Name: Gaurav Dhavale **NET ID: gdd160130**

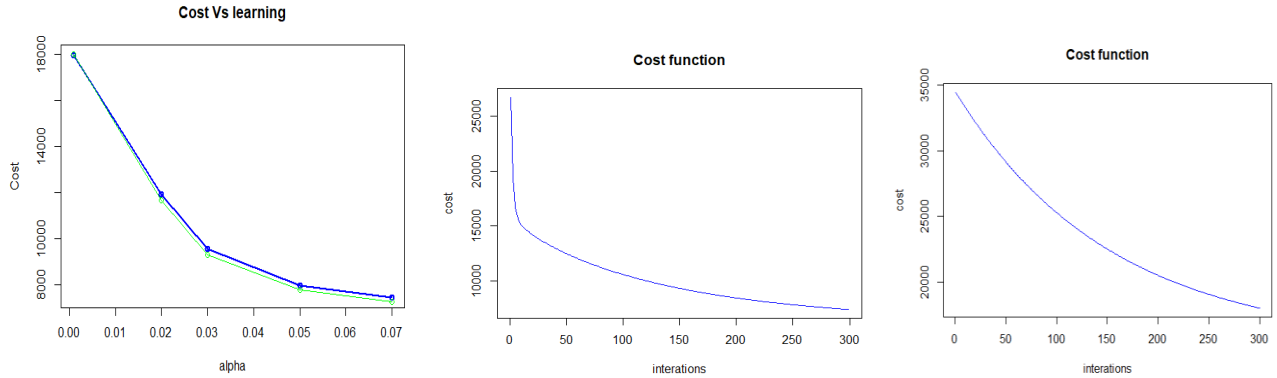
Task 1: Divided main dataset into train and test dataset with the ratio off 70/30.

Task 2: regression model at the end of report

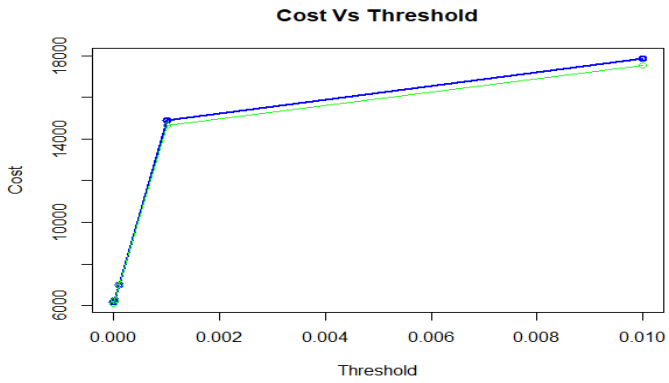
Task 3: Initial parameter: alpha <- 0.01, Threshold <- 0.00001, theta <- matrix(c(0,0), nrow=53)

Experiment 1:

Calculated cost for train and test dataset for 5 different alpha values which are 0.001, 0.009,0.01,0.05,0.07. The green line for test and blue line for train dataset. The next two graph shows cost behavior for alpha 0.001 (middle) and 0.07 (right) against no of iterations. **When alpha is smaller, the cost reduces slowly as compare to when alpha is 0.07**

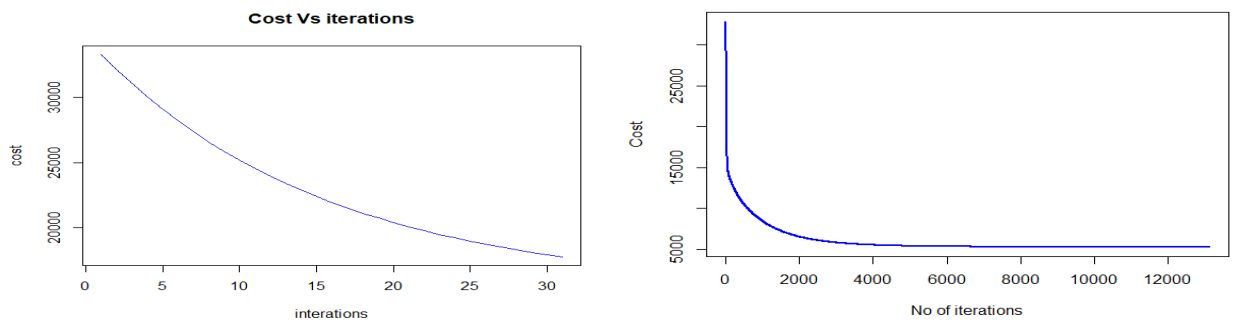


Experiment 2: Graph of cost Vs Threshold Observation: Calculated the cost for 6 different thresholds stating from 0.01 to 0.000001. The model convergence value is lower for lower threshold values. Green line: test dataset, Blue line: train dataset

