COMP 5411 FB

Fall 2019

Project Report

Project Title: Crime Analysis and prediction on Vancouver Crime reports.

Group Members:

Name	**Student No.**	**Section**
Bhargav Lad	1117021	FB
Pathikkumar Patel	1117477	FB

Abstract:

A typical Big Data Project Pipeline consists of Data Engineering, Data Preprocessing and performing analytics on that data. However, the importance for each of these steps varies according to the dataset chosen for the project. The Dataset chosen for this project is Vancouver Crime Reports Dataset. Analysing and Forecasting crime incidents can help the Police Department decrease the crime activities in the city making it a safer place for the residents to live in. The main Objective of this project is to provide a reliable and a highly interpretable model for crime forecasting which can not only predict the type of crime but also explain as to which factors led to such activity in an area. Here, we grouped the crime types based on the similarity between them to generalize the Feature values. This grouping was done based on extensive research and domain knowledge. We added new features to the Dataset that are highly correlated with our dataset and can help us gain more insights. These new features were "weekday" and "holiday". These features tell us what day of the week did the crime happen and whether it was a public holiday or not. The next step was to pre-process the data and find missing values and deal with the missing data in an appropriate manner. After that we used two classification models, namely: Random Forest and XGBoost for predicting the crime type. These models use a set of features obtained by Feature Selection techniques as input data. The accuracy obtained by the two models is 62.85% and 69.15% respectively. We then try to locate the most probable locations of crime given date, time, neighbourhood and crime type based on random forest predictions. These locations can be used by the police department as a reference to guard the area with appropriate protection measures.

Introduction

Problem Definition

Analysing crime record data of Vancouver city using various Classification techniques and Exploratory Data analysis. The goal of the project is to analyse, model and forecast criminal activity in Vancouver city. Furthermore, we will also try to predict location and create hotspots for a specific type of crime. We will try to apply various machine learning models and compare their performance and try to reason for its better or worse performance.

Problem Significance

Crime is one of the major concerns in our society as it affects various socio-economic factors such as population, income, unemployment etc. Therefore, its prevention is crucial for the growth and development of society. As the number of crimes reported each day is growing rapidly, the police department has to solve the cases in much faster way. With significant increase in criminal activities, police department has to come up with some method with which they can predict and identify crimes at a much faster rate. There is a need of technology with which the cases could be solved in faster and efficient way.

Literature Review

Crime can be defined as an act of offense that is illegal in nature and is punishable by law. It is a significant threat to the people and is increasing and spreading at a higher rate. Crime is not constrained to a specific community or a place. It can happen in a distant town or at a major urban city. Although the types of crime occurring in these areas might different, it is equally harmful for both of the communities. The major crimes happening in these areas are – robbery, murder, assault, B&E(Breaking and Entering), kidnapping and theft. Thus it has become necessary to monitor and control such activities. This is possible through analysis of the data collected for each of the reported crime and maintaining that database. This data can become a valuable asset in monitoring the crime related activities in the city and also forecasting these activities based on the historical data. This kind of analysis requires building models and applying analysis techniques that take into consideration all the relevant aspects of the data and provide us with valuable insights generated from it.

Related work

There has been some similar work done by researches in the crime analysis and prediction field. One such instance is the work done by[1] on the Chicago Crime dataset, where they have used different classification techniques for prediction purposes. The techniques used in the article [1] are KNN, GaussianNB, MultinomialNB, SVC and DecisionTree Classifier. The results show that KNN outperforms all the other methods in terms of performance and accuracy for the given dataset. Another such work has been done by Ian Santillan on the Toronto Crime dataset[2] where the author has used different classification techniques to forecast the crimes happening in Toronto. Suhong Kim *et al.*[3] have done data analysis on the Vancouver crime dataset using different techniques. They have used two Machine Learning techniques and have provided results and accuracy for these methods. The authors in [6] have proposed a method for crime prediction from demographics and mobile data. The findings by the authors in [6] suggest that aggregated human behavioral data captured from the mobile network infrastructure, in combination with basic demographic information, can be used to predict crime. The authors in [7] show how point process models of crime can be extended to include leading indicator crime types, while capturing both short-term and long-term patterns of risk, through a marked point process approach. Several years of data and many different crime types are systematically combined to yield accurate hotspot maps that can be used for the purpose of predictive policing of gun-related crime. The authors in [8] compares the two different classification algorithms namely, Naïve Bayesian and Decision Tree for predicting 'Crime Category' for different states in USA. The paper[9] focuses on finding spatial and temporal criminal hotspots. It analyses two different real-world crimes datasets for Denver, CO and Los Angeles, CA and provides a comparison between the two datasets through a statistical analysis supported by several graphs. Then, it clarifies how we conducted Apriori algorithm to produce interesting frequent patterns for criminal hotspots. In addition, the paper shows how the authors used Decision Tree classifier and Naïve Bayesian classifier in order to predict potential crime types. To further analyse crimes' datasets, the paper introduces an analysis study by combining our findings of Denver crimes' dataset with its demographics information in order to capture the factors that might affect the safety of neighborhoods. The approach proposed by authors in [10] is to construct a decision tree based classification model for a crime prediction. Proposed model assists law enforcement agencies in discovering crime patterns and predicting future trends. Chen et al. [11] categorized different crime types and proposed some techniques to mine crime data such as entity extraction and association rule mining. They also developed a

