Apply machine learning

Suresh Gopalakrishnan

February 18, 2018

1. How do you frame your main question as a machine learning problem? Is it a supervised or unsupervised problem? If it is supervised, is it a regression or a classification?

New Yorkers take thousands of Taxi ride every day. Yellow Taxi's are dominating the market though we have Uber and Lyft in the market. However ridesharing apps gaining popularity now days with convenient apps on mobile phones. It is increasingly important for taxi companies to provide visibility to their estimated ride duration, since the competing apps provide these metrics upfront. Machine Learning algorithms are answer to this problem. We have used supervised learning and since it's a time prediction, we are using regression techniques.

---

2. What are the main features (also called independent variables or predictors) that you'll use?

We have used 12 independent variables to predict "Trip Duration". Main file just provided Trip data and time along with pickup/drop off coordinates. Using the data, we have derived features like, weekday, hour of day, direct distance, distance and time calculated using OSRM and whether it is an Airport Trip. We have used these derived fields for our prediction.

---

3. Which machine learning technique will you use?

We have used Linear Regression, Random Forest and XG Boost algorithms for prediction. Among those, Random Forest algorithm performed well based on and Pseudo R2.

---

4. How will you evaluate the success of your machine learning technique? What metric will you use?

We have evalutated model performance by RMSLE and Pseudo R^2

$$PseudoR^2 = 1 - \frac{MSE}{Var(y)}$$