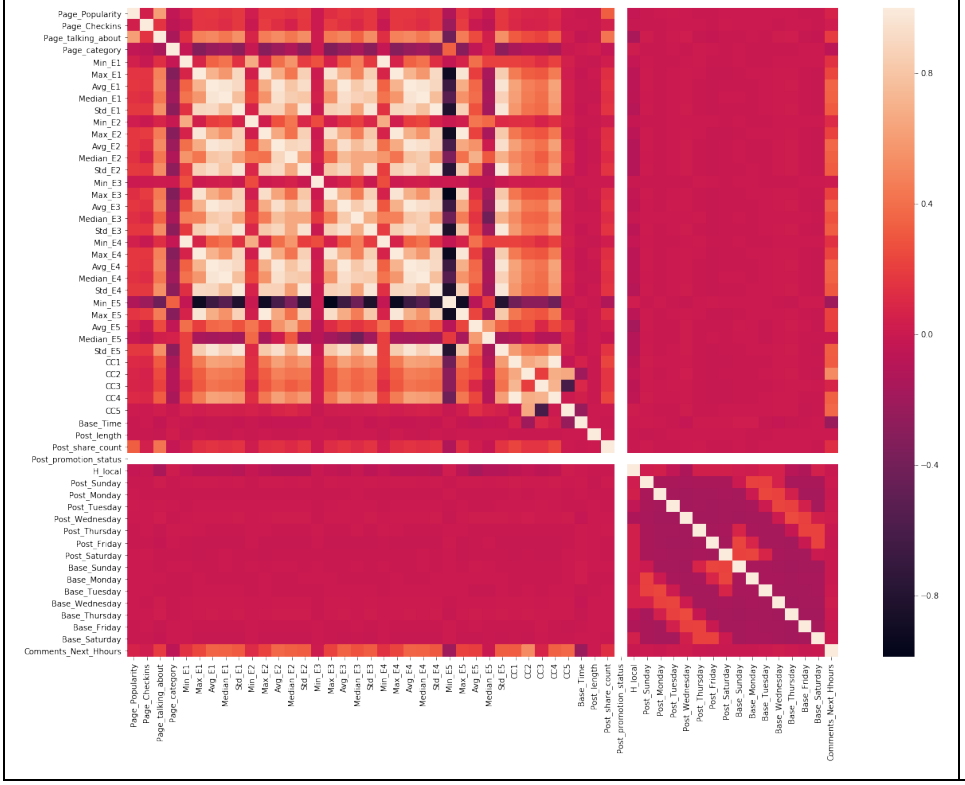
ML ASSIGNMENT HOMEWORK-1 REPORT

CHINAR ARORA

Exploratory Data Analysis:

Trying to understand the relationship between different predictors and response variable.

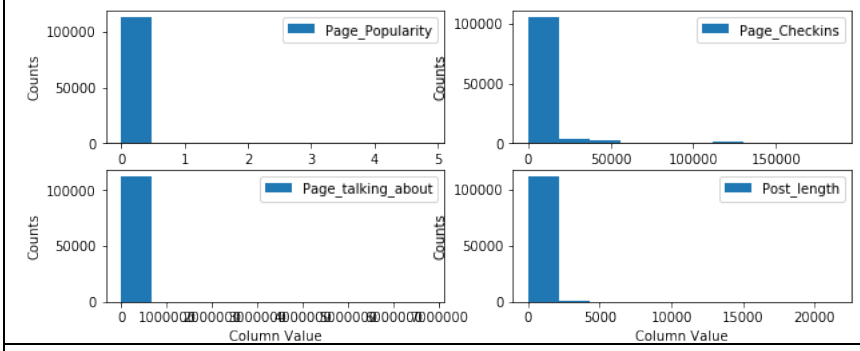


From the correlation plot obtained, it can be interpreted that **Min E5**,**Base_Time** is **negatively correlated** with the target variable.

