Predicting Cab Fare

Final Report

The Advantage of Deep Learning

There is true value in predicting taxis fares in advance. When traveling from one destination to another, one has many options to choose from: Uber/Lyft, public transportation, personal vehicles, or taxis. In this day and age, taxis are at a disadvantage because the cost of the ride is not known in advance.

But the cost can be known in advance within reason. Since the mileage is known, and the base fares are known, all that is left is to predict traffic. Far from a trivial problem, there are many unpredictable variables when predicting traffic including traffic lights and other drivers. Yet estimates will only improve with more data, and live time series analysis.

Given a past New York Taxi Fare dataset, I was able to predict realistic fares within approximately 2.8 dollars. The models that I tuned included Random Forests, LightGBM, and Sequential Deep Learning. I scrubbed the dataset to eliminate unrealistic outliers including rides that traveled no distance, and latitude/longitudes well outside of the general New York area.

A root mean squared error of 2.8 suggests room for improvement. With deep learning, more data will make a substantial difference, and live time series analysis will help all models. The general idea of taxi services offering ride estimations within a couple dollars of accuracy should benefit the industry as a whole since potential customers can better make informed decisions. For instance, even though a taxi may be a couple of dollars more, it may be worth it to take one rather than wait five an extra minutes for an Uber/Lyft. It's up to the individual to decide how much their time is worth.

My recommendations are to expand this research project to include more metropolitan areas. New York is one of the most congested areas in the world, so estimations should improve in other areas. Live time series analysis can also help tremendously.

In addition, the root mean squared error of 2.8 dollars should not be taken as a final result, but as a baseline for improvement.

All previous reports, in addition to Python code, pipelines, graphs and data analysis may be found at https://github.com/coreyjwade/NYC_Cab_Fare.

Corey Wade
November 6, 2018