Kaiwen Yang

Classification of Robbery Crimes in Urban Toronto

1 Introduction

1.1 Research Background

Understanding the complex interplay of socio-economic, temporal, and spatial factors in shaping crime dynamics within urban environments is essential for developing effective crime prevention strategies and safeguarding community well-being. In the bustling metropolis of Toronto, where crime incidents unfold daily, there is a pressing need to explore the underlying determinants of criminal behavior, particularly in high-risk scenarios such as robbery crimes. As a student residing in Toronto, the urgency to comprehend these dynamics is palpable, driven by a desire to mitigate personal risk and contribute to broader efforts aimed at fostering safer neighborhoods.

Drawing from insights gleaned from existing literature, our research endeavors to elucidate the multifaceted nature of robbery crimes in Toronto, with a keen focus on the intricate relationship between neighborhood quality, temporal patterns, and spatial factors. Building upon seminal studies conducted in diverse urban contexts, such as the investigation into temperature fluctuations and crime rates in New South Wales, Australia [1], and the analysis of temporal variations in homicide rates during weekends and holidays in the United Kingdom [2], our study seeks to contextualize these findings within the unique socio-cultural landscape of Toronto.

1.2 Research Question

Do temporal and spatial patterns, along with the socio-economic conditions of neighborhoods, significantly impact the likelihood of robbery crimes occurring in Toronto across different neighborhoods?

Temporal factors include season, month, day of the week, hour of the day, and the presence of sunset at the time of the crime.

Spatial patterns encompass longitude, latitude, and premise type of the location of the crime.

Socio-economic indicators of neighborhoods such as income levels, housing density, number of rental properties, permanent job and labor force ratio, transportation service worker and population ratio, healthcare service worker and population ratio, and the difference in individuals' median and average income, will also be considered.

1.3 Data Source

The first dataset utilized in this research is derived from the Census of Toronto City's 158 neighborhoods for the year 2021, obtained from the Toronto Open Data Catalogue. This census, conducted as part of the broader Population Census in Canada, gathers data every five years on demographics, social dynamics, and economic indicators. It comprises information such as the name of each neighborhood and the average and median incomes within Toronto's

neighborhoods. The data originates from Statistics Canada and covers census years 2001, 2006, 2011, 2016, and 2021.

The second dataset employed in this study consists of seven crime datasets sourced from the Toronto Police Service Public Safety Data Portal. These datasets encompass recorded incidents of robbery, break and enter, homicides, shootings, and various types of theft spanning the period from 1995 to 2023, with the majority of recorded data falling between 2015 and 2023. Given that all datasets originate from the same organization, they exhibit uniformity in their structure, featuring columns that store unique crime IDs, occurrence and recorded dates, years, months, days of the week, hours, division types, location types, premises types, offense types, crime classifications, neighborhood information, status indicators, as well as longitude and latitude coordinates of the incidents.

2 Methodology

2.1 Data Collection

The dataset of 2021 census could be downloaded directly through Toronto Open Data Catalogue[3], with keyword search Neighbourhood Profiles.

The dataset of seven crimes could be downloaded directly through Toronto Police Service Public Safety Data Portal[4].

2.2 Data Preporcessing

Preprocessing steps aimed to streamline the datasets, enhance their interpretability, and generate new variables that could potentially serve as predictors in the subsequent analysis of robbery crime prevalence across Toronto neighborhoods.

2.2.1 Census

Column Selection: The necessary columns related to income, demographic, and housing characteristics were selected from the original dataset. This included variables such as neighborhood name, individual median income, individual average income, population, household sizes, number of rental properties, and employment-related variables.

Column Renaming: The selected columns were renamed to improve clarity and consistency in column names, making them more intuitive and informative for analysis purposes.

Creation of Ratio Variables: Several ratio variables were created to capture additional insights from the data. *Permanent_Job_and_Labour_Force_Ratio, Transportation_Service_Worker_and_Population_Ratio* and *HealthCare_Service_Worker_and_Population_Ratio* were computed to analyze the proportion of permanent job positions, transportation service workers, and healthcare service workers relative to the total labor force and population, respectively.

Calculation of Difference in Income: *Difference_in_Individual_Median_Average_Income* was created to calculate the absolute difference between individual median and average incomes, providing insight into income distribution within neighborhoods.

2.2.2 Crime and Robbery

Creation of Indicator Variable for Robbery: A binary indicator variable *Robbery* was created to identify instances of robbery crimes within the dataset, based on the *MCI_CATEGORY* column, which is a categorical variable storing names of different types of crime.

Conversion of Month and Day of the Week to Numeric Representation: The month and day of the week was converted from character to numeric representation: *Month*, *Day_Of_Week_Numeric* to facilitate analysis.

Selection and Renaming of Necessary Columns: Only the necessary columns related to crime category, occurrence time, location, and neighborhood were retained in the dataset. These columns were renamed for clarity and consistency.

Creation of Season Variable: Categorical variable *Season* was generated based on the month of the crime incident, categorizing each observation into one of the four seasons..

Creation of Presence of Sunset Indicator: Indicator variable *Darkness* was created to identify whether a crime occurred between sunset and sunrise, based on predefined average sunset and sunrise times.

