Cab Fare Prediction

Problem Statement

Cab rental start-up company has successfully run the pilot project and now want to launch cab service across the country. Company has collected the historical data from pilot project and now have a requirement to apply analytics for fare prediction. Need to design a system that predicts the fare amount for a cab ride in the city.

Business Understanding

For the cab rental start-up company, deciding fare of cab for a trip is one of the important challenge. Cab fare for the trip is decided considering parameters like time, distance, booking fee, busy time, area or load on the system (number of trip requests and available cabs), etc.

Dataset Details

Data consist of 16067 trips with the following attributes –

pickup_datetime – timestamp value indicating when the cab ride started

pickup_longitude – float value for longitude coordinate of where the cab ride started

pickup_latitude – float value for latitude coordinate where the cab ride started

dropoff_longitude – float value for longitude coordinate of where the cab ride ended

dropoff_latitude – float value for latitude coordinate where the cab ride ended

passenger_count – an integer indicating the number of passengers in the cab

Data Cleaning

Dataset is checked for missing values, invalid values and outliers.

Missing Value Data Treatment

Missing value percentage is very low. So ignoring this data and deleting this data from further analysis.

Attribute Name	Missing Value Count	Missing Value Percentage
fare_amount	24	0.14
passenger_count	55	0.34

Invalid Value Data Treatment

Incorrect values data affects the prediction so incorrect data is removed from further processing. Checks applied on individual features are as follows –

- passenger_count –
  1. Passenger count value must be positive and integer value. (float value not allowed)
  2. Passenger count value can not be greater than 6 (considering SUV also)
- Pickup, dropoff - longitude and Latitude values
  1. Longitude Value should be in the range - -180 degree to + 180 degree
  2. Latitude value should be in the range - -90 degree to +90 degree

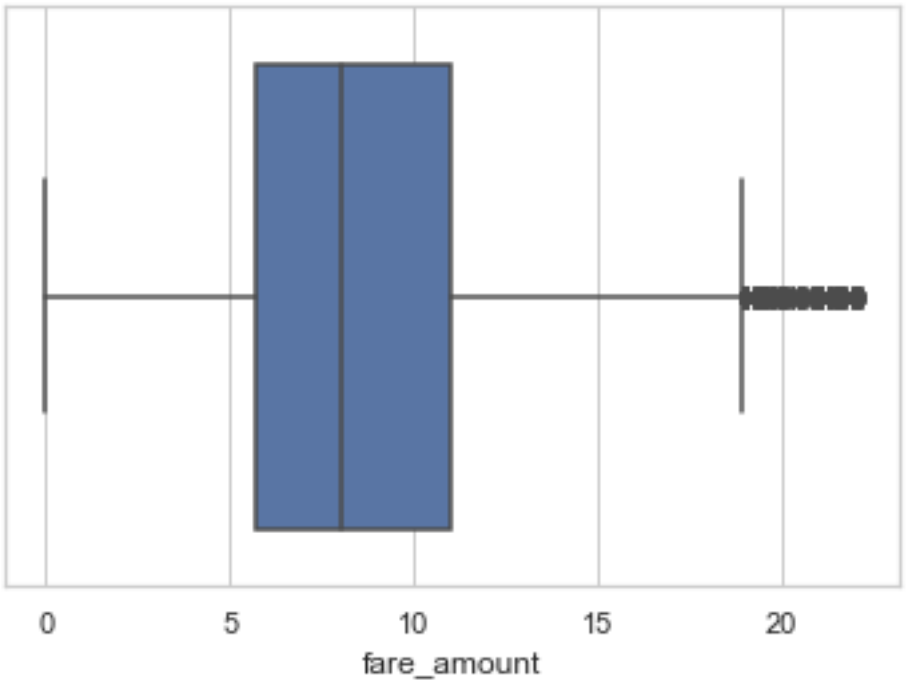
3. Longitude and Latitude both value should not be zero (default data)

- Pickup_datetime –
  1. Value should be convertible into datetime format.
  2. 43 is one invalid value in the data which is ignored.
- Fare_amount –
  1. Fare amount must be positive

