CSCI 6612 Visual Analytics

CRIME IN VANCOUVER

PROJECT REPORT

DATA MINERS

Aakash Patel [ak553155@dal.ca]

Sampark Pradhan [sm459977@dal.ca]

Bhumi Patel [bh842792@dal.ca]

Ganrong Tan [gn485145@dal.ca]

CONTENTS

1 Abstract

We studied the intricacies of Crime Dataset obtained for Vancouver in detail as part of the project to get insights, trends and patterns of crime happening in the city. It's aimed to visualise and get a better picture of crime data with the help of powerful visualizations which foster advanced analytical decisions. With multiple ways of visualizing crime information, we present charts that improve cognitive mental model of its users to derive crisp and clear understanding of the data. The interactivity of the charts enable the users to analyze sub-section of the data in no time. The idea is based on the vision to provide efficient quick insights and analytics not only to the legal authorities but to the layman person as well. We also applied machine learning model to the data which can be used to predict crime occurrence with a fair amount of accuracy. This intellect one gets is critical, as it relates to the lives of each and every person in the city. Lastly we tried to justify model predictions through model interpretability, which helped us to explain model decisions with the help of features in favour and those opposing it.

2 Introduction

Crime is one of the major issues of the modern world because of technological advancements and population growth [6]. Along with the difference between the rich and the poor continues to widen, the number of crimes is showing a dramatically increasing in the past few years, and it starts to affect the economic development and people's daily life, which people have to take more actions to protect themselves away from the unexpected risks. Crime can be divided into different types such as crime against properties like theft, burglary, and robbery, and crime of aggregation which are homicides, assaults, and rape. With the increasing number of crimes, crime analysis has become one of the vital techniques to measure and predict the risks, based on the analytic result, policies or the government can take the necessary steps to reduce the risk of crime in different areas. Crime analysis has two important components, one is quantitative methods, and the other is qualitative methods [6]. Qualitative methods are more focusing on the profile or environment scanning, but quantitative methods more rely on the machine learning approach to predict the underlying risk and use a visualization approach to visualize the results. In this project, quantitative methods will be discussed.

3 Background

3.1 Project Objective

In order to start working on crime analysis, we collected a Vancouver crime dataset from the open data of the Vancouver Police Department (VPD). This crime dataset has 530,652 records with different types of crimes, and the time crossed from 2003 to 2017. Vancouver, located in the Province of British Columbia, as one of the biggest and one of the highest population density cities in Canada, is a perfect city sample to work with.

The aim of this project is to discover which type of crimes has the highest frequency in the past thirteen-years and where each type of crime happened, and we will combine with the bar chart, pie chart, and heat map to visualize them. Other than that, the machine learning algorithm helps us to identify which type of crime has the highest possibility to occur and where is the high-risk crime zones in the following year. The prediction result is likely to provide the government a better insight of this city and will also benefit to VPD to dispatch their officers more efficiently and deliver a feasible plan for next year.

4 Project Details

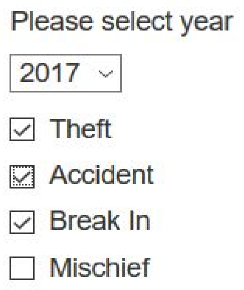
4.1 Data Preprocessing

The dataset is well structured and doesn’t contain much missing or error values, which is a huge benefit to us since we do not have to spend much time on this step. However, there were still a few things we did to the dataset. Firstly, we removed 4187 crime records involved in two types of crimes which were an offense against a person and homicide, because the detailed information had been wiped out by the law enforcement for the protection of the victim's privacy. Secondly, we concatenated all time values (Year, Month, Day, Hour, and Minutes) and geographic coordinates (Latitude and Longitude) together. We took advantage of that when doing the time series analysis and the crime occurrence in Vancouver next year.

4.2 Crime Visualization and EDA

4.2.1 Crime Trends Data Visualization

Crime Trends Visualization is mainly aimed to give its users, quick trends of crime in the city. The filter of crime type and year enable the user to see various types of crime taking place over a particular year and its distribution in months. Users can understand the crime count year after year and also analyze seasonal variation in data. Crime Type filter is to allow the user to view the prevalence of all the crime types happening over the year. Law Enforcement agencies can formulate new laws/rules to curb a particular type of crime majorly occurring, as suggested by this visualization. For instance, formulating traffic laws in case of the prevalence of “Accident” type crime in the city.



