Project Proposal

Vincent Morin

11/13/2019

{Disclaimer}

Coming from my background in urban tech startups / venture capital and still in my first semester of MPP, I wanted to do a introductory project into this space by looking at general infrastructure between two cities. Conscious of the 4-week deadline, I kept it high-level. More than open to any recommendations around approach / data used / and strategy.

Statement of Purpose:

The objective of this final proposal will be to analyze and compare the precicted average total fare per trip of New York Taxi companies in 2015 with those of taxi companies in Chicago from the same year. New York data will serve as the training data, while Chicago data will serve as the test dataset.

Data Sources:

This final proposal will be conducted utilizing open source datasets from the City of New York and Chicago. The following problem sets are accessed from these sites:

- **Yellow Taxi 2015 Trip Data** - https://data.cityofnewyork.us/Transportation/2015-Yellow-Taxi-Trip-Data/ba8s-jw6u
- **Green Taxi 2015 Trip Data** - https://data.cityofnewyork.us/Transportation/2015-Green-Taxi-Trip-Data/gi8d-wdg5
- **Chicago Taxi Trips Data** - https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew

The New York datasets can be accessed via the NYC OpenData site and can be readily downloaded in CSV format. Likewise, the Chicago dataset will be accessed via the Chicago Data Portal and can also be downloaded in CSV format.

Strategy, Analysis, & Machine Learning:

The datasets will require some working and general overhaul to prepare for proper data analysis. In order to get an accurate model of New York Taxi companies, both the Green Taxi 2015 dataset and Yellow Taxi 2015 dataset will need to be combined. Additionally, some missing observations exist and some extranneous columns will be removed such as: pickup_longitude, pickup_latitude, RateCodeID, dropoff_long, etc.. The key columns which will be selected are pickup_datetime, dropoff_datetime, fare_amount, total_amount, etc.. The Chicago dataset will require additional cleaning as well, as it contains data ranging from

2009 (188 million rows). Data from 2015 will therefore be pulled out of the dataset for the testing model.

Specific variables which will be used in the predictor models will be total fare, number of passengers, distance, time, etc. Each will require some manipulation of the data included in the datasets to be properly utilized.

Visiualizations included in the model will be illustrations of different variables of the data, including peak hours of trips, distribution of fare across number of riders, miles per trips, etc..

The packages utilized in this project will be:

- tidyverse
- lubridate
- ggthemes - visualizations
- textdata - data manipulation
- topicmodels
- recipes
- caret

While conducting analysis of the data, the folowing models will be utilized:

- Traditional linear model
- KNN model
- Decision Trees model
- Random Forest model

Based on their applicability, the best will be chosen as a measure of the relationship between the two datasets. Results of each will be included in the final report.

Measuring Success:

The measure of success in this project will be to draw a conclusion of whether or not fare per trip, based on variables of distance, passengers, time, etc., will be an accurate predictor of fare per trip in Chicago.