New York City Taxi Trip Duration Prediction

Weiling Deng

Deakin University, Australia

Introduction

This project required to 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, picktime and dropoff time, and several other variables.

Dataset Loading and Description

- In this section, After we read, clearn the data and check NAs, The train data which we have 1458644 Rows and 11 columns. The test data which we have 625134 Rows and 9 columns.We built the linearregression model, using distance to predict the trip duration and test.
- Data Types

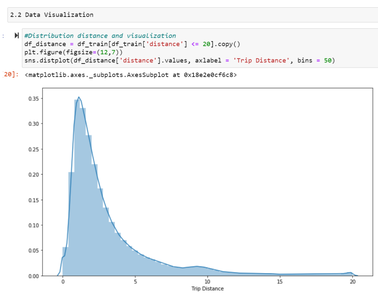
Attribute	Description
vendor id	a code indicating the provider associated with the trip record
pickup datetime	date and time when the meter was engaged
dropoff datetime	date and time when the meter was disengaged
pickup longitude	the longitude where the meter was engaged
pickup latitude	the latitude 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	Y=store and forward; N=not a store and forward trip
trip duration	duration of the trip in seconds

Data Visualization

- Distance characteristic analysis

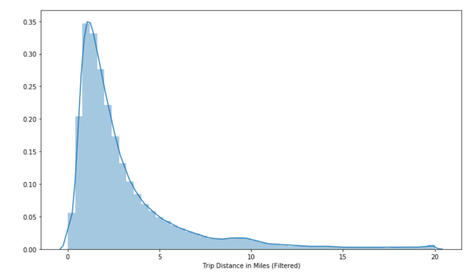
```
2.2 Data Visualization

#Distribution distance and visualization
df_distance = df_train[df_train['distance'] <= 20].copy()
plt.figure(figsize=(12,7))
sns.distplot(df_distance['distance'].values, axlabel = 'Trip Distance', bins = 50)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x18e2e0cf6c8>
```




```
df_distance = df_test[df_test['distance'] <= 20].copy()
plt.figure(figsize=(12,7))
sns.distplot(df_distance['distance'].values, axlabel = 'Trip Distance in Miles (Filtered)', bins = 50)
```

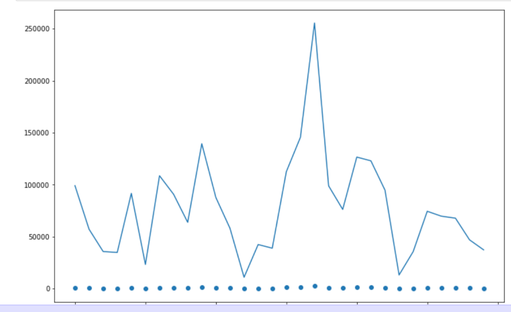
```
<matplotlib.axes._subplots.AxesSubplot at 0x18e3254f888>
```



Visualize Predict Result

- Trip Duration Visualize Predict Result

```
#Visualization the predict data
plt.figure(figsize=(12,8))

plt.scatter(np.arange(len(y_test)),y_test,label = 'true value')
plt.plot(np.arange(len(y_test)),lr_y_predict,label = 'predict value')
plt.show()
```

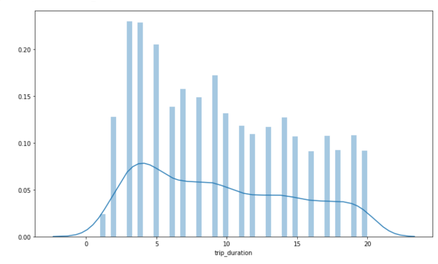


Data Visualization

- Trip Duration characteristic analysis

```
In [22]: df_duration = df_train[df_train['trip_duration'] <=20].copy()
         plt.figure(figsize=(12,7))
         sns.distplot(df_duration['trip_duration'].values, axlabel = 'trip_duration', bins = 50)
```

```
Out[22]: <matplotlib.axes._subplots.AxesSubplot at 0x18e3278d388>
```



Select features and gropby data

Select trip duration and distance as features, groupby data to train data and test data in the train dataframe, set 0 to 150 as train data, 150 to 180 as test data, named as train and test

```
#groupby data to train data and test data in the train dataframe
df=data_train
train = df[:150]
test = df[150:180]
train
```

```
;]:
```

	trip_duration	distance
0	455	1.498521
1	663	1.805507
2	2124	6.385098
3	429	1.485498
4	435	1.188588
...	...	...
145	972	2.211689
146	318	1.136076
147	625	1.897957
148	859	2.304592
149	141	1.578737

150 rows × 2 columns

Built Modeling and Predict Result

Model: LinearRegression Model

RSMLE: 4.616529350404572

Conclusion

The Propose of this project which is to find the data features, select the attruibuates to built the linearregression model, find the best parameter using distance data to predict the trip duration and test.But , from the chart we can see the model is not very good, maybe we an explore other model to improve the accuracy.

TULIP Team for Universal Learning and Intelligent Processing

The 11^{th} International Conference on Knowledge Science, Engineering and Management (KSEM 2018), 17-19/08/2018, Changchun, China