Crime Rate Prediction using fbprophet

Saamiya Newrekar [1]

[1] *Student, Department of Information Technology, Vidyalankar Institute of Technology, Mumbai, India*

Abstract— *Cities like Chicago have had high crime rates for a long time now. Understanding the trends and basis of which may help us to reduce the same in the future. Fbprophet[1] is open-source software released by Facebook which is highly accurate and fast, fully automated with a tunable forecast. Fbprophet is robust in the sense that it handles missing data, and dramatic changes in the dataset well and is efficient in handling outliers. Statistics show it outperforms almost any other approach when it comes to forecasting and making accurate future predictions. Using the fbprophet library, we can forecast any number of days in the future. For this project, we've forecasted 730 days; that is 2 years into the future. The primary purpose of the paper is to develop an accurate and robust fbprophet model to predict the crime rates of Chicago and map out the trends to help better understand which crimes are committed the most and during which time of the year. The underlying trends, along with the expected highs and lows have been noted and seaborn library is used to visualize the outcomes.*

Keywords—fbprophet, prediction, trends, machine learning, seaborn, forcasting.

1. INTRODUCTION

Crime rates are increasing in every part of the world daily. Arsons, burglaries, rapes, murder, domestic violence, etc are on the rise. Technologies along with their many advantages have had a sure hand in the recent increase in violent crimes these days. And technology is also the best weapon in our arsenal to prevent such violent uprisings. Even if we can't predict who exactly will be affected; by predicting the area, or the time at which maximum crimes take place, we can definitely control the number of crimes occurring and provide better prevention methods to promote safety among the people living in such high crime areas. In our paper, we discuss the various crimes committed in Chicago and focus on understanding the trends of the same to help make it a better, safer city for its residents.

Chicago ranks 42nd **[2]** in all of the United States for its crime index. It is considered one of the most unsafe cities in the USA. Knowing how many crimes were committed, why the crimes were committed, around what time of the year, and where they were committed could help us to prevent similar situations in the future. We use the fbprophet library to make a machine learning model for predicting the crime rates in the future. As mentioned in the abstract the fbprophet library developed by Facebook is known for its high accuracy, robust nature, full automation, and tunable forecasts. Forecasting is a data science tool**[3]** for goal setting, planning, allocating resources, and preventing the risks which can be understood by studying the trends. One of the areas where fbprophet shines is with human-scale seasonal data and crime rates committed by humans is one such example of the same. The dataset used for our prediction model is incredibly compatible with how fbprophet works, making it an ideal solution to our given problem statement.

2. PROPOSED METHODOLOGY

In the previous sections, we identified the goal of our project and the need for this research. The methodology includes the following steps-

2.1. Data Collection

The dataset used for this prediction is obtained from the data world website. The dataset consists of 6017767 rows and 23 columns consisting of all the crimes committed from 2005 to 2017 spanning across 12 years.

2.2. Data Preprocessing

The real-world data collected from various sources is incomplete, inconsistent with noise. To provide uniformity and improve the accuracy and performance of the machine learning model, the dataset needs to be cleaned, parsed, and standardized.**[4]**

The various steps involved in data preprocessing are as follows-

2.2.1. Data Cleaning

Missing or noising data is required to be cleaned or it would result in inconsistency and low accuracy. Two ways of handling missing data are- the tuples with missing values are ignored or the missing values are filled using the mean, median mode of the respective column. Noisy data on the other hand is smoothed out using regression or clustering techniques.

In our dataset, we use `error_bad_lines=False`

when we import the dataset. This ignores all missing values.

2.2.2. Data Transformation

Data needs to be transformed in appropriate format to make data mining techniques effective. One such technique used for transforming data is normalization i. It is used to scale the data values in a specified range for example -1.0 to 1.0. Another technique often used is called standardization which standardizes the data values in the range -3 to 3

2.2.3. Data Reduction

Large volumes of data are difficult to handle and manage. Hence data reduction techniques are used to increase the storage efficiency and analysis of the dataset. The various techniques used for data reduction are dimensionality reduction, attribute subset selection, and data cube aggregation. Each technique reduces the shape of the dataset by selecting the most relevant or important features and discarding the features which do not have a direct relation to the dependent variable.

