Al Alipour
September 17, 2021

Time Series Analysis of Crime Rate in Chicago

Problems Statement

Crime is one of the main problems that municipalities of different sizes, cultures, and resources face. Chicago is the most affected city by crime in the United States, with a murder rate that is four times higher than the national average. Violent crime rate in Chicago is higher than the US average as well (Wikipedia, accessed August 31, 2021). Previous studies have shown that crime rate in Chicago is associated with seasonal as well as trending patterns. Forecasting crime rate in Chicago can provide useful information for municipalities and other governmental departments to identify ways to address this issue and develop programs to lower crime rate in the city. It can also provide useful information to individuals living in or visiting the city.

Data and Preparation

Main source of data for the project was City of Chicago's data portal (https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2). The dataset included more than 7,000,000 records of criminal activities in the city between January 1, 2001 and August 12, 2021 (date the data were acquired). For each record, there were data on the location and date/time of the activity as well as its criminal type and whether an arrest was made. Less than 10% of the data included missing information on some of the location features (Fig. 1).

Data were explored, organized and cleaned through the data wrangling phase of the project. It was investigated whether missing information on some of the location features could be imputed using the rest of the location information. It was, however, discovered that the majority of missing values could not be imputed accurately through this approach. Since the

location information were likely not to be used in the analysis at all, imputation of the missing location information was not addressed further.

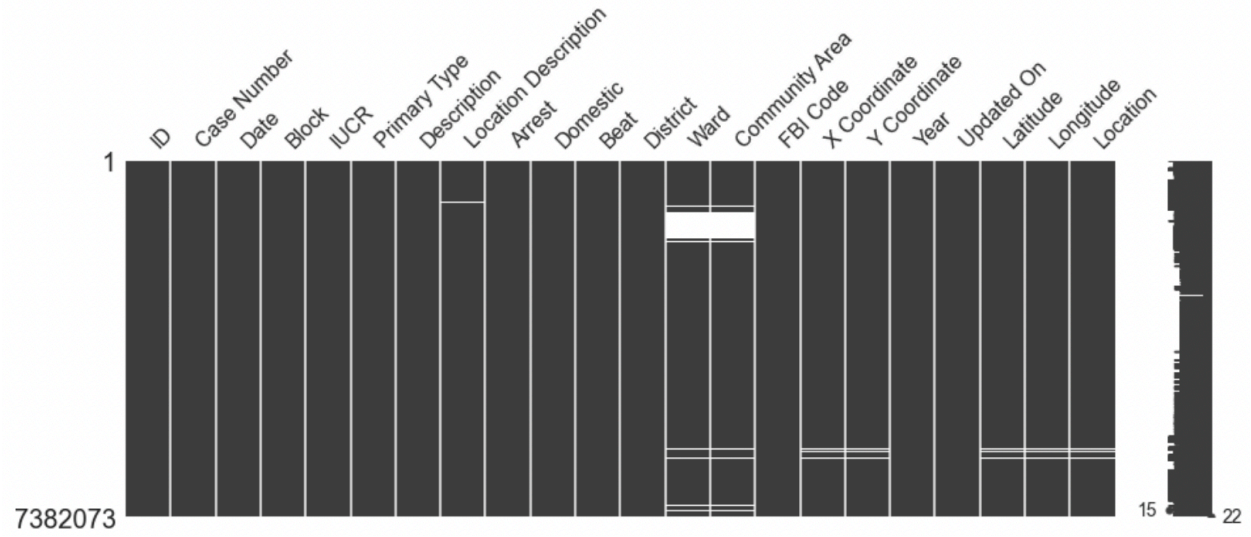

Figure 1. Distribution of missing values

Exploratory Data Analysis

Data were further explored using different visualizations to find potential patterns, identify required corrections, and discover associations. Visualizing monthly distribution of crime rate in Chicago clarified that crime rate had a seasonal pattern and was associated with a decreasing trend (Fig. 2). Seasonality and trend were thus identified as two of the factors to be considered during preprocessing and modelling phases. Distribution of different types of crime in Chicago was visualized as well using a bar chart, and it was discovered that historically theft followed by battery were the two most widespread crime types in the city (Fig. 3). Another visualization was created using Folium heat map to identify how crime distributed across different locations in the city (Fig. 4). In terms of location information of the criminal activity such as street, sidewalk, etc., it was discovered that street followed by residence and apartment witnessed the highest number of criminal activities (Fig. 5). Finally, it was observed that most criminal actives were not domestic and that most of them did not lead to an arrest (Fig. 6).

