NYC Taxi Trip Duration Prediction

Yifan Yang
Kon Woo Kim
Outline

1. **Problem Description.**
2. Features Extraction
3. Models
4. Future works

Reference
Problem Description

Given a set of taxi trip attributes, such as

- pickup location;
- drop off location;
- date time and etc.

our goal is to predict the duration of the trip.

Evaluation Metric:

Root Mean Squared Logarithmic Error (RMSLE):

$$\epsilon = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( \log(p_i + 1) - \log(a_i + 1) \right)^2}$$
Outline

1. Problem Description.

2. **Features Extraction**

3. Models

4. Future works

Reference
Feature Extraction

To extract features for each trip, we use following datasets:

- Dataset provided by the competition;
- NYC Taxi and Limousine Commission (TLC) Dataset;
- Open Source Routing Machine (OSRM) Data;
- 2016 NYC Weather Dataset;
- Holiday Data;
- 2016 Extreme Weather Event Data;
Feature Extraction

**Date time** features:
- Weekday
- Hour
- Minute

**Location** features:
- L2 distance between pickup location and drop-off location;
- Direction from pickup location to drop-off location;
- Mid-point between pickup location and drop-off location. (~0.005 improvement);
- Manhattan distance computed with manhattan coordinates.

**Location clustering** features:
- Use K-means to do clustering on all the pickup and drop-off location.
Feature Extraction

Coordinate Transformation (beneficial to tree-based method generating splits):

- PCA (sensitive to outliers however.)
- Manhattan coordinates: we use the 5th Ave. as the Y axis in the reference frame.
Feature Extraction

**Weather** features (From Weather dataset and 2016 Extreme Weather Event Data):

- Average temperature
- Precipitation
- Extreme weather (Boolean)

**Holiday** features (`import` holidays):

- 2016 US holiday (Boolean)

**OSRM** features (From OSRM dataset):

- Number of steps: Number of route changing (turning).
- Estimated Travel Time: shortest travel time from pickup location to drop-off location by considering speed limitation and trip distance.
Feature Extraction

TLC features (From TLC dataset):

- Trip distance: actual trip distance for each data sample;

- Fare Amount: Money cost for each trip, has strong correlation to time duration

NYC Taxi Standard Metered Fare:
50 cents per 1/5 mile when traveling above 12mph or
per 60 seconds in slow traffic or when the vehicle is stopped.
Outline

1. Problem Description.

2. Features Extraction

3. Models

4. Future works

Reference
Model

Regression model: XGBoost

Regression Target: \( \bar{a}_i = \log(a_i + 1) \)

Objective function: mean-square error (MSE).

Prediction: \( p_i = e^{\bar{p}_i} - 1 \)

Validation set:

- 10% training data,
- roughly the same distribution with test set in terms of mean and variance
- learning the weights for each model in the ensemble.
Model Ensemble

We have trained different models on the data with different distributions.

Specifically, for each model, we sampled training data from the training set with certain criterion.

Distance Division (long / medium / short):

- Trip samples with long distance dist > 10km;
- Trip samples with medium distance 3km <= dist <= 10km;
- Trip samples with short distance dist < 3km;

We fit a XGBoost for each case.

Note: Inside each division, the partition is not strict and there are some shared samples.
Model Ensemble

Trip Duration (Estimated) Division (long / short):

- Trip samples with long duration time > 600s;
- Trip samples with short duration time <= 600s;

We fit a XGBoost for each case.

Time Division (midnight / midday / evening):

- midnight: Trip samples with pickup hour from 23 pm to 7 am;
- midday: Trip samples with pickup hour from 7 am to 18 pm;
- evening: Trip samples with pickup hour from 18 pm to 0 am;

We fit a XGBoost for each case.
Model Ensemble

**Workday / Holiday Division** (work / dayoff):

- **Workday**: Trip samples happened during Mon. to Fri.;
- **Holiday**: Trip samples happened during Sat. to Sun. and traditional holiday;

We fit a XGBoost for each case.

**Trip Type Division** (H2H, H2L, L2H, L2L):

- **H2H**: Trips with both pickup and dropoff at high density region.
- **H2L**: Trips with pickup at high dens. region and dropoff at low dens. region.
- **L2H**: Trips with pickup at low dens. region and dropoff at high dens. region.
- **L2L**: Trips with both pickup and dropoff at low density region.
Model Ensemble

We also specifically tuned a ‘base’ model with all the training samples.

The model hyperparameters are tuned with 5-fold cross validation (fixed lr. and iters.)

Prediction with ensemble:

1. Weighted Average (Linear regression trained on validation set)

2. Average

<table>
<thead>
<tr>
<th>Methods</th>
<th>Private Score</th>
<th>Public Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble (average)</td>
<td>0.22397</td>
<td>0.23049</td>
</tr>
<tr>
<td>Base model</td>
<td>0.22472</td>
<td>0.23173</td>
</tr>
</tbody>
</table>
Outline

1. Problem Description.

2. Features Extraction

3. Models

4. Future works

Reference
Future works

1. Explore how to config the ensemble models.

2. Investigate how to combine the predictions effectively.

3. We can do the representation learning on the traffic data by (geometric) deep learning to learn some meaningful features, such as spatial-temporal features for the traffic so to give some information about the whether or not the traffic is busy at certain time.
Reference

1. TLC dataset: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
2. Taxi fare: https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page
3. Holiday data: https://pypi.org/project/holidays/