CRIME DATASET ANALYSIS AND PREDICTION

Prajakta Rodrigues, Samarth Parikh, Sangeetha Chandrashekar, Tej Bhavsar

CS6220 Data Mining Techniques (Fall 2019), Northeastern University

1. ABSTRACT

We live in a world where we constantly must be alert of our surroundings and the lower crime rate is the most desirable quality while searching for a place to live. But, the crime rate around the world has been steadily increasing. It has become necessary for authorities to take measures to reduce the crime rate and protect the general public. Analysis of available crime dataset is pertinent in finding the commonalities between different types of crimes, various factors influencing higher crime rate, detecting unique patterns and recognizing the diverse crimes and their relations with several criteria like demographic, locations of the crime, time, year, nature of crime, etc. We are going to analyze the comprehensive crime dataset made available by the City of Chicago, the Chicago Police Department and Boston Police Department to find patterns in a variety of crimes. The analysis of this dataset will provide insights which can be used by the respectable authorities and officials to provide better security to their citizens. Crime data analysis can be used to control the crimes in the areas, increase task force and avail intelligent agents like criminologists to monitor as well as keep a check on the crime rate, based on the patterns revealed. We plan to use various machine learning techniques, association mining algorithms and clustering algorithms, classifiers and time series to find associations between different features, performing exploratory analysis and forecasting crime based on the results obtained. These rules and patterns will help to understand the correlation between the different features and types of crimes, palpable in empowering the relevant agencies.

2. INTRODUCTION

One of the most important factors that predict the quality of life in a city is the crime rate. Every citizen would want to live in a safe and friendly neighborhood. However, crime in some form or the other exists in society. Hundreds of crimes are recorded daily by the data officers working alongside law enforcement authorities throughout USA. It might not always be possible to control whatever happens in our surroundings however, we can certainly attempt to recognize the threats in the society, identifying the distinct factors that drive up the crime rate which would help us be more aware about our surroundings and take preventive actions to make certain crimes defunct and increase patrolling in certain areas based on the deductions. With the advancements in data mining and different types of machine learning algorithms, analysis and predictions employment on huge data set, which are heterogeneous and multi-sourced, has abundant influence in various applications. The insights revealed can be further analyzed by experts to develop models that drive solutions to be more effective and accurate. Although seemingly erratic data sets, like crime, they would reveal patterns in a broader perspective when used with appropriate learning and association

algorithms. The developments revealed by the application of data mining techniques on crime dataset can complement the efforts of the authorities in extraction and understanding the different types of crimes, their influencers and ultimately help in assisting crime prevention and management. Hence, taking inspiration from the statements above, we have decided to examine the data provided in the Chicago and Boston crime data sets and analyze them to identify as well as categorize the trends in crime over the years. We then proceed to develop a forecast for certain types of crimes in both the cities and compare the accuracies. We also attempt to detect similarities in crime patterns between the two major cities in the USA to report any major commonalities.

Historically, Chicago has been a city with a high crime rate. The homicide rate in the city is significantly higher than that of other large cities in the USA but lower in comparison to the other smaller cities. In order to address the problems that lead to a high crime rate in the city, it is essential to identify the trends that lead to these crimes.

By identifying areas and neighborhoods with high crime rates in Chicago and with the help of other historical, cultural and demographic data, it would become possible to point out patterns like demographic responsible for these crimes, motives behind such crimes etc. This information could turn out to be pertinent for the Police in order to set up various rules and regulations to decrease the crime rate and promote awareness to general public. These solutions could be applicable to other cities facing similar issues, hence can have an advantage in decreasing the crime rate in different cities.

In order to execute the analysis of the crime setting in Chicago and Boston, the problem we are trying to address, and the steps taken to bring about the solution for our problem statement above can be described in two parts:

1. Performing exploratory analysis of the data to mine patterns of crimes in Chicago

   The first step is determining the rate of crime in certain locations in Chicago for the years mentioned in the dataset. Once we figure out the areas with high crime rate, the next step is to determine if there are any trends in the crimes committed in these areas. These trends could include identifying types of crimes that are most frequently committed in each location or how a certain crime has changed over the years or finding out the hotspots for a certain crime in the city. After effective identification of trends in crimes, it would be useful to determine the relation between the arrests and the crimes committed. These could include the relation between number of arrests and crimes and would be essential in getting more insight about how these crimes have been addressed so far. Final step is to classify the types of crimes, determine feature association and forecast the crime outlook for the next year using time series model.

2. Performing exploratory analysis of the data to mine patterns in crime in Boston and its comparison with Chicago

   Boston, just like Chicago is one of the major cities in the USA. Owing to different demographics, culture, ethnicity etc. the crime scenario in Boston would be different from that of Chicago. However, as both are mega cities, there would be some commonalities

between the two. In the second section we would be comparing the patterns observed in the crime rate between the two cities. Having already found some of the common trends in crime in the city of Chicago, the next step would be to examine the crime scenario in Boston and then compare it to the trends we identified in Chicago.
We would be comparing the crimes that are more prevalent in Boston with those in Chicago which will help us understand and help improve some aspects of safety in both the cities. Apart from the trends in crime, we would also be comparing the number of arrests made in both the cities which would give us a fair idea about the current systems in place for safety in each city and how successfully the crimes have been combated.

As a part of our literature survey, we read through many effective models that have been used to apply for different sizes of data and efficiently using the machine learning algorithms to classify, regress and predict the outcomes. One such paper by Mingchen Feng , Jiangbin Zheng, Jinchang Ren, Amir Hussain, Xiuxiu Li, Yue Xi and Qiaoyuan Liu [4] use Big data analysis and applications of Neural Network to predict crime trends in San Francisco and forecast it, thereby revealing the different, effective types of time series models that can be used for effective forecast involving big dataset. Other papers by Isha Pradhan[5] also effectively use Apache Spark for crime analysis and generating trends of crimes in downtown San Francisco. The most effective application of crime data analysis is "Series Finder" derived from a paper by Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri [6] which is successfully using machine learning algorithm that is trained to detect patterns in and prevent burglary by assisting the Cambridge Police Department, Massachusetts develop a modus operandi (M.O) of the offender and find patterns. There are many more research papers and applications that reveal interesting patterns of the crime and other big data analysis, that we draw inspirations from for our research today. Any other papers referenced will be judiciously noted in the upcoming sections.

3. DATASET

We are utilizing the 'Crimes in Chicago' dataset for this project. This dataset reflects reported incidents of crime (except for murders where data exists for each victim) that occurred in the City of Chicago from 2001 to 2017. Its features consist of information on Crime type, Case number, Location, Latitude, Longitude, whether it is domestic or not etc.
The dataset is availed from - https://catalog.data.gov/dataset/crimes-2001-to-present-398a4

Apart from this we will be also utilizing 'Crimes in Boston' dataset to compare general crime trends between Chicago and Boston. This dataset is comparable to the dataset of crimes in Chicago.
Boston crime dataset is availed from - https://www.kaggle.com/AnalyzeBoston/crimes-in-boston

Attributes in 'Crimes in Chicago' dataset

Feature	**Description**
ID	a unique ID for the particular crime data
Case Number	The Chicago PD Records Division Number, which is unique to the incident
Date	Date when the incident occurred. this is sometimes a best estimate

Block	The partially redacted address where the incident occurred, placing it on the same block as the actual address
IUCR	The Illinois Uniform Crime Reporting code. This is directly linked to the Primary Type and Description
Primary Type	The primary description of the IUCR code
Description	The secondary description of the IUCR code, a subcategory of the primary description
Location Description	Description of the location where the incident occurred
Arrest	Indicates whether an arrest was made
Domestic	Indicates whether the incident was domestic related as defined by the Illinois Domestic Violence Act
Beat	Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts. See the beats at https://data.cityofchicago.org/d/aerh-rz74
District	Indicates the police district where the incident occurred
Ward	The ward (City Council district) where the incident occurred
Community Area	Indicates the community area where the incident occurred. Chicago has 77 community areas
FBI Code	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS)
X Coordinate	The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block
Y Coordinate	The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block
Year	Year the incident occurred
Updated On	Date and time the record was last updated
Latitude	The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block
Longitude	The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block
Location	The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block

4. METHODOLOGY

4.1. Data Preprocessing

Data preprocessing is an important step in data mining and analysis. Real world data are generally incomplete, noisy and inconsistent. Data preprocessing step ensures that the data set we are working on will be free of the above that allow further exploratory or predictive analysis on the data can be effectively relied on. Data preprocessing consists of data cleaning, data integration, data reduction and data transformation. In the datasets we are using above, we will only be applying data cleaning, data reduction and data transformation steps since the data integration step does not apply to our data set. For our dataset, data cleaning is essential since any invalid or null values can affect the analysis. In our dataset, we found missing values in features like location, x coordinate, y coordinate, latitude, longitude, community area and ward. These are features that cannot be filled with any random or aggregate values since it would lead to inconsistent data. Hence, we dropped the rows that contained missing values using the *dropna( )* method. Next, we handled the duplicate values by retaining a single copy of these duplicates and removing the other similar rows using the *drop_duplicates( )* method and setting the *subset* parameter to 'Case number'. We then proceed to data reduction and data transformation steps by discarding the features that are not relevant to the exploratory analysis of the crime dataset. The discarded features do not add any value in generating predictive results in the future classification models. These features are '*location(lat, long)*' , '*X coordinate*, '*Y coordinate* '*ID*' , '*Updated On*', '*FBI Code*', '*Description*'. Further, we transform the values of '*Arrest*' and '*Domestic*' features from *true* or *false* to 0 and 1 values, and date from object to date time format.

4.2 Exploratory analysis and narrative visualization

After cleaning the dataset, we performed an initial exploration of the dataset to understand the dataset. (More in depth observations will be drawn up later)

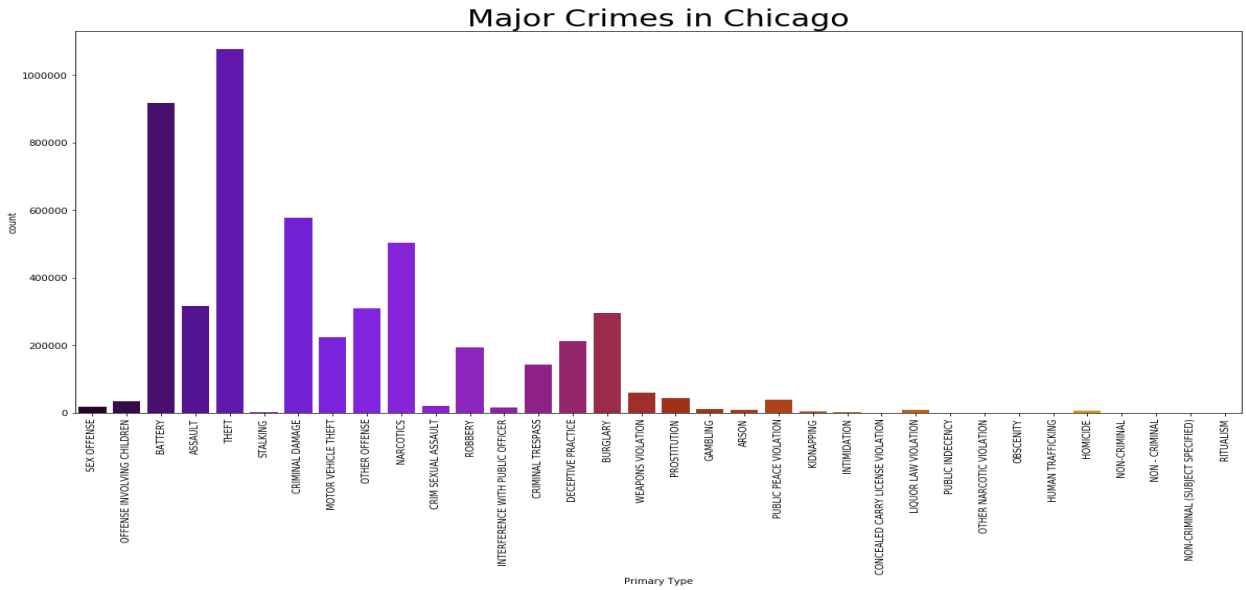


Figure 1 Major Crimes

Plotting the number of crimes against their primary type we observed that the major crimes types of the Chicago crime dataset are theft, battery, narcotics and criminal damage. Certain crimes like rituality, intimidation and indecency are very infrequent.

