Yellow Taxi Drivers' Revenue Analysis Based on Linear Model

Jiakai Ni

Introduction

Taxi is a main form of transportation vehicles in people's daily life. Over 200,000 drivers who have TLC licences complete more than 1,000,000 trips in total per day (Taxi & Limousine Commission, 2019). And as growing demand of pick-up servers, it's time to rethink what might be the factors that potentially affect drivers' revenue. It is very likely that once drivers discover there is some kind of patterns (i.e. longer trips or a certain pickup location may lead to less revenue), they might become unwilling to pick up passengers under some circumstance. And this might result in wasting lots of taxis resources. Hence the aim of this report is to analyse what might be the factors affecting drivers' revenue, and to give TLC some suggestions about how to produce a better taxi fare calculator, hence drivers could focus on how to pick up a passenger in shorter time intervals instead of how to pick up passengers that might end up with higher revenue. Hopefully, this analysis is beneficial in terms of helping taxi companies or governments to better distribute taxis resources around the NYC city.

Data

Taxi & Limousine Commission (TLC) collected trip records for both taxis and the For-Hire Vehicles ("FHV"). In order to analyse the latest revenue, the yellow taxi record in February 2019 (Taxi & Limousine Commission, 2019) is used (since the data in 2020 is heavily impacted by COVID-19, it's not representative, green taxi has limitation for pick up location and FHV dataset don't contain lots of attributes). The dataset we selected has 18 features and more than 7 million instances (see more details for dataset on TLC website).

Since weather condition may also be a factor affects taxi drivers' revenue, weather condition data during the same time period is collected from (Weather Underground, 2019). To simplify the problem, we only investigated whether it's precipitated.

Data cleaning

Most data are atomically collected, so there should be no missing data caused by input error. But from a rough observation, some outliers are noticed. Hence some reasonable constraint is implemented to filter out these outliers. The constraints are shown in the table below.

Attribute	Lower bound	Upper bound	Attribute	Lower bound	Upper bound
Passenger count	1	4	Fare amount (USD)	0	80
Duration (min)	0.25	100	Tip amount (USD)	0	20

Total amount (USD)	0	100	Trip distance (mile)	0.01	30
Revenue (USD/hour)	0	200	Extra	0	∞
Mta tax	0.5	0.5	Congestion surcharge	0	∞
Tolls amount	0	∞	Improvement surcharge (USD)	0.3	0.3
Revenue (USD/min)	0	5			

Attribute selection

1. Based on real-world experience, ‘Vendor ID’ and ‘store and fwd flag’ is irrelevant to drivers’ revenue. So even if there is evidence suggesting that there could be a relationship between ‘Vendor ID’ and ‘store and fwd flag', there is no causation. Therefore, these attributes are deleted.

2. Based on definition, 'congestion surcharge', 'extra', 'mta tax', 'improvement surcharge', 'tolls amount', 'total amount' can be derived from other attributes, so they are not useful. Hence, these attributes are deleted.

3. Find start hour, date of week and weather condition based on 'pickup datetime' and 'drop off datetime'. Since we extract enough temporal information from these two attributes, they are not useful anymore. Therefore, these attributes are deleted.

4. As mentioned in the introduction, we only concern about the dataset where payment type is credit card and rate code ID is standard rate. So these two attributes are consistent throughout the whole dataset, which means they are not useful. Therefore, these attributes are deleted.

5. The aim of this task is to analyse what factors might affect revenue, yet not to over-complicate the calculations. Since there are 263 different location IDs, if we include this attribute in the calculations, it would become very messy. So delete pick up location and drop off location.

6. From the Pearson correlation figure shown in below, we notice that trip distance is highly related to fare amount and which is reasonable. Because in the real world, we calculate fare amount based on trip distance and waiting time. Since the assumption of a linear model includes all the factors are independent to each other, so we delete this attribute from the dataset.

