

Regression Project

Executive Summary

In this project we have tried to predict the energy consumption of a house using Linear Regression from the SKlearn package and a custom implementation of the gradient descent for linear regression function and we have classified the houses into high consumption houses and low consumption houses using the Logistic regression

Conclusion

Had to do better Exploratory data analysis on the given data and preprocess the data better to get a better prediction result

Based on the given data the best alpha value is 0.2 and the efficient number of iterations to run the gradient descent is 1000

Dataset Preparation

A review of the data tells us that there are no NULL values and we do not have any missing values.

The data column as it not used by my Regression.

Part-1

I have randomly distributed the dataset into training and testing sets using the split percentage as 0.7 which means that 70% of data will be used in the training phase and 30% of the data will be used to validate the model

Part-2 and 3

I have designed the Linear Regression to model the energy usage of appliances. I have done it in 2 ways

- Using the Library of Sklearn to implement the Linear Regression model

The model equation is

$$\text{Appliances} = \text{Theta}_0 + \text{Theta}_1 X_1 + \text{Theta}_2 X_2 + \text{Theta}_3 X_3 + \dots + \text{Theta}_{27} X_{27}$$

- I have built a custom implementation of the gradient descent function which implements the Linear model. The initial parameters that I have got are

Intercept=98.0129	X1=0.0428664	X2=63.1533	X3=-42.0937	X4=-57.0755	X5=51.5884	X6=15.5299	X7=4.2924
X8=10.2659	X9=-0.664645	X10=2.25913	X11=39.7546	X12=8.28865	X13=4.11049	X14=-10.6005	X15=18.55

X16=-31.2992	X17=-45.4888	X18=6.08119	X19=-33.0298	X20=0.820169	X21=-1.65478	X22=5.48564	X23=2.28621
X25=6.37683	X26=-0.163698	X27=-0.163698					

I have included all the features available to except the date feature which I have removed.

Part-4

I have convert this problem into a binary classification problem using the “Appliances” column in the dataset. I have taken the mean value of the column and classified the home’s which consume electricity more then this particular mean value are classified as high consumption houses and the houses electricity lesser then this particular mean value are considered as low energy consumption houses. The accuracy I have got is 79.412 %

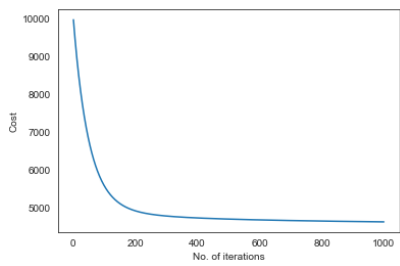
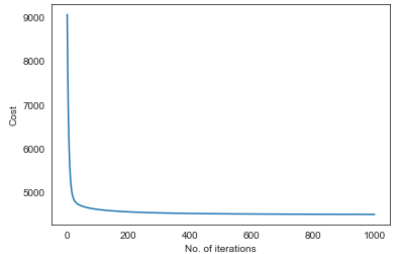
Experiments

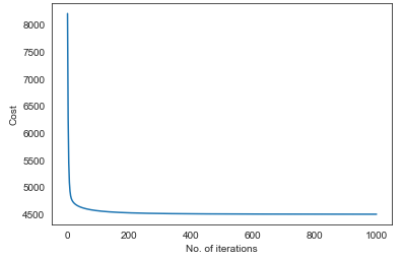
• Experiment-1

(1) Linear Regression

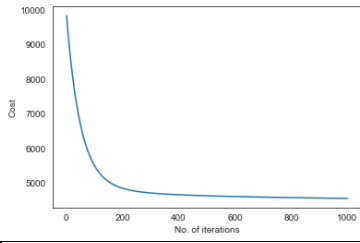
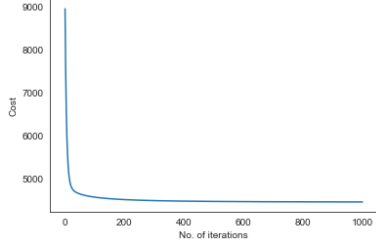
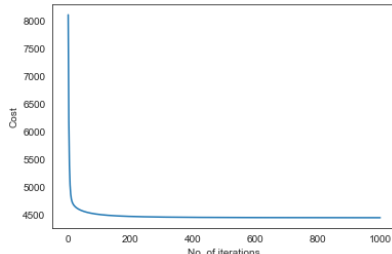
For this experiment, the maximum number of iterations for gradient descent is fixed at 1000 iterations. The learning rate experimented with are 0.01,0.04,0.005,0.1,0.2. The variations of train and test error for different values of alpha (learning rate) is plotted below:

