Capstone Project

NYC Taxi Trip Time Prediction

Kiran Mamtani

Content

- Summary of Data
- Exploratory Data Analysis
- Feature Creation
- Cleaning and Transforming Data
- Machine Learning Models
- Conclusion
- Challenges

Summary of Data

About Data - The dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform.

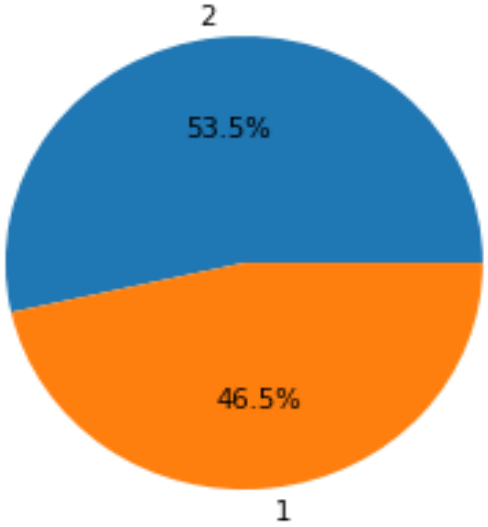
Size of Data -

Rows - 1458644
Columns - 11

Source: https://www.freepik.com/

Exploratory Data Analysis

Vendor ID

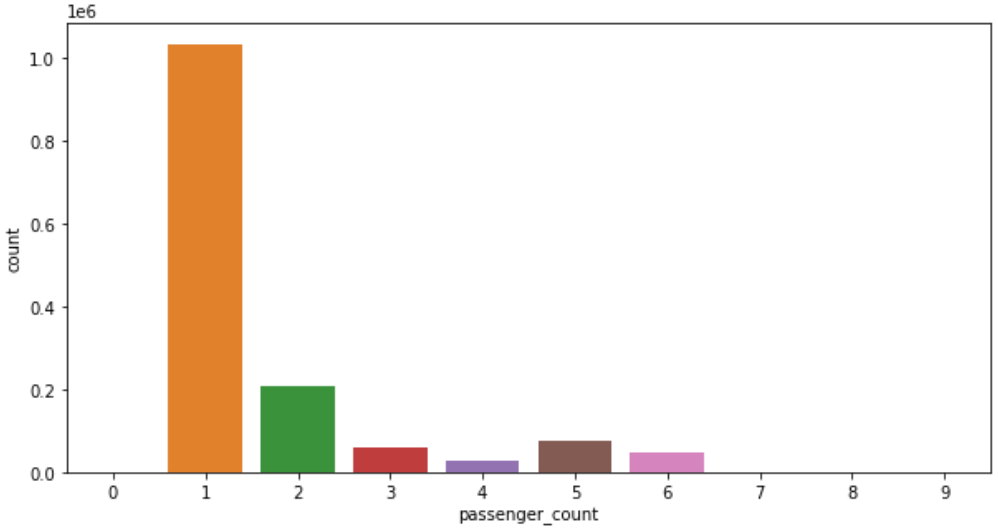


We have two unique
Vendor id i.e. 1 & 2.

46.5% data is of vendor id 1
53.5% data is of vendor id 2

Exploratory Data Analysis

Passenger Counts



There are some trips with 0 passenger count. There is only 1 trip for 9 & 8 passengers and 3 trips for 7 passengers

The highest amount of trips is with 1 passenger. Here we will not remove 0, 9, 8, 7 passenger rows as this number is very small compared to the data.

Feature Creation

Had created few new features from the existing features

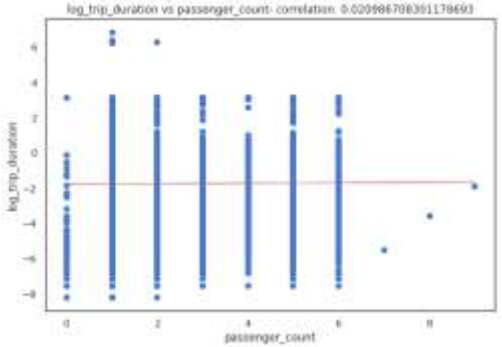
- From pick-up and drop-off date time column we have created pick-up & drop-off day, date, month, hour, minutes, time zone of day (Morning, Evening, Afternoon, late-night)
- From latitude and longitude, created distance feature
- From time and distance, created speed feature

