Predicting Taxi Trip Duration in NYC

Zihan Yin
Student ID: 1149307
Github repo with commit

August 25, 2023

1 Introduction

With the advent of the technology age, mobile phones and the internet have become integral to our daily lives. Nowadays, when people need to get around, they may be more inclined to summon a taxi directly through a mobile app rather than the traditional hand waving on the street. Yellow Taxi, New York's most iconic and traditional taxi service company, is facing a turning point of the times. If they fail to adapt and adopt new tech trends on time, they are likely to be replaced by emerging online taxi platforms.

Considering this challenge, it is assumed that Yellow Taxi has decided to transition online and plans to develop a mobile app to serve its customers better. To this end, our report examines Yellow Taxi's trip duration to provide accurate trip duration forecasts for the company's app. Our report is aimed at the following people interested in trip length: passengers, taxi drivers, Yellow Taxi's management team, and relevant agencies responsible for traffic and road management in New York (e.g., the New York City Transit Authority).

1.1 TLC Data

This study uses Yellow Taxi trip data provided by New York City, available at NYC TLC Trip Record Data[1]. Each record describes a trip in detail, including various features such as boarding location, boarding time, and trip distance.

We analyzed records from January 2019 to June 2019 specifically for data selection. There were two primary considerations for the selection of this period:

1. **COVID-19 impact**: Considering trip duration could be affected by the subsequent COVID-19 outbreak, we wanted to avoid this potential bias by selecting data before the outbreak.
2. **Rise of online taxi market**: We see the first half of 2019 as a period of rapid growth for online taxi market, which could be a critical turning point for Yellow Taxi if they plan to expand online.

1.2 External Data

In addition to Yellow Taxi's trip data, we used an external dataset from Visual Crossing's Weather Data Services covering January 2019 to July 2019 weather information for New York City[2]. We chose one more month of data than the TLC data to facilitate the testing of the model part.

Each row in this weather dataset describes in detail the weather conditions in New York City on a particular day, including essential features such as actual temperature, body temperature, humidity, etc. The main reason we chose to introduce weather as an external dataset is that, intuitively, weather conditions significantly impact trip duration.

Below is a brief overview of the raw TLC data and external weather data.

Datasets	**num of instances**	**num of features**
TLC Data	44658561	19
External Data	212	33

Table 1: Dataset Shapes

2 Preprocessing

2.1 Feature Engineering

We derived a new set of features from the original TLC data to explore the factors associated with *trip duration*. Although these features were not present in the original data, we believe they strongly correlate with trip duration.

- **Temporal features**: We extracted a feature named *date*, which includes only month and day information for subsequent merging with external data.
- **Continuous features**: There are 2 continuous features, *trip duration* (in s) and *average speed* (in miles/h).
- **Discrete features**: We generated the following discrete features:
  - *if weekend*: Indicates whether it is a weekend.
  - *if morning peak* and *if evening peak*: Indicate whether it is the morning and evening peak hour, respectively. These 2 features are further used to generate the *if peak hour* and are excluded in the subsequent steps.
  - *if overnight*: Indicates whether it is late night.
  - *if airport*: Indicates whether it is to or from the airport.

  These discrete features are binary, where 1 means "yes" and 0 means "no".

With the above feature engineering, the number of features in the TLC data was increased from 19 to 28, while the number of instances remained the same.

For the external dataset of weather, it covers a wide range of different features. We noticed the existence of a feature called *conditions*, which provided a large amount of helpful information and integrated some information from other features. From this, we extracted vital information and created several new features. Firstly, we constructed a temporal feature *date*, extracting only the month and day from the original temporal feature to facilitate subsequent integration with the TLC data. Second, we created five discrete features based on the conditions feature: *if rain*, *if snow*, *if overcast*, *if cloudy*, and *if clear*. Similar to the newly created discrete features in the TLC data, these five features take the value of either 0 or 1, where 1 represents "yes" and 0 represents "no". After the above feature engineering, the number of features for the external data increased from 33 to 39, while the number of instances remained the same.

