CRIME PREDICTION AND ANALYSIS

Data Mining Techniques

Nikhil Nikhil and Maxwell King

Northeastern University | Summer 2019

Table of Contents

ABSTRACT: **1**
INTRODUCTION: **1**

LITERATURE REVIEW: 2

METHODOLOGY: **3**

DATA COLLECTION: 4
PRE-PROCESSING AND DATA CLEANING: 4

Cleaning Instances: *4*
Cleaning Features: *5*

DATA VISUALIZATION: 5
CLASSIFICATION & PREDICTION MODELS 6
PATTERN IDENTIFICATION 6

RESULTS: **7**

CHICAGO DATASET 7

Arrest Feature Prediction: *8*
District Feature Prediction: *10*
Hour Feature Prediction: *13*
Primary Type Prediction: *15*
Apriori and FP-Growth: *16*

BOSTON DATASET 19

Shooting Feature Prediction: *20*
District Feature Prediction: *23*
Offense Level Prediction: *25*
Apriori and FP-Growth: *26*

DISCUSSION: **27**
FUTURE WORK: **28**
CONCLUSION: **29**
REFERENCES: **31**

Abstract:

Crime has been an ever-present issue that has negatively impacted people's lives and societies around the world. Crime is one of the major concerns for many Americans. More than one-third of the Americans fear walking alone at night in their neighborhoods, and even greater percentages worry about particular types of crimes. As modern times have brought about the age of information, this project focuses on using the newly found and ever-growing data that exists about crime in the United States to try and detect trends and make meaningful conclusions.

To complete the project's goal two datasets of major American cities were looked at, Chicago IL and Boston MA. The datasets from these cities were evaluated separately but had very similar attributes. Both of the datasets contained meaningful information about instances of crime that had taken place within the city over a course of time. To gain insight from these datasets they were first cleaned and then evaluated to try and find correlations between features and eventually to try and build out prediction models of various features in the data set. In the end the prediction models that were built out had a relatively high success rate of being able to predict various features and lots of insight was gained about crime in these cities and the united states as a whole.

Introduction:

United States is one of the biggest and most powerful countries in the world but still it is highly susceptive to crime. Crime continues to be one of the major concerns for the American public. According to 2018 survey from Gallup, just under half (49%) of Americans believe the problem of crime in the United State is very or extremely serious (Norman, 2018). In the same poll, 47 percent said they worry about home burglaries, and 44 percent said they worry about thefts of or from their motor vehicles. Corresponding figures for other crimes were: experiencing identity theft, 67 percent; getting mugged, 34 percent; getting attacked while driving your car, 19 percent; being

sexually assaulted, 22 percent (out of which 37 percent were women); and getting murdered, 20 percent (among the least figures in this poll, but one that still translates to 42 million adults worrying about being murdered) (Barkan, 2016). According to statistics obtained from the FBI, a property crime was reported about every three seconds in the U.S., and a violent crime was reported about every 22 seconds.

The above statistics clearly indicate that crime and various types of crimes are a grave concern for the American Society. Although there is a uniform concern about crime throughout the United States, some of these concerns might exceed what the facts about crime would justify. As we just mentioned that most of the public thinks the crime rate has been rising, but if we look at trends this rate has actually been declining since the early 1990s (Barkan, 2016).

The aim of this project is to shed light upon and analyze such trends. By using data mining techniques this project will shed light on two major cities in the United States, Boston, Massachusetts and Chicago, Illinois. The goal being to analyze patterns, correlate and make associations between different factors contributing to crimes. The hope being that by the end of the project there is a clearer understanding of what features of the data correlate to crime and to see if the use of various machine learning algorithms can be used to try and predict some of these features.

Literature Review:

Chung-Hsien Yu, Max W. Ward, Melissa Morabito and Wei Ding in their 2012 paper, *Crime forecasting using data mining techniques*, discussed the preliminary results of a crime forecasting model they developed in collaboration with the police department of a United States city in the Northeast. They architected datasets from original crime records along with the time and location of the events. Then they used an ensemble of data mining techniques such as SVM and Neural Networks to forecast the hotspots for crime in the city. (Yu, Ward, Morabito, & Ding, 2012)

Ginger Saltos and Mihaela Cocea, in their 2017 paper, An Exploration of Crime Prediction Using Data Mining on Open Data, used instance-based learning, regression and decision trees to explore 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 (Saltos & Cocea, 2017).

Shaobing Wu and Changmei Wang in their 2019 paper, Crime Prediction Using Data Mining and Machine Learning, obtained decision rules for criminal variables by adopting data mining and machine learning techniques such as Random Forest, Bayesian Networks, and Neural Network methods. They also attempted to observe the validity and accuracy of the random tree algorithm in predicting crime data. (Wu & Wang, 2019)

Above are some of the many related research works we looked at before starting this project. Our goal would be to build upon and learn from these and other previous works to analyze and improve our data models and accurately predict crime rates for near future.

Methodology:

This project involved many different steps in coming to the completion of its goal. In order to use data on crime in a meaningful way it first had to be collected and processed. After the collection and processing of the data the data was then analyzed by visualizations, prediction models and feature pattern analysis.

Data Collection:

The Boston crime dataset used for this project was obtained from data.boston.gov which is the official source for all the datasets related to the city of Boston [1] and the Chicago crime dataset was obtained from Kaggle.[2]

For this project two separate datasets were used and worked with to try and gain a better understanding of the crime in USA cities. These datasets reported large instances of crime over a given time frame for both Boston (2005 – 2018) and Chicago (2015 – 2017). The features in both the datasets varied slightly but overall the features were pretty similar with the exception of one or two features.

Pre-Processing and Data Cleaning:

Before either the Boston or Chicago Datasets could be used to extract meaningful information, they had to be cleaned. To do this various data mining techniques were used to clean up misreported and bad data instances as well as remove and clean up the features.

Cleaning Instances:

To clean the instances of each dataset the duplicate values and null values were removed from the dataset. In the case of these datasets, the null values were not replaced due to the reason that there was no way to calculate a meaningful replacement value and that the datasets were large enough compared to the data to be removed that it would not have a significant effect on how the features correlated or associated with each other.

[1] https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system
[2] https://www.kaggle.com/currie32/crimes-in-chicago

Cleaning Features:

After the instances were cleaned for each dataset the features were looked at. In order to clean up both datasets, a subset of unwanted features ware removed from each of the datasets as well as some of the wanted datasets were re-organized and their datatypes were re-adjusted.

We removed the features that did not add any value to our data, for example the Boston crime data set had a feature named Offense Code which was redundant as Offense Type and Offense Description were already a part of the dataset.

We also converted some of the labels in order to fill out null values and have more descriptive labels. For example, the Shooting feature in the Boston dataset had null values for all the instances of crime where shooting did not take place. We replaced all these null values to ‘0’ or ‘No’ based on our requirements for applying different data mining techniques.

We also needed to convert the label-based data into numeric data and vice versa while applying different classification and association techniques. For example, while using Apriori algorithm we had to convert numerical features such as hour into categorical labels of morning, evening and night. Similarly, while applying classification and prediction algorithms we had to convert label-based features into numerical features.

Data Visualization:

After doing the preprocessing and cleaning of the various datasets some data visualizations were built to get a better understanding of some of the trends that were present in the data. This information was then used to compare against the prediction models to see if any insights could be found.

Classification & Prediction Models

For the prediction section of this project various features from both the Boston and Chicago datasets were chosen to predict on. The goal was to see how different prediction models performed at predicting various features in the dataset.

All of the features that were selected to be used as the labels in the models were classification data types. So, to keep consistency throughout the project 5 different classification models were built out for each feature and then evaluated.

1. Decision Tree Classifier
2. Extra Tree Classifier
3. Random Forest Classifier
4. K-Neighbors Classifier
5. Gaussian NB Classifier

These models were all evaluated using confusion matrices, log loss and accuracy scores. The features chosen from each dataset differed slightly but an attempt to keep some consistency was taken into account. The results of these prediction models can be seen in the results section of this paper.

Pattern Identification

We used Apriori and FP Growth algorithm to obtain the association rules and frequent patterns. We proceeded by identifying the frequent individual labels in the database and extending them to larger set of labels given that these labels appear enough times in the dataset. These frequent label sets that were determined by the Apriori algorithm were used to come up with association rules that highlight general trends present in the dataset. These associations can help determine what factors play a key role in the instances of crime.

Similarly, the Frequent Pattern Growth (FP-Growth) algorithm was used to count the occurrences of attribute value pairs in the pre-processed dataset with mostly label-based data. FP growth does not require candidate generation and uses frequent pattern trees which is an advantage when compared to Apriori. The main goal behind using these algorithms was to identify the frequent crime patterns and see how features are associated with each other.

Results:

The results of this project are broken down by their respective datasets. Though there was an effort to compare similar datasets the results of the prediction models on those datasets are unique to the dataset and therefore separated here. In the discussion part of this paper an examination of how the two datasets did with respect to each other can be found.

