Qualitative Analysis Of NYC Yellow Taxi's

Optimizing Tip Amounts

Dion Doanh Tran
Student ID: 994765
The University of Melbourne
MAST30034 – Applied Data Science
Tutorial: (Hossein Alipour, Thursday 3:15pm – 5:15pm)

Abstract

All around the globe, tipping has become such a social norm in to where it is constantly expected. Tipping, considered by many, an act of kind gesture and free will has now become such an important part of daily life to the American society. Today we will be looking at one of America's most iconic cities and their mascot Yellow Taxis, New York. With coverage spanning 15 million trips in the first 3 months of 2020, the data provided by the Taxi & Limousine Commission (TLC) will aim to help see if there are any affects that go into the odds of tipping or its amount. From the previous exploration with the help of visualization, we found a surprisingly low percentage of rides were not tipped, a staggering '~3%' in fact and that when they did tip, people were more likely to tip more when their pickup or drop-off location were in close vicinity of a cemetery or costly recreational hobbies. With those areas varying between zoos, golf parks and hiking areas. Time was also an aspect that showed consistent but small results. This report aims to see if there are any or none specific features that impact tip rate and tip amount. Many techniques will be used to explore the mystery surrounding tipping.

We believe that any knowledge on tip amounts will benefit taxi drivers in the future even though tip amounts are not in their control.

1 Looking back at Phase 1

The initial phase used visualization techniques to map out different areas of pickup and drop-off locations along with their individual average tip amounts. We realized slight patterns in higher average tip amounts when taxi rides had started or ended in areas with a large density of cemeteries. However, due to the lack of statistical tests, there wasn't any clear evidence in whether such were the case because there was a lack of trips occurring in such areas as well. Other notable findings included which hour or day were people more likely to tip a larger amount. We found that through all three months, Thursday had the higher average with late nights also recurring more tip amounts. Notable features were before early mornings, 4-7am, the average tip amount was at its lowest, and thus, though people may be more appreciative of taxi drivers working such hours, it will not necessarily mean higher tip average.

1.1 General January Visualization

Figure 1 shows the Pick Up Locations and their average tip amounts during the first month of 2020. With closer inspection. Areas around the airport near to top left are gaining traction. The areas in darker shade of green towards the centre of NYC contain many cemetaries with some location ID's housing up to 5 cemetaries. Figure 2 and Figure 3 are show trends mentioned above.

