International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

An end-to-end Novel Forecasting Model for Crime Prediction based on Big Data

M. Vinodhini, Dhruv Srivastava, Sameer Hirani, Shubham Arora

Abstract : ***Big data analytics is a field in which we analyse and process information from large or convoluted data sets to be managed by methods of data-processing. Big data analytics is used in analysing the data and helps in predicting the best outcome from the data sets. Big data analytics can be very useful in predicting crime and also gives the best possible solution to solve that crime. In this system we will be using the past crime data set to find out the pattern and through that pattern we will be predicting the range of the incident. The range of the incident will be determined by the decision model and according to the range the prediction will be made. The data sets will be non-linear and in the form of time series so in this system we will be using the prophet model algorithm which is used to analyse the non-linear time series data. The prophet model categories in three main category and i.e. trends, seasonality, and holidays. This system will help crime cell to predict the possible incident according to the pattern which will be developed by the algorithm and it also helps to deploy right number of resources to the highly marked area where there is a high chance of incidents to occur. The system will enhance the crime prediction system and will help the crime department to use their resources more efficiently.***

Key points: Crime data, Prophet model, Arima model, decision tree system.

I.I NTRODUCTION

Big data technique is the analysis of the data set which are big in sizes the mathematical, statically approach is made to analyse and extract the required information's. Today, big data analysis is been used by every company to understand there consumers what they are viewing what they are liking what all things they are not liking. We can use these data to make strategies to make profit for the company and can increase there consumer base. Big data can also be used by the department of police to predict the areas of crime in the region.

Manuscript received on March 15, 2020.
Revised Manuscript received on March 24, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
M. Vinodhini*, Assistant professor of Information technology in SRM IST,Chennai,Tamil Nadu,India.Email: chandraleka.t@rmp.srmuniv.ac.in
Dhruv Srivastava, Department of Information Technology,SRM IST,Chennai,Tamil Nadu,India. Email: srivastavadhruv999@gmail.com
Sameer Hirani, Department of Information Technology,SRM IST,Chennai,Tamil Nadu,India.Email: sameerhirani4287@gmail.com
Shubham Arora, Department of Information Technology,SRM IST,Chennai,Tamil Nadu,India.Email: Shubham.arora309@gmail.com

As we know with the growth in urban areas of the country will lead to increase in the rate of crimes such as signal breaching, robbery, cybercrimes and etc in the highly rated area of the region. But department is not having so much of resources to tackle these crimes. So we need some technical approachto control the crime and helpthe department to use their resources effectively.

So, to overcome this problem we can use the big data analytics to analyse the past data sets and predict the high areas where there is a possibility of crime happening. And in addition we can use the decision system i.e. "if-then" to make the best possible decision at the time of prediction. Decision system can use the past as well as the present data to make the judgement. The values can converted into the common index and that index value can be compared with the range value and can decide the range of the crime whether the range is high, medium, or low. Prophet algorithm is used in predicting time series data sets which are non-linear in nature and are arranged on the bases of years, months, holidays. Prophet algorithm is best used in handling the complicated data.

ARIMA model Auto regressive moving average is used in the time series data to figure out data or to predict the indications in the prediction. Auto regressive moving model is used in cases where data shows the non-stationarity in which we apply the differencing steps to remove the non-stationarity.

II. LITERATURE SURVEY

A. "A Data-driven Approach for Spatio-Temporal Crime Predictions in Smart Cities"

In the past 10 years there is a speedy growth in the urban regions of this country. The economic and social status of the area has increased very significantly. With this growth in urban region causes the increase in the crime rate of the area. Crimes involving loot, signal breaching and etc causing this growth a very negative effect. So, to control this negative effect of we can use the past data and can analyses and detect the highly effected crime area. We can use the combination of time series forecasting and autoregressive model for analyses and prediction. This approach can be consider as temporary as we cannot detect the level of crime and we can decide what kind of resources will be required in those areas. The authority can use this prediction to use their resources very effectively and can deploy more in the highly marked crime area.

B. "Classify Interval Range of Crime Forecasting for Crime Prevention Decision Making"

Predicting crime and locating the areas which are highly effected of crime activities is not enough to control the crimes. The authorities some system which can give them decision on the situation. So through the prediction of past data if we can provide the decision option by using the forecasting data and interval forecasting. Decision option will use simple "if-then" to decide level of crime. We can convert crime data into the index value and then make a judgement whether the crime is of high, medium or low level. This will help the authority to decide which region needs what kind of resources. The ranging will be pre decided and it will be set in the algorithm and when the algorithm has to make any judgement then the index value will be checked with the range value. Whatever range is detected then according to that authority will use their resources.

