Bharath Kumar Natarajan (BXN180005)
BUAN 6341.003 – Applied Machine Learning (Assignment 1)

Objective:

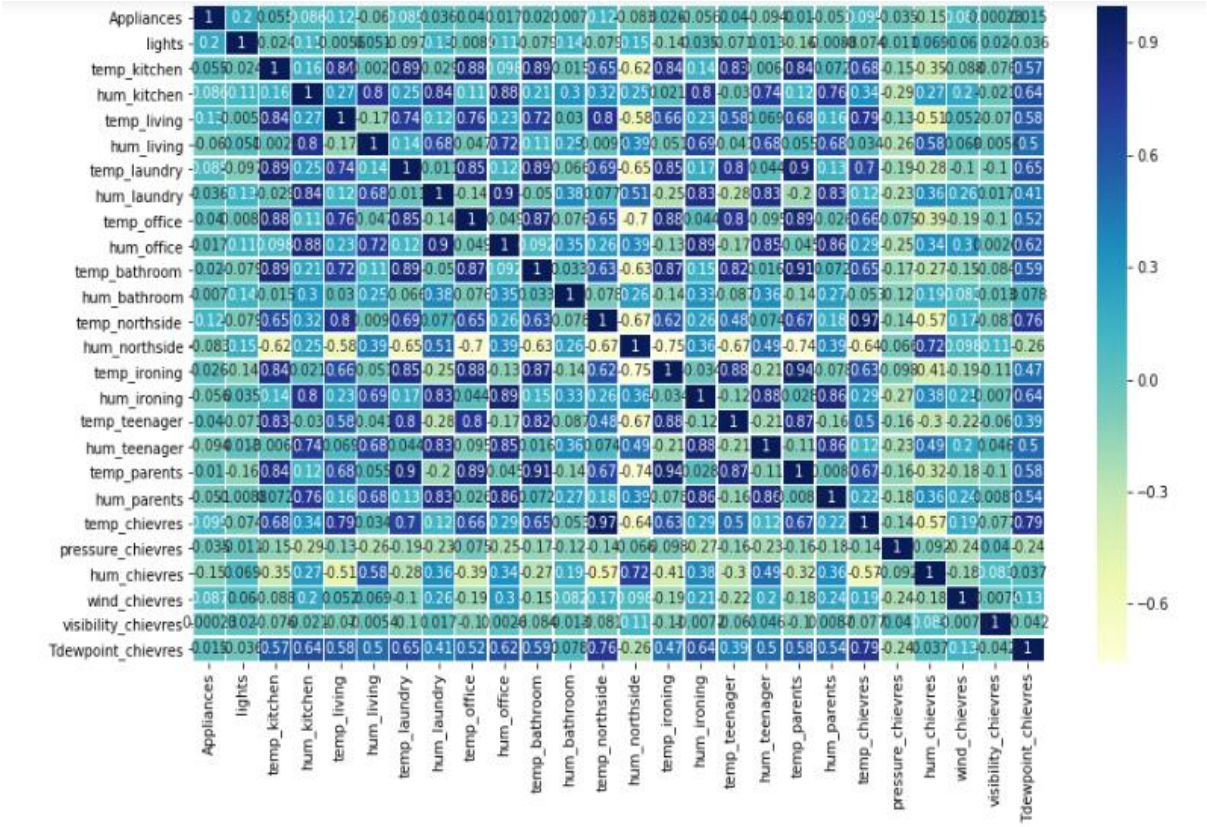
To implement a linear regression and logistic regression model on the given dataset and various experiments are performed with the design and feature choices.

Dataset Description:

The dataset is downloaded from UCI Machine Learning repository. The Dataset consists of 19735 observations on 29 variables. None of the column contains any missing value, so no missing value imputation is required.

