MAP 4191- Mathematical Modeling of Data

December 13,2021

PREDICTIVE MODEL FOR ONLINE NEWS POPULARITY

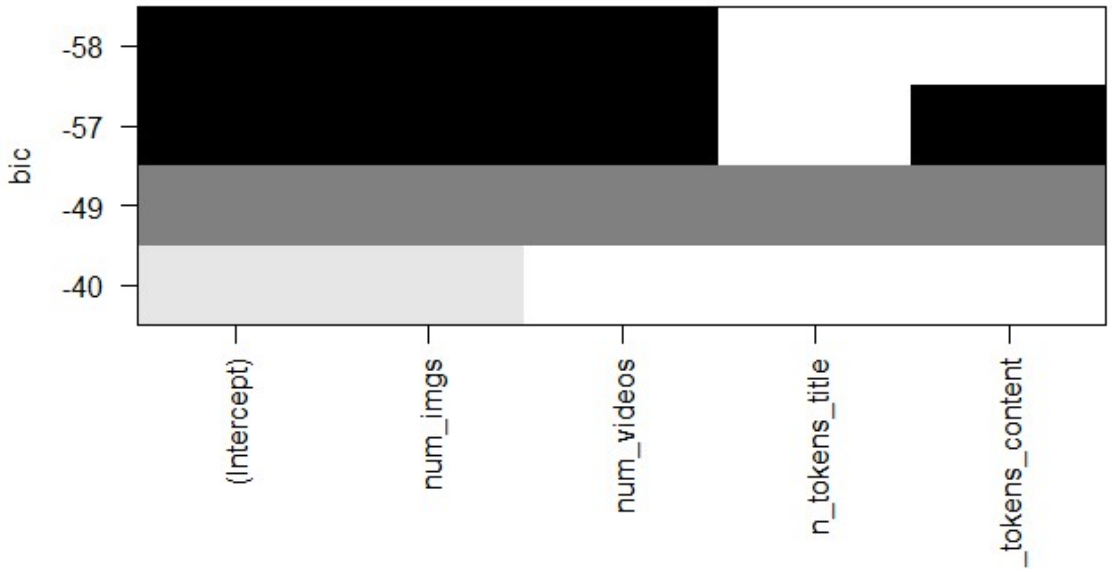
The Online News Popularity dataset was analyzed . It contains 61 characteristics of 39797 articles shared on Mashable.com .Some of the characteristics measured were the number of images , number of videos, title and article length and the number of shares each article received. This data set had 61 variables and 39797 samples. Processing the entire dataset led to memory issues and severe lagging in the code so the data set was reduced and the number of variables was decreased to 5. The number of images , number of videos, title and article length and the number of shares each article received were chosen since those values were easily measurable by the average person . The purpose of the predictive model was to predict the number of article shares using the article characteristics chosen.

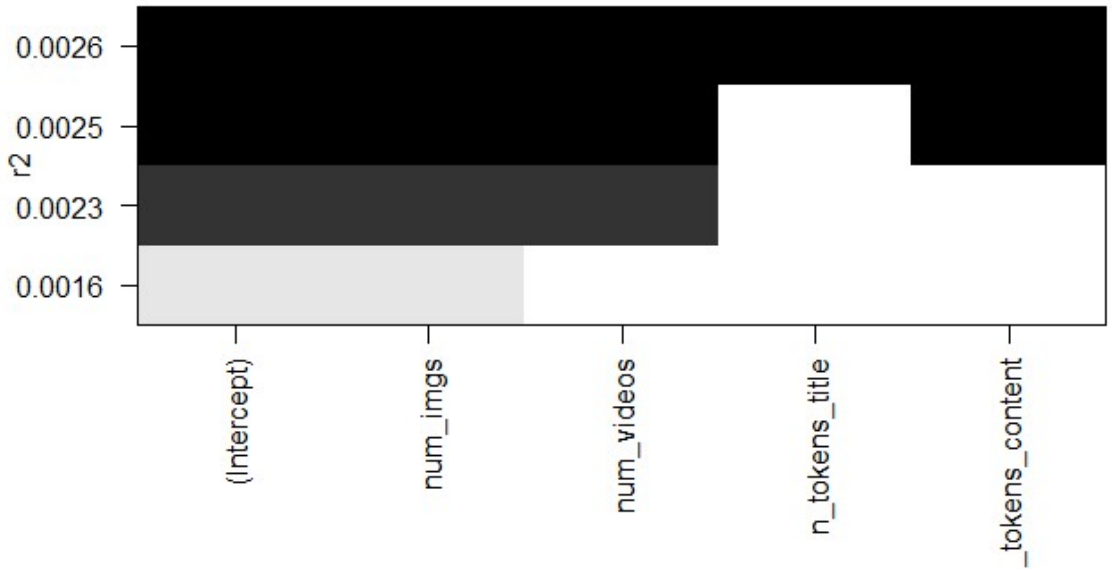
The distribution of these variables were plotted ,summary tables were made and a correlation matrix was also created. This matrix showed that all of the variables had a low correlation and similar correlation with shares so the four variables chosen were not worse predictors than the other variables. To test the best prediction model , best subset selection and forward and backward stepwise selection was implemented. 10-fold Cross Validation was also performed on each selection technique. Prior to this, cross validation was also performed on models that contained only one variable and a model with all variables .

