Major Crime Incidence in Toronto, Ontario

By Matthew Chen, Mark Faynboym, Jasper Tsai

Today we will be presenting our preliminary report on the incidence of major crimes in Toronto Canada.

Data Description

- Source: Toronto Public Safety Data Portal
- Relevant predictor variables:
  - Neighborhoods (158)
  - Occurrence date and hour
  - Location type (home, apartment, etc..)
  - GPS longitude and latitude
- Main objective: learn different patterns of crime in Toronto

We sourced our data from the Toronto Police Department.

The data set has roughly 350,000 reported crimes recorded between 2014 and 2022.

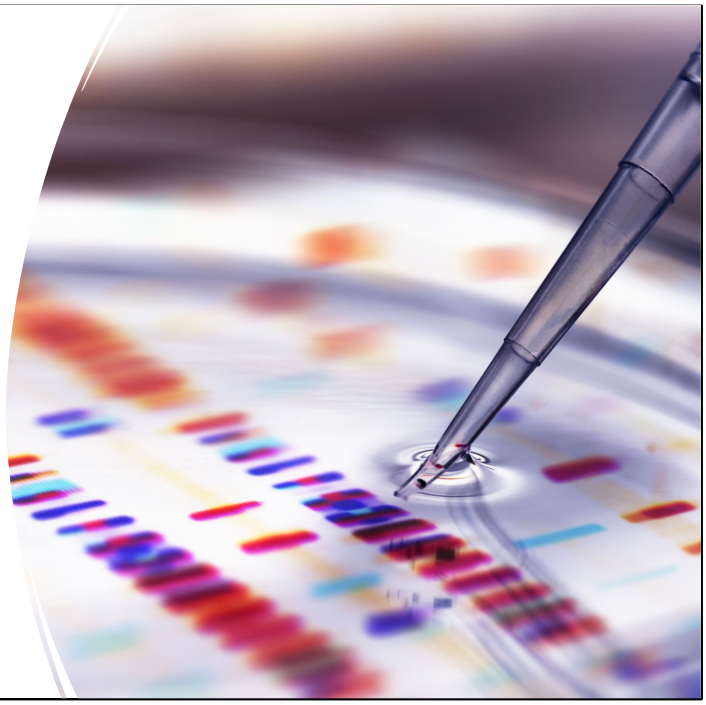
Other variables include 33 location type, 158 neighborhoods , nautical coordinates etc...

Toronto is the most populated city in Canada.

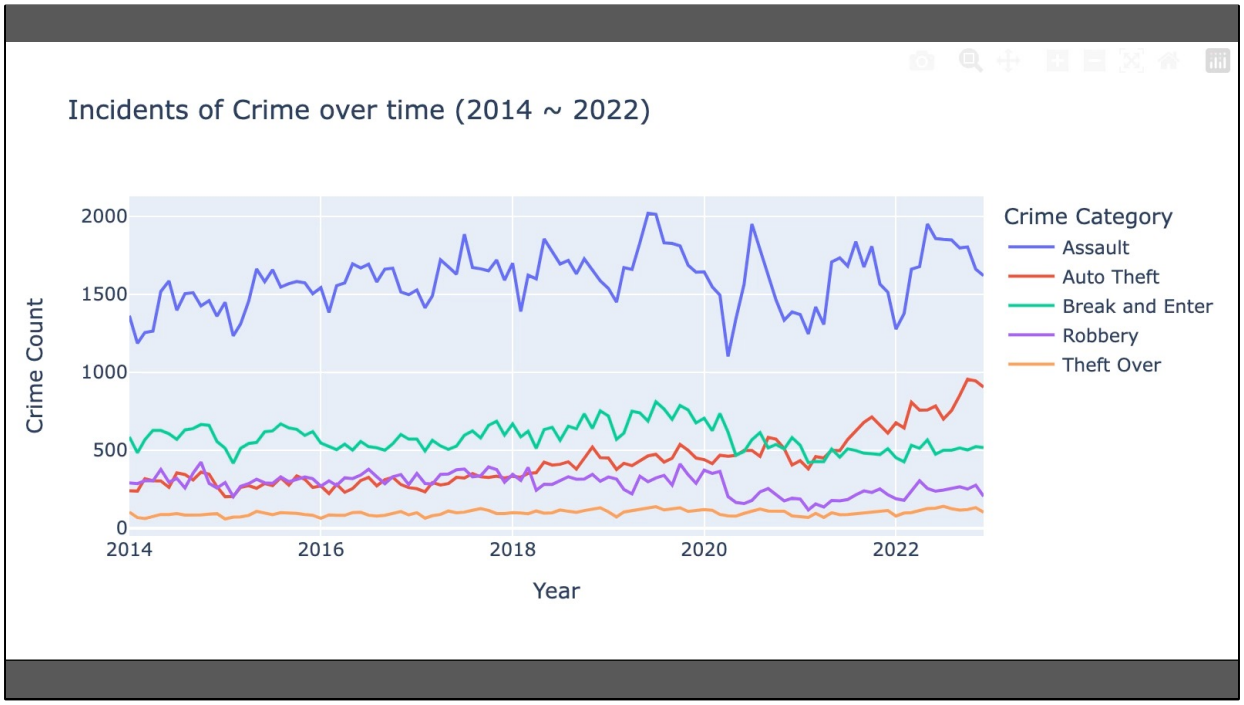
Projects like this are significant because analyzing crime data has many uses. It can help police departments gain insight into crime patterns or influence the real estate market as house prices get affected by local crime rates