Subset Data for Year 2020: A new dataset *crime_2020* was created by filtering crime incidents that occurred specifically in the year 2020, preparing for the merging step.

Data Cleaning: Crime incidents with non-recorded or NA information were removed from the dataset to ensure data quality and consistency.

2.3 Exploratory Data Analysis (EDA)

The EDA is conducted directly on the merged dataset from census data and crime data by the sharing column *Neighbourhood*, after modifying incoherent neighborhood spelling in the two datasets.

2.3.1 Basic Data Exploration

Kable:
- created a summary statistics table for continuous variables such as longitude, latitude, population, average household size, number of rental properties, and ratios.

GGPlot2:
- created a sequence of bar plots for discrete numerical and categorical variables such as season, month, day of the week, and hour.

2.3.2 Advanced Data Exploration

Plotly:

- Three-dimensional scatter plots: created to visualize the relationship between robbery counts and socioeconomic indicators across different neighborhoods in Toronto.
- Box plots with facets: generated to examine the distribution of robbery counts by premises type, categorized based on the presence of sunset, *Darkness*.
- Line plots with facets: utilized to illustrate changes in robbery counts in different premises type across different seasons, months, days of the week, and hours of the day.
- Bar plots: utilized to show proportions and changes of robbery crime in total crime across different seasons, months, days of the week, and hours of the day.

Leaflet:

- Interactive Map: displayed the geographical distribution of robbery incidents across Toronto. Each marker on the map represented a robbery incident, with color indicating the type of premises where the incident occurred. Popup information provided details such as neighborhood name, premises type, and population.

2.4 Modelling

First, I splitted the merged data into testing and training data with a 70:30 ratio. Then, I built six models: Generalized Linear Model (GLM), Classification Tree, Bagging, Random Forest, Boosting, and XGBoost. I trained them using the training data, tuned the optimal hyperparameter if possible, and calculated their training accuracy. Next, I tested them using the test data and calculated their performance metrics including test accuracy, precision, recall, and F1-score. Finally, I created a summary table to display and compare the performance of all models.

3 Results

3.1 EDA

3.1.1 Basic Data Exploration

Longitude	Min. :-79.64	1st Qu.:-79.48	Median :-79.40	Mean :-79.41	3rd Qu.:-79.35	Max. :-79.13
Latitude	Min. :43.59	1st Qu.:43.66	Median :43.70	Mean :43.71	3rd Qu.:43.75	Max. :43.85
Population	Min. : 6260	1st Qu.:14960	Median :20080	Mean :19963	3rd Qu.:24305	Max. :33300
Average Household Size	Min. :1.500	1st Qu.:2.100	Median :2.500	Mean :2.412	3rd Qu.:2.700	Max. :3.800
Number of Rental Properties	Min. : 350	1st Qu.: 2665	Median : 3705	Mean : 4293	3rd Qu.: 5110	Max. :11325
Permanent Job and Labour Force Ratio	Min. :0.4893	1st Qu.:0.6400	Median :0.6774	Mean :0.6696	3rd Qu.:0.7070	Max. :0.7522
Transportation Service Worker and Population Ratio	Min. :0.005034	1st Qu.:0.015821	Median :0.021148	Mean :0.024502	3rd Qu.:0.032030	Max. :0.063964
HealthCare Service Worker and Population Ratio	Min. :0.03873	1st Qu.:0.05312	Median :0.05785	Mean :0.05973	3rd Qu.:0.06452	Max. :0.11151
Difference in Individual Median Average Income	Min. : 4320	1st Qu.: 8440	Median : 14300	Mean : 21491	3rd Qu.: 24200	Max. :165600

Table 1: Summary statistics for numerical variables

The summary of continuous numerical variables is shown in Table 1. The difference in individuals' median and average income among neighborhoods varies significantly, with a minimum of 4320 and a maximum of 165600. Similarly, the population, average household size, number of rental properties, permanent job labor force ratio, and healthcare service worker population ratio show significant variation among different neighborhoods. The only feature that all neighborhoods share similar statistics for is the transportation service worker and population ratio, which appears to be guaranteed by the government to ensure transportation efficiency within society.

