Chicago Crime Forecast

Vijaya Lakshmi Pothula
50442307
vpothula@buffalo.edu

Rohit Basamgari
50442133
rohitbas@buffalo.edu

Sai Deep Juluri
50442134
saideepj@buffalo.edu

Statement Problem: Chicago being one of the most popular cities over, the years the criminal activity in such a prominent city has to be analyzed in order to predict when a crime will be committed.

I. Introduction

One of our society's most significant and pervasive issues is a crime. Numerous crimes are perpetrated often each day. In this project, the criminal activities reported are based on several circumstances such as theft, murder, assault, homicide, burglary, etc. These crimes happen in small villages, towns to big cities.

In order to curtail this issue, we require the power to predict the trends of crimes and to analyze the trend so that we can predict the criminal activities have occur. We can also find how crime rate has changed in Chicago over the years.

II. Dataset Description

This dataset contains the reported criminal activity that took place in Chicago in the year 2020. It is classified under the category of Public Safety. The CLEAR system of the Chicago Police Department is used to extract data. The dataset is a Time Series specific through which we can forecast. The dataset contains 17 features and up to 200,000 records. Among the crime cases we are concerned about thefts that have happened.

III. Theft Class variation

We are highlighting theft class because we notice that theft activities have more frequency among the others from the bar plot.

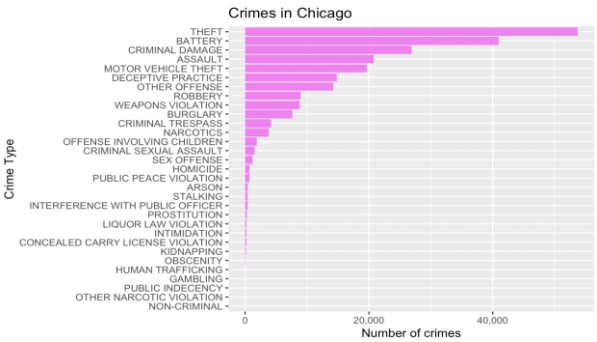

Fig.1 Bar plot showing the number of crime cases

So, we are going to use theft class to forecast in the city Chicago.

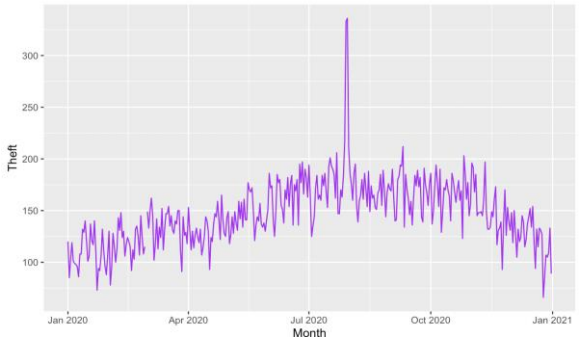

Fig.2 Theft variation monthly

This autoplot in fig.2 is between date and time which choose the trend of the theft happened in Chicago. The trend is increasing till the August month, thereafter we can observe decreasing trend. Moreover, from the plot we can infer that at the month of august we can observe most number of theft cases.

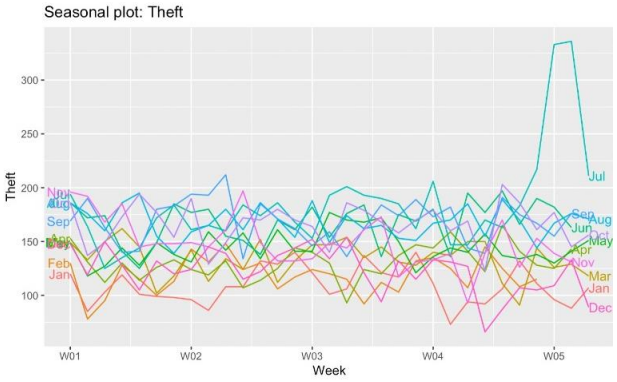

Fig.3 Theft variation weekwise

The above fig.3 is a autoplot is between weeks and theft, we can notice that at the start of week 5 of July has maximum theft case.

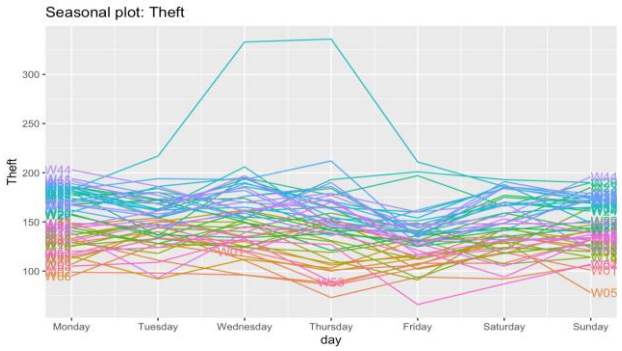

Fig.4 Theft variation daywise

The plot from fig.4 is a seasonality plot which plots the seasonality against date. Here each curve represents each week of all the months. From this plot we can infer that the data is having the seasonality, this can be observed by the similar patterns of each week curve.