framework that identifies the relationships between crime types and data mining techniques applied in crime data analysis. The developed framework helps investigators to find associations and identify patterns to make predictions. They implemented some case studies and showed their developed framework can help increase efficiency and reduce errors. The authors in [12] have an approach between computer science and criminal justice to develop a data mining procedure that can help solve crimes faster. Instead of focusing on causes of crime occurrence like criminal background of offender, political enmity etc the authors are focusing mainly on crime factors of each day.

Data

Data Source

The dataset chosen for this project is from www.kaggle.com . The dataset is maintained by Kangbo Lu and is expected to be updated annually. The data is collected from City of Vancouver's Open Data Catalogue.

Link : https://www.kaggle.com/agilesifaka/vancouver-crime-report

Data Description

Content

This dataset contains the Vancouver Police Department's crime records from 2003 to 2019. The dataset contains the crime type, time data (year, month, day, and hour), and location data from street location to x, y GPS locations.

Features:

1. **TYPE -** Crime type
2. **YEAR** - Recorded year
3. **MONTH -** Recorded month
4. **DAY -** Recorded day
5. **HOUR -** Recorded hour
6. **MINUTE** - Recorded minute
7. **HUNDRED_BLOCK** - Recorded block
8. **NEIGHBOURHOOD** - Recorded neighborhood
9. **X** - GPS longtitude
10. **Y** - GPS latitude

Dimension of dataset

The dataset has 10 features and 609000 instances.

Missing Values

The following 4 columns has missing values NEIGHHOURHOOD (10.42% missing), MINUTE(10.02% missing), HOUR(10.02% missing), and HUNDRED_BLOCK(0.002% missing).

Attributes Description [4]

- **TYPE:** The type of crime activities

  ✦**BNE Commercial:** (Commercial Break and Enter) Breaking and entering into a commercial property with intent to commit an offence.
  ✦**BNE Residential/Other:** (Residential Break and Enter) Breaking and entering into a dwelling/house/apartment/garage with intent to commit an offence.
  ✦**Vehicle Collision or Pedestrian Struck (with Fatality):** Includes primarily pedestrian or cyclist struck and killed by a vehicle. It also includes vehicle to vehicle fatal accidents, however these incidents are fewer in number when compared to the overall data set.
  Note: There is no neighbourhood information.
  ✦**Vehicle Collision or Pedestrian Struck (with Injury):** Includes all categories of vehicle involved accidents with injuries. This includes pedestrian and cyclist involved incidents with injuries.
  Note: There is no neighbourhood information.
  ✦**Homicide:** A person, directly or indirectly, by any means, causes the death of another person.
  ✦**Mischief:** A person commits mischief that willfully causes malicious destruction, damage, or defacement of property. This also includes any public mischief towards another person.
  ✦**Offence Against a Person:** An attack on a person causing harm that may include usage of a weapon.
  ✦**Other Theft:** Theft of property that includes personal items (purse, wallet, cellphone, laptop, etc.), bicycle, etc.
  ✦**Theft from Vehicle:** Theft of property from a vehicle.
  ✦**Theft of Vehicle:** Theft of a vehicle, motorcycle, or any motor vehicle.
  ✦**Theft of Bicycle:** Theft of a bicycle.
  Note: There is no neighbourhood information.
- **YEAR:** A four-digit field that indicates the year when the reported crime activity occurred.
- **MONTH:** A numeric field that indicates the month when the reported crime activity occurred.
- **DAY:** A two-digit field that indicates the day of the month when the reported crime activity occurred.
- **HOUR:** A two-digit field that indicates the hour time (in 24 hours format) when the reported crime activity occurred.
  Note: This information is based on the findings of the police investigation. No time information will be provided for Offences Against a Person crime type.
- **MINUTE:** A two-digit field that indicates the minute when the reported crime activity occurred.
  Note: This information is based on the findings of the police investigation. No time information will be provided for Offences Against a Person crime type.
- **HUNDRED_BLOCK:** Generalized location of the report crime activity.
  Note: Locations for reported incidents involving Offences Against a Person have been deliberately randomized to several blocks and offset to an intersection. No time or street location name will be provided for these offences. For property related offences, the VPD has provided the location to the hundred block of these incidents within the general area of the

block. All data must be considered offset and users should not interpret any locations as related to a specific person or specific property.
Coordinates data for records with “Offset to Protect Privacy” was not disclosed to provide privacy protection.
X NK_LOC ST is default location value used for incidents with unknown location and is geolocated to 312 Main Street.

