Project 1: Applied Machine Learning

Abhishek Pandey (AXP180002)

Dataset description and Summary

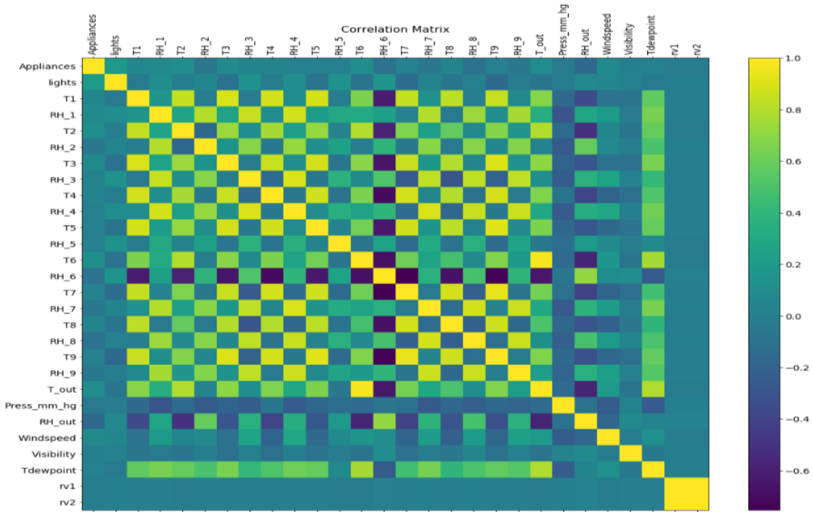
The dataset used for implementing linear regression using batch gradient descent and logistic regression using scikit learn library has been downloaded from UCI Machine Learning data repository. The link to download the dataset is Energy Dataset.

A brief description about the dataset:

1. Dataset consists of 19735 number of instances or rows
2. Dataset has a total of 29 attributes
3. The dependent or target variable of the dataset is "Appliances" representing energy consumption in Wh
4. More detailed description about all the variables can be found in the link provided above
5. None of the columns has any missing values so no missing value imputation is required

Correlation Plot

Our next step as an EDA would be to identify the correlation between independent and target variables and between all the variables amongst themselves. The correlation between variables amongst themselves increases standard errors thus we would want to identify such variables and ignore them from our final parametric equation. Also, identifying correlation between independent and target variables will help us define the best features that can be used for our final parametric model.



T9 is highly correlated with T7 and so is T_out with T6 and rv2 with rv1.Thus, while implementing regression models we will ensure that these highly correlated variables (**T9, T_OUT, rv2**) are not included as parameters in the final model.

Algorithm Preparation and Parameters

Multivariate Linear Regression:

1. **Functions**

- The key things to keep in mind while modelling linear regression are:
  - How the Cost is defined
  - How the algorithm learns to create this regression with different learning rates

- For this experiment, we have used squared error cost function. The cost function used was:

$$\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2$$

- This experiment uses the batch gradient descent function. The goal of the gradient descent is to reduce our cost function the most and each time the parameters (B1, B2, …, Bn) are updated.

- Updating of gradient descent takes place using the below formula:

$$\theta j := \theta j - \alpha \frac{\partial}{\partial \theta j} J(\theta)$$

2. **Parameters Experimentation**

   Alpha – This is our learning rate which determines that at every gradient descent whether we are taking big steps or little steps towards attaining an optimal cost function

   Threshold – This determines when can we stop iterating or descending. We can determine how small the increment in regression changes that we want in order to minimize the computations while still obtaining the best possible information

   Features – Experimentation on the features will be seen on the third and fourth part

Logistic Regression:

1. **Functions**

- The key things to keep in mind while modelling logistic regression are:
  - How the decision boundary line is decided
  - How the algorithm learns to create this regression with different hyperparameters

- For this experiment, we have used log error cost function. The cost function used was:

$$-\frac{1}{m}\sum\left[y^{(i)}\log(h\theta(x(i)))+\left(1-y^{(i)}\right)\log(1-h\theta(x(i)))\right]$$