To choose the best models, linear regression was performed on each variable individually then on a model that included all four variables. From this analysis, it was predicted that the best model was a four variable model since this model had the lowest MSE. However, the testing was

not sufficient to conclude that the four variable model was the best choice. Forward and backward selection were also performed to choose the best prediction model for shares . Forward and Backward selection also determined that the four variable model was the best model to predict shares. The R^2 ,adjusted R^2, Cp ,BIC and RSS values were the best for the four variable model. The four variable model chosen was :Shares= 2761.5554047 + 65.9358207 num_imgs + 80.4838476 num_videos +   44.2298988 n_tokens_title + -0.4141284 n_tokens_content.

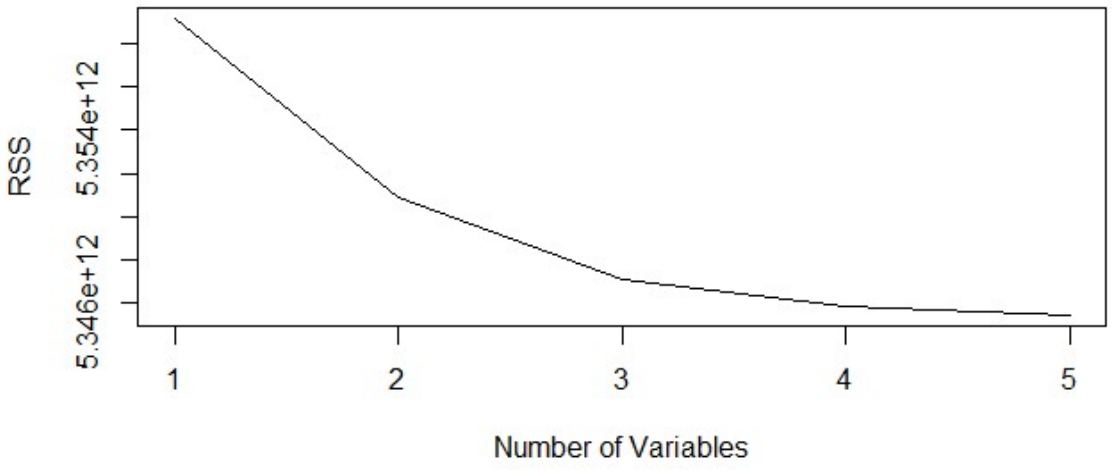
BACKWARD


bic
-58
-57
-49
-40
(Intercept)
num_imgs
num_videos
n_tokens_title
_tokens_content