Outlier Data Treatment

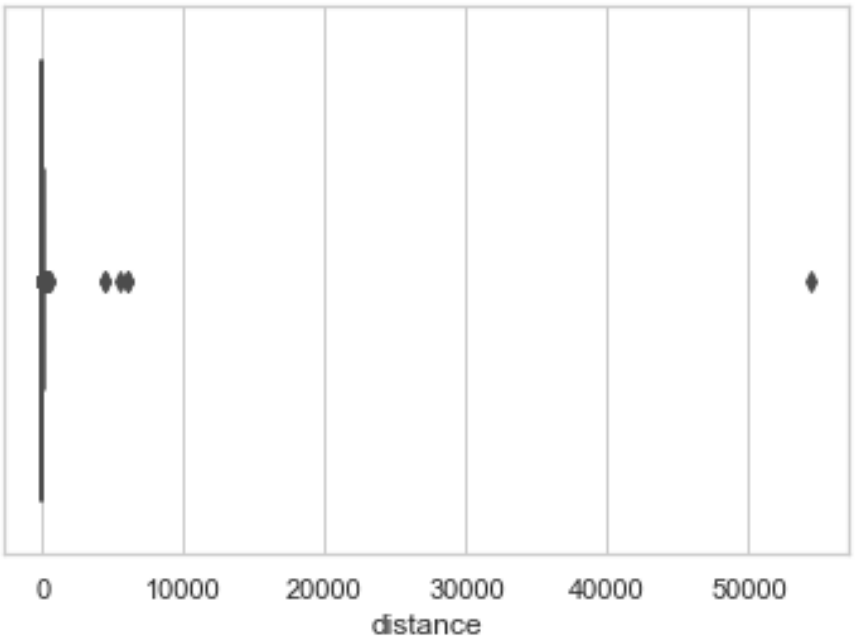
Fare Amount –

There are outliers present in the fare amount feature. Data with outliers has been removed from further processing.



Distance –

There are outlier present for the distance of a trip. So we removed the data having outlier distance.



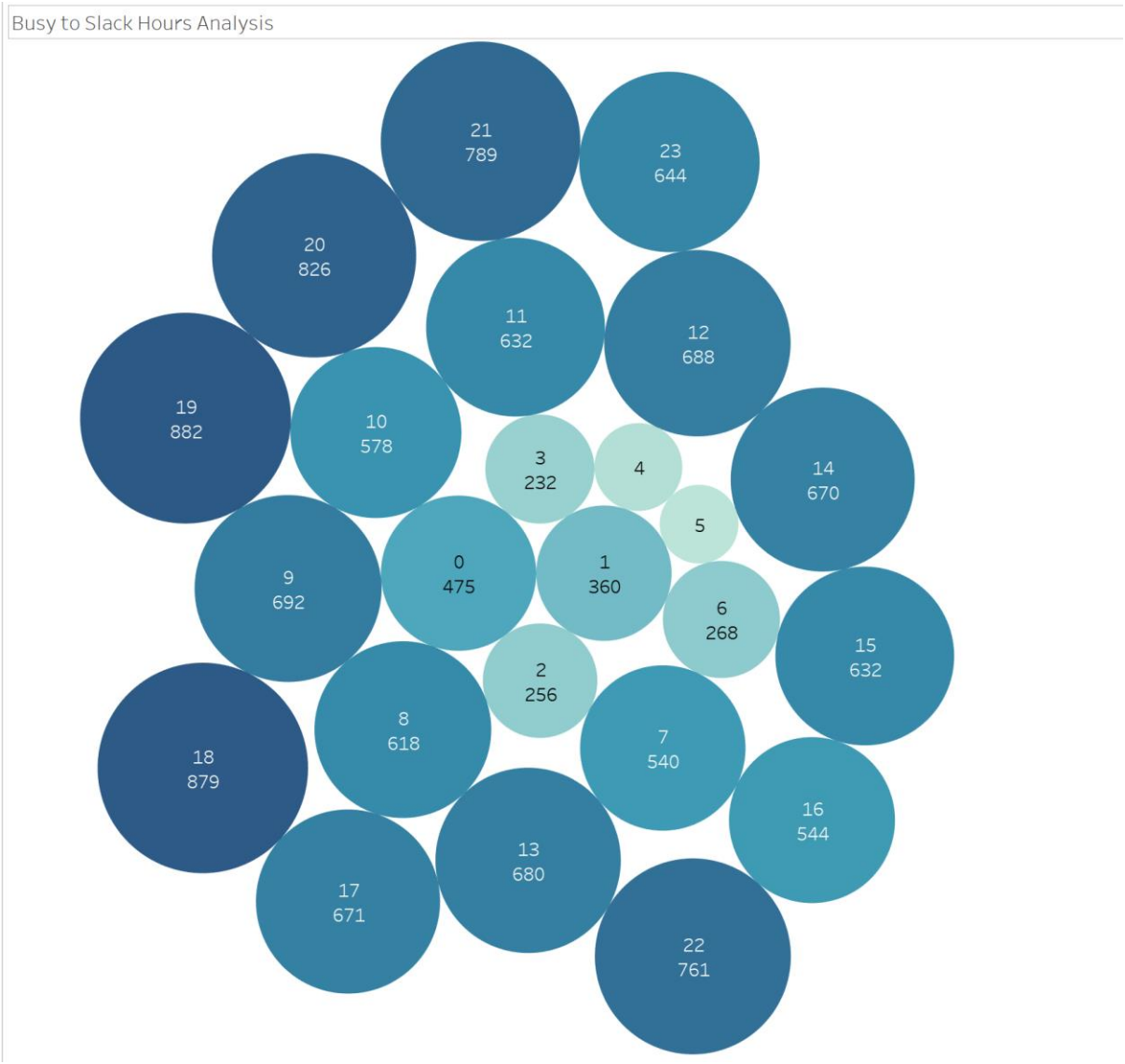
Deriving New Features and Feature Analysis

