New York Taxi Trip Tip Analysis and Predict

Junbo Hu
Student ID: 1038361
Github repo with commit

August 21, 2022

1 Introduction

Gig economy is becoming more and more popular in recent years, and it has become one of income for many people in United States. Besides, there is a culture of tipping throughout the United States, while the tip is an important part of income for service industry practitioners. Taxi drivers especially Lyft or Uber drivers is groups of people who benefit from gig economy, and tip from passenger contributes a lot to their income. For drivers, estimating tip income accurately is important for planning works. With the development and application of big data technology, more and more companies, and organizations such as Uber, Lyft and the Taxi and Limousine Commission (TLC) would like to share the trip data to public, these data typically include essential data related to a trip, among them contains the amount of tip.[1].

In this work, we selected Yellow Taxi Trip Data from NYC and the objective of this work is analyzing which factors will influence the tip for driver in a trip and finding a model that can predict the amount of tip for a trip. It might be beneficial for drivers and help them works more efficiently and it also might help platforms like Uber design better algorithm to improve drivers'income.

2 Preprocessing, Analysis, and Geospatial Visualisation

2.1 Dataset

1. Yellow Taxi Trip Data

   The data is provided by New York city government, it includes fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. This work picks the data from September to December in 2021. The data downloaded is in Apache Parquet format, which can be loaded by pandas package directly. It also includes taxi zoon maps to reference.[2].

2. Weather Data

   The weather data is from Visual Crossing, a website provides weather history data around the world and allows user to query data by different conditions. It includes the average, maximum, minimum temperature for a day, wind speed, visibility, humidity, and other weather attributes. This work picks the hourly data in New York from September to December in 2021, and data downloaded is in csv format.

2.2 Pre-process

1. Data Cleaning

   After calculating the share of empty data in the dataset, the result shows that it is reasonable to delete all empty data because the share of them is low which may not influence the whole dataset. Another step took to clean the data is dropping outliers with the help of box plot. In the taxi trip data, fare amount, tip amount and total amount are variables which contains lots of outliers. In this work, the outliers are removed with the help of interquartile range. The results of this process are showed in Figure 1 and Figure 2.

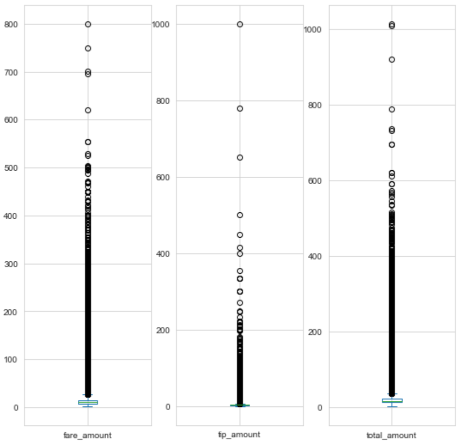

Figure 1: Outliers of amounts before cleaning

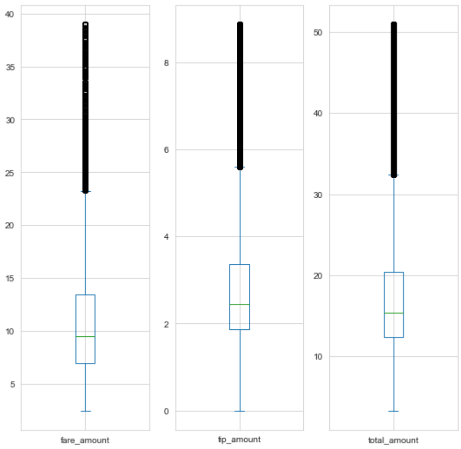

Figure 2: Outliers of amounts after cleaning

   Finally, data which may not helpful for this work is removed by following conditions:

   - Records whose passenger count or trip distance is 0.
   - Records whose pick-up or drop-off data time is out of range.
   - Records whose fare amount is pretty low (less than 2.5 dollars).