III. PROPOSED SYSTEM

The crime in our country has increased very rapidly traffic signal breaching, road accidents and accidents to control these kinds of incidents is very difficult job to do. As we all know that in our country we are not having so much of resources to stop or to control these incidents if it occur in many parts of the area. So we are proposing a system which will be used for crime prevention and help the authorities to use their resources more efficiently and effectively. This crime prevention system will provide information on the bases of crime prevention decision option.Each time the selection of decision in this system is based on the crime limit range and predicting values. Limit range is determined through decision frequency on the level of high, medium or low. These limits will be determined by the "if-then rules" in the decision process. This system can give the accurate location of the crime by managing large volumes of data. Advantages of this system

- System can analysis the similar data and multi sourced data.
- System can figure out the similarities among the incidents of same kind and can propose the prediction accordingly.
- It supports the time series modelling.
- It increases the forecasting limit by classification.

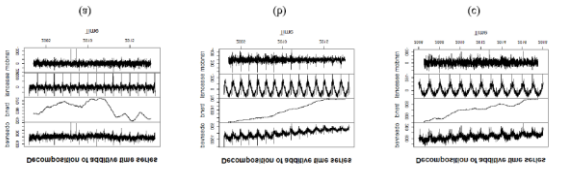

Fig: 1 Decomposition of effective time series

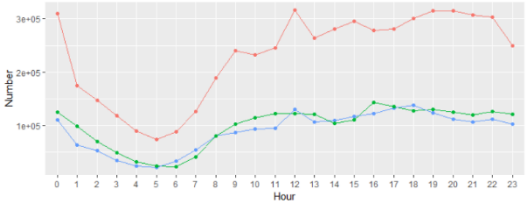

Fig: 2 Time Series graph

A. Prophet Algorithm

Prophet algorithm is used in the process of predicting the time series data which is a non-linear in nature and are arranged on the basis of years, months, holidays. Prophet algorithm is a time series algorithm which works best with the time series data. Prophet is best in handling the lost data and it manages outlier of the process very well. Prophet model is used in such a way that it can handle the complicated features of time series. The prophet model categories time series into three main parts which are trends, seasons, and breaks.

They are merged in the below equations

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \qquad (1)$$

"In this, g(t) is the trend of any non-periodic changes in the time series, s(t) represents periodic changes and h(t) represents the holiday effects of any potentially irregular schedules over one or more days. The error term 't represents any random effects which are not accommodated by the model. For the trend function g(t), we utilized a linear trend with limited change points, where a piece-wise constant rate of growth provides a parsimonious and useful model".

$$g(t) = (k + a(t)^T\delta)t + (m + a(t)^T\gamma) \qquad (2)$$

$$a_j(t) = \begin{cases} 1, & if\ t \geq s_j, \\ 0, & otherwise. \end{cases} \qquad (3)$$

"In this, k is the growth rate, m is the offset parameter, and is set to -*si*, s*j* to make the function continuous".

B. ARIMA model

Auto regressive moving average is used in the time series data to figure out data or to predict the indications in the prediction. Auto regressive moving model is used in cases where data shows the non-stationarity in which we apply the differencing steps to remove the non- stationarity.

In ARIMA, AR part of ARIMA indicates the regressive values of the variables in the series. The MA part of the ARIMA indicates the regressive error or it describe the linear combination error of errors of values which have occur repletely in the past. ARIMA process defines the drift with:

$$\frac{\delta}{1 - \sum \phi_i}.$$

IV. SYSTEM ARCHITECTURE

This system comprises four different steps to analyse and predict the data from the past data sets. The system uses the technique of data cleansing, data extraction and data processing.

First step is to adapt the data sets from the sources and then that data sets is been cleansed that is the data which is important for the process is been taken and remaining data is been removed. After that in the next step data is been pre-processed in which the corrupted or encrypted files is been removed and the data files is converted into the readable format and then the data which is generated is passed to next step. So, now we got the datasets in the readable format and know we have to start the analysis according to the information which is required from the past. For example, if we need data of signal breaching then all the data related to signal breaching will be extracted from the data sets and then that data will be analysed further.
The extracted data will be processed in this the present data will be entered and forecasting comparison will done which means the present data and the past data values will be combined and the final value will be determined and the value will be compared with the limit range then the crime level will be determine i.e., high, medium, low crime level after that the visualization of the value will be done points will be plotted in the graph through which the decision will be finalised and the final output will sent to authority to take the necessary step.

