MASHABLE POPULARITY NEWS ASSIGNMENT

Training Data Set: alldata_onlinenews_33

STATISTICS FOR BUSINESS ANALYTICS I MSc IN BUSINESS ANALYTICS

TABLE OF CONTENTS

1. Introduction

1.1. Description of the problem and study aim

In the present big data period in which the web is rapidly expanding, the prediction of online news popularity is becoming a more and more trendy research topic. The purpose of this study is to investigate and identify the decisive factors which result in the sharing of articles published in the Mashable website (www.mashable.com). The goal to be achieved is the creation of a prediction model for post sharing by performing multiple regression analysis methods (stepwise methods and 10-fold cross validation).

1.2. Data Characteristics

The dataset sample used for model training consisted of 3,000 observations while this of the model evaluation consisted of 10,000 ones. The primary variable of the study is the number of shares as the indicative measurement of the site's/post's popularity followed by 61 additional variables as shown below:

- shares: Number of shares (target response)
- url: URL of the article (non-explanatory)
- timedelta: Days between the article publication and the dataset acquisition (non-explanatory)
- n_tokens_title: Number of words in the title
- n_tokens_content: Number of words in the content
- n_unique_tokens: Rate of unique words in the content
- n_non_stop_words: Rate of non-stop words in the content
- n_non_stop_unique_tokens: Rate of unique non-stop words in the content
- num_hrefs: Number of links
- num_self_hrefs: Number of links to other articles published by Mashable
- num_imgs: Number of images
- num_videos: Number of videos
- average_token_length: Average length of the words in the content
- num_keywords: Number of keywords in the metadata
- data_channel_is_lifestyle: Is data channel 'Lifestyle'? (1=Yes, 0=No)
- data_channel_is_entertainment: Is data channel 'Entertainment'? (1=Yes, 0=No)
- data_channel_is_bus: Is data channel 'Business'? (1=Yes, 0=No)
- data_channel_is_socmed: Is data channel 'Social Media'? (1=Yes, 0=No)
- data_channel_is_tech: Is data channel 'Tech'? (1=Yes, 0=No)
- data_channel_is_world: Is data channel 'World'? (1=Yes, 0=No)
- kw_min_min: Worst keyword (min. shares)
- kw_max_min: Worst keyword (max. shares)
- kw_avg_min: Worst keyword (avg. shares)
- kw_min_max: Best keyword (min. shares)
- kw_max_max: Best keyword (max. shares)
- kw_avg_max: Best keyword (avg. shares)
- kw_min_avg: Avg. keyword (min. shares)
- kw_max_avg: Avg. keyword (max. shares)
- kw_avg_avg: Avg. keyword (avg. shares)
- self_reference_min_shares: Min. shares of referenced articles in Mashable
- self_reference_max_shares: Max. shares of referenced articles in Mashable
- self_reference_avg_sharess: Avg. shares of referenced articles in Mashable
- weekday_is_monday: Was the article published on a Monday? (1=Yes, 0=No)
- weekday_is_tuesday: Was the article published on a Tuesday? (1=Yes, 0=No)
- weekday_is_wednesday: Was the article published on a Wednesday? (1=Yes, 0=No)
- weekday_is_thursday: Was the article published on a Thursday? (1=Yes, 0=No)
- weekday_is_friday: Was the article published on a Friday? (1=Yes, 0=No)

- weekday_is_saturday: Was the article published on a Saturday? (1=Yes, 0=No)
- weekday_is_sunday: Was the article published on a Sunday? (1=Yes, 0=No)
- is_weekend: Was the article published on the weekend? (1=Yes, 0=No)
- LDA_00: Closeness to LDA topic 0
- LDA_01: Closeness to LDA topic 1
- LDA_02: Closeness to LDA topic 2
- LDA_03: Closeness to LDA topic 3
- LDA_04: Closeness to LDA topic 4
- global_subjectivity: Text subjectivity
- global_sentiment_polarity: Text sentiment polarity
- global_rate_positive_words: Rate of positive words in the content
- global_rate_negative_words: Rate of negative words in the content
- rate_positive_words: Rate of positive words among non-neutral tokens
- rate_negative_words: Rate of negative words among non-neutral tokens
- avg_positive_polarity: Avg. polarity of positive words
- min_positive_polarity: Min. polarity of positive words
- max_positive_polarity: Max. polarity of positive words
- avg_negative_polarity: Avg. polarity of negative words
- min_negative_polarity: Min. polarity of negative words
- max_negative_polarity: Max. polarity of negative words
- title_subjectivity: Title subjectivity
- title_sentiment_polarity: Title polarity
- abs_title_subjectivity: Absolute subjectivity level
- abs_title_sentiment_polarity: Absolute polarity level

2. Descriptive Analysis & Exploratory Data Analysis

The software tool used for conducting this study is RStudio. A number of data preprocessing steps was initially followed to clear and transform the data so as to proceed to the required analysis of the current section. More specifically:

- The URL and timedelta columns have been omitted since they are meta-data and cannot be treated as features.
- The is_*weekend* column has been omitted since it is a duplicate of the existent *is_saturday* and *is_sunday* columns.
- The first column imported via the training data set file has been omitted since it had no title and was not included in the field description of this data set.
- The records related to articles with no words included have been omitted since they have no meaning for the analysis.
- NAs, NaN and infinite values were checked but none of these existed.
- The variables weekday_is_monday, weekday_is_tuesday, weekday_is_wednesday, weekday_is_thursday, weekday_is_friday, weekday_is_sunday and weekday_is_saturday were converted to factor variables.
- The variables weekday_is_monday, weekday_is_tuesday, weekday_is_wednesday, weekday_is_thursday, weekday_is_friday, weekday_is_sunday and weekday_is_saturday were converted to factor variables.
- The variables data_channel_is_lifestyle, data_channel_is_entertainment, data_channel_is_bus data_channel_is_socmed, data_channel_is_tech and data_channel_is_world were converted to factor variables.

After the above process is completed, 58 variables in total remain in the data set, 45 numeric and 13 factors. The histograms of Tables 1,2 and 3 were used for the visualization of the most important numeric variables. By looking at these diagrams, it is observed that:

- The n_tokens_title, n_unique_tokens and n_non_stop_unique_tokens variables (See Diagram 1) seem to follow normal distribution.
- None of the rest variables indicates such high symmetry in its distribution, though.

  HEAVILY RIGHT-SKEWED VARIABLES

  - The n_tokens_content , num_hrefs, num_imgs and num_videos variables (See Diagram 1).
  - The LDA_00, LDA_01, LDA_02, LDA_03, LDA_04, title_subjectivity and shares variables (See Diagram 2).
  - The global_rate_positive_words, global_rate_negative_words and rate_negative_words variables (See Diagram 3).

  HEAVILY LEFT-SKEWED VARIABLES

  - The n_non_stop words variable (See Diagram 1) .
  - The rate_positive_words variable (See Diagram 3).
- The majority of the numeric variables have outliers such as shares and LDA_00.

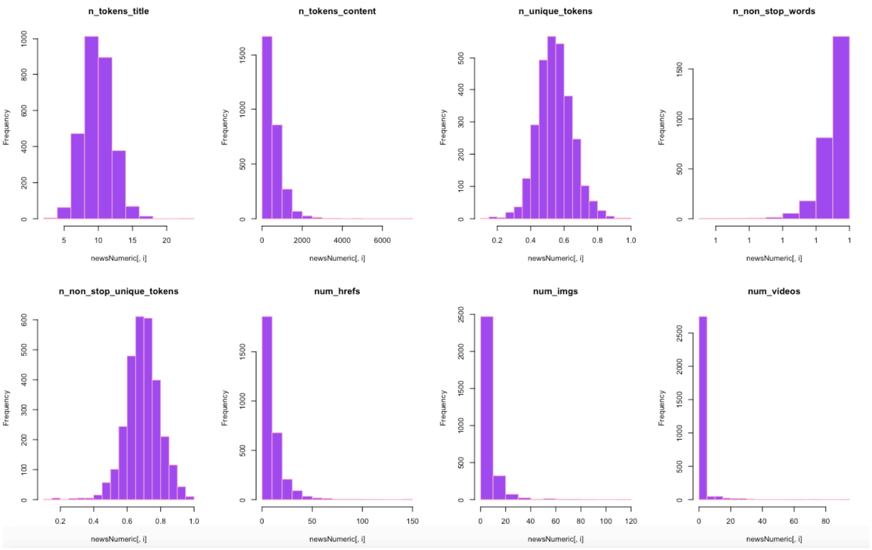


Diagram 1 – Numerical Variables Distribution (I)

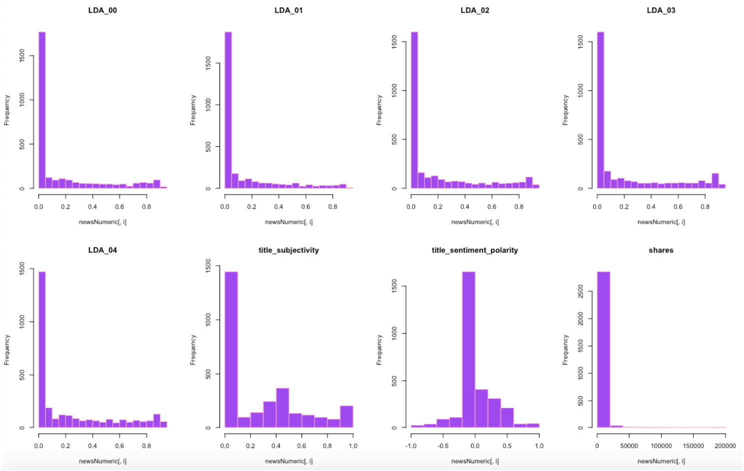


Diagram 2 – Numerical Variables Distribution (II)

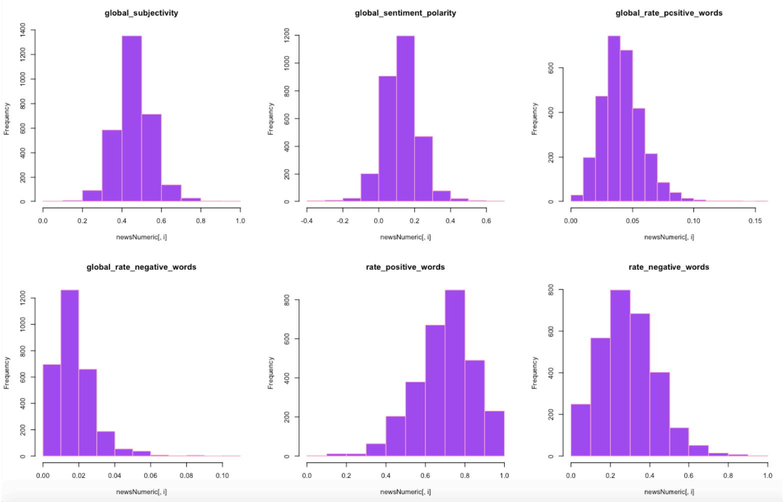
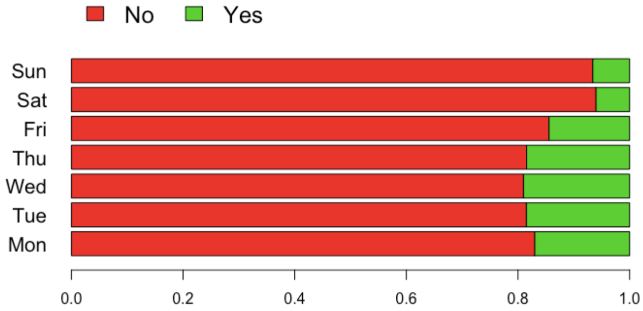


Diagram 3 – Numerical Variables Distribution (III)



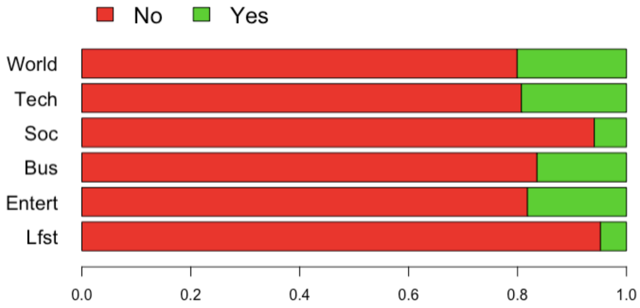


Diagram 4 – Categorical Variables (I)

Diagram 5 – Categorical Variables (II)

The above bar plots depict the factor variables which are related to the day of the published article (See Diagram 4) and the article topic (See Diagram 5). The articles published during the majority of working days (i.e. Tuesday, Wednesday and Thursday) tend to be more shared than these of the other days while the hottest subjects of most shares seem to be technology, entertainment and world-wide news.