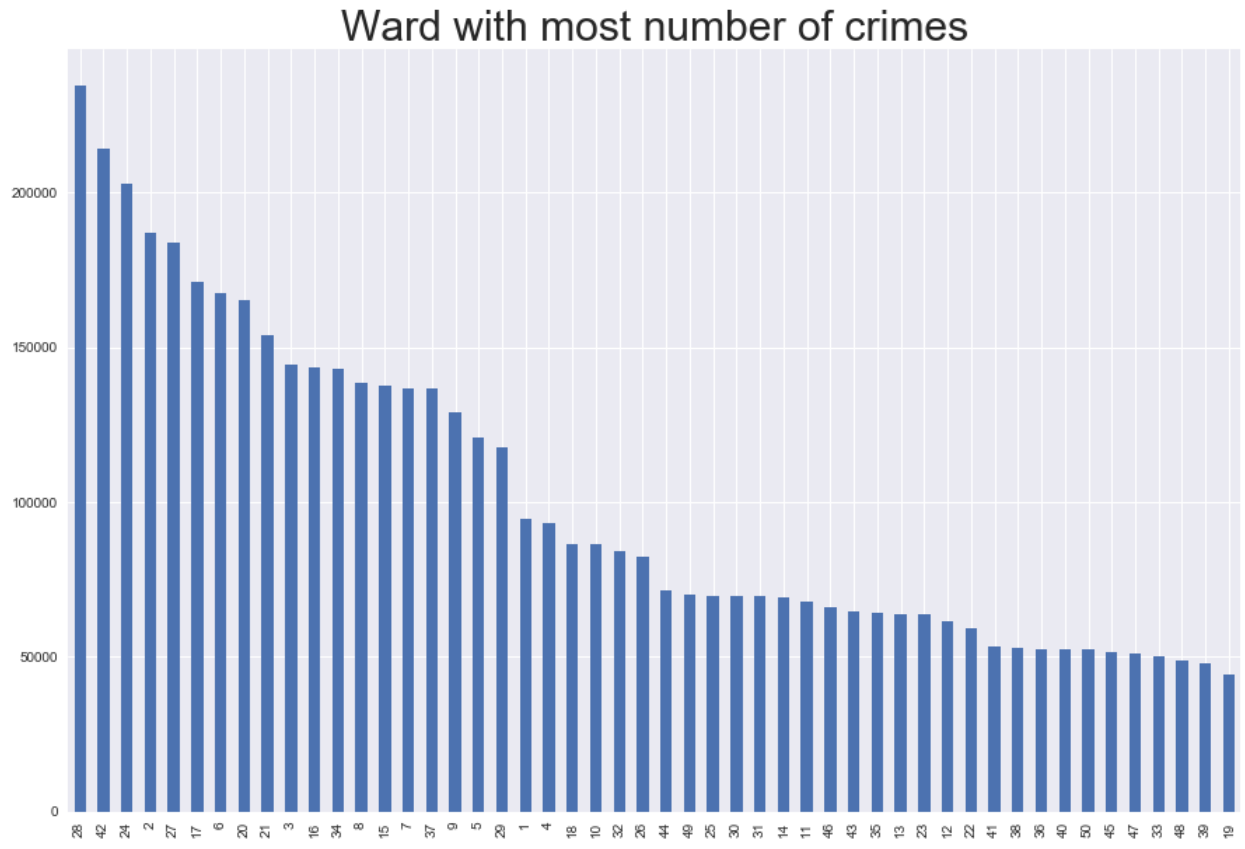


Figure 2 Wards with most number of crimes

We plotted the number of crimes per ward and noticed that a few wards like 28, 42, 2 and 24 show a higher crime rate than other wards. Also, the ward wise distribution of crimes follows a skewed distribution.

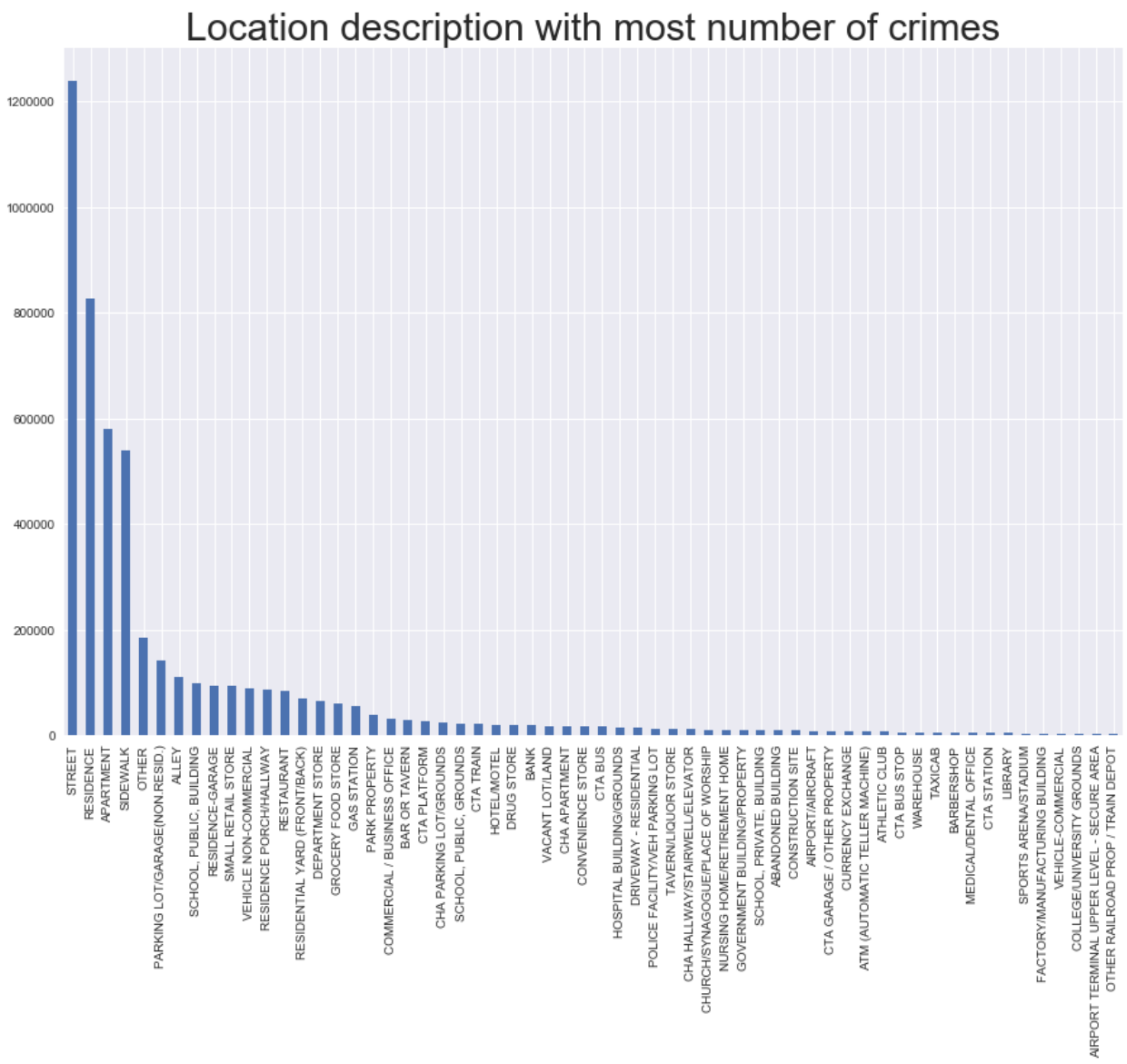
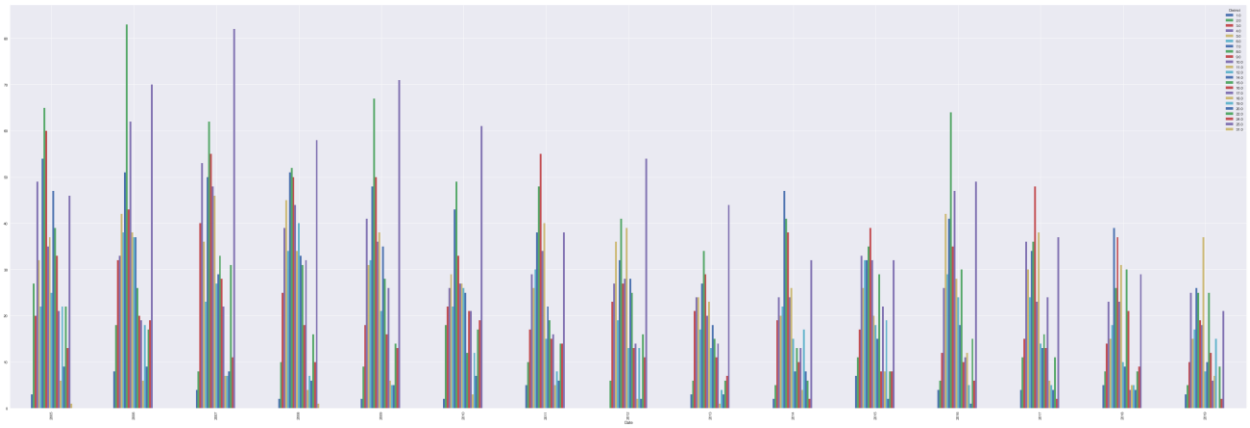


Figure 3 Location description with most number of crimes

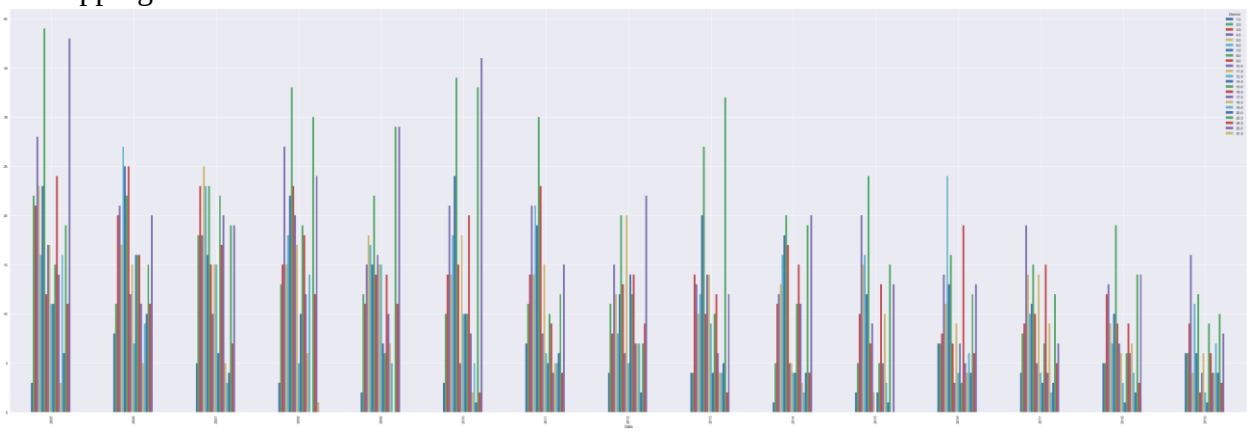
We plotted the number of crimes to the locations and discovered that maximum crimes occurred on the street, in residences, apartments and sidewalks. Crimes at any other location are quite low in number to these locations.