2.2 Remove Invalid Instances and Useless Features

Based on Yellow Taxi's data dictionary[3], we found many instances in TLC data that did not match business rules or were considered invalid. For the accuracy of data, we cleaned or adjusted these instances:

- In order to ensure the standardization of *payment_type*, we kept the values (1 and 2) corresponding to the payment type "credit card" and "cash".
- Removed *RatecodeID* values that were not in the range of 1 to 6.
- Only *store_and_fwd_flag* with values 'Y' and 'N' are retained.
- Deleted *fare_amount*, *tip_amount*, *total_amount*, *number of passengers*, *trip_distance*, *extra_fee*, *trip_duration*, and *average_speed* that are less than or equal to 0.
- Since *mta_tax* and *improvement_surcharge* are fixed fees, we retained only the values of 0.5 and 0.3, corresponding to these two features, respectively.
- In *congestion_fee*, lots of instances are null, which we assume the congestion fee was not incurred during the trip, or was not recorded for some reason, so we replace these null values with 0.

After the above processing, the number of instances of TLC data decreased from 44658561 to 26163008.

After the initial data cleaning, we further removed 15 features from the TLC data:

1. We removed 9 features that were not related or causally related to travel time, which cannot be used for prediction. The main ones were different types of fees: *VendorID*, *RatecodeID*, *store_and_fwd_flag*, *payment_type*, *fare_amount*, *mta_tax*, *improvement_surcharge*, *tip_amount*, and *total_amount*.
2. 4 features, from which essential information was previously extracted to create new features, were discarded: *pickup_time*, *dropoff_time*, *if_morning_peak*, and *if_evening_peak*.
3. Two features, deemed redundant due to overlaps with others, were also eliminated: *airport_fee* and *extra_fee* were found to coincide with *if_airport*. Moreover, all entries under *airport_fee* were null, prompting its removal.

2 features that were considered redundant due to overlap with other features were also removed: *airport_fee* and *extra_fee* were found to overlap with *if_airport* and *toll_fee*. In addition, all instances under *airport_fee* were null, so they were removed.

After feature removing, the number of features in the TLC data is reduced from 28 to 13.

After examining the external data, we found that the dataset had high quality with no invalid instances. Therefore, we directly remove the features that are not useful or have had information extracted from them. In the end, we saved 9 features:

- The 6 features generated during our feature engineering step above.
- 3 continuous features: *temperature* (temperature body feels like), *uv_index* (ultraviolet ray intensity), *visibility* (visual clarity).

2.3 Remove Outliers

There were 5 features in TLC data that we thought may have outliers: *trip_distance*, *trip_duration*, *average_speed*, *congestion_fee*, and *toll_fee*. By plotting the box plot, we found that *congestion_fee* did

not have outliers (Figure 1).

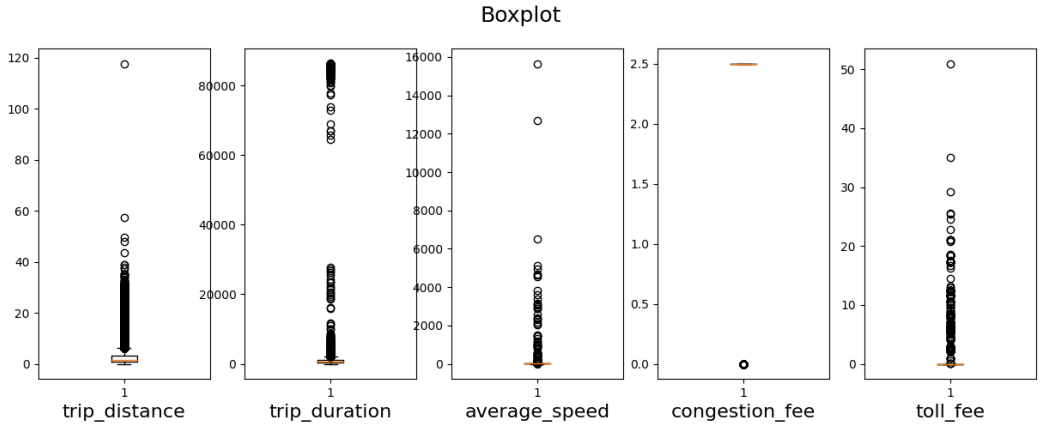

Figure 1: Boxplot for Detecting Outliers (Sample size = 0.01)

For the other 4 features, we used the IQR method to remove outliers.