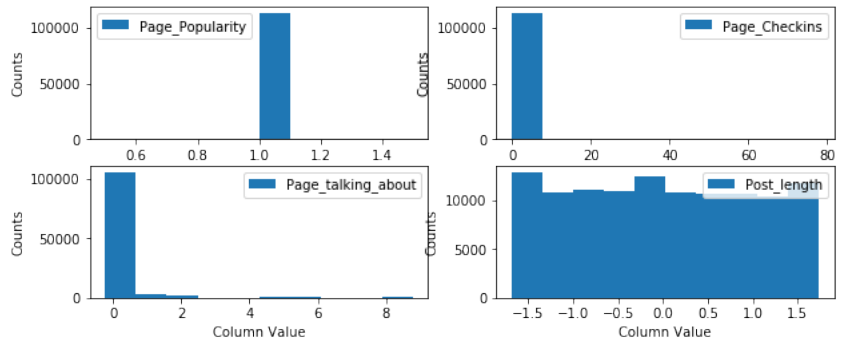
There seems to be **positive correlation** between **CC5,CC4** with target variables also.

Other features do not seem to be correlated strongly with the target variable.

Data Pre Processing:



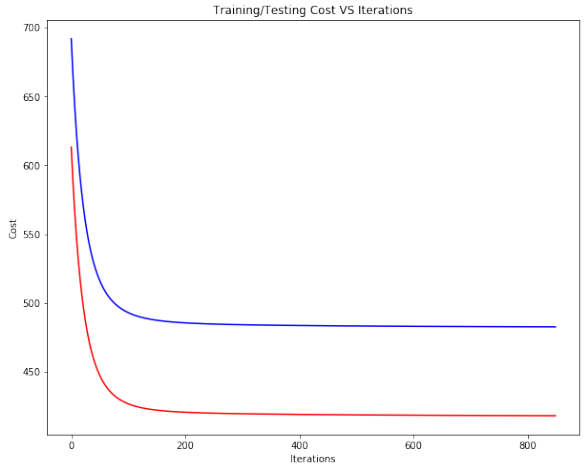
All features have different scales as it can be seen in this plot.



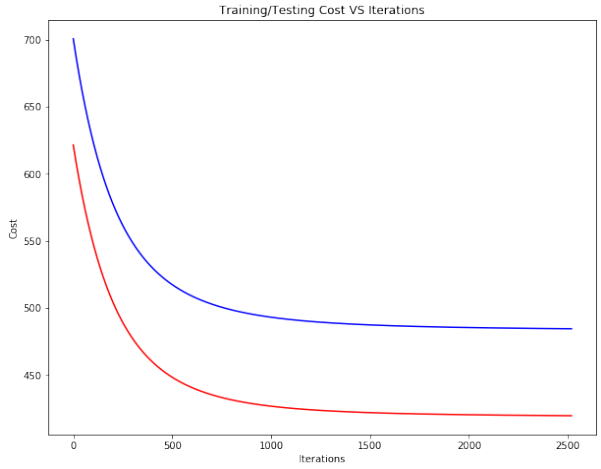
In order for the gradient descent to converge faster, all features should be normalized. After, normalization, we see that values are scaled.

From exploratory Data Analysis 13 features have been chosen.
They are: 'Page_Popularity','Page_Checkins','Page_talking_about','Base_Time','Post_length', 'Post_share_count','Post_promotion_status','CC1','CC2','CC3','CC4','CC5','H_local'.

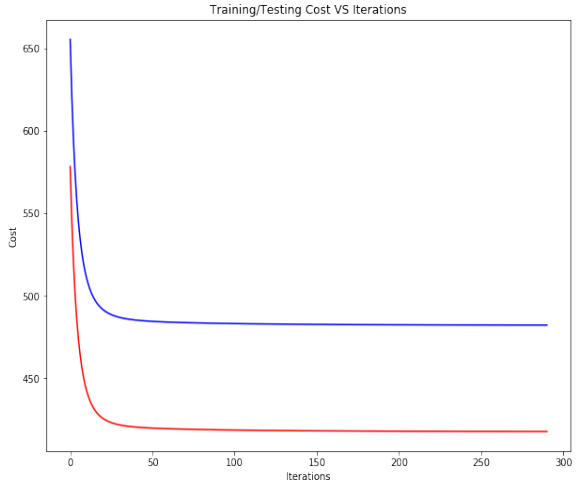
Experiment 1: Identifying the correct value for learning parameter.
In gradient descent, the size of the steps is determined by the learning parameter alpha.
If the learning rate is too small, then the algorithm will have to go through many iterations to converge, which will take a long time. If the learning rate is too high, you might jump across the valley and end up on the other side, possibly even higher up than you were before.
*The blue line represents the learning curve for the test set and the red line represents the learning curve for the training set.



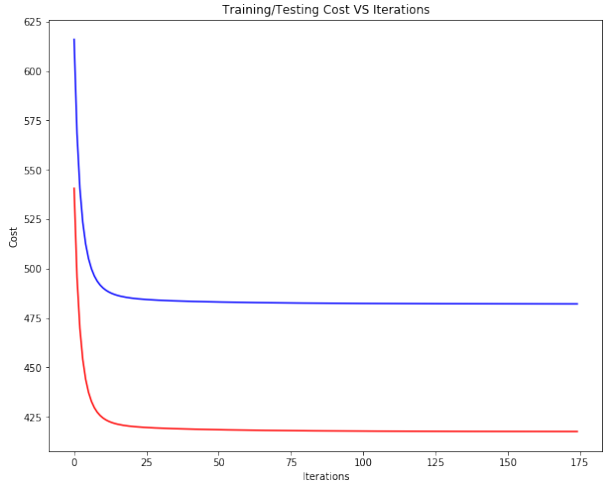
Alpha=0.01
Iterations=850
Convergence Criteria=0.001
Train Set Cost:417.94
Test Set Cost:482.53



Alpha=0.001
Iterations=2519
Convergence Criteria=0.001
Train Set Cost:419.76
Test Set Cost:484.57



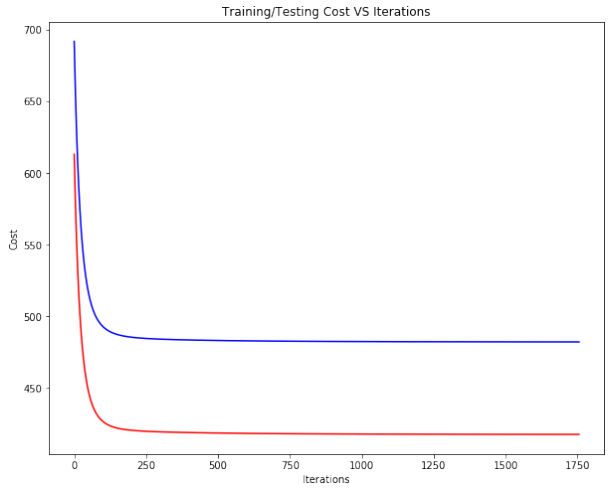
Alpha=0.05
Iterations=291
Convergence Criteria=0.001
Train Set Cost:417.64
Test Set Cost:482.21



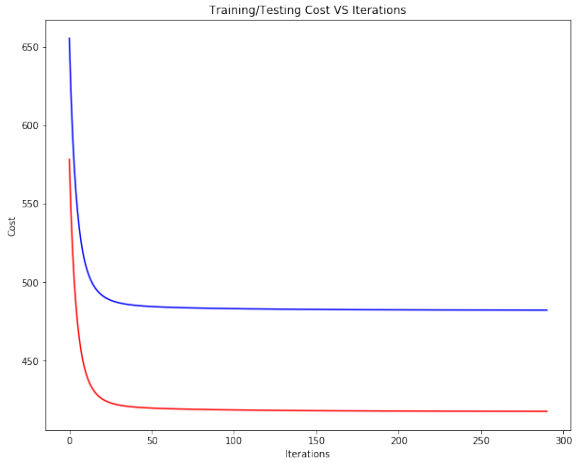
Alpha=0.1
Iterations=175
Convergence Criteria=0.001
Train Set Cost:417.60
Test Set Cost:482.17

