Application of Big Data techniques in Crime Analysis

Sriram Srinivasan *MS Computing in Big Data and Artificial Intelligence*
Letterkeny Institute of Technology
Letterkenny, Ireland
L00156509@student.lyit.ie

***Abstract*—Big Data Analysis is one of the seamless approach for unfolding the mystery behind the overabundant data. These analysis will ultimately lead us to understand the underlying relationships between the data.In this paper we are going to apply the Big Data Techniques to visualize the pattern and trends of the crimes happened over the last 20 years. With the advancement of technologies, the percentage of the crime rate has also increased over the decades and it is imperative that these analysis would not only help the Law Enforcement Authorities to act on solving such crimes in no time but would also proactively prevent such crimes in the future. The data will be pre-processed and some of the machine learning algorithms will be used to identify the trends**

I. Introduction

An increase in the crime rate has become the talking point augmenting wider threat both to the governing authorities and to the common civilian in the country. As the population increases , the number of crimes and the frequency of the crimes has spiked to a different level. In order to handle such offenses, it's high time to get some assistance from technological experts who can drill down the data further and narrow down the rationale behind the offenses [1]. Big Data refers to datasets that are very huge, very high in variety and having high velocity. It is a great challenge to handle such Big Data using conventional tools, techniques and hardware/software platforms [2]. Crimes are neither structured nor unspecific and it is impossible to find out any correlation between the crimes without any advanced technology or tools that can read such huge data. Big Data makes the job easier to visualize the analogy of the crime events . The prime purpose of crime analysis is the study of crime(eg: Theft, Rape, Burglary etc) and disorder problems (eg; noise complaints, burglar alarms and suspicious activities) and information related to the nature of the incidents,offenders and victims of targets (i.e , inanimate objects such as buildings or property) of these problems [3]. We may not be able to identify the criminals nor the victims of the crime directly but this analysis proves vital in finding the patterns of the crimes and the crime location. Big Data Analytics can effectively address the challenges of data that are too vast, too unstructured, and too fast moving to be managed by traditional methods [4]

Identify applicable funding agency here. If none, delete this.

II. Related works

Mingchen Feng et all has used some of the mining techniques to explore the data and for some effective visualization and trending of criminal activities [5] .Chhaya Chauhan and Smriti Sehgal has completed a similar work on the data mining techniques and algorithms [6].In human behavioural data derived from mobile network activity combined with demographic information using real crime data were used to predict crime hotspots in London, UK [7] [6]. A study by Khushboo Sukhija, Shailendra Narayan Singh and Mukesh Kumar using regression model to understand and identify the correlation of parameters associated with Rape to determine the significant variables which helps the police authorities in preventing crime more efficiently [8] Romika Yadav and Savita Kumari Sheoran have explored the Auto Regression Techniques to accurately predict the crime with minimum error for such time series data by identifying the relationship among crimes attributes [9]

III. Crime Data Analysis

The use of statistical methods for categorical data has increased dramatically, particularly for applications in the biomedical and social Sciences [10]. As one of the fundamental techniques of BDA, data mining is an innovative, interdisciplinary, and growing research area, which can build paradigms and techniques across various fields for deducing useful information and hidden patterns from data [11].While analysing the crime data, the connection between the crimes such as the periodicity of the crimes, time interval between such events, history of similar activities etc, will be considered to understand the pattern of the crimes. The dataset that is used in this research is available for public access online. BDA can effectively address the challenges of data that are too vast, too unstructured, and too fast moving to be managed by traditional methods. This dataset contains 1,048,576 crime incidents that happened in chicago which comprises data from the year 2001 till 2020.