Chicago Dataset

The Chicago crime dataset contained 15 features and 4,273,756 instances after we cleaned and preprocessed the dataset. The data types of the data can be seen in the figure on the right.

```
RangeIndex: 4273756 entries, 0 to 4273755
Data columns (total 15 columns):
Date                    object
ID                      int64
Block                   int64
Primary Type            int64
Description             int64
Location Description    int64
Arrest                  bool
Domestic                bool
District                float64
Year                    int64
Latitude                float64
Longitude               float64
Month                   int64
Day                     int64
Hour                    int64
```

Figure 1: Info of the Chicago dataset once preprocessing as completed

The feature breakdown can be represented by the following:

1. ID: Unique identifier for the record.
2. Date: Date when the incident occurred.
3. Block: The clock in Chicago where the crime occurred.
4. Primary Type: The primary description of the crime.
5. Description: The secondary description of the crime.

6. Location Description: Description of the location where the incident occurred.
7. Arrest: Indicates whether an arrest was made for that crime.
8. Domestic: Indicates whether the incident was domestic-related for that crime.
9. District: Indicates the police district where the incident occurred.
10. Year: Year the incident occurred.
11. Latitude: The latitude of the location where the incident occurred.
12. Longitude: The longitude of the location where the incident occurred.
13. Month: The month the incident occurred
14. Day: the day of the month the incident occurred.
15. Hour: the hour of the day the incident occurred.

The features that were chosen to be predicted on in the Chicago dataset were arrest, district, hour, location description and primary type. All of these features are classification features and therefore the log loss, accuracy and confusion matrix were used to evaluate how successful the various prediction models were at being able to predict the feature.

Arrest Feature Prediction:

The arrest feature is a binary feature that indicates whether the instance resulted in someone being arrested or not. The overall spread of the feature is around 3 non-arrests to 1 arrest. The features that were used to predict on the arrest label can be seen in the figure to the right.

Block	int64
Primary Type	int64
Description	int64
Location Description	int64
Domestic	bool
District	float64
Year	int64
Month	int64
Day	int64
Hour	int64

Figure 2: Features used in predicting the arrest feature

When looking at the results it is important to take into account the skew of the arrest data because it is seen in the confusion matrix. As you can see in the figure below, the model predicts with a high accuracy when the arrest label was false and had a harder time predicting the arrest label to be true. Part of this can be contributed to the fact that the dataset contained many more instances of the arrest label being false then being true and the model has taken that into account.

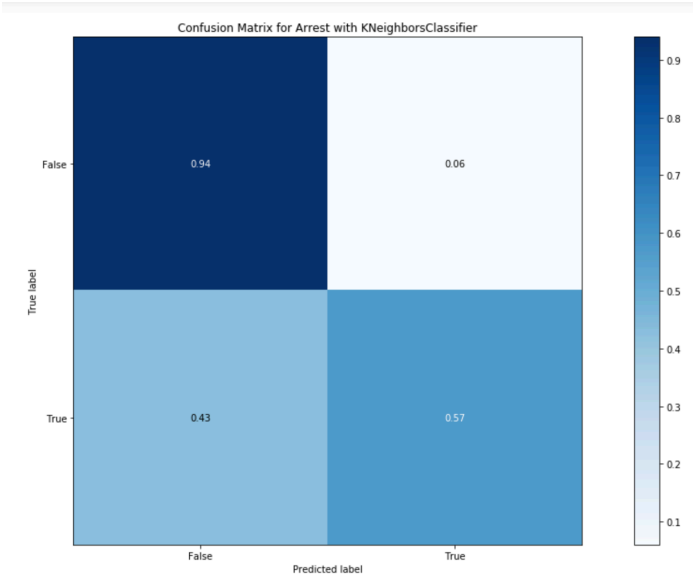

Figure 3: Confusion matrix for arrest feature

For the arrest feature the prediction model that gave back the best prediction score was the K-Nearest Neighbors method. It was able to predict with an accuracy of 84% and a log loss of only 2.10. Some of the other models had slightly better log loss scores but were unable to achieve the same accuracy. A full list of those models and their scores can be seen below.

Feature Predicted	Prediction Model Technique	Precesion	Recall	f1-Score	Log-Loss	Accuracy
Arrest	DecisionTreeClassifier	0.7	0.68	0.69	11.06	0.6795
	ExtraTreeClassifier	0.66	0.63	0.64	12.72	0.6317
	RandomForestClassifier	0.74	0.75	0.75	1.41	0.7463
	KNeighborsClassifier	0.83	0.83	0.83	2.01	**0.8340**
	GaussianNB	0.67	0.69	0.68	**0.69**	0.6944

Figure 4: Full prediction scores for arrest feature

For the results of the models it seems that the arrest feature can be predicted with an above average score which would indicate that there is a decent correlation between the features that were used to predict the models and arrest label.

District Feature Prediction:

The district feature was also used to build out prediction models. It is a classification label that countians 25 different classes. The spread of instances through these classes is slightly skewed with a few of the classes having very little data (districts: 13, 23 and 31).

Block	int64
Primary Type	int64
Description	int64
Location Description	int64
Arrest	bool
Domestic	bool
Year	int64
Month	int64
Day	int64
Hour	int64

Figure 5: Features used in predicting the district

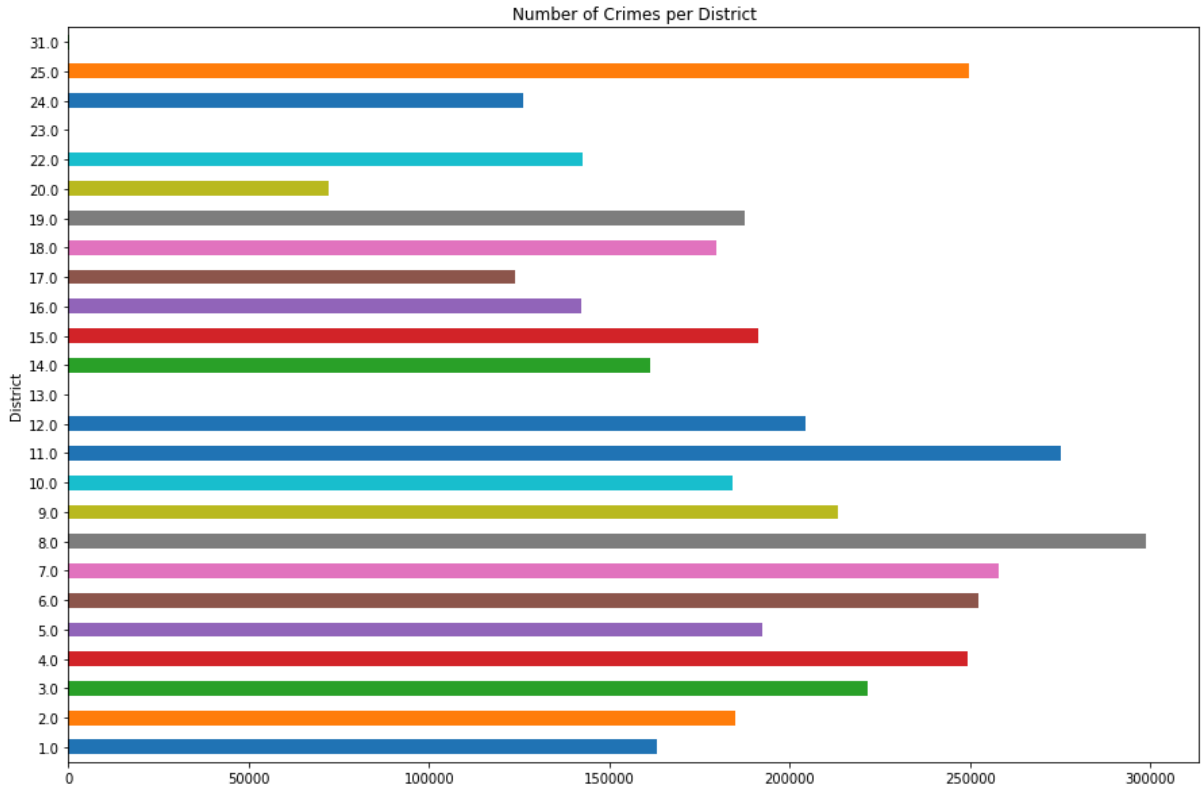

Figure 6: Visual of the number of instances per district

For the first iterations of trying to predict the districts the instances where the small valued districts were seen were left in the dataset. This resulted in a relatively high success rate with the best model being the decision tree classifier, which resulted in an accuracy of 96.1% and a log loss score of 1.35.

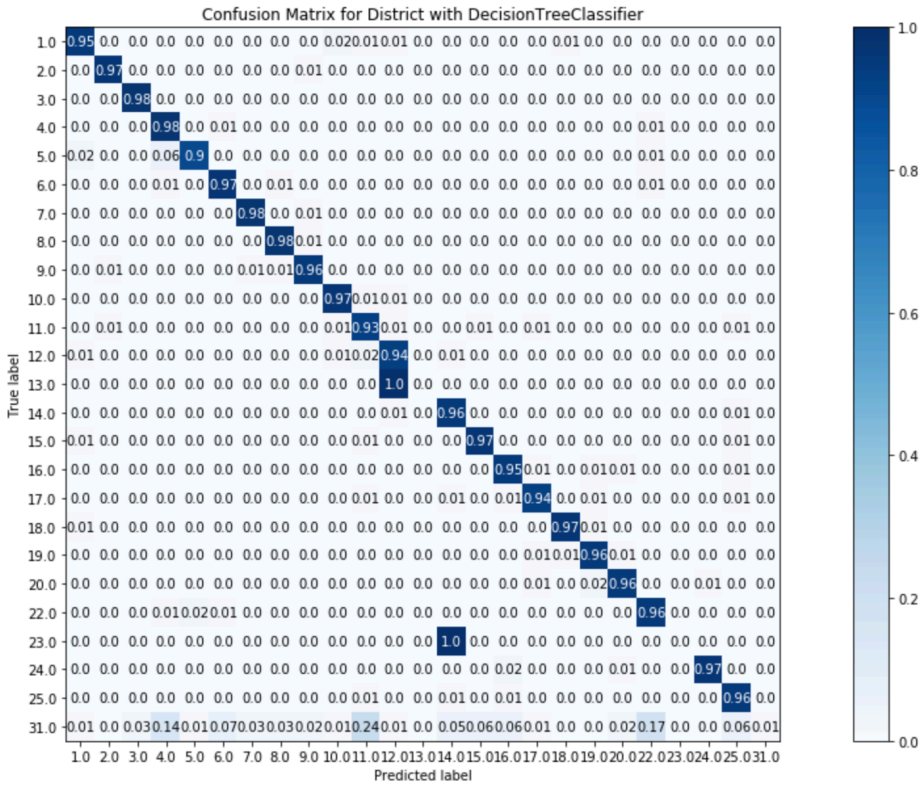

Figure 7: Confusion matrix based of the results of running all the districts on a decision tree model

However, as seen in the confusion matrix the districts that had little to no data played an effect in throwing off the prediction model.

Therefore, another iteration of prediction models were built where the data set being feed into the models did not include any instance from districts 13, 23 or 31. By doing this there was a slight uptick in the success level of all of the models. The decision tree classifier received the best score again, with a accuracy of 96.1% and a log loss of 1.34. The results from both models can be seen below.

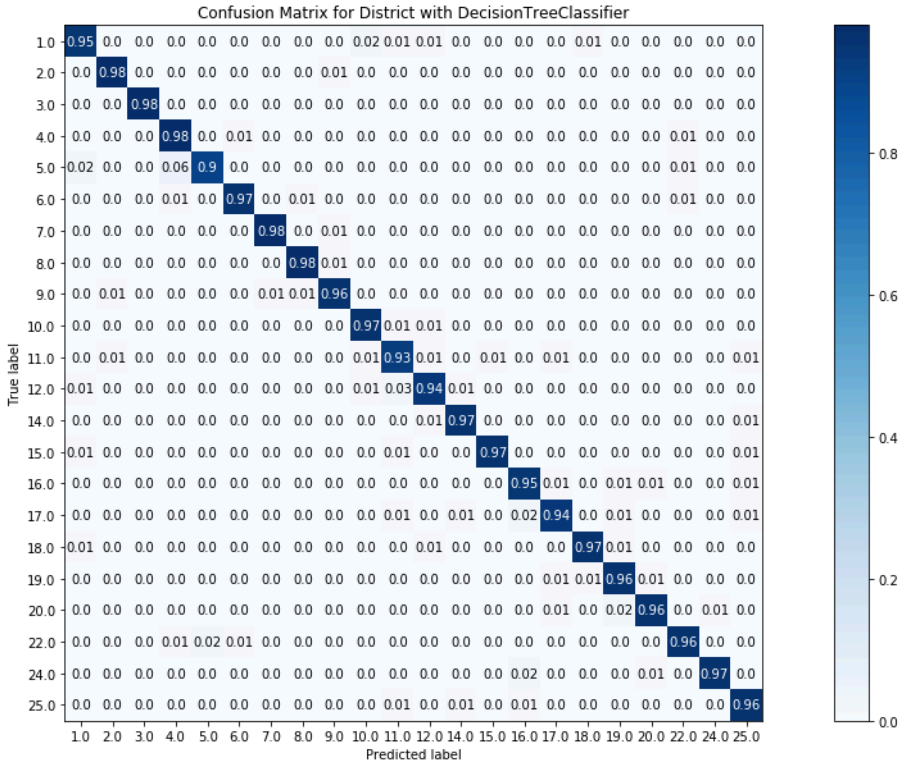

Figure 8: Confusion matrix based of the results of running all districts except for 13, 23 and 31 on a decision tree model

Feature Predicted	Prediction Model Technique	Precision	Recall	f1-Score	Log-Loss	Accuracy
District	DecisionTreeClassifier	0.96	0.96	0.06	1.35	0.9607
	ExtraTreeClassifier	0.18	0.17	0.17	28.80	0.1662
	RandomForestClassifier	0.34	0.15	0.15	7.66	0.1529
	KNeighborsClassifier	0.67	0.46	0.53	3.58	0.4642
	GaussianNB	0.12	0.1	0.07	3.03	0.0978

Figure 9: Results from running prediction models with all districts

Feature Predicted	Prediction Model Technique	Precision	Recall	f1-Score	Log-Loss	Accuracy
District	DecisionTreeClassifier	0.96	0.96	0.96	1.34	0.9613
	ExtraTreeClassifier	0.21	0.2	0.2	27.58	0.2016
	RandomForestClassifier	0.27	0.24	0.22	12.58	0.2350
	KNeighborsClassifier	0.64	0.64	0.64	3.90	0.6444
	GaussianNB	0.14	0.15	0.11	2.76	0.1513

Figure 10: Results from running prediction models without districts 13, 23 and 31

In summary the results from trying to predict which district the crime took place came back with a high success rate for the decision tree classifier with all the districts and with removing a handful of them.

Hour Feature Prediction:

The hour the crime took place was a feature that did not do well when it was attempted to be used as a label. Its results were quite underwhelming and didn't show much insight into the data.

Block	int64
Primary Type	int64
Description	int64
Location Description	int64
Arrest	bool
Domestic	bool
District	float64
Year	int64
Month	int64
Day	int64

Figure 11: List of features used to predict the hour that the crime occurred

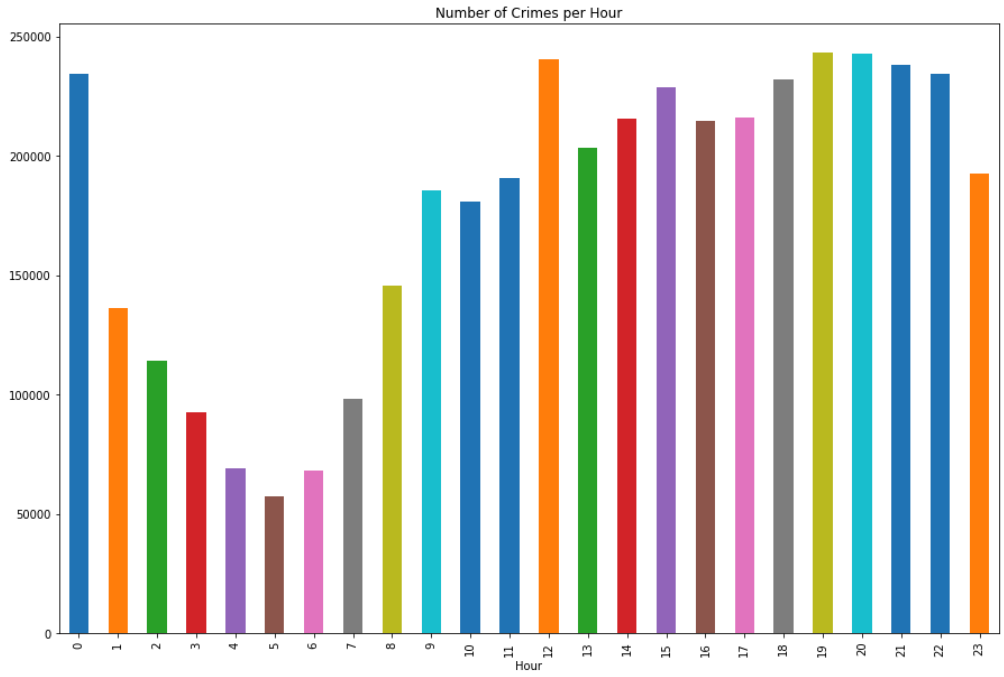

Figure 12: Visual representation of the spread of instances over the hour of the day.

All of the resulting scores were very low with the best classifier being the Naïve Bayes classifier with an accuracy of only 5.6% and a log loss of 3.32. As you can see below in the confusion matrix, the model predicted the majority of the instances to be in either hour 0 or hour 9.