We will explore further using EDA as well as some Machine Learning models to predict crime type based on time and location.

Exploratory Data Analysis



In this section we will get into some exploratory data analysis to help us understand the data better.



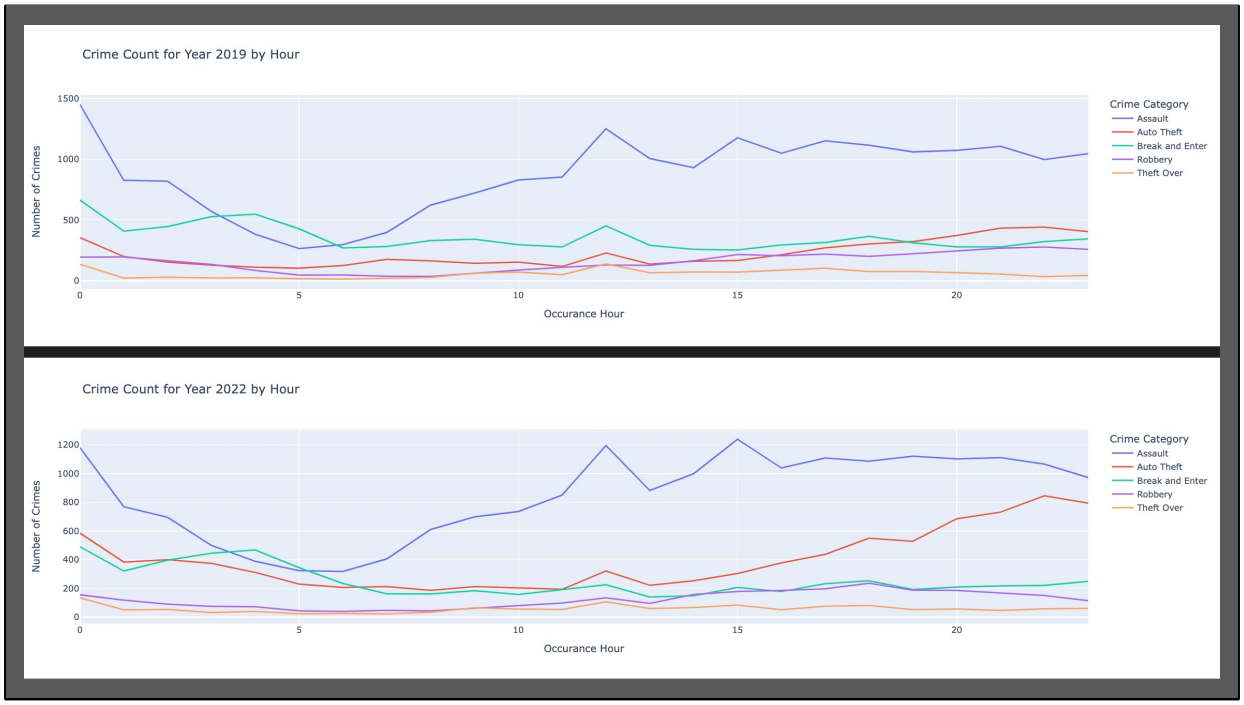
Let's begin with some Time Series observations.

We see Assault dominates the 5 categories of major crime.

We notice that the beginning of every year gives us the minimum amount of assaults.

What else we notice is that Break and Enter and Robbery noticeably drop after 2020, probably due to quarantine lockdowns.

Interestingly enough, Auto Theft surprisingly doubled.
This spike in auto thefts occurs at the same time as the chip shortage we experienced and the consequential spike in car prices so there may be some causation there.