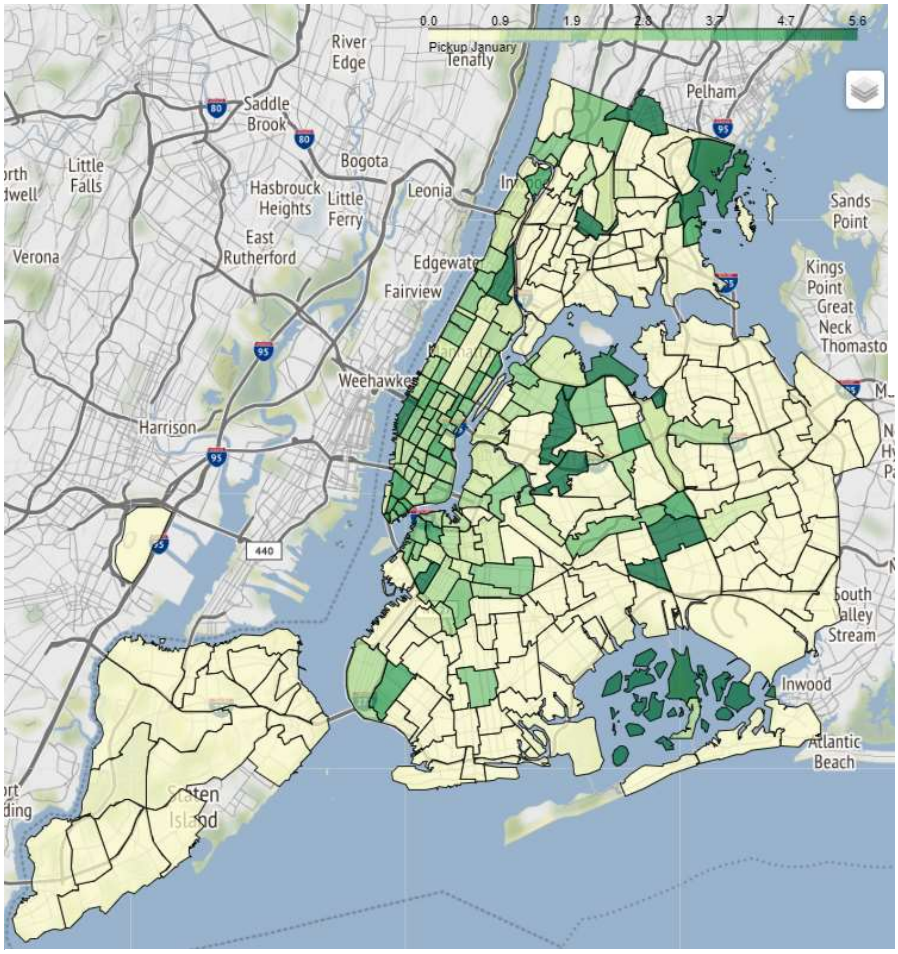

Figure 1: Average Tip Amount in each Pick Up Location ID for Yellow Taxi rides in January 2020

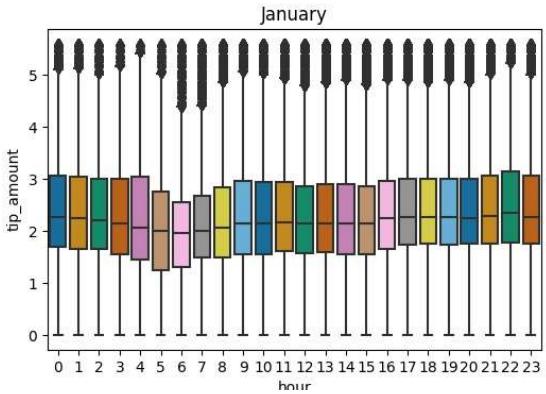

Figure 2: Average Hourly Tip Amount in January 2020

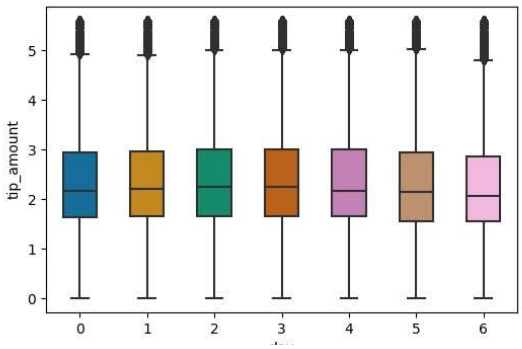

Figure 3: Average Tip Amount per Day in January.

2 Data Cleaning

Raw data stayed consistent with the data dictionary provided by the TLC. The data was cleaned similarly to phase 1 however all three months were merged into one large dataset. The dataset was then cleaned together ultimately reaching almost 1.15 million instances remained.

1. Invalid Data points were dropped. These included trips with any N/A values.

2. As tip amount were only recorded with payment made by credit card, payment_type = 1, other instances were removed.

3. There were heavy outliers that were removed from data points where tip amount were greater than $5.8 and with trip distances that were greater than 4km. Data was extremely left skewed.

4. Date-Time attributes were used to create two separate attributes, day of the week and hour of the day.

5. As there is a minimum of $2.50 hailing fee, instances with total amount being less then that were dropped as well.

3 Attribute Analysis

3.1 Initial Manual Feature Selection

- All Features include: VendorID, PickUp Date/Time, DropOff Date/Time, Passenger Count, Trip Distance, PickUp Location ID, DropOff Location ID, RateCode ID, Store and Forward Flag, Payment_Type, Fare Amount, Extra, MTA Tax, Improvement Surchage, Tolls Amount, Total Amount.

- Features that correlate: Fare Amount, Tolls Amount, Total Amount, MTA Tax. Will be represented by Fare Amount.