As highlighted before, the shares variable which will be the dependent variable for the prediction model has a heavily skewed distribution. We are now going to investigate the existent outliers of shares.

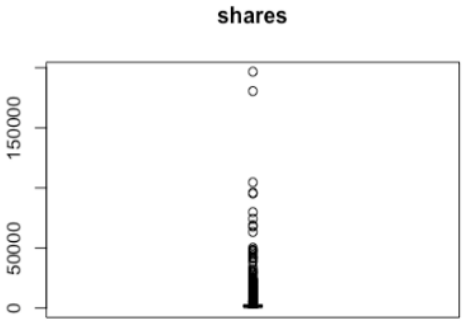


Diagram 6 – Shares Boxplot

```
## [1] "Outliers of shares"

##   [1]   13200   22000    5600   10900    9900   95300    9400    8700    5900    5500
##  [11]    5400    5600    8600   15200   48500   10200    5800    9800   16200   13000
##  [21]    8200   39000    8400    9800   15200    8400    5900   10700   24800   12500
##  [31]    5500    7300    6900    6500    6100    6100    9700    7300   14700   17600
##  [41]  12400   9700  16300   6400  21000   7200   9200  63300   9500  10400
##  [51]  13100  50200   9000   5600  23600  11500  13000  16400   9700   8200
##  [61]  20700  11700   6000 196700  30200  10500   9300  11600   6400  14300
##  [71]  11300   7500  34600   5900   8700   9800  48000  12800   8800  12900
##  [81]  17100   7300   6900  67800   5400  11900  11700   9100   7700   8400
##  [91]  12600  18600   7300  10700  22600  10000   9000   9500   7900   7000
## [101]  11700   6500   6600   5400  11400  10100  20800   6600  10600  16100
## [111]   7300   8600   6800  11800   5800  24300   5700  11400  13300  15600
## [121]  12300  14900   7500   5500   7300  17600   6600  10200  22300  11500
## [131]  29500   9700   8900   9600   9700  12300  16200   8700   7600   6200
## [141]   8200  18600  16900  18000   7400   8500  16300  20200  10300   5600
## [151]  15800  44300  13300   8500   7300  32000   6300   7700  15000   7600
## [161]  69000   6800   7500   8200   7400   6600  18400   7400  16800  21500
## [171]   5800  11000  21700   6200  10700  20800   7300   5800  13300 104600
## [181]   7900   6000  41600  18000   7900   6600   5800   5800   6800  10400
## [191]   5900  11900   5700  23600   8300   6700  10400  20700   7100  32200
## [201]   9200  10700   6900   5500  11000   5600   8300  10500  47300   7300
## [211]   5600  16700   7400  14600   5400   6600   9100   6600   9200  20300
## [221]   7400  10100   8200   5500   7900   6200   5700  13100   5400   8900
## [231]   5700   7100  13000   5400  14700  38900   7100  79900  23700  12400
## [241]  14800   5600  30200  27400   9000  16900   9700  45900   5700   5600
## [251]   5700  11200  11500   8300   7800   6100  13900   8800  14400  10800
## [261]  25900   5500   6300  12600   6200  28600   8900   8900  10100   5400
## [271]  12000  17500  13100   7200  74300   6100   7900   6400   5800  10800
## [281]  17900   9700   9500   8000  41800   6900   7000  13300   7700   6300
## [291]  19400   5400   9100   6900  22300   5600   6800   6700  22500  23600
## [301]  18800   6600  15900  14000   7200  15200  42500  10700  16800   9500
## [311]  13600  23100  22800   5600   7800   9300  12200   5800   7400   9000
## [321]  10600   5500   8100  16100   5400   8100   5800 180600   5400   6600
```

The skewed distribution of the dependent variable in combination with the existence of the outstanding number 330 of outliers suggests reason for log transformation to reduce the influence of those observations in the model prediction, action in which we will proceed. After this change, the distribution of the log_shares is depicted in Diagram 7, as shown below.

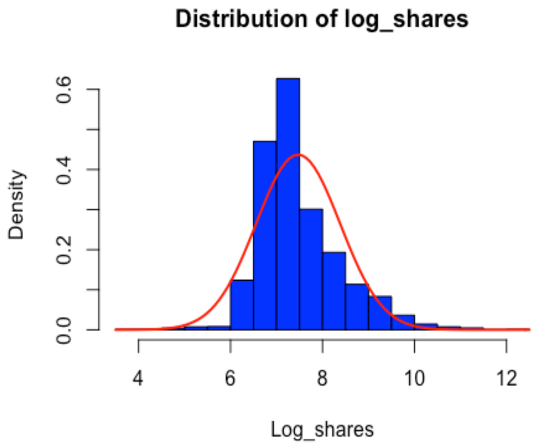


Diagram 7 – Shares' Logarithm Distribution

3. PAIRWISE COMPARISONS

In this section the variable correlations are analyzed and interpreted between (i) the log_shares and the most important numerical variables and (ii) the log_shares and the categorical variables. By examining the Diagrams 8,10 and 12 thoroughly, the first conclusion to be excluded is that the log_shares variable seems to have a non-linear relationship with the other variables, something which will prevent the data linearity assumption of the prediction model to be valid when any of these variables is selected as coefficients. Moreover, the existence of outliers in the created pairwise associations is again observed and verified raising a concern about potentially influential points that might affect the model.

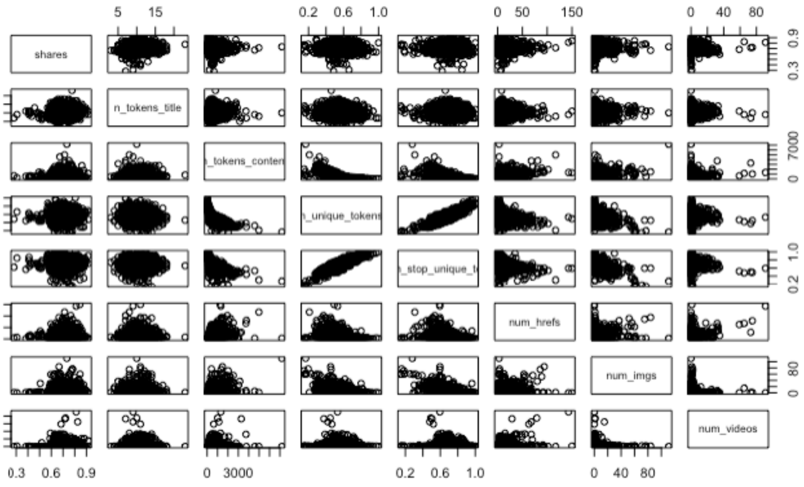


Diagram 8 – Numerical Variables Pairs (I)

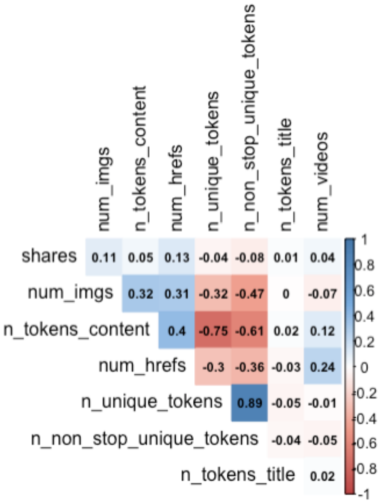


Diagram 9 – Numerical Variables Correlation (I)

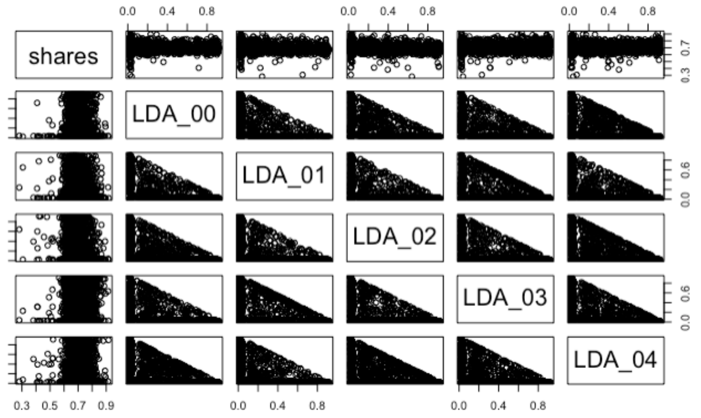


Diagram 10 – Numerical Variables Pairs (II)

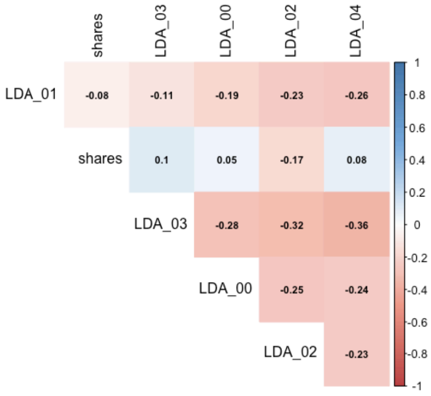


Diagram 11 – Numerical Variables Correlation (II)

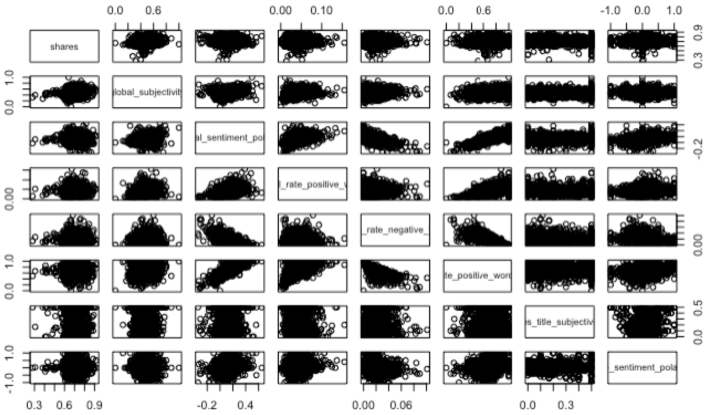


Diagram 12 – Numerical Variables Pairs (III)

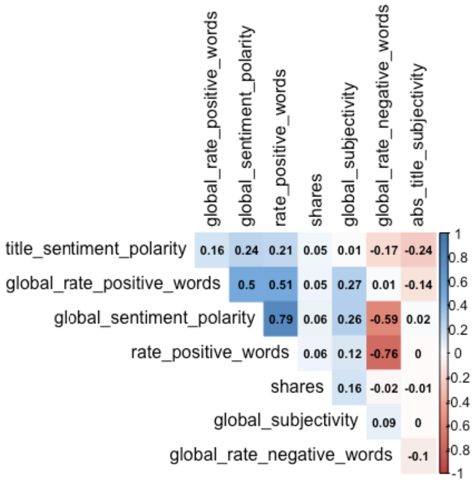


Diagram 13 – Numerical Variables Correlation (III)

On top of that as shown in Diagrams 9,11 and 13 where the correlation percentages fluctuate in extremely low levels, log_shares variable is not highly correlated with none of these variables. However, the variables which the log_shares variable has the highest correlation with are LDA_02 (17%), global_subjectivity (16%), num_hrefs (13%), num_imgs (11%), n_non_stop_unique_tokens (8%) and LDA_04 (8%). The relation between log_shares-num_hrefs and log_shares-num_imgs is almost expected since most people tend to be more affected when reading articles/posts which contain visual clues that are tangible and more easily understandable. The same also applies to the log_shares-global_subjectivity relation taken into account that an article with subjectivity characteristics has a greater impact on the readers' psyche making them willing to share it with other readers.