- **NEIGHBOURHOOD:** The Vancouver Police Department uses the Statistics Canada definition of neighbourhoods within municipalities. Neighbourhoods within the City of Vancouver are based on the census tract (CT) concept within census metropolitan area(CMA).

Note: The Musqueam Indian Band is located in the southwest corner of the City of Vancouver. There is a service agreement between Musqueam and the City of Vancouver, where the City provides municipal services such as policing. As a result, Musqueam crime data is included with the VPD Open Data.

- **X:** Coordinate values are projected in UTM Zone 10. All data must be considered offset and users should not interpret any locations as related to a specific person or specific property.
- **Y:** Coordinate values are projected in UTM Zone 10. All data must be considered offset and users should not interpret any locations as related to a specific person or specific property.

Methods and Tools

Data Preprocessing

Four columns had missing values in our dataset, namely: HUNDRED_BLOCK, NEIGHBOURHOOD, X and Y. The number of missing values for the features HUNDRED_BLOCK, X and Y was negligible with respect to the size of the dataset. Thus we decided to eliminate those instances. Further, the Feature NEIGHBOURHOOD had 64574 missing values in the Dataset. The imputation strategy used for filling these values was KNN algorithm which imputed the neighbourhood values based on the X & Y coordinates of the feature. The distance measure used for calculating distance between the coordinates is Haversine Distance. Haversine distance is used here instead of Euclidean distance because it takes the spherical surface of the earth into consideration.The value of k was set to be five and the imputation was done by majority voting from the nearest five neighbours.

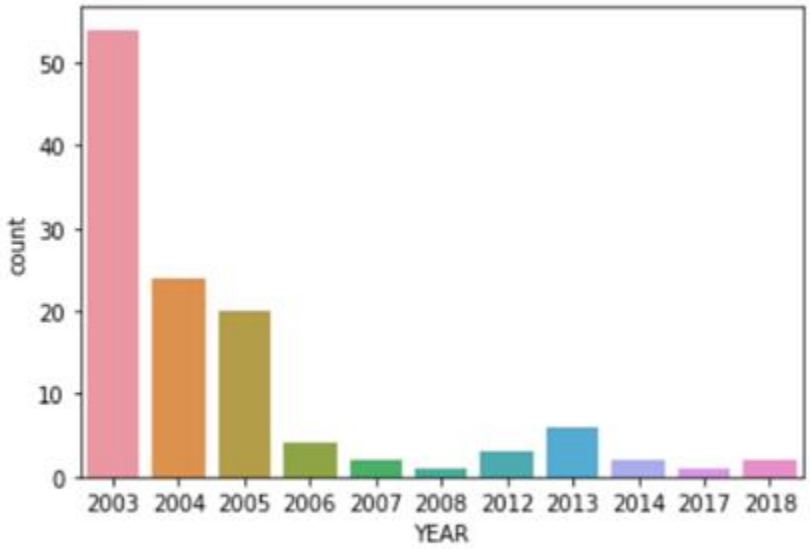
After completing the preprocessing step, Feature Engineering was performed where new features were added to the Dataset. The features that were added were Holiday and Weekday. The Holiday Feature has binary values (0 and 1) which states whether the given day is a holiday or not. The Weekday Feature provides information about the day of the week for a given date. The target feature had a total of 11 class values i.e 11 different types of crime to be identified. We grouped them into four total classes, Theft, B&E, Mischief and SevereCrime. This grouping was done based on the domain knowledge and some research into the crime categories.