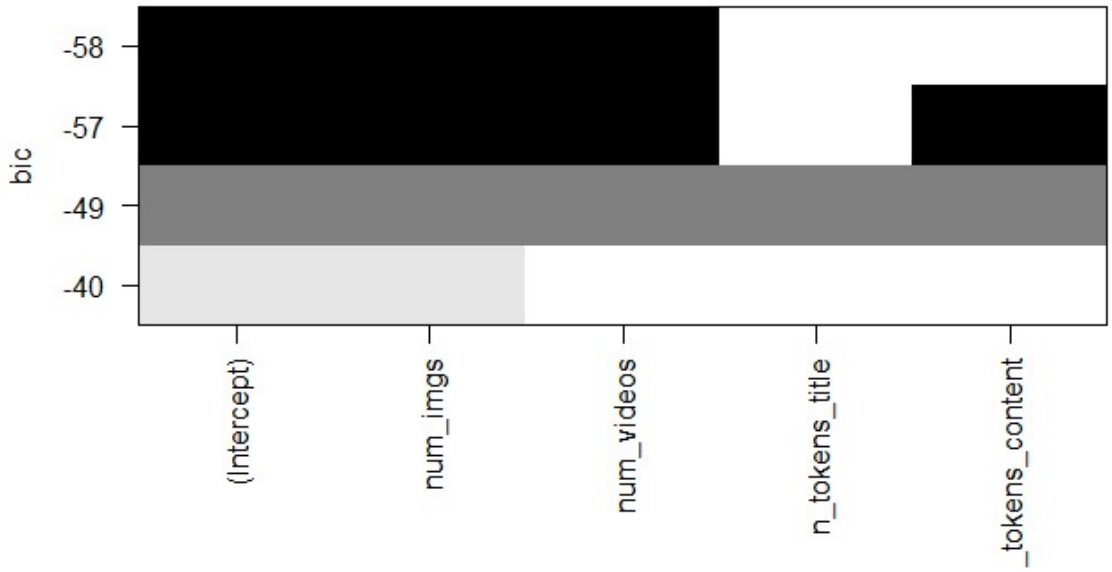

r2
0.0026
0.0025
0.0023
0.0016
(Intercept)
num_imgs
num_videos
n_tokens_title
_tokens_content

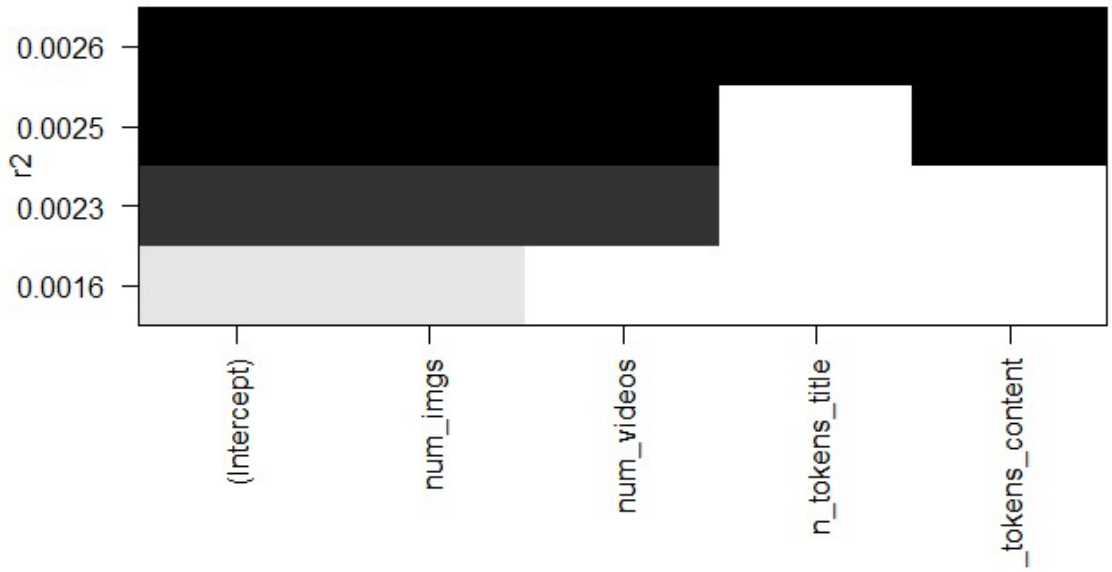
adjr2
0.0025
0.0024
0.0022
0.0015
(Intercept)
num_imgs
num_videos
n_tokens_title
_tokens_content

