NEW YORK CITY TAXI TRIP DURATION

SIYI YU

Abstract. In this competition, Kaggle is challenging you to build a model that predicts the total ride duration of taxi trips in New York City.Train.csv is provided by kaggle to help you to bulid the model,and test.csv is used to test the ability of your model.

Contents

1. Introduction 2
2. Related Work 3
3. Method 3
3.1. Data Processing 3
3.2. Model selection 8
4. Experiments 9
5. Conclusions 9
Acknowledgement 9
References 10
List of Todos 10

Date: 2021-04-26.
2020 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. predict, duration, model ...

1. Introduction

The Kaggle Competition is aim to build a model that predicts the total ride duration of taxi trips in New York City.Kaggle provides us three files to help us complete this competition,which are train.csv,test.csv,submission.csv.

The train.csv is aim to help us bulid a good model to make a prediction.There are some details about this file,seeing Figure 1

```
<class 'pandas.core.frame.DataFrame'>
Index: 1458644 entries, id2875421 to id1209952
Data columns (total 10 columns):
 #   Column              Non-Null Count    Dtype
---  ------              --------------    -----
 0   vendor_id           1458644 non-null  int64
 1   pickup_datetime     1458644 non-null  object
 2   dropoff_datetime    1458644 non-null  object
 3   passenger_count     1458644 non-null  int64
 4   pickup_longitude    1458644 non-null  float64
 5   pickup_latitude     1458644 non-null  float64
 6   dropoff_longitude   1458644 non-null  float64
 7   dropoff_latitude    1458644 non-null  float64
 8   store_and_fwd_flag  1458644 non-null  object
 9   trip_duration       1458644 non-null  int64
dtypes: float64(4), int64(3), object(3)
memory usage: 122.4+ MB
```

Figure 1. Attributes of train data

The test.csv is aim to test the capability of the model you have trained.There are some details about this file,seeing Figure 2

```
<class 'pandas.core.frame.DataFrame'>
Index: 1458644 entries, id2875421 to id1209952
Data columns (total 10 columns):
 #   Column              Non-Null Count    Dtype
---  ------              --------------    -----
 0   vendor_id           1458644 non-null  int64
 1   pickup_datetime     1458644 non-null  object
 2   dropoff_datetime    1458644 non-null  object
 3   passenger_count     1458644 non-null  int64
 4   pickup_longitude    1458644 non-null  float64
 5   pickup_latitude     1458644 non-null  float64
 6   dropoff_longitude   1458644 non-null  float64
 7   dropoff_latitude    1458644 non-null  float64
 8   store_and_fwd_flag  1458644 non-null  object
 9   trip_duration       1458644 non-null  int64
dtypes: float64(4), int64(3), object(3)
memory usage: 122.4+ MB
```

Figure 2. Attributes of test data

The submission.csv is aim to provide format for submitting answers.

In this competition,Our work is mainly divided into two steps,which are Data Processing and Model Selection.

During data processing, we first do the data visualization,which will help us to find the stucture of the data and the law of data development.Throung this step, we can have a deeper sight into the data.

After that,we can have better ideas for data processing.We will clean the data,since there are some items against common sense or missing attributes.Removing the oulters and Dealing with missing vlues is necessary.We also need to change type of some attributes,since some attributes have no quantitative relationship.Converting types of this attribute into string is a good choice.In some cases,we need to add some attributes to gain a better prediction.

There are some good models to select,such as Ridge,Bagging,Boosting,RandomForest,Lightgbm,Xgboost.In this process,the most crucial step is the choice of hyperparameters. There is no shortcut, only do constantly experimentations to find the best value during this process.

2. Related Work

There are some popular models to trian the model.

Boosting Bagging puts a lot of small classifiers together, a random part of the data for each train, and then combines their final results (majority voting system).

Bagging Boosting is theoretically more advanced than Bagging, and it is also a classifier. But arrange them linearly. The next classifier adds a higher weight to the parts that were not well classified by the previous classifier, so that the next classifier can learn more "deeply" in this part.