Exploratory Data Analysis

1. Total number of missing values in the Dataset

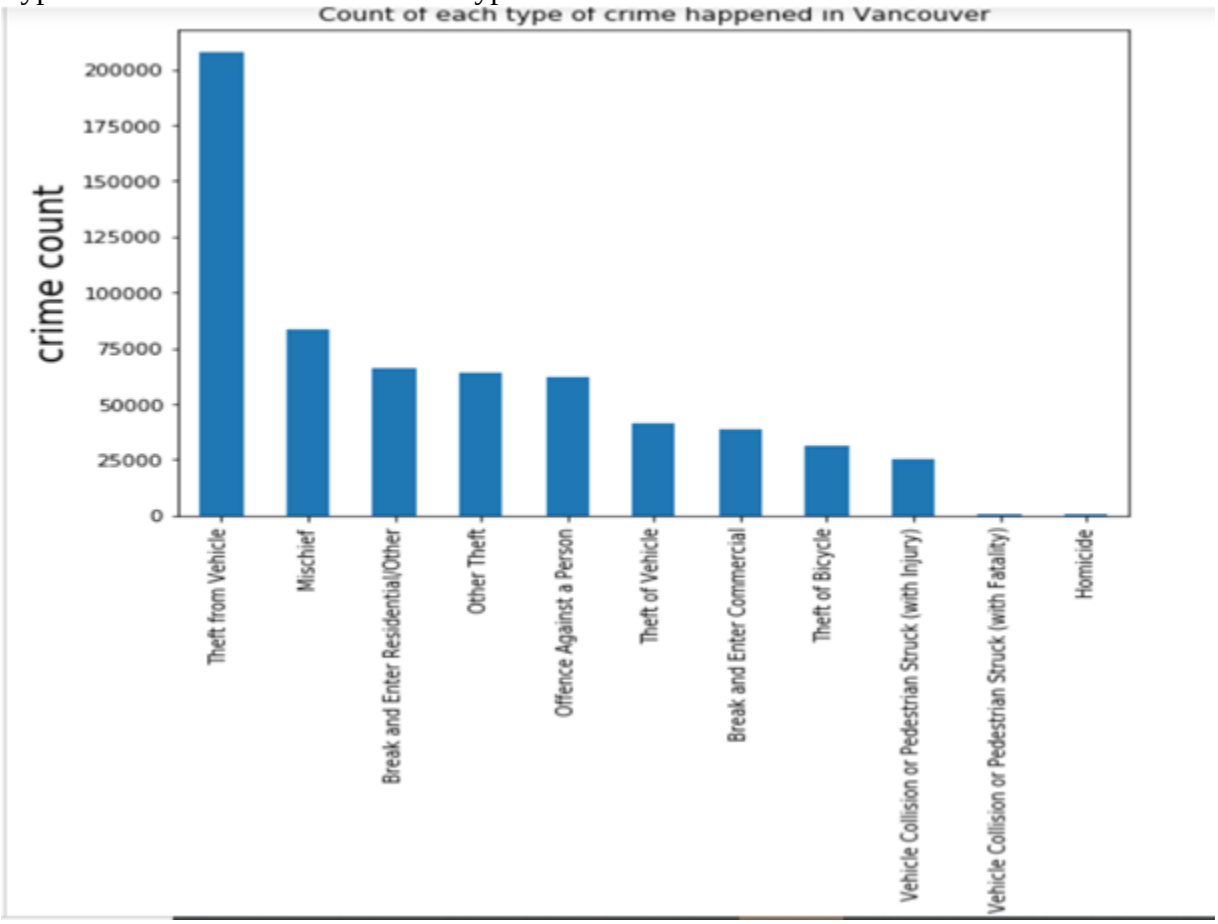
```
TYPE                0
YEAR                0
MONTH               0
DAY                 0
HOUR                0
MINUTE              0
HUNDRED_BLOCK      13
NEIGHBOURHOOD   64574
X                 119
Y                 119
dtype: int64
```

2. Number of missing values for each year in the Dataset

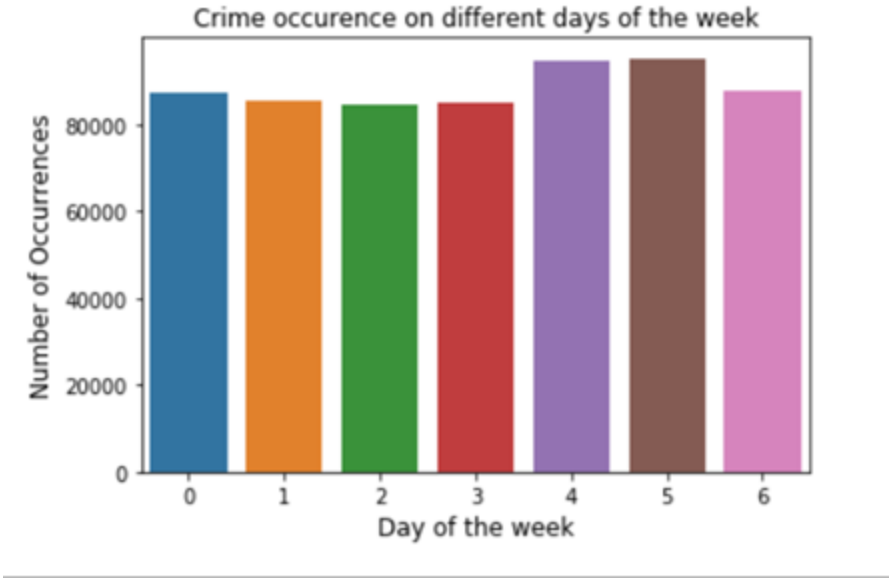


This shows that most of the missing coordinate values are in year 2003 (almost half)