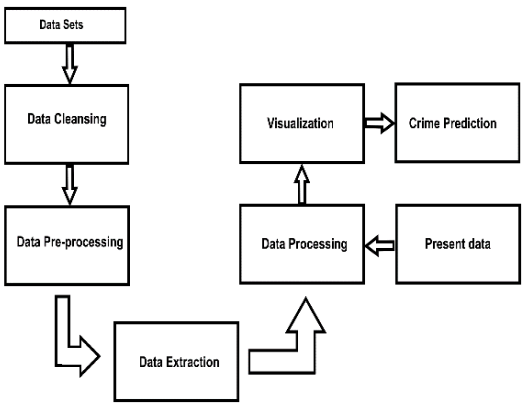

Fig 3 Architecture diagram

V. IMPLEMENTATION

Module 1 : Exploratory data analysis

Exploratory data analysis is a procedure of filtering through information and looking for data which is unique. Analyst current tools is used for interpret data which includes “database management systems, statistical analysis packages, data mining tools, visualization tools, and report generators”. Since this procedure uses data driven manner, which makes it flexibly which also helps in composing data at each phase of analysis. It utilizes a design that “monitors the mapping of visual items to data in shared databases”.

EDA mainly depends on visual and graphical translations of information. While statistical modelling provides a basic low dimensional which requires advanced knowledge of statistical methods and various “mathematical principles. visualization and graphical representation are much easier to understand and also to generate, which also helps in continuous exploring various aspects of a dataset”. The ultimate role is to generate simplest form of data that is easily understood. Besides this is not the end of data science pipeline, but still a significant one.

Module 2: Pre-processing

Data pre-processing is a significant advance to set up the information to frame a model. There are numerous significant steps in data pre-processing, for example, data cleaning, data transformation and feature selection. Data cleaning and transformation are techniques used to eliminate anomalies and arrange the information with the goal that they take a structure that can be effortlessly used to make a model. A data set may contain several factors (descriptors); notwithstanding, a significant number of these factors will contain unwanted and irrelevant data. So as to rearrange the model, it is essential to choose just factors that contain one of a kind and relevant data. data mining systems can be utilized to evacuate factors that don't add any factor to the model.

Before feature extraction and classifications is done a series of pre-processing steps are carried out.

Tasks in data pre-processing

Data cleaning: filling the empty value, identifying the smooth noisy data, eliminate outliers, and improve inconsistencies.

Data integration: using multiple databases, data cubes, or files for gathering various information.

Data transformation: Standardization and cumulation.

Data reduction: Eliminating the excess data and preparing the similar analytical results.

Data discretization: Subset of data reduction, supplanting numerical attributes with nominal ones.

Module 3: Feature Selection

Feature selection process can be differentiate into three types: filters, wrappers, and embedded/hybrid technique. Wrappers techniques perform better than filters strategies since feature selection procedure is enhanced with the classifier to be utilized. While, wrapper strategies have high value to be utilized for vast component space due to excessive computational cost and each list of capabilities must be guided with the prepared classifier that in the end make procedure moderate. filter techniques have less computational expense and quicker, but it is inefficient reliable in differentiate as measured with wrapper technique and better appropriate for lots of dimensional data collections. embedded strategies are as mixture filters and wrappers. As it is cross breed approach utilizes best circumstances of both filters and wrappers technique.

filters technique can be additionally classified into two sets, to be feature weighting programs and subset search programs as appeared above.Feature weighting programs appoint loads to feature independently and rates them dependent on their significance to the objective.

Feature extraction can also be utilized in eliminating complication and provide a uncomplicated representation of data and representing each element in feature as a sequential combination of genuine input
variable.

The most relevant and pre-owned feature extraction technique is Principle Component Analysis.

Module 4: Prediction

When training, each tree in an irregular forest gains from an arbitrary example of the data points. The samples are drawn with substitution, known as bootstrapping, which implies that a few examples will be utilized on numerous occasions in a single tree. The thought is that via training each tree on various examples, each tree may have high variance concerning a specific arrangement of the training data, in general, the whole forest will have lower variance but not at the expense of expanding the bias.

During examine time, the predictions are made by cumulating the prediction of each tree. This method of training each individual learner on various bootstrapped subsets of the data and then cumulating the prediction is known as bagging.

