Time Series Analysis and Trends Forecasting of Crime Data

Project Final Report
ITCS 5156 - APPLIED MACHINE LEARNING

Submitted by

Rama Sri Saladi
(801254656)

Submitted to

Minwoo "Jake" Lee

Github link: https://github.com/slramasree/Applied_ML_project

Paper Information:

Title: Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data.
Author: Mingchen Feng, Jinchang Ren
Year: July 2019
Conference/Journal:
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.

Link:
https://www.researchgate.net/publication/334617796_Big_Data_Analytics_and_Mining_for_Effective_Visualization_and_Trends_Forecasting_of_Crime_Data

Introduction:

Crime consists of intentional acts that can cause harm, physical and psychological damage to people and property. The rate of criminal activities is steadily increasing, prompting authorities to develop new and efficient ways of controlling measures. Producing a model or a machine learning algorithm that can help the crime investigation department in predicting these crimes will be the best solution to prevent crimes; however, the development is not without its challenges. It is difficult to procure and manage such a large amount of data. Developing new techniques and methods to analyze this complex data from multiple sources will help technicians track past events or crimes, recognize similarities from those events and take immediate action or decision which further helps in understanding both current and historical incidents. Improving crime prevention models will lead to a secure living environment, improved quality of life and economic growth. The recent surge in cloud computing and storage technologies has led to an increased opportunity to collect data and make it public. It is important to extract meaningful insights from this big data to understand the complex patterns. Big data analytics is emerging as an efficient approach that can address such complex, unstructured, and vast data.

Data Mining is one of the fundamental techniques and a growing research field of big data analytics which can build patterns across various fields for exploring useful information from the data. Data mining is not only used to discover new patterns from the data but also to gain understanding of the existing ones. With the help of such techniques, big data can help in identifying the different patterns in crimes that occur in specific areas and how they are related. The lack of conventional standards to record crimes across the globe is the reason engineers or data scientists spend 70 to 80 percent of their time in data mining process which is focused on collecting, storing, and pre-processing the data. Crime forecasting depends on the evidence such as time, location, type, the weapon used, etc. Considering the data that is available, it is not easy to extract the necessary fields as the data does not have an appropriate structure. To query and retrieve the required fields, the data should be in a structured format. This process of extracting and cleaning the data to suit a particular storage format is necessary to obtain good predictions. In addition, the selection of attributes is an integral part of this process; dropping the attributes that are not necessary for analysis helps in increasing performance and reduces the risk of overfitting. To predict crimes based on past crime data, time series forecasting is the best suitable method. It can make predictions based on the historical time-stamped data. Time series forecasting is being used in different sectors including weather forecasting, stock prediction, healthcare, environmental forecasting, etc. There are two types of time series forecasting: Univariate, Multivariate. Univariate is based on only one variable varying over time, whereas multivariate relies on multiple variables varying over time. Crime data belongs to the multivariate data category. The results of this study revealed that some of the time series analysis models that suites this data are Long short-term memory (LSTM), Prophet model, and Ensemble of classifiers. Details of the results obtained using these models are explained in detail in further sections.

Problem Statement:

Crime prevention has always been a top priority for governments to ensure a safe living environment for their citizens. Accurately forecasting crimes will help in reducing the crime rate. As this is an active research area, many researchers applied several machine learning, deep learning, and time series forecasting algorithms on real-world crime data sets from major cities like Chicago, San Francisco, Brisbane, etc. To date, Prophet, LSTM, and Ensemble of classifiers are some of the most successful models. Prediction results from these studies illustrate varying degrees of accuracy. This study focused on leveraging the effective strategies of big data analytics in present crime prevention research as well as applying improvements to the current methods.

Motivation & Challenges:

Till date all the papers were focused on predicting crimes on a yearly, monthly, daily basis, but my idea is to predict crimes based on time of the day (i.e., Morning, Afternoon, evening, night). This prediction helps in distribution of troops based on the crime type and severity. For this purpose, new columns were needed like Day of the week, Day of month, Day of year. As the selected dataset contains the data based on city level there is an Insufficiency of Data based on district and street level so, I have extracted the location based on latitudes and longitudes using the Geopandas library.

Open questions in the domain:

1. What is the best way to handle the missing values in data before predictive modelling?
2. Does resampling of data will add weight to features.
3. Does Time series analysis accept calibrating parameter to control the usage of past data?

Brief overview of the approach to address the challenges:

- Initial Data Analysis
- Data pre-processing
- Data manipulation
- Predictive Modelling
- Performance Analysis

<u>Backgrounds:</u>

Summary of other related researches:

Paper 1:

- Title: A Comparative Study on Crime in Denver City Based on Machine Learning and Data Mining.
- Author: Md. Aminur Rab Ratul
- Year: January 2020
- Conference/Journal: Researchgate
- Why:
  - Md. Aminur Rab Ratul analyzed the Denver County crime dataset and applied various classification algorithms like Random Forest, Decision Tree, AdaBoost classifier, Extra tree classifier, K-Neighbors classifier, 4 ensemble models to classify 15 different classes of crimes and concluded that Ensemble models produce high accuracy when compared to other classification models.
- Link: https://www.researchgate.net/publication/338500247_A_Comparative_Study_on_Crime_in_Denver_City_Based_on_Machine_Learning_and_Data_Mining

Paper 2:

- Title: Modeling Daily Crime Events Prediction Using Seq2Seq Architecture
- Author: Mingchen Feng, Jinchang Ren
- Year: July 2019
- Conference/Journal: Researchgate
- Why:
  - Jawaher Alghamdi built ARIMA, RNN, Conv1D, (Seq2Seq) based LSTM models to predict crimes week ahead of occurrence. By comparing the results of these models, we concluded that Seq2Seq model is highly effective.
- Link: https://www.researchgate.net/publication/349191312_Modeling_Daily_Crime_Events_Prediction_Using_Seq2Seq_Architecture

Pros and Cons

- Data transformations are used to pick key attributes which eliminated noisy data, but it also produces anomalous outputs in decision trees.
- Although the accuracy in every evaluation method is above 90%, the time complexity of the above machine learning models is high.
- Accuracy in short-term forecasting (over a week) is high for seq2seq based LSTM model.

How the other work is related to the main method:

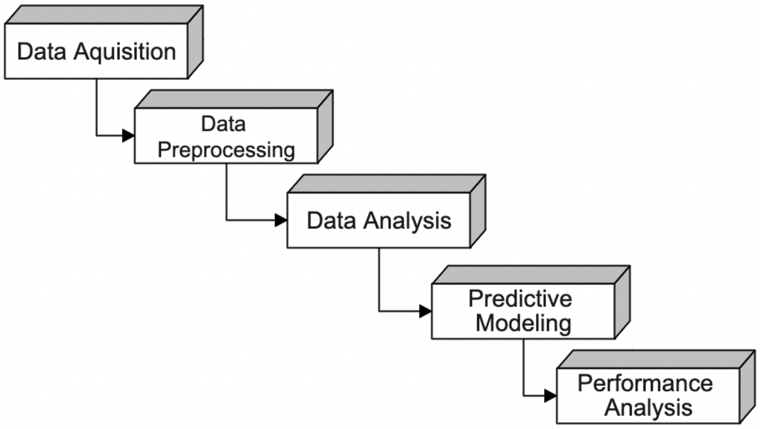
Big data analytics has been extensively studied and applied in crime data analysis. Machine learning and big data are equally useful in gaining incites from large volumes of data, although they do present some challenges. Mingchen Feng [1] described the promise and potential of BDA in crime data and compared the time series models to Neural networks and concluded that time series models produced better prediction than neural networks. Jawaher Alghamdi [2] built ARIMA, RNN, Conv1D, (Seq2Seq) based LSTM models to predict crimes week ahead of occurrence. By comparing the results of these models, I concluded that Seq2Seq model is highly effective. Md. Aminur Rab Ratul [3] analyzed the Denver County crime dataset and applied various classification algorithms like Random Forest, Decision Tree, AdaBoost classifier, Extra tree classifier, K-Neighbors classifier, 4 ensemble models to classify 15 different classes of crimes and concluded that Ensemble models produce high accuracy when compared to other classification models.

Methods:

Details of the algorithms and methods:

Different methods were used at different steps of the entire process for effective prediction results. The following is a detailed overview of methods/implementations followed at each step.

Framework:



Data Acquisition:

The datasets are taken from the Kaggle and the links for the datasets and the details are provided below.

Los Angeles Dataset

In Los Angeles dataset there are 21 columns. The columns are *DATE OCC, TIME OCC, AREA, AREA NAME, Rpt Dist No, Crm Cd, Crm Cd Desc, Mocodes, Vict Age, Vict Sex, Vict Descent, Premis Cd, Premis Desc, Weapon Used Cd, Weapon Desc, Status Desc, Crm Cd 1, Crm Cd 2,* *LOCATION, LAT,* and *LON*. While most of the column names and description are obvious, some are not; for example, *Rpt Dist No* (4-digit sub area code), *Crm Cd* (code of crime committed), *Mocodes* (Modus Operandi), *Vict Descent* (race of the victim), *Premis Cd* (code that describes the type of structure), and *Crm cd 1* (primary crime codes). I have used two datasets, the first one is the data from the year 2020 to 2021 which gives insights about the latest trends and the second one is from 2012 to 2016 which can be used for training.

Link: *https://www.kaggle.com/cityofLA/crime-in-los-angeles*

B. Chicago Dataset

In Chicago dataset there are 21 columns. The columns are *ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordinate, Year, Updated On, Latitude, Longitude, Location, Neighborhood, Municipality* and *County*. While most of the column names and description are obvious, some are not, including IUCR (Illinois Uniform Crime Reporting Code), Primary type (description of IUCR code), description (secondary description of IUCR code), Domestic (if the crime comes under domestic violence or not), and Beat (the smallest police geographic area). The dataset consists of data from the year 2001 to 2021.

Link: *https://www.kaggle.com/currie32/crimes-in-chicago*

Data Analysis and Visualization:

Inspecting and experimenting with the statistical details of data before beginning the classification helps in revealing some interesting facts about the data like the hourly, monthly, yearly trend in crimes. This section describes the methodologies used to extract detailed periodic insights on Chicago and Los Angeles datasets.

A. Los Angeles Dataset Analysis

In the initial phase of this study, the columns with most null values were checked and removed, then the top 10 crimes were analyzed. The results showed that the vehicle stolen is the most prevalent crime type.

