TORONTO CRIME ANALYSIS

By

Jils Joseph

Shivangi Sheth

Sachin Jacob

Abstract

The objective of this work is to analyse past data on crime for the city of Toronto. Insights from the data can be used to help the police become proactive and reduce the crime rate. The findings can also help civilians become more aware of the crime hotspots in the city so they can stay safe. In this project we use various visualizations to break down the data into useful insights, with the help of popular time series techniques and using machine learning we will forecast crime rate and predict where potential crimes are more likely to occur in future.

1 Introduction

Crime is a major problem in major cities, it’s been around since the beginning of time. In this project we use past data from Toronto police open data portal to come up with insights about the major crimes taking place in Toronto. We can also visualize the data into graphs and tables which are easier to understand.

The Toronto police have been tracking crime for numerous years. We have past data from all the way dating to the late 90’s, however the older seems to be incomplete and vague so for the purpose if this project we are working with data from 2014 to 2019. The dataset mainly consists of the Date on which the crime took place along with the location, Crime type and other crime related attributes. Toronto police have categorised Crime into five major Types. Assault, Break and Enter, Robbery, Theft Over and Auto theft. All crimes are subsets of these five.

Using various visualization techniques, we transformed the data into comprehensible graphs and interactive maps so the data could be easily interpreted. Many statistical techniques were used for understanding trend and seasonality of the data. Finally, we used Machine learning to build models to predict crimes and Deep learning to predict future crime rates.

2 Related Work

Many papers were published based on crime analysis, however most of them focused on exploratory data analysis. There were a handful of papers which used Machine learning along with exploratory data analysis.

- A study was conducted for the city of Chicago where the researcher analysed past data to predict crimes in Chicago. The study mainly used multi label classification problems to run the models. A binary model was initially constructed using one Crime type, later a multiclass model was built using four major crimes. The accuracy of the multiclass model was finally compared against the binary model. In conclusion the binary model seems to do a far better job than the multiclass model.

- Another paper analysed the crime in Vancouver city. The crimes were broken down into five categories. The study concluded that Assault was the most common crime type and mostly too place on the weekends.

- Another study focused on using data mining from social media and news to achieve data about crimes. The data was stored in a database from it was fed into a Naïve Bayes model to learn keywords about specific crime types so the model could learn unknown inputs. Naïve Bayes was preferred because it performs well for smaller datasets.

3 Methods

The dataset was downloaded from Toronto police open data portal. The dataset had record dating from 1998 to 2019. We filtered the data from 2014 to 2019 for the purpose of the project. We have around 16000 rows and 15 columns. There were no null or missing values. Each record includes a date, crime type and area where the crime took place.

Initially we visualized the data using Tableau, crimes were broken down Yearly, Monthly, Weekly, Daily and hourly to attain insights. Crimes were also compared against Temperature data to understand the relationship between Crime rate and Temperature, turns out temperature has a direct impact on criminal activities.

Secondly, we removed unnecessary columns, created a new column called frequency which essentially contained the total number of crimes that took place per day. After that columns were scaled using Minmax scaling and encoded using one hot encoding and split into Test and Train sets. Modelling was the next step with “Neighbourhood” as the target variable.

We plotted an ARIMA model for time series analysis. First, we started with taking two columns date and FC (frequency count). Plot the time series using the Data Frame’s. plot () method. Run Dicky-Fuller test. Import ACF and PCF functions and plot their graphs for analysis weather they are stationary or not. Import the SARIMAX model using loop method fit SARIMA model. Create the 4 diagnostics plots. At last plot the mean predictions to find the best pdq values.

Logistic regression, Random Forest, Decision Tree and KNN models were implemented initially without any optimization. After evaluating the models, we realized that precision and recall were really low for all the models because of a lack of records for two crime categories which brought the overall accuracy down so in order to increase the accuracy we ran the models again using the top three crime categories. This improved the accuracy of all the models substantially. After that each model was separately optimized, example hyper parameter tuning was done on Random Forest to estimate the number of n_estimators. An LSTM model was also built by trying various optimisers and epochs to achieve the optimal model.

4 Modelling

A good model represents the dataset. We use different type of models and time series methods to predict the future, i.e. crimes which are going to occur in different neighbourhood.

Our modelling methods are: -

K-Nearest Neighbors (KNN): is the simplest and most used classification algorithm. It classifies based on the feature similarity. It uses the whole data to process which make it a high memory consuming and slow, as n value increases.

Logistic Regression: We use an ordinal logistic regression as we had more than three prediction categories. Hyperparameter tuning was performed on the regressor. The dataset was split into 70:30 and 80:20 ratio to run the model.

Random Forest: Random forest model gave us the best accuracy amongst the others, parameter tuning was performed to obtain the optimal number of n_estimator for building the forest and the criterion was set to “Gini”.