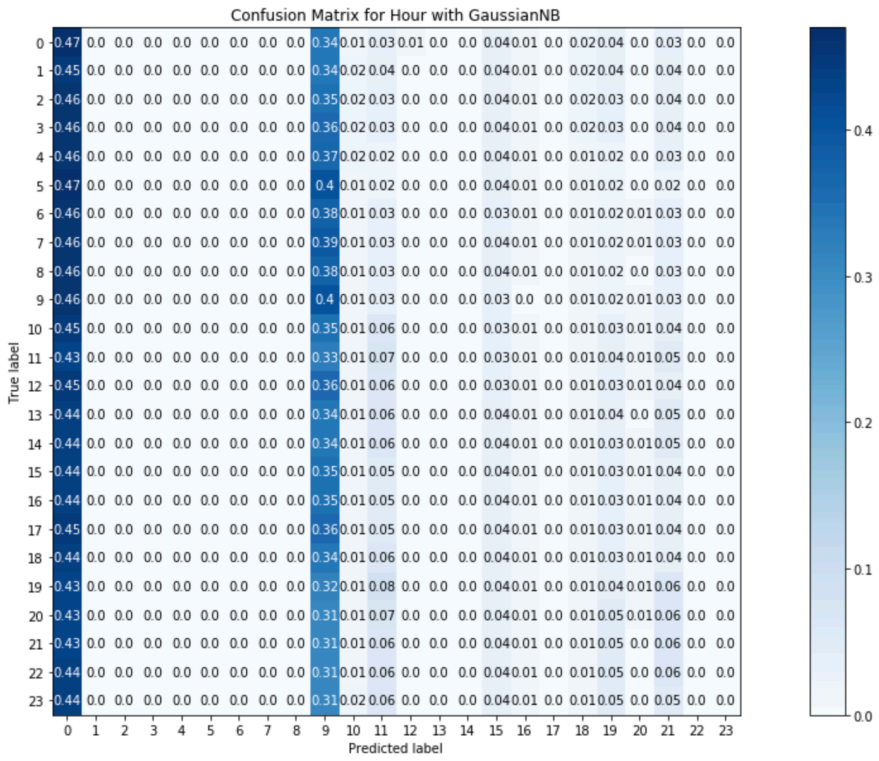

Figure 13: Confusion matrix of the Naive Bayes model for predicting the hour the crime took place

Feature Predicted	Prediction Model Technique	Precision	Recall	f1-Score	Log-Loss	Accuracy
Hour	DecisionTreeClassifier	0.05	0.05	0.04	32.83	0.0494
	ExtraTreeClassifier	0.05	0.05	0.05	32.70	0.0531
	RandomForestClassifier	0.05	0.05	0.04	25.51	0.0520
	KNeighborsClassifier	0.06	0.05	0.05	26.58	0.0546
	GaussianNB	0.05	0.06	0.03	**3.32**	**0.0561**

Figure 14: Results from all of the various prediction models that were used to try and predict the hour of day that the crime took place

The results from trying to predict what hour in the day the crime happened were extremely poor and concludes that there is no strong correlation between the features we chose to use to predict and the hour in which the crime occurred. This is something that will need to be addressed and reworked future plans of this project.

Primary Type Prediction:

Block	int64
Description	int64
Location Description	int64
Arrest	bool
Domestic	bool
District	float64
Year	int64
Month	int64
Day	int64
Hour	int64

Figure 15: Features used to predict the primary type of crime

One of the more important features that was attempted to be predicted was the primary type of crime. This feature is a classification feature and can be one of 34 categories. Though due to how skewed the classes were the bottom 15 classes were combined together during the preprocessing to form one class called 'Other'. So that reduced the number of classes that the crime could be down to 20.

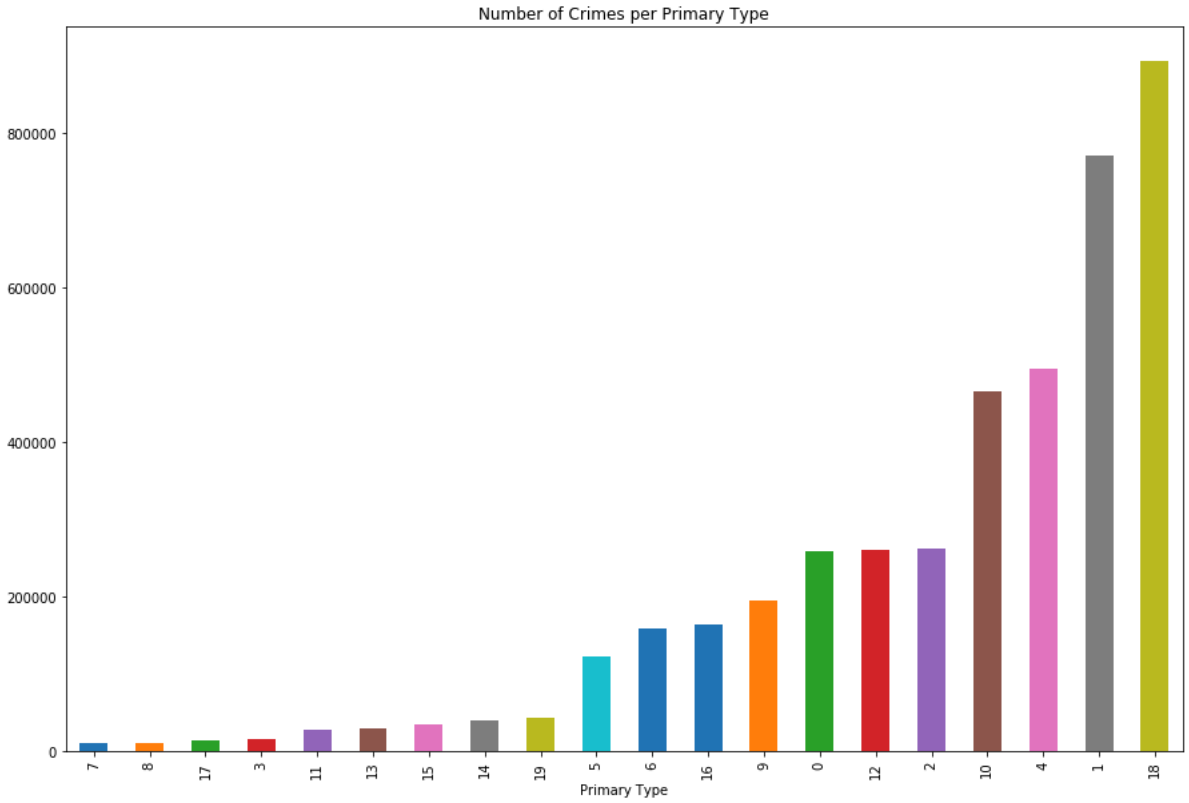

Figure 16: Visual representation of the primary types of crimes and the number of instances for each type.

The models made to predict the primary type of crime for each instance came back with varying success levels. The best model, the decision tree classifier, had an accuracy of 91.0% and a log loss score of 3.10. This is a very good score for a prediction model and it shows that there is a strong correlation between the features used to predict the primary type of crime and the label itself.

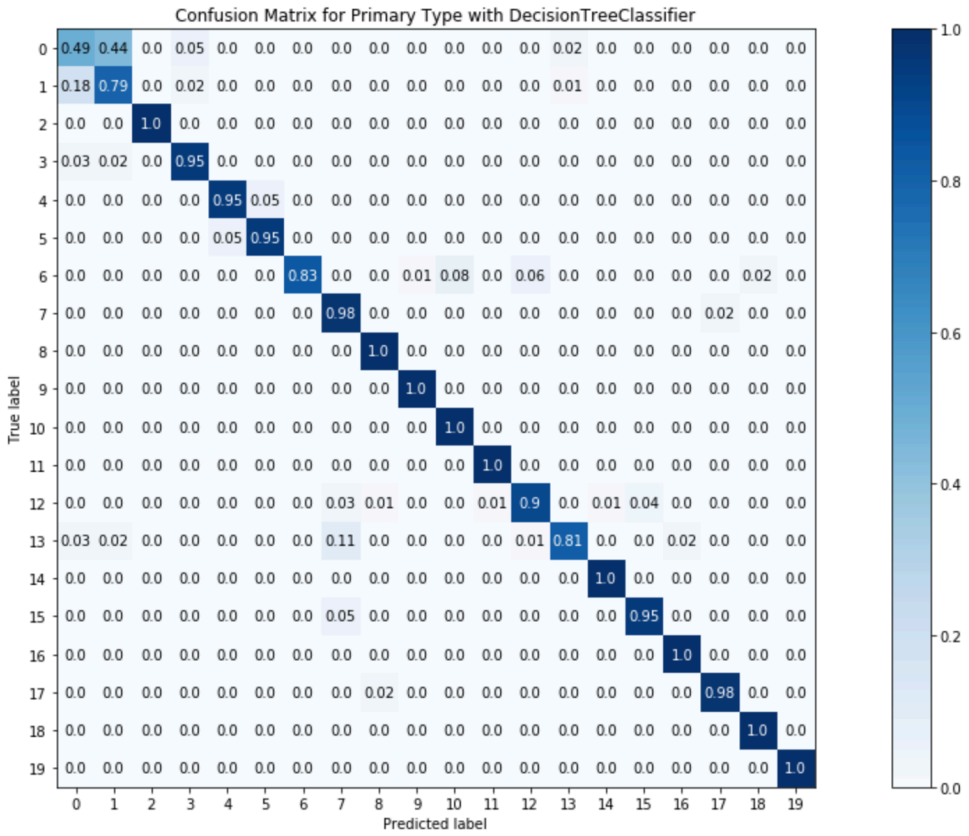

Figure 17: Confusion matrix for predicting primary type of crime using the decision tree classifier model