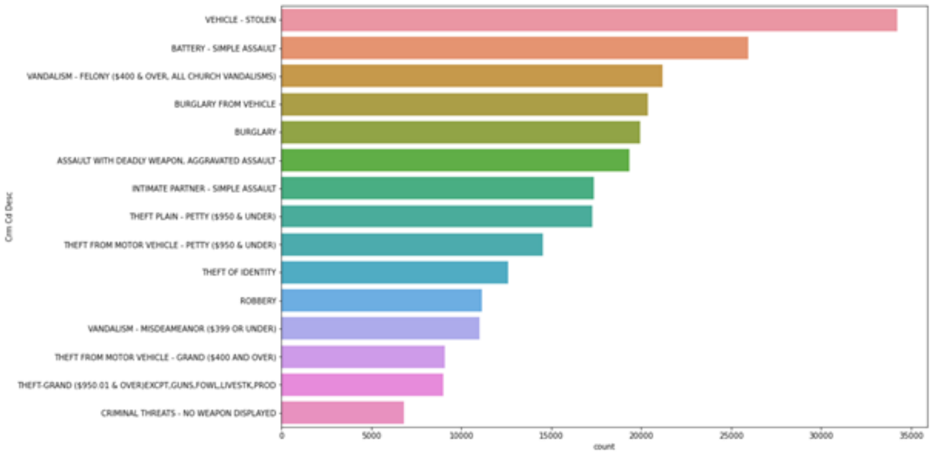


Fig 1: Top 10 crime types in Los Angeles

Fig 2 shows the top ten crimes location types in Los Angeles by filtering the location type where the crime has occurred, illustrating that the highest crime rates were in the streets (more than 80000), with single family dwelling as second highest.

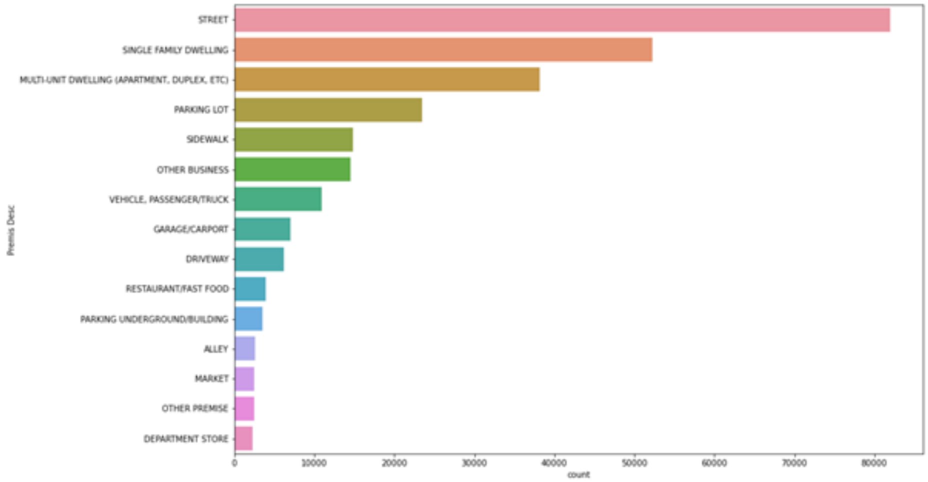


Fig 2: Top 10 crime location types in Los Angeles

To analyze crime count on a quarterly basis, the *date occ* column was string formatted to date and time. Segregating the number of crimes quarterly revealed that the number of crimes remained almost consistent, with a steep decline at the end in Fig 3 due to incomplete final quarter data.

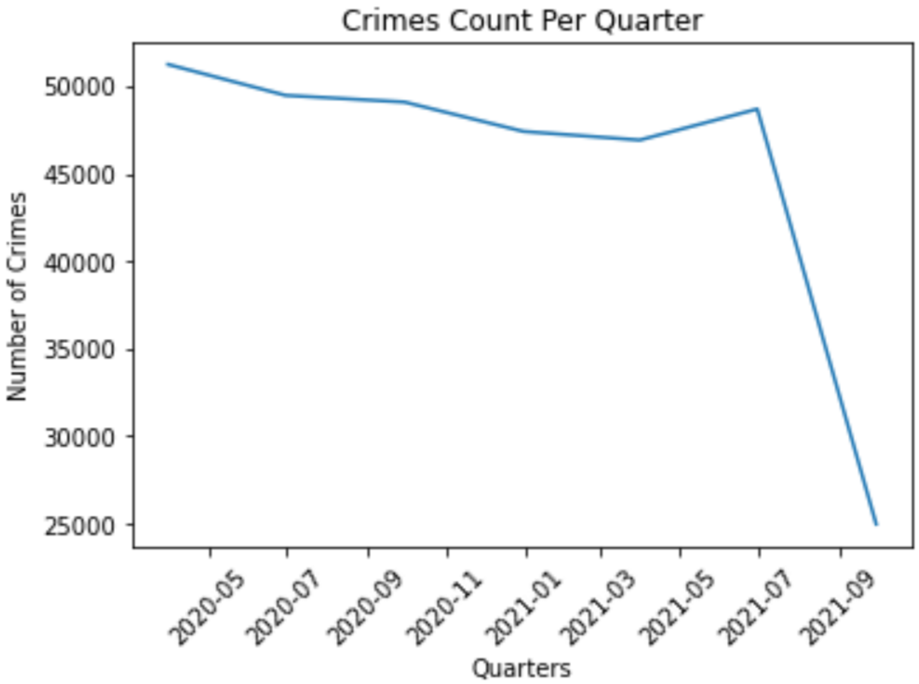


Fig 3: Number of crimes per Quarter in Los Angeles

To visualize the yearly trend for crimes in Los Angeles, another dataset was used from 2012 to 2016. Analysis showed an increase in the number of crimes from 2012 to 2014 and a decrease from 2014 to 2016.

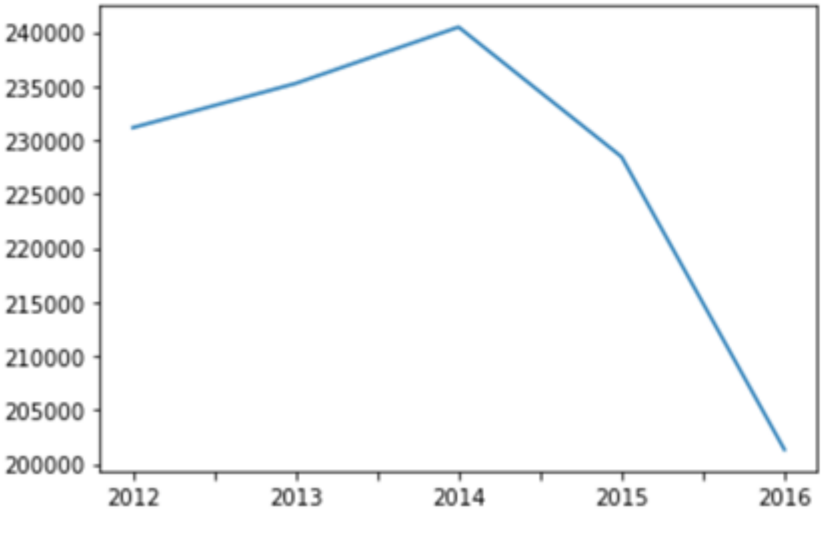


Fig 4: Number of crimes per year in Los Angeles

To visualize the top 10 crimes every year, I separated the crime data for each year and filtered the highest number of crimes in that year and discovered that Burglary, battery theft, vehicle stolen, and traffic DR remained consistently in the top four.

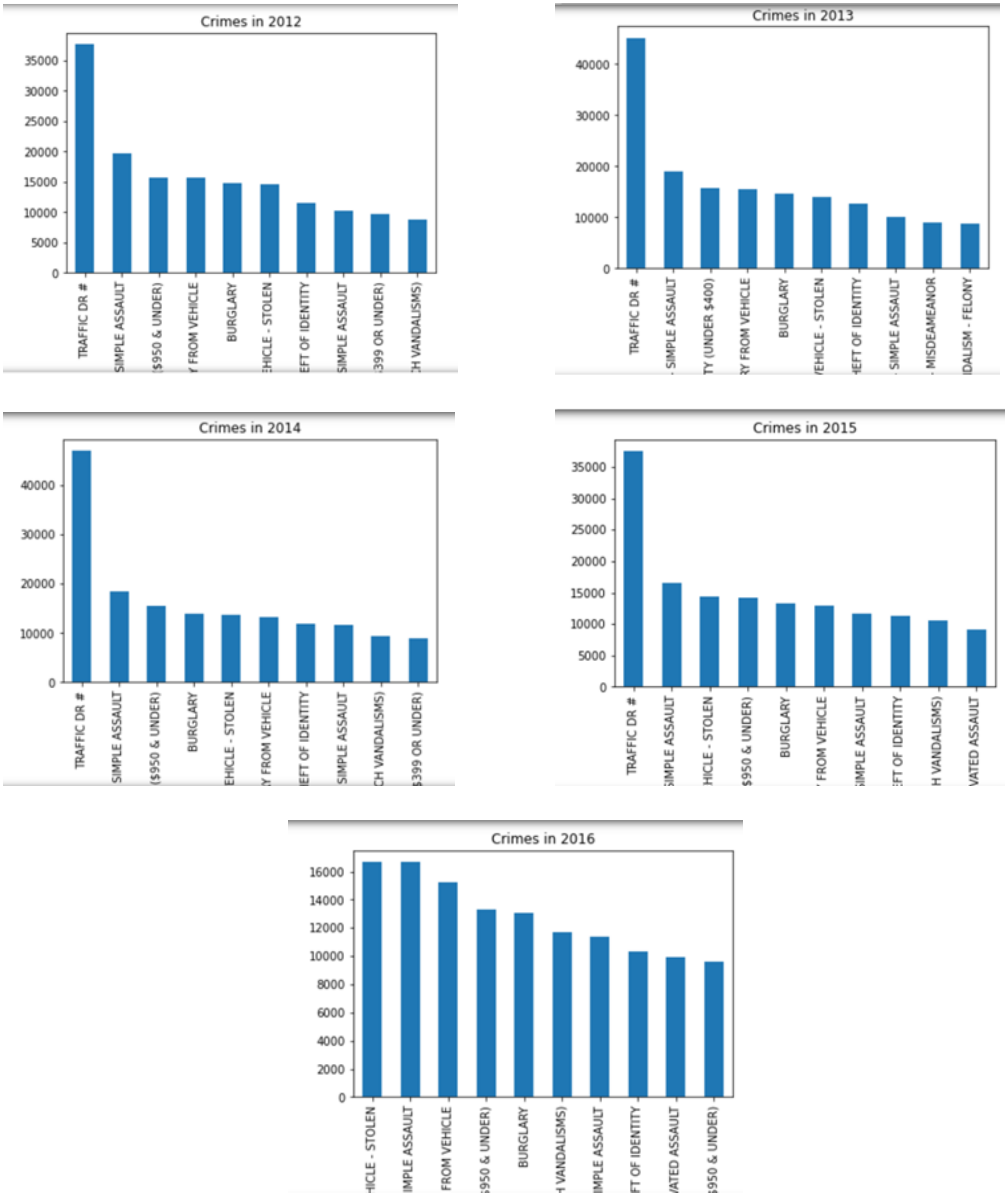
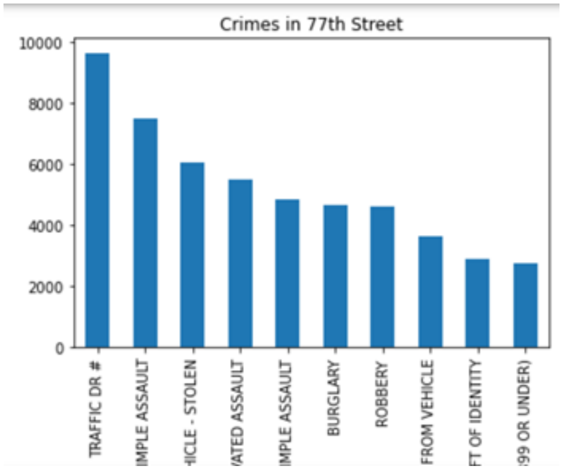
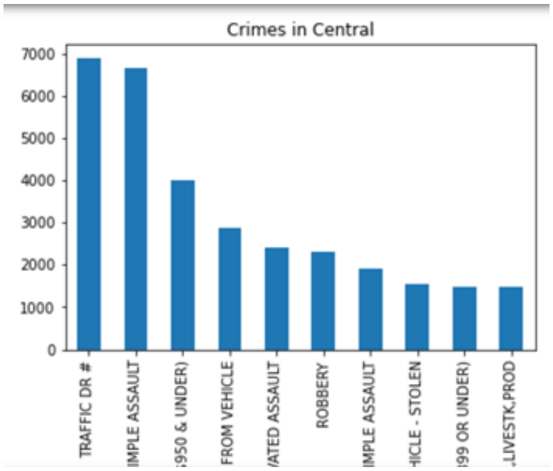
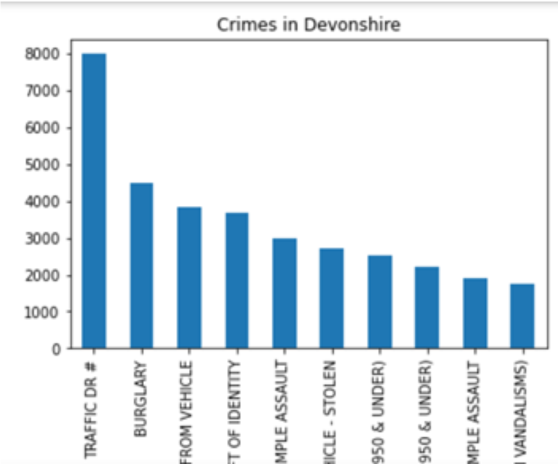