Below are the plots of different crime types throughout the years in various districts for Chicago dataset –

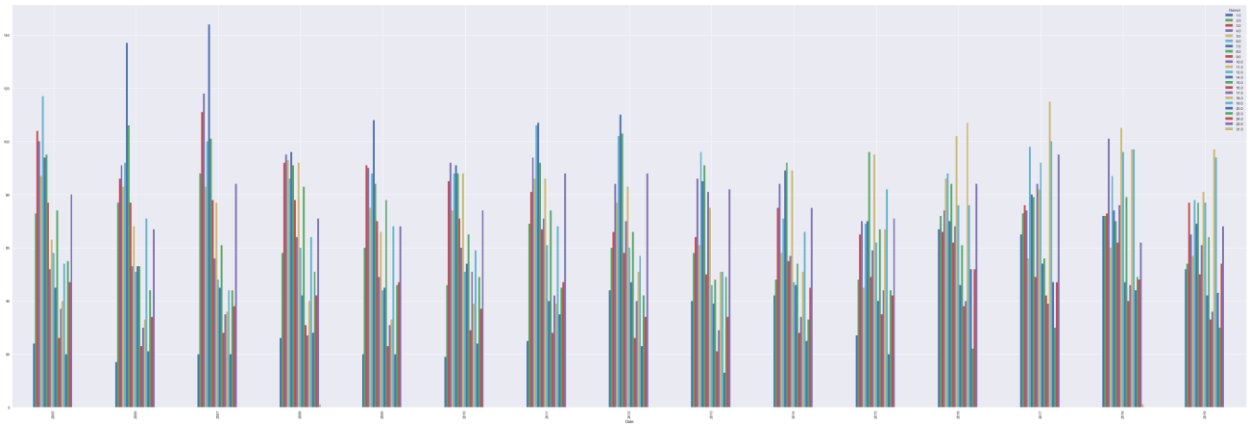
1. Arson



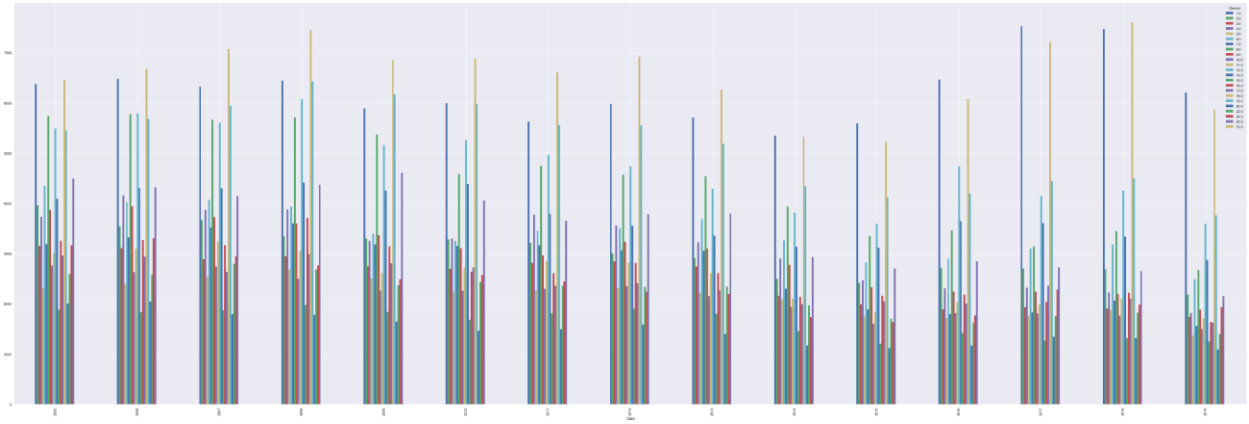
2. Kidnapping



3. Crim Sexual Assault

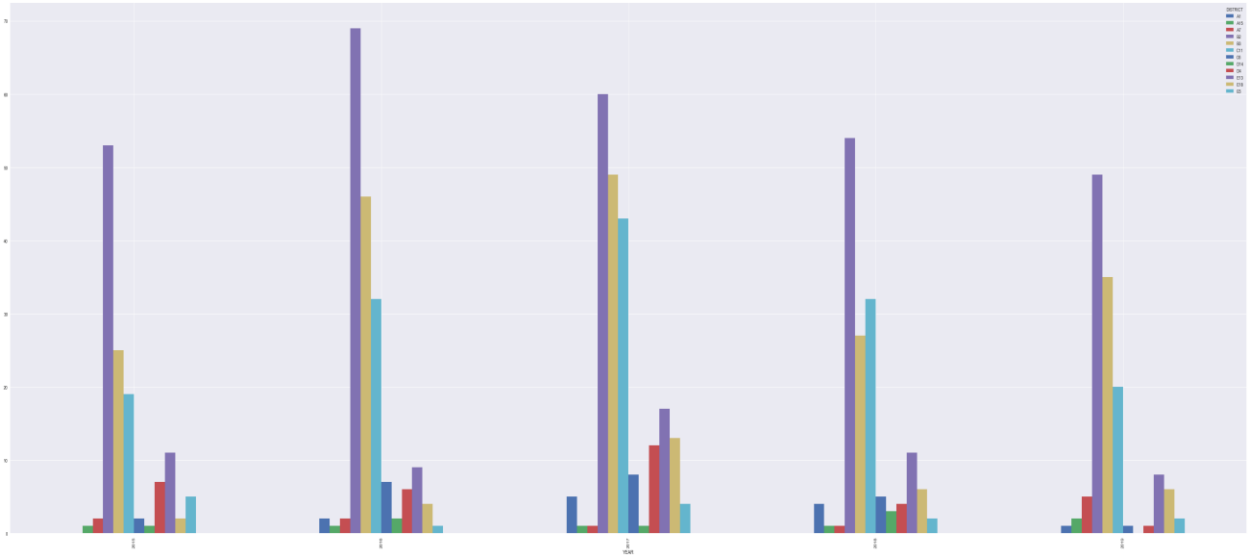


4. Theft

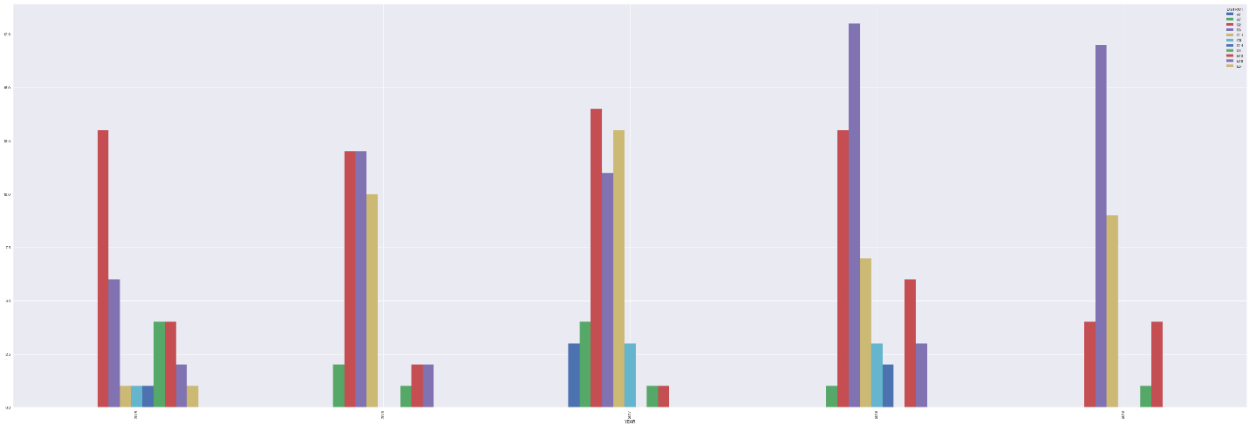


Below are the plots of different offense code types throughout the years in various districts for Boston dataset -

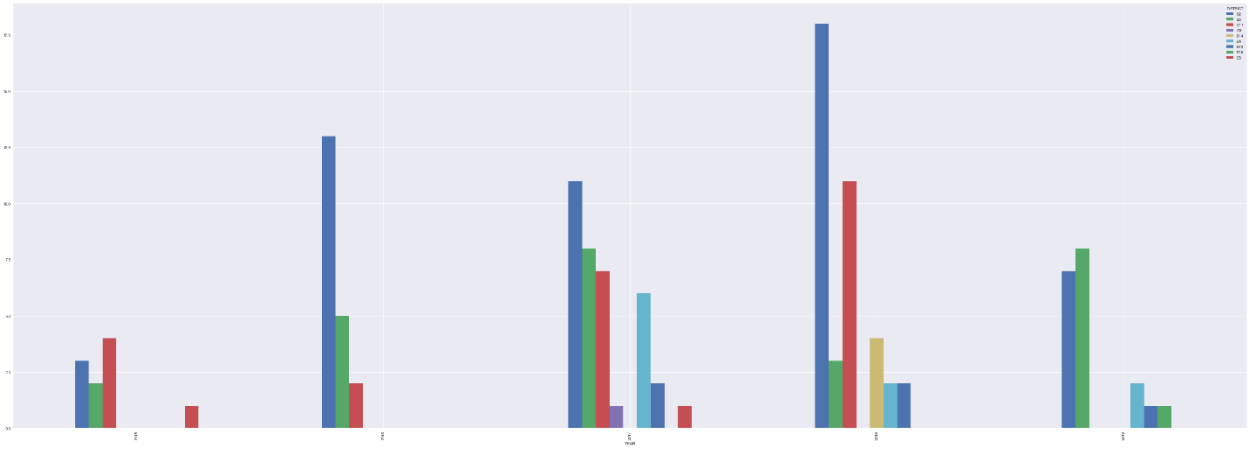
1. Aggravated Assault



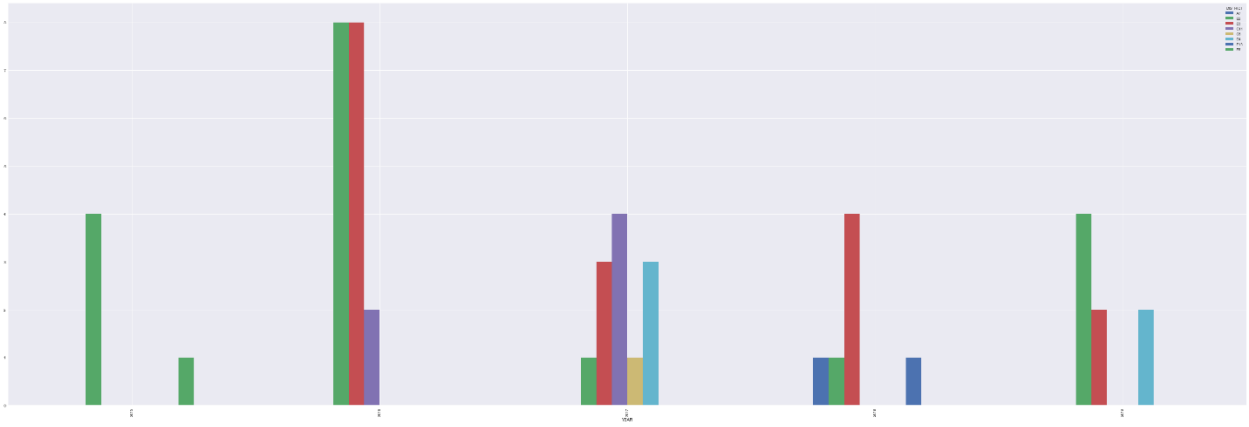
2. Homicide



3. Firearm Violations



4. Drug Violation



4.3 Model

We apply two types of analysis in our dataset to understand the dataset, determine patterns and perform predictions to tackle the problem of crime trends forecasting and analyze the dependency in features. These are classification and time series model.

4.3.1 Classification

Classification models predict a target class or category for a given set of data and their features based on the observations drawn from the training set with same set of features. For our crime dataset, supervised classification is used to predict the category of the crime incident reported. The classifiers we used are Decision Trees and Random Forest.

4.3.1.1 Decision Trees

Decision trees model is a learning method for classification to iteratively partition the instant space. Decision trees are a directed tree with the first node as “root” node. The leaves of the tree are called decision nodes or terminal nodes and are associated to a categorical value. The decision node can

also have probability vector that supports the truth probability of the prediction. The splitting of the tree using the input attributes is based on discrete functions. If one were to start from the root of the node with a specific set of inputs for the given features that have been used to generate the tree, the person is redirected through a series of nodes to a determined leaf based on the inputs. This node provides the decision of which class the set of inputs belong to. Decision tree uses entropy and information gain to split the nodes and arrive at a complete decision tree.

The process of obtaining decision tree is:

Step 1. Calculate the entropy using frequency table of one attribute using the below formula

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

where S is the attribute and p_i is the probability of different values of the given attribute

Step 2. Calculate the entropy based on frequency table using two attributes.

