Time Series Analysis and Trend Forecasting of Crime Data

ITCS5156-Applied Machine Learning Project Proposal

Project proposal by ***Ramasri Saladi***

(Niner Id: 801254656)

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, Los Angeles, 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. My study focuses on leveraging the effective strategies of big data analytics in present crime prevention research as well as applying improvements to the current methods.

Journal Details:

- ***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:*** *IEEE*
- *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.*

Motivation:

Till date all the papers were focussed 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.

DataSet:

- *https://www.kaggle.com/currie32/crimes-in-chicago*
- *The dataset consists of data from the year 2001 to 2021.*
- *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).*

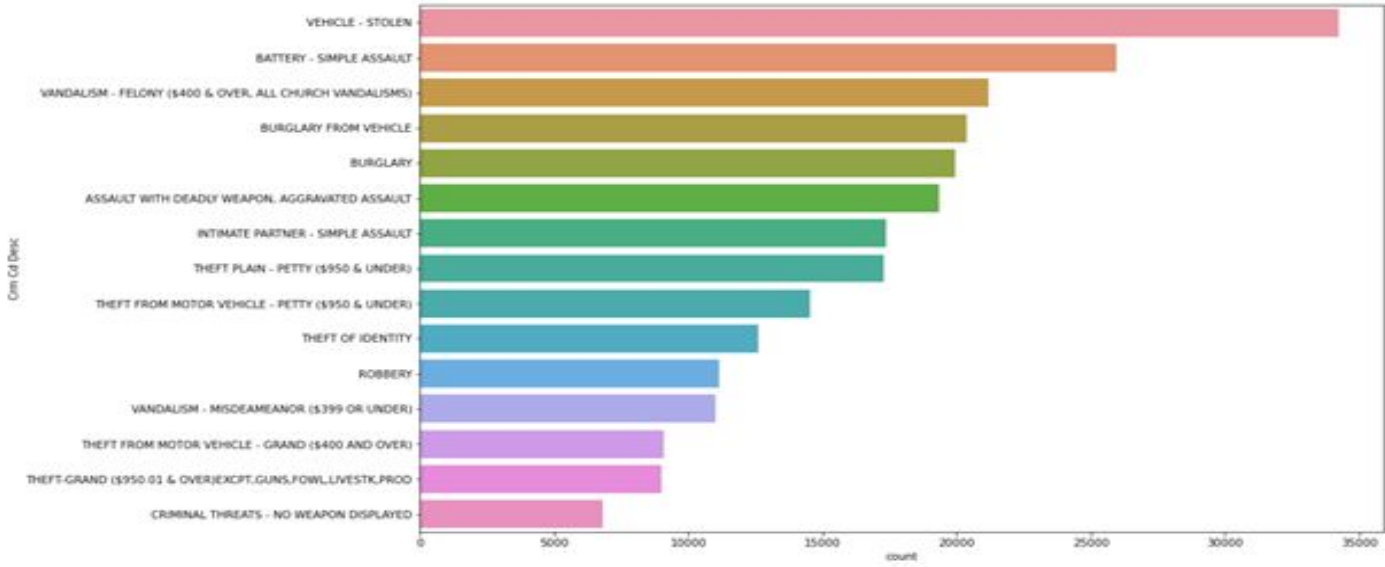
Approach

- *Objective*
  Building effective models that helps in crime prevention.
- *Initial Analysis*
- *Data Preprocessing*
- *Predictive Modeling*
  a) *Time series Analysis*
     i) *LSTM*
     ii) *Prophet*