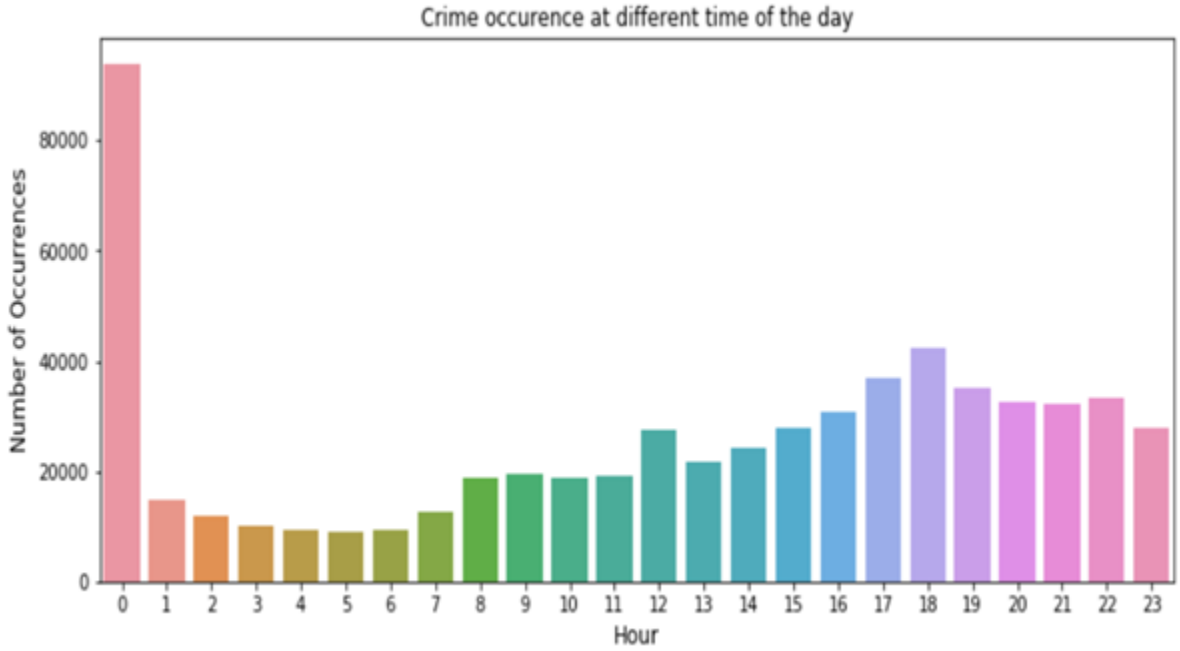
3. Types of Crime with count for each type



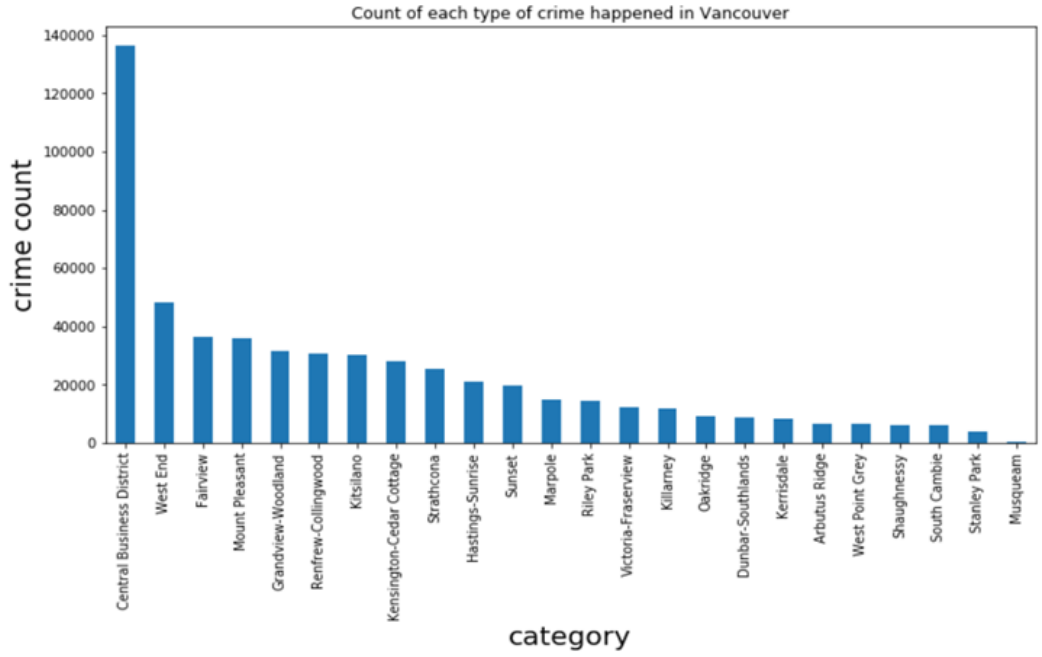
4. Crime Occurrence on different days of the week



