Factor Analysis of Article Shares

ISE/IDA 5103 Intelligent Data Analytics

Final Project Report

Group 10

John Osborn

Lince Rumainum

University of Oklahoma, Data Science and Analytics, Norman

Executive Summary

The news industry has turned into a popularity contest. Long gone are the days of paper routes and buying a newspaper at the coffee shop. The majority of news is now consumed online. With a new way to consume news, a new way to analyze how an article performs must be developed. Shares, or rather the number of shares, has emerged as a clear indicator of an article's success. Since traffic into news website generate profit (i.e.: through ads), the more successful an article does will bring in more money.

The goal of this paper is not only to be able to have an accurate prediction of an articles success but also to give controllable characteristics of the articles production. These will give writers and editors the power to have a head start when in comes to article success. Luckily, because all the news is produced online, we have an abundant amount of data that helps predict the number of shares.

Our data, provided by the University of California Irvine (UCI), was all collected from a single website, Mashable.com. We are assuming the data was collected correctly and not tampered with. The dataset contains variables such as, what day the article was posted, number of words, and positivity of the article. In total, there are 58 predictive variables with the target variable, shares.

With so many variables, dimensionality was a concern.Through a combination of feature selection and feature creation, we were able to reduce the number of variables to 24. Another concern was the existence and the amount of outliers present in the data. The range of shares is between 2 and 900,000. We removed the top 0.15% of the data because of the leverage it would have in the model. The other outliers weer kept because we believe they represent the goal - to garner the most shares. That large of a spread can often indicate skewness as well as outliers in the data. To normalize the data further, we performed a log transform, including the target variable. We created several predictive models using linear and non-linear models. For the model we split the original data 70:30.

The best model was a k-Nearest Neighbor (kNN). This model produced a RMSE of 0.8063. This model shows that the two most important variables are the subjectivity and the number of images in the article. Therefore, we recommend writing your opinions and including more pictures in your article.

Problem Description

Background

The news industry has turned into a popularity contest. Since before the turn of the century, the sales of physical newspapers are on the decline. At the peak of sales, 63 million people bought newspapers a year in 1970 – 1990 (Pew Research Center's Journalism Project 2019). News readers have turned to the internet and digital new sources for their daily news. Digital news sources provide fast, up-to-date news, and accessible at the tip of the readers fingers. News companies adapted their production methods and now the same stories a reader can find in the newspaper are available online. Publishing online offers new and improved ways to analyze the articles. Actions, such as, when and where a reader clicked on your article is now available. Companies can now truly understand the reach certain article have. Traditional news source does not offer any industry changing like this. However, because of its newness, the news industry had to develop a way to comprehend and implement all the data that they are getting.

Shares quickly became a favorite way to judge the success of an article. Contains with the number of shares was the reach of the article as well as the popularity of the article. It was apparent that the more popular an article become the more valuable and successful it is. Since traffic into news website generate profit (i.e.: through ads), the more successful an article does will bring in more money. Thus, the number of shares equal increased profitability.

Individuals also realize the huge impact of the number of shares can have. There has been an increase in independent articles and website created. Scott DeLong started an independent blog that is now worth 100 million dollars. Scott would repost article he took from other websites and use Facebook as a way to proliferate the content. While it is not the most ethical practice, it does demonstrate the power of shares. Scott was using the idea of virality to grow his business (Miguel, 2019). Viral articles are articles that get outrageous number of shares, views, comments or clicks. The more exposure an article receives the more revenue it will produce for the

company. Thousands of articles have been written on how to go viral. Why? Because going viral can bring you fame and fortune.

News companies now driven by profits, but with newspapers sales on the decline, leaders in the industry are looking for new ways to turn a profit. The question was simple: How can as article generate more shares? The solution was not that simple. Many factors are involved with the writing and publication of these articles. This reports effort is to provide a model as a way to predict shares and well and help writers understand what factors involved in an article will generate more shares.

Problem Definition

The goal of this report is to produce a model that will be able to predict the number of shares articles will elicit. The model produced in turn will be used to increase profits. Knowing the predicted number of shares will be helpful in some scenarios, but for forward thinking writers and editors, it does them little to no good. The second goal of the report is to describe factor importance. Factor importance can then be used to improve the writing and production of the articles. Thus, this project will include not only a model to predict article success but what makes articles successful.