Cp
5
5.6
13
39
(Intercept)
num_imgs
num_videos
n_tokens_title
_tokens_content


RSS
5.346e+12
5.354e+12
1
2
3
4
5
Number of Variables

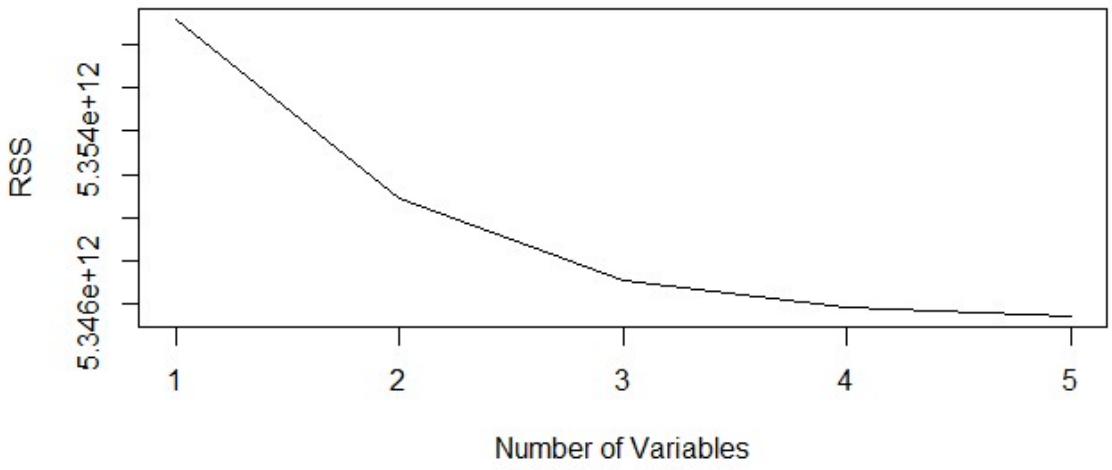
FORWARD


bic
-58
-57
-49
-40
(Intercept)
num_imgs
num_videos
n_tokens_title
_tokens_content

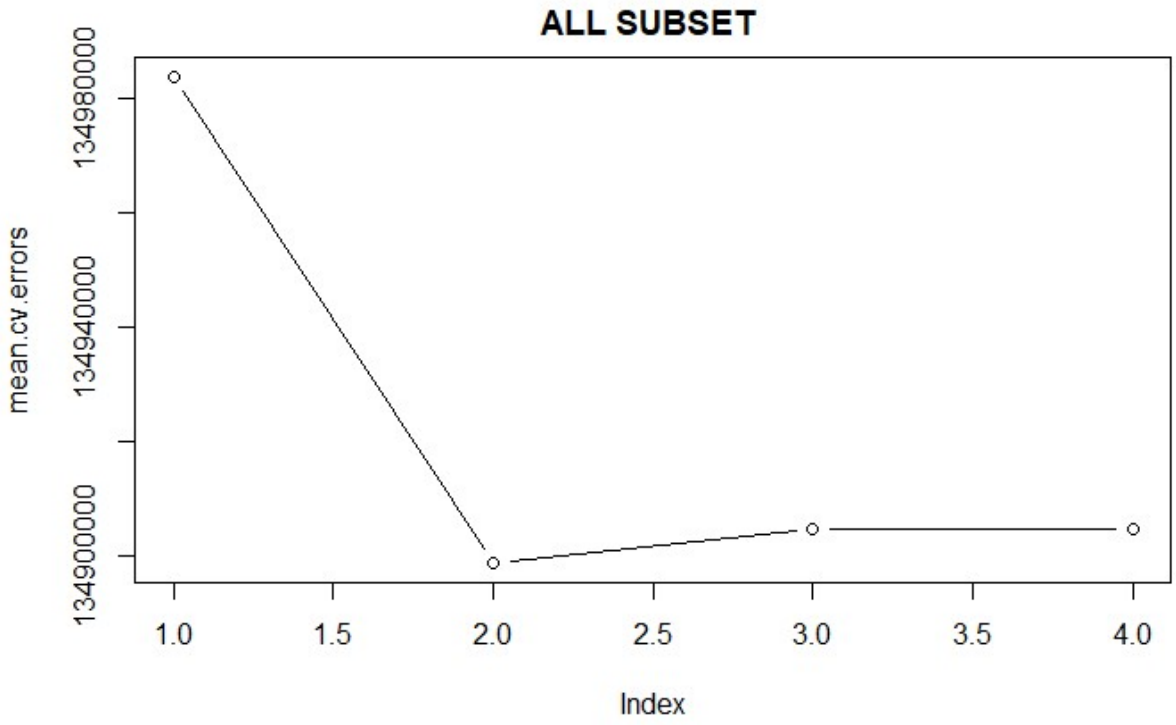

r2
0.0026
0.0025
0.0023
0.0016
(Intercept)
num_imgs
num_videos
n_tokens_title
_tokens_content