5. Crime Occurrence during different time of the day



6. Number of Crimes Reported in each Neighbourhood



7. Absolute Change in crime over the years 2003-2019

Classification methods

As discussed in the previous section the target variable for our classification model is Crime Type. This feature has four class labels which were stated in the previous section. The classification was performed with two different methods Random Forest Classifier and XGBoost Classifier. The experiments were performed both with and without the feature selection techniques. Random forests are an ensemble learning method for classification and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) of the individual trees. XGBoost (Extreme Gradient Boosting) belongs to a family of boosting algorithms and uses the gradient boosting (GBM) framework at its core. Boosting is a sequential technique which works on the principle of an ensemble. It combines a set of weak learners and delivers improved prediction accuracy. At any instant t, the model outcomes are weighed based on the outcomes of previous instant t-1. The outcomes predicted correctly are given a lower weight and the ones miss-classified are weighted higher. Rather than training all of the models in isolation of one another, boosting trains models in succession, with each new model being trained to correct the errors made by the previous ones. Models are added sequentially until no further improvements can be made.
Note: The categorical features in the dataset have been encoded and labeled to fit into the model. We have used default parameters for both of the models.

For the forecasting part we try to predict the most probable location for the given datetime, neighbourhood and crime type. Our goal is to predict the coordinates based on our previous reported instances. The main issue is we cannot predict X and Y coordinates separately as the

coordinates are not independent themselves hence we cannot use two different models to predict them. Therefore, we need to combine both the coordinates into a single feature which can be predicted by the model. Hence, we use a technique known as geohashing to encode the coordinates into a single feature.Geohashing is a geocoding method used to encode geographic coordinates (latitude and longitude) into a short string of digits and letters delineating an area on a map, which is called a cell, with varying resolutions. The more characters in the string, the more precise the location. We then use a random forest model trained on specific neighbourhood data to predict the most probable crime location.