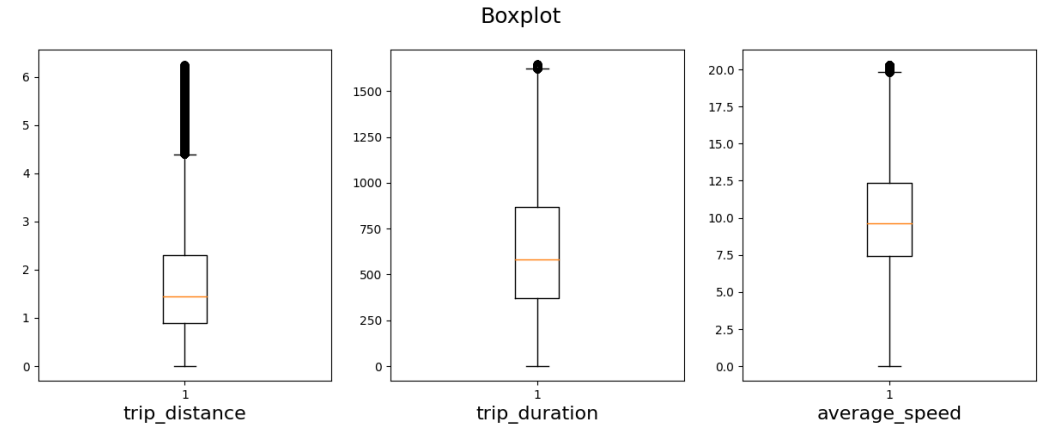

Figure 2: Boxplot After Removing Outliers (Sample size = 0.01)

In fact, after applying this method to *toll_fee*, we found that all values were considered as outliers and removed. Considering that this method may be too strict for *toll_fee* feature, we decide not to remove its outliers. After the above processing, we left-joined curated TLC data with external data based on *date* feature. The merged data had 22,015,147 instances and 21 features.

3 Geospatial Visualisation

We initially hypothesised that *trip duration* has a strong correlation with *location id*. To verify this, we decided to use geospatial plots to check whether different locations have a significant effect on *trip duration*. To do this, we created aggregated data for *up_location_id* and *off_location_id* using the merged data and calculated the *average trip duration* corresponding to each location. Afterwards, we merged these 2 sets of aggregated data with taxi zones data based on *location id* separately. Based on these two aggregated data, we plotted 2 choropleth maps: Average Trip Duration for Pickup Location (Figure 3) and Average Trip Duration for Dropoff Location (Figure 4). Finally, we marked the top 10 locations with the longest *average trip duration* on these 2 plots.