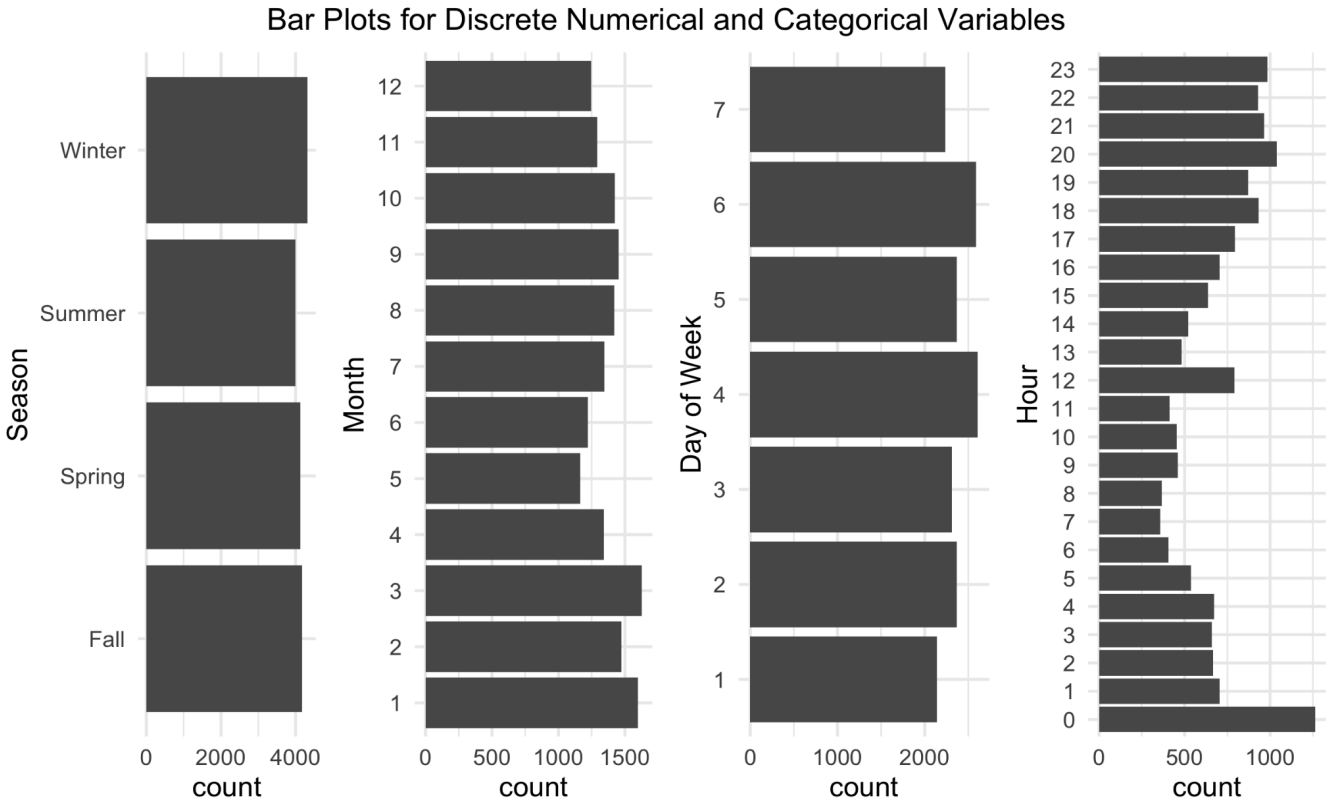

Figure 1: Bar Plots for Discrete Numerical and Categorical Variables

The distribution of discrete numerical variables and categorical variables is shown by bar plots in Figure 1: Bar Plots for Discrete Numerical and Categorical Variables. Glancing from left to right, the Season is comparably evenly distributed, while the count of months is higher between January and March, and July and October. Saturday and Thursday have relatively higher counts compared to other days of the week, and there seems to be a normal distribution within the hour of the day, centered exactly at midnight, with the highest count of around 1250. The count of crime decreases from both 1 AM to 11 AM and from 1 PM to 11 PM, except for a jump in count at noon.

3.1.2 Advanced Data Exploration

Plotly: 3D Scatter Plot

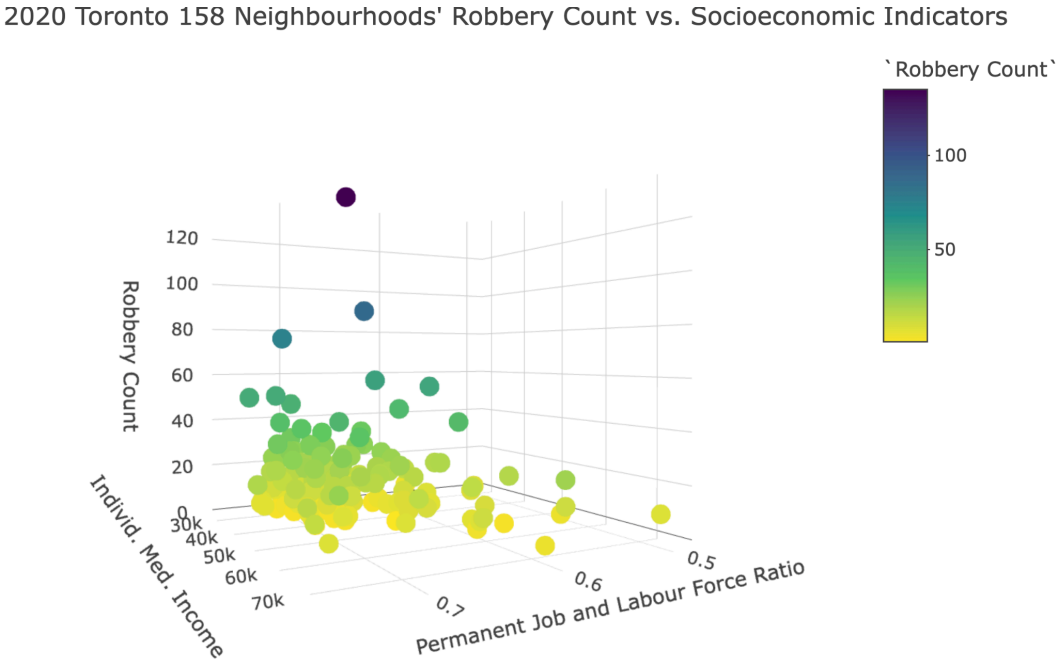

Figure 2: Toronto 158 Neighbourhoods' Robbery Count vs. Socioeconomic Indicators

The relationship between *Permanent_Job_and_Labour_Force_Ratio, Individual_Median_Average_Income* and robbery count among neighbouhoods is shown in Figure 2: Toronto 158 Neighbourhoods' Robbery Count vs. Socioeconomic Indicators. There is a noticeable cluster at low individual median income, and surprisingly with a high permanent job labour force ratio.

