Analysis, Visualization and Prediction of Vancouver's Criminal Activity

Vighnesh Mane
Data Science
City, University of London
London, United Kingdom
vighnesh.mane@city.ac.uk

***Abstract*—Crime analysis and prediction is a methodological technique to analysing and discovering various patterns, relationships, and trends in crime. The exploratory data analysis is used in this paper to examine criminal data in Vancouver. To increase the accuracy of the forecasts, the criminal category is projected by performing feature selection. Naïve Bayes and Decision Trees models are utilized. The data was then analysed using a time-series approach, and future crime trends were predicted for the following years. As a result of these encouraging results, police departments and law enforcement agencies will be able to solve crimes more quickly and gain insights into criminal activity.**

Keywords— data analysis, crime prediction, time series analysis.

I. INTRODUCTION

It's a problem for both the police and the law enforcement agencies because of the rising number of crimes. Different communities and societies have different ways to deal with criminal activity. In order to find out what causes a crime to occur, security analysts need to do a thorough examination of the events that occurred. Crime data analysis helps law enforcement officers retain track of criminal behaviour, detect resemblance between instances, set priorities, and make rapid decisions by enabling them to evaluate crime data. New insights regarding crime statistics can be gleaned from the immense quantities of crime data which have been made public due to the fast expansion of computer system and storage technology.

On the premise of crime data, and more explicitly time series data, the potential ramifications of machine learning and analytic models will help us better understand the pattern and developments of crimes in a country, and this comprehension will further assist society in making plans for the prevention and curtailment of crime. In this article, we looked at crime data from the city of Vancouver. It is the focus of this research to investigate criminal data from Vancouver through exploratory methodologies. The criminal group is forecasted through doing feature selection in order to increase the precision of the forecasts. We apply the models of Naïve Bayes and Decision Trees. Afterwards, the data is analysed through into the lens of a time-series approach, and future crime trends were anticipated for the years to come.

II. DATA

In this research, just one primary source of data is available. All the portion of the information came from the Vancouver Police Department's official website. From 2003 to 2021, there were 8,10,983 crimes in this dataset.

A. Features

Almost every record in our information gathering corresponds to a particular crime, and each record has the following characteristics:

- Type - This feature represents the category of the crime. In the classification step, we must forecast the target/label that we will use in the classification stage.
- Year - Crimes happened throughout the course of the year.
- Month - On a monthly basis, criminal activity was documented.
- Day - Criminal conduct was seen on a daily basis and recorded.
- Hour - Criminal activity was observed and documented on a 24-hour basis.
- Minute - On a minute-to-minute period, malfeasance was monitored and recorded.
- Hundred block - Location of the crime scene on the street.
- Neighbourhood - The area of the police department where the offence has been apportioned.
- X - The location of a crime in terms of longitude.
- Y - The location of a crime in terms of latitude.

III. ANALYTICAL QUESTIONS

Police and law enforcement agencies are dealing with an increasing number of crimes, which is a challenge. Criminal conduct is dealt with in a variety of ways depending on the society and the people in consideration. This dataset provides exact information require for analysis. The study of crime patterns, correlations, and trends is accomplished via the use of crime analysis and forecasting as a method. The following collection of questions will help organise your examination of the dataset and broaden its reach.

1. What is the trend in Vancouver's crime rate?
2. From 2003 to 2021, what is the percentage change in each form of crime in Vancouver?
3. On which months in which more than 5000 crimes occurred?
4. Which are the most dangerous neighbourhoods in Vancouver?
5. At what hour the theft from vehicle is at peak?
6. What can we expect in terms of crime in Vancouver in the future?

In the first question, we'll see if the rate of crime is increasing or decreasing. As % change provides an overview of changes in crime, the second question will assess all crime types' percent change. It is vital to maintain track of the most serious crimes committed over time. So, in the third question, we'll see which months had more than 5,000 crimes. People do not want to live in risky areas, thus in the fourth question we will find out which areas are the most dangerous. As auto theft is a major concern in Vancouver, we will examine the fifth question. Finally, predicting future trends is critical, therefore in the sixth question we will examine Vancouver's crime forecast.

IV. ANALYSIS

A. Data Preparation

The data is obtained from the Vancouver Police Department website and it requires some pre-processing, thus the column names are adjusted to make them more understandable. The address and neighbourhood columns have the missing data, therefore entering any random location is not an option due to the presence of the missing variables. Therefore, the missing data are removed from the table. The missing values discovered in the longitude and latitude columns are replaced by the mean of the values of it.