Data Pre-processing:

Correlation Plot



All the temperature variables from T1-T9 and T_out have positive correlation with the target Appliances. For the indoor temperatures, the correlations are high as expected, since the ventilation is driven by the HRV unit and minimizes air temperature differences between rooms. Four columns have a high degree of correlation with T9 - T3, T5, T7, T8 also T6 & T_out has high correlation (both temperatures from outside.

Also, from the above correlation plot, it is evident that the variables “T6 Temperature outside the building (north side)”, “T9 Temperature in parents’ room” and “Visibility (from Chievres weather station)” are highly correlated. So, to implement the regression model we will drop temp and season from the training.

Also, the random variables columns have been removed from the data set as it does not help much for the model prediction. The date column was also removed for the same reason.

Tasks:

The data has been transformed initially using Standard_Scaler () to standardize features by removing the mean and scaling to unit variance. The dataset has been partitioned into training and test datasets with split percentage of 70 as training data and 30 as testing data.

With the variables finally taken for analysis to model the Linear regression equation, the following equation can be framed: -

The following Linear Regression model was designed with the existing parameters:

$$\text{Appliances}=\beta_0+\beta_1*x1+\beta_2*x2+\beta_3*x3+\beta_4*x4+\beta_5*x5+\beta_6*x6+\beta_7*x7+\beta_8*x8+\beta_9*x9+\beta_{10}*x10+\beta_{11}*x11+\beta_{12}*x12+\beta_{13}*x13+\beta_{14}*x14+\beta_{15}*x15+\beta_{16}*x16+\beta_{17}*x17+\beta_{18}*x18+\beta_{19}*x19+\beta_{20}*x20+\beta_{21}*x21$$

β0	**βtemp_kitchen**	**βhum_kitchen**	**βtemp_living**	**βhum_living**
-0.00539144	-0.04573451	0.56813491	-0.25499545	-0.50779914
βtemp_laundry	**βhum_laundry**	**βtemp_office**	**βhum_office**	**βtemp_bathroom**
0.46013182	0.19517236	-0.04727051	0.06646581	-0.10244412
βhum_bathroom	**βhum_northside**	**βtemp_ironing**	**βhum_ironing**	**βtemp_teenager**
0.03387902	0.06991268	-0.17741055	-0.00942403	0.15964828
βhum_teenager	**βhum_parents**	**βtemp_chievres**	**βpressure_chievres**	**βhum_chievres**
-0.35896397	-0.10143236	-0.03497141	0.02332409	-0.03810698
βwind_chievres	**βtdewpoint_chievres**			
0.05659019	0.0548662			

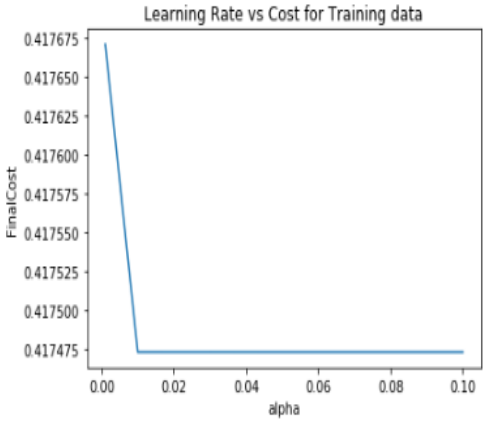
Considering this as the original set of features implementation of Linear regression model with alpha=0.001 cost values computed from the model are as follows:

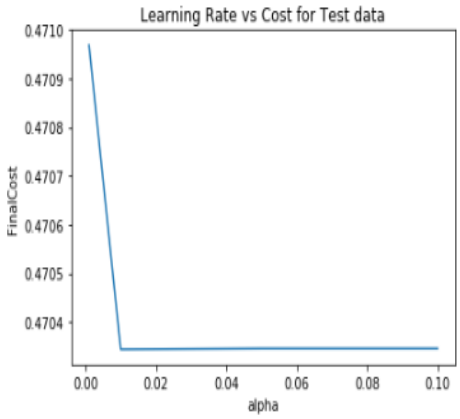
- Final Cost-Train: 0.417671095300688
- Final Cost-Test: 0.470968283354324

