Dheeraj Goud Borlla (dxb160130)

ASSIGNMENT – #6: (BUILDING A PREDICTIVE MODEL)

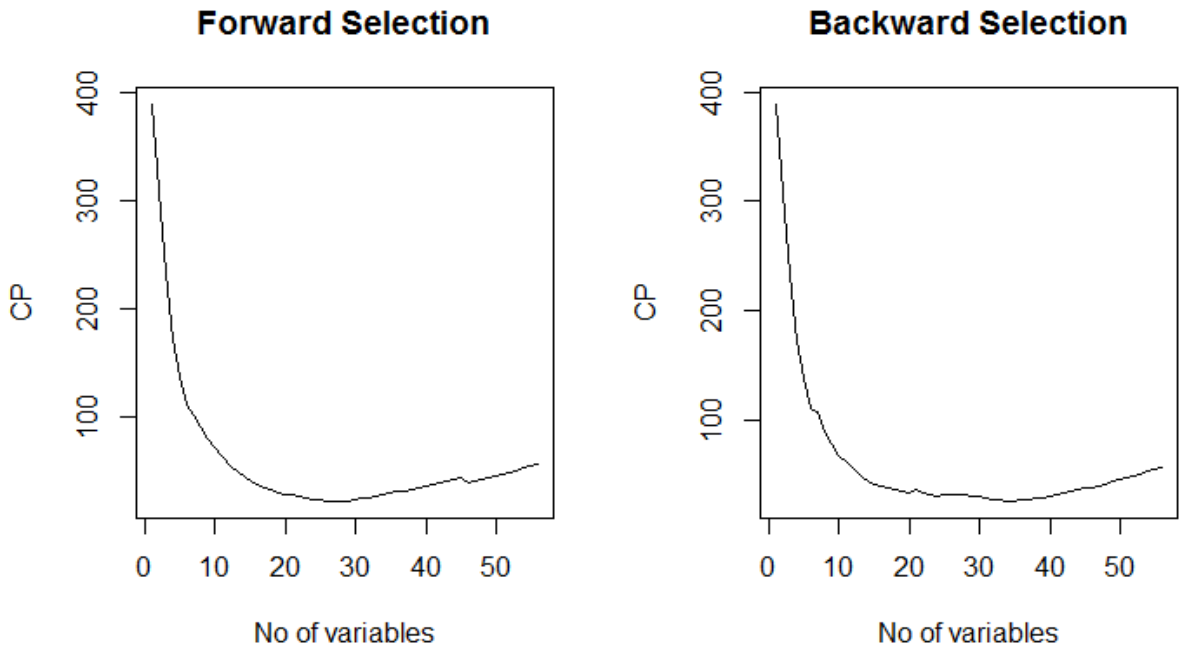
***PURPOSE*:** In this assignment, we analyse the online news popularity dataset and apply different techniques like forward & Backward subset selections and Lasso & Ridge regression, try to create a model that will predict the number of shares for an article based upon other predictors.

DATASETS: Online News Popularity Data Set (Link).

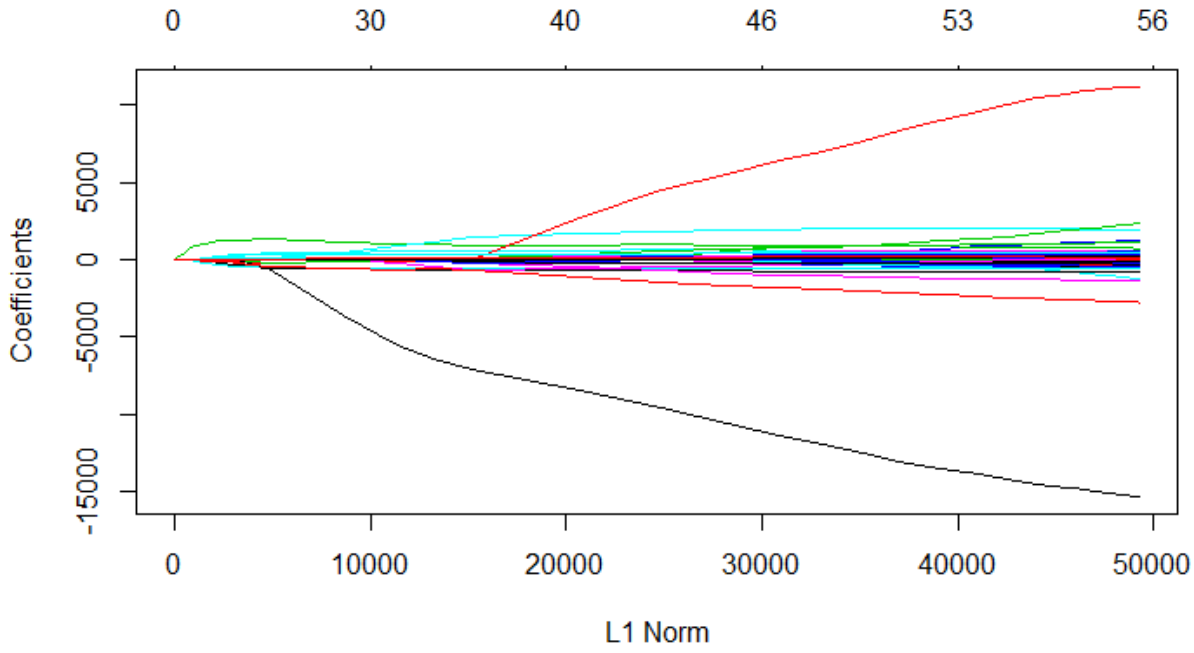
APPROACH:

- The dataset has 61 variables, '*shares'* is the target attribute and the other are used to predict the target but few of them which are unpredictable are removed.
- First I have fitted a linear model with *shares* as response variable and others as predictors. When we did the summary on the linear model, we observed the 'Adjusted R-squared' is 0.0217 and few variables gave '*NA*' in coefficient calculation, indicating those variables are linear combinations of others. So, these variables are removed from the dataset.
- Now we tried to apply the '*regsubsets*' function for model selection. We try to select the subsets of the predictors by doing forward and backward stepwise selection.
- First, I have applied the forward subset selection on the dataset. The summary of this model gives the value of '*cp*' and other values for each possible model with subset of predictors.
- Similarly, backward subset selection is applied on the dataset. And from the summary of the model gives us information about *cp, adjr2* and other values.
- From the graph 1, plots of '*cp*' values from forward selection model and backward selection model, the number of predictors with the minimum value of *cp* is selected. For forward selection model, 28 set of predictors have minimum '*cp*' and for backward selection model, 34 set of predictors has the least value of '*cp*'.
- There is difference in the results of backward and forward selection subsets. We will go for further analysis doing Lasso and Ridge, and try to find the more accurate model.
- Now, we used the *glmnet* package. We used ridge regression, where we add a penalty term to the residual sum of squares and try to reduce the combination. We remove the variables that does not add much in reducing the value of the combination.
- Finally, we did the Lasso regression, this method zero out the co-efficient that seem not useful in the model. The plot of the lasso model is shown in graph 2. And we used the cross validation to get us the coefficients of best model and the test MSE.
- The model that we predicted has the *adjusted R square* around 0.2 (starting it was 0.02). This means that the dataset isn't enough for building a model.

GRAPHS:



(Graph 1)



(Graph 2)

SUMMARY:

- In this assignment, we want to predict a best model for the given dataset. Here we want predict '*shares*' attribute from the other predictors.
- I have seen that; linear dependencies can make it hard for algorithms like forward and backward selection. So, we should remove those variables before applying subset selection. And the exhaustive search is not a good idea for dataset with many variables.
- Lasso and Ridge regression gave the model better than the previous applied techniques. The MSE found out by cross validation was mostly equal in both cases. But it was large which tells us that the dataset is not enough or is not appropriate to analyse the news popularity.