Data Exploration

Our data, provided by University of California Irvine (UCI), was collected all from a single website, Mashable.com. Taking data from a single source will reduce this skewness and make the data applicable to all new source websites. Popular websites would generate more shares and thus skewing the dataset. The dataset contains 39,644 observation and 61 variables. The 61 variables contain the target, two non-predictive factors and 58 predictive factors (**Appendix A – Figure 7**). The majority of the variables are meta data from the publishing the article: day, number of words, number of images, etc. The dataset also includes factors of the writing of the article. For example, using the pattern web mining module, the creators of the dataset were able to compute the subjectivity and the polarity of the articles (Fernandes, Pedro, Cortez 2015). The accuracy of the data is assumed to be very good. All the data was collected

within the Mashable website and through trusted algorithms. Due to the data being collected from meta-data, there is no missingness in the data and will not need to be addressed. Due to the large dimensionality, factor selection will be important to model success. Outliers are a concern with in the dataset.

```
> md.pattern(dat)
 /\     /\
{  `---'  }
{  o   o  }
==>  V <==  No need for mice. This data set is completely observed.
 \  \|/  /
  `-----'
```

Figure 1 Result from Missing Data Analysis using MICE package in R

Outliers

In the dataset there is an abundance of outliers, especially within the shares variable. The range of the shares variable was 2 – 900,000. Based on the boxplot, we were able to cover that almost 25% of the up-reaching data are considered outliers. The upper bound of the data turned out to be 2900 shares.

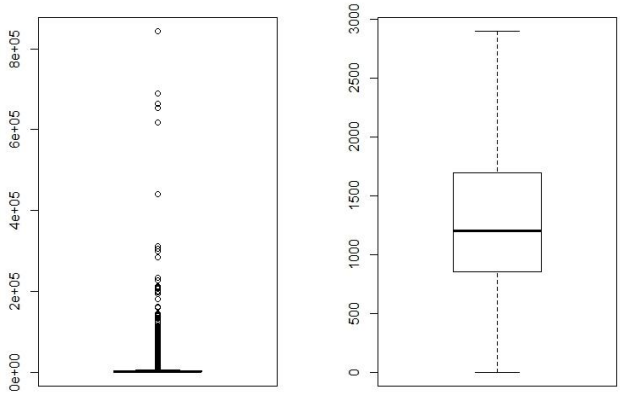

Figure 2 The Box Plot with All Data Present (Left) and The Box Plot with the "shares" > 2900 Removed (Right)

Additionally, these points we determined to be outliers by the Rosner Outliers test. We determined that some outliers might be important to the dataset. The goal of the model is to predict what produces an article with a lot of shares. Simply removing the outliers would cause us to lose out of important data. Thus, we determined to keep the outliers, at least partially. In further investigation, we determined that some articles, viral articles, will be shared regardless of the data they produce. Viral articles will have a

greater pull on the data than desired. Additionally, only 0.15% of the shares value was above 100,000. Such a small amount of data controlling the outcome of the predicting model would cause inconsistencies and produce a poor model. Thus, we decided that to remove shares above 100,000 within out data.

Feature Selection

To reduce the 58 dimensions of the dataset, we needed to evaluate the scope of our efforts. Our scope is the article itself. We wanted to investigate the data of the article and remove any external data that might be present. Additionally, we are going to produce the tangible features that can be used to improve article success. We were able to remove all the keyword variables (**Appendix A – Figure 7**) because they were determined to be external to the article. The keyword variables were all the keyword presented in the separate articles that were linked in the actual article. They are not the keywords with the article, but keywords in separate article, thus are external to our scope. Secondly, the LDA variables are described as the subject of the article. While this is a part of the article, there variables would violate the second part of our scope. There is no description provided of the LDA variables. The topics of the LDA variables could be anything. If the articles of LDA 1 produce the most shares, we would be unable to say what topic that was. Additionally, the topics of the articles are already encapsulated with the data_channel, and thus would be duplication of the data. Therefore, LDA variables have little meaning to our scope and were removed.

