# Appliances Energy Prediction

## Introduction:

In the era of smart homes, the ability to predict energy consumption not only save money for users but also help in generating money for the user by giving excess energy back to Grid (in case of solar panels usage). In this case regression analysis will be used to predict Appliance energy usage based on data collected from various sensors.

The energy prediction will come under supervised machine learning task aiming to Appliance energy consumption for a house based on factors like temperature, humidity & pressure. Many techniques, Gradient descent algorithm, and linear regression (in built function) have been applied to credit predict the energy consumption.

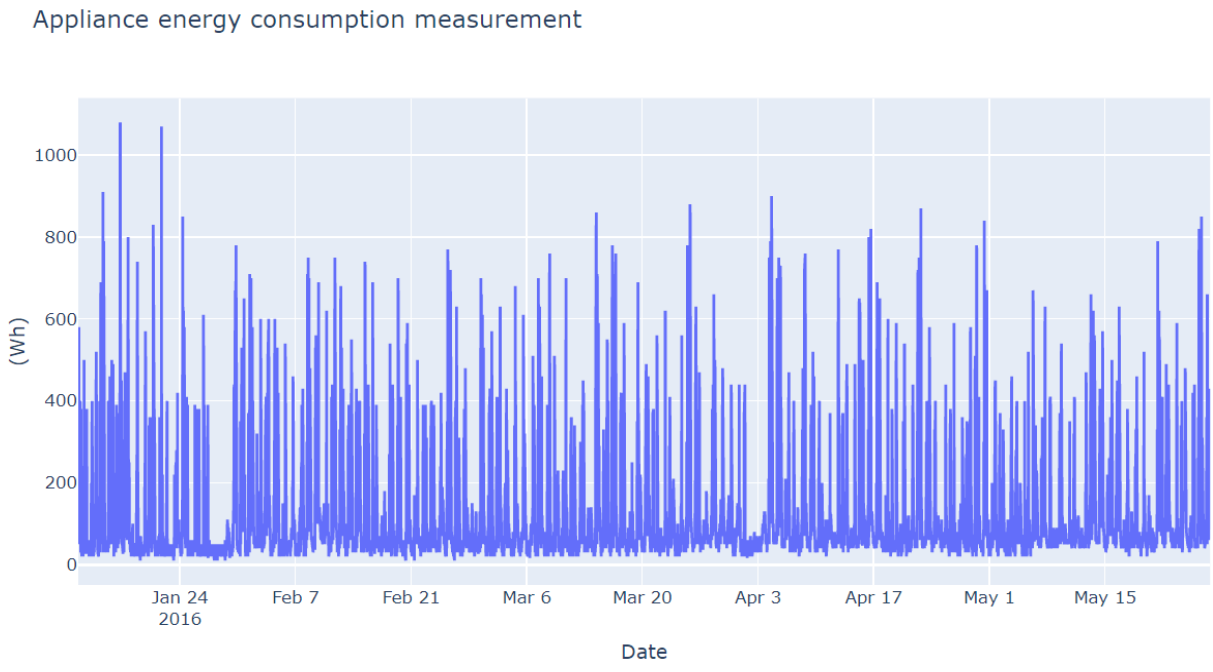
## Dataset:

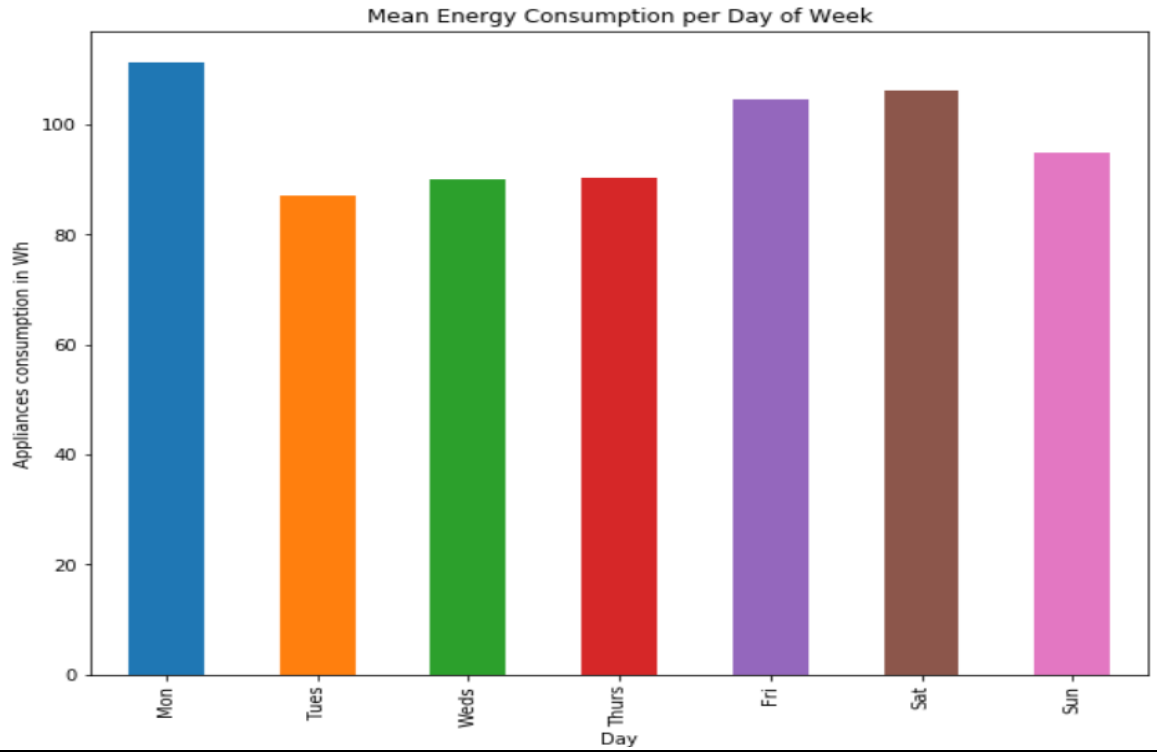
The dataset (Appliances Energy Prediction) dataset is download at:
https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction

The data set is at 10 min for about 4.5 months. The house temperature and humidity conditions were monitored with a ZigBee wireless sensor network. Each wireless node transmitted the temperature and humidity conditions around 3.3 min.

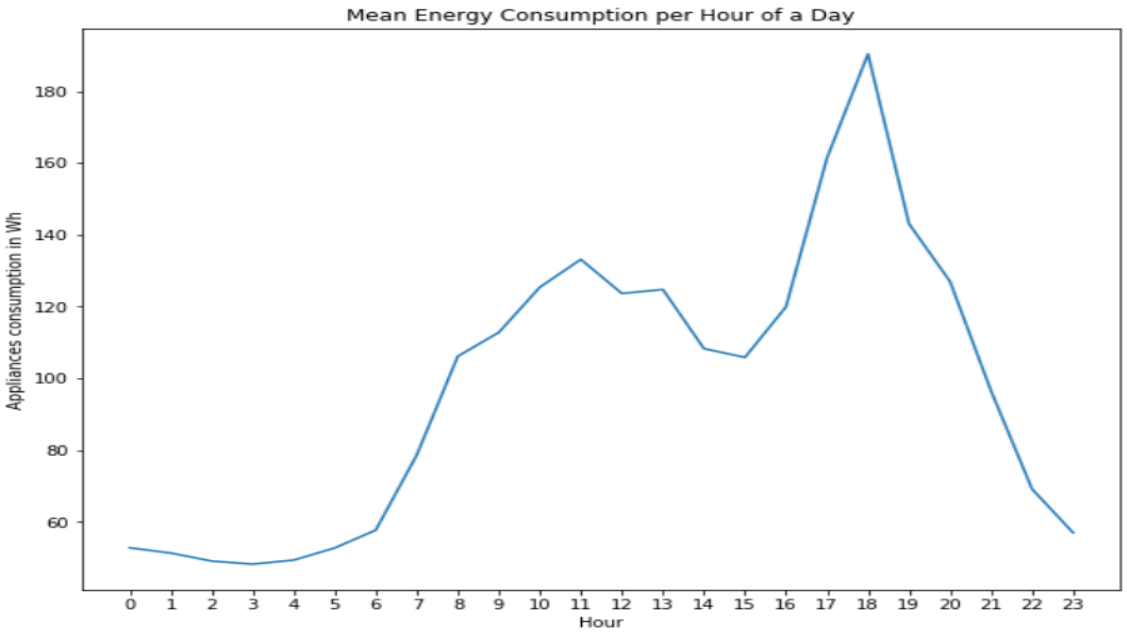
The dataset has 19735 rows and 29 variables include lights, date, Temperature and Humidity in various places in the house, pressure etc., The number of missing values and null values is zero. Number of weekdays is 14263 and weekend (Saturday and Sunday) is 5472.