Data Analysis & Visualization Los Angeles

Top Crimes in LA

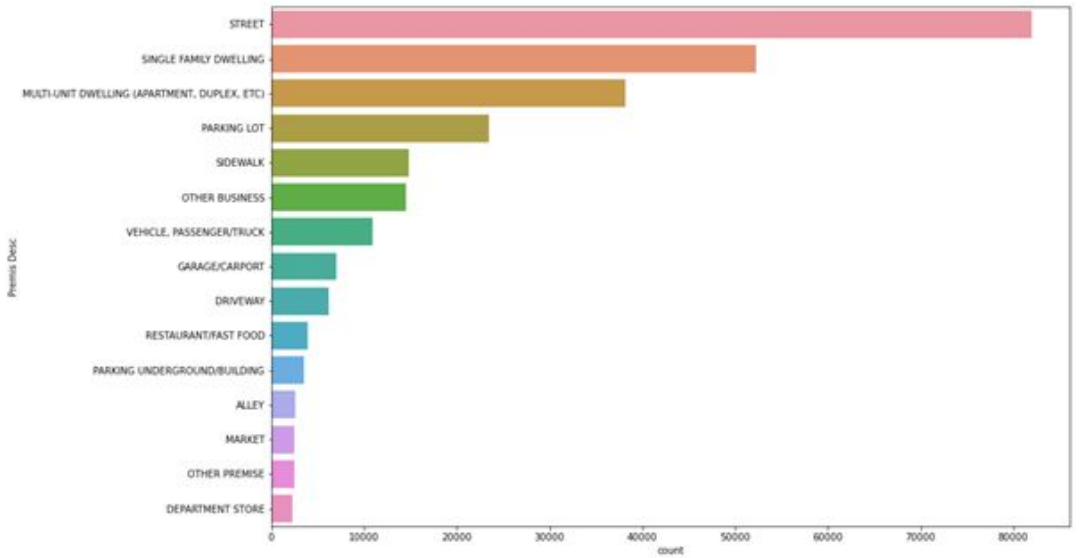
- *Vehicle stolen*
- *Battery Assault*
- *Vandalism*
- *Burglary*



Data Analysis & Visualization Los Angeles Contd.

Top crime locations in LA

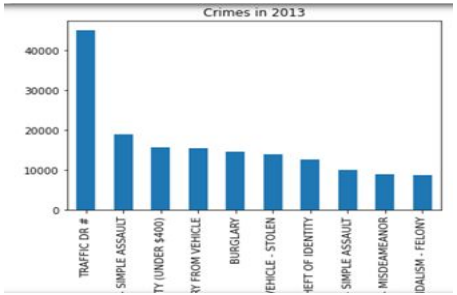
- *Street*
- *Single Family Dwelling*
- *Multi Unit (Apt, Duplex)*
- *Parking Lot*
- *Side Walks*

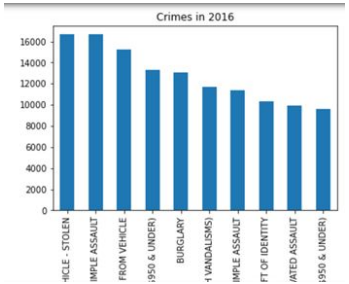


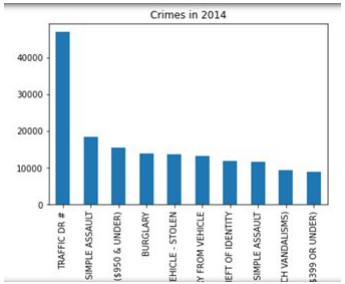
Data Analysis & Visualization Los Angeles Contd.

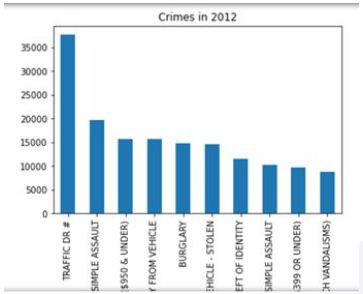
Top 10 crimes every year

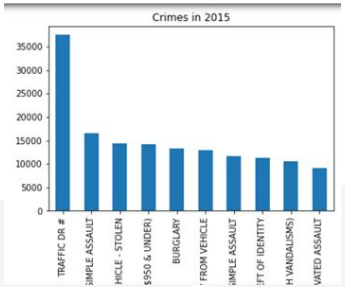
- *Burglary*
- *Battery Theft*
- *Vehicle Stolen*
- *Traffic DR#*











Common trends b/w LA & Chicago

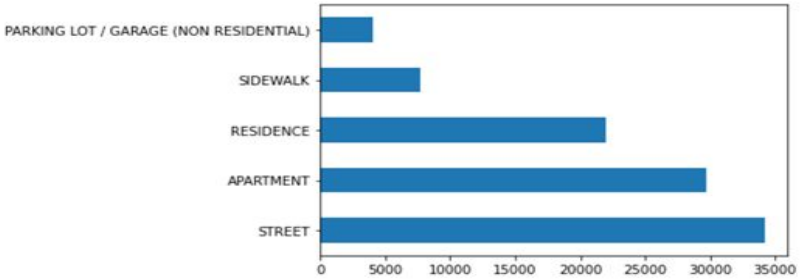
- *Data analysis results from Chicago & Los Angeles has shown some similar trends.*
- *Top crime locations are almost same:*
  - *Street,*
  - *Apartment,*
  - *Sidewalks*
- *Top crimes are similar:*
  - *Battery,*
  - *Burglary,*
  - *Assault*
- *Decrease in the number of crimes across the years (2000 – 2021)*

