# New York City Taxi Trip Duration Project

Qin Zhang
Chongqing Technology and Business University

2022/8/21

# Overview



## Project Description

Project Description

## Data Description

Data Description

## Feature Engineering

Data Preprocessing

Feature Filtering

## Model introduction

Model introduction

## Evaluation Results and Analysis

Experimental result

## Conclusion

Conclusion



# Project Description

# Project Description



- The competition 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 project provided 1458644 pieces of effective data for the construction of machine learning model. The purpose of the project is to predict the trip-duration time of passengers through "characteristics" such as pickup-datetime and dropoff-datetime.



# Data Description

# Data Description



- Data fields:
  - id : a unique identifier for each trip
  - 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
  - passenger-count : the number of passengers in the vehicle (driver entered value)
  - 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 : This 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
  - trip-duration : duration of the trip in seconds

# Data Description



## 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 |

- 1458644 rows × 11 columns.
- 10 features, and trip-duration is the prediction target.

# Data Description



- Visualize the distribution of trip-duration values.

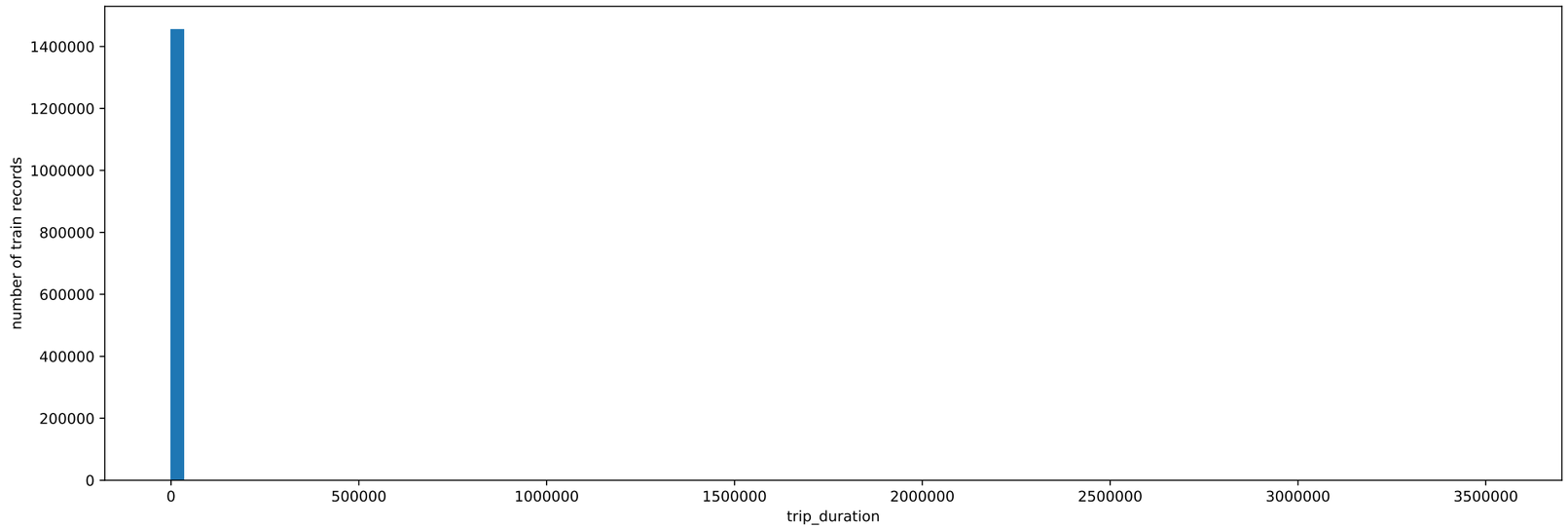



Figure 1: Trip-duration



# Feature Engineering

# Data Preprocessing



Table 2: Number of non empty samples and field type of each feature

| | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | id | 1458644 non-null | object |
| 1 | vendor_id | 1458644 non-null | int64 |
| 2 | pickup_datetime | 1458644 non-null | object |
| 3 | dropoff_datetime | 1458644 non-null | object |
| 4 | passenger_count | 1458644 non-null | int64 |
| 5 | pickup_longitude | 1458644 non-null | float64 |
| 6 | pickup_latitude | 1458644 non-null | float64 |
| 7 | dropoff_longitude | 1458644 non-null | float64 |
| 8 | dropoff_latitude | 1458644 non-null | float64 |
| 9 | store_and_fwd_flag | 1458644 non-null | object |
| 10 | trip_duration | 1458644 non-null | int64 |

# Data Preprocessing



- Analyze outliers:
  - Select an appropriate trip-duration range according to longitude and latitude, and delete the interference data.
  - According to the analysis, select the data with trip-duration less than 1 million, and delete the remaining data.