Random Forest Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

3. Method

3.1. **Data Processing.** There are two main steps.

Data Vasulization Frist,we analyze the correlation of attributes.We use Spearman Correlation instead of Person corrlation. The results can be seen in Figure 3.From the figure, we can see that latitude has the greatest impact on duration, and we need to emphasize on this attribute greatly.

For the attribute named vendor_id,it's a code indicating the provider associated with the trip record.We can see it's distribution in Figure4.It's worth noting that it's type is int.We need to convert it into string for one-hot encoding.

For the attibute named store_and_fwd_flag,it's a 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.Proportion of different values of this attribute can be seen in Figure 5

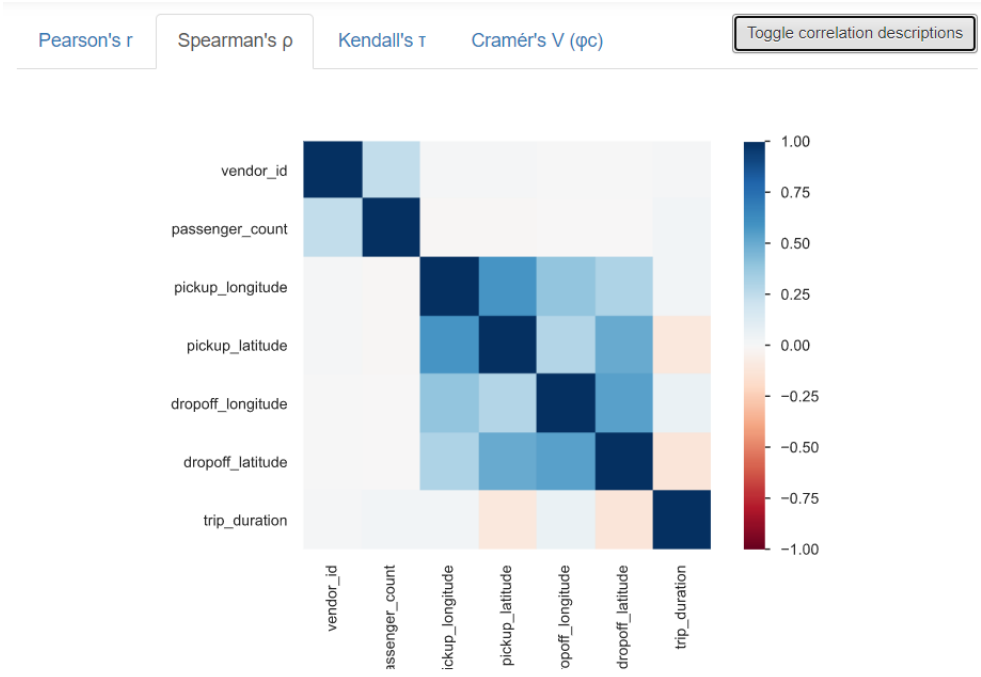

FIGURE 3. Correlation of Attributes

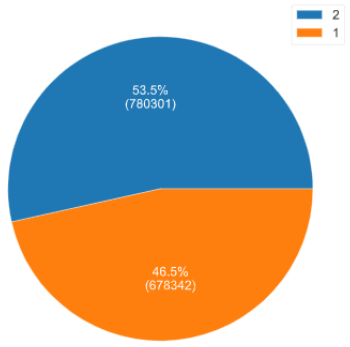

FIGURE 4. Proportion of different values of vendor_id

Value	Count	Frequency (%)
False	1450599	99.4%
True	8045	0.6%

FIGURE 5. Proportion of different values of store_and_fwd_flag

For the attibute named passenger_count,it means the number of passengers in the vehicle (driver entered value).From figure 6,we can find there is some data whose passenger_count is more than 6.Howerver,such a large-capacity taxi does not exist.Therefore,we need to remove these items.

```
1    1033540
2     210318
5      78088
3      59896
6      48333
4      28404
0         60
7          3
9          1
8          1
Name: passenger_count, dtype: int64
```

FIGURE 6. Number of trips for different passenger_count

For the attibute named dropoff_datetime,which doesn't exist in test,it means it's useless for prediction based on test.csv.There,we need to remove this attribute.

