Final Project - Economy in the Big Data World

In the final project you are required to analyze a dataset by Uber. You will predict the demand for Uber rides within 1-kilometer distance from the New York Stock Exchange, for each quarter of hour, during September 2014 evenings. Especially, you'll focus on the correlation between Uber rides and the S&P 500 daily return. Your predictions will be evaluated based on a list of time intervals, when you do not have access to the real number of Uber pickups that already have been occurred during those times.

The Data

For this work you have 3 different datasets:
1. uber_train - This is the fundamental data you are going to analyze and use for modeling. This file contains data on over 4.5 million Uber pickups in New York City from April to July 2014.
Note: You have a separate file for each month and you can use as many data as you like in order to train your model. Think carefully how much data to use.
2. uber_test - For the list of time intervals in this data, you will make your predictions, based on the model you created.
3. External data source – The S&P 500 daily return is not supplied by Uber, so you have to download this data source from the Internet by yourself. In addition, you are allowed to use any other data source from the Internet /anyplace else that you think will improve your prediction accuracy.

Project Description

Section A: Prediction *without* the S&P 500 daily return

This section will be divided into 3 parts:

1. **Data Rearrangement** - Your training data is not in the same format as the test data. In addition, it contains data on pickups outside the target zone that do not took place over the relevant hours (17:00 – 00:00). Thus, your first mission is:

- Use New York Stock Exchange coordinates (latitude = 40.706913, longitude = -74.011322) in order to subset the data for pickups within 1-kilometer distance from the New York Stock Exchange. The figure at the end of this document illustrates this target zone.
  Tip: use 'distm' function (with fun = distHaversine) from 'geosphere' package in order to calculate distance between two geo-points.

- Transform your training data in a way that it will be in the same format as the test data.
  Tip: the appendix may help you with this step.

- Keep only time intervals between 17:00 – 00:00.
  - Tip: 'hour' function from 'lubridate' package may help you with this step.

- If you choose to use additional data source, merge it with your training data and extract as many features as you like from the data.

2. **Exploratory Analysis** - present and explain descriptive statistics (including charts) about the dataset. Emphasize in your analysis:
   i. What actions were done on the data in order to reach its final form.
   ii. The relationship between the dependent variable and other variables you will later use in your modeling as predictors.
   As we showed in class, data analysis is used in order to gain intuition for better predictions and models.

3. **Model Estimation** - build your best model. After evaluating few different models with different variables, choose your best model. Explain why you decided to choose this one, explain the model and how you built the different variables. Describe clearly how you will eventually estimate the demand for Uber rides at any time interval in the target zone.
   Notice: In this part, you are not allowed to use the S&P 500 daily return for your predictions.

Section B: Prediction *with* the S&P 500 daily return

Again, this section will be divided into 3 parts:

1. **Data Rearrangement** – Merge the S&P 500 daily return with the data you generated in the previous section.

2. **Exploratory Analysis** - Present and explain descriptive statistics (including charts) about the new variable. Emphasize in your analysis the relationship between this variable and the dependent variable / other variables in your dataset.

3. **Model Estimation** - Repeat part 3 from the previous section. Notice that this time you have to use the S&P 500 daily return as a predictor.

Section C: Submission

For each time interval (row) in the data called 'uber_test.csv' – predict the demand for Uber rides, once with the model you built in section A and second with the model you built in section B. Your output should be in the form of a table with 3 columns:

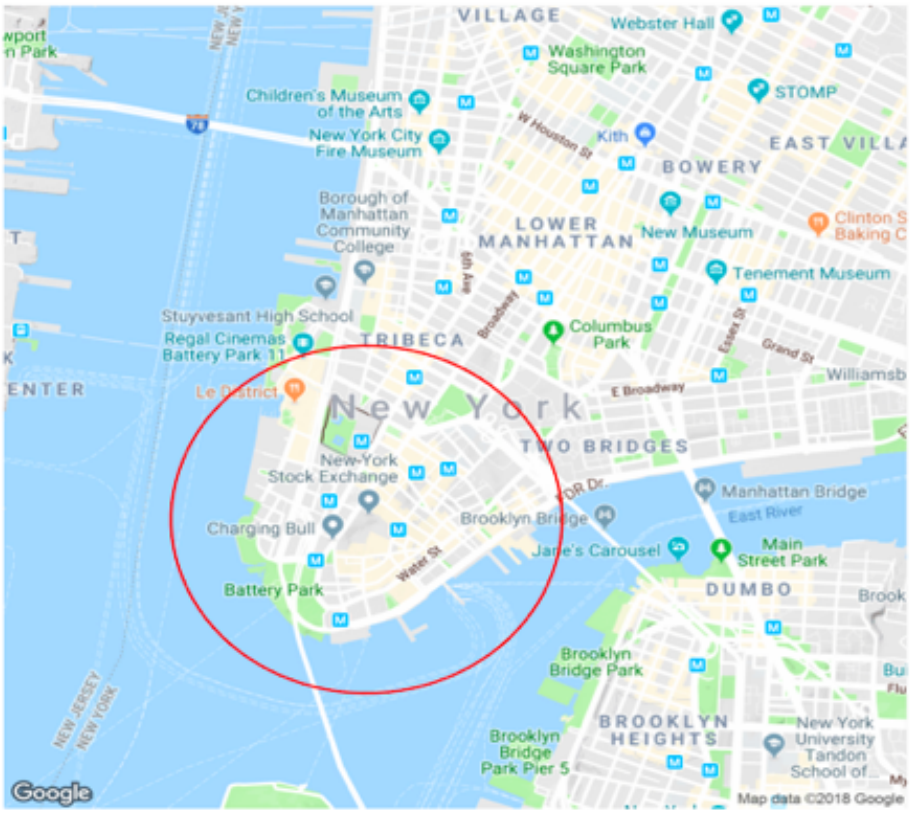
Time_Interval	pick_num_withoutSP	pick_num_withSP
2014-09-24 18:00:00	13	11
2014-09-24 18:15:00	61	63

Submitted documents:
You should submit the following 3 files (zip files aren't allowed):
1. The 'uber_test.csv' file that looks exactly as the table above (**same columns names**).
- Make sure that the file name is 'uber_test.csv'
- Note that those predictions must be based on the model you selected and showed in the two other documents.
2. The R code containing your script / commands.
3. A PDF document (up to 7 pages) that shows:
    - Main points you have found in your exploratory analysis (including charts and descriptive statistics).
    - A summary and explanation of your selected models for prediction, including coefficients values, model outputs, and an evaluation of the model on validation set.
    - A short brief regarding other models you examined and ruled out.
- If you have used modeling methods that weren't taught in this course, you must add an appendix of one page (which is on top of the other 6 pages) which describes these methods.
- DON'T add code / command to the PDF file.

Due dates
By 01-March-2020 9:00 submit on moodle (the course's website) all of the files listed above.

Grading
The grading will be done by the following key:
- Quality of explanations and analysis (40%)
- Accuracy of prediction in comparison to other teams (for both sections) (60%)

Illustration of the target zone



Good luck!