Crime prediction in Austin,Texas using FbProphet

Sruti Dutt
Department of Computer Engineering andTechnology
Dr. Vishwanath Karad MIT World Peace UniversityPune, India

Pradnya S. Kulkarni
Department of Computer Engineering and Technology
Dr. Vishwanath Karad MIT World Peace UniversityPune, India

***Abstract*— Crimes are treacherous and a common problem faced worldwide. Crimes affect the quality of life, economic growth, and reputation of a nation. The high prevalence of crime in modern times presents a significant problem for any city's police force. Every year, a vast amount of information about various criminal activity occurring in various locations isgathered and stored. Data analysis is vital if future predictions of comparable incident patterns and prospective solutions for solving and minimising criminal episodes are to be made.**

Then, it can be done utilising a combination of big data and several machine learning methods. There is a need for cutting-edge systems and fresh ideas for enhancing crime analytics in order to protect communities from criminal activity. We suggest a system that can analyse, find, and forecast different crime probabilities in a given area. Using different data mining approaches, this study describes various forms of criminal analysis and crime prediction. The purpose of this article is to use the Prophet Model and Random Forest Model to uncover patterns of prior crimes or the most common crimes in a specific location. The accuracy achieved using Fb Prophet Model is 92% whereas while using Random Forest Crime prediction can aid in preventing such recurring crimes and in detecting them faster. The r2 score of the random forest regressor is ~0.484 which is relatively low.

Keywords— Crime Detection; Crime Prevention; Big Data Analysis; Machine Learning

I. INTRODUCTION

Urban crime poses significant challenges for cities worldwide, affecting public safety, community trust, and effective governance. In recent years, the city of Austin, Texas, has witnessed dynamic shifts in crime patterns, necessitating the use of advanced analytical methods to support preventive strategies. Accurately forecasting future crime incidents can enhance decision-making for law enforcement agencies and city officials by enabling better deployment of resources and early intervention. Big Data is just an enormous amount of information that hasbeen gathered from numerous sources and may or may not be organised. Such a large amount of data may be difficult for older processing systems to handle. Big data analytics (BDA) combines a variety of methods and technologies to examine sizable, comprehensive data sets and get valuable insights from them. Our nation's population growth is causing crime to rise, which creates a vast amount of data that may be studied to help the government make vital decisions regarding the upkeep of law and order. This is becoming really critical as concerns about the crime rate grow. It aims to integrate the previously dispersed discourseon what constitutes enormous data, what characteristics define bulk data, and what tools and technologies are available to take advantage of the promise of bulk data.

Machine learning is a subfield of data science that deals with algorithms able to learn from data and make accurate predictions [1].

Machine learning is a critical component of the rapidly expanding field of data science. Algorithms are trained to make classifications or predictions using statistical methods, revealing key insights in data mining projects. These insights then influence decision making within applications and businesses, ideally influencing key growth metrics.

Crime consists of conduct that is in violation of federal, stateor local laws. When a law is broken, there is a penalty imposed. The penalty can include a loss of one's freedom oreven one's life. The types of crime that can be found most commonly are Drug Crimes, Street Crimes, Organized Crimes, Political Crimes, Victimless Crimes, White-Collar Crimes.

Thus study uses Austin crime dataset in order to analyze and predict crime. As of April 2021, murders in Austin are up significantly this year. According to the City of Austin Data, Austin’s population is currently 1,010,835. Five years ago, in 2017, itwas 967,629.[10]

This paper has been divided into 3 major sections namely; Literature Review, System Architecture and Working. Working is further subdivided into Data Collection, Data Preprocessing, Model Training, Results and Conclusion.

II. LITERATURE REVIEW

The purpose of the research by Menkudle et al [1] is to identify patterns of crime that occur frequently using knowledge discovery and its prediction. The LSTM Method and Prophet Model was used for future crime prediction. This work will be beneficial to local police stations in terms of crime suppression. Aarathi et al. [2] provides a thorough analysis of crime by integrating approaches, incidents, and their importance in literature. Sqoop is used for Data Migration, Hive and Map Reduce can be used for data analytics but over here it is found that Map Reduce performs better. It has been demonstrated by MINGCHEN et al. [3] that the Prophet model and Keras stateful LSTM perform better than Conventional Neural Network (CNN) models, where the optimal size of the training data is found to be three years. The Prophet Model is robust to missing data and shifts in trends. In paper [5], K-means clustering is applied, yielding crime hotspots. Then, a crime ratio matrix is constructed leading to the prediction of crime probability when subjected to a machine learning model. Here, crime monitoring is performed with the help of two methods: a. Crime transition probability computes the connection of one crime to another. b..Vulnerability of an area indicates safety of the area