We continued our efforts to reduce the features of the dataset by combining the day the article was published. Originally, each day was a separate binary variable. Because the days are mutually exclusive, we were able to combine the days into a single factor column using the unite function available in caret.

```
57
58  dat <- unite(dat, "day", weekday_is_monday:weekday_is_sunday, remove = TRUE)
59
```

Figure 3 Unite Function for Creating New Feature from Day Related Variables

We also combined the data_channel variable since it was binary as well and mutually exclusive.

Data Transformation

Now that the attributes has been selected, we need to look at each of them to see if the data is normally distributed. For each attribute that is not normally distributed, data transformation is necessary to create a reliable model. There are ways to transform skewness in the data, such as, scaling them into z-score or box cox transformation. There are two functions to create a normalized data with box cox:

$$x = \frac{x^{\lambda} - 1}{\lambda} \text{ if } \lambda \neq 0 \quad or \quad x = log(x) \text{ if } \lambda = 0$$

Transforming the data as z-score might seem to be a good idea since it does normalized the data and although it does, the normal distribution between the training and testing data might be a lot different since it depends on the mean and standard deviation of those data. So, using box cox transformation is a better option. To choose whether to find an optimum lambda or not, it is necessary to keep in mind that the same exact transformation on the training data must also be done in the testing data. So, although optimized lambda value gives the most normalized data, finding an optimized lambda might not result in the same lambda value in training and testing data. Therefore, it best to transform the skewed attributes by taking the log function of the data, where lambda is zero. The data distributions of each selected attribute in our model before data transformation are shown in **Appendix A – Figure 8** and **Appendix A – Figure 9**.

Model Selection

To predict the number of shares five models were chosen: linear model, k-Nearest Neighbors (kNN), Support Vector Machines, Multivariate Adaptive Regression Splines (MARS), and bagged Classification and Regression Tree (CART). We chose to not repeat any models because each model will impact the data differently. Using all tree models might neglect certain factors or be strongly impacted by a factor that skews the results. Linear modeling was chosen because of its simplicity. A linear model would provide a simple solution that would allow us to use it as a benchmark of model success. All the model are well-known regression models. Specifically, we chose SVM

because of the higher dimensionality of the dataset and to combat overfitting of the training data. Due to the existence of our outliers we needed a way to challenge the noise they present. K-nearest neighbors is that model. Some of our models require a form of re-sampling. Each model that required re-sampling using a K-fold cross validation to ensure that the most accurate and consistent models. There were five folds and it was repeated three times.

Our models were built with 70 percent of the original data, which leave 30 percent of the data for testing. Our training data had 27,710 observations and our test data contained 11,876 observations. We remove the shares in the test data to be predicted later and calculated the Root Mean Square Error of the result from that model. We are going to predict the shares using linear model and non-linear model.

Our models will be validated primarily using root mean square error (RMSE).The RMSE will be computed with the predicted number of shares and the actual number of shares an article receives. RMSE will be a good indicator of success because due to the inclusion of our outlier we want to reduce the effect that their magnitude may include. Additionally, Q-Q Normal plot will be used to see if the factors we chose are helpful in predicting the share values. Once we validate the accuracy of our model we will then use the model to find the variance importance.

Results

The first result is of the linear model. The linear model is used because of its simplicity and because it is a good benchmark model to compare with the non-linear models. From the twenty-six variables that were selected, two variables, which are avg_negative_polarity and title_sentiment_polarity, were removed due to the error it created while running the linear model. Since the new created features are non-numeric data, they also had to be removed from the linear model. For this model, the RMSE result for the linear model is 0.8547. The plots that shows the behavior of the linear model is in **Appendix A – Figure 10**. Since we have our benchmark model, we created several non-linear models. Those models are Multivariate Adaptive Regression Spline (MARS), k-Nearest Neighbor (kNN), Support Vector Machine (SVM), Classification and

Regression Tree (CART), and Neural Net. The model behavior for MARS model is shown in **Appendix A**. All of the RMSE results of these models is shown in table below:

Model	RMSE
kNN	0.8063
MARS	0.8296
SVM	0.8520
OLS	0.8547
treeBag	0.8656
avNNet	6.4702

Figure 4 RMSE Comparison Between Different Fit Models

Since kNN model creates the best model out all of the other models, it is useful to see the variable importance according to that model to be able to make sense of the data and help further with creating a better model. The variable importance rates from the most to least important according to kNN model is shown below in **Figure 5**.

