NEW YORK CITY TAXI TRIP DURATION PROJIECT

QIN ZHANG

Abstract. The objective of this project is to establish a machine learning model to predict the total travel time of taxis 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, and several other variables. By analyzing these data sets, we can not only predict the travel time of passengers, but also mine more detailed information about urban travel behavior. Experiments show that the proposed method can effectively predict travel time.

Contents

Date: (None).
2020 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. Machine Learning, Data Mining, ...

1. Introduction

Once, the yellow taxi shuttling through the streets and lanes of New York City was a symbol of the metropolis. The yellow taxi witnessed the changes of New York and was also one of the important transportation modes for every new Yorker. In recent years, with the rise of the sharing economy, the taxi industry in New York has begun to lose its glory. In November 2016, according to the statistics of the taxi driver license system in New York, there were about 13000 taxis in New York. Since 2015, according to weekly statistics, Uber has increased from 10000 to nearly 40000, and LYFT has also increased rapidly, approaching the total number of taxis. Therefore, we hope to effectively analyze taxi data and find out the travel rules and characteristics of citizens. It is helpful to analyze the future market of the taxi industry in New York.

Based on the travel records of taxis in New York from January to June 2016, this paper analyzes the travel data of taxis in New York, and explores the relationship between the travel time of each trip of taxis and the company where the taxis are located, the number of passengers, the boarding date, whether it is a weekend and the travel distance. Firstly, based on the data, the average daily taxi travel time, average travel distance, travel peak and other data characteristics are calculated. Then, the features are further processed. Finally, the LGBMRegressor model, LinearRegression model and DecisionTreeRegressor model are established respectively, and the effect is greatly improved.

In this paper, experiments are carried out on real data sets. We analyzed and modeled the original data, tested four evaluation indicators of three models: LGBMRegressor, LinearRegression and DecisionTreeRegressor, and analyzed the experimental results. The main contributions of this paper are as follows:

- In this paper, experiments on real data sets prove the reliability of the model prediction.
- The comparison of several models shows the stability of the model selected in this paper.

The remainder of this paper is structured as follows. In Section 2, this paper briefly describes the experimental data. In Section 3, the data processing method and feature engineering of the experiment are described in detail. In Section 4, we describe and analyze the experimental results. In Section 5, summarize the experiment.

2. Data Description

The dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform. The data was originally published by the NYC Taxi and Limousine Commission (TLC). The data was sampled and cleaned for the purposes of this playground competition. Based on individual trip attributes, participants should predict the duration of each trip in the test set.

The data set contains 1458644 trip records and 10 features, which is a regression item.

3. Method

The method in this paper mainly focuses on data processing and feature analysis of the New York City Taxi Trip Duration, and then establishes multiple baseline

TABLE 1. Dataset description

	id	vendor_id	pickup_datetime	dropoff_datetime	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	store_and_fwd_flag	trip_duration
0	id2875421	2	2016-03-14 17:24:55	2016-03-14 17:32:30	1	-73.982155	40.767937	-73.964630	40.765602	N	455
1	id2377394	1	2016-06-12 00:43:35	2016-06-12 00:54:38	1	-73.980415	40.738564	-73.999481	40.731152	N	663
2	id3858529	2	2016-01-19 11:35:24	2016-01-19 12:10:48	1	-73.979027	40.763939	-74.005333	40.710087	N	2124
3	id3504673	2	2016-04-06 19:32:31	2016-04-06 19:39:40	1	-74.010040	40.719971	-74.012268	40.706718	N	429
4	id2181028	2	2016-03-26 13:30:55	2016-03-26 13:38:10	1	-73.973053	40.793209	-73.972923	40.782520	N	435

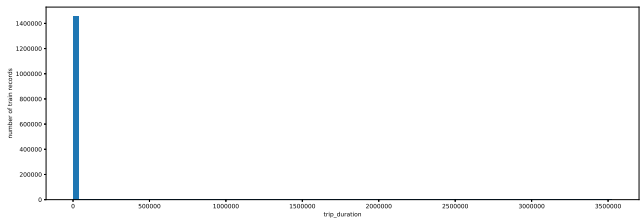

FIGURE 1. Visualize the distribution of trip-duration values

models for comparison. The best prediction model is obtained according to the evaluation index.

3.1. **Outlier Processing.** The processing of outlier values should have different methods for each feature.

- According to the relationship between the position of longitude and latitude and the trip-duration, we can filter out the abnormal data. It can be clearly seen from Figure 1 that the travel time is mostly distributed within 1 million. Therefore, this article will delete the interference data that is not within 1 million. Selecting appropriate time data will help to train the model and improve the accuracy of the experiment.
- Passenger-count is the number of passengers in the vehicle. According to the analysis of Fig. 2, the number of passengers varies from 0 to 9. Therefore, we need to delete the data with 0 passengers.

3.2. **Feature filtering.** Feature processing mainly deals with several special features of this paper.

