Final Assignment - Predict Uber demand in New York
Economy in the Big Data World 2024
Dr. Roy Sasson; Ophir Betser

Submit no later than the date:
2024-10-27 23:59

Important details:

1. Submission is in **groups of 2**.
   a. Only one member of each group should submit the assignment in the moodle.
   b. Please clearly state who the group members are within the report and in the name of the report file as follows:
   EinBDW24_final_(name #1)_(name #2).pdf
   example:
   EinBDW24_final_RoySasson_OphirBetser.pdf
2. The assignment's share is **70%** o f the course grade.

Files to submit:

1. A detailed report with your answers to the questions, along with relevant graphs and tables. The report should be focused - do not submit more than 6 pages, and emphasis will be placed on the aesthetics of the graphs and / or tables attached to it.
   a. Each graph should be accompanied by titles explaining its contents.
   b. Make sure the graphs you attach are legible, clear, understandable and aesthetic.
   c. There's no need to write a lot of text, answer in a short and focused manner.
   d. The file format for the report should be **pdf**.
   e. The language of the report should be **Hebrew**, font David 12, space between lines 1.5.
2. csv files that contain your prediction for Section C. The files should be exactly in this format (only this one sheet, with those column names, filled with your prediction - other format will cause loss of grade.
3. your R code.

Grading
The grading will be done by the following key:

1. Quality of explanations and analysis in the **report** (60%)
2. **Accuracy of prediction** in comparison to other teams, base on **MSE** (lower is better) (40%)
   a. prediction in part c.1.
   b. prediction in part c.2.

About
In the final assignment you are required to analyze a dataset by Uber. You will predict the demand for Uber rides within 1000-meters radius distance from Empire State Building (lat-lng: 40.7484, -73.985), for each 15 minutes interval, during 10 to 30 of September 2014. 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 have already occurred during those times.

For this project you have 2 different datasets, and one test_set that you will use 2 times, once with each train dataset:

1. train_sets - This is the fundamental data you are going to analyze and use for modeling. This file contains raw data on over 3.5 million Uber pickups in New York City from 1st April to 9 of September 2014.
   a. a version with all the raw uber pickups - train_raw_data.csv.
   b. a version with all the raw uber pickups that are not in the radios of 2000 meters from the Empire State Building - train_raw_data_dists_more_then_2000.csv.

2. test_set - contains aggregated data of 10 of September to the end of September. We keep the data in time intervals of 15 minutes, and keep only the hour of the day above 17:00 to 00:00.
   you will make your predictions, based on the model you created, and fill the 'number_of_pickups' for each time interval.

Your goal in this assignment is to provide a quality data analysis of Uber's operations in New York and give a forecast that is as accurate as possible for the future. The grade on the part of the forecast is **competitive** between the course groups.

Research Questions:

Section A: 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 take place over the relevant hours (17:00 - 00:00). Thus, your first mission is to create an aggregate data set that is relevant to your prediction mission.

a. those function can help you for your goal:
    i. geosphere::distHaversine - will help you to calculate distance from two points.
    ii. purrr::map2_dbl - will help you to apply functions on your data.
    iii. lubridate::floor_date - will help you to create time intervals.
    iv. you can use 'leaflet' to create great maps in r.
b. make sure that your dataset is clean.
c. for this section - chatGPT & stack overflow are your best friends.

Section B: Telling story with the data

In order to provide a quality forecast for the future, our data must be prepared for analysis, and we need to create EDA. Note that this section is iterative, and it requires updating every time you learn something new from the data.

1. **Data Preparation**

- Add to your data set columns that will help you in the analysis.
  Date base data:
  You can start by adding date-base columns, like the day of week or the week of the year for each row.
  Data from the Web:
  You are required to **add at least 3 explanatory variables** to your dataset from the web.
  Data that can help in forecasting can be very diverse. You can use stock data, weather data, data from Google Trends or any data source you find and find relevant.
  merge what you find with your data-sets.
  Explain why you think the variables you have added will be relevant to the prediction.
- Clean your data set if you think it is a good idea. Whether you perform a cleansing or not, explain your thought process regarding the matter.
- Handling missing values (if there are any).

