Data Science Capstone Project Predicting NYC Taxi Trip Times

Greg DeVore and Ryan Blosser

Deriving Knowledge from Data at Scale, Fall 2017

Can we use data from various sources to predict the length of time for a taxi trip?

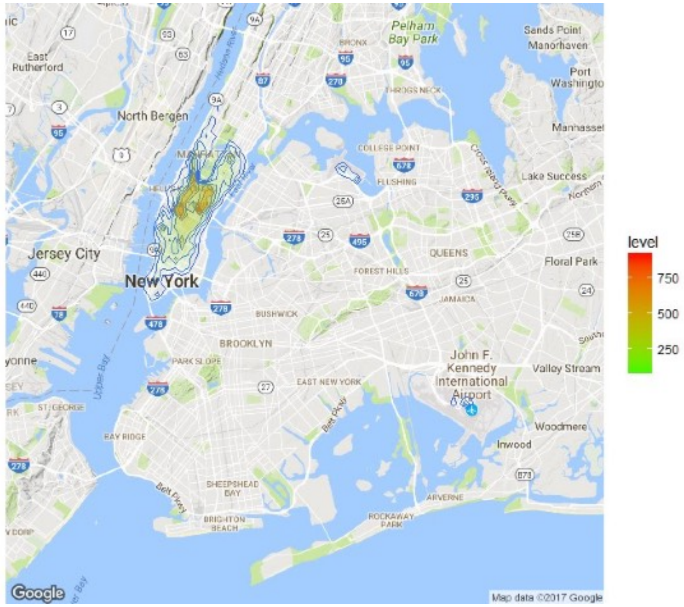
- Significant impact in driver scheduling: Over 250,000 taxi trips a day in New York City
- Can be used to predict fare cost: Over $1 billion in fares paid per year in the city

Data sources used

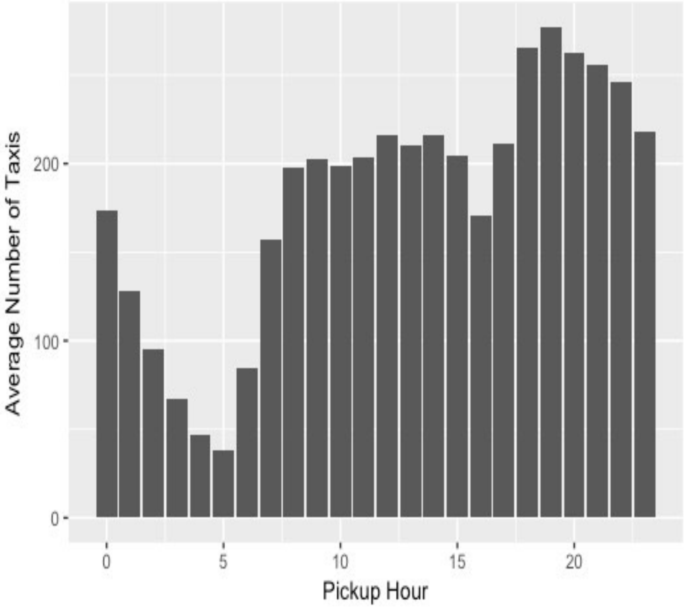
- Class-provided set of taxi trips from 2013, 1.7 million trips from a larger set of 173 million
- Hourly climatological data from JFK airport found at NOAA’s site https://www.ncdc.noaa.gov/cdo-web/
- Latitude and Longitude associated with NYC ZIP Codes https://gist.github.com/erichurst/7882666
- Data dictionary to understand some features in taxi dataset from class https://data.cityofnewyork.us/