B. Data derivation

It may be essential to add a few additional columns to a dataset, or to generate a subset of the current columns or dataset, at various points in time. All of these things fall under the category of data derivation. To begin with, there are three new columns: date, time, and dayofweek, which will be used for data extraction. The date column is developed by merging the columns for the year, month, and day. The hour and minute columns are combined, due to which the time column is born! It is necessary to extract the day from the date column in order to build the dayofweek column. In addition, a data frame called crime year is created, which provides the total number of crimes committed in a given year. Finally, for time series analysis, a data frame called date count is generated, which contains the date as well as the number of crimes committed on a daily basis.

C. Data Visualization

Data visualisation, which makes use of maps or graphs, aids in the understanding of what information is being presented and how it connects to us. First and foremost, the sort of crime is planned. Then, using a line plot, the pattern of crimes is investigated. Following that, the distribution of each offence is shown graphically. The percentage change in crime is then shown on a line plot to show the progression of the crime. After then, the most serious crimes are shown graphically. A heatmap is also used to illustrate the number of crimes that occurred on a daily, monthly, and hourly basis. Following that, a bar plot is used to investigate the riskiest neighbourhoods and blocks. Then, by the use of a bar plot, the crimes committed from a vehicle were represented. Then, using a heatmap, the correlation between variables is determined in order to choose the most appropriate features for machine learning models. Line plots are used to depict the future trend of the crimes, which is the final step.

D. Construction of models

Naïve Bayes - Predictors are assumed to be independent in Bayesian categorization. It is presumed that the presence of a particular feature in a class is independent to the presence of any other feature in the same class. Apples are red in colour, round, and about 3 inches across. They're also around the same size. Fruits that have all of these characteristics are referred to be "naive," even if some of them are dependent on each other or on the occurrence of additional attributes. First, the features are chosen and the type variable is used as an output variable in model creation. The data is then split in half, with eighty percent going to training and twenty percent going to testing. After that, the model is set up. Decision Tree - Each node in the decision tree has a branch that is dependent on a certain event. When a predetermined cut off value is achieved, data is divided into branches. There are three sorts

of nodes in a decision tree: root nodes, child nodes, and leaf nodes. Selected characteristics are turned into decision nodes to begin the process of building a decision tree model. Feature selection is then repeated for each kid until the last attribute has been selected. Model generation begins by selecting a set of characteristics and using the type variable as an output. Eighty percent of the data is used for training, while twenty percent is used for testing. The model is then ready for use.

E. Time series analysis

A time series forecast is a prediction of what will happen in the future assuming the same patterns continue to hold true based on previous observations. The Holt-Winters forecasting approach is used to model and predict the behaviour of a sequence of values through time in this study's time series analysis. For this model, the date column is grouped, then resampling is carried out, and the first day of each month is taken into account. The four seasons are then examined to see whether there is a seasonality. 80% of the data is used for training, whereas just 20% is used for testing purposes. The model is now ready to be used.

F. Validation of results

The accuracy of the Naïve Bayes model is 37 percent, the macro average is 17 percent, and the weighted average is 30 percent. The Decision Tree model has an accuracy of 44 percent, the macro average has an accuracy of 32 percent, and the weighted average has a precision of 45 percent.

V. FINDINGS, REFLECTIONS AND FURTHER WORK

A. Findings and reflections

1) Trend analysis

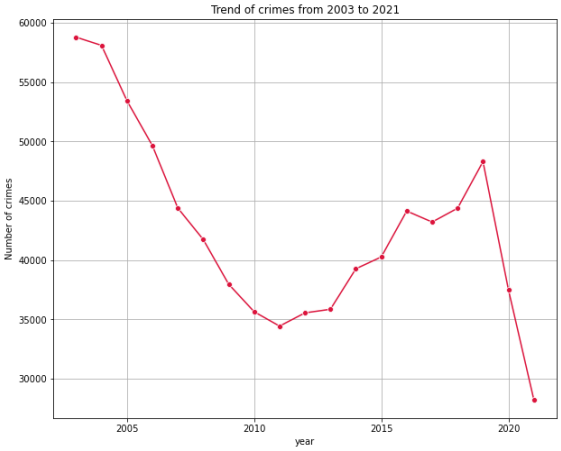

Fig a. Trend of crimes from 2003 to 2021

The declining trend in crime is noticed [Fig. a.]. The crime rate peaked in 2003, then began to decline. However, it began to rise again in 2011-2012, and it is on track to hit about 50,000 in 2019. Then it began to plummet at a quick pace.