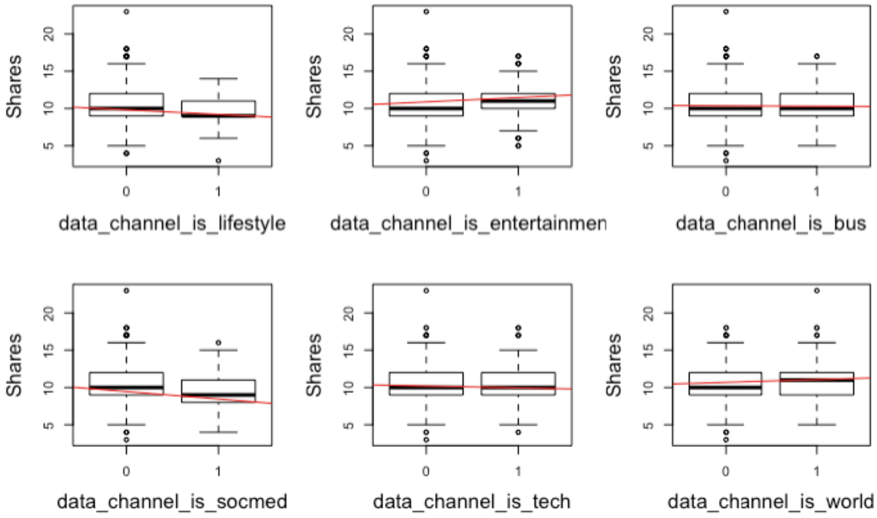


Diagram 14 – Shares on Factor Variables (I)

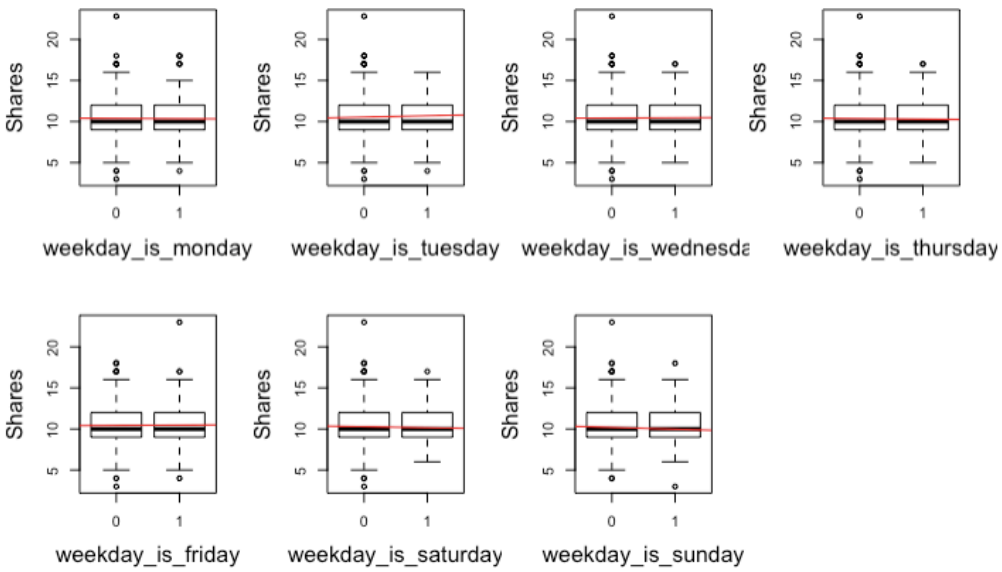


Diagram 15 – Shares on Factor Variables (II)

In the above Diagrams 14 and 15, the degree in which the log_shares are affected by article topic (Diagram 14) and article published day (Diagram 15) are depicted. More specifically, it is seen that shares have a negative relation with lifestyle, social media and technology topics; when the article refers to subjects other than these (i.e. entertainment, business, world-wide news), its shares are increased, otherwise they are decreased. It is also clear that as the weekend approaches during which articles are posted in Sunday and Saturday, the shares of these articles are reduced in comparison with the other days. The last conclusion seems reasonable as people have more free time at weekends which is usually and mostly spent on various activities other than browsing the web for informative purposes, a common habit of working days. Last but equally important is that no influential points appear to exist for any of the existent factors since the lines of all the diagrams fit relatively well in the created boxes.

4. PREDICTIVE/DESCRIPTIVE MODELS

4.1. PREDICTION MODEL (TRAINING DATASET)

In this subsection a multiple regression model will be constructed that will be the most suitable for predicting the popularity of a post. As mentioned before, the response variable is shares (for which a log transformation has already be done in section 2 and the other variables are the predictors. The model searching process will be initiated for a linear model where $y_i = b_0 + b_1x_1 + b_2x_2 + ... +b_px_p + \varepsilon_i$ for i = 1,2, ... n.

The initial model is the full model; it includes all the 58 variables of the training data set (see Table 1 – Initial full model). In this model:

- A great number of variables is considered insignificant (P_r(>|t|)>0.05).
- The P_r(>|t|) value of the n_non_stop_words, LDA_04, weekday_is_sunday variables is calculated as NA implying multicollinearity problem for these variables.
- The adjusted R-squared is low (0.1572) and the residual standard error is high (0.8381).

Since multicollinearity has been identified, the relevant 3 columns will be deleted from the training data set to fix this problem. Then the stepwise method (backward and forward) is applied to this later model in order to decrease the number of predictors by selecting both AIC and BIC criteria to compare the results.

- When selecting the AIC criterion, 24 variables (and the constant term) are included in the full model out of which three (global_rate_positive_words, title_sentiment_polarity, kw_min_max) are not statistically important. The adjusted R-squared counts to 0.161 and the residual standard error to 0.8362 (See Table 2 - Full model (AIC criterion)).
- When selecting the BIC criterion, 16 variables (and the constant term) are included in the full model all of which are statistically important. The adjusted R-squared counts is slightly lower to 0.154 and the residual standard error is 0.8397 (See Table 3 - Full model (BIC criterion)).

Comparing all three models, it is derived that the BIC model has the least variables with a difference of 8 and 42 variables from the AIC full model and full model respectively. It has a greater residual standard error and less adjusted R-squared than the other two models which leads to the conclusion that it may be the best fitting model for now, something which will also be verified when these models are evaluated in a subsequent section. The low R-squared and adjusted R-squared (0.1586 and 0.154) values imply high-variability and high-variance data. Although such data can cause problems when precisions are required for the predictions since data points fall further from the regression line, the predictor variable still provides information about the response. What should also be highlighted is that the p-value of the F-Statistic is less than 0.05 so the H_o hypothesis can be rejected, i.e. the fit of the intercept only model and the current model are not the same indicating that additional variables do not provide value when taken together.

4.2. MODEL ASSUMPTIONS

In this subsection the assumptions of the linear regression model will be validated. These assumptions are:

- Linearity of residuals
- Homoscedasticity of residuals variance
- Normality of residuals
- Outliers and leverage points

In Diagram 12, it is clear that residuals do not follow a normal distribution because the points except those located in the center do not lie closely to the line. This is also verified from the result of the Lilliefors and Shapiro-Wilk normality tests (p-value<0.05). In addition, what it is observed is that in the Residuals vs. Fitted plot, the red line is not approximately horizontal at zero supporting the absence of a linearity relationship between the predictor and the outcome variables, as already highlighted in Section 2. Moreover, it can be seen that the variances of the residual points increase with the value of the fitted outcome variable, suggesting non-constant variances in the residuals' errors, meaning that heteroscedasticity exists. Levene test and ncvTest confirm this conclusion as well ($p< 0.05$ and $P_r<0.05$ respectively). Last but not least, the Residuals vs. Leverage plot argues that no outliers are present, as all values fall well within the 0.5 bands (the bands are outside of the visible area).

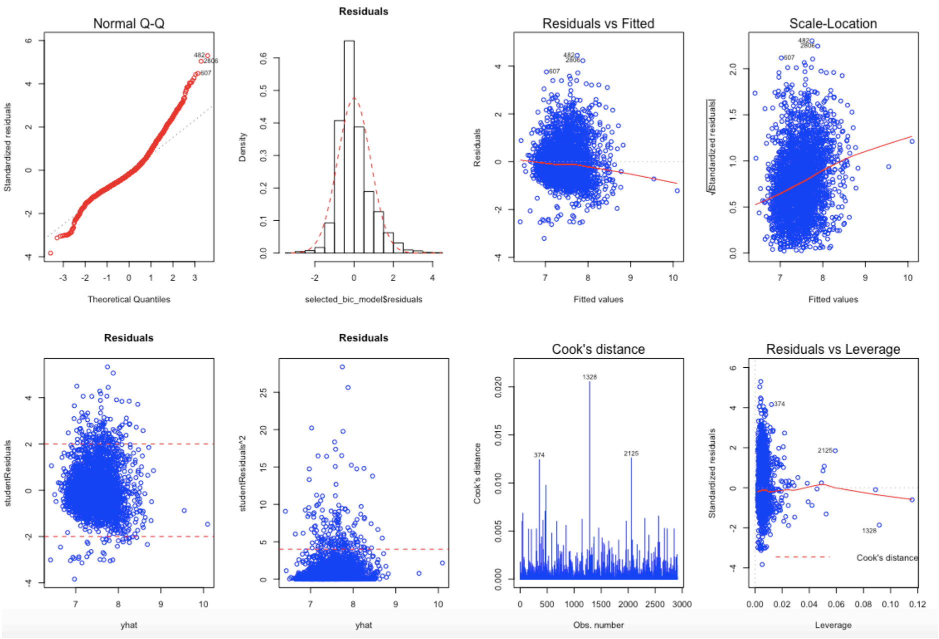


Diagram 16 – Normality, linearity, homoscedasticity, outliers

Since normality, linearity and homoscedasticity problems are identified, a significant effort is put into fixing them. More specifically, polynomial terms (up to fifth grade) and logarithms for dependent variables participated in the full BIC model. As far as the polynomial terms are concerned, a model is created with better fitting (See Table 4 – Polynomial model (AIC criterion)) but with an extremely difficult interpretation. On top of that, many coefficients not statistically significant are included leading to a probable overfitting of the model.

The second practice chosen related to algorithms doesn't return any better results (neither for the fitting nor for the fixing of regression assumptions) due to the fact that logarithms could be applied only for a few variables of the model for which non-zero and non-negative values exist (See Table 5 – Logarithms model, Table 6 – Logarithms model (AIC criterion) and Table 7 – Logarithms model (BIC criterion)).

To conclude, it is not possible to find a model for which all the regression assumptions will be valid simultaneously. From the trials made, the polynomial terms improved the normality, linearity and homoscedasticity problems a little but not to the extend required for the relevant problems to be completely faced.