For the attibute named pickup_datetime,we first divide it into month, day, hour. From Figure7,Figure8,we can see month,day and hour truely affect trip duration.However,if we direclty use one-hot encoding on the attributes(day and hour), it will cause disaster of dimensionarity.Therefore,we need to divide them into different categories.Moreover,We can deduce the day of the week from the date.From Figure9,it's useful to add this attribute.

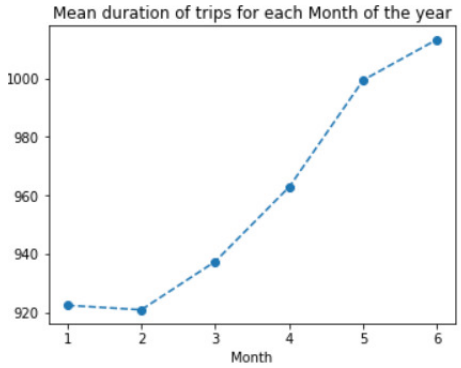

FIGURE 7. Mean trip duration of month

As we all know, time is equal to distance divided by speed.Therefore,I introduced the distance between Euclidean and Manhattan through latitude and longitude in order to predict better.

On the other hand,the speed of taxis is different in different areas, such as urban areas and suburbs.So I divided the city into five clusters in term of longitude and latitude.You can see the result in Figure10.

I also introduced the azimuth angle through the latitude and longitude, because it will affect the speed to a certain extent.

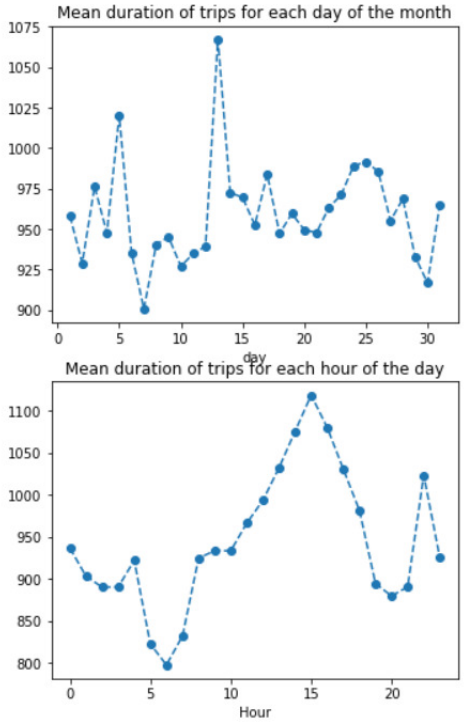

FIGURE 8. Mean trip duration of day and hour

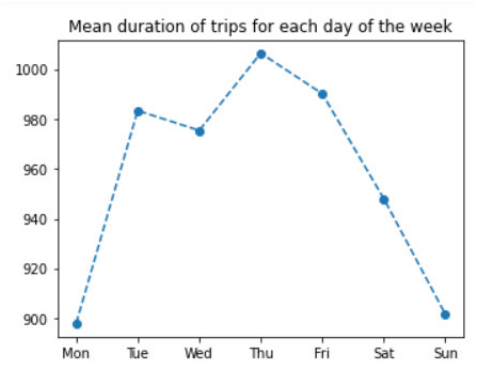

FIGURE 9. Mean trip duration of weekday

I marked the pick-up and drop-off location by latitude and longitude,in Figure 11.From Google,we know New York City longitude vary from -74.03 to -73.75,and latitude vary from 40.63 to 40.85.In term of this, we remove some items.

For the attibute named trip_duration,it's distribution can be seen in Figure12.We can use power conversion to make it more like Gaussian distribution. This can improve our prediction accuracy.The distribution can be seen in Figure13 after power conversion.