- This experiment uses the batch gradient descent function. The goal of the gradient descent is to reduce our cost function the most and each time the parameters (B1, B2, …, Bn) are updated.
- Updating of gradient descent takes place using the below formula:

$$\theta j := \theta j - \alpha \frac{\partial}{\partial \theta j} J(\theta)$$

2. **Parameters Experimentation**

   Decision boundary – This is our boundary that will help us separating one class from the other. We have taken mean of the target variable to have a distinguishable decision boundary

   Hyperparameters – These act as parameters that help us optimizing our cost function and improve the accuracy

   Features – Experimentation on the features will be seen on the third and fourth part

   Sigmoid – It is used to restrict the range of the final value between 0 and 1

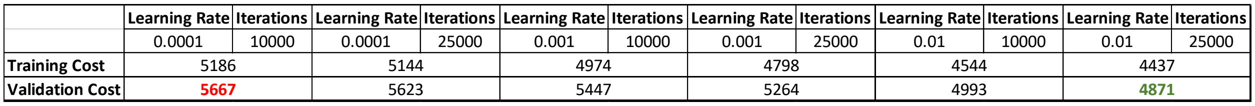
Experimentation

Experiment 1: Linear Regression

Here we would be trying to optimize the cost function by changing the values of learning rate and iteration thus trying to identify the best values of the parameters stated below. The linear regression to build a parametric model would be:

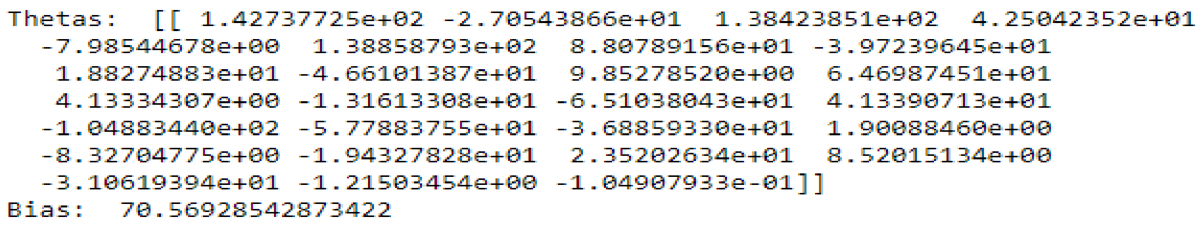
Energy = Bo + B1 *(lights) + B2*(T1) + B3*(RH_1) + B4*(T2) + B5*(RH_2) + B6*(T3) + B7*(RH_3) + B8*(T4) + B9*(RH_4) + B10*(T5) + B11*(RH_5) + B12*(T6) + B13*(RH_6) + B14*(T7) + B15*(RH_7) + B16*(T8) + B17*(RH_8) + B18*(T9) + B19*(RH_9) + B20*(T_out) + B21*(Press_mm_hg) + B22*(RH_out) + B23*(Windspeed) + B24*(Visibility) + B25*(Tdewpoint) + B26*(rv1) + B27*(rv2)

Test and Train costs for different values of learning rate and iterations:



	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations
	0.0001	10000	0.0001	25000	0.001	10000	0.001	25000	0.01	10000	0.01	25000
Training Cost	5186		5144		4974		4798		4544		4437	
Validation Cost	5667		5623		5447		5264		4993		4871	

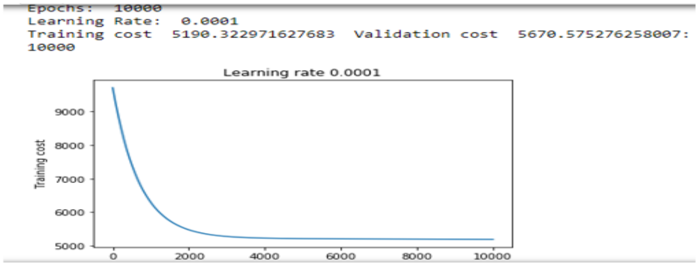
The best value for learning rate and iterations was measured to be **0.01 and 25,000** respectively. The value of parameters during the best case observed was:



```
Thetas:  [[ 1.42737725e+02 -2.70543866e+01  1.38423851e+02  4.25042352e+01
  -7.98544678e+00  1.38858793e+02  8.80789156e+01 -3.97239645e+01
   1.88274883e+01 -4.66101387e+01  9.85278520e+00  6.46987451e+01
   4.13334307e+00 -1.31613308e+01 -6.51038043e+01  4.13390713e+01
  -1.04883440e+02 -5.77883755e+01 -3.68859330e+01  1.90088460e+00
  -8.32704775e+00 -1.94327828e+01  2.35202634e+01  8.52015134e+00
  -3.10619394e+01 -1.21503454e+00 -1.04907933e-01]]
Bias:  70.56928542873422
```