- Features that Taxi Drivers can control, and such deemed more useful: PickUp Location ID, PickUp Date/Time

- Features that Taxi Drivers cannot control yet maybe important: Trip Distance and Passenger Count.

3.2 Initial Skewness

Skewness is the measure of asymmetry in distributions and is a not a favorable trait when working with large datasets. The skewness shows in this dataset is very negative where a majority of the rides tip amounts, trip distance, fare amount is closer to 0. Thus, it can indicate there are many significant outliers in the data and that there are much more higher counts of shorter trips then long ones.

```
VendorID                 -0.721745
passenger_count           2.441252
trip_distance             3.826656
RatecodeID              100.683703
PULocationID             -0.294898
DOLocationID             -0.344003
payment_type              1.629602
fare_amount            3916.413441
extra                  3944.027962
mta_tax                3944.601097
tip_amount               16.260268
tolls_amount             69.938405
improvement_surcharge   -16.655201
total_amount           2931.103197
congestion_surcharge      -3.398369
dtype: float64
```

Figure 4: Initial Attribute Skew before Data Cleaning

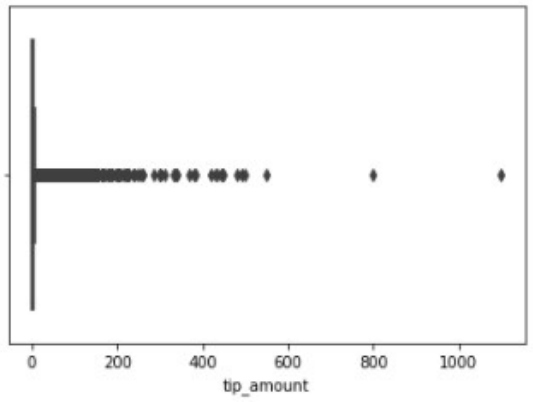

Figure 5: A boxplot of tip amount before Data Cleaning

3.3 Initial Attribute Distributions

Figure 4 and Figure 5, show large negative skew with fare amount being 3916.4, total amount approximating 2931, with tip amount reaching 16.

Table 1: Data description before removing outliers.

	tip_amount	fare_amount	trip_distance	passenger_count	PULocationID	DOLocationID
count	1.15135e+07	1.15135e+07	1.15135e+07	1.15135e+07	1.15135e+07	1.15135e+07
mean	2.94875	12.4659	2.86919	1.49494	166.058	163.956
std	2.66754	10.936	3.68284	1.1413	65.2085	69.3545
min	0	2.5	0.01	0	1	1
25%	1.75	6.5	1	1	132	113
50%	2.32	9	1.65	1	162	162
75%	3.26	13.5	2.91	2	234	234
max	1100	804	310.9	9	265	265

Table 2: Data description after removing outliers.

	tip_amount	fare_amount	trip_distance	passenger_count	PULocationID	DOLocationID
count	9.34697e+06	9.34697e+06	9.34697e+06	9.34697e+06	9.34697e+06	9.34697e+06
mean	2.19875	8.58946	1.55005	1.49199	168.947	167.697
std	0.945023	3.25139	0.814341	1.14059	65.189	66.9445
min	0	2.5	0.01	0	1	1
25%	1.58	6	0.9	1	137	125
50%	2.1	8	1.4	1	163	163
75%	2.76	10.5	2.07	2	234	234
max	5.88	21.5	3.72	9	265	265

From comparing Table 1 and Table 2, Removing outliers though IQR has decreased max tip amount from $1100 to $5.8. $671100 to $19.5 for fare amount, with trip distance lowering from 369.94km to 3km. This will remove any large one off tip amounts from the data.