# Data Preprocessing



- Analyze outliers:
  - Delete data with passenger-count of 0.

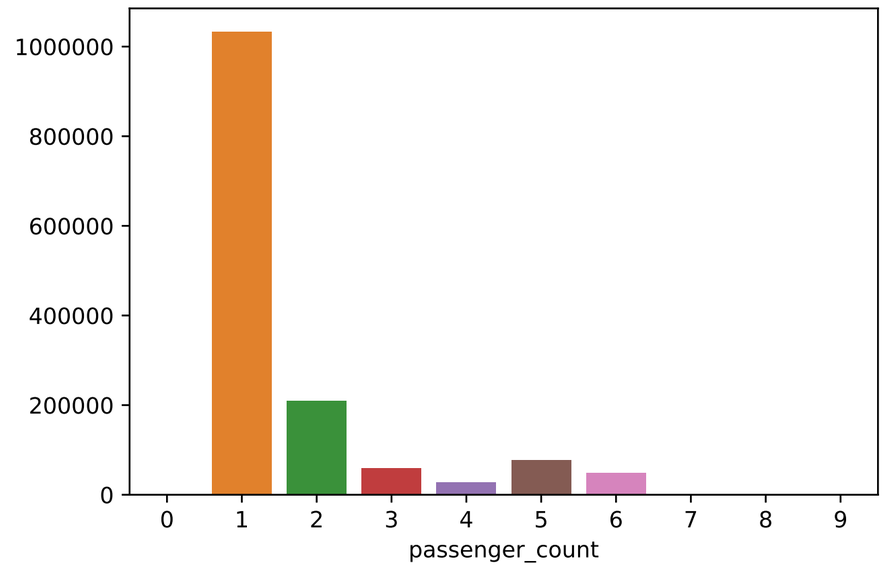
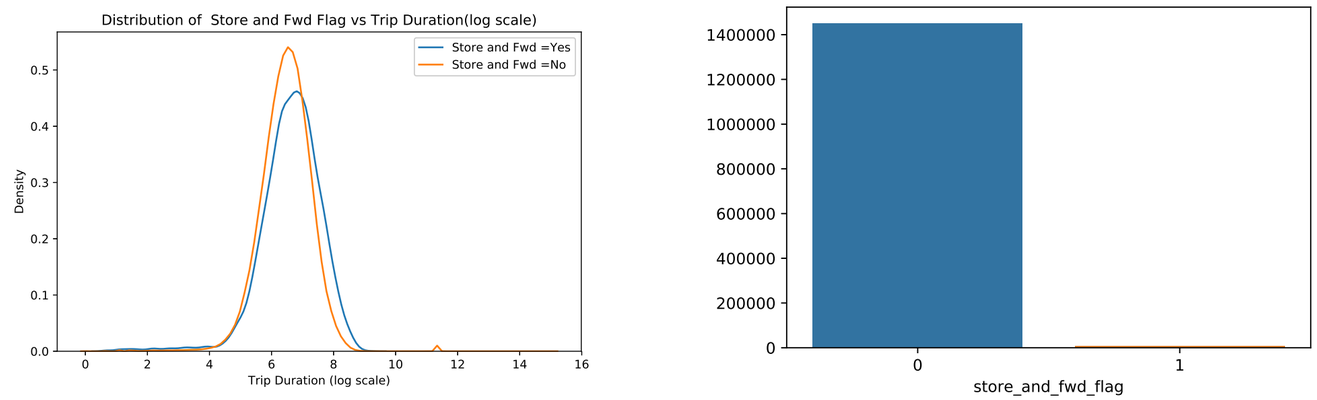



Figure 2: Impact of Embarked on Survival

# Data Preprocessing



- store-and-fwd-flag column: string to numeric




(a) a (b) b

Figure 3: Store-and-fwd-flag

# Feature Filtering



- Datetime: object->datetime-> hour, day of week, month

Table 3: Data after time processing

| | pickup_datetime | day_of_week | hour_of_the_day | month |
|---|---|---|---|---|
| **0** | 2016-03-14 17:24:55 | 1 | 17 | 3 |
| **1** | 2016-06-12 00:43:35 | 7 | 0 | 6 |
| **2** | 2016-01-19 11:35:24 | 2 | 11 | 1 |
| **3** | 2016-04-06 19:32:31 | 3 | 19 | 4 |
| **4** | 2016-03-26 13:30:55 | 6 | 13 | 3 |

# Feature Filtering



- Lat-long to distance by haversine function.