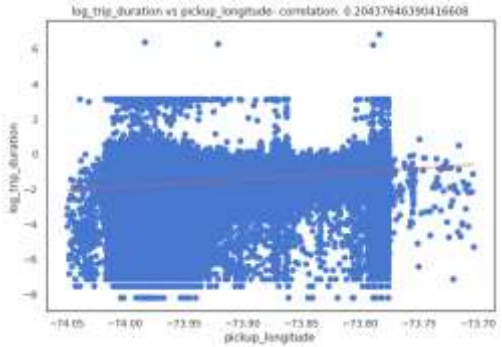
Cleaning & Transforming Data

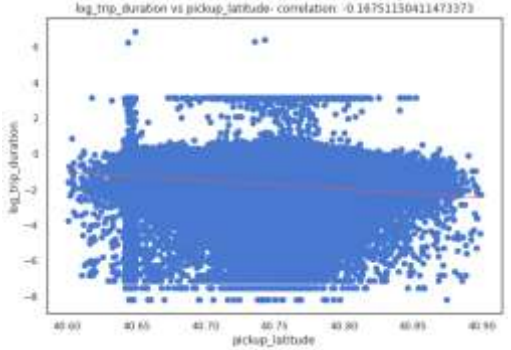
After creating features we have cleaned few features and transformed few…

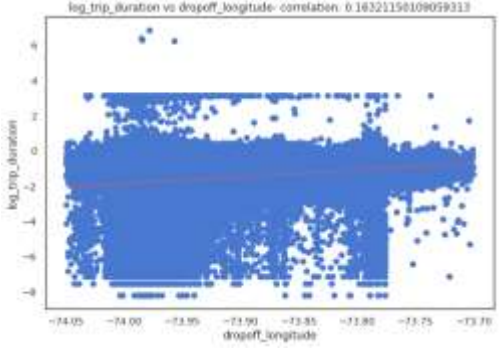
- We had applied a range in longitude & latitude feature and removed all the outliers.
- In the distance feature, there was few rows with 0 km so replace those rows with the mean.
- Trip duration and distance features was extremely right skewed so applied a log transform to normalize the feature.

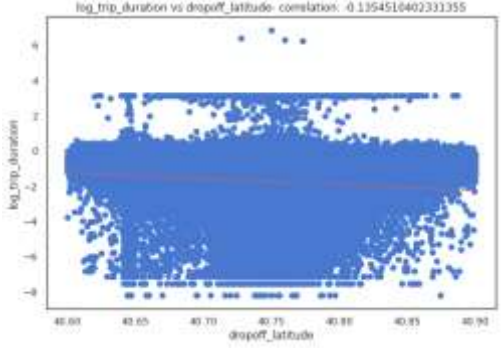
Correlation between target & independent features