Experiment1:

For this experiment, the maximum number of iterations for gradient descent is fixed at 100000 iterations. The learning rate experimented with are 0.001,0.01,0.05 and 0.1. The variations of train and test error for different values of alpha (learning rate) is plotted below:

Bharath Kumar Natarajan (BXN180005)
BUAN 6341.003 – Applied Machine Learning (Assignment 1)





Given the number of iterations taken by the gradient descent algorithm to converge, we can see that 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. For this dataset and with the columns that I have considered, the best learning rate seems to be 0.1 or 0.05 as the train and the test error are minimum.

Alpha	0.001	0.01	0.05	0.1
Train Error	0.417671095300688	0.417473311214347	0.417473288042090	0.417473288042090
Test Error	0.470968283354324	0.470344061280054	0.470345704051668	0.470345704051802

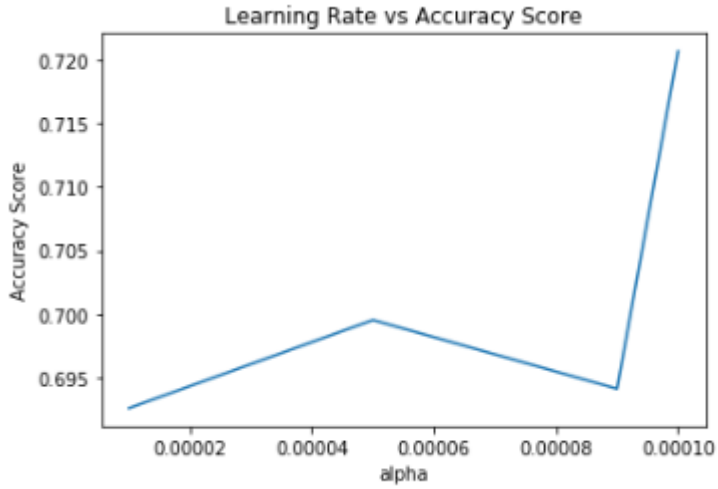
Logistic Regression Model:

The Logistic Regression model has been implemented with the same columns that was used for implementing Linear Regression model. The Logistic Regression model was implemented using the package SGDClassifier available in the Scikit-learn. The dependent variable “Appliances” was converted to a categorical variable of two classes using the median value which was found to be 60. The target variable with values less than 60 was taken as 0 and values greater than 60 was considered as 1.

The accuracy score is then calculated with the trained model using the predicted and actual value of the target variable. The learning rate and accuracy scores obtained are plotted as in the bottom figure. It can be noted that the highest accuracy is with the alpha value of 0.0001 for the given dataset.

Alpha	0.00001	0.00005	0.00009	0.0001
Accuracy Score	0.692619489951021	0.699543995946630	0.694139503462253	0.720655294713730

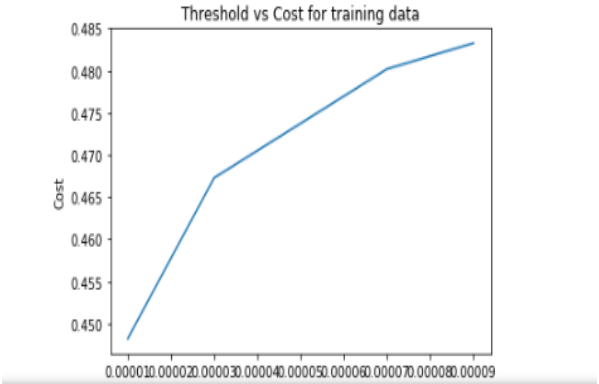
Bharath Kumar Natarajan (BXN180005)
BUAN 6341.003 – Applied Machine Learning (Assignment 1)