Plotly: Boxplots with Facets

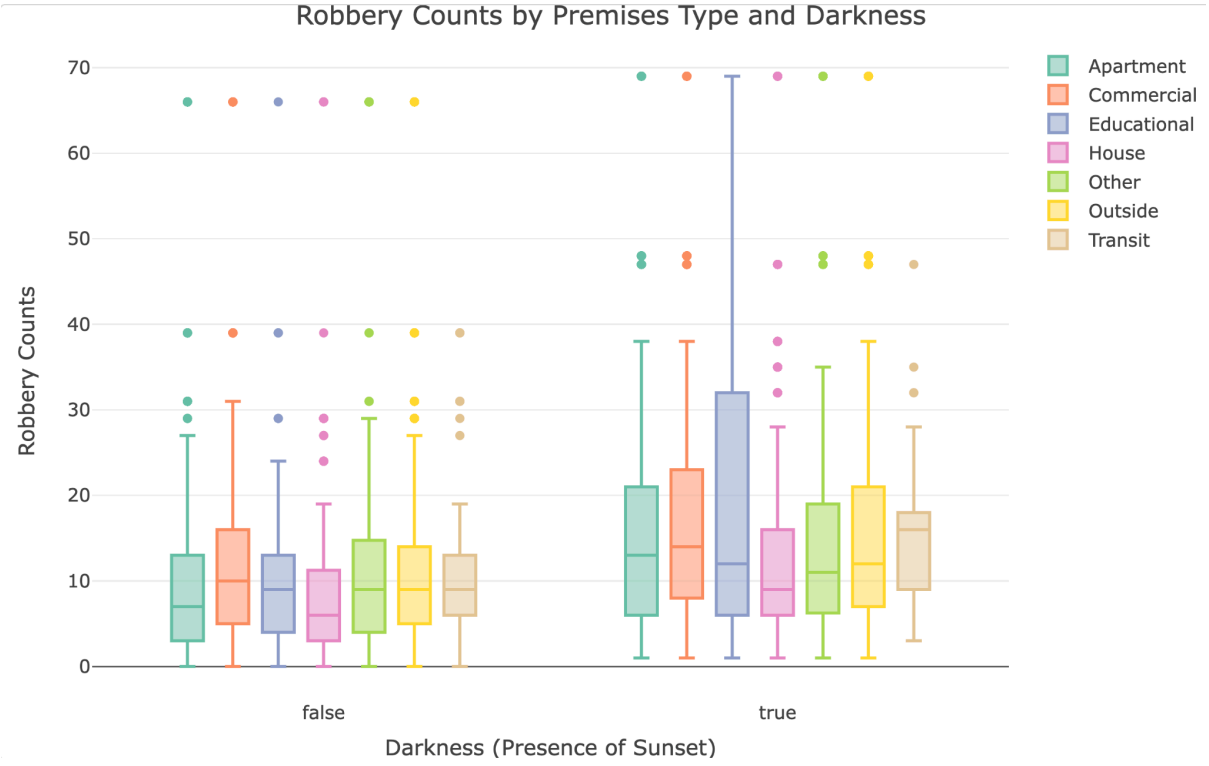

Figure 3: Robbery Counts by Premises Type and Darkness

The distribution of robbery counts by *Premises_Type*, categorized based on the presence of sunset, *Darkness* is shown in Figure 3: Robbery Counts by Premises Type and Darkness. We can see that the mean and variance of robbery count in darkness are clearly higher than those not in darkness, regardless of premises type. Additionally, the darkness group has more outliers and a higher robbery count than the other group.

Looking through premises types, we can see that commercial and educational areas have much bigger ranges in robbery count for the 1st quartile and 3rd quantile in the darkness group. This may imply that people who appear in those areas may have a higher probability of being involved in a robbery crime during darkness compared to the other group.

Plotly: Line Plots with Facets

Figure 4: Robbery Count by Premises Time and Season / Month / Day / Hour

The changes in robbery counts in different *Premises_Type* across different *Season*, *Month*, *Day_of_Week*, and *Hour* of the day is shown in Figure 4: Robbery Count by Premises Time and Season / Month / Day / Hour. We can see that commercial and outdoor areas have higher counts than all other types of premises.

Regarding the season, the count of robbery increases sharply in winter, which is consistent with existing studies that show crime rates increase during national holidays like Christmas and New Year.

For the month, the count of robbery shows an increasing trend during the summer months, which also matches the conclusion of existing research suggesting a correlation between temperature and crime rates.