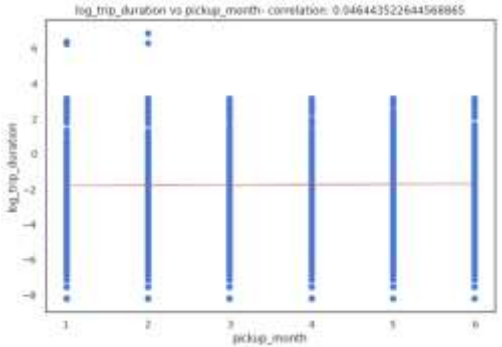


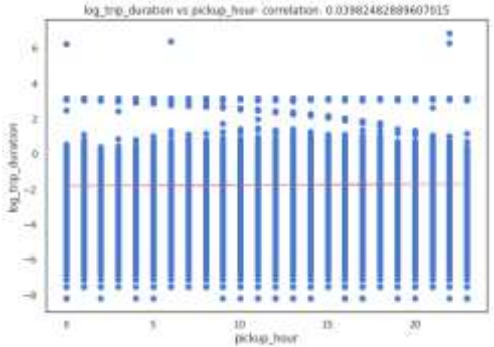


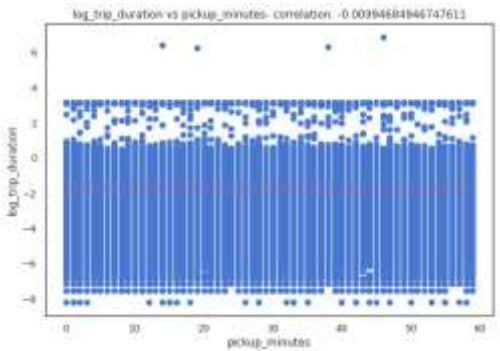


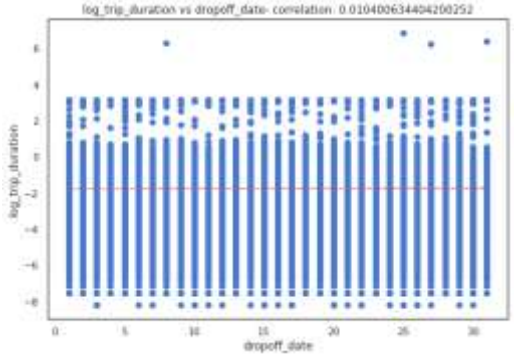


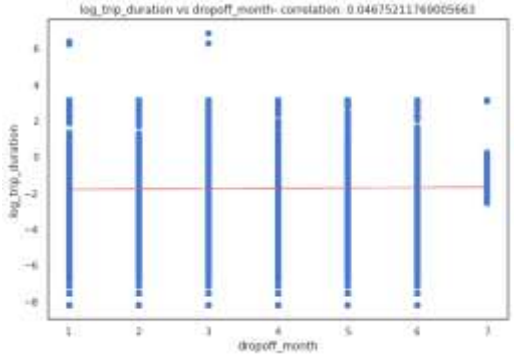
Correlation between target & independent features

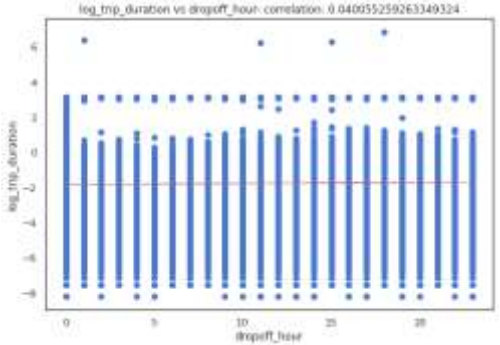






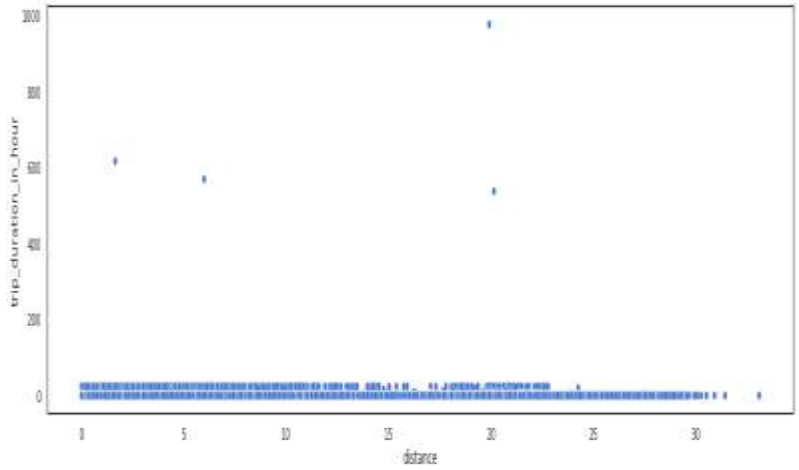




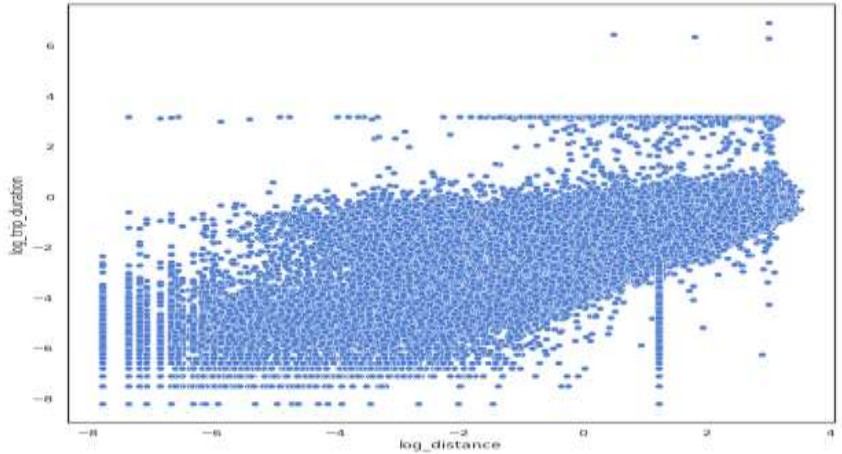


Relation between Trip Duration & Distance

Trip Duration and Distance before applying log transformation



Trip Duration and Distance after applying log transformation



Machine Learning Models

Test Train Split

- Divided the data into two parts
- Train data set which has 80% of data
- Test data set which has 20% of data

