Qualitative Analysis on Tipping for Yellow Taxi Based in New York City

Andrew Tjen
939206
The University of Melbourne
atjen@student.unimelb.edu.au

Abstract

With the ongoing Covid-19 pandemic causing travel restrictions, it becomes ever so difficult for the taxi industry, specifically its drivers, to generate revenue. The pandemic however, did not impact the strong gratuity culture in America (Heiler, 2020). This report is an extension of Andrew Tjen's previous report "New York City Yellow Taxi Data Analysis" , using starting point attributes that are concluded in the paper (Andrew, 2020). The aim of the report is to provide a reliable model to estimate of profitability measure (In this report Tip Amount) based on different factors and relationship of between each factors of each trips using for Yellow Taxi Drivers of New York City. We will discuss the briefly the result of the previous report, data preparation, along with choice of models and feature selection methods, and finally, conclude our report with a final model and suggestions on interpretation of the result.

1 Problem Formulation

This report will target New York state Yellow Taxi drivers in the pandemic season with goal to provide them with a approximate estimation of amount of tips they will receive for each trip based on several factors. Few factors considered is a result from Andrew Tjen's previous report, which is summarized below. The next section will be details and reasoning behind our data and attribute selection.

1.1 Summary of Previous Findings

- 75% of payment types are credit card payments.
- Covid-19 pandemic significantly affect the record frequency of the dataset, showing a huge decrease from 3 million to 200,000 records between the month of February to March.
- Finding that the Daily Change of Covid case has a negative correlation towards rate of tips
- Precipitation level indicates slight positive correlation towards rate of tips, but cannot be concluded as a strong
- The geographical factors that affects rate of tips can be extended with a measure of quality in the area.
- Faster trip times correlates positively to amount of tips given.

1.2 Data and Attribute Selection

Datasets to be used:

1. **"Yellow Taxi New York TLC Dataset" First Half of 2020**: Contains all trip records about yellow taxi in New York City every month. To asses the impact of the pandemic,

we will again use dataset from the First Half of the year 2020. Justification for this, instead of using one year worth (Second Half of 2019 to First Half of 2020), is due to significant impact of Covid-19 that affects the frequency of records before and after the pandemic hits. Using one Year worth may lead Covid-19 related feature to be insignificant as Covid-19 did not exist pre-2020. Source: `https//www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page`

2. **"Historical Weather Report in New York City" First Half of 2020**: Contains the average weather level report everyday. We will use the First half of 2020. Relevant attributes from this dataset is *"Percipitation Level (in)"*, which represents amount of rain/snow that day. Source: `https://www.wunderground.com/history/monthly/us/nj/newark/KEWR/date/2020-6`
3. **"Covid Case Report" First Half of 2020"**: Contains complete covid-19 case report such as total number of cases and deaths for everyday since the beginning of the pandemic for all states in USA. We will filter this to the state of New York. Source: `https://data.humdata.org/dataset/nyt-covid-19-data?#metadata-0`

Features to be selected:

1. **Tip amount**: Amount of tips given by the passenger. This attribute only consist card payment tips. This will be the attribute to be predicted.
2. **Duration**: The time duration of every trip taken. This attribute is chosen as findings from previous report finds positive correlation towards tip amount.
3. **Passenger count**: Number of passengers for every trip. This is chosen as hypothesis that higher passenger counts will lead to a higher tip.
4. **Fare amount**: The fare price paid for every trip. This is also chosen as a hypothesis that a strong tipping culture exist in USA. Expected to have a positive correlation to prove the 10% tipping culture (Marina, 2018).
5. **Trip distance**: The distance travelled of each trips. This attribute is chosen with hypothesis that longer trips leads to higher tips due to sentimental reasoning.
6. **Daily change**: The increase or decrease of Covid-19 case per day. This attribute is feature engineered from the Covid-19 data. Further description in data pre-processing section. This attribute is chosen as previous findings suggest strong correlation.
7. **Precipitation**: The total level of Precipitation (rain or snow) in the state of New York for everyday. This attribute is selected as previous findings shows some correlation.

We will consider the first 6 months of 2020 as the data period for the analysis with reasons as mentioned in Yellow Taxi data description. Further explanation of the reasoning, 2019 data has an approximately 6-9 million data per month. Since the covid case is 0 in those months, there will be around at least 30 million records that will have 0 in the feature. Therefore, may result insignificance of the Covid-19 attribute. We will then pre-process the data points and transform it into an appropriate data format to be learned by our model planned. The goal of the report is to implement a deeper analysis with statistical measure for Taxi Drivers to have an approximate expectation of amount of tips with trips and everyday conditions such as weather and ,sadly, Covid-19 cases. The depth of analysis that will be achieved in this report is transferable to other areas of the world with similar conditions.

2 Data Pre-Processing and Cleansing

The pre-processing is done in python using several filtering methods. Firstly we use data of each month from January to June of 2020 specifically for Yellow Taxi from the New York TLC dataset. We are dealing with 16,499,407 rows of data in total. Some filtering is needed to obtain a clean data to run our model.

2.1 Initial Cleansing

Firstly, we remove what we consider "invalid" data. We will generate a new feature called duration, which is calculated by using the difference between pick up and drop off times in minutes. Records that has negative duration will be considered invalid as this is accounted as human data entry error. Next, we will filter the datasets that only have payment type of "1", which means credit payment only. This is because tip amount in the TLC dataset only records tips that are paid by credit card. This is justified as previous report shows that 75% of the dataset consist of credit card payments (Andrew, 2020). Next we need to merge the external datasets of weather and covid-19.

2.2 Feature Engineering and Merging

For weather attribute, We will take the precipitation level every day from January to June of 2020 from our weather report dataset source. We first copy the date and precipitation level from January to June of 2020 to our excel. We then convert it into CSV format with ";" separation and load it into our Virtual Machine Python 3 framework. Since excel has a different DateTime format, we need to convert the DateTime format to fit python's format. Due to lack of knowledge in weather conditions, we only can interpret one attribute. Precipitation level represents snow or raining weather. Concern arises of the accuracy of the data as we only take per day and not per hour. There is no publicly available source more detailed than per day analysis, therefore we are working on what we have then left join this Precipitation level into our taxi data by their date with new formatted pick up time of taxi dataset.

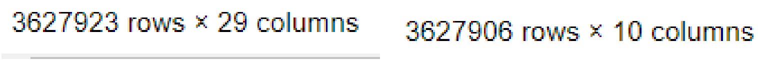
Finally, we need to prepare the dataset of COVID-19 cases. The dataset starts on March 1 of 2020. Thus, we need to self-sample data to have a range of data starting from January 1 of 2020. We will sample data from January 1 until February 29 with 0 COVID-19 and death cases. We will then generate a new attribute where we have the number of new COVID-19 cases per day to represent COVID-19. This is a better measurement because daily changes of COVID-19 are more volatile and we can observe how daily changes can affect the profitability of drivers. It's also due to social bias where we judge the state of COVID-19 by the number of new cases. We generate this attribute by subtracting the number of total cases today with the previous day. Then we ensure that the dataset is within the first 6 months of 2020.