To classify another example, every decision tree gives a grouping to input data; random forest gathers the classification and select the most predicted forecast as the solution. The input of each tree is examined data from the pre-owned dataset. Also, a subset of feature is arbitrarily selected from the first features to improve the tree at one and all stage. Each tree is improved in the absence of pruning. primarily, arbitrary forest enables countless powerless or feebly connected classifiers to create a solid classifier.

VI. RESULTS

In the system we have used the data set of Chicago and firstly we have cleansed the unwanted data from the data sets. After that we have mined the data and we have extracted the required information. In the below figure we have extracted the block, primary type, description of the crime, location of the crime, arrest in the particular crime and location of the crime.

Date	Block	PrimaryType	Description	LocationDescription	Arrest	Domestic	Latitude	Longitude
01/01/2017 01:00:00 AM	056XX S MAYFIELD AVE	OTHER OFFENSE	HARASSMENT BY TELEPHONE	RESIDENCE	False	True	NaN	NaN
01/01/2017 01:00:00 AM	019XX N HAMLIN AVE	OTHER OFFENSE	HARASSMENT BY TELEPHONE	RESIDENCE	False	True	41.916211	-87.721638
01/01/2017 01:00:00 PM	023XX W 115TH ST	OTHER OFFENSE	OTHER VEHICLE OFFENSE	STREET	False	False	41.684460	-87.679939
01/01/2017 01:00:00 PM	021XX W ARMITAGE AVE	OTHER OFFENSE	HARASSMENT BY ELECTRONIC MEANS	OTHER	False	True	41.917689	-87.680898
01/01/2017 01:10:00 AM	075XX S DAMEN AVE	OTHER OFFENSE	PAROLE VIOLATION	STREET	True	False	41.757026	-87.673325

Table : 1 Crimes with description and location

Now, we have got the total crimes which happened in the area located and matches the above criteria. In the below figure the list of crimes is listed and mentioned which crime has how many number of times. So, at last we got total offense equal to 41.

```
                                    PrimaryType
Description
TELEPHONE THREAT                           2625
HARASSMENT BY TELEPHONE                    1861
HARASSMENT BY ELECTRONIC MEANS             1761
VIOLATE ORDER OF PROTECTION                1053
OTHER VEHICLE OFFENSE                       960
PAROLE VIOLATION                            599
FALSE/STOLEN/ALTERED TRP                    576
VEHICLE TITLE/REG OFFENSE                   415
LICENSE VIOLATION                           323
OTHER CRIME AGAINST PERSON                  259
...
Total Other Offense Crime Descriptions: 41
```

Fig 4: List of crimes

In the below figure, total crime description is visualised in the graphical in which the telephone threat is been marked highly dangerous.

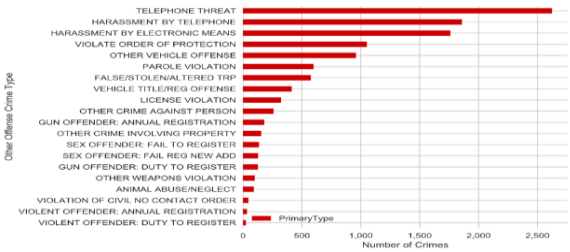

Fig: 5 Visualization of different crimes

VII. FUTURE ENHANCEMENT

Further advancement will stretch out to the displaying of increasingly point by point situations to encourage forecast dependent on itemized input wrongdoing. The point here was to remove a fundamental summed up model of wrongdoing occurrences. Notwithstanding, explicit areas best displayed freely of different information on explicit seasons. Despite the fact that there contemplations can be displayed independently if adequate measure of excellent information is supporting it. Just as numerous factual instruments can examine the remarkable occasions that will give more noteworthy knowledge to how these occasions change the degrees of wrongdoing and eventually characterizing the guidelines that will alter frequency tally fundamentally.

In future, we intend to finish our on-going stage for conventional huge information investigation which will be fit for preparing different sorts of information for a better use. We additionally join chart mining systems and fine-grained spatial examination to reveal progressively potential examples and patterns inside these datasets. In addition, we expect to lead increasingly practical contextual analyses to additionally assess the viability and adaptability of the various models in our framework.

VIII. CONCLUSION

Big Data allude as a job in changing crude information into significant choice emotionally supportive network for the law making body and legal executive to find a way to deal with the everyday violations and keep a check. In a present situation, the expanding populace and shooting crime percentages, it is of at most need to recognize the enormous wrongdoing informational indexes as to utilize them to distinguish patterns. Also the architecture proposed in this paper will reduce cost and efforts of the police department by patrolling the cops in area which are more sensitive and predictive for crime as per the analysis.