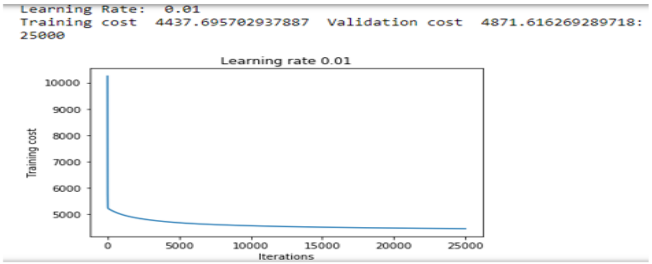
The below graphs show the best and worst cost measured during the experiment.

Worst plot



```
Epochs:  10000
Learning Rate:  0.0001
Training cost  5190.322971627683  Validation cost  5670.575276258007:
10000
```

Best Plot



```
Learning Rate:  0.01
Training cost  4437.695702937887  Validation cost  4871.616269289718:
25000
```

Logistic Regression:

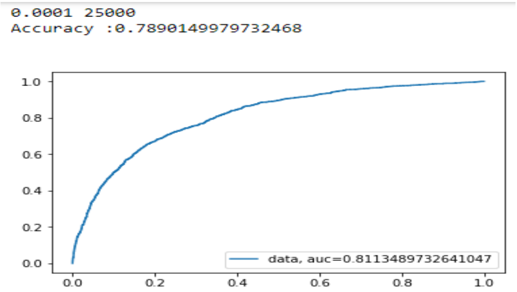
The target variable has continued values and in order to classify the target variables into a class we must modify the variable and define a decision boundary line which here has been defined as the mean of the target variable. Any value below the mean would be classified as zero and above mean would be classified as one.

The experiment was done with multiple values of tolerance and iterations with decision boundary being the mean of target variable.

Test and Train costs for different values of tolerance and iterations:

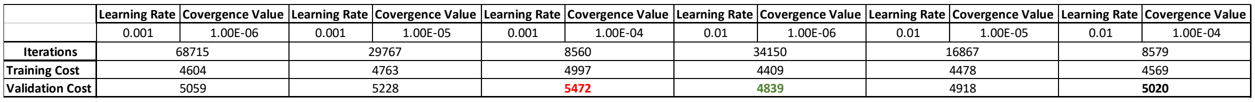
	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations
	0.0001	10000	0.0001	25000	0.001	10000	0.001	25000	0.01	10000	0.01	25000
Test Accuracy	0.789		0.789		0.788		0.788		0.789		0.789	
Validation Accuracy	**0.794**		0.794		0.794		0.794		0.794		0.794	

Since the accuracy and AUC for all the iterations have almost similar values thus, we can take the best values to be any learning rate (0.0001) with 10,000 iterations. The AUC curve shown below when plotted had a stable 81% value throughout all the curves.



```
0.0001 25000
Accuracy :0.7890149979732468
```

Experiment 2: In this experiment we will try to see where the model converges by changing the values of the threshold and learning rates.



	Learning Rate	Covergence Value	Learning Rate	Covergence Value	Learning Rate	Covergence Value	Learning Rate	Covergence Value	Learning Rate	Covergence Value	Learning Rate	Covergence Value
	0.001	1.00E-06	0.001	1.00E-05	0.001	1.00E-04	0.01	1.00E-06	0.01	1.00E-05	0.01	1.00E-04
Iterations	68715		29767		8560		34150		16867		8579	
Training Cost	4604		4763		4997		4409		4478		4569	
Validation Cost	5059		5228		**5472**		**4839**		4918		**5020**	

As observed in the image above the least cost for the model was at convergence value of 1e-6 with 0.01 learning rate.