IV. Methods

Time Series Forecasting helps us to analyze and forecast or compute the probability of an incident, based on data stored with respect to changing time. In a nut shell, time series forecasting involves forecasting and extrapolating future trends or values based on old data points, clustering them into groups, and predicting future patterns. We consider the events of theft w.r.t to date as the measure that repeats at a specific frequency. We have used

- Mean
- Naïve
- Seasonal Naïve
- Drift and
- ARIMA models to forecast the trends of theft.

V. Modelling

- In mean method the forecasts of all future values are equal to the average of the historical data.
- The naive approach considers what happened in the previous period and predicts the same thing will happen again.
- Seasonal naive approach is beneficial for highly seasonal data. In this scenario, we set each forecast to be equal to the last observed value from the same season.
- Model drift represents a change in a model's predictions over time. Prediction drift also represents a shift in predictions from new values compared to pre-produced predictions.

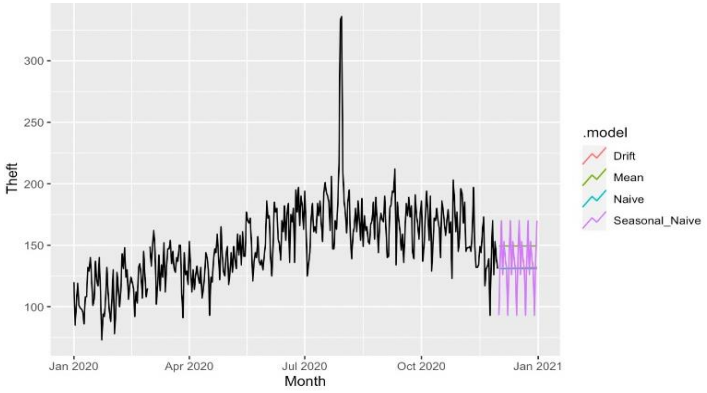

Fig.5 Theft forecasts from Dec to Jan 2021

Results: RMSE values

	Train set	Test set
• Mean	31.48564	22.20629
• Naïve	24.76405	33.26289
• Seasonal Naïve	29.63292	21.88091
• Drift	24.76403	34.59955

ARIMA Model:

ARIMA makes use of lagged moving averages to smooth time series data. They are frequently utilized in technical analysis to predict upcoming asset price trends. Autoregressive models implicitly presume that the future will mimic the past.

p=order of the autoregressive part.
d= degree of first differencing involved.
q=order of the moving average part.

These are the p,d,q=(1,1,0) values that are determined.

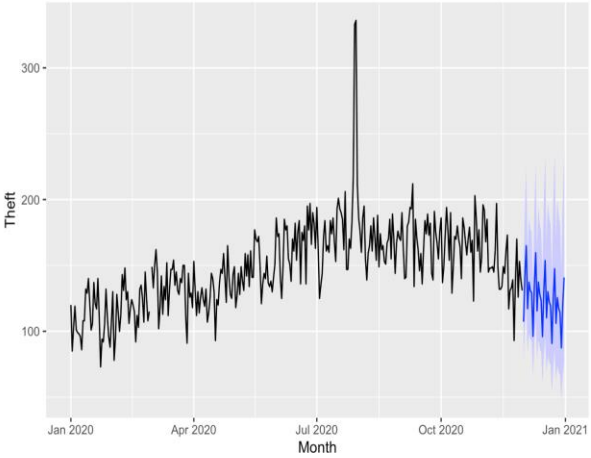

Fig.6 ARIMA model now forecasts theft activity from Dec to Jan 2021

Result:

RMSE for train set: 23.60133

RMSE for test set: 24.88797

VI. Conclusion

- The plots forecasts that the theft activity trend is decreasing.
- From the results we infer that **ARIMA** model gives us the best result because it produces the lowest RMSE values. A model with least RMSE value gives the best prediction.
-

VII. References

1. *Dataset source:* *https://data.cityofchicago.org/api/views*
2. *ARIMA:* *https://otexts.com/fpp3*
3. Professor Nikolay A Simakov lecture slides.
4. *http://r-statistics.co/Time-Series-Analysis-With-R.html*.

Github_link:*https://github.com/vijaya-ally/Group-13-Chicago-crime-forecasting/blob/main/Project%20(1).Rmd*