Assessment of Toronto Crime through Exploratory Data Analysis and Classification

Katherine Ault
501092397

- INTRODUCTION & PROJECT RATIONALE
- RESEARCH QUESTIONS
- PROJECT APPROACH
- EXPLORATORY DATA ANALYSIS
- FEATURE SELECTION & SMOTE OVERSAMPLING
- TRAINING, TESTING & MODEL SELECTION
- CONCLUSIONS
- RECOMMENDATIONS

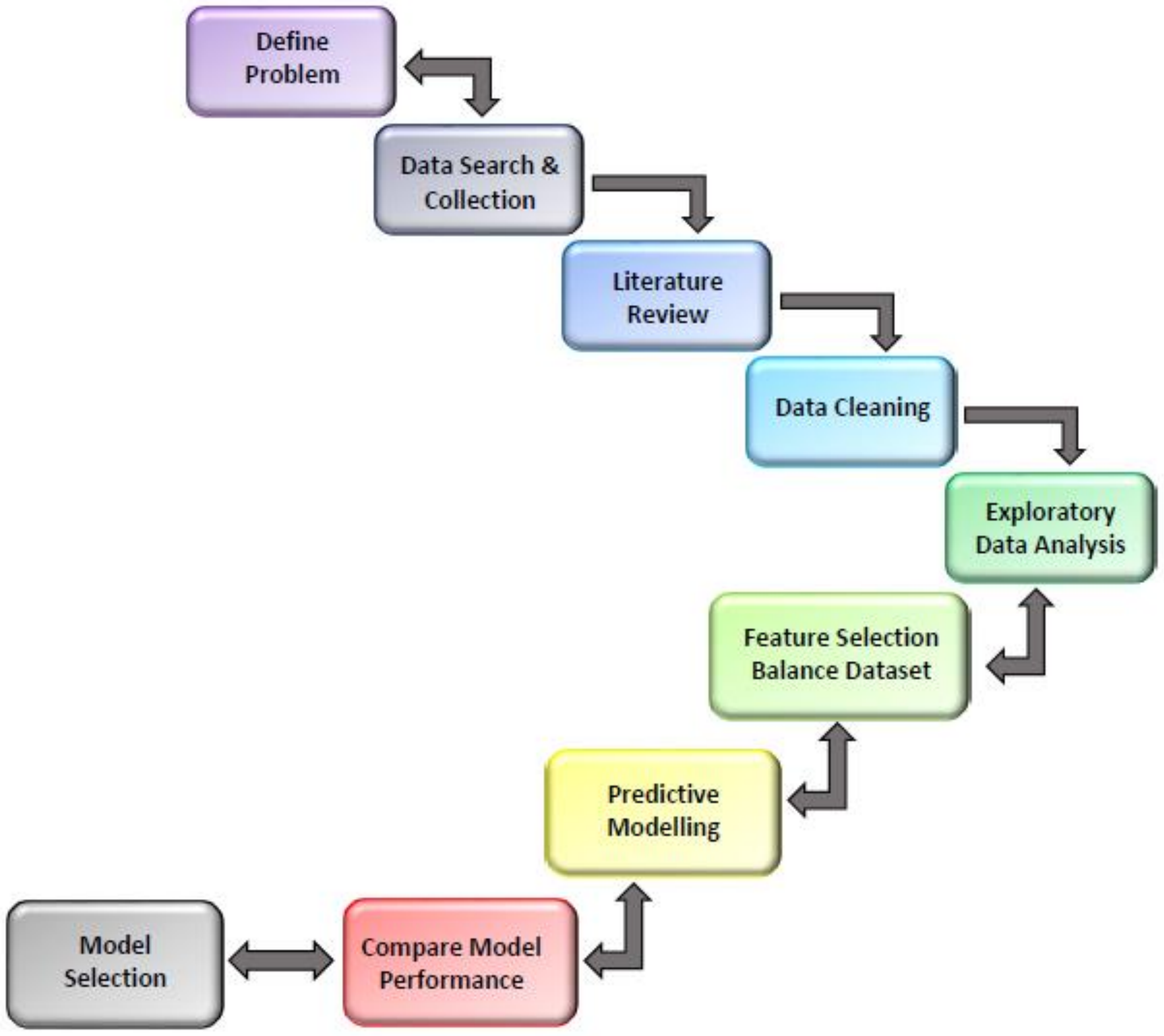
Toronto Metropolitan University

- The potential use of predictive analytics in the field of crime analysis and forecasting first recognized in the 1990s.
- 'Pre-Emptive Policing' was recognized by Time Magazine as one of the 50 best inventions of 2011 (Grossman *et al.*, 2011).
- The availability of open crime datasets has allowed for the field of crime analysis and detection to expand; numerous studies conducted over the past decade on the application of machine learning models in crime prediction.
- ML model results used to support evidence-based decisions by law enforcement agencies such as informing choices regarding resource allocation, deployment, divisional staffing, and patrol plans.