## Energy Consumption:

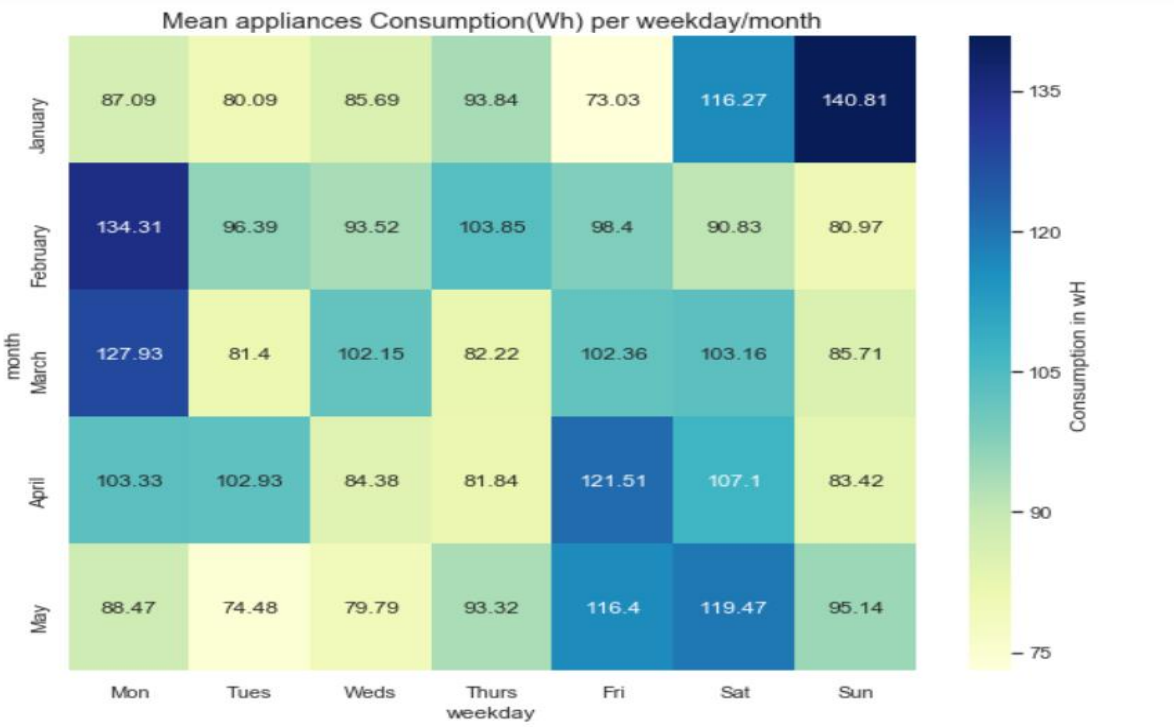





**The power load is a bit higher on Monday, Friday, Saturday and Sunday than the other days.**




At night hours from 22:00-7:00 the power load is below 80Wh, meaning that most appliances are off or standby. Between 9:00-13:00 the power load is 120-135Wh and after launch reduces again to 110Wh. At afternoon, the energy consumption ranges from 130-185Wh as family members are at home and many devices are on.




We can see from the heatmap above that the assumption that more power is consumed on Monday, Friday, Saturday and Sunday is valid for each month. However, in our data set we have only 4,5 months and therefore we cannot use months as feature for our model.