Please select year

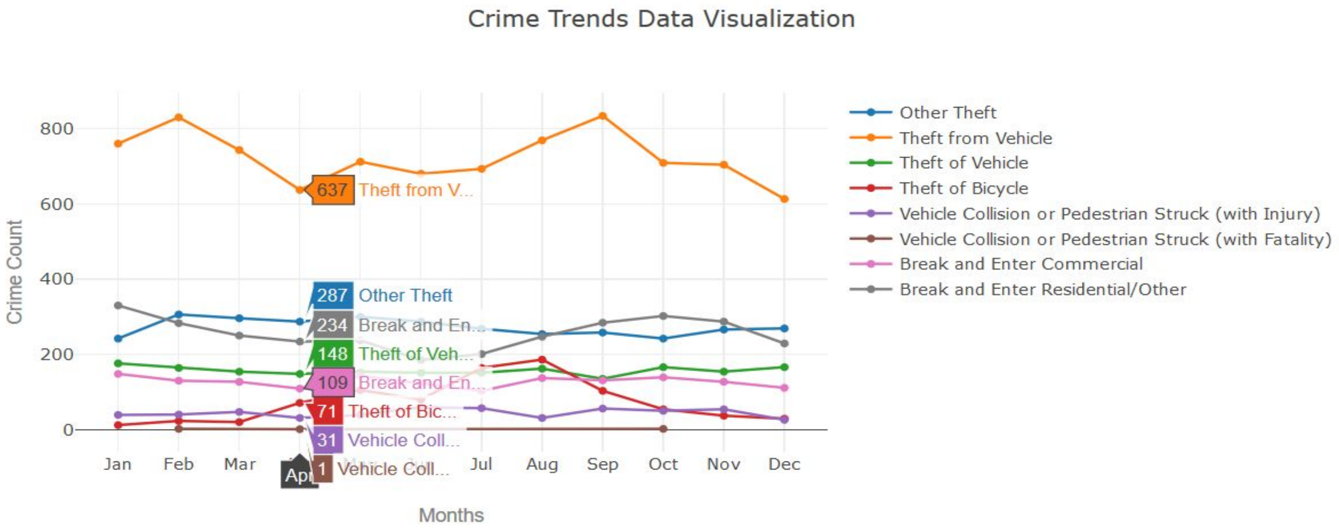
2017

☑ Theft

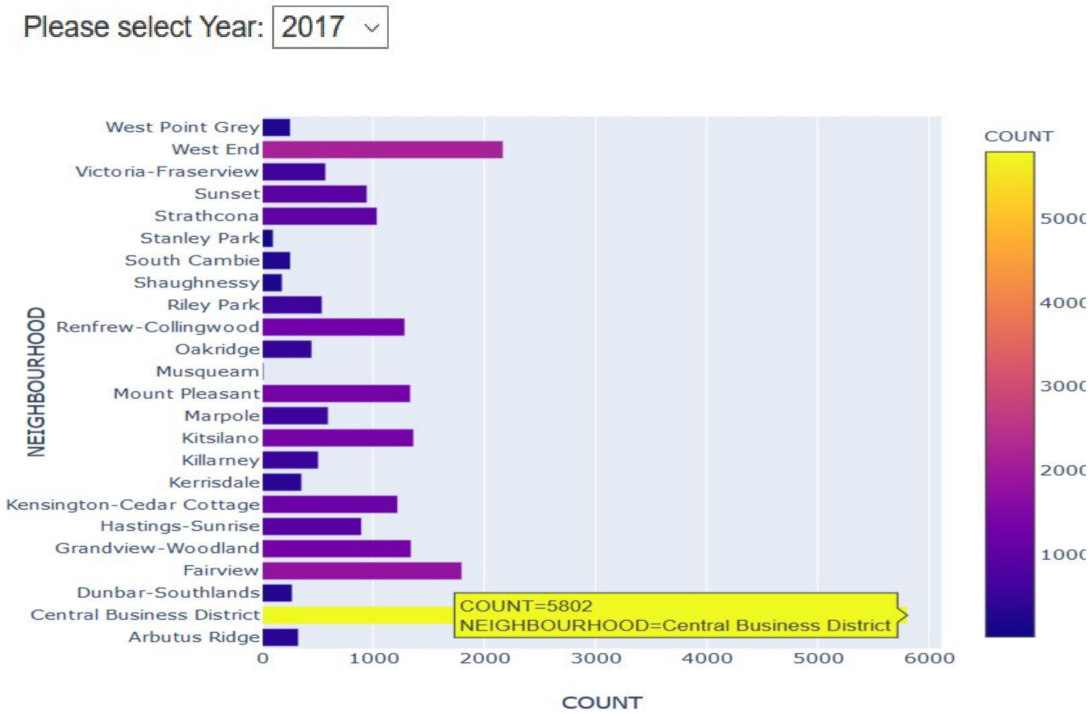
☑ Accident

☑ Break In

□ Mischief



4.2.2 Neighbourhood Data Visualization