2. Exploratory Analysis

present and explain rich 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.

iii. You can add insights from additional sources of information or from your personal knowledge to interpret the data and describe the business environment.

As we showed in class, data analysis is used in order to gain intuition for better predictions and models. This part is very similar to the first assignment of the course.

Section C: Forecast for the future

1.

In this section you have to provide a forecast with the help of a model. Explain how you chose to train your model, how you choose which variables to put into it, and how you checked that it does not suffer from overfitting. Write this in the report.

After selecting a final model, use it to make a **prediction** on the test set.

save your model predictions in a csv with the same format as the test set and call it

a. EinBDW24_final_pred_c1_(name #1)_(name #2).csv

2.

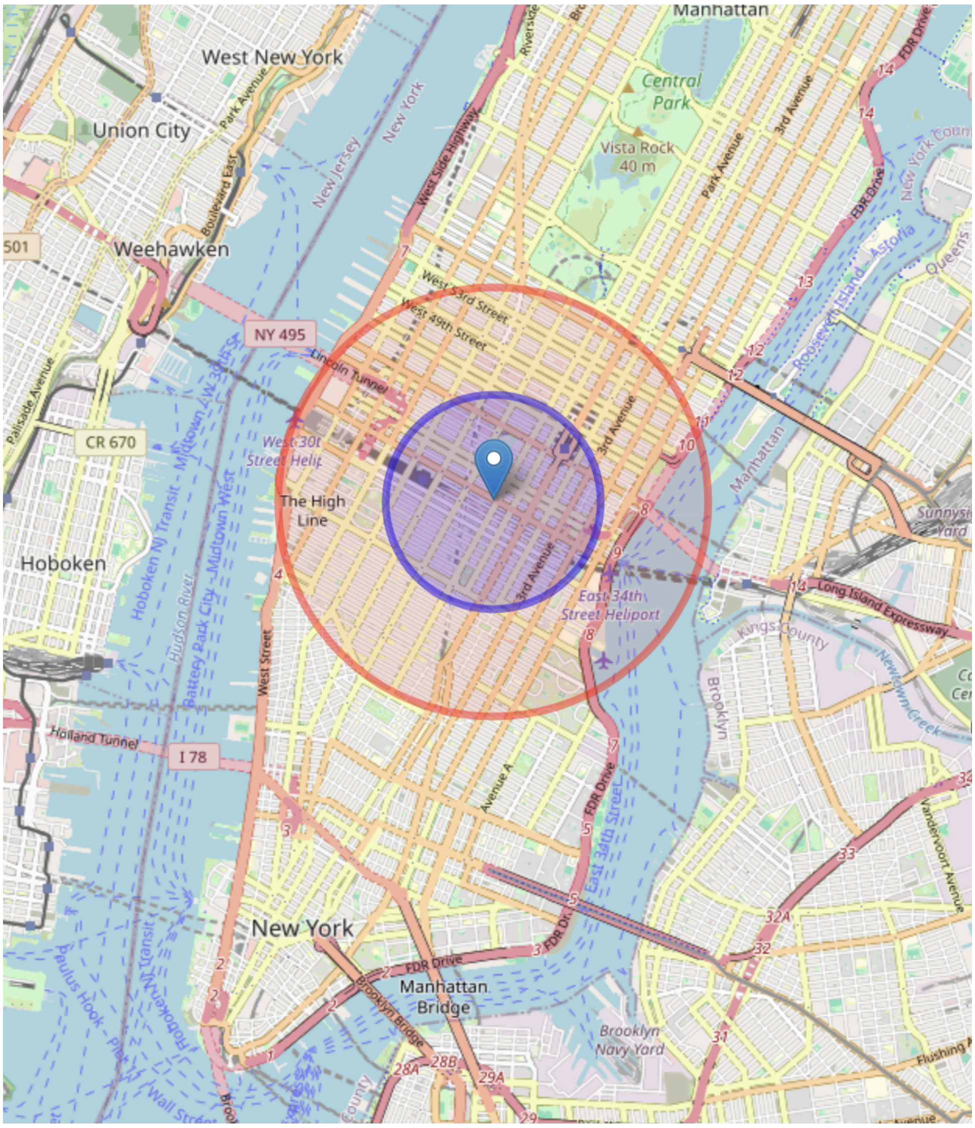
Now, use only train_raw_data_dists_more_then_2000.csv to create your prediction.