**Feature Range:**

1. Temperature : -6 to 30 deg

2. Humidity : 1 to 100 %

3. Windspeed : 0 to 14 m/s

4. Visibility : 1 to 66 km

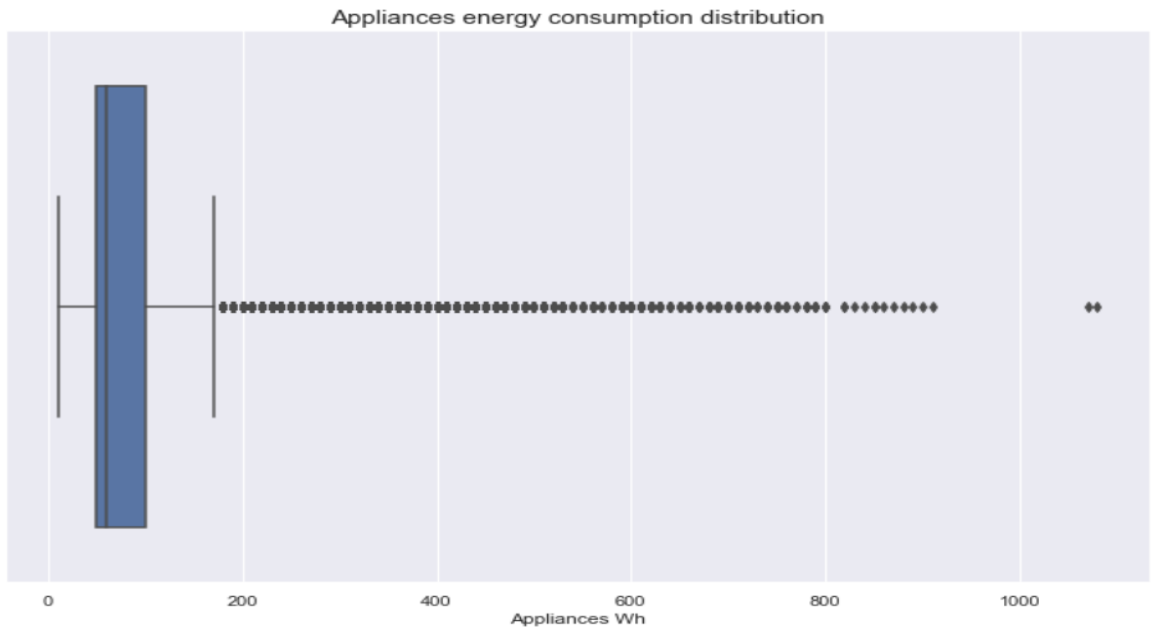
5. Pressure : 729 to 772 mm Hg

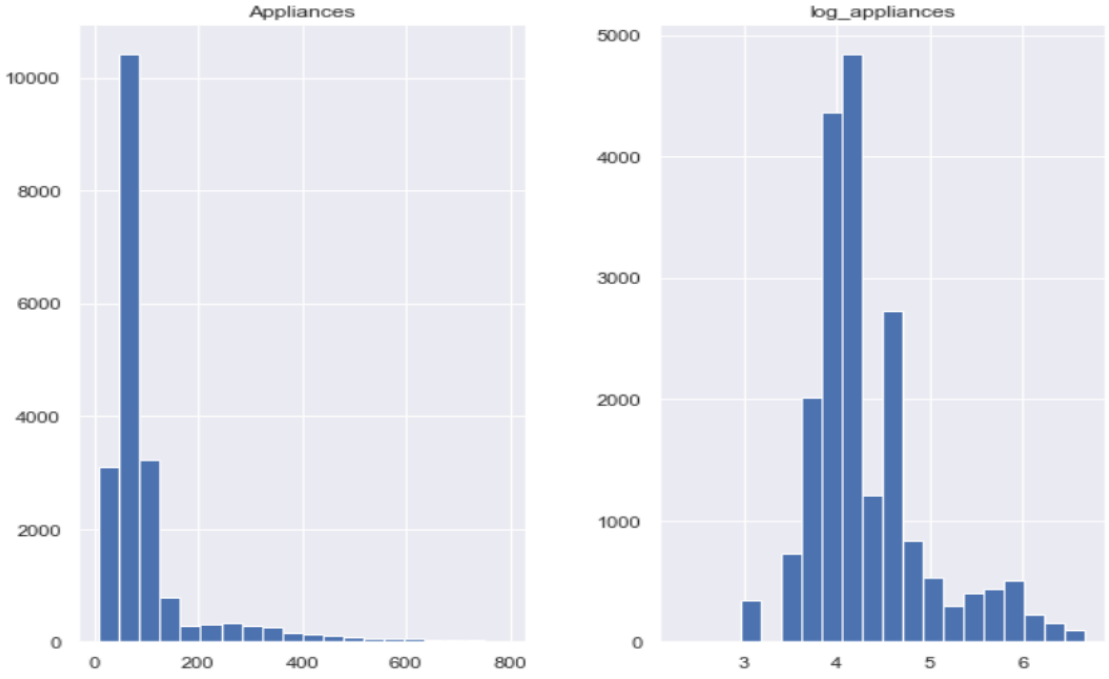
6. Appliance Energy Usage : 10 to 1080 Wh

**Outlier Detection for Appliance Energy Usage:**

The number of the 0,1% top values of appliances' load is 19 and they have power load higher than 790 Wh. So, we remove the instances above than 790Wh.

The idea is that appliances' load is hour, day, week, month dependent. It is logical assumption that in night hours the load is low or at weekends the energy consumption is higher than the weekdays because more people are at home.

