Capstone Abstract: Toronto Vision Zero Road Safety Data
Jason Kim

Context

Earlier this year, the City of Toronto launched the Vision Zero Challenge, which was an 8-week hackathon to create a solution for the City's Vision Zero plan. Vision Zero is a global movement of cities to eliminate all pedestrian deaths from vehicle collisions. It began in Sweden in 1997 and has since been adopted by multiple global cities including New York, London, and Toronto.

In Toronto's case, its Vision Zero plan began in 2017 and its goal is to eliminate pedestrian deaths by 2021. Recently with the spate of cyclist and pedestrian deaths, the City Council increased funding by $22m in order to accelerate its Vision Zero plan. In 2017, Toronto experienced 162 pedestrian or cyclist KSIs (Killed or Seriously Injured). Pedestrian fatalities are on an uptrend with 37 deaths last year and 2018 on pace to be a record year for pedestrian and cyclist deaths.

I participated in the Vision Zero Challenge, but felt the work needs much improvement and much use of data analysis skills. Because the Challenge had very restrictive requirements that caused me to include various uncorrelated variables on an a priori basis and because not enough time or guidance was given, I ended up creating a very barebones solution in the form of the Toronto High Injury Network. This solution maps the high risk areas and streets for KSIs, but has many weaknesses from a statistical or data analysis point of view.

I would like to continue my work on preventing pedestrian and cyclist deaths in Toronto for my capstone so that with your guidance and without the restrictions of the hackathon I can create a truly predictive model the City can use in its Vision Zero strategy.

In October, I will be making a presentation to the General Manager of Transportation Services related to Vision Zero based on the preliminary work I've already done creating the Toronto Highly Injury Network.

This means this project has a very real chance to be actually implemented by the City, which is why I require your further guidance and the time to improve and expand my preliminary work during the capstone.

Capstone Abstract: Toronto Vision Zero Road Safety Data
Jason Kim

Research Questions

1) What micro and macro factors are correlated to a KSI, including age, location, speed limit, street, etc.?
2) What intersections, streets, and neighbourhoods are the most at-risk for KSIs?
3) Can KSIs be predicted on a per street or per intersection level? How accurate can the model be based on very sparse data?
4) What techniques are there for working with sparse data? Because there are many more non-KSIs vs. KSIs (obviously non-collisions are not reported), this means a collision event is actually rare, and a death or serious injury even more rare.

Data Sources

Because of my preliminary work on the problem, I have the advantage of having worked with several relevant data sources, but there are potentially more that could be added using join operations.

The primary dataset of interest is a dataset created by the Transportation Services Big Data team which contains every reported vehicle, pedestrian, and cyclist collision in Toronto from 2008-2017. This dataset is open access and available here.

I also plan to use many of the open access datasets available through the City's open data portal, including its Centrelines geospatial dataset, which contains the lengths, classification, names, and geo coordinates for every street in the City.

Techniques

The premise of the Vision Zero strategy contains both a classification problem and a regression problem:

- It is a classification problem because whether a certain street is more at risk for a KSI requires classification
- It can also be considered a regression problem if you want to predict the expected KSIs on a given street

Capstone Abstract: Toronto Vision Zero Road Safety Data
Jason Kim

I will also need to use techniques for working with sparse data (pedestrian deaths and hospitalizations from car collisions is sparse even though the dataset contains over 20,000 rows of collisions) and datasets that contain many hundreds of columns but less than 200 rows.

Although the data is reasonably clean, a lot of ETL techniques need to be used since there are so many potential datasets to join and many variables that can created out of existing variables pulled across different datasets (feature engineering).

I will need to use R for classification and data cleaning, Hadoop/Hive for working with the primary dataset, QGIS or ArcGIS for working with geospatial data (I learned how to use these during my preliminary work), and Tableau for creating visualizations.