As we've seen, our dataset has 23 columns which can be referred to as features describing the data. The features namely are- Index, ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community, FBI Code, Coordinate, Y Coordinate, Year, Updated, Latitude, Longitude, Location. For training our model, we don't need to take 23 of these as our input, that would lead to more processing time and memory inefficiency. After carefully considering the dependencies to give a final accurate output, we decided to only consider 8 out of the 23 features mentioned above.

```
chicago.drop(['Index', 'Case Number', 'Case Number',
'IUCR', 'X Coordinate', 'Y Coordinate','Updated
On','Year', 'FBI Code', 'Beat','Ward','Community Area',
'Location', 'District', 'Latitude' ,

'Longitude'], inplace=True, axis=1)
```

2.3. Changing the datetime format to standard type

Fbprophet requires a specific type of format**[5]** for the datetime field. Data scraped from the internet may not be in the format required by fbprophet as well as pandas and hence it is important for it to be transformed in that specific format. The format required should be YYYY-MM-DD HH:MM:SS .

```
chicago.Date=pd.to_datetime(chicago.Date,
format='%m/%d/%Y %I:%M:%S %p')
```

2.4. Indexing and Resampling

Fbprophet requires the dataset to be indexed on the datetime field as it uses this for the basis of identifying trends and mapping them. The below code indexes the dataset on the datetime field which in our case is Date.

```
chicago.index = pd.DatetimeIndex(chicago.Date)
```

The pandas.DataFrame.resample() method is used for frequency conversion and resampling of time series data**[6]**. time series means data listed or indexed in time order. Resampling is used to generate a unique sampling distribution on the basis of the actual dataset. Using this, we can analyze the data based on the frequencies- weekly, monthly, semi-monthly, and quarterly.

```
chicago.resample('Y').size()  #resampling based on Year
```

```
Date
2005-12-31    455811
2006-12-31    794684
2007-12-31    621848
2008-12-31    852053
2009-12-31    783900
2010-12-31    700691
2011-12-31    352066
2012-12-31    335670
2013-12-31    306703
2014-12-31    274527
2015-12-31    262995
2016-12-31    265462
2017-12-31     11357
Freq: A-DEC, dtype: int64
```

Fig -1

Fig. 1 shows how the data looks when it is resampled on yearly frequency.

```
chicago.resample('M').size()  #resampling based on Month
```

```
: Date
  2005-01-31    33983
  2005-02-28    32042
  2005-03-31    36970
  2005-04-30    38963
  2005-05-31    40572
                ...
  2016-09-30    23235
  2016-10-31    23314
  2016-11-30    21140
  2016-12-31    19580
  2017-01-31    11357
  Freq: M, Length: 145, dtype: int64
```

Fig -2

Fig. 2 shows how the data looks when it is resampled on monthly frequency.

2.5. Data Exploration and Visualization

Understanding the dataset is an important factor to be able to fine tune it as well as to analyze the data. In our project, we explored the date on one specific feature, namely- counting the number of crimes by their type,

```
chicago['Primary Type'].value_counts()  #counting the number of crimes by their type
```

```
THEFT                                1245111
BATTERY                              1079178
CRIMINAL DAMAGE                       702702
NARCOTICS                             674831
BURGLARY                              369056
OTHER OFFENSE                         368169
ASSAULT                               360244
MOTOR VEHICLE THEFT                   271624
ROBBERY                               229467
DECEPTIVE PRACTICE                    225180
CRIMINAL TRESPASS                     171596
PROSTITUTION                           60735
WEAPONS VIOLATION                      60335
PUBLIC PEACE VIOLATION                 48403
OFFENSE INVOLVING CHILDREN             40260
CRIM SEXUAL ASSAULT                    22789
SEX OFFENSE                            20172
GAMBLING                               14755
INTERFERENCE WITH PUBLIC OFFICER       14009
LIQUOR LAW VIOLATION                   12129
ARSON                                   9269
HOMICIDE                                5879
KIDNAPPING                              4734
INTIMIDATION                            3324
STALKING                                2866
OBSCENITY                                422
PUBLIC INDECENCY                         134
OTHER NARCOTIC VIOLATION                 122
NON-CRIMINAL                              96
CONCEALED CARRY LICENSE VIOLATION         90
NON - CRIMINAL                            38
HUMAN TRAFFICKING                         28
RITUALISM                                 16
NON-CRIMINAL (SUBJECT SPECIFIED)           4
Name: Primary Type, dtype: int64
```

Fig -3

Fig. 3 showing the various crimes committed and the number associated with each of them.