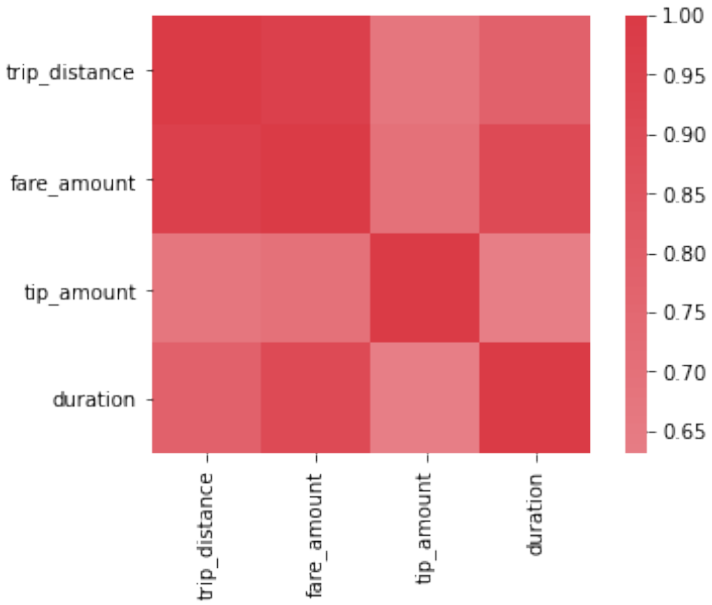

Figure 1 Pearson correlation matrix for few continuous factors

After data cleaning and selection, we end up with 7 attributes which might be relative to revenue. The name and types of attributes are list in the table 1 below. And there are still more than 4 million instances could be used to train and evaluate.

Attributes	Type	Unit
Trip distance	Continues	Miles
Tip amount	Continues	USD
duration	Continues	Minutes
Start hour	Category	
Day of week	Category	
Weather	Category	
Number of passengers	Ordinal	
Revenue	Continues	USD/hour

Table 1 Type of attributes

Transformation

In linear model (LM), we assume all the variables follow normal distribution. But after deleting outliers, obvious skew is observed from the distribution graph and QQ plot (figure 2 and figure 3) So Box-Cox transformation is implemented. Since Box-Cox only works when all the data is greater than 0 and we got lots of 0 tip amount in our dataset. So before applying the Box-Cox transformation, we transferred all 0 to 0.001 cause it's smaller than the smallest currency unit and also reasonably close to 0. After transformation, most sample quantiles fit the line of theoretical value (figure 4 and figure 5). So we recognize these features as normally distributed.

