has in making its prediction using Sklearn’s `predict_proba` method and other Python packages embedded in Streamlit. Therefore, the final model was able to generate a reliable prediction with a tool for quicker crime preventive responses.

The distinctive characteristic in this research lies in the use of a web application that can be accessed on a mobile device such as a smart phone. The automatic extraction of information from the live portal and the use of computed geographic and demographic information give users a tool to understand the underlying causes of crime and have quicker proactive responses. The other advantage of this proposal is its predictive ability, catered to the demographics of the user. The method described predicts crime probability using variables that capture the dynamics and characteristics of a past victim profiles, rather than only making extrapolations from previous crime location, date and time. Operationally, this means that the proposed model is specialized for the geographic boundaries of Cincinnati, its residents and commuters, and can be expanded in detail to become a powerful artificial intelligence tool. In the model building, the importance of features found for tree building are information about the user’s demographic information. The highest ranking was age being probable cause of crime activity.

Finally, recommendations for the Streamlit web application were summarized. Suggestions for re-analysis were given, along with the advice to maintain user discretion since the application itself does not reduce criminality nor does not take full control over the decision making of the user and the police. It is evident that law enforcing agencies, real estate agents and tourists, just to mention a few, can take great advantage using model in the fight against crime. The whole thesis includes scripts of analysis hosted on GitHub <sup>[42]</sup> and on Streamlit share <sup>[50]</sup> for download and reuse. See <https://share.streamlit.io/arthurga/thesis-streamlit/app.py>

## **5.2 Future Work**

Despite the limitations discussed above and the need to validate the approach and robustness of the indicators, this research creates many avenues for research and computational approach to deal with the crime problem. As an extension, more classification models can be added increase crime prediction and enhance overall performance. It is also helpful to gather information about neighborhoods in order to see if there is a relationship between neighborhood income level and or amenities and their crime rate. Lastly, since the data analysis and final product is made public, hopefully a trend for developing mobile apps for predicting crimes will start, which can help law enforcements and keep the community safe for everyone.

## REFERENCES

- [1] R. Arulanandam, B. Savarimuthu and M. Purvis, 'Extracting Crime Information from Online Newspaper Articles', in Proceedings of the Second Australasian Web Conference — Volume 155, Auckland, New Zealand, 2014, pp. 31–38
- [2] NeighborhoodScout Crime Risk Report - Cincinnati.  
<https://www.neighborhoodscout.com/oh/cincinnati/crime>. Accessed (07/12/2021)
- [3] Smart City <https://www.techrepublic.com/article/in-cincinnati-mission-to-become-a-smart-city-public-data-is-critical-to-its-success/> Accessed (09/01/2020)
- [4] Andrew G. Ferguson, 'Policing Predictive Policing' Washington University Law Review. Volume 94. Issue 5. UDC David A. Clarke School of Law (2017)
- [5] PDI (Police Data Initiative) Crime incidents <https://dev.socrata.com/foundry/data.cincinnati-oh.gov/k59e-2pvf>. Accessed (1/14/2021)
- [6] Walter L. Perry, Brian McInnis, Carter C. Price, Susan C. Smith, John S. Hollywood, 'Predictive Policing. The Role of Crime Forecasting in Law Enforcement Operations'. National Institute of Justice. RAND Corporation (2013)
- [7] R. Stein and C. Griffith, Community policing strategies need to take into account police and residents' different perceptions of neighborhood crime. USApp—American Politics and Policy Blog (2015)

- [8] Naik, N., et al. "Streetscore-Predicting the Perceived Safety of One Million Streetscapes." *Streetscore-Predicting the Perceived Safety of One Million Streetscapes*, 2014, pp. 779-785. *SCOPUS*, [www.scopus.com](http://www.scopus.com).
- [9] M. Tayebi, F. Richard and G. Uwe, 'Understanding the Link Between Social and Spatial Distance in the Crime World', in Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '12), Redondo Beach, California, 2012, pp. 550–553.
- [10] Rountree, Pamela Wilcox, and Kenneth C. Land. "Perceived Risk versus Fear of Crime: Empirical Evidence of Conceptually Distinct Reactions in Survey Data." *Social Forces* 74, no. 4 (1996): 1353-376. Accessed July 14, 2021. doi:10.2307/2580354
- [11] Janet L. Lauritsen, Daniel L. Cork. 'Modernizing Crime Statistics: Report 1: Defining and Classifying Crime (2016). Chapter: 3 Users (and Uses) of Crime Statistics'. The National Academies of Sciences, Engineering, Medicine, pp 85 – 88
- [12] J. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. Do, O. Dousse, J. Eberle, and M. Miettinen. From big smartphone data to worldwide research: The mobile data challenge. *Pervasive and Mobile Computing*, 9:752–771, 2013.