From the above mentioned 34 felonies, we chose the 15 highest ranking ones and visualized them using the seaborn library.

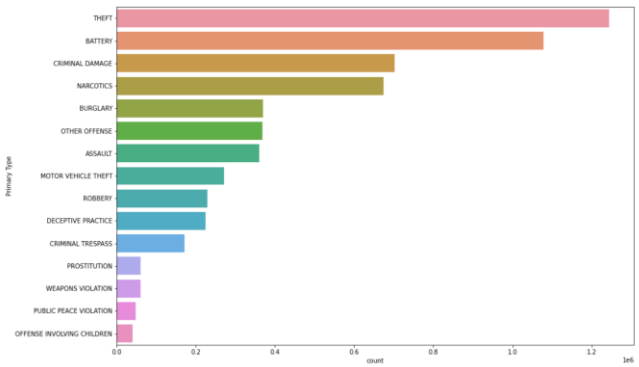

Fig -4

Fig. 4 shows the top 15 crimes committed in Chicago along with their respective count

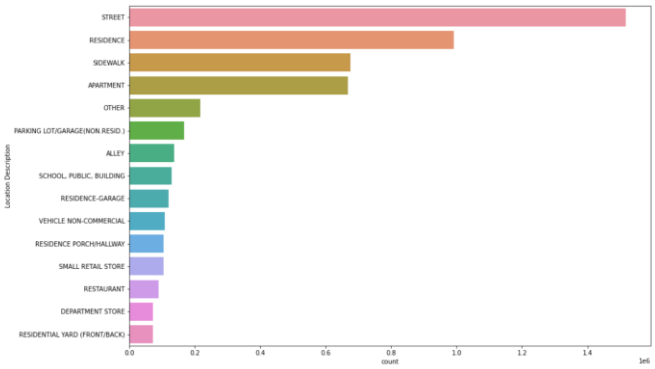

Fig -5

Fig. 5 shows the top 15 places where crimes were committed.

2.6. Preparing the dataframe for forecasting

The dataframe that is given as an input to fbprophet consists of only two columns- datestamp which is a date time field and measurement. which has to be a numeric field. Currently the column names are Date, 0 which needs to be changed to ds, y.

The steps for carrying out the transformation mentioned above along with the output is mentioned below. Fig. 6 is the current appearance and fig. 7 is the required form.

	Date	0
0	2005-01-31	33983
1	2005-02-28	32042
2	2005-03-31	36970
3	2005-04-30	38963
4	2005-05-31	40572
...	...	...
140	2016-09-30	23235
141	2016-10-31	23314
142	2016-11-30	21140
143	2016-12-31	19580
144	2017-01-31	11357

145 rows × 2 columns

Fig - 6

	ds	y
0	2005-01-31	33983
1	2005-02-28	32042
2	2005-03-31	36970
3	2005-04-30	38963
4	2005-05-31	40572
...	...	...
140	2016-09-30	23235
141	2016-10-31	23314
142	2016-11-30	21140
143	2016-12-31	19580
144	2017-01-31	11357

145 rows × 2 columns

Fig -7

2.7. Using fbprophet

We call the Prophet() function from the fbprophet library and fit our model using it. We can forecast any number of days in the future and for our project we have forecasted 730 days or 2 years in the future.

```
# Fitting the model.
model = Prophet()
model.fit(chicago_final)
# Forecasting into the future
future = model.make_future_dataframe(periods=730)
#forecasting two years into the future
forecast = model.predict(future)
```

3. RESULTS

When the data was resampled yearly, the distribution of the data points is shown below.

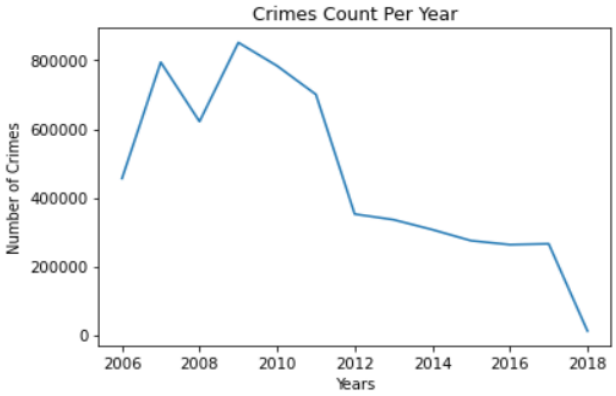

Fig -8

Fig. 8 shows the yearly distribution of the number of crimes committed when data is resampled according to year.

