Linear and Logistic Regression

Dataset Description and Summary Statistics

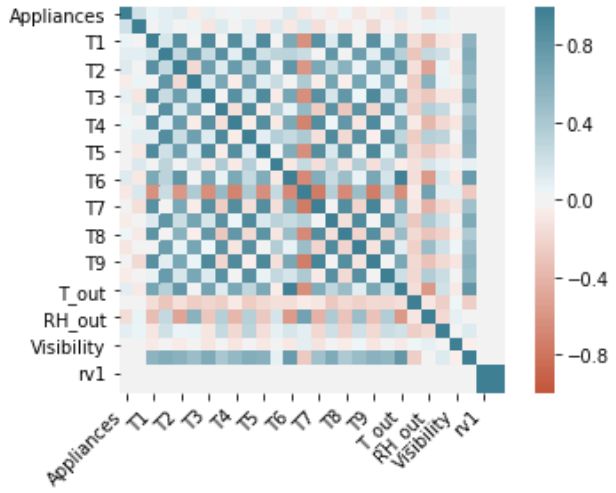
A brief description of the data set is as following:

• Dataset consists of 19735 observations on 29 attributes.

• The dependent variable for the linear regression model is 'Appliances': Energy use in Wh

• None of the column contains any missing value, so no missing value imputation is required

Correlation: The Correlation Plot below shows us the correlation between the independent variables.



Correlation with more than 0.88		
Corr Value	Fearture1	Feature 2
0.8924022851105534	T3	T1
0.8803585351891977	RH_4	RH_1
0.8989782902830947	RH_4	RH_3
0.8852468734266851	T5	T1
0.8881689127749756	T5	T3
0.8943012488817869	RH_7	RH_4
0.8821232262844226	T8	T7
0.8839839643177134	RH_8	RH_7
0.9013235850825261	T9	T3
0.8894391511402565	T9	T4
0.9110551178067879	T9	T5
0.9447764235687138	T9	T7
0.9747866900664592	T_out	T6

With Correlation threshold at 0.88 we find the above features highly corelated. Hence, we drop:

```
'date', 'T3', 'RH_3', 'T_out', 'rv1', 'rv2', 'T4','RH_4', 'T7', 'RH_7', 'T5'
```

Additional features Dropped:

Lights: High volume of values as '0' does not help the prediction. [80% of values]

Visibility: Correlation of the feature with Dependent Variable is too low. [0.000230]

Value Counts of Lights	
0	15252
10	2212
20	1624
30	559
40	77
50	9
60	1
70	1

Name: lights, dtype: int64

Box Plot of Dependent Variables [Appliances]

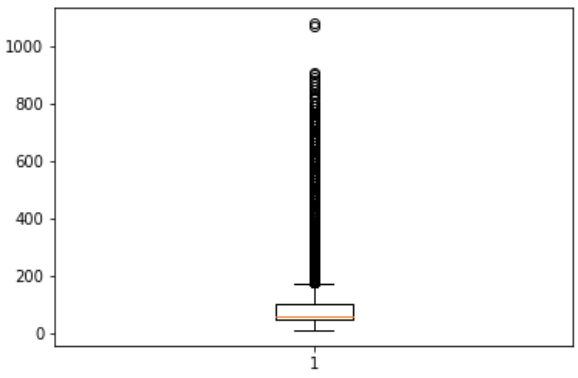
Box Plot:

The IQR for Appliances using the Box plot is at 175 ie

75% -100 50%- 60 25% -50

IQR = (100 -50) *1.5 = 175

Outliers threshold is set at 175



Tasks

The linear regression equation to model the Energy use:

$$App = \beta_0 + \beta_{T1} * T1 + \beta_{RH\_1} * RH\_1 + \beta_{T2} * T2 + \beta_{RH\_2} * RH\_2 + \beta_{RH\_5} * RH\_5 + \beta_{T6} * T6 + \beta_{RH\_6} * RH\_6 + \beta_{T8} * T8 + \beta_{RH\_8} * RH\_8 + \beta_{T9} * T9 + \beta_{RH\_9} * RH\_9 + \beta_{Press} * Press + \beta_{RH\_out} * RH\_out + \beta_{Windspeed} * Windspeed + \beta_{Tdewpoint} * Tdewpoint$$

The initial parameters are: Beta: 0.5 Number of iterations: 1000 Learning Rate: 0.01

Experiments

Experiment 1:

Linear Regression:

Experimenting with various paraments such as learning rate:

Choosing three learning Rates: [0.007, 0.005, 0.010]

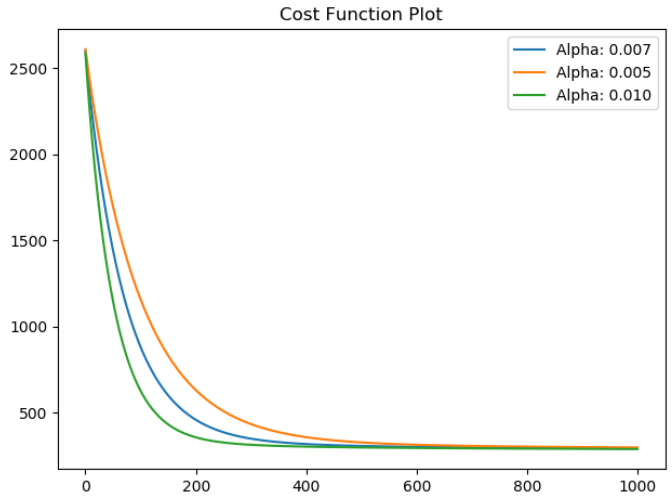
Training Dataset:

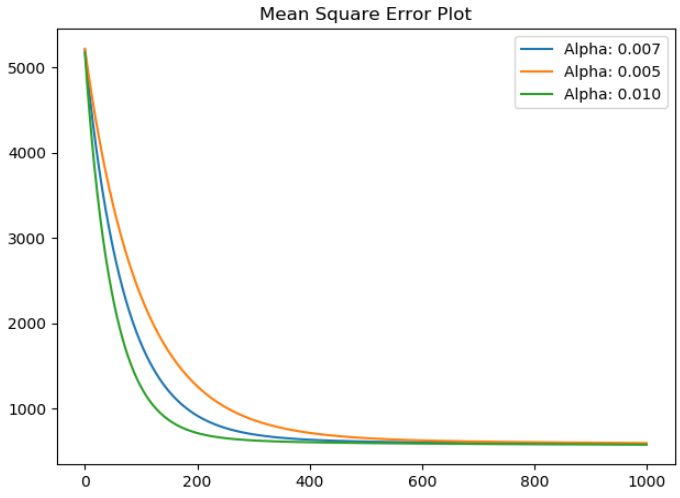
The Model was tested with varying learning rates and the best learning rate was found to be 0.01 as the number of iteration taken to complete the regression is the least while producing the best accuracy with its R square being the highest among the rest of the Learning Rates and the Cost Function being the Least.

```
Learning rate: 0.007, Iteration of Convergence: 630
Learning rate: 0.005, Iteration of Convergence: 882
Learning rate: 0.010, Iteration of Convergence: 440
```

```
Optimal Values for Training Data
Measures for Training Data
              CF        MSE       MAE       R²
0.005   297.5127   595.0255   18.1099   0.2726
0.007   292.2345   584.4690   17.9992   0.2855
0.010   288.1791   576.3583   17.8707   0.2954
```

Varying Learning Rate





Test Dataset:

The varying learning rates are tested with the test data set, in this case too as you can seem from the below images, 0.01 was the best Rsquare and the lowest Cost Function thus making it the best Rsquare

You can also observe that as the Learning Rate increases , the vost fuction decreses.

Measures for Test Data

	CF	MSE	MAE	R^2
0.005	295.9177	591.8353	18.0782	0.2555
0.007	290.8510	581.7021	17.9728	0.2682
0.010	286.8035	573.6070	17.8356	0.2784

Logistic Regression:

Classes of the Dependent variable were classified using the median.

```
[0 if x <= 60 else 1
```

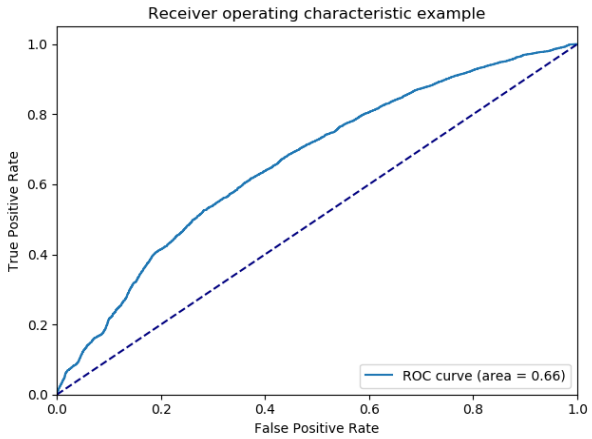
Logit Model applied on the Training Datasets has given me the following Metrics:

The learning rate was also verying with values such as 0.001, 0.005, 0.010, 0.050

The Value for the lowest Cost Function was 0.01 which converged after 2000 iterations.