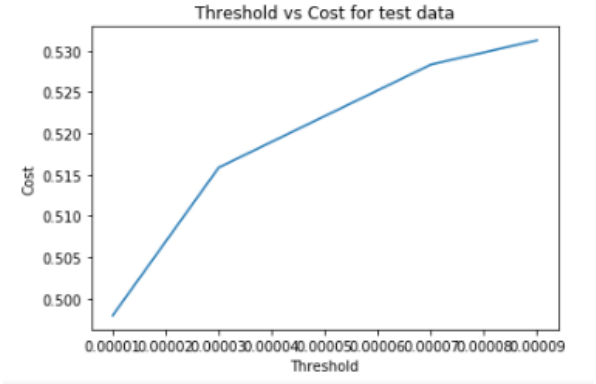


Experiment2:

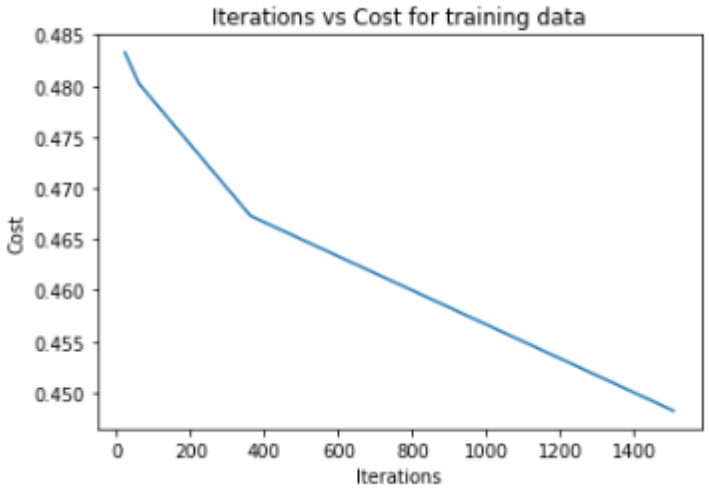
The threshold values chosen for this experimentation are 0.00001, 0.00003, 0.00007, 0.00009 at the fixed learning rate of alpha 0.001. The plot of train and test error as function of threshold are given at the bottom. As seen from the plot of training cost, the lower value of threshold the lesser iteration it needs to decide the convergence conditions to get the minimum cost. The best threshold for this dataset seems to be 0.00001.

Threshold	0.00001	0.00003	0.00007	0.00009
Train Error	0.448221859758169	0.467291376603866	0.480196330761827	0.483270109787755
Test Error	0.497958048051661	0.515871645656337	0.528333558516817	0.531274001912191





The plot below shows the cost/error as a function of number of iterations:



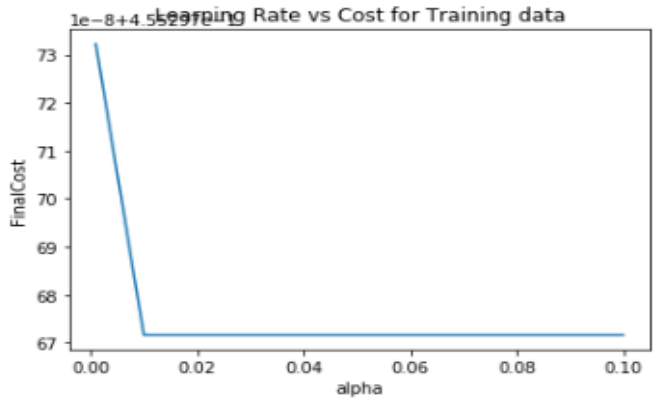
Experiment 3

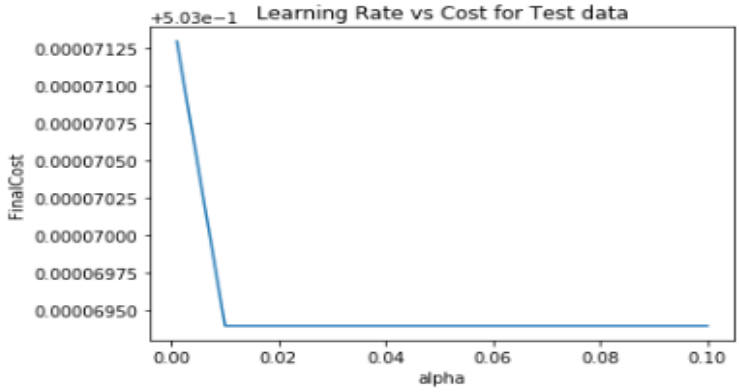
The columns that were chosen randomly are as follows:

```
['temp_kitchen', 'hum_living', 'temp_laundry', 'hum_office','hum_bathroom', 'hum_northside', 'hum_ironing', 'temp_teenager', 'hum_parents', 'temp_chievres']
```

Alpha	0.001	0.01	0.05	0.1
Train Error	0.455297732257167	0.455297671673410	0.455297671673410	0.455297671673410
Test Error	0.503071297010942	0.503069397201322	0.503069397201322	0.503069397201322

For this experiment, the maximum number of iterations for gradient descent is fixed at 100000 iterations. The learning rate experimented with are 0.001,0.01,0.05 and 0.1. The variations of train and test error for different values of alpha (learning rate) is plotted below:





The comparison between the train and the test error between model containing 10 random features and the model containing the original set of features is given below:

Model Type	Train	Test
10 Random Features	0.455297732257167	0.503071297010942
Original Model	0.417671095300688	0.470968283354324

The model with 10 random features has greater value for both the train and the test case as compared to the model which uses original set of features to predict the target variable. This is expected as the features were picked up randomly without any statistical explanations and the odds of randomly picked three features performing better than the full feature dataset is very low for any supervised model.

Experiment 4

The ten features which I think will be best suited for the analysis that will be used to train the model are as follows:

```
['temp_kitchen', 'temp_living', 'hum_northside', 'temp_ironing','hum_ir
oning', 'temp_teenager', 'hum_parents', 'temp_chievres', 'pressure_chie
vres', 'hum_chievres']
```

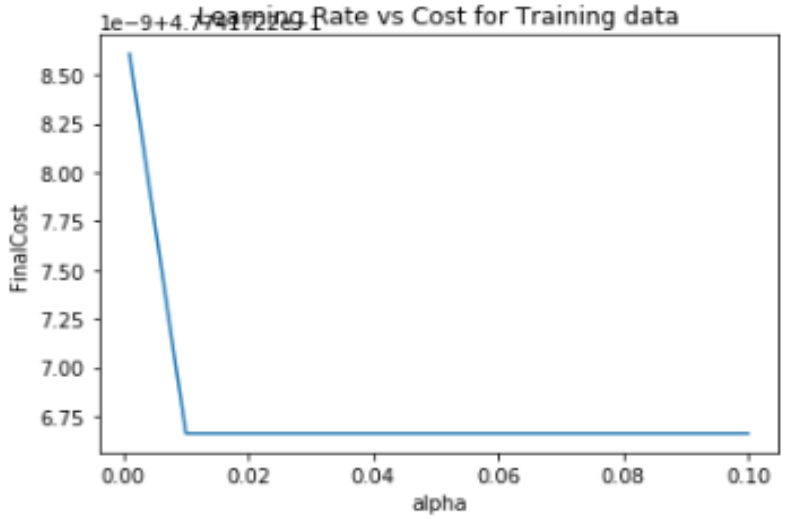
The above 10 variables were selected of my choice based on the correlation values and with the impact of the variables as seen so far on the target variable.