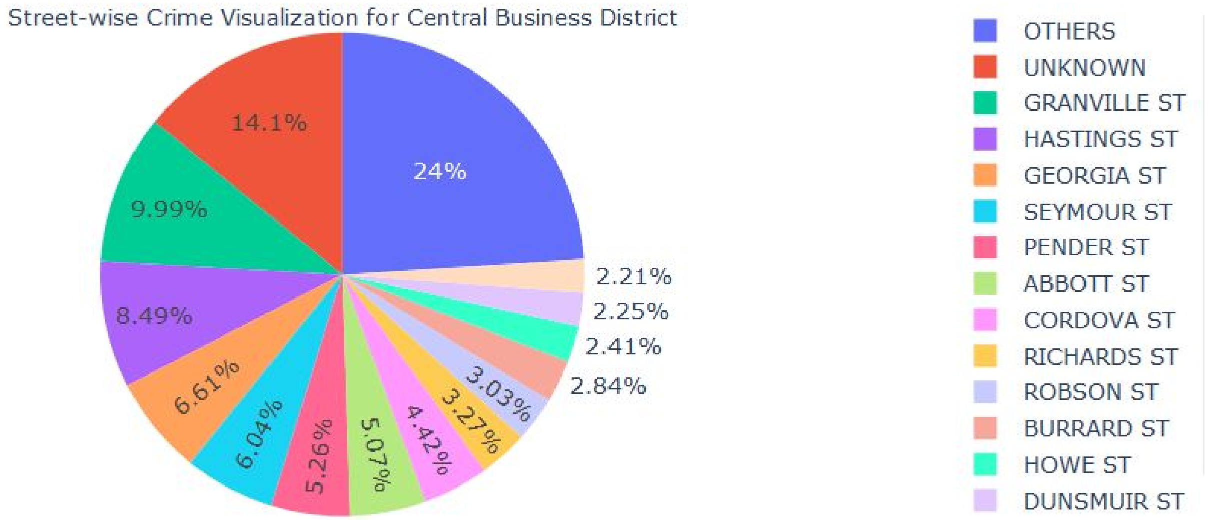
Please select Year: 2017

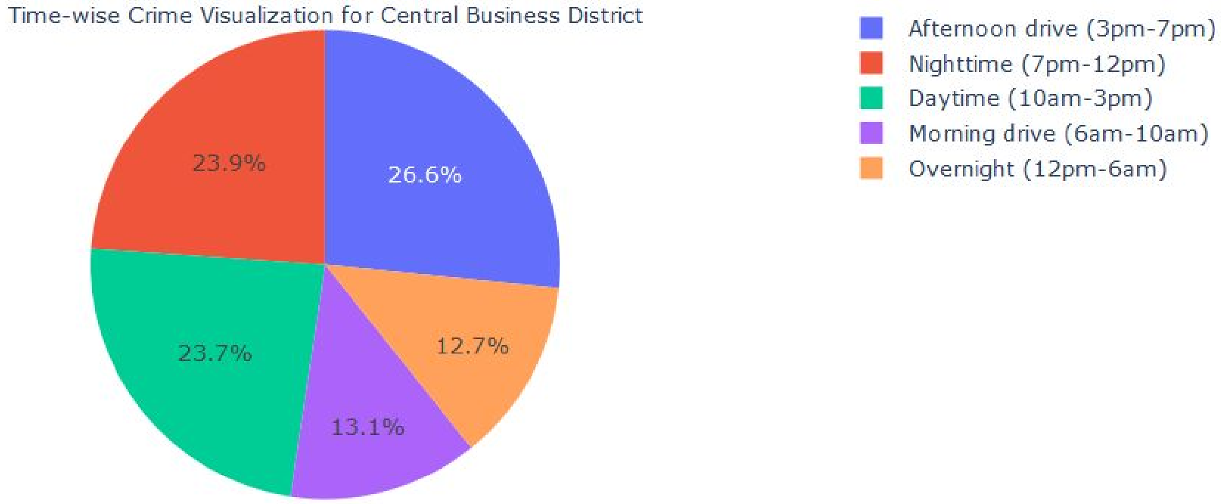
COUNT=5802
NEIGHBOURHOOD=Central Business District