(a) Training data

Alpha	Error	Graph	Iteration
0.01	4500.9913		1000
0.1	4505.6286		1000

0.2	4500.9876		1000
-----	-----------	--	------

(b) Testing Data

Alpha	Error	Graph	Iteration
0.01	4566.2996		1000
0.1	4453.7380		1000
0.2	4449.8531		1000

The best value for alpha (learning rate) is 0.2

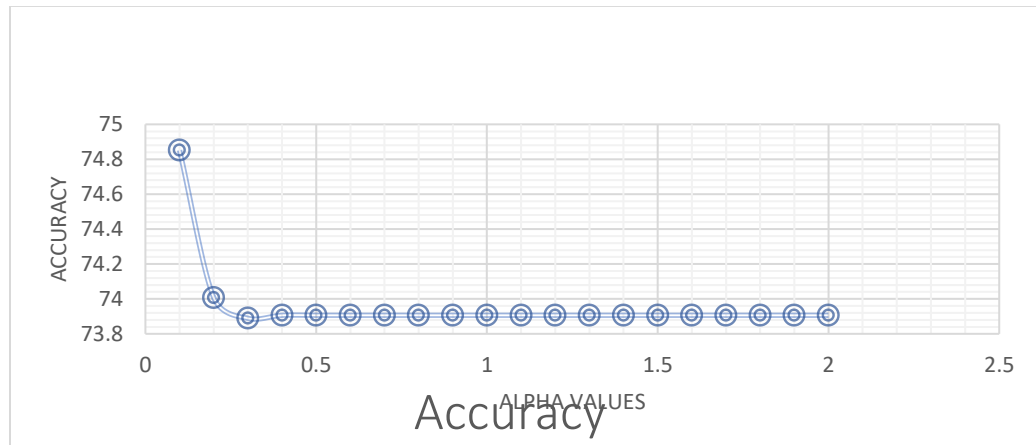
(2) Logistic Regression

As I have used the library implementation of Logistic Regression. I am not able to retrieve the cost vector to plot the graph.

Alpha	Accuracy	iterations
0.25	74.0077	100

0.75	73.9064	100
1	73.9064	100
1.25	73.9064	100

Below I have attached a graph in which the alpha(learning rate) is along the x-axis and the Accuracy is along the y-axis . From this graph we can infer that Accuracy is constant around 73% to 74%

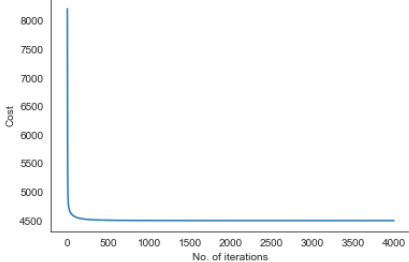


- Experiment-2

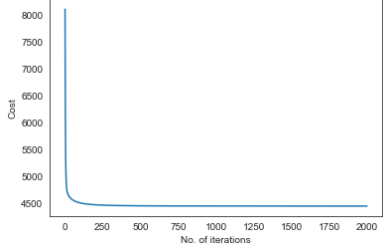
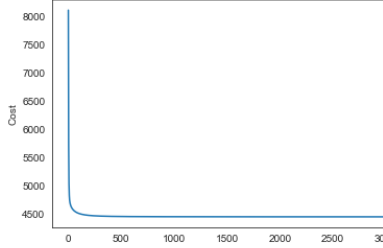
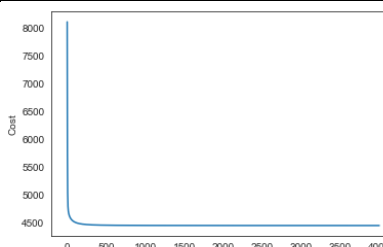
Linear Regression

(a) Training data

Threshold	Alpha	Iterations	Graph	Error
-0.00065	0.2	2000		4499.3388
-0.00014	0.2	3000		4499.000

$-3.26 \cdot 10^{-4}$	0.2	4000		4498.924
-----------------------	-----	------	--	----------

(b) Testing data

Threshold	Alpha	Iterations	Graph	Error
-0.000594	0.2	2000		4448.4374
-0.000140	0.2	3000		4448.1239
$-3.3224 \cdot 10^{-5}$	0.2	4000		4448.0497

Experiment-3

I have picked the first 10 columns as random data and built the 3 models on that and I have reported the Error/Accuracy with comparisons.