2.3 Random Sampling

One drawback of this devised experiment is that the lack of computation power. As the model we will run will be runned in R-studio, dealing with millions rows of data may not be the most ideal. This is also could be due to lack of experience using R-studios. A solution approached is to random sample the dataset to reduce the dimensions. For every month between January to March, we take a random sample of 1 million dataset due to having more than 3 million data. The month of April, May and June has a data less than 1 million, so instead we will be using the whole dataset. These months is when covid-19 has a significant impact on New York , leading to Travel Restriction regulation and lock downs (Chris F, 2020). An upside to this random sampling is that we can possibly observe an increase in significance of our Covid Data as we use less data from Pre-covid period.

2.4 Further Cleansing

Further refinement and dealing with missing data is needed to be fitted to the model planned. The predictors attribute for our model are mostly continuous. Therefore, missing attributes can be vital to our model. Negative values in fare amount and tip amount will be considered as human error and be discarded. The reason of it being discarded instead of formulated is because the amount of data removed is insignificant compared to the size of data.



3627923 rows × 29 columns

3627906 rows × 10 columns

Figure 1: Before and After Final Cleansing.

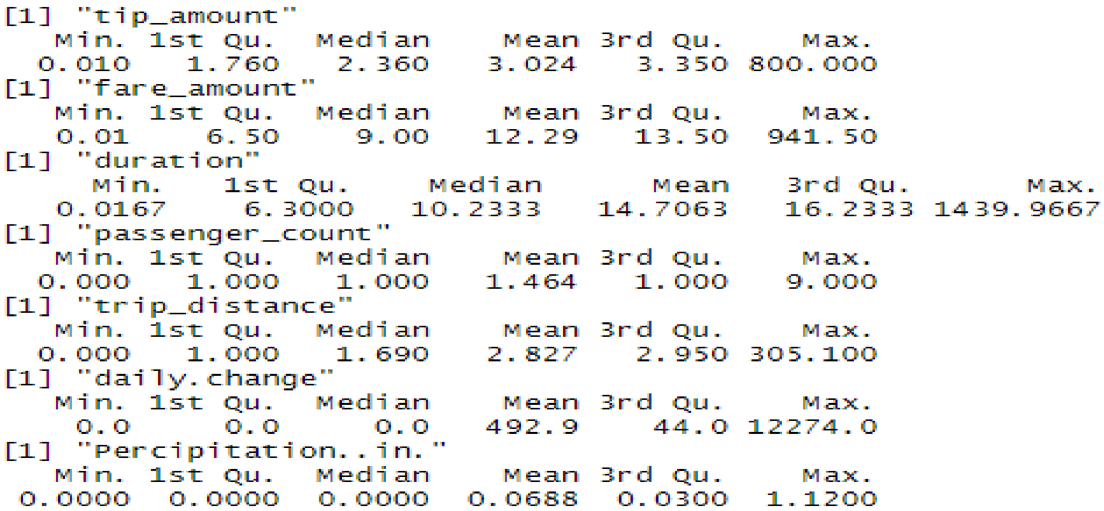
We then filter the attributes to be analyze and ,finally, have a clean dataset to use for our model. Final data we are dealing with 3,627,906 rows of dataset. We will do train-test split of 80% to 20% to test our model.

3 Analysis

3.1 Descriptive Analysis

Before fitting the model, it is good practice to statistically analyze attributes before fitting models. This can lead to adjustment and expectations before fitting to the model and get our results. We will use R-studio for section 3.1.1 and 3.1.3 and use python correlation function on section 3.1.2.

3.1.1 5 Number Summary Statistics



```
[1] "tip_amount"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.010   1.760   2.360   3.024   3.350 800.000 
[1] "fare_amount"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.01    6.50    9.00   12.29   13.50  941.50 
[1] "duration"
     Min.    1st Qu.    Median      Mean    3rd Qu.      Max. 
   0.0167    6.3000   10.2333   14.7063   16.2333 1439.9667 
[1] "passenger_count"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   1.000   1.000   1.464   1.000   9.000 
[1] "trip_distance"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   1.000   1.690   2.827   2.950 305.100 
[1] "daily.change"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0     0.0     0.0   492.9    44.0 12274.0 
[1] "Percipitation..in."
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.0688  0.0300  1.1200 
```

Figure 2: 5 number summary statistics.

The 5 number summary statistics is a basic statistics showing minimum , 1st Quartile, median, mean, 3rd Quartile, and maximum value of each features. There are some information that can be taken from this data.

1. The significant increase from 3rd Quartile and maximum value shows a possibility of an outliers. This is prevalent in few attributes (Tip amount, fare amount, duration, and trip distance). We can further look into possible outliers into our diagnostic plots. Rest of attributes looks to be as expected.
2. Disadvantage of using 5 number summary is that it does not take account unit of measurements. Thus, unable to compare one attributes to another.

To take care of the last point, we will try to observe the correlation of each attribute.

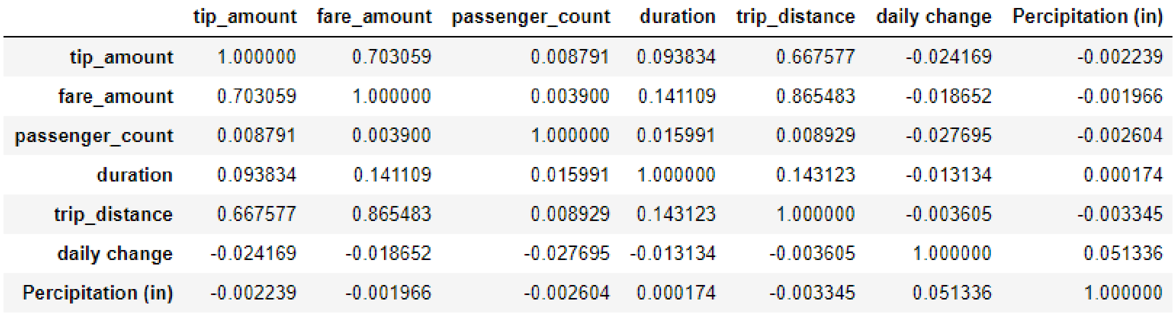
3.1.2 Correlation Analysis

The aim of this analysis is to identify relationships and the strength of the relationships among the attributes. This is necessary, so we can expect which predictor attributes may show possible interaction to be taken into account to our model. In this case, we are using Pearson Correlation Coefficient.

$$\gamma = \frac{\sum(x-\hat{x})\sum(y-\hat{y})}{\sqrt{\sum(x-\hat{x})^2\sum(y-\hat{y})^2}} \tag{1}$$

Equation 1: Pearson Correlation Coefficient.