4.3. MODEL EVALUATION

In this subsection the models created in subsection 4.2 are evaluated in order for their out-of-sample predictive ability to be assessed by using another sample of 10,000 observations. These models are the full model, the AIC and BIC full models, the AIC and BIC polynomial models (See Table 8-Full model (10-fold cross validation), Table 9-Full AIC model (10-fold cross validation), Table 10-Full BIC model (10-fold cross validation), Table 11-Polynomial AIC model (10-fold cross validation) and Table 12-Polynomial BIC model (10-fold cross validation)).

The first model to be tested is the full model. This model presents much more predictors that are statistically significant ($P_r(>|t|)<0.05$) with the adjusted R-squared to be decreased to 0.1338 as opposed to 0.1572 of the training data set. The residual standard error is higher fluctuating to 0.8607.

The prediction ability of this AIC full model is significantly lower in the testing data set with the adjusted R-squared to be 0.1253 in comparison to 0.161 that existed in the training data set. A few predictors are not significant ($P_r(>|t|)>0.05$) with the residual standard error to seem slightly increased to 0.8649 towards 0.8362.

The BIC full model is the one with the most significant differences identified between the test and the training data set. The adjusted R-squared counts to 0.1092, a high difference from 0.154. Moreover, it includes three predictors not statistically significant ($P_r(>|t|)>0.05$) which in the training data appeared to be significant ($P_r(>|t|)<0.05$). The residual standard error increase is calculated around 4% ((0.8729 - 0.8397) / 0.8397) for this model.

The AIC and BIC polynomial models not only have much higher adjusted R-squared value in the test data set (0.1198 and 0.05273 respectively) but have also much higher residual standard error (0.8677 and 0.9001 respectively).

4.4. PREDICTION MODEL (TEST DATASET)

In this subsection the multiple regression model will be constructed by using the test dataset of 10,000 observations. The initial model is again the full model; it includes all the 55 variables of the training data set (see Table 13-Full model (test data)) since three have already been removed due to multicollinearity problems (n_non_stop_words, LDA_04, weekday_is_Sunday). In this model:

- A great number of variables is considered insignificant ($P_r(>|t|)>0.05$).
- The adjusted R-squared is low (0.1338) and the residual standard error is high (0.8607).

When applying the stepwise methods:

- And the AIC criterion is selected, 37 variables (and the constant term) are included in the full model out of which three (global_rate_positive_words, title_sentiment_polarity, kw_min_max) are not statistically important. The adjusted R-squared value counts to 0.1349 and the residual standard error to 0.8602 (See Table 14 - Full AIC model (test data)).

- And the BIC criterion is selected, 21 variables (and the constant term) are included in the full model all of which are statistically important. The adjusted R-squared value counts slightly lower to 0.1299 and the residual standard error to 0.8627 (See Table 15-Full BIC model (test data)).

After applying 10-fold cross validation on the above models, it is concluded that:

- The full model presents much more predictors that are statistically significant ($P_r(>|t|)<0.05$) with the adjusted R-squared value to be decreased to 0.1572 as opposed to 0.1338 of the test data set. The residual standard error is higher, fluctuating to 0.8381.
- The prediction ability of this AIC full model is significantly increased in the training data set with the adjusted R-squared value to be 0.156 in comparison to 0.1349 that existed in the test data set. A great number of predictors are not significant ($P_r(>|t|)>0.05$) with the residual standard error to seem decreased to 0.8387 towards 0.8602.
- In the BIC full model, the adjusted R-squared counts to 0.1526, a great difference from 0.1299. Moreover, it includes four predictors not statistically significant ($P_r(>|t|)>0.05$) which in the test data appeared to be significant ($P_r(>|t|)<0.05$).

Based on the above, it seems that in the test data set where many more observations existed for the analysis, the significance of the predictors, the residual standard errors and the adjusted R-squared values were calculated in a stricter way and this is why these differences are observed. Still the BIC model appears to be the most suitable if taking into account that although it has a higher residual standard error (around 1.98%) and less variability (around 0.05%), it includes a significant number of fewer variables in comparison to the other two models (AIC, full) leading to a better prediction fit.

4.5. FINAL MODEL & INTERPRETATION

In this subsection, the final prediction model will be selected for the posts popularity. During this decision process, all the findings of section 2 and subsections 4.1, 4.2, 4.3 and 4.4 should be taken into account.

The ***full model*** is extremely complicated with a great number of insignificant variables ($P_r(>|t|)>0.05$) which count to 55.
The ***AIC model*** includes 24 variables out of which only three variables are not statistically significant with its prediction abilities to be quite similar in the evaluation dataset, though.
On the other hand, the ***BIC model*** has relatively fewer variables (16 in total) but it presents noticeable differences when evaluated in the test dataset implying serious doubts about its accuracy.
The polynomial models (either the AIC or the BIC) do not provide much greater prediction results than these of the other three models especially if their interpretation complexity is taken for granted.

Based on the aforementioned points, the AIC model is the most appropriate one for the analysis. The difference in the number of the variables is small and does not suggest a critical reason for excluding this model option since it does have a better goodness of fit, more precision and a rather simple interpretation of the factors which result in the popularity of a post. Therefore, the following model is selected:

$$\begin{aligned}\log(\text{shares})_i = {} & 6.07 + 2.3 * \text{n\_tokens\_title}_i + 1.19 * \text{n\_tokens\_content}_i + 3.86 * \text{num\_hrefs}_i \\ & - 2.10 * \text{data\_channel\_is\_entertainment}_i + 1.39 * \text{data\_channel\_is\_socmed}_i \\ & - 2.10 * \text{data\_channel\_is\_tech}_i + 8.95 * \text{kw\_min\_min}_i + 1.39 * \text{kw\_min\_max}_i \\ & - 6.96 * \text{kw\_min\_avg}_i - 4.10 * \text{kw\_max\_avg}_i + 3.94 * \text{kw\_avg\_avg}_i \\ & + 2.33 * \text{self\_reference\_min\_shares}_i - 1.70 * \text{weekday\_is\_monday}_i \\ & - 3.26 * \text{weekday\_is\_tuesday}_i - 2.51 * \text{weekday\_is\_wednesday}_i \\ & - 3.30 * \text{weekday\_is\_thursday}_i - 1.37 * \text{weekday\_is\_friday}_i + 2.24 * \text{LDA\_00} \\ & + 7.60 * \text{global\_subjectivity}_i - 5.11 * \text{min\_positive\_polarity}_i \\ & - 5.84 * \text{avg\_negative\_polarity}_I + 2.72 * \text{min\_negative\_polarity}_i\end{aligned}$$

$+ 1.16 * \text{title\_sentiment\_polarity}_i - 1.81 * \text{global\_rate\_positive\_words}_i + \varepsilon_i$
where $\varepsilon_i = N(0, 0.8362^2)$

The following should be highlighted regarding the selected model:

- Since the dependent variable of shares is a log transformed variable, then its change is interpreted as a percentage change in terms of a one-unit x_i change. For instance, when the number of words included in an article/post is increased by one word, then the shares of this post will be increased by 1.19%.
- The strongest positive effect to the shares is this of the worst keyword variable (8.95%) while the strongest negative effect comes from the average keyword's variable (- 6.96%)
- As mentioned before, the adjusted R-squared value points to 0.161 which provides valid information about the relationship between shares and the relevant independent variable with a lack of precision when this is needed for the prediction, though. Further analysis would be required out of this study to identify another model better for predictions.

5. Conclusions & Viral Post Characteristics

After significant effort the final model for online news popularity is selected. As initially thought, the article posting day and the article subject play an important role determining the sharing of posts since the majority of the available factors were included in the selected multiple regression model. People do not also prefer to share posts related to technology or entertainment and are affected by the text subjectivity and the title polarity. Moreover, it seems that the increasing extent of the article's content and title, predetermine positively the readers to share it with others. As far as the characteristics of a viral post are concerned, it seems that an article including a great number of links and worst keywords, a high degree of subjectivity in its content and a high average number of keywords could suggest crucial factors for boosting the shares of an article to the top.

Another point worth mentioning is that the assumptions of the model could not be impressively corrected as part of this research, something crucial for accurate prediction models. A linear simpler model is selected that interprets the popularity prediction in a limited way. However, this prediction ability could be further enhanced in the future if more train and evaluation data also existed for a much more detailed research.

6. TABLE OF FIGURES

7. APPENDIX I (LIST OF TABLES)

Table 1 – Initial full model

```
## Call:
## lm(formula = shares ~ ., data = onlineNews)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1596 -0.5267 -0.1471  0.3477  4.3479
##
## Coefficients: (3 not defined because of singularities)
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     6.304e+00  9.405e-01   6.703 2.46e-11 ***
## n_tokens_title                  2.487e-02  7.669e-03   3.243 0.001195 **
## n_tokens_content                7.751e-05  6.008e-05   1.290 0.197104
## n_unique_tokens                -2.765e-01  5.114e-01  -0.541 0.588743
## n_non_stop_words                       NA         NA      NA       NA
## n_non_stop_unique_tokens        5.072e-02  4.396e-01   0.115 0.908157
## num_hrefs                       4.403e-03  1.774e-03   2.483 0.013101 *
## num_self_hrefs                 -7.585e-03  4.962e-03  -1.529 0.126474
## num_imgs                        1.653e-03  2.507e-03   0.660 0.509580
## num_videos                      3.788e-03  3.767e-03   1.006 0.314645
## average_token_length           -2.477e-02  6.512e-02  -0.380 0.703695
## num_keywords                   -1.742e-03  1.031e-02  -0.169 0.865795
## data_channel_is_lifestyle1     -1.624e-02  1.127e-01  -0.144 0.885457
## data_channel_is_entertainment1 -2.335e-01  6.900e-02  -3.384 0.000723 ***
## data_channel_is_bus1            3.656e-03  1.077e-01   0.034 0.972921
## data_channel_is_socmed1         1.280e-01  1.047e-01   1.223 0.221577
## data_channel_is_tech1           2.136e-01  1.023e-01   2.088 0.036888 *
## data_channel_is_world1         -1.015e-02  1.047e-01  -0.097 0.922779
## kw_min_min                      6.853e-04  4.389e-04   1.561 0.118553
## kw_max_min                      1.390e-05  2.580e-05   0.539 0.590003
## kw_avg_min                     -1.503e-04  1.584e-04  -0.949 0.342719
## kw_min_max                     -3.907e-07  3.104e-07  -1.259 0.208242
## kw_max_max                     -5.058e-08  1.571e-07  -0.322 0.747565
## kw_avg_max                     -2.450e-07  2.374e-07  -1.032 0.302153
## kw_min_avg                     -6.827e-05  2.170e-05  -3.146 0.001675 **
## kw_max_avg                     -3.926e-05  9.615e-06  -4.083 4.56e-05 ***
## kw_avg_avg                      3.999e-04  4.521e-05   8.845  < 2e-16 ***
## self_reference_min_shares       3.014e-06  2.008e-06   1.501 0.133451
## self_reference_max_shares       7.140e-07  9.945e-07   0.718 0.472850
## self_reference_avg_sharess     -1.526e-06  2.706e-06  -0.564 0.572730
## weekday_is_monday1             -1.557e-01  7.261e-02  -2.145 0.032052 *
## weekday_is_tuesday1            -3.092e-01  7.203e-02  -4.293 1.82e-05 ***
## weekday_is_wednesday1          -2.316e-01  7.138e-02  -3.244 0.001192 **
## weekday_is_thursday1           -3.133e-01  7.176e-02  -4.366 1.31e-05 ***
## weekday_is_friday1             -1.214e-01  7.422e-02  -1.635 0.102078
## weekday_is_saturday1            4.316e-02  8.848e-02   0.488 0.625776
## weekday_is_sunday1                     NA         NA      NA       NA
## LDA_00                          1.116e-01  1.244e-01   0.897 0.369713    
## LDA_01                         -9.713e-02  1.381e-01  -0.703 0.481918    
## LDA_02                         -1.655e-01  1.266e-01  -1.307 0.191273    
## LDA_03                         -1.379e-01  1.338e-01  -1.030 0.302885    
## LDA_04                                 NA         NA      NA       NA    
## global_subjectivity             7.018e-01  2.260e-01   3.105 0.001924 ** 
## global_sentiment_polarity      -4.568e-01  4.523e-01  -1.010 0.312650    
## global_rate_positive_words     -1.470e+00  1.806e+00  -0.814 0.415742    
## global_rate_negative_words      2.888e-01  3.592e+00   0.080 0.935920    
## rate_positive_words             3.565e-01  8.616e-01   0.414 0.679098    
## rate_negative_words             6.413e-02  8.846e-01   0.072 0.942216    
## avg_positive_polarity           3.837e-01  3.770e-01   1.018 0.308805    
## min_positive_polarity          -5.936e-01  3.046e-01  -1.948 0.051459 .  
## max_positive_polarity          -1.494e-01  1.144e-01  -1.306 0.191555    
## avg_negative_polarity          -7.751e-01  3.408e-01  -2.274 0.023032 *  
## min_negative_polarity           3.298e-01  1.222e-01   2.699 0.007000 ** 
## max_negative_polarity           3.186e-01  2.844e-01   1.120 0.262602    
## title_subjectivity              4.116e-03  7.496e-02   0.055 0.956217    
## title_sentiment_polarity        1.095e-01  6.748e-02   1.623 0.104693    
## abs_title_subjectivity          6.845e-02  9.831e-02   0.696 0.486330    
## abs_title_sentiment_polarity    8.428e-02  1.090e-01   0.773 0.439341    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.8381 on 2854 degrees of freedom
## Multiple R-squared:  0.1729, Adjusted R-squared:  0.1572 
## F-statistic: 11.05 on 54 and 2854 DF,  p-value: < 2.2e-16
```

