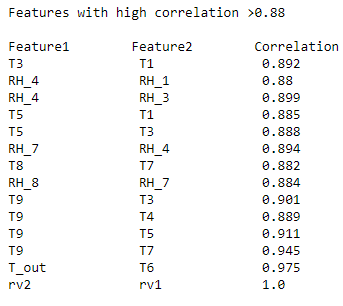
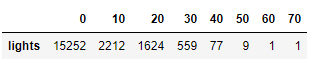
**Exploratory Data Analysis**

Factors considered for dropping the features

Features date, rv1, rv2 are just random and do no have any relation to the target variable – energy usage.

****

Count of Values for Feature = lights

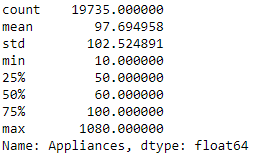
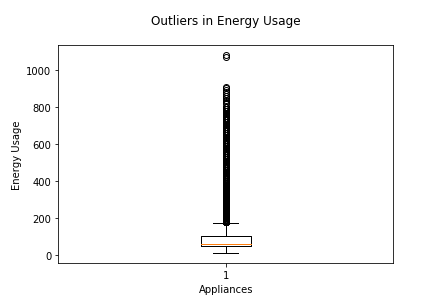
****

Lights has almost 80% (15252) of its value to be 0. Lights feature can be removed.

From the above interpetation, columns

'date','lights','T3','T\_out' ,'RH\_4','T4','T7','RH\_7','T5', 'rv1','rv2' can be dropped from the dataset.

**Outliers Detection and Removal**



Thresold for Outlier Point

Upper Bound = (1.5\*IQR) + 75th percentile, Lower Bound = (1.5\*IQR) - 25th percentile

Upper Bound = 1.5\*(100-50) + 100 = 175, Lower Bound = 1.5\*(100-50) - 50 = -25

Number of outliers = (count of values above the upper bound) + (count of values below the lower bound)

Number of outliers = 2138 which is 10.8% of the total observations

***We may not consider deleting all the outlier observations, since 10.8% of the total observations is considerably high. But the model accuracy and metrics drastically increased after the removal of these outliers. So, we are deleting the outlier observations in this situation.***

**Experimentation 1: Linear Regression**

Accuracy/Error variation with different Learning Rates

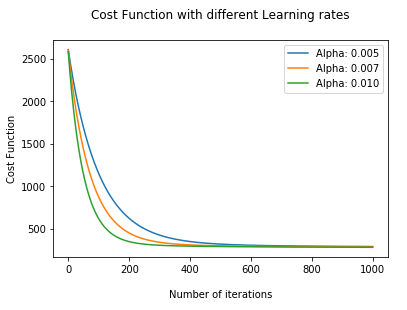
**Learning Rates used: 0.005, 0.007, 0.01, Initial betas: 0.5, Number of iterations: 1000**

Metrics considered: Cost Fuction, Mean Squared Error, Mean Absolute Error, R-Squared

Below are the Gradient Descent Results

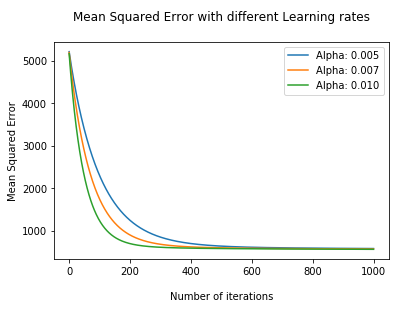
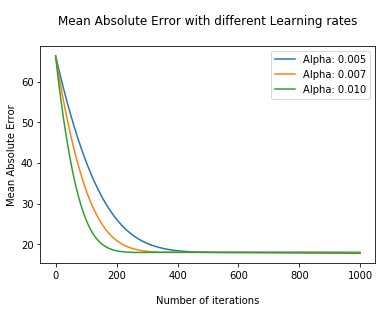
Cost Function Convergence (When difference in Cost Fuction for each iteration was <=0.1)

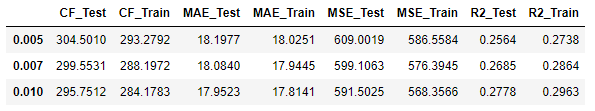


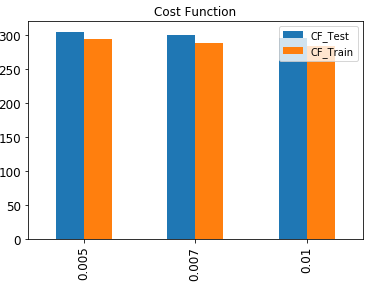


Cost Function converges quickly with increasing learning rates or increasing step sizes

Model with Alpha 0.01 performs the best as it converges quickly.





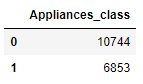
From the results, Cost Fuction and the other errors are the lowest when learning rate = 0.01 and hence the best model.