Long short-term memory (LSTM): is a recurring neural network. It has a design in which it prevents from the decay of the network. It is widely used because of its high-performance index. We can use different optimizer and epochs to tweak the output to get prediction

ARIMA: It is popular time series model. ARIMA means autoregressive integrated moving average. It is combination of two models AR model and MA model. There are different types of ARIMA to make it more accurate AUTO-ARIMA and SARIMA.

Decision tree: It is predictive model used for classification and regression.it has tree-like graph which basically do splitting, pruning and tree selection.it predict outcomes and identify target groups. The goal of a decision tree is encapsulating the training data in the smallest possible tree.

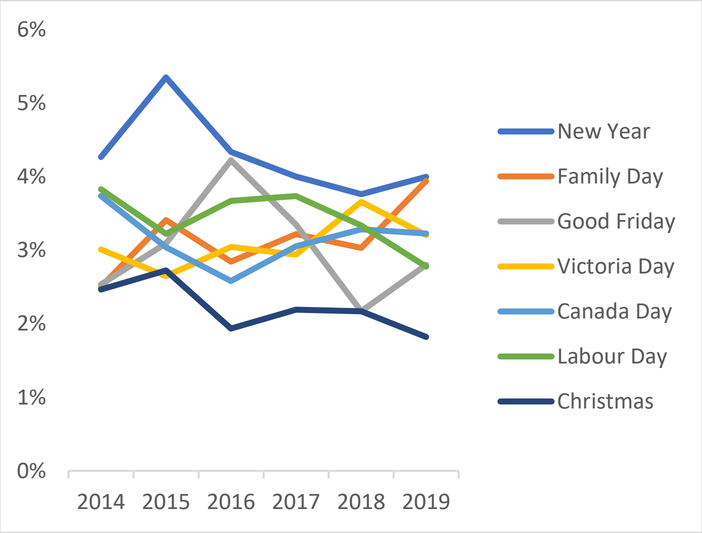
5 Results

model	Train Accuracy	Test Accuracy
Logistic Regression	61.3	58.6
Random Forest	73.2	69.5
Decision Tree	67.1	64.2
KNN	81	65

model	Crime type	precision	recall	f1-score
Random Forest	Assault	0.71	0.92	0.80
	Auto Theft	0.68	0.14	0.24
	Break and Enter	0.64	0.40	0.50
Logistic Regression	Assault	0.71	0.88	0.79
	Auto Theft	0.57	0.27	0.37
	Break and Enter	0.58	0.40	0.47
KNN	Assault	0.72	0.76	0.74
	Auto Theft	0.41	0.39	0.40
	Break and Enter	0.50	0.45	0.47

From the above tables we can see that Random Forest model have the best overall precision recall score. All three categories have good scores. We believe optimization have a positive result on the model, hence Random Forest can be chosen as the best model.

The relation between crime and event is a basic question when doing an analysis. We use default 7 holidays in Canada to verify our speculation. We see an upwards trend for crime during the years in our dataset. We take crime occurred at each holiday and compare it with the total crime for that month for the different years.



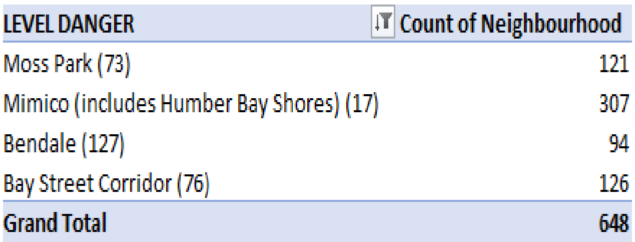
We cannot find any common trend with the data, throughout the years data shows both increasing and decreasing in percentage. So, the impact on these days to crime is non-existent.

Figure 1 shows the result of our calculation for the 7 holidays and the percentage of crime with respect to that month.

Multiple events

During the years we have multiple crime occurring in a single place. We were able to determine the thread levels of these areas by calculating the number of crimes occurred in a single place. We split theses into different criteria's as per the count of crime in that point.

Levels ranges are 0 as safe,1 to 9 as low,10 to 100 as medium,100 to 499 as high and danger from 500 and above



LEVEL DANGER	Count of Neighbourhood
Moss Park (73)	121
Mimico (includes Humber Bay Shores) (17)	307
Bendale (127)	94
Bay Street Corridor (76)	126
Grand Total	648

As we can see, although there is some similarity between top neighbourhood with crime to neighbour with danger, we cannot say that both these have a relation with each other.

Fig 2 shows the top Danger area in Toronto

The response time is the time taken by the police o response to an emergency. We calculate the response time by finding the difference between occurrence time and response time. From the data we were able to see the delay in response and how much it is. More than 10k records are inside response time which is half of our record.