Regarding the day of the week, we can see that robbery counts increase during weekends and drop on weekdays.

For hours of the day, it is clear that the robbery count increases from the afternoon to midnight and decreases sharply at 2 AM, which aligns perfectly with my intuition in creating the *Darkness* variable.

Plotly: Barplots with Facets

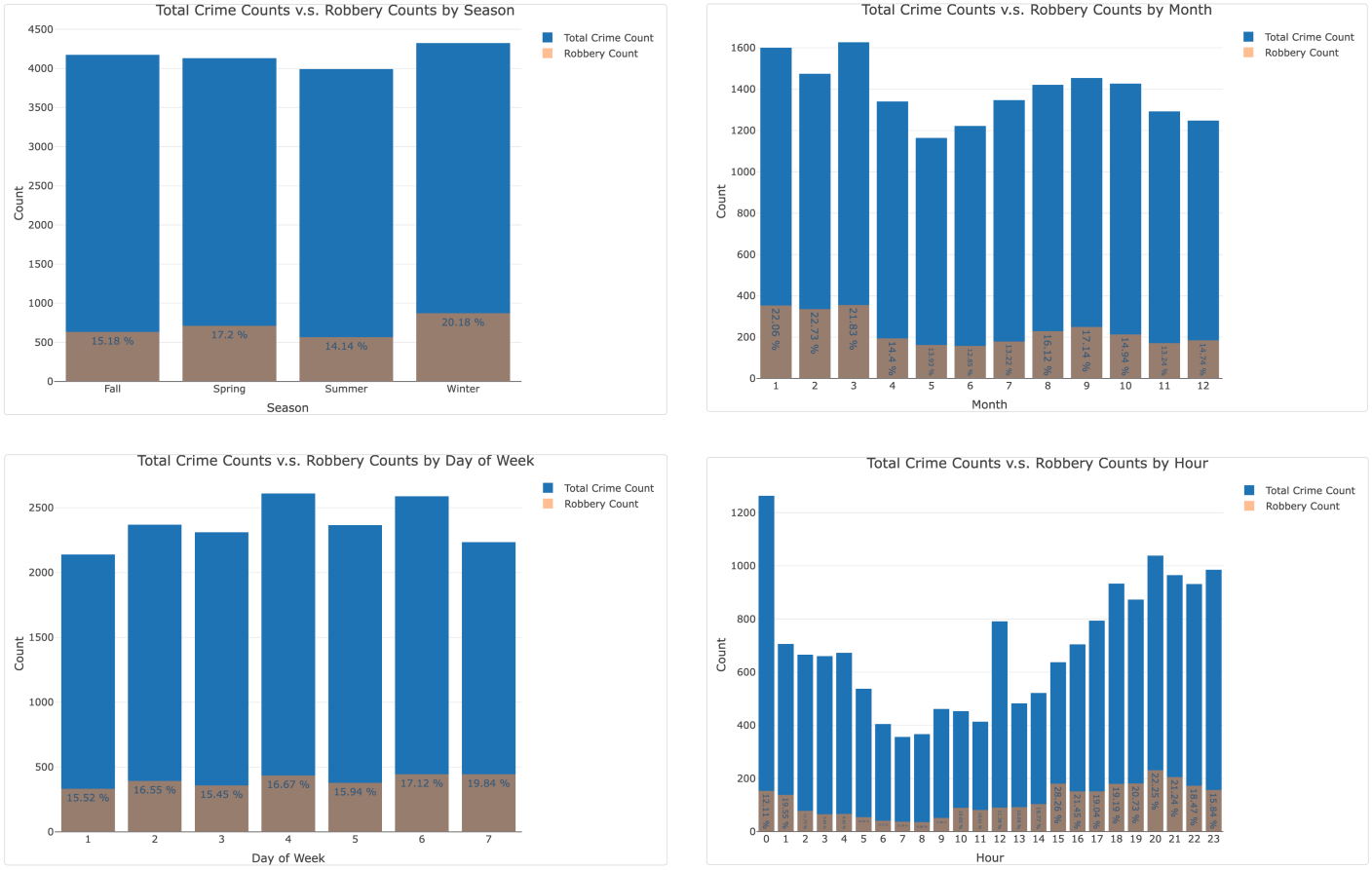

Figure 5: Total Crime Counts and Robbery Proportion by Season / Month / Day / Hour

The total crime counts and proportion of robbery crime across different seasons, months, days of the week, and hours of the day are shown in Figure 5. We can observe that not only do the total crime counts increase during holidays, weekends, and darkness, but also the proportion of robberies.

