Toronto Crime Analysis & Prediction

CKME136 Capstone Project

An effort to predict crime with supervised machine learning algorithms

Presented by:
Vidyasankar Sundar
Student ID: 500737042
Aug 2020

Ryerson University

Presentation Content

- Introduction & Project Set up
- Approach
- Results from Exploratory Data Analysis
- Feature Selection
- Class Imbalance via SMOTE
- Model Selection
- Conclusion & Future Work

Ryerson University

Introduction

- ✓Crime prediction is a law-enforcement technique that uses data, and machine learning for the identification of crimes most likely to occur.

- ✓Examples:
  - ✓PredPol is a software that is currently used by more than 60 police departments in the US to identify areas in a neighborhood where serious crimes are likely to occur during a particular period;
  - ✓ShotSpotter detects 90% of gunfire incidents with a precise location in less than 60 seconds to improve response times;
  - ✓COMPAS—or the Correctional Offender Management Profiling for Alternative Sanctions—purports to predict a defendant's risk of committing another crime. It works through a proprietary algorithm that considers some of the answers to a 137-item questionnaire.

Ryerson University

Can we predict crime before it happens?

- ✓ Toronto is one of the most populous, ethnically diverse, and multicultural urban cities in Canada. In this capstone project we attempt to tackle this key question, analyze crime data and build predictive models to predict crime in Toronto.

- ✓ Supervised machine learning algorithms such as classification will be utilized for predicting the category of crimes in Toronto. The techniques will also cover to include pattern identification, prediction, and visualization.

- ✓ More specifically, four algorithms namely Decision tree, KNN Classifier, Naïve Bayes, and Random Forest will be tested, compared and evaluated in order to identify the best performing model for crime prediction.

- ✓ This work does not focus on the victim and the offender, but on the prediction of occurrence of a certain crime type per location and time using past data.

Ryerson University

Key Exploratory Research Questions

Aside from the prediction task, the following questions are also addressed using exploratory/descriptive statistical data analysis methods:

1. Top crimes committed by type - which crimes are most frequently committed and what are the trends in the crimes being committed?
2. When do these crimes occur and is there a pattern?
3. Crime hotspots in Toronto - which locations are these frequent crimes being committed to?
4. What are the safest neighborhoods in Toronto?

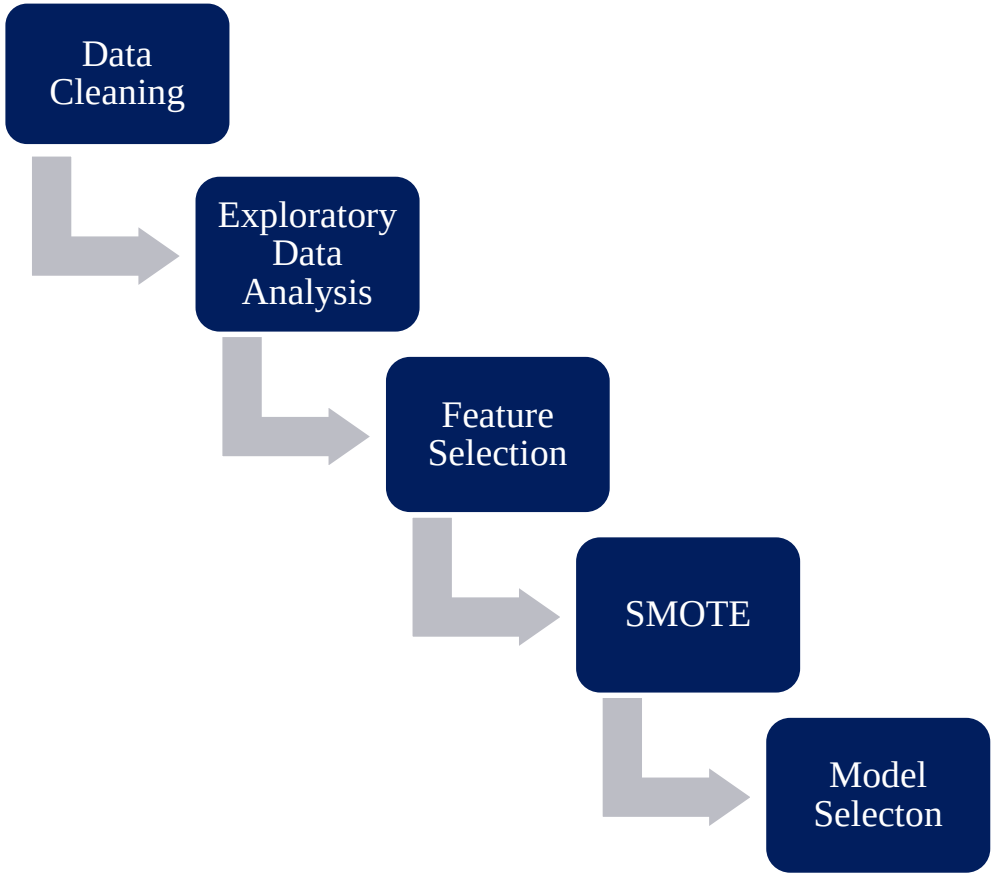
Ryerson University

The Dataset and Project Set up

- **Project Set Up**
  - Toronto crime data from 2014-2019 is made available/ open by the Toronto Police Service public safety data portal. Link to this dataset can be found here.
  - The GitHub repository for this project where all the files including R code is stored can be found here.
- **Dataset**
  - The dataset is structured and contains all Major Crime Indicators (MCI) and offences between 2014 and 2019 including location, neighborhood, day, month, and time.
  - The MCI categories are Assault, Break and Enter, Auto Theft, Robbery and Theft Over (act of stealing property in excess of $5,000).
  - Note that the dataset excludes Homicides and Sexual Assaults.
  - The location of crime occurrences have been deliberately offset to the nearest road intersection node to protect the privacy of parties involved in the occurrence. All location data must be considered as an approximate location of the occurrence henceforth in the analysis.
  - The original dataset contained 27 features (attributes or variables) and 206435 observations prior to data cleaning. After cleaning the data, we end up with 205321 observations of 27 variables.