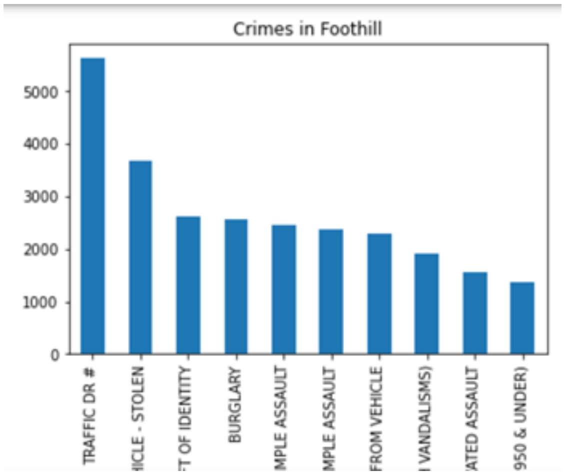
Figure 5: Yearly top 10 crime types in Los Angeles

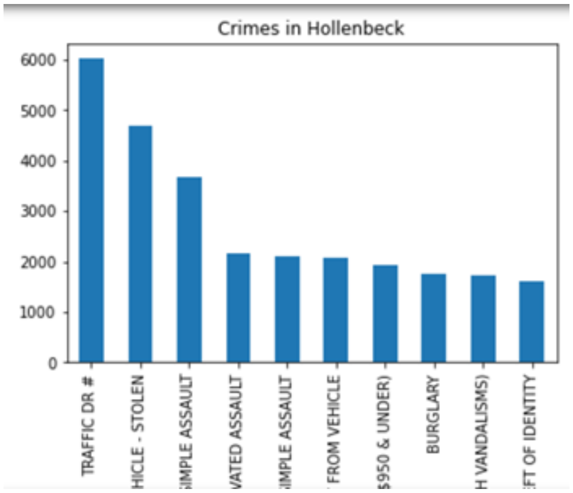
Street level crime details were determined by filtering the top 10 streets with the highest crime rates and then filtering the top 10 crimes in each street. Areas such as the Central city, Devonshire, Foothill, Hollenbeck, Hollywood, and Pacific which are tourist places had the most prevalent crime is theft, while poorer areas such as Rampart and Southeast showed spousal abuse as the most prevalent.

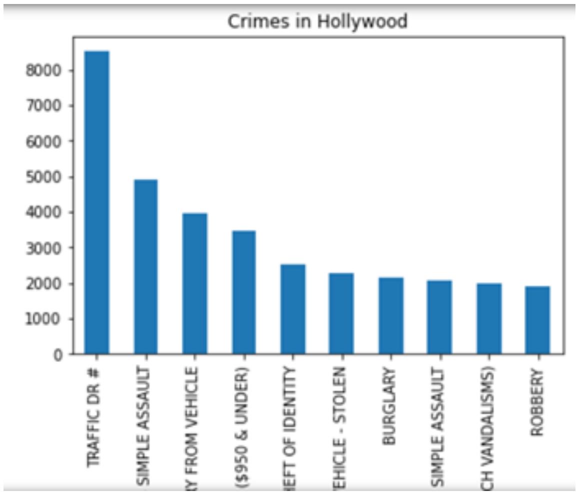












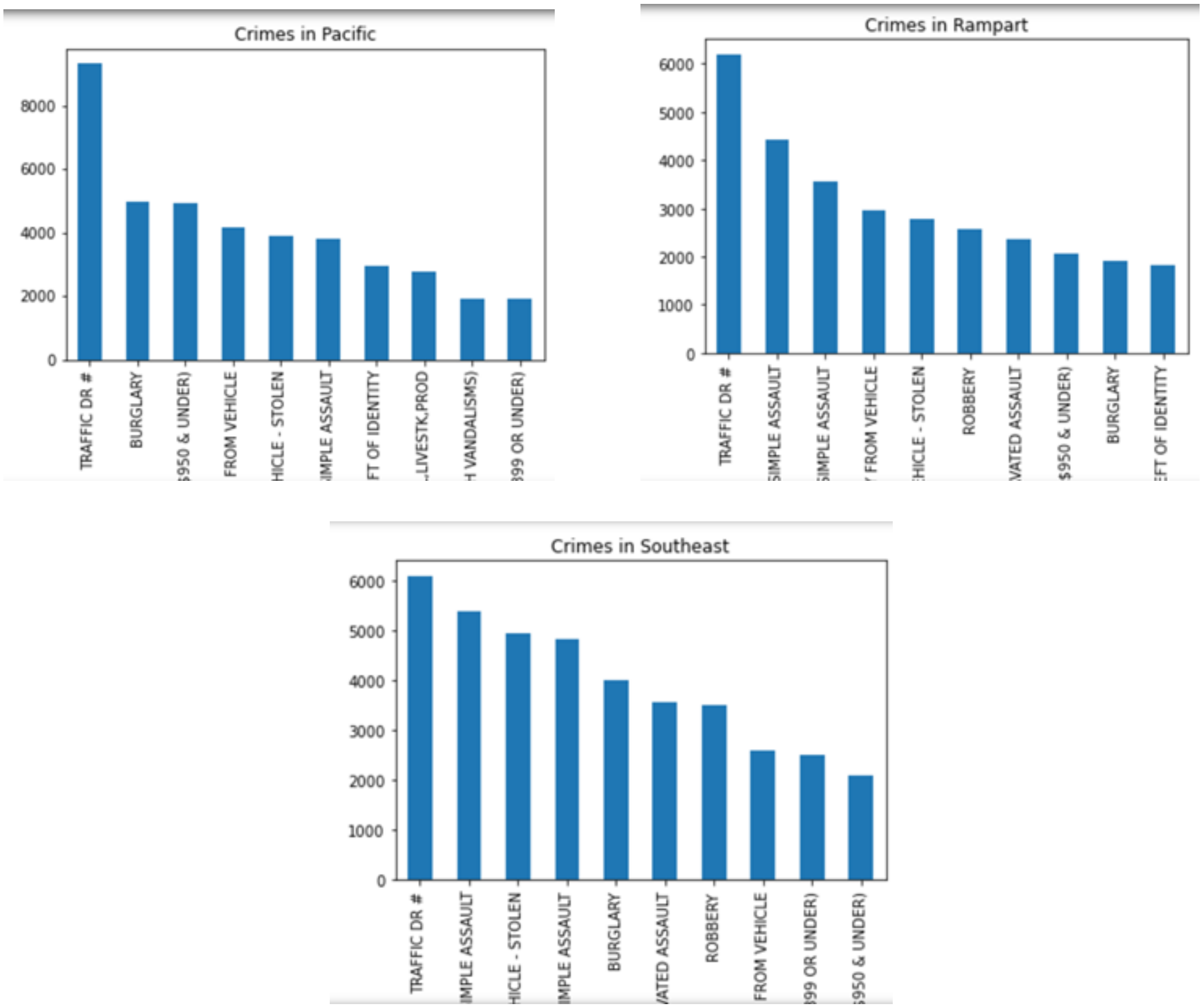


Fig 6: Street wise crime types in Los Angeles

B. Chicago Dataset Analysis:

The top five crimes and locations in Chicago were determined by plotting a horizontal bar graph based on crime count with respect to *Primary Type* and *Location Description* columns. The top five crimes (Fig 7) are battery, theft, criminal damage, assault, and deceptive practice and the top five crime locations are street, apartment, residence, sidewalk, and parking lot/garage.

