IDS 575 PROJECT REPORT

PROJECT FINAL REPORT

RICHA GUPTA | rgupta46@uic.edu | UIN: 678751252

Project Summary

I. OBJECTIVE

Using the dataset (released by the NYC Taxi and Limousine Commission) to predict what the duration of the trips in the test set will be. The output of the prediction is of the form "<tripid>,<duration in seconds>"

II. CREATIVE CONTRIBUTIONS AND INFERENCES

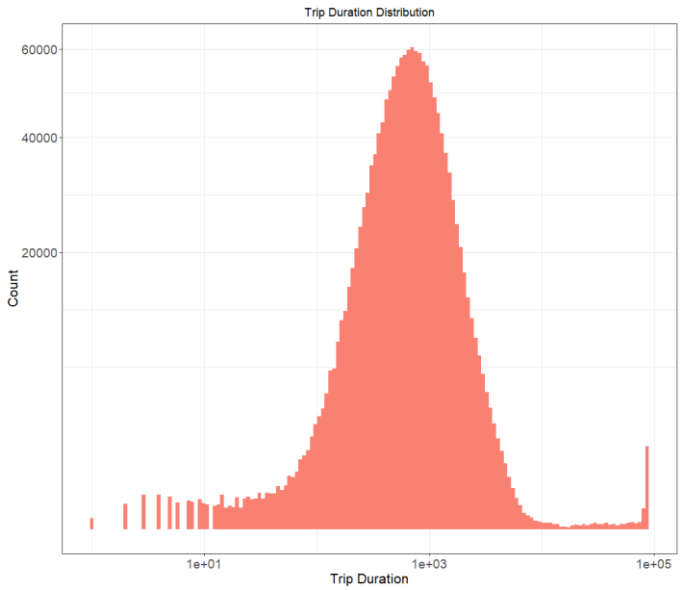
I. FEEL OUT THE DATASET:

I started by summarizing the training and test data sets.
The training set contains 1,458,644 observations, the test set contains 625,134. The training data set contains 11 variables while the test set contains 9 (the target variable 'trip_duration' is to be predicted, and based on that the "dropoff_datetime" is obviously not present as well).
I also observed certain interesting facts about the dataset:

- Average trip duration is 1000 seconds (i.e 17 minutes)
- There are 2 different vendors driving the cabs, 1 being used more than the other

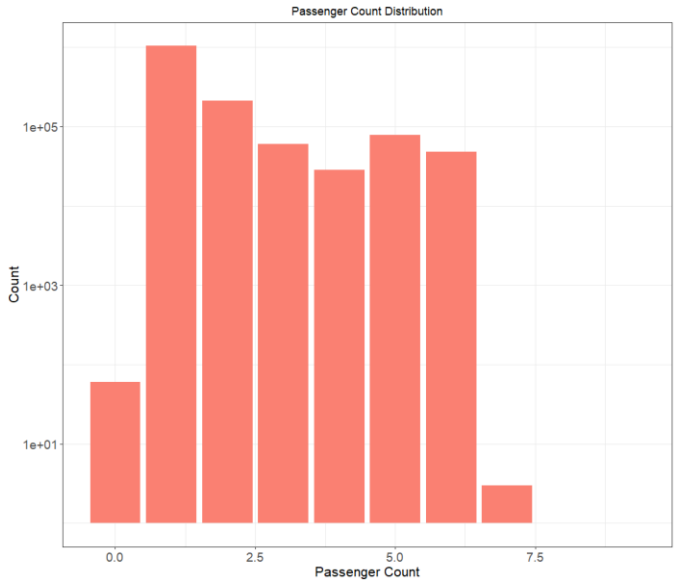
II. FINDING OUTLIERS:

To detect outliers I checked the distributions of the various attributes . Based on the summary data I notice some odd values for the target variable 'trip_duration' and 'passenger_count'. So I plotted their distributions to analyze them further.



From the distribution of 'trip_duration' above, I saw that there are certain trips which seem to last more than 24 hours! These are obviously outliers and so I decided that they should not be part of the modelling training set.

Next we focus on the 'passenger_count' variable.