- [13] Sebastian Rath. “Prediction of Assaults with a Combination of Retrospective and Prospective Analyses on the City of Graz” UNIGIS MSc Jahrgang 2014. University of Salzburg. Interfaculty Department of Geoinformatics – Z\_GIS, Nuremburg 2016
- [14] Friendly, M. (2007). A.-M. Guerry’s Moral Statistics of France: Challenges for Multivariable Spatial Analysis. *Statistical Science*, 22(3), 368–399
- [15] Beach S.R., Greenberg M.S. (2001) Paradoxical reactions of property crime victims. In: Martinez M. (eds) *Prevention and Control of Aggression and the Impact on its Victims*. Springer, Boston, MA. [https://doi.org/10.1007/978-1-4757-6238-9\\_47](https://doi.org/10.1007/978-1-4757-6238-9_47)
- [16] See, e.g., Martin Hildebrand et al., *Predicting Probation Supervision Violations*, 19 PSYCHOL. PUB. POL’Y & L. 114, 115 (2013)
- [17] Predictive Policing: Don’t Even Think About It, *THE ECONOMIST* (July 20, 2013), <http://www.economist.com/news/briefing/21582042-it-getting-easier-foresee-wrongdoing-and-spotlikely-wrongdoers-dont-even-think-about-it>; Leslie A. Gordon, *Predictive Policing May Help Bag Burglars—But it May Also be a Constitutional Problem*, A.B.A. J. (Sept. 1, 2013, 8:40 AM), [http://www.abajournal.com/mobile/mag\\_article/predictive\\_policing\\_may\\_help\\_bag\\_burglars--but\\_it\\_may\\_also\\_be\\_a\\_constitutio/](http://www.abajournal.com/mobile/mag_article/predictive_policing_may_help_bag_burglars--but_it_may_also_be_a_constitutio/) [https://perma.cc/A9PX-2JJD].
- [18] Andrew G. Ferguson, *Policing Predictive Policing*, 94 WASH. U. L. REV. 1109 (2017). Available at: [https://openscholarship.wustl.edu/law\\_lawreview/vol94/iss5/5](https://openscholarship.wustl.edu/law_lawreview/vol94/iss5/5)

[19] See, e.g., Andrew Guthrie Ferguson, Big Data and Predictive Reasonable Suspicion, 163 U. PA. L. REV. 327, 329 (2015) [hereinafter Big Data]; Andrew Guthrie Ferguson, Predictive Policing and Reasonable Suspicion, 62 EMORY L.J. 259, 265–69 (2012) [hereinafter Predictive Policing]; Fabio Arcila Jr., Nuance, Technology, and the Fourth Amendment: A Response to Predictive Policing and Reasonable Suspicion, 63 EMORY L.J. ONLINE 87, 89 (2014).

[20] Dubois, P. F. (2007). Python: Batteries Included. Computing in Science Engineering, Volume 9(3).

[21] Greenberg, M. S., and Ruback, R. B., 1992, *After the crime: Victim decision making*. Plenum Press, New York.

[22] Norris, F. H., Kaniasty, K., and Thompson, M. P., 1997, The psychological consequences of crime: Findings from a longitudinal population-based study. In *Victims of Crime* (2nd ed) (R. C. Davis, A. J. Lurigio, and W. S. Skogan, eds.), Sage Publications, Thousand Oaks, CA, pp. 146–166.

[23] Frieze, I. H., Hymer, S., and Greenberg, M. S., 1987, Describing the crime victim: Psychological reactions to victimization. *Professional Psychology: Research and Practice* **18**: 299–315

[24] Beach S.R., Greenberg M.S. (2001) Paradoxical reactions of property crime victims. In: Martinez M. (eds) Prevention and Control of Aggression and the Impact on its Victims. Springer, Boston, MA. [https://doi.org/10.1007/978-1-4757-6238-9\\_47](https://doi.org/10.1007/978-1-4757-6238-9_47)

[25] Udo Schlegel, “Toward Crime Forecasting Using Deep Learning” Universitat Konstanz. Research Gate (2018)

[26] Nick Malleson and Martin A Andreson, “Spatio-temporal crime hotspots and the ambient population” Crime Science, a SpringerOpen Journal (2015)

[27] Andresen, M., Jenion, G. Ambient populations and the calculation of crime rates and risk. *Secur J* **23**, 114–133 (2010). <https://doi.org/10.1057/sj.2008.1>

[28] Andrey Bogolomov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, Alex (Sandy) Pentland, “Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data”. Research Gate. September (2014)

[29] ITUNews. “Does almost everyone have a phone?” <https://news.itu.int/almost-everyone-phone/> Accessed (07/15/2021)

[30] W. Dong, B. Lepri, and A. Pentland. Modeling the co-evolution of behaviors and social relationships using mobile phone data. In MUM 2011, 2011.