After being able to view the crime count of a particular type of particular duration one may want to understand how exactly is crime distributed in major areas of the city. This visualization helps us achieve this objective. It just helps us visualize which neighborhood contributes to the most number of crimes occurring in the city. Clicking on any “Neighbourhood” column displays further details in the other two visualizations, namely streetwise crime distribution and time-wise crime distribution.

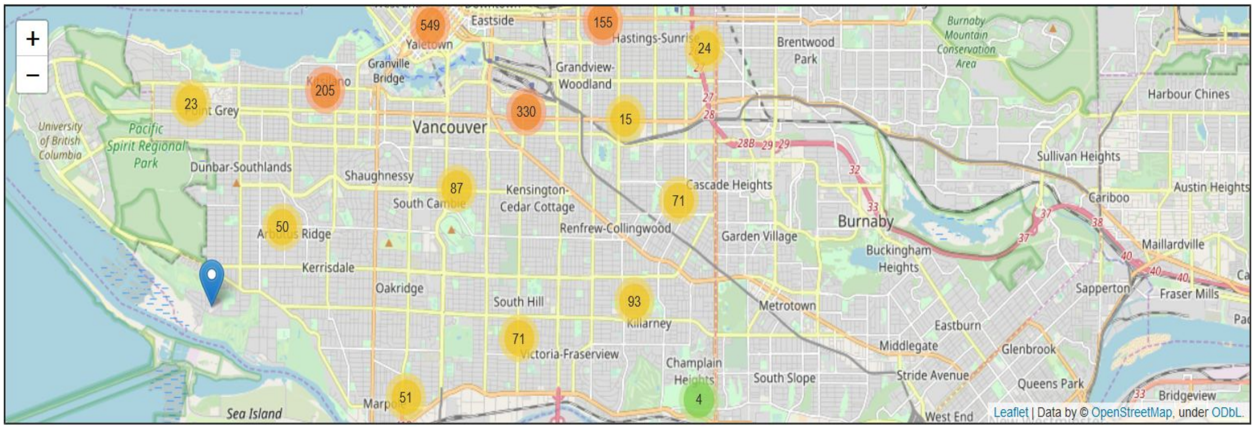
Understanding Streetwise crime distribution further helps authorities pinpoint to the streets of major interest, by indicating potential crime hit streets. This type of visualization can thus indicate any unlawful activity (like drug smuggling) going on, in the particular street under study.

Knowing the crime location alone can too be very abstract or unfruitful in terms of controlling the crime. This is where timewise distribution of crime in a particular neighborhood, comes in handy. Authorities can understand which time of the day has the maximum probability of crime occurrence. A major contribution to the crime of a particular part of the day like midnight, for instance, can indicate the authorities when to increase patrolling and security of the area and thus optimizing the availability of their workforce.