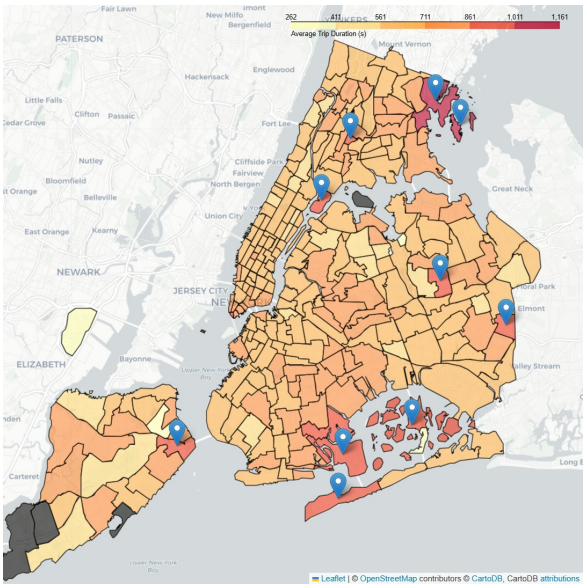

Figure 3: Top 10 Average Trip Duration for Pickup Location

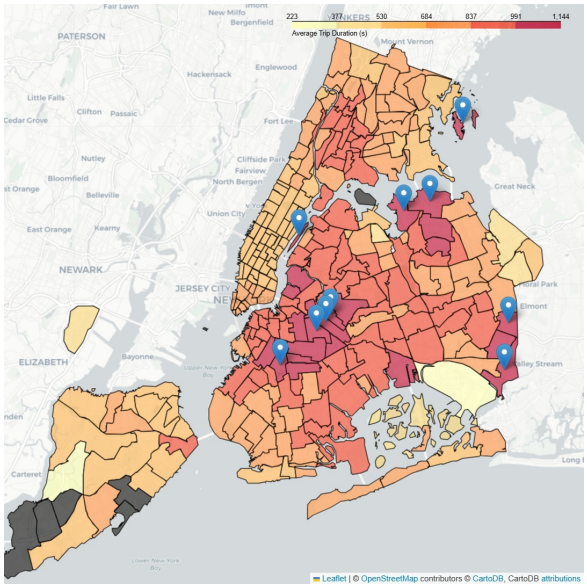

Figure 4: Top 10 Average Trip Duration for Dropoff Location

After observing the geospatial maps, we noted an interesting phenomenon in Figure 3: 7 of the top 10 locations with the longest average trip duration are seaward or island areas: Fort Wadsworth, Floyd Bennett Field,Riis Beach, Jamaica Bay, Randalls Island, Pelham Bay Park, and City Island. Similarly, in the map based on *dropoff_location_id*, 4 of the locations, Roosevelt Island, College Point, Whitestone, and City Island, are also by the sea or islands (Figure 4). This observation is discussed in depth in Discussion section. In general, these 2 maps demonstrate that there is a significant difference in the average trip duration between different locations, which further demonstrates the strong relationship between *trip duration* and *location id*. In further work, we can also consider *pickup_location_id* and *dropoff_location_id* separately and investigate whether they have an interaction effect on *trip duration*.

4 Plotting and Analysis

We first plotted the distribution of *trip_duration*. After observation, we found the distribution of the feature showed a positive skewness. To get a shape close to the normal distribution, we performed a log transformation on *trip_duration*. However, the transformed distribution showed a left skew.

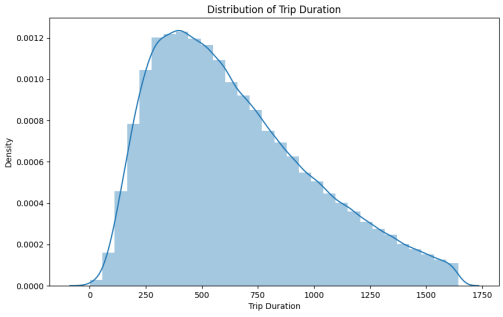

Figure 5: Distribution of Trip Duration (sample size = 0.01)

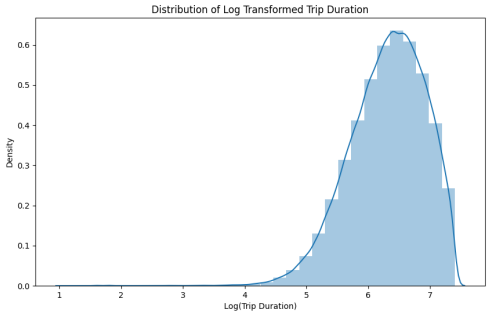

Figure 6: Distribution of Log Transformed Trip Duration (sample size = 0.01)

To deeply explore the correlation between continuous features and *trip duration*, we plotted a pair plot. Through observation, we found that *trip distance* shows an approximately positive correlation with *trip duration*, while *average speed* shows an approximately negative correlation with *trip duration*. However, the other continuous features do not demonstrate significant correlations.