Energy Usage = 67.1349 + β1\*T1 + β2\*RH\_1 + β3\*T2+ β4\*RH\_2 + β5\*RH\_3 + β6\*RH\_5 + β7\*T6 + β8\*RH\_6 + β9\*T8 + β10\*RH\_8 + β11\*T9 + β12\*RH\_9 + β13\*Press\_mm\_hg + β14\*RH\_out + β15\*Windspeed + β16\*Windspeed + β17\*Tdewpoint

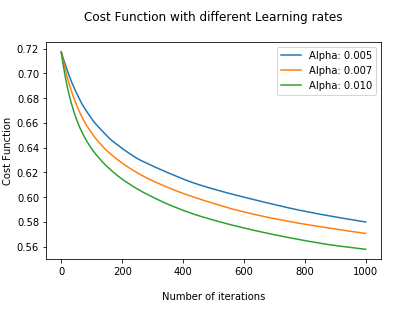
β0=67.1349 β1=2.9932 β2=11.2539 β3=3.4322 β4=-0.553 β5=1.1047, β6=2.7928, β7=2.7835,β8=1.9768,β9=8.9667,β10=-9.7271,β11=-10.2668,β12=-8.9756,β13=-0.7374,β14=-0.5899, β15=2.2409,β16=0.1185,β17=0.9451

**Logistic Regression**

Transforming the energy usage into classes based on its median. Median is 60, so if energy usage <= 60 then class 0, else class =1



Gradient Descent

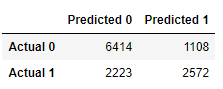
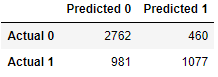
****

Plot shows the Cost Fuction for different learning rates after 1000 iterations. Cost Function converges quickly with increasing learning rates.

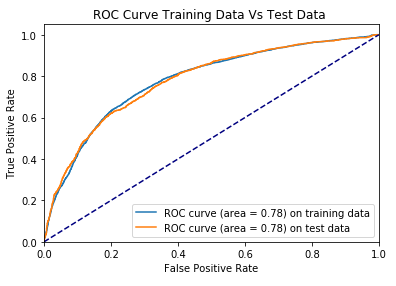
Model with Alpha = 0.01 performs the best.

**Confusion Matrix**

**Training Data Test Data**

**Accuracy: 0.729 Accuracy: 0.727**

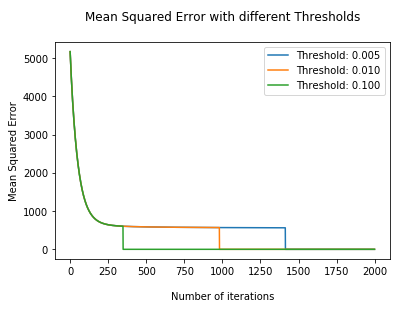
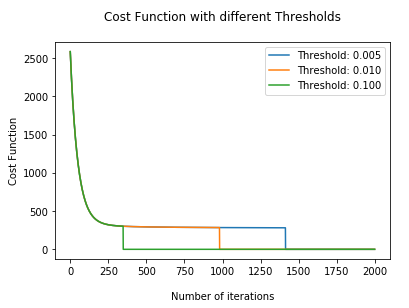


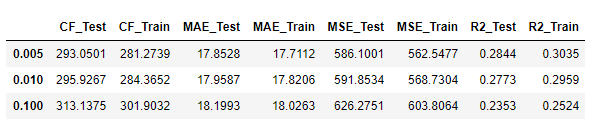
β0=-0.437 β1=0.0798 β2=0.3029 β3=0.2156 β4=-0.0554 β5=0.0445,β6=0.2512,β7=0.1891,β8=0.0643,β9=0.2423,β10=-0.3875,β11=-0.2166,β12=-0.3899,β13=-0.106,β14=-0.1542, β15=0.1746,β16=0.0261,β17=-0.0095

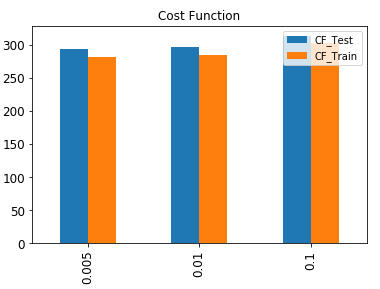
**Experimentation 2: Varying thresholds for Cost Function**

**Thresholds used: 0.005, 0.01, 0.1, Initial betas: 0.5, Alpha: 0.01 Number of iterations: 2000**

****

****

****

****

As the threshold decreases, the number of iterations required to achieve it increases.

Model with the lowest threshold 0.005 performs the best as it takes the highest number of iterations.

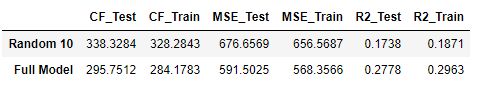
**Experimentation 3: Choosing 10 Random Variables**

**Linear Regression:**

**Initial betas: 0.5, Alpha: 0.01 Number of iterations: 1000**

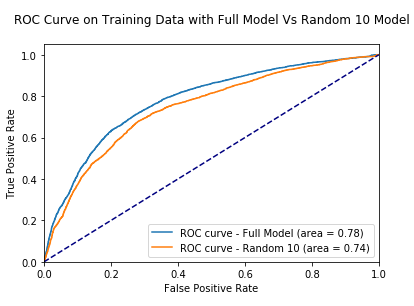
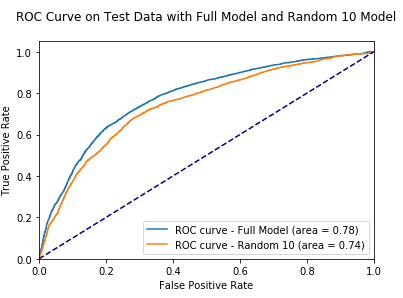
The ten random features selected are 'RH\_2', 'RH\_9', 'Tdewpoint', 'RH\_1', 'RH\_4', 'T2', 'T8', 'RH\_7', 'Press\_mm\_hg', 'RH\_6'. These features are selected using the Random Fuction.

Comparison of Train and Test results for Exp: 1 Model and Exp: 3 Model



**Logistic Regression**

|  |  |  |
| --- | --- | --- |
| Model | Training Accuracy | Test Accuracy |
| Experiment 1 Model | 0.729 | 0.727 |
| Experiment 3 Model | 0.706 | 0.708 |

For both Linear and Logistic Regressions, Experiment 1 Model (>15 features) performs better than the Experiment 3 Model (10 Features randomly selected)

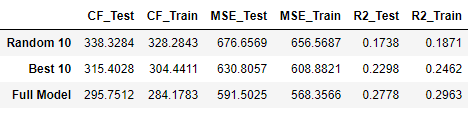
**Experimentation 4**

Best 10 features were selected based on each feature correlation with the target variable, based on increasing order of collinearity.

The best selected features are 'T1', 'T2', 'RH\_2', 'T6', 'RH\_6', 'T8', 'RH\_8', 'T9', 'RH\_9', 'RH\_out'

Full Model here represents the experiment 1 model, with greater than 15 features.

**Linear Regression Metrics**



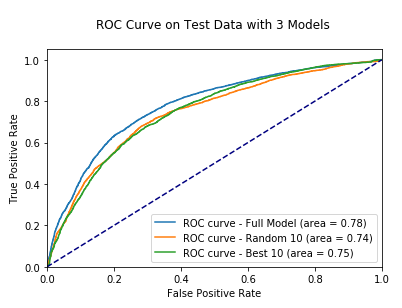
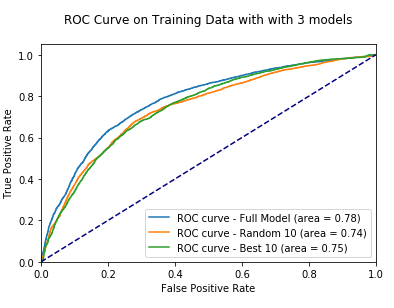
**Logistic Regression Metrics**

|  |  |  |
| --- | --- | --- |
| Model | Training Accuracy | Test Accuracy |
| Experiment 1 Model | 0.729 | 0.727 |
| Experiment 3 Model | 0.706 | 0.708 |
| Experiment 4 Model | 0.6985 | 0.695 |

For both Linear and Logistic Regression, Experiment 1 model (>15 features) performs better than the Random 10 model and the Best 10 model.

This is because the Experiment 1 model has more features that can contribute towards the variation in Energy Usage. Or in other words we can say that the dataset has more than 10 features that are correlated with energy usage.

And with the best 10 model, the number of highly correlated features with target variable has been restricted to 10. So, it is always a good practice to include as many correlated features in the model to better the model metrics.



**Discussion**

**Linear Regression:**

When the learning rate = 0.01, the model gave us the best results on the testing data with Cost Function: 295.75 and R-Squared: 0.28.

With R-squared to be only 0.28, only 28% of the variation in Energy Usage is casued by the features in the dataset. This also means, there could be other factors that are contributing towards the energy usage but not present in the model. These unobserved features are captured in the error term leading the cost funtion contribution.

To improvise the model metrics and accuracy, transformation on some of the feature variables can be performed like including the logarithm, squaring, cubing of the features. In conclusion, this dataset does not perform the best on Linear Regressions. Other regression algorithms like Lasso, Ridge RandomForestRegressor can be implemented to improvise the accuracy and the metrics.

Some of the major contributors of energy usage were Temperature in teenager room (T8), Temperature in Living Room area (T2), Humidity Outside (RH\_out), Temperature in Kitchen area (T1), Humidity outside the building (RH\_6)

**Logistic Regression**

With respect to Logistic Regression, model with learning rate of 0.01 gave us the best results with accuracy close to 73%. The model has two classes: 0, when Energy Usage was less than or equal 60 and class 1, when Energy Usage was greater than 60. Although Logistic Regression gave us a good accuracy, other classification models like RandomForestClassifier, Support Vector Machines, Nearest Neighbours models could have given even more better results.