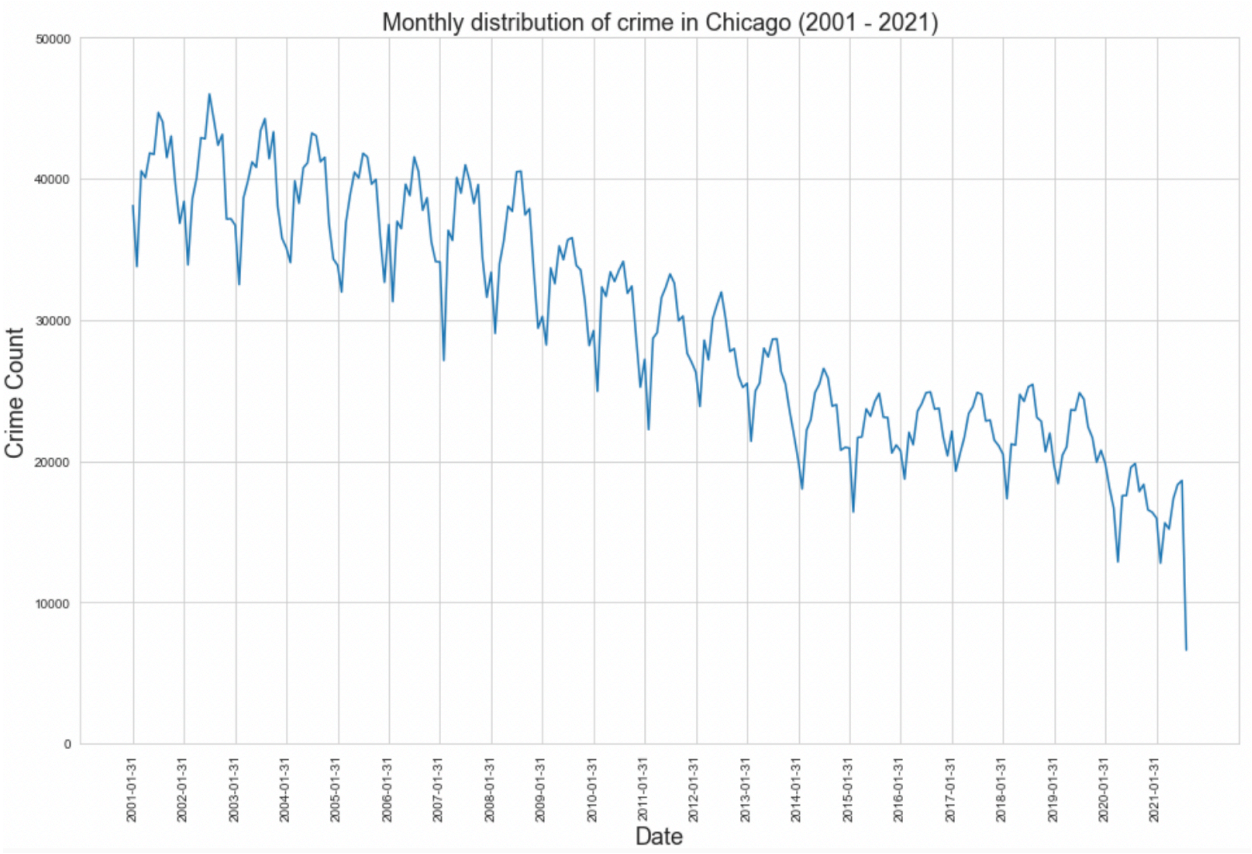

Figure 2. Monthly distribution of crime rate in Chicago

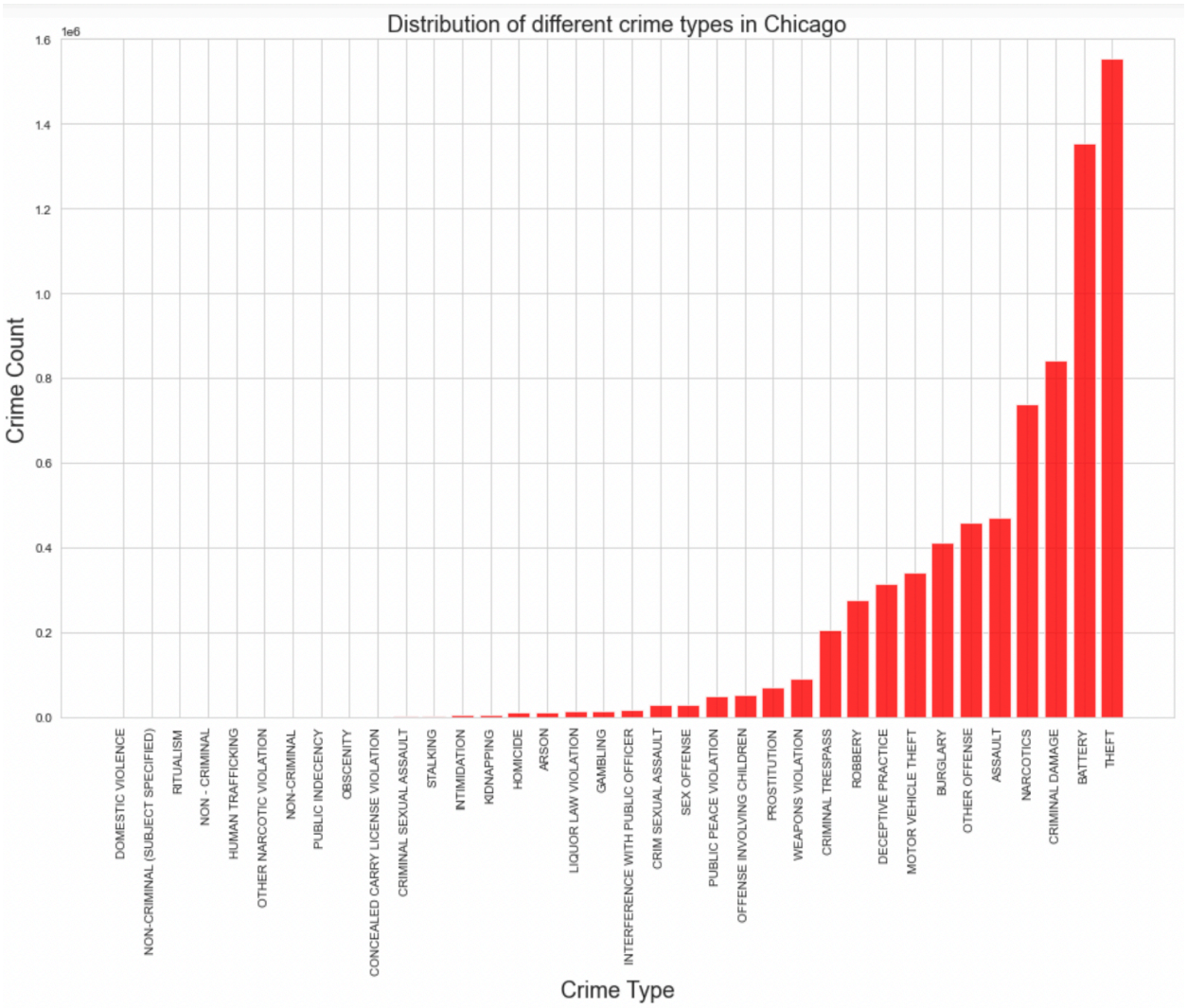

Figure 3. Distribution of different crime types in Chicago

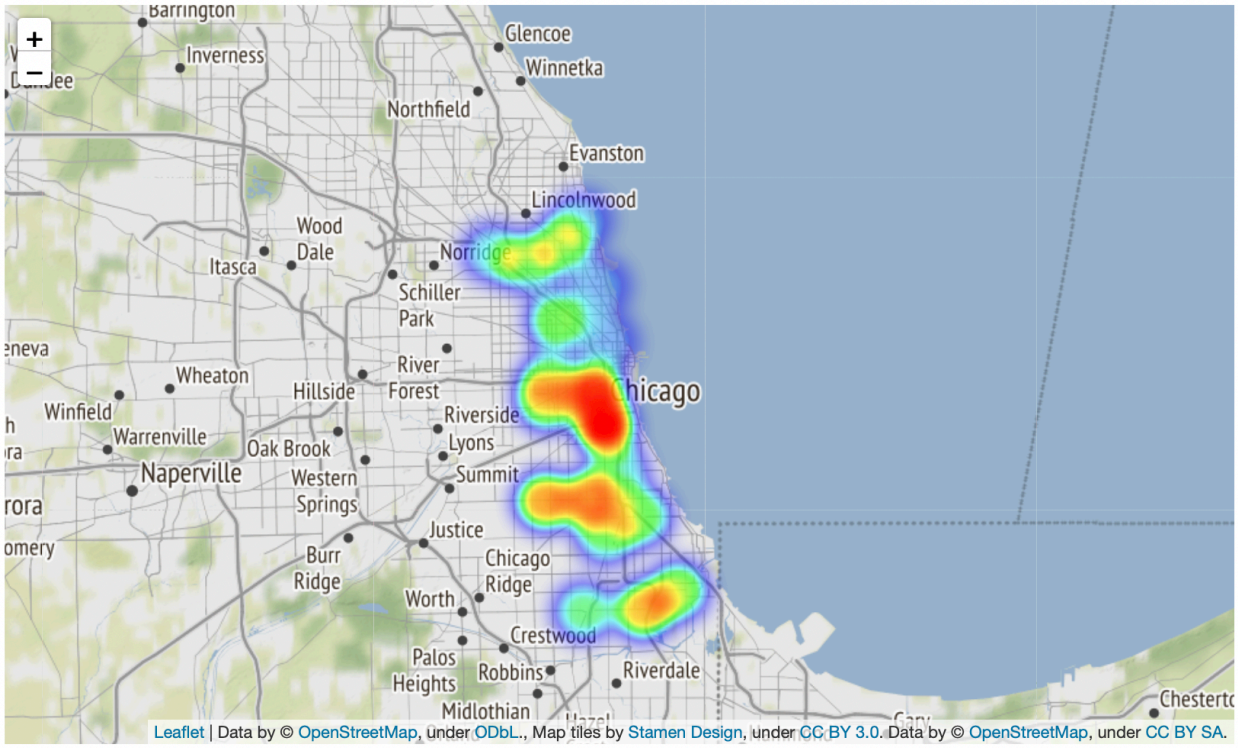

Figure 4. Location distribution of crime in Chicago

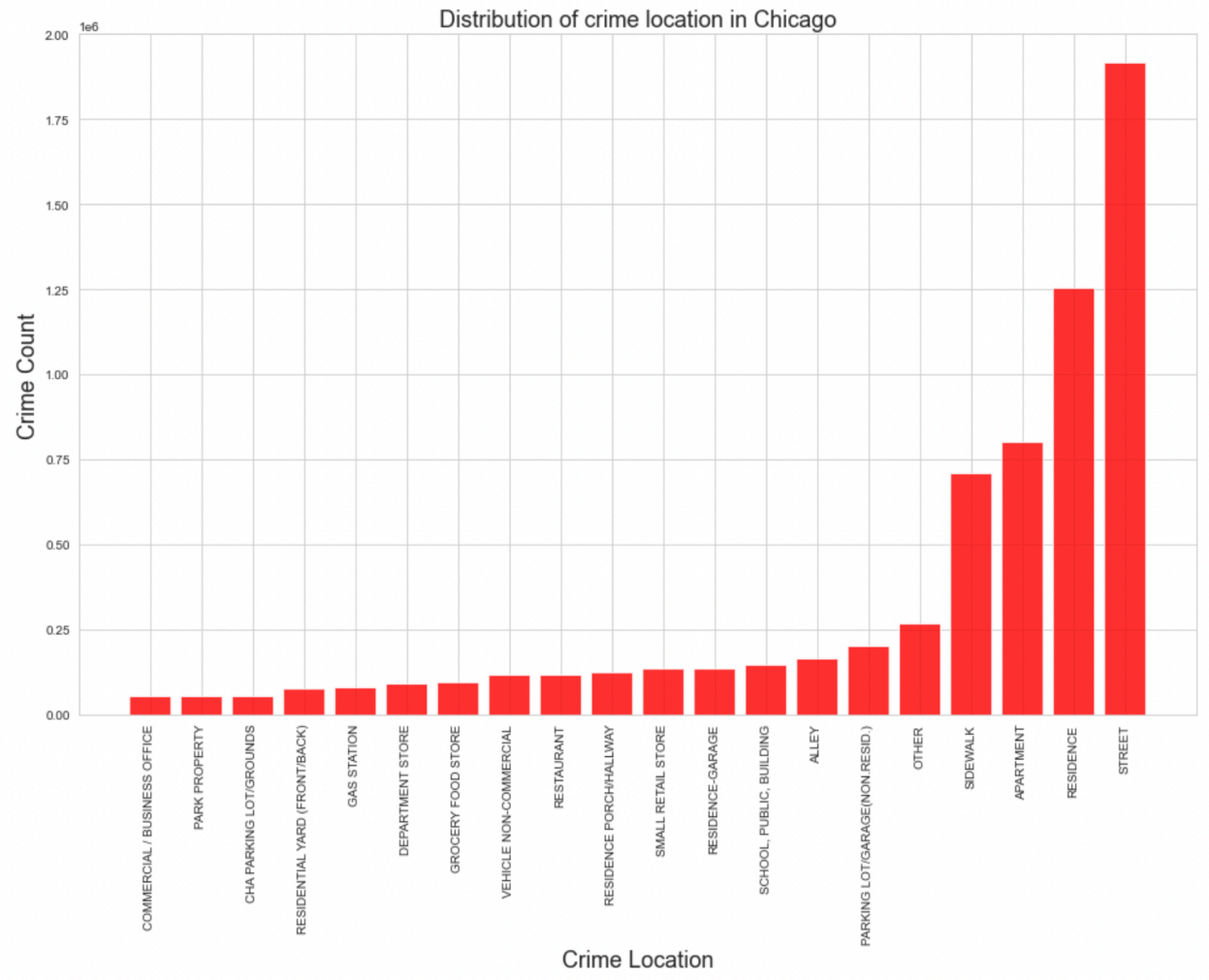

Figure 5. Most frequent locations of criminal activities in Chicago