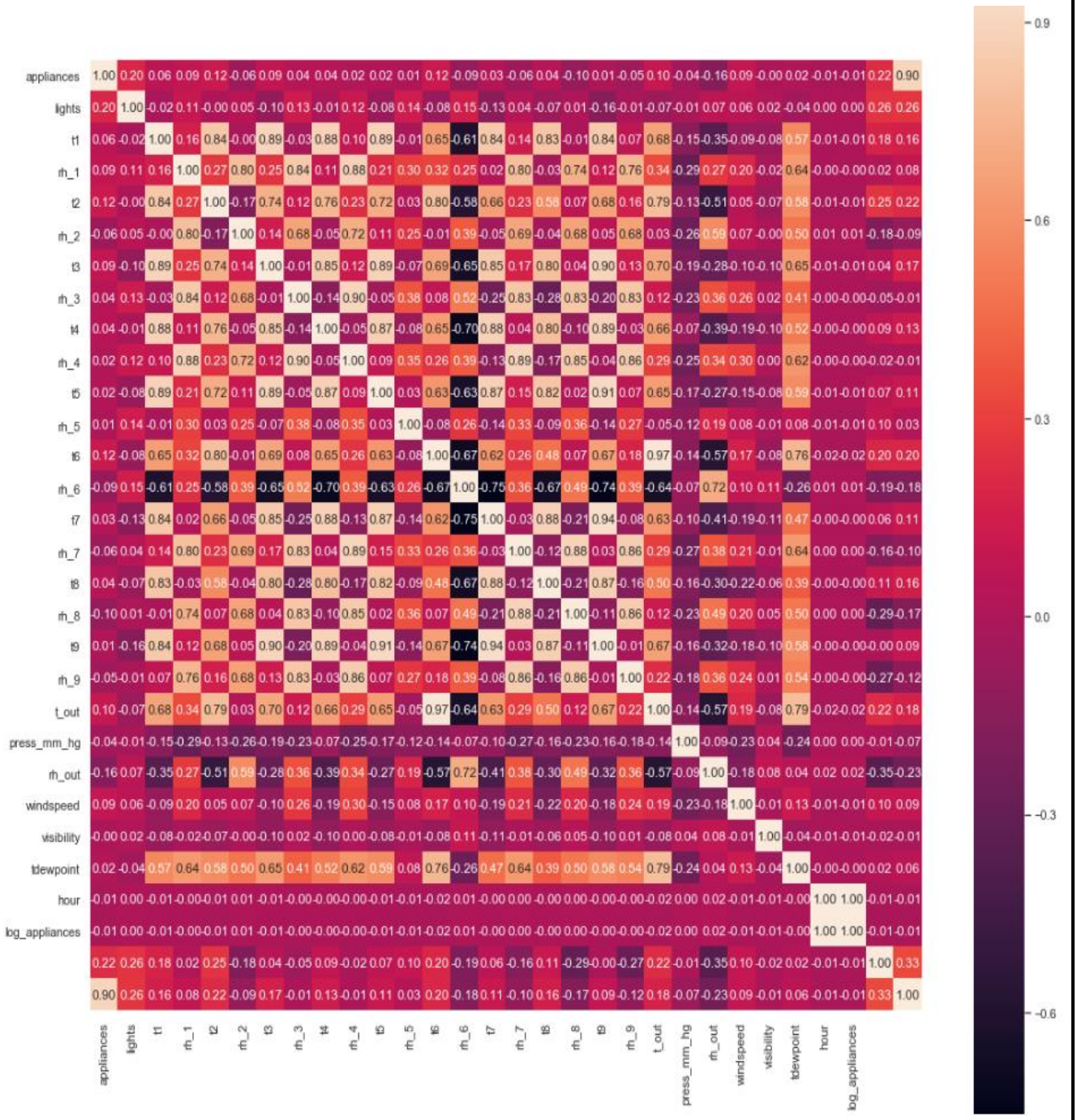


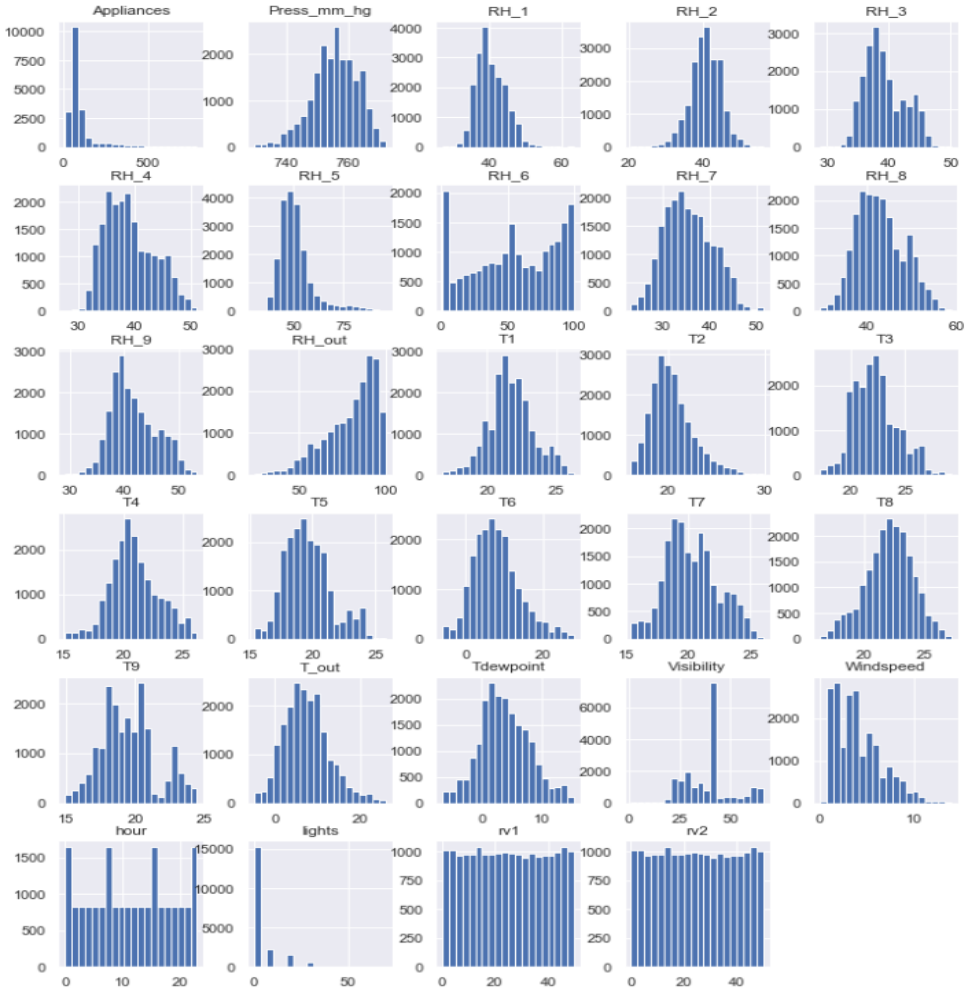

The distribution of power load is not normal as we have left asymmetry, for this reason I will use in my analysis log (power load) which distribution is more normal.