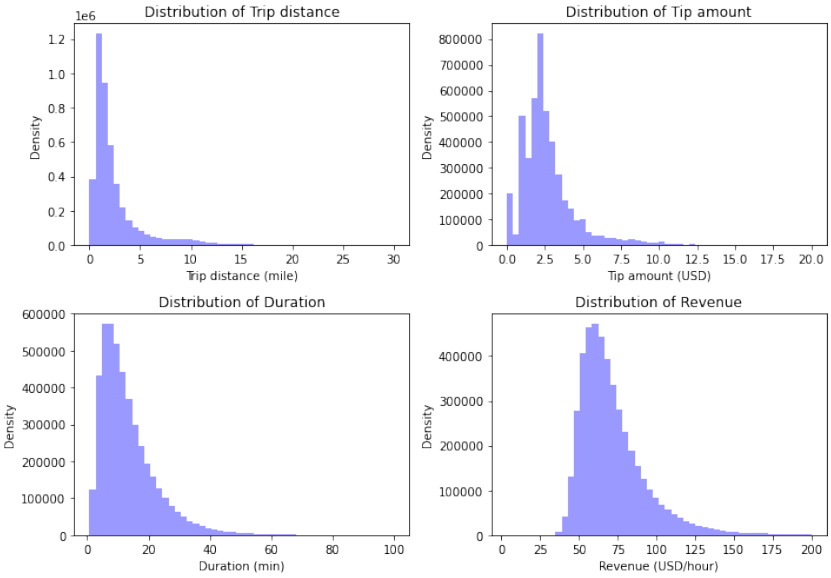

Figure 2 Distribution for continuous predictors before transformation

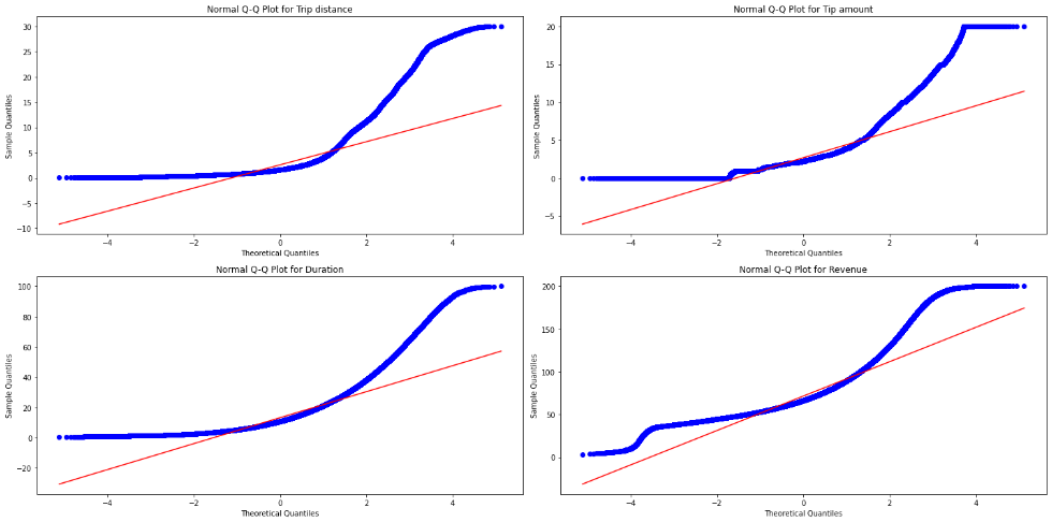

Figure 3 QQ plot for continuous predictors before transformation

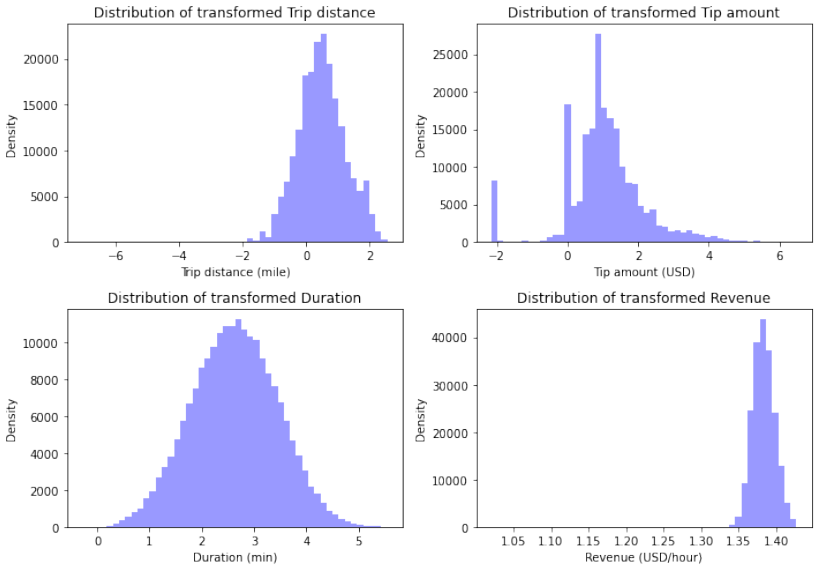

Figure 4 Distribution for continuous predictors after transformation

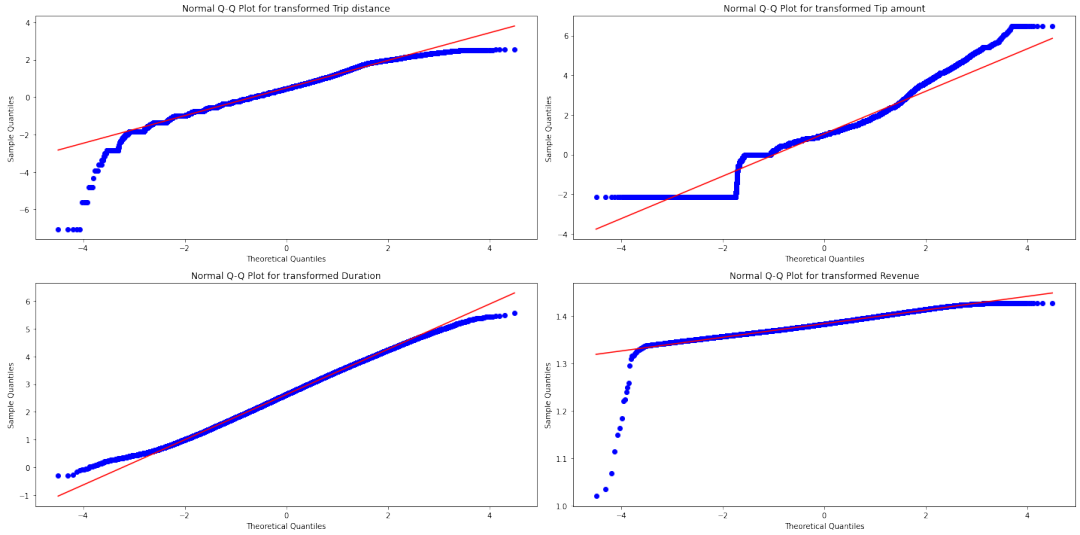

Figure 5 QQ plot for continuous predictors after transformation

Models and Methodology

Revenue is defined as

$$revenue = 60 \times \frac{fare\ amount + tip\ amount}{duration}\ (USD/hour)$$

Box-Cox transformation

Box-Cox is introduced in order to get normally distributed data. The equation is shown in the below.

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln y, & \lambda = 0 \end{cases}$$

Assumptions for linear model after transformed:

1. There is a linear and additive relationship between response and predictors.
2. All predictors are normally distributed and independent to each other.
3. Error of linear model is normally distributed and intendent with all predictors.

Distribution and independency are already checked in the pre-pressing stage. Homoscedasticity and normality of the error will be checked in the analysis part.

Model

Assumption:

There trip distance, tip amount, duration, start hour, day of week and weather might linearly affect drivers' revenue. Besides, duration might has interaction with start hour, day of week and weather.

Linear model is one of the most common models in mathematics and machining learning area. Since lots of objects have linear relationship (or could be transferred into linear relationship) in real life, it has relative decent performance in many areas. In this report, general linear models will be mainly used and evaluated. Other method(s) (such as XGBoost) will be used to evaluate performance of the linear model.

The model after transformation is shown below.

$$\left(\frac{y^{\lambda_y}-1}{\lambda_y}\right)=\mu+\sum_{i=1}^{3}\left(\frac{{x_i}^{\lambda_i}-1}{\lambda_i}\right)\beta_i+\sum_{j=4}^{7}x_i\tau_i+\sum_{k=4}^{6}\xi_{3k}+\varepsilon$$

Where λ_i is lambda for Box-Cox transformation for correspond factors, β_i is coefficient in linear model for correspond factors and ξ_{ij} is interactions for factor i and factor j (see index and corresponding predictors in the table below).

