What is the problem I am attempting to solve?

New York City is arguably the taxi capital of America and home of the classic yellow taxicab. I am interested to work on yellow-taxi (medallion) demand prediction. They are the only vehicles that have the right to pick up street-hailing and prearranged passengers anywhere in New York City. Taxicabs are operated by private cab companies and licensed by the New York City Taxi and Limousine Commission (TLC). They lease it to their drivers who in turn get to keep 100% of the fares and tips (some companies charge less for the lease, but retain a portion of the fares). These fares are set by the TLC, and the amount that a cab company can lease the vehicle to the driver is also set.

Ride-sharing services such as Uber and Lyft have disrupted the taxi industry. Uber has become hugely popular in New York, and its trips outpaced yellow taxis for the first time last year. The biggest problem with taxis is they are not easy to find. Riders might have to wait longer when demand exceeds supply. There are Currently there are about 13,500 medallion cabs in NYC. They may not fulfill all the demand but it is possible to improve their utilization.

How is my solution valuable?

The model can make real-time demand predictions. It can be deployed inside the cars. Drivers could go to a location where there is higher demand in the next hour. As drivers get more revenue, they will be less likely to move to Uber/Lyft. Taxi cab companies could maintain their lease revenue for long period of time. If companies retain portion of the fares, they can work with the drivers to make a plan on how to deploy their taxis throughout the day to maximize profit.

What is my data source and how will I access it?

A recent dataset for Uber is not publicly available. NYC OpenData maintains yellow-taxi trip record data. It contains data from 2010 up to 2018. It is updated recently. The trip data for 2017 contains 113M rows, and 2018 contains 60M rows.

What techniques from the course do I anticipate using?

Time Series Analysis

I will experiment classic time-series forecasting methods. AR, MA, ARMA, ARIMA, SARIMA, VAR. I will deal with non-stationary data, seasonality, ACF, PACF, one-step forecast, multi-step forecasts, finding a model with the least AIC.

Supervised Learning

I can model a time-series problem as a supervised learning problem. Adding lagged variables, binary indicators. I will use popular machine learning techniques like Linear Regression, RandomForest, XGBoost.

Tensor Flow and Keras

Sequence Models: RNN, LSTM

What do I anticipate to be the biggest challenge I'll face?

I already started working with the data. It is a huge dataset. Loading everything to memory is not possible. It took me days to learn and use a new framework called Dask. It is relatively new framework and it doesn't have many references. My plan was to specialize Time-series analysis and Deep Learning. The unforeseen problems allowed me to learn another skill-set: Big Data Analysis. Still I am struggling to visualize the raw data. Hopefully, I can reduce the dataset by grouping, aggregate and resampling the data.

I know ARIMA and its variations. I noticed that ARIMA works fine for one step forecast but struggles to forecast multiple steps accurately. They can only be used for univariate time-series forecasting. I have to research about VAR and other multivariable time-series forecasting methods.

Thanks to Keras, LSTM is just 11 lines of code but preprocessing time-series data is tricky. I need GPU to accelerate model training and test. Getting a high-performance machine is a challenge. I don't want to wait 3 hours to train a model. It was happening when I was doing some of my previous projects. I managed to learn how to configure and use Google Cloud Platform. I revised my unsupervised learning capstone using google platform. I have already used half of the trial limit. Google Collab is another option.