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

Where T,X are the two attributes and P is the probability of every value of X and E is entropy of that value.

To break the different values obtained from entropy, we use information gain that analyses the decrease in entropy after the dataset is split on an attribute. The decision tree is based on finding attributes that return the highest information gain. Formula for information gain is as given below:

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

4.3.1.2 Random Forest

Random forest model is an ensemble learning method for classification that operates by constructing a multitude of decision forests at training and outputting the class based on observed behaviors. Random forest uses random sampling of training data points when building trees as well as random subset of features are considered when splitting up the features. The result based on different prediction decision trees solve the problem of overfitting that is often seen in Decision tree classification. While constructing trees with Random Forest, we use the Gini index as an impurity measure. Gini index gives the separation measure between the probabilistic measure of the target attribute's values. In a decision tree generated with data D of n samples and feature vector $\{\mathbf{X}_i\}$, $i = 1$ *to* d [1]

$$D = \{(\mathbf{X}_1 , y_1), \ldots\ldots, (\mathbf{X}_n , y_n) \}$$

where each feature vector

$$\mathbf{X}_k = (x_{k1}, \ldots\ldots, x_{kd})$$

the top node contains all the examples ($\mathbf{X}_k$, y_k) and the set of examples is sub divided among children in each node. This continues till the leaf nodes are assigned to a class. At each node, random feature x_i and threshold *a* are chosen to minimize diversity in children nodes. This diversity reduction is measured by Gini index. Python provides RandomForestClassifier in it's scikit-learn library, that we will apply to our crime dataset to predict the crime types/classes by using the data from other features.

4.3.2 Time Series

One of the objectives in this report is to determine the crime trends in Chicago and Boston over a period. We can achieve this objective using the time series model. Time series is a collection of data points taken over a constant time interval to determine long time trend to forecast the future. Our datasets are collected over a period and are perfect examples for a time series prediction. With the results from the time series model, we can also observe any seasonal trends that arises in different types of crimes. For our time series prediction, we use the Prophet model which is a procedure for forecasting time series data based on additive model where non-linear trends are fit with yearly, weekly, daily seasonality plus holiday effects[2]. It is designed to handle complex features in the time series, has intuitive features which can be adjusted without knowing the details of the underlying model. Crime data can be classified under multi-period seasonality (daily, weekly, and annually). The seasonal functionality relies on the Fourier Series defined as below, where P denoted regular period(daily, weekly, and annually). [3]

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\frac{2\pi nt}{P} + b_n \sin\frac{2\pi nt}{P} \right)$$

Prophet model then combines the trend, seasonality and holidays to generate time series using the below equation:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

where *g(t)* represents the trend of non-periodic changes in the time series, *s(t)* represents the periodic or seasonal changes in the time series and *h(t)* refers to the holiday effects. ε_t is the error term. Prophet model is available on the sklearn and we can use the *fit( )* and *predict( )* methods to generate the time series and the forecast using the training and testing data.

4.3.3 Association Mining

Association rule learning is a rule-based machine learning algorithm for discovering interesting patterns between items(features) in datasets. The form of an association rule is $I \rightarrow j$, where I is a set of items and j is an item. The implication of this association rule is that if all the items in I appear in some basket, then j is "likely" to appear in that basket as well. The term "likely" is formalized using confidence given which is the probability of j given $I = \{i_1, \ldots\ldots, i_k\}$

Support (I) = Number of occurrences of I / Total number of items

Confidence (I -> j) = Support (I U j) / Support (I)

Lift (I -> j) = Support (I U j) / Support (I) * Support (j)

Although association mining is largely used with Market-Basket model, it has many other applications as well. So, we are going to be using association mining to generate rules for our dataset using each crime record as a basket and all the relevant features as items.

5. CODE

The code consists of the reading the csv files and converting them to data frame using the panda library. The data then undergoes data pre-processing(finding missing values, removing the missing values, removing duplicates, reducing the cardinality of the dataset by removing unnecessary features, modifying the feature data type to best suit the analysis, etc.) and exploratory analysis, as revealed above. The applications of the different classifiers and time series model will soon be added, and the patterns revealed will be analyzed through visualizations and metrics . The code will contain the description and in depth use of classifiers and time series models made available by the Python's panda and sci-kit library.

The GitHub(CCIS) link containing the code and detailed code description:

https://github.ccs.neu.edu/sangeethac/CS6220-DMT-Project

6. RESULTS

6.1 Classification

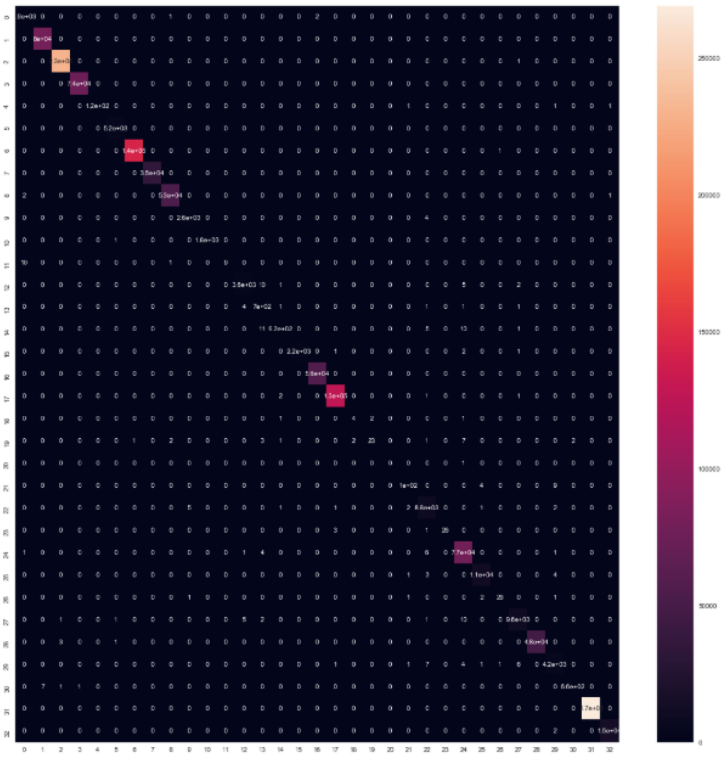
Below are the results obtained from the classification applied on the Chicago and Boston dataset. The primary objective for applying classification algorithm for our datasets is to identify the different crime types based on a partial feature set for a training dataset and to predict the crime types for the testing test. We also run metrics on the classifiers to determine the accuracy and reliability of classifiers model for our datasets.

6.1.1 Classification: Chicago

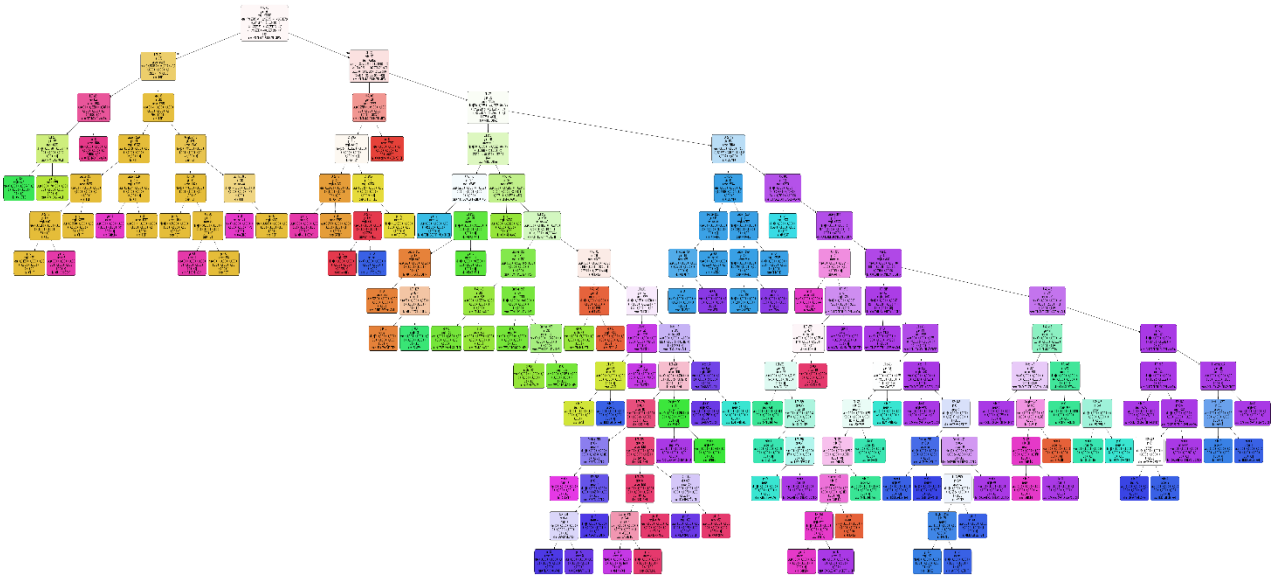
Chicago police have identified 34 different crime types in their records. These unique values are found under 'Primary Type' feature in the Chicago dataset. The features applied for the classifiers are ID, IUCR, Description, Location Description, Arrest, Domestic, District, Ward, Community Area, Year, Latitude, Longitude.

First, we transformed the certain multiclass label using Label Encoder from sklearn library to fit them into the classifiers. The features that underwent transformations were IUCR, Primary Type, Description and Location Description.

Next, we fit the model into the Decision Tree and Random Forest classifier from the sklearn library. Each of the classifiers were fit with three variations of the feature set and analyzed. The first fit was with the feature set including the IUCR column. This model gave the accuracy of 0.9999952438139855 or 99.99% for decision trees and 0.9997525308087245 or 99.97% for random forest classifier. This model would be deemed unfit for real life prediction since IUCR is directly related to the primary type and would not yield any useful results since IUCR would not have been assigned to unopen cases/reports. The heatmap obtained from the confusion matrix generated for this result is as shown below.



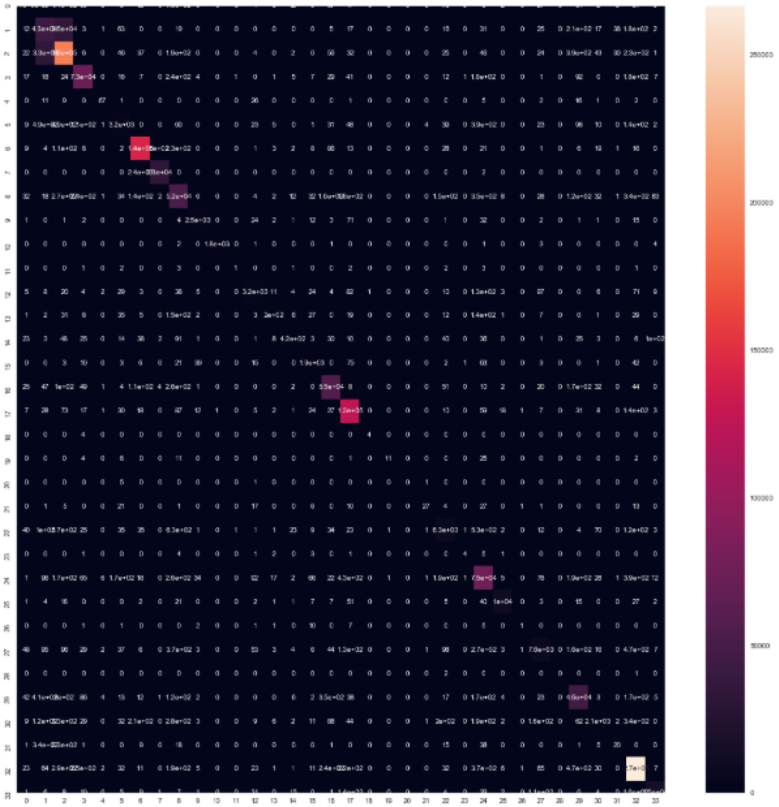
The decision tree generated for the first variation is as below:



Therefore, we discarded IUCR from the features and applied the remaining to the classifiers. The accuracy result obtained from fitting the data to decision trees is 0.6855138464465346 or 68.55% and the result from obtained from random forest classifier is 0.9224866211968467 or 92.24%. The heatmap obtained from the confusion matrix for this result is as shown below.