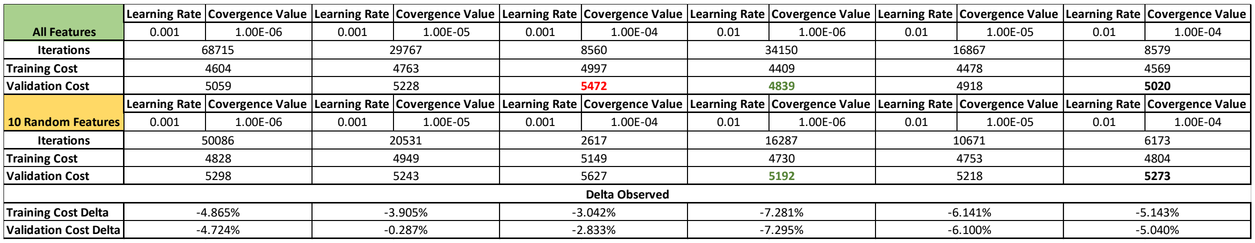
Experiment 3:

The ten randomly picked features are:

1	2	3	4	5	6	7	8	9	10
RH_4	RH_out	Windspeed	Visibility	T8	lights	T_out	T9	Press_mm_hg	RH_7

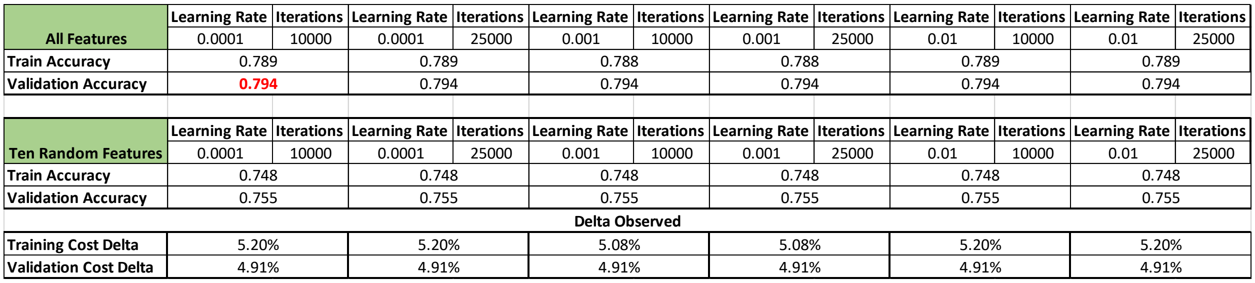
Multiple Linear Regression:

The delta observed when a comparison was made for cost calculated during randomly selected features model and when cost was calculated with all the variables, we found the cost has increased by ~ 5% for model with random variables showing that if you do not pick relevant features to build a model than the cost and errors incurred and would be higher.



	Learning Rate	Covergence Value	Learning Rate	Covergence Value	Learning Rate	Covergence Value	Learning Rate	Covergence Value	Learning Rate	Covergence Value	Learning Rate	Covergence Value
All Features	0.001	1.00E-06	0.001	1.00E-05	0.001	1.00E-04	0.01	1.00E-06	0.01	1.00E-05	0.01	1.00E-04
Iterations	68715		29767		8560		34150		16867		8579	
Training Cost	4604		4763		4997		4409		4478		4569	
Validation Cost	5059		5228		**5472**		**4839**		4918		**5020**	
	Learning Rate	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**
10 Random Features	0.001	1.00E-06	0.001	1.00E-05	0.001	1.00E-04	0.01	1.00E-06	0.01	1.00E-05	0.01	1.00E-04
Iterations	50086		20531		2617		16287		10671		6173	
Training Cost	4828		4949		5149		4730		4753		4804	
Validation Cost	5298		5243		5627		**5192**		5218		**5273**	
Delta Observed												
Training Cost Delta	-4.865%		-3.905%		-3.042%		-7.281%		-6.141%		-5.143%	
Validation Cost Delta	-4.724%		-0.287%		-2.833%		-7.295%		-6.100%		-5.040%	

Logistic Regression: The delta observed when a comparison was made for cost calculated during randomly selected features model and when cost was calculated with all the variables, we found the cost has increased by almost 5% for model with random variables showing that if you do not pick relevant features to build a model than the cost and errors incurred and would be higher.