Data Analysis & Visualization Chicago

Top 5 crimes & crime locations in Chicago

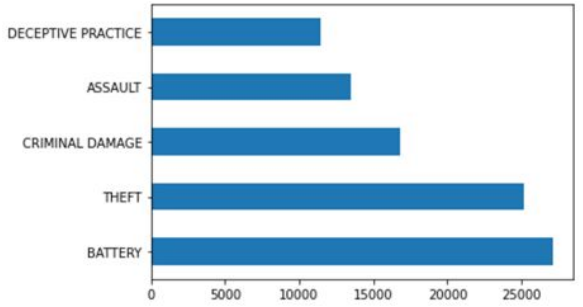
Locations

- *Street*
- *Apartment*
- *Residence*
- *Sidewalk*
- *Parking Lot*



Crimes

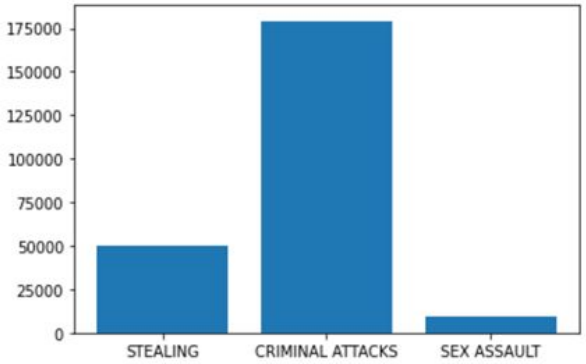
- *Battery*
- *Theft*
- *Criminal damage*
- *Assault*
- *Deceptive Practice*

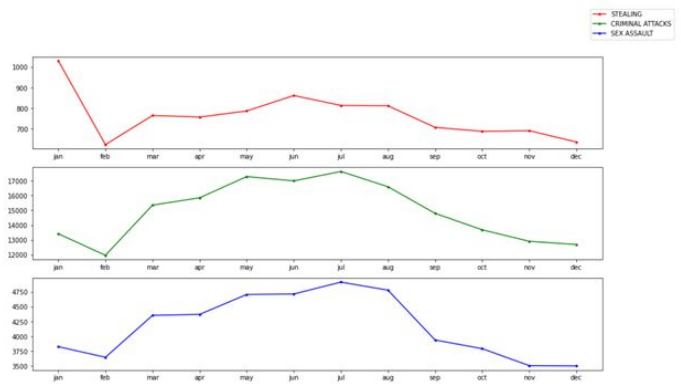


Data Analysis & Visualization Chicago Contd.

Number of domestic crimes & monthly crimes in Chicago

- *Criminal attacks, Stealing, Sex assault are the domestic crimes among which criminal attacks have been more during the mid year.*

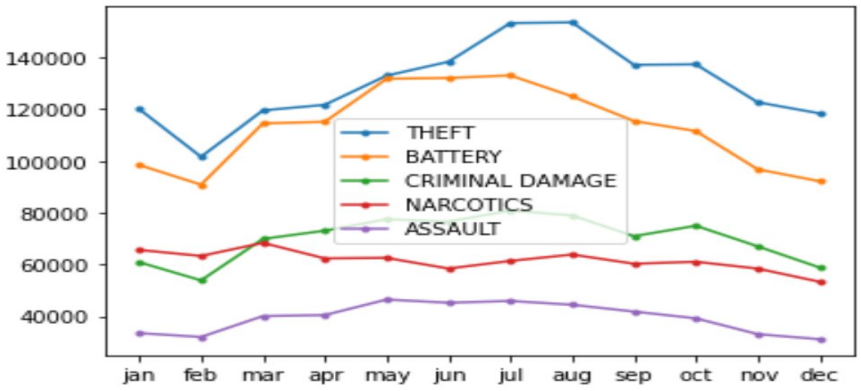




Data Analysis & Visualization Chicago Contd.