Machine Learning Models

Linear Regression

- MSE for Train set – 0.2270
- MSE for Test set – 0.2225
- RMSE for Train set – 0.4765
- RMSE for Test set – 0.4717
- R2 for Train set – 0.6406
- R2 for Test set – 0.6477

As we can see the error is high and the accuracy is less so lets regularize our model for better result

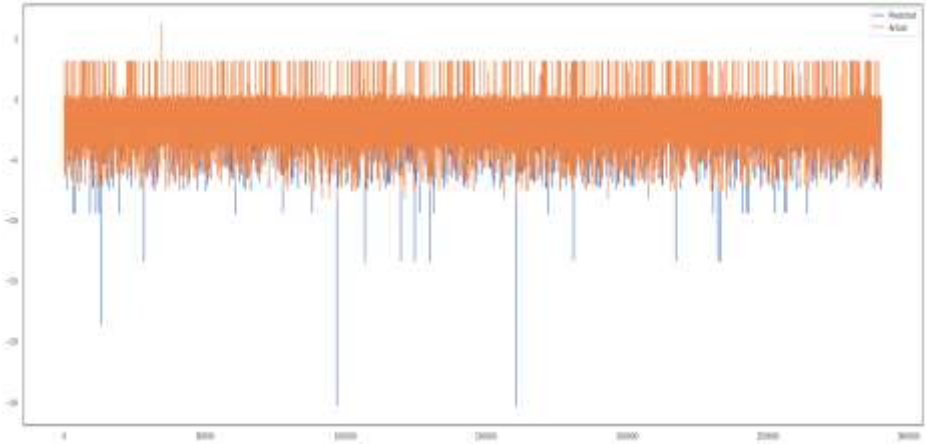
Machine Learning Models

Lasso Regression

MSE : 0.24515744022236624
RMSE : 0.49513375992994685
R2 : 0.6118330911857324
Adjusted R2 : 0.6118170547650201

Cross validation

MSE : 0.24474684760854845
RMSE : 0.4947189582061197
R2 : 0.6124831977684373
Adjusted R2 : 0.6124671882057142

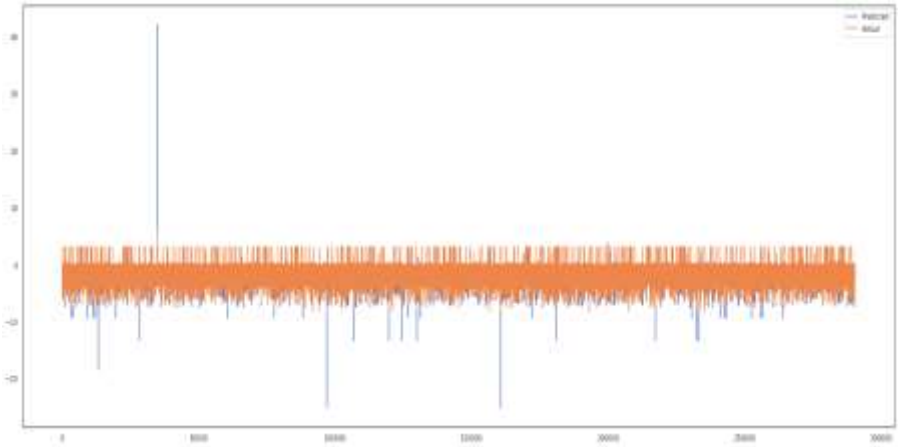


As in lasso regression we are not getting good result. It is similar to one we got in linear regression above. And as seen in above graph, the predicted and actual results differ that means we have residuals.