[31] M. Gonzalez, C. Hidalgo, and L. Barabasi. Understanding individual mobility patterns. *Nature*, 453(7196):779–782, 2008.

[32] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: concepts, methodologies, and applications. *ACM Transaction on Intelligent Systems and Technology*, 2014.

[33] Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. “Learning to Detect Patterns of Crime” Massachusetts Institute of Technology, Cambridge, MA 02139, USA, Cambridge Police Department, Cambridge, MA 02139, USA

[34] Babak Shahian Jahromi, ‘Predicting Neighborhood Safety Using Boosting Machine Learning Algorithm’. Safety Prediction in Chicago.

<https://medium.com/@BabakShah/introduction-ec19c481e6d2>. (Accessed 3/22/2021)

[35] G. Saltos and M. Cocea, An Exploration of Crime Prediction Using Data Mining on Open Data. <https://core.ac.uk/download/pdf/83937056.pdf> (Accessed 6/9/2021)

[36] Felix Bode, Florian Stoffel, and Daniel Keim. “Variabilität und Validität von Qualitätsmetriken im Bereich von Predictive Policing”. In: (2017), pp. 1–14 (cit. on p. 10).

[37] John Gramlich, “What the data says (and doesn’t say) about crime in the United States”. Pew Research Center. November 2020. <https://www.pewresearch.org/fact-tank/2020/11/20/facts-about-crime-in-the-u-s/> (Accessed 7/16/2021)

[38] United States Census Bureau, Quick Facts, Cincinnati city, Ohio.  
<https://www.census.gov/quickfacts/cincinnati-cityohio> (Accessed 7/16/2021)

[39] City of Cincinnati, Wholton - <http://www.openstreetmap.org/> (Accessed 7/16/2021)

[40] Streamlit sharing. <https://streamlit.io/sharing> (Accessed 5/23/2021)

[41] Streamlit web application. <https://streamlit.io/> (Accessed 7/16/2021)

[42] GitHub, 'thesis-streamlit/app.py', 2020. Available: <https://github.com/arthurga/thesis-streamlit> (Accessed 7/16/2021)

[43] Genetec, City of Cincinnati Security and Surveillance. "Cincinnati- collaborating on security" <https://www.genetec.com/customer-stories/city-of-cincinnati-security-and-surveillance> (Accessed 07/17/2021)

[44] U.S. Department of Justice Federal Bureau of Investigation, Criminal Justice Information Services Division, "Crime in the United States", Offenses Cleared. <https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/clearances>. (Accessed 7/17/2021)

[45] X. Zhang, L. Liu, L. Xiao and J. Ji, "Comparison of Machine Learning Algorithms for Predicting Crime Hotspots," in IEEE Access, vol. 8, pp. 181302-181310, 2020, doi: 10.1109/ACCESS.2020.3028420.

[46] Martin, Damien. Are you sure that's a probability? (2019) <https://kiwidamien.github.io/are-you-sure-thats-a-probability.html> (Accessed 6/1/2021).

[47] Scikit-learn.org, '3.3. Model evaluation: quantifying the quality of predictions — scikit-learn 0.17.dev0 documentation', 2015. [Online]. Available: [http://scikit-learn.org/dev/modules/model\\_evaluation.html](http://scikit-learn.org/dev/modules/model_evaluation.html). (Accessed: 7/18/2021).

[48] BestPlaces. <https://www.bestplaces.net/crime/?city1=53918000&city2=53915000>  
(Accessed 5/25/2021)

[49] Eleanor Klibanoff, “Of 194 rapes reported in Louisville in 2017, only four were ultimately convicted of rape.” Prosecution declined. <https://kycir.org/2019/12/05/prosecution-declined/>  
December 2019. Accessed (7/19/2021)

[50] Gifty Arthur, “Neighborhood Safety Prediction App”. Streamlit share. Available:  
<https://share.streamlit.io/arthurga/thesis-streamlit/app.py> Accessed 7/19/2021)