Variable x and y represents the different attributes that are being compared. $\hat{x}$ $\hat{y}$ represents the mean of each attributes. The result is a range between -1 (which indicates negative correlation) and 1 (which indicates positive correlation).



	tip_amount	fare_amount	passenger_count	duration	trip_distance	daily change	Percipitation (in)
tip_amount	1.000000	0.703059	0.008791	0.093834	0.667577	-0.024169	-0.002239
fare_amount	0.703059	1.000000	0.003900	0.141109	0.865483	-0.018652	-0.001966
passenger_count	0.008791	0.003900	1.000000	0.015991	0.008929	-0.027695	-0.002604
duration	0.093834	0.141109	0.015991	1.000000	0.143123	-0.013134	0.000174
trip_distance	0.667577	0.865483	0.008929	0.143123	1.000000	-0.003605	-0.003345
daily change	-0.024169	-0.018652	-0.027695	-0.013134	-0.003605	1.000000	0.051336
Percipitation (in)	-0.002239	-0.001966	-0.002604	0.000174	-0.003345	0.051336	1.000000

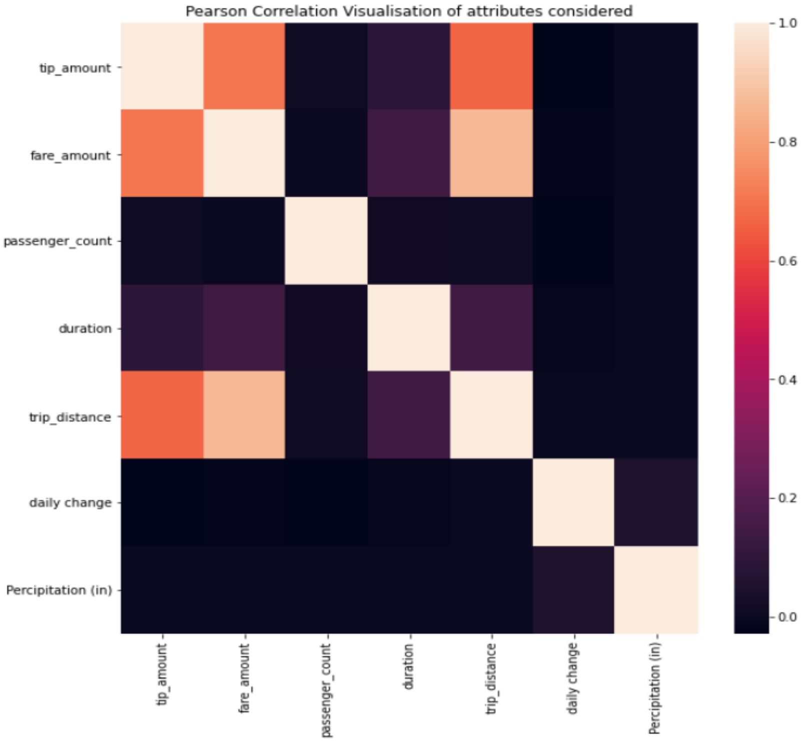


Figure 3: Pearson Correlation Table and heatmap.

These are the information we can take from this plot.

1. Fare amount(0.703059) and trip distance (0.667577) shows a strong correlation towards the response feature Tip amount. This may indicate the strong 10% tipping culture in America to be valid. Trip distance is a little more vague, but can be interpreted from a sentiment perspective where riders are more appreciative to drivers after a long ride.
2. Percipitation shows the lowest absolute correlation(0.002239) which may indicate insignificance of the attribute towards predicting tip amount.
3. Strong correlation observed from tip amount and trip distance (0.865483) as expected, which may indicate possible interaction between each other.
4. Duration shows positive correlation towards fare amount (0.141109) and duration (0.143123) as expected, again, may indicate interaction between predictor variables
5. Daily Change shows a relatively small correlation (-0.024169) , which may indicate insignificance in predicting tip amount.

Things to keep in mind when analyzing correlation is that correlation does not equal to causation. Our hypothesis obtained above is currently just a possible explanation. We can prove some of the hypothesis above by plotting the pairwise plot.

3.1.3 Pairwise Analysis

Pairwise comparison is a method of analyzing multiple attributes to see if they are significantly different from each other. One way to do it is to plot one attribute in one axis and the other in the next axis. We will show the pairwise comparisons of the attributes to be analysed.