The following features are available for each incident of the crime :

1) ID - Unique identifier for the record 2) Case Number - Unique Chicago Police Department RD number 3) Date - Date of the incident 4) Block - The block in which the incident happened 5) IUCR - The Illinois Uniform Crime Reporting code 6) Primary Type : Primary description of the

crime 7) Description : Sub category of the crime 8) Location Description : Description of the location of the incident 9) Arrest : Whether an arrest was made 10) Domestic : Indicates a domestic violence 11) Beat : The Beat where the incident happened 12) District : Indicates the district where the incident occurred 13) Ward : Ward where the incident occurred 14) Community Y Area: Indicates the community area where the incident occurred 15) FBI code : FBI code used in the National Incident- Based reporting system 16) X coordinate : The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block. 17) Y coordinate: The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block. 18) Year : Year of the incident 19) Updated on : Date and time when the record was last updated 20) Latitude : Latitude of the location 21) Longitude : Longitude of the location 22) Location : The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data

IV. Census Dataset

In addition to the crimes dataset, census data with six socioeconomic indicators are used to find out if there is any correlation between the crime events and the income and the poverty status in the location of the crime. The census data was pre-processed and the fields that are empty were removed from the dataset. Columns such as ”PERCENT AGED UNDER 18 OR OVER 64” and ”Hardship Index” were deleted. The data in the crimes dataframe is then mapped with the data in the census dataframe with the help of Community Area Number column being the common between the two dataframes.

V. Data Preprocessing

1) The fields with null values have been removed from the dataset as there were only a few blank rows and columns

2) Some of the columns with less value add like ID, X coordinates, Y coordinates were removed

3) The date attribute has been further divided into five different fetaures: Year(2001 - 2020, month(1-12), date (1-31), hour(1-23), minutes(1-59).

4) Columns like Date and Updated on are converted to pandas datetime format and some of the columns like Description , Primary Type, Location Description are converted to category type.

VI. Data Visualization

Data visualization is the demonstration of data with illustrations and graphical representations. There are ample of softwares that can do the art of data visualization. Visualization has proven effective for not only presenting essential information in vast amounts of data but also driving complex analyses [12]. When it comes to data visualization, there are a number of visualization techniques involved that can classify and present the data in various

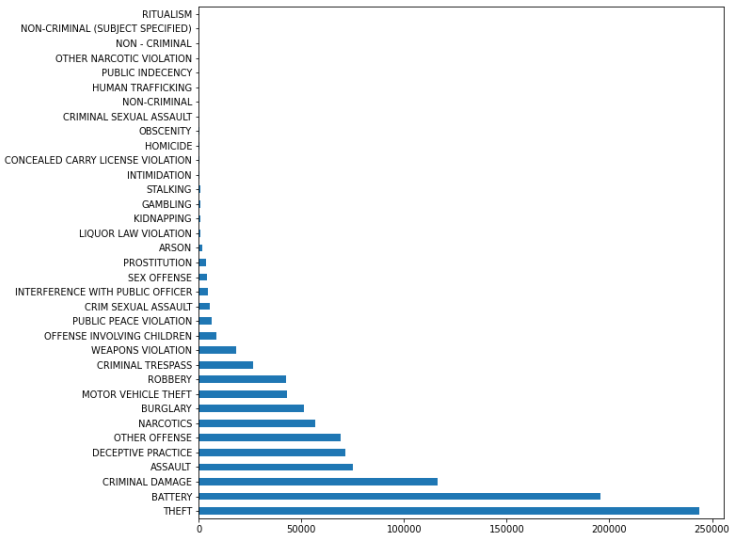


Fig. 1: Major contributors

formats. Some of the categories of visualizations are Scientific Visualization, GIS, Multi-dimensional Plots, Multi-dimensional Tables,Information Landscapes and Spaces, Node and Link,Trees, and Text Transforms [13].

The chart in Fig 1 exhibits the different types of crimes from the dataset. From the chart it’s clear that theft has contributed more in numbers than any other crime, followed by Battery and Criminal Damage at the second and third spot respectively. With too much of data the chart might look a bit skewed. In order to make it look more accurate, let’s take only the list of top 10 crimes.The Pie chart above showcase the top 15 crimes since 2001.