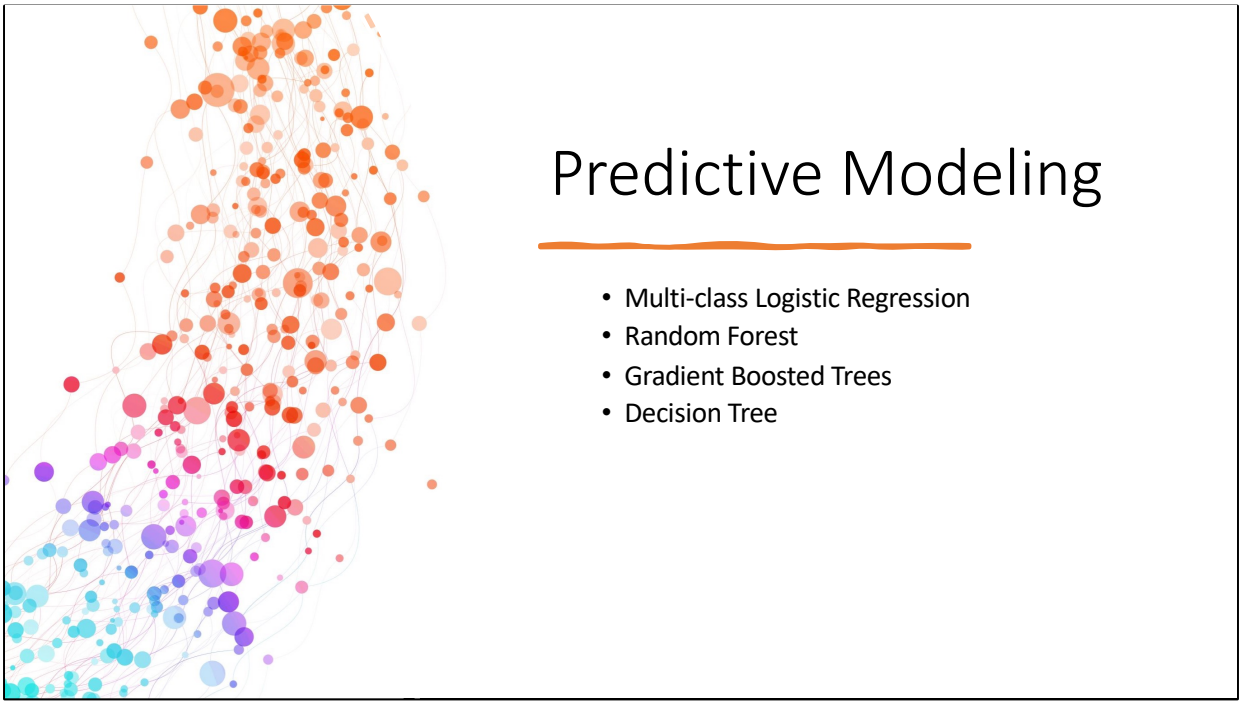


When plotting by hour in the years before and after COVID we see a significant increase in auto thefts after 3pm and an overall decrease in Breaking and Entering and Robbery overall. All else seems to be about the same

We move on to our heat map which gives us crime incidence by percentages and overall crime count.

When selecting for Assault, the map seems to indicate a very large incidence around Moss Park.

When selecting for Auto Theft, the map leads us to West-Humber Clairville.



Next, we move on to our predictive modeling section.
Specifically, we are trying to predict crime type based on the aforementioned variables about a crime's time and location.

To evaluate our model's performance, we focus on overall accuracy, precision, recall, and f1 score (which is the harmonic mean between precision and recall).

Multiclass Logistic Regression

Multi-logistic regression performance					
	Assault	Auto Theft	Break and Enter	Robbery	Theft Over
f1	0.733	0.533	0.442	0.184	0.001
precision	0.637	0.576	0.535	0.575	0.500
recall	0.862	0.497	0.377	0.109	0.000
accuracy			0.614		

As an initial model, we fit a multi-class logistic regression (i.e. multinomial loss)

This assumes a linear relationship between the log-ratio of probabilities from different classes and the predictor variables

This is unlikely to do well because of potential nonlinearities in the model

For example, crime predictions are likely to be nonlinear in GPS coordinates

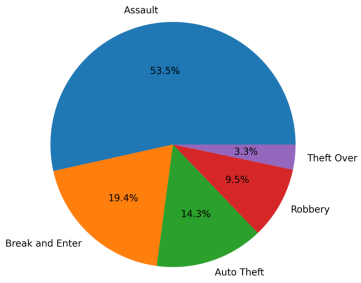
We begin with a multi-class logistic regression. We use this as a good initial model to try because it can fit more quickly than many other methods.
However, the drawback is that the model assumes a linear relationship between input features, which may not be the case for our data.