Research by Chauhan et al [6] focuses on crime mapping using recorded data and cutting-edge technologies like R Tool, Hadoop, and Artificial Neural Networks. It consists mainly of three phases: Distribution of data geographically and creating clusters, Cluster analysis of created clusters and Prediction of crime. In study [7], LSTM and Stacked LSTM deeplearning models were build and used to predict the type of crime. The accuracy of crime prediction is used to compare the models. The outcome demonstrates that the Stacked LSTM outperforms the LSTM in terms of prediction accuracy anda better model can be found using this comparison.Paper [8] focuses on crime mapping using recorded data and validation technologiessuch as K-cross fold validation which helps us check thevalidation of the predictions given by the model. This validation method is easier to implement and gives an accurate result. Rony et al. [9] used Clustering algorithms and Linear regression to map out the crime rates of a specified area. This algorithm finds patterns and attributes that are similar to the already recorded crime andgives us a prediction based on that.

III. SYSTEM ARCHITECTURE

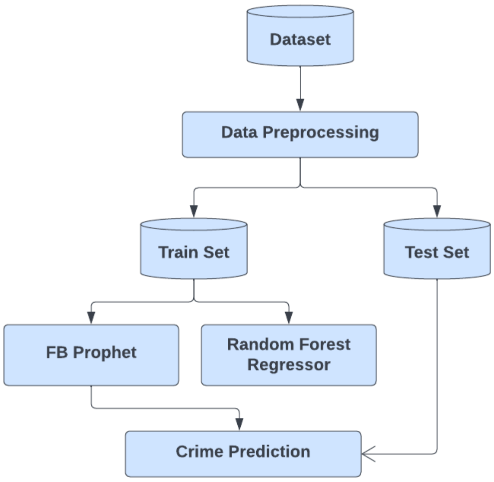

Figure 1 System Architecture Diagram

Figure 1 shows the proposed system’s architectural diagram. Below are two methods we used to predict the crime rate

a) Random Forest Method:

Random Forest is a classifier that contains a number ofdecision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and predicts the final output based on the majority votes of predictions. One of the most important characteristics of the Random Forest Algorithm is that it can handle data sets with both continuous and categorical variables, as in regression and classification. It outperforms other algorithms in classification problems.

b) Prophet Method:

Prophet is an additive regression model with a piecewise linear or logistic growth curve trend. It is a method for forecasting time series data that is based on additive models that match non-linear trendswith yearly, weekly, and daily seasonality, as well as the holiday impact.

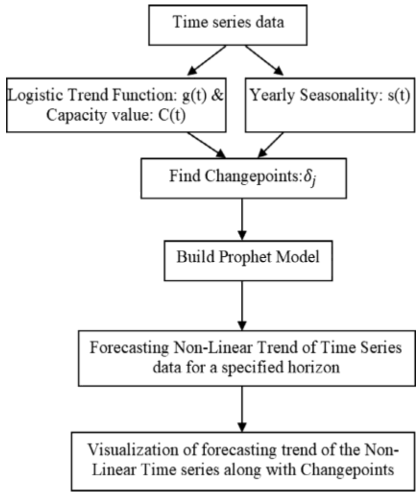

Figure 2 Flowchart for using FbPRophet

Prophet employs a decomposable model with three major components: trend, seasonality, and holidays, which are summarised below:

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

g(t) is the trend function to model non-periodicchanges;
s(t) is a function that represents periodic changes(e.g., weekly and yearly seasonality);
h(t) is a function that represents the effects ofholidays that occur on potentially irregular schedules and the error term represents any idiosyncratic changes that the model does notaccount for. Prophet performs best with time series with substantial seasonal effects and data from severalseasons.