The final variation applied to the classifiers was with both IUCR and description removed from the feature set. The accuracy result obtained from fitting the data to decision trees is 0.4857846689365 or 48.57% and the result from obtained from random forest classifier is 0.39764650453733175 or 39.76%. Since we already obtained better results with the random forest in the previous variation, we will not be using the first and last variation result for further metrics.

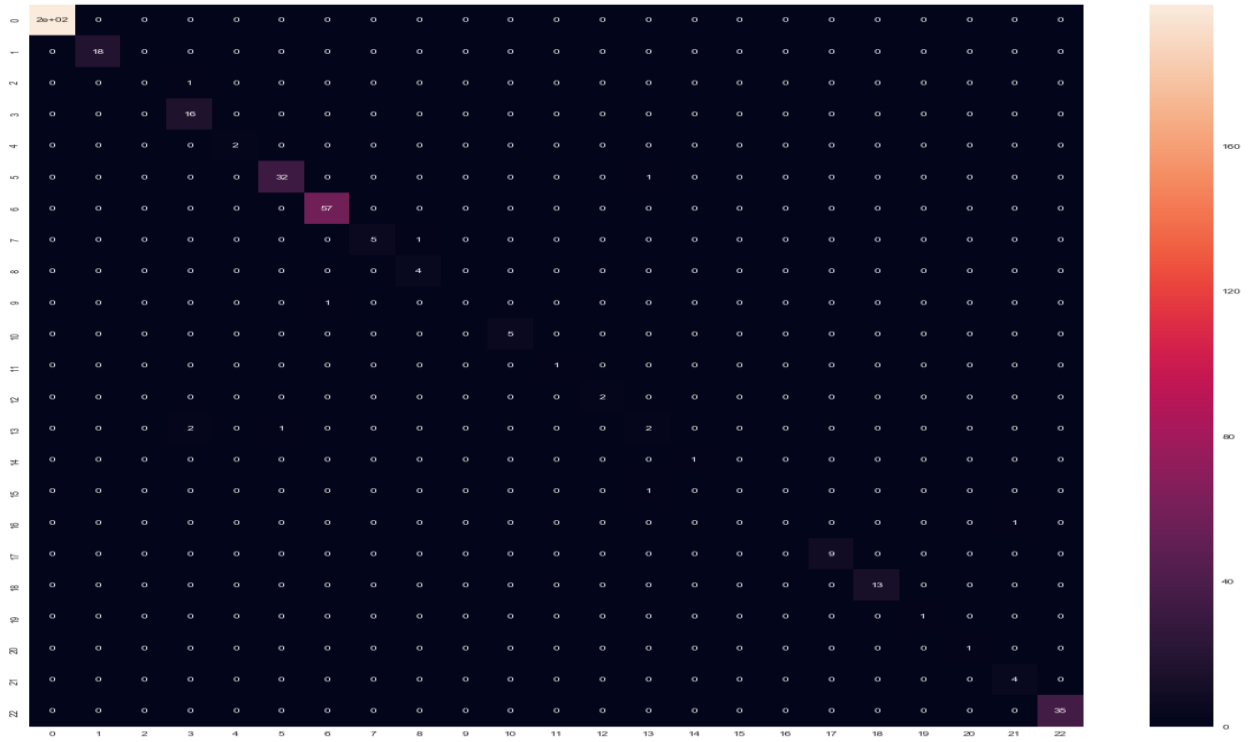
The model of applying the data features without IUCR to random forest gives the best result i.e. 92.24%. This model is obtained without any inadvertent overfitting since the random forest determines the classes using evaluations from various decision trees and thus eliminating bias. We further calculated the metrics to support our model's prediction results. The root mean squared error was 2.4698460195028735. This is a good result for our Chicago dataset features and shape. The F1 score was 0.9214697805854445 and supports our recognition that the model obtained by using random forest without IUCR feature is a reliable prediction method for generating the crime type given a feature set. A detailed classification report can be found in the code.



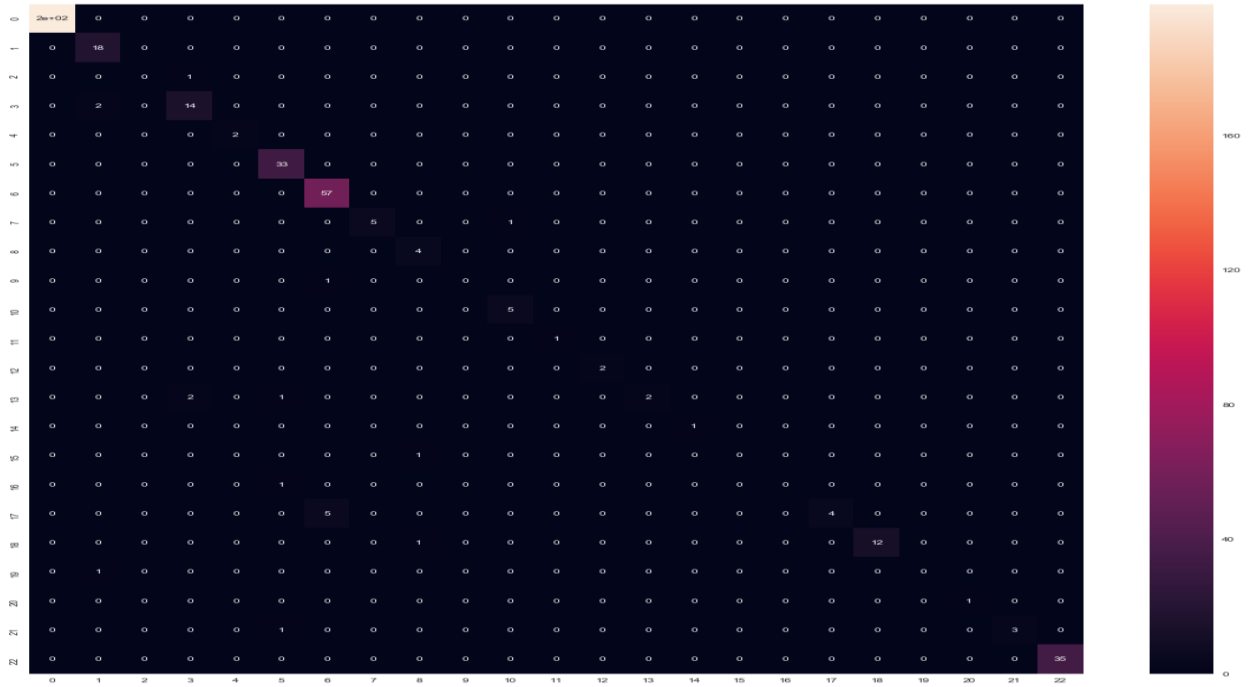
6.1.2 Classification: Boston

There are 27 different types of crimes under the 'offense_code_group' feature. We tried to classify given data under one of these crime types.

The offense_code_group feature is directly related to 'offense_code'. Hence when we include the 'offense_code' in our set of features to train we get very high accuracy for the model i.e. 0.9783653846153846 or 97.8%. The heatmap obtained of confusion matrix for the above model is shown below.



When we applied Random Forest Classification without using the 'offense_code' we got an accuracy of 0.9567307692307693 i.e. 95.6%. Below is the heatmap of confusion matrix for this model.



Again, when we classified the data without both 'offense_code' and 'offense_description' we got an accuracy of about 0.6105769230769231 i.e. 61.05%.

Since we have applied Random Forest Classification 3 times, the model without 'offense_code' and with 'offense_description' is the best one to give good results. The metrics related to this model are shown below

F1 score: 0.9481022485736115

Root Mean Squared Error: 2.318922992717491

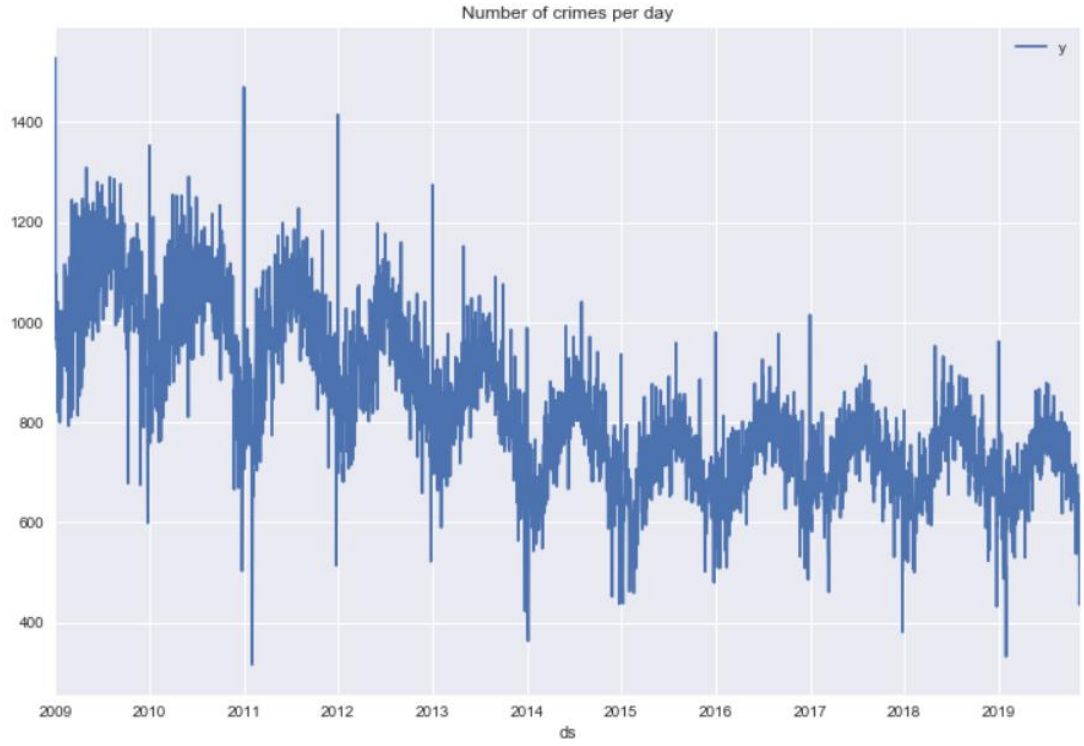
6.2 Time series Model

In the two sections, we demonstrate the results i.e. the forecast obtained by applying prophet model to our datasets for the year 2020. The metrics to further support prophet's prediction is also mentioned.

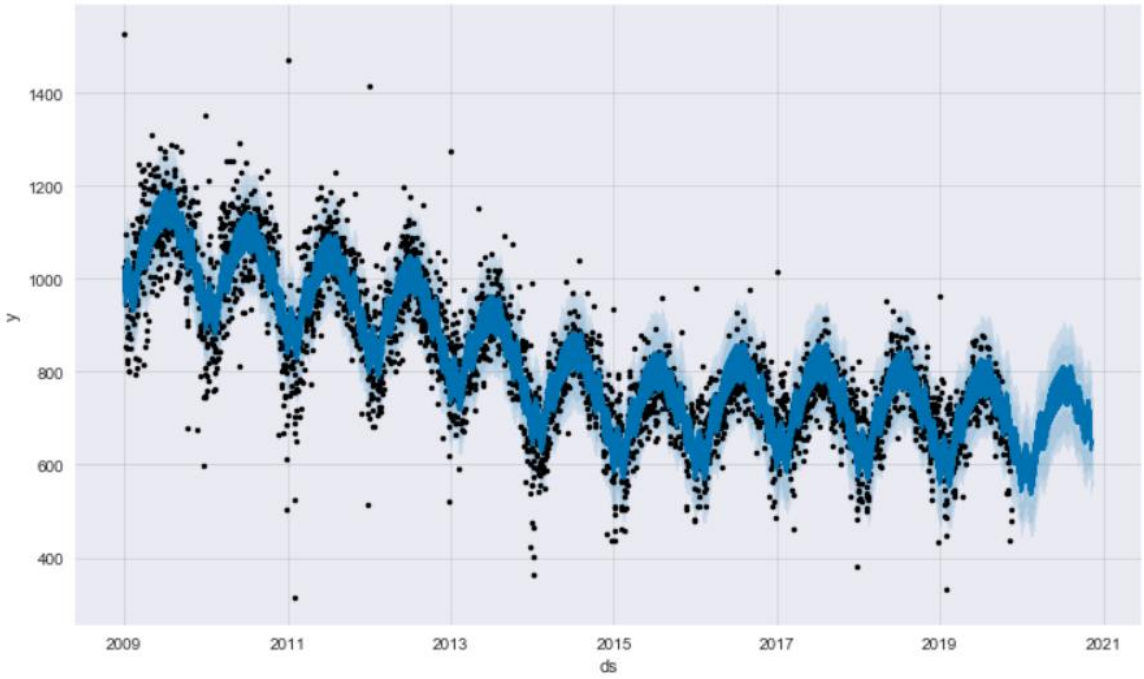
6.2.1 Time Series Model for Chicago

For furthering our goal of forecasting the Chicago crime reports for the year 2020, we applied the Chicago dataset to the Facebook's open source model, Prophet which can be installed and imported using fbprophet library in python notebooks.

