FLIP 00 Presentation

Jiaqi Liu

Jilin University

2021-04-25

Overview

Problem

Description and Evaluation

Description	Predict taxi trip fares according to attributes of time and postion.
Evaluation:	Root mean squared error(RMSE) $$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Data Process

Basic Information of Data

Attribute	Meaning
fare_amount	Cost of trip and meanwhile the predition target
pickup_datetime	The specific time when the driver picks the passenger
pickup_longitude	The longitude where the driver picks up the passenger
dropoff_longitude	The longitude where the driver drops off the passenger
pickup_latitude	The latitude where the driver picks up the passenger
dropoff_latitude	The latitude where the driver drops off the passenger
passenger_count	Number of passengers

- Over 55M lines in train set.

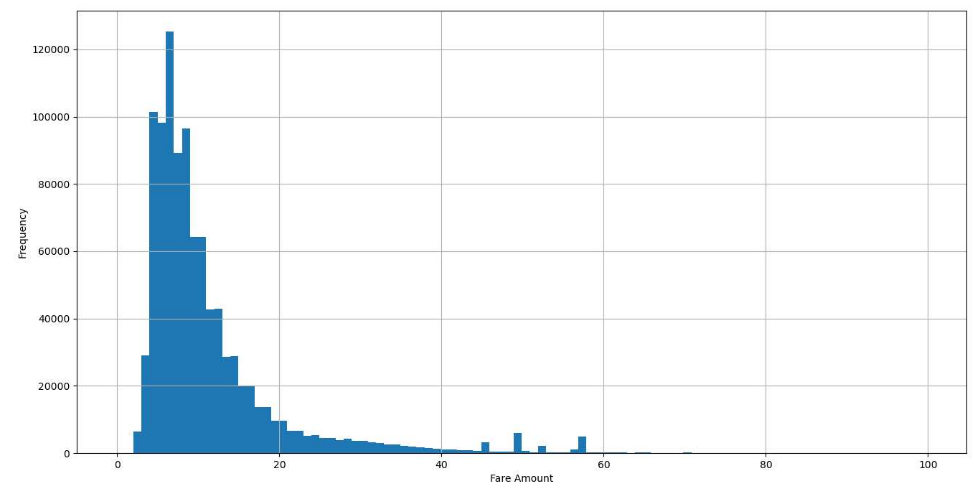
Missing Values


```
Out[10]:
key                   0
fare_amount           0
pickup_datetime       0
pickup_longitude      0
pickup_latitude       0
dropoff_longitude    10
dropoff_latitude     10
passenger_count       0
dtype: int64
```

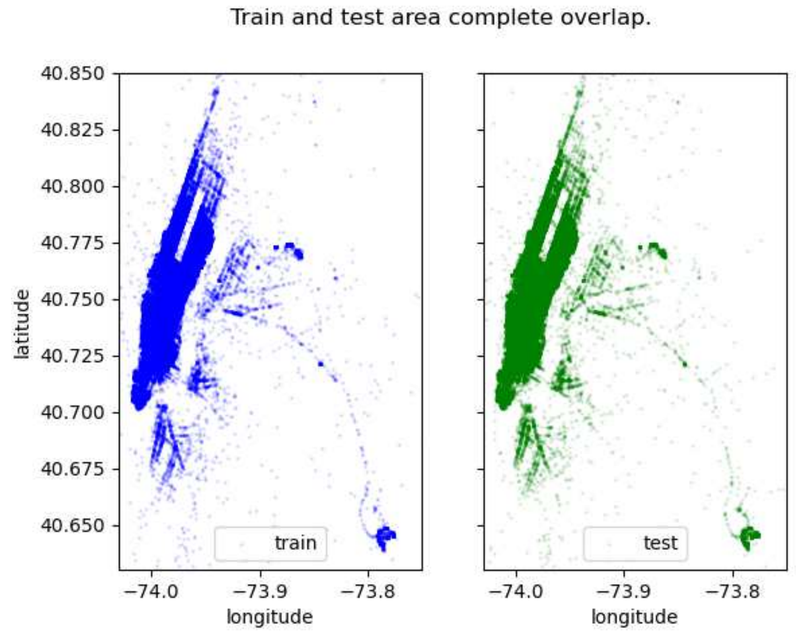
Figure 1: missing values

Outlying Numbers

Abnormal Numbers	■ Minus fare_amount values ■ Outlying passenger_count
Position Restriction	■ Restrict longitude into [-75,-72] ■ Restrict latitude into [40,43]



(a) fare amount distribution



(b) pickup position map

Distance and Fare

- Calculate Haversine distance according to pickup and dropoff positions.