Table 4: Data after distance processing

| | pickup_longitude | pickup_latitude | dropoff_longitude | dropoff_latitude | distance |
|---|---|---|---|---|---|
| **0** | -73.982155 | 40.767937 | -73.964630 | 40.765602 | 1.498521 |
| **1** | -73.980415 | 40.738564 | -73.999481 | 40.731152 | 1.805507 |
| **2** | -73.979027 | 40.763939 | -74.005333 | 40.710087 | 6.385098 |
| **3** | -74.010040 | 40.719971 | -74.012268 | 40.706718 | 1.485498 |
| **4** | -73.973053 | 40.793209 | -73.972923 | 40.782520 | 1.188588 |

# Feature Filtering



- Normalization: trip-duration -> log(trip-duration)

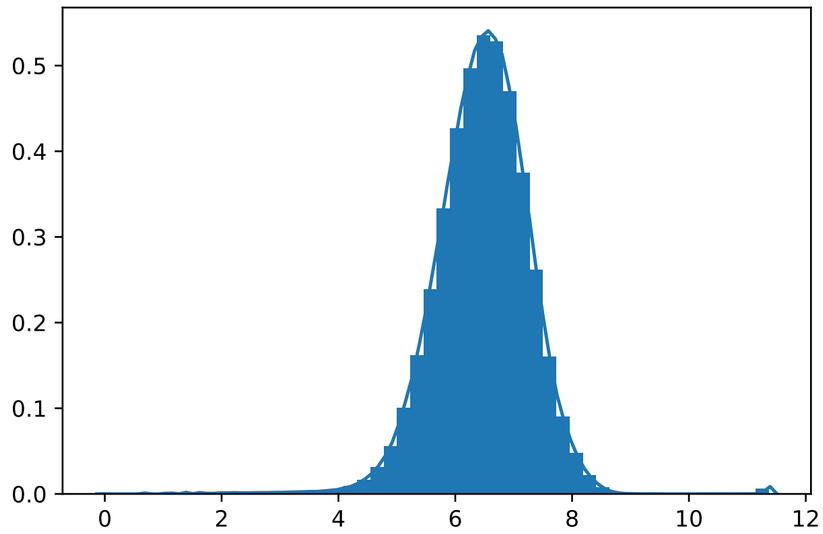



Figure 4: Trip-duration

# Feature Filtering



- The final data result after deleting invalid data

Table 5: Dataset description

| | vendor_id | passenger_count | store_and_fwd_flag | day_of_week | hour_of_the_day | month | distance |
|---|---|---|---|---|---|---|---|
| **0** | 2 | 1 | 0 | 1 | 17 | 3 | 1.498521 |
| **1** | 1 | 1 | 0 | 7 | 0 | 6 | 1.805507 |
| **2** | 2 | 1 | 0 | 2 | 11 | 1 | 6.385098 |
| **3** | 2 | 1 | 0 | 3 | 19 | 4 | 1.485498 |
| **4** | 2 | 1 | 0 | 6 | 13 | 3 | 1.188588 |

# Feature Filtering



## Feature relationship heatmap

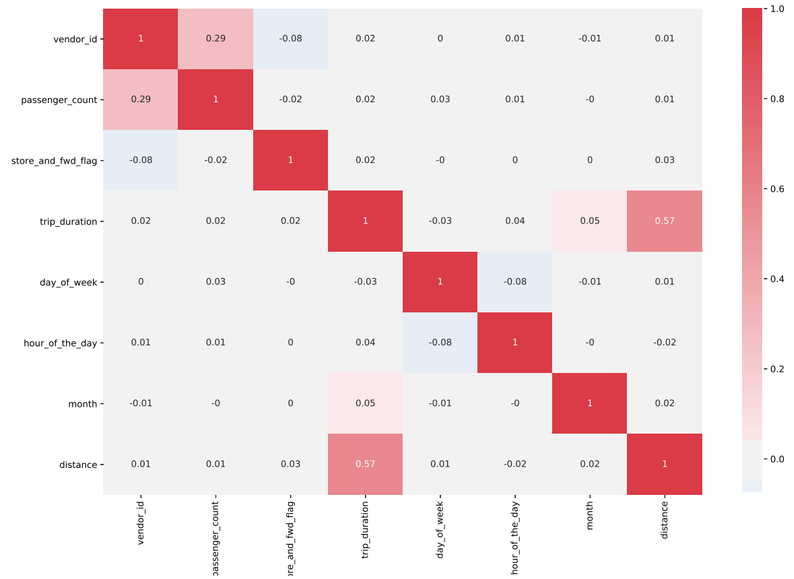
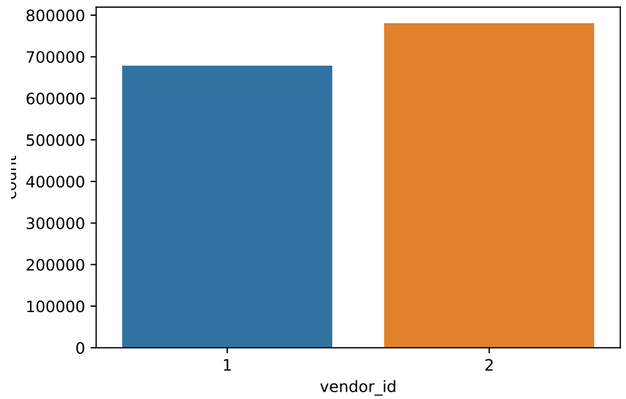