Toronto Metropolitan University

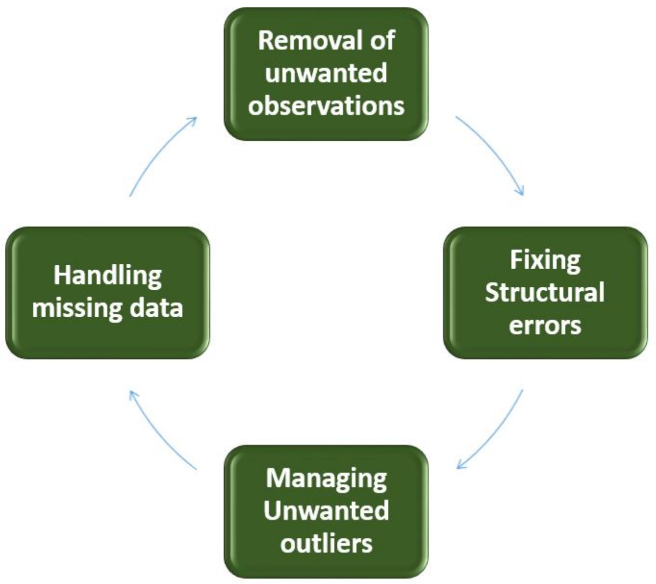
TORONTO MAJOR CRIME INDICATORS DATASET (2014 – 2021): PRIMARY RESEARCH QUESTIONS

- Can major crime indicator categories be accurately predicted?
- Which predictive model exhibits the most potential to forecast crime in Toronto?
- Which Toronto neighbourhoods have the highest/lowest incidence of crime?
- Are there recognizable temporal and spatial trends in overall crime and MCI categories?



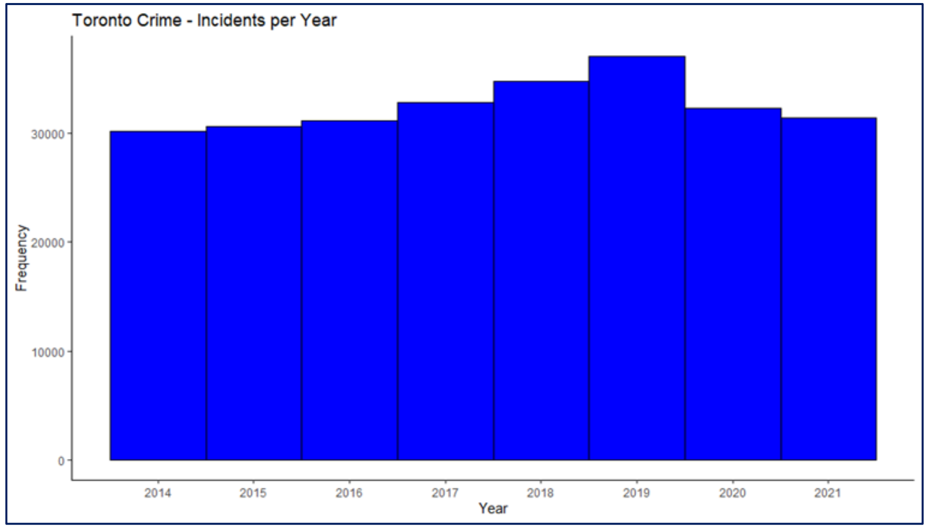
Toronto Metropolitan University

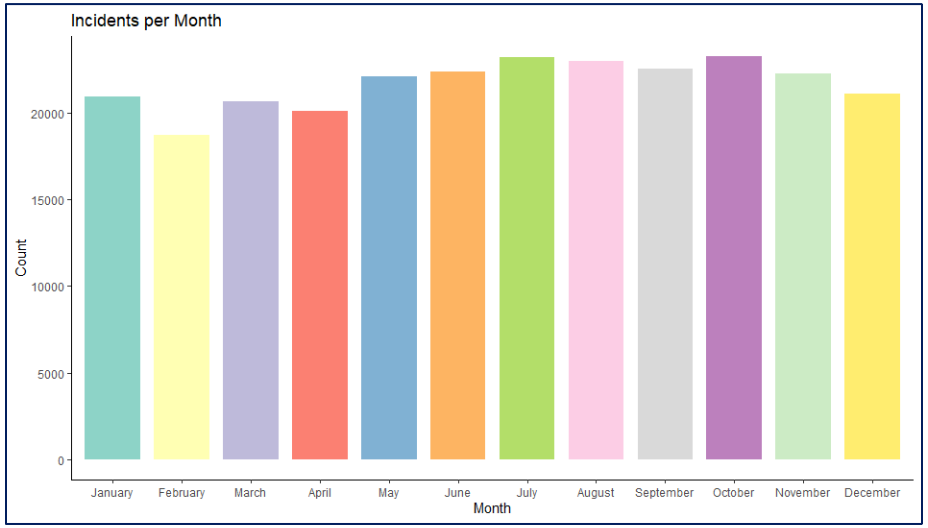
- Crime analysis and prediction was conducted using the 2014 – 2021 Toronto Major Crime Indicator (MCI) dataset
- The MCI dataset contained 281,692 records and 30 variables for founded incidents categorized as Assault, Break and Enter, Auto Theft, Robbery, and Theft over $5000.
- Data types were mixed: character (14), integer (12), and numeric (4).
- The dataset contained no NA values, 20,233 duplicated rows (based on event ID and offense), and 1372 entries for occurrences prior to 2014