   Weather history data query from database has already been examined by website owner so that there is no empty value for important attributes such as temperature, visibility. Bases on this opinion, data cleaning is not applied on weather data.

2. Feature Engineering

   To better research the influence of datetime, this work adds these features: day, week, month, weekday, starting hours, ending hours. In addition, the trip time calculated by timestamp of picking and dropping is also added.

   For weather data, some features may be helpful are picked: feels like temperature, humidity, windspeed, visibility and conditions. Because conditions attribute is category label, one-hot encoding is applied on it.

   The processed taxi data and weather data are combined in a new dataset. The weather data for a trip is based on hour of pick-up time.

2.3 Aalysis and Visualisation

To better understand how features interact, correlation matrix is computed for the dataset, and it is represented in a heat map (Figure 3-4). The heat maps illustrate that tip could be significantly influenced by trip distance, trip time and total fare, besides, weather also influence it in some degree.

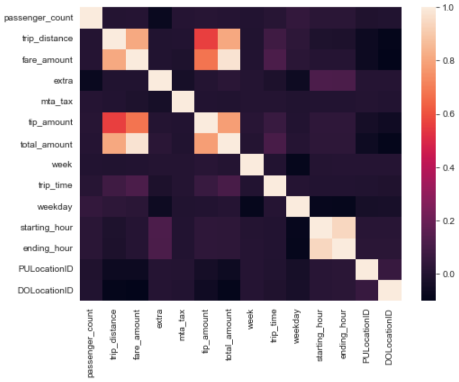

Figure 3: Heat map for taxi trip data

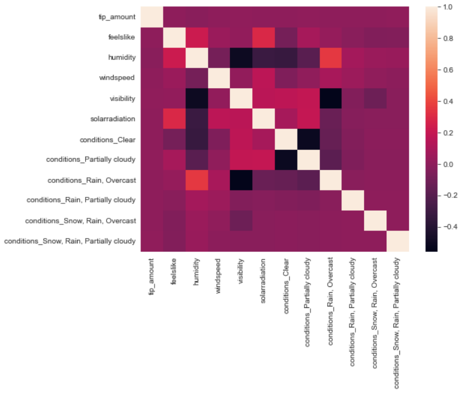

Figure 4: Heat map for weather data

In the next step, the distributions of features are visualized using histogram plot. Log transformation is applied on the data (Figure 5-6). Besides, this work also analysis the average tip for a trip in different weather conditions in Figure 7. In the period selected in this work, the temperature in New York is mainly in 0 to 30 Celsius degrees, the wind speed is slower than 30 meters/s and visibility is mainly higher that 10 kilometers.

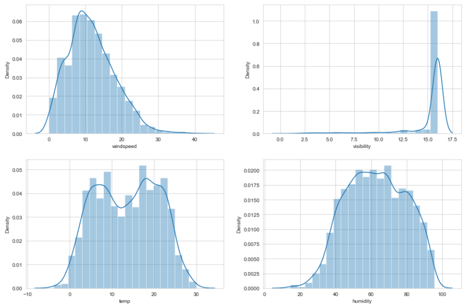

Figure 5: Weather features distribution

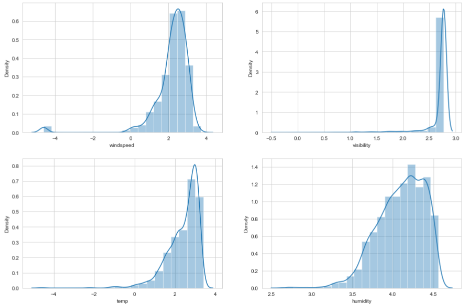

Figure 6: Weather features distribution(With log transformation)

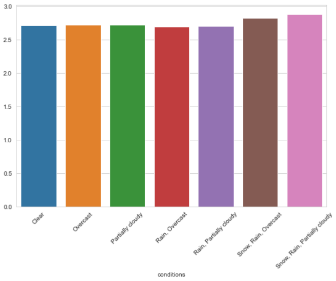

Figure 7: Average tip by weather conditions

Figure 7 shows that for special weather conditions with rain or snow, the average tip is appreciably higher than other conditions.