Ryerson University

The Approach



Ryerson University

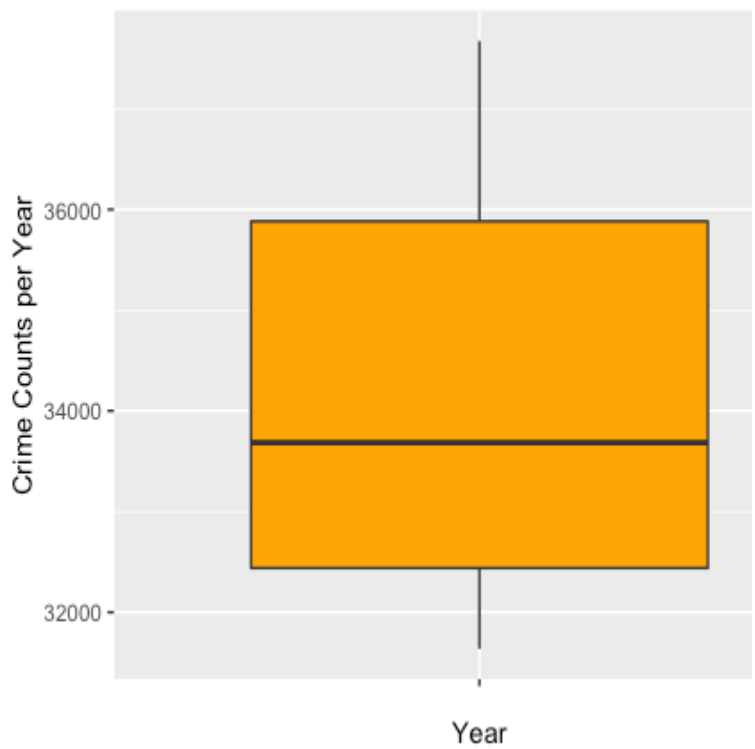
Data Cleaning

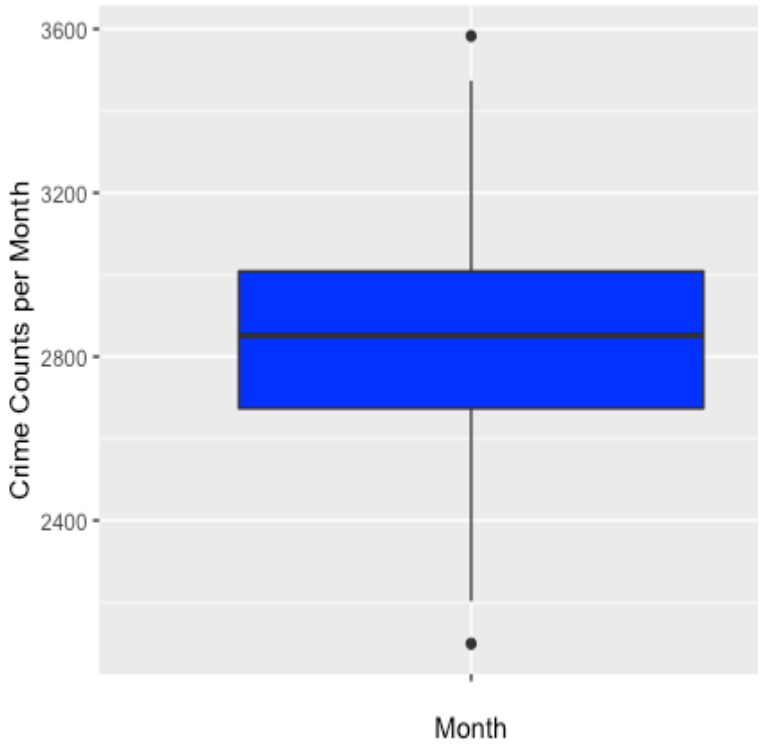
Cleaning up of the dataset involved the following steps:

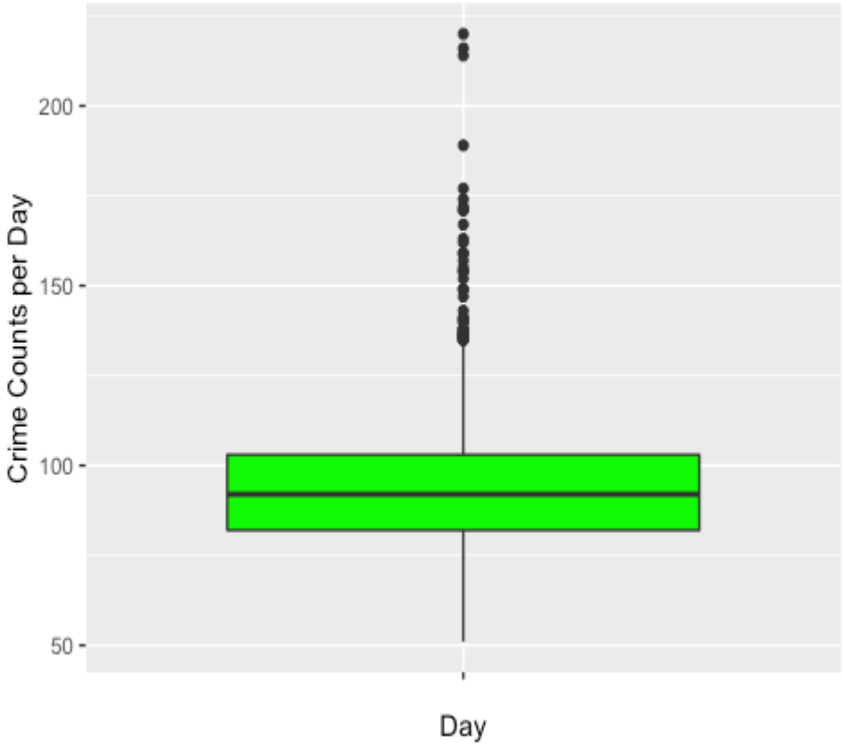
- ✓Total number of records in the dataset was 206435
- ✓Missing values, about 59, were omitted from the raw data
- ✓In addition, some crime data records (about 1055) going back to 2000 was also in the dataset. Therefore the dataset was filtered to include only the occurrences from 2014 to 2019.
- ✓After all this cleaning, we ended up with 205321 observations
- ✓Factor variables such as occurrence month and occurrence day of week were converted to order factors as the levels earlier were arranged by alphabetic order. We wanted to avoid the possibility of this causing issues during the exploratory phase.

Ryerson University

In Toronto, the average number of crime incidents is 33685 per year, 2852 per month, and 92 per day.