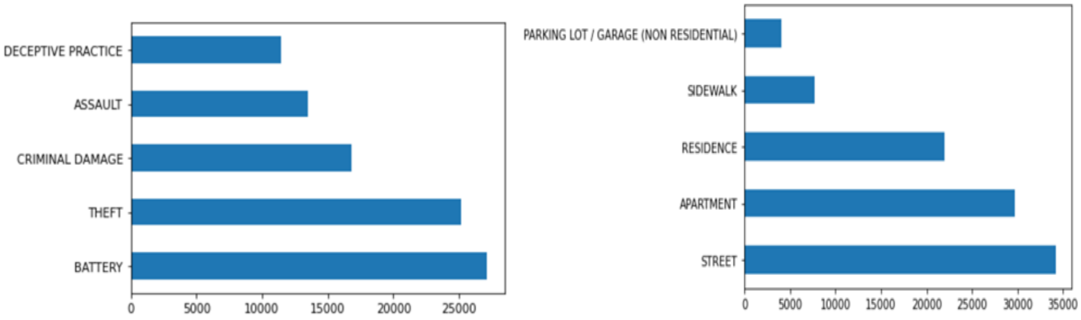


Fig 7: Top 5 crimes in Chicago (Left), Top 5 crime location in Chicago (Right)

Fig 8 demonstrates the crime rate across the years in Chicago i.e., from the year 2000 to 2020 (ignored 2021 as its full data for the year is not available) and plotted it with the help of a line graph. I see an overall decrease in the number of crimes that occurred over the 20-year period.

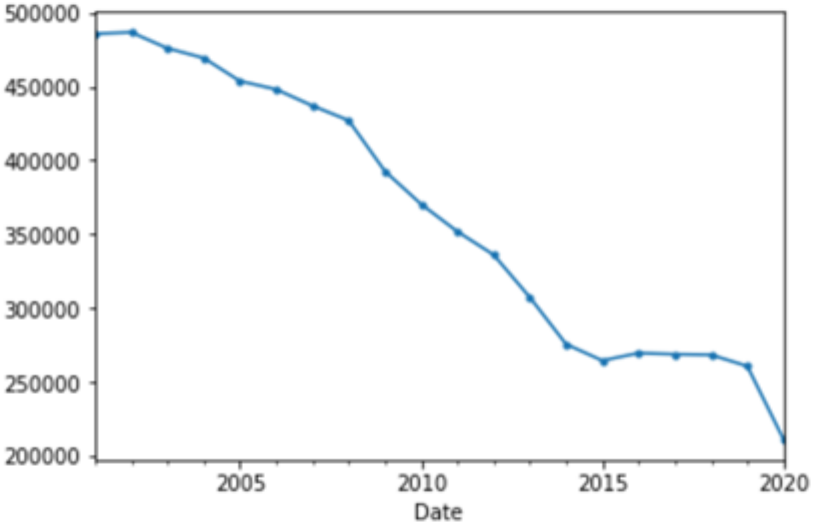


Fig 8: Number of crimes yearly in Chicago

The *Arrest* column has around 90% values as false i.e., the columns from the year 2001 to 2021 are all false.

Fig 9 shows the monthly count of top five crime types in Chicago and found out that the crime occurrence increases during summer.

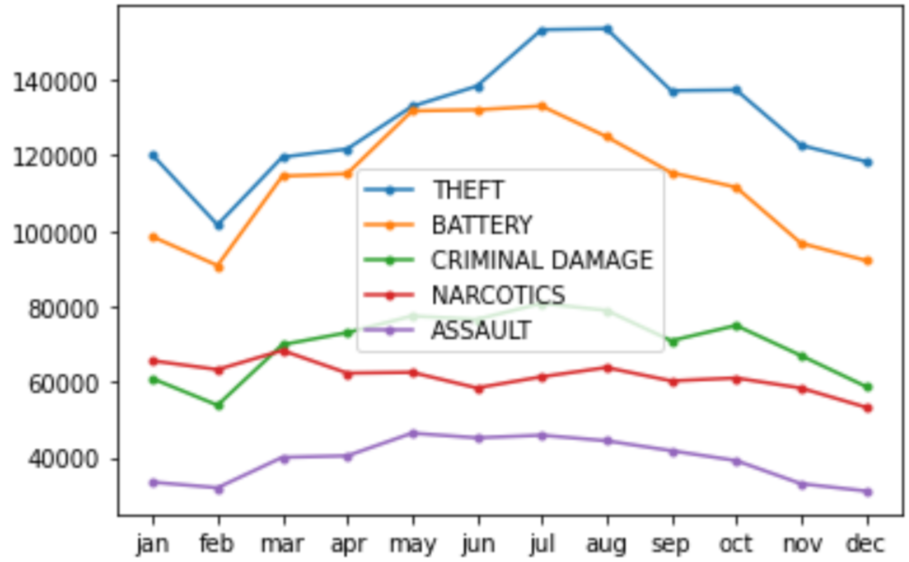


Fig 9: Month wise count of crimes in Chicago

Domestic crimes were detailed by classifying different crimes in the *Primary type* column i.e., robbery, theft, burglary, and motor vehicle theft as stealing whereas homicide, kidnapping, assault, criminal damage, and human trafficking were classified as criminal attacks and sex offense, crime sexual assault, stalking, obscenity, domestic violence, and intimidation were classified as sex assault. The bar graph in Fig 10 plots the number of crimes and shows that criminal attacks were the highest and when checked the same on monthly basis and plotted the below line graph Fig 11 and discovered that stealing and criminal attacks have been more in the mid-year and less around starting and ending times of the year.

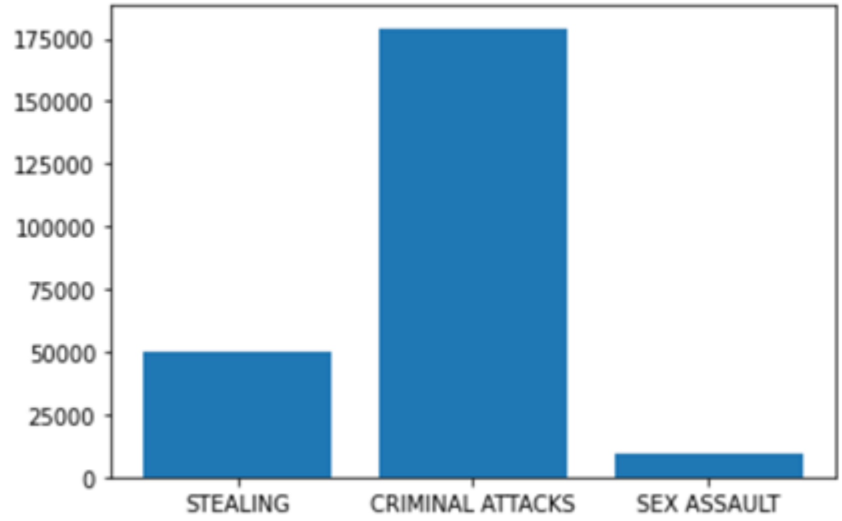


Fig 10: Number of Domestic crimes in Chicago

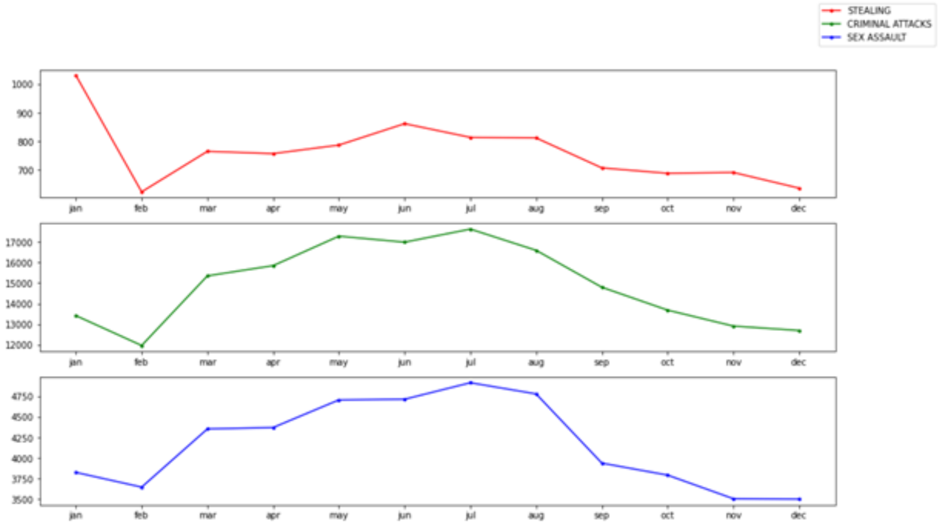


Fig 11: Number of monthly crimes in Chicago

A geolocator library was used to generate neighborhood, county, and municipality details with the help of longitude and latitude. The new data revealed that Hyde Park Township, Jefferson Township, and Lake Township are the top three municipalities with the highest number of crimes. All three places have tourist attractions I.e., Hyde Park Township has a museum and Lake Township has several lakes. This analysis concluded that the high crime rates may be attributed to high tourist foot fall.

As shown in fig 12, crimes based on time (I.e., Morning, Afternoon, Evening, Night) reveals that the highest number of crimes occurred in the afternoon. Further analysis of crime types based on time show that Assault, Criminal damage, Battery, and Theft are top four crimes across all the times (Fig 13).