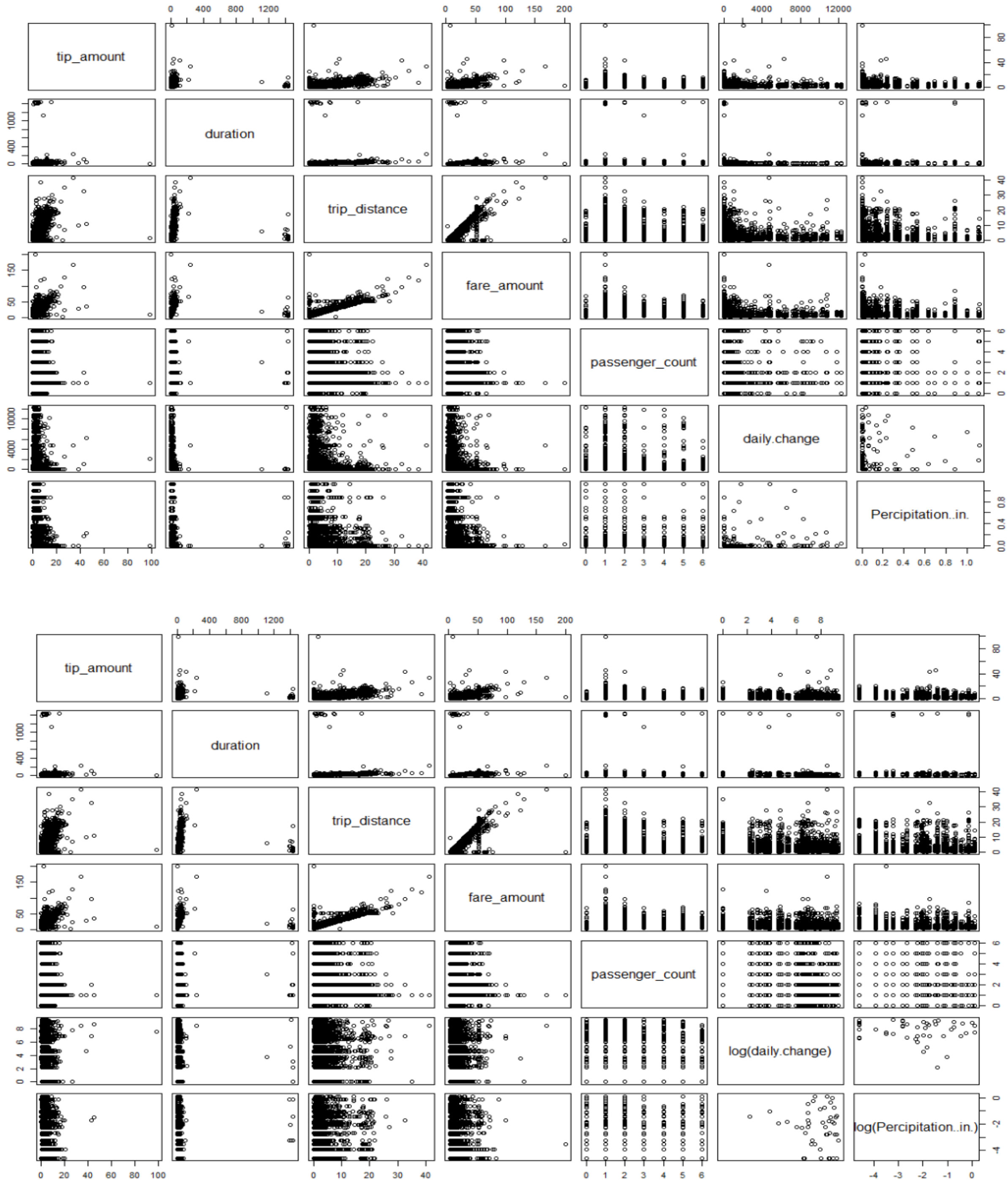


Figure 4: Pairwise plot of all attributes before and after log transformation on daily change and percipitation.

The difference between the first and second part of figure 4 is the log transformation on daily change and precipitation level. The reasoning behind this is that, after achieving the first part of figure 4, slight pattern is observed on daily change and precipitation towards other taxi data predictors. Our goal of introducing external data is to provide an independent set of variables that may have a potential effect on predicting the response. Therefore, log transformation is done with hope to have a more normalized data, which have less interaction to taxi predictors. As you can see in the second part , both log(daily change) and log (precipitation level) has less clear interaction towards other attributes.

These are the information we can take from the plot.

1. Trip distance and fare amount shows a clear linear relationship, the interaction between the variables might be worth considering.
2. Duration, although less clear due to the size of the data, shows a linear trends towards fare amount and trip distance as well. Indicating possible interaction.
3. Passenger counts seems to be a discrete data. We need to tune and re adjust our parameters before fitting the model to take account passenger count discreteness.
4. Linear trend observed from tip amount towards duration, fare amount and trip distance. Further enforcing our argument in the previous section.

As we have enough information regarding our predictor variables. We can start to model the relationship.

3.2 Models

Slight readjustment needed to be made as we need to take account passenger count attribute as discrete. This will be done in R with the factor function.

This section will explain step by step attempt on making a model and refining the model with several concepts and techniques. We will use R-squared (R2) and Root Mean Square Error (RMSE) as an evaluation metric for our models. RMSE is defined as the square root of the averaged squared difference between the actual value and the value predicted by the regression model. While R2 is defined as the coefficient of determination of how much better our data's minimal error compared to the baseline model. This metric is between 0-1. The model will be trained with 80% total data and test with 20% of the remaining.

3.2.1 Linear Regression Model

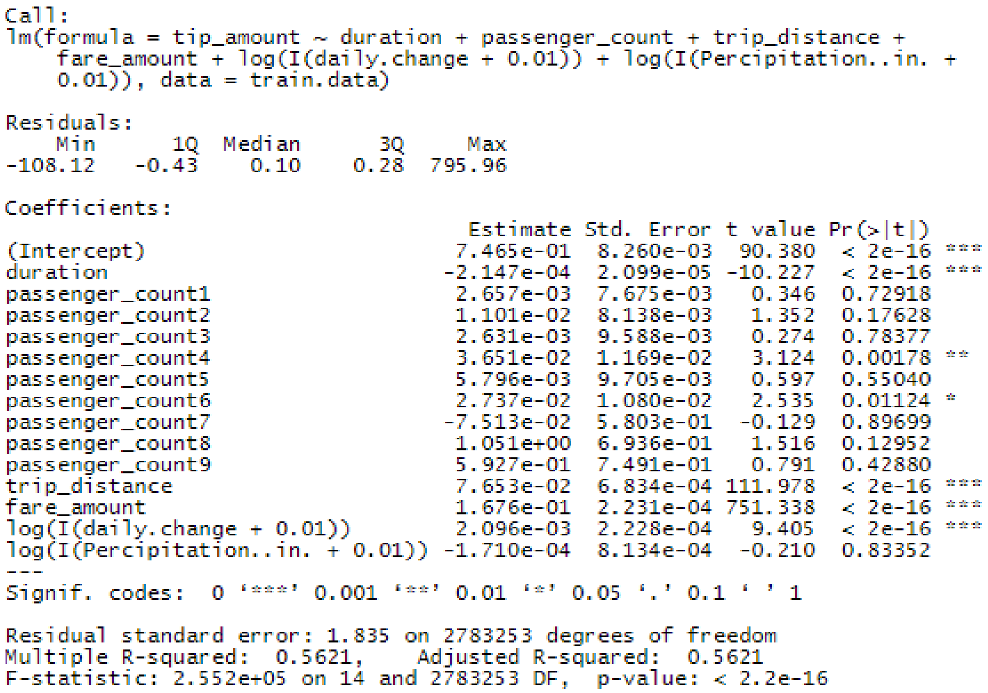
Firstly, we will consider a linear regression model to fit our data. Our model assumption is that by the large dimension of data and central limit theorem, we can assume a normal distribution on the dataset. And also due to the linearity observed from the pairwise plotted in figure 4.

$$y = X\beta + \epsilon \tag{2}$$

Equation 2: Linear Regression Equation.

Where y represents the response variable, X represents the explanatory variables, and $beta$ as the slope of each explanatory variable. While ϵ represents the error term. Assumption for linear model is that the error term has a normal distribution with mean 0 and co-variance matrix. If attributes follows a linear relationship, we should observe homoskedasticity fit.

This is an appropriate model as we can observe the response of tip amount based on the predictors variables of our choice. This will help us achieve our research goal on achieving reliable model to estimate profitability based on multiple factors.