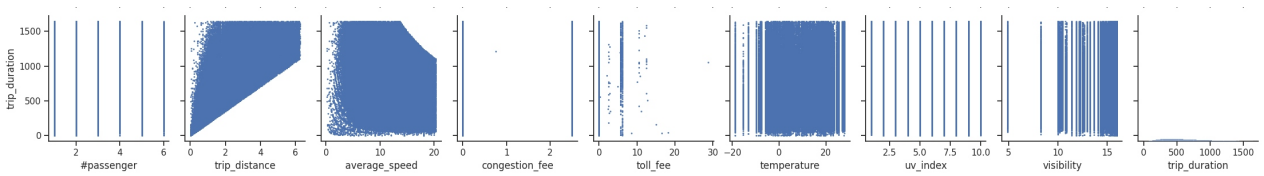

Figure 7: Part of Pair Plot for Continuous Features (sample size = 0.01)

Next, we plotted Pearson and Spearman coefficient based heat maps for the continuous features. In the Pearson based heat map, *trip distance* and *trip duration* show a strong positive correlation, while *average speed* shows a weak negative correlation. Meanwhile, the correlations of *number of passengers* and *visibility* are close to zero, and the other features show only weak positive correlation. In the Spearman-based heat map, although the overall correlation has increased, the overall trend has not changed significantly.

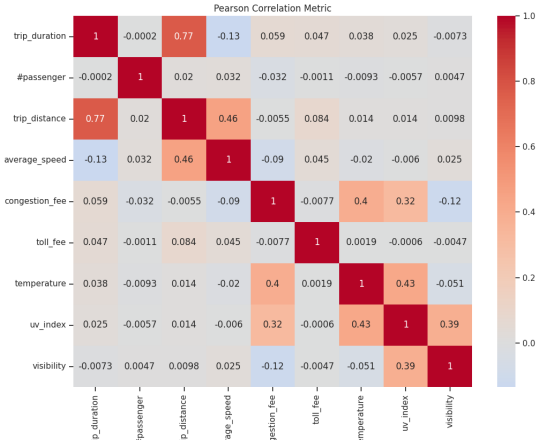

Figure 8: Heat Plot for Continuous features (sample size = 0.01)

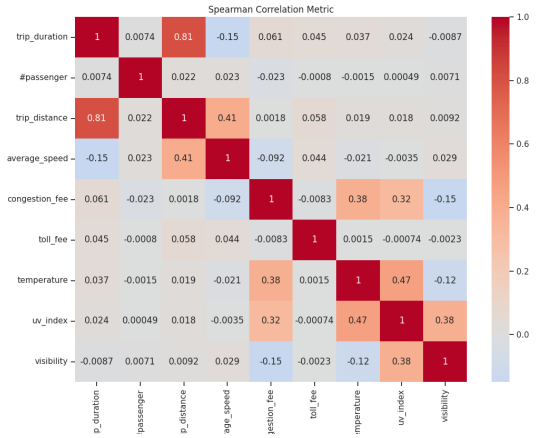

Figure 9: Heat Plot with Spearman for Continuous features (sample size = 0.01)

We further explored the correlation between *trip duration* and the discrete features. Firstly, 1-way ANOVA was performed on the discrete features except for the 2 location features. The results showed that the p-values of these discrete features were all much less than 0.05, which means that they have a significant relationship with *trip duration*(Table 2 (a)). Subsequently, we conducted a 2-way ANOVA with interaction for *up_location_id* and *off_location_id*. The results show that the p-values of both features and their interaction term are also much less than 0.05, indicating that they are significantly related to *trip duration* not only have a significant relationship, but also an interaction effect(Table 2 (b)).

5 Modelling

We selected the yellow taxi data of 2019.07 to be merged with the external data and then used as merged data for test. Merged data was sampled with the sample size of 0.004 to get the train set,

(a) 1-way ANOVA

Feature	p-value
if_overnight	0.000000e+00
if_peak_hour	1.283358e-128
if_weekend	2.001433e-62
if_snow	1.071452e-30
if_airport	1.947484e-17
if_rain	1.986263e-11
if_clear	1.102846e-03
if_overcast	2.390249e-03
if_cloudy	2.405049e-01

(b) 2-way ANOVA with Interaction

Feature	p-value
up_location_id	1.500442e-21
off_location_id	9.166038e-100
up_location_id:off_location_id	1.065981e-286
Residual	NaN