Table 2 – Full model (AIC criterion)

```
## 
## Call:
## lm(formula = shares ~ n_tokens_title + n_tokens_content + num_hrefs + 
##     data_channel_is_entertainment + data_channel_is_socmed + 
##     data_channel_is_tech + kw_min_min + kw_min_max + kw_min_avg + 
##     kw_max_avg + kw_avg_avg + self_reference_min_shares + weekday_is_monda
y + 
##     weekday_is_tuesday + weekday_is_wednesday + weekday_is_thursday + 
##     weekday_is_friday + LDA_00 + global_subjectivity + min_positive_polari
ty + 
##     avg_negative_polarity + min_negative_polarity + title_sentiment_polari
ty + 
##     global_rate_positive_words, data = onlineNews)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.1583 -0.5230 -0.1495  0.3702  4.5042 
## 
```

```
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                    6.079e+00  1.372e-01  44.291  < 2e-16 ***
## n_tokens_title                 2.312e-02  7.405e-03   3.123 0.001811 ** 
## n_tokens_content               1.191e-04  4.407e-05   2.702 0.006935 ** 
## num_hrefs                      3.859e-03  1.536e-03   2.512 0.012069 *  
## data_channel_is_entertainment1 -2.104e-01  4.408e-02  -4.773 1.91e-06 ***
## data_channel_is_socmed1        1.387e-01  6.892e-02   2.013 0.044247 *  
## data_channel_is_tech1          2.905e-01  4.415e-02   6.580 5.56e-11 ***
## kw_min_min                      8.954e-04  2.333e-04   3.837 0.000127 ***
## kw_min_max                     -4.859e-07  2.764e-07  -1.758 0.078785 .
## kw_min_avg                     -6.965e-05  1.964e-05  -3.546 0.000397 ***
## kw_max_avg                     -4.107e-05  8.037e-06  -5.109 3.44e-07 ***
## kw_avg_avg                      3.944e-04  3.364e-05  11.722  < 2e-16 ***
## self_reference_min_shares       2.328e-06  7.785e-07   2.990 0.002813 **
## weekday_is_monday1             -1.699e-01  5.803e-02  -2.928 0.003434 **
## weekday_is_tuesday1            -3.260e-01  5.717e-02  -5.702 1.30e-08 ***
## weekday_is_wednesday1          -2.508e-01  5.676e-02  -4.418 1.03e-05 ***
## weekday_is_thursday1           -3.301e-01  5.701e-02  -5.789 7.85e-09 ***
## weekday_is_friday1             -1.375e-01  6.011e-02  -2.287 0.022269 *
## LDA_00                          2.242e-01  6.549e-02   3.423 0.000627 ***
## global_subjectivity             7.608e-01  2.018e-01   3.770 0.000166 ***
## min_positive_polarity          -5.111e-01  2.499e-01  -2.045 0.040925 *
## avg_negative_polarity          -5.836e-01  2.010e-01  -2.903 0.003729 **
## min_negative_polarity           2.725e-01  9.475e-02   2.876 0.004059 **
## title_sentiment_polarity        1.157e-01  6.092e-02   1.898 0.057750 .
## global_rate_positive_words     -1.810e+00  1.052e+00  -1.720 0.085591 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8362 on 2884 degrees of freedom
## Multiple R-squared:  0.1679, Adjusted R-squared:  0.161
## F-statistic: 24.24 on 24 and 2884 DF, p-value: < 2.2e-16
```

Table 3 – Full model (BIC criterion)

```
## Call:
## lm(formula = shares ~ n_tokens_title + n_tokens_content + num_hrefs +
##     data_channel_is_entertainment + data_channel_is_tech + kw_min_min +
##     kw_min_avg + kw_max_avg + kw_avg_avg + weekday_is_tuesday +
##     weekday_is_wednesday + weekday_is_thursday + LDA_00 + global_subjectiv
ity +
##     avg_negative_polarity + min_negative_polarity, data = onlineNews)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2084 -0.5276 -0.1459  0.3804  4.4437
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.897e+00  1.273e-01  46.319  < 2e-16 ***
## n_tokens_title                  2.165e-02  7.389e-03   2.930 0.003412 **
## n_tokens_content                1.369e-04  4.325e-05   3.166 0.001560 **
## num_hrefs                       4.453e-03  1.531e-03   2.909 0.003655 **
## data_channel_is_entertainment1 -2.267e-01  4.385e-02  -5.171 2.49e-07 ***
## data_channel_is_tech1           2.863e-01  4.369e-02   6.553 6.66e-11 ***
## kw_min_min                      8.551e-04  2.327e-04   3.674 0.000243 ***
## kw_min_avg                     -7.557e-05  1.919e-05  -3.939 8.39e-05 ***
## kw_max_avg                     -4.105e-05  8.006e-06  -5.127 3.14e-07 ***
## kw_avg_avg                      3.970e-04  3.338e-05  11.895  < 2e-16 ***
## weekday_is_tuesday1            -2.100e-01  4.335e-02  -4.846 1.33e-06 ***
## weekday_is_wednesday1          -1.417e-01  4.287e-02  -3.307 0.000956 ***
## weekday_is_thursday1           -2.204e-01  4.325e-02  -5.096 3.69e-07 ***
## LDA_00                          2.459e-01  6.385e-02   3.851 0.000120 ***
## global_subjectivity             6.659e-01  1.917e-01   3.473 0.000522 ***
## avg_negative_polarity          -5.991e-01  2.011e-01  -2.979 0.002916 **
## min_negative_polarity           2.706e-01  9.469e-02   2.857 0.004303 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8397 on 2892 degrees of freedom
## Multiple R-squared:  0.1586, Adjusted R-squared:  0.154
## F-statistic: 34.08 on 16 and 2892 DF, p-value: < 2.2e-16
```