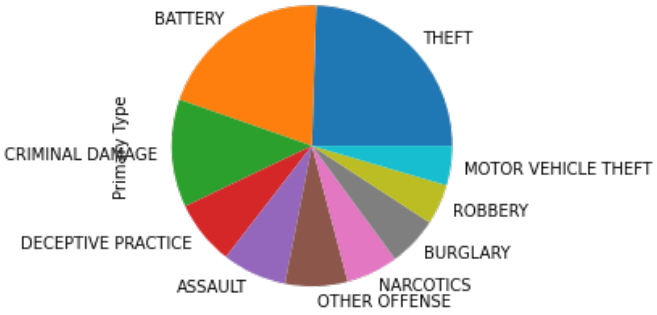


Fig. 2: Top 10 Crimes

A major challenge facing all law-enforcement and intelligence-gathering organizations is accurately and efficiently analyzing the growing volumes of crime data [5].The ability to make timely decisions based on available data is crucial to business success, clinical treatments, cyber and national security, and disaster management [14]. This chart in Fig 3 shows the crimes that are spread across twenty four hours in a day. The graph shows some interesting fact that the majority of the crime events had occurred after 12pm in a day. The graph unveils that 5am appears to be the most safest time in a day as the crime incidents are very much lower during

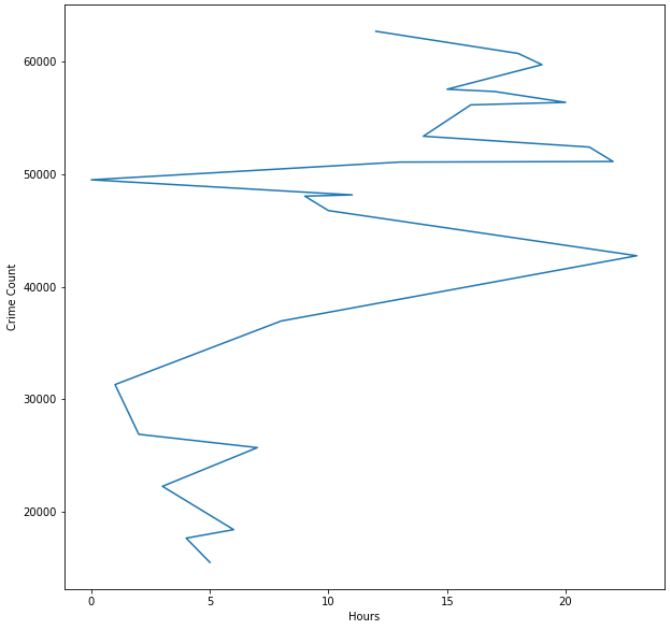


Fig. 3: Crimes Vs Hours in a day

dawn whereas afternoon and evening have contributed more number of incidents. The worst time in a day happens to be 12pm, 6pm and 7pm as they are the top 3 with around 60000 criminal activities each. More police force should be deployed during theses hours in order to bring down the number of offenses. Morning 1 to 8 are the hours with least number of criminal activities.

Looking at the monthly data August, September, October and January are the most vulnerables with January topping the list for the most number of Criminal offenses. This shows that tha criminals are very active during the beginning of the year. The least number of crimes recorded were during the month of Nov, Dec and Feb. Probably the offenders take more time during Nov and Dec to create a master plan that they could execute during the month of Jan.

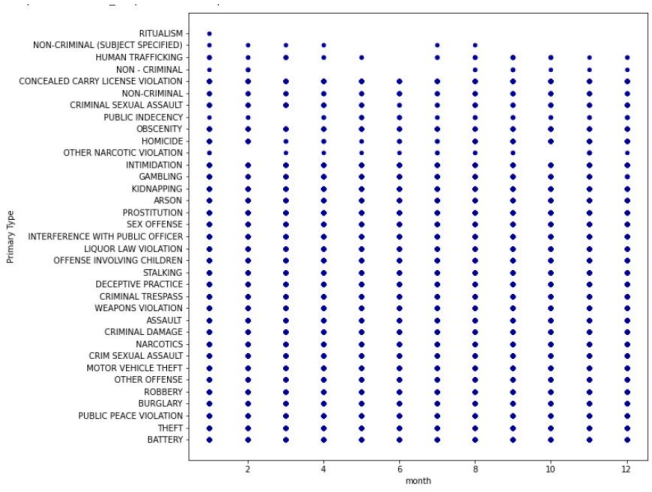


Fig. 4: Crimes - Monthly view

The month of Feb is one of the least affected month during a year and it could be the time they give a partial break for all their crimes. Big data is extensively used to transform large unstructured or structured raw data into crucial and meaningful information which helps in forming a healthy decision support system for the judiciary and legislature to enforce law and order towards keeping crimes in check and making strategic decisions for safety and well being of the society [15].

