Numerical Lab Taxi Trip Duration Prediction

Plagiarism will result in 0 in your project and all Lab Tasks.

Introduction

In this competition, We are challenging you to build a Approximation theory based model or that predicts the total ride duration of taxi trips in New York City.
This is a online competition arranged among all the **students of FAST NU** currently enrolled in Numerical course. This competition is live and can only be accessed with link below.

https://www.kaggle.com/t/1f6283d301b344f5bb2a39649ee4cc1b

Tutorial

Tutorial about registering and working on this competition and submitting your work will be given in Lab so kindly make sure you are present in your next Numerical Lab.

Queries

In case of any query about this competition/project feel free to contact on **i150077@nu.edu.pk**

Dataset

Data is splitted in training and testing files. Each file has 30000 samples.

Data Fields

1. **id** - a unique identifier for each trip
2. **vendor_id** - a code indicating the provider associated with the trip record
3. **pickup_datetime** - date and time when the meter was engaged
4. **passenger_count** - the number of passengers in the vehicle (driver entered value)
5. **pickup_longitude** - the longitude where the meter was engaged
6. **pickup_latitude** - the latitude where the meter was engaged
7. **dropoff_longitude** - the longitude where the meter was disengaged
8. **dropoff_latitude** - the latitude where the meter was disengaged
9. **store_and_fwd_flag** - This 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
10. **trip_duration** - duration of the trip in seconds (**Labels / To Predict**)

Tip

(As we studied Python in Numerical Lab as our basic language. So it is highly recommended to use python in project. Reason for recommendation is because you are working on large dataset. It will be difficult to manage such data in other languages and

perform your numerical analysis on this data. Guide / Sample code is also given in Python.)

Procedure

Reading Data

Data is given in CSV (Comma Splitted). To read each file use following code..

Training_data = pd.read_csv(path_of_your_training_file)

Processing Data

You will need to separate labels / true values from above variable.

True_values = Training_data.pop("trip_duration")

As there are some columns that are not important if you are working with Numerical Theorems. Such columns are id, pickup_datetime and dropoff_datetime.

So can drop these columns from your data by using following code.

Training_data.pop("id")

Training_data.pop("pickup_datetime")

Training_data.pop("dropoff_datetime")

As your data has a feature with name of store_and_fwd_lag. This feature does not have integer values. It contains strings and have only 2 unique values ('N', 'Y').

So can replace 'N' by 0 and 'Y' by 1.

Training_data["store_and_fwd_lag"][Training_data["store_and_fwd_lag"] == 'N'] = 0

Training_data["store_and_fwd_lag"][Training_data["store_and_fwd_lag"] == 'Y'] = 1

Training_data["store_and_fwd_lag"] = Training_data["store_and_fwd_lag"].astype('int8')

Numerical Model

Now you can use Approximation theory to predict Taxi Duration.

Where **x1, x2, x3, x4 ….. xn** are data fields discussed above like vendor_id, passenger_count etc.

Predictions

Now you can read test data same as training data. And simple remove id, pickup_datetime and dropoff_datetime if you removed in Training data and simply make predictions.

Submissions

Submissions are also live and online and your score will be evaluated as you make submission and you will be ranked accordingly. If you have minimum loss then you will be on top and will be on position 1. You have to make submission of your final code on Google Classroom as well so that we can verify If you completed it on your own.

HAPPY CODING (BONNE CODAGE) :)

Feel free to ask questions. :)

All further details will be discussed in Numerical Lab.