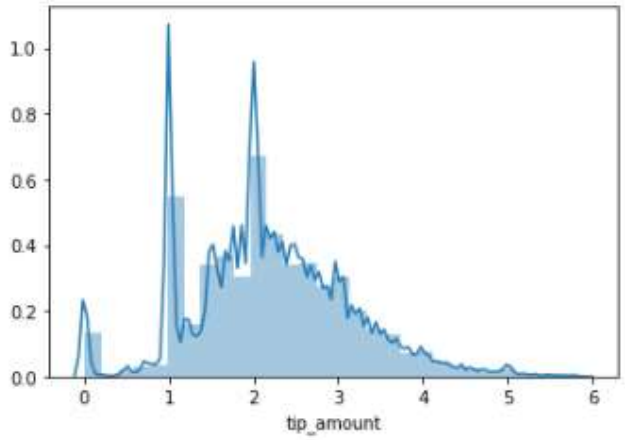

Figure 6: Distribution of Tip Amounts

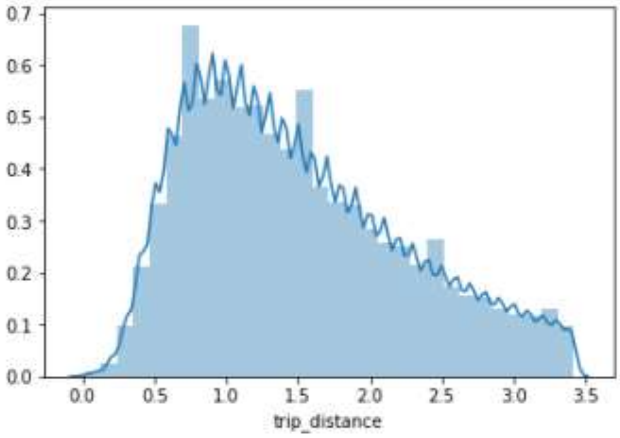

Figure 7: Distribution of Trip Distance

Figure 6 and Figure 7 now show slight negative skew, it is now much more plausible under realistic situations. In Figure 6, there are three peaks at no tip, $1 and $2 tips which is also realistic. Have a cut off at trip distance at 3km seems a bit much.

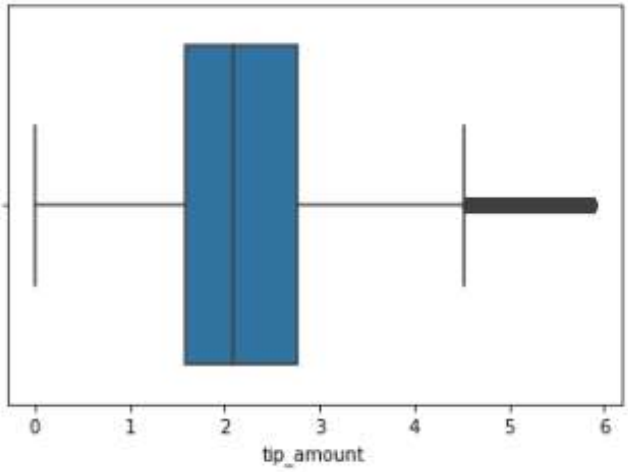

Figure 8: Boxplot of Tip Amount after Pre-processing

Figure 8 shows Tip Amount after preprocessing and when compared to Figure 5, is much easier to read.

Overall a total of 1.15 million instances were removed from the total dataset which contains rides between January to March 2020. Which means we have lost approximately 20% of our data to outliers across the data frame which is much more usual.

This information will be kept in mind further during the analysis.

3.5 Pearson Correlation Statistic

The person correlation is used to measure interrelationship linear correlation between two attributes. The statistic is sensitive to outliers, and due to the large removal of many outliers, statistic may not be able to represent the entire dataset.