Feature Predicted	Prediction Model Technique	Precision	Recall	f1-Score	Log-Loss	Accuracy
Primary Type	DecisionTreeClassifier	0.92	0.91	0.91	3.10	0.9102
	ExtraTreeClassifier	0.59	0.4	0.45	20.88	0.3953
	RandomForestClassifier	0.76	0.6	0.65	2.96	0.5995
	KNeighborsClassifier	0.73	0.73	0.72	3.75	0.7323
	GaussianNB	0.37	0.29	0.3	2.985	0.2938

Figure 18: Results from all of the various prediction models that were used to try and predict the primary type of crime that took place

Apriori and FP-Growth:

The initial results that were looked at in this project were that of how the features related to each other. The Apriori and FP-Growth algorithms were used to try and build out associated rules between the features and build a base level of understanding.

```
{STREET} ---> {Not Arrested, Not Domestic}:  conf = 0.668, sup = 0.169
{Noon, Not Domestic} ---> {Not Arrested}:  conf = 0.703, sup = 0.123
{Not Arrested, Noon} ---> {Not Domestic}:  conf = 0.869, sup = 0.123
{Noon} ---> {Not Arrested, Not Domestic}:  conf = 0.621, sup = 0.123
{Evening, Not Arrested} ---> {Not Domestic}:  conf = 0.875, sup = 0.132
{Evening, Not Domestic} ---> {Not Arrested}:  conf = 0.678, sup = 0.132
{Evening} ---> {Not Arrested, Not Domestic}:  conf = 0.602, sup = 0.132
{CRIMINAL DAMAGE, Not Domestic} ---> {Not Arrested}:  conf = 0.939, sup = 0.102
{Not Arrested, CRIMINAL DAMAGE} ---> {Not Domestic}:  conf = 0.921, sup = 0.102
{CRIMINAL DAMAGE} ---> {Not Arrested, Not Domestic}:  conf = 0.86, sup = 0.102
{THEFT, Not Domestic} ---> {Not Arrested}:  conf = 0.892, sup = 0.183
{Not Arrested, THEFT} ---> {Not Domestic}:  conf = 0.972, sup = 0.183
{THEFT} ---> {Not Arrested, Not Domestic}:  conf = 0.868, sup = 0.183
{Morning, Not Domestic} ---> {Not Arrested}:  conf = 0.741, sup = 0.165
{Not Arrested, Morning} ---> {Not Domestic}:  conf = 0.86, sup = 0.165
{Morning} ---> {Not Arrested, Not Domestic}:  conf = 0.646, sup = 0.165
{NARCOTICS, Not Domestic} ---> {Arrested}:  conf = 0.995, sup = 0.11
{Arrested, NARCOTICS} ---> {Not Domestic}:  conf = 1.0, sup = 0.11
{NARCOTICS} ---> {Arrested, Not Domestic}:  conf = 0.995, sup = 0.11
```

Figure 19: Sample of some of the associated rules that were produced using the Apriori method

```
{NARCOTICS} ---> {Arrested}:  conf = 0.995, sup = 0.11
{Not Domestic} ---> {Not Arrested}:  conf = 0.712, sup = 0.62
{Not Arrested} ---> {Not Domestic}:  conf = 0.857, sup = 0.62
{Morning} ---> {Not Domestic}:  conf = 0.873, sup = 0.223
{Morning} ---> {Not Arrested}:  conf = 0.751, sup = 0.192
{THEFT} ---> {Not Domestic}:  conf = 0.973, sup = 0.206
{THEFT} ---> {Not Arrested}:  conf = 0.893, sup = 0.189
{Night} ---> {Not Arrested}:  conf = 0.731, sup = 0.239
{Night} ---> {Not Domestic}:  conf = 0.852, sup = 0.278
{BATTERY} ---> {Not Arrested}:  conf = 0.771, sup = 0.14
{BATTERY} ---> {Not Domestic}:  conf = 0.579, sup = 0.105
{Domestic} ---> {Not Arrested}:  conf = 0.803, sup = 0.103
{Evening} ---> {Not Domestic}:  conf = 0.889, sup = 0.195
{Evening} ---> {Not Arrested}:  conf = 0.688, sup = 0.151
{CRIMINAL DAMAGE} ---> {Not Domestic}:  conf = 0.915, sup = 0.109
{CRIMINAL DAMAGE} ---> {Not Arrested}:  conf = 0.934, sup = 0.111
{SIDEWALK} ---> {Not Domestic}:  conf = 0.915, sup = 0.105
{STREET} ---> {Not Domestic}:  conf = 0.94, sup = 0.238
{STREET} ---> {Not Arrested}:  conf = 0.718, sup = 0.182
```

Figure 20: Sample of some of the associated rules that were produced using the FP-Growth method

The results from running these algorithms showed some interesting relationships in the data. For instance, in the FP-Growth results it was shown that when the primary type of crime was ‘Narcotics’ the incident usually ended up with an arrest. However, in the rules there was some issue with the binary features. For example, both the Domestic feature

and Arrest feature appeared frequently in the rules because they are binary and highly skewed. Despite that being the case lots of insight was gained from running the Apriori and FP-Growth algorithms on the dataset.

Boston Dataset

The Boston crime dataset contained 17 features and 408,514 instances after we cleaned and preprocessed the dataset. We generated a profile report for the dataset before and after the data cleaning which gives an overview of the dataset. Here's what the dataset looked like after cleaning and preprocessing:

Dataset info

Number of variables	18
Number of observations	408514
Total Missing (%)	0.2%
Total size in memory	56.1 MiB
Average record size in memory	144.0 B

Variables types

Numeric	6
Categorical	10
Boolean	2
Date	0
Text (Unique)	0
Rejected	0
Unsupported	0

Figure 21 A brief summary of the Boston Crime Dataset

The features in the dataset mentioned on the right. The features can be explained as follows:

```
Incident Number          object
Offense Type              int64
Offense Description       int64
District                  int64
Reporting Area            int64
Shooting                  int64
Date                     object
Year                      int64
Month                     int64
Day of Week               int64
Hour                      int64
UCR Offense Level         int64
Street                    int64
Latitude                float64
Longitude               float64
Coordinates              object
Is Dark                   int64
```

Figure 22 Features in the Boston Crime Dataset

- Incident Number: The identification number for the registered crime
- Offense Type: Specifies the type of offense
- Offense Description: A brief description that explains the offense
- District: The district where the crime occurred
- Reporting Area: Code for the reporting area
- Shooting: Indicates if shooting occurred as a part of the crime
- Date: Date of the crime incident in 'yyyy-mm-dd hh:mm:ss a' format
- Year: Year when the crime occurred
- Month: Month when the crime occurred
- Day of Week: Day of week when the crime occurred
- Hour: The hour when the crime occurred (24 hour format)
- UCR Offense Level: Specifies the level of offense

- Street: The street where the crime occurred
- Latitude: The recorded latitude for where the crime occurred
- Longitude: The recorded longitude for where the crime occurred
- Coordinates: The latitude and longitude pair for where the crime occurred
- Is Dark: Specifies if the crime happed during nighttime

Shooting Feature Prediction:

The shooting feature is a binary feature which indicates if shooting occurred as a part of the crime or not. The ratio for occurrence of shooting to non-occurrence of shooting in the dataset is 99.7 to 0.3, which means there were very rare occasions when shooting occurred during a crime. The features that were used to predict Shooting are mentioned on the right.

```
Offense Type           int64
Offense Description    int64
District               int64
Reporting Area         int64
Month                  int64
Year                   int64
Day of Week            int64
Hour                   int64
Is Dark                int64
Street                 int64
UCR Offense Level      int64
```

Figure 23 Features used to predict the Shooting feature

Because the data for Shooting is skewed, the model predicts with a high accuracy when the shooting did not occur and has a harder time predicting the shooting to occur. The reason behind this can be the fact that the dataset contained many more instances of the shooting not occurring than the opposite and the model has taken that into account.

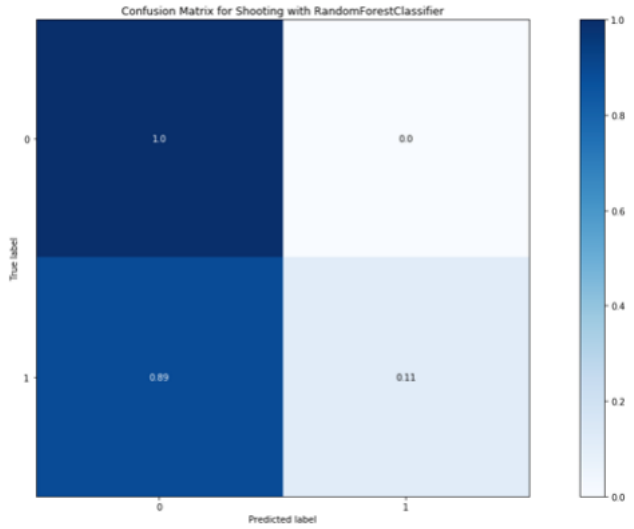

Figure 24 The confusion matrix for predicting Shooting feature using Random Forest Classifier

The results obtained by using different prediction model to predict Shooting are as mentioned below. As you can see the highest accuracy was achieved by Random Forest Classifier which is often referred to as resilient in dealing with skewness.