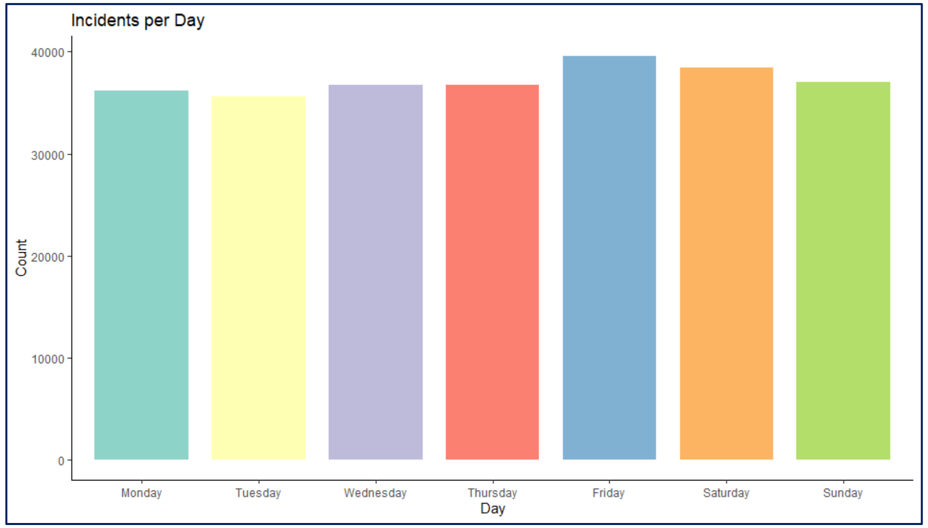


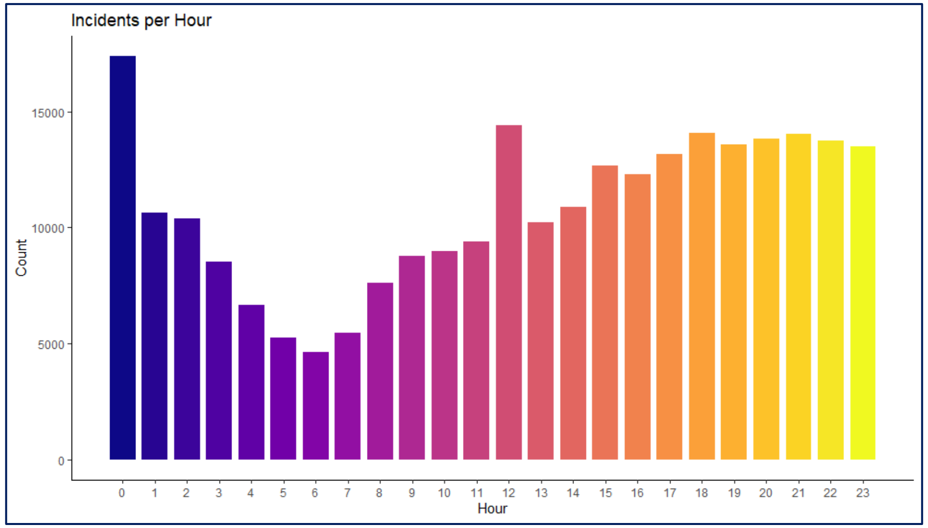
- Removal of 20,233 duplicated rows (based on event ID and offense), 1,372 entries for occurrences prior to 2014, and 15 redundant variables.
- Addition of 2 columns: weights associated with UCR codes and season.
- Convert categorical variables (e.g., day, month, crime time, premises type) to factors to avoid potential complications during analysis.
- The cleaned Toronto MCI dataset contained 260,175 observations and 17 variables.