As geographic-related data is provided, in this work, average drop-off, average tip and average fare are visualized in different destination region in a map graph. Most passengers pay the tip at the destination so that this work mainly analysis the effects of destination on tip. As the Figure 8-10 represent, high average fare and high average tip may happen in the same region, but both may not associate with average drop-off.

For region includes airport such as JFK and Newark, it could have both high fare and high tip. A surprise finding is that in rich region such as Manhattan, the average tip is not really high, although it is the highest drop-off region.

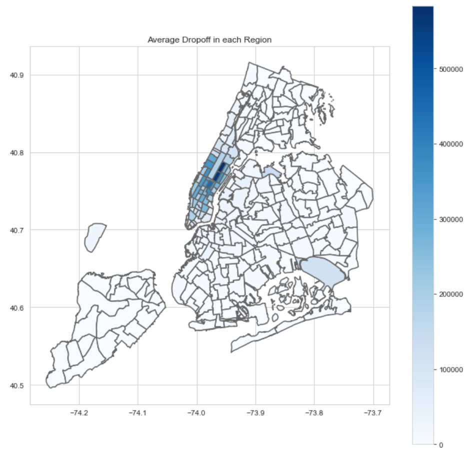

Figure 8: Average drop-off by regions

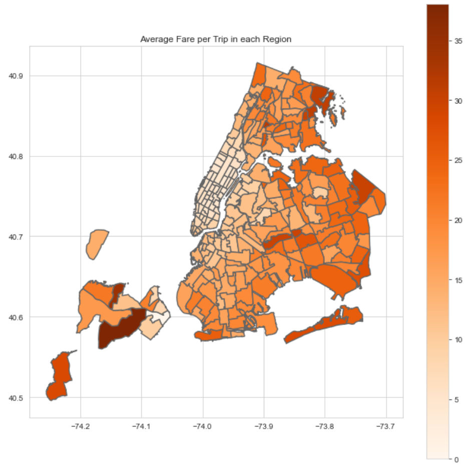

Figure 9: Average fare by regions

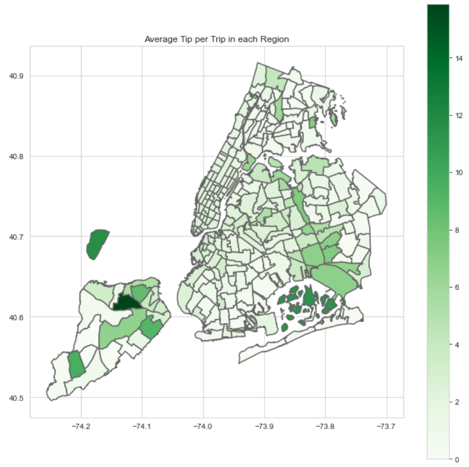

Figure 10: Average tip by regions

By inspecting the relationship between the time of a day or day of a week and the trip frequency, trips occur more frequently around 12:00 to 18:00 in the afternoon between Tuesday to Friday. This work finds that the average tip in work days is a little higher then weekends, but there is no significant difference in different days of week. Furthermore, the results shows that passengers may pay more tip in the nigh, especially in period from 22:00 to 01:00 and from 04:00 to 05:00.

3 Modelling

Regression methods are used to building a model which can predict the tips for each trip. Linear regression and XGBoost regression are tested in this work. Before training the model, features:

’tip_amount’,’total_amount’,’tpep_pickup_datetime’,’tpep_dropoff_datetime’,’VendorID’,’RatecodeID’,

’conditions’,’store_and_fwd_flag’ and ’payment_type’ are removed from dataset.