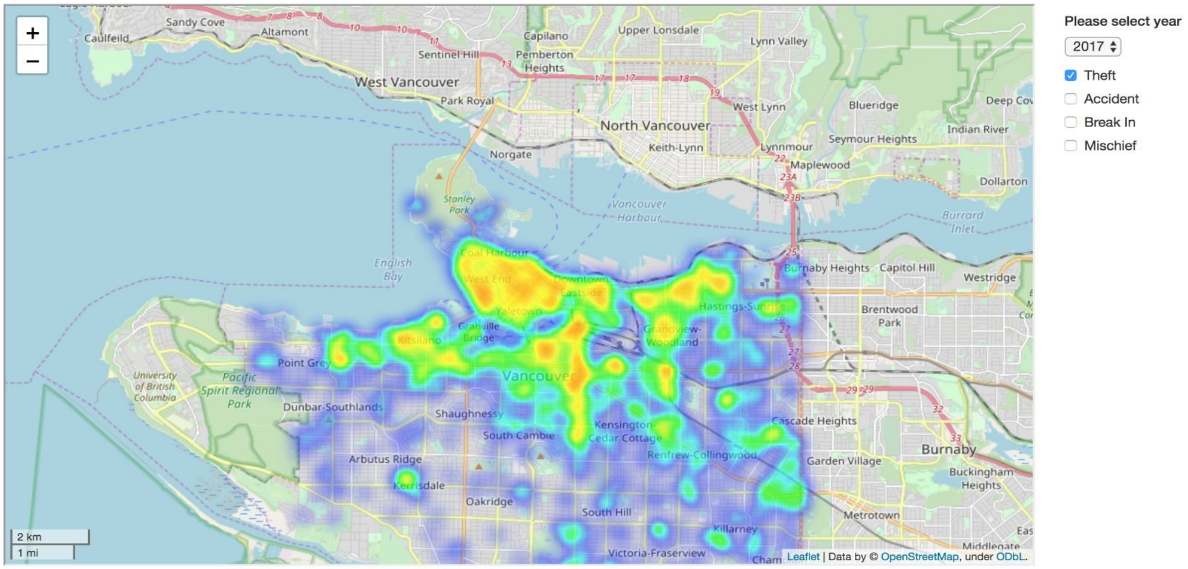


4.2.3 Crime Zones Data Visualization



The main objective of the crime zones data visualization is to chart the exact geospatial locations of crime onto a map to derive insights from it. With the help of this kind of geospatial visualization of crime locations not only the clustered crime zones with their counts are displayed, but also the correlations of crime occurrence in two zones. An investigative team of Crime Experts can relate to the crime occurred in a particular area, which later escalated to a major event. For instance, the recent accumulation of protestors in Hong Kong to protest against the Chinese government. [3]

4.2.4 Crime Severity Visualization



Please select year

2017

☑ Theft

☐ Accident

☐ Break In

☐ Mischief

The main objective of Heat Map visualization is to enable users to visually identify crime hot spots and zones in the city. It provides a very abstract picture of crime in the city. The user can just pan in to get further information about a particular area of the city. As naturally suggested by the redness of the particular area, authorities can straight away focus onto a particular zone/area of the city prior to even looking at other plotted visualizations our product puts forward. We also implemented the filter based on year and crime type suggested to us in demo thus allowing the user to selectively see crime intensive areas based on selection.

4.3 Crime Prediction and Analysis

4.3.1 Machine Learning Approach

Our problem statement was to correctly classify the different types of crime that can occur at a given place in Vancouver. We tried with many classifiers such as Logistic Regression, SVM, XGBoost, RandomForest, MLPClassifier, etc. We moved forward with Random Forest because it gave us better accuracy. Random Forest is an ensemble classifier that combines Breima's idea of bagging and random selection of features. This creates a forest of decision trees and output's the class that is most voted by individual trees. The Random Forest implementation in scikit uses CART trees.[10] The advantages behind using Random forest are : -