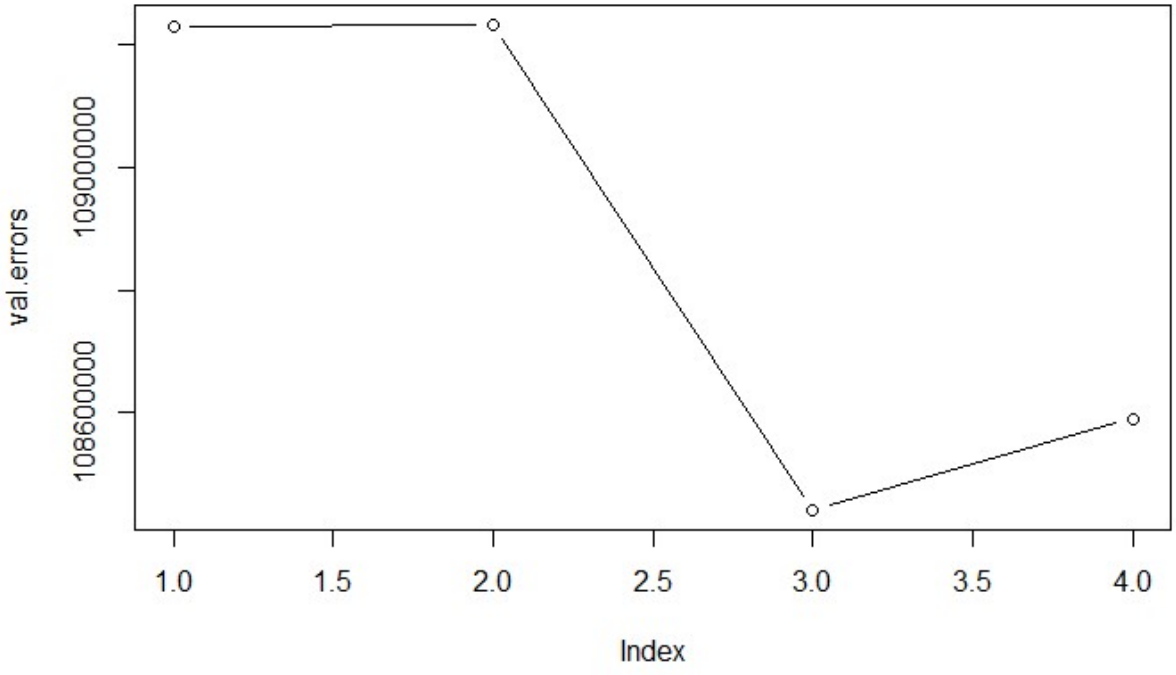
adjr2
0.0025
0.0024
0.0022
0.0015
(Intercept)
num_imgs
num_videos
n_tokens_title
_tokens_content

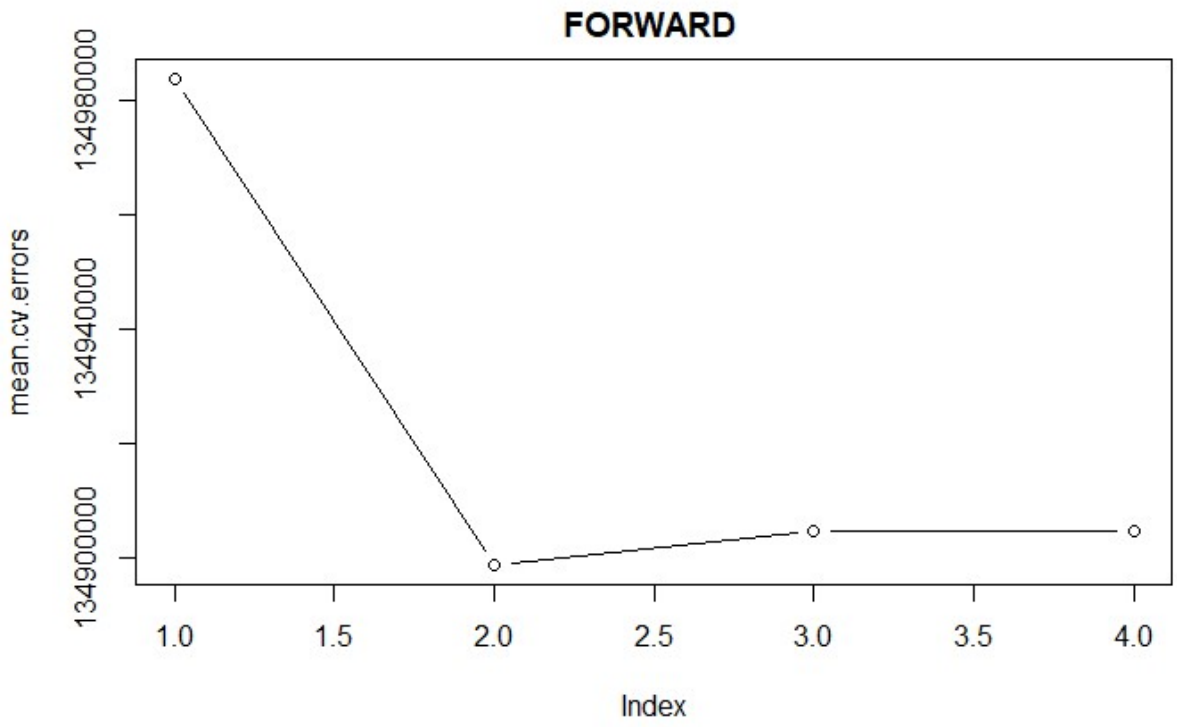
Cp
5
5.6
13
39
(Intercept)
num_imgs
num_videos
n_tokens_title
_tokens_content