Feature Predicted	Prediction Model Technique	Precision	Recall	f1-Score	Log-Loss	Accuracy
Shooting	DecisionTreeClassifier	0.99	0.99	0.99	0.26	0.9925
	ExtraTreeClassifier	0.99	0.99	0.99	0.21	0.9941
	RandomForestClassifier	1	1	1	0.06	0.9969
	KNeighborsClassifier	0.99	1	1	0.10	0.9966
	GaussianNB	0.99	0.99	0.99	0.03	0.9925

Figure 25 Results from all of the various prediction models that were used to try and predict the Shooting feature

Other interesting insights that we derived from the Shooting feature are:

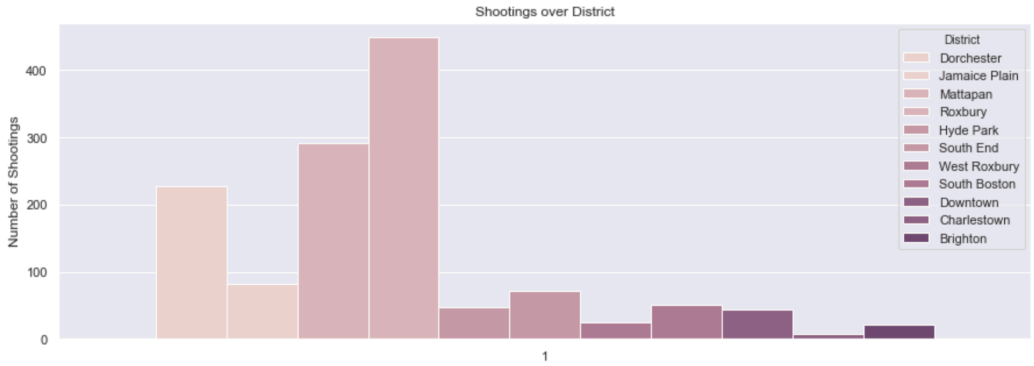

Figure 26 Number of Shootings per District

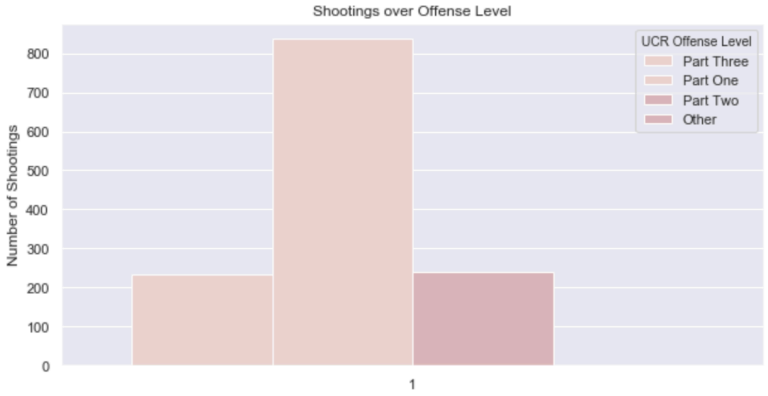

Figure 27 Number of Shootings grouped by UCR Offense Level

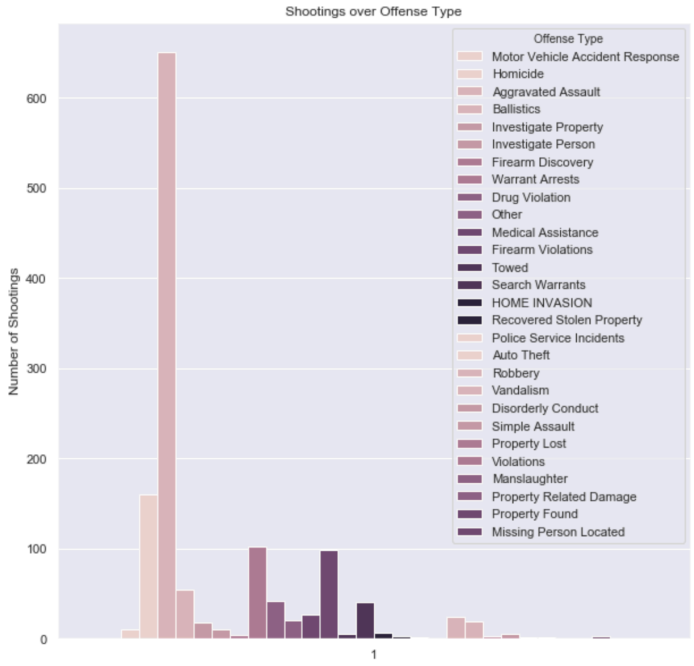

Figure 28 Number of Shootings grouped by different Offense Types

District Feature Prediction:

The District feature specifies the district that the crime occurred in. Originally, this feature consisted of district codes, but we decided to replace them with actual district names because there are a limited number of districts mentioned in the dataset and it is a categorical feature. We created a lookup dictionary for this purpose and replaced the district codes without any extra effort. The features used to predict the district feature are mentioned on the right.

```
Offense Type        int64
Day of Week         int64
Reporting Area      int64
Month               int64
Year                int64
Shooting            int64
Hour                int64
UCR Offense Level   int64
```

Figure 29 Features used to predict the District where the crime occurred

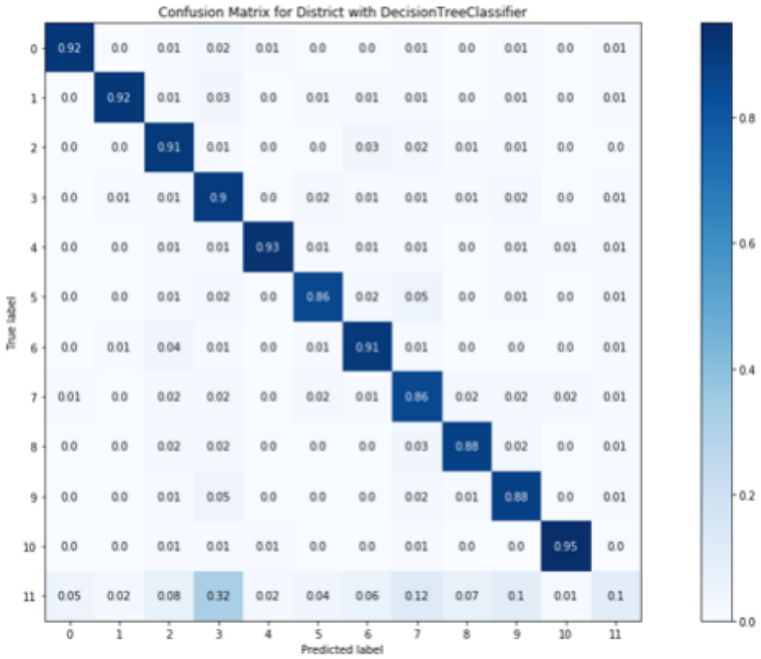

Figure 30 The confusion matrix for predicting the District where the crime occurred using Decision Tree Classifier

The highest accuracy score was achieved by Decision Tree Classifier with an accuracy of 89.10%. K-Nearest Neighbor Classifier and Random Forest Classifier also performed well when predicting the District feature. This shows that the features used to predict the

District, even though not directly associated but can provide enough information to predict the district where the crime would occur.

Feature Predicted	Prediction Model Technique	Precesion	Recall	f1-Score	Log-Loss	Accuracy
District	DecisionTreeClassifier	0.9	0.89	0.89	3.72	0.8910
	ExtraTreeClassifier	0.38	0.37	0.37	21.69	0.3710
	RandomForestClassifier	0.67	0.66	0.67	3.70	0.6647
	KNeighborsClassifier	0.8	0.8	0.8	2.91	0.7976
	GaussianNB	0.28	0.23	0.17	2.10	0.2340

Figure 31 Results from all of the various prediction models that were used to try and predict the District where the crime took place

Some of the interesting insights obtained from this feature are:

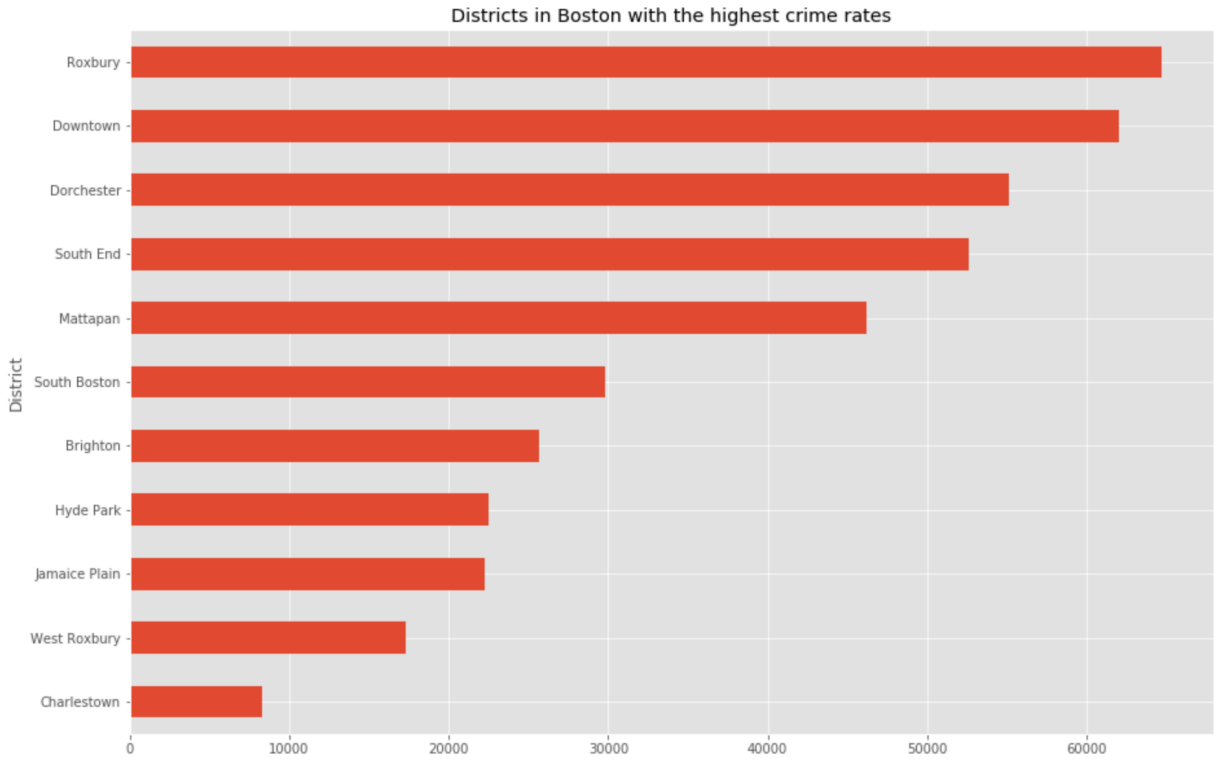

Figure 32 Districts in Boston with the highest rate of crime

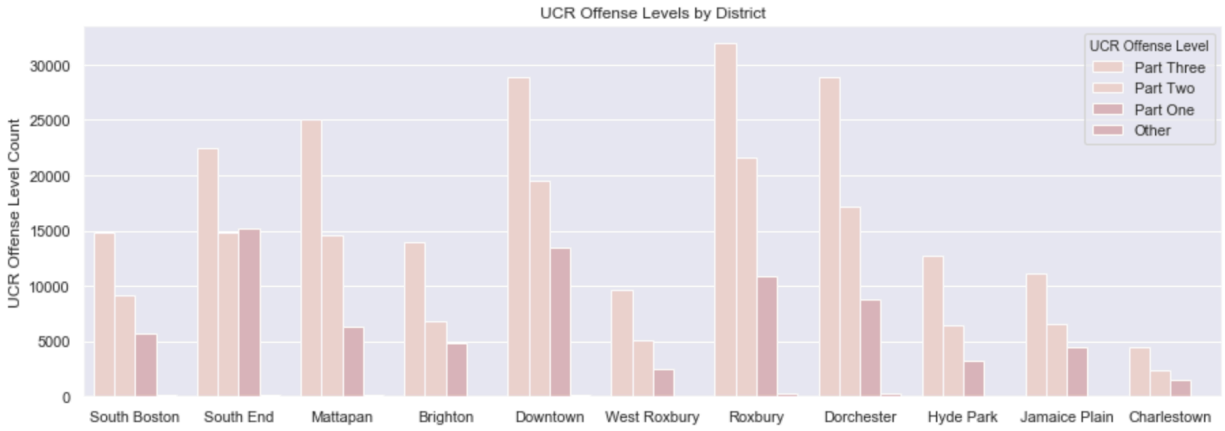

Figure 33 Different UCR Offense Levels by Districts

Offense Level Prediction:

The Offense Level feature specifies the level of offense for a crime. There are four distinct values for this feature in the dataset. The features used to predict the Offence Level are mentioned on the right.

```
Day of Week       int64
Reporting Area    int64
Month             int64
Street            int64
District          int64
Year              int64
Shooting          int64
Hour              int64
Street            int64
```

Figure 34 Features used to predict the offense level for the crime

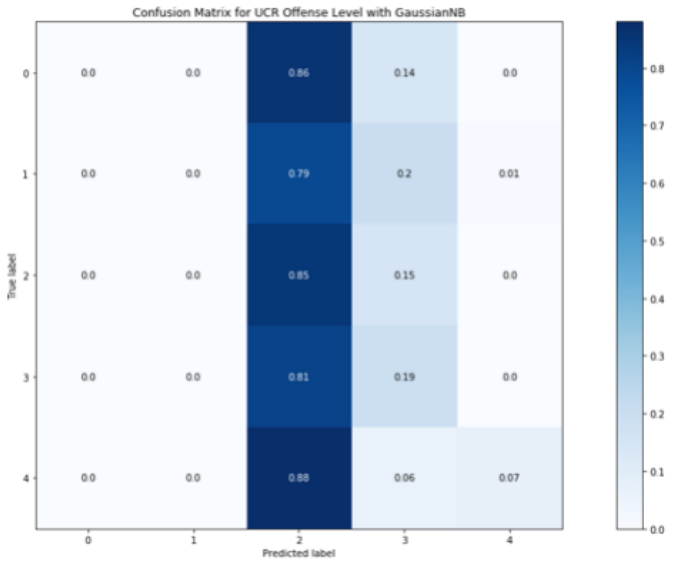

Figure 35 The confusion matrix for predicting the UCR Offense Level for the crime occurrence using Gaussian Naive Bayes Classifier

The highest accuracy for predicting this feature was obtained using the Gaussian Naïve Bayes Classifier. Gaussian Naïve Bayes performs well in case of continuous features which is the case here, but the results are still not too favorable. Overall, all the models that we used to try and predict this feature gave moderate results.

Feature Predicted	Prediction Model Technique	Precision	Recall	f1-Score	Log-Loss	Accuracy
Offense Level	DecisionTreeClassifier	0.38	0.34	0.35	18.8776	0.33791
	ExtraTreeClassifier	0.37	0.35	0.35	17.45	0.3494
	RandomForestClassifier	0.39	0.39	0.38	4.49	0.3853
	KNeighborsClassifier	0.43	0.45	0.43	5.81	0.4467
	GaussianNB	0.36	0.48	0.39	1.22	0.4818

Figure 36 Results from all of the various prediction models that were used to try and predict the UCR Offence Level for the crime occurrence

Apriori and FP-Growth:

We also applied the Apriori and FP-Growth algorithms on this dataset to mine the frequent patterns for crimes that exist in the dataset. Below are the results from Apriori:

```
{Part Three, Mattapan} ---> {No}:  conf = 0.998, sup = 0.06
{Part Three, Saturday} ---> {No}:  conf = 0.999, sup = 0.072
{Evening, Part Two} ---> {No}:  conf = 0.998, sup = 0.076
{Part Three, Jul} ---> {No}:  conf = 0.999, sup = 0.053
{Investigate Person, No} ---> {Part Three}:  conf = 1.0, sup = 0.058
{Part Three, Investigate Person} ---> {No}:  conf = 0.999, sup = 0.058
{Investigate Person} ---> {Part Three, No}:  conf = 0.999, sup = 0.058
{Evening, Part Three} ---> {No}:  conf = 0.999, sup = 0.111
{Part Three, Wednesday} ---> {No}:  conf = 0.998, sup = 0.073
{Part Three, South End} ---> {No}:  conf = 0.999, sup = 0.056
{Part Three, Friday} ---> {No}:  conf = 0.999, sup = 0.076
{Part Three, Motor Vehicle Accident Response} ---> {No}:  conf = 0.999, sup = 0.118
{Motor Vehicle Accident Response, No} ---> {Part Three}:  conf = 1.0, sup = 0.118
{Motor Vehicle Accident Response} ---> {Part Three, No}:  conf = 0.999, sup = 0.118
{Dorchester, Part Three} ---> {No}:  conf = 0.999, sup = 0.071
{Part Three, Monday} ---> {No}:  conf = 0.999, sup = 0.072
{Roxbury, Part Two} ---> {No}:  conf = 0.995, sup = 0.052
{Part Two, Other} ---> {No}:  conf = 0.999, sup = 0.053
{Night, Part Two} ---> {No}:  conf = 0.997, sup = 0.074
{Part Three, Noon} ---> {No}:  conf = 0.999, sup = 0.108
{Part Three, Thursday} ---> {No}:  conf = 0.998, sup = 0.074
{Roxbury, Part Three} ---> {No}:  conf = 0.997, sup = 0.079
{Medical Assistance, No} ---> {Part Three}:  conf = 1.0, sup = 0.076
{Part Three, Medical Assistance} ---> {No}:  conf = 0.999, sup = 0.076
{Medical Assistance} ---> {Part Three, No}:  conf = 0.999, sup = 0.076
{Part Three, Tuesday} ---> {No}:  conf = 0.999, sup = 0.074
{Part Three, Night} ---> {No}:  conf = 0.998, sup = 0.126
{Part Three, Downtown} ---> {No}:  conf = 0.999, sup = 0.073
{Noon, Part Two} ---> {No}:  conf = 0.998, sup = 0.069
{Drug Violation, No} ---> {Part Two}:  conf = 1.0, sup = 0.052
{Drug Violation, Part Two} ---> {No}:  conf = 0.998, sup = 0.052
{Drug Violation} ---> {Part Two, No}:  conf = 0.998, sup = 0.052
{Part One, Larceny} ---> {No}:  conf = 1.0, sup = 0.081
{Larceny, No} ---> {Part One}:  conf = 1.0, sup = 0.081
{Larceny} ---> {Part One, No}:  conf = 1.0, sup = 0.081
{Part Three, Sunday} ---> {No}:  conf = 0.998, sup = 0.065
{Part Three, Morning} ---> {No}:  conf = 1.0, sup = 0.159
```