Table 4 – Polynomial model (AIC criterion)

```
## Call:
## lm(formula = shares ~ poly(n_tokens_title, 5, raw = TRUE) + poly(num_hrefs
,
##     5, raw = TRUE) + poly(kw_min_min, 5, raw = TRUE) + poly(kw_min_avg,
##     5, raw = TRUE) + poly(kw_max_avg, 5, raw = TRUE) + poly(kw_avg_avg,
##     5, raw = TRUE) + poly(LDA_00, 5, raw = TRUE) + poly(global_subjectivit
y,
##     5, raw = TRUE) + data_channel_is_entertainment + data_channel_is_tech
+
##     weekday_is_tuesday + weekday_is_wednesday + weekday_is_thursday,
##     data = onlineNews)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3422 -0.5137 -0.1296  0.3946  4.3666
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             2.737e+00  3.032e+00   0.902
## poly(n_tokens_title, 5, raw = TRUE)1    1.825e+00  1.378e+00   1.324
## poly(n_tokens_title, 5, raw = TRUE)2   -3.186e-01  2.526e-01  -1.261
## poly(n_tokens_title, 5, raw = TRUE)3    2.684e-02  2.215e-02   1.212
## poly(n_tokens_title, 5, raw = TRUE)4   -1.088e-03  9.272e-04  -1.174
## poly(n_tokens_title, 5, raw = TRUE)5    1.714e-05  1.477e-05   1.160
## poly(num_hrefs, 5, raw = TRUE)1         5.635e-04  1.263e-02   0.045
## poly(num_hrefs, 5, raw = TRUE)2         1.270e-04  8.352e-04   0.152
## poly(num_hrefs, 5, raw = TRUE)3         2.337e-06  2.052e-05   0.114
## poly(num_hrefs, 5, raw = TRUE)4        -7.318e-08  2.005e-07  -0.365
## poly(num_hrefs, 5, raw = TRUE)5         3.801e-10  6.545e-10   0.581
## poly(kw_min_min, 5, raw = TRUE)1        1.502e-02  5.582e-01   0.027
## poly(kw_min_min, 5, raw = TRUE)2       -1.174e-02  2.179e-01  -0.054
## poly(kw_min_min, 5, raw = TRUE)3        4.452e-04  7.703e-03   0.058
## poly(kw_min_min, 5, raw = TRUE)4       -5.192e-06  8.549e-05  -0.061
## poly(kw_min_min, 5, raw = TRUE)5        1.561e-08  2.514e-07   0.062
## poly(kw_min_avg, 5, raw = TRUE)1       -1.207e-03  6.096e-04  -1.980
## poly(kw_min_avg, 5, raw = TRUE)2        1.762e-06  1.260e-06   1.398
## poly(kw_min_avg, 5, raw = TRUE)3       -1.177e-09  9.241e-10  -1.274
## poly(kw_min_avg, 5, raw = TRUE)4        3.772e-13  2.856e-13   1.321
## poly(kw_min_avg, 5, raw = TRUE)5       -4.562e-17  3.160e-17  -1.444
## poly(kw_max_avg, 5, raw = TRUE)1       -8.972e-06  1.008e-04  -0.089
## poly(kw_max_avg, 5, raw = TRUE)2       -2.521e-09  1.313e-08  -0.192
## poly(kw_max_avg, 5, raw = TRUE)3        1.273e-13  7.023e-13   0.181
## poly(kw_max_avg, 5, raw = TRUE)4       -2.437e-18  1.581e-17  -0.154
## poly(kw_max_avg, 5, raw = TRUE)5        1.276e-23  1.234e-22   0.103
## poly(kw_avg_avg, 5, raw = TRUE)1       -7.082e-04  8.369e-04  -0.846
## poly(kw_avg_avg, 5, raw = TRUE)2        6.236e-07  3.986e-07   1.564
## poly(kw_avg_avg, 5, raw = TRUE)3       -1.487e-10  8.615e-11  -1.727
## poly(kw_avg_avg, 5, raw = TRUE)4        1.509e-14  8.432e-15   1.789
## poly(kw_avg_avg, 5, raw = TRUE)5       -5.374e-19  2.969e-19  -1.810
## poly(LDA_00, 5, raw = TRUE)1            2.067e+00  2.371e+00   0.872
## poly(LDA_00, 5, raw = TRUE)2           -1.864e+01  1.838e+01  -1.014
## poly(LDA_00, 5, raw = TRUE)3               6.031e+01  5.377e+01   1.122
## poly(LDA_00, 5, raw = TRUE)4              -7.728e+01  6.650e+01  -1.162
## poly(LDA_00, 5, raw = TRUE)5               3.430e+01  2.932e+01   1.170
## poly(global_subjectivity, 5, raw = TRUE)1 -1.976e+00  9.726e+00  -0.203
## poly(global_subjectivity, 5, raw = TRUE)2  2.410e+01  4.799e+01   0.502
## poly(global_subjectivity, 5, raw = TRUE)3 -7.502e+01  1.114e+02  -0.674
## poly(global_subjectivity, 5, raw = TRUE)4  9.812e+01  1.200e+02   0.817
## poly(global_subjectivity, 5, raw = TRUE)5 -4.584e+01  4.818e+01  -0.951
## data_channel_is_entertainment1            -2.537e-01  4.528e-02  -5.603
## data_channel_is_tech1                      2.668e-01  4.486e-02   5.948
## weekday_is_tuesday1                       -2.120e-01  4.335e-02  -4.890
## weekday_is_wednesday1                     -1.453e-01  4.294e-02  -3.383
## weekday_is_thursday1                      -2.212e-01  4.327e-02  -5.113
##                                           Pr(>|t|)    
## (Intercept)                               0.366887    
## poly(n_tokens_title, 5, raw = TRUE)1      0.185451    
## poly(n_tokens_title, 5, raw = TRUE)2      0.207403    
## poly(n_tokens_title, 5, raw = TRUE)3      0.225776    
## poly(n_tokens_title, 5, raw = TRUE)4      0.240533    
## poly(n_tokens_title, 5, raw = TRUE)5      0.246004    
## poly(num_hrefs, 5, raw = TRUE)1           0.964407    
## poly(num_hrefs, 5, raw = TRUE)2           0.879145    
## poly(num_hrefs, 5, raw = TRUE)3           0.909301    
## poly(num_hrefs, 5, raw = TRUE)4           0.715105    
## poly(num_hrefs, 5, raw = TRUE)5           0.561522    
## poly(kw_min_min, 5, raw = TRUE)1          0.978540    
## poly(kw_min_min, 5, raw = TRUE)2          0.957058    
## poly(kw_min_min, 5, raw = TRUE)3          0.953917    
## poly(kw_min_min, 5, raw = TRUE)4          0.951579    
## poly(kw_min_min, 5, raw = TRUE)5          0.950495    
## poly(kw_min_avg, 5, raw = TRUE)1          0.047749 *  
## poly(kw_min_avg, 5, raw = TRUE)2          0.162137    
## poly(kw_min_avg, 5, raw = TRUE)3          0.202885    
## poly(kw_min_avg, 5, raw = TRUE)4          0.186686    
## poly(kw_min_avg, 5, raw = TRUE)5          0.148961    
## poly(kw_max_avg, 5, raw = TRUE)1          0.929087    
## poly(kw_max_avg, 5, raw = TRUE)2          0.847761    
## poly(kw_max_avg, 5, raw = TRUE)3          0.856139    
## poly(kw_max_avg, 5, raw = TRUE)4          0.877524    
## poly(kw_max_avg, 5, raw = TRUE)5          0.917701    
## poly(kw_avg_avg, 5, raw = TRUE)1          0.397463    
## poly(kw_avg_avg, 5, raw = TRUE)2          0.117814    
## poly(kw_avg_avg, 5, raw = TRUE)3          0.084351 .  
## poly(kw_avg_avg, 5, raw = TRUE)4          0.073728 .  
## poly(kw_avg_avg, 5, raw = TRUE)5          0.070379 .  
## poly(LDA_00, 5, raw = TRUE)1              0.383362    
## poly(LDA_00, 5, raw = TRUE)2              0.310456    
## poly(LDA_00, 5, raw = TRUE)3              0.262083    
## poly(LDA_00, 5, raw = TRUE)4              0.245235    
## poly(LDA_00, 5, raw = TRUE)5              0.242151    
## poly(global_subjectivity, 5, raw = TRUE)1 0.839048    
## poly(global_subjectivity, 5, raw = TRUE)2 0.615519    
## poly(global_subjectivity, 5, raw = TRUE)3 0.500621    
## poly(global_subjectivity, 5, raw = TRUE)4 0.413787    
## poly(global_subjectivity, 5, raw = TRUE)5 0.341507    
## data_channel_is_entertainment1            2.31e-08 ***
## data_channel_is_tech1                     3.04e-09 ***
## weekday_is_tuesday1                       1.06e-06 ***
## weekday_is_wednesday1                     0.000727 ***
## weekday_is_thursday1                     3.39e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8347 on 2863 degrees of freedom
## Multiple R-squared:  0.1769, Adjusted R-squared:  0.164
## F-statistic: 13.67 on 45 and 2863 DF, p-value: < 2.2e-16
```

Table 5 – Logarithms model

```
##
## Call:
## lm(formula = shares ~ log(n_tokens_title) + log(n_tokens_content) +
##     num_hrefs + kw_min_min + kw_min_avg + kw_max_avg + kw_avg_avg +
##     log(LDA_00) + global_subjectivity + avg_negative_polarity +
##     min_negative_polarity + data_channel_is_entertainment + data_channel_i
s_tech +
##     weekday_is_tuesday + weekday_is_wednesday + weekday_is_thursday,
##     data = onlineNews)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2357 -0.5290 -0.1430  0.3776  4.4284
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    5.372e+00  2.649e-01  20.276  < 2e-16 ***
## log(n_tokens_title)            2.004e-01  7.436e-02   2.695 0.007070 **
## log(n_tokens_content)          8.475e-02  2.922e-02   2.900 0.003756 **
## num_hrefs                      4.531e-03  1.544e-03   2.934 0.003371 **
## kw_min_min                     8.795e-04  2.337e-04   3.763 0.000171 ***
## kw_min_avg                    -7.603e-05  1.925e-05  -3.950 8.01e-05 ***
## kw_max_avg                    -4.192e-05  8.050e-06  -5.208 2.04e-07 ***
## kw_avg_avg                     4.009e-04  3.370e-05  11.896  < 2e-16 ***
## log(LDA_00)                    3.934e-02  1.259e-02   3.125 0.001797 **
## global_subjectivity            6.760e-01  1.924e-01   3.513 0.000450 ***
## avg_negative_polarity         -5.705e-01  2.021e-01  -2.823 0.004797 **
## min_negative_polarity          2.770e-01  9.895e-02   2.799 0.005154 **
## data_channel_is_entertainment1 -2.366e-01  4.346e-02  -5.443 5.68e-08 ***
## data_channel_is_tech1          2.763e-01  4.345e-02   6.359 2.35e-10 ***
## weekday_is_tuesday1           -2.127e-01  4.343e-02  -4.899 1.02e-06 ***
## weekday_is_wednesday1         -1.409e-01  4.293e-02  -3.283 0.001041 **
## weekday_is_thursday1          -2.219e-01  4.330e-02  -5.123 3.20e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8408 on 2892 degrees of freedom
## Multiple R-squared:  0.1563, Adjusted R-squared:  0.1517
## F-statistic:  33.5 on 16 and 2892 DF, p-value: < 2.2e-16
```

Table 6 – Logarithms model (AIC criterion)

```
##
## Call:
## lm(formula = shares ~ log(n_tokens_title) + log(n_tokens_content) +
##     num_hrefs + kw_min_min + kw_min_avg + kw_max_avg + kw_avg_avg +
##     log(LDA_00) + global_subjectivity + avg_negative_polarity +
##     min_negative_polarity + data_channel_is_entertainment + data_channel_i
s_tech +
##     weekday_is_tuesday + weekday_is_wednesday + weekday_is_thursday,
##     data = onlineNews)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2357 -0.5290 -0.1430  0.3776  4.4284
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    5.372e+00  2.649e-01  20.276  < 2e-16 ***
## log(n_tokens_title)            2.004e-01  7.436e-02   2.695 0.007070 **
## log(n_tokens_content)          8.475e-02  2.922e-02   2.900 0.003756 **
## num_hrefs                      4.531e-03  1.544e-03   2.934 0.003371 **
## kw_min_min                     8.795e-04  2.337e-04   3.763 0.000171 ***
## kw_min_avg                    -7.603e-05  1.925e-05  -3.950 8.01e-05 ***
## kw_max_avg                    -4.192e-05  8.050e-06  -5.208 2.04e-07 ***
## kw_avg_avg                     4.009e-04  3.370e-05  11.896  < 2e-16 ***
## log(LDA_00)                    3.934e-02  1.259e-02   3.125 0.001797 **
## global_subjectivity            6.760e-01  1.924e-01   3.513 0.000450 ***
## avg_negative_polarity         -5.705e-01  2.021e-01  -2.823 0.004797 **
## min_negative_polarity          2.770e-01  9.895e-02   2.799 0.005154 **
## data_channel_is_entertainment1 -2.366e-01  4.346e-02  -5.443 5.68e-08 ***
## data_channel_is_tech1          2.763e-01  4.345e-02   6.359 2.35e-10 ***
## weekday_is_tuesday1           -2.127e-01  4.343e-02  -4.899 1.02e-06 ***
## weekday_is_wednesday1         -1.409e-01  4.293e-02  -3.283 0.001041 **
## weekday_is_thursday1          -2.219e-01  4.330e-02  -5.123 3.20e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8408 on 2892 degrees of freedom
## Multiple R-squared:  0.1563, Adjusted R-squared:  0.1517
## F-statistic:  33.5 on 16 and 2892 DF,  p-value: < 2.2e-16
```

Table 7 – Logarithms model (BIC criterion)

```
## Call:
## lm(formula = shares ~ num_hrefs + kw_min_min + kw_min_avg + kw_max_avg +
##     kw_avg_avg + log(LDA_00) + global_subjectivity + data_channel_is_enter
tainment +
##     data_channel_is_tech + weekday_is_tuesday + weekday_is_wednesday +
##     weekday_is_thursday, data = onlineNews)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3292 -0.5278 -0.1399  0.3795  4.5144
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    6.354e+00  9.529e-02  66.683  < 2e-16 ***
## num_hrefs                      5.779e-03  1.393e-03   4.147 3.46e-05 ***
## kw_min_min                     7.675e-04  2.319e-04   3.309 0.000947 ***
## kw_min_avg                    -7.263e-05  1.920e-05  -3.783 0.000158 ***
## kw_max_avg                    -4.023e-05  8.017e-06  -5.019 5.52e-07 ***
## kw_avg_avg                     3.904e-04  3.309e-05  11.797  < 2e-16 ***
## log(LDA_00)                    4.096e-02  1.257e-02   3.260 0.001128 **
## global_subjectivity            6.967e-01  1.859e-01   3.747 0.000182 ***
## data_channel_is_entertainment1 -2.172e-01  4.323e-02  -5.024 5.37e-07 ***
## data_channel_is_tech1          2.789e-01  4.301e-02   6.484 1.04e-10 ***
## weekday_is_tuesday1           -2.150e-01  4.346e-02  -4.948 7.94e-07 ***
## weekday_is_wednesday1         -1.400e-01  4.304e-02  -3.253 0.001155 **
## weekday_is_thursday1           -2.254e-01  4.341e-02  -5.192 2.22e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8432 on 2896 degrees of freedom
## Multiple R-squared:  0.1505, Adjusted R-squared:  0.147
## F-statistic: 42.76 on 12 and 2896 DF, p-value: < 2.2e-16
```