- Restrict distance values into (0,200] by equation: $distance = (fare - 2.5)/1.56$
- Adjust fare_amount of zero values by equation: $fare = 2.5 + 1.56 * distance$

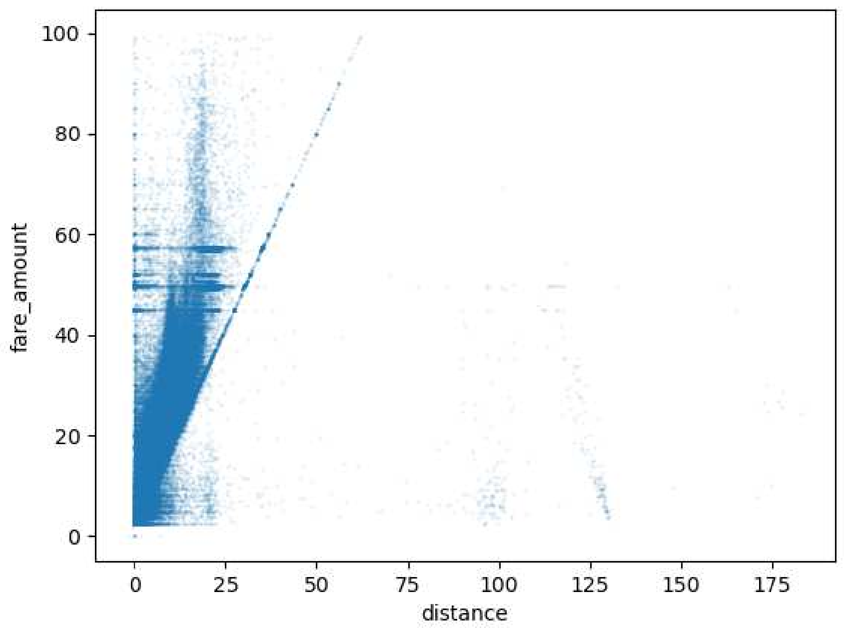


Figure 2: fare-distance scatter after the process

Datetime Process

Break datetime into

- year
- month
- weekday
- hour

Feature Selection

Feature Correlations

The correlation between fare_amount and other features.(Sorted according to absolute value)

Feature	Correlation
distance	0.838918
pickup_longitude	0.378179
dropoff_longitude	0.291588
pickup_latitude	-0.193441
dropoff_latitude	-0.171066
year	0.118953
month	0.026073
hour	-0.019402
passenger_count	0.016048
weekday	0.003206
day	0.001230

Feature Selection

So it is proper to drop weekday and day and then use other features to train the model. But according to experiment, weekday is a better feature than month, which brings better score. This may result from different distribution over years and needs further discussion.

Modeling and Predicting

Model

- Random forest model
- The random forest is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample.
- Parameters:
  - Number of trees:n_estimator = 100
  - Node size:depth = 30
  - Number of features sampled:See Feature Selection

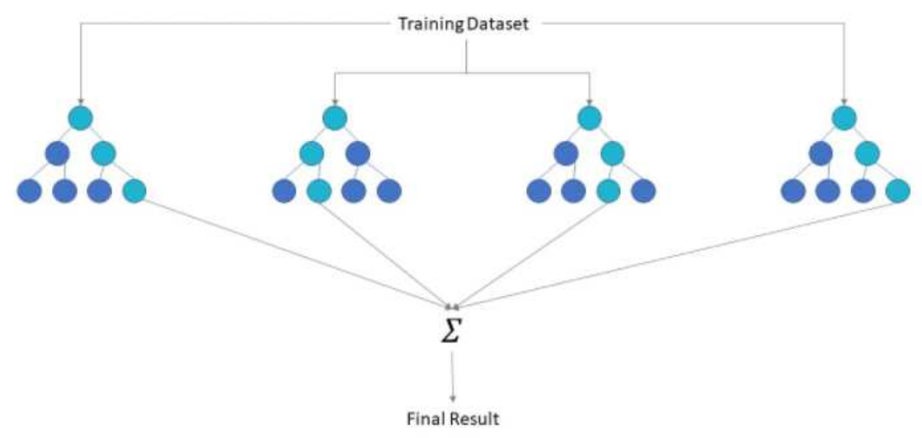


Figure 3: random forest

Feature Engineering

Feature	Importance
distance	0.791355
dropoff_longitude	0.059132
pickup_longitude	0.037138
dropoff_latitude	0.035333
pickup_latitude	0.026133
year	0.025464
hour	0.013933
weekday	0.007623
passenger_count	0.003890

Prediction Result

- Score:3.23791
- Rank:474/1483