Final Project Report

New York Taxi Fare Prediction

Yeshen Liu: 7310174348

TaiHsuan Wu: 2524355409

Wei Suo: 1588018218

Anjie Chen: 2857438520

ISE 529 Wednesday Section

2019 December 09

Project Description

In this competition, we are tasked with predicting the fare amount for a taxi ride in New York City given the train data set with 5.7 gigabyte. Having basic feature of pickup and drop off locations (longitude and latitude), pickup datetime, and passenger count. The result is scoring by RMSE.

Steps to Complete the Project

Step 1: Data Preprocessing

After using describe function, we found out there are some illogical data and we have to preprocess them.

a. Since the dataset is really large, we decided to take 500 millions datasets to make train.
b. The minimum value of fare amount should not be less than 0, and after google the initial charge taxi fare is $2.5. Therefore, remove the fare amount below $2.5.
c. The minimum value of passenger count should not have 0.
d. After compare with test dataset, New York city longitude is around -74.5 to -72.8 and with latitude around 40.5 to 41.8.
e. We check that there is no nah value.

Step 2: Explore New Features

From the experience of taking the taxi, we have known that the fare is affected by distance, weather, hot spot and the amount of requirement. In order to make data more precisely, we want to explore new features by the given features.

a. Changing pickup date time to year, month, day, weekday and hour.
b. In order to get the distances between two spots, we can use haversine formula to compute the distance by using latitude and longitude.
c. Usually, the hot spot like airport will have higher fare or fixed price. In this case, we pick up three spots: Newark Liberty International Airport, JFK Airport and LaGuardia Airport. Therefore, we decided to compute the distance coordinates to the three spots above, to see if the pickup and dropoff is near those spots.

After adding new features, we have 14 features now.

Step 3: Tune and do model Training

Splitting train dataset to train and test by 30%.

A. Decision Tree Regressor: We utilize grid search to find out the best max_depth parameter which is 8 for decision tree regressor.
B. Random Forest Regressor: Doing same grid search,max_depth which is # for random forest regressor, and set n_estimators to 100.
C. Kneighbors Regressor: In this case, we try to find out best neighbors parameter, #, for Kneighbors regressor.
D. XGB Regressor: Finding best learning rate, #, and then set n_estimators as 200.
E. Light GBM: Setting their max_depth to 3, learnig_rate to 0.75, and use regression application.

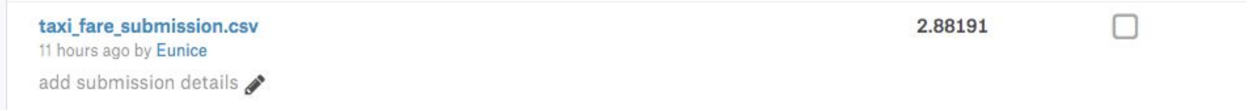
In order to find out the best model to predict the taxi fare, we compare the RMSE between those models, KNN: 5.023; Random Forest: 4.311; XGB: 4.083; Decision Tree: 4.4014; Light GBM: 3.622.

F. Neural Network: Using sequential from keras model, set kernel initializer as “normal” and activation based on “relu”. About the compile part, since the result is scored by RMSE, set the loss by mean squared error, decide the optimizer as “adam” and select the metrics based on rmse. During the training process, we use ModelCheckpoint to save the good weights of model.
   a. First try: Setting two layers with Dense 12 and 1. Aftering using kFold and compute using cross validation score the average rmse is about 5.151.
   b. Second try: Standardize the features by using StandardScaler and Pipeline. In this case, we found out the average rmse comes to 4.1.
   c. Third try: Try to add layers of model, we add the dense 12, 6 and 1 and also do the standardize. In this case, we get the average rmse as 4.01.
   d. Forth try: Try to expand the width of model, we set the dense as 20 and 1. In this case, we get the average rmse as 3.2.

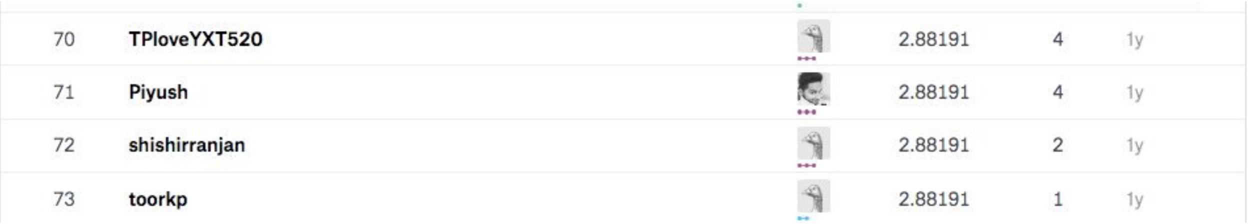
   Since I have saved the best weights of each model, we let the model learn again by the best weights. And after comparing, we decide to take the final model which gave us the best result.

Step 4: Clean the Test data and make the Prediction

We did the same process of train data to clean the test data, and make the prediction by using NN model. We get the result of score 2.88, which is on the rank about 70 of 1483 teams.



taxi_fare_submission.csv 2.88191 □
11 hours ago by Eunice
add submission details



70	TPloveYXT520	2.88191	4	1y
71	Piyush	2.88191	4	1y
72	shishirranjan	2.88191	2	1y
73	toorkp	2.88191	1	1y

Reference

1. https://www.kaggle.com/breemen/nyc-taxi-fare-data-exploration
2. https://medium.com/analytics-vidhya/machine-learning-to-predict-taxi-fare-part-one-exploratory-analysis-6b7e6b1fbc78