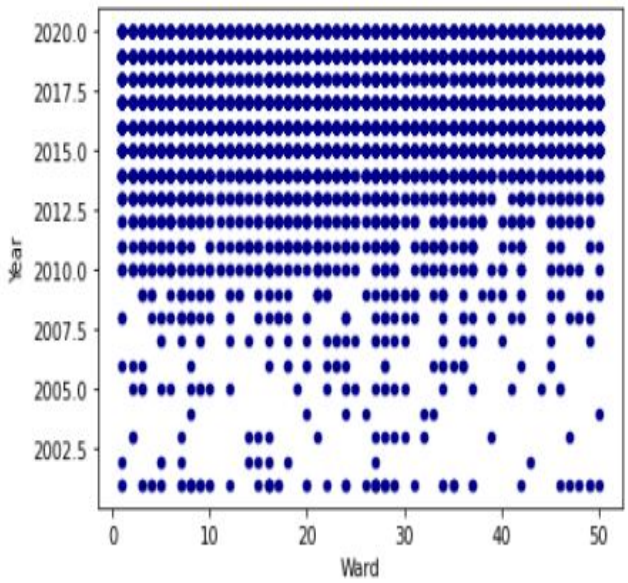


Fig. 5: Ward V Year data

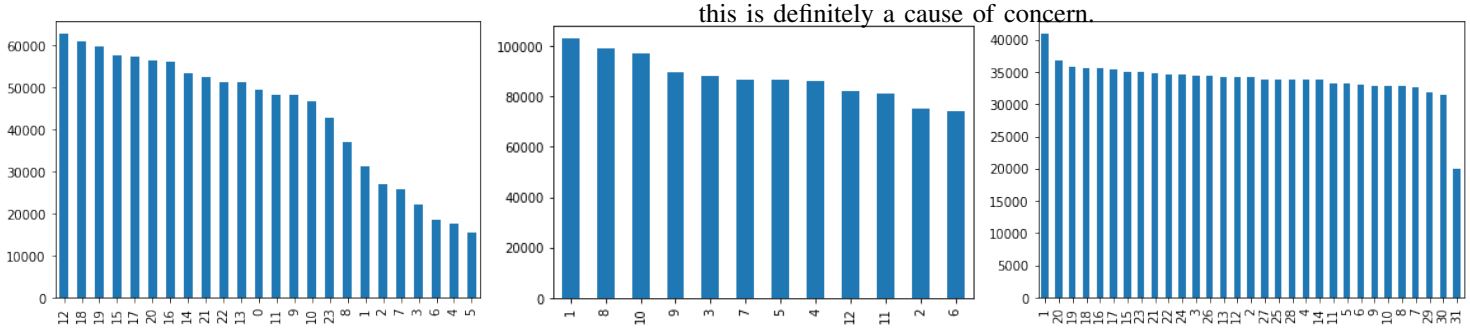
Scatter plot in the above figure displays the offenses happened since 2001 and from the chart, year 2001 to 2010 is sparsely populated. Not much of criminal activities during the first 10 years of the dataset. Not all of the wards were impacted during this time whereas from year 2011 to 2020, the scatter dots are densely populated. Each and every ward was being affected and impacts would have been huge. Especially the years between 2014 to 2020 looks horrible as no wards were spared durig these 6 years.

VII. Machine Learning Techniques

Another type of analysing the data is by using machine learning based techniques. Some of the analysis performed in the traditional machine learning techniques predominantly performs well on structured data. But the modern day techniques can drill down on the huge dataset even if it is unstructured or unorganized. Recently, the Vancouver Police Department (VPD) introduced a crime-predictive model to predict crimes related to property break-ins and, once implemented, the city of Vancouver witnessed a 27 percentage drop in residential break-ins [16]

Linear Regression algorithm, aslo known as statistical algorithm as it originated from statistics, predicts the relationships between the input and the output variables.The purpose of the linear regression model is to improve the accuracy of the predictions. In other words this model produces low error rate and hence this is one of the best machine learning algorithm

this is definitely a cause of concern.