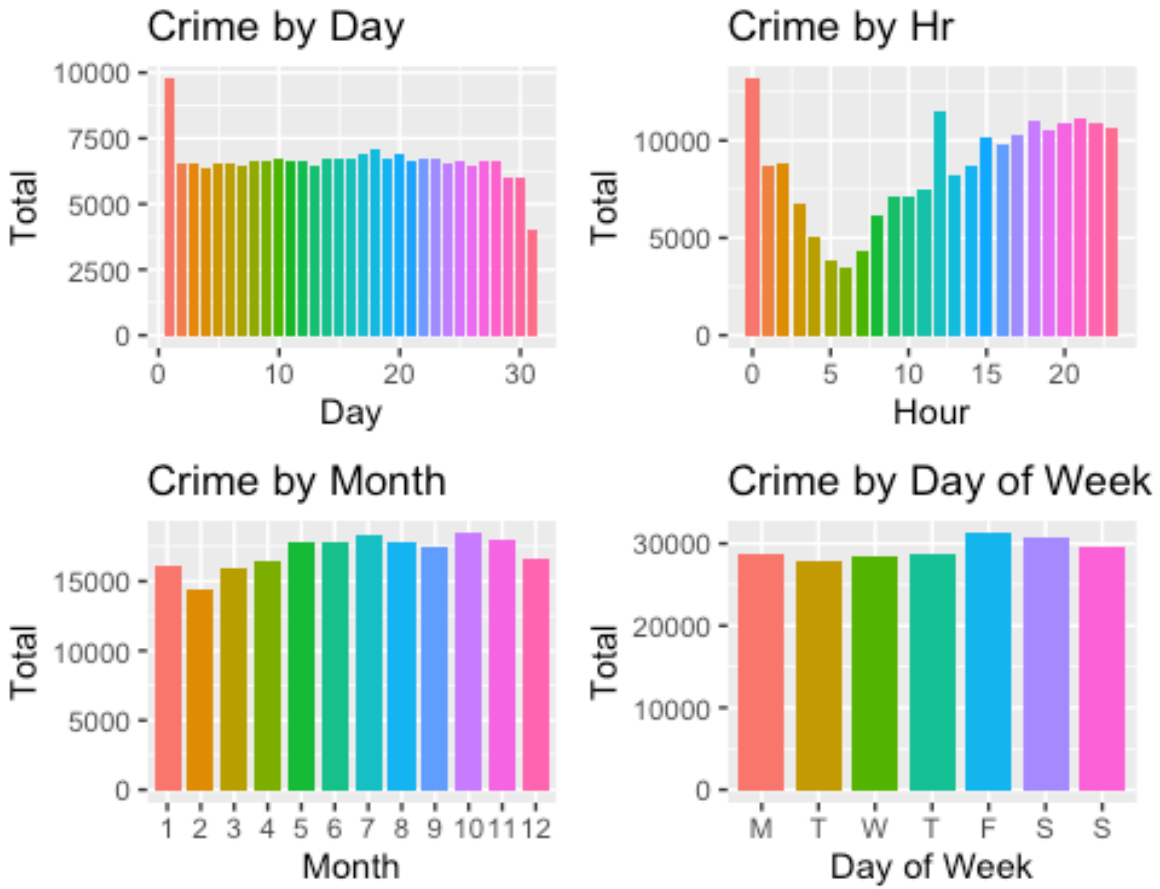






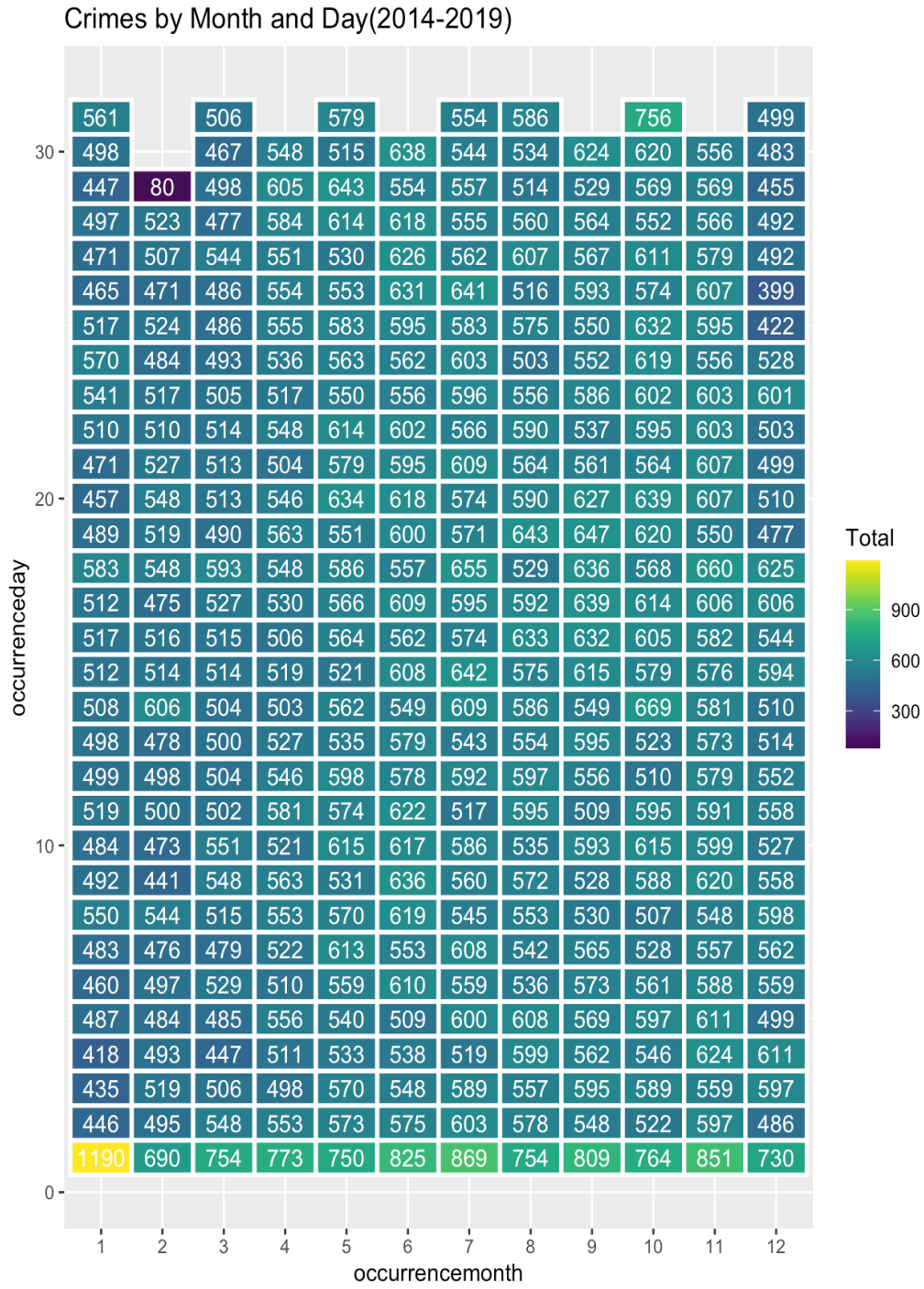
Ryerson University

Crime incidents are most during summer and fall (May-October), with frequencies peaking on the first of every month. Crimes are also most around Fridays and weekends.

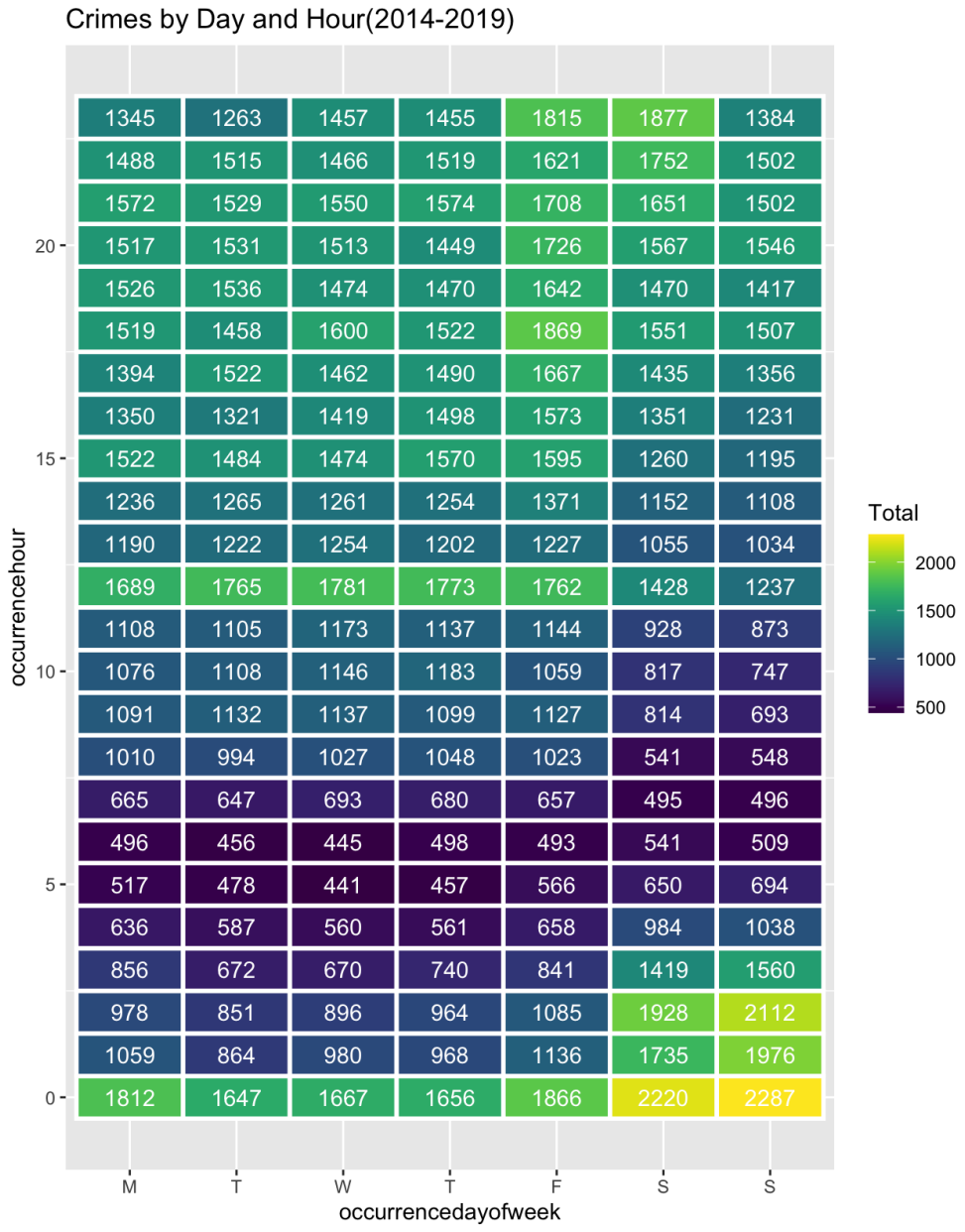


Ryerson University

The first and the last week of any month is seen to have the most incidents. The peak observed is on the first day of every month.