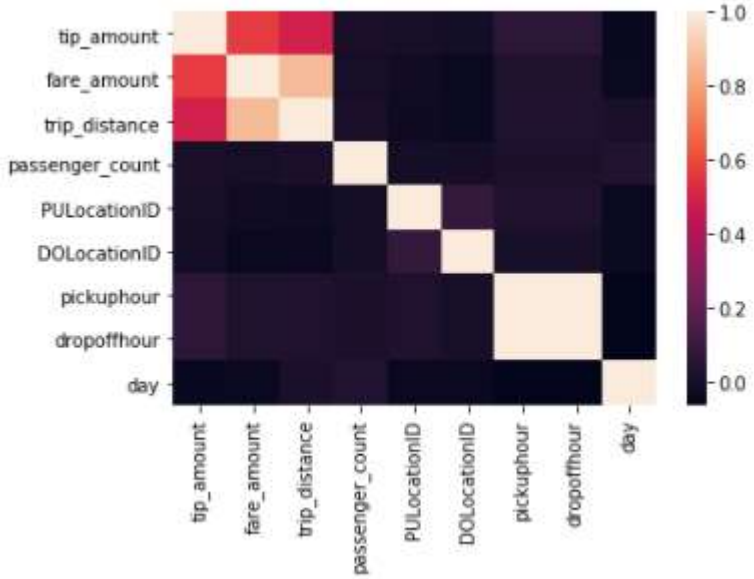

Figure 9: Correlation Heatmap of Manually chosen Features

- Figure 9 reveal that tip amount highly correlates with 'fare_amount' and 'trip_distance'.
- There is little to no correlation between 'tip_amount' and all other attributes.
- As 'trip_distance' is used to calculate 'fare_amount' there is innate correlation between the two attributes, thus may seem redundant but will be used for analysis to predict the odds of tipping.

3.6 Exploring Correlation between Tip Amount vs Fare Amount and Trip Distance

High correlation inspired some quick look at the scatter plots between to see how consistent the correlation is and how far the spread is.

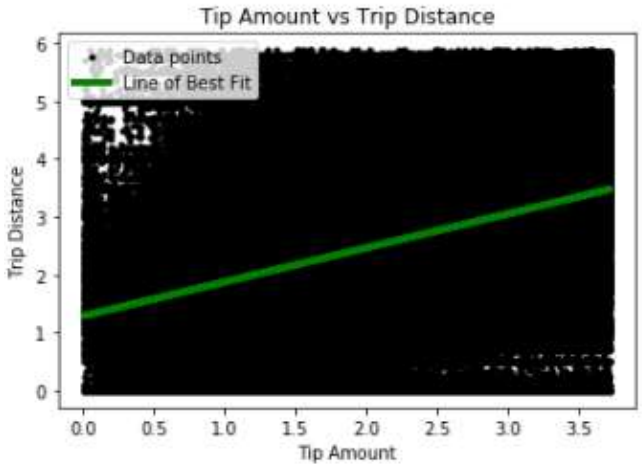

Figure 10: Scatter plot of Tip Amount vs Trip Distance

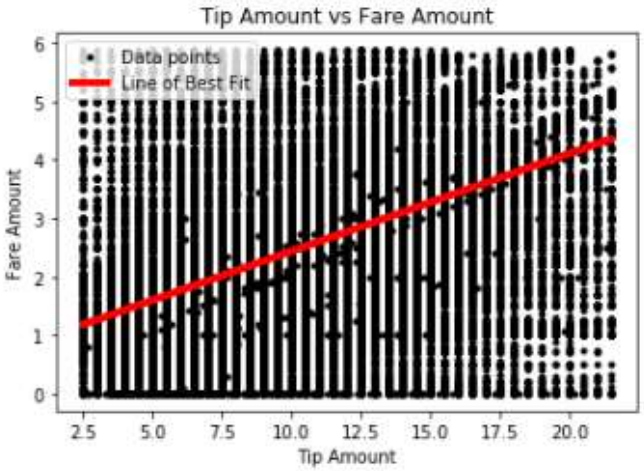

Figure 11: Scatter plot of Tip Amount vs Fare Amount

Figure 10 and Figure 11 show a large spread of data points across the scatter plot and though it may be difficult to see visually, the line of best fit show that there is high variance when comparing between Tip Amount vs Trip distance and Trip Amount vs Fare Amount however not substantial.

4 Evaluation Metrics

4.1 Mean Absolute Error

Average difference between the predicted value and the tested values. Interpreted by comparing to the scale of our variables.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{\theta}_i - \theta_i\right| \quad (1)$$

4.2 Mean Squared Error

The average of the squares of the errors.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2 \quad (2)$$

4.3 Root Mean Squared Error

The square root of MSE. Known as the standard error

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2} \quad (3)$$

4.4 MAE vs RMSE

MAE and RMSE express average model prediction error through units of the certain variable and such both metrics can range from 0 to infinite. For both, the lower the values, the 'better' the model is. However, since RMSE takes the square root of the average squared errors, it has the benefit of penalizing larger errors. This may make RMSE a desirable statistic due to the high variance already seen throughout the data.