Accuracy of the model in training was 26.59%

```
Area under the ROC curve: 0.663
```



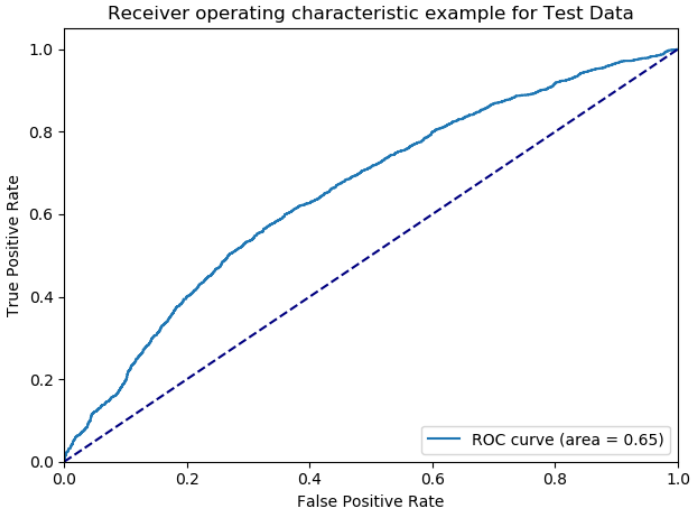

```
Confusion Matrix
[[4832 2689]
 [1919 2877]]
Sensitivity:      0.5998748957464554
Specificity:      0.6424677569472145
     Accuracy Score:
     0.6258829260371844
```

Test:

The learning rate used when testing was 0.01, which produced the best cost function.

Accuracy of the Model on Test Data was 62.1%

```
Confusion matrix for Test
[[2034 1189]
 [ 812 1245]]
Sensitivity: 0.6052503646086533
Specificity: 0.6310890474713
     Accuracy Score:
     0.62102272727272
```



Experiment 2:

The different thresholds of cost function were implemented on the train dataset.

As the threshold decreses the accuracy of the model increases. The Cost Function decreases with decreasing threshold.

The iterations required to converge significantly decreased with Threshold value.

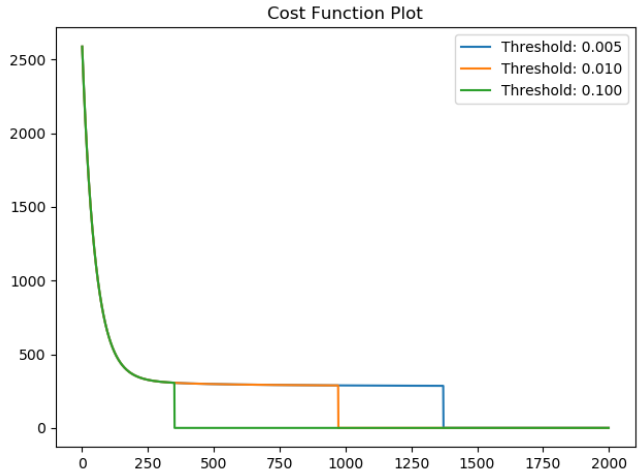
Train:

```
Learning rate: 0.010, Iteration of Convergence: 1371,
Threshold: 0.005
19:23:22.516451
Learning rate: 0.010, Iteration of Convergence: 972,
Threshold: 0.010
19:23:40.444699
Learning rate: 0.010, Iteration of Convergence: 350,
Threshold: 0.100
```

Optimal Values for Training Data Experiment 2

Measures for Training Data Experiment 2

	CF	MSE	MAE	R^2
0.005	285.5945	571.1889	17.7862	0.3018
0.010	288.4433	576.8866	17.8793	0.2948
0.100	306.1593	612.3187	18.1112	0.2515