We hypothesize this because our prediction is based on Longitude and Latitude coordinates which we suspect to not be linear.

After fitting this model on the training data, we get an accuracy score of roughly 61% on the holdout test dataset. We will keep this in mind as a baseline compared to the other models that we will fit later.

Random Forest

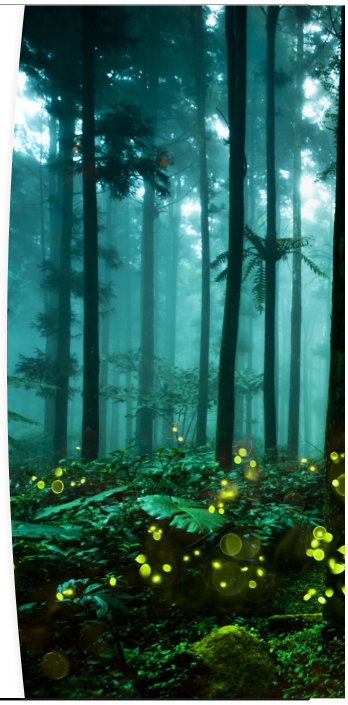
- Benefits of Random Forest:
  - Learns nonlinear relationships
  - Performs automatic feature selection
  - Not prone to overfitting
  - Robust to outliers



Random Forest Performance

	Assault	Auto Theft	Break and Enter	Robbery	Theft Over
f1	0.773	0.622	0.566	0.496	0.056
precision	0.706	0.644	0.633	0.654	0.201
recall	0.855	0.601	0.512	0.399	0.032
accuracy			0.680		

We have poor performance on the minority classes!



Next, we try fitting a random forest. Random Forest is a good model to try because it is capable of handling mixed data types, performs automatic feature selection, is robust to outliers, and discovers non-linear relationships.

Additionally, random forest is not prone to overfitting and doesn't require much hyper-parameter tuning, if at all.

Note that performance is poor for most crime types, except for assault. At the same time, the dataset is a bit unbalanced towards assaults as seen in the pie chart.

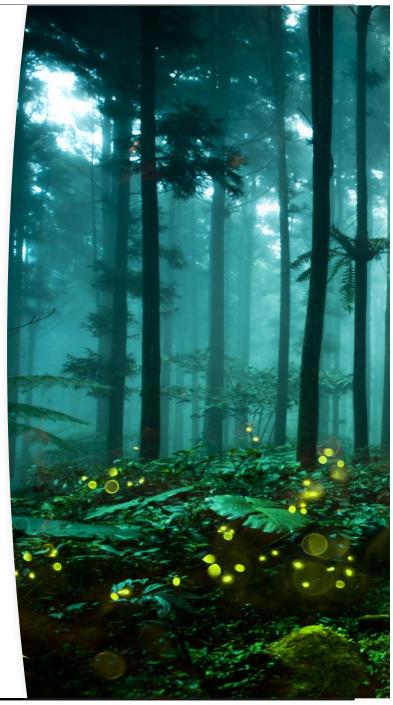
Weighted Random Forest

- To deal with the poor performance on the minority classes, we employ a weighted random forest
- We weight the impurity calculations for the constituent decision trees toward the minority classes
- Intuitively, this penalizes missclassifications on the minority classes more.

Weighted Random Forest Performance

	Assault	Auto Theft	Break and Enter	Robbery	Theft Over
f1	0.772	0.619	0.561	0.500	0.057
precision	0.701	0.647	0.637	0.657	0.206
recall	0.857	0.593	0.502	0.404	0.033
accuracy	0.679				

Unfortunately, we still have poor performance on the minority classes!



To deal with the poor performance on the minority classes, we employ a weighted random forest.

Essentially, by weighting the impurity calculations for the constituent decision trees toward the minority classes, we can penalize misclassifications on the minority classes more.

Unfortunately in this case we did not get any improvement in our accuracy.

Gradient Boosted Trees (xgboost)

- Gradient boosting may be a good algorithm to consider because of its sequential nature
- Each subsequent model focuses more on the previous model's misclassifications
- May help reduce bias for the minority classes

XGBoost Performance					
	Assault	Auto Theft	Break and Enter	Robbery	Theft Over
f1	0.771	0.617	0.587	0.372	0.069
precision	0.696	0.640	0.651	0.629	0.342
recall	0.866	0.595	0.534	0.264	0.038
accuracy	0.677				

Unfortunately, we still have poor performance on the minority classes!