Toronto Metropolitan University

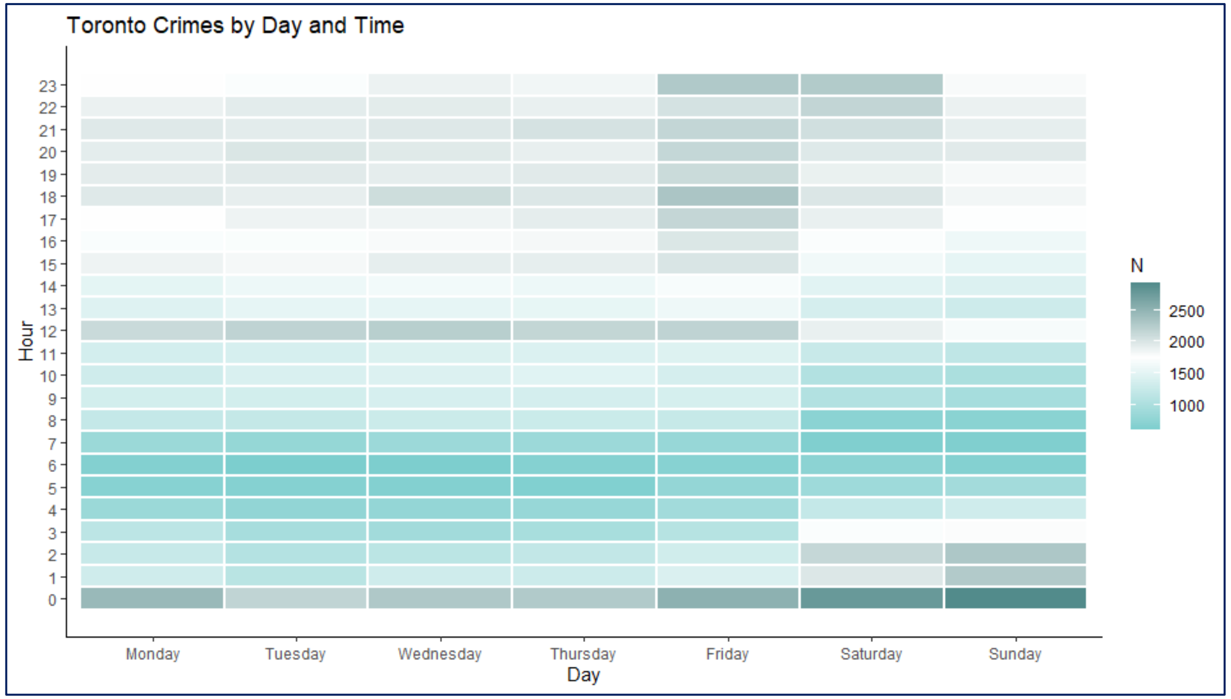


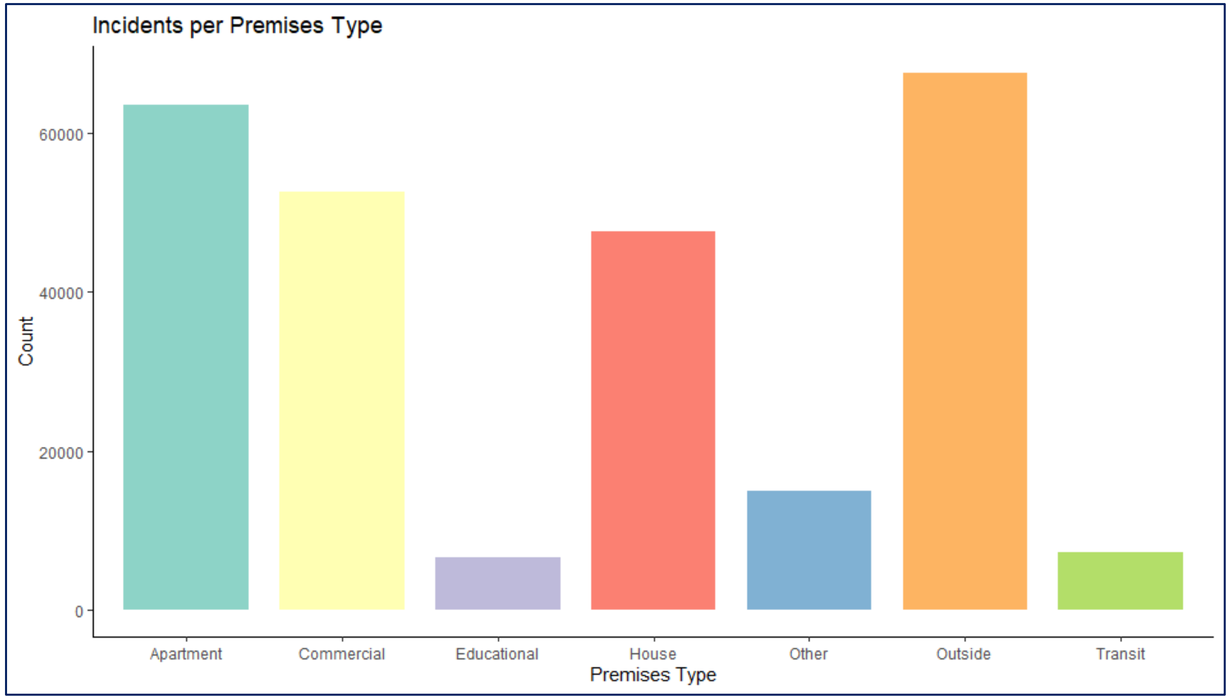




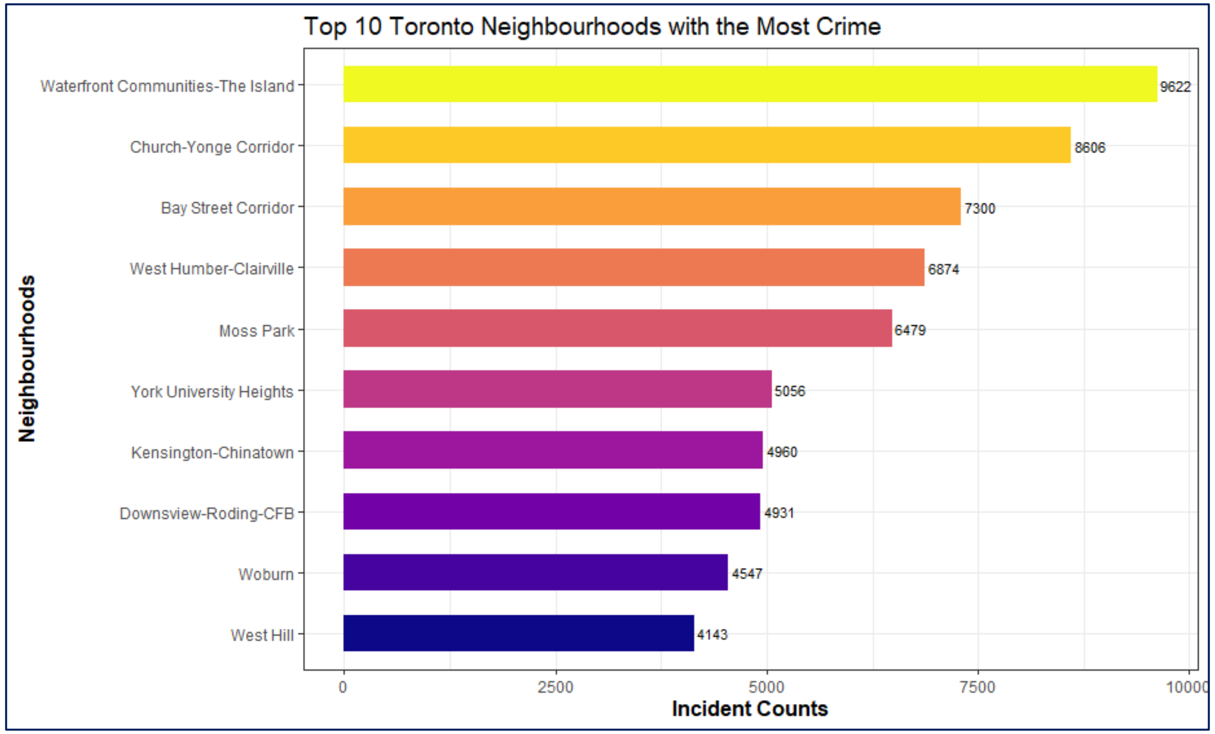


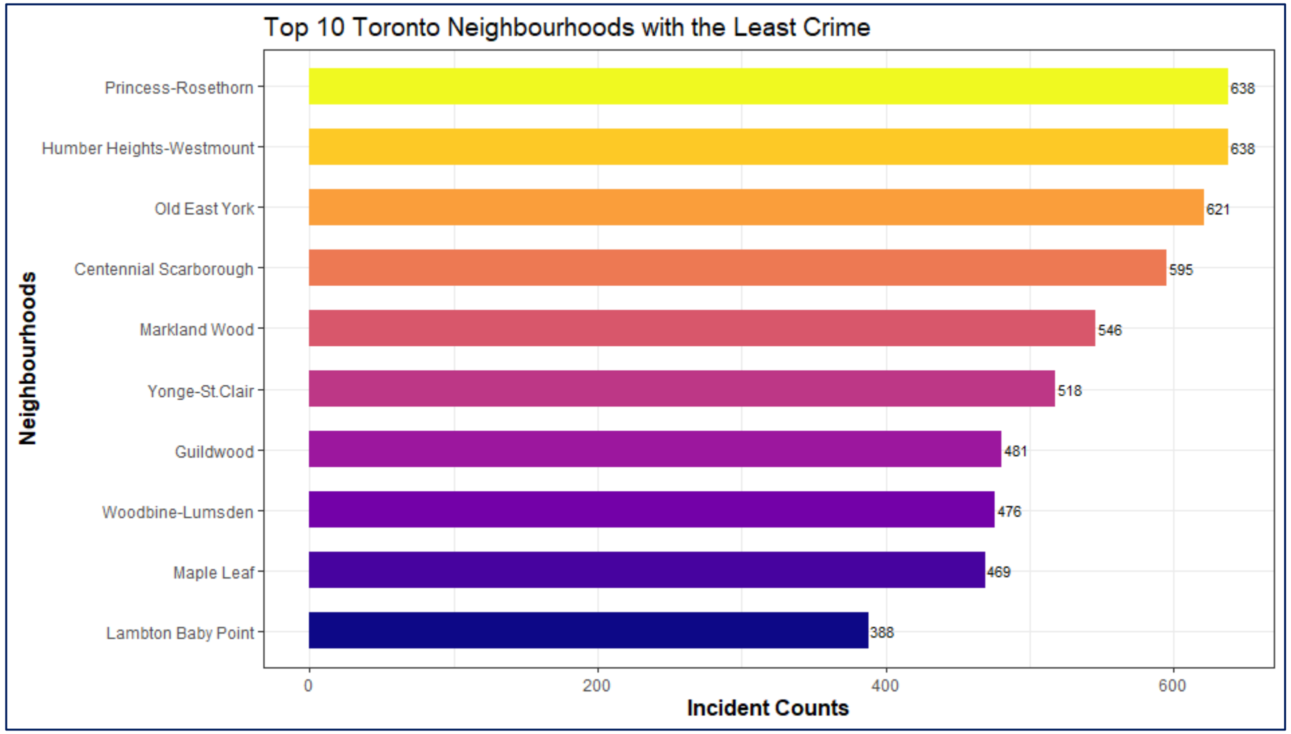
Toronto Metropolitan University



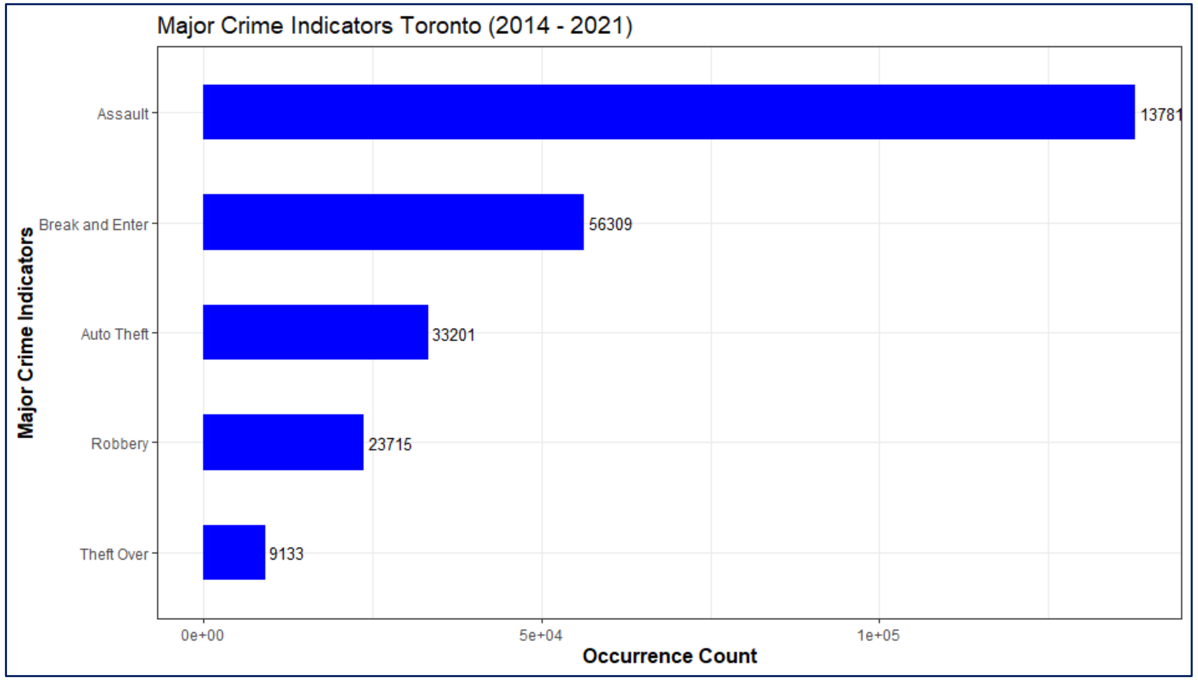


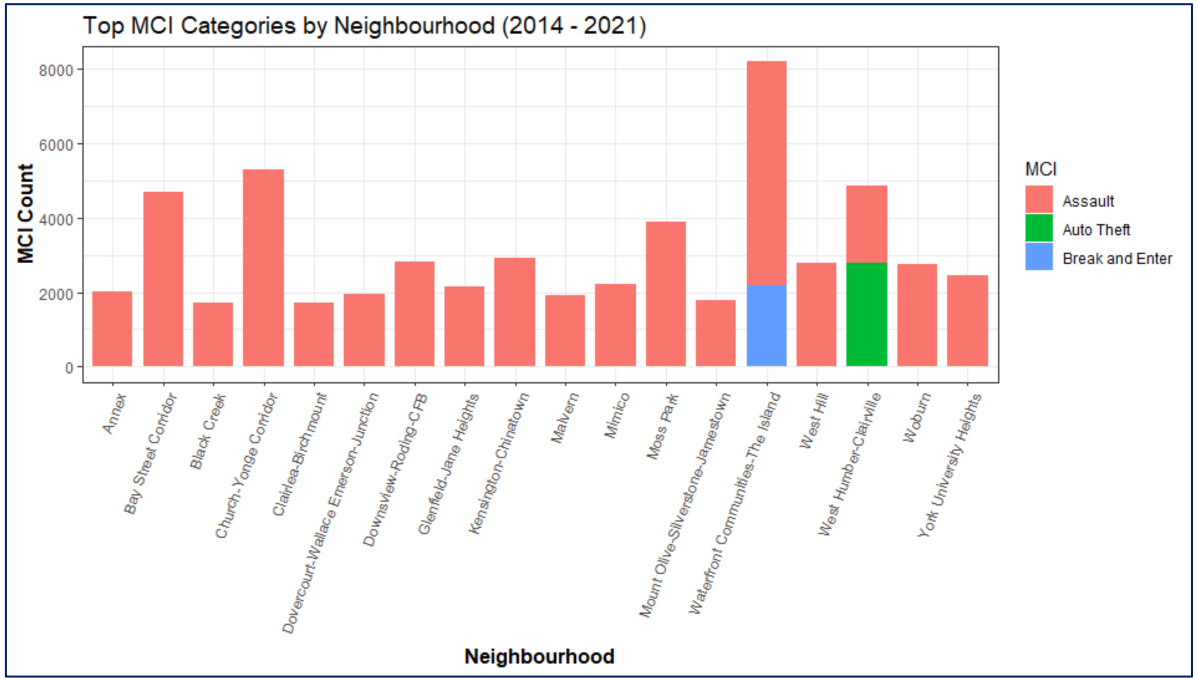
Toronto Metropolitan University

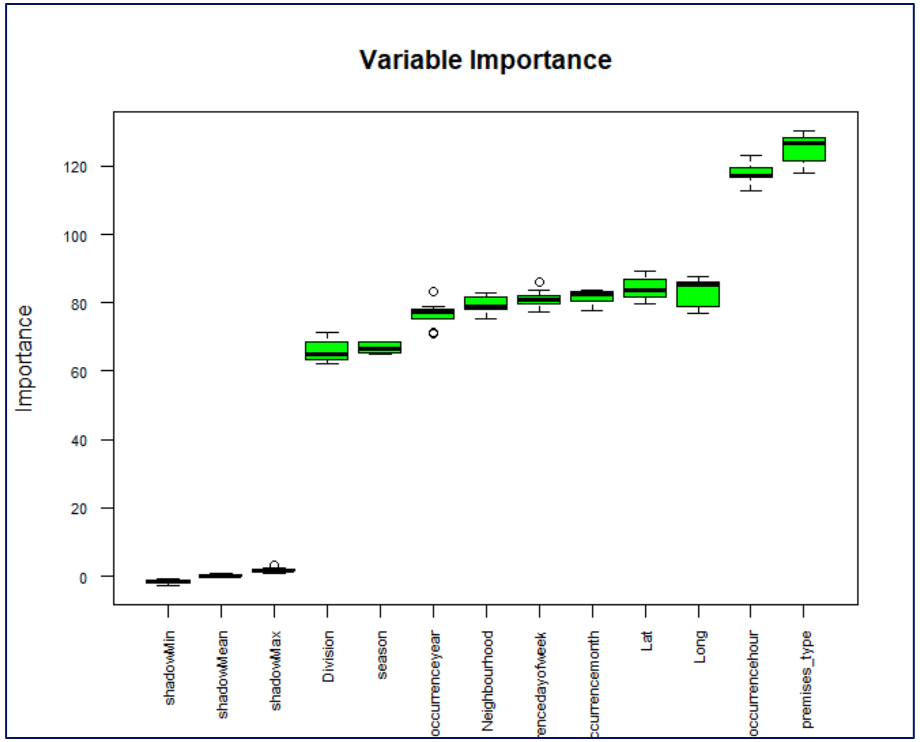




Toronto Metropolitan University


Major Crime Indicators Toronto (2014 - 2021)
Major Crime Indicators
Assault
Break and Enter
Auto Theft
Robbery
Theft Over
13781
56309
33201
23715