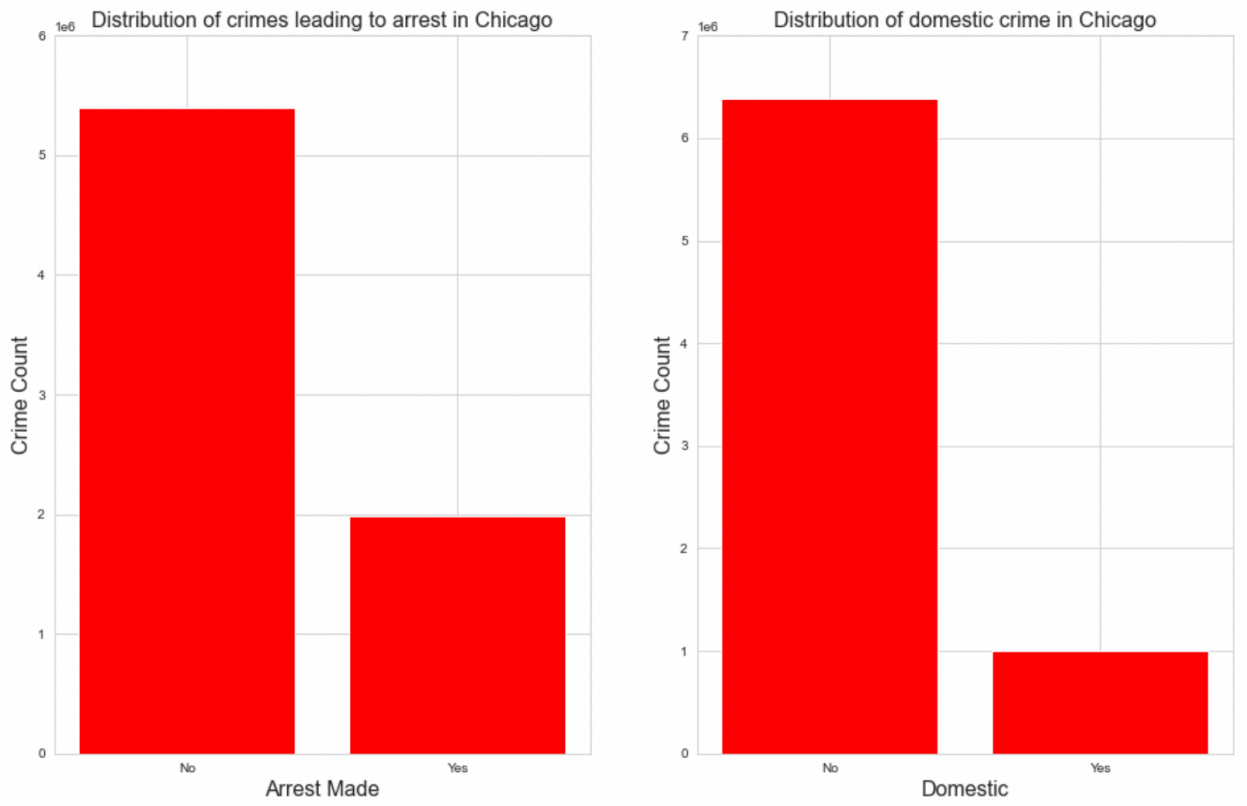

Figure 6. Percentage of domestic criminal activities and percentage of crimes leading to an arrest

Preprocessing

Feature engineering was performed to create a number of useful features for the analysis. First, a column was added to the dataset indicating the day of the year when the criminal activity had happened. Day of the year was then converted into two sine and cosine features to show how different years were associated (Fig. 7). A column for the day of the week of the criminal activity was created as well and two sine and cosine features were engineered similarly to those for day of the year (Fig. 8). Next, a feature was created to indicate whether or not a criminal activity had happened on a holiday. Finally, the number of criminal activities on each date were counted and added as a column to the dataset. Thus, the dataset for modelling included 6 features: number of criminal activities on each date as the target feature, and sin/cosine of day of week, sine/cosine of day of year, and holiday (or not holiday) as the other features.