Prophet is especially useful for datasets that:

- Contain an extended time period (months or years)of detailed historical observations (hourly, daily, or weekly) Have multiple strong seasonalities
- Include previously known important, but irregular, events
- Have missing data points or large outliers
- Have non-linear growth trends that are approachinga limit

IV. EXPERIMENTS

1) *Dataset:* A dataset from data. world which is called AUSTIN POLICE DEPARTMENT DATA DISCLAIMER is used for this study. Thisdataset has over 20,00,000 entries with columns such as Offense code, Offense description, Report time, Location, Zipcode etc.[4]

2) *Data Preprocessing:* The second stage is to use the crime dataset's input crime data and enhance the input data quality by using filters to remove extraneous noise. Various highlighted attributes are included in the crime dataset and In preprocessing step the unwanted attributes such as PRA, Census Tract, Clearance Status, Clearance Date, UCR Category, X- Coordinate, Y-Coordinate, Latitude and Longitude are removed.

Having unnecessary columns in the dataset will lead to inaccurate predictions. The null valued entries are also removed from the dataset, which reduced the number of entries to 16,54,528. Next, visualisations were made in order to understand the most prevalent crimes. Figure 2 provides the visualization depicting the share of top 10 crimes in Austin from 2004- 2018. Preprocessing steps were same for both the models PbProphet and random forest regressor.

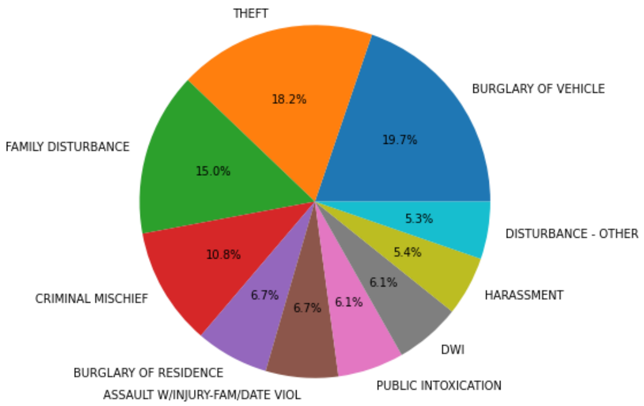

Figure 3Pie chart depicting the share of top 10 crimes in Austin from 2004- 2018

3) *Analysing their correlation:* Correlation heatmaps can be used to identify potential correlations between variables and to assess their strength. Correlation plots can also be used to find outliers and linear and nonlinear correlations. The cell colour coding makes it simple to detect correlations between variables at a glance. Correlation heatmaps are useful for identifying both linear and nonlinear connections between data. Correlation heatmaps for this project were made using the matplotlib and seaborn library of python. Figure 4 shows correlation heatmap for all columns

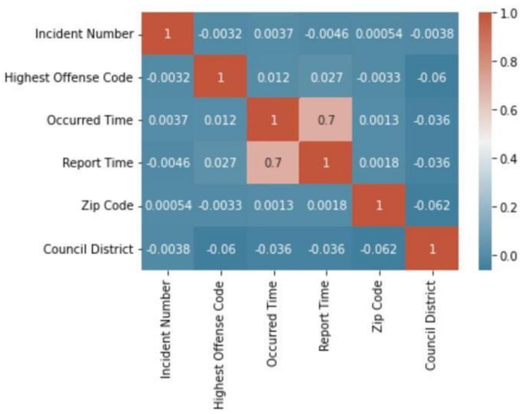

Figure 4 The correlation heatmap for Austin Crime Database

present in the Austin crime database. It depicts strong correlation between the occurred time and reported time attributes. Other attributes have very weak corelation

4)Model Training:

Random forest Regressor : For the Random Forest Classifier algorithm, the metrics for it has to be calculated from the confusion matrix as shown earlier. Precision implies the percentage of instances classified as positive thatare actually positive. Recall (True Positive Rate) is the ability of the algorithm to identify only the positive instances as positive. An ideal model should have a high Recall (True Positive Rate) and high precision for accurate predictions. The F-1 Score is the weighted harmonic mean of precision and recall values as calculated from the confusion matrix. For creating the confusion matrix, we imported the confusion matrix function from the "sklearn.metrics" library. For training the model we used the parameters n_estimators, and min_samples_split to ensure maximum accuracy. n_estimators is the number of decision trees we want the model to make before taking the average of the predictions to give a final output. We settled on 100 because it showed us the most optimal results without over-convoluting the model. min_samples_split parameter helps us to specify the number of records it should have before splitting it into 2 nodes. The default value for the same is 2 but due to the large number of records we have, we found 100 to give us the most optimal results. After specifying these parameters, we fit the train data into the model. Then we generated the importance each feature has on the predictions made by the model which can be seen in the graph below. For the final step, we tested the model with our test dataframe which gave us the following results.