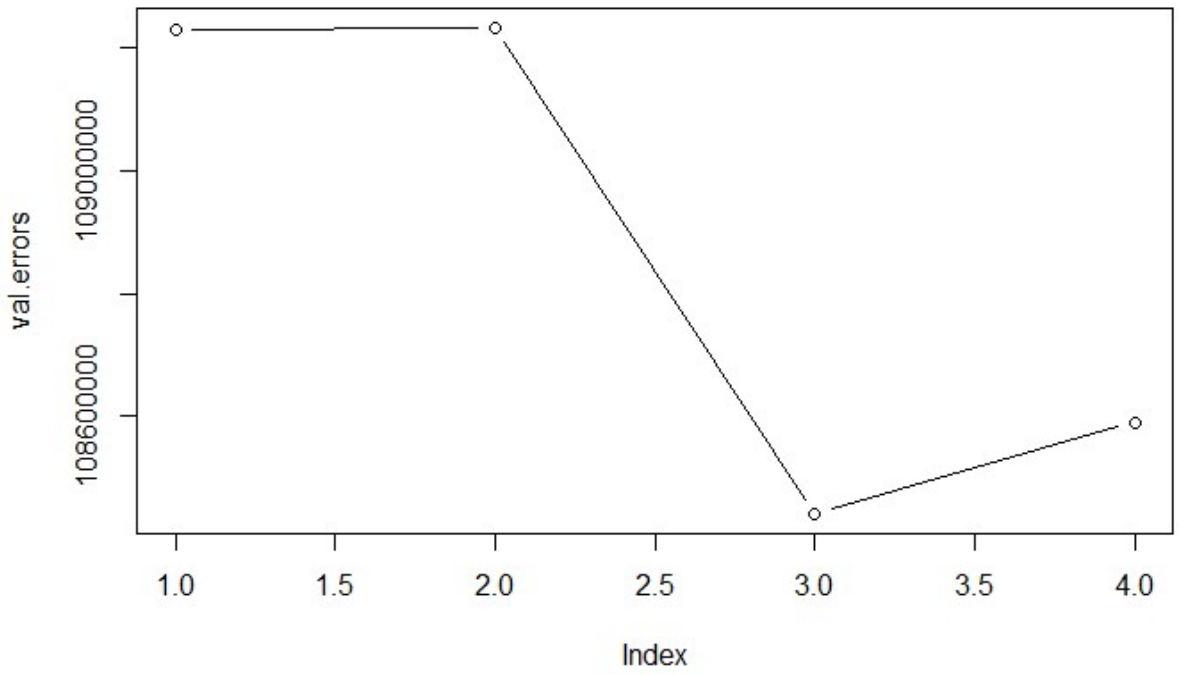


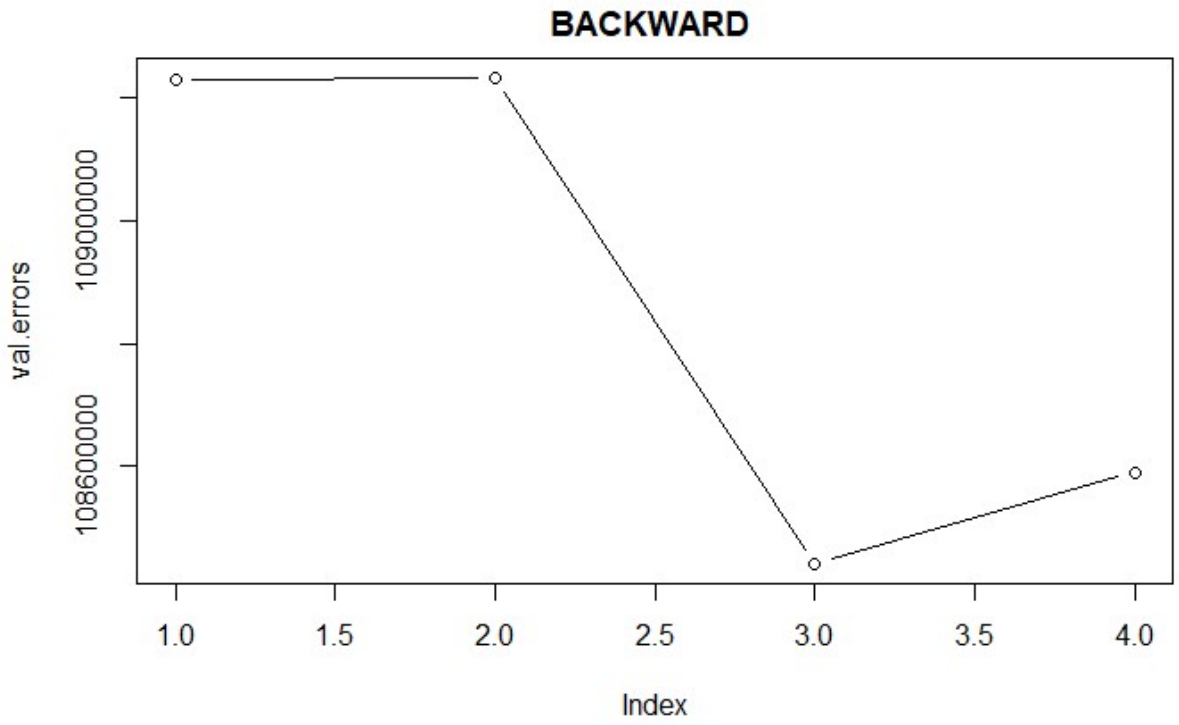
Best subset selection and forward and backward selection were also performed with 10-fold cross-validation . The selection done without cross-validation determined that the four model was the best model . However, when cross-validation was performed the three and two variable model were also deemed as strong models. The three variable model had the lowest MSE and the two variable model had the lowest cross validation error . This model had the lowest