	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations
All Features	0.0001	10000	0.0001	25000	0.001	10000	0.001	25000	0.01	10000	0.01	25000
Train Accuracy	0.789		0.789		0.788		0.788		0.789		0.789	
Validation Accuracy	**0.794**		0.794		0.794		0.794		0.794		0.794	
	Learning Rate	**Iterations**	**Learning Rate**	**Iterations**	**Learning Rate**	**Iterations**	**Learning Rate**	**Iterations**	**Learning Rate**	**Iterations**	**Learning Rate**	**Iterations**
Ten Random Features	0.0001	10000	0.0001	25000	0.001	10000	0.001	25000	0.01	10000	0.01	25000
Train Accuracy	0.748		0.748		0.748		0.748		0.748		0.748	
Validation Accuracy	0.755		0.755		0.755		0.755		0.755		0.755	
Delta Observed												
Training Cost Delta	5.20%		5.20%		5.08%		5.08%		5.20%		5.20%	
Validation Cost Delta	4.91%		4.91%		4.91%		4.91%		4.91%		4.91%	

The randomly selected features for logistic regression were –

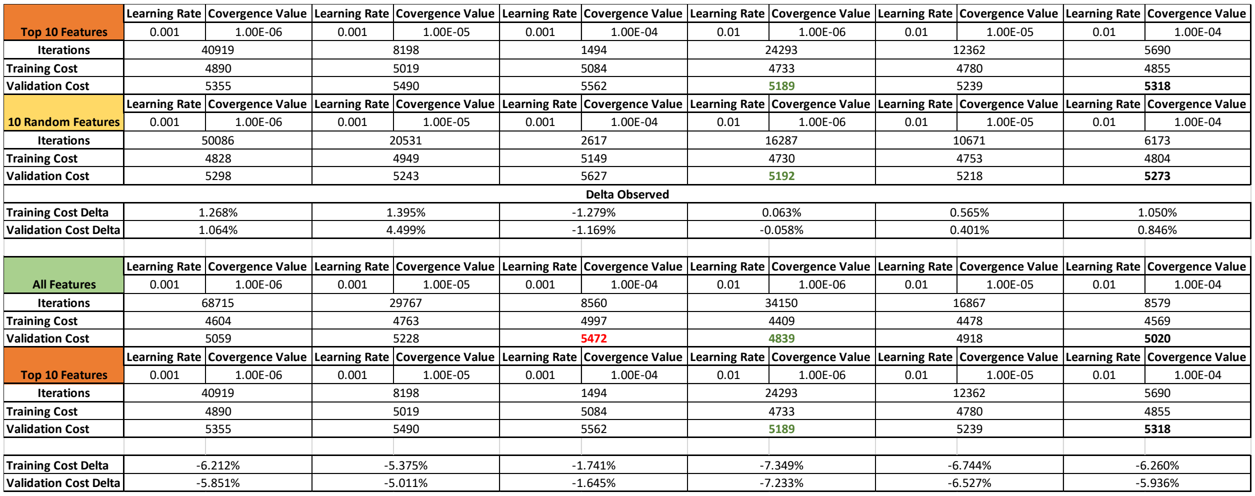
1	2	3	4	5	6	7	8	9	10
RH_4	RH_out	Windspeed	Visibility	T8	lights	T_out	T9	Press_mm_hg	RH_7

Experiment 4: The top ten features had to be selected to carry out this experiment. The three variables namely (**T9, T_OUT, rv2)** which posses high collinearity with other variables were excluded and then the Pearson correlation was carried out to find out the best features responsible for to influence the result of Appliance energy consumption. The top ten features to be found are reported below:

1	2	3	4	5	6	7	8	9	10
lights	T2	T6	Windspeed	RH_1	T3	T1	T4	T8	RH_3

Linear Regression: The cost comparison for regression model with top 10 features against random and all the variables is depicted below. We can see that the best cost for top model with top 10 variables is a little less against the model with ten random variables depicting that it is important to weed out the variable from the model that are collinear to each other and include those variables that are highly correlated to the target variables.