It can be observed that for alpha=0.0001, the convergence occurs after 2519 iterations. As the time taken to converge will be large in this case this value is not chosen. For alpha=0.01, 850 iterations are required to converge to global minimum and for alpha=0.05, 291. The lowest number of iterations are taken by alpha=0.01 that is 175. Also, the train and test cost is considerably lower than for alpha=0.001,0.01 or 0.05. **So, I have chosen alpha=0.1 as it converges fast and has lower cost also.**

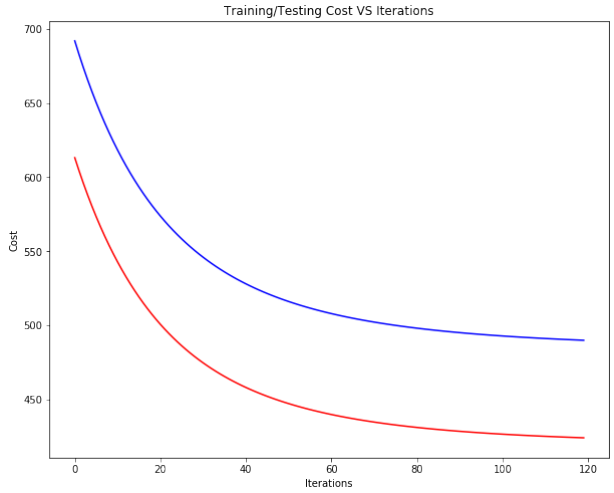
Experiment 2: Experimenting with different values of convergence criteria(threshold).



Alpha=0.01
Iterations=1756
Convergence Criteria=0.0001
Train Set Cost:417.64
Test Set Cost:482.53



Alpha=0.01
Iterations=850
Convergence Criteria=0.001
Train Set Cost: 417.94
Test Set Cost: 482.53

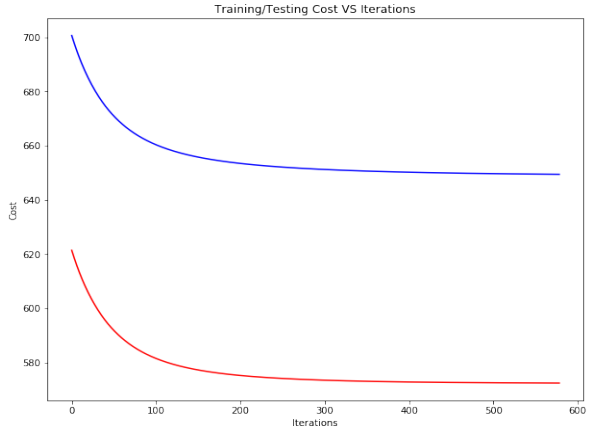


Alpha=0.01
Iterations=291
Convergence Criteria=0.1
Train Set Cost:424.03
Test Set Cost:489.82

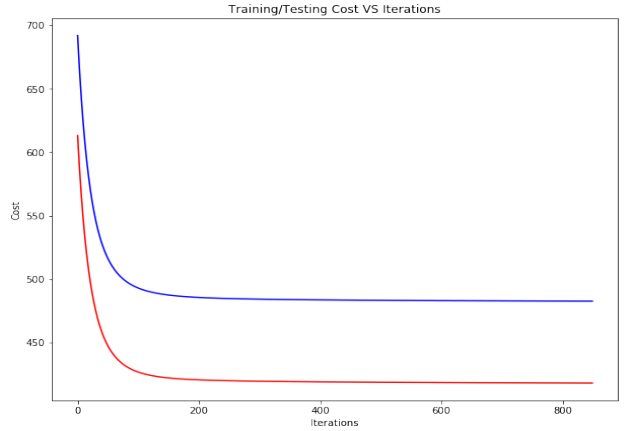
It can be observed that for threshold=0.0001, we need 1756 iterations which is quite large. For threshold=0.01, we get lower iterations (291). However, as compared to threshold=0.001, it has a higher train and test cost. So, threshold chosen is **0.001.**

Experiment 3:

The 5 Random features chosen are: Page Popularity, Page Checkins, Page_talking_about, H_local, Post_length.