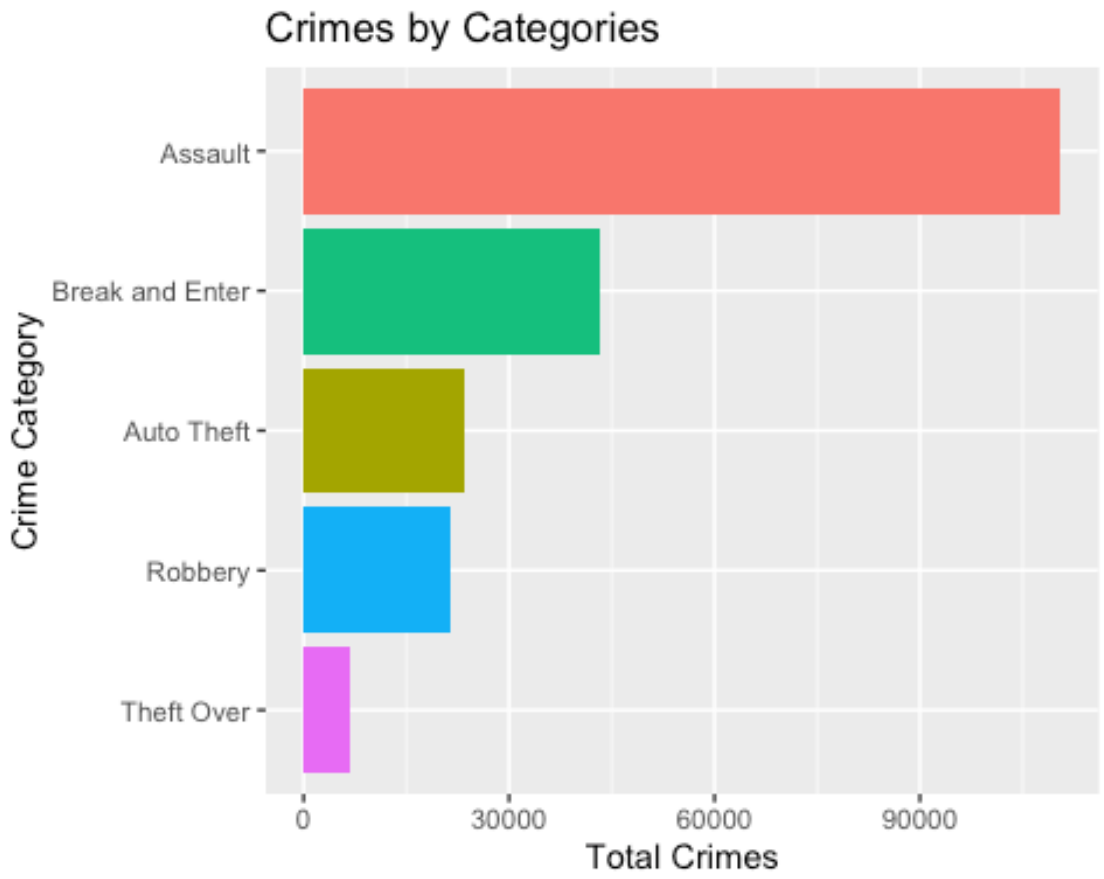


During weekdays, peak is observed at noon and then starts to gradually build from 3PM. During weekends, peak is observed at midnight.



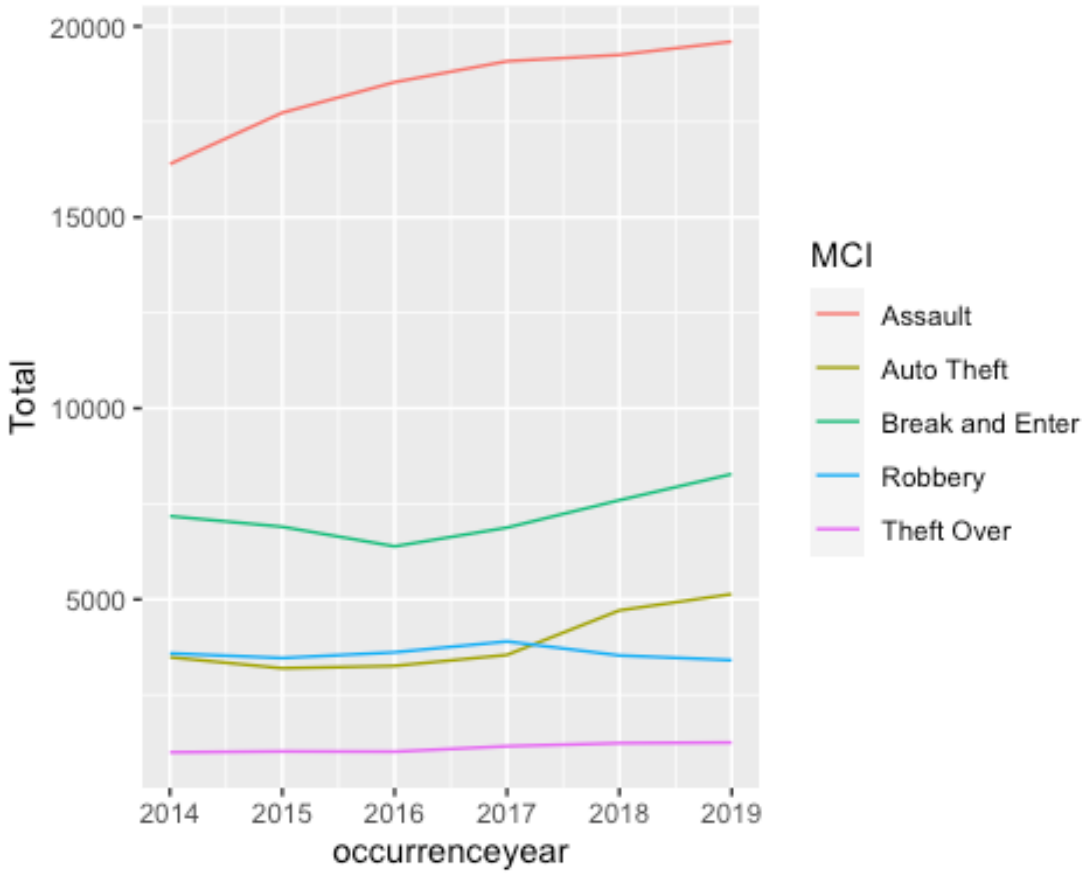
Ryerson University

Assault is the most prevalent form of crime in Toronto followed by home/commercial break and enter and auto theft.



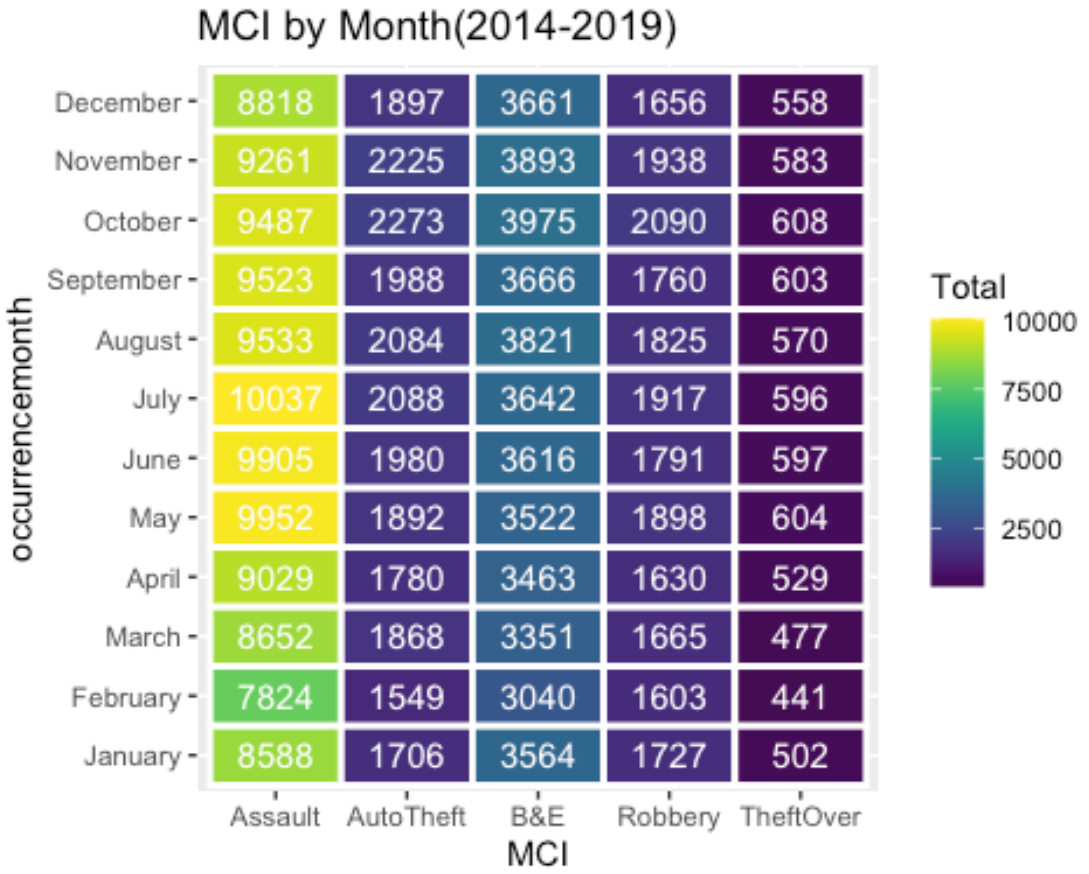
Ryerson University

All crime types have seen an increasing trend since 2014 with the exception of robbery.

