Assignment1

Appliances Energy Prediction

Dataset Description

The dataset for implementing the gradient descent algorithm for linear and logistic regression is Appliance Energy Prediction Dataset. The dataset for implementing the gradient descent algorithm for linear regression is downloaded from UCI Machine Learning repository. The link for downloading the dataset is [appliances+energy+prediction]

We have cross-sectional dataset with 19735 observations and 29 variables.The variables are Appliances and lights energy consumption in the house, 9 temperatures and 9 relative humidity in 9 different locations in the house, 6 weather variables such as outside temperature, outside humidity, Pressure, Windspeed, Visibility and Dewpoint. We also have date column, and two random variables that are being ignored in this model. Additional information about the dataset can be found out in the above mentioned link.

Dependent variable is Appliances and Date variable is dropped from the dataset for convenience

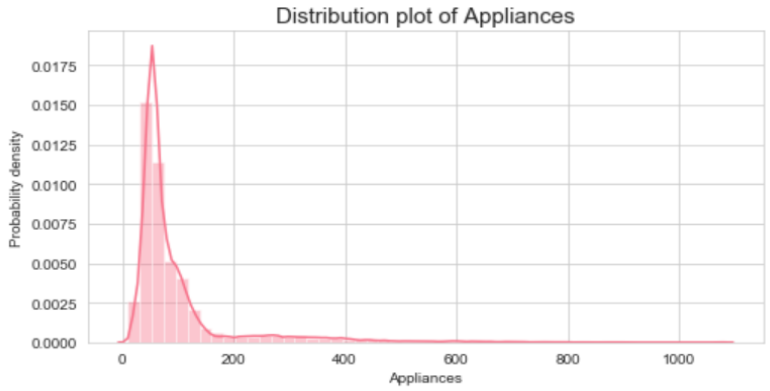
None of the variables consists of missing attributes so no missing value imputation is required

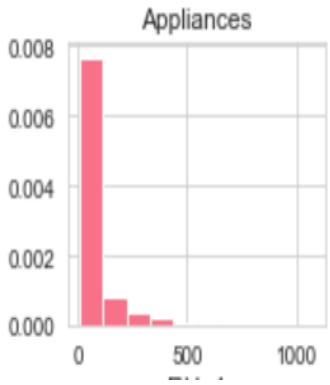
Data is partitioned randomly into train and test using 70:30 split percentage(13814 and 5921 observations respectively)

All the features are scaled to the range [0,1] by subtracting each feature from its mean and dividing by standard deviation in order to maintain evenness in the data

Exploratory Data Analysis

Target variable - Appliances – the dependent variable is approximately normally distributed around its mean