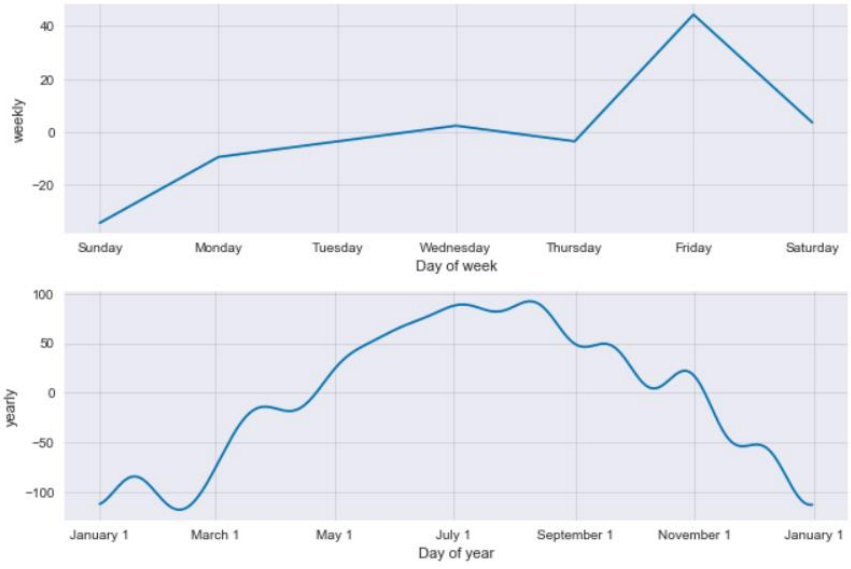
To fit the dataset to prophet model, we generated a data frame consisting of the number of incidents that occurred in a given day from the year 2009 to 2019. We plot a graph first with this data to understand the pattern of the all the crimes that happened in Chicago per day from 2009 to 2019. The graph is as below. The x-axis represents the crime that happened in a day, y-axis represents the dates from year 2009-2019.



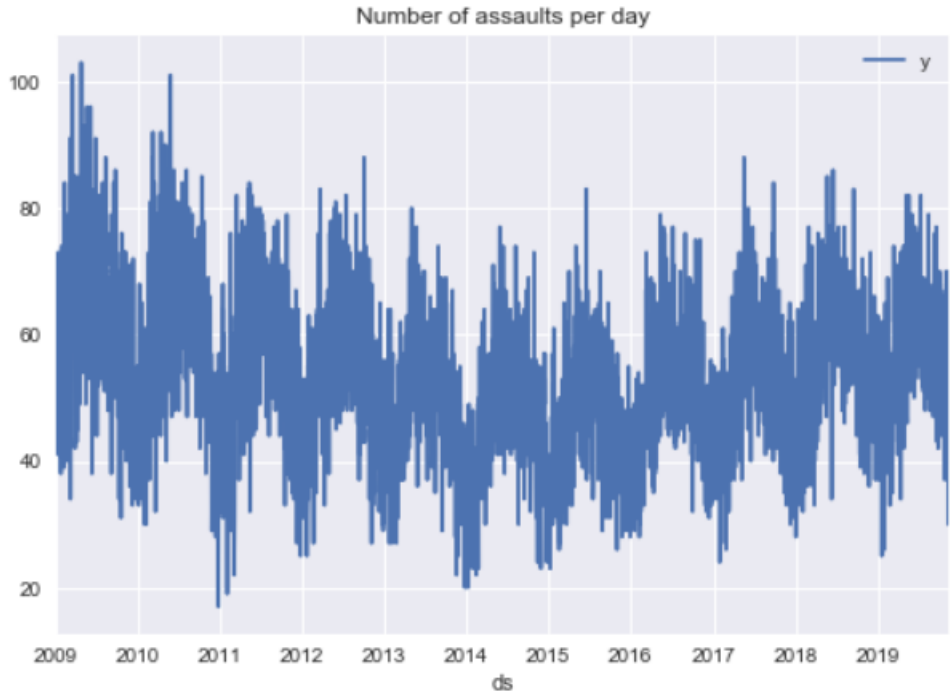
This table was then fit to the prophet model to obtain the below result and forecast for the year 2020. The table was specifically fit to generate forecast for 365 days. The dark blue lines are the predicted and general patterns identified by prophet. The dark blue line continues in the end showing predictions for the year 2020. The light blue lines above and below the dark blue line are the range of error/variation for the predicted values. The black dots are real incidents that do not fit in the model generated by the prophet and are called outliers.



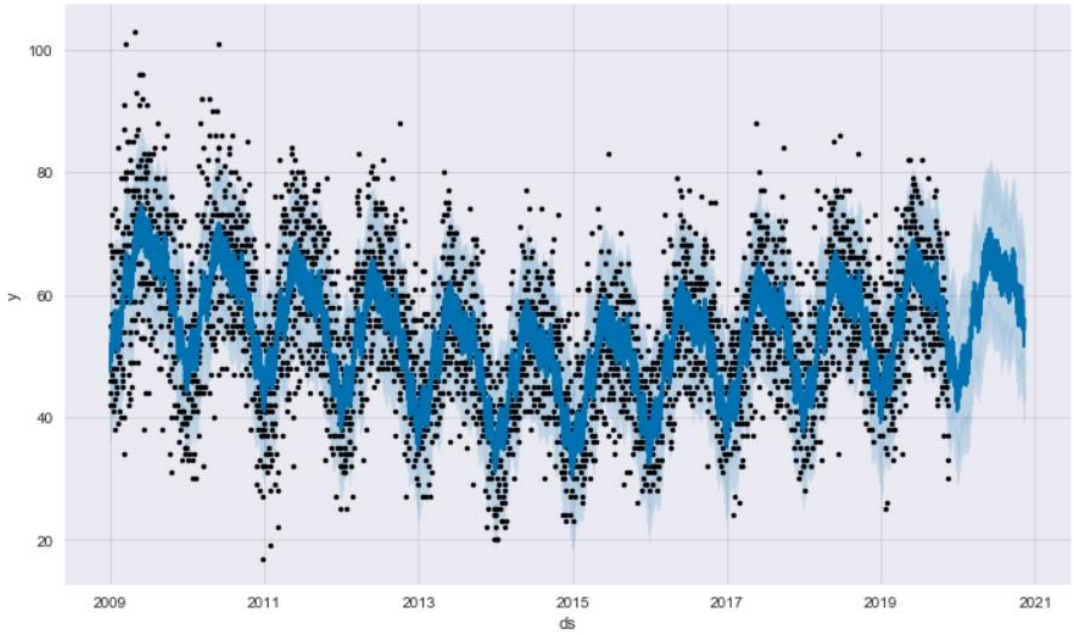
The prophet model also generates data for the trends in the weekly and monthly forecast for the predicted 365 days. This graph is as below.



We also plotted graph for frequently occurred violent crime type i.e. Assault. To fit this in the model, we, first, filtered rows with Primary type = 'Assault'. Below is the prediction for the assault throughout the years to understand the general pattern.



The forecasted result from fitting the above filtered data is as below.



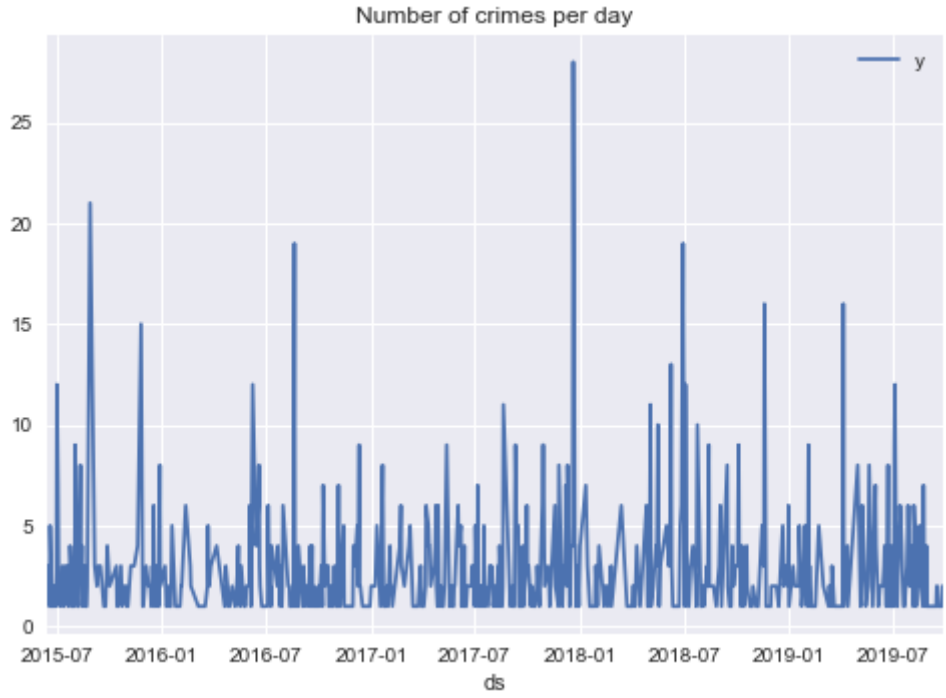
Similarly, we have generated Prophet forecasts for theft , narcotics, assault, prostitution, crim sexual assault, homicide, arson, gambling, kidnapping, stalking. These were the five violent and five non-violent crimes we identified for predicting forecasts.

6.2.2 Time Series Model for Boston

We implemented Time Series analysis and forecasting for the Boston crime data set in order to determine the trends in crimes over a period. We used the Prophet Python library for future forecasting. Using this library, we were able to carry out the forecasting for specific crimes in Boston for a period of 365 days.