- Sklearn Linear Regression

Models	R ² values
Intial model that was built	14%
Model built using the 10 random features	9.4%

- Custom Gradient Descent for Linear Regression

Training data

Models	Error
Intial model that was built	4500.9876
Model built using the 10 random features	4772.4227

Testing data

Models	Error
Intial model that was built	4449.8531
Model built using the 10 random features	4729.2233

- Logistic Regression

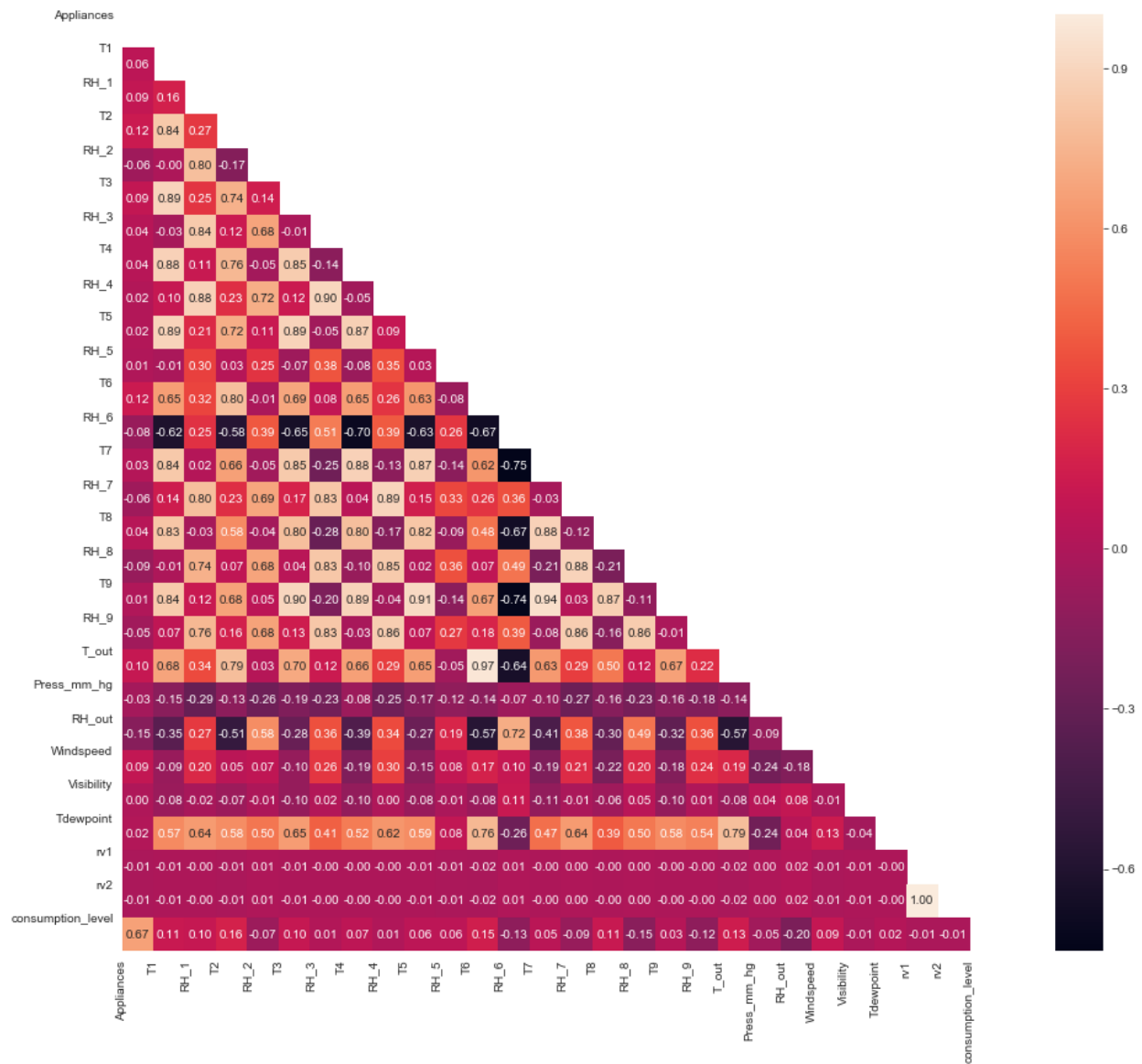
• Models	Accuracy score
Intial model that was built	74%
Model built using the 10 random features	74%

Experiment-4

The correlation matrix will help us choose the best features. The correlation matrix for the dataset is. We will the top 10 highly correlated features with the

target variable. These 10 features are taken from the correlation matrix given below.

T2	T6
T_out	Windspeed
RH_1	T3
T1	T4
T8	RH_3



- Sklearn Linear Regression

Models	R^2 values
Intial model that was built	14%
Model built using the 10 picked features	5%

- Custom Gradient Descent for Linear Regression

Training data

Models	Error
Intial model that was built	4500.9876
Model built using the 10 picked features	4609.2233

Testing data

Models	Error
Intial model that was built	4449.8531
Model built using the 10 picked features	4613.23

- Logistic Regression

• Models	Accuracy score
Intial model that was built	74%
Model built using the 10 picked features	73%