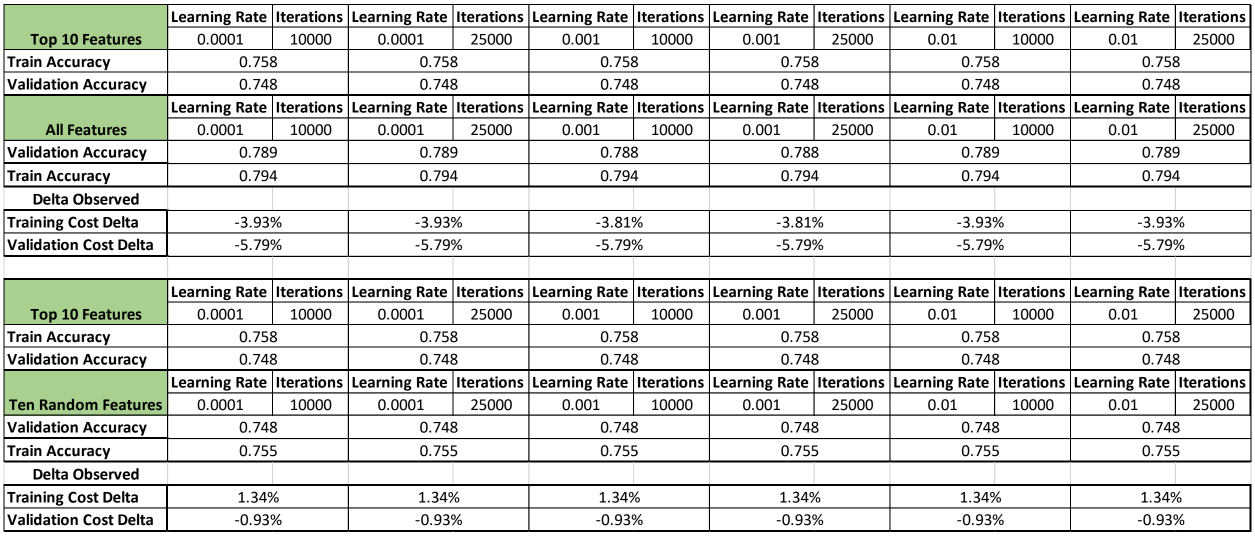
While in the comparison case where top 10 features model was compared to model with all the features it was found out that the cost in model with top 10 features was a little more in comparison to the model with all the variables thus showing that only correlation is not the criteria for choosing best features for a model. We could have considered experimenting with some other high-end models like random forest, regularization regression which helps us selecting the best features in a better way. The comparison metrics are shown below for linear regression.



Top 10 Features	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**
	0.001	1.00E-06	0.001	1.00E-05	0.001	1.00E-04	0.01	1.00E-06	0.01	1.00E-05	0.01	1.00E-04
Iterations	40919		8198		1494		24293		12362		5690	
Training Cost	4890		5019		5084		4733		4780		4855	
Validation Cost	5355		5490		5562		**5189**		5239		**5318**	
10 Random Features	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**
	0.001	1.00E-06	0.001	1.00E-05	0.001	1.00E-04	0.01	1.00E-06	0.01	1.00E-05	0.01	1.00E-04
Iterations	50086		20531		2617		16287		10671		6173	
Training Cost	4828		4949		5149		4730		4753		4804	
Validation Cost	5298		5243		5627		**5192**		5218		**5273**	
Delta Observed												
Training Cost Delta	1.268%		1.395%		-1.279%		0.063%		0.565%		1.050%	
Validation Cost Delta	1.064%		4.499%		-1.169%		-0.058%		0.401%		0.846%	
All Features	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**
	0.001	1.00E-06	0.001	1.00E-05	0.001	1.00E-04	0.01	1.00E-06	0.01	1.00E-05	0.01	1.00E-04
Iterations	68715		29767		8560		34150		16867		8579	
Training Cost	4604		4763		4997		4409		4478		4569	
Validation Cost	5059		5228		**5472**		**4839**		4918		**5020**	
Top 10 Features	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**	**Learning Rate**	**Covergence Value**
	0.001	1.00E-06	0.001	1.00E-05	0.001	1.00E-04	0.01	1.00E-06	0.01	1.00E-05	0.01	1.00E-04
Iterations	40919		8198		1494		24293		12362		5690	
Training Cost	4890		5019		5084		4733		4780		4855	
Validation Cost	5355		5490		5562		**5189**		5239		**5318**	
Training Cost Delta	-6.212%		-5.375%		-1.741%		-7.349%		-6.744%		-6.260%	
Validation Cost Delta	-5.851%		-5.011%		-1.645%		-7.233%		-6.527%		-5.936%	

