Predicting Cab Fare

Milestone Report

What does the data tell us?

There is a problem in determining how much a cab fare will cost in advance. Unknown parameters include the traffic, the route, and often the rate per mile and while sitting in traffic, in addition to any base fees. The total cost is something that many of us have wondered about at one time or another.

Google and Coursera have offered a 55 million row New York Taxi Fare dataset through a Kaggle competition that will end on September 25, 2018. It includes NaN values that I eliminated, latitudes and longitudes nowhere near New York, and unrealistic rides that required deeper analysis to uncover.

There are posted rules of New York Taxis that include a base rate of $2.50 plus $2.00 per additional mile. Using Euclidean Distance, all taxi rides should be underneath the line $y = 2x + 2.5$. Yet tens of thousands of rows were not. These unrealistic rides were eliminated from the dataset.

Furthermore, there were many rides that traveled not distance with a range of fares that tended rather right. These may be due to passengers going to multiple locations, or returning to where they started. In these cases, the starting and ending point is not appropriate to determine the fare. Such rows were also eliminated.

I created new columns based on the timestamp, and latitude-longitude determinations. I used an interactive latitude-longitude map to circumscribe Manhattan in a quadrilateral to determine if a given drop-off or pick-up occurred in Manhattan. The same can also be done for airport drop-offs and other target locations. (This was very time-consuming to apply to 50+ million rows.) In addition to Euclidean Distance, I also converted latitude-longitude into taxicab distance, a more realistic metric, for comparison.

I created many other time columns including 15-minute intervals, summer months, cold months, weekends, rush hour, night rush, night charge, and weekday surcharge; the latter two are additional fees posted on all taxis.

My initial findings are as follows:

1) The mean fare is $11.35
2) The median far is $8.50
3) The fares are right skewed.
4) The original dataset includes unrealistic rides of nearly a million dollars and negative 300.
5) Bus fares may have been included due to a high number of passengers and long distances traveled.
6) Latitude and longitude requires great care and further development

Corey Wade
August 29, 2018