Machine Learning Models

Ridge Regression

MSE : 0.22251127121554737
RMSE : 0.4717110039161132
R2 : 0.6476896142914113
Adjusted R2 : 0.6476750592187396



Ridge regression outcome is better than lasso regression but still our accuracy is less

Cross validation

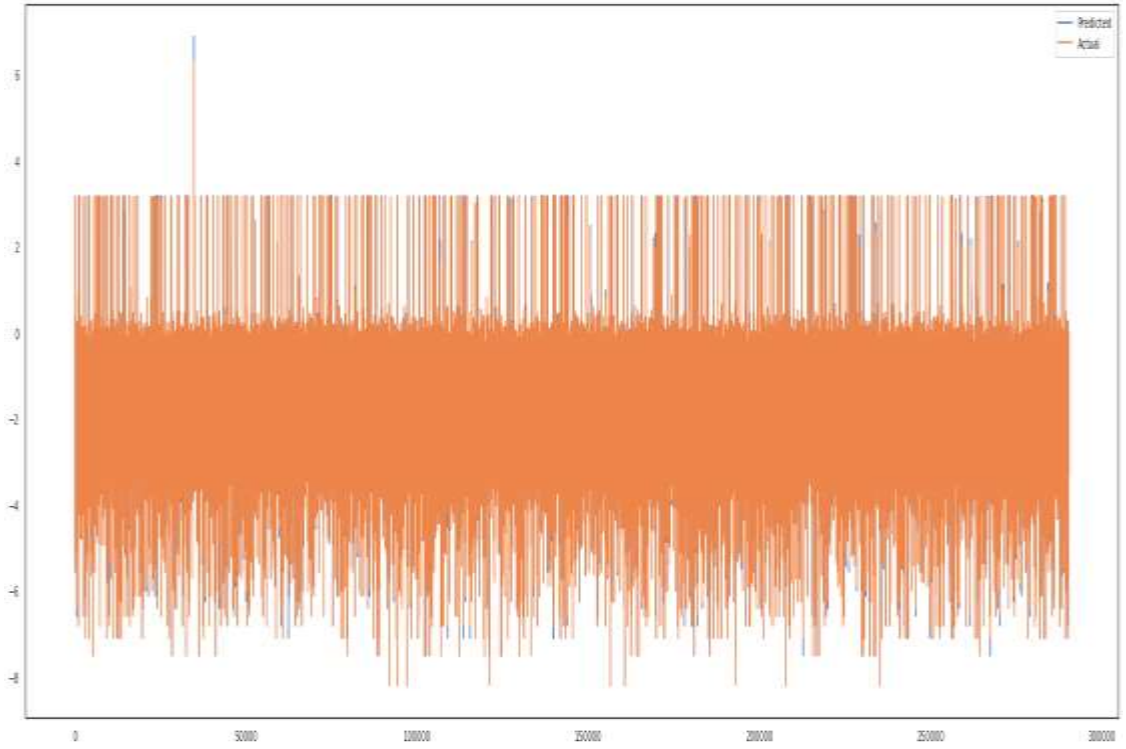
MSE : 0.22353128418087587
RMSE : 0.472790951881353
R2 : 0.6460745897612852
Adjusted R2 : 0.6460599679667672

Machine Learning Models

Decision Tree Regressor

MSE : 0.00016784323389092853
RMSE : 0.012955432601458297
R2 : 0.9997342475545277
Adjusted R2 : 0.9997342365754412

The mean square error in decision tree method is very less in compared to lasso, ridge and linear regression and the accuracy is also higher. This model seems fit. Now lets try last method i.e. Random Forest