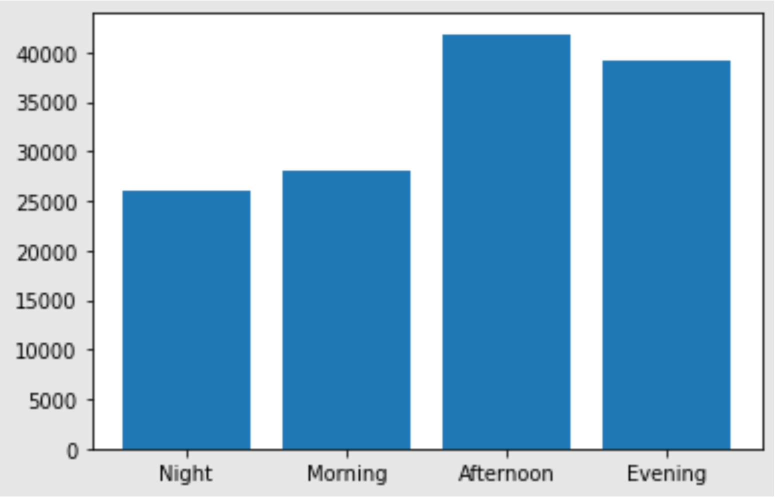


Fig 12: Crimes based on time in Chicago

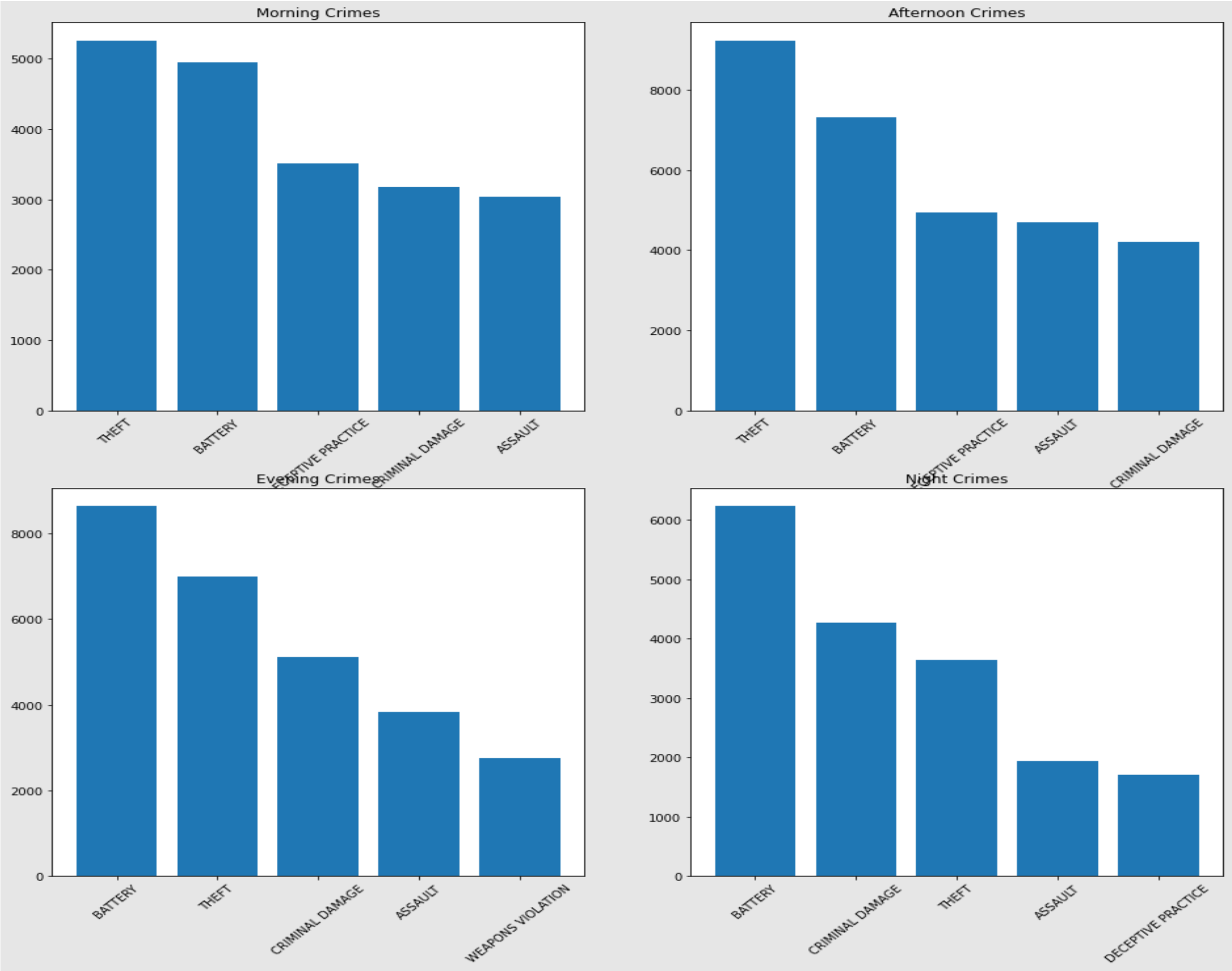


Fig 13: Type of crimes based on time in Chicago

Analyzed and plotted the different crimes on a yearly basis to observe the trends (increasing or decreasing) from the year 2006 to 2016. Fig 14, 15, 16, illustrates a significant decrease in crime rate in Chicago from 2006 to 2016.

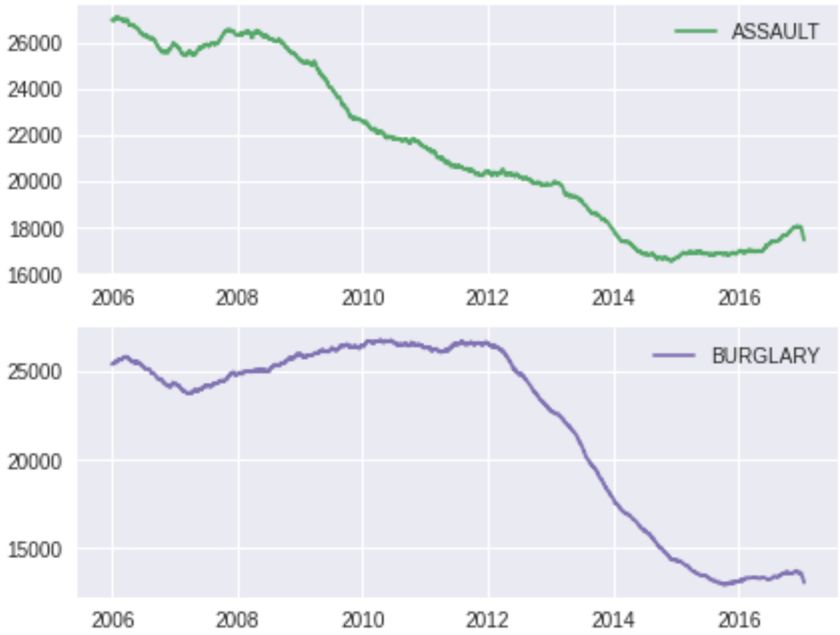


Fig 14: Assault & Burglary crimes trend.

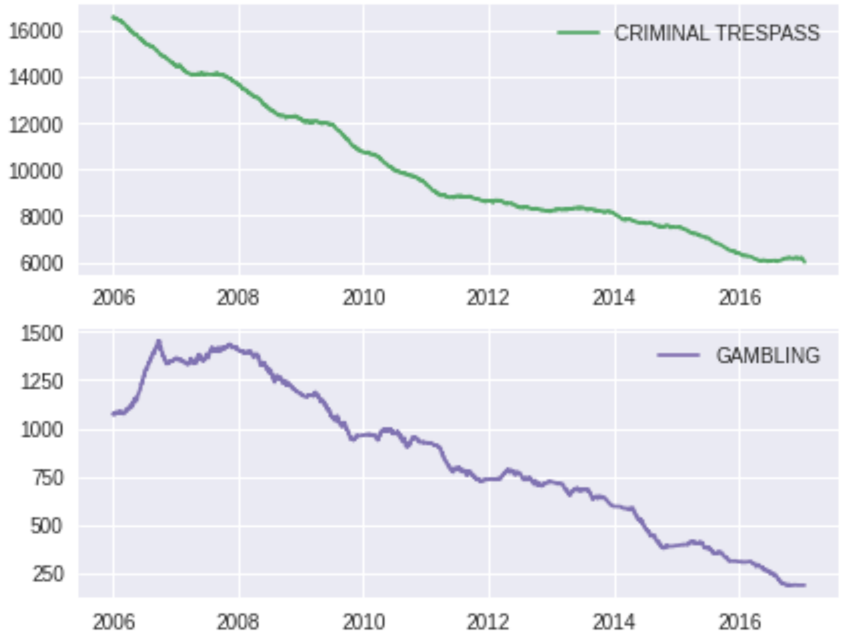


Fig 15: Criminal Trespass & Gambling trend.

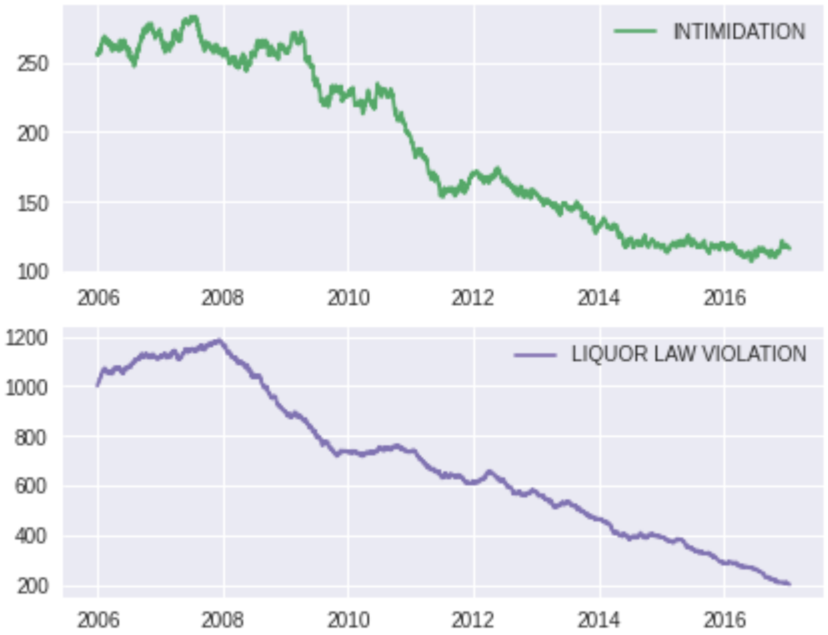


Fig 16: Intimidation & Liquor trend.

Fig 17 shows a significant increase in crime rates among different crimes in Chicago from 2006 to 2016.

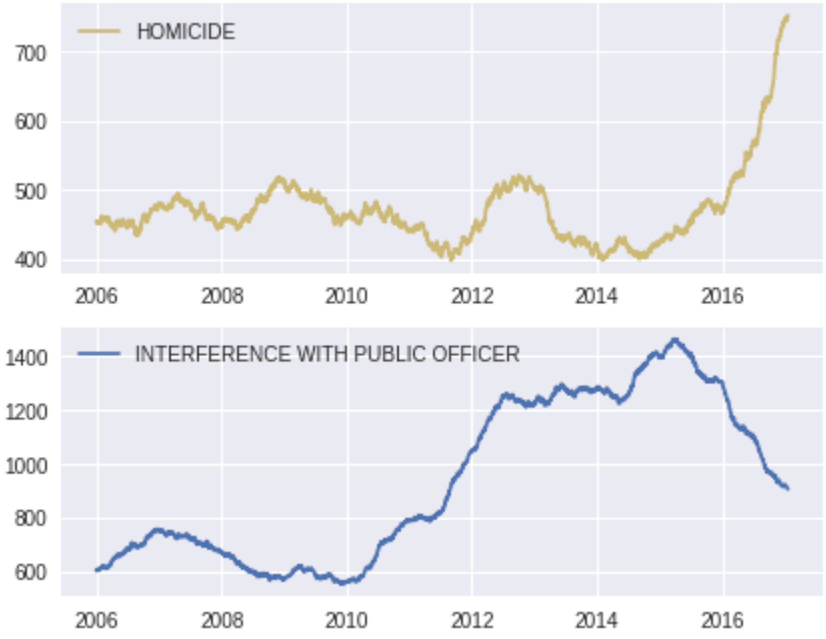


Fig 17: Homicide & Interference with public officer trend.