- Store-and-fwd-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. We convert its character identification into numbers for easy analysis. It can be clearly seen from Fig. 3 that the distribution of 0,1 is uneven.
- Add month, week and day. From the perspective of the trend, the overall taxi time has been increasing from January to June 2016. It may be that users are gradually accustomed to taking a taxi from a longer distance, or it may be that more and more vehicles are driving on the road or the weather is bad, causing traffic congestion. In this article, the original pickup-datetime is divided into hour, day of week and month.
- Create a distance feature. According to the longitude and latitude in the known data, the distance between the starting position and the ending position of the taxi is calculated according to the formula.

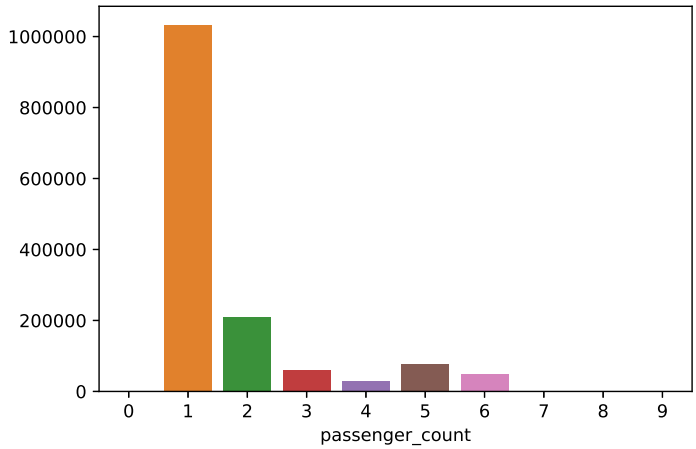

FIGURE 2. Visualize the distribution of passenger-count values

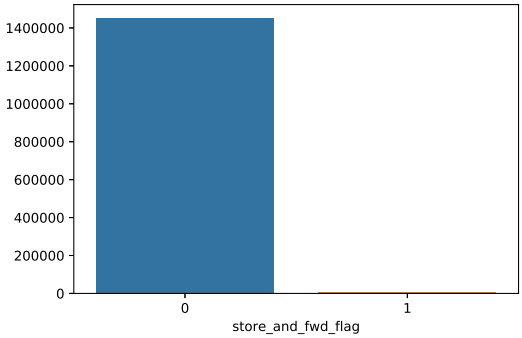

FIGURE 3. Visualize the distribution of store-and˙fwd-flag

4. Experiment and Analysis

Three models were selected in this experiment: LGBMRegressor, LinearRegression and DecisionTreeRegressor. The division ratio of training set and test set is set to 0.3.

4.1. **Evaluation indicators.** In this paper, four evaluation indicators are used: R2-score, MAE, MSE, RMSE.

- $MSE = \frac{1}{m}\sum_{i=1}^{m}\left(f\left(x_i\right) - y_i\right)^2$
- $MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - f\left(x_i\right)\right|$
- $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - f\left(x_i\right)\right)^2}$

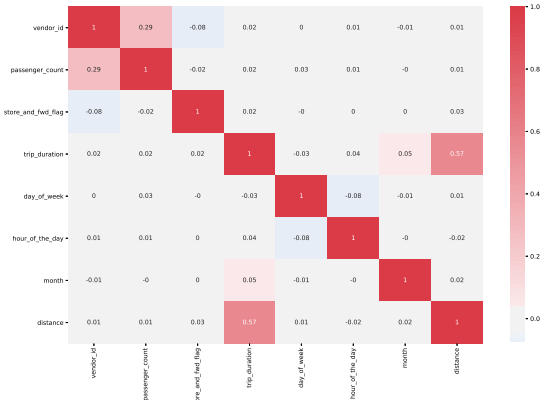

FIGURE 4. Features relationship heatmap

TABLE 2. The experiment result on New York City Taxi Trip Duration

Model	LGBMRegressor	LinearRegression	DecisionTreeRegressor
R2_score	0.66	0.35	0.3
MSE	0.22	0.42	0.45
MAE	0.32	0.46	0.46
RMSE	0.47	0.65	0.67

4.2. **Experimental analysis.** From table 2, it can be clearly seen from the experimental results that there are obvious differences in the evaluation effects of the three baseline models on the four evaluation indexes. Generally speaking, the evaluation value of LGBMRegressor model is better than other baseline models, and better evaluation results can be obtained.

5. CONCLUSIONS

The data of the simulation experiment of this project is from the real data of New York City Taxi Trip Duration, which is conducive to verifying the effectiveness of the model. In this project, the effects of 3 baseline models on 4 evaluation indexes are compared. Finally, it is found that the prediction effect of LGBMRegressor model is the best as a whole.

The analysis of this experiment can not only accurately predict the travel time of taxis, but also mine more detailed information about urban travel behavior. For example, it is possible to analyze which areas in which periods of time are more prone to orders, and where people generally go from which places - this is an effective data for rental scheduling. It can be inferred from the abnormal value brought by the blizzard that the weather is closely related to the order quantity. According to the weather data corresponding to the date, the impact of the weather and the

order quantity can be further analyzed. Combined with the location data, it can also analyze which areas are greatly affected by the weather.

Acknowledgement

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

The authors would like to thank ...

References

List of Todos

(A. 1) School of Computer Science,, Chongqing Technology and Business University
Email address, A. 1: qzhang@tulip.academy