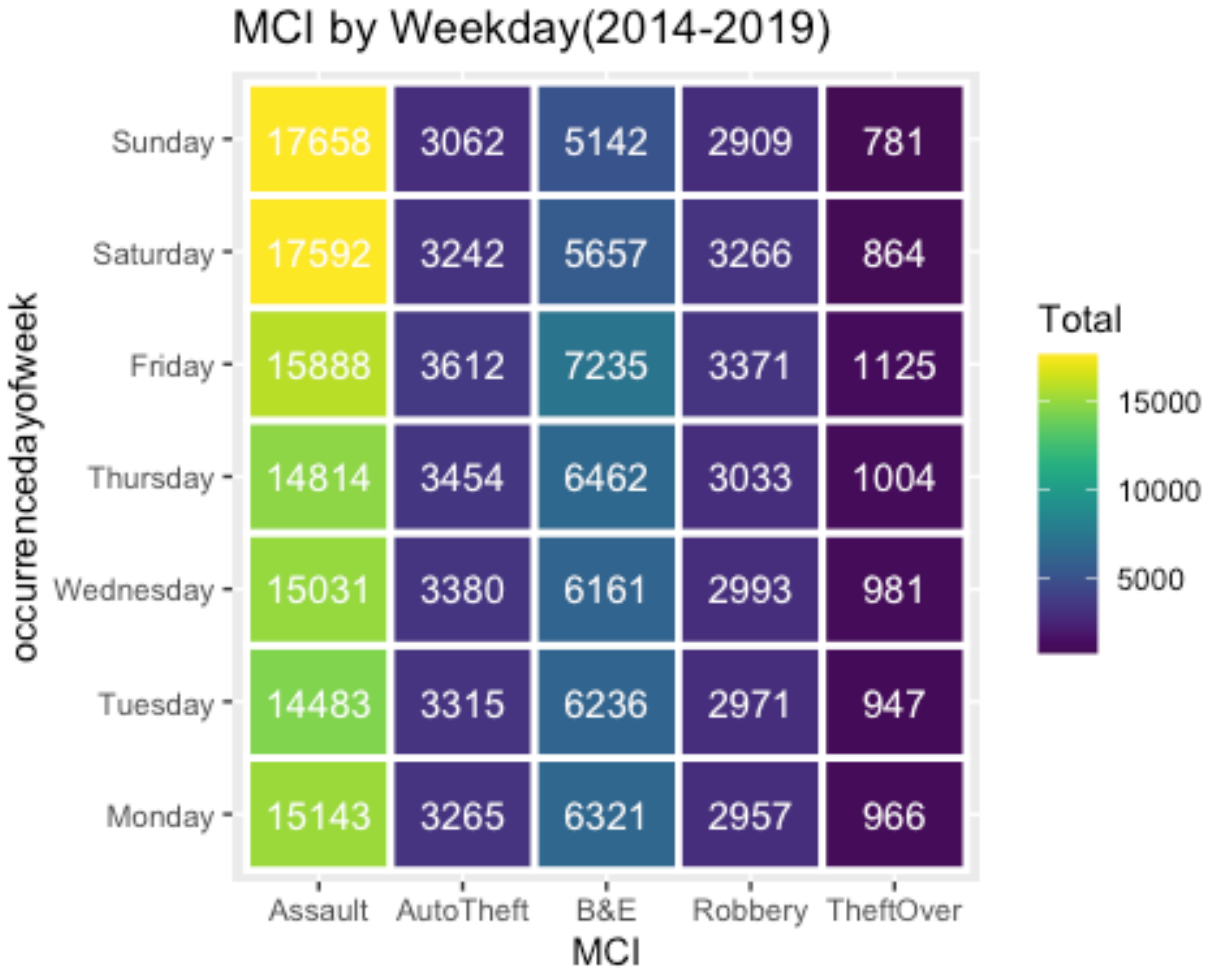


Ryerson University

Most assault incidents happen between May and July, auto theft between July and November while other types are fairly consistent across all months.



Assault peak is observed usually on weekends, while other types peak on Fridays.

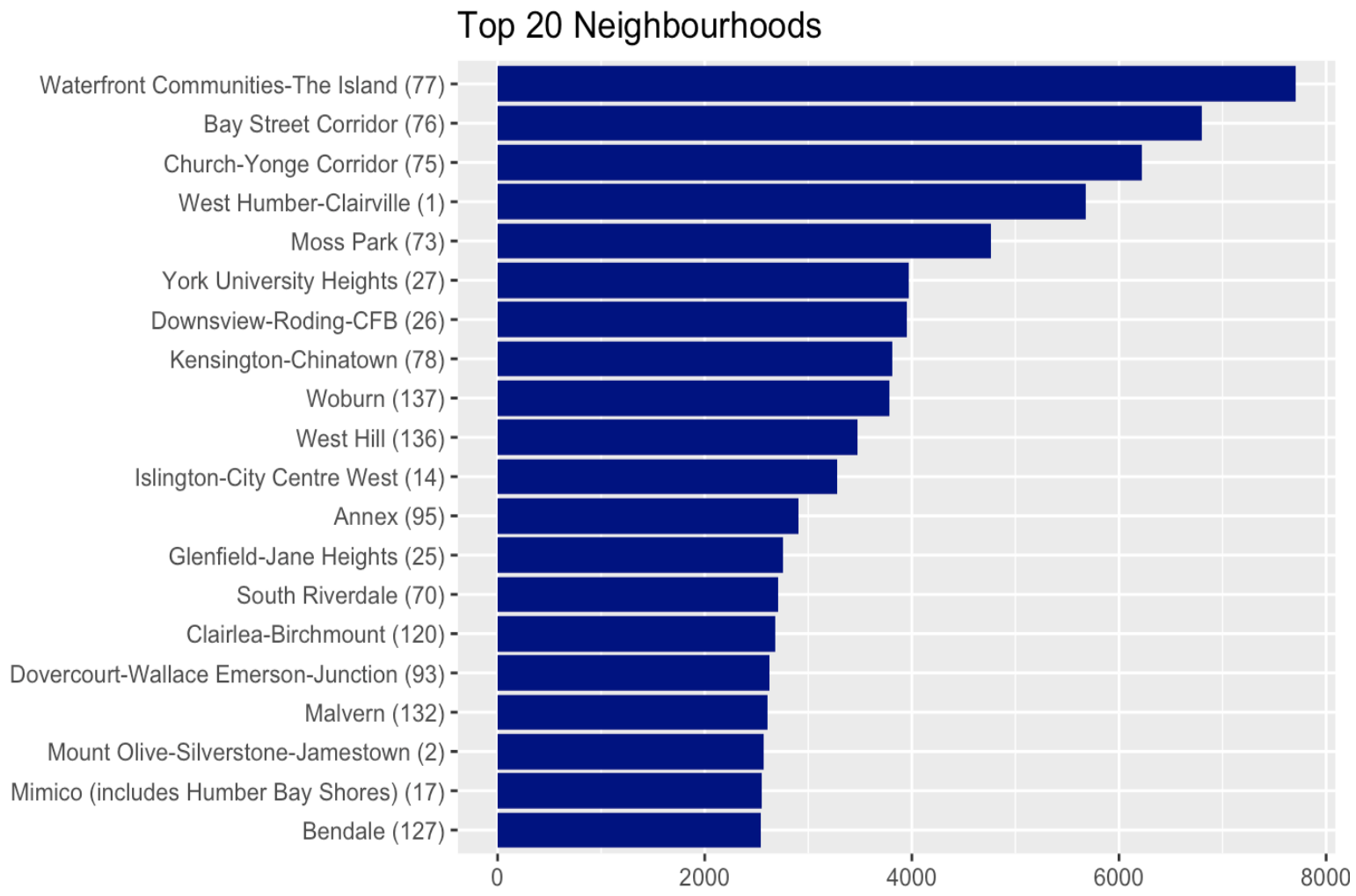


Ryerson University

Assault, Break and Enter, Theft over peaks are observed at both midnight and noon, auto theft at 10PM, robbery at 9PM.

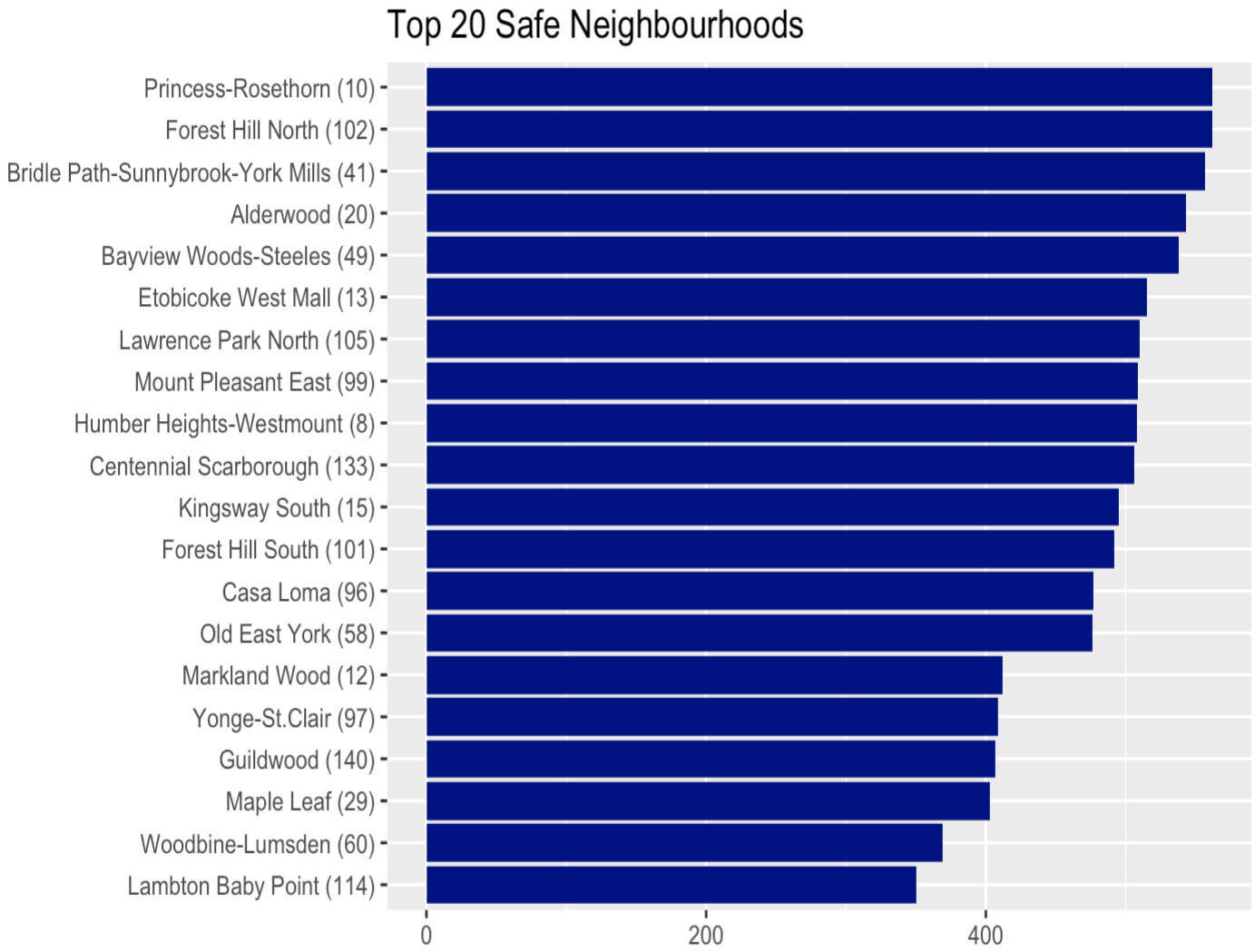
Ryerson University

The most dangerous neighborhood is the Waterfront. The other two most dangerous hoods are the Bay Street Corridor and Yonge-Church Corridor.