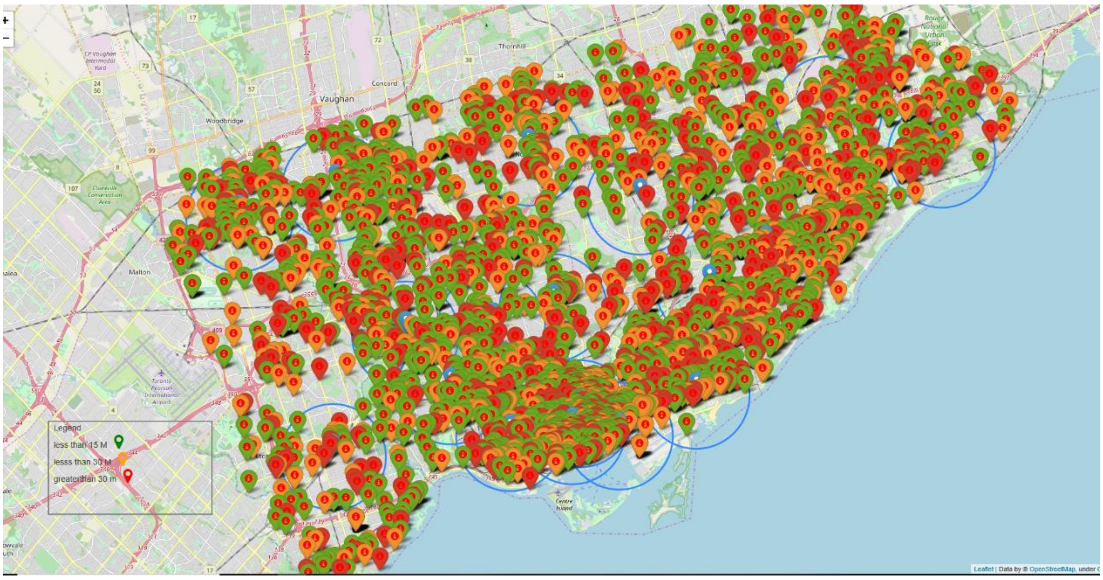


Fig: 3 Delay

We have plotted different divisions in the map and created a 5 km radius circle around it. We have plotted the points with different delays.

6 Discussion

Our main objective is predicting the crime occurring in different neighbourhood at different time and provide basic analysis of crime occurring in the city of Toronto. From the models we created we were able to get a model with an accuracy of more than 70%. We can say our model is good at predicting, so we are satisfied with the output.

We have different analysis for the data. We were able to find out that there is no correlation between holiday event and crime. From the data we were able to identify that Response time of Toronto police is lacking, the rate at which crime occurring in the city is increasing and crime is repeatedly occurring at certain places.

The whole project is affected with the case of low accuracy as there is little data for Robbery, Theft Over. So, in order to increase the accuracy, we are removing these two crimes. Except for KNN all our models predicted best by doing this.

Plotting the data by using different methods like matplotlib, seaborn, folium etc. Identifying the best methods for the plotting is one of the challenges. Due to high volume of data plotting into maps can cause confusion and are overcrowded.

7 Conclusion

In our exploratory data analysis, we discovered that there are mainly five types of crimes that take place in Toronto. We found out that Assault mostly takes place during 11am to 1pm. More than 80% of assaults took place during weekends. Another interesting observation was Temperature had a direct impact on the crime rate. February has the lowest crime rate throughout the year since its been recorded as the coldest month of the world, whereas crimes usually spike in summer months.

From our model evaluation we selected Random Forest as our best model since it gave us the best precision recall scores amongst the other models. We performed various optimization techniques such as hyperparameter tuning to improve the accuracy of the model. Other models such as LSTM failed to perform well due to the lack of computational power. We believe with more data and computational power we can achieve better results in future.

8 Contribution

Sachin Jacob:

- Reshaped the dataset by removing irrelevant columns.
- Built Random Forest, Logistic Regression and LSTM model on the data
- Optimized the models using various parameter tuning techniques.

Shivangi Sheth:

- Visualized the data using Tableau
- Built an Arima model to forecast the crime rate
- Constructed a Decision tree classifier and optimized using hyper parameter tuning.

Jils Joseph:

- KNN modelling with different neighbours.
- Analysis using excel for delay and event levels and using python to visualize and different analysis.
- Using FOLIUM package to plot different data's in maps.

9 Reference

1. https://open.toronto.ca/
2. https://github.com/Tbhangale/Chicago-Crime-Analysis
3. https://www.researchgate.net/publication/280722606_Crime_Analysis_and_Prediction_Using_Data_Mining
4. https://nbviewer.jupyter.org/github/python-visualization/folium_contrib/tree/master/
5. https://www.kaggle.com/daveianhickey/how-to-folium-for-maps-heatmaps-time-data/
6. https://www.cs.ubc.ca/~tmm/courses/547-17F/projects/alexandra-amon/report.pdf

10 Appendices

All the code that has been used in this project is uploaded on this GitHub page, please access the page to view/download the code.

https://github.com/karnevar/Toronto_Crime_ML_Analysis