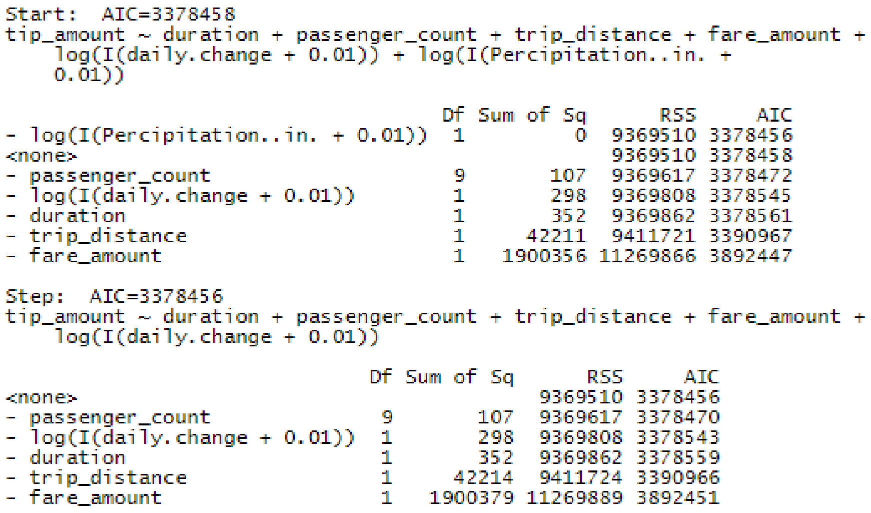
```
Call:
lm(formula = tip_amount ~ duration + passenger_count + trip_distance +
    fare_amount + log(I(daily.change + 0.01)) + log(I(Percipitation..in. +
    0.01)), data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-108.12   -0.43    0.10    0.28  795.96

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    7.465e-01  8.260e-03  90.380  < 2e-16 ***
duration                      -2.147e-04  2.099e-05 -10.227  < 2e-16 ***
passenger_count1               2.657e-03  7.675e-03   0.346  0.72918
passenger_count2               1.101e-02  8.138e-03   1.352  0.17628
passenger_count3               2.631e-03  9.588e-03   0.274  0.78377
passenger_count4               3.651e-02  1.169e-02   3.124  0.00178 **
passenger_count5               5.796e-03  9.705e-03   0.597  0.55040
passenger_count6               2.737e-02  1.080e-02   2.535  0.01124 *
passenger_count7              -7.513e-02  5.803e-01  -0.129  0.89699
passenger_count8               1.051e+00  6.936e-01   1.516  0.12952
passenger_count9               5.927e-01  7.491e-01   0.791  0.42880
trip_distance                  7.653e-02  6.834e-04 111.978  < 2e-16 ***
fare_amount                    1.676e-01  2.231e-04 751.338  < 2e-16 ***
log(I(daily.change + 0.01))    2.096e-03  2.228e-04   9.405  < 2e-16 ***
log(I(Percipitation..in. + 0.01)) -1.710e-04  8.134e-04  -0.210  0.83352
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.835 on 2783253 degrees of freedom
Multiple R-squared:  0.5621,    Adjusted R-squared:  0.5621
F-statistic: 2.552e+05 on 14 and 2783253 DF,  p-value: < 2.2e-16
```

Figure 5: Additive Model of All Attribute

The model shows how each attribute contribute to the estimation of tip amounts. The estimate of the coefficients represent on how the attribute affects tip amounts. The result may slightly differ from correlation analysis due to the model estimates coefficient by controlling other attributes, while correlation estimate only between response (tip amount) and that specific predictor. From figure 5, we can get some information.

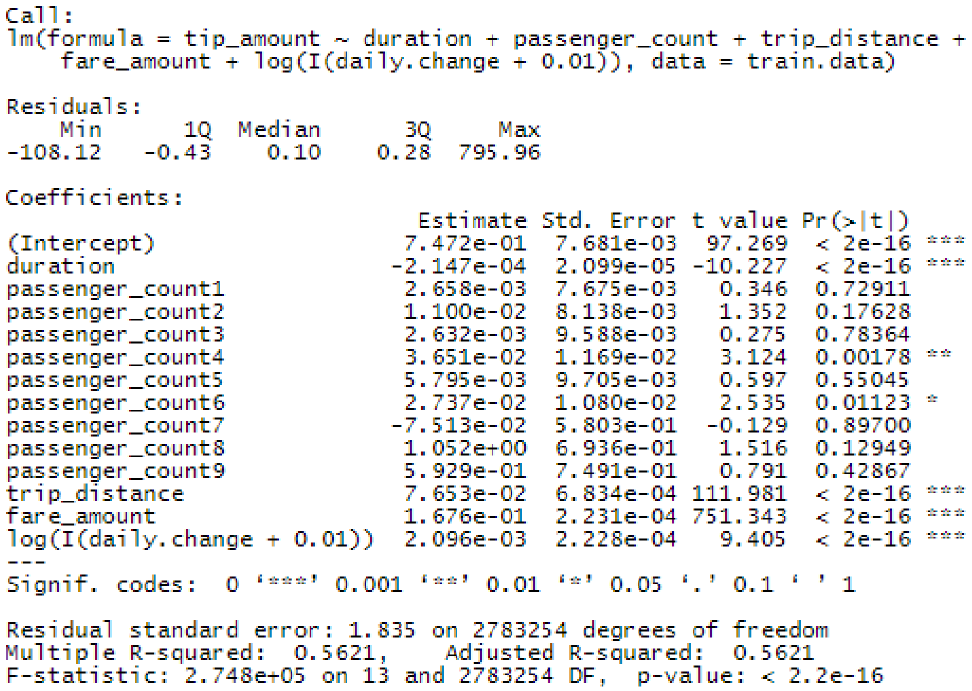
1. The p-value of the model is less than 0.05. Meaning that we are 0.95 confident that the linear model is a better fit than the null model. (A null model is an empty model fitted with 1 vector as predictors and tip amount as response).
2. 0.01 addition on log(daily change) and log(precipitation level) attributes is needed as both attributes consist of zeroes. Logarithmic of zero causes mathematical problem. So 0.01 is introduce as a salt to prevent mathematical calculation.
3. Passenger count 1-9 attribute is caused by the nature of passenger count being discrete.
4. The p value of the t distribution result of precipitation level is larger than 0.05. This is further evidence that the attribute is insignificant.

Although, the current model measures what we want. We have not achieved the most optimal number of predictor variables to achieve a "Parsimonious model" (Vandekerckhove, Matzke & Wagenmakers, 2015). We can refine this model by implementing step wise model selection with AIC as criterion.