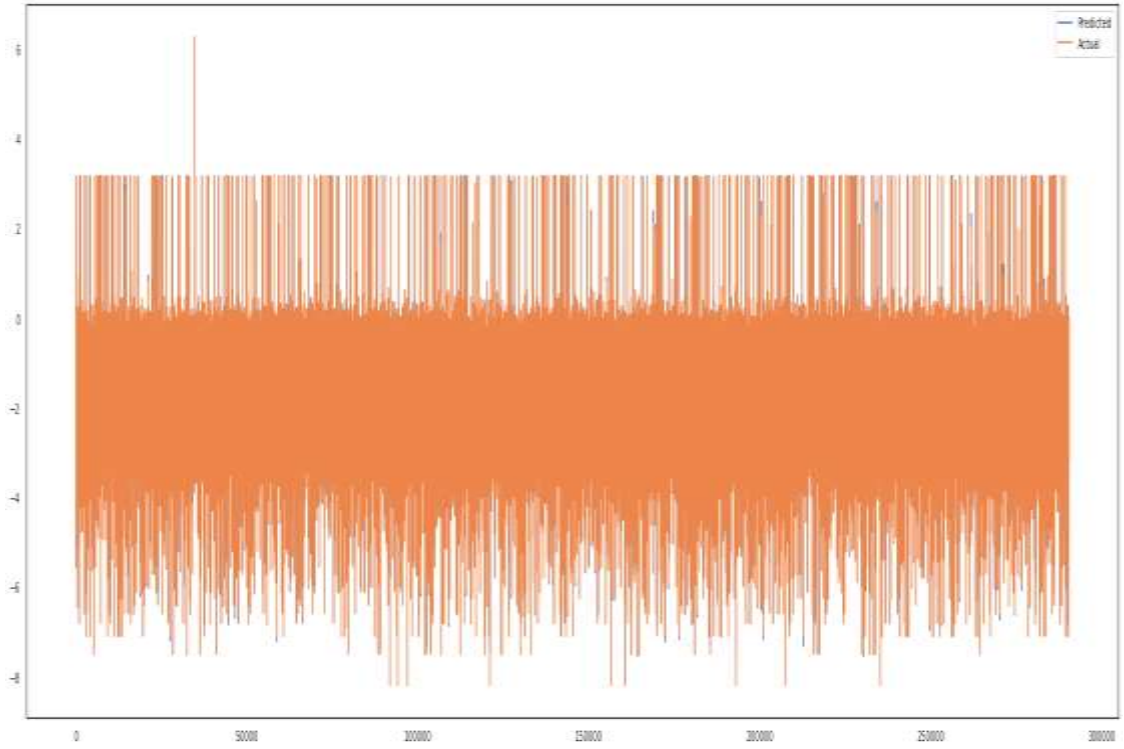


Machine Learning Models

Random Forest Regressor

MSE : 7.838811591655448e-05
RMSE : 0.008853706337831 32
R2 : 0.9998758851756019
Adjusted R2 : 0.9998758800480202

In random forest model, the accuracy is high i.e. 99.98% and its seems similar to Decision tree model. Here the value of MSE is 7.83.



Conclusion

First of all, importation of all the necessary libraries including Sklearn library is done. Dataset gets imported accordingly in order to analyze the various attributes of the taxi trip duration.

Exploratory Data Analysis was done to investigate dataset patterns and outliers.

- 45.5 % is vendor id 1 data and 53.5% is of vendor id 2 data
- There are some trips with even 0 passenger count. There is only 1 trip for 9 & 8 passengers and 3 trips for 7 passengers. The highest amount of trips is with 1 passenger.
- 99.4% is N i.e. not a store and forward trip and 0.6% is Y i.e. store and forward trip

Created few new features from pick-up and drop-off datetime. Created distance feature from longitude and latitude columns. And created speed feature from distance and trip duration.

In evening, maximum pick-ups and drop-offs has occurred followed by late night, morning afternoon and early morning. The maximum pick-up and drop-off has occurred on Friday followed by Saturday and Thursday. Maximum trips happened in March followed by April and May.

After creating features we have cleaned few features and transformed few. We had applied a range in longitude & latitude feature and removed all the outliers. In the distance feature, there was few rows with 0 km so replace those rows with the mean. Trip duration and distance features was extremely right skewed so applied a log transform to normalize the feature.

Conclusion

The cleaned data is then analyzed deeply for more feature extraction by finding out the correlation in the data which ensures maximum coverage.

Data was split into Train and Test data set and different machine learning regression models were applied.

In linear regression, the r2 value of train and test data is almost same so we can say that our model is correct but the accuracy of predicted values is around 64% only. So lets regularize our model for better result by applying different regression models.

As in lasso regression we are not getting good result. It is similar to one we got in linear regression. Here the predicted and actual results differ that means we have residuals. r2 is 61%.

As in ridge regression, the outcome is better than lasso regression but still our model does not fit as r2 is still 64%.
The mean square error in decision tree method is very less in compared to lasso, ridge and linear regression and the accuracy is also higher.

In random forest model, the accuracy is high i.e. 99.98% and its seems similar to Decision tree model. Here the value of MSE is 7.83.

So the best model for our data is Random Forest since it has smallest MSE and r square is 99.98%.

Challenges

The few challenges I faced during this projects are…

- It was difficult to guess the log transformation of drip duration. Without transformation the model was failing.
- Deciding to split the data whether in 80:20 ratio or 70:30 ratio. Done some trial and error and then decided to take 80:20 ratio.
- Analyzing all the linear regression models
- Choosing the alpha value for lasso and ridge regression.

Thank You!