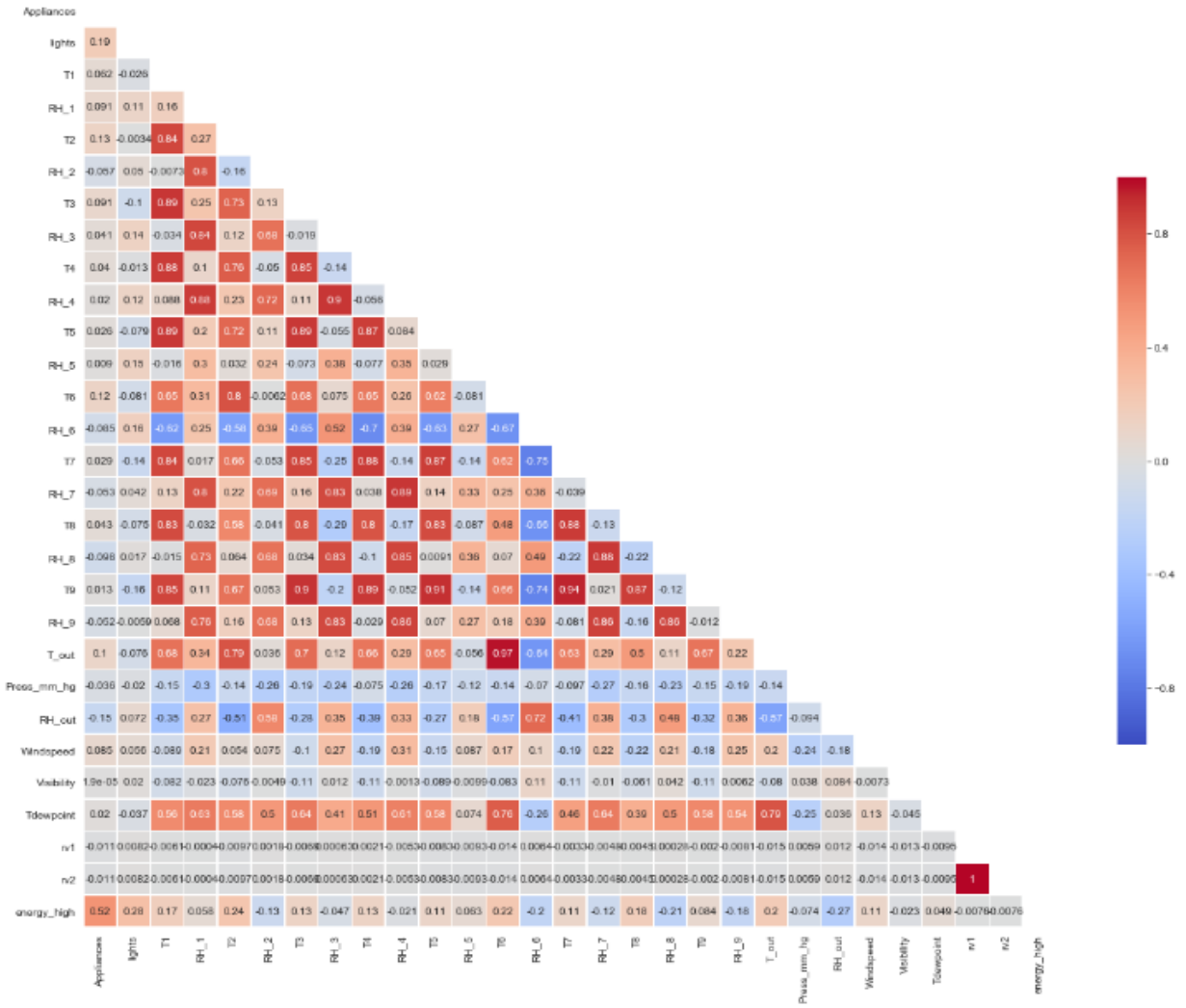




Distribution of the features

Correlation Matrix

Next, we look out for any correlation among the dependent variables as the correlation between the dependent variable can increase the standard errors of the regression coefficients. Below is the correlation plot:



Correlation between rv1 and rv2 is 1 which ideally means they are the same

T1 is highly correlated with T2, T3, T4, T5, T6, T7, T8, T9, T_out

T9 is showing high correlations with all the other area temperatures (T's)

Similarly, all the RH's are highly correlated

Tdewpoint is also showing significant correlations with all the other variables

Variable Selection

Three types of variable selection:

- Experiment1 and 2 requires 15 variables. They are selected by removing the features having very small correlations (approximately less than 0.03) with the target variable
  T4, T5, T7, T8, T9, RH_3, RH_4, RH_5, Visibility, Tdewpoint, rv1, rv2 and T_out are removed
  The 15 remaining variables being ('lights', 'T1', 'RH_1', 'T2', 'RH_2', 'T3', 'T4', 'T6', 'RH_6', 'RH_7', 'RH_8', 'RH_9', 'Press_mm_hg', 'RH_out', 'Windspeed')
- Experiment 3 requires 10 randomly selected variables. All the 10 variables are obtained using a simple random function. During the experimentation the random variables used are ['lights', 'RH_1', 'T2', 'RH_2', 'T3', 'T6', 'RH_6', 'RH_8', 'RH_out', 'Windspeed']
- Experiment 4 requires 10 best suited variables. They are selected based on the criteria that each of the features has a correlation more than or equal to 0.6. The best suited features being ['lights', 'RH_1', 'T2', 'RH_2', 'T3', 'T6', 'RH_6', 'RH_8', 'RH_out', 'Windspeed']