**Correlation:**

The most correlated features with energy consumtion(log_appliances) are: hour=0.33, lights=0.26, t6=0.198, t2=0.215, t3 = 0.168,t_out = 0.177, rh_out = -0.227, rh_8 = -0.17, rh_6 = -0.18, windspeed = 0.09.



**Feature Distribution:**




**<u>Experiment 1:</u>** (Changing learning rates)

**Model:**

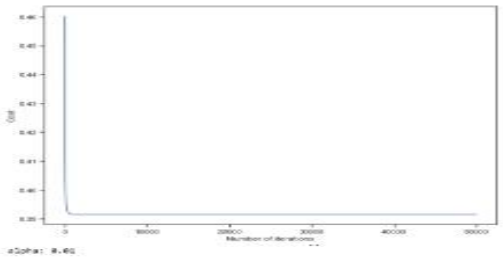
Appliance=b0+b1*T1+b2*T2+b3*T3+b4*T4+b5*T5+b6*T6+b7*RH_1+b8*RH_2+b9*RH_3+b10*RH_5+b11*H_6+b12*Tdewpoint+b13*RH_out+b14*Press_mm_hg+b15*Windspeed+b16*lights

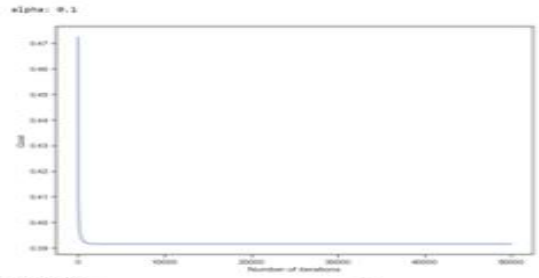
Initial Coeeficients – 0, initial alpha 0.5, Data Split: 80:20

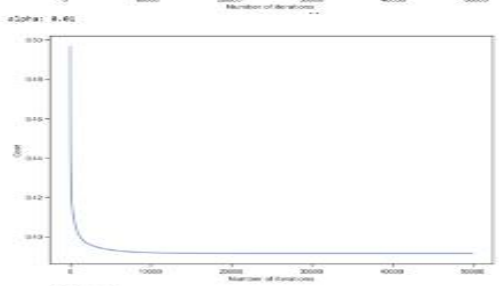
I have fixed the threshold as "0.0000001" and varied learning rate 0.2, 0.1, 0.01, 0.05, 0.001, 0.005, 0.0001, 0.0005. To find best alpha value, I ran linear regression using in-built function and matched the coefficients.

**Appliance=0.31475457 -0.45527275 *T1+0.36473878 *T2-0.20422269 *T3-0.09200132 *T4+0.29578117*T5+ 0.78138302 *T6-0.61675319 *RH_1-0.17220266 *RH_2+0.17220266 * RH_3-0.04142052 * RH_5-0.02864974 * H_6-0.29513445* Tdewpoint + 0.03476882 * RH_out -0.03131108* Press_mm_hg +0.02511575 * Windspeed +0.27005184* lights**

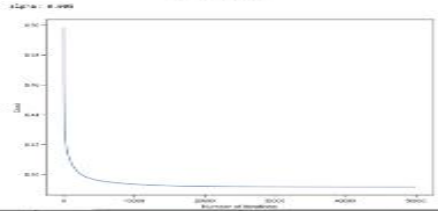







| Learning Rate | Cost Value | Converging Point | Train MSE | Test MSE |
|---|---|---|---|---|
| 0.2 | 0.39155033 | 998 | 0.78306872 | 0.72764619 |
| 0.1 | 0.391566245 | 1778 | 0.78306872 | 0.72764619 |
| 0.05 | 0.391598083 | 3117 | 0.78306873 | 0.7276462 |
| 0.01 | 0.391850964 | 10500 | 0.78306872 | 0.72764619 |
| 0.005 | 0.392164638 | 16651 | 0.78307543 | 0.72765026 |

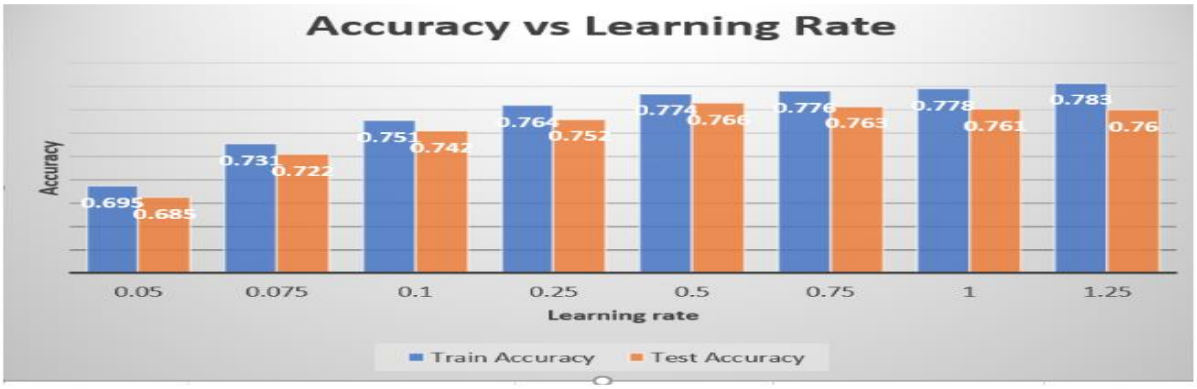
- The best alpha value is 0.2 and converged at 998, model cost is 0.3915503, MSE is 0.7830 for training data.
- When alpha decreased, all the cost value, converging point and MSE increased

**Logistic Regression:**

I created a new variable called Appliance_class. It is a categorical variable and it takes “1” for appliance energy greater than 60Wh (median value) and “0” for less than 60Wh.