cross-validation error in forward, best subset and backward selection :Shares= 3039.01447 + 57.62953 num_imgs + 75.59873 num_videos . This model had the lowest MSE in forward, best subset and backward selection: Shares= 3268.898882 + 65.517764 num_imgs + 79.859560num_videos -0.432109n_tokens_content.

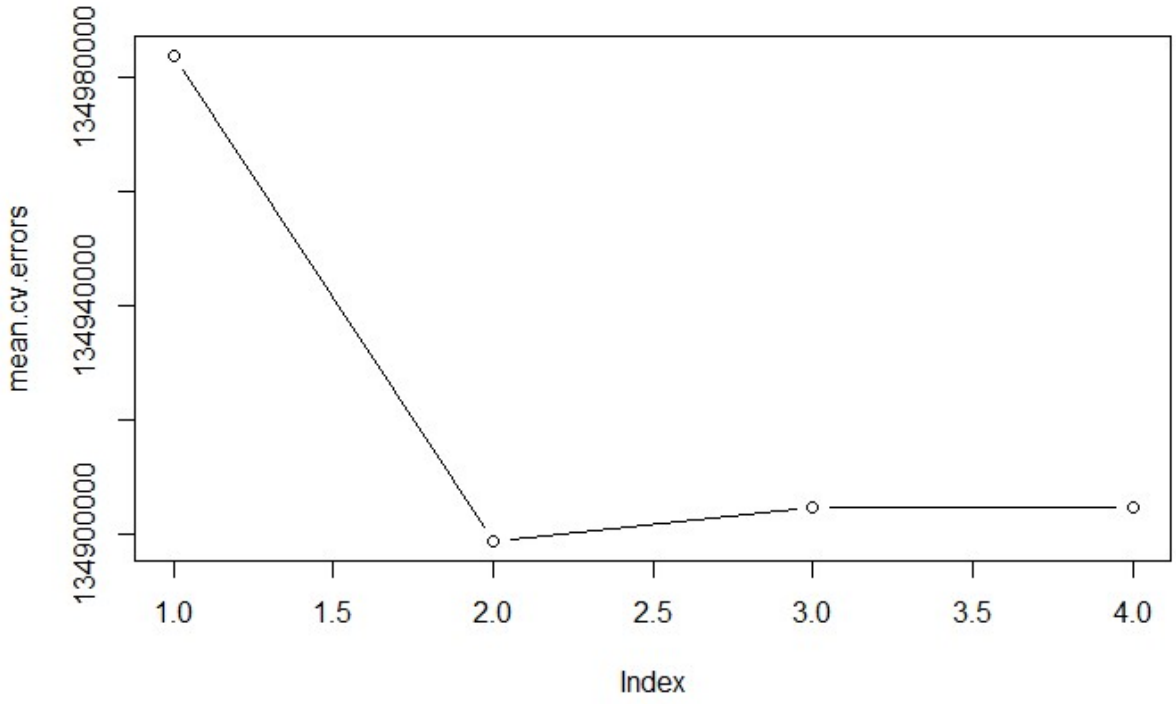

ALL SUBSET
mean.cv.errors
134900000
134940000
134980000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Index


val.errors
108600000
109000000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Index