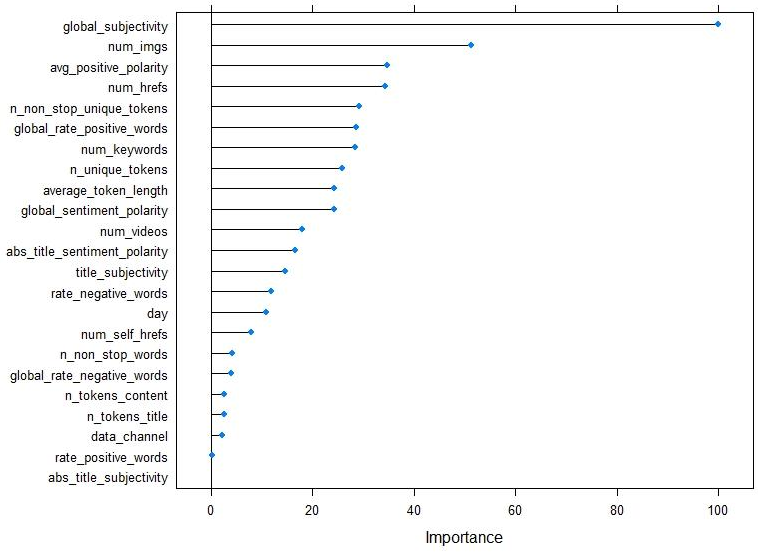

Figure 5 Variable Importance According to k-Nearest Neighbor Model

Conclusion

The aim of the report was to understand the controlling factors of shares and produce a model to predict the number of shares and article will generate. In the end we were able to produce a model with okay predicting power. A lower RMSE would be desired to have more accurate models. Higher RMSE would likely indicate that there is data with high leverage. This is confirmed since k-Nearest Neighbor is the best performing model, which is supposed to reduce the amount to leverage in the data. We believe that the feature selection is good for the scope of our report, but it is the outliers to have caused the issues in the modeling. While we believed that outliers were important to analysis, that could have had more leverage than we intended.

One way to see how well the model behave is to look at the Normal Q-Q plot of each model. The goal is to create a well normal distributed data, which is indicated by a positive linear trend line. The Normal Q-Q plots for each model is shown below in **Figure 6**.

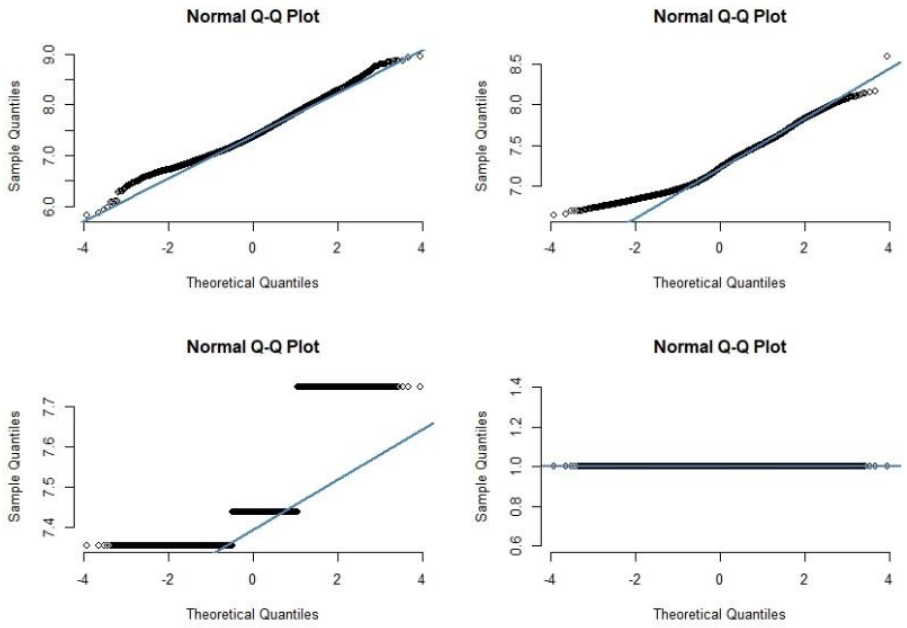

Figure 6 Normal Q-Q Plot of kNN (top-left), SVM (top-right), CART (bottom-left), and Neural Net (bottom-right)

As we know from the RMSE results, kNN model (**Figure 6**) is the best model, which is the one with the closest to normal distribution than any other models. The MARS (**Appendix A – Figure 11**) is the next best model, SVM model (**Figure 6**) is the third best, the linear model (**Appendix A – Figure 10**) is ranked fourth, CART (**Figure 6**) follows, while Neural Net model (**Figure 6**) shows that it is not a good model to use for this data set. If compared, the closer it is to normal distribution, the better it is in predicting the model.