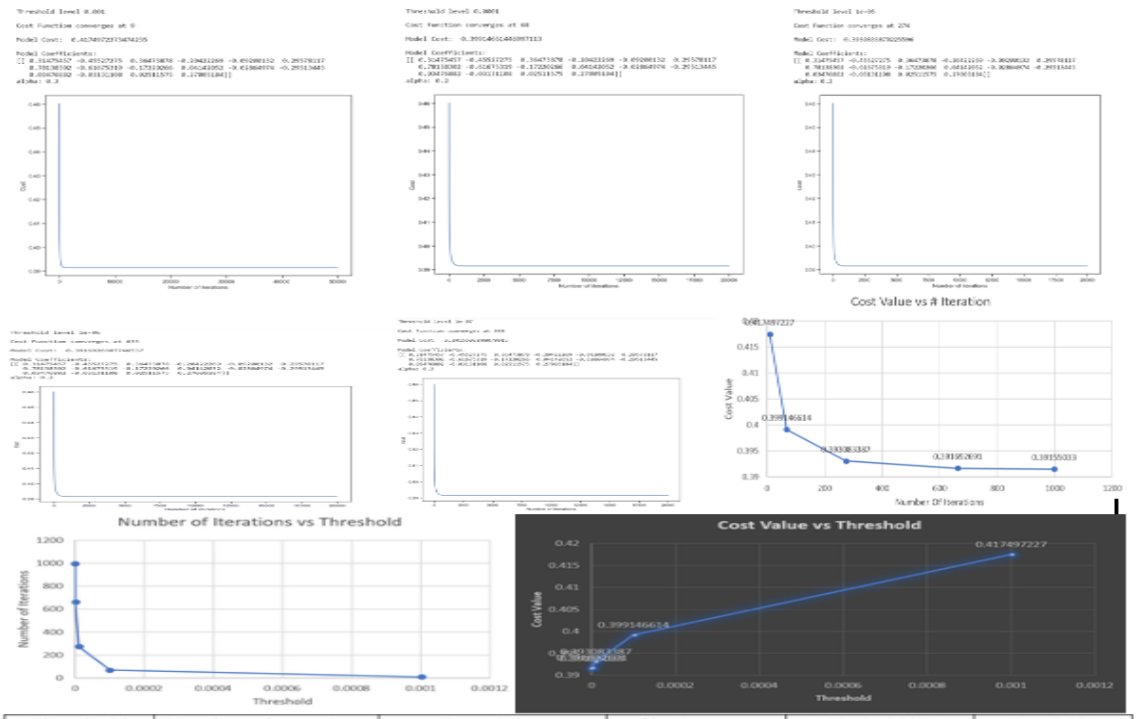
I used **Gradient Boosting Classifier** to classify the Appliance Class for different learning rates (0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1,1.25).

- Learning rate 1.25 gives the best train and test accuracy.
- Precision, Recall and F1-value for test data is higher for learning rate 1.25.
- The top 3 important features are humidity attributes (RH_out, RH_8 and RH_1) to help to identify the appliance category.




| | | | Weighted AVG | | |
|---|---|---|---|---|---|
| Learning rate | Train Accuracy | Test Accuracy | Presicion | Recall | F1-Score |
| 0.05 | 0.695 | 0.685 | 0.75 | 0.68 | 0.62 |
| 0.075 | 0.731 | 0.722 | 0.74 | 0.72 | 0.69 |
| 0.1 | 0.751 | 0.742 | 0.75 | 0.74 | 0.72 |
| 0.25 | 0.764 | 0.752 | 0.75 | 0.75 | 0.75 |
| 0.5 | 0.774 | 0.766 | 0.76 | 0.77 | 0.76 |
| 0.75 | 0.776 | 0.763 | 0.76 | 0.76 | 0.76 |
| 1 | 0.778 | 0.761 | 0.76 | 0.76 | 0.76 |
| 1.25 | 0.783 | 0.76 | 0.76 | 0.76 | 0.76 |

**Experiment 2:** (Changing Threshold levels):




| Threshold | Number of Iterations | Cost Value | % Improve | Train MSE | Test MSE |
|---|---|---|---|---|---|
| 0.001 | 9 | 0.417497227 | 0 | 0.78306872 | 0.72764619 |
| 0.0001 | 68 | 0.399146614 | 4.395385568 | 0.78306872 | 0.72764619 |
| 0.00001 | 274 | 0.393083387 | 1.519047695 | 0.78306872 | 0.72764619 |
| 0.000001 | 663 | 0.391692691 | 0.353791662 | 0.78306872 | 0.72764619 |
| 0.0000001 | 998 | 0.39155033 | 0.036345066 | 0.78306872 | 0.72764619 |

I have fixed the learning rate as "0.2" and varied learning rate 0.001, 0.0001, 0.00001, 0.000001, 0.0000001. To find best Threshold value, I ran linear regression using in-built function and matched the coefficients. Initial Coefficients – 0, initial alpha 0.5

• For training data, threshold "0.0000001" has lowest Cost function and converged at 998.