Index	Predicators
1	Trip distance
2	Tip amount
3	duration
4	Start hour
5	Day of week
6	Weather
7	Number of passengers
y	Revenue

Table 2 index and corresponding predictors

Training strategy

holdout method would be used in this analysis and two hundred thousand instances (away more than enough to see whether the predictors are significant and already take few minutes to train) will be used to train the linear model and the rest will be used to evaluate the model.

Evaluation

There are lots of matrices that can be used to evaluate how good the model fits the dataset. In this report, R square, mean square error (MSE) and mean absolute error (MAE) are being used to analyse how many patterns in the model is captured and how large the error is. And in order to have some baseline to compare against with our model performance, we also calculate the values for these metrics by using the mean, median and XGBoost model.

Analysis

Feature selection:

Stepwise is implemented to select features from the full model. Normally we would use AIC as a goodness-of-fit statistic. But from the formula below, we can see that as sample size increase, AIC would mostly depend on $-2\ ln\ ln\ (likelihood)$ since only this term depends on sample size. These will result as all features are maintained which defeat the purpose of feature selection.

$$AIC = -2\ ln\ ln\ (likelihood) + 2p$$
$$= n\ ln\ ln\left(\frac{SS_{Res}}{n}\right) + 2p + const.$$

So we use BIC instead. From the equation below, notice the penalty term in BIC also includes sample size. This means penalty is still important as sample size increases, which is exactly what we want.

$$BIC = -2\ ln\ ln\ (likelihood)\ + p\ ln\ ln\ n$$
$$= n\ ln\ ln\left(\frac{SS_{Res}}{n}\right) + p\ ln\ ln\ n\ + const.$$

Diagnosis

In the pre-processing section, we already checked the statistical independence across each factor and normality, we still need to check homoscedasticity and linearity between response and factors. Here is what comes up

- From figure 6 top left we can see that most fitted data points located in the range between 1.35 to 1.45, and there is no obvious trend in these points. Besides, the residuals line nestles and is fairly parallel to the line y=0, which means so far this model has no major flaws.

