CMSC 12300 Project Proposal

Group Name: Fast & Furious
Group Member: Hyun Ki Kim, Andi Liao, Zunda Xu, Weiwei Zheng

1. Data set

The main data set our group will work on is New York City (NYC) Yellow Taxi trip records data. Unlike green taxi and For Hire Vehicles (FHVs), yellow taxi provide transportation exclusively through street-hails, so it will reveal New Yorkers' behavior well. The data was collected and provided to the NYC Taxi and Limousine Commission by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs. The data set is available from 2009 to 2017 from the NYC government website and we are planning to use data from January 2009 to June 2016, with file size of average 2.4 GB per month (total 217 GB). The data set includes nearly 10,000,000 taxi trips for each month including pickup and drop-off time, location (longitude and latitude), passenger count, trip distance, fare amount, and tip amount.

There are some supplementary datasets we might use to complement the analysis. Firstly, we might refer to fix-yahoo-finance API to get historical data on different market indices on yahoo finance. We can get historical daily information such as opening price, closing price, trade volumes. Secondly, we will combine taxi trips with weather data to control for weather factors. We will use APIs such as WunderWeather to collect hourly weather information at zip code level. Lastly, we are going to use taxi-zone data which contains 265 major spots affiliated to the NYC Taxi dataset to get specific area of each ride's locations.

2. Hypotheses

Our project goal is to predict macroeconomic outcomes by using large-scale micro data. Specifically, we will use New York City's yellow taxi trip record to forecast stock market. We believe market index as a proxy for how well economy is doing, so we can predict market index by analyzing individual's economic activity. Firstly, we will categorize taxi trips by its purpose (going to airport, dating, dining, shopping, etc.). We will seek for correlation between trip records and weekly or monthly market indices for different sectors (beverages, insurance, oil & gas, airlines, etc.). Secondly, we will categorize taxi trips by different pick up and drop off location. These places could be large districts or landmarks such as Time Square, Broadway, Wall street, or more specific locations such as Empire State Building, Museum of Modern Art, New York Stock Exchange.

3. Algorithms

Based on the dataset and hypothesis mentioned above, we plan to use the following algorithms:

1. Random Forest & Neural Network: For predicting the macroeconomic outcomes, we will use random forest and neural network algorithms to train the model as we have plenty of features in the Taxi dataset.

2. Matching Pair Analysis: As for detecting traffic patterns, we will pair taxi trips based on similar features, like pickup location and drop locations, and then compare their difference in total time. Ideally, we will find the traffic pattern between two specific locations changing with different levels of time.