The whole data are spited into training set and test set. The size of training set is 70 percent of the origin data. To eliminate the influence of different range for features, min-max scale are applied on the data.

Linear regression analysis is widely used in data science which predicts the value of a variable based on the values of another variables. The variable to be predicted is called the dependent variable while variables used to predict the other variable’s value is called the independent variable.[3]

As this algorithm describe the relationship between independent variables and dependent variable using linear equation, it estimates the coefficients of such an equation. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. Simple linear regression calculators that use a least squares method to discover the best-fit line for a set of paired data.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework.[4]

To evaluate the performance of models, we calculated the score and MSE using API in scikit-learn. And this work takes sample of predicted values and real values, plots them in the same graph for comparison. The blue line is real value while the yellow line is predicted value. It is clearly that both models can predict accurate value, but for condition that there is no tip, they may give error predictions.

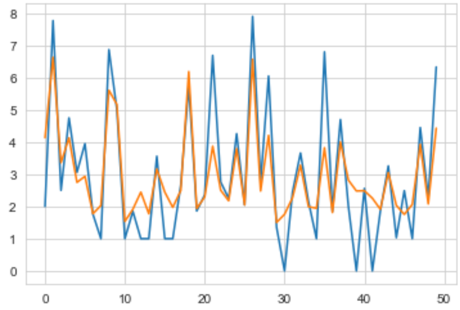

Figure 11: Linear regression model

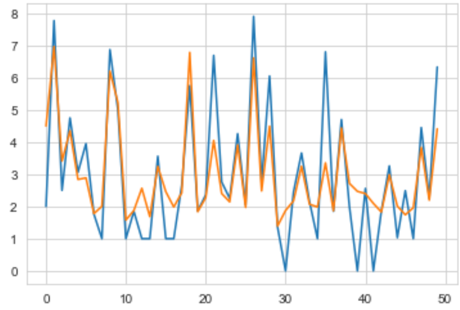

Figure 12: XGBoost regression model

The following table shows details about performance in two models. They both have analogous performance and XGBoost model is a little bit better.

Model	Score-Train set	Score-Test set	MSE
Linear Regression	0.4736	0.4736	1.177
XGBoost Regression	0.5107	0.5061	1.10

In this work, the sum of predicted value in test set is also calculated, and it is compared with real value. The real sum is 7167450, the sum by Linear Regression model is 7166491 and for XGBoost it is 7166592.

4 Recommendations

In this work, only data in four months are used to research, long tern data could help to improve the accuracy. For external data, only weather is considered. Moreover, some other external data set such as the traffic conditions, oil price can be also considered in this kind of work.

Regression model may have tiny error when predicting tip for a single trip, however, as there are enough records in a big sample, the sum of predicted values in close to the real value. In further, for a single trip, it may be helpful if using classifier to implements classification on whether passenger will pay tip or the level of tip such as tiny, general, much.

Furthermore, only two regression models is tested in this work, other regression algorithms such as SVR (Support Vector Machine Regression), Logistic Regression and neural network could be considered.

5 Conclusion

In conclusion, there are many factors which may influence tip in a trip, and it is predictable. Visualization plays an important role in data analysis in this work. It is meaningful to company like Uber in designing an algorithm which helps to schedule drivers efficiently and improve their tip income.

References

[1] Stan Malos, Gretchen Lester, and Meghna Virick. *Uber Drivers and Employment Status in the Gig Economy: Should Corporate Social Responsibility Tip the Scales?* Employee Responsibilities and Rights Journal Vol30 2018,12. 10.1007/s10672-018-9325-9.

[2] *TLC Trip Record Data by NYC.* https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[3] IBM. *What is logistic regression?* https://www.ibm.com/topics/logistic-regression.

[4] Tianqi Chen and Carlos Guestrin. *XGBoost: A Scalable Tree Boosting System.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794, 2016. 10.1145/2939672.2939785.