Fig 18 and Fig 19 suggest that the crime data from 2006 to 2013 were not recorded (or might have been considered as a different category).

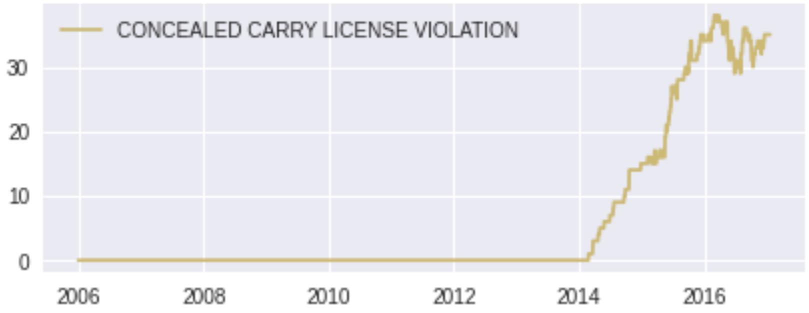


Fig 18: Concealed License Violation trend.

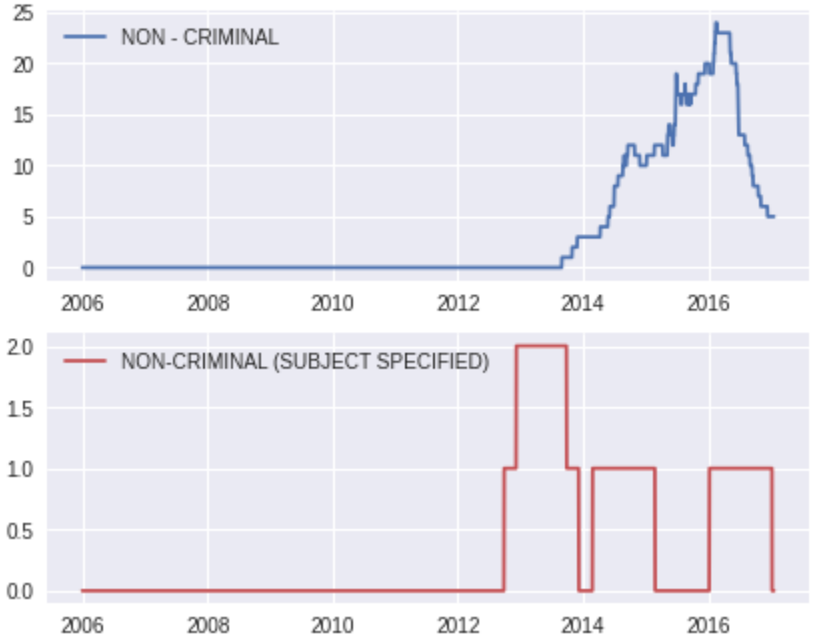


Fig 19: Non-Criminal trends.

Fig 20 illustrates a heatmap and location frequency for each crime in various places. The Y-axis represents the type of crime, and the X-axis represents the places. The color scale represents the severity of crime in a particular area; for example, Burglary is more severe in the streets.

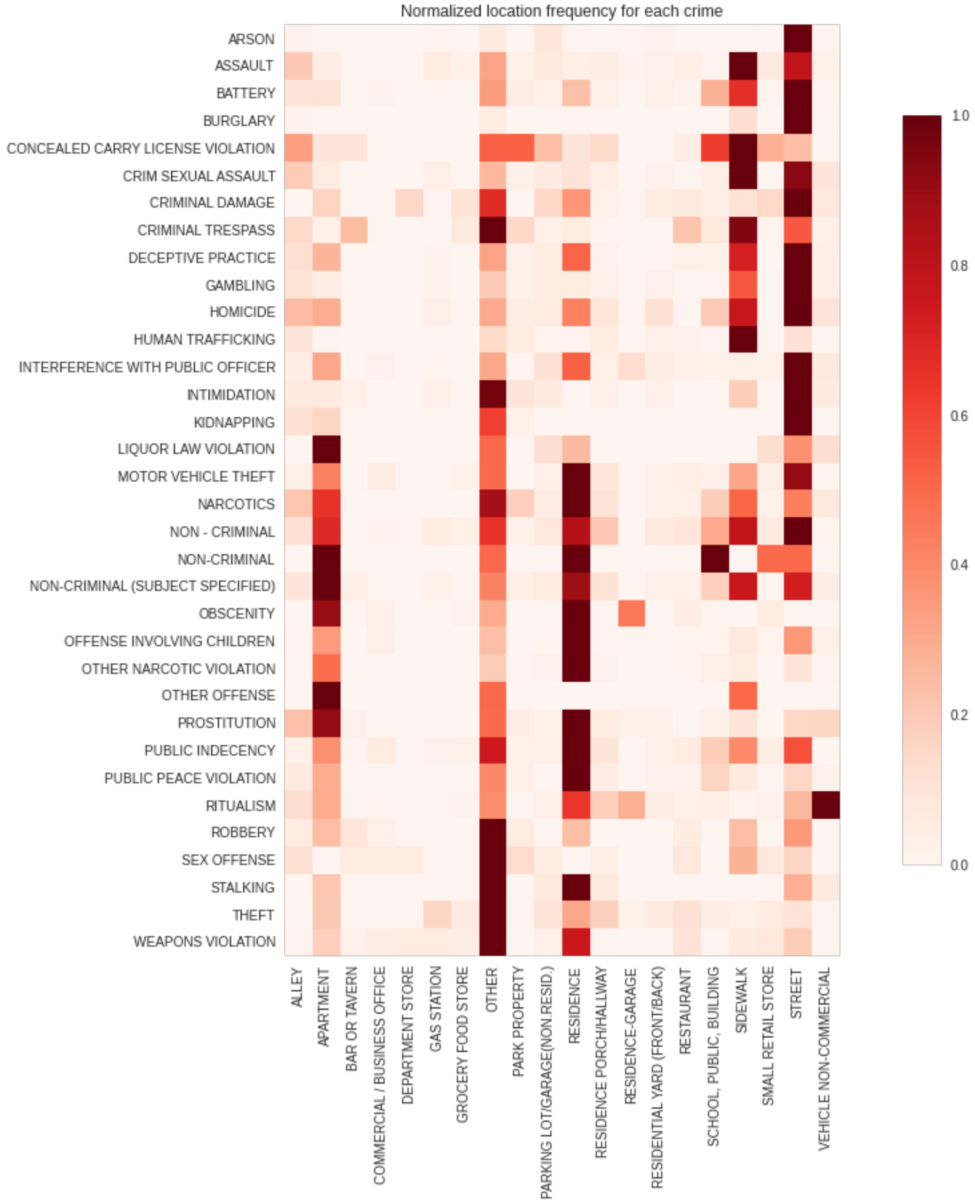


Fig 20: Heatmap describing the relation between the type of crimes and the type of locations

Projecting crimes on a map as a marker (Fig 21), based on its latitude and longitude, shows crimes at street level, and recognizes the hotspots. Based on this visualization, city authorities can increase the patrol units during peak crime hours in the hotspots. Marker cluster algorithm has been used on Chicago dataset to manage markers based on zoom level (When zoomed-in into the maps, the granularity of crime numbers based on region/street level increases).

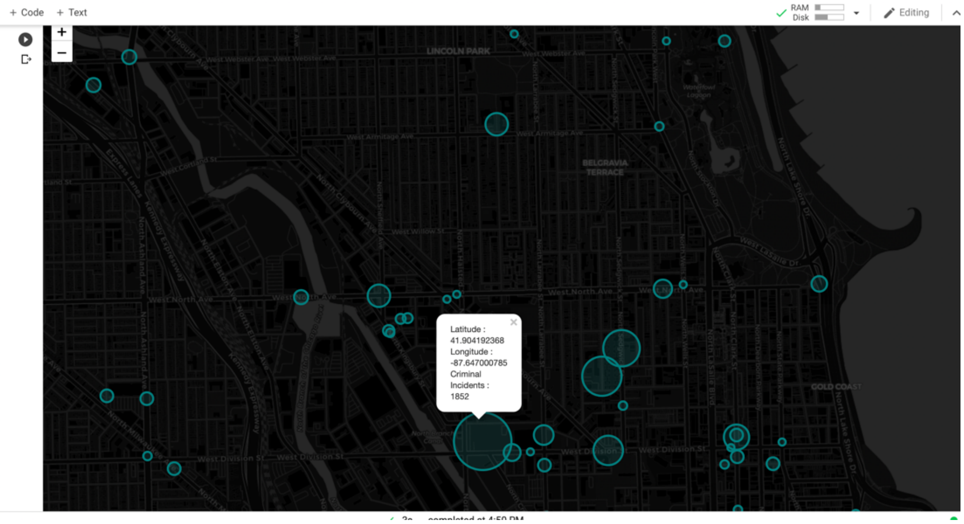


Fig 21: Folium map for Chicago city.

A comparison of Chicago and Los Angeles crimes shows that the number of crimes has decreased over a period. The top five crimes in both the cities are the same, namely, battery theft, theft, criminal damage, assault, and deceptive practice, and the top three crime locations also seen to be same: Most of the crimes occur on streets, apartments, and single-family residence.

Data Preprocessing:

Data pre-processing is the phase where data is transformed or encoded to a level where clearly understood by the algorithm. The quality of model depends on the quality of data. Below are the various data preprocessing methods performed.

The main objective of this analysis is to predict future crime based on past trends. So, using different techniques to fill in the missing values proved an unsatisfactory approach, as missing values can compromise the model.

These issues were addressed by

- Removing all the rows and columns with missing values from the dataset.
- Analysis of crimes at the district level in the city was done by identifying all the unique districts in Chicago city.
- The next step was creating separate columns for Month, Day of Week, Day of Month, Day of Year, Week of Month, Week of Year for filtering the crimes on monthly, weekly, yearly basis to check the frequency of crimes.
- Resampled the data to get the crime count based on date.
- Unnecessary attributes were omitted from the dataset.
- Converted longitude and latitude into area names using location generation API using multi-processing.

- Crimes with less frequency were grouped under other crime types.
- Using Pearson correlation, removed irrelevant columns.