Experimentation1:

Linear Regression:

The regression equation to model the Appliances energy is:

$$Appliances = \beta 0 + \beta 1\ lights + \beta 2\ T1 + \beta 3\ RH\_1 + \beta 4\ T2 + \beta 5\ RH\_2 + \beta 6\ T3 + \beta 7\ T4 + \beta 8\ T6 + \beta 9\ RH\_6 + \beta 10\ RH\_7 + \beta 11\ RH\_8 + \beta 12\ RH\_9 + \beta 12\ Press\_mm\_hg + \beta 14\ RH\_out + \beta 15\ Windspeed$$

The initial parameter values from the implementation of the linear regression model to the above model equation is:

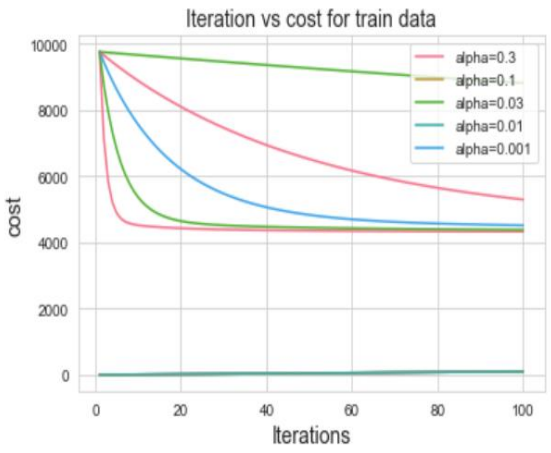
$\beta 0 = 96.93426958$, $\beta 1 = 18.18618179$, $\beta 2 = 3.26829553$, $\beta 3 = 68.77797069$, $\beta 4 = -29.27189955$, $\beta 5 = -54.1651787$, $\beta 6 = 44.09221557$, $\beta 7 = -21.82412024$, $\beta 8 = 7.24558783$, $\beta 9 = 12.39250603$, $\beta 10 = -5.67494124$, $\beta 11 = -22.34662532$, $\beta 12 = -5.05496434$, $\beta 13 = 1.51924159$, $\beta 14 = -2.56329643$, $\beta 14 = 4.24072548$

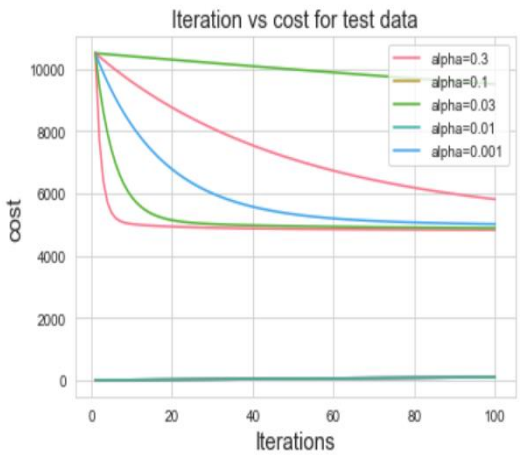
The Root mean squared error of the train model = 92.81876

The Root mean squared error of the test model = 97.9562

T2, Pressure and RH are significant features

For this experiment, the maximum number of iterations for gradient descent is fixed at 100000 and threshold of 0.0001. The learning rate experimented with are 0.3,0.1,0.03,0.01 and 0.001. The variations of train and test error for different values of alpha (learning rate) is plotted below:





Further the number of iterations to converge (maximum iteration if doesn't converge before the number of iterations exhaust), train error and test error is as given below:

Alpha	0.3	0.1	0.03	0.01	0.001
Iterations	827	2160	6026	14859	81231
Train RMSE	92.8185	92.8187	92.8194	93.8215	92.8499
Test RMSE	97.9560	97.9562	97.9569	97.9589	97.9861

As evident from the train and test plots and the number of iterations taken by the gradient descent algorithm to converge, we can see that the number of iterations to converge and the train and the test error are functions of learning rate alpha. For smaller learning rate, the number of iterations required is more and in case of learning rate 0.001 the algorithm fails to converge within 100000 iterations if the learning rate goes below 0.001. For this dataset, the best learning rate seems to be 0.1 from the minimum number of iterations for convergence. Also, the train and the test error are minimum and after that there is not much variation.