Data cleansing procedures

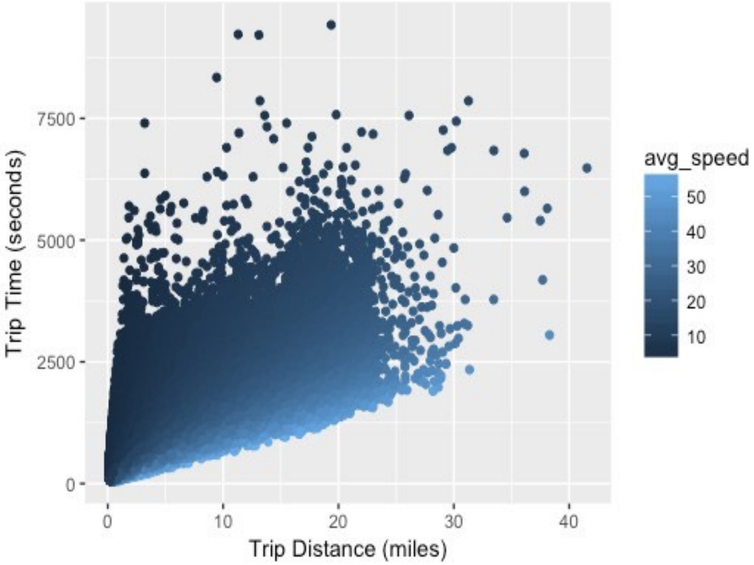
- Removing duplicate trips in data (“Medallion”, “License” and “Pickup_Datetime” was a composite key to join the two vertically-partitioned taxi trip datasets and had a small number of duplicates)
- Removing observations with extremely low or high values for both Trip Distance and a generated feature, Speed (we are given Trip Distance and Trip Times)
- Cleansing to either swap Latitude and Longitude values if they were reversed or remove the observations that had impossible values
- Filling in blank observations for weather data based on values of nearby observations (hourly data given but some hours are blank but “sandwiched” in between hours with valid values)



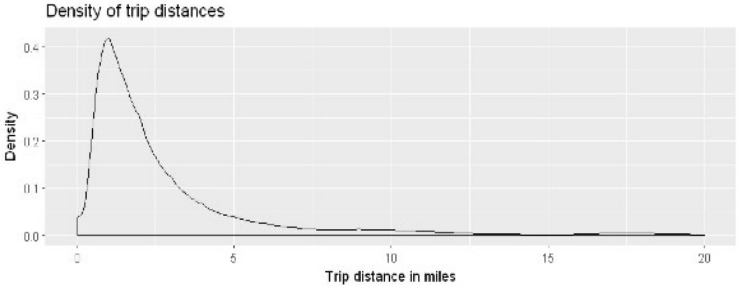
A heat map of all the pickups in the dataset (the drop-offs is very similar). There are two heavy clusters in the center of Manhattan, with a few pockets to the north and south of them. There is another pocket east of Manhattan, by the airports.



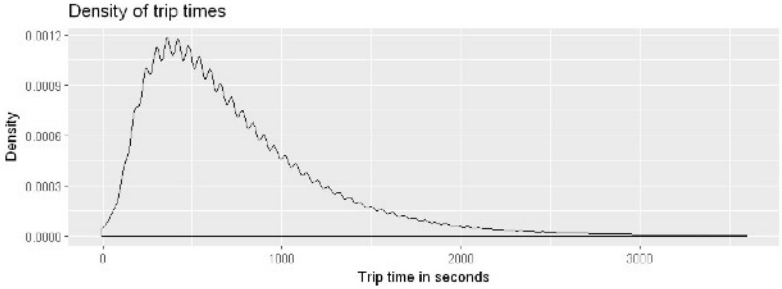
Traffic distribution by hour; this observation was used to determine “rush hour”.



A plot of trip distance versus trip time, showing a strong correlation between the two variables.



