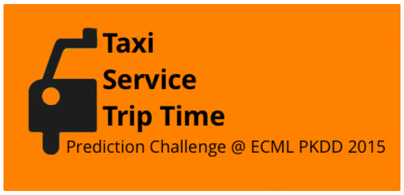
kaggle Search kaggle Competitions Datasets Kernels Discussion Jobs



ECML/PKDD 15: Taxi Trip Time Prediction (II)

Predict the total travel time of taxi trips based on their initial partial trajectories

$250 · 345 teams · 2 years ago

Overview Data Kernels Discussion Leaderboard Rules Late Submission

✓ You have accepted the rules for this competition.

Competition Data

- evaluation_script.r
- metaData_taxistandsl...
- sampleSubmission.csv...
- test.csv.zip
- train.csv.zip

train.csv.zip 508.89 MB

Download

Data Description

I. Training Dataset

We have provided an accurate dataset describing a complete year (from 01/07/2013 to 30/06/2014) of the trajectories for all the 442 taxis running in the city of Porto, in Portugal (i.e. one CSV file named "train.csv"). These taxis operate through a taxi dispatch central, using mobile data terminals installed in the vehicles. We categorize each ride into three categories: A) taxi central based, B) stand-based or C) non-taxi central based. For the first, we provide an anonymized id, when such information is available from the telephone call. The last two categories refer to services that were demanded directly to the taxi drivers on a B) taxi stand or on a C) random street.

Each data sample corresponds to one completed trip. It contains a total of
9 (nine) features, described as follows:

1. **TRIP_ID**: (String) It contains an unique identifier for each trip;
2. **CALL_TYPE**: (char) It identifies the way used to demand this service. It may contain one of three possible values:

   a. ‘A’ if this trip was dispatched from the central;
   b. ‘B’ if this trip was demanded directly to a taxi driver on a specific stand;
   c. ‘C’ otherwise (i.e. a trip demanded on a random street).

3. **ORIGIN_CALL**: (integer) It contains an unique identifier for each phone number which was used to demand, at least, one service. It identifies the trip’s customer if CALL_TYPE=’A’. Otherwise, it assumes a NULL value;
4. **ORIGIN_STAND**: (integer): It contains an unique identifier for the taxi stand. It identifies the starting point of the trip if CALL_TYPE=’B’. Otherwise, it assumes a NULL value;
5. **TAXI_ID**: (integer): It contains an unique identifier for the taxi driver that performed each trip;
6. **TIMESTAMP**: (integer) Unix Timestamp (in seconds). It identifies the trip’s start;
7. **DAYTYPE**: (char) It identifies the daytype of the trip’s start. It assumes one of three possible values:
   a. ‘B’ if this trip started on a holiday or any other special day (i.e. extending holidays, floating holidays, etc.);
   b. ‘C’ if the trip started on a day before a type-B day;
   c. ‘A’ otherwise (i.e. a normal day, workday or weekend).
8. **MISSING_DATA**: (Boolean) It is FALSE when the GPS data stream is complete and TRUE whenever one (or more) locations are missing
9. **POLYLINE**: (String): It contains a list of GPS coordinates (i.e. WGS84 format) mapped as a string. The beginning and the end of the string are identified with brackets (i.e. [and], respectively). Each pair of coordinates is also identified by the same brackets as [LONGITUDE, LATITUDE]. This list contains one pair of coordinates for each 15 seconds of trip. The last list item corresponds to the trip’s destination while the first one represents its start;

The total travel time of the trip (the prediction target of this competition) is defined as the (number of points-1) x 15 seconds. For example, a trip with 101 data points in POLYLINE has a length of (101-1) * 15 = 1500 seconds. Some trips have missing data points in POLYLINE, indicated by MISSING_DATA column, and it is part of the challenge how you utilize this knowledge.

II. Testing

Five test sets will be available to evaluate your predictive framework (in one single CSV file named "test.csv"). Each one of these datasets refer to trips that occurred between 01/07/2014 and 31/12/2014. Each one of these data sets will provide a snapshot of the current network status on a given timestamp. It will provide partial trajectories for each one of the on-going trips during that specific moment.

The five snapshots included on the test set refer to the following timestamps:

14/08/2014 18:00:00
30/09/2014 08:30:00
06/10/2014 17:45:00
01/11/2014 04:00:00
21/12/2014 14:30:00

III. Sample Submission Files

File sampleSubmission.csv uses the average travel time of all trips in the training set.

IV. Other Files

Along with these two files, we have also provided two additional files. One contains meta data regarding the taxi stands metaData_taxistandsID_name_GPSlocation.csv including id and location.

The second one includes an evaluation script for both problems developed in the R language ("evaluation_script.r").

Our Team Terms Privacy Contact/Support