CIND 820: Big Data Analytics Project: The Effects of Resource Accessibility on Toronto Crime Rates

Alyzeh Jiwani
501106857
25 July 2022

Abstract

Research Question:

Ontario Government spending on adult correctional facilities is sizeable and rapidly increasing yearly, hitting $1,116,561,000 in the 2020/2021 period ($806,764,000 per year with an average increase per period of $42,913,250, adjusted in 2002/2003 dollars) (Government of Canada, Statistics Canada, 2022) . What factors are contributing to high crime rates, and where should our resources be focused in order to potentially see a reduction in these values in the future?

Context:

I intend to answer these questions by investigating the relationship between crime rates and socio-economic factors in Toronto by neighbourhood. I will also look at what resources each neighbourhood has and which ones are most significant in attempting to predict crime rates. I plan on looking at two scenarios in particular; how the socio economic factors and resources of a neighbourhood effect crime rates in the same year, and how they effect crime rates three years in the future. This should provide insight into what initiatives should be implemented in communities so as to reduce the probability of major and violent crimes occurring later on. If effective, given levels of significant factors at a certain point in time, future likelihoods of crimes may be accurately predicted. Moreover a better understanding of what factors need to be targeted by government resources, specific to individual neighbourhoods should be achieved.

Data Utilized:

I aim to use a vast cohort of datasets from the City of Toronto open data repositories.

There are numerous datasets available providing insight into an extensive number of

potential features that may be used in my analysis. These datasets have standardized IDs

for individual neighbourhoods and districts, which facilitates analysis across the different

data sources. Additionally, much of the data has been consistently collected periodically

throughout the past decade, thus easing the process by which we can compare the effects

of the same variables over a long period of time. The data comes in a wide variety of

formats, including csv, xsls, geoJSON, and unstructured text files.

The two most notable datasets I plan on using are:

https://open.toronto.ca/dataset/neighbourhood-profiles/(2001, 2006, 2011, 2016 Census

data)

https://open.toronto.ca/dataset/neighbourhood-crime-rates/ (dataset consisting of counts

of each type of crime by neighbourhood, as well as population for 2021, similar datasets

for previous years will also be used)

While these provide the core data that I require, there are numerous other relevant data

sources in the repository that I intend to utilize as well.

Techniques and Tools:

The primary themes of my project are predictive analytics, data mining and knowledge discovery. I feel that these will be the most practical in resolving my proposed question. As mentioned earlier, I will be investigating many data sources from the repository to conduct my research. This is because the City of Toronto has individual datasets per socio-economic/demographic attribute per five year period, thus there is no one data source that contains all the information I require. Thus, I will compile the relevant information from these sources into a single relational database for ease of access. From there I will be able to export my data to my software of choice, where I will conduct an exploratory data analysis to evaluate the statistical significance of the features under consideration and reduce them to those that are the most applicable. I will create charts to visualize the effects the features have on the response variable, calculate and examine their correlation coefficients and variances. I will implement other feature selection and extraction algorithms if applicable to further cut down on the dimensionality of my data. I will then create my predictive model using the earlier dated data for the test and train sets and, once content with the quality of my model, use it to predict future crime rates of each Toronto neighbourhood based on the current values of statistically significant features.