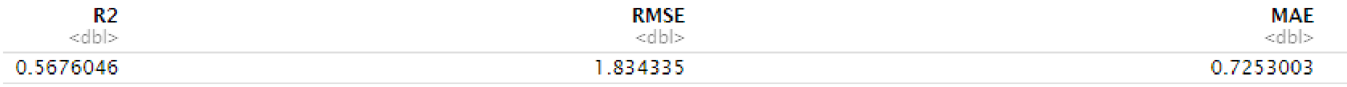
```
Start:  AIC=3378458
tip_amount ~ duration + passenger_count + trip_distance + fare_amount +
    log(I(daily.change + 0.01)) + log(I(Percipitation..in. +
    0.01))

                                   Df Sum of Sq      RSS     AIC
- log(I(Percipitation..in. + 0.01))  1         0  9369510 3378456
<none>                                           9369510 3378458
- passenger_count                   9       107  9369617 3378472
- log(I(daily.change + 0.01))       1       298  9369808 3378545
- duration                          1       352  9369862 3378561
- trip_distance                     1     42211  9411721 3390967
- fare_amount                       1   1900356 11269866 3892447

Step:  AIC=3378456
tip_amount ~ duration + passenger_count + trip_distance + fare_amount +
    log(I(daily.change + 0.01))

                             Df Sum of Sq      RSS     AIC
<none>                                     9369510 3378456
- passenger_count             9       107  9369617 3378470
- log(I(daily.change + 0.01)) 1       298  9369808 3378543
- duration                    1       352  9369862 3378559
- trip_distance               1     42214  9411724 3390966
- fare_amount                 1   1900379 11269889 3892451
```

Figure 6: Step Wise Feature Selection Based on AIC



```
Call:
lm(formula = tip_amount ~ duration + passenger_count + trip_distance +
    fare_amount + log(I(daily.change + 0.01)), data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-108.12   -0.43    0.10    0.28  795.96

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  7.472e-01  7.681e-03  97.269  < 2e-16 ***
duration                    -2.147e-04  2.099e-05 -10.227  < 2e-16 ***
passenger_count1             2.658e-03  7.675e-03   0.346  0.72911
passenger_count2             1.100e-02  8.138e-03   1.352  0.17628
passenger_count3             2.632e-03  9.588e-03   0.275  0.78364
passenger_count4             3.651e-02  1.169e-02   3.124  0.00178 **
passenger_count5             5.795e-03  9.705e-03   0.597  0.55045
passenger_count6             2.737e-02  1.080e-02   2.535  0.01123 *
passenger_count7            -7.513e-02  5.803e-01  -0.129  0.89700
passenger_count8             1.052e+00  6.936e-01   1.516  0.12949
passenger_count9             5.929e-01  7.491e-01   0.791  0.42867
trip_distance                7.653e-02  6.834e-04 111.981  < 2e-16 ***
fare_amount                  1.676e-01  2.231e-04 751.343  < 2e-16 ***
log(I(daily.change + 0.01))  2.096e-03  2.228e-04   9.405  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.835 on 2783254 degrees of freedom
Multiple R-squared:  0.5621,	Adjusted R-squared:  0.5621
F-statistic: 2.748e+05 on 13 and 2783254 DF,  p-value: < 2.2e-16
```

Figure 7: Model After AIC



R2	RMSE	MAE
<dbl>	<dbl>	<dbl>
0.5676046	1.834335	0.7253003

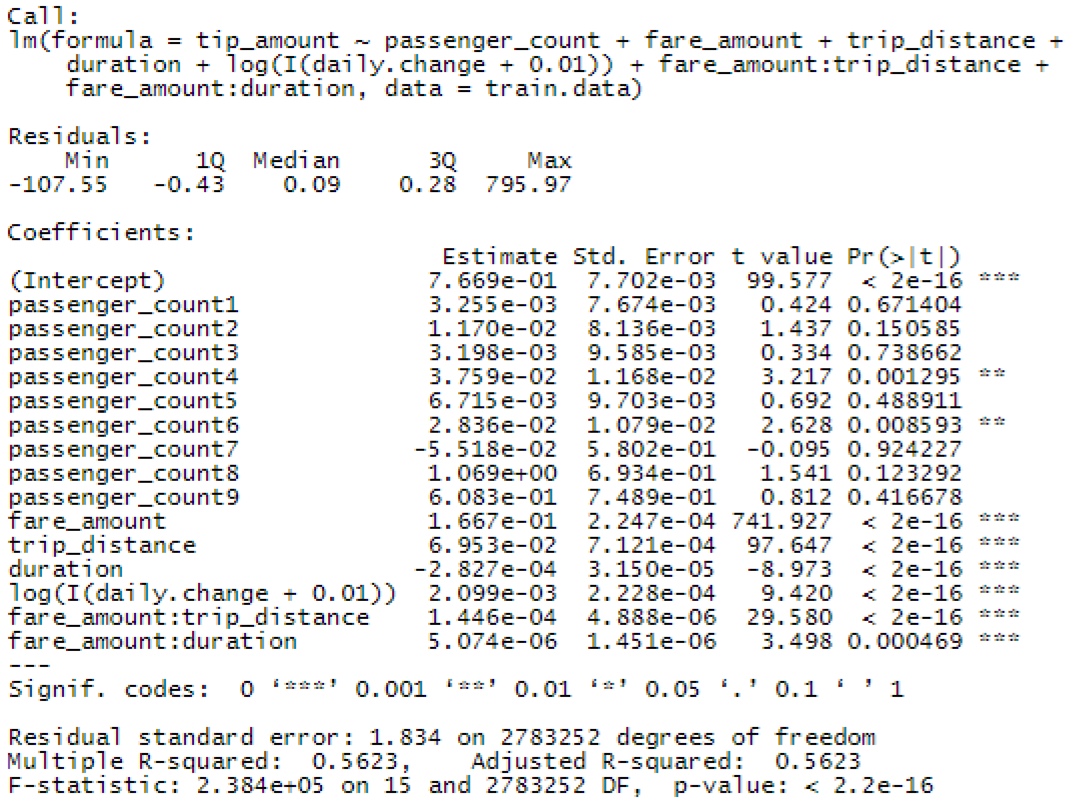
Figure 8: Result of Model Predicting test data

The AIC feature selection method has decided to remove Precipitation level as it is deemed to be insignificant which we hypothesis in previous sections. Also as we achieve the same R-square for our adjusted model further solidify it. Our model results in an R2 of 0.5676046 and RMSE of 1.834335 on the test set. Although in the world of science, an R2 of 0.5 and a relatively small RMSE is generally considered acceptable. We want to explore different models and tuning of parameters to achieve a higher R2 and lower RMSE. We can consider fitting the interaction to the model of predictor attributes to the variable.

3.2.2 Interaction Model

From the pairwise plot we consider several predictor attributes that may have strong interaction between each other. Thus, we will fit the interaction of fare amount towards duration and trip distance.