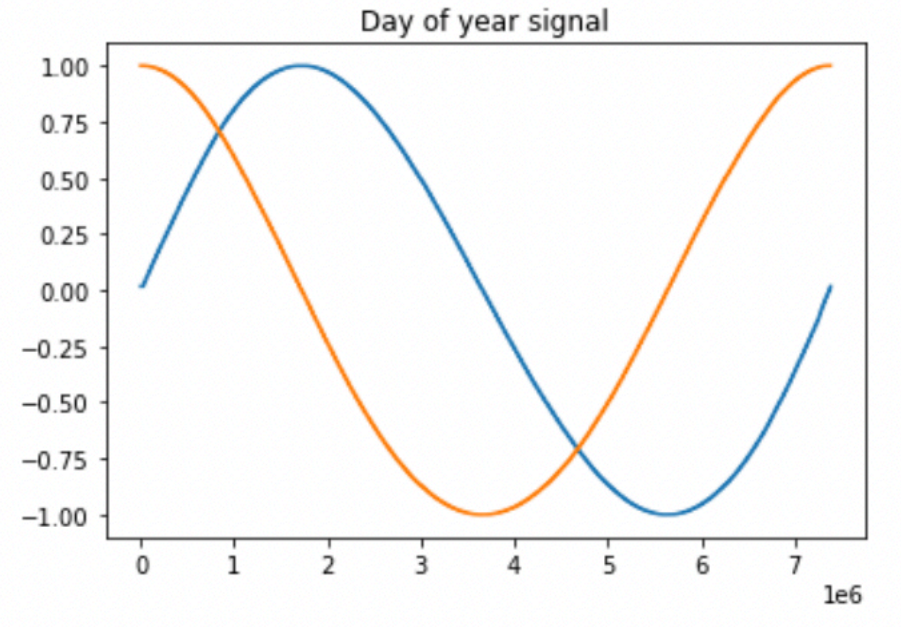

Figure 7. Day of the year signal

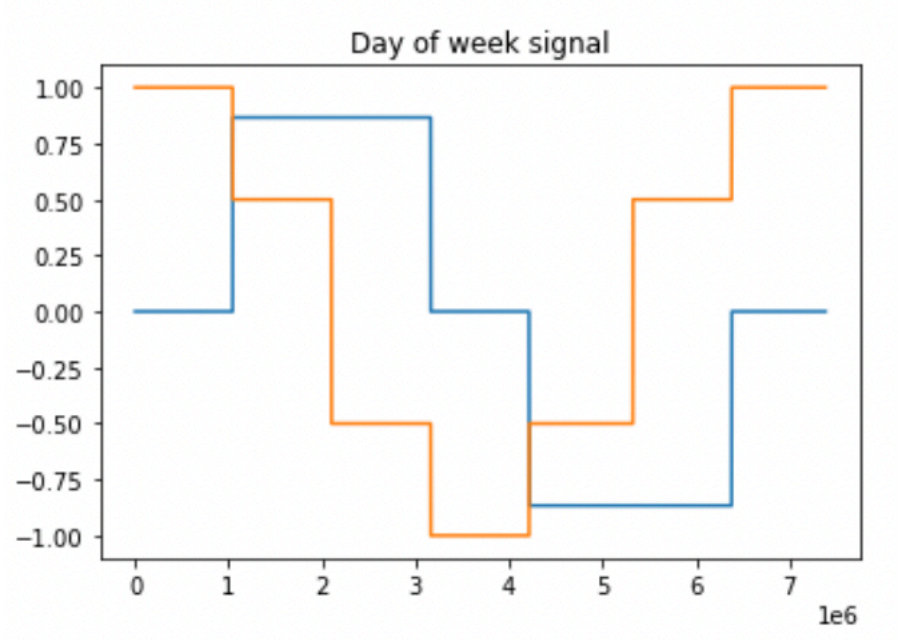

Figure 8. Day of the week signal

Augmented Dickey-Fuller test was performed on the target feature (number of criminal activities on each day) to identify whether the data were stationary. Since the p-value of the test was quite high (0.4), the null hypothesis that the data were non-stationary could not be rejected. Therefore, difference between number of criminal activities for successive days was computed and added to the dataset to be used as the alternative target feature. The p-value for the new target feature through augmented Dickey-Fuller test was very small (2.4e-29) and thus the data were considered to be stationary.

Data were then normalized and split into training (70%), validation (20%) and test (10%) sets. Validation set was used during training to tell the model when to stop and the test set was used to analyze the performance of the model for data it had never observed.

Modelling

Modelling was performed by trying different time windows. For this purpose, a window generator class was defined to sample batches of data from training, validation, and test sets for training and testing the model. The main model for the analysis was a Long Short_Term Memory (LSTM) neural network model. LSTM models have proven to be very useful for time series analysis. Given the limited computational capacity available, we tried only one particular number of batches as well as window width for modelling. We used 32 batches of a window size of 120 for inputs and 60 for labels. We then performed a manual grid search using different parameters of our LSTM model. Performance of different models of the grid were recorded and the best performing model(s) was identified. While the best performing models on the validation and test set batches were not the same, we chose the model with the best overall performance on the training, validation, and test set batches as our selected model (Fig. 9).

In order to analyze the performance of the selected model further, we performed modelling using the well-known Facebook Prophet time series analysis API as well and compared its performance against the selected LSTM model. To do so, we created six windows each with 120 days as inputs (historical crime rate and the other 5 features) and 60 days as labels (crime rate to be predicted). The six prediction windows covered the last 360 days of data in the original crime rate dataframe. Facebook Prophet only requires historical crime rate data as inputs and does not

use the other features. However, the entire historical crime data (before prediction period) were used for training the Facebook Prophet model for each of the six prediction windows.

