Crime prediction in Austin,Texas using FbProphet

Sruti Dutt
School of Computer Engineering and Technology
MIT World Peace University
Pune, India
sruti.dutt16@gmail.com

Atharva Patil
School of Computer Engineering and Technology
MIT World Peace University
Pune, India
atharvapatil128@gmail.com

Anushka Chattaraj
School of Computer Engineering and Technology
MIT World Peace University
Pune, India
anushkaxchattaraj@gmail.com

Tanvi Moholkar
School of Computer Engineering and Technology
MIT World Peace University
Pune, India
tanvi.moholkar@gmail.com

***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 is gathered 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

Big Data is just an enormous amount of information that has been 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 discourse on 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. As big data expands and grows, so will the market demand for data scientists, who will be required to assist in the identification of the most relevant business questions and, ultimately, the data to answer them.

Crime consists of conduct that is in violation of federal, state or local laws. When a law is broken, there is a penalty imposed. The penalty can include a loss of one's freedom or even one's life.

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, it was 967,629.[10]

The types of crime that can be found most commonly are Drug Crimes, Street Crimes, Organized Crimes, Political Crimes, Victimless Crimes, White-Collar Crimes. 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 paper 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.[1]

In this paper, we were provided with 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.[2]

Th predictive results show 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.[3]

In this paper, 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. As part of the proposed methodology, crime monitoring is performed with the help of the following methods:

- Crime transition probability computes the connection of one crime to another.
- Vulnerability of an area indicates how safe an area is.[5]

The proposed research in this paper 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. [6]

In this study, we build two LSTM and Stacked LSTM deep learning models and use them 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 and a better model can be found using this comparison. [7]

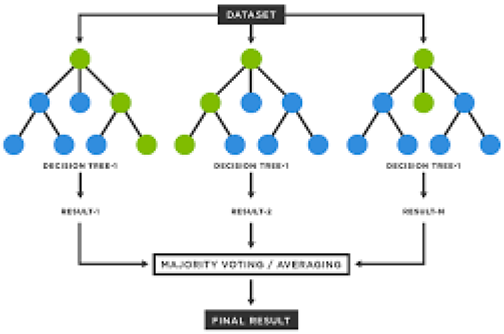
The proposed research in this paper focuses on crime mapping using recorded data and validation technologies such as K-cross fold validation which helps us check the validation of the predictions given by the model. This validation method is easier to implement and gives an accurate result.[8]

In this paper, this proposed methodology uses 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 and gives us a prediction based on that. [9]

III. SYSTEM ARCHITECTURE

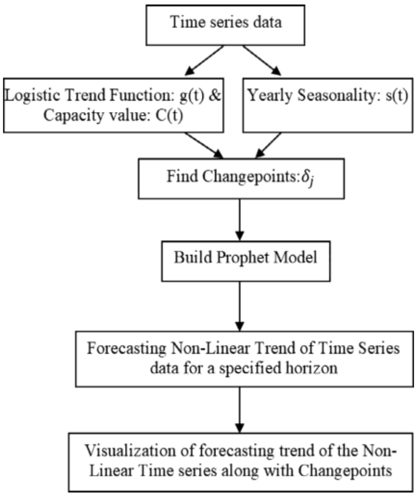
a) *Random Forest Method:*

Random Forest is a classifier that contains a number of decision 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 trends with yearly, weekly, and daily seasonality, as well as the holiday impact.


Time series data
Logistic Trend Function: g(t) & Capacity value: C(t)
Yearly Seasonality: s(t)
Find Changepoints: δ_j
Build Prophet Model
Forecasting Non-Linear Trend of Time Series data for a specified horizon
Visualization of forecasting trend of the Non-Linear Time series along with Changepoints

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-periodic changes;
s(t) is a function that represents periodic changes (e.g., weekly and yearly seasonality);
h(t) is a function that represents the effects of holidays that occur on potentially irregular schedules and the error term represents any idiosyncratic changes that the model does not account for.
Prophet performs best with time series with substantial seasonal effects and data from several seasons.