The true strength of the model was our ability to provide factors that will help the writer generate more shares. Increasing the amount of media included in your article is crucial to increasing shares, as both the number of images and number of videos have factor importance. Writers who are opinionated and positive also garner more shares. Interestingly enough the length of the article has little importance. These factor trends all make sense. Reader like a positive voice in the article because happier stories increase their happiness. Increasing the media provided in the article also make sense because it allows the reader to visualize the story the article is trying to tell.

Future Investigation

Since the contents of the story have been looked at, the next set would be to evaluate how articles are treated on social media. Social media is increasingly becoming a place for news and news promotion.

Another model could be instead created as a classification model, and not a regression model. The actual number of shares may in fact be irrelevant. One additional share for an article, may be statistically insignificant. Instead of using the real values, collect the shares into groups. Since the goal is not to reach a specific number of shares. Reaching a range of shares might be a better indicator of success because of the lack of separation between the actual values.

References

Fernandes, Kelwin, Pedro Vinagre and Paulo Cortez. “A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News.” *EPIA* (2015).

Miguel, Claudio. “ViralNova.com Case Study: How One Man Built a $100M Viral Content Website.” Outbound.net, March 15, 2019. https://outbound.net/how-viralnova-com-became-a-100m-viral-content-website/.

“Trends and Facts on Newspapers: State of the News Media.” Pew Research Center's Journalism Project, July 9, 2019. https://www.journalism.org/fact-sheet/newspapers/.

APPENDIX A

Variables	
0. url: URL of the article (non-predictive)	31. weekday_is_monday: Was the article published on a Monday?
1. timedelta: Days between the article publication and the dataset acquisition (non-predictive)	32. weekday_is_tuesday: Was the article published on a Tuesday?
2. n_tokens_title: Number of words in the title	33. weekday_is_wednesday: Was the article published on a Wednesday?
3. n_tokens_content: Number of words in the content	34. weekday_is_thursday: Was the article published on a Thursday?
4. n_unique_tokens: Rate of unique words in the content	35. weekday_is_friday: Was the article published on a Friday?
5. n_non_stop_words: Rate of non-stop words in the content	36. weekday_is_saturday: Was the article published on a Saturday?
6. n_non_stop_unique_tokens: Rate of unique non-stop words in the content	37. weekday_is_sunday: Was the article published on a Sunday?
7. num_hrefs: Number of links	38. is_weekend: Was the article published on the weekend?
8. num_self_hrefs: Number of links to other articles published by Mashable	39. LDA_00: Closeness to LDA topic 0
9. num_imgs: Number of images	40. LDA_01: Closeness to LDA topic 1
10. num_videos: Number of videos	41. LDA_02: Closeness to LDA topic 2
11. average_token_length: Average length of the words in the content	42. LDA_03: Closeness to LDA topic 3
12. num_keywords: Number of keywords in the metadata	43. LDA_04: Closeness to LDA topic 4
13. data_channel_is_lifestyle: Is data channel 'Lifestyle'?	44. global_subjectivity: Text subjectivity
14. data_channel_is_entertainment: Is data channel 'Entertainment'?	45. global_sentiment_polarity: Text sentiment polarity
15. data_channel_is_bus: Is data channel 'Business'?	46. global_rate_positive_words: Rate of positive words in the content
16. data_channel_is_socmed: Is data channel 'Social Media'?	47. global_rate_negative_words: Rate of negative words in the content
17. data_channel_is_tech: Is data channel 'Tech'?	48. rate_positive_words: Rate of positive words among non-neutral tokens
18. data_channel_is_world: Is data channel 'World'?	49. rate_negative_words: Rate of negative words among non-neutral tokens
19. kw_min_min: Worst keyword (min. shares)	50. avg_positive_polarity: Avg. polarity of positive words
20. kw_max_min: Worst keyword (max. shares)	51. min_positive_polarity: Min. polarity of positive words
21. kw_avg_min: Worst keyword (avg. shares)	52. max_positive_polarity: Max. polarity of positive words
22. kw_min_max: Best keyword (min. shares)	53. avg_negative_polarity: Avg. polarity of negative words
23. kw_max_max: Best keyword (max. shares)	54. min_negative_polarity: Min. polarity of negative words
24. kw_avg_max: Best keyword (avg. shares)	55. max_negative_polarity: Max. polarity of negative words
25. kw_min_avg: Avg. keyword (min. shares)	56. title_subjectivity: Title subjectivity
26. kw_max_avg: Avg. keyword (max. shares)	57. title_sentiment_polarity: Title polarity
27. kw_avg_avg: Avg. keyword (avg. shares)	58. abs_title_subjectivity: Absolute subjectivity level
28. self_reference_min_shares: Min. shares of referenced articles in Mashable	59. abs_title_sentiment_polarity: Absolute polarity level
29. self_reference_max_shares: Max. shares of referenced articles in Mashable	60. shares: Number of shares (target)
30. self_reference_avg_sharess: Avg. shares of referenced articles in Mashable	