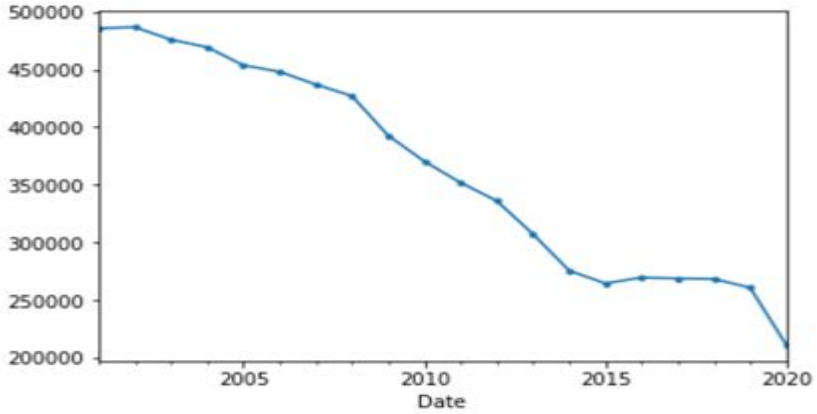
Monthly count of top 5 crimes

Monthly trend shows significant increase during summer.



Yearly count across years

Yearly trend shows significant decrease in crime count from 2001 to 2020.

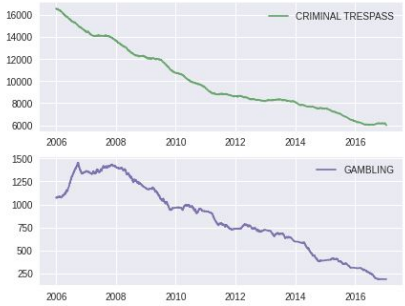


Domestic crimes were grouped by 'Primary Type'
Ex: Kidnapping, Homicide, human trafficking classified as 'Criminal Attacks'

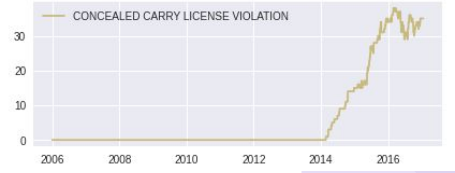
Data Analysis & Visualization Chicago Contd.

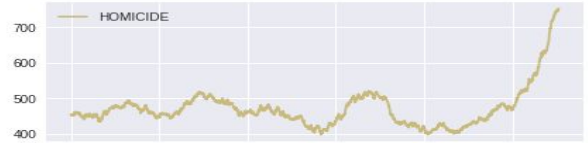
Yearly trend for Individual crime types

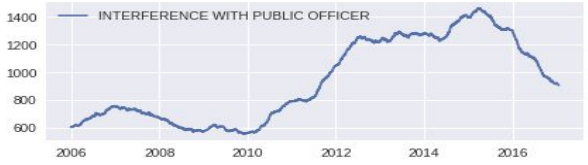
Decrease in crime count



Increase in crime count



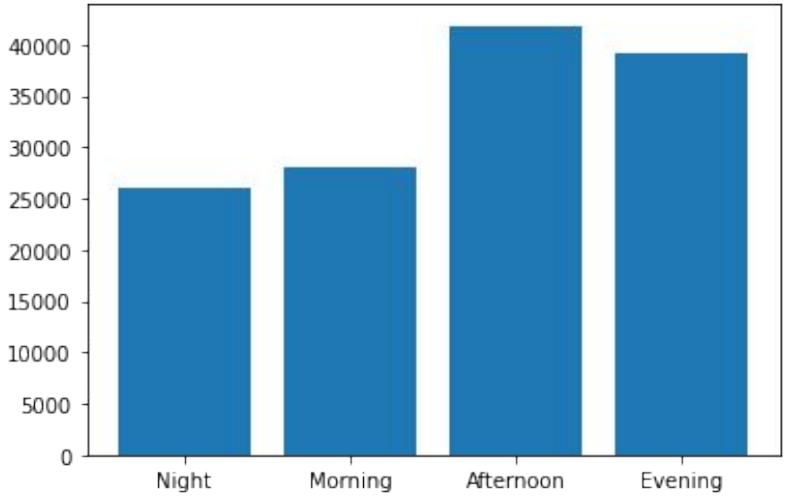




Data Analysis & Visualization Chicago Contd.