FbProphet: For this part we extracted the date and time of the crimes and the number of crimes that happened on each day to feed into our FbProphet model. Then we changed their names to ds and y so that the model knows where to map each column. Further to analyse the working and accuracy of the model, we split this new dataframe into test and train dataframes and trained 2 different models. 1 modelwas trained on the entire dataset and the other was trained on the train dataframe and then was tested on the test dataframe. We then compared the findings from these models to analyse the accuracy of the model and it proved tobe ~96% accurate.

V RESULTS AND DISCUSSION

We discovered from EDA research that there is seasonality on a monthly and quarterly basis but not on an annual basis. If the time series is longer than two cycles, Prophet bydefault fits weekly and yearly seasonality. Using the 'add seasonality' method, users can add seasonality such as hourly, monthly, and quarterly.

Create a new Prophet object and call the fit method to train on the data to create a forecast. A confidence interval around the forecast is produced specifically by Note "interval width=0.95". Prophet approximates periodic signal using a partial Fourier sum. How quickly the seasonality canchange depends on the number of Fourier orders.

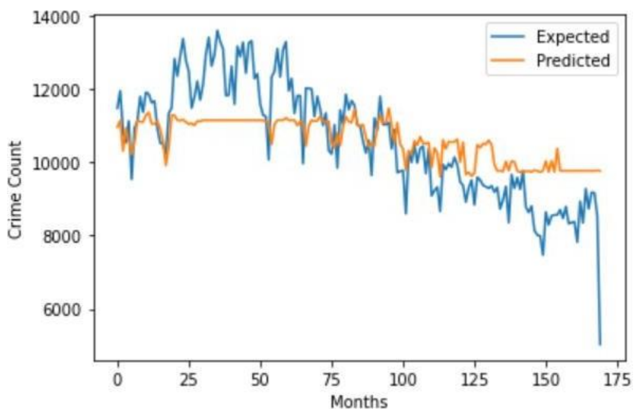

Figure 5 Line Graph depicting our Random Forest Regressor model comparing the predictions and the expected values

Figure 5 depicts Random Forest Regressor is not able to predict the crime accurately.

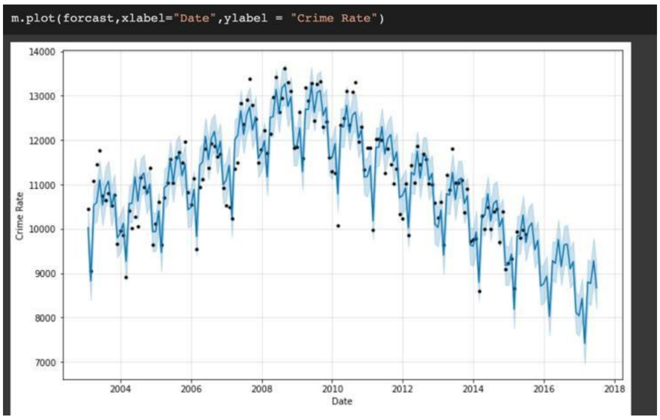

Figure 6 The results predicted by FbProphet, where deep blue indicates the prediction and the light blue area indicates a 5% error room

Figure 6 shows the prediction plot for the FbProphet model. The deep blue line represents model projections, and the black dots represent historical facts, as seen in the image. A 95% confidence interval around the forecasts is represented by the light blue shadow. The blue line displays a strong correlation with the prior pattern, indicating a reliable forecast based on historical data.

