Final Project Proposal--Group 1

Anthony Pagan, Tommy Jenkins, Violeta Stoyanova, Todd Lisa, Peter Kowalchuk, Eleanor Secoquian

Introduction:

“The Partnership for New York City's one-page study asserted that “excess congestion” deprives the five boroughs and the suburbs of Long Island, Westchester and Rockland counties and northern New Jersey $20 billion annually,” according to Crain’s [1]. Moreover, emergency services reports life or death consequences based on NYC traffic conditions, according to multiple sources [2]. We seek to analyze publicly accessible datasets in the service of better understanding the factors that influence traffic as encapsulated in the Kaggle competition: New York City Taxi Trip Duration. In a recent article, Emil Skandul, writing for City and State NY, cited the importance of Mayor De Blasio hiring a technologically savvy commissioner to lead the Taxi and Limousine Commission to rival Lyft and Uber [3,4]. In all of these regards, we believe being able to predict traffic through the proxy of taxi trip duration can be highly relevant to the welfare of the NYC population.

Objective:

The purpose of this project is to build various models in an attempt to predict the trip duration of yellow taxis in New York City. We will be using the Kaggle data that was part of a Playground Predict Competition. The data was originally published in NYC Taxi and Limousine Commission (TLC) [5,6].

Methodology:

Using the techniques we've learned in the class, like classification, model diagnostics and transformation, we will explore data to find new patterns. And just like what is required in the Kaggle contest, we will try to predict the duration of each trip in the test set.
We will build multiple linear regression modeling and then summary to interpret the results. We’ll further analyze the results by adding discrimination in the model and then assess the discrimination with ROC curve.

We will consider following the competition evaluation metrics and results format, only in cases where it fully agrees with the recommendations of our DATA 621 coursework [5]:

The evaluation metric for this competition is Root Mean Squared Logarithmic Error.

The RMSLE is calculated as

$$\epsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i + 1) - \log(a_i + 1))^2}$$

Where:

ϵ is the RMSLE value (score)
n is the total number of observations in the (public/private) data set,
p_i is your prediction of trip duration, and
a_i is the actual trip duration for i.
$\log(x)$ is the natural logarithm of x

Submission File

For every row in the dataset, submission files should contain two columns: id and trip_duration. The id corresponds to the column of that id in the test.csv. The file should contain a header and have the following format:

```
id,trip_duration
id00001,978
id00002,978
id00003,978
id00004,978
etc.
```

Data Description:
The features we will initially consider can be found under the data tab in the competition page -- and we may choose to augment the data with information such as weather [5, 7]:

File descriptions

- train.csv - the training set (contains 1458644 trip records)
- test.csv - the testing set (contains 625134 trip records)
- sample_submission.csv - a sample submission file in the correct format

Data fields

- id - a unique identifier for each trip
- vendor_id - a code indicating the provider associated with the trip record
- pickup_datetime - date and time when the meter was engaged

- dropoff_datetime - date and time when the meter was disengaged
- passenger_count - the number of passengers in the vehicle (driver entered value)
- pickup_longitude - the longitude where the meter was engaged
- pickup_latitude - the latitude where the meter was engaged
- dropoff_longitude - the longitude where the meter was disengaged
- dropoff_latitude - the latitude where the meter was disengaged
- store_and_fwd_flag - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
- trip_duration - duration of the trip in seconds

Data Sources:

1. https://www.kaggle.com/c/nyc-taxi-trip-duration/data#
2. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
3. https://www.kaggle.com/c/nyc-taxi-trip-duration/discussion/37192#latest-572412 (external data to be considered as ways to augment our data)

References:

1. Traffic congestion costs metro economy $20 billion a year: study. Crain's New York Business. https://www.crainsnewyork.com/article/20180118/POLITICS/180119895/traffic-congestion-costs-metro-economy-20-billion-a-year-study. Published January 18, 2018. Accessed November 13, 2019.
2. FDNY: Traffic — Not Bike Lanes — is to Blame for Increased Response Times. Streetsblog New York City. https://nyc.streetsblog.org/2019/09/19/fdny-traffic-not-bike-lanes-is-to-blame-for-increased-response-times/. Published September 19, 2019. Accessed November 13, 2019.
3. Skandul, E. How New York City taxis can get ahead of Uber and Lyft. CSNY. https://www.cityandstateny.com/articles/opinion/commentary/how-new-york-city-taxis-can-get-ahead-of-uber-and-lyft.html. Published October 17, 2019. Accessed November 13, 2019.
4. Sanders A. De Blasio struggles to find new Taxi and Limousine commissioner after first pick's disastrous performance at Council Hearing. nydailynews.com. https://www.nydailynews.com/news/politics/ny-city-council-de-blasio-official-tlc-commissioner-jeff-lynch-20191003-cfx27yihafcebj4e62wytsprpu-story.html. Published October 3, 2019. Accessed November 13, 2019.
5. New York City Taxi Trip Duration | Kaggle. Kaggle.com. https://www.kaggle.com/c/nyc-taxi-trip-duration/data#. Published 2017. Accessed November 13, 2019.

6. About TLC - TLC. Nyc.gov. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Published 2019. Accessed November 13, 2019.
7. Use holidays as a feature | Kaggle. Kaggle.com. https://www.kaggle.com/c/nyc-taxi-trip-duration/discussion/37192#latest-572412. Published 2017. Accessed November 13, 2019.