Number of crimes per day

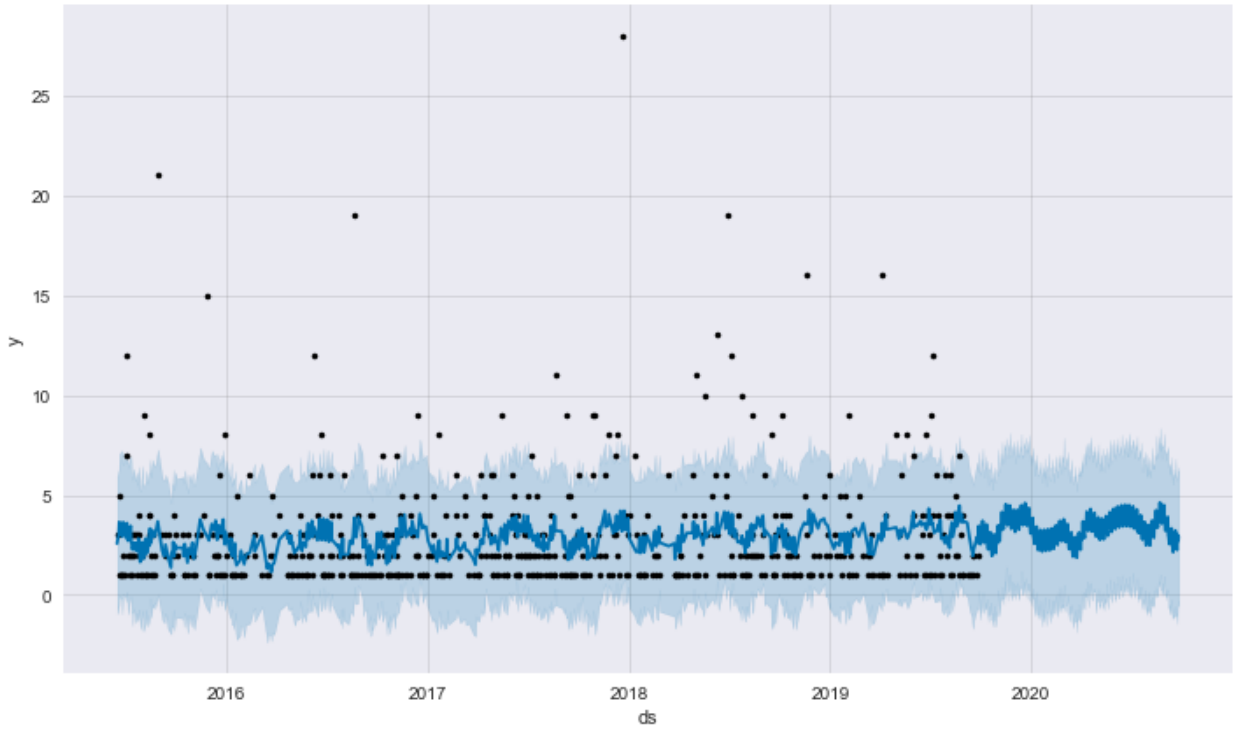
We grouped the crime data according to date and generated a plot to view the number of crimes per day in Boston after 2015:



As seen from the plot, there was a spike in the number of crimes on a few days but mostly the number of crimes ranged around 10-12. The most number of crimes were observed on a day during the later part of 2017.

Forecast of crimes for next 365 days

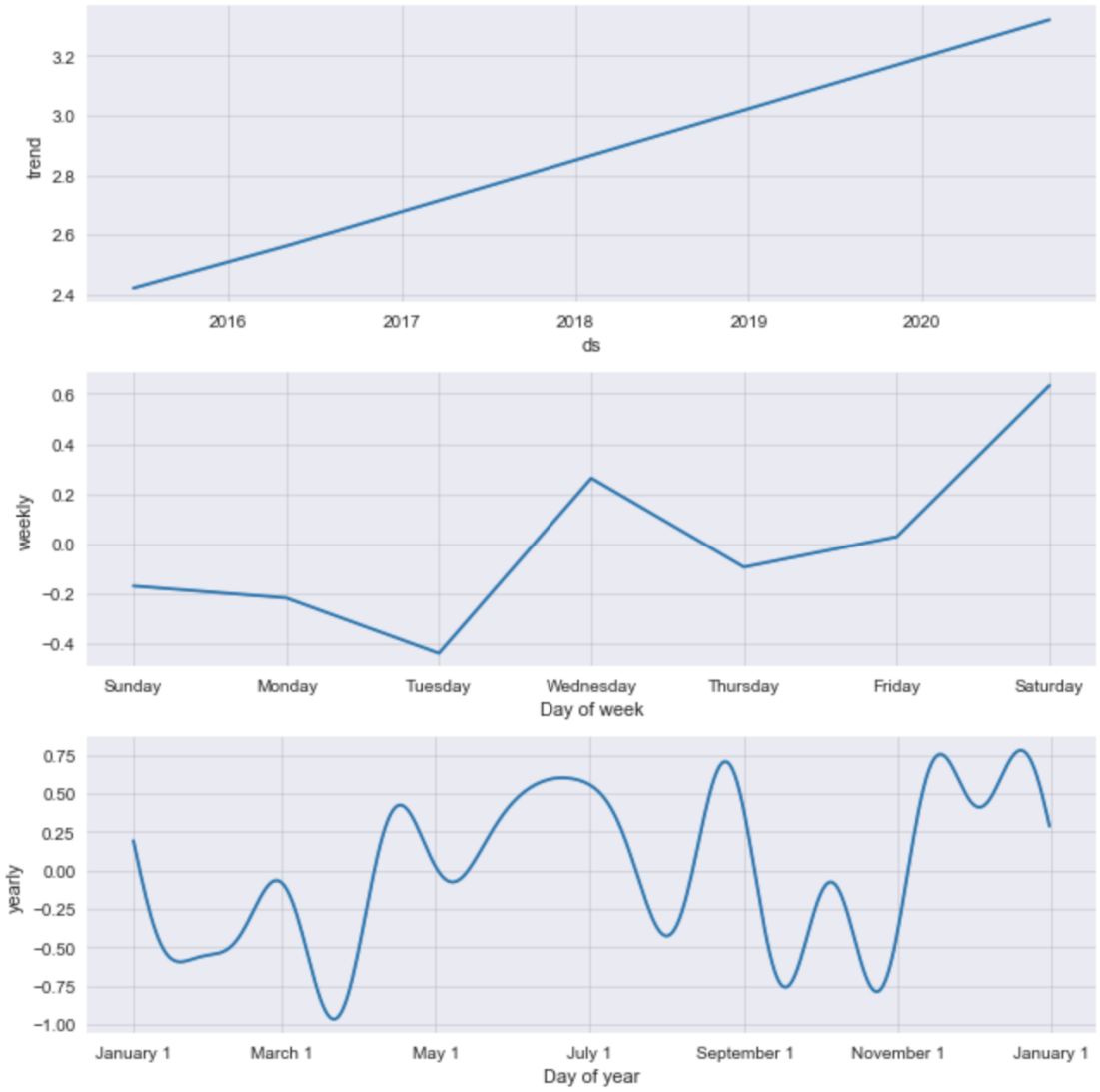
In order to forecast the crimes for a period, we executed model training by calling the fit function of Prophet and passing in the Pandas data frame. We did future prediction by executing the predict function and passing the parameter to describe how many days into the future to predict (365 days in our example).



Here the black points are actual data, the blue line is the prediction and the blue shaded area is the uncertainty in the prediction. As it can be seen from the plot, a prediction is made for the crimes for the year 2020 and we can observe it to be almost the same as the trends observed in the recent years.

Yearly, weekly and daily forecasts

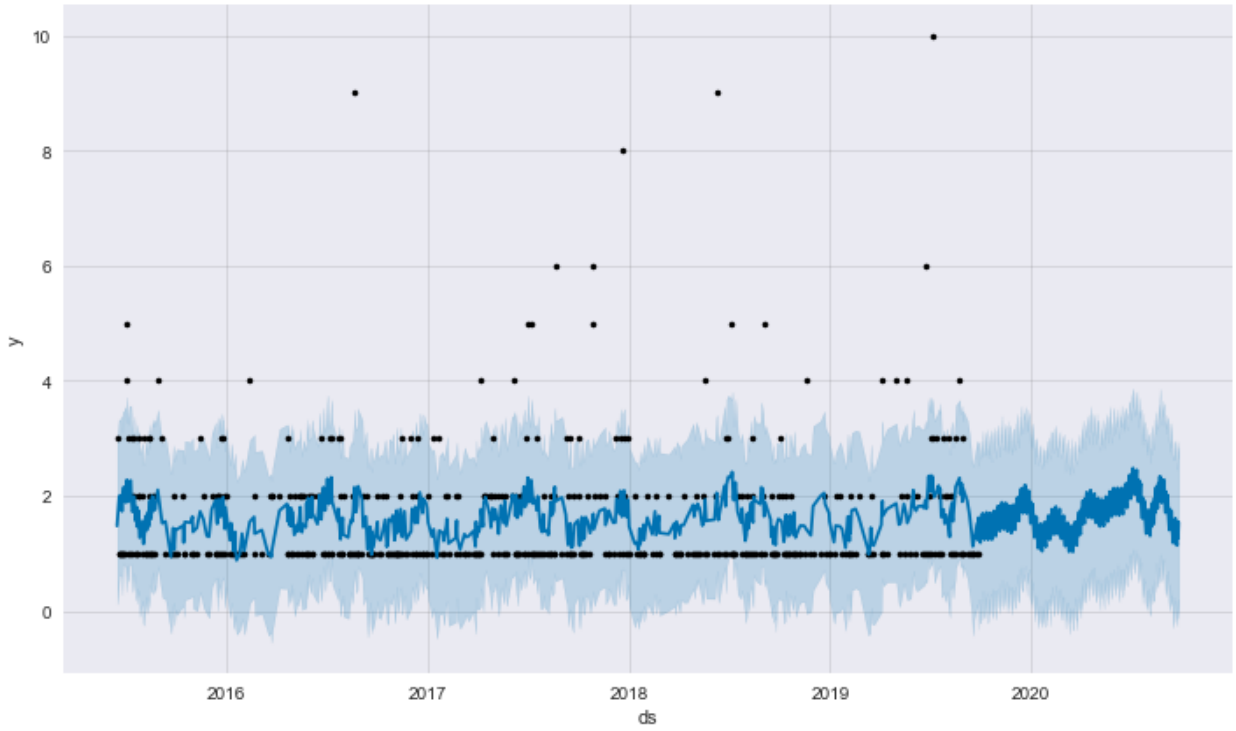
We broke the forecast down a bit further by calling the plot_components() method of the Prophet library to inspect the yearly, weekly and daily forecast components.



We observed from plotting these components that a linear trend was observed for yearly component, whereas Wednesday and Saturday were predicted to be days which high chances of crime. Also, from the daily component plot, it was observed that chances of crime were higher for the later part of the next year which could be accounted to harsh winters in Boston.

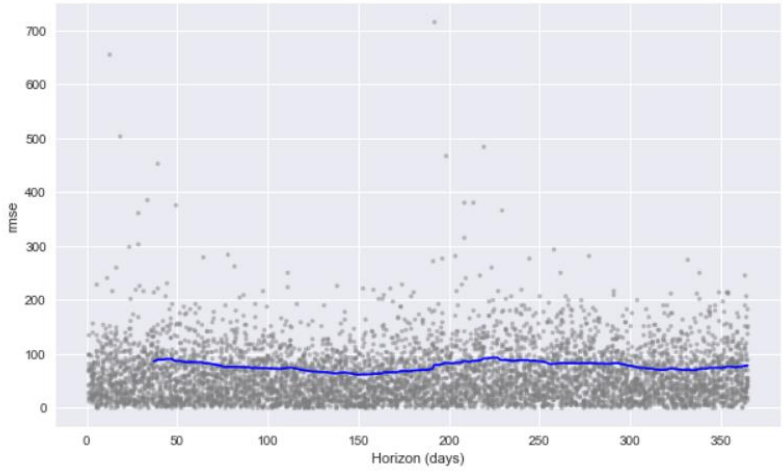
Forecast for most common crime

After carrying out an exploratory analysis for the Boston crime data set, we observed that "Aggravated Assault" was the most common crime that was recorded in Boston. We narrowed down the crime data set for the aggravated assault crimes to find out the number of crimes per day and get a forecast for these crimes for the next 365 days



Similarly, we forecasted the crime results for next year for the four next highest crimes observed in Boston namely Homicide, Firearms violation, Vandalism and drug violation. Using these results, we were able to identify yearly, weekly and daily trends in crimes in Boston.

The metrics is also provided by the fbprophet library to determine the reliability of the Prophet's results and we generated a root mean square error table for the predictions and plotted the metrics which can be seen below. The light blue line indicates that the generated results is highly reliable and is within the actual results space for the predicted 365 days.