- From figure 6 top right we can see that most standardized values are close to theoretical quantiles, which is relatively good.

- From figure 6 bottom left indicates there is no obvious trend of residual is observed, this means variance is stable when factors change, which is a pleasant finding.

- In figure 6 bottom right, there is even no line for Cook's distance, which indicates that no point has high influential outliers to the model. This is exactly the desired outcome.

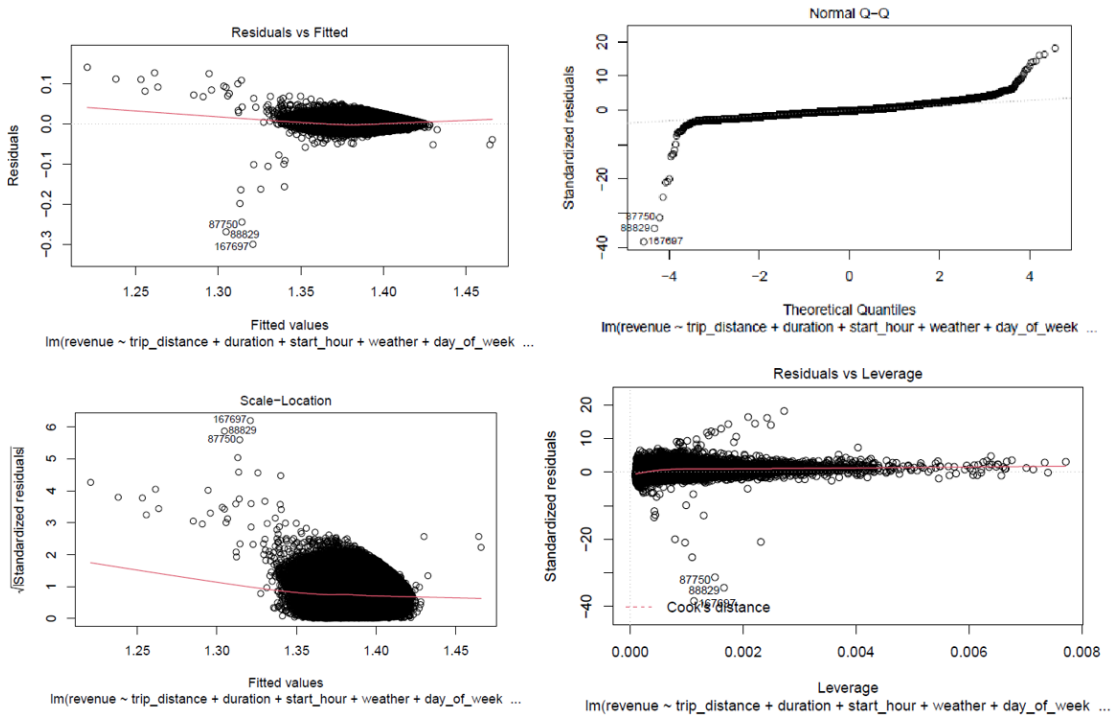

Figure 6 diagnosis plots for linear model

Till now, it seems the model satisfied all the assumptions for the linear model and all the diagnosis plots suggest the model works well for the problem.

Transform test dataset

Use lambda from train dataset, implement Box-Box transformation on test dataset. From figure 7 and figure 8 below, we can see that the distribution for each factor is reasonably close to normal distribution. Next these predictors will be used to calculate response. And evaluate the perforce of model on test dataset.

