Machine Learning Assignment-1 Sathvik Reddy Junutula (sxj174230)

Data Definition: The data consists of 61 variables which have one predicted, 58 independent and 2 indicators to provide the metrics of an online site that has articles which are shared by the users. The idea of the problem is to predict the number of shares upon giving the new data. The data consists of mixed independent variables such as continuous and categorical.

Agenda: To be able to download, manipulate, divide, and experiment based on the questions posted as part of the assignment using R software.

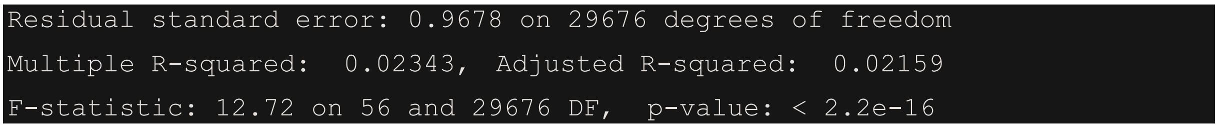
Workflow:

1) Importing and cleaning: The whole dataset is imported to the software and is labelled accordingly. Unnecessary variables that aren't useful in the experimentation, like the date&time stamp and the website URL have been removed.

2) Scaling the data: For better results all the variables are mean normalized using the formula Xj := (xj-$\overline{x}$)/SD(x) as to bring all the data units in accordance and normalize it.

Experiment1: To run the linear regression model to predict the shares.

The data has been divided randomly to training and testing to predict the shares. The ration to train and test has been used as 3:1 and below are the summary statistics for the predictions. Below are the summary statistics of the model:



```
Residual standard error: 0.9678 on 29676 degrees of freedom
Multiple R-squared:  0.02343,  Adjusted R-squared:  0.02159
F-statistic: 12.72 on 56 and 29676 DF,  p-value: < 2.2e-16
```

RMSE on train data : 0.99

RMSE on test data : 1.05

Goind ahead as questioned by the assignment the following procedures were implemented to aachieve the goal. Also, the experimentations are listed below.

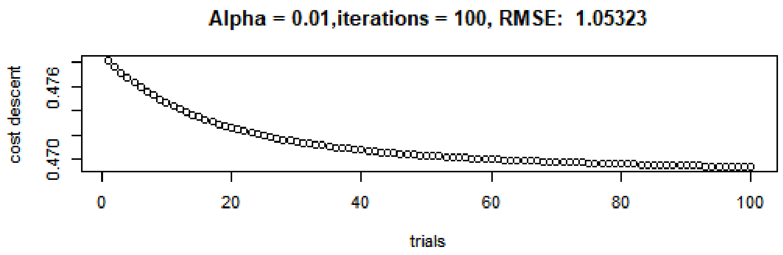
1) Alpha, learning rate
2) Iterations = number of times the coefficients are modified

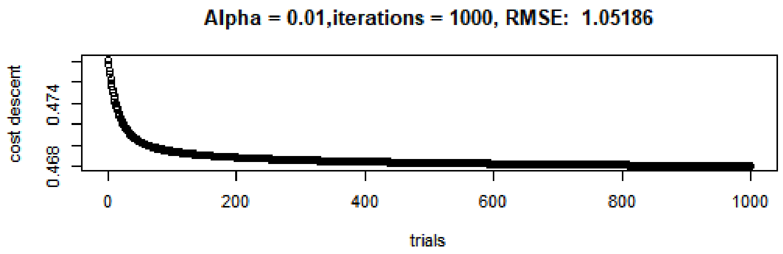
3) Gradient descent to be achieved in the final iteration is left out and the only parameters were changed were iterations and alpha to leave it to machine to calculate the values. Both of these methods were applied to Linear Regression as well as logistic regression in further part of the assignment.

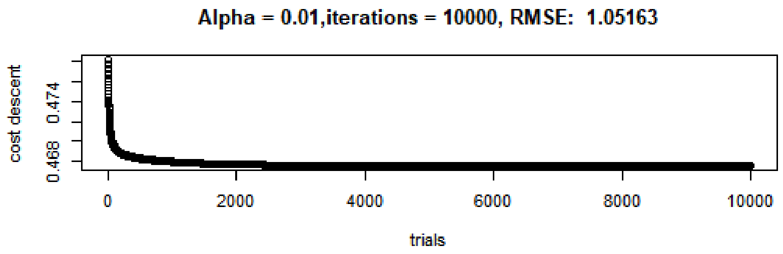
Expreriment 2:

a)To run the gradient descent algorithms on the model's to achieve the best coefficients the data is converted to a gradient descent problem and the code is generated to achieve the best coefficients, and predict the test set data based on the coefficients and calculate the RMSE. Upon running the model with all the variables below are the graphs of cost function versus the number of iterations based on the experiments. The experiments included were alpha(learning rate) and the number of iterations which are mentioned on the graph's subject line.