As we understood some of the features which decides the fare of trip are distance, busy hours, region, etc. So we will derive following features

1. Distance
   Distance of the trip is calculated using Haversine formula considering Pickup latitude, longitude and Dropoff latitude , longitude values. Formula is sated below –

   *Distance, d = 3963.0 * arccos[(sin(lat1) * sin(lat2)) + cos(lat1) * cos(lat2) * cos(long2 – long1)]*

2. Busy Hours
   To find busy hours, pickup hour is calculated from pickup_datetime and number of trips at each hour is calculated.

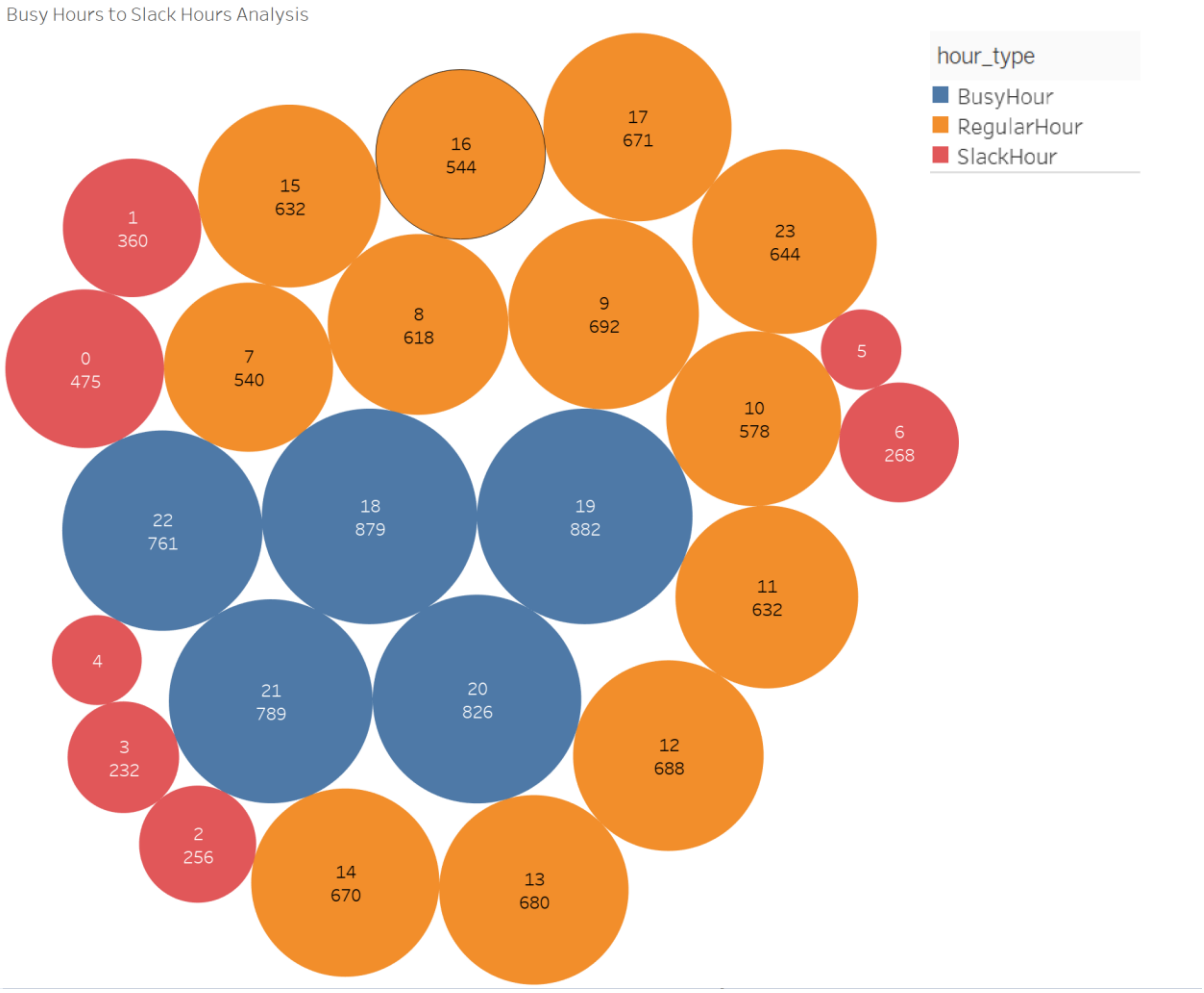


Visualization shows the hour time and number of trips in that hour. Based on the analysis, all the trips data is divided into three categories

1. Busy Hours
2. Regular Hours
3. Slack Hours

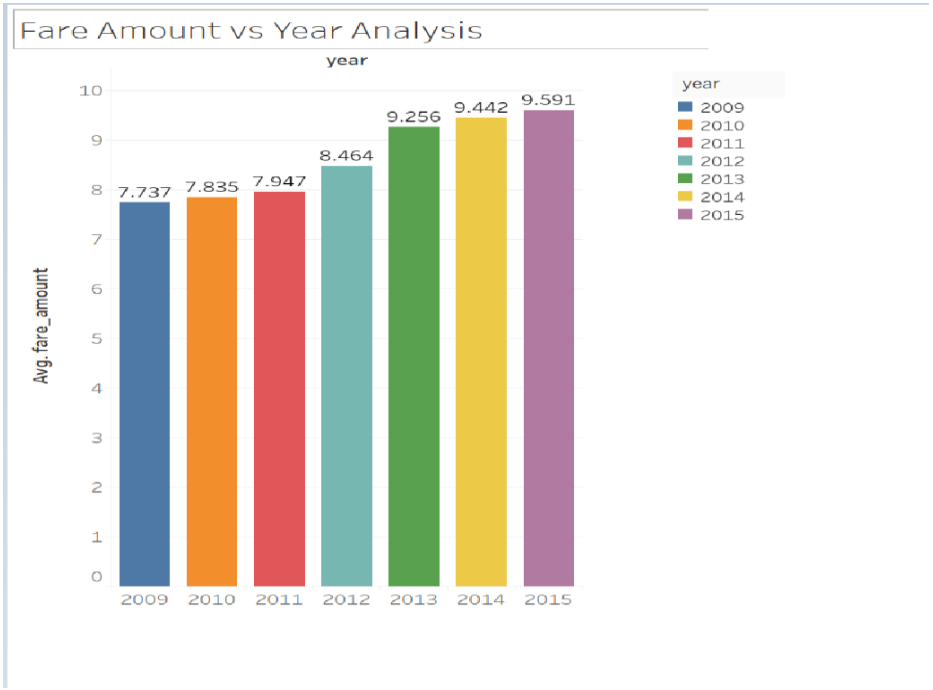
For the modelling purpose, this variable is converted from categorical to integral using dummies.