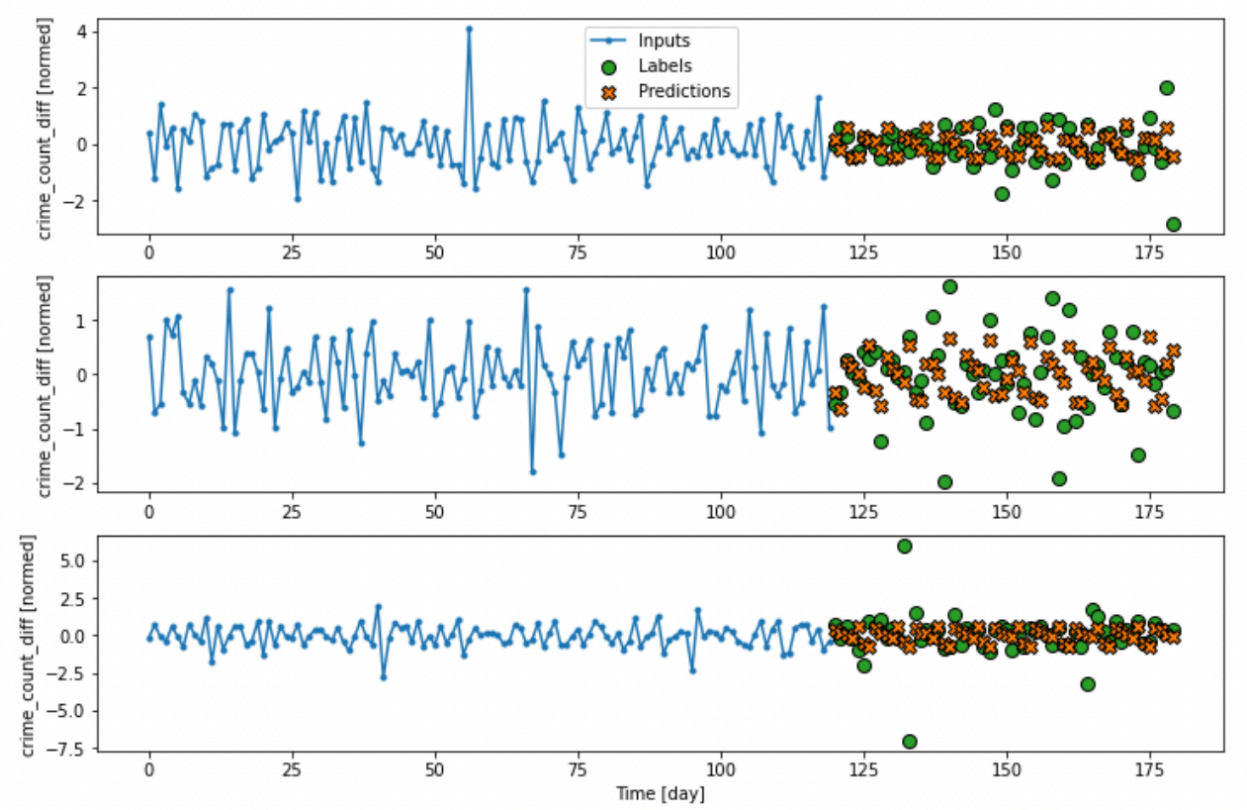

Figure 9. Model performance on three of the training batches

The LSTM model proved more adept at forecasting the future crime rates than the Facebook Prophet model for four of the six prediction windows (Fig. 10 and 11). Overall, the LSTM model proved more reliable in forecasts with an average mean absolute error of 7.78 less than the Prophet model (Table 1).

Time Period	MAE for LSTM	MAE for Facebook Prophet
1	68.646848	78.217711
2	75.028736	70.562665
3	54.398343	55.233943
4	53.458665	46.158135
5	48.034061	57.068427
6	52.550800	91.567163

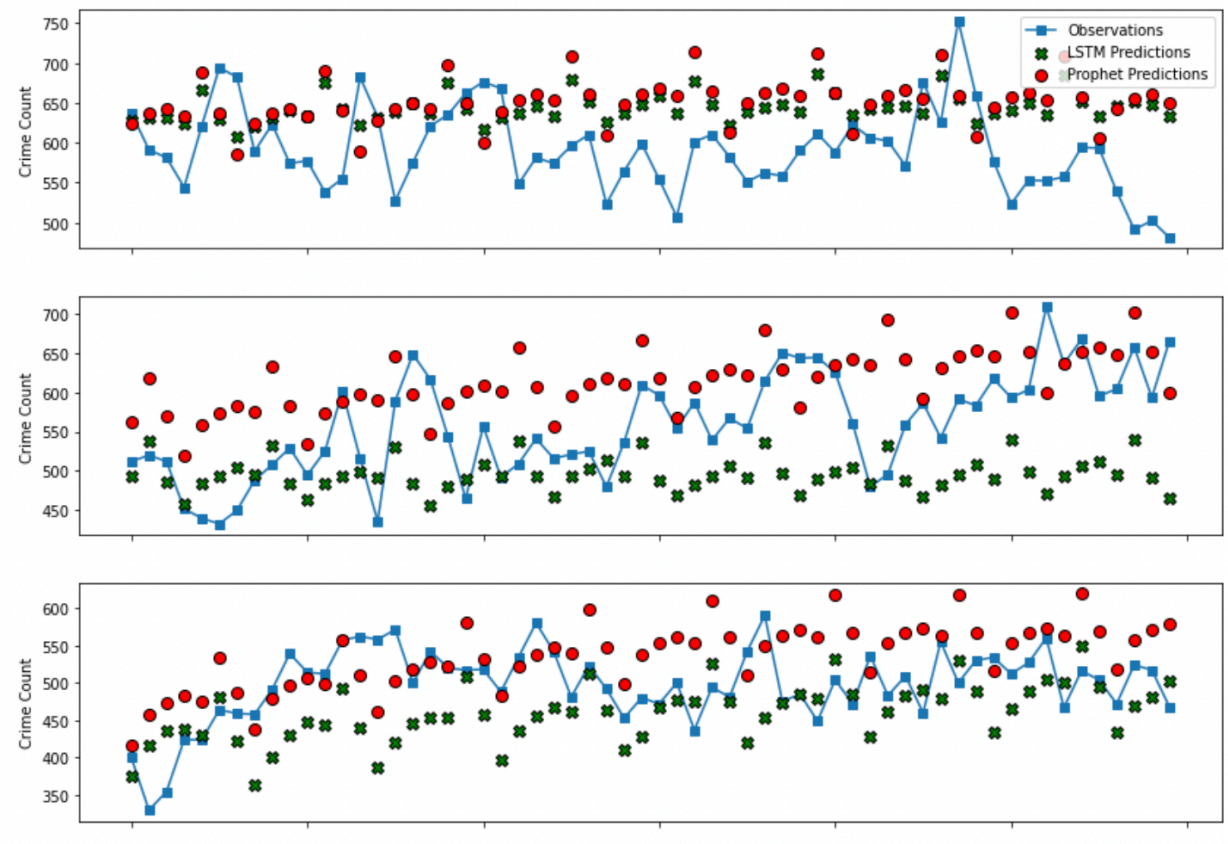

Figure 10. Performance of the LSTM and Facebook Prophet models in forecasting crime rate