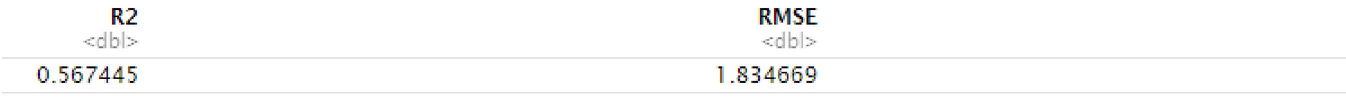
First we fit all the predictors attributes with addition of the interaction to be considered. We then run step wise AIC feature selection again to see if there is a change in relevancy of attributes.



```
Call:
lm(formula = tip_amount ~ passenger_count + fare_amount + trip_distance +
    duration + log(I(daily.change + 0.01)) + fare_amount:trip_distance +
    fare_amount:duration, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-107.55   -0.43    0.09    0.28  795.97 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  7.669e-01  7.702e-03  99.577  < 2e-16 ***
passenger_count1             3.255e-03  7.674e-03   0.424 0.671404    
passenger_count2             1.170e-02  8.136e-03   1.437 0.150585    
passenger_count3             3.198e-03  9.585e-03   0.334 0.738662    
passenger_count4             3.759e-02  1.168e-02   3.217 0.001295 ** 
passenger_count5             6.715e-03  9.703e-03   0.692 0.488911    
passenger_count6             2.836e-02  1.079e-02   2.628 0.008593 ** 
passenger_count7            -5.518e-02  5.802e-01  -0.095 0.924227    
passenger_count8             1.069e+00  6.934e-01   1.541 0.123292    
passenger_count9             6.083e-01  7.489e-01   0.812 0.416678    
fare_amount                  1.667e-01  2.247e-04 741.927  < 2e-16 ***
trip_distance                6.953e-02  7.121e-04  97.647  < 2e-16 ***
duration                    -2.827e-04  3.150e-05  -8.973  < 2e-16 ***
log(I(daily.change + 0.01))  2.099e-03  2.228e-04   9.420  < 2e-16 ***
fare_amount:trip_distance    1.446e-04  4.888e-06  29.580  < 2e-16 ***
fare_amount:duration         5.074e-06  1.451e-06   3.498 0.000469 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.834 on 2783252 degrees of freedom
Multiple R-squared:  0.5623,	Adjusted R-squared:  0.5623 
F-statistic: 2.384e+05 on 15 and 2783252 DF,  p-value: < 2.2e-16
```

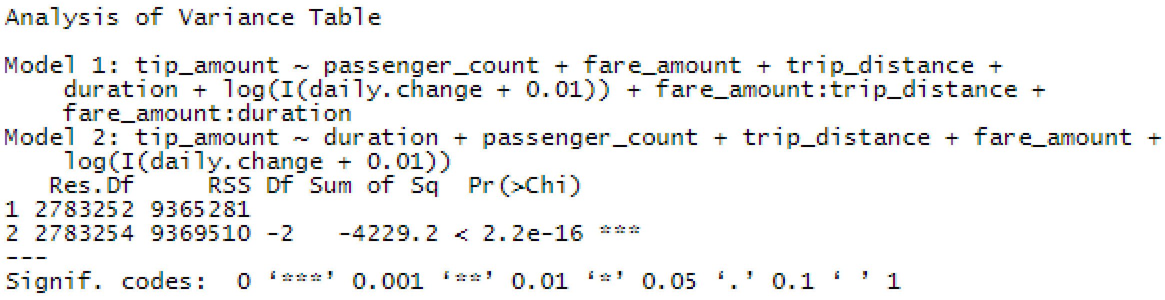
Figure 9: Final Model of Interaction

This model again shows that Precipitation Level is insignificant to predict tip amount as it is removed after step-wise AIC method. As the interaction is not removed from the model (the last two coefficient in the final model), shows the interaction is relevant. We can enforce this by running a likelihood ratio test with the additive model.



R2 <dbl>	RMSE <dbl>
0.567445	1.834669

Figure 10: Result of Model predicting test data



```
Analysis of Variance Table

Model 1: tip_amount ~ passenger_count + fare_amount + trip_distance +
    duration + log(I(daily.change + 0.01)) + fare_amount:trip_distance +
    fare_amount:duration
Model 2: tip_amount ~ duration + passenger_count + trip_distance + fare_amount +
    log(I(daily.change + 0.01))
   Res.Df     RSS Df Sum of Sq  Pr(>Chi)
1 2783252 9365281
2 2783254 9369510 -2   -4229.2 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11: Likelihood ratio test between Additive and Interaction Model

Since the Likelihood ratio test indicates that the P-value (bottom right) is less than 0.05, it is concluded that fitting interaction indicates a significant difference in model. However, we observe in figure 10 that we have a slight decrease in R2 compared to the previous model. This may be caused by the nature of the test being only one random hold out of train test split. Train data may not represent the total dataset. Especially due to random sampling, this issue is a possibility. One other possibility is that as we deal with large amount of data, it is more proned to errors and outliers that are not screened in the pre-processing steps. We can observe the diagnostic plot to conclude this hypothesis.

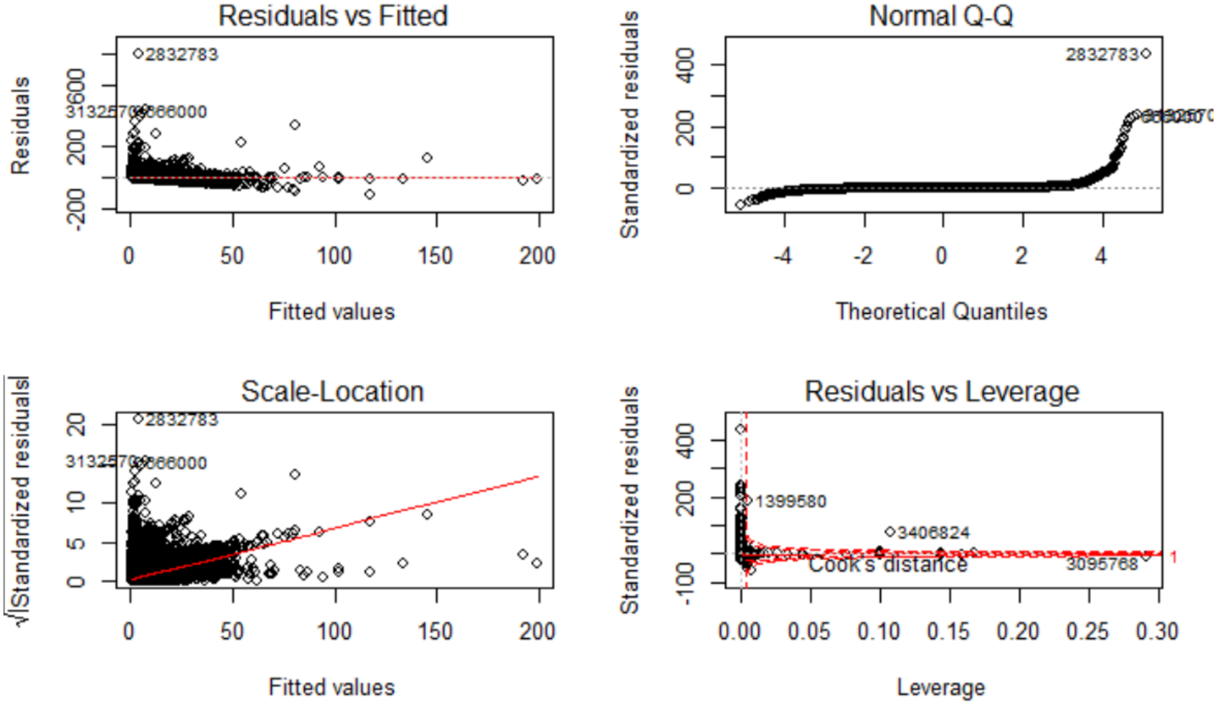


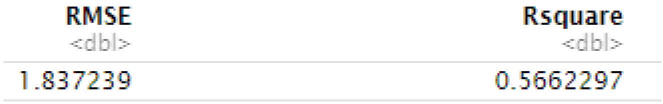
Figure 12: Diagnostic Plot of Interaction Model

An ideal diagnostic plot for residual vs fitted is that it is randomly dispersed with no pattern. However, we observe data that are clustered to the left. This indicates the existence of outliers that is skewing the dataset. Further explained by the normal QQ plot showing heavy skew in the right tail, as supposedly we observe a good homoscedasticity pattern. Standadized residual also shows left clusters with indication of linear best fit line with positive slope. Residuals and Leverage plots seems to be good overall with a cook distance less than 0.5. We can observe several outliers points however as only few outliers go beyond the cluster.

Another possibility, as an increase in predictor leads to a less generalize model. To many predictors are trying to predict tip amount which leads to a lower R2. We can explore ways to generalize the model by exploring ridge regression.

3.2.3 Ridge Regression Model

Ridge Regression Model is a model that implements L2 Regularisation techniques on Linear Model. It works by adding L2 penalty which limits the size of the coefficient by square root of the magnitude of the coefficient. We implemented this model by first fitting the model with training data of all predictor attributes and sequences of lambda. Lambda is a hyper-parameter that controls the strength of the penalty term. We then tune the hyper-parameter of the learner by calculating the most optimal Lambda by Cross-Validation. Then finally fitting the model with the train data with the optimal hyper parameter. This is the result of the model.



RMSE <dbl>	Rsquare <dbl>
1.837239	0.5662297

Figure 13: Ridge Regression Model Result.

The results shows a slightly lower R2 than the previous models with a higher RMSE. This indicates a worse performance. This performance can be caused by the penalty on predictor variables decreases the predictability of the model. As regularisation techniques are to solve over fitting of data. If the data is not overfitted to begin with, it is possible that it decreases performance.

4 Discussion

1. Although generalizing techniques is a good strategy overall to increase model performance, in this case it is not appropriate as we are dealing with large and unrefined dataset that has not the best performance.
2. The evaluation result of all the models indicates the difficulty of predicting tip amounts with just the attributes named. In fact, shows the difficulty of predicting tip amount overall as there is a qualitative side of tipping culture as much or even more than there is the quantitative side. Meaning that there are factors that are unexplored.
3. The sample size of the data , although still to be considered huge, is not exactly representative to the whole dataset. Which may be a possibility of weak performance issue.
4. Using testing procedure of 1 random sub sample is not an ideal test methodology as test may not reflect the actual representation of the data. A usage of more than 1 or implementation of cross validation is ideal.
5. The Diagnostic plot of the interaction model seems to show the effect of outliers affecting the data. This is a hard to overcome as removing too much of an outliers may not show a good representation of the population. And specific rules that needed to be explored to know which data is outliers. E.g. Is fare amount of $1000 possible.
6. The increase of complexity of the interaction model may result in the slight decrease of performance, as the model gets more complex than more general.
7. All evaluation of model only shows a small decrease in performance, which is an issue as we don't know how significant is the decrease. Further methods needs to be explored in future studies to statistically analyze the performance.
8. Although likelihood ratio test shown a significant result in adding interaction to the model. We cannot conclude that the impact is negative or positive due to the limitations that we have.

5 Conclusion

In conclusion, the final model metric to be fitted will be measured by the model that gives the highest performance. Since the refinement leads to a lower performance. The additive models will be final model suggested as it provides the highest R2 with the smallest RMSE.

$$\begin{aligned} tipamount &= 0.0747 - 0.0002Duration + coef_i passenger_i \\ &+ 0.00765tripdistance + 0.0167fareamount \\ &+ 0.0021log(I(daily.change + 0.01)) \end{aligned}$$

Equation 3: The final model equation with it's coefficient.

Where $coef_i$ is a variable that gets the coefficient of number of passenger count in the taxi.

Using the equation above, drivers can approximately regress expected amount of tips per given ride every day. E.g. given that a particular trip has 5 passengers , last around 30 minutes with a final fare of $20 and trip distance of 10 miles with the covid-19 case today having 119 increase. By plugging the values to the formula above, we can regress that the amount of tips to be expected are approximately atleast $0.5.

Stake holders can use the model to regress the amount of tips received for each trip or have an approximation of the impact of the factors based on the coefficient generated by the linear regression. The coefficient generated is based on optimization method, which produces the best estimate possible for that given model. In summary, predicting factors that affects tips is no easy task as there are many factors that can affect this. This shows the importance generating good features to regress your model well. Regularisation and adding interaction to the model in most good cases increase model performance. Improvements from Andrew's previous report is that we implemented descriptive analysis and implemented different models to indicate relationship of each attribute and performance.

Improvements needed to be made are using a higher computational power to observe all the dataset and even larger scope. Use a different test procedure to achieve a balance evaluation. And pre-define limits to the outliers as it is significantly affecting the model.

References

[1] Abraham, Zina, "The Influence of National Culture on Tipping Behavior" (2014). UNLV Theses, Dissertations, Professional Papers, and Capstones. 2626. https://digitalscholarship.unlv.edu/thesesdissertations/2626

[2] Helier C, "Coronavirus and tipping: Will the outbreak make us more generous?" (2020) https://www.bbc.com/news/world-us-canada-52627734

[3] Andrew T, "New York City Yellow Taxi Data Analysis" (2020)

[4] Marina Z, "How To Tip In America (According To An American)" (2018) https://www.theurbanlist.com/melbourne/a-list/how-to-tip-in-america

[5] Chris F, "Timeline: The first 100 days of New York Gov. Andrew Cuomo's COVID-19 response" (2020) https://abcnews.go.com/US/News/timeline-100-days-york-gov-andrew-cuomos-covid/story?id=71292880

[6] Vandekerckhove, J., Matzke, D., & Wagenmakers, E. J. "Model comparison and the principle of parsimony." (2015) Oxford handbook of computational and mathematical psychology.