6.3 Association Mining

After initial preprocessing Chicago dataset to apply Apriori algorithm to it, we found following interesting rules:

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(Location Description APARTMENT)	(no arrest)	0.132719	0.787342	0.113643	0.856268
1	(Location Description RESIDENCE)	(no arrest)	0.163834	0.787342	0.144772	0.883648
2	(Location Description STREET)	(no arrest)	0.222171	0.787342	0.164915	0.742291
3	(Primary Type ASSAULT)	(no arrest)	0.081629	0.787342	0.066977	0.820514
4	(Primary Type BATTERY)	(no arrest)	0.194940	0.787342	0.154862	0.794407
5	(Primary Type CRIMINAL DAMAGE)	(no arrest)	0.105125	0.787342	0.098588	0.937813
6	(Primary Type THEFT)	(no arrest)	0.237840	0.787342	0.214634	0.902432
7	(Time Period 13)	(no arrest)	0.076540	0.787342	0.062175	0.812322
8	(Time Period 29)	(no arrest)	0.081738	0.787342	0.061926	0.757614
9	(Time Period 33)	(no arrest)	0.087471	0.787342	0.067745	0.774488
10	(Time Period 37)	(no arrest)	0.092568	0.787342	0.073247	0.791274
11	(Time Period 41)	(no arrest)	0.093545	0.787342	0.072265	0.772522

12	(Time Period 45)	(no arrest)	0.087534	0.787342	0.065737	0.750986
13	(Time Period 49)	(no arrest)	0.106120	0.787342	0.086308	0.813305
14	(Time Period 5)	(no arrest)	0.078916	0.787342	0.061898	0.784353
15	(Time Period 9)	(no arrest)	0.078299	0.787342	0.062925	0.803655

Rules generated for year 2019 confidence > 0.6

As we observe the rules for year 2019, provide various primary types of crimes like theft, battery, assault which have high confidence and support of no arrest being made. There are also crimes occurring in various time periods like 5, 9, 41, 45 for which there is high confidence and support of no arrest being made. There are crimes occurring in various location descriptions like street, residence and apartments where there is high confidence of no arrest being made.

antecedents	**consequents**	**antecedent support**	**consequent support**	**support**	**confidence**	**lift**
0	(Location Description APARTMENT)	(no arrest)	0.132045	0.738316	0.109086	0.826129
1	(Location Description RESIDENCE)	(no arrest)	0.153595	0.738316	0.132833	0.864823
2	(Location Description STREET)	(no arrest)	0.234527	0.738316	0.169942	0.724617
3	(Primary Type BATTERY)	(no arrest)	0.189444	0.738316	0.144601	0.763293
4	(Primary Type CRIMINAL DAMAGE)	(no arrest)	0.110929	0.738316	0.103817	0.935884
5	(Primary Type NARCOTICS)	(arrest)	0.083838	0.261684	0.083807	0.999630
6	(Primary Type THEFT)	(no arrest)	0.219984	0.738316	0.193747	0.880730

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
7	(Time Period 33)	(no arrest)	0.089798	0.738316	0.066913	0.745150
8	(Time Period 37)	(no arrest)	0.098862	0.738316	0.074689	0.755485
9	(Time Period 41)	(no arrest)	0.096499	0.738316	0.068966	0.714676
10	(Time Period 45)	(no arrest)	0.088626	0.738316	0.061042	0.688761
11	(Time Period 49)	(no arrest)	0.102447	0.738316	0.078165	0.762981

Rules generated for year 2015 confidence > 0.6

As we observe the rules for year 2015, provide various primary types of crimes like theft, battery, narcotics and criminal damage which have high confidence and support of no arrest being made. There are also crimes occurring in various time periods like 33, 37, 41, 45 for which there is high confidence and support of no arrest being made. There are crimes occurring in various location descriptions like street, residence and apartments where there is high confidence of no arrest being made.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(Location Description APARTMENT)	(no arrest)	0.125680	0.715213	0.102727	0.817367
1	(no arrest)	(Location Description APARTMENT)	0.715213	0.125680	0.102727	0.143631
2	(Location Description RESIDENCE)	(no arrest)	0.164833	0.715213	0.141653	0.859373

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
3	(no arrest)	(Location Description RESIDENCE)	0.715213	0.164833	0.141653	0.198057
4	(Location Description SIDEWALK)	(arrest)	0.110180	0.284787	0.060667	0.550614
5	(arrest)	(Location Description SIDEWALK)	0.284787	0.110180	0.060667	0.213025
6	(Location Description STREET)	(no arrest)	0.247687	0.715213	0.179500	0.724706
7	(no arrest)	(Location Description STREET)	0.715213	0.247687	0.179500	0.250974
8	(no arrest)	(Primary Type BATTERY)	0.715213	0.179067	0.135507	0.189463
9	(Primary Type BATTERY)	(no arrest)	0.179067	0.715213	0.135507	0.756739
10	(Primary Type BURGLARY)	(no arrest)	0.070647	0.715213	0.067560	0.956308
11	(no arrest)	(Primary Type BURGLARY)	0.715213	0.070647	0.067560	0.094461
12	(no arrest)	(Primary Type CRIMINAL DAMAGE)	0.715213	0.108093	0.100007	0.139828
13	(Primary Type CRIMINAL DAMAGE)	(no arrest)	0.108093	0.715213	0.100007	0.925188
14	(Primary Type NARCOTICS)	(arrest)	0.118753	0.284787	0.118620	0.998877

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
15	(arrest)	(Primary Type NARCOTICS)	0.284787	0.118753	0.118620	0.416522
16	(Primary Type THEFT)	(no arrest)	0.197373	0.715213	0.174560	0.884415
17	(no arrest)	(Primary Type THEFT)	0.715213	0.197373	0.174560	0.244067
18	(Time Period 33)	(no arrest)	0.093807	0.715213	0.067093	0.715230
19	(no arrest)	(Time Period 33)	0.715213	0.093807	0.067093	0.093809
20	(no arrest)	(Time Period 37)	0.715213	0.097593	0.070773	0.098954
21	(Time Period 37)	(no arrest)	0.097593	0.715213	0.070773	0.725186
22	(no arrest)	(Time Period 49)	0.715213	0.107227	0.079060	0.110540
23	(Time Period 49)	(no arrest)	0.107227	0.715213	0.079060	0.737317

Rules generated for year 2011 lift > 1.0

Rules which have lift higher than 1 are likely to appear together. So, as we observe the rules for year 2011, there are various primary types of crimes like theft, battery, narcotics, burglary, and criminal damage and no arrest are very likely. Also, crimes occurring in various time periods like 33, 37, 49 and no arrest being made is also very likely. There are crimes occurring in various location descriptions like street, sidewalk, residence and apartments and no arrest is very likely.

7. DISCUSSION

The results generated using the classification and time series model are what we wished to obtain from the overall project. The results and its metrics back our proposal for improving the quality of safety by empowering the authorities with patterns in the crime for both Chicago and Boston. The qualities that were observed in Chicago and Boston is very similar since both the cities have similar characteristics in terms of population and demography as well as climatic conditions. General crime patterns determine that the crime rate increases in both the cities during the summer season and reached peak during June/July months. The daily trend also sheds light on the similarities of the datasets. Weekends and Fridays depict high crime rates. The classification helped in categorizing the values using powerful supervised learning algorithms like decision trees and random forest. With different variations in feature set we were able to eliminate overfitting and generate pure classification model with accuracy rate of ~92%.

Even with all this data analysis, we were still a step behind since there was no effective strategy plan for the authorities in the lower regions. The crime patterns all depicted in the city level and not in a smaller region. To overcome this problem, we wrote a method to generate forecasts for every crime type in every police district that has been identified by the Chicago and Boston authorities. The code snippet is as below.

```
def predictCrimeTrend(policeDistrict, crimeType):
    newPredict = chicago[chicago['Year'] >= 2009]
    newPredict = newPredict[newPredict['Primary Type'] == crimeType]
    newPredict = newPredict[newPredict['District'] == policeDistrict]
    newPredict['Date Only'] = newPredict['Date'].dt.date
    ts_newPredict = newPredict.groupby('Date Only').count()['Primary Type'].to_frame()
    ts_newPredict.reset_index(inplace=True)
    ts_newPredict.columns = ['ds','y']
    print("Crime trends for {0} in police district {1}".format(crimeType,policeDistrict))
    ts_newPredict.plot(x='ds', title='Current crime trend for given inputs')
    plt.show()
    prophet_newPredict = ts_newPredict.copy()
    prophet_model = Prophet()
    prophet_model.fit(prophet_newPredict)

    # Let's try a forecast for 365 days
    future = prophet_model.make_future_dataframe(periods=365)
    forecast_df = prophet_model.predict(future)
    ts_newPredict = prophet_model.plot(forecast_df)
    ts_newPredict.show()
    PC = prophet_model.plot_components(forecast_df)
```

This code helps generate forecast for every crime type i.e. 34 crime type for every police district i.e. 28 police districts. Therefore, we can obtain 952 different forecast graphs using the above code. This completes our goal of helping the authorities be abreast with a plan of action by understanding trends and creating plan of action to minimize the crime and keeping the citizens aware and protected. The probability of success of controlling crime when concentrated on police district level is immensely higher than trying to control the crime from a whole city point of view. Thus, we believe with this part of the analysis and prediction, we have completed our problem statement.

8. FUTURE WORK

As a part of the future work, we believe the time series forecast can be added to every beat inside a police district so that the preventive measures can be applied at the lowest level of the authorities to improve and increase the chances of reducing and handling the control of crime.

We can also perform data analysis and reveal more trends and patterns by combining a dataset of household income per community area to understand how the crime affects the richer communities as compared to the poorer communities.

We can also combine the crime dataset with the education levels and job opportunities dataset to understand the mindset of the criminals and perpetrators.

9. CONCLUSIONS

In this project, we have used various data mining techniques like classification, association mining, time series prophet model to analyze the Chicago and Boston crime dataset. For classification, we have used Random Forest algorithm and the decision tree model. For association mining, we have used the Apriori Algorithm.

For the Chicago dataset, the accuracy obtained from Decision Tree classification is ~68%. Through various experiments and application of combinations of data features, we found that Random Forest applied for feature set without IUCR feature gives the best prediction results with accuracy ~93%. Although Random Forest eliminates the issue of overfitting by consulting multiple Decision Trees to classify a given data, removing IUCR which is dependent on other attributes ensures that the algorithm is not subjected to inadvertent overfitting. Using association mining, we found many rules which were evaluated based on high support, confidence and lift greater than 1. These rules can help the police to find areas more prone to crimes where arrests are uncommon, times when crimes are more frequent and types of frequent crimes which do not have arrests made.

For the Boston dataset, we found that Random Forest applied for feature set without Offense code feature gives the best prediction results with accuracy ~95%.

The metrics for both the classification and prophet model have revealed that the results obtained for the final models are almost ideal and have the least error margin compared to other initial models we used. In order to satisfy our goal and achieve a solution for our problem statement of providing an analysis and predicting the crime trends for future to help the authorities plan preventive measures, we extended the time series prediction to forecast the trend for year 2020 for every crime type in every police district. Hence, we successfully completed our goal to bring about analysis and prediction to help authorities take preventive measures.

10. REFERENCES

[1] Louppe, Gilles. (2014). Understanding Random Forests: From Theory to Practice. 10.13140/2.1.1570.5928.

[2] Taylor SJ, Letham B. 2017. Forecasting at scale. PeerJ Preprints 5:e3190v2

[3] Feng, Mingchen & Zheng, Jiangbin & Ren, Jinchang & Hussain, Amir & Li, Xiuxiu & Xi, Yue & Liu, Qiaoyuan. (2019). Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2930410.

[4] M. Feng et al., "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," in IEEE Access, vol. 7, pp. 106111-106123, 2019. doi: 10.1109/ACCESS.2019.2930410

[5] Pradhan, Isha, "Exploratory Data Analysis And Crime Prediction In San Francisco" (2018). Master's Projects. 642.

[6] Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. 2013. Detecting patterns of crime with series finder. In Proceedings of the 17th AAAI Conference on Late-Breaking Developments in the Field of Artificial Intelligence (AAAIWS'13-17). AAAI Press 140-142.

[7] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar. Introduction to Data Mining (Second Edition)

[8] Craig Silverstein, Sergey Brin, Rajeev Motwani. Beyond Market Baskets: Generalizing Association Rules to Dependence Rules.1998 Kluwer Academic Publishers, Boston

[9] Lior Rokach and Oded Maimon. 2014. Data Mining with Decision Trees: Theory and Applications (2nd ed.). World Scientific Publishing Co., Inc., River Edge, NJ, USA.

[10] S. Sathyadevan, D. M. S and S. G. S., "Crime analysis and prediction using data mining," 2014 First International Conference on Networks & Soft Computing (ICNSC2014), Guntur, 2014, pp. 406-412.