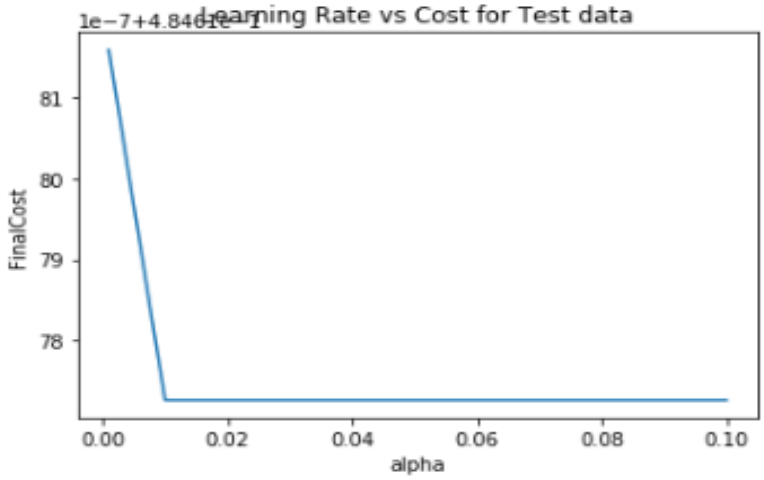
Bharath Kumar Natarajan (BXN180005)
BUAN 6341.003 – Applied Machine Learning (Assignment 1)

Below

Alpha	0.001	0.01	0.05	0.1
Train Error	0.477417228607256	0.477417226661205	0.477417226661205	0.477417226661205
Test Error	0.484618158722170	0.484617726761419	0.484617726761419	0.484617726761419

For this experiment, the maximum number of iterations for gradient descent is fixed at 100000 iterations. The learning rate experimented with are 0.001,0.01,0.05 and 0.1. The variations of train and test error for different values of alpha (learning rate) is plotted below:





The comparison between the train and the test error between model containing 10 features of choice, 10 random features and the model containing the original set of features is given below:

Bharath Kumar Natarajan (BXN180005)
BUAN 6341.003 – Applied Machine Learning (Assignment 1)

Model Type	Train	Test
10 Features of Choice	0.477417228607256	0.484618158722170
10 Random Features	0.455297732257167	0.503071297010942
Original Model	0.417671095300688	0.470968283354324

As expected, using a smaller feature set perform poorer as compared to using the full feature set. That's why we see the smallest Cost values for the original model which had around 21 columns and the 10 feature sets have comparatively greater cost values. The reason for 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.

Also, surprisingly in this case, the feature of choice model performs poorer in comparison to the random feature set. The random feature set has variables that have good level of impact on the target variable compared to the variables in the 10 feature of choice model.

Discussion:

Based on the model implemented above, we can see that the most important features for the given dataset are 'Humidity in kitchen area', 'Humidity in laundry room area', 'Temperature in laundry room area' and 'Temperature in teenager room 2'. These variables play an important role in predicting the target variable.

Also, from the correlation plot that was done before analysis, it is evident that the variables "T6 Temperature outside the building (north side)", "T9 Temperature in parents' room" and "Visibility (from Chievres weather station)" are highly correlated, due to which it was dropped before analysis.

As expected, using a smaller feature set perform poorer as compared to using the full feature set. That's why we see the smallest Cost values for the original model which had around 21 columns and the 10 feature sets have comparatively greater cost values. The reason for 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.

Also, surprisingly in this case, the feature of choice model performs poorer in comparison to the random feature set. The random feature set has variables that have good level of impact on the target variable compared to the variables in the 10 feature of choice model.

Further steps that can be considered to improve the model are as follows:

1. Firstly, by build simple models and using many independent variables need not necessarily mean that your model is good. Next step is to try and build many regression models with different combination of variables. Then you can take an ensemble of all these models.
2. To understand the relationship between dependent variable and all the independent variables and whether they have a linear trend. Only then we can afford to use them in the model to get a good output.
3. It's also important to check and treat the extreme values or outliers in your variables. This could be one reason why your predicted estimate values might vary as they are getting skewed by the outlier values.