FLIP00 REPORT

JIAQI LIU

ABSTRACT. Finishes the kaggle project New York City Taxi Fare Prediction.Practiced data process, visualization and feature selection skills.

CONTENTS

Date: 2021-04-25.
1991 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. Random Forest, Data Mining, ...

1. Introduction

The prediction task involves multiple time and postion features.It expects to get the prediction value of taxi fares.The raw dataset contains taxi trip datas of over 55M

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

2. Data Process

- Drop the missing value.
- Drop minus fare_amount values.
- Drop large passenger_count values.
- Restrict longitude values into [-75,-72]
- Restrict latitude values into [40,43]
- Calculate distance based on latitudes and longitudes.
- Break datetime into year,month,day,weekday,hour.
- Restrict distance values into (0,200] by equation:
  $distance = (fare\_amount - 2.5)/1.56$
- Adjust fare_amount of zero value byu equation:
  $fare\_amount = 2.5 + 1.56 * distance$

3. Data Visualization

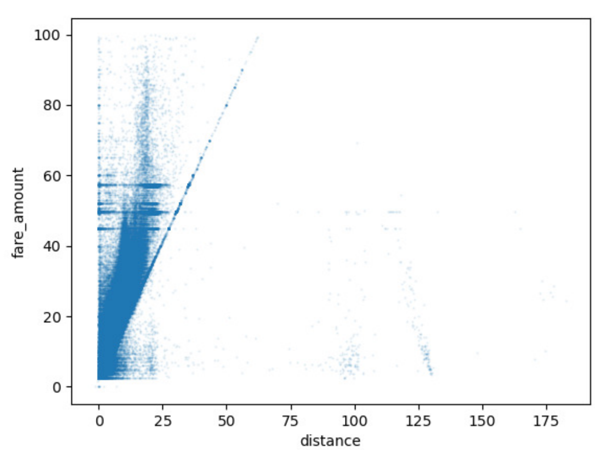

Figure 1. Scatter chart between distance and fare_amount

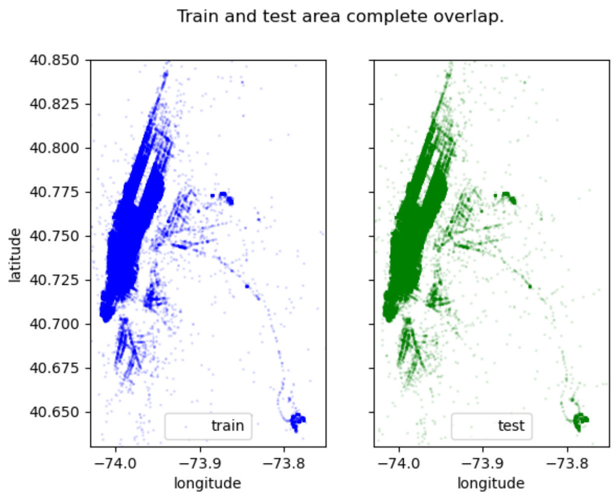

FIGURE 2. Pick up postion map

4. Feature Selection

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

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.

5. Modeling and Result

- Model:random forest
- Score:3.23791
- Rank:474/1483

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

The importance rank consistents with the correlation rank in general, but the importance of postion features and feature weekday differs.

6. Conclusion

- In all the features involved, distance plays the most vital role.
- The fare hardly changes with time.

(A. 1) Department of Computer Science and Technology,, Jilin University, China
Email address, A. 1: `jqliu@tulip.academy`