(a) Hours of crimes occured (b) Monthly view of Crimes (c) Crime splitup - Daywise

Fig. 6

when it comes to dealing with huge data. In this paper linear regression is used to check the accuracy of the model.

The chart in Figure 8 is a comparison of the crimes that had happened, split into 2 decades. The crimes from the year 2001 to 2010 is highlighted as the first decade and the crimes between 2011 and 2020 is considered as the second decade. Suprisingly the crime rate in the second decade is over 1 million compared to the ones that had happened in the first decade which is around twenty eight thousand only. This is quite alarming as the crime rate has blown up to a all time high. However this could be due to various reasons. Either not much data is available for the first decade as the evolution of computer technology in the last 10 years is humongous. But if all of the crimes were recorded and the data is accurate, then

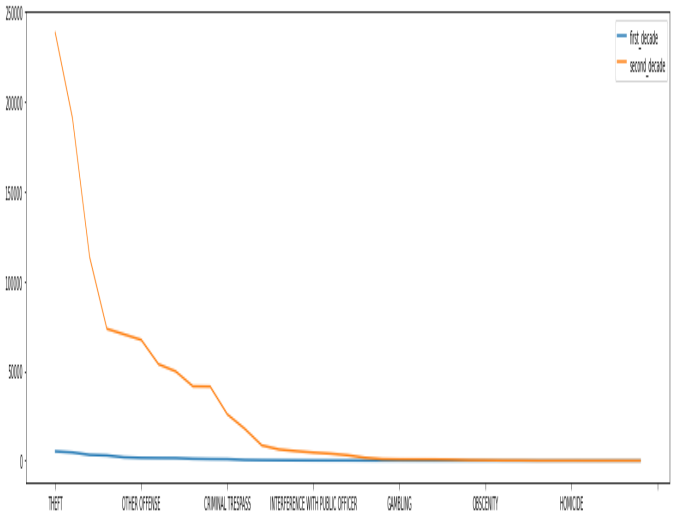
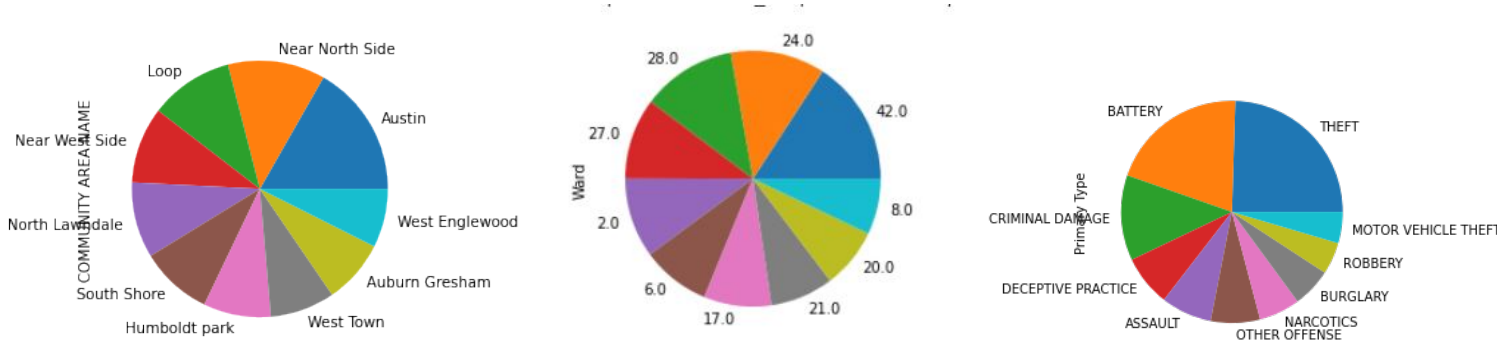


Fig. 8: Decade comparison

In order to check the accuracy of the model, first the x and y variables were defined, x being the Primary Type column and Y being the Beat column. Then the training and testing data were defined. For this model 80 percent of the data from the dataset is taken for training purpose and the remaining 20 percent will be used for testing purpose to check the accuracy. Once fitting the data into the model the regression score for the training and the testing set is as shown below.



(a) Community Area - Top 10 contributors (b) Ward - Top 10 Contributors (c) Top 10 Crimes

Fig. 7

Linear regression deals with static values whereas classifi-

cation deals with discrete values

The accuracy score for the linear regression model for the training set was 0.0007136838865832162 The accuracy score for the linear regression model for the testing data was 0.0006854550786575819

Upon further analysis, the data was scrutinized using Random forest Regression model.

Random forest uses ensemble method of learning. Random forest works by producing multiple decision trees and uses the mean value of all the classes to narrow down on the accuracy of the model. Random forest at times not only used for classification but also used for regression models.

The accuracy score while using Random Forest Regression model was 85 Percent.

VIII. Conclusions

We have used both the Linear regression model and the Random forest model to identify the different types of crimes, crime pattern, frequency of the crimes, location impacted more often, the connection between several of these crimes and so on. The results are better with the Random Forest Regression as the accuracy rate is 85 percent while using Random forest. The idea behind this analysis is to get an understanding on the crime pattern which will in turn alert the investigation authorities and take proactive measures and to get an idea on how to handle or prevent such crimes from happening.It's high time that we need to develop a system that can report such incidents real time rather than waiting for it to be identified after the damage is done.It's a positive sign that even the government authorities are coming forward to fund these kind of research activities and it can only get better with crime rate possibly coming down in the near future

References

[1] H. Hassani, X. Huang, E. S. Silva, and M. Ghodsi, "A review of data mining applications in crime," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 9, no. 3, pp. 139–154, 2016.

[2] A. Londhe and P. P. Rao, "Platforms for big data analytics: Trend towards hybrid era," in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pp. 3235–3238, IEEE, 2017.

[3] R. B. Santos, *Crime analysis with crime mapping*. Sage publications, 2016.

[4] J. Zakir, T. Seymour, and K. Berg, "Big data analytics.," *Issues in Information Systems*, vol. 16, no. 2, 2015.

[5] H. Chen, W. Chung, J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: a general framework and some examples," *Computer*, vol. 37, no. 4, pp. 50–56, 2004.

[6] C. Chauhan and S. Sehgal, "A review: Crime analysis using data mining techniques and algorithms," in *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pp. 21–25, 2017.

[7] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland, "Once upon a crime: towards crime prediction from demographics and mobile data," in *Proceedings of the 16th international conference on multimodal interaction*, pp. 427–434, 2014.

[8] K. Sukhija, S. N. Singh, and M. Kumar, "Using linear regression to investigate parameters associated with rape crime in haryana," in *2020 10th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, pp. 107–111, 2020.

[9] R. Yadav and S. Kumari Sheoran, "Crime prediction using auto regression techniques for time series data," in *2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)*, pp. 1–5, 2018.

[10] A. Agresti, *An introduction to categorical data analysis*. John Wiley & Sons, 2018.

[11] U. Thongsatapornwatana, "A survey of data mining techniques for analyzing crime patterns," in *2016 Second Asian Conference on Defence Technology (ACDT)*, pp. 123–128, IEEE, 2016.

[12] D. Keim, H. Qu, and K.-L. Ma, "Big-data visualization," *IEEE Computer Graphics and Applications*, vol. 33, no. 4, pp. 20–21, 2013.

[13] E. H.-h. Chi, "A taxonomy of visualization techniques using the data state reference model," in *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, pp. 69–75, IEEE, 2000.

[14] D. Keim, H. Qu, and K.-L. Ma, "Big-data visualization," *IEEE Computer Graphics and Applications*, vol. 33, no. 4, pp. 20–21, 2013.

[15] K. Kattankulathur, "Crime analysis and prediction using big data," *International Journal of Pure and Applied Mathematics*, vol. 119, no. 12, pp. 207–211, 2018.

[16] J. Kerr, "Vancouver police go high tech to predict and prevent crime before it happens," *Vancouver Courier*, 2017.