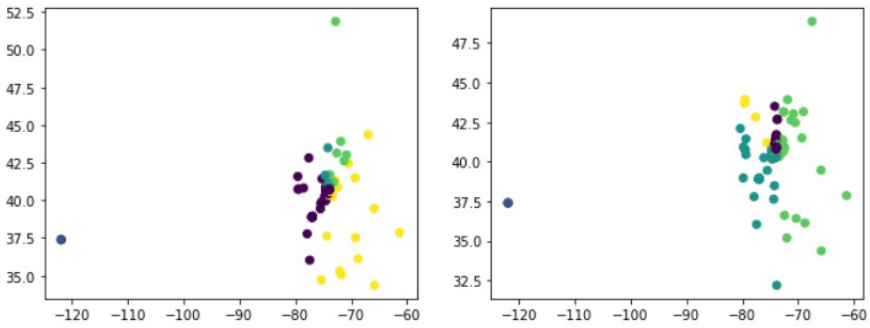

FIGURE 10. Clusters of pick_up and drop_off locations

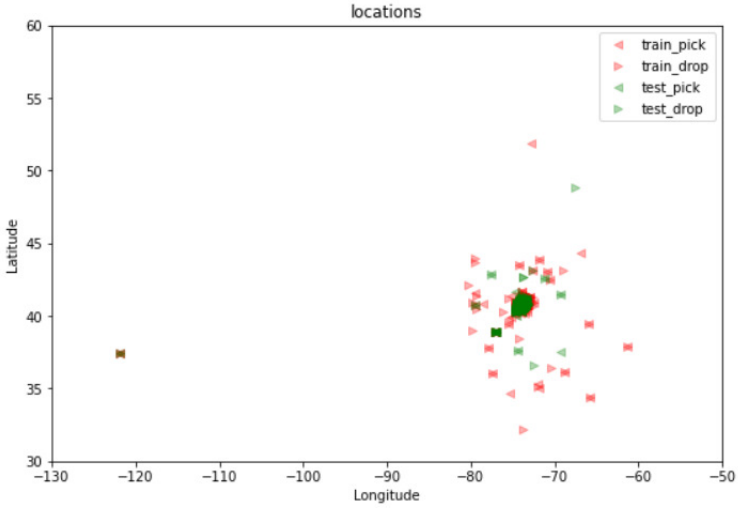

FIGURE 11. Clusters of pick_up and drop_off locations

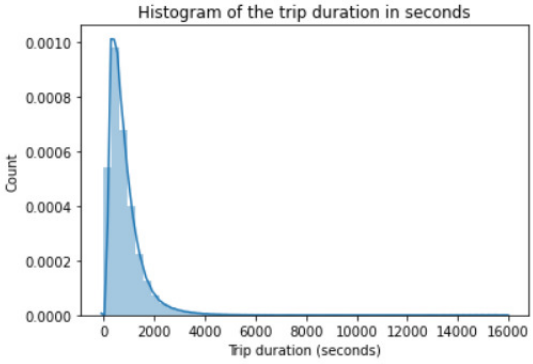

FIGURE 12. Distribution of trip_duration

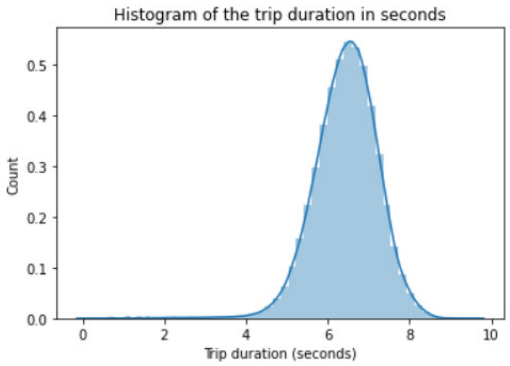

FIGURE 13. Distribution of trip_duration after power conversion

3.2. **Model selection.** There are some popular Models,such as Ridge,Bagging, Boosting,RandomForest,Lightgbm,Xgboost.In our Competition, Lightgbm and Xgboost perform best.It's results can be seen in Figure15,Figure14. Finally, we used these two models at the same time to fit a training data,where the superparameter of Lightgbm is 13 and the superparameter of Xgboost is 11.In the prediction phase, I give the two models 50% of the voting rights.

