CSC311 Project Proposal | Fall 2024
Jun Hee Lee | Yuxin Fan | Hyun Jo Jang

Type of project: Application project

Description of problem and motivation
When emergency services, specifically the police, receive a report or call about a crime the initial information delivered to them is rarely detailed enough to paint a complete picture. The type of crime is a crucial piece of information to make proper response measures from the police department, yet the report may not be detailed enough or delivered late. Therefore, we wanted to identify the intricate relationships and patterns between the different features of the crime reports in Toronto to assist with this problem. The information that can immediately be captured by a police report is the time and date of the call, and the location where the call is made.

Precise Objectives
Our main objective is to implement a model that can accurately predict the crime type based on the features of the crime report such as the time of day, day of the week, month of the year, neighbourhood and precise location. We will model the relationship between the [time, date, and location data] and [crime type (e.g. gunshot, theft, etc.)].

We will leave the addition of features regarding the report's location as a potential project expansion for the sake of task management. By including these factors, the neighbourhood data may gain additional context, which could improve the model's prediction accuracy. The possible factors to consider for each neighbourhood would be:

- Neighbourhood Improvement Areas - https://open.toronto.ca/dataset/neighbourhood-improvement-areas/
- Housing Density - https://open.toronto.ca/dataset/social-housing-unit-density-by-neighbourhoods/
- Green Spaces - https://open.toronto.ca/dataset/wellbeing-toronto-environment/

Feasibility & Evaluation of the Project
We will first implement a KNN and logistic regression model as our baselines to evaluate the initial accuracy of crime-type predictions based using the raw features available on the dataset such as the time and location of the report. This will provide us with a foundation of which algorithms work best with our dataset and guide improvements in feature engineering and model tuning.

To measure the success we will plot training curves with a confusion matrix to help us understand which crime types are frequently misclassified, indicating potential areas for model improvement.

Based on the feedback from the above methods, we may either: deploy new models (e.g. Neural Nets) or perform additional feature mapping and/or manipulation to further improve performance.

Dataset

Three datasets for this project are Toronto public datasets collected and processed by the Toronto Police Service (TPS) and Statistics Canada. The final training/testing dataset will contain the crime event description (e.g. case occurred date and time, crime type, neighbourhood code, coordinates), population, crime rate per year and crime rate per 100,000 population (includes all crime categories). The precise location of all reported cases will be anonymised to ensure privacy.

Sources of raw datasets:

1. Toronto Police Service (2024), *Major Crime Indicators Open Data*, https://data.torontopolice.on.ca/pages/open-data
2. Open Data - City of Toronto (2024), *Neighbourhood Profiles,* https://open.toronto.ca/dataset/neighbourhood-profiles/
3. Open Data - City of Toronto (2024), *Neighbourhood Crime Rates,* https://open.toronto.ca/dataset/neighbourhood-crime-rates/

Background Research

Dong and Liu (2022) conducted a regression analysis examining the relationship between green space and assault rates in Toronto, using TPS data. Their findings indicate an inverse relationship: higher percentages of green space are associated with lower assault rates. They further suggested that walkability, which tends to reduce green space, is correlated with higher assault rates.

Similarly, Mohammadi et al. (2022) used kernel density estimation (KDE) to study homicide rates across Toronto neighborhoods, analyzing built environment and socioeconomic factors through datasets from TPS and Toronto's Open Data portal. Their study revealed multiple ways in which these factors relate to variations in homicide rates.

Additionally, Law et al. (2014) applied a Bayesian spatiotemporal model to analyze crime hotspots, finding that the Bayesian approach provided superior predictions compared to frequentist methods. This is largely because the Bayesian model could offer forward-looking estimations rather than simply summarizing historical data.

References

Dong, Y., & Liu, H. (2022). Does green space influence assaults? evidence from Toronto, Canada. *Urban Forestry & Urban Greening*, *68*, 127481. https://doi.org/10.1016/j.ufug.2022.127481

Law, J., Quick, M., & Chan, P. W. (2014). Analyzing hotspots of crime using abayesian spatiotemporal modeling approach: A case study of violent crime in thegreatertorontoarea. *Geographical Analysis*, *47*(1), 1–19. https://doi.org/10.1111/gean.12047

Mohammadi, A., Bergquist, R., Fathi, G., Pishgar, E., de Melo, S. N., Sharifi, A., & Kiani, B. (2022). Homicide rates are spatially associated with built environment and socio-economic factors: A study in the neighbourhoods of Toronto, Canada. *BMC Public Health*, *22*(1). https://doi.org/10.1186/s12889-022-13807-4