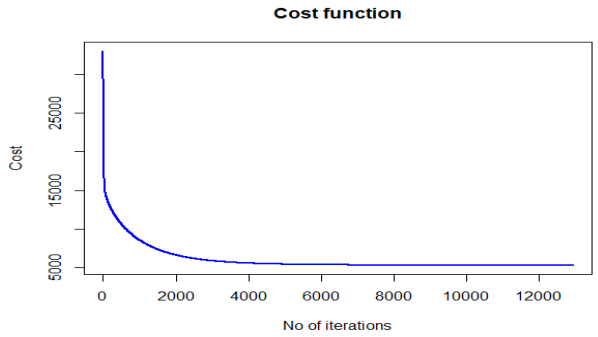


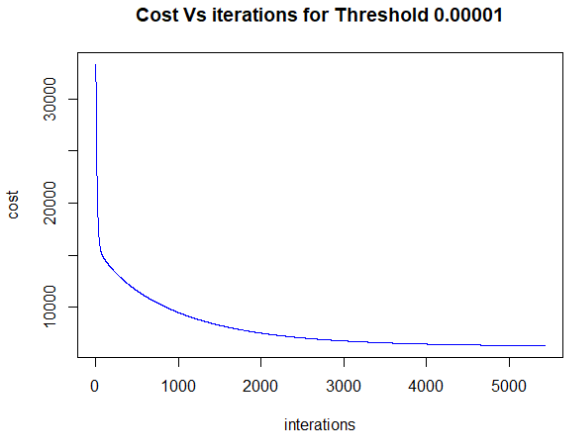
The following three graphs shows cost functions against no off iterations. 1st one is with highest threshold value: 0.01 and 2nd graph is with threshold 0.000001. Both the graph plotted for train dataset.

The 2nd graph shows model is converge after 10k iteration.



1st graph is for 0.000001 threshold but for test dataset. **Found better convergence for threshold: 0.00001**.

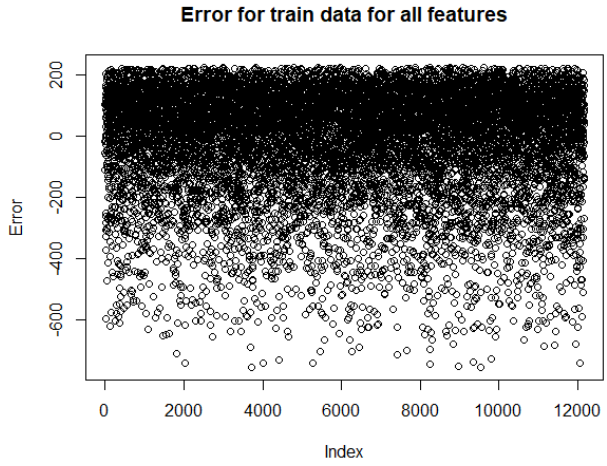


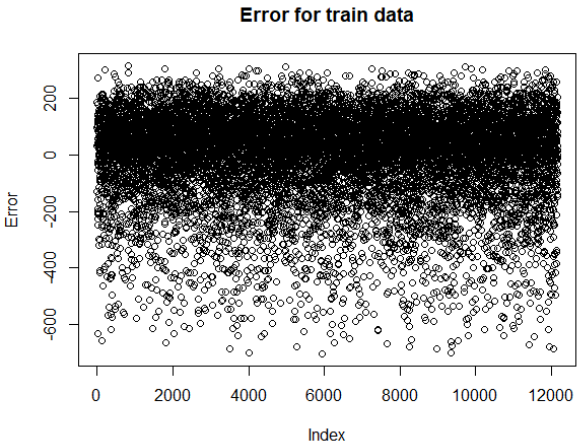


Experiment 3:

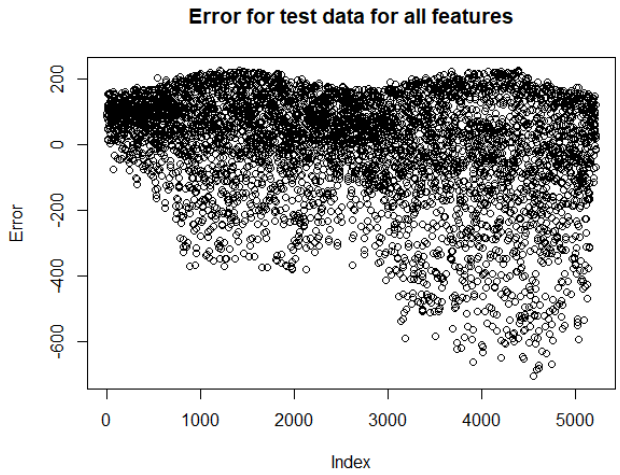
Randomly selected values: hum, windspeed, temp