• Train MSE is same for different threshold levels, so do the Test MSE.

• As Threshold decreases, cost value also decreases.

**Experiment 3 (Random Variables):**

**Model:**

Appliance=b0+b1*T1+b2*T2+b3*T5+b4*T7+b5*T8+b6*RH_2+b7*RH_4+b8*RH_5+ +b9*Tdewpoint+ b10*Windspeed

| | Alpha - 0.2 | | | Threshold- 0.0000001 | | |
|---|---|---|---|---|---|---|
| | Cost Value | Convering Point | Train MSE | Test MSE | Train R-square | Test R-square |
| Model - 10 random variables | 0.454830332 | 432 | 0.84786555 | 0.84429808 | 0.082616498 | 0.082616498 |
| Best Model from Experiment 1 &2 | 0.39155033 | 998 | 0.78306872 | 0.72764619 | 0.216931277 | 0.212692842 |

- Model with 10 random variables performed very bad. As its Cost value, Train MSE, and Test MSE are higher than the best model for 16 variables.
- Training R-square and Test R-square values are lower than the best model in experiment 1 & 2.

**Experiment 4 (Random Variables):**

- I have created a variable called "hour"- time at which the readings are taken.
- I found the correlation between the dependent variable and all independent variables to find the variables which possibly explains the Appliance variable.

**Model:**

Appliance=b0+b1*T2+b2*HL+b3*T3+b4*T6+b5*RH_3+b6*lights+b7*hour+b8*T_out+ +b9*RH_OUT+ b10*RH_8

*HL – Interaction between Hour and Lights

| | Alpha - 0.2 | | | Threshold- 0.0000001 | | |
|---|---|---|---|---|---|---|
| | Cost Value | Convering Point | Train MSE | Test MSE | Train R-square | Test R-square |
| Model - 10 random variables | 0.454830332 | 432 | 0.84786555 | 0.84429808 | 0.082616498 | 0.082616498 |
| Best Model from Experiment 1 &2 | 0.39155033 | 998 | 0.78306872 | 0.72764619 | 0.216931277 | 0.212692842 |
| Model - 10 best variables | 0.386401829 | 781 | 0.7727871 | 0.7245353 | 0.227212 | 0.216062 |
| Model -all variables | 0.357642726 | 1702 | 0.71515165 | 0.68073202 | 0.284848348 | 0.263453589 |

**Logistic Regression Results:**

| | Learning Rate:1.25 | | | | |
|---|---|---|---|---|---|
| | Train Accuracy | Test Accuracy | Precision | Recall | F1-Score |
| Best Model from Experiment 1 &2 | 0.783 | 0.76 | 0.76 | 0.76 | 0.76 |
| Model - 10 random variables | 0.781 | 0.731 | 0.73 | 0.73 | 0.73 |
| Model - 10 best variables | 0.808 | 0.775 | 0.77 | 0.78 | 0.77 |
| Model -all variables | 0.823 | 0.773 | 0.78 | 0.78 | 0.78 |

- The model with 10 best variables performed better than the other two (model – 10 random variables and Best model in experiment 1 & 2).
- It has lower cost value, train MSE and test MSE.
- It has better Test and Train R-square values.
- Model with 10 best features gave better Accuracy, Precision, Recall, F1-Score than other 2 models.
- But the Model with all variables out performed all the other model in terms of Cost Value, Train MSE, Test MSE, Train and Test R-square and Accuracy, Precision, Recall, F1-Score.
- **My choice of features did not provide better results than using all features**.
- Some of the variables in the model with all features are highly correlated and 2 random variables may influence the model. I did not add those features with high correlation with other variable and random variables.
- Higher R-square may be addition of more variables, in this case R-square is not the correct metric to evaluate the model. Adjusted R-square is a better metric to evaluate.

**What do you think matters the most for predicting the energy usage?**

- The top important features are humidity attributes (RH_out, RH_8 and RH_1), lights, and hour to help to predict the appliance energy usage.

**What other steps you could have taken with regards to modeling to get better results?**

- We can use Dimension reduction technique (Principal Component Analysis) that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables
- We can do hyperparameter tuning deploy better algorithm such Support Vector Regressor, Decision Tree regressor, Gradient Boosting, Neural Nets etc., to get lower MSE values.
- We can approach this as “Time Series Analysis” problem to forecast the energy usage based on the previous trend. We can use Arima or time series Decomposition methods to forecast the energy usage.