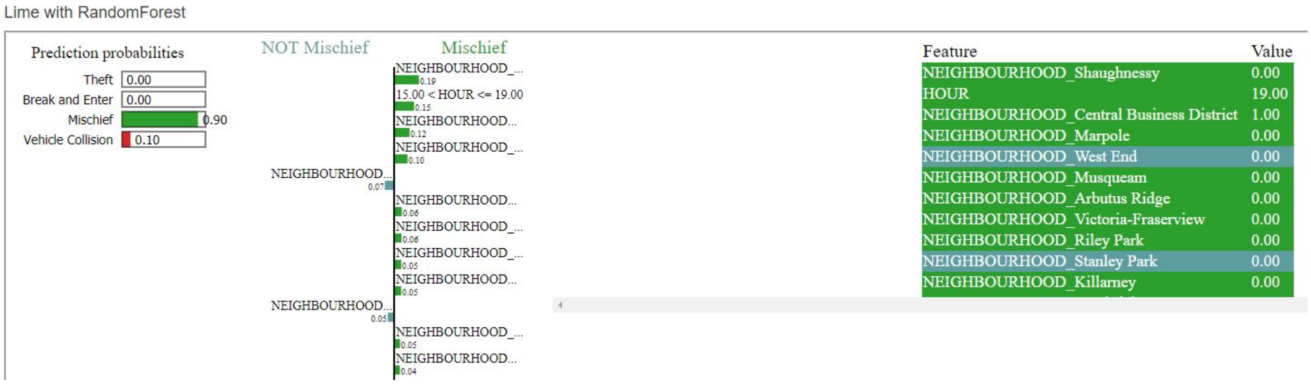
- We had around 4,50,000 records to train the classifier and random forest runs efficiently on large datasets such as ours.
- The random forest has got methods in balancing errors in class population unbalanced data sets.
- Random forest generates an internal unbiased estimate of the generalization error as the forest building progresses.
- It is fast to build, predict and is fully parallelizable.
- It also offers us the instability feature i.e, if we change the data a little, the individual tree may change but the forest is relatively stable as it is the combination of many trees.

We achieved an accuracy of 62.57% while predicting crime types in Vancouver.

4.3.2 Evaluation Metric

We have different “CRIME TYPE” labels to classify our data against. But in the context of our data, we're only concerned with crime occurrence if any, without taking the Crime Type into consideration. We thus took Accuracy as our model evaluation metric. As we know F1-Score metric is only used when we want to understand the underlying Crime Types participating in classification, it seemed to be unsuitable for our objective of just visualizing if the crime is happening or not.

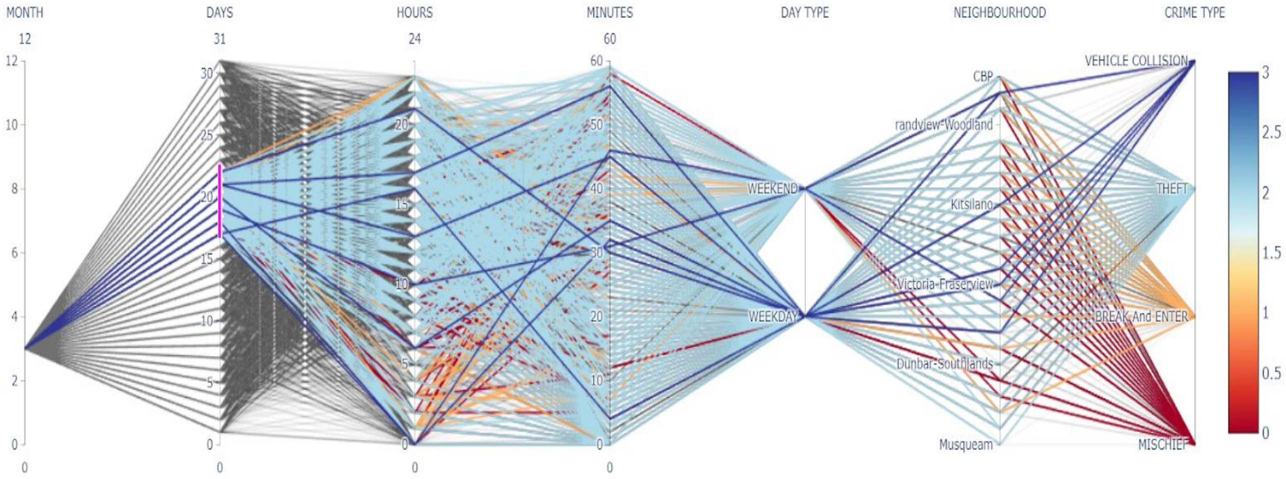
4.3.3 Model Interpretability using LIME



Feature	Value
NEIGHBOURHOOD_Shaughnessy	0.00
HOUR	19.00
NEIGHBOURHOOD_Central Business District	1.00
NEIGHBOURHOOD_Marpole	0.00
NEIGHBOURHOOD_West End	0.00
NEIGHBOURHOOD_Musqueam	0.00
NEIGHBOURHOOD_Arbutus Ridge	0.00
NEIGHBOURHOOD_Victoria-Fraserview	0.00
NEIGHBOURHOOD_Riley Park	0.00
NEIGHBOURHOOD_Stanley Park	0.00
NEIGHBOURHOOD_Killarney	0.00