Validation Method

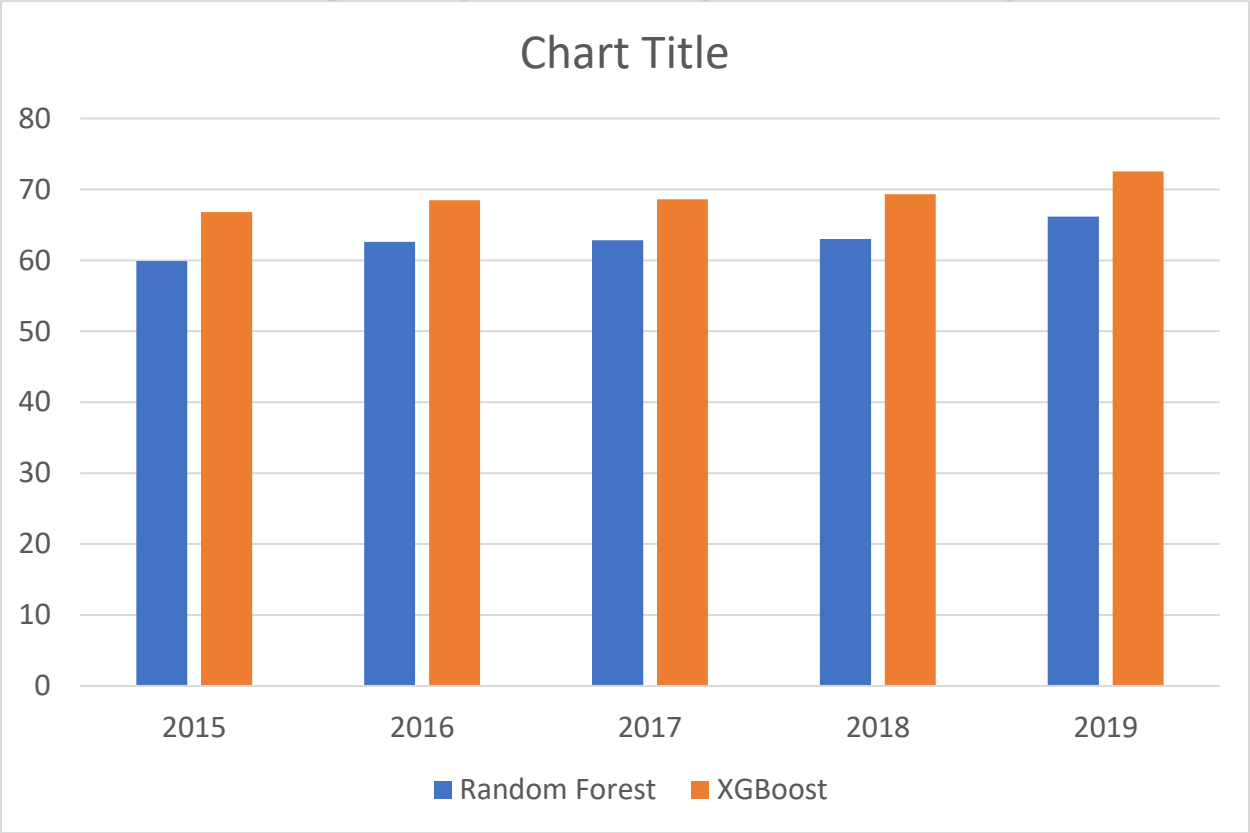
The splitting of data into training and testing parts cannot be done randomly, thus k-fold cross validation cannot be used for validation purpose. This is because we try to predict the crime type happening for a future date and other set of features. Hence, we treat our dataset as time series and take testing steps accordingly. The training and testing sets have been separated by performing the following steps:
Step-1: Sort the dataset with respect to the Date Feature
Step-2: Take the data from the year 2003 to 2014 as the training set for predicting the crime in the year 2015.
Step-3: Repeat the above step for all the subsequent years after 2015 and add the previous years data to the training set.

This method will provide us with different testing accuracies for the years 2015 - 2019. We take an average of these accuracies which can be considered as the accuracy for a trained model.

The location prediction model generates a map as an output which has probable locations where there is a high probability of crime.