Figure 37 Some of the association rules generated using Apriori Algorithm

```
{Dorchester} ---> {No}:  conf = 0.996, sup = 0.136
{Motor Vehicle Accident Response} ---> {No}:  conf = 0.999, sup = 0.118
{Motor Vehicle Accident Response} ---> {Part Three}:  conf = 1.0, sup = 0.118
{South End} ---> {No}:  conf = 0.998, sup = 0.129
{Wednesday} ---> {No}:  conf = 0.997, sup = 0.148
{Evening} ---> {No}:  conf = 0.997, sup = 0.233
{Evening} ---> {Part Three}:  conf = 0.474, sup = 0.111
{Jul} ---> {No}:  conf = 0.996, sup = 0.106
{Morning, No} ---> {Part Three}:  conf = 0.546, sup = 0.159
{Part Three, Morning} ---> {No}:  conf = 1.0, sup = 0.159
{Morning} ---> {Part Three, No}:  conf = 0.546, sup = 0.159
{Night, No} ---> {Part Three}:  conf = 0.502, sup = 0.126
{Part Three, Night} ---> {No}:  conf = 0.998, sup = 0.126
{Night} ---> {Part Three, No}:  conf = 0.499, sup = 0.126
{Noon, No} ---> {Part Three}:  conf = 0.491, sup = 0.108
{Part Three, Noon} ---> {No}:  conf = 0.999, sup = 0.108
{Noon} ---> {Part Three, No}:  conf = 0.49, sup = 0.108
{No, Motor Vehicle Accident Response} ---> {Part Three}:  conf = 1.0, sup = 0.118
{Part Three, Motor Vehicle Accident Response} ---> {No}:  conf = 0.999, sup = 0.118
{Motor Vehicle Accident Response} ---> {Part Three, No}:  conf = 0.999, sup = 0.118
{Evening, No} ---> {Part Three}:  conf = 0.475, sup = 0.111
{Part Three, Evening} ---> {No}:  conf = 0.999, sup = 0.111
{Evening} ---> {Part Three, No}:  conf = 0.473, sup = 0.111
```

Figure 38 Some of the association rules generated using FP-Growth Algorithm

We observed that UCR Level Three offenses took place mostly on Sundays and during the morning time and did not involve shooting. UCR Level Three offenses were prevalent in Downtown, Dorchester and South End district areas. We can see from the association rules that UCR Level Three offenses, no Shooting occurrence and Motor Vehicle Accident Response Offense are closely related features. Overall, the results were slightly meaningful and helped us in observing what features are closely associated.

Discussion:

We observed that most of our experiments resulted in moderate to highly accurate classification and prediction models. We realized early on that Linear Regression would not be suitable for these datasets because the data is mostly non-linear. Principal Component Analysis turned out to be helpful in choosing features to use for predictions. We experimented with various classifier models in order to improve the accuracy for our predictions. Overall, we were able to get pretty accurate results using the Decision Tree Classifier. In cases when Decision Tree Classifier gave poor results, Random Forest

Classifier and Gaussian Naïve Bayes Classifiers helped improve the accuracy to acceptable levels. K-nearest Neighbors also helped improve the accuracy for some features. Below are the average accuracies that we were able to achieve using different classifier models:

Overall Results		
Prediction Model Technique	**Log Loss Avg.**	**Accuracy Avg.**
DecisionTreeClassifier	9.7933	0.7012
ExtraTreeClassifier	19.2175	0.4249
RandomForestClassifier	6.1777	0.5058
KNeighborsClassifier	5.3605	0.6228
GaussianNB	1.7887	0.4155

Figure 39: Averages of the prediction models that were run on both the Boston and Chicago datasets

As you can see from the above table, the Decision Tree Classifier and K-nearest Neighbors Classifier outperformed the other classifier models that we used. Random Forest Classifier generated almost as accurate results as the two classifier models mentioned, while Extra Tree Classifier and Gaussian Naïve Bayes resulted in moderate average accuracies. For different features, different models predicted results with varying accuracies and we think this is because all of these features had idiosyncratic properties in terms of how the data was distributed.

Future Work:

We would like to expand upon this project by including data for more cities around the United States. We also plan to experiment with other classification models such as Support Vector Machines and XG Boost. The datasets also contain the coordinates for the crime locations which can be used to create a heatmap for active crime locations. These heatmaps would be really helpful for general public and various agencies to understand the changes in crime rate and crime locations, and help promote a safer environment and society. We tried to do that as a part of that project, but it ended up taking a big chunk of our time, so we had to move on. Below is an example for one of the heatmaps we generated:

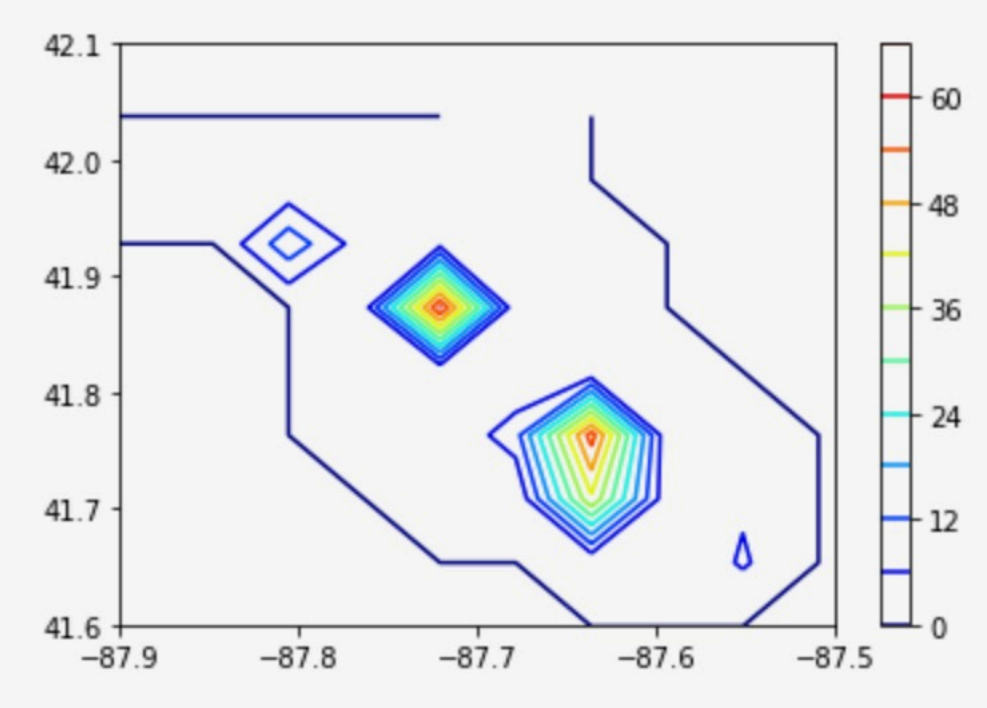

Figure 40 An attempt at generating a heat map for active crime locations

Moving further we would also like to create an interactive web app where we can present our findings and observations to make it available for the general public.

Conclusion:

The data needs to be highly preprocessed before we can apply any of the classification models or association rule algorithms. Principal Component Analysis helps in deciding what features should be used while predicting a feature and different models should be explored based on what type of features we're trying to predict and the nature of the data.

From all the experiments that we did for this project, we concluded that not all of the features for a dataset can be predicted in the same way. Each feature in the dataset is

unique and has to be predicted using suitable classification methods which can be decided based on data distribution, skewness and number of classes etc.

References:

Barkan, S. E. (2016). Social Problems: Continuity and Change. In S. E. Barkan, *Social Problems: Continuity and Change.* Flat World Knowledge, Inc.

Norman, J. (2018). *Americans' concern about national crime abating*. Retrieved from Gallup: https://news.gallup.com/poll/244394/americans-concerns-national-crime-abating.aspx

Saltos, G., & Cocea, M. (2017). An Exploration of Crime Prediction Using Data Mining on Open Data . *International Journal of Information Technology and Decision Making*.

Wu, S., & Wang, C. (2019). Crime Prediction Using Data Mining and Machine Learning. *Advances in Intelligent Systems and Computing*, 1-3.

Yu, C.-H., Ward, M. W., Morabito, M., & Ding, W. (2012). Crime Forecasting Using Data Mining Techniques. *IEEE*, 1.