Figure 7 List of All the Variables from the Data Set

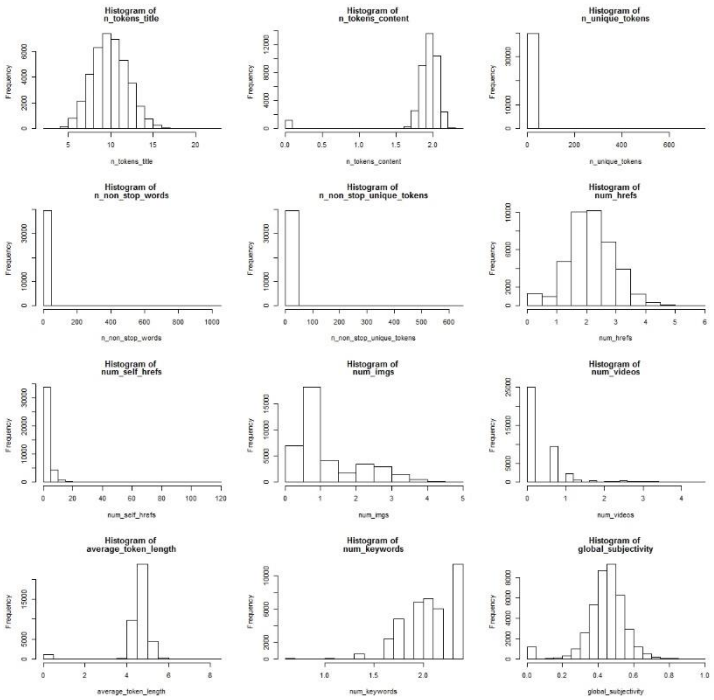

Figure 8 Histogram of the Selected Variables - Part 1

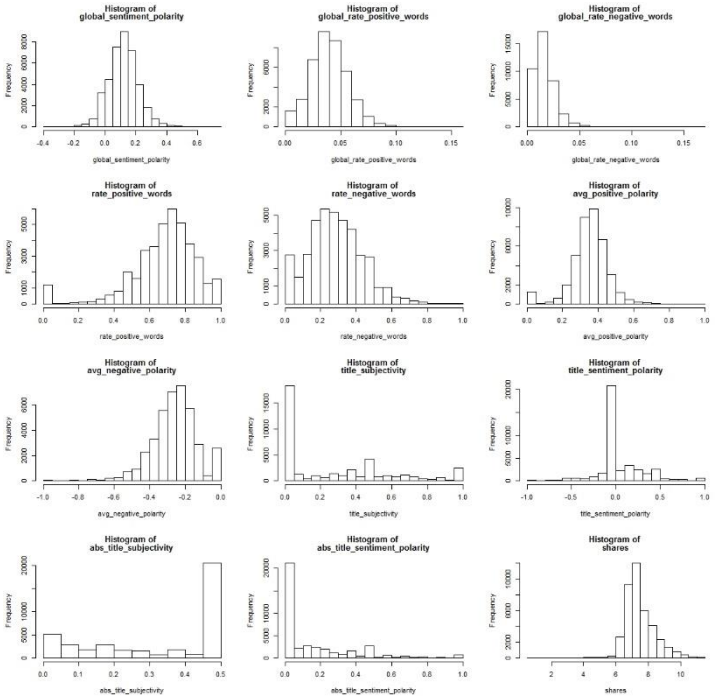

Figure 9 Histogram of the Selected Variables - Part 2