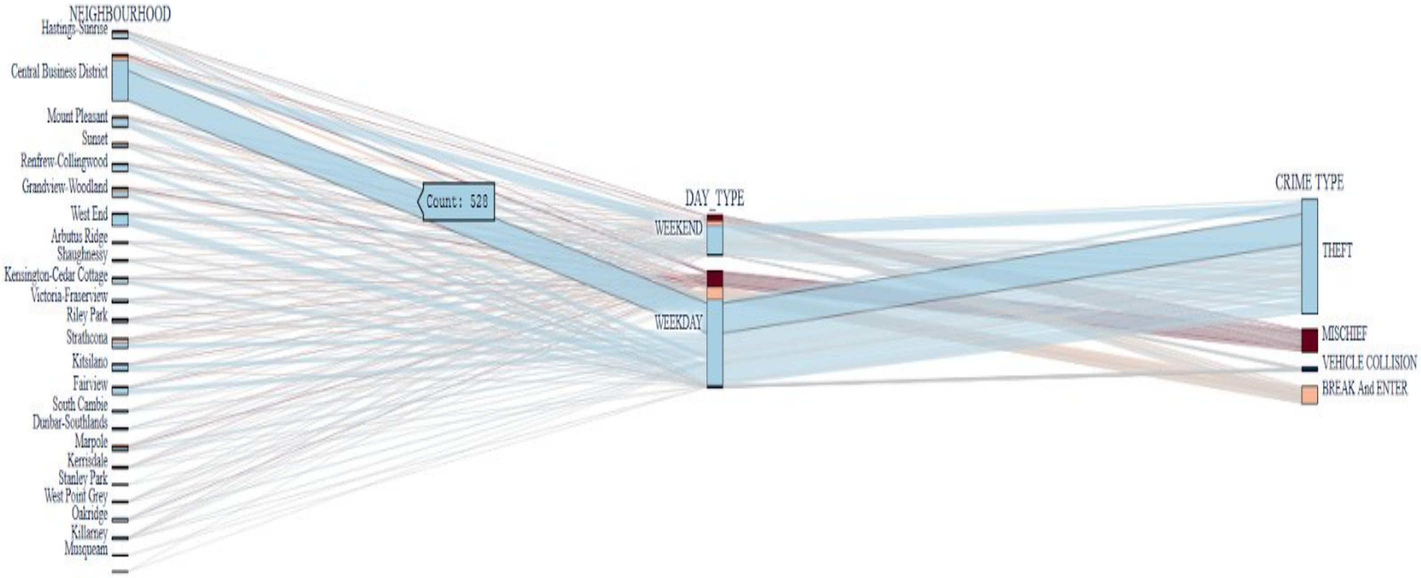
LIME (Local Interpretable Model-agnostic Explanations) stands for explaining partial parts of an unknown model, and it is a tool that can help us to understand and explain how a complex machine learning model makes a decision [13]. More specifically, the explanation given by LIME is from a local sample, the current input data, or by establishing a local model to predict the most important features. This provides us a basic understanding of which features contribute more to a sample and which features contradicted it. In our case, mischief has the highest probability happening next year, and the hour contributes the highest weight in contrast to other features.

4.3.4 Parallel Coordinates



Having essentially a multi-dimensional time-varying data in our case, we looked into ways of neutralizing the complexity by having a better visual approach to understand correlation among various attributes in data. With the use of a Parallel Coordinate System, users can better visualize how attributes of data are varying to finally being related to crime in the city. The user has the flexibility to switch between the sequence of parallel axis and select range in every dimension to perform a detailed analysis of its impact on the CRIME TYPE attribute.

4.3.5 Sankey Diagram



With a lot of data being charted at one place, we visualized the neighborhood crime occurrences according to DAY_TYPE plotted against the crime type. Sankey Diagrams like these particularly showcase dominant contributors stand out by differentiating them in various colors [4]. Users have the benefit of viewing through multiple levels. Users can get a high-level view, see specific details or generate interactive views by hovering through or dragging DAY_TYPE, CRIME TYPE and NEIGHBOURHOOD attributes. These diagrams allow showcasing complex processes visually while allowing users to focus on a single aspect or resource that he/she may want to highlight.