5 Predictive Models

As Tip Amount's range is limited from \$0 to \$2.9, Figure 10 and Figure 11 showed large spread between datapoints, and thus fitting a general linear model will not be the strongest model. A regression tree model will also be considered and implemented.

5.2 Linear Regression Models (LR)

5.2.1 Fitting a Model - Backward elimination

Backward elimination starts with all independent variables in the equation with every step a variable is deleted one at a time if it is not able to contribute to the regression equation. Backward elimination was chosen over forward selection and stepwise selection as with the given computational power, backward elimination should help compromise accuracy for lesser time complexity.

OLS Regression Results

Dep. Variable:	tip_amount	R-squared:	0.336
Model:	OLS	Adj. R-squared:	0.336
Method:	Least Squares	F-statistic:	6.769e+05
Date:	Fri, 09 Oct 2020	Prob (F-statistic):	0.00
Time:	06:38:29	Log-Likelihood:	-1.0818e+07
No. Observations:	9346974	AIC:	2.164e+07
Df Residuals:	9346966	BIC:	2.164e+07
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.6378	0.001	446.534	0.000	0.635	0.641
fare_amount	0.1660	0.000	1020.931	0.000	0.166	0.166
trip_distance	0.0068	0.001	10.511	0.000	0.006	0.008
pickuphour	0.0078	4.39e-05	177.761	0.000	0.008	0.008
day	-0.0089	0.000	-65.910	0.000	-0.009	-0.009
passenger_count	0.0060	0.000	26.982	0.000	0.006	0.006
PULocationID	8.825e-05	3.88e-06	22.738	0.000	8.06e-05	9.59e-05
DOLocationID	0.0001	3.78e-06	28.331	0.000	9.97e-05	0.000

Omnibus:	1166417.586	Durbin-Watson:	1.979
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2969999.534
Skew:	-0.715	Prob(JB):	0.00
Kurtosis:	5.363	Cond. No.	1.42e+03

Table 3: OLS Regression Results with All independent variables in the equation.

Table 3 shows a low R-squared value along with an average high AIC/BIC. However as the p-values of each variable are all < 0.005, no variable will be removed and the model will be built upon all features.

5.2.2 Assumptions

1. The predictors are accurately observed.
2. Linear relationship between the response and predictors
3. Independent errors and constant variance
4. Full rank of X: n>p and the $\{xi\}_{i=1}^{n}$ are not linear combinations of each other

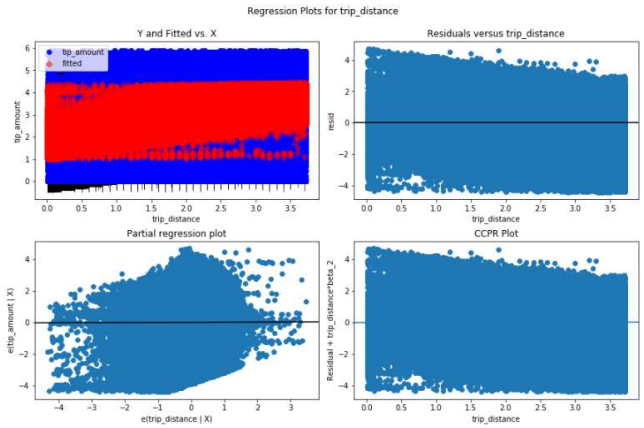

Figure 12: Regression Plots for Trip Distance

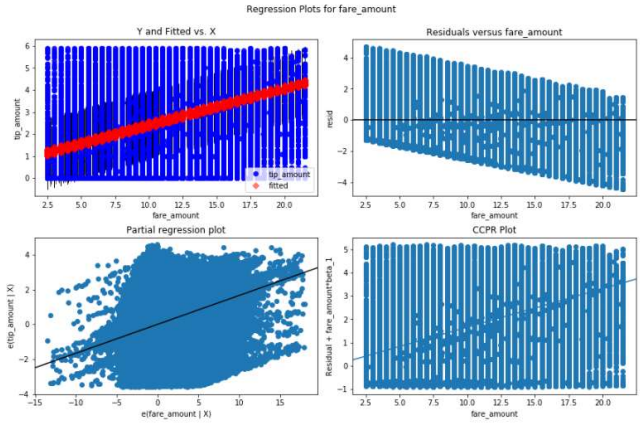

Figure 13: Regression plots for Fare Amount

Residual plots show linear relationship and constant variance.