Table 8 – Full model (10-fold cross validation)

```
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1921 -0.5404 -0.1681  0.3860  5.2097
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    6.906e+00  8.925e-01   7.738 1.11e-14 ***
## n_tokens_title                 9.145e-03  4.351e-03   2.102  0.03560 *
## n_tokens_content               6.474e-05  3.448e-05   1.878  0.06044 .
## n_unique_tokens                5.116e-01  2.927e-01   1.748  0.08053 .
## n_non_stop_unique_tokens      -6.756e-01  2.485e-01  -2.718  0.00657 **
## num_hrefs                      2.221e-03  1.028e-03   2.161  0.03071 *
## num_self_hrefs                -8.822e-03  2.838e-03  -3.109  0.00188 **
## num_imgs                       2.459e-03  1.404e-03   1.752  0.07984 .
## num_videos                     1.434e-03  2.448e-03   0.586  0.55798
## average_token_length          -8.630e-02  3.632e-02  -2.376  0.01751 *
## num_keywords                  -5.212e-03  5.604e-03  -0.930  0.35233
## data_channel_is_lifestyle     -1.250e-01  5.932e-02  -2.107  0.03512 *
## data_channel_is_entertainment -1.738e-01  3.908e-02  -4.447 8.82e-06 ***
## data_channel_is_bus           -1.556e-01  5.753e-02  -2.705  0.00685 **
## data_channel_is_socmed         1.458e-01  5.661e-02   2.576  0.01001 *
## data_channel_is_tech           9.045e-02  5.619e-02   1.610  0.10751
## data_channel_is_world         -4.405e-02  5.694e-02  -0.774  0.43923
## kw_min_min                     1.066e-03  2.376e-04   4.485 7.38e-06 ***
## kw_max_min                     2.595e-05  7.761e-06   3.344  0.00083 ***
## kw_avg_min                    -2.241e-04  5.533e-05  -4.051 5.15e-05 ***
## kw_min_max                    -5.413e-09  1.813e-07  -0.030  0.97618
## kw_max_max                     1.751e-07  8.509e-08   2.058  0.03965 *
## kw_avg_max                    -7.802e-07  1.286e-07  -6.064 1.38e-09 ***
## kw_min_avg                    -7.865e-05  1.178e-05  -6.675 2.61e-11 ***
## kw_max_avg                    -4.716e-05  4.293e-06 -10.986  < 2e-16 ***
## kw_avg_avg                     3.924e-04  2.295e-05  17.096  < 2e-16 ***
## self_reference_min_shares      2.959e-06  1.116e-06   2.651  0.00804 **
## self_reference_max_shares     -2.662e-08  5.969e-07  -0.045  0.96443
## self_reference_avg_sharess     1.189e-07  1.517e-06   0.078  0.93755
## weekday_is_monday             -2.400e-01  3.876e-02  -6.193 6.13e-10 ***
## weekday_is_tuesday            -3.289e-01  3.773e-02  -8.716  < 2e-16 ***
## weekday_is_wednesday          -3.418e-01  3.797e-02  -9.003  < 2e-16 ***
## weekday_is_thursday           -3.370e-01  3.810e-02  -8.844  < 2e-16 ***
## weekday_is_friday             -2.433e-01  3.937e-02  -6.180 6.68e-10 ***
## weekday_is_saturday           -1.133e-02  4.772e-02  -0.237  0.81236
## LDA_00                         1.935e-01  6.768e-02   2.859  0.00426 **
## LDA_01                        -1.716e-01  7.585e-02  -2.262  0.02373 *
## LDA_02                        -2.680e-01  6.865e-02  -3.905 9.50e-05 ***
## LDA_03                        -1.595e-01  7.237e-02  -2.204  0.02756 *
## global_subjectivity            2.720e-01  1.258e-01   2.161  0.03069 *
## global_sentiment_polarity      4.864e-02  2.462e-01   0.198  0.84341
## global_rate_positive_words    -2.020e+00  1.075e+00  -1.880  0.06015 .
## global_rate_negative_words     1.136e+00  2.049e+00   0.554  0.57926
## rate_positive_words            5.024e-01  8.664e-01   0.580  0.56205
## rate_negative_words            4.291e-01  8.726e-01   0.492  0.62291
## avg_positive_polarity          3.674e-02  2.010e-01   0.183  0.85501
## min_positive_polarity         -1.375e-01  1.675e-01  -0.821  0.41172
## max_positive_polarity         -4.700e-02  6.459e-02  -0.728  0.46685
## avg_negative_polarity         -2.657e-01  1.863e-01  -1.426  0.15400
## min_negative_polarity         -2.057e-04  6.890e-02  -0.003  0.99762
## max_negative_polarity          1.413e-01  1.550e-01   0.912  0.36197
## title_subjectivity             8.297e-02  4.239e-02   1.958  0.05031 .
## title_sentiment_polarity       1.035e-01  3.856e-02   2.684  0.00728 **
## abs_title_subjectivity         1.382e-01  5.519e-02   2.503  0.01233 *
## abs_title_sentiment_polarity  -1.668e-02  6.034e-02  -0.276  0.78226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8607 on 9655 degrees of freedom
## Multiple R-squared:  0.1386, Adjusted R-squared:  0.1338
## F-statistic: 28.78 on 54 and 9655 DF,  p-value: < 2.2e-16
```

Table 9 – Full BIC model (10-fold cross validation)

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2491 -0.5474 -0.1708  0.3957  5.2905
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   6.252e+00  7.217e-02  86.623  < 2e-16 ***
## n_tokens_title                5.226e-03  4.281e-03   1.221 0.222265
## n_tokens_content              8.047e-05  2.494e-05   3.226 0.001261 **
## num_hrefs                     3.264e-03  9.210e-04   3.544 0.000396 ***
## data_channel_is_entertainment -1.078e-01  2.501e-02  -4.309 1.66e-05 ***
## data_channel_is_tech          2.468e-01  2.520e-02   9.793  < 2e-16 ***
## kw_min_min                    1.156e-03  1.301e-04   8.890  < 2e-16 ***
## kw_min_avg                   -8.700e-05  1.048e-05  -8.302  < 2e-16 ***
## kw_max_avg                   -4.673e-05  3.013e-06 -15.508  < 2e-16 ***
## kw_avg_avg                    3.879e-04  1.656e-05  23.427  < 2e-16 ***
## weekday_is_tuesday           -1.577e-01  2.421e-02  -6.516 7.56e-11 ***
## weekday_is_wednesday         -1.800e-01  2.457e-02  -7.326 2.57e-13 ***
## weekday_is_thursday          -1.670e-01  2.481e-02  -6.733 1.76e-11 ***
## LDA_00                        1.846e-01  3.614e-02   5.109 3.30e-07 ***
## global_subjectivity           4.152e-01  1.089e-01   3.811 0.000139 ***
## avg_negative_polarity        -1.710e-01  1.172e-01  -1.459 0.144487
## min_negative_polarity         2.003e-03  5.464e-02   0.037 0.970752
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8729 on 9693 degrees of freedom
## Multiple R-squared:  0.1107, Adjusted R-squared:  0.1092
## F-statistic: 75.38 on 16 and 9693 DF,  p-value: < 2.2e-16
```

Table 10 - Full AIC model (10-fold cross validation)

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1809 -0.5431 -0.1706  0.3848  5.3228
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    6.465e+00  7.704e-02  83.924  < 2e-16 ***
## n_tokens_title                 6.586e-03  4.257e-03   1.547  0.12192
## n_tokens_content               7.660e-05  2.528e-05   3.030  0.00245 **
## num_hrefs                      2.608e-03  9.182e-04   2.840  0.00452 **
## data_channel_is_entertainment -8.064e-02  2.501e-02  -3.224  0.00127 **
## data_channel_is_tech           2.642e-01  2.529e-02  10.447  < 2e-16 ***
## data_channel_is_socmed         2.696e-01  3.960e-02   6.807 1.05e-11 ***
## kw_min_min                     1.103e-03  1.297e-04   8.505  < 2e-16 ***
## kw_min_max                    -4.081e-07  1.659e-07  -2.460  0.01390 *
## kw_min_avg                    -8.266e-05  1.074e-05  -7.698 1.52e-14 ***
## kw_max_avg                    -4.534e-05  2.999e-06 -15.117  < 2e-16 ***
## kw_avg_avg                     3.753e-04  1.654e-05  22.691  < 2e-16 ***
## self_reference_min_shares      3.176e-06  4.799e-07   6.617 3.85e-11 ***
## weekday_is_monday             -2.481e-01  3.248e-02  -7.640 2.38e-14 ***
## weekday_is_tuesday            -3.328e-01  3.117e-02 -10.678  < 2e-16 ***
## weekday_is_wednesday          -3.521e-01  3.151e-02 -11.176  < 2e-16 ***
## weekday_is_thursday           -3.433e-01  3.165e-02 -10.847  < 2e-16 ***
## weekday_is_friday             -2.494e-01  3.321e-02  -7.508 6.54e-14 ***
## LDA_00                         1.750e-01  3.676e-02   4.761 1.96e-06 ***
## global_subjectivity            4.530e-01  1.150e-01   3.938 8.28e-05 ***
## min_positive_polarity         -3.724e-02  1.382e-01  -0.269  0.78762
## avg_negative_polarity         -1.602e-01  1.164e-01  -1.376  0.16875
## min_negative_polarity         -7.578e-03  5.441e-02  -0.139  0.88922
## title_sentiment_polarity       1.058e-01  3.379e-02   3.131  0.00174 **
## global_rate_positive_words    -1.709e+00  6.163e-01  -2.773  0.00556 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8649 on 9685 degrees of freedom
## Multiple R-squared:  0.1275, Adjusted R-squared:  0.1253
## F-statistic: 58.96 on 24 and 9685 DF,  p-value: < 2.2e-16
```