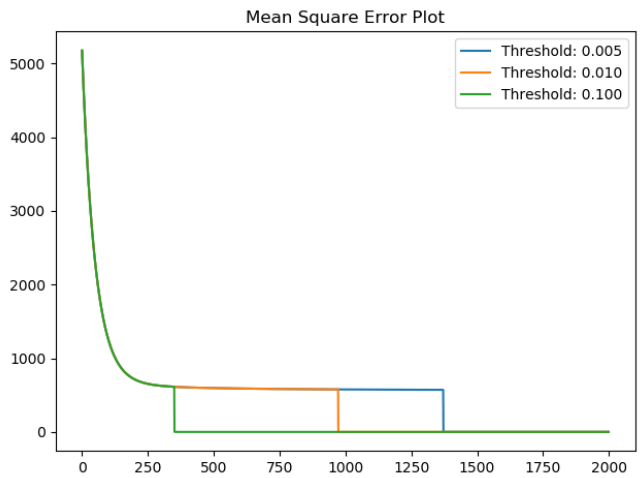


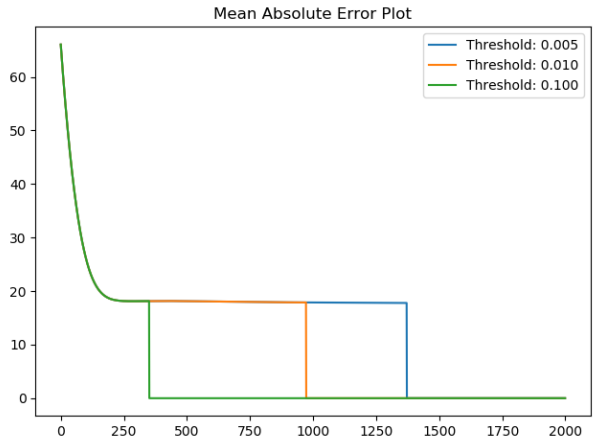
The graph shows the Cost Function with different thresholds

X- Axis: No. Of Iterations

Y: Axis: Cost Function

As you can see in the graph the Cost Function speets to an almost zero after it hits the Threshold.





Tha above graph also shows the best threshold and its graph plotted as a function of the number of iterations.

Test Data:

Using the different threshold values the measure of cost function and the r square value.

We can interpret that as we vary the threshold the cost function decresases with decrease in Threshold value

The R square on the other hand increases with decreasing threshold value, which holds true to the training data as well.

On varying different thresholds

```
Measures for Test Data
              CF       MSE      MAE      R²
0.005  284.1312  568.2624  17.7356  0.2851
0.010  287.0727  574.1455  17.8455  0.2777
0.100  304.0246  608.0492  18.0629  0.2351
```

Experiment 3:

Linear regression:

Randomly picking the values:

```
Random Column picked are:
Index(['T2', 'T9', 'Visibility', 'T1', 'T4', 'RH_1', 'RH_7
', 'T7', 'T_out',
       'T6'],
      dtype='object')
```

The R square when randomly picked when compared to the 15 features picked shows a decrease ie the model is less accurate than the Optimized model.

Though the outliers were removed in both the cases, the features had a lot to do with the accuracy being down.

As the Visibilty, T_out are some feature which has highly corelated and decrease the accuracy of the model.

```
Optimal Values for Training Data for Experiment 3
Measures for Training Data for Experiment 3
            CF       MSE      MAE      R²
0.01  331.7969  663.5938  19.4359  0.1888
Testing on test data for Experiment 3
Measures for Test Data for experiment 3
           CF       MSE      MAE    R²
0.01  329.883  659.7659  19.4493  0.17
```

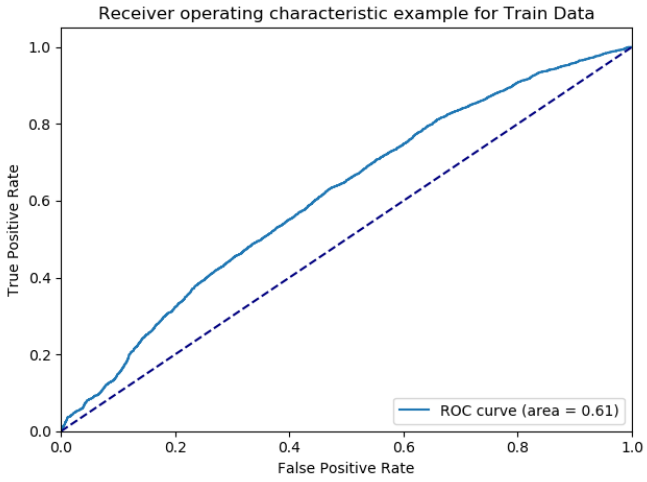
Logit:

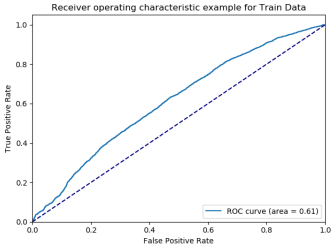
The Learing rate for the random variables was 0.01

The Rsquare like expected is less than the train dataset and the previously modeled data.

```
Optimal Values for Training Data for Experiment 3
Measures for Training Data for Experiment 3
             CF       MSE      MAE      R²
0.01   331.7969   663.5938   19.4359   0.1888
Testing on test data for Experiment 3
Measures for Test Data for experiment 3
            CF       MSE      MAE     R²
0.01   329.883   659.7659   19.4493   0.17
[4, 18, 24, 2, 8, 3, 15, 14, 20, 12]
The Random features picked are:
Index(['T2', 'T9', 'Visibility', 'T1', 'T4', 'RH_1', 'RH_7
', 'T7', 'T_out',
       'T6'],
      dtype='object')
```

Logistic Regression





Experiment 4

Linear Regression:

Train':

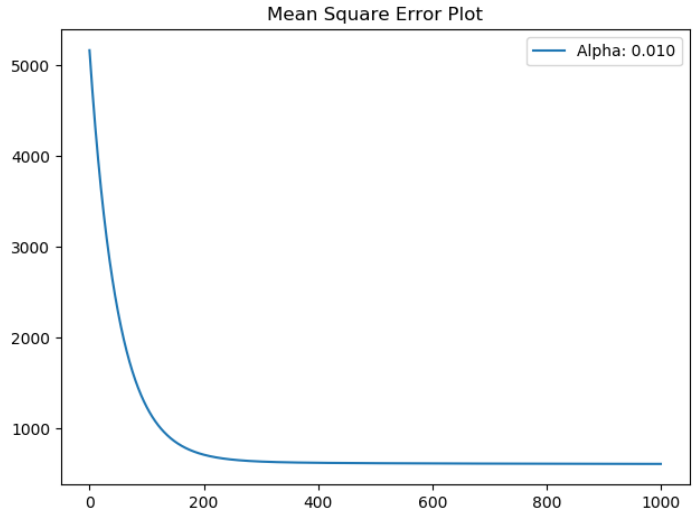
```
Optimal Values for Training Data
Measures for Training Data
            CF       MSE       MAE       R²
```

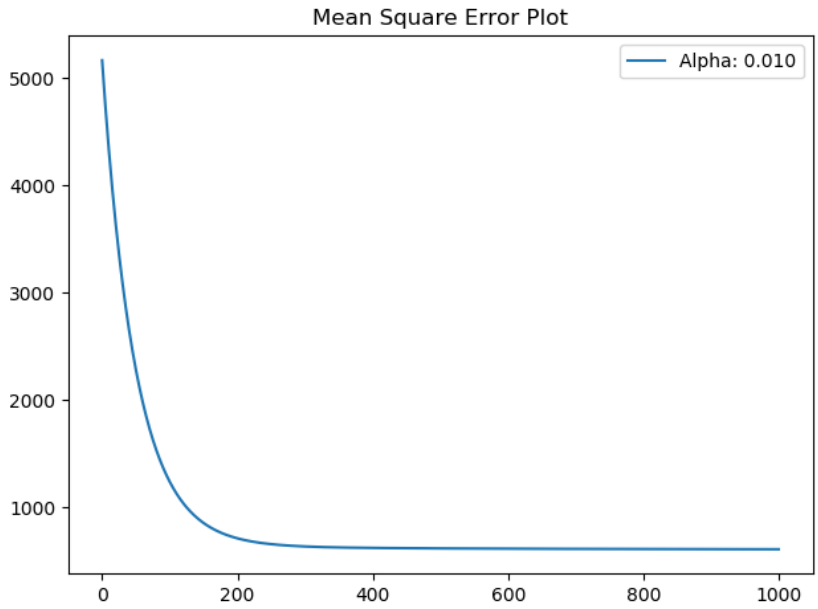

known

```
0.01   305.351   610.702   18.2914   0.2535
```

Test:

```
Testing on test data
Measures for Test Data
             CF        MSE       MAE       R²
0.01   303.3763   606.7527   18.4842   0.2367
```




Mean Square Error Plot
Alpha: 0.010
5000
4000
3000
2000
1000
0
200
400
600
800
1000