It assists with keeping up propriety in the general public by withstanding lawfulness. The precautious measures and additional security could be given territory explicit to focus on the wrongdoing around that bit of the region to guarantee wellbeing and prosperity. For ex-adequate in a specific region with a specific wrongdoing being overwhelmingly present, security could be authorized at the prime spots to guarantee exacting watchfulness by dispatching required police power contingent on the force of the wrongdoing.

It helps in building a feeling of wellbeing in the brains of the residents of a dwelling nation.

REFERENCE

1. Mingchen feng 1, jiangbin zheng1, jinchang ren 2,3, (senior member, ieee), amir hussain 4, (senior member, ieee), xiuxiu li5, yue xi 1, and qiaoyuan liu 6 “Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data”
2. K. Chitra Lekha and Dr. S. Prakasam, “An analysis of finding the Influencing Factors of supporting for the “GiveitUp” LPG Subsidy for the Government using Data mining Techniques”, IJCA, Vol. 143, Issue 5, June 2016, pp.34-39.
3. R.M.Rani, Dr.M. Pushpalatha," Generation of Frequent sensor epochs using efficient Parallel Distributed mining algorithm in large IOT", Computer Communications, Volume 148, 15 December 2019, Pages 107-114
4. R.Mythili, Revathi Venkataraman, T.Sai Raj,"An attribute-based lightweight cloud data access control using hypergraph structure", The Journal of Supercomputing(JoS),Published online: 02 Jan 2020 DOI: 10.1007/s11227-019-03119-7
5. S.Sivamohan, Liza.M.K, R.Veeramani, Krishnaveni.S, Jothi.B, "Data Mining Techniques for DDOS Attack in Cloud Computing", IJCTA International Scoience Press, Pg: 149-156
6. S Pandiaraj, Aishwarya, Surbhi, Alisha Minj, Priyanshu Singh, "Enabling Cloud Database Security Using Third Party Auditor", International Journal of Engineering and Advanced Technology (IJEAT), Volume-8 Issue-4, April, 2019
7. R.Veeramani,Dr.R.Madhan Mohan, "Iot Based Speech Recognition Controlled Car using Arduino", International Journal of Engineering and Advanced Technology,Volume-9 Issue-1, October 2019
8. T.H. Feiroz khan, N.Noor Alleema, Narendra Yadav, Sameer Mishra, Anshuman Shahi "Text Document Clustering using K-Means and Dbscan by using Machine Learning",International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019
9. S.Babeetha, B. Muruganantham, S. Ganesh Kumar, A. Murugan, "An enhanced kernel weighted collaborative recommended system to alleviate sparsity", International Journal of Electrical and Computer Engineering (IJECE), Volume 10, February 2020, Page No. 447-454
10. Kavitha.R ,K.Malathi,"Recognition and Classification of Diabetic Retinopathy utilizing Digital Fundus Image with Hybrid Algorithms", October 2019,International Journal of Emgineering & Advanced Technology(IJEAT), Volume 9, Isssue 1, 109-122
11. T.Chandraleka,Jayaraj R, " Hand Gesture Robot Car using ADXL 335 ", International Journal ofEngineering and Advanced Technology (IJEAT)’, Volume-8 Issue-4, Nov 2019
12. H.Sangeetha,S.Abinayaa, "Smart Irrigation Systems using Sensors and GSM" in ‘International Journal of Recent Technology and Engineering (IJRTE)’, Volume-8 Issue-1, May 2019. Page No.:884-886
13. B.Sathya Bama,,Y.Bevish Jinila, "Attacks in Wireless sensor networks- A Research" ,International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8, Issue- 9S2, July 2019
14. Vellingiri, J., S. Kaliraj, S. Satheeshkumar and T. Parthiban , "A Novel Approach for User Navigation Pattern Discovery and Analysis for Web Usage Mining", Journal of Computer Science 2015, vol 11 (2): Page no 372.382

AUTHORS PROFILE

M. Vinodhini is Assistant Professor in Department of Information Technology, SRMIST,Chennai, TamilNadu,India

Dhruv Srivastava is currently pursuing bachelors of technology in information technology from SRM IST,Chennai,Tamil Nadu,India

Sameer Hirani is currently pursuing bachelors of technology in information technology from SRM IST,Chennai,Tamil Nadu,India

.

Shubham Arora is currently pursuing bachelors of technology in information technology from SRM IST,Chennai,Tamil Nadu,India