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 approaching a limit

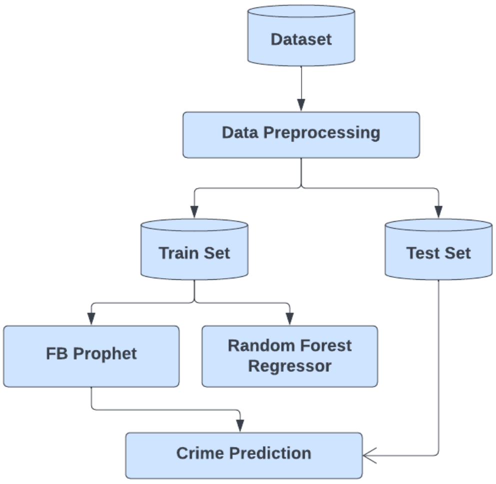


Fig 1: System Architecture Diagram

IV. WORKING

1) *Data Collection:* The first stage is to collect a suitable crime dataset which has substantial amount of entries to train and test our model. We have used a dataset from data.world which is called AUSTIN POLICE DEPARTMENT DATA DISCLAIMER[]. This dataset 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 we need to process the data by removing unwanted attributes such as PRA, Census Tract, Clearance Status, Clearance Date, UCR Category, X-Coordinate, Y-Coordinate, Latitude and Longitude.
The dataset that we collected had 21,75,274 entries and 27 columns. We first dropped the columns that were not important for our model such as X-co-ordinates and Y-co-ordinates of the crime scene, crime code etc. Having unnecessary columns in the dataset will lead to inaccurate predictions. To increase the efficiency of our model we dropped the less relevant columns. After that we dropped the null valued entries from our dataset, which reduced the number of entries to 16,54,528. Next, we extracted the top 10 crimes and top 10 crime locations and made visualisations on it to understand the most prevalent crimes. We also changed the datatype of our date and time record to the python DateTime64 datatype so that it complies with the model.
The data pre-processing was similar to what we have done for the FbProphet algorithm, hence we used the same cleaned and pre-processed dataframe for this algorithm as well.

We can visualize crime by various visualization techniques like bar graphs. In these techniques we take into account the previous crime data and analyze the data after which we can make several predictions.

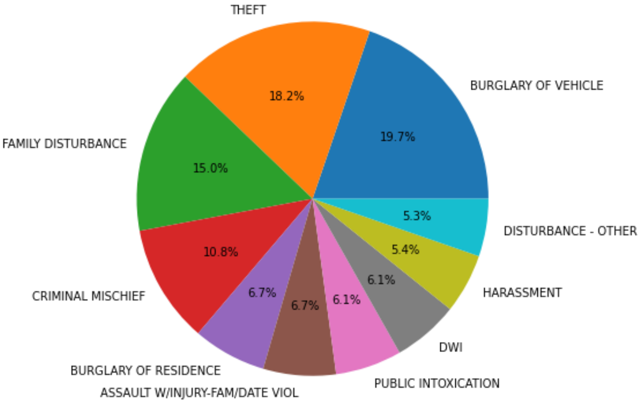


Fig 2: Pie chart depicting the share of top 10 crimes in Austin from 2004-2018

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.

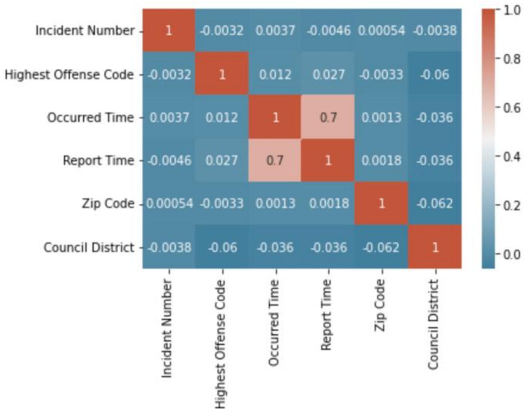


Fig 3: The correlation heatmap for all the columns present in our dataframe and their relevance

6)*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 that are 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 model was 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 to be ~96% accurate.