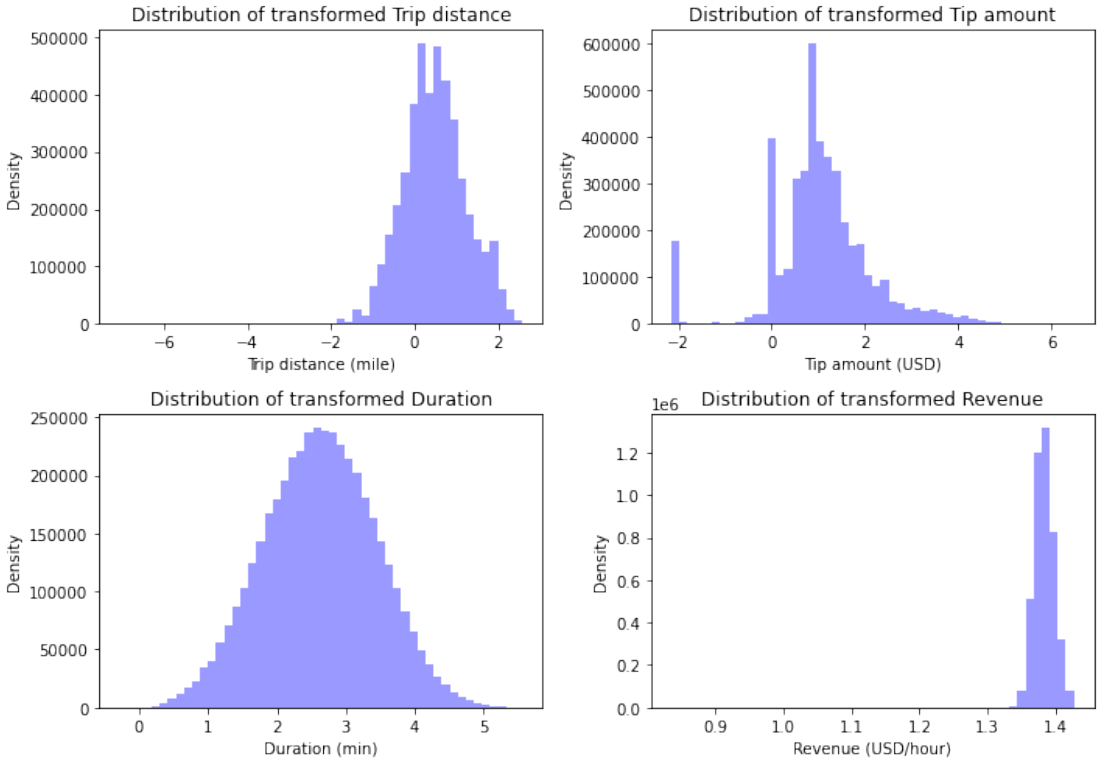

Figure 7 Distribution for continuous predictors (test dataset) after transformation

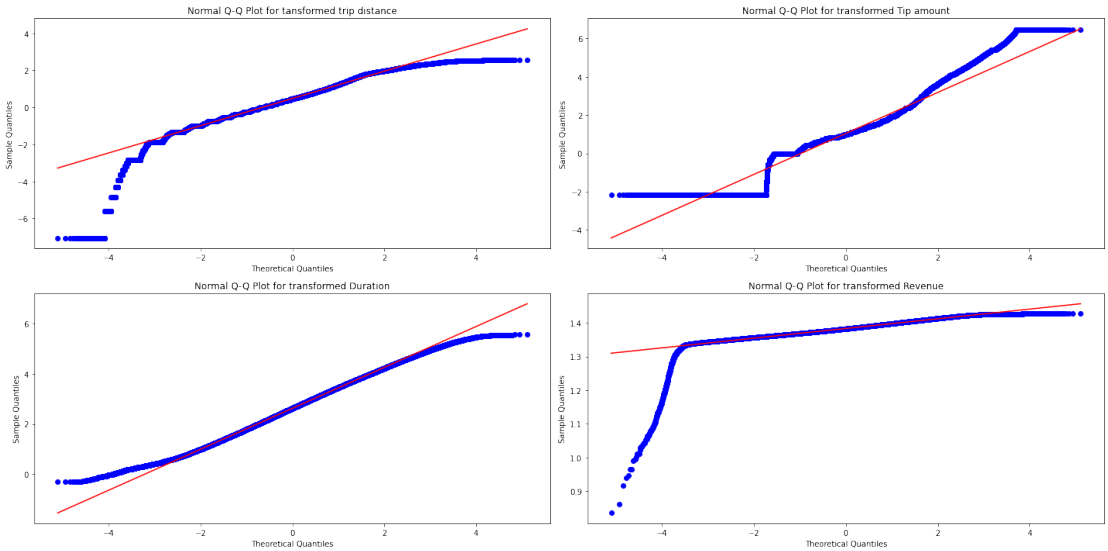

Figure 8 QQ plot for continuous predictors (test dataset) after transformation

Model accuracy

From the table 3 below, we can see that both mean and median have very small R square. This is because we only use a constant value instead of a predicted value, which means there is no pattern encapsulated from the model. And the R square of the linear model is about 0.68, which means we do capture some pattern from the dataset although this information obtained is much less than XGBoost. The error metrics indicate the error for the linear model is much smaller than the error by using mean or median although it underperforms XGBoost.