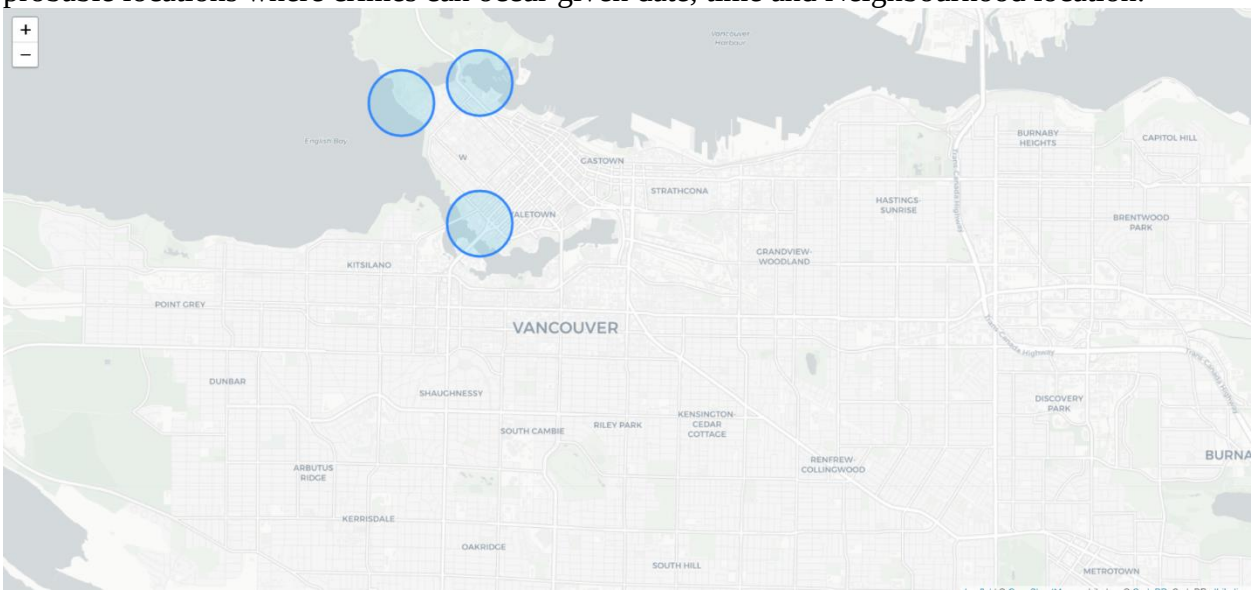
Results

1. The results obtained after performing the Validation steps are as shown in the figure below:



These results show that the XGBoost Classifier performs better than the Random Forest Classifier. It can also be observed that the crime prediction accuracy increases from the year 2015 to 2019. The reason behind this is that as we move towards the current year the training dataset for the model keeps on increasing. Thus the model learns better from the data and gives better results. The model takes the following parameters as input: 'YEAR', 'MONTH', 'DAY', 'HOUR','WEEKDAY', 'NEIGHBOURHOOD','HOLIDAY'. Thus by providing new information about these features, type of crime can be predicted at a specific location.

2. The following figure shows a map of Vancouver city where the circles highlight the probable locations where crimes can occur given date, time and Neighbourhood location.



Conclusion and Future Work:

In this project, we did analysis and prediction of crime on the Vancouver crime Dataset. The analysis and insights gathered from the data are as described in the Exploratory analysis section. Data Preprocessing was also an important part of our analysis. The Classification methods used in the project for predicting Crime type and location are standard Machine Learning algorithms. The results obtained provides us with the information which the police department can act upon to keep the neighbourhoods safe from criminal activities. Future work for this project can be headed in the direction of using Deep Network Models for classification purpose. LSTM-RNN can be an interesting approach for this dataset as we treat the dataset as a Time-Series one. Other Time-Series analysis methods can be applied and their results can be compared to find their performance with respect to each other.

References:

[1] https://www.irjet.net/archives/V5/i9/IRJET-V5I9192.pdf
[2] https://github.com/datascience-ninja/Toronto-Crime-Analysis
[3] Kim S, Joshi P, Kalsi PS, Taheri P. Crime Analysis Through Machine Learning. In2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) 2018 Nov 1 (pp. 415-420). IEEE.
[4] https://data.vancouver.ca/datacatalogue/crime-data.html
[5] Rotton J, Frey J. Air pollution, weather, and violent crimes: Concomitant time-series analysis of archival data. Journal of personality and social psychology. 1985 Nov;49(5):1207.
[6] Bogomolov A, Lepri B, Staiano J, Oliver N, Pianesi F, Pentland A. Once upon a crime: towards crime prediction from demographics and mobile data. InProceedings of the 16th international conference on multimodal interaction 2014 Nov 12 (pp. 427-434). ACM.
[7] Mohler G. Marked point process hotspot maps for homicide and gun crime prediction in Chicago. International Journal of Forecasting. 2014 Jul 1;30(3):491-7.
[8] Iqbal R, Murad MA, Mustapha A, Panahy PH, Khanahmadliravi N. An experimental study of classification algorithms for crime prediction. Indian Journal of Science and Technology. 2013 Mar 1;6(3):4219-25.
[9] Almanie T, Mirza R, Lor E. Crime prediction based on crime types and using spatial and temporal criminal hotspots. arXiv preprint arXiv:1508.02050. 2015 Aug 9.

[10] Nasridinov A, Ihm SY, Park YH. A decision tree-based classification model for crime prediction. InInformation Technology Convergence 2013 (pp. 531-538). Springer, Dordrecht.
[11] Chen H, Chung W, Xu JJ, Wang G, Qin Y, Chau M. Crime data mining: a general framework and some examples. computer. 2004 Apr 1(4):50-6.
[12] Sathyadevan S, Gangadharan S. Crime analysis and prediction using data mining. In2014 First International Conference on Networks & Soft Computing (ICNSC2014) 2014 Aug 19 (pp. 406-412). IEEE.

Appendix:

The program contains 3 file

1. Exploratory_Analysis.ipynb - contains exploratory data analysis which was done on the dataset
2. Crime_Prediction.ipynb - contains the crime type prediction with Random Forest and XGBoost
3. Location_Prediction.ipynb - contains the location prediction with Random Forest.

The files are jupyter notebook files hence require jupyter notebook to be installed on your system.
link : https://www.anaconda.com/distribution/

To execute the program successfully you will need the following packages to be installed in your environment
- Numpy
- Pandas
- Scikit learn
- Matplotlib
- Seaborn
- Plotly
- holidays
- xgboost
- pyproj
- folium
- geohash_hilbert

Steps to run:

1. Open Anaconda Navigator
2. Start jupyter notebook or jupyter labs
3. open the three files
4. Follow cell by cell execution in a sequential manner
5. The output of each cell will be generated and shown below it