New York City Taxi Trip Duration

Tianjiao Wang

Beijing Technology and Business University

Introduction

The project will build a model that predicts the total ride duration of taxi trips in New York City. The primary dataset is one released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables. Accordingly, this project problem is *taxi trips duration*, which is a *outlier detection*.

Data loading and overview aims to Loading the data and take a overview of the data.

Data cleaning aims to identify duplicated and missing values, and deal with outliers.

Features engineering aims to visualize the distribution of trip-duration values, deal with categorical features, deal with dates, create distance and speed, do correlations and dimensionality reductions.

This project can predict the duration of each trip in the *test set* , after *model selecting*, and *Hyperparameters tuning* .

Data loading and overview

- At first, I *quickly look at the first 5 lines of a dataset to understand the structure, format, and content of the data* . Then I take a overview of *the type and amount and other information of df and test data.*
- *Group Outlying Aspects Mining*, *Outlying Aspects Mining* and *Outlier Detection* are different with each other.

Colonne	Description
id	a unique identifier for each trip
vendor_id	a code indicating the provider associated with the trip record
pickup_datetime	date and time when the meter was engaged
dropoff_datetime	date and time when the meter was disengaged
passenger_count	the number of passengers in the vehicle (driver entered value)
pickup_longitude	the longitude where the meter was engaged
pickup_latitude	the latitude where the meter was engaged
dropoff_longitude	the longitude where the meter was disengaged
dropoff_latitude	the latitude where the meter was disengaged
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)
trip_duration	duration of the trip in seconds

overview of data

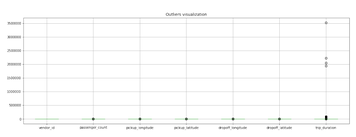
Data cleaning

I do the *data learning* to check the *Duplicated and missing values* and *Deal with outliers*

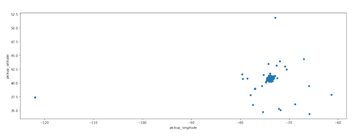
```
id                  0
vendor_id           0
pickup_datetime     0
dropoff_datetime    0
passenger_count     0
pickup_longitude    0
pickup_latitude     0
dropoff_longitude   0
dropoff_latitude    0
store_and_fwd_flag  0
trip_duration       0
dtype: int64
```

There are no duplicated or missing values.

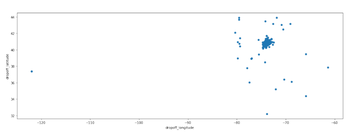
Visualize outliers There are outliers. I can't find a proper interpretation and it will probably damage our model, so I choose to get rid of them.



outliers for trip-duration



outliers for pickup positions



outliers for dropoff positions

Outlying Degree Scoring *In this step, I only keep trips that lasted less than 5900 seconds, and only keep trips with passengers.*

Features engineering

Visualize the distribution of trip-duration value,

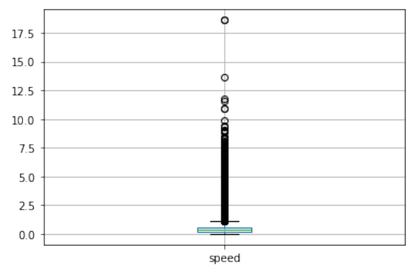
The distribution of trip-duration values

The distribution is right-skewed so we can consider a log-transformation of trip-duration column.

Deal with categorical features

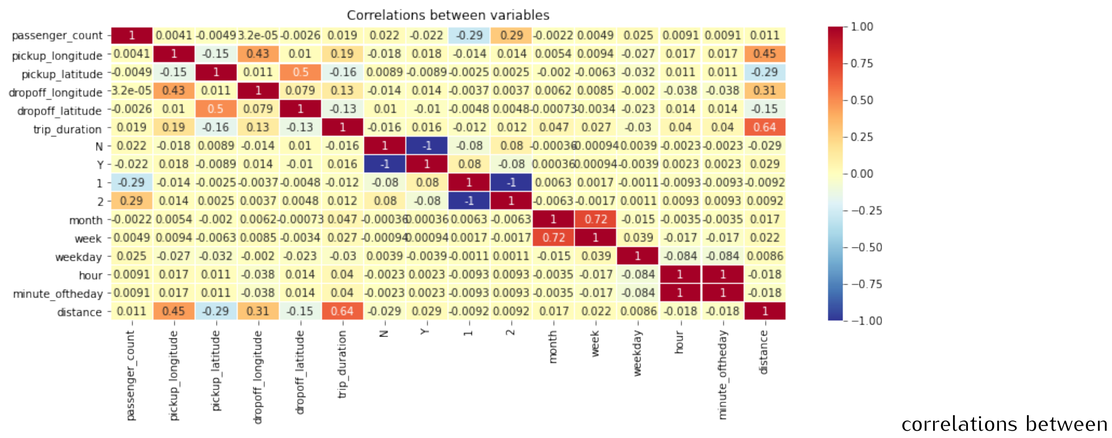
Deal with dates

Distance and speed creations



speed

Correlations and dimensionality reductions



correlations between variables

Model selection

LightGBM is blazingly fast compared to RandomForest and classic Gradient-Boosting, while fitting better. It is our clear winner.

Our LightGBM model is stable.

Training and predictions

Training Training on all labeled data using the best parameters

Prediction Make predictions on test data frame

	id	trip_duration
0	id3004672	716.070826
1	id3505355	672.125770
2	id1217141	455.368356
3	id2150126	938.637832
4	id1598245	354.432595

Acknowledgement

- Flip00 learning-kaggle project