New York City Taxi Trip Duration

Qin Zhang[1]

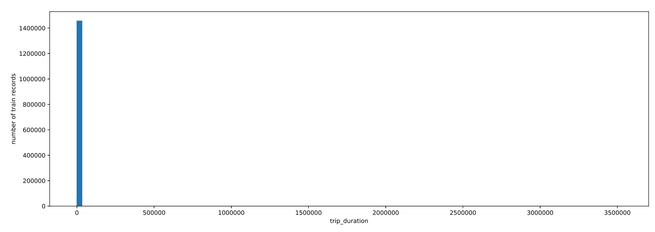
[1] Chongqing Technology and Business University, China

Introduction

The competition dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform. The data was originally published by the *NYC Taxi* and Limousine Commission (TLC). The data was sampled and cleaned for the purposes of this playground competition. Based on individual trip attributes, participants should predict the duration of each trip in the test set. The project provided 1458644 pieces of effective data for the construction of machine learning model. The purpose of the project is to predict the trip-duration time of passengers through "characteristics" such as pickup-datetime and dropoff-datetime.

Data Description

- The data set has a total of *1458644 lines and 10 features*, which is a regression item.
- *10 features*: ID, Vendor-id, Pickup-datetime, Dropoff-datetime, Passenger-count, Pickup-longitude, Pickup-latitude, Dropoff-longitude, Dropoff-latitude, Store-and-fwd-flag.
- *Trip-duration*: Duration of the trip in seconds.



Feature Engineering

Outlier Processing

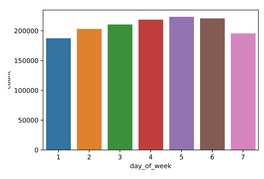
- According to the relationship between longitude and latitude position and travel duration, abnormal data can be filtered out. Therefore, according to the analysis, the interference data within 1 million will be deleted. Selecting appropriate time data will help to train the model and improve the experimental accuracy.
- *Passenger-count* is the number of passengers in the vehicle. According to the analysis of Fig. 2, the number of passengers varies from 0 to 9. Therefore, we need to delete the data with 0 passengers.

Feature Processing

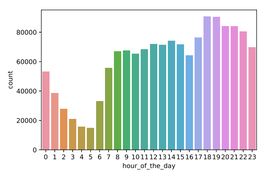
- Store-and-fwd-flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server. Y=store and forward, N=not a store and forward trip. We convert its character identification into numbers for easy analysis.
- Add month, week and day. From the perspective of the trend, the overall taxi time has been increasing from January to June 2016. It may be that users are gradually accustomed to taking a taxi from a longer distance, or it may be that more and more vehicles are driving on the road or the weather is bad, causing traffic congestion. In this article, the original pickup-datetime is divided into hour, day of week and month.
- Create a distance feature. According to the longitude and latitude in the known data, the distance between the starting position and the ending position of the taxi is calculated according to the formula.

Feature Visualization

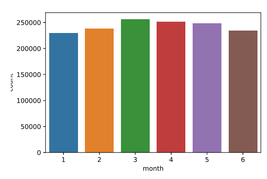
- Deleting invalid features can improve the experimental effect. The following is a visualization of three randomly selected features:



day_of_week



$hour_of_the_day$

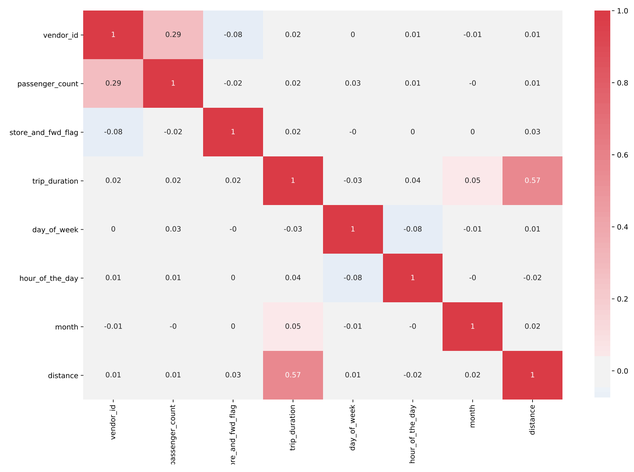


month

Feature Engineering

Feature filtering

- Through the analysis of feature related data and heat map, the relationship between features can be obtained. It can be seen from the figure that the selected features are important.



Experiment

Model Introduction

- Three models were selected in this experiment: LGBMRegressor, LinearRegression and DecisionTreeRegressor.
- The division ratio of training set and test set is set to 0.3.

Evaluation Indicator

- $MSE = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2$
- $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - f(x_i)|$
- $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2}$

Experimental Result

Model	LGBMRegressor	LinearRegression	DecisionTreeRegressor
R2_score	0.66	0.35	0.3
MSE	0.22	0.42	0.45
MAE	0.32	0.46	0.46
RMSE	0.47	0.65	0.67

It can be clearly seen from the experimental results that there are obvious differences in the evaluation effects of the three baseline models on the four evaluation indexes. Generally speaking, the evaluation value of LGBMRegressor model is better than other baseline models, and better evaluation results can be obtained.

Conclusion

- The data of the simulation experiment of this project is from the real data of New York City Taxi Trip Duration, which is conducive to verifying the effectiveness of the model. In this project, the effects of 3 baseline models on 4 evaluation indexes are compared. Finally, it is found that the prediction effect of LGBMRegressor model is the best as a whole.
- The analysis of this experiment can not only accurately predict th… of taxis, but also mine more detailed information abou… For example, it is possible to analyze which … more prone to orders, and where pe… is an effective data for rental schedu… value brought by the blizzard that t… quantity.

Acknowledgement

- International Cooperation Project (Y7Z0511101) of IIE, Chinese Academy of Sciences