Logistic Regression: The cost comparison for regression model with top 10 features against random and all the variables is depicted below. We can see that the best cost for top model with top 10 variables is a little less or same against the model with ten random variables depicting that it is important to weed out the variable from the model that are collinear to each other and include those variables that are highly correlated to the target variables.

While in the comparison case where top 10 features model was compared to model with all the features it was found out that the cost in model with top 10 features was a little more in comparison to the model with all the variables thus showing that only correlation is not the criteria for choosing best features for a model. We could have considered experimenting with some other high-end models like random forest, regularization regression which helps us selecting the best features in a better way. The comparison metrics are shown below for logistic regression.



Top 10 Features	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations
	0.0001	10000	0.0001	25000	0.001	10000	0.001	25000	0.01	10000	0.01	25000
Train Accuracy	0.758		0.758		0.758		0.758		0.758		0.758	
Validation Accuracy	0.748		0.748		0.748		0.748		0.748		0.748	
All Features	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations
	0.0001	10000	0.0001	25000	0.001	10000	0.001	25000	0.01	10000	0.01	25000
Validation Accuracy	0.789		0.789		0.788		0.788		0.789		0.789	
Train Accuracy	0.794		0.794		0.794		0.794		0.794		0.794	
Delta Observed												
Training Cost Delta	-3.93%		-3.93%		-3.81%		-3.81%		-3.93%		-3.93%	
Validation Cost Delta	-5.79%		-5.79%		-5.79%		-5.79%		-5.79%		-5.79%	

Top 10 Features	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations
	0.0001	10000	0.0001	25000	0.001	10000	0.001	25000	0.01	10000	0.01	25000
Train Accuracy	0.758		0.758		0.758		0.758		0.758		0.758	
Validation Accuracy	0.748		0.748		0.748		0.748		0.748		0.748	
Ten Random Features	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations	Learning Rate	Iterations
	0.0001	10000	0.0001	25000	0.001	10000	0.001	25000	0.01	10000	0.01	25000
Validation Accuracy	0.748		0.748		0.748		0.748		0.748		0.748	
Train Accuracy	0.755		0.755		0.755		0.755		0.755		0.755	
Delta Observed												
Training Cost Delta	1.34%		1.34%		1.34%		1.34%		1.34%		1.34%	
Validation Cost Delta	-0.93%		-0.93%		-0.93%		-0.93%		-0.93%		-0.93%	

Discussion:

- Linear and logistic regression are the basic models and we could try working around with other non-parametric or regularized models which would have taken less novice approach to reach the final solution. For instance, ensemble-based methods like XG boosting takes previous errors into consideration and eventually helps in making better predictions
- According to me internal temperatures and weather parameters should be the best parameters to gauge what the energy usage at an instance would be because most of the time people stay alert based on the weather predictions and use AC, heaters and other appliance which consume more energy

- The number of data points available should have been more to make better predictions
- As discussed above we should have tried other methods for finding out the best features to build a model rather than correlation.
- Tuning more hyperparameters might have helped models achieved better accuracy and less cost as well
- Most of the temperature and weather-related parameters values were equal to mean thus making the variables useless. Thus, sanity of data should also be taken into consideration
- In order to make classification in logistic regression we could have considered developing a boundary using other methods like median, frequency etc as well
- Regularization methods like lasso and lego regression could have been used to penalize the errors more. In case of logistic regression method, we could have worked around to build the model from the scratch and used advanced optimization methods like newton optimization instead of gradient descent and see if the results improve or not.
- Weather recorded from the nearest airport station might not be that accurate and it could be checked
- Energy and humidity parameters were recorded for 3.3 mins and then were averaged at 10 mins which might not be able depict the true picture region wise accurately