Logistic Regression:

The regression equation to model the Appliances energy (target variable is energy_high) is:

$$Appliances = \beta 0 + \beta 1\ lights + \beta 2\ T1 + \beta 3\ RH\_1 + \beta 4\ T2 + \beta 5\ RH\_2 + \beta 6\ T3 + \beta 7\ T4 + \beta 8\ T6 + \beta 9\ RH\_6 + \beta 10\ RH\_7 + \beta 11\ RH\_8 + \beta 12\ RH\_9 + \beta 12\ Press\_mm\_hg + \beta 14\ RH\_out + \beta 15\ Windspeed$$

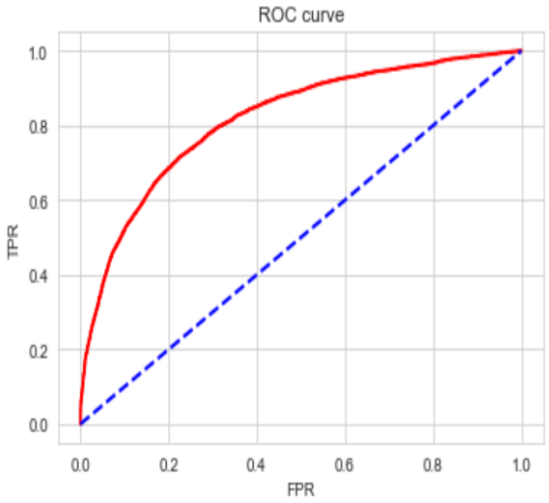
The initial parameter values from the implementation of the linear regression model to the above model equation is:

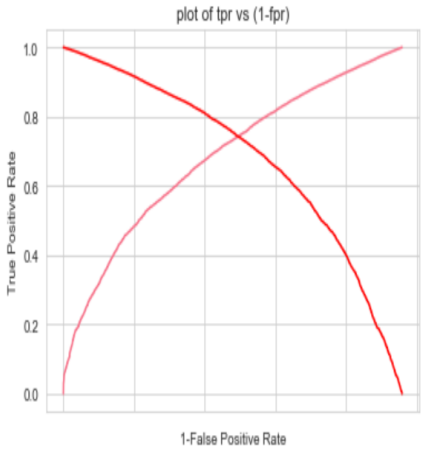
$\beta 1$ = 0.6854, $\beta 2$ = 1.0861, $\beta 3$ = 2.2675, $\beta 4$ = -0.8055, $\beta 5$ = -1.3185, $\beta 6$ = 0.3942 , $\beta 7$ = -0.7829, $\beta 8$ = 0.4664, $\beta 9$ = 0.3568, $\beta 10$ = -0.1981, $\beta 11$ = -0.7838, $\beta 12$ = -0.7290, $\beta 13$ = -0.0399, $\beta 14$ = -0.0667, $\beta 15$ = 0.1747

The Root mean squared error of the train model = 1.0866

The Root mean squared error of the test model = 1.0954

By computing the false positive rate and the false negative rate, area under the ROC is obtained as 0.816926





The optimal cut off would be where true positive rate is high and false positive rate is low

Optimal cut off point is when (True positive rate) – (1 – false positive rate) = zero or close to zero

Experimentation2:

The threshold values chosen for this experimentation are 0.0001, 0.00001, 0.000001, 0.0000001 at the best learning rate of best alpha 0.1. The plot of train and test error as function of threshold, the lower value of threshold the lesser iteration it needs to decide the convergence conditions. Whereas in case of test set we see that the lower value of threshold requires more iterations for convergence. The best value of threshold for this dataset seems to be 0.0000001