Based on the division the above graph displays as

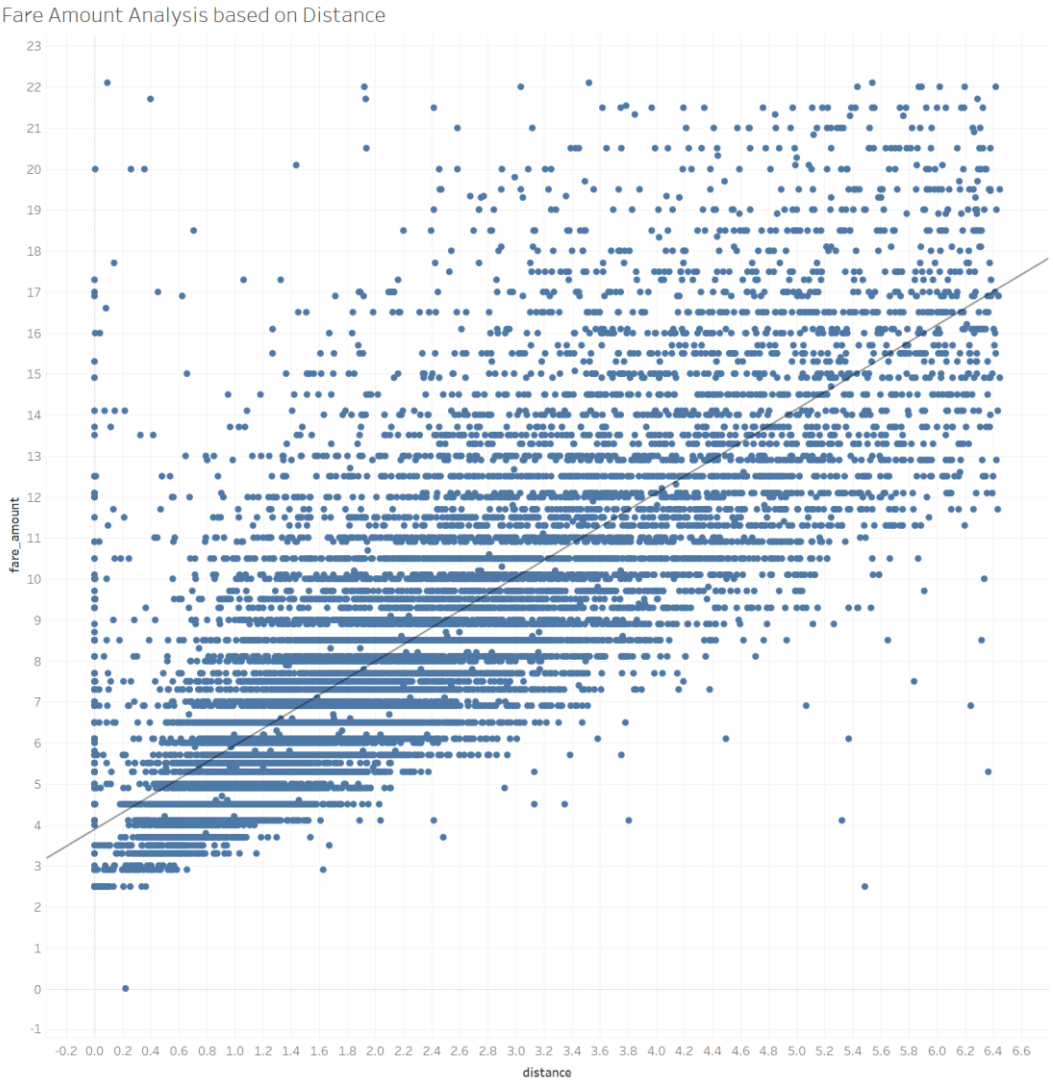


3. Year of the trip

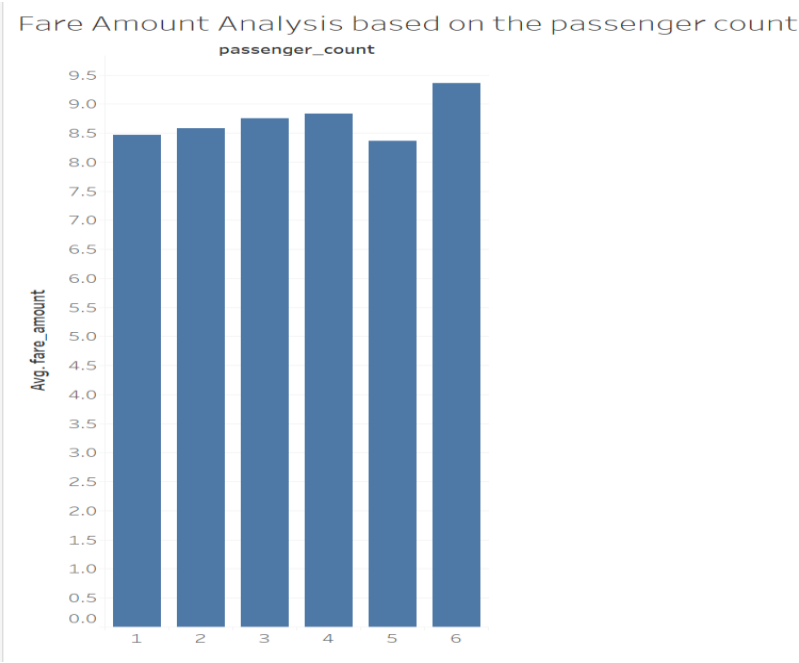
Following graph shows that average fare amount in every year is increasing. This factor can be related with the booking fee amount which is increasing every year. Year of the trip variable is converted from categorical to dummies.



4. Fare Amount Analysis based on Distance
   Basic trend shows that fare amount increases with increase in distance.

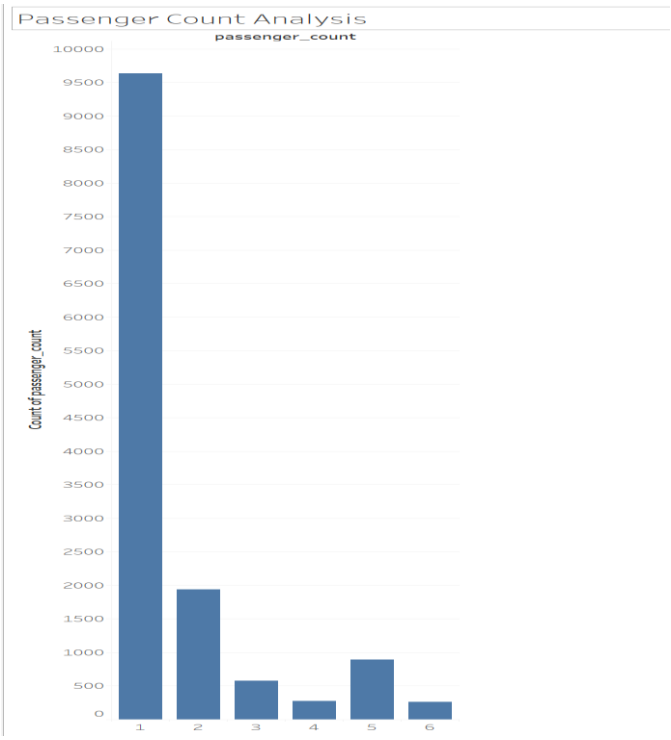


5. Fare amount analysis based on passenger count
   Following graph shows that average price for 6 passenger count is higher than others this can be related with SUVs have higher fare charges than other vehicle.



6. Passenger Count Analysis

Majority of the trips are done by single passenger.



Model Selection and Analysis

Problem statement is clearly of regression types. Following are the regression models which are developed –

1. Multiple Linear Regression
2. Decision Tree
3. Random Forest
4. Gradient Boosting

To train and test the model, we have divided the data into 80-20 split i.e. 80% of data is used for training purpose and 20% of data is used for testing purpose. The features which are used to derive new features are dropped from the modelling.