Distribution of trip distance, modal around 1 mile with a long right tail



Distribution of trip time, modal around 7 minutes with a long right tail

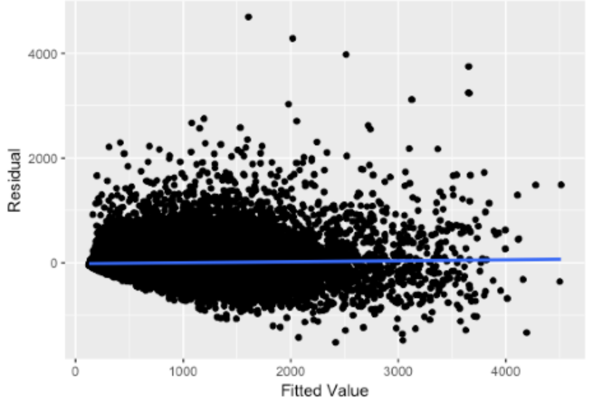
Feature selection and engineering

- We could only use data that would have been known to the driver at the time of pickup
- Date and Time engineering: splitting the Pickup_DateTime column into granular buckets
    - Day of week, Weekend Day or not, Holiday (US Federal) or not, Month
    - Time was split into single hours, as well 7 buckets per day and a binary Rush Hour value (based on observation of highest number of taxis per hour for dataset)
- Pickup and drop-off ZIP Code generated by choosing the one with the closest centroid to the coordinates of the observation
- Used k-means clustering for pickup and drop off locations
- Created direction (16-way) based on pickup to drop off location
- Created “traffic” by counting number of taxis with a pickup or drop off in a region over the last hour (drop off location was based on time at pick up)
- Binary feature for rain based on hourly weather data (used value of hour of pick up)
- Binary feature for the existence of a “toll” fee or “surcharge” fee
- We could not use money paid for the trip, this would have been target leakage

Model selection

- Trip time varies nonlinearly due to many factors, tree based model chosen due to ability to capture nonlinear behavior
    - Compared single, random forest and boosted regression models
- Random forest chosen - low RMSE, low variance, trees can be grown in parallel
    - 200 trees, 16 features, 5 features per tree, minimum 6 rides per node

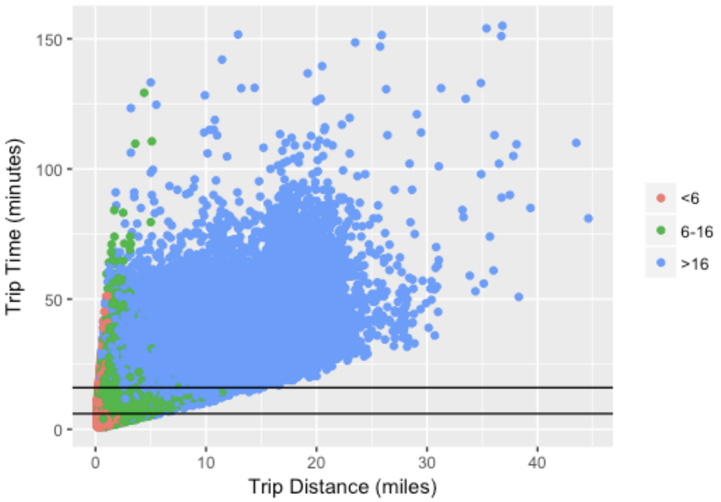
Residual vs. Fitted Values (Regression)



Model performance

- Regression (predict trip time in seconds)
    - Validation RMSE ~4 minutes, R^2 value 0.8
    - Residuals centered about zero, slight heteroscedasticity observed
    - RMSE is high given median trip time of 10 minutes
- Classification (predict trip as being short, medium, or long)
    - Trip times binned using 4-quantiles (< 6 min, 6-16 min, > 16 min)
    - Average prediction accuracy ~80% across three categories
    - Highest misclassification on short trips in heavy traffic
- Accurately predicting trip time is difficult, mostly due to effects of traffic
    - A fixed trip distance can take 10x as long depending on location and time of day

Predicted trip length with horizontal lines drawn at categorical boundaries



We decided to use the classification approach for predicting trip times due to the high RMSE on the regression model, and because our particular binning of trip times produced favorable metrics in terms of overall prediction accuracy.