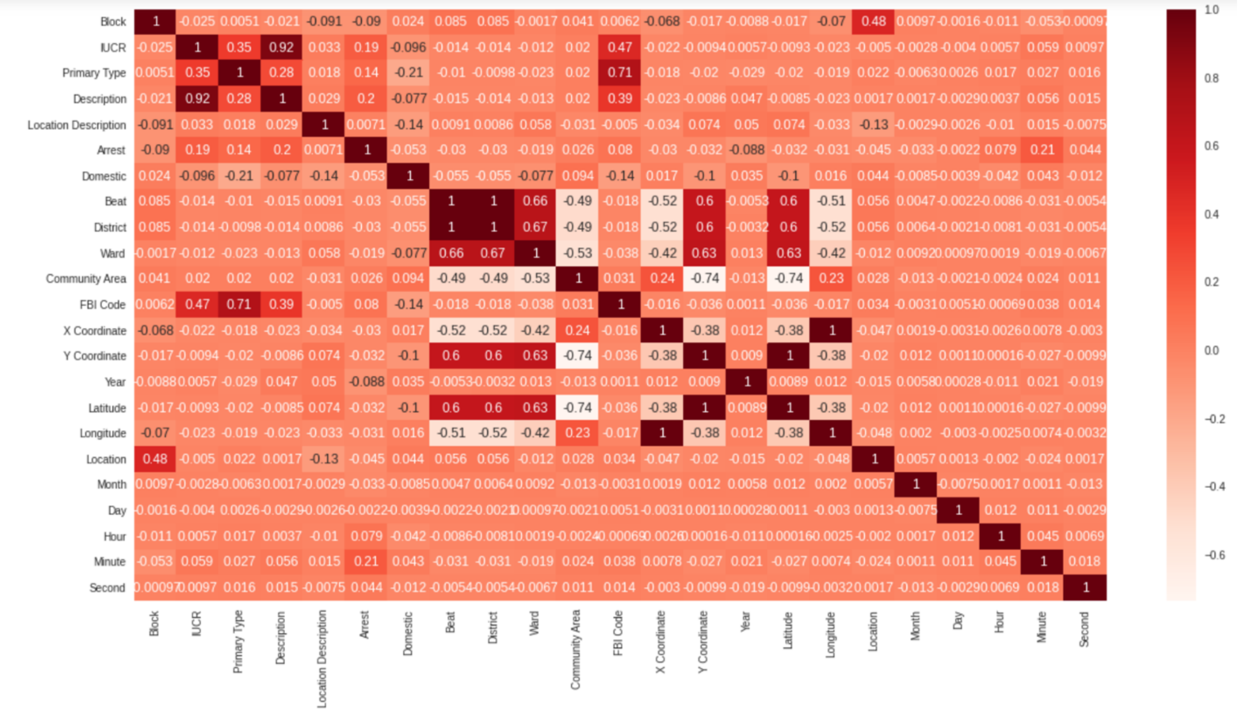


Fig 22: Heat map representing Pearson correlation

For classification models, some of the continuous values in a column such as the longitude and latitude were converted into area names I.e., categorical values. We were facing some issues, such as requirement of additional software and dependencies, while we were trying to run the location generation API in the college virtual systems using Spark, so we used multi-processing to do the same. Some of the columns which were found to be irrelevant by the Pearson correlation were removed. Some of the crimes which occurred less frequently were grouped under others.

Prediction Models:

To tackle the problem of crime forecasting, I explored several state-of-the-art machine learning models, and time series models. Time series models are a sequence of numerical data points successively indexed or listed in time order.

Below are the models that brought more insights into the crime data.

A. LSTM

Long Short-Term Memory (LSTM) is a special type of recurrent neural network (RNN) which has 4 layered architectures to overcome the long-term dependency problem in RNN.

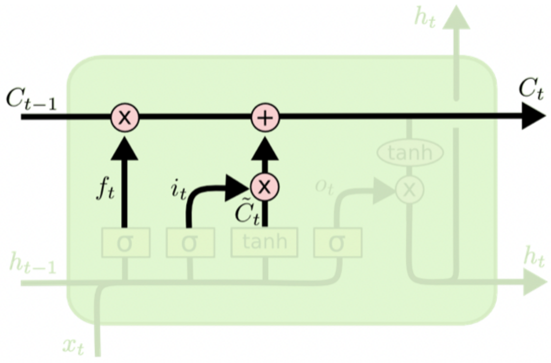


Fig 23: LSTM Module

The core of LSTM is its cell state (The highlighted horizontal line at the top of Fig 22) which runs straight down the entire chain with some linear interactions. Information was added or removed to the cell state based on the gates/layers. The information is then passed through a series of layers i.e., forget gate layer, input gate layer, tanh function layer and sigmoid function layer. f_t (Forget gate layer) is a sigmoid function that decides what information needs to pass. f_t takes h_{t-1} and x_t as inputs and outputs, a number between 0 and 1. A '0' means remove all data while a '1' means keep all data.

$$f(t) = \sigma(W_f.[h_{t-1}, x_t] + b_f)$$

This step served to identify the latest information will be added to the state. Initially, i_t a sigmoid layer (input gate layer) was used to decide new values. Then $\tilde{C}_t$, a tanh layer, was used to create a vector of new feature values to be added to the state.

$$i(t) = \sigma(W_i.[h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = tanh(W_C.[h_{t-1}, x_t] + b_C)$$

In this step, we will update old cell state (C_{t-1}) to a new cell state (C_t) as shown below.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Finally, the output will be based on cell state with some linear alterations. We pass our input data separately through sigmoid and tanh layers and multiply those outputs to get the desired information.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * tanh(C_t)$$

LSTM Training Procedure:

The LSTM-model training procedure involved creating a list to train each district and group based on month and year. To train the LSTM model at district level I divided the data into training and testing sets where the training set contains data from 2012 to 2018 and the testing set contains data from 2019 to 2020. The number of layers in LSTM is 50 and trained for 200 epochs.

B. Prophet [1]

Prophet model is designed for forecasting on univariate time series datasets. It uses a decomposable time series model which involves 3 main model components i.e., seasonality, trend, and holidays. They are combined in the following equation

$$Y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

where g(t) is the trend function for non-periodic changes in time series, s(t) speaks about periodic changes (e.g., hourly, weekly, and yearly seasonality), and h(t) represent the effects of holidays which occur on potentially irregular schedules over one or two days. ϵ_t is the error term representing the changes that are not accommodated by the model.

For trend function, they have implemented a piecewise linear model with a limited change points and constant growth rate. Trend changes are incorporated into the model by explicitly defining the change points.

$$g(t) = (k + a(t)^T \delta)t + (m + a(t)^T \gamma)$$

$$\sigma(s, i) = \begin{bmatrix} 1, & t \geq s_j, \\ 0, & otherwise. \end{bmatrix}$$

where k is the growth rate, δ is for rate adjustment, m is the offset parameter and γ is set to $-s_j\delta_j$ to make the function continuous. δ_j is the change in rate that occurs at time s_j. s_j will give us the number of change points at times j=1,2,3…S.

Time series data often have multi period seasonality because of human behaviors. For example, vacations and school breaks can produce effects that repeat each year. We rely on the Fourier series defined below for our seasonal function.

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

where P is the period (e.g., P=365 for yearly data and P=7 for weekly data). Holidays and special events often vary with the trends in periodic patterns, so their effects are not well modeled by a smooth cycle. By adding an indicator function (l), that figures out time (t) during holiday (i) and allocate each holiday (i), a parameter k_i which is proportional to the change in forecast. This is achieved by generating a matrix of regressors Z(t).

$$Z(t) = [l(t \in D_1, ..., l(t \in D_L))]$$

$$h(t) = Z(t)k$$

Where, D_i is set of past and future holidays (i). k~Normal $(0, v^2)$.

Training Procedure:

Training procedure involves resampling and splitting of data. Data based on day was resampled and a new data frame was created with 2 columns (ds (date), y (count)) as per prophet requirement. Then, the prophet model was trained by dividing the data into training and testing sets where the training set contained data from 2012 to 2018 and the testing set had data from 2019 to 2020.

Experimental Results and its Analysis:

Using the Chicago dataset from 2012 to 2018, LSTM & Prophet model were trained. Testing was done using the data from 2019 to 2020. When comparing the results from both the models (LSTM & Prophet), I can see the root mean square error (RMSE) for LSTM is 121.4 whereas the Prophet model got an RMSE of 143.91. From these initial prediction results, I can conclude that the LSTM performed better when compared to the Prophet model.

Test Results of the Proposed Method:

LSTM Prediction results:

The prediction accuracy of the result was evaluated using Root Mean Square Error (RMSE) which is calculated by the root of the mean of the squared differences between the ground truth and predictions. An RMSE of 121.4 was produced when the model was trained with 50 layers using Adam optimizer, Relu activation function and an epoch of 200. Figure 23 to 26 illustrates some of the district crime predictions and the ground truth graphs.

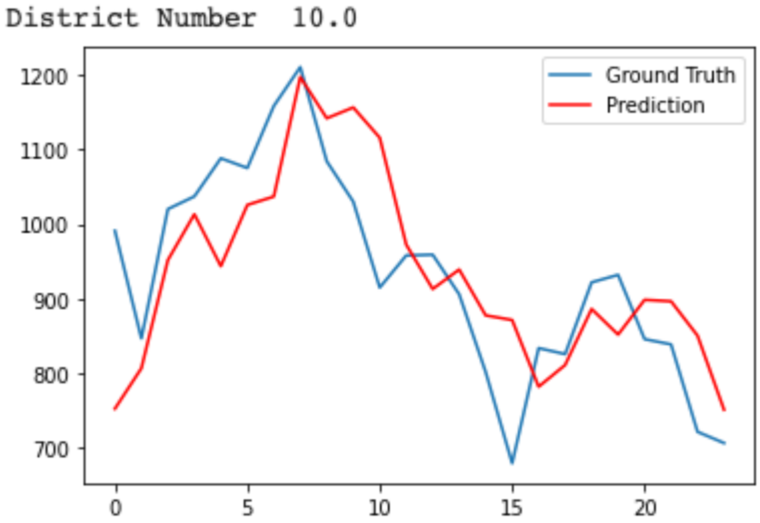


Fig 24: District 10 crime prediction.