Figure 5: Feature relationship diagram

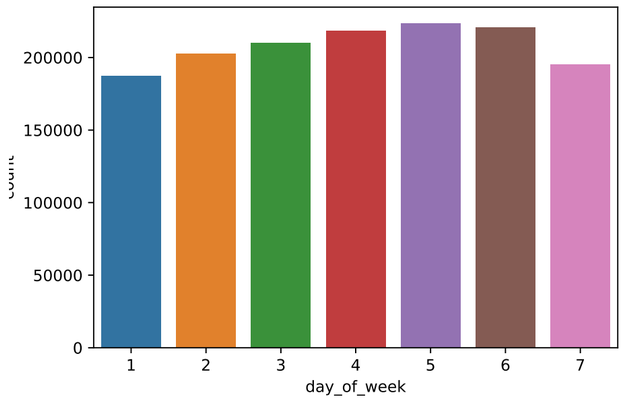
# Feature Filtering



- Visual analysis of important features




(a) vendor-id



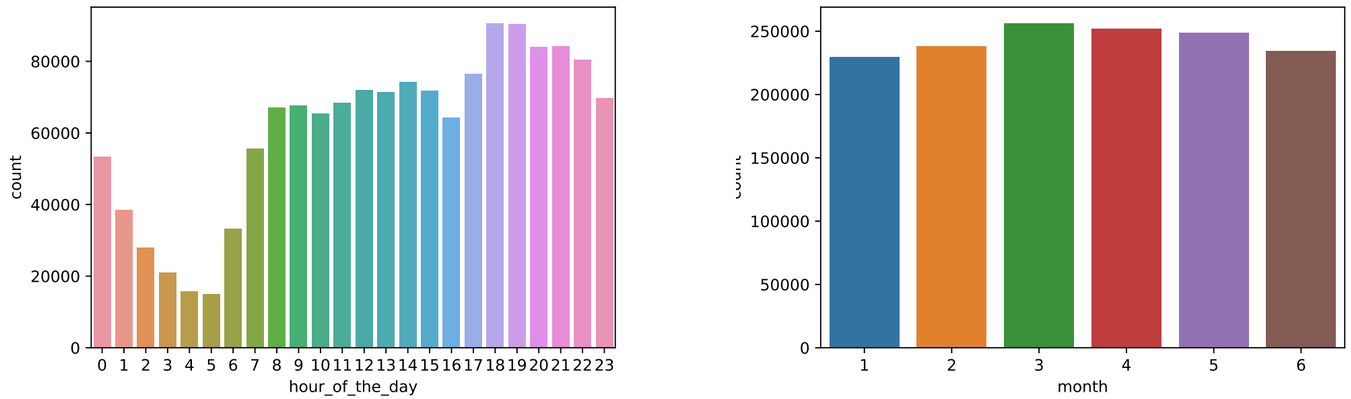

(b) day-of-week

Figure 6: Feature analysis

# Feature Filtering



- Visual analysis of important features




(a) hour-of-the-day (b) month

Figure 7: Feature analysis



# Model introduction

# Model introduction



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



# Evaluation Results and Analysis

# Experimental result



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

# Experimental result



Table 6: The experiment result on synthetic dataset

| 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 |



# Conclusion

# Conclusion



- The competition dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform. The real data set is helpful to verify the authenticity and reliability of the test results;
- 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;
- By analyzing this data set, we can not only predict the travel time of passengers, but also mine more detailed information about urban travel behavior. For example, you can analyze which areas in which periods of time are more prone to orders, and where people generally go from. 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, etc.

# Questions?



# Contact Information


Qin Zhang

School of Artificial Intelligence

Chongqing Technology and Business University, China

QZHANG@TULIP.ACADEMY

TEAM FOR UNIVERSAL LEARNING AND INTELLIGENT PROCESSING