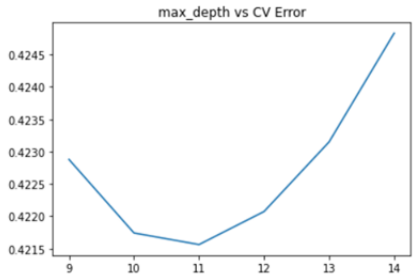

FIGURE 14. Xgboost

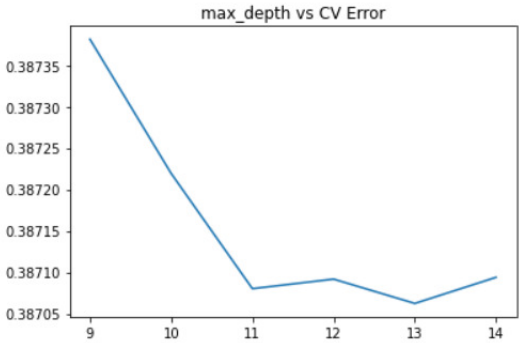

FIGURE 15. Lightgbm

4. Experiments

I conducted a total of 18 experiments.There are three main developments.I start with a prototype that only contains latitude,whose private score is 0.59486 and public score is 0.59596.After data processing completed,the private score is 0.42527 and public score is 0.42716.After that, I introduced two models Xgboost and Lightgbm, and the prediction accuracy has also been improved. My best grade is that the private score is 0.40978 and public score is 0.41166.The results of the experiment can be seen in Figure 16

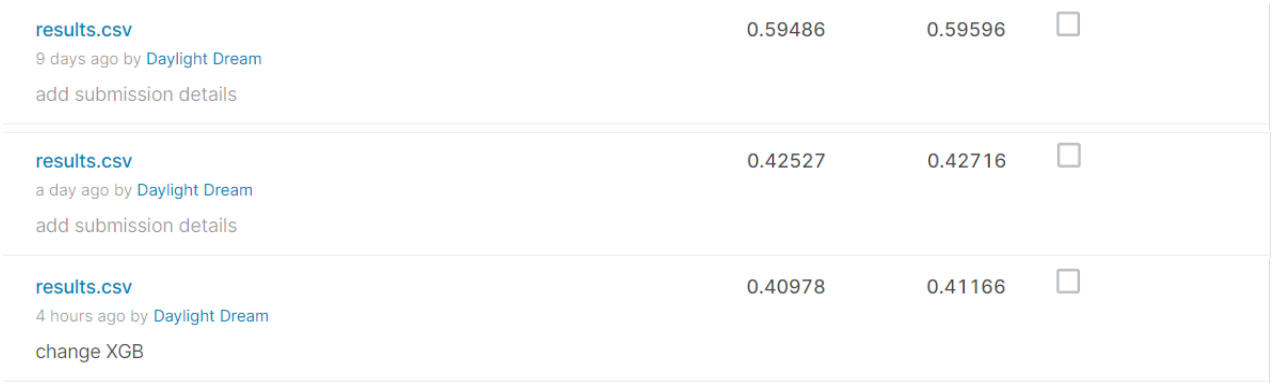

Figure 16. Results of experiments

5. Conclusions

New York taxi trip duration is a prediction problem.The key part of this competition is Data processing,which largely determines the predictive ability of the model.In the process of data processing, what separates you from others is mainly your newly added attributes.In this experiment,The most important thing is your processing of latitude and longitude and the hidden attributes behind it.Model selection is the most time-consuming steps.You have to experiment many times to find the most suitable hyperparameters.

Acknowledgement

References

List of Todos

(A. 1) School of Computer Science,, Jilin University, Jilin, 130000, China
Email address, A. 1: yusiyicsat@163.com