Big Data Analytics

Final Project: Taxi Fare Prediction in NYC

Dataset: TLC Trip Record Data

(http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml)

Submitted by

Md Shohidul Haque, Rabbil Bhuiyan, Md Zahidul Khan

Submitted to Leonardo A. Espinosa (PhD)

Introduction

New York City taxi rides paint a vibrant picture of life in the city. The millions of rides taken each month can provide insight into traffic patterns, road blockage, or large-scale events that attract many New Yorkers. With ride sharing apps gaining popularity, it is increasingly important for taxi companies to provide visibility to their estimated fare and ride duration, since the competing apps provide these metrics upfront. Predicting fare and duration of a ride can help passengers decide when is the optimal time to start their commute, or help drivers decide which of two potential rides will be more profitable.

Data access and download

For our final project, we wanted to work with a really big dataset – one whose size would make it unfeasible to analyze and visualize using any analytical program. Fortunately we have the powerful Python tools to analyze big data. However, downloading big data set e.g 114 GB is challenging task for execution. We downloaded tens of gigabytes of data from the New York City Taxi and Limousine Commission and set out to produce some prediction analyses of tax ride patterns. However, the local environment e.g personal laptop is totally outclassed to explore such big data on it. As a data scientist we need to test our skills on such big dataset in alternative computing environment. Here, CSC supercomputer turned out to be the solution.

We accessed multiple CSV files uploaded by NYC Taxi & Limousine Commission here: www.nyc.gov/html/tlc/html/about/trip_record_data.shtml. We applied 'wget command' to download the data as URL in puhti CSC environment and applied 'cat command' to combined several CSV files. Finally, we moved all the data from puhti to pouta CSC environment using allas module.

Research question

In this task, we are going to predict the fare amount for a taxi ride in New York City, given the pick up, drop off locations and the date time of the pick up.

Data preprocessing

The data used in this study are all subsets of New York City Taxi and Limousine Commission's trip data, which contains observations on around 1 billion taxi rides in New York City between 2013 and 2019 (114 GB). For the main analyses of this study, the data for yellow taxi rides during the month of January 2019 were used. Since each month consists of about 916272426

observations, and there were computational limitations, a random subset of 10,0000 observations from Januray 2019 were used for model building. The original dataset contains features as pickup and dropoff locations, as longitude and latitude coordinates, time and date of pickup and dropoff, ride fare, tip amount, payment type, trip distance and passenger count as well as others.

In order to predict the estimated fare amount, we utilized few features to determine the expected fare amount. In the data preprocessing we observed the following features to correlate significantly in determining taxi fare. These were pickup time or date, drop off time or date, passenger count, fare amount (see correlation graph in Notebook). We have stick to these features for the simplicity of our analysis. We have also filtered these variables to remove the possible outliers (see Visualization in the notebook). We have also checked the missing values for these variables as data preprocessing.

We also did some feature engineering task such as extracted separate features for year, month, day, weekday, hour and minute from the date and time of each ride, as well as trip duration as the difference between drop off and pickup time. As taxi fares might change during different hours of the day and on weekdays/weekends/holidays as well. Although, it is logical that the longer the trip (the distance from pick up to drop off), the higher the price, still the different hours of the day and different days of the week will affect the taxi fare. Thus, we considered those variables into our analysis. The data were explored and analyzed using Jupyter notebook.

Machine learning modelling and model evaluation

Once we preprocessed our data by removing any outliers and featured engineering for datetime variable, we implemented our data into a Machine Learning Model. First we used linear regression model as baseline model. Then we applied regularization techniques in order to solve the overfitting problem. We used sklearn's Lasso Regression regularization in this case.

We used RMSE, which stands for root mean square error, and R^2 to identify which model preformed the best. Root Mean Square Error (RMSE) gives the standard deviation of the difference between actual fare amount and predicted fare amount in $ and can be calculated as the square root of sklearn's mean_squared_error(). The R-square gives how much percentage of the actual fare amount*'s* variation is predicted by the model and can be directly obtained from sklearn's r2_score().

Furthermore, to improve the model RMSE we applied hyper-parameter tuning. For example we tuned alpha parameter for Lasso regression assuming that smaller alpha will make Lasso regression similar to linear regression results. We applied both sklearn's

generic GridSearch CV and sklearn's own LassoCV for tuning. The former one is mostly used for tuning any machine learning model while the later one is specific to just lasso regression. We used both and compared the best alpha. We have also applied more complex model e.g the random forest regression model in order the compare the accuracy. In our model analysis we found LassoCV model to give better accuracy compared to other models. Finally, we have plotted the actual fare vs the predicted fare from the LassoCV model to visually check the correlation (see Jupiter notebook).

Lastly, we applied our full dataset in puhti.csc.fi environment in order to explore our skills on machine learning using big data set. That was eventually our aim to learn the knowledge and skills needed to work on supercomputer environment, in our case Puhti super computer and analyze big data on that environment. After applying the same model with full dataset at puhti we found the model accuracy pretty good (0.92) (see the Jupiter notebook) with as lower as alpha hyper parameter (0.003) and we were successful with big data analysis in supercomputer environment.

Conclusions

Considering the input features used in this study, their predicting results were fairly accurate for all the models applied. However, the LassoCV method, with best alpha of 0.003, was the best model as it produced the lowest RMSE score (3.7) and highest R-squared score (0.92), which explains the highest variability and tells us how well the model fits in this data. The similar result was also observed while we run our full dataset at Puhti -with highest R-squared score (0.92) for LassoCV. As a further scope, we can apply our dataset to more complex model e.g gradient boost model or even neural network model. Additionally, more variability in the dataset can be applied in order to improve the prediction accuracy. For example different zones of New York City, drivers speed or average speed in hour etc could affect the taxi fare.

References

Panda G. and Panda S.P. (2019) Machine learning using exploratory analysis to predict taxi fare. International Journal for Research in Applied Science and Engineering Technology (IJRASET).ISSN:2321-9653; Volume 7 Issue VIII (www.ijraset.com)

Taxi and Limousine Commission of New York City (2020) TLC trip record data. Retrieved from http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

Stephen M.L. (2019) New York Taxi data set analysis-predicting taxi fare using regression models. Retrieved from https://towardsdatascience.com/new-york-taxi-data-set-analysis-7f3a9ad84850