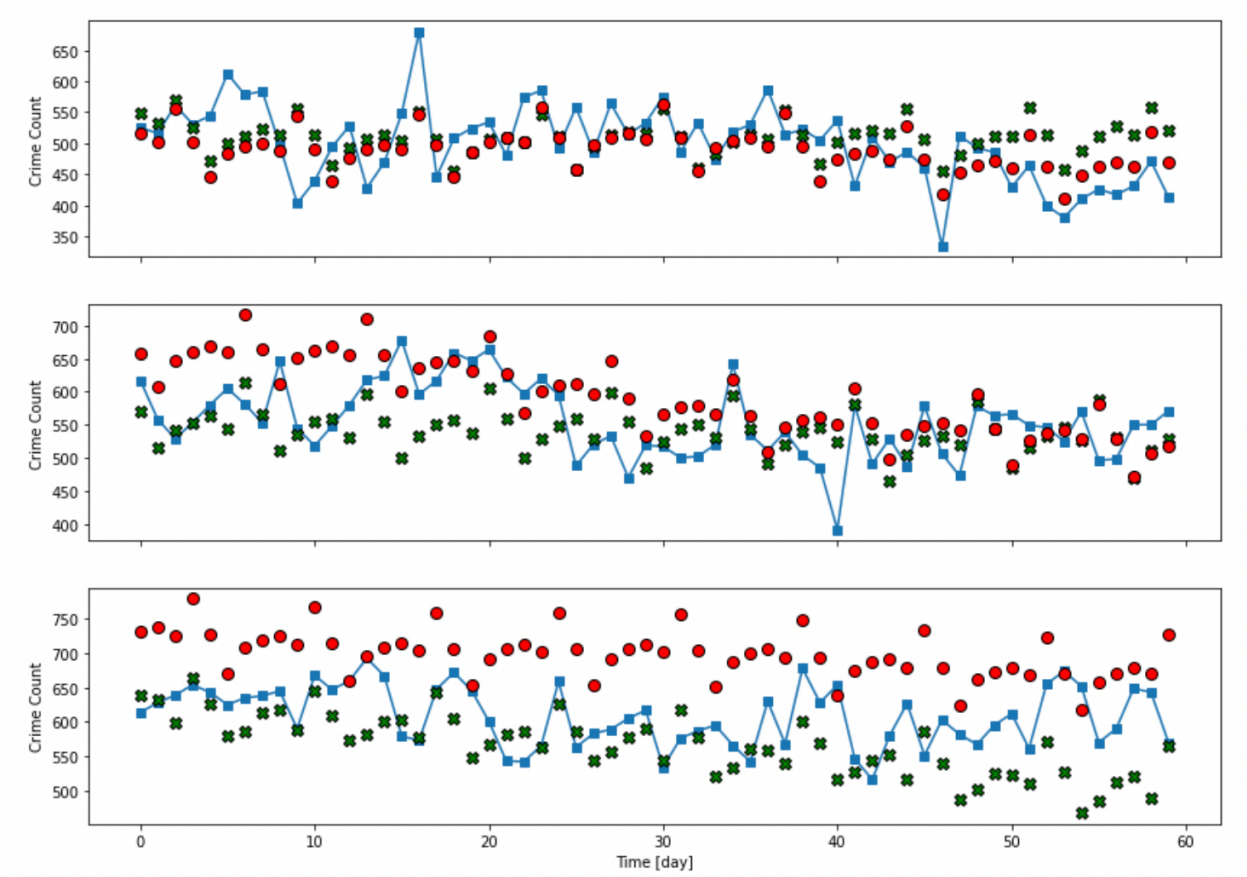

Figure 11. Performance of the LSTM and Facebook Prophet models in forecasting crime rate

Recommendations

1. The selected LSTM model can be used to forecast future crime rates for any length of time. However, given the computational capacity available, the model was fit to data batches of 120 days for inputs and 60 days for labels. Thus, the model has been tested for forecasting future crime rates for 60 day periods.
2. The selected LSTM model proves more reliable in forecasting future crime rates than Facebook Prophet. The performance of the model probably can even be improved further by applying a more rigorous hyper parameter tuning.

Future Research

It would be very interesting to perform a more rigorous grid search including trying out varying window sizes and see how the performance of the model improves. It would also be interesting to engineer some other useful features and analyze the model performance. One important addition could be through incorporating location information in the analysis and performing a neighborhood- or district-wise analysis.