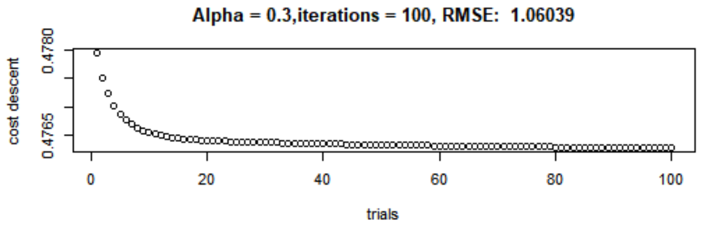
Below is the graph that's generated for running gradient descent for linear regression on all the variables.

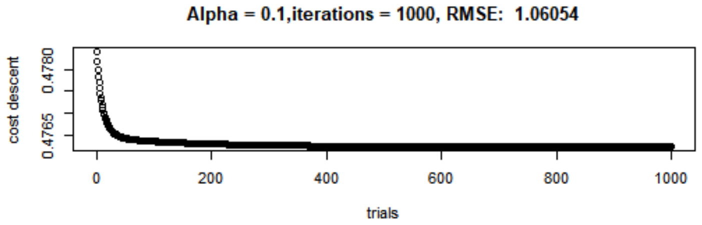


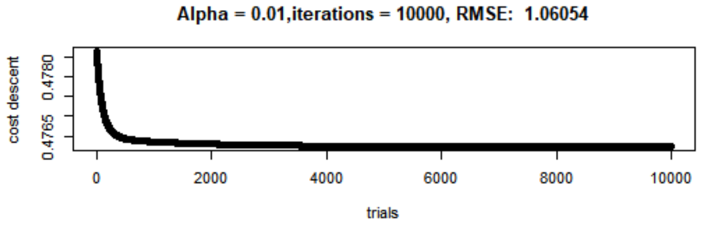




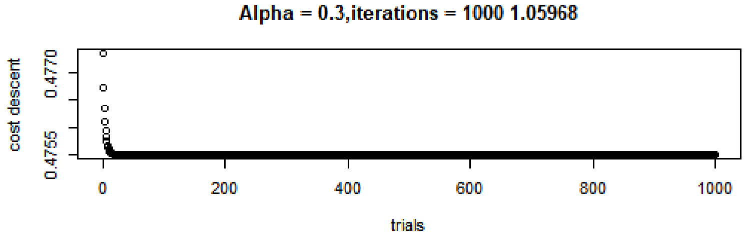
b) Below are the graphs for running the gradient descent algorithms on ten randomly selected models.

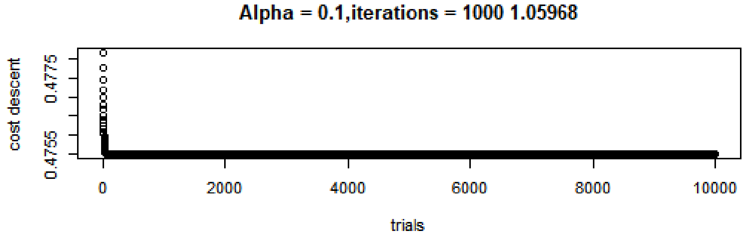

Alpha = 0.3,iterations = 100, RMSE: 1.06039

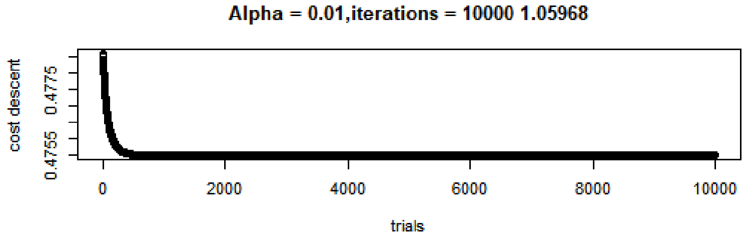

Alpha = 0.1,iterations = 1000, RMSE: 1.06054


Alpha = 0.01,iterations = 10000, RMSE: 1.06054

c) Below are the graphs for the linear regression for the best 10 variables:


Alpha = 0.3,iterations = 1000 1.05968


Alpha = 0.1,iterations = 1000 1.05968

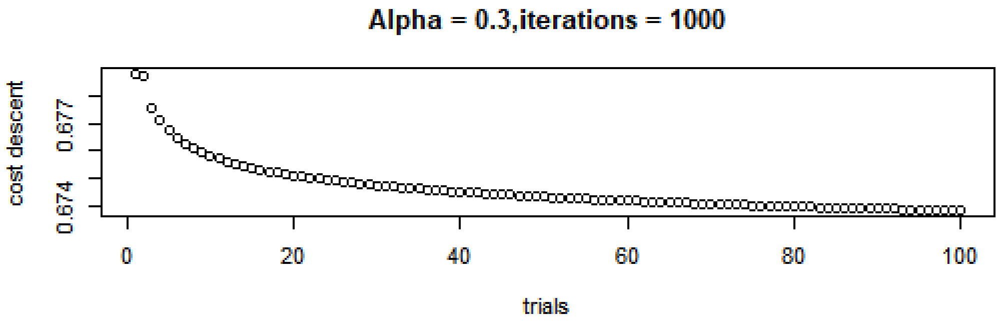

Alpha = 0.01,iterations = 10000 1.05968

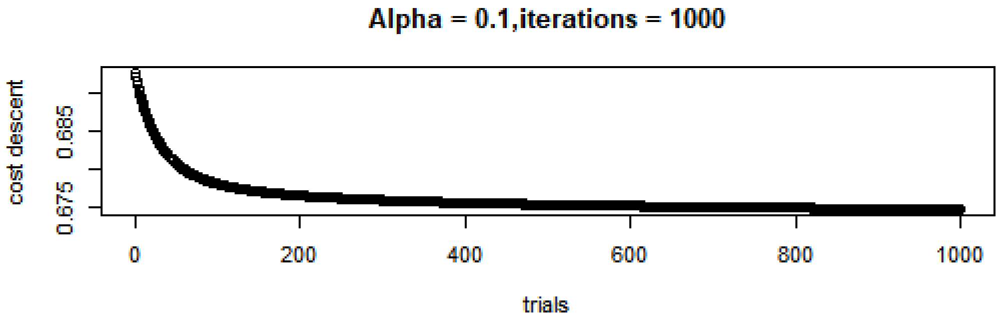
Gradient Descent for Logistic Regression:

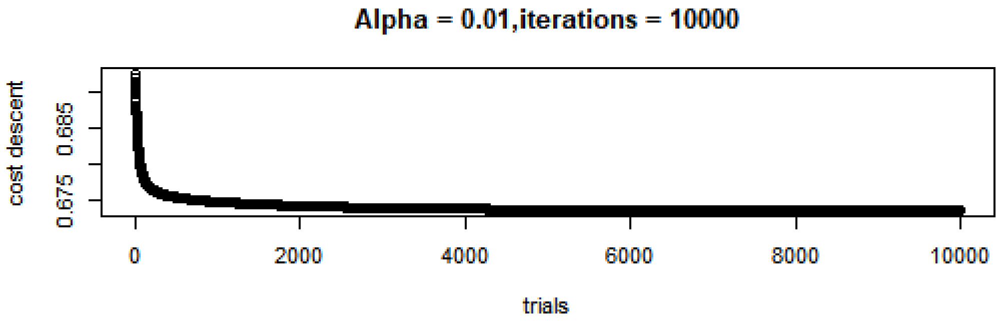
Experiment3: Logistic Regression is carried on the train and test set for all the variables by creating a boundary at the values where the median of the logistic prediction takes places and in this case it is .14. All the values that are above 0.14 were assigned a value of 1 and less than it were allowed 0.

Gradient Descent for Logstic Regression for all variables

Note: Only 100 labels misread them as 1000 in header of plot

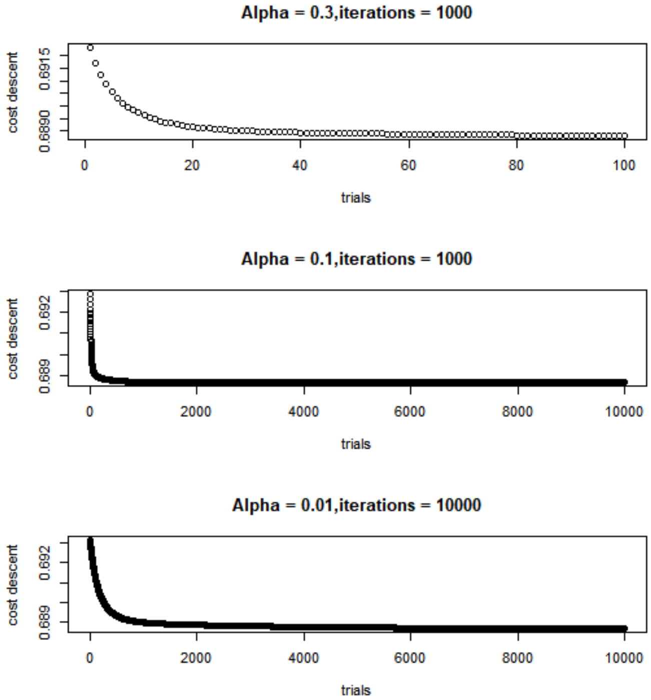

Alpha = 0.3,iterations = 1000


Alpha = 0.1,iterations = 1000

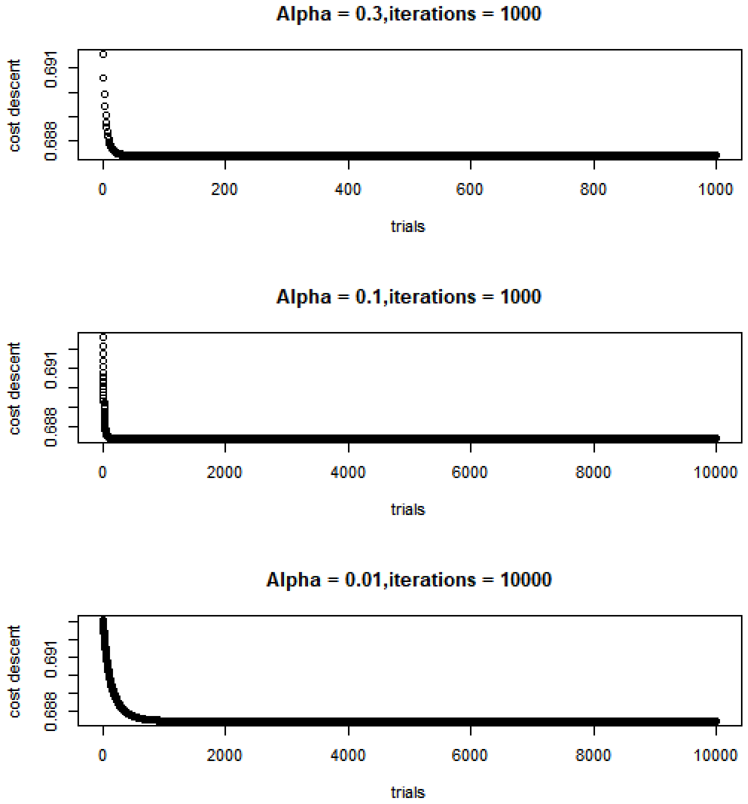

Alpha = 0.01,iterations = 10000

Below is the graph for Logistic Gradient Descent for 10 random variables:

Note: Only 100 iterations, misread them as 1000 in the main plot.



Below is the graph for logistic Gradient descent for best 10 variables:



Observations::

1) For linear Regression the cost function decreases as the number of iterations increases. The rate of decreasing the cost function increases rapidly as we go from all the variables to random 10 variables to top 10 variables that intuitively are best in their performance(Hints taken from running basic linear regression significant variables and by comparing the basic AIC BIC tests(forward and backward).
2) For logistic regression the sigmoid function is passed into the cost function so that it is calculated accordingly. In this case also the rate of decreasing cost function decreases rapidly from all variables to random 10 variables to top 10 variables.

3) The cost function had an extremely erratic behaviour when the data wasn't scaled initially hence the scaling had to be done.

4) When started with alpha value greater than 1, the cost function dropped down so dramatically that even after increasing the number of iterations to 1M the bend in the graph was very less, so I had to stick up to only 3 alpha values and to show the values, I considered 3 different values for different observations.

References: In the process of generating the code and understanding the concepts further, I had to take help online from a few resources and use their ideas to understand the concepts. Below is the list of the websites that I visited for the same.

1) https://www.ocf.berkeley.edu/~janastas/stochastic-gradient-descent-in-r.html
2) https://www.r-bloggers.com/linear-regression-by-gradient-descent/

5) Future work: Instead of the redundant code I will work on writing the functions for the same to improve the performance of the code and also to reduce the computational cost.