**NYC YELLOW TAXI FARE TIP PERCENTAGE PREDICTION**

**INTRODUCTION**

The tipping behavior reflects customer satisfaction/dissatisfaction with their rides. Here we are trying to analyze the percentage of tips, the NYC yellow taxi drivers get per ride. The tip amount depends on several factors such as fare amount, trip distance, pickup and drop off locations, travel speeds etc., We have taken these factors including the other factors in the NYC Yellow taxi dataset and predicted the percentage of tip amount for every ride.

As per the data dictionaries provided by NYC Taxi and Limousine Commission for the yellow taxi, the tip amount data has been included only for the payment made by the CREDIT CARD. We do not have the data for tips paid by cash. Hence, we must aggregate the data accordingly with the sample data of 1,20,000 transactions.

There were many data errors and lot of feature extractions required to be performed on the dataset to achieve the goal.

**OBJECTIVE:**

Our Objective is to predict the percentage of tips based on the several factors on every ride that the NYC YELLOW TAXI Drivers would receive.

**METHODOLGY:**

To predict the tips percentage, we have created two models with the training data and evaluated the performance based on the test data. We have used two regression techniques.

1. Multiple linear Regression

2. Random Forest Regression

**DATA PREPROCESSING:**

1. We have considered only the properties that will be helping our analysis from the dataset.

2. Tested for the missing values and dropped the variables which had higher number of missing-ness as they will not be useful in our analysis.

3. We have noticed data errors such as unrealistic fares, negative duration’/trip distance, Wrong GPS Coordinates, more than 100% tip amount etc., considering all those as outliers, we have removed those records and retained the cleaned data for further analysis

4. A custom feature called tip percentage has been created with python code as part of our analysis, which will be result of various factors involved in the tipping percentage. Below is the formula used for tip percentage.

**EXPLANATORY DATA ANALYSIS:**

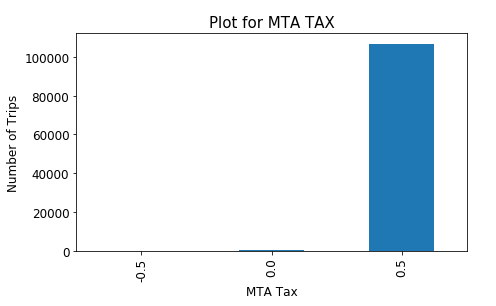
We have done the explanatory data analysis in some important features to understand the nature of the data we have collected.

**1. Visualization on the density of pickup and drop off locations requested for the trip inside NYC.**



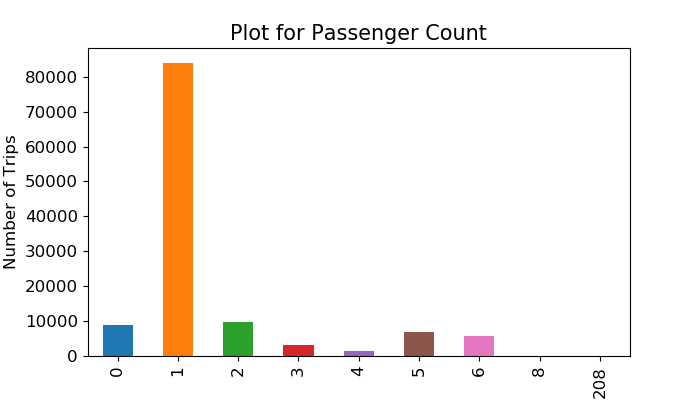
The plot shows the higher density of drop off in Manhattan and Brooklyn areas.

**2. Plot for MTA TAX Vs Number of trips**



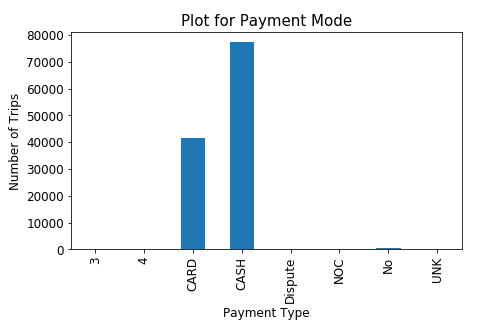
MTA TAXis automatically triggered based on the metered rate in use which is defaulted to $0.50. so, we focus on only that amount for our analysis