Crimes based on time

- *Visualized the number of crimes based on time i.e., (Morning, Afternoon, Evening, Night)*
- *Highest number of crimes occurred in the Afternoon.*

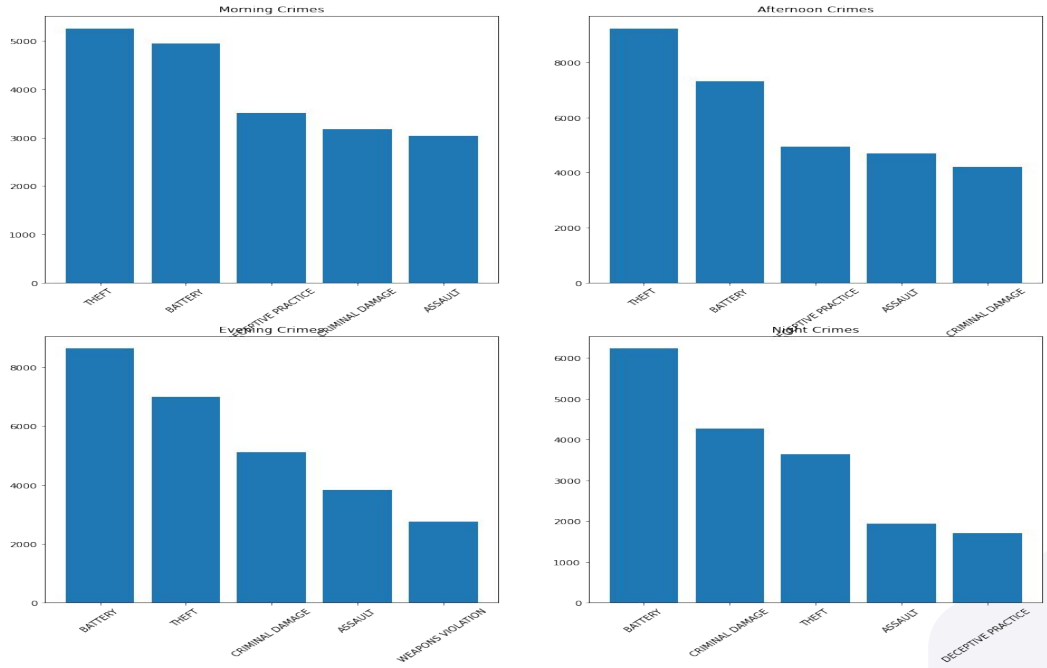


Data Analysis & Visualization Chicago Contd.

Crimes types based on time

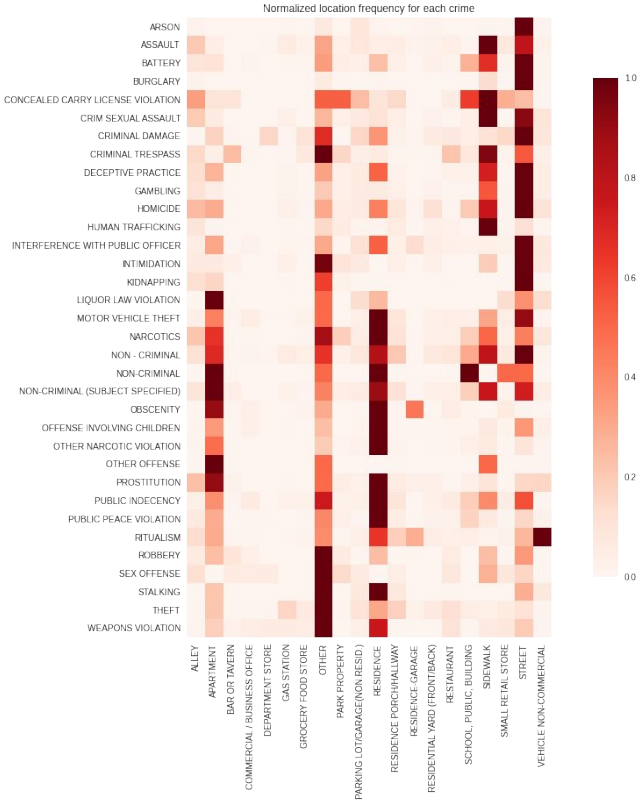
- *Visualized the type of crimes based on time i.e., (Morning, Afternoon, Evening, Night)*
- *Assault, Criminal damage, Battery theft are top crimes during all times.*

Data Analysis & Visualization Chicago Contd.



Data Analysis & Visualization Chicago Contd.