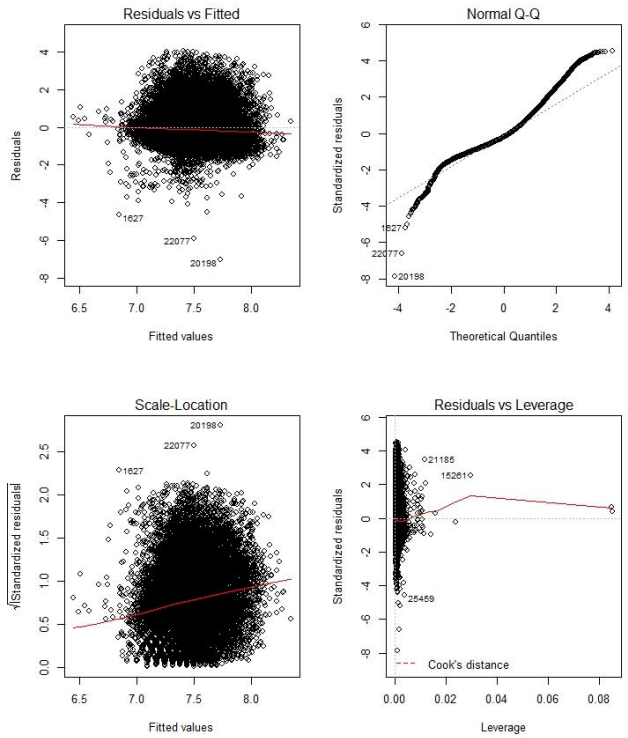

Figure 10 Linear Model Behavior with the Selected Variables Data

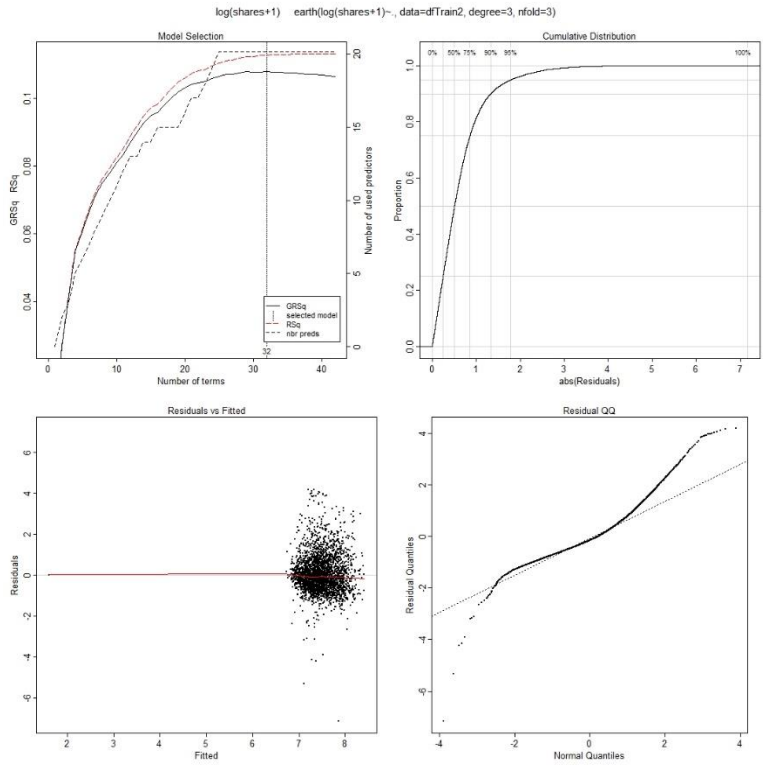

Figure 11 Plot of MARS Model with Degree Hyperparameter of 3 and 3 number of folds