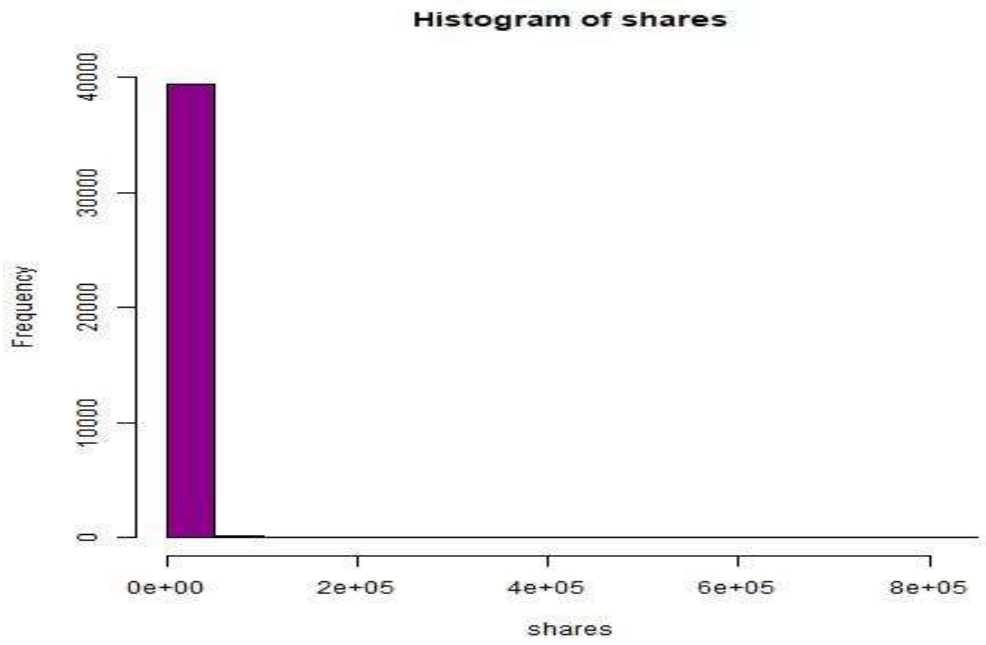

FORWARD
mean.cv.errors
134900000
134940000
134980000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Index


val.errors
109000000
108600000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Index


BACKWARD
val.errors
108600000
109000000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Index


mean.cv.errors
134900000
134940000
134980000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Index

The prediction models created can be a good foundation for determining the characteristics that affect number of article shares. However, a more accurate prediction model for shares would include more variables . Only five of 61 variables were examined . Other variables may have been better predictors for number of shares. However, due to consistent error with processing the large data in the csv file all the variables could not be tested and only four variables were chosen to be tested as possible predictors for shares. In future research, more variables would be tested and compared. Another issue that may have impacted the prediction model was the distribution of shares . The shares distribution was positively skewed. To change the shares distribution to a normal distribution ,the number of shares were logged. However, when using log shares, the MSE value for a model that used all the variables as a predictor was -0.005016253. This seemed incorrect.



It can be concluded that a one variable model is not a suitable prediction model for shares since the one variable models consistently had higher errors than the models with more variables. However, 10-fold cross-validation, cross validation, forward, best subset and backward stepwise selection sometimes chose different variable models as the best predictor. The three models that were favored were a two variable model, three variable model and a four variable model. This model had the lowest cross-validation error in forward, best subset and backward selection :Shares= 3039.01447 + 57.62953 num_imgs + 75.59873num_videos . This model had the lowest MSE in forward, best subset and backward selection: Shares= 3268.898882 + 65.517764 num_imgs + 79.859560num_videos -0.432109n_tokens_content.The R^2 ,adjusted R^2, Cp ,BIC and RSS values were the best for the four variable model,Shares= 2761.5554047 + 65.9358207 num_imgs + 80.4838476 num_videos + 44.2298988 n_tokens_title -0.4141284 n_tokens_content.

I HAVE ATTACHED MY R CODE,MY R CODE THAT CONSIDERS LOG SHARES AND MY KNIT FILE FOR LINE 1 TO 390 OF THE CODE WITH THIS SUBMISSION.