9133
0e+00
5e+04
1e+05
Occurrence Count


Top MCI Categories by Neighbourhood (2014 - 2021)
MCI Count
8000
6000
4000
2000
0
Annex
Bay Street Corridor
Black Creek
Church-Yonge Corridor
Clairlea-Birchmount
Dovercourt-Wallace Emerson-Junction
Downsview-Roding-CFB
Glenfield-Jane Heights
Kensington-Chinatown
Malvern
Mimico
Moss Park
Mount Olive-Silverstone-Jamestown
Waterfront Communities-The Island
West Hill
West Humber-Clairville
Woburn
York University Heights
Neighbourhood
MCI
Assault
Auto Theft
Break and Enter



MCI Category	% - Before	% - After
Assault	53	28
Auto Theft	14	7
Break & Enter	21	11
Robbery	8	4
Theft Over $5000	4	49

- Applied Boruta feature selection method to dataset followed by stepwise regression for comparison.
- Premises type, occurrence hour, latitude and longitude were the most significant features associated with the major crime indicators.

- Dataset heavily imbalanced toward the Assault category.
- SMOTE oversampling served to re-balance the dataset by significantly increasing the minority class and reducing the majority class.

Toronto Metropolitan University

Classifier	Accuracy	Kappa	Training Time
Random Forest	79.9%	0.692	61 min.
Decision Tree	76.2%	0.635	8 min.
k-NN	72.1%	0.578	4 hrs.
Naïve Bayes	61.6%	0.348	10 sec.
Multinomial Logistic Regression	49.2%	0.0039	3 min.

- Selected classifiers that were typically used in studies of crime data: decision tree, random forest, k-NN, Naïve Bayes, and multivariate logistic regression.
- Dataset split into 75% training and 25% testing subsets.
- Settings applied were either default or recommended by previous studies to get a general idea of performance.
- Algorithms executed twice to ensure consistency and evaluated using 10-fold cross validation.

