New York City Taxi Fare Prediction

Runsha Pan

Hunan University

(None)

Overview

Problem Description

- Overview
- Dataset Description

Exploration and Preprocessing

- Exploratory data analysis
- Data Visualization
- Processing and Computing

Data Modeling

- Data Modeling
- Conclusion

Problem Description

Overview

Problem

- Predicting the taxi fare according to the known features of the passengers taking taxis, which including longitude coordinate and latitude coordinate of where the taxi ride started and so on..
- Do better than this using Machine Learning techniques: an RMSE of 5–8

Evaluation

RMSE measures the difference between the predictions of a model, and the corresponding ground truth.

- RMSE is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Dataset Description

- Files: train.csv; test.csv; sample_submission.csv

- Data fields
  - ID: key - Comprised of pickup_datetime plus a unique integer.
  - Features
    - pickup_datetime - timestamp value indicating when the taxi ride started.
    - pickup_longitude - float for longitude coordinate of where the taxi ride started.
    - pickup_latitude - float for latitude coordinate of where the taxi ride started.
    - dropoff_longitude - float for longitude coordinate of where the taxi ride ended.
    - dropoff_latitude - float for latitude coordinate of where the taxi ride ended.
    - passenger_count - integer indicating the number of passengers in the taxi ride.
  - Target: fare_amount - float dollar amount of the cost of the taxi ride.

Problem Description

Exploration and Preprocessing

Exploratory data analysis

Data Visualization

Processing and Computing

Data Modeling

Exploration and Preprocessing

Exploratory data analysis

Problem Description

Exploration and Preprocessing

Exploratory data analysis

Data Visualization

Processing and Computing

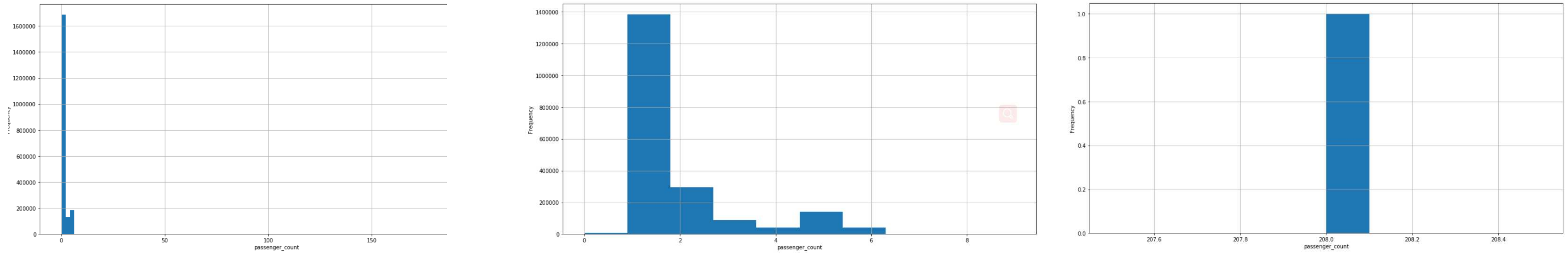
Data Modeling

- Data import and preliminary processing.
  - Import the data and view it.
  - Calculating eigenvalues.minimum value, etc.
  - Preliminary processing of date types and sorting.
- Handling missing values.
  - Deleting 14 rows with missing values, resulting in 1999986 remaining rows.
- Handling outliers.Remaining 1999822 rows.
  - The fare should be positive.Delete 77 lines with negative fares.
  - latitude:[-90,90],longitude:[-180,180].Deleting rows outside the range.
  - The number of passengers has an outlier of 208, which should be removed.

Data Visualization

- Plot the passenger fare histogram.

  Combined with the data description results of the training set, the histogram of passenger consumption in the training set was made, and the outliers of passenger consumption were screened and removed.



(a) figure 1:100 groups

(b) figure 2:10 groups,x<10

(c) figure 3:10 groups,x>=10

Figure 1: Histogram of different groups

Processing and Computing

- Data processing and computing.
  - Frame the training set range.Find the longitude and latitude range of the test set. The training dataset contains 1957,917 records after being processed.
  - Calculate the distance between pick-up and drop-off points according to the Haversine Equation.

    The records with zero distance between taxi fare and pick-up and drop-off location are invalid, and 1957913 training set data are obtained after deletion.

Data Modeling

Data Modeling

Problem Description
Exploration and Preprocessing
Data Modeling
Data Modeling
Conclusion

- Data modeling.
  - multiple linear regression models matrix representation:

    $$Y = X\beta + \mu$$

    The multiple regression equation is obtained as follows:

    $$Y_i = 4.46 + 2.10X_{0_i} - 0.05X_{1_i} + 0.04X_{2_i}$$

    where X_{0_i} is the value of "H_Distance", X_{1_i} is the value of "weekday+1", X_{2_i} is the value of "passenger_count".

Conclusion

- Data to predict.
  - Putting the relevant fields of the prediction set into the regression equation yields the predicted value of "fare_amount" :

Table 1: Prediction for the field "fare_amount"

	key	fare_amount
0	2015-01-27 13:08:24.000000200	9.29
1	2015-01-27 13:08:24.000000300	9.50
2	2011-10-08 11:53:44.000000200	5.51
...	...	...

Contact Information

Runsha Pan
School of Hunan University

✉ qq2724493252@163.com