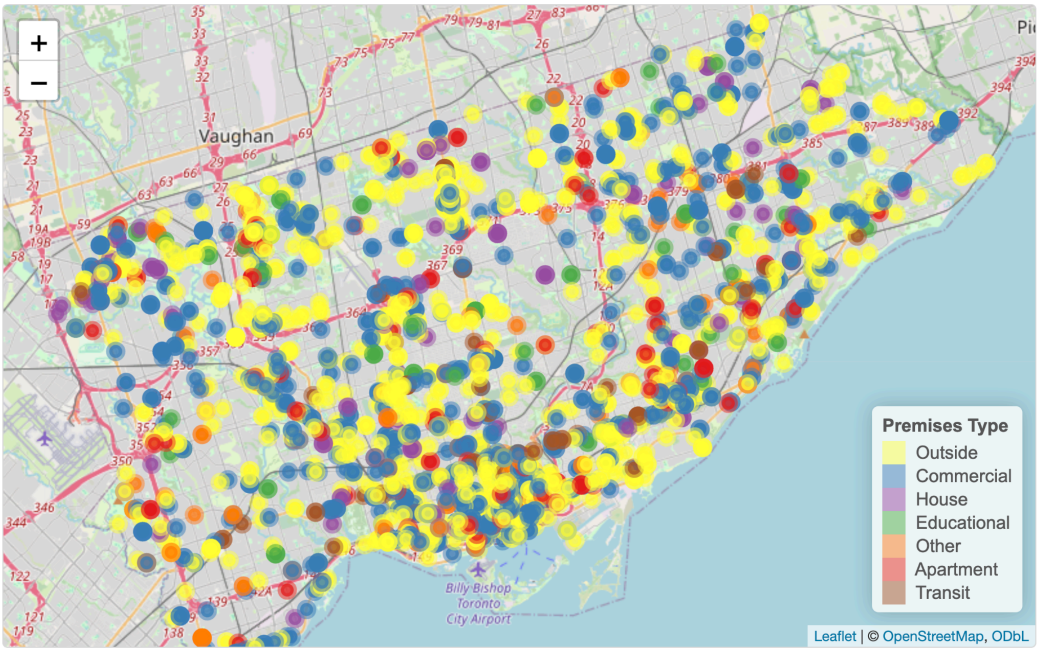

Figure 6: Map of Toronto Robbery Crimes by Premises Type in 2020

The geographical distribution of robbery incidents across Toronto is shown in Figure 6. We can observe clusters of robbery crimes on the map, as well as areas that do not have any reported robberies. This suggests that there may be a spatial influence on the probability and occurrence of robbery crimes.

3.2 Statistical Model

3.2.1 Generalized Linear Model (GLM)

In this section, a logistic regression model was built using the `glm` function. The model was trained on the training data and evaluated using various performance metrics on both training and test datasets. The parameters used for this model were as follows:

- Family: Binomial
- Link Function: Logit

3.2.2 Classification Tree

A decision tree model was constructed using the `rpart` package. The model was trained with the training data and evaluated using metrics such as accuracy, precision, recall, and F1 score. The following parameters were employed:

- Minimum Split: 150
- Minimum Bucket: 5
- Optimal Complexity Parameter: 0.002796479

3.2.3 Bagging

The bagging ensemble method was applied using the `randomForest` package. The model was trained and evaluated.

3.2.4 Random Forest

The random forest algorithm was implemented using the `randomForest` package. The model's performance was assessed on both training and test datasets.

3.2.5 Boosting

Gradient boosting was performed with the `gbm` package. The model was trained and evaluated, and the following parameters were utilized:

- Number of Trees in the Ensemble: 100
- Maximum Depth of the Tree: 1
- Learning Rate: 0.1

3.2.6 XGBoost

Extreme gradient boosting was conducted using the `xgboost` package. The model was trained and evaluated, and the following parameters were specified:

- Number of Boosting Iterations: (1:10) * 50
- Maximum Depth of the Tree: (1, 3, 5, 7)
- Learning Rate: 0.5

3.2.7 Model Performance

Each model underwent rigorous evaluation, including training accuracy, testing accuracy, precision, recall, and F1 score assessments, providing a comprehensive understanding of model performance. The performance scores are shown in Table 2: Performance Metrics of Different Models.

Performance Metrics of Different Models

Model	Train_Accuracy	Test_Accuracy	Precision	Recall	F1_Score
GLM	0.8382214	0.8324639	0.6756757	0.0294811	0.0564972
Classification Tree	0.8417477	0.8342697	0.5916667	0.0837264	0.1466942
Bagging	0.9980218	0.8733949	0.6817420	0.4799528	0.5633218
Random Forest	0.9973338	0.8836276	0.7723577	0.4481132	0.5671642
Boosting	0.8395975	0.8338684	0.7631579	0.0341981	0.0654628
XGBoosting	0.8879333	0.8430979	0.5993976	0.2346698	0.3372881

Table 2: Performance Metrics of Different Models

In Table 2, Bagging has the highest training accuracy, and Random Forest has the second highest training accuracy, which is reasonable due to their design nature. Comparing test accuracy, we can see that both Random Forest and Bagging perform better than other models, achieving accuracies of 0.8836276 and 0.8733949, respectively.

Upon examination of precision, recall, and F1 Score, it is evident that Random Forest and Bagging still exhibit better performance than other models. Random Forest demonstrates comparatively better performance in precision and F1 Score, while Bagging excels in recall. Although Bagging achieves a higher recall score than Random Forest, and recall is an important measure for identifying actual robbery cases, which aligns with our primary objective, Random Forest achieves much higher precision than Bagging, despite the

difference in recall for both models. Since our goal is to help people avoid becoming victims of robbery, it is worth noting that higher precision may be more effective in reducing unnecessary panic, as precision measures the proportion of actual robberies among predicted robberies. Consequently, I selected Random Forest as the final model.