2) Percentage variation in crimes

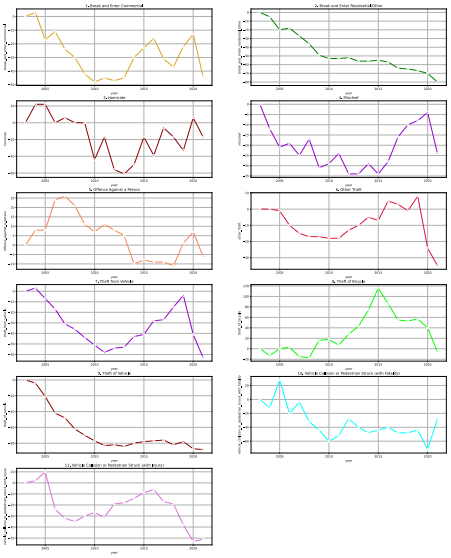

Fig b. Percentage change in crimes

From 2003 to 2021, the vast majority of crime types experienced a precipitous fall. In accordance with [Fig. b.], the number of break and enter commercial crimes has fallen by 40%, while the number of break and enter residential/other crimes has decreased by 70%. The number of homicides and mischiefs has dropped by about 20 and 25 percent, respectively. Following a nearly 25% increase in 2007, the offence against people has decreased by 5 percent in 2021, while the other theft has decreased by 30% in the same period. Theft from vehicles has decreased by 60%, which is a significant reduction. Theft of bicycles has decreased by around 5 percent in 2021, following an increase of over 120 percent in the previous year. Vehicle collisions resulting in death or serious injury have decreased by around 25 and 50 percent, respectively. [Fig c.] shows the distribution of each crime type. The largest crime rate, which is 28.2 percent, is shown by the pie chart, and the lowest crime rate, which is 0.0355 percent.

3) Observation of crimes on monthly basis

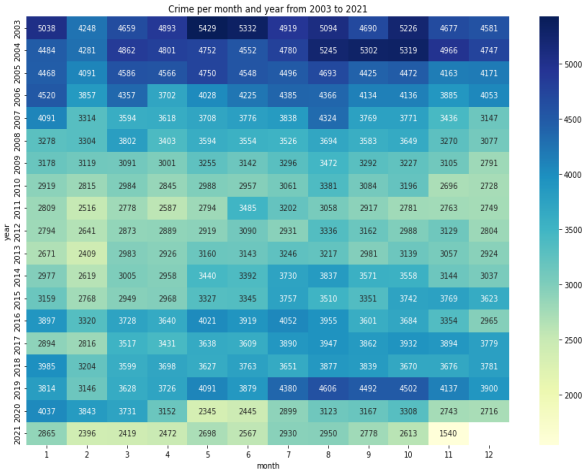

Fig d. Crimes per month from 2003 and 2021

If you compare 2003 to other years, you will notice that the crime rate was exceptionally high. Criminal activity surpassed 5000 in the months of January, May, June, August and October of that year. In addition, during the months of August, September, and October in the year 2004, the number of crimes surpassed 5000 [Fig d.].

4) Crime analysis of neighbourhood

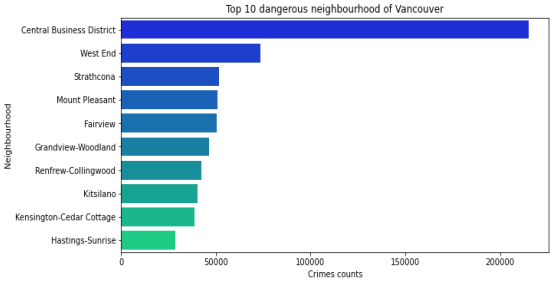

Fig e. Dangerous neighbourhoods

According to the Offenses in neighbourhood Visualization, the central business area, which is located in the heart of the city, has more than 200000 crimes. West End is one of the most accessible neighbourhoods for travellers to explore, and it has more than 80,000 offences on record. Strathcona is the following neighbourhood, and it has a somewhat higher crime rate than Mount Pleasant and Fairview. Grandview-Woodland is followed by Renfrew-Collingwood, Kitsilano, Kensington-Cedar Cottage, and Hastings-Sunrise, each of which has a crime rate less than 50000.

5) Theft from vehicle

Because of this, thefts from vehicles were at their highest in [Fig f]. Therefore, it occurs on a regular basis and can be detected. At 6 p.m., vehicle theft is at its peak.

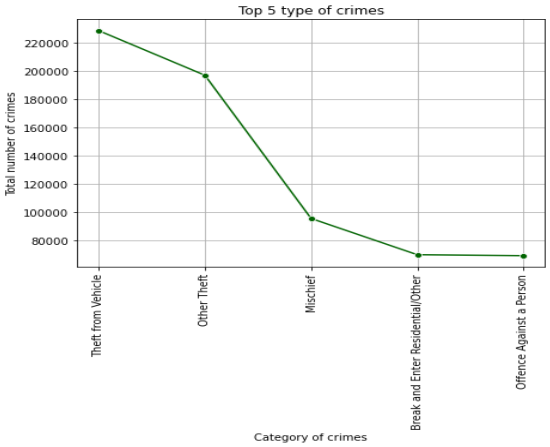

Fig f. Top five type of crimes

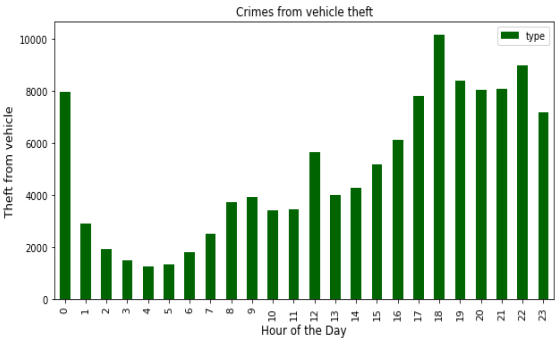

Fig g. Theft from vehicle in hour of the day

6) Future trend analysis

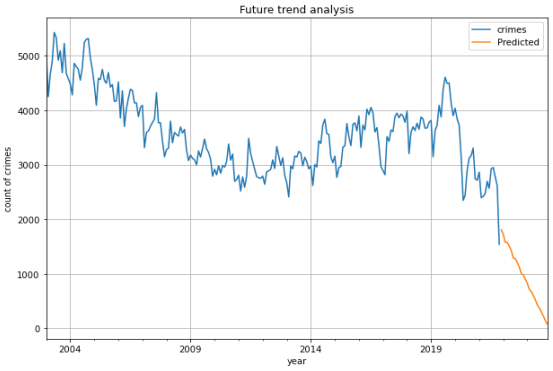

Fig h. Future trend analysis

Crime forecasting is critical because it provides a fundamental framework for planning. Holt-Winters forecasting is used to anticipate the future trend for 2 years starting from Dec 2021 till Nov 2023, and the declining tendency is seen.

B. Further work

The information used in this study was gathered from a total of 18 years of Vancouver's crime statistics records. It is possible to accurately predict between 37 percent and 44 percent of crimes using

machine learning models such as Nave Bayes and Decision Trees, depending on the model used. It will be attempted to increase the accuracy of models by modifying their parameters, and it will also be attempted to discover more details by experimenting with other techniques.

VI. REFERENCES

[1] S. Kim, P. Joshi, P. S. Kalsi and P. Taheri, "Crime Analysis Through Machine Learning," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2018, pp. 415-420, doi: 10.1109/IEMCON.2018.8614828.

[2] A. Shukla, A. Katal, S. Raghuvanshi and S. Sharma, "Criminal Combat: Crime Analysis and Prediction Using Machine Learning," 2021 International Conference on Intelligent Technologies (CONIT), 2021, pp. 1-5, doi: 10.1109/CONIT51480.2021.9498397.

[3] S. Yadav, M. Timbadia, A. Yadav, R. Vishwakarma and N. Yadav, "Crime pattern detection, analysis & prediction," 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), 2017, pp. 225-230, doi: 10.1109/ICECA.2017.8203676.

[4] A. J. Park, V. Spicer, H. H. Tsang, K. Behiels and J. Song, "Discovering Crime Trends and Patterns Using Three-Dimensional Visual Analytics," 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2019, pp. 0545-0549, doi: 10.1109/IEMCON.2019.8936251.

[5] Kleber Henrique de Jesus Prado, Methanias Colaço Júniora ," Data Science Applied to Crime Analysis Based on BrazilianOpenGovernmentData",https://www.tandfonline.com/doi/full/10.1080/19361610.2020.1848057.

[6] Devarakonda, D. S. (n.d.). "Time Series Analysis And Forecasting Of Crime Data",https://csu-csus.esploro.exlibrisgroup.com/esploro/outputs/99257831172001671.

[7] G. Borowik, Z. M. Wawrzyniak and P. Cichosz, "Time series analysis for crime forecasting," 2018 26th International Conference on Systems Engineering (ICSEng), 2018, pp. 1-10, doi: 10.1109/ICSENG.2018.8638179.

TABLE 1

Serial No.	*Section*	*Word count*
	Abstract	113
I	Introduction	250
II	Data	188
III	Analytical Questions	275
IV	Analysis	847
V	Findings, Reflections and Further Work	599
VI	References	199