**3. Plot for Passenger Count Vs Number of trips**



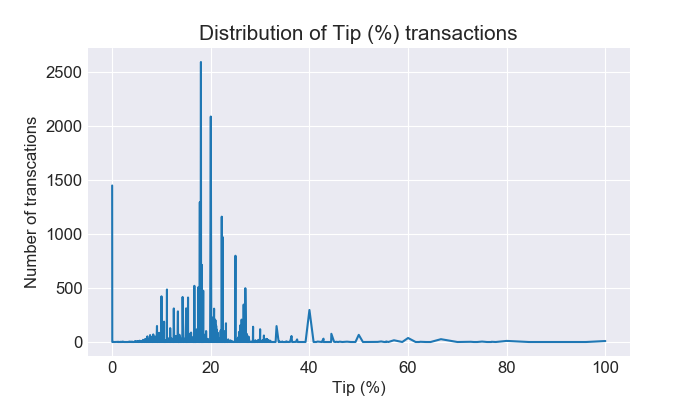
A usual trip has 1 to 6 passengers. Here the passenger count seems to be one in maximum number of trips. Rest of them (0) can be discarded as a data error

**4. Plot for Payment Mode Vs Number of trips**



The most common payment from the sample collected turned out to be cash as per the above analysis. However, as per data dictionaries provided by NYC Taxi and Limousine Commission, cash tips are not included in the dataset. So, we are considering the CARD payment mode for our analysis.

**5. Plot for the distribution of tip percentage over the trip fare transactions**



This plot becomes one of the most important key phase of our analysis. Here we can notice that the tip percentage shows more than 60% people are not tipping during the fare payment even in credit card transaction. Only rest 40% has customary gratitude to the NYC taxi which fluctuates between 10% to 40% provides tip. We can also notice that the highest percentage of tip most often given around 20%.

**MODEL CREATION FOR PREDICTING THE TIP PERCENTAGE**

After all the data preprocessing and explanatory analysis, now we are ready to start with the model creation. We ae using two machine learning techniques to create the model for our objective. We have split our sample into 80% training and 20% testing.

* Multiple Linear Regression
* Random Forest Regression

**MULTIPLE LINEAR REGRESSION:**

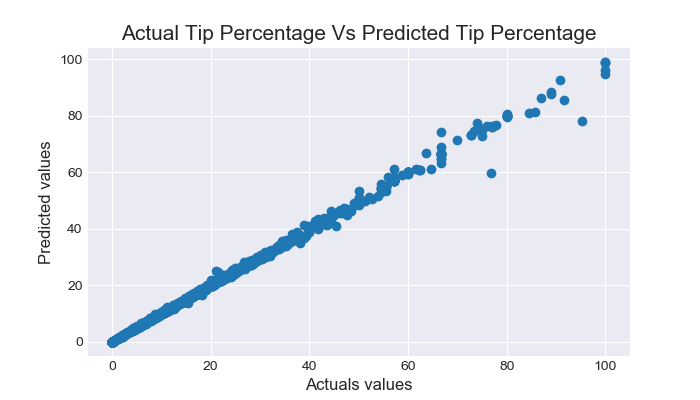
The model created with the multiple linear regression gave us a very bad performance with highest mean squared error rate of 134.743 and the R2 value of 44.2% which explains the goodness fit of the model is very bad. So, we have decided to go for Random Forest Regression

**RANDOM FOREST REGRESSION:**

Random forest is one of the ensemble machine learning technique to perform regression. This ensures to give the high-performance model with the enormous number of decision tree construction.

As expected, as a result of the random forest regression, we have achieved a R2 value of 99.9% with our sample data. This means this model can be act as best model to predict the tip percentage based on the factors provided.

Its performance on the testing data is calculated using Mean Squared Error which turned to be 0.184 which is very minimal compared to the linear regression model. Also, we plotted the actual values of tip percentage to the predicted tip percentage, which shows a high linear relationship between them. The predicted values are almost closed to the actual tip percentage.



**CONCLUSION:**

Therefore, we can conclude that the model created with the random forest regression can provide us the best prediction for the NYC YELLOW TAXI TRIP data fare tip percentage.