Threshold	0.01	0.0001	0.000001	0.0000001
Iterations	814	2160	3508	4180
Train RMSE	92.498	92.8187	92.818448	92.818445
Test RMSE	97.9859	97.9562	97.9559	97.9559

The best value of threshold for this dataset seems to be 0.0000001

Experimentation3:

A simple random number generator is used to pick 10 random variables. All the 10 variables are obtained using a simple random function. During the experimentation the random variables used are ['lights', 'RH_1', 'RH_2', 'T3', 'RH_4', 'T6' 'RH_7', 'T8', 'T9', 'Visibility']

Model Type	Train	Test
10 Random Features	93.7218	98.7106
15 Featured Model	92.8184	97.9559

The model with 10 random features has bigger value for both the train and the test datasets as compared to the model which uses 15 selected features to predict the Appliances energy consumption. This is expected as the features were picked up randomly without any statistical explanations and the odds of randomly picked ten features performing better than the selected feature dataset is very low for any supervised model. The regression equation is:

$$Appliances = \beta 0 + \beta 1\ lights + \beta 2\ RH\_1 + \beta 3\ RH\_2 + \beta 4\ T3 + \beta 5\ RH\_4 + \beta 6\ T6 + \beta 7\ RH\_7 + \beta 8\ T\_8 + \beta 9\ T\_9 + \beta 10\ Visibility$$

The initial parameter values from the implementation of the linear regression model to the above model equation is:

$\beta 0 = 96.93426958$, $\beta 1 = 15.42773157$, $\beta 2 = 49.39496614$, $\beta 3 = -20.65801875$, $\beta 4 = 35.9998126$, $\beta 5 = -12.42305447$, $\beta 6 = 10.45291248$, $\beta 7 = -26.71546762$, $\beta 8 = 7.14730113$, $\beta 9 = -46.15101434$, $\beta 10 = 0.71987784$

Experimentation4:

The 10 features that are best suited to predict the output are: ['lights', 'RH_1', 'T2', 'RH_2', 'T3', 'T6', 'RH_6', 'RH_8', 'RH_out', 'Windspeed'].

Best suited features are picked based on the criteria that all the features having correlation with Appliances greater than or equal to 0.6 are picked. The features having multicollinearity are removed from the dataset.

The comparison of the train and test error between 10 best feature case, 10 random feature case and using all the features to train the model is given below:

Model Type	Train	Test
10 Random Features	93.7218	98.7106
15 Featured Model	92.8184	97.9559
10 Best suited features	93.2812	98.174

As expected, using a smaller feature set perform poorer as compared to using the 15 feature set. The reason behind such an output is the smaller feature set is not able to capture the variance in the dependent variable as good as the full feature set does.

Surprisingly, in this case the model performs poorer in comparison to the random feature set.

$$Appliances = \beta 0 + \beta 1\ lights + \beta 2\ RH\_1 + \beta 3\ T2 + \beta 4\ RH\_2 + \beta 5\ T3 + \beta 6\ T6 + \beta 7\ RH\_6 + \beta 8\ RH\_8 + \beta 9\ RH\_out + \beta 10\ Windspeed$$

The initial parameter values from the implementation of the linear regression model to the above model equation is:

$\beta 0 = 96.93426958$, $\beta 1 = 16.89493667$, $\beta 2 = 73.15952374$, $\beta 3 = -41.26837015$, $\beta 4 = -63.01083481$, $\beta 5 = 37.63332687$, $\beta 6 = 7.85231437$, $\beta 7 = 15.87809363$, $\beta 8 = -28.37273321$, $\beta 9 = -2.13849244$, $\beta 10 = 5.88142486$

Discussion:

Out of all the three models, the initial model with 15 features has best cost and error measures. There could be high multicollinearity in the data because of the more number of features and high correlations among them. The reason might be because all the rooms are in the same house so there ought to be multicollinearity.

The best case squared error for the regression model on the training dataset is 92.8184 with learning rate 0.1 and threshold 0.0000001. This shows that the correct modeling algorithm for this dataset should be non-linear in nature.