Multiple Linear Regression

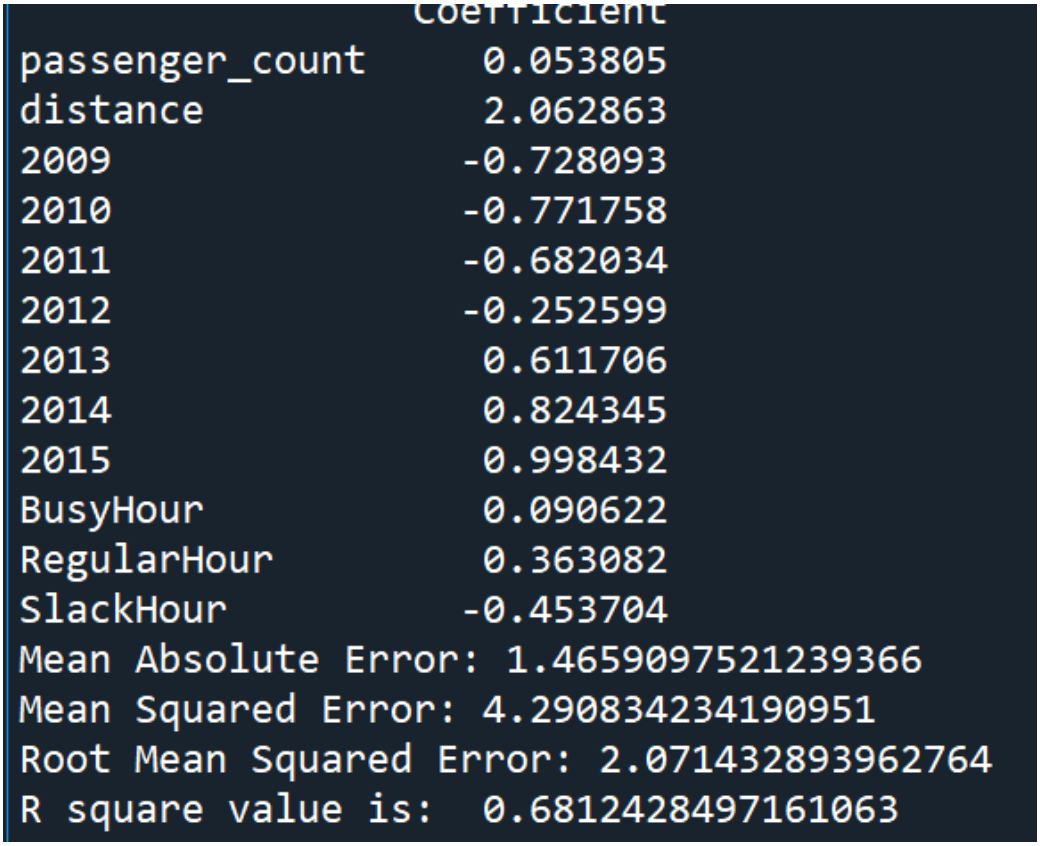
First passenger_count, distance, year of trip and hour_type are selected as features. Following is the result of that model. Root mean squared error is 2.07 and R square value is 0.68.

Positive Coefficient value shows as the value of independent variable increases value of dependent variable increases and negative value shows as the value of independent variable increases value of dependent variable decreases.

Coefficient value shows the change in the dependent variable based on the change in independent variable.

Based on the results, features affecting to calculate fare_amount

1. Distance – Major affecting feature. As the distance of the trip increases fare_amount increases.
2. Year of trip – Based on the coefficient of 2009, 2010, 2011, 2012, 2013, 2014, 2015 we can clearly say that fare amount is affected by year. The more recent year the more fare amount will be. This coefficients we can also understand that we already seen avg.fare amount value comparison across years.
3. Hour Type – Slack Hour decreases the value of fare amount. We can also understand this because if it is slack hour then companies provide some discount or there is no surcharge applied on trips.
   Hyper tuning of this model is done based on the p-value for feature. p-value signifies the feature is significant in predicting the value or not.



Decision Tree Regressor

Root Mean squared error for this model is 2.42 and R squared value is 0.56.

The hyper tuning of parameters for this model is done by varying the max_depth of the tree. Value was varied from 2 to 15. Optimum value of max depth is selected based on the low Root Mean Square Value and High R score.

Random Forest Regression

Root Mean Square Error for this model is 2.24 and R squared value is 0.62.

Hyper tuning of paramters for this model is done by varying the estimator value. Value was varied from 10 to 100 with the step of 10.

Summary Table

Based on the RMSE and R – squared value for all the three types of models. Multiple linear regression model is selected based on the low root mean squared error value and high R- squared value.

Linear Regression		Decision Tree Regression		Random Forest Regression	
RMSE	R-Squared	RMSE	R - Squared	RMSE	R - Squared
2.07	0.68	2.42	0.56	2.24	0.62

Predicting Fare Amount for the Test Data

- So we selected multiple linear regression model from our model analysis to predict the fare amount for test data.
- As per our process in analysis, features are derived – Distance, Type of Hour which are used during training of the model.
- Fare amount for the test data is predicted and file is stored in csv “predicted_cabfare_data_python.csv” .