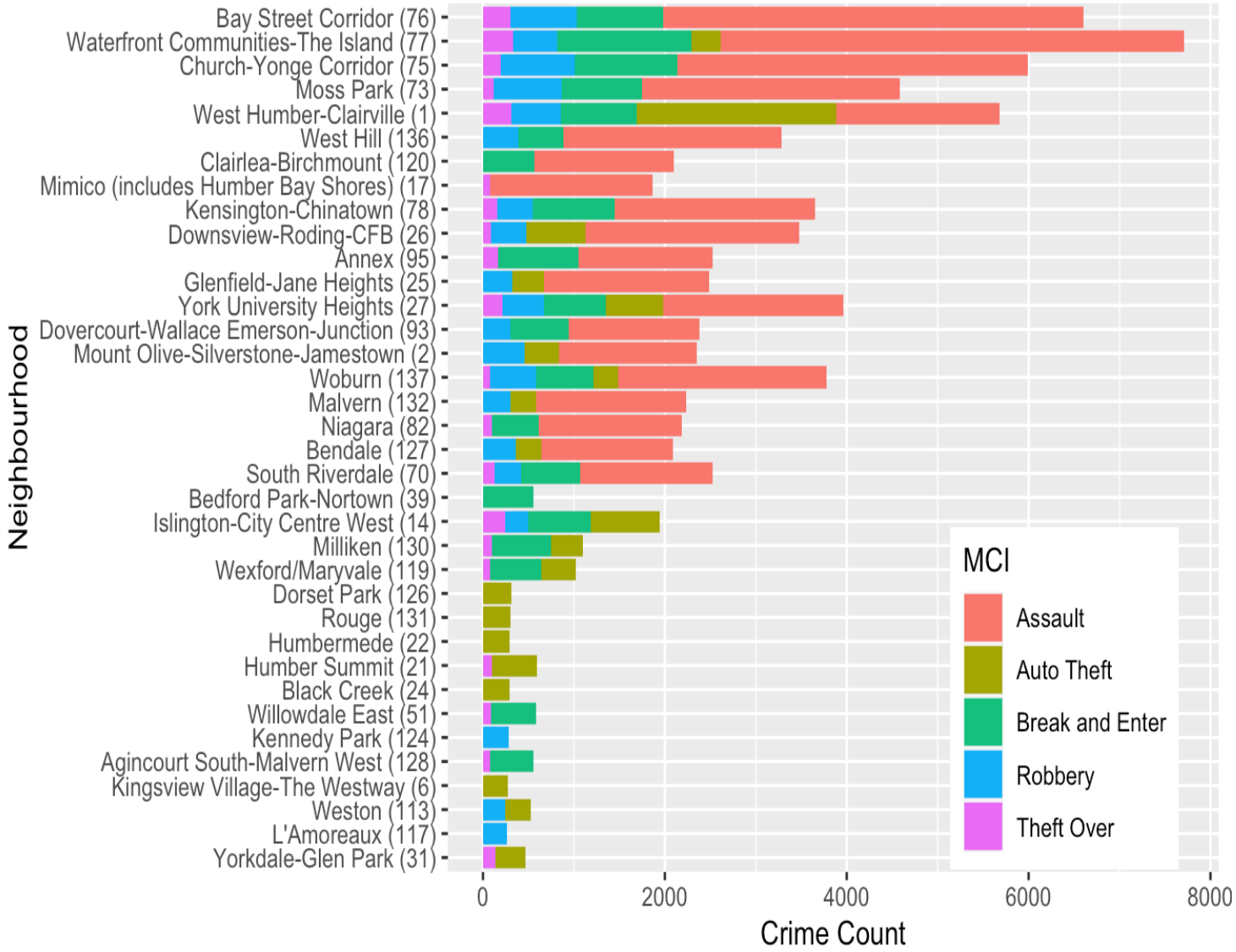


Ryerson University

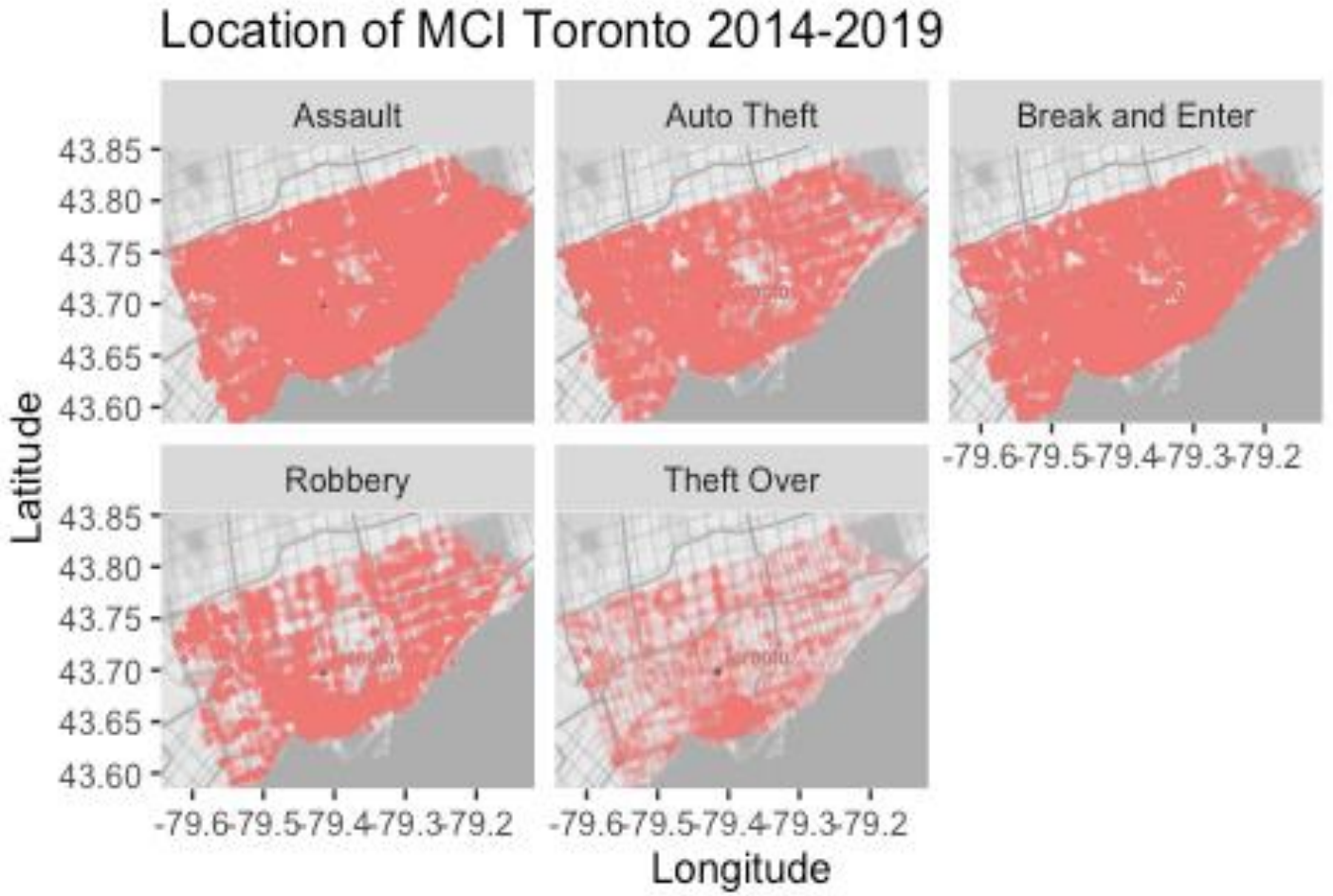
The safest hoods are Lambton Baby Point, Woodbine-Lumsden, Maple Leaf, and Guildwood.



Besides assaults, Bay Street Corridor, Church-Yonge Corridor and Waterfront had the most break and enter crimes while West Humber- Clairville had the most vehicle stolen crimes.



Assaults, and Break and Enter occur all over the city, with a concentration in the Waterfront areas. Other crimes, such as Auto Theft has more points on the west side than the east side. Robbery and Theft Over are primarily in the Waterfront areas.



Feature Selection

- Feature selection is a data preprocessing technique for selecting a subset of the best variables prior to constructing a model. It helps to remove irrelevant variables (i.e., variables that do not share a strong relationship with the target variable.
- The dataset is split into 70% training set and 30% test set. Feature selection is performed on the training set.
- Although there are a wide variety of feature selection techniques i.e. filter-based, wrapper-based, and embedded families, in this project we attempt filter-based feature selection techniques to search for the best subset of variables prior to our model construction. Notably, the use of filter-based feature selection techniques is considered low cost.
- There are many variants of filter-based feature selection techniques such as correlation-based, information gain, chi-square based and consistency based feature selection.
- In this project, we use information gain feature selection method which ranks variables according to the information gain with respect to the outcome. This is implemented by utilizing the FSelector package in R on the training dataset.
- Manual feature selection was also performed to drop off the attributes that were considered redundant or irrelevant for the analysis.

Feature Selection Results – Run with all variables

Applying information gain filter based method:

- Run using all variables
- **print**(weights)
- ## attr_importance
  ## Ã¯..Index_ 0.3600815277
  ## event_unique_id 1.2399845410
  ## occurrencedate 0.0022664022
  ## reporteddate 0.0023671202
  ## premisetype 0.1747254739
  ## ucr_code 1.2568644048
  ## ucr_ext 0.8768522014
  ## offence 1.2568644048
  ## reportedyear 0.0017143021

reportedmonth 0.0007058475
reportedday 0.0003986175
reporteddayofyear 0.0007608399
reporteddayofweek 0.0028992584
reportedhour 0.0409393476
occurrenceyear 0.0017017727
occurrencemonth 0.0006335547
occurrenceday 0.0018686390
occurrencedayofyear 0.0013338778
occurrencedayofweek 0.0021671882
occurrencehour 0.0250131790
Division 0.0293952571
Hood_ID 0.0491091607
Neighbourhood 0.0543388334
Long 0.0374646908
Lat 0.0207975019
ObjectId 0.3564296349

Ryerson University

Feature Selection Results – Run with select variables

Manually remove variables that are not appropriate given the definitions and perform the same run again:

- Data frame contains 27 variables. Drop the variables (14) that are not meaningful. This includes the following: #Index, event unique id, occurrence date and reported date in unix formats,ucr_code, ucr_ext, offence, reported year, month, day, day of year, day of week, hour, and objectid.

- Hood_ID is an identifier, and the occurrence year, month, day, day of year, day of week are relatively weaker attributes compared to the others based on the importance scores.
- We can, therefore, select the following variables for building the models based on the attribute importance scores. This includes: premisetype, occurrencehour, division, neighbourhood, longitude and latitude.

```
##                  attr_importance
## premisetype          0.1737349122
## occurrenceyear       0.0018372615
## occurrencemonth      0.0006536711
## occurrenceday        0.0017255687
## occurrencedayofyear  0.0013449538
## occurrencedayofweek  0.0024220349
## occurrencehour       0.0246724036
## Division             0.0294742407
## Hood_ID              0.0489552797
## Neighbourhood        0.0546586372
## Long                 0.0407133003
## Lat                  0.0202343335
```

Ryerson University

SMOTE

- The training dataset is heavily skewed to the ‘Assault’ category within MCI. Classification using class-imbalanced data is biased in favor of the majority class. Therefore, this class imbalance problem must be addressed in the training set prior to building the model.
- SMOTE is an oversampling technique that generates synthetic samples from the minority class. It is used to obtain a synthetically class-balanced or nearly class-balanced training set, which is then used to train the classifier.
- SMOTE was implemented in R using the UBL library to balance the classifications in the training dataset.

	Assault	Auto Theft	B&E	Robbery	Theft Over
# of Counts	77103	16428	30230	15137	4721
% of Counts	54%	11%	21%	11%	3%

Ryerson University

Model Selection

Since this is a classification problem, we chose to run the following algorithms in R using the RWeka package:

- Decision Tree
- KNN classification algorithm
- Naïve Bayes
- Random Forest

KNN, by far, took the longest time for training and evaluation followed by Random Forest. Naïve Bayes and Decision Tree algorithms took the shortest training time.