5.2.3 Predictive Regression Model

OLS Regression Results

Dep. Variable:	tip_amount	R-squared (uncentered):	0.797
Model:	OLS	Adj. R-squared (uncentered):	0.797
Method:	Least Squares	F-statistic:	4.387e+06
Date:	Fri, 09 Oct 2020	Prob (F-statistic):	0.00
Time:	08:20:03	Log-Likelihood:	-1.2686e+07
No. Observations:	7836017	AIC:	2.537e+07
Df Residuals:	7836010	BIC:	2.537e+07
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
fare_amount	0.1717	0.000	1282.097	0.000	0.171	0.172
trip_distance	0.0482	0.001	73.991	0.000	0.047	0.049
pickuphour	0.0172	6.74e-05	254.632	0.000	0.017	0.017
day	0.0145	0.000	66.571	0.000	0.014	0.015
passenger_count	0.0310	0.000	83.172	0.000	0.030	0.032
PULocationID	0.0009	5.95e-06	158.449	0.000	0.001	0.001
DOLocationID	0.0009	5.8e-06	158.867	0.000	0.001	0.001

Omnibus:	28596819.689	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	152018311815577.062
Skew:	75.940	Prob(JB):	0.00
Kurtosis:	21580.202	Cond. No.	374.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Table 4: Predictive Regression Model with all variable.

5.2 One Rule (OR)

One Rule model is Regression Tree which has a maximum depth of one. It is a simple baseline used to compare with other models. It is expected that OR will get the highest average error in its prediction.

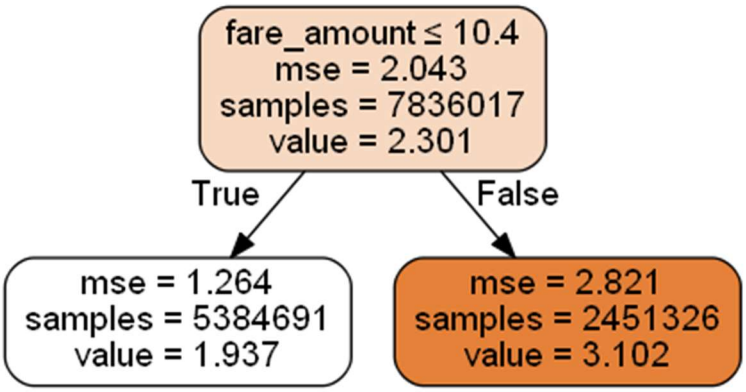

Figure 14: One Rule Tree

5.3 Regression Tree (RT)

A decision tree used for regression, is similar to a normal decision tree however its evaluation metrics will differ from trees build for classification. It goes through many binary recursive partitioning which splits the data into many different branches, ending it to the leaf. This will be evaluated similar to LR.