5 Challenges and Bottlenecks

- Working with Dash to render web app.
- Charting every graph into single webapp.
- Integrating global intractability across all the charts.
- Overcoming Limitations of Dash for loading large amounts of data.
- Deployment across cloud options.
- Porting everything in Flask and Bringing everything together.

- Depicting Modelling Visualization into appropriate format.
- Optimizing on Modelling time efficiency.
- Transforming Vbox widget to integrate in Flask Frontend.

6 Conclusion

In this project, we introduced a high-level visualization approach, parallel coordinates, to visualize the future crime activities that predicted by random forest.

With the parallel coordinates graph attached above, the law enforcement and the city will be able to the establishment and implement counterplan to potential high risk-zone of next year more intuitively. The past year’s crime plots can also provide an insight into law enforcement when reviewing the past crime data.

7 Future Work

Not just for law enforcement and the city, we plan to make the results in public to everyone who is interested. The traveler or the local residents will benefit from our results because they can avoid traveling to a certain area or take necessary precautions if it is a necessary trip. Moreover, we plan to combine with the census data in Vancouver and the crime dataset to discover the relationship between crime rate and house prices in different communities. One last thing is to predict the possibility of crime occurrence on Holiday and discover whether or not a holiday is playing a major factor in crime occurrence.

8 References

[1] "Flask", *PyPI*, 2019. [Online]. Available: https://pypi.org/project/Flask/. [Accessed: 23-Nov- 2019].

[2] "Plotly Python Graphing Library", *Plot.ly*, 2019. [Online]. Available: https://plot.ly/python/. [Accessed: 29- Nov- 2019].

[3] "China rebukes US for backing Hong Kong protesters", *BBC News*, 2019. [Online]. Available: https://www.bbc.com/news/world-asia-china-50584928. [Accessed: 29- Nov- 2019].

[4] "The What, Why, and How of Sankey Diagrams", *Medium*, 2019. [Online]. Available: https://towardsdatascience.com/the-what-why-and-how-of-sankey-diagrams-430cbd4980b5. [Accessed: 29- Nov- 2019].

[5] Jayaweera, I., Sajeewa, C., Liyanage, S., Wijewardane, T., Perera, I. and Wijayasiri, A. (2015). Crime analytics: Analysis of crimes through newspaper articles. 2015 Moratuwa Engineering Research Conference (MERCon).

[6] Shamsuddin, N., Ali, N. and Alwee, R. (2017). An overview on crime prediction methods. 2017 6th ICT International Student Project Conference (ICT-ISPC).

[7]"Parallel Coordinates Plot", *Plot.ly*, 2019. [Online]. Available: https://plot.ly/python/parallel-coordinates-plot/. [Accessed: 01- Dec- 2019].

[8]"Parallel Categories Diagram", *Plot.ly*, 2019. [Online]. Available: https://plot.ly/python/parallel-categories-diagram/. [Accessed: 01- Dec- 2019].

[9]"3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.21.3 documentation", *Scikit-learn.org*, 2019. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Accessed: 01- Dec- 2019].

[10]"Why Random Forest is My Favorite Machine Learning Model", *Medium*, 2019. [Online]. Available: https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706. [Accessed: 01- Dec- 2019].

[11]"Model Interpretation Strategies", *Medium*, 2019. [Online]. Available: https://towardsdatascience.com/explainable-artificial-intelligence-part-2-model-interpretation-strategies-75d4afa6b739. [Accessed: 01- Dec- 2019].

[12]"Understanding model predictions with LIME", *Medium*, 2019. [Online]. Available: https://towardsdatascience.com/understanding-model-predictions-with-lime-a582fdff3a3b. [Accessed: 01- Dec- 2019].

[13]Elsinghorst, S. and Elsinghorst, S. (2019). Looking beyond accuracy to improve trust in machine learning - codecentric AG Blog. [online] codecentric AG Blog. Available at: https://blog.codecentric.de/en/2018/01/look-beyond-accuracy-improve-trust-machine-learning/?utm_content=65461834&utm_medium=social&utm_source=twitter [Accessed 1 Dec. 2019].

[14] Sidr. (2019). Retrieved 2 December 2019, from https://www.berriart.com/sidr/.(This reference was used to develop web page slider functionality and css code).