1. find the best way to create a prediction model using this limitation
2. create a model
3. use it to make a **prediction** on the test set.
4. save your model predictions in a csv with the same format as the test set and call it
5. EinBDW24_final_pred_c2_(name #1)_(name #2).csv

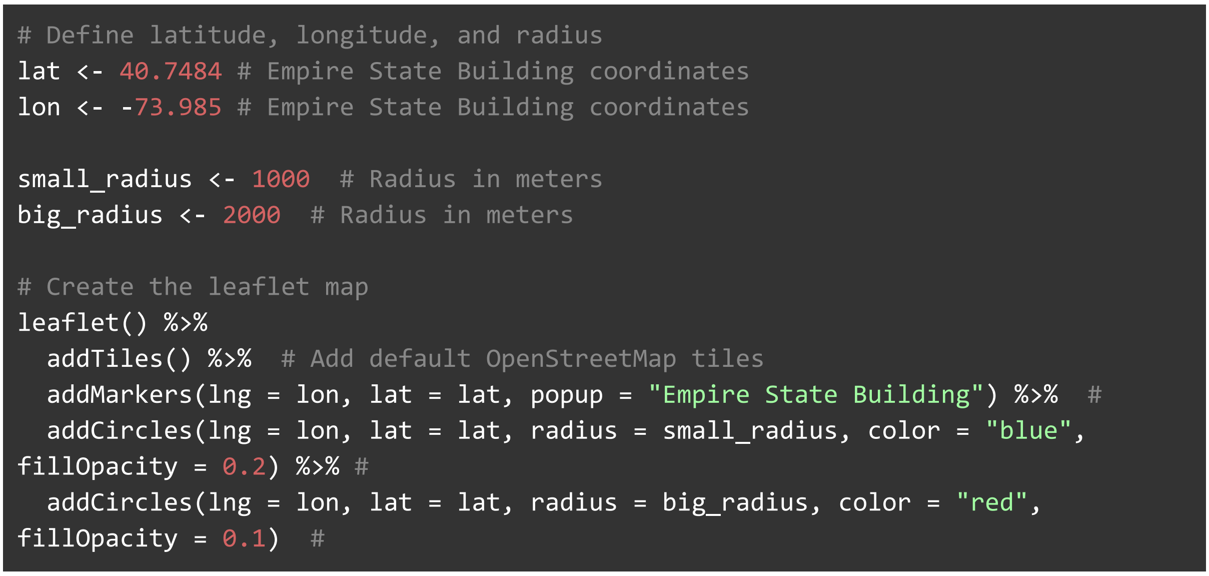
add in this section explanation on how you trained your model and why.

Appendix

Area of the data: 1000& 2000-meters radius distance from Empire State Building



code to create the map:



```
# Define latitude, longitude, and radius
lat <- 40.7484 # Empire State Building coordinates
lon <- -73.985 # Empire State Building coordinates

small_radius <- 1000  # Radius in meters
big_radius <- 2000  # Radius in meters

# Create the leaflet map
leaflet() %>%
  addTiles() %>%  # Add default OpenStreetMap tiles
  addMarkers(lng = lon, lat = lat, popup = "Empire State Building") %>%  #
  addCircles(lng = lon, lat = lat, radius = small_radius, color = "blue",
fillOpacity = 0.2) %>% #
  addCircles(lng = lon, lat = lat, radius = big_radius, color = "red",
fillOpacity = 0.1)  #
```