5.3.1 Regression Tree with a maximum depth of 4

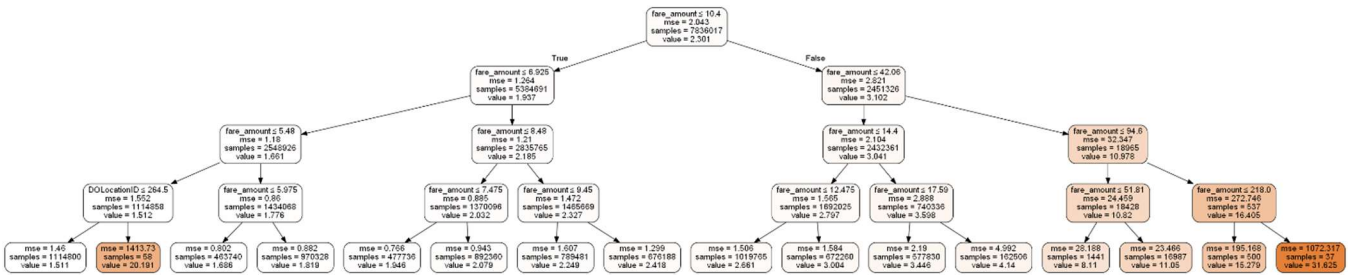

Figure 15: RT4

The tree in Figure 14 show skewness similar to the trends found previously in the total data, as the color is mainly on the right side, except for the single node on the left.

5.3.2 Regression Tree with a maximum depth of 6

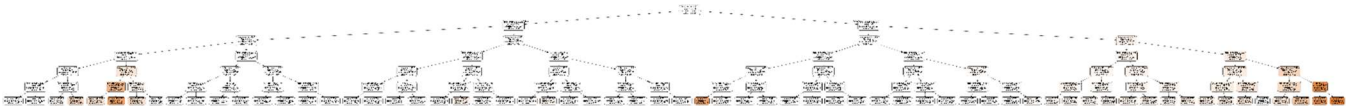

Figure 16: RT6

Figure 16 also show a similar color range to Figure 15, with similar trend in patterns, however the leaf directly under the root is also standing out. The different regression trees show not many irregularities as the colors stayed consistent shades.

6 Model Error Analysis

Training and testing will be split into 80:20 ratio. All models were trained with the same training set and tested against the same testing set.

Model	MAE	MSE	RMSE	Runtime
LR	0.605629	2.237167	1.495716	4.817040
RT4	0.598311	2.223338	1.491086	17.348381
RT10	0.593017	2.287643	1.512495	38.741930
OR	0.684182	2.494061	1.579260	4.813544

Table 4: Model Evaluation

Table 4 shows the Mean Absolute Error (MAE), Mean Square Error (MSE) and the Root Mean Square Error (RMSE) of the four models shown today. LR represents Linear Regression, RT4 represents the regression tree with a maximum depth of 4 with RT10 having a maximum depth of 10. OR is the baseline model that the other models will compare itself towards. The table also reveals very similar numbers as the Tip amount has a range of $0 ~ $5.88, from Table 2, a approximately 60 cent average error is decently poor. When comparing the models, OR had the highest MAE and RMSE, which makes it the worst model, as expected. The LR had barely worse results compared to the RT4 and RT10 however the difference was not too large as to discard the model. Runtime was very minor with LR almost running as long as OR. Altogether, both models showed an approximately 60 cents error average which is by no means indicative of a weak linear relationship, however.
.

7 Conclusion

The results from this analysis found that when not taking into account the large outliers, tip amounts had quite large variance and were not heavily impacted by any variables other than fare amount and trip distance. From the initial visualization phase, we saw trends with specific sites in the different pick up and drop off locations as well as trends in the time or weekday people were being picked up at. However, when further exploring the specificities surrounding pickup and drop off time and location, there seemed little to no correlation with tip amount at all.

To test if tip amount was not completely random, we fitted three types of models and tested their average errors to see how difficult predicting tip amount would be. With the average error of approximately 60 cents the models showed that predicting tip amount was not a feat too far.

References

1) Quick Analysis of Tipping Yellow Taxi Drivers in New York : Phase 1
2) "TLC Trip Record Data." About TLC - TLC. accessed Oct 9th, 2020. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
3) https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d
4) "How to interpret error measures?" StackExchange.com accessed Oct 9th, 2020.https://stats.stackexchange.com/questions/131267/how-to-interpret-error-measures
5) "Decision Trees in Python with Scikit-Learn" Stack Abuse accessed Oct 9th, 2020. https://stackabuse.com/decision-trees-in-python-with-scikit-learn/
6) StatsModels.org accessed Oct 9th, 2020. https://www.statsmodels.org/v0.10.2/examples/notebooks/generated/predict.html

Code referenced from MAST30034 workshops and statsmodels.org