From fig. 8 we can conclude that the overall crime rate is decreasing over time and it should be expected to decrease further in the future.

We will now see the results of using fbprophet and verify our initial conclusion of the crime rate decreasing over time.

Using fbprophet we have successfully predicted the crime rate of the next two years. The maximum as well the minimum value for each data point in the time series dataset is also shown in fig. 9.

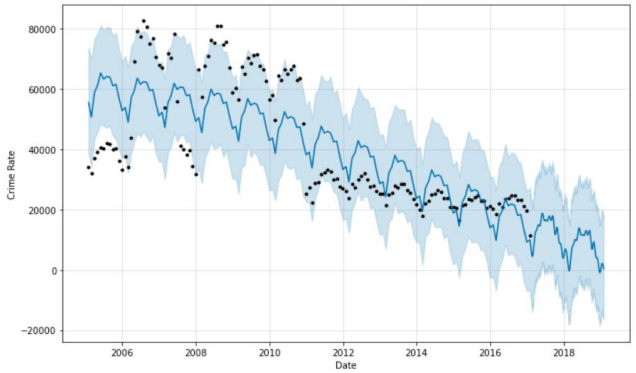

Fig -9

Fig. 9 shows the underlying trend of crime rates according to the month.

Looking at fig. 9 we can see the trends in crime rate with time. The highest was in 2005 and it is shown to be decreasing till 2017. Our forecast which was done 2 years in the future shows that for the years 2018 and 2019, the crime rate will fall resulting in Chicago becoming a safer city to live in as compared to 2005. If we see the datasets which include the crime rate from well before 2005, we will see that it was considerably higher than 2005. This confirms the accuracy of the fbprophet. We can also conclude from the above fig. that fbprophet handles outliers well and is adaptive, resulting in an overall excellent model for forecasting on time series data.

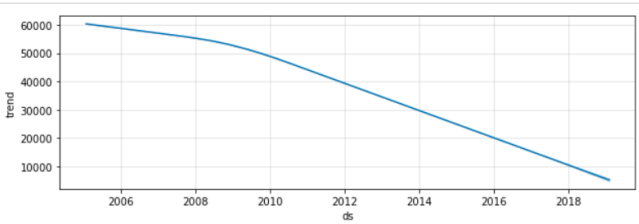

Fig -10

From Fig. 10 we can see that the crime rate has decreased over time and is expected to fall further.

4. CONCLUSION

In this paper we've explored how fbprophet works and predicted accurately the trend of the crimes committed in Chicago for the next two years, that is 2018 and 2019. From fig. 9 and fig. 10 it is clear that the number of

crimes committed in Chicago has decreased and is expected to fall further, which is a good thing since it means that Chicago is becoming safer with time. Our initial conclusion drawn from fig. 8 which says that the crime has decreased over time is verified by the fbprophet model, confirming its accuracy. As crime rates in Chicago are falling, we can further decrease it by implementing safety measures and making sure the felonies are reported properly. Fbrpophet is considered as a very convenient approach for making predictions on time series data and this paper verifies its claim. The model can be used to predict the crime rate for any other city in the world, given that the dataset is available on the internet. With the help of this model, we can know which crimes are committed the most, where they are committed the most and which need immediate attention so as to decrease the number of crimes committed in the future.

5. REFERENCES

[1]https://facebook.github.io/prophet/

[2]List of United States cities by crime rate - Wikipedia

[3]https://research.fb.com/blog/2017/02/prophet-forecasting-at-scale/.

[4]Data Preprocessing in Data Mining - GeeksforGeeks

[5]Quick Start | Prophet (facebook.github.io)

[6]https://medium.com/analytics-vidhya/time-series-analysis-a-quick-tour-of-fbprophet-cbbfbffdf9d8

[7]'Naresh Kumar', 'Seba Susan', COVID-19 Pandemic Prediction using Time Series Forecasting Models, 11th ICCCNT 2020 conference.

[8]'Devan M.S', 'Surya S. Gangadharan',Crime Analysis and Prediction Using Data Mining, 1st International Conference on Networks & Soft Computing, 2014.

[9]'Sean J. Taylor', 'Benjamin Letham', Forecasting at Scale, Facebook, Menlo Park, United States of America, 2017.