Table 2: ANOVA for Discrete Features (sample size = 0.01)

while merged data for test was sampled with the sample size of 0.01 to get the test set. The inputs to the model have 6 continuous features and 12 discrete features: *num of passengers*, *trip_distance*, *congestion_fee*, *toll_fee*, *temperature*, *uv_index*, *up_location_id*, *off_location_id*, *if_weekend*, *if_peak_hour*, *if_overnight*, *if_airport*, *if_rain*, *if_snow*, *if_overcast*, *if_cloudy*, *if_clear*, *location_interaction_term*. The output of the model is *trip_duration*. *date*, *average_speed*, *visibility* were removed because the former was a feature used to merge the data, and the latter two had correlation coefficients close to 0 in the previous heatmap. *location_interaction_term* was added because it proved to be significant in the previous anova test.

5.1 Linear Regression Model

The linear regression model learnt in machine learning is chosen because the prediction label is continuous variable. linear regression is highly interpretable, fast and suitable as a base line mode. The R^2 of the linear regression model is 0.6314 and 0.6041 on the train and test data, respectively, which means that the model explains more than 60% of the variation in trip duration in the data and performs consistently on the unknown data. The RMSE for the test data $RMSE = \sqrt{(MSE)} = \sqrt{(49433.9984)} = 222.34$ seconds, which is approximately 3.71 minutes. We believe that this error is acceptable for taxi trip duration prediction.

5.2 Random Forest Regressor

The random forest regressor is chosen because it can capture the nonlinear relationship between features and output and is robust to outliers[4][5]. The Random Forest regressor has an R^2 of 0.6953 and 0.6478 on the train and test sets, respectively, meaning that it explains about 5% more variation than the linear regression model. The MSE of the Random Forest Regressor on the train and test sets is 37583.7094 and 43978.8330, respectively, which means that the Random Forest is more accurate on unknown data, with an error of about 209.71 seconds on the test set.

6 Discussion

In Geospatial Visualisation section, we found that more than half of the 20 locations with the highest average trip durations are either sea-facing or islands. Upon closer inspection, we found that these locations were located far from the bridge. It is possible that these locations are a short straight line

distance from their destination on a trip, but they need to go around the bridge that is far away. Therefore we reasonably suspect that the travelling needs of these locations are difficult to meet.

In Plotting and Analysis section, we find that all discrete features are significant for *trip duration*, which contains 5 weather-related discrete features. We also found that *pickup_location* and *dropoff_location* have interactive effects on *trip duration*.

In Modelling section, we think the features for predicting trip duration are generally efficient because the results of prediction are acceptable.

7 Recommendations

For the NYC department of transportation and roads: Travel demand on islands or coasts with high average trip duration could be met by building more bridges or providing more public transport such as ships.

For drivers and passengers who care about the trip duration: pay attention to the weather forecast, as weather conditions can and do affect the trip duration. However, temperature, UV intensity and visibility have a negligible effect on trip duration.

For yellow taxi companies looking to transition to online: focus on trip distance, weather, pick-up and drop-off locations and time of day as variables in predicting trip duration.

8 Conclusion

In general, in the data preprocessing step, we created new features for TLC data and external data, removed invalid or non-business rule rows and irrelevant or already extracted columns. Removed outliers for TLC data, and finally merged TLC data with the external data. In the analysis step, we plotted geopatial plots to see the *average trip duration* for different locations. The distribution of "trip duration" was plotted. Then pair plot and heat map were plotted for continuous features and ANOVA was performed for discrete features to detect feature correlation. Finally, two models were developed to predict *trip duration*.

References

[1] New York City Government. *NYC TLC Trip Record Data*. Accessed: 2023-08-17. 2019. URL: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[2] Visual Crossing. *Weather Data Services*. Accessed: 2023-08-19. 2019. URL: https://www.visualcrossing.com/weather/weather-data-services#.

[3] *Yellow Taxi Data Dictionary*. 2019. URL: https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf.

[4] Scikit-learn developers. *RandomForestRegressor*. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html. Accessed: 2023-08-20. 2021.

[5] Onesmus Mbaabu. *Introduction to Random Forest in Machine Learning*. Accessed: 2023-08-20. 2021.