10 fold cross validation	Training Time
J48 Decision Tree	60 secs
KNN Classifier	6.5 hours
Naïve Bayes	30 secs
Random Forest	65 mins

Ryerson University

Model Results Summary

Accuracy: It calculates the proportion of correctly classified instances. The Random Forest method scores better in terms of overall accuracy compared to the other methods on both the train and test sets.

Kappa: The Kappa coefficient or statistic is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance). It measures the reliability of a model by measuring the inter-rater agreement for qualitative items. Decision tree and Naïve Bayes achieved fair agreement on the Kappa metric while KNN and Random Forest achieved moderate agreement on the Kappa metric on the training set. In the test set, all algorithms achieved only fair agreement on the Kappa metric.

No Information Rate: It is the proportion of classes that we could accurately guess if we randomly allocated them. Overall accuracy must be greater than the no information rate in general and this seems to be the case for all the algorithms.

	Training Set		Test Set		
Algorithms	Accuracy	Kappa	Accuracy	Kappa	No Inf Rate
Decision Tree	46.9%	0.3357	43.2%	0.2557	0.2637
KNN	60.4%	0.5047	48.8%	0.2807	0.3707
Naïve Bayes	40.3%	0.2532	40.0%	0.213	0.2593
Random Forest	63.7%	0.5466	51.7%	0.3275	0.351

Ryerson University

Measurements for Performance Evaluation

- The **confusion matrix** helps to understand where the classifier is going wrong. The table below shows the confusion matrix after feeding a test set through the decision tree classifier. As a simple rule, the more zeroes or smaller the numbers on all cells but the diagonal, the better the classifier is doing.
- **Precision** (or positive predictive value (PPV)): Given all the predicted labels (for a given class say Assault), how many instances were correctly predicted? It refers to the percentage of results that are relevant. Precision for Assault i.e. PPV = TP/TP+FP

  12227/(12227+4043+6009+6534+4176) = 12227/32989 = 0.37
- **Recall** (or sensitivity or true positive rate (TPR))**:** For all instances that should have a label Assault, how many of these were correctly captured? It refers to the percentage of total relevant results correctly classified by the algorithm. Recall (sensitivity) for Assault i.e. TPR = TP/TP+FN

  12227/(12227+307+2554+886+267) = 12227/16241 = 0.7528

Decision Tree Test Set	Actual Class				
Prediction	Assault	Auto Theft	B&E	Robbery	Theft Over
Assault	12227	4043	6009	6534	4176
Auto Theft	307	3649	726	1625	729
B&E	2554	1043	7192	609	1680
Robbery	886	1191	713	2968	726
Theft Over	267	333	468	348	583

Ryerson University

Classifier Evaluation Results on the Test Set - Summary

- The table below captures the precision and recall for the various classes for each of the algorithms implemented on the test set.
- Of all the algorithms evaluated, the Random Forest method appears to perform better on both precision and recall for all the classes studied within the crime category.
- That said, none of the algorithms seem to have good precision or recall scores for the 'Theft Over' class.

Decision Tree	Assault	Auto Theft	B & E	Robbery	Theft Over
Recall	0.7528	0.35569	0.476	0.24561	0.073854
Precision	0.3706	0.51862	0.5499	0.45774	0.291646

Naïve Bayes	Assault	Auto Theft	B & E	Robbery	Theft Over
Recall	0.7363	0.29172	0.4126	0.23047	0.06119
Precision	0.3517	0.53564	0.5031	0.34638	0.239856

KNN	Assault	Auto Theft	B & E	Robbery	Theft Over
Recall	0.7212	0.35878	0.462	0.30356	0.077503
Precision	0.4923	0.48682	0.5365	0.45875	0.190036

Random Forest	Assault	Auto Theft	B & E	Robbery	Theft Over
Recall	0.7605	0.40019	0.4942	0.3226	0.091859
Precision	0.4916	0.54318	0.6125	0.5255	0.198767

Ryerson University

Conclusions

- Assault is the most prevalent form of crime followed by home/commercial break and enter and auto theft. All crime types have seen an increasing trend since 2014 with the exception of robbery.
- Crime occurs the most during summer and fall (May-October). The first and the last week of any month is seen to have the most incidents. Notably, the peak observed is on the first day of every month. During weekdays, peak is observed at noon and then starts to gradually build from 3PM onwards increasing into the night. During weekends, peak is observed at midnight.
- The most dangerous neighborhood is the Waterfront. The other two most dangerous hoods are the Bay Street Corridor and Yonge-Church Corridor. The safest hoods are Lambton Baby Point, Woodbine-Lumsden, Maple Leaf, and Guildwood.
- The Random Forest method shows the most promise and offers potential to predict crime in Toronto. This method also performs better on all evaluation measures such as overall accuracy, kappa as well as both precision and recall for all the classes studied within the crime category.
- The results from this analysis could be used to elevate people's awareness regarding the dangerous locations in Toronto and attempt to help the Toronto Police Service to predict future crimes in a specific location within a particular time. Overall, the capstone project report publication could start a trend of crimes prediction whereby we can anticipate that the law enforcing agencies throughout Canada can start to take advantage of machine learning algorithms to effectively fight crime to keep our country safer for everyone.

Future Work

- In this project, we only attempted the information gain filter-based technique for feature selection. In future, we could not only try the other filter-based methods but also the wrapper-based and embedded techniques to see the features selected for this data set and analyze the performance of these algorithms.
- We balanced the dataset using SMOTE. In future, we could try under sampling methods and oversampling methods and see how these algorithms perform on the dataset.
- None of the algorithms studied had good precision or recall scores for the minority class within the crime category i.e. ‘Theft Over’ class. This must be investigated in future.
- Of all the algorithms evaluated, Random Forest observed the highest differences or gap in accuracy between the training set and testing set. Though the results are not provided here, testing and running the algorithm without balancing the classes i.e. running the algorithm on the imbalanced dataset and dropping the division variable we were able to close the gaps in accuracy between the training and test sets. That said, there still exists an opportunity to investigate this formally in future.
- Lastly, there exists an opportunity to secure conclusions by means of formal statistical testing procedures. This would be the application of nonparametric testing including the combination of a Friedman test with post-hoc test to compare the classifiers used in the project and statistically finalize the selection of the best performer.

References

- Y. L. Lin, L. C. Yu, and T. Y. Chen, "Using machine learning to assist crime prevention," IEEE 6th Intl. Congr on Advanced Appl. Inform. (IIAIAAI), Hamamatsu, Japan, Jul. 2017.
- Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland, "Once upon a crime: towards crime prediction from demographics and mobile data," Proc. of the 16th Intl. Conf. On Multimodal Interaction, pp. 427-434, 2014.
- V. Grover, R. Adderley, and M. Bramer, "Review of current crime prediction techniques," Intl. Conf. on Innovative Techn and Appl. Of Artificial Intel. pp. 233-237, Springer, London, 2007.
- R. Iqbal, M. A. A. Murad, A. Mustapha, P. H. Shariat Panahy, and N.Khanahmadliravi, "An experimental study of classification algorithms for crime prediction," Indian J. of Sci. and Technol., vol. 6, no. 3, pp. 4219-4225, Mar. 2013.
- L. McClendon and N. Meghanathan, "Using machine learning algorithms to analyze crime data," Mach. Learn and Appl.: an Intl. J. (MLAIJ), vol.2, no.1, Mar. 2015.
- H. W. Kang, H. B. Kang, "Prediction of crime occurrence from multimodal data using deep learning," PLoS ONE, vol. 12, no. 4, Apr. 2017.
- M. V. Barnadas, Machine learning applied to crime prediction, Thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, Sep 2016.
- Kim, Suhong, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri. "Crime Analysis through Machine Learning." In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 415-420. IEEE, 2018.
- Jiarpakdee, Jirayus & Tantithamthavorn, Chakkrit & Treude, Christoph. (2018). AutoSpearman: Automatically Mitigating Correlated Metrics for Interpreting Defect Models.
- Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1)
- Landis, J.R.; Koch, G.G. (1977). "The measurement of observer agreement for categorical data". Biometrics. 33 (1): 159–174. DOI: 10.2307/2529310. JSTOR 2529310. PMID 843571.

Ryerson University

Thank you.
Questions?

Ryerson University