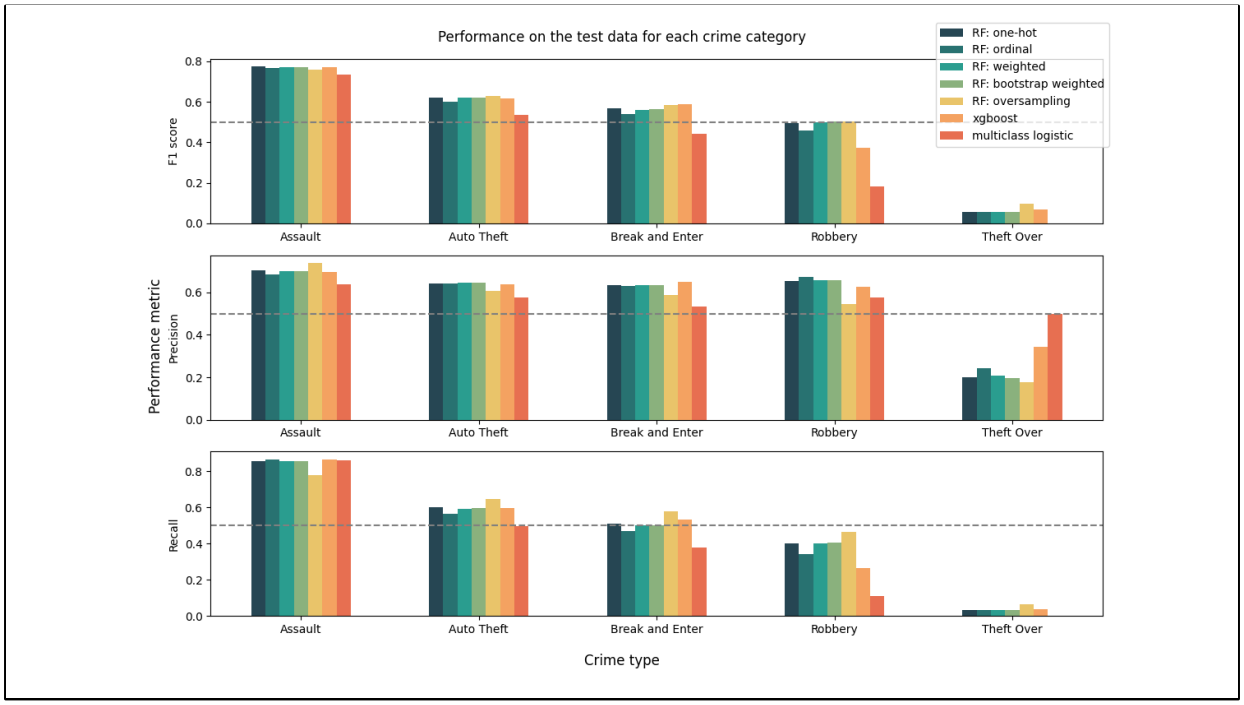
Here, we try a different algorithm to deal with the low performance on the minority classes.

Gradient boosting may be a good algorithm to consider because of its sequential nature, where each subsequent model focuses more on the previous model's misclassifications.

This is usually a good way to reduce bias, and in our case, it may help with the low performance on the minority classes.

For our implementation, we conducted hyperparameter tuning using a 5-fold cross validated random search on the given parameter space. After tuning the hyperparameters, we fit the overall model using the training set, and evaluate the performance on the test dataset.

The resulting performance is as shown (refer to table). Unfortunately, the performance on the minority classes are largely unchanged.



Here is a visualization of the performance on some of the models we tried. As we emphasized before, while overall accuracy is acceptable, we have poor performance on the minority classes as seen by the f1 score, precision, and recall values.

This is true even after we employed strategies to mitigate this like trying higher capacity models or
weighting the sample cases in the random forest. It is reasonable to conclude that our X variables just do not have enough discriminatory power to accurately predict the different
crime types that are not assaults.

Moving forward, we will try building a higher performance model for the binary classification case, specifically, predicting if a crime is an assault or not.

We aim to build a model with good interpretability so next we will fit a decision tree model as it is able to model nonlinearities unlike logistic regression yet still be interpretable at the same time.

References:

- Hastie, T., Tibshirani, R., & Friedman, J. (n.d.). Elements of statistical learning: Data mining, inference, and prediction. 2nd edition
- *Major crime indicators open data*. Toronto Police Service Public Safety Data Portal. (2023, March 31) https://data.torontopolice.on.ca/maps/major-crime-indicators-open-data
- Goodfellow, I., Courville, A., & Bengio, Y. (2017). *Deep learning*. Deep Learning