3.2.8 Final Model

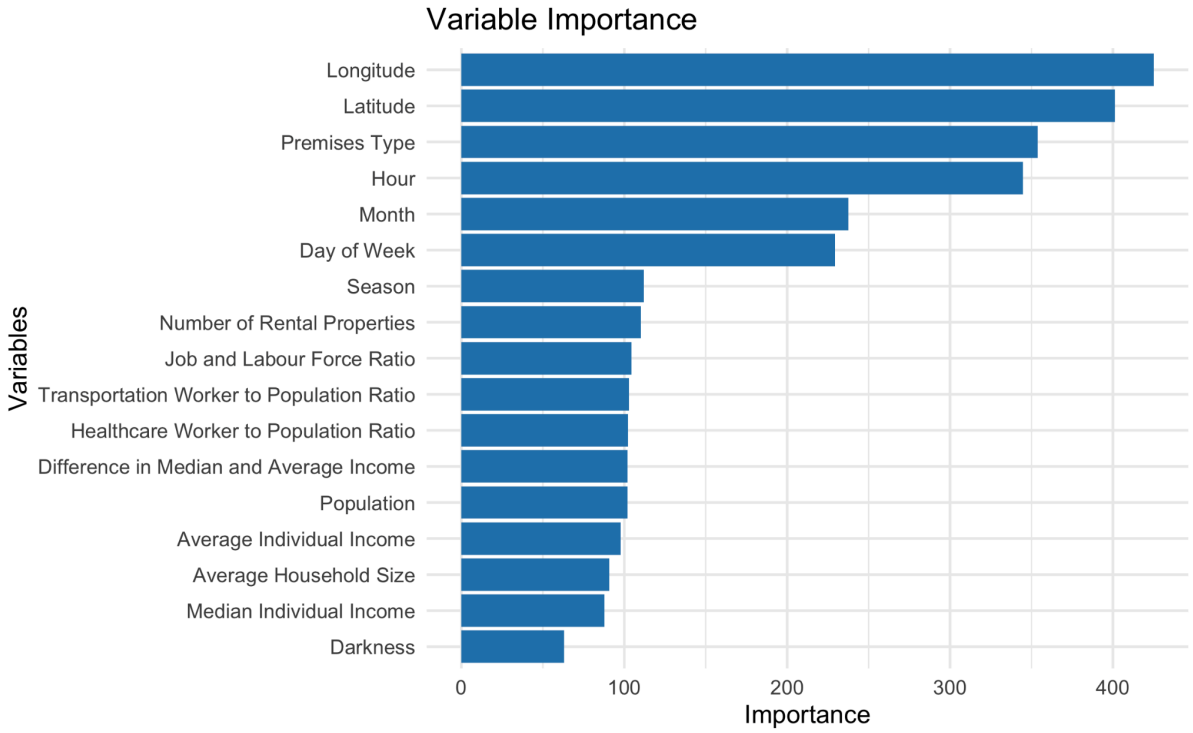

Figure 7: Variable Importance of Random Forest

The variable importance of the fitted Random Forest model is shown in Figure 7.

Spatial indicators such as *Longitude*, *Latitude*, and *Premises_Type* are ranked as the three most important variables in the model.

Following spatial indicators, the second most important factors are temporal factors, including *Hour*, *Month*, *Day_of_Week*, and *Season*.

Finally, we come to the socio-economic factors of neighborhoods, ranked in descending order of importance: *Number_of_Rental_Properties*, *Permanent_Job_and_Labour_Force_Ratio*, *Transportation_Service_Worker_and_Population_Ratio*, *HealthCare_Service_Worker_and_Population_Ratio*, and so on. It is worth noting that the newly created variable *Difference_in_Individual_Median_Average_Income* is ranked higher than both *Individual_Median_Income* and *Individual_Average_Income*, implying success in variable creation.

4 Conclusion

4.1 Summary of Findings

4.1.1 Advanced EDA Findings

A notable cluster of high robbery counts emerges in neighborhoods characterized by low median income but high permanent job labor force ratios. Moreover, robbery counts surge in darkness across all premises types, with commercial and educational areas showing particularly wide ranges of counts during dark hours, indicating heightened risk. Commercial and outdoor premises consistently exhibit higher robbery counts compared to other types, suggesting specific vulnerabilities in these areas. Temporal trends also unveil distinct patterns, with robbery counts peaking during winter months, coinciding with holiday-related crime rate increases, and surging in summer, aligning with higher temperatures. Additionally, weekends witness elevated robbery counts compared to weekdays, and robbery incidents rise from afternoon to midnight, sharply declining at 2 AM, reflecting dynamic temporal dynamics shaping criminal activity.