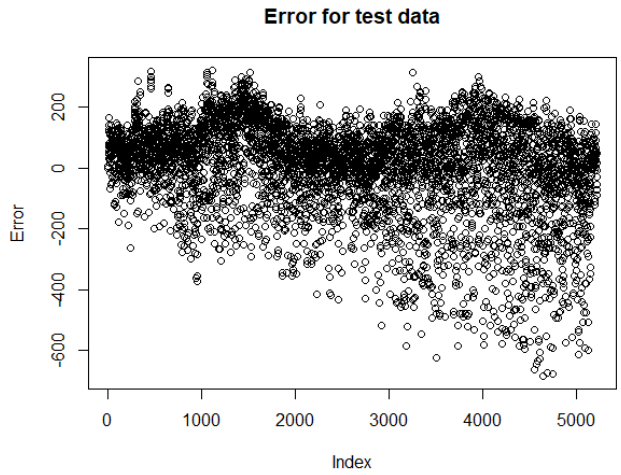
<u>Error plot for train dataset for all features and for 3 features</u>





<u>Error plot for test dataset for all features and for 3 features</u>



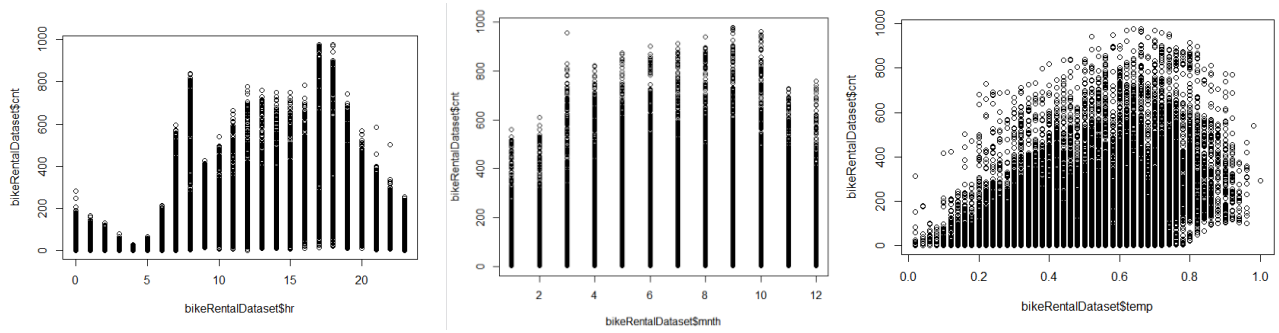


Randomly selected variables error graph's pattern is almost similar to all features graph for both dataset

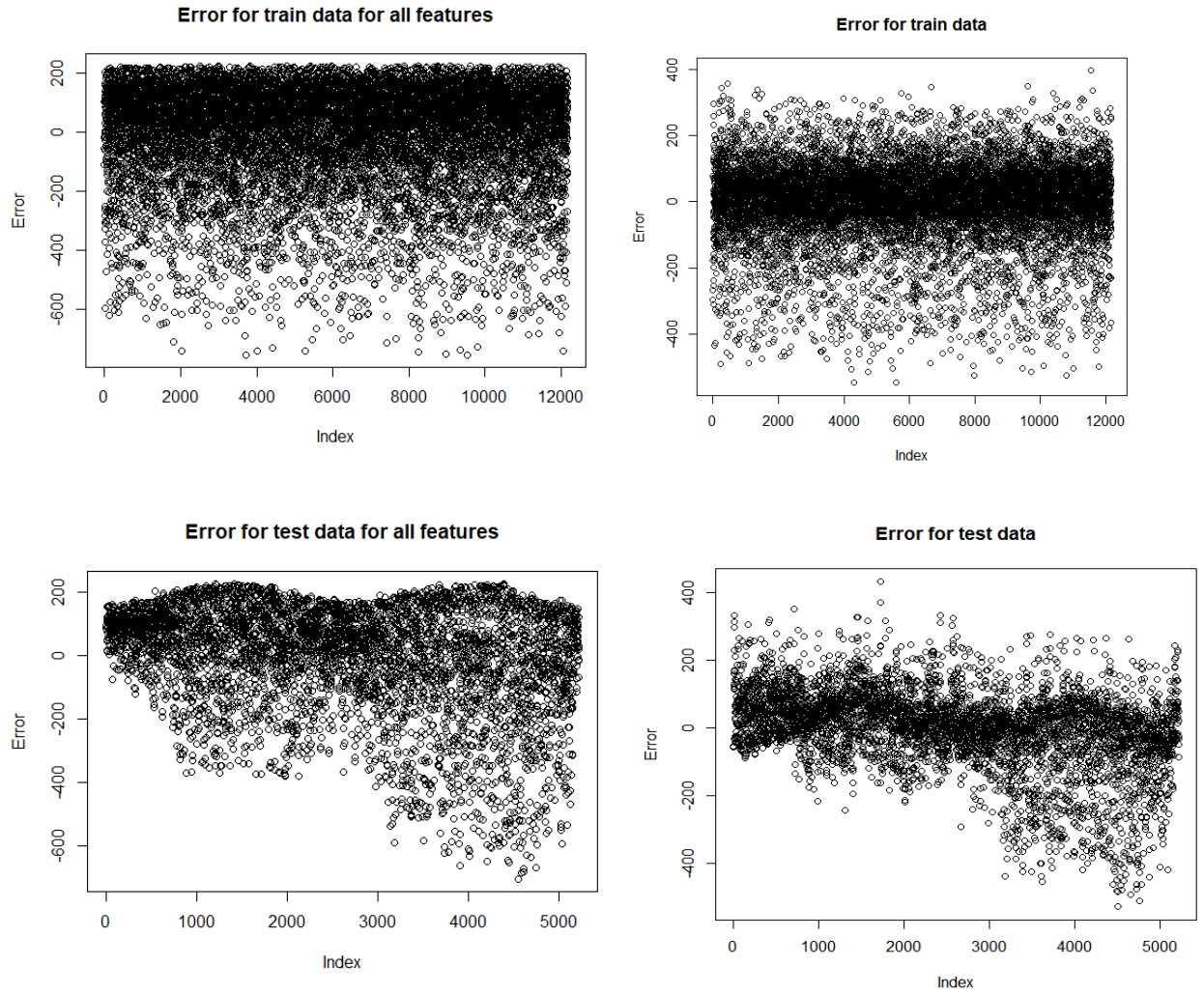
Experiment 4:

Based on correlation plotted for variable hr, mnth, temp verses target variable, I observed target variable varies for hr, mnth, temp variable.

The relation for these three variables with target variable is as follows:



Error graph for train and test dataset for all features and 3 features.



In the above graphs, we can see that the points for 3 features train and test graphs are concentrated near 0 which means predicted and actual y values are nearby.

Discussion:

To improve model, the following steps can be taken:

- Generate model based on optimal learning rate and threshold found during the experimentation
- Check correlation between features if any and then we can add correlation term in the equation. We can also improve model by removing features which are not impacting the target.
- Check the outliers and try to exclude these values
- Perform featured scaling if required

Interpretation of results: bike rentals measuredly depends on season, temp, hr. I found there are no much variations in the result for humidity. Bike rental increases during evening hours 5-6pm and morning hour 7am.

To improve the results, I retrained my models new alpha and thresholds obtained from experimentation results. I have also compared the train and test error results**. Model equation with alpha <- 0.01 & Threshold <- 0.00001:**

Regression model Equation:

79.50616551

- 39.75053313 * B1 - 1.53033597 * B2 - 19.23613674 * B3 - 7.84560326 * B4 - 7.33710949 * B5 + 4.65402575 * B6 - 4.66759031 * B7 + 11.75778519 * B8 - 3.91185228 * B9 - 21.229589 * B10 + 3.04190112 * B11 + 38.89436248 * B12 + 38.10335209 * B13 + 10.16550131 * B14 - 97.89199256 * B15 - 117.81477302 * B16 - 121.42239708 * B17 - 131.23782628 * B18 - 134.5988241 * B19 - 118.41800024 * B20 - 70.58612826 * B21 + 51.86229318 * B22 + 170.44382337 * B23 + 44.18518664 * B24 - 7.55592054 * B25 + 14.91109662 * B26 + 54.93593565 * B27 + 45.73160861 * B28 + 23.77638205 * B29 + 37.0381297 * B30 + 92.90233114 * B31 + 226.32415011 * B32 + 204.2749751 * B33 + 105.92858804 * B34 + 36.96138456 * B35 - 4.73763984 * B36 - 38.01238831 * B37 - 11.05901794 * B38 - 9.34628387 * B39 - 7.10273038 * B40 - 1.5187081 * B41 - 0.01388317 * B42 - 3.97658936 * B43 + 3.41592116 * B44 + 1.86302809 * B45 + 50.78581356 * B46 + 43.0367632 * B47 - 14.62117622 * B48 + 138.74031845 * temp + 127.87815398 * atemp - 109.83253334 * hum + 23.05902765 * windspeed

Variable Meaning: B1 to B3: Season, B4 to B14: mnth, B15 to B37: hr, B38: holiday, B39 to B44: weekday, B45: workingday, and B46 to B48: weathersit

Regression model Line for above equation: