SEOUL BIKE DATA DEMAND

INTRODUCTION

Public bike-sharing systems have been gaining momentum only in the last decade. The main purpose other than convenience and easy-to-use service for customers is the mobility. More people are turning to healthier life styles and locations where bike riding can be easily available. There are many benefits in bike riding. Therefore, it is important to have rental bikes available to the customers (in our case, the public) to reduce their waiting time.

In this project, we implement an algorithm using gradient descent to check on the rental bikes available to the customers. It helps to give us the best model selected through experimentation and then evaluate the model for prediction

DATASET

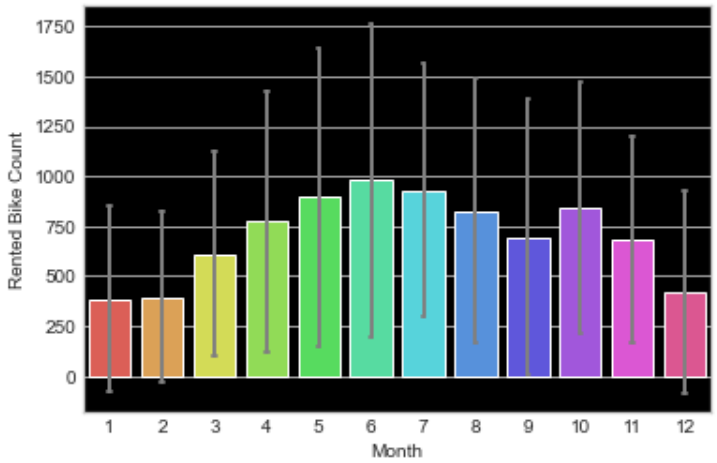
This dataset has 8760 records comprising the details of every hour each day and also 14 columns. The data contains the hourly and daily count of rental bikes. It also contains the weather information (Rainfall, Snowfall, Temperature, Humidity, Visibility etc). Rented Bike Count is taken as the independent variable for our analysis.

We plan to implement the algorithm using linear regression through gradient descent. We wil change the learning rate, convergence threshold, restarts and epoch for our algorithm implementation

DATA EXPLORATION

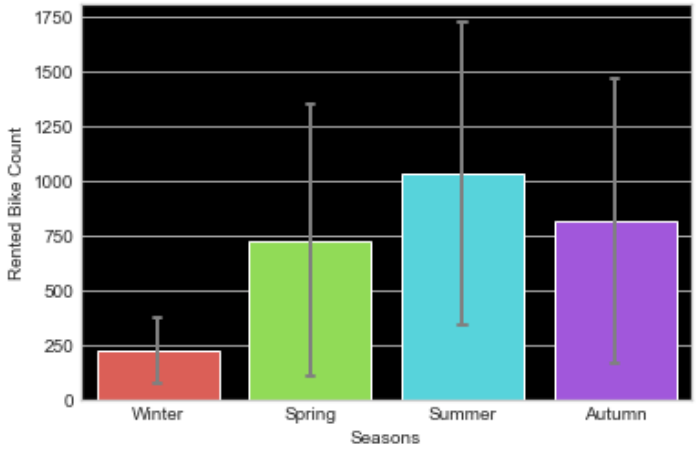
Exploratory Data Analysis is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model.

We will use our dependent variable, Rented Bike Count to find dependency of this column with other variables.



In Fig 3.1, months 5,6 and 7 shows high demand in bike count compared to the other months.

Fig 3.1 Count of Rented Bikes According to Month



In Fig 3.2, Summer Season shows the rented bike count demand to be the highest and Winter has the least bike count.

Fig 3.2 Count of Rented Bikes According to Seasons

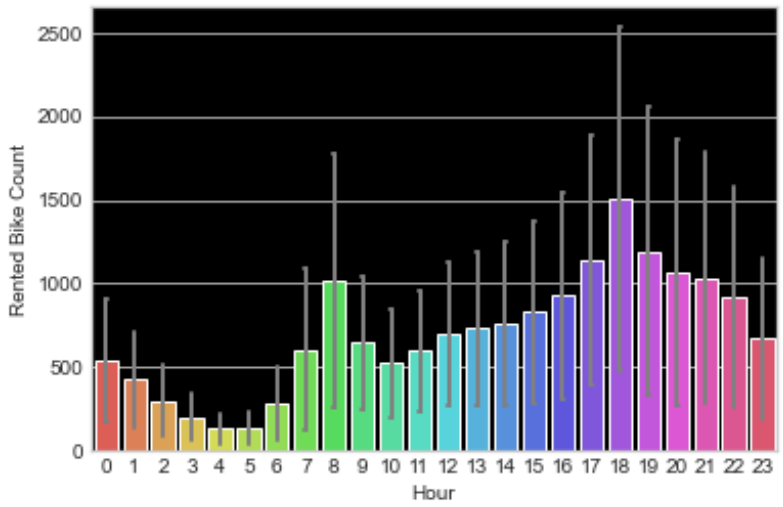


Fig 3.3 Count of Rented Bikes According to Hour

In the above Fig 3.3, which shows the use of rented bike according the hours and the data are from all over the year. We can see that people use rented bikes during their working hour from 7am to 9am and 5pm to 7pm.

Apart from the above dependency of dependent variable with categorical variable, the dependency with respect to numerical variables were also analysed. In below plots Fig 3.4, the following observations are noted.

If there is heavy rain , the demand for bikes are not less. For 20mm of rain , we get a spike in the demand for bikes.

Dependency of wind speed with rented bike count shows some uniformity. It can be observed that there for a spike for 7 m/s, there is a huge demand for bikes which shows that people do ride bikes when it is windy.

As temperature increases, the demand for bikes also increases. We can see that for temperature between 25-30 degree celsius, the demand is very high.

During snowfall, the demand for bikes are low. At 4cm of rainfall, we can see a decrease in bike counts that the demand is very low during snowfall.

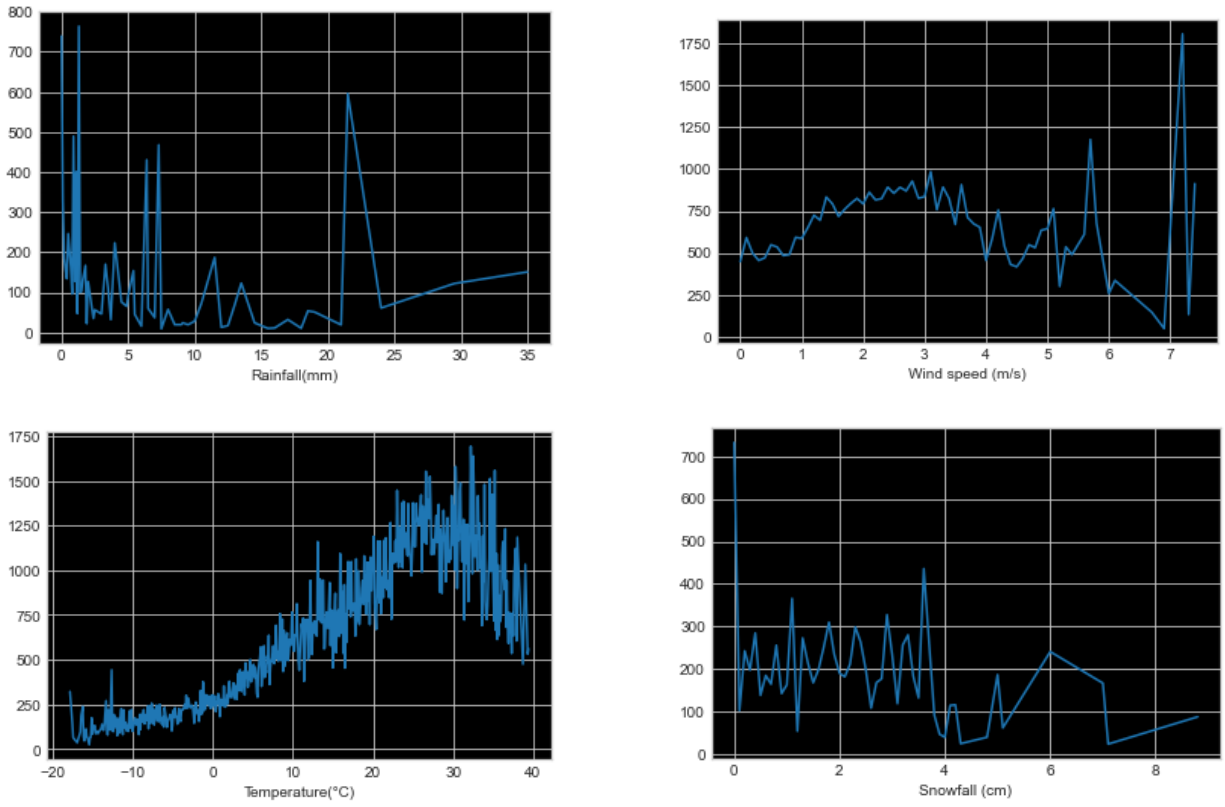


Fig 3.4 Dependency of Rented Bike Count with numerical variables

In Fig 3.5, the target variable, Rented Bike Count has positive correlation with the temperature (0.54), dew point temperature (0.38) and solar radiation (0.26). Most negatively correlated variables are humidity (-0.2) and rainfall (-0.12).

In Fig 3.6, we can see a strong positive correlation between columns ‘Temperature’ and ‘Dew Point Temperature’, which shows 0.91. So, we can drop ‘Dew Point Temperature’ as they will have the same variations in our analysis.

Also in Fig 3.6, based on the correlation data and dependency between the variables,we can see that Humidity and Dew Point Temperature has a positive correlation of 0.54 and Visibility and Humidity has a negative correlation of -0.54.

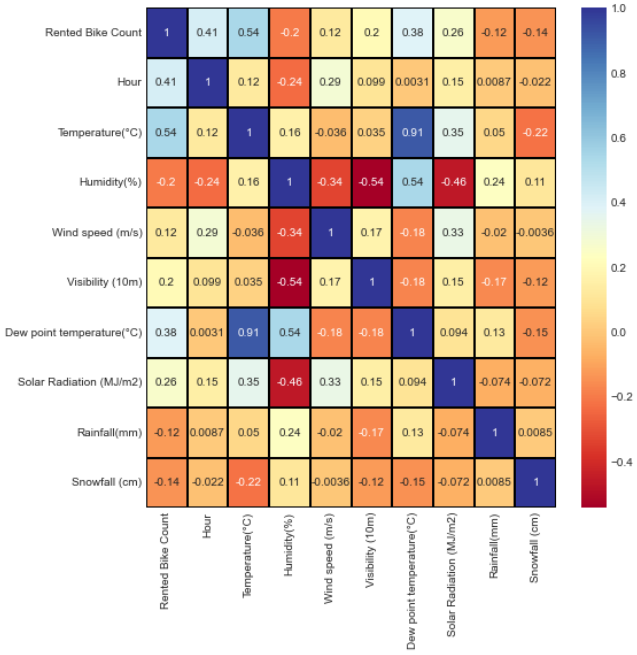


Fig 3.5 Correlation between variables

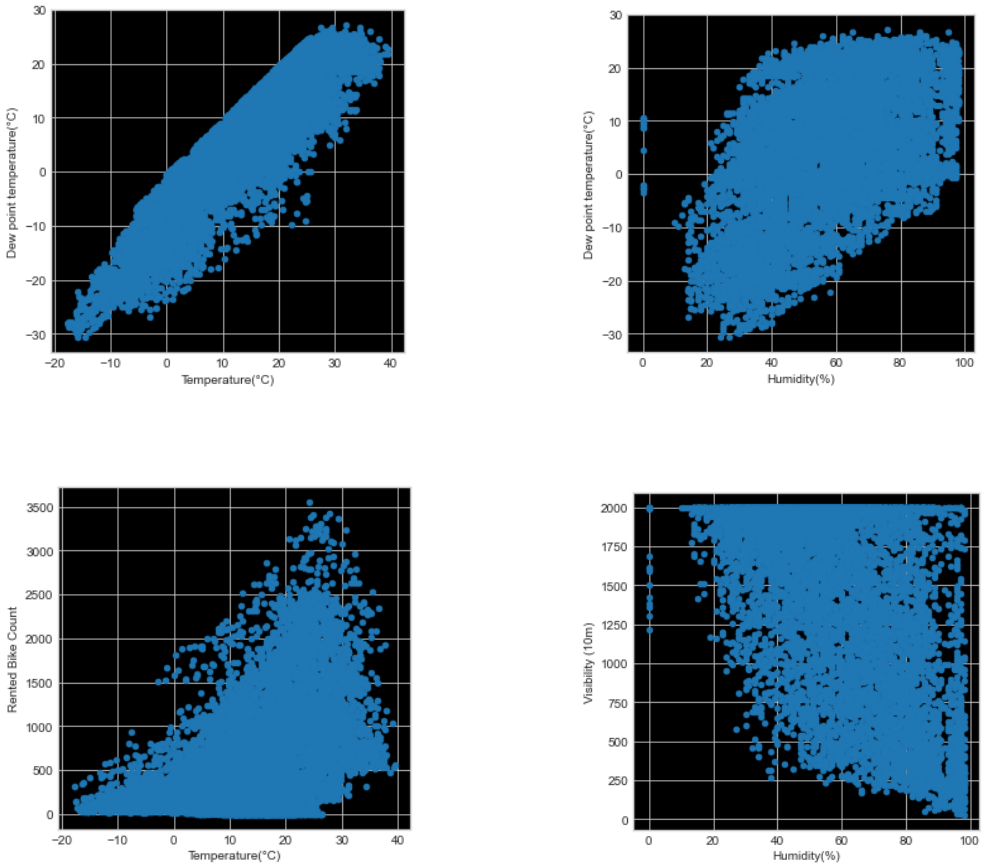
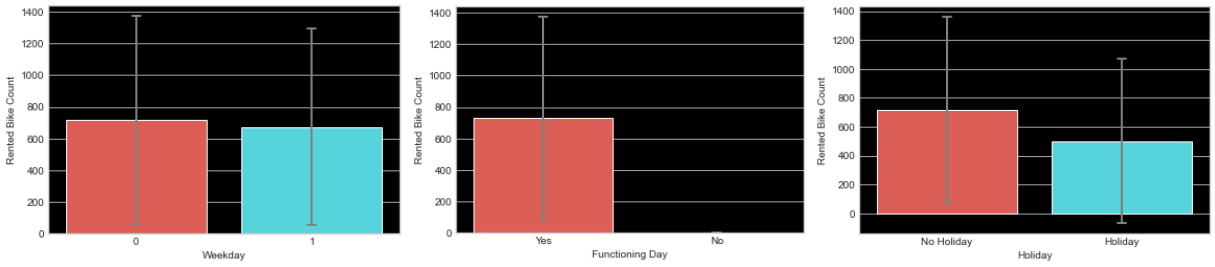


Fig 3.6 Variation between strong variables

We also check the bike count with respect to weekdays, functioning day and holiday



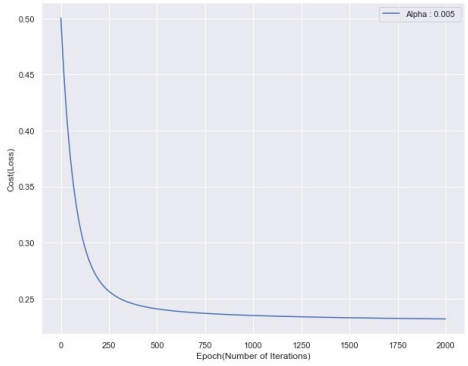
GRADIENT DESCENT

Gradient Descent algorithm has been performed in this dataset. For this, the features have been preprocessed accoridng to the analysis from EDA and Correlation. The featurs have also been standardised using StandardScalar (subtract from mean and divide by sd). Dummy variables have been created for categrical variables which will help in out analysis.

MODELLING

The data has been split and we have used gradient algorithm for linear regression for finding coefficients.(β values). The hyperparameter used and we plan to tune is learning rate (∝).

y_hat = β0 + β1 * Hour + β2 * Temperature(°C) + β3 * Humidity (%) + β4 * Wind speed (m/s) + β5 * Visibility (10m) + β6 * Dew point temperature(°C) + β7 *Solar Radiation (MJ/m2) + β8 * Rainfall(mm) + β9 * Snowfall (cm) + β10 * Month + β11 * Day + β12 * Weekday + β13 * Seasons_Autumn + β14 * Seasons_Spring + β15 * Seasons_Summer + β16 * Seasons_Winter + β17 * Holiday_ Holiday + β18 * Holiday_No Holiday + β19 *Functioning Day_No + β20 * Functioning Day_Yes

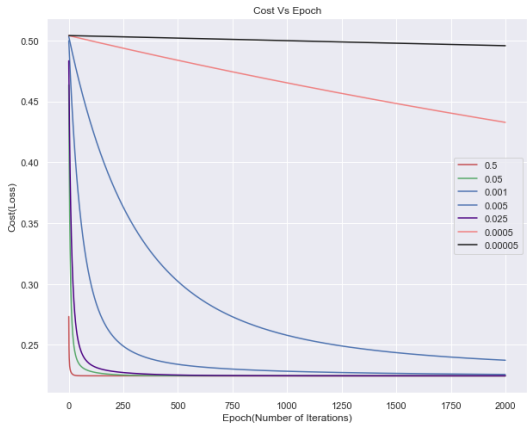



```
In [61]: beta_value_new

Out[61]: intercept                  0.007192
         Hour                       0.308761
         Temperature(°C)            0.293184
         Humidity(%)               -0.265913
         Wind speed (m/s)           0.024252
         Visibility (10m)           0.046082
         Dew point temperature(°C)  0.156765
         Solar Radiation (MJ/m2)   -0.067071
         Rainfall(mm)              -0.105022
         Snowfall (cm)              0.012802
         Month                      0.013267
         Day                        0.004077
         Weekday                   -0.026621
         Seasons_Autumn             0.101035
         Seasons_Spring             0.029909
         Seasons_Summer             0.024818
         Seasons_Winter            -0.156541
         Holiday_Holiday           -0.022317
         Holiday_No Holiday         0.022317
         Functioning Day_No        -0.129734
         Functioning Day_Yes        0.129734
         dtype: float64
```

EXPERIMENTATION

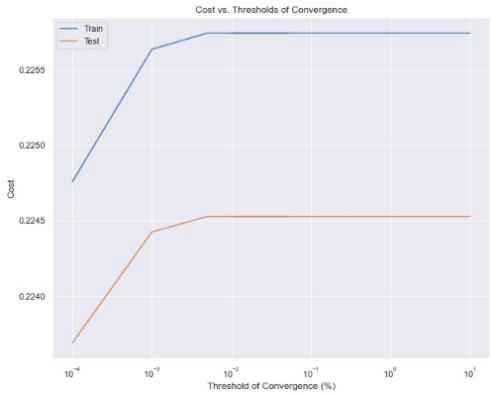
1.Change learning rate



For very low learning rate, the number of iterations required to reach convergence is high. We can see that for learning rate of 0.5, the algorithm will diverge and not reach the minimum point. The lower the **loss,** the better a model (unless the model has over-fitted to the training data). The loss is calculated on **training** and its interpretation is how well the model is doing for the set. Unlike accuracy, loss is not a percentage. It is a summation of the errors made for each example in training or validation sets.

Here, we can see that for each iteration, the cost is decreasing. The convergence has been met within the threshold.

2. Change Threshold For Convergence



Threshold specifies the minimal allowed movement in each iteration. If the cost function update in the current iteration is less than or equal to tolerance, the iteration is stopped as this is near the minimum, where gradients are usually very small. it can also happen near a local minimum.

Here, gradient descent is set to be converged if the change in cost (J) is within the threshold. We can see here at increasing threshold, the train as well as test data also seems to have an increase in the cost. Ideally, the algorithm also converge fast at higher threshold.

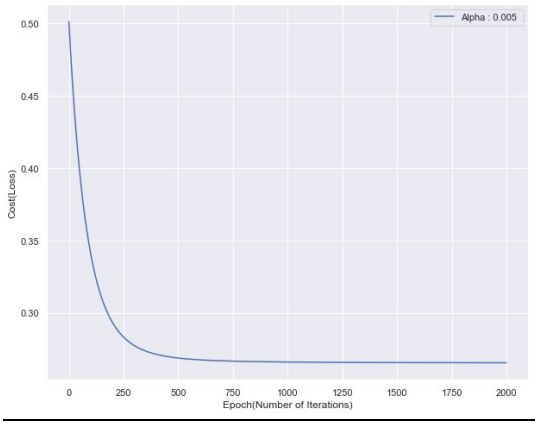
Tolerance	Train Error	Test Error
0.0001	**0.22475**	**0.2236**
0.001	0.22563	0.2244
0.005	0.2257	0.2245

Minimum error was found for tolerance at 0.2236.

3. Selection of 8 Random Features

8 features were selected at random. The train and test errors are compared with the original analysis (that had all the features). We could see that random feature selection tend to have more error than other features that comprosed 21 variables.

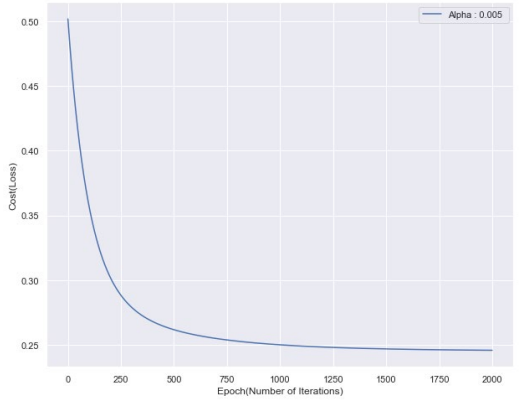
y_hat = β0 + β1 * Hour + β2 * Temperature(°C) + β3 * Humidity (%) + β4 * Dew point temperature(°C) + β5 * Rainfall(mm) + β6 * Snowfall (cm) + β7 * Holiday_ Holiday + β8 * Holiday_No Holiday



	Random Variables	Original Features	% Change
Train Error	0.53384	0.4514	18.26%
Test Error	0.53385	0.449	18.89%

4.Selection of 8 Suited Features

y_hat = β0 + β1 * Hour + β2 * Temperature(°C) + β3 * Humidity (%) + β4 * Wind speed (m/s) + β5 * Weekday + β6 * Seasons_Summer + β7 * Holiday_ Holiday + β8 * Functioning Day_Yes



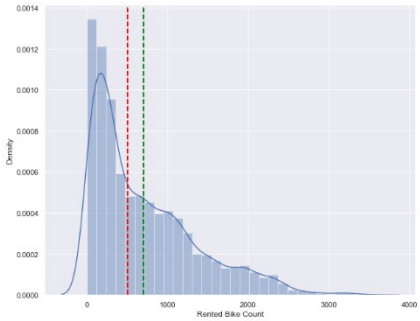
	Random Variables	Original Features	% Change	Suited Features	% Change
Train Error	0.53384	0.4514	18.26%	0.4917	8.92%
Test Error	0.53385	0.449	18.89%	0.4869	7.99%

The train and the test set have a percent change of 8.92% and 7.99% respectively. This shows that the selected (suited) features have performed well when compared to randomly picked features. This shows that features chosen as to data analysis (which shows how they are correlated with one another, domain knowledge etc) plays factor for good algorithm.However, we could see that the performance in terms of original features is slightly less as we have not taken all features and reduced the number of variables. This can lead to not having good interpretation of the data. Ultimately, this can be an impact for error to go high and not have a better algorithm as original feature selection.

CONCLUSION

Data Pre-Processing is very important in the analysis of having a good algorithm. This can help in reducing overfitting and having better accuracy of the model. It is very important in optimizing bias variance trade off in machine learning. But it is always important to not have restricted data as well. Features play a vital role in our analysis, and it is important that the features chosen are not by random and are chosen based on our knowledge or analysis of data. EDA can definitely help in influencing a positive selection of the features. Hyperparameters like learning rate and threshold are parameters that can be set by the user before starting training. They can have a big impact on model training as it relates to training time, infrastructure resource requirements (and as a result cost), model convergence and model accuracy.

From our analysis , we can see that **Temperature** and **Hour** are the major factors that help in identifying the increase/decrease in bike counts. Another primary factor is the Functioning Day which helps to identify that the bikes have an impact. As temperature is high, there is an expected increase in bike usage within the area. Snowfall and Rainfall fields have been ignored as it is positively skewed to 0. Dew Point Temperature field has been omitted as to high correlation with Temperature field.



Regarding dependent variable Rented Bike Count, we could see from below figure that Rented Bike Count variable is skewed and we have normalized the data as to which the analysis has been conducted. Correlation between variables and dummy variables for categorical variables have been conducted for this analysis. In addition, the Date field has been split to check to check on which day (weekend/weekday) the bike count is high. In addition to this, we can perform analysis using different models like Logistic Regression, Random Forest to check which model can provide a better accuracy and better information for the data set. Currently, we have split the dataset to 70-30 configuration. Training data can be increased as to which it can have more information on our analysis. Finding the optimum value of each parameter helps a lot in the analysis. For example, learning rate, tolerance etc can be varied and tested to get the best value that fits the model. Regularization can be a better factor in feature selection. Outliers have not been considered here. Better handling of outliers can also help in the analysis.