Classifier	Accuracy	Kappa	NIR%
Random Forest	80.5%	0.697	48.1
Decision Tree	76.4%	0.637	48.0
k-NN	72.5%	0.583	49.6
Naïve Bayes	61.9%	0.356	64.8
Multinomial Logistic Classification	49.2%	0.0038	99.4

MCI Category	DT*	ML	NB	RF	KNN
Assault	0.689	0.016	0.557	0.734	0.622
Auto Theft	0.413	NA	NA	0.482	0.334
Break & Enter	0.494	NA	0.254	0.553	0.439
Robbery	0.159	NA	NA	0.268	0.252
Theft Over $5000	0.956	0.660	0.766	0.977	0.947

***F1-Score**

- Results on test set nearly identical compared to training set.
- The MCI categories of Assault and Theft Over $5000 were correctly classified most often compared to Auto Theft, Break & Enter, and Robbery which were either poorly classified or not classified at all.
- The random forest classifier consistently outperformed all other algorithms, with decision tree and k-NN returning comparable results.

RF: MCI Category	Sensitivity	Specificity	Accuracy	Precision	F1-Score
Assault	0.659	0.928	79.3%	0.829	0.734
Auto Theft	0.551	0.955	75.3%	0.428	0.482
Break & Enter	0.598	0.939	76.9%	0.515	0.553
Robbery	0.497	0.962	73.0%	0.183	0.268
Theft Over $5000	0.988	0.968	97.9%	0.967	0.977

DT: MCI Category	Sensitivity	Specificity	Accuracy	Precision	F1 Score
Assault	0.627	0.902	76.5%	0.765	0.689
Auto Theft	0.454	0.950	70.2%	0.378	0.413
Break & Enter	0.524	0.933	72.9%	0.468	0.494
Robbery	0.285	0.959	62.1%	0.110	0.159
Theft Over $5000	0.965	0.950	95.8%	0.948	0.956

kNN: MCI Category	Sensitivity	Specificity	Accuracy	Precision	F1 Score
Assault	0.622	0.855	73.9%	0.622	0.622
Auto Theft	0.350	0.946	64.8%	0.319	0.334
Break & Enter	0.432	0.929	68.1%	0.446	0.439
Robbery	0.256	0.965	61.0%	0.249	0.252
Theft Over $5000	0.943	0.953	94.8%	0.951	0.947

NB: MCI Category	Sensitivity	Specificity	Accuracy	Precision	F1 Score
Assault	0.528	0.836	68.2%	0.590	0.557
Auto Theft	0.00	0.925	46.2%	0.00	NA
Break & Enter	0.458	0.903	68.1%	0.176	0.254
Robbery	0.00	0.955	47.8%	0.00	NA
Theft Over $5000	0.673	0.842	75.8%	0.888	0.766

MLR: MCI Category	Sensitivity	Specificity	Accuracy	Precision	F1 Score
Assault	0.404	0.723	56.4%	0.008	0.016
Auto Theft	0.00	0.925	46.2%	0.00	NA
Break & Enter	NA	0.887	NA	NA	NA
Robbery	NA	0.955	NA	NA	NA
Theft Over $5000	0.493	0.727	61.0%	0.997	0.660

- In classifying each MCI category, the random forest model typically provided the highest values for each performance metric.
- The highest sensitivity values were returned for the majority classes while the highest specificity values were typically returned for the minority classes; none of the models were able to classify the minority classes particularly well.

Toronto Metropolitan University

- Most crime in Toronto occurred from Friday to Sunday between midnight and 1 am; least number of incidents between 3 am and 11 am each day. Crime counts were the highest between May and October.
- Lambton-Baby Point and Maple Leaf neighbourhoods had the lowest incidence of crime and the Waterfront Communities, and the Church-Yonge Corridor the highest.
- Assault was the most prevalent of the MCI categories followed by break and enter, auto theft, robbery, and theft over $5000; assault was also the highest MCI category per neighbourhood.
- The random forest model outperformed decision tree, k-NN, Naïve Bayes, and multivariate logistic classifiers, demonstrating the most potential for use as a crime forecasting tool.
- Those MCI categories with the highest proportions (i.e., Assault and Theft Over $5000) were most often correctly classified compared to Auto Theft, Break & Enter, and Robbery which were either poorly classified or not classified at all.

- The majority classes were predicted with the highest accuracy - consider combining some crime categories to produce general classes such as property crime or crimes against person which could serve to increase general model performance and predictive capability for minority classes.
- Consider introducing some of the higher ranking ‘tentative’ independent variables to see how they influence model performance.
- Compare performance results using upsampled, down sampled, and original imbalanced data to those returned using SMOTE oversampled data.
- Models should be re-developed through the adjustment or addition of hyperparameters, and comparative performance be evaluated to determine whether this optimization significantly influenced classification results.
- Investigate whether a Gradient Boost classifier outperforms the random forest model.
- Conduct cluster analysis to further detect patterns in the data.

Toronto Metropolitan University