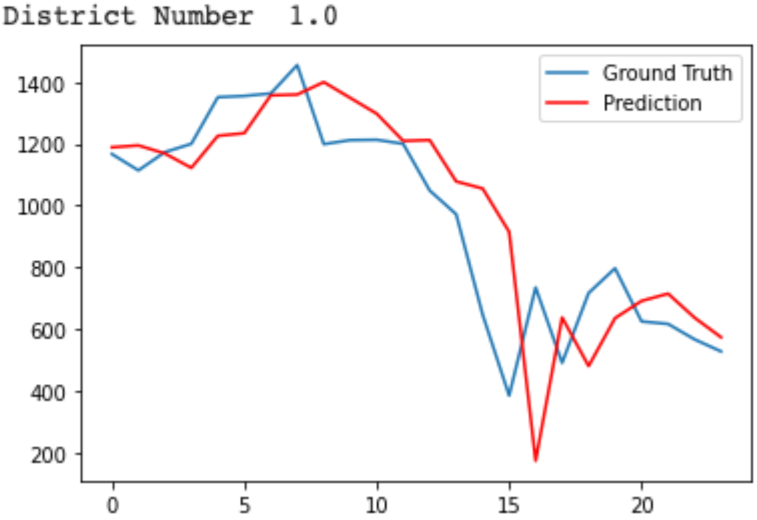


Fig 25: District 1 crime prediction.

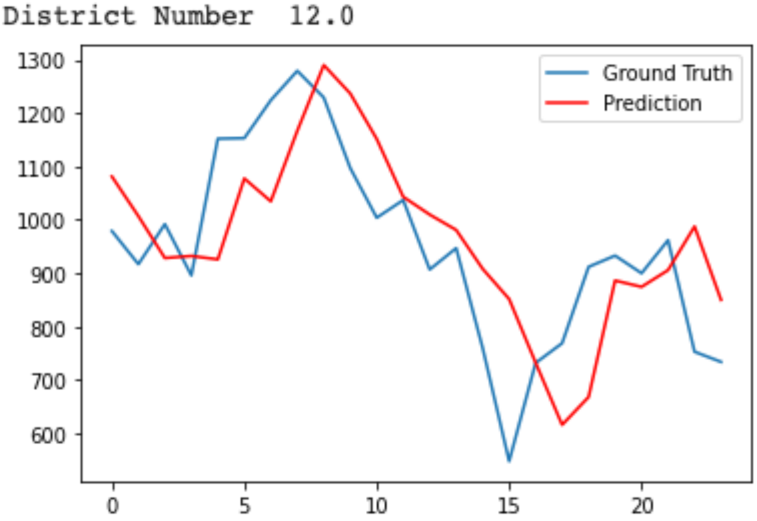


Fig 26: District 12 crime prediction.

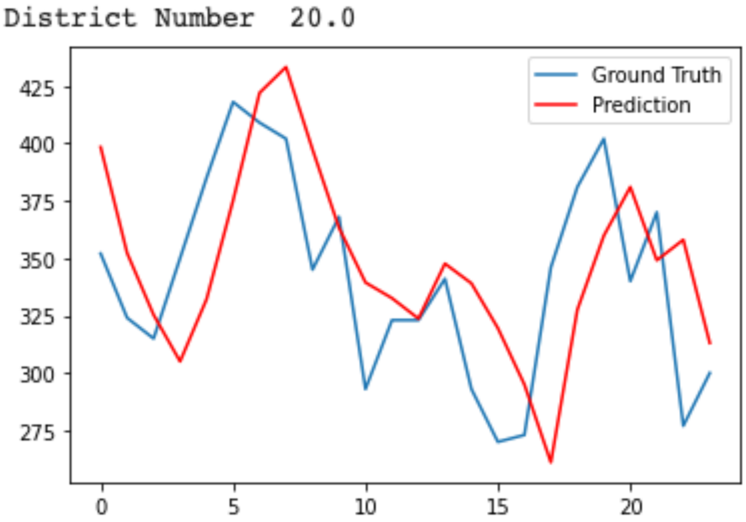


Fig 27: District 20 crime prediction.

Prophet Prediction Results:

To present a clear visualization, plotted the monthly crimes by adding all the crimes from that month. The prediction results were evaluated using Root Mean Square Error (RMSE) which is calculated by the root of the mean of the squared differences between the ground truth and predictions. An RMSE of 143.91 was obtained when predicting the monthly data (Evaluation of prophet model on daily crime data is in progress).

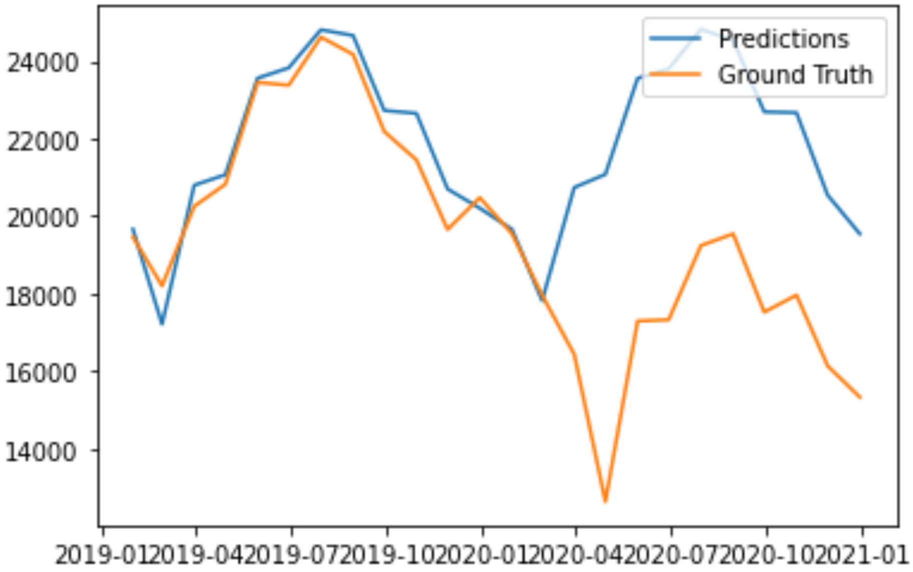


Fig 28: Prophet model crime prediction.

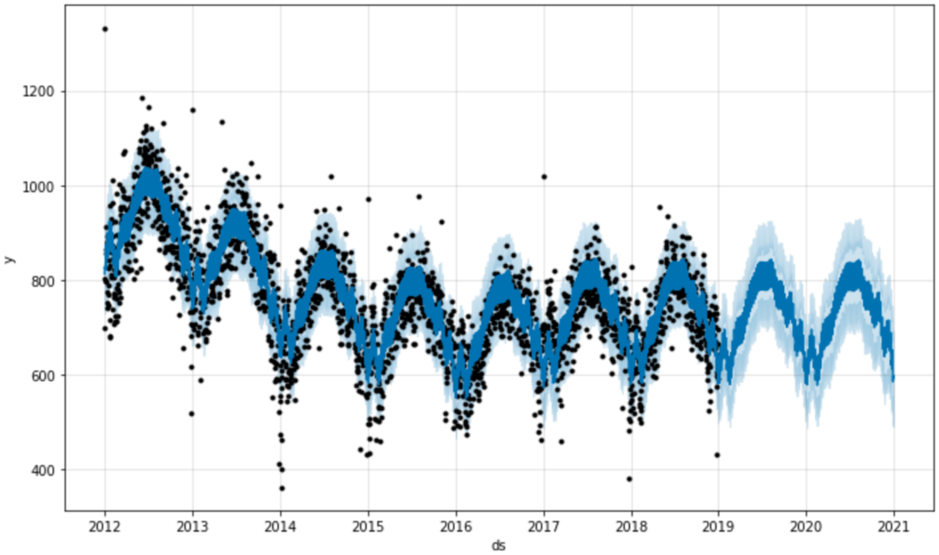


Fig 29: Cross validation of prophet model.

Conclusion:

In this paper, various machine learning and visualization techniques on the Chicago and Los Angeles dataset helped me identify patterns and trends in the crimes. Data analysis results of Chicago dataset clearly indicates a decline in the number of crimes for certain types (Assault, Burglary, Criminal trespass, Gambling, Intimidation etc.) from 2006 to 2020. Also, I can see an incline in the number of crimes for certain types (Homicide, Interference, License violation, non-criminal trends) from 2006 to 2020. I was able to identify the crime type based on the location and time. This will help the authorities to deploy the forces accordingly and help in reducing the crime rates.

Challenges:

As the selected dataset contains the data based on city level there is an Insufficiency of Data based on district and street level so, I have extracted the location based on latitudes and longitudes using the Geopandas library.

Contributions:

1) Till date all the papers were focused on predicting crimes on a yearly, monthly, daily basis, but my idea is to predict crimes based on time of the day (i.e., Morning, Afternoon, evening, night). This prediction helps in distribution of troops based on the crime type and severity. For this purpose, new columns were needed like Day of the week, Day of month, Day of year.
2) The other research papers have helped me to get an overview of the process to be followed, various levels of predictive modeling and to get a clear understanding of the various steps

like data pre-processing, predictive modeling, and alternative methods like using the pipelines involved within the method.

Future Work:

- In future, I plan to build a Reinforcement Learning model where agents can identify the crime patterns in the data with a proper reward system and decision-making process (exploration and exploitation) in place.
- Building a classifier with current data along with spatial and surrounding information can significantly improve the results.

References:

[1] 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.

[2] Jawaher Alghamdi, Zi Huang, Modeling Daily Crime Events Prediction Using Seq2Seq Architecture in ResearchGate, doi:10.1007/978-3-030069377-0_16.

[3] Md. Aminur Rab Ratul, A Comparative Study on Crime in Denver City Based on Machine Learning and Data Mining, CoRR abs/2001.02802 (2020).

[4] W. Safat, S. Asghar and S. A. Gillani, Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques, in IEEE Access, vol. 9, pp. 70080-70094, 2021, doi: 10.1109/ACCESS.2021.3078117.

[5] Pratibha, A. Gahalot, Uprant, S. Dhiman and L. Chouhan, Crime Prediction and Analysis, 2nd International Conference on Data, Engineering and Applications (IDEA), 2020, pp. 1-6, doi: 10.1109/IDEA49133.2020.9170731.

[6] TaylorSJ, Lethan B.2017. Forecasting at scale, Peer J Preprints 5:e3190v2, https://doi.org/10.7287/peerj.preprints.3190v2.

[7] Anahita Ghazvini, Siti Norul Huda Sheikh Abdullah, Mohammad Kamrul Hasan, Datuk Zainal Abidin Bin Kasim, Crime Spatiotemporal Prediction with Fused Objective Function in Time Delay Neural Network, Access IEEE, vol. 8, pp. 115167-115183, 2020.

(Anonymous) Sharing agreement:

Do you agree to share your work as an example for next semester?
Yes

Do you want to hide your name/team if you agree?
Yes