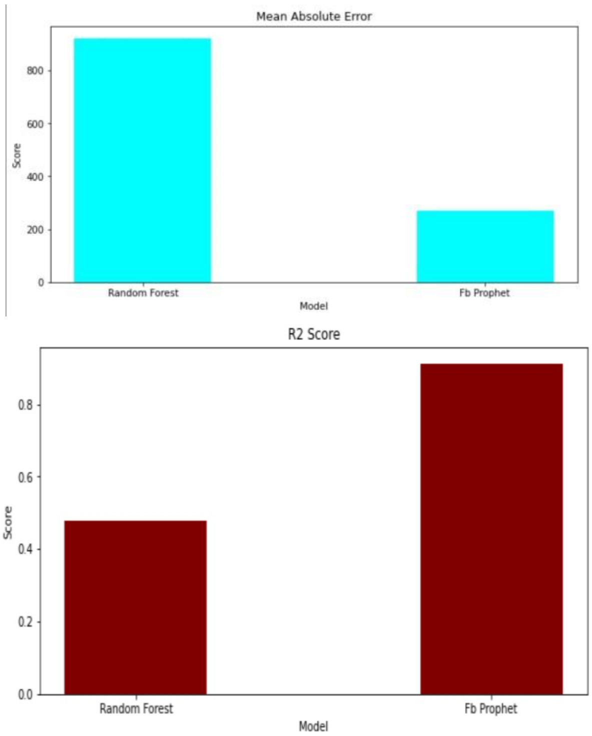

Figure 7 Comparison based on Mean absolute error and R2 score

Figure 7 shows comparative analysis of Random Forest Regressor and FbProphet based on Mean absolute error and r2 score. The r2 **score** is a statistical metric used to evaluate the accuracy of a regression model. It indicates how well the predicted values from a model approximate the actual data**.**

A higher r2 score indicates that the accuracy of our model ishigh and a mean absolute error of 269 where we are working with data that ranges in lakhs is also considered to be good.

VI CONCLUSION

This research highlights the use of FbProphet as a practical forecasting tool for analyzing and predicting crime trends in Austin, Texas. The model effectively identified recurring patterns and temporal fluctuations within the crime dataset, providing meaningful projections that could support proactive decision-making. Its ability to model seasonality and trends with minimal configuration makes it especially useful for time series problems in the public safety domain. It is observed from our experiments that FbProphet is able to achieve higher accuracy than the random forest model for crime prediction. Mean absolute error as well as R2 score show significant improvement as compared to random forest predictor.

Although the results are promising, this approach has limitations. Factors such as socio-economic conditions, policy changes, and real-time incidents were not included in the model, which could affect predictive accuracy. Future enhancements could involve integrating external variables and experimenting with hybrid models to improve performance and capture more dynamic influences on crime.

In summary, FbProphet serves as a valuable starting point for crime forecasting and opens the door for more advanced predictive systems aimed at enhancing community safety and strategic planning.

VII References

[1] Mayuri M. Menkudle, 2 Rachana S. Potpelwar "Big Data Analytics and Crime Patterns Detection and Prevention", Volume 8, Issue 9 September 2020.

[2] Aarathi Srinivas Nadathur, Gayathri Narayanan, Indraja Ravichandran, Srividhya.S, Kayalvizhi.J "CRIME ANALYSIS AND PREDICTION USING BIG DATA", Volume 119 No. 12 2018, 207- 211.

[3] MINGCHEN FENG 1, JIANGBIN ZHENG1, JINCHANG REN "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data", Received May 8, 2019, accepted July 9, 2019, date of publication July 22, 2019, date of current version August 15, 2019.

[4] [cityofaustin]. ([2019, March]). [Crime Reports], [2fa88b8f]. Retrieved [2022] from [https://data.world/cityofaustin/fdj4- gpfu/activity].

[5] Ashokkumar Palanivinayagam , 1 Siva Shankar Gopal , 1 Sweta Bhattacharya , 2 Noble Anumbe , 3 Ebuka Ibeke , 4 and Cresantus Biamba "An Optimized Machine Learning and Big Data Approach to Crime Detection", Volume 2021, Article ID 5291528

[6] Tirthraj Chauhan1,*, Rajanikanth Aluvalu2, "Using Big Data Analytics For Developing Crime Predictive Models"

[7] 1Anjana Ravi, 2,1Praseetha V.M "CRIME PREDICTION AND ANALYSIS USING BIG DATA", JETIR July 2021, Volume 8, Issue 7

[8] P. Kaur, G. Rani, T. Sharma and A. Sharma, "A Comparative Study to analyze crime threats using data mining and machine learning approach," 2021 International Conference on System, Computation, Automation and Networking (ICSCAN), 2021, pp. 1-4, doi: 10.1109/ICSCAN53069.2021.9526489.

[9] Rony, Sumon and Bakchy, Sagor Chandra and Rahman, Hadisur, Crime Detection Using Data Mining Techniques (January 21, 2021). Computer Science & Engineering: An International Journal (CSEIJ), Vol. 10, No. 5, October 2020.

[10] https://www.kvue.com/article/news/investigations/defenders/austin-crime-rate-population-growth/269-4bf6284e-6c23-45b4-9fdf-1f8328d30b64

[11] https://facebook.github.io/prophet/docs/quick_start.html