In general, the goodness-of-fit statistics show that linear models do capture some information from the dataset (the compression is based on predict and true value of revenue after inverse transform).

Model	R square	Mean absolute error	Mean squared error
Mean	0	15.8335	462.8469
Median	-0.0543	15.2751	488.0914
Linear model	0.6312	7.6470	170.6920
XGBoost	0.9893	1.5776	4.9609

Table 3 performance for different model

Conclusion and future work

Conclusion

All the coefficients for linear models every close to 0, this means all the discrete factors are negligible. For the continuous variables, the general relationship between predictors and responds are not linear, so we need to inverse Box-Cox transformation to see how these factors affect revenue. From the figure 9 below, we can see the relationship between revenue before inverse transformation and after transformation. Combine figure 9 and table 4 below, we can see that the influence caused by trip distance is not very significant (since Δx is still relatively small) and the influence caused by duration is relatively significant.

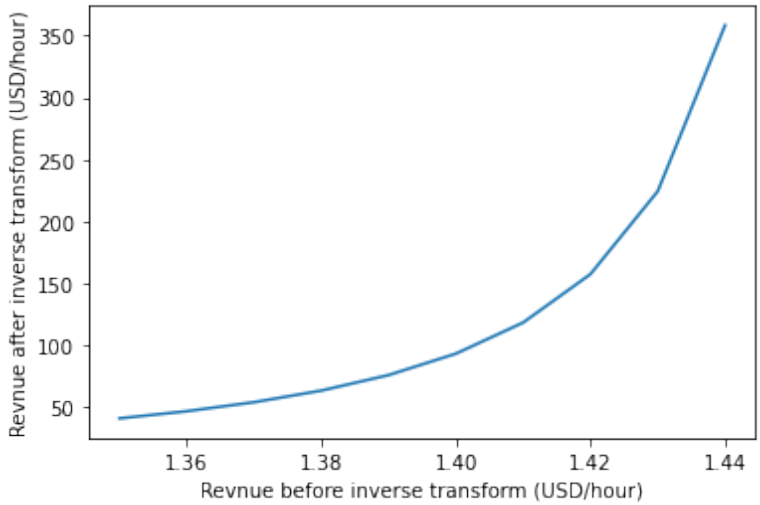

Figure 9 relationship between revenue before inverse transform and after inverse transform

Factor	Coefficient	Maximum - Minimum	effect on revenue (Δx)

Trip distance	0.01918	around 4	0.07672
Duration	-0.02534	around 5	-0.1267

Table 4 how the trip distance and duration will affect revenue (before inverse transform)

This might raise the problem that drivers might be less willing to take on passengers that take long duration since there is an inversely proportional relationship with earned revenues. In order to increase drivers' motivation to deliver these passenger(s) take long trip duration, the government should encourage to develop policies/compensation to incentivize the drivers to take passengers that are travelling long-duration. However, whichever solution is feasible or the exact amount of money can only be determined with further investigation.

Future work

This model fitting only used one-month of data, which has the potential risk of unrepresentative data selection. Future improvement can be made by sampling trip records through an extensive period of time.

In the previous analysis, we already notice XGBoost might perform better than linear models. This might because in real life, revenue = income/duration, which means the relationship between attributes isn't simply linear. Hence for future analysis, we could introduce other models and evaluate if they performance better. Then analysis what factors might significantly include revenue based on these models.

Reference

Weather Underground. (2019). *New York City, NY Weather History* [Daily Observations]. Retrieved from https://www.wunderground.com/history/daily/us/ny/new-york-city/KLGA

Taxi & Limousine Commission. (2019). *TLC Trip Record Data* [2016 January Yellow Taxi Trip Record]. Retrieved from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Data Dictionary - Yellow Taxi Trip Records. (2018). *Yellow Taxi Trip Records*. Retrieved from https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

Taxi Fare Calculation in New York City (2020). Retrieved from https://www.taxi-calculator.com/taxi-fare-new-york-city/259