CIND 820: Big Data Analytics Project

Alyzeh Jiwani
501106857
16 May 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 demographic and socio-economic factors in Toronto by neighbourhood. I will also look at what resources each neighbourhood has, how accessible to the populations they are, and if they are well utilized. In particular, I plan on investigating how the levels of these factors during an individuals childhood/youth whilst living in these neighbourhoods affect crime rates years later during adulthood. This should provide insight into what initiatives should be implemented in communities when its minor populations are still young, so as to reduce the probability of major and violent crimes occurring when they are older. If effective, given levels of significant factors at a certain point in time, future likelihoods of crimes 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. Further, there is substantial data available on demographics of individual neighbourhood populations enabling me to make inferences on these individual populations with their unique contexts in mind. 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. It is possible that there will be the slight use of text mining and sentiment analysis, especially as there are some relevant semi-structured data sources in the repository, but this would be the extent of its use.

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 likely 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.

References

Government of Canada, Statistics Canada. (2022, April 20). Operating expenditures for adult correctional services. Retrieved May 16, 2022, from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3510001301&pickMembers%5B0%5D=1.8&cubeTimeFrame.startYear=2016%2B%2F%2B2017&cubeTimeFrame.endYear=2020%2B%2F%2B2021&referencePeriods=20160101%2C20200101