4.1.2 Statistical Model Findings

Random Forest and Bagging models demonstrate superior performance across various metrics, including training and testing accuracy, precision, recall, and F1 score, surpassing other models in predictive efficacy. Particularly, Random Forest achieves higher precision, essential for accurately identifying actual robbery cases and reducing unnecessary panic. Variable importance analysis underscores the significance of spatial indicators such as *Longitude*, *Latitude*, and *Premises_Type*, followed by temporal factors like *Hour*, *Month*, *Day_of_Week*, and *Season*. Socio-economic factors, including *Number_of_Rental_Properties* and *Permanent_Job_and_Labour_Force_Ratio*, also play pivotal roles in the model, with newly created variables showing promise in enhancing prediction accuracy and understanding underlying trends in robbery incidents.

4.2 Limitations

4.2.1 Time Range Constraint

The study's reliance on data solely from 2020 introduces limitations stemming from temporal constraints. While using a single year streamlines analysis, it may lead to overfitting in models and susceptibility to the influence of unique events specific to that year, such as the COVID-19 pandemic. The census data, collected every five years, presents challenges in capturing nuanced fluctuations within each interval, potentially impacting the robustness and generalizability of findings.

4.2.2 Area / City Constraint

Focusing exclusively on Toronto and its 158 neighborhoods restricts the generalizability of the study's outcomes beyond this specific urban context. Spatial factors like longitude and latitude are constrained to Toronto's geography, limiting the applicability of the developed models to other cities with different environmental and infrastructural configurations. This

narrow focus may overlook variations in crime dynamics across diverse urban landscapes, compromising the broader utility of the study's insights.

4.2.3 Variable Creation Constraint

The approach to variable creation introduces uncertainties and potential biases. For instance, the manual creation of the *Darkness* variable based on fixed sunrise and sunset times disregards variations in daylight duration throughout the year, potentially misrepresenting crime occurrences during transitional periods. Similarly, the computation of ratios, such as the *Permanent_Job_and_Labor_force_Ratio*, overlooks the complexities of employment dynamics, potentially leading to skewed assessments of socio-economic factors' impact on crime. The decision not to normalize variables like the *Number_of_Rental_Properties* may obscure nuanced relationships with crime, as it fails to account for differences in population density or neighborhood area size, limiting the depth of insights into mobility patterns and social cohesion within neighborhoods.

4.3 Possible Future Development

4.3.1 Time Range Constraint

To address the limitations posed by the temporal constraint, future studies could explore longitudinal data spanning multiple years to capture temporal trends and mitigate the risk of overfitting to specific years. Utilizing advanced statistical techniques such as time series analysis could provide insights into evolving crime patterns, considering factors like seasonal variations, economic fluctuations, and societal changes. Additionally, leveraging data from multiple sources beyond the census, such as crime incident reports and socioeconomic indicators, could enrich analyses and enhance the robustness of findings by providing a more comprehensive understanding of the socio-temporal dynamics influencing crime.

4.3.2 Area / City Constraint

To enhance the generalizability of findings, future research could adopt a multi-city or regional approach, encompassing diverse urban contexts with varying socio-economic and environmental characteristics. This would enable the development of models and interventions that are more adaptable and transferable across different urban landscapes. Moreover, incorporating geospatial analysis techniques could facilitate the identification of common spatial patterns and risk factors associated with crime, allowing for the implementation of targeted interventions tailored to specific geographical contexts.

4.3.3 Variable Creation Constraint

Future studies could employ more sophisticated methods for variable creation that capture the complexities of socio-economic and environmental factors influencing crime. For instance, utilizing machine learning algorithms or spatial analysis techniques could help derive more accurate measures of variables like darkness or ratios of service workers to population. Moreover, exploring alternative approaches, such as clustering analysis or principal component analysis, could enable the identification of latent variables that better capture the underlying dynamics of crime. Additionally, incorporating stakeholder perspectives through

community engagement and participatory research approaches could ensure the relevance and validity of variables considered in predictive models and interventions, enhancing their effectiveness in addressing real-world challenges.

5 Ackowledgements

This project is conducted by Kaiwen Yang as part of the course JSC370H1: Data Science II, offered by the University of Toronto in Winter 2024, instructed by Professor Meredith Franklin, with teaching assistance from Jun Ni (Jenny) Du.

In addition, I extend my sincere gratitude to Professor Meredith Franklin in the Department of Statistical Sciences and the School of the Environment at the University of Toronto for her invaluable supervision and support, which greatly contributed to the completion of this project.

6 References

[1] “Hot and bothered? Associations between temperature and crime in Australia”, Heather R. Stevens & Paul J. Beggs & Petra L. Graham & Hsing-Chung.
Available at:
https://www.researchgate.net/publication/331499161_Hot_and_bothered_Associations_between_temperature_and_crime_in_Australia

[2] “Do homicide rates increase during weekends and national holidays?”, Alison Baird, David While, Sandra Flynn, Saied Ibrahim, Navneet Kapur, Louis Appleby & Jenny Shaw.
Available at:
https://www.researchgate.net/publication/332279464_Do_homicide_rates_increase_during_weekends_and_national_holidays

[3] Toronto Open Data Catalogue.
Available at:
https://open.toronto.ca/dataset/neighbourhood-profiles/

[4] Toronto Police Service Public Safety Data Portal.
Available at:
https://data.torontopolice.on.ca/pages/open-data

7 Appendix

Source of the analysis available at:
https://github.com/KaiwenYangUT/Robbery_Crime_Classification