7)Results:

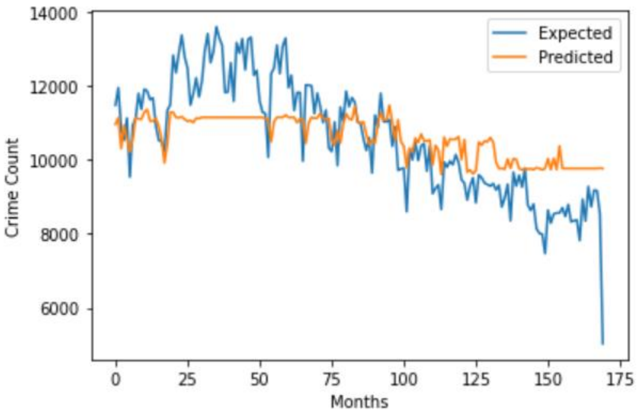
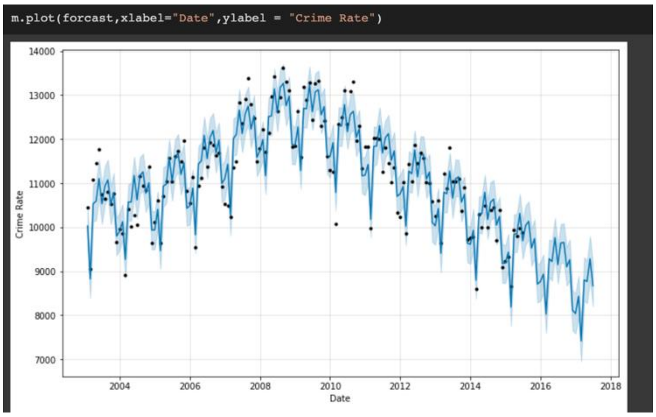


Fig 4: Line Graph depicting our Random Forest Regressor model comparing the predictions and the expected values

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 by default fits weekly and yearly seasonality. Using the 'add seasonality' method, users can add seasonality such as hourly, monthly, and quarterly.

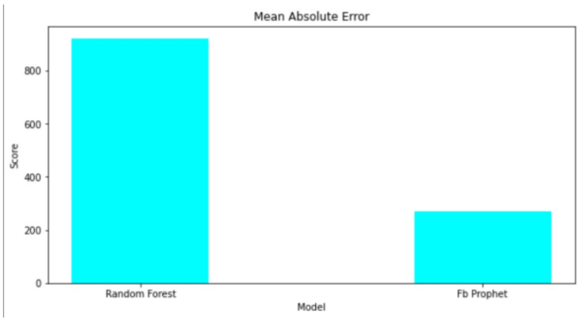
Create a new Prophet object and call the fit method to train on the data to create a forecast. A confidence interval around the forecast is produced specifically by Note "interval width=0.95". Prophet approximates periodic signal using a partial Fourier sum. How quickly the seasonality can change depends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

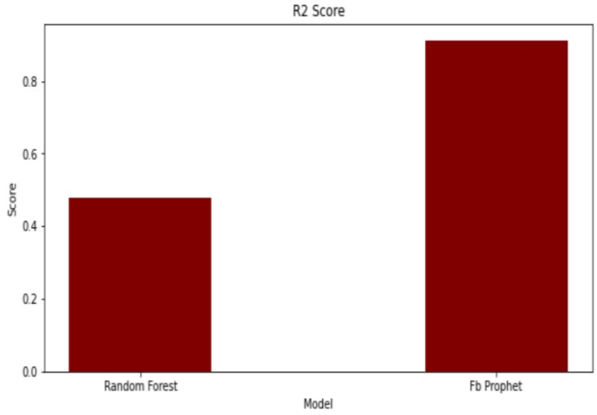
The deep blue line represents model projections, and the black dots represent historical facts, as seen in the image. A 95% confidence interval around the forecasts is represented by the light blue shadow. The blue line displays a strong correlation with the prior pattern, indicating a reliable forecast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



```
m.plot(forcast,xlabel="Date",ylabel = "Crime Rate")
```

Fig 9: The results predicted by the model, where deep blue indicates the prediction and the light blue area indicates a 5% error room





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

8)Conclusion:

Hence, due to a higher accuracy of FbProphet model, it is found to be more effective for predicting the results in case of problems related to time series forecasting and especially crime rate analysis and prediction.

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