<table>
<tr>
<td> 

Learning curve for random 5 features:

Alpha=0.01
Iterations=579
Convergence Criteria=0.001
Train Set Cost:572.47
Test Set Cost:649.47</td>
<td> 

Learning curve for 14 selected features:

Alpha=0.01
Iterations=850
Convergence Criteria=0.001
Train Set Cost:417.94
Test Set Cost:482.53</td>
</tr>
</table>

The randomly chosen 5 features have higher train and test cost as compared to the model initially chosen with 14 features.

Experiment 4:

Feature Selection using Random Forest.

The ExtraTreeClassifier helps to find the most significant variables. This is because random forests will always a choose a subset of random variables and so it removes correlation from other trees. The random forest uses the mean decrease accuracy for feature selection. This mean it selects the predictor and cutpoint such that it leads to the minimum decrease in the sum of residual errors.

In the given case, we find that the important variables are:

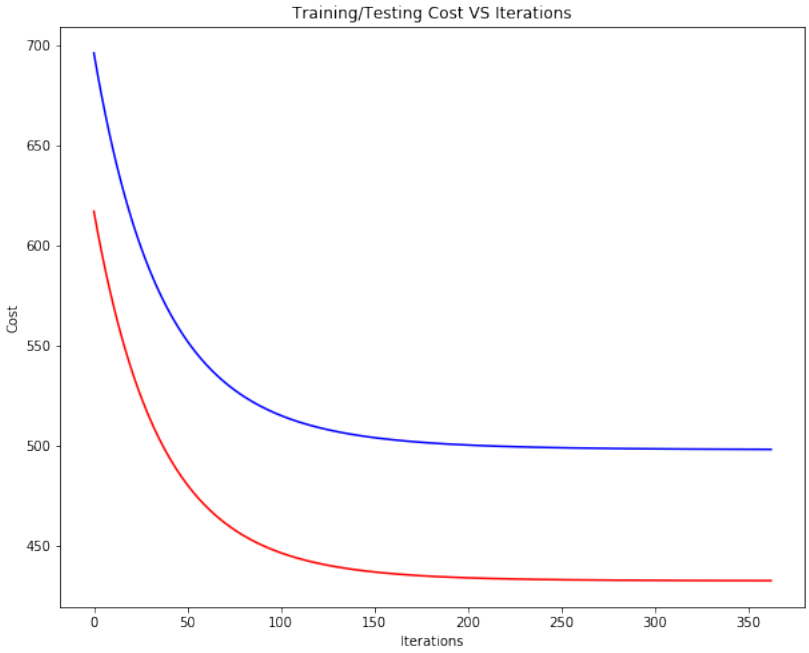
CC1-Number of comments before selected base time

CC5- Difference between the number of comments in last 24 hours and last 48 hours

CC4-The number of comments in the last 24 hours

Post length-Character count in the post

Base time-Selected time



Learning curve for 5 most relevant features:

Alpha=0.1
Iterations=47
Convergence Criteria=0.001
Train Set Cost:432.75
Test Set Cost:498.12

We get the lowest train and test cost with the 5 most relevant features.

- Compared to the random features which takes 579 iterations to converge, this model takes only 47 iterations to converge. Hence it is much faster to reach the global optimum. Also, the train and test costs are lesser.
- Compared to the model initially selected with 14 features, this model takes lesser time to converge and the train/test costs are also low.

Final Equation:

Comments_In_Next_Hhours= 7.15+11.43*CC5+5.73*CC1+6.26*CC4+0.145*Post_Length -5.69*Base_time

Learning Outcomes:

- Our experimentation has revealed that much of the comment volume of a post is determined by the **length of the post** and the **number of posts in the preceding 24 hours. Other factors like page likes, page checkins are relatively less important.**
- I also learnt how to implement gradient descent to find the global maximum of the cost function(mean squared errors) for linear regression.