PREDICTING TRANSIT RIDERSHIP FOR A CITY

Capstone Project 1: Project Proposal Siri Surabathula

Problem

As urban population levels rise, they lead to increasing traffic congestion, which in turn worsens mobility and carbon-emissions problems for cities. Public transit systems can be key in managing the urban mobility and emissions crises. However, the success of transit systems depends on their ability to scale and their responsiveness to regional variations in demand. Both short-term and long-term prediction models are hence vital to effectively solving problems of urban mobility.

This project aims at predicting transit ridership at different stations across a large city over yearly and multiple-year time frames using weather, traffic, taxi/cab, gas price and demographic data. The spatial aspect of transit, taxi/cab, traffic and demographic data will be taken into account while building the model for prediction.

Client

Prospective clients for the model could be various city transportation authorities (like NYC DOT, NYC MTA) and city planning authorities (like City of New York)

City planning authorities could use the ridership predictions to efficiently allocate resources and modulate the frequency of trains and buses along specific routes. Multiple-year data could be used to plan effectively for temporary and permanent capacity expansions along certain routes.

Data

The following data will be used for New York City over the years 2010-2018

- Hourly transit ridership across NYC. Transit type could be one or more of - bus or light-rail / tram / train
    - Turnstile data (no. of persons entering and exiting a station) for NYC MTA Transit. From May 2010 to Present collected every 4 hours across ~500 stations in NYC
      http://web.mta.info/developers/turnstile.html

- Weather data for NYC
    - NYC Central Park weather station data from the National Climatic Data Center

- Taxi/Cab ridership data
    - The official TLC trip record dataset containing data for over 2 billion taxi trips from January 2009 through July 2017, covering both yellow and green taxis http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
    - From FiveThirtyEight, publicly available data covering nearly 19 million Uber rides in NYC from April–September 2014 and January–June 2015. https://github.com/toddwschneider/nyc-taxi-data

- Traffic data
    - Traffic speed data from NYC Open data https://data.cityofnewyork.us/Transportation/Real-Time-Traffic-Speed-Data/qkm5-nuaq

- Gas Price data for NYC
    - Average monthly gas prices for NYC https://www.nyserda.ny.gov/Researchers-and-Policymakers/Energy-Prices/Motor-Gasoline/Monthly-Average-Motor-Gasoline-Prices

The following data will also be investigated to study multiple-year trends

- Population data (spatial) for NYC https://www1.nyc.gov/site/planning/data-maps/nyc-population/about-data.page

- Economic Indicator data (spatial) for NYC
    - Housing rates/prices/rents
    - Employment https://www1.nyc.gov/site/planning/data-maps/open-data.page

Approach

- The datasets will be cleaned and preliminary exploration studies will be conducted to reveal correlations of transit ridership with various factors (traffic speed, cab ridership, weather, gas prices, demographics)
- Different time frames will be considered for aggregating the data (weekly, monthly, yearly and multiple-year periods) and the relationships with factors will be studied over each aggregation interval
- The spatial attributes (provided by geographic coordinates and GIS shapefiles) of the data will be leveraged to study and determine spatial relationship of factors and predictors with transit ridership at a specific station.

- Regression models will be developed for one or more aggregation interval / time frames (the available transit ridership data will be partitioned into training and test sets). Prediction will be performed on the test set.
- Trends and results will be plotted.

Deliverables

- Slide Deck - Presentation slides describing the problem, value proposition for prospective clients, method employed to solve the problem, data sources, results in tabular (aggregate) and graphical format and inferences.

- Code repository (github)