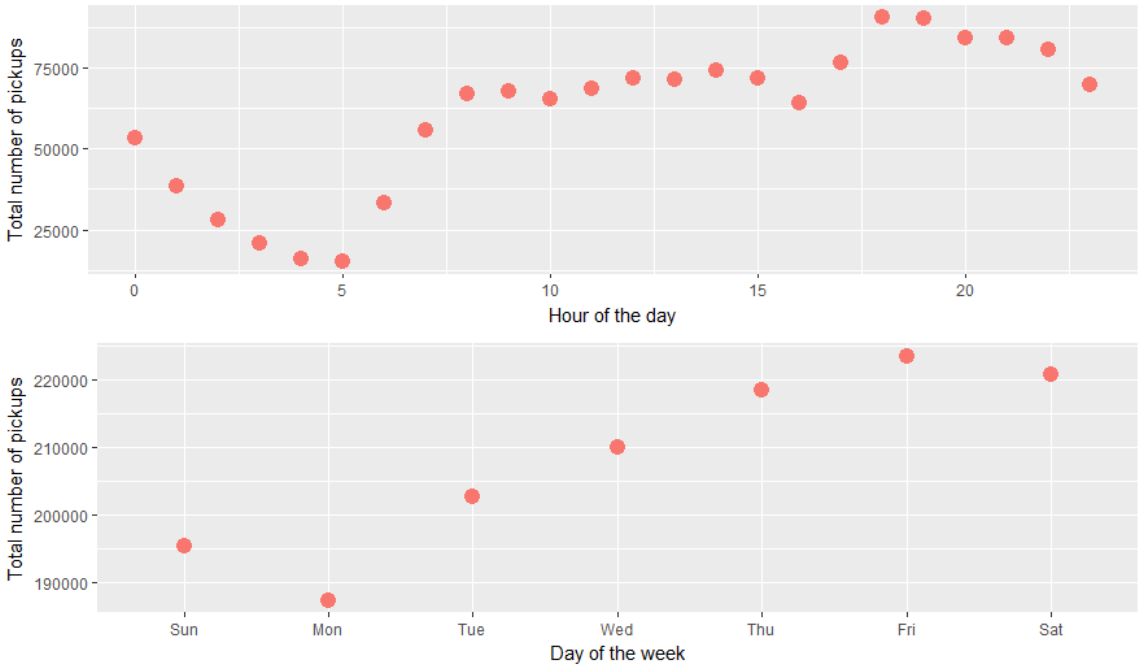


I observed that there were 60 trips with 0 passengers. Although this is an obvious outlier, it is possible that the test set may also contain similar data. However, since the goal is to make an effective and more importantly ***generalizable*** model, I chose to remove these from the training set.
This graph also showed an interesting observation- most people take taxis alone. (The highest number of trips is for passenger count=1). This insight could be extremely useful while trying to introduce ride-sharing services in NYC, or allowing taxis to participate in ride-sharing.

III. CREATING USEFUL VARIABLES:

To enable effective modeling, I created additional variables from the given pickup and drop-off datetimes and longitude and latitudes.

TEMPORAL FEATURES: I created variables to show the day of the week, the month and hour of the day from both pickup and drop-off datetimes. The motivation to do this was to separate out the influence of these individual factors on the trip duration. For instance, whether the pickup day was a weekday, or a weekend would surely influence the trip time. Similarly, depending on whether the hour of pickup was late at night or in peak traffic hours, the trip time would vary greatly. After creating these, I used plots to see how they affect trip duration.

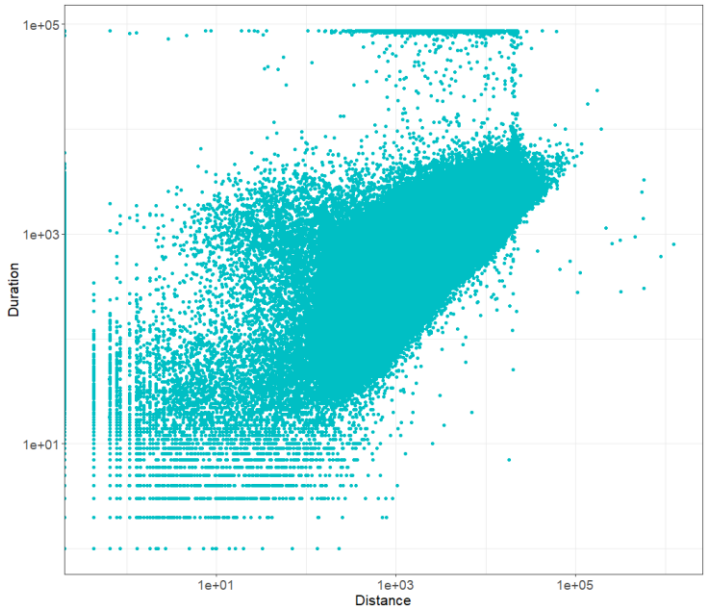


Looking at the hour of the day, we observe that the highest number of cabs are taken at around 1800 hours, i.e 6pm when people are leaving from work. You would imagine that morning rush hour peak would be similar to that in the evening. But the delta increase indicates that there are additional taxis being taken in the evening. Perhaps people going out post work to an entertainment venue? (especially if they involve alcohol, since no one would drive to those)
This information could be used to introduce more cabs in the evening to help deal with the surge in demand.

The day of the week also shows an interesting pattern. The weekends are the busiest, with Friday being the peak. The number of cabs taken taper off as the week starts, with the least being taken on Monday.

Next, I looked at the pickup and drop data over the year as a whole. Although at this level of rollup, you would expect a consistent pattern of taxi pickups. However, I observed something interesting here too. There was a sharp drop in late Jan/early Feb '16. That's winter in NYC, so it makes sense that this could be a weather-related incident. A quick Google showed me that the "Jan 22-23,'16 blizzard was NYC's Biggest Snowstorm on Record since 1869!"

SPATIAL FEATURES: The most important variable that would influence trip duration is distance. Since this was not outright present in the dataset, I created it using the longitude and latitude values of the pickup and drop-off points. The 'geosphere' library supports various geographical distance calculations. Based on research on these, I chose the distGeo() function using the WGS84 reference system which is essentially what GPS trackers use. Hence this would give the most accurate distance calculation for the trips.



Intuitively, with increasing distance, the trip duration should also increase. I plotted these variables (as above) and observed that this intuition was indeed correct.

IV. MODELING

Since the expected output- trip duration- is numerical, I performed regression analysis using OLS regression, ridge and lasso regression and compared their performance to pick the best performing one. Since this data set on Kaggle was given to be an EDA exercise, there was no actual test error available to compare to. Hence I used validation error as a metric to pick the best model. The error metric I used was RMSLE (root mean square logarithmic error) Based on the RMSLE, I found OLS to be the best performing model. I applied this model to the test set to predict the trip durations.

	Validation RMSLE
OLS	0.679359
LASSO	0.679648
Ridge	0.681861

APPENDIX

R CODE & DATA FILES

The following are the data files associated with the project.
They were submitted along with the project mid report and remain unchanged.

- Rcode- IDS575_Proj.R
- Input data files : train.csv,test.csv
- Output data file : test_prediction.csv