Heat map for location frequency of each crime



- *X – axis represents location (place)*
- *Y – axis represent crime type*
- *Color scale represents severity of crime in that area*
- *Ex: Burglary is severe in streets*

Data Preprocessing:

Steps:

- *Missing values can compromise the model so, removed all the rows and columns with missing values.*
- *Identified unique districts to analyze crimes at district level.*
- *Break down of the crime date by adding additional columns like Month, Day of the week, Day of the month etc. to filter the crimes on monthly, yearly, weekly basis.*
- *Resampled the data to get the crime count based on date.*
- *Omission of unnecessary attributes.*
- *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.*

Time series Analysis – LSTM:

- *Long Short-Term Memory (LSTM) is an artificial neural network architecture used in the field of deep learning*
- *Developed to deal with the vanishing gradient problem*
- *Information will be passed through series of layers i.e., forget gate layer, input gate layer, tanh function layer, sigmoid function layer.*

Time series Analysis - LSTM

Training Procedure:

- *Created a list based on month and year to train the model for each district.*
- *No. of layers in LSTM is 50 and trained for 200 epochs.*

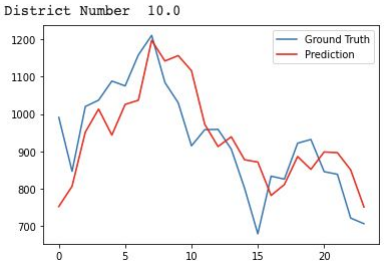
- *Split the data into training and testing data sets:*
  - *Training data – 2012 to 2018*
  - *Testing data – 2019 to 2020*

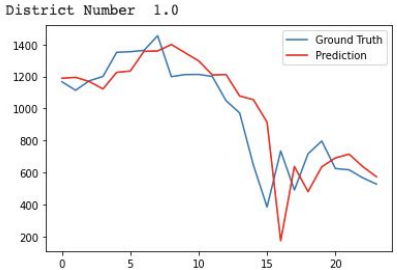
Time series Analysis - LSTM

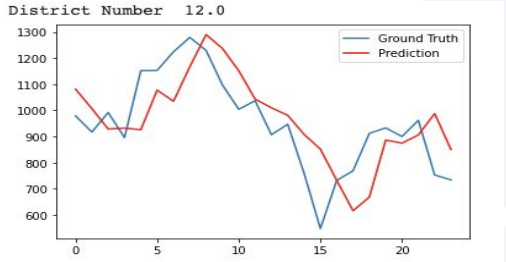
LSTM Prediction results:

- *Accuracy was evaluated using Root Mean Square Error (RMSE).*
- *RMSE obtained was 121.4*

District crime predictions:







Time series Analysis - Prophet

- *It involves 3 main model components i.e., seasonality, trend, and holidays*

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

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

Time series Analysis - Prophet

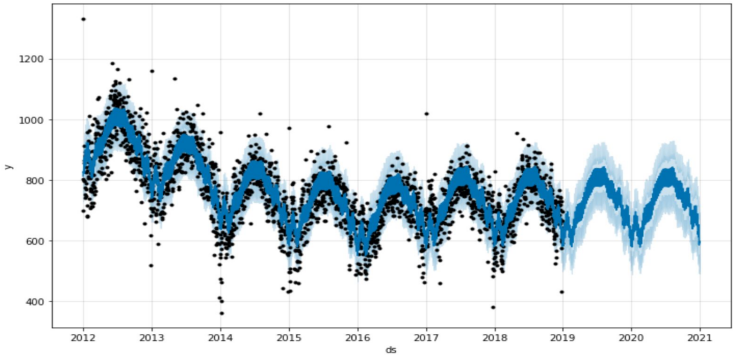
Training Procedure:

- *Resampled the data based on day and created a new data frame with 2 columns (date, count).*

- *Split the data into training and testing data sets:*
  - *Training data – 2012 to 2018*
  - *Testing data – 2019 to 2020*

Time series Analysis - Prophet

Prediction Results:



Conclusion:

- *Analysis results from Chicago dataset indicates a decline in number of crimes for certain types (Burglary, Assault, Criminal trespass, Gambling.) from 2006 to 2020.*
- *Incline in certain crime types (Homicide, Interference, License violation, non-criminal trends) from 2006 to 2020.*
- *Model was able to identify the crime type based on the location and time which will help in deploying the forces accordingly.*

Future Work:

- *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 in place.*
- *Building a classifier with current data along with the spatial and surrounding information can significantly improve the results.*

Related Papers:

- ***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.

Related Papers (Contd):

- ***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.

Timelines:

Task	***Start Date***	***End Date***
Data Selection	*01/15/2022*	*01/22/2022*
Project Proposal	*01/22/2022*	*01/29/2022*
Initial Analysis	*01/29/2022*	*02/05/2022*
Data Pre-Processing	*02/05/2022*	*02/12/2022*
Predictive Modeling	*02/12/2022*	*02/19/2022*
Validation and evaluation	*02/19/2022*	*02/26/2022*
Final Report	*02/26/2022*	*03/01/2022*

Thank you