Revised Project Proposal

Vincent Morin

11/13/2019

Note

A few changes to note: after considering the notes provided following my initial proposal, I decided to trim down the size of the data I'm analyzing, and also bring in an additional dataset for added variables.

Due to the size of this project, I will only focus on New York City Green Taxi data, rather than both Yellow, and Chicago data as well. Further, I'm bringing in a 2016 weather dataset for additional predictor variables – I will also be using Green Taxi 2016 Data instead of 2015. My methodology on this switch, is that this project is designed for me to communicate a comprehension of the class material, not necessarily solve complex answers to some particular data, as I am still in my first semester of MPP.

Further, my definition of success is centered around what the completed project should look like, regardless of a "successful" outcome.

Statement of Purpose:

This project will be an exploratory analysis into the relationships between tip-per-ride, and different external variables such as weather, trip duration, fare cost, time/day, etc.

While analyzing relationships, attention will be given to opportunity for further exploration.

Data Sources:

- **Green Taxi 2016 Tip Data** - https://data.cityofnewyork.us/Transportation/2016-Green-Taxi-Trip-Data/hvrh-b6nb - obtained via New York Open Data.
- **2016 New York Weather Data, Central Park** - obtained via Kaggle, available open source from weather.gov. http://w2.weather.gov/climate/xmacis.php?wfo=okx

Methods:

Data Wrangling:

- **Green Taxi 2016 Trip Data**

This dataset is massive, relative to the size of this project. In order to conduct analysis on this data set, it will be scalled and subsetted in the following way:

- Data will only cover date range 1/1/2016 - 6/30/2016

- Data will filter out entries without tips: the goal of this analysis is to analyze the relationship of tip-size with external factors, not frequency or likeliness to tip. Although, given available resources, this would be an interesting analysis to conduct with this data.
- Several columns which contain unknowns, constant values, or irrelevant data will be removed.

-2016 Central Park Weather Data

This weather only covers daily averages of observations. It will be inner-joined with the Taxi data, and matched with each row based on date of observation.

Several additional variables will be created based on the data, such as logical values for days where snow and rain did, or did not occur.

Visualizations:

- Histograms - to understand the distribution of data.
- Barplot - for different analysing methods.
- Lineplot - for comparisions.

Machine Learning:

This project will be conducting supervised machine learning processes to discover what variables best effect tip-size. Models used will include:

- Linear Regression (lm)
- K-nearest Neighbors (knn)
- Random Forest (ranger)

In order to conduct this analysis, the data will need to be transformed via the recipes package; categorical variables will be converted to dummies and ranges for continuous varibles will be normalized/standardized.

Measuring Success:

In this case, a successful project will be one that conducts analysis on the relationships inherent to the different variables, related to tip-size per trip. This will include providing professional visualizations of the data to assist in reader comprehension. Supervised machine learning models run on the data will also be explored, detailed, and illustrated. Finally, the project will consist of a section related to recommendations for next steps in conducting further analysis, as well as a conclusion section which outlines what conclusions were derived from the analysis/models.

{Original Proposal}

{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.