Table 11 – Polynomial AIC model (10-fold cross validation)

```
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3526 -0.5508 -0.1621  0.3934  5.3582
##
## Coefficients:
##                                               Estimate Std. Error t value
## (Intercept)                                  7.149e+00  2.427e+00   2.946
## `poly(n_tokens_title, 5, raw = TRUE)1`      -4.997e-01  1.184e+00  -0.422
## `poly(n_tokens_title, 5, raw = TRUE)2`       1.072e-01  2.339e-01   0.458
## `poly(n_tokens_title, 5, raw = TRUE)3`      -1.102e-02  2.234e-02  -0.493
## `poly(n_tokens_title, 5, raw = TRUE)4`       5.492e-04  1.032e-03   0.532
## `poly(n_tokens_title, 5, raw = TRUE)5`      -1.059e-05  1.849e-05  -0.573
## `poly(num_hrefs, 5, raw = TRUE)1`            6.264e-03  6.826e-03   0.918
## `poly(num_hrefs, 5, raw = TRUE)2`            8.318e-05  4.331e-04   0.192
## `poly(num_hrefs, 5, raw = TRUE)3`           -4.216e-06  9.982e-06  -0.422
## `poly(num_hrefs, 5, raw = TRUE)4`            4.909e-08  9.011e-08   0.545
## `poly(num_hrefs, 5, raw = TRUE)5`           -1.790e-10  2.710e-10  -0.661
## data_channel_is_tech                         2.387e-01  2.536e-02   9.411
## weekday_is_tuesday                          -1.540e-01  2.411e-02  -6.386
## weekday_is_wednesday                        -1.734e-01  2.449e-02  -7.080
## `poly(kw_min_min, 5, raw = TRUE)1`          -1.317e-03  4.551e-03  -0.289
## `poly(kw_min_min, 5, raw = TRUE)2`          -3.818e-04  3.577e-04  -1.067
## `poly(kw_min_min, 5, raw = TRUE)3`           7.101e-06  5.532e-06   1.283
## `poly(kw_min_min, 5, raw = TRUE)4`          -3.780e-08  2.805e-08  -1.348
## `poly(kw_min_min, 5, raw = TRUE)5`           6.186e-11  4.504e-11   1.373
## `poly(kw_min_avg, 5, raw = TRUE)1`          -1.158e-03  3.449e-04  -3.357
## `poly(kw_min_avg, 5, raw = TRUE)2`           1.339e-06  7.108e-07   1.884
## `poly(kw_min_avg, 5, raw = TRUE)3`          -6.313e-10  5.199e-10  -1.214
## `poly(kw_min_avg, 5, raw = TRUE)4`           1.369e-13  1.604e-13   0.853
## `poly(kw_min_avg, 5, raw = TRUE)5`          -1.146e-17  1.775e-17  -0.646
## `poly(kw_max_avg, 5, raw = TRUE)1`          -8.347e-06  1.686e-05  -0.495
## `poly(kw_max_avg, 5, raw = TRUE)2`          -1.409e-09  9.781e-10  -1.440
## `poly(kw_max_avg, 5, raw = TRUE)3`           3.575e-14  1.975e-14   1.810
## `poly(kw_max_avg, 5, raw = TRUE)4`          -2.874e-19  1.436e-19  -2.002
## `poly(kw_max_avg, 5, raw = TRUE)5`           5.913e-25  2.819e-25   2.098
## `poly(kw_avg_avg, 5, raw = TRUE)1`           3.575e-04  1.007e-04   3.549
## `poly(kw_avg_avg, 5, raw = TRUE)2`           1.013e-08  3.251e-08   0.312
## `poly(kw_avg_avg, 5, raw = TRUE)3`          -3.233e-12  4.428e-12  -0.730
## `poly(kw_avg_avg, 5, raw = TRUE)4`           1.205e-16  2.483e-16   0.485
## `poly(kw_avg_avg, 5, raw = TRUE)5`           9.471e-23  4.942e-21   0.019
## `poly(LDA_00, 5, raw = TRUE)1`              -2.803e+00  1.318e+00  -2.127
## `poly(LDA_00, 5, raw = TRUE)2`               2.259e+01  1.016e+01   2.224
## `poly(LDA_00, 5, raw = TRUE)3`              -6.253e+01  2.963e+01  -2.110
## `poly(LDA_00, 5, raw = TRUE)4`               7.234e+01  3.658e+01   1.977
## `poly(LDA_00, 5, raw = TRUE)5`              -2.972e+01  1.610e+01  -1.846
## `poly(global_subjectivity, 5, raw = TRUE)1`  8.474e-01  8.355e+00   0.101
## `poly(global_subjectivity, 5, raw = TRUE)2`  2.107e+00  3.853e+01   0.055
## `poly(global_subjectivity, 5, raw = TRUE)3` -1.269e+01  8.509e+01  -0.149
## `poly(global_subjectivity, 5, raw = TRUE)4`  2.034e+01  8.937e+01   0.228
## `poly(global_subjectivity, 5, raw = TRUE)5` -1.041e+01  3.568e+01  -0.292
## data_channel_is_entertainment               -9.712e-02  2.544e-02  -3.818
## weekday_is_thursday                         -1.617e-01  2.473e-02  -6.539
##                                             Pr(>|t|)
## (Intercept)                                 0.003230 **
## `poly(n_tokens_title, 5, raw = TRUE)1`      0.672959
## `poly(n_tokens_title, 5, raw = TRUE)2`      0.646632
## `poly(n_tokens_title, 5, raw = TRUE)3`      0.621769
## `poly(n_tokens_title, 5, raw = TRUE)4`      0.594789
## `poly(n_tokens_title, 5, raw = TRUE)5`      0.566891
## `poly(num_hrefs, 5, raw = TRUE)1`           0.358842
## `poly(num_hrefs, 5, raw = TRUE)2`           0.847695
## `poly(num_hrefs, 5, raw = TRUE)3`           0.672751
## `poly(num_hrefs, 5, raw = TRUE)4`           0.585931
## `poly(num_hrefs, 5, raw = TRUE)5`           0.508863
## data_channel_is_tech                         < 2e-16 ***
## weekday_is_tuesday                          1.79e-10 ***
## weekday_is_wednesday                        1.54e-12 ***
## `poly(kw_min_min, 5, raw = TRUE)1`          0.772359
## `poly(kw_min_min, 5, raw = TRUE)2`          0.285816
## `poly(kw_min_min, 5, raw = TRUE)3`          0.199351
## `poly(kw_min_min, 5, raw = TRUE)4`          0.177774
## `poly(kw_min_min, 5, raw = TRUE)5`          0.169689    
## `poly(kw_min_avg, 5, raw = TRUE)1`          0.000790 ***
## `poly(kw_min_avg, 5, raw = TRUE)2`          0.059562 .  
## `poly(kw_min_avg, 5, raw = TRUE)3`          0.224636    
## `poly(kw_min_avg, 5, raw = TRUE)4`          0.393416    
## `poly(kw_min_avg, 5, raw = TRUE)5`          0.518490    
## `poly(kw_max_avg, 5, raw = TRUE)1`          0.620590    
## `poly(kw_max_avg, 5, raw = TRUE)2`          0.149789    
## `poly(kw_max_avg, 5, raw = TRUE)3`          0.070312 .  
## `poly(kw_max_avg, 5, raw = TRUE)4`          0.045317 *  
## `poly(kw_max_avg, 5, raw = TRUE)5`          0.035938 *  
## `poly(kw_avg_avg, 5, raw = TRUE)1`          0.000388 ***
## `poly(kw_avg_avg, 5, raw = TRUE)2`          0.755375    
## `poly(kw_avg_avg, 5, raw = TRUE)3`          0.465316    
## `poly(kw_avg_avg, 5, raw = TRUE)4`          0.627439    
## `poly(kw_avg_avg, 5, raw = TRUE)5`          0.984711    
## `poly(LDA_00, 5, raw = TRUE)1`              0.033471 *  
## `poly(LDA_00, 5, raw = TRUE)2`              0.026173 *  
## `poly(LDA_00, 5, raw = TRUE)3`              0.034875 *  
## `poly(LDA_00, 5, raw = TRUE)4`              0.048039 *  
## `poly(LDA_00, 5, raw = TRUE)5`              0.064971 .  
## `poly(global_subjectivity, 5, raw = TRUE)1` 0.919217    
## `poly(global_subjectivity, 5, raw = TRUE)2` 0.956391    
## `poly(global_subjectivity, 5, raw = TRUE)3` 0.881476    
## `poly(global_subjectivity, 5, raw = TRUE)4` 0.820000    
## `poly(global_subjectivity, 5, raw = TRUE)5` 0.770543    
## data_channel_is_entertainment               0.000135 ***
## weekday_is_thursday                         6.53e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.8677 on 9664 degrees of freedom
## Multiple R-squared:  0.1239, Adjusted R-squared:  0.1198 
## F-statistic: 30.36 on 45 and 9664 DF,  p-value: < 2.2e-16
```

Table 12 – Polynomial BIC model (10-fold cross validation)

```
## 
## Call:
## lm(formula = .outcome ~ ., data = dat)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.8180 -0.5820 -0.1962  0.4113  5.3164 
## 
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         7.561e+00  1.874e-02 403.469  < 2e-16
## `poly(kw_min_avg, 5, raw = TRUE)1` -1.112e-03  3.564e-04  -3.122   0.0018
## `poly(kw_min_avg, 5, raw = TRUE)2`  1.261e-06  7.338e-07   1.719   0.0856
## `poly(kw_min_avg, 5, raw = TRUE)3` -5.203e-10  5.362e-10  -0.971   0.3318
## `poly(kw_min_avg, 5, raw = TRUE)4`  9.856e-14  1.653e-13   0.596   0.5510
## `poly(kw_min_avg, 5, raw = TRUE)5` -6.645e-18  1.826e-17  -0.364   0.7160
## data_channel_is_entertainment      -1.321e-01  2.446e-02  -5.401 6.79e-08
## data_channel_is_tech                1.159e-01  2.422e-02   4.783 1.75e-06
## weekday_is_tuesday                 -1.676e-01  2.495e-02  -6.719 1.94e-11
## weekday_is_wednesday               -1.860e-01  2.532e-02  -7.346 2.21e-13
## weekday_is_thursday                -1.694e-01  2.557e-02  -6.627 3.61e-11
## 
```

```
## (Intercept)                        ***
## `poly(kw_min_avg, 5, raw = TRUE)1` **
## `poly(kw_min_avg, 5, raw = TRUE)2` .
## `poly(kw_min_avg, 5, raw = TRUE)3`
## `poly(kw_min_avg, 5, raw = TRUE)4`
## `poly(kw_min_avg, 5, raw = TRUE)5`
## data_channel_is_entertainment      ***
## data_channel_is_tech               ***
## weekday_is_tuesday                 ***
## weekday_is_wednesday               ***
## weekday_is_thursday                ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9001 on 9699 degrees of freedom
## Multiple R-squared:  0.0537, Adjusted R-squared:  0.05273
## F-statistic: 55.04 on 10 and 9699 DF, p-value: < 2.2e-16
```

Table 13 – Full model (test data)

```
## Call:
## lm(formula = shares ~ ., data = testOnlineNews)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1921 -0.5404 -0.1681  0.3860  5.2097
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    6.906e+00  8.925e-01   7.738 1.11e-14 ***
## n_tokens_title                 9.145e-03  4.351e-03   2.102  0.03560 *
## n_tokens_content               6.474e-05  3.448e-05   1.878  0.06044 .
## n_unique_tokens                5.116e-01  2.927e-01   1.748  0.08053 .
## n_non_stop_unique_tokens      -6.756e-01  2.485e-01  -2.718  0.00657 **
## num_hrefs                      2.221e-03  1.028e-03   2.161  0.03071 *
## num_self_hrefs                -8.822e-03  2.838e-03  -3.109  0.00188 **
## num_imgs                       2.459e-03  1.404e-03   1.752  0.07984 .
## num_videos                     1.434e-03  2.448e-03   0.586  0.55798
## average_token_length          -8.630e-02  3.632e-02  -2.376  0.01751 *
## num_keywords                  -5.212e-03  5.604e-03  -0.930  0.35233
## data_channel_is_lifestyle     -1.250e-01  5.932e-02  -2.107  0.03512 *
## data_channel_is_entertainment -1.738e-01  3.908e-02  -4.447 8.82e-06 ***
## data_channel_is_bus           -1.556e-01  5.753e-02  -2.705  0.00685 **
## data_channel_is_socmed         1.458e-01  5.661e-02   2.576  0.01001 *
## data_channel_is_tech           9.045e-02  5.619e-02   1.610  0.10751
## data_channel_is_world         -4.405e-02  5.694e-02  -0.774  0.43923
## kw_min_min                     1.066e-03  2.376e-04   4.485 7.38e-06 ***
## kw_max_min                     2.595e-05  7.761e-06   3.344  0.00083 ***
## kw_avg_min                    -2.241e-04  5.533e-05  -4.051 5.15e-05 ***
## kw_min_max                    -5.413e-09  1.813e-07  -0.030  0.97618
## kw_max_max                     1.751e-07  8.509e-08   2.058  0.03965 *
## kw_avg_max                    -7.802e-07  1.286e-07  -6.064 1.38e-09 ***
## kw_min_avg                    -7.865e-05  1.178e-05  -6.675 2.61e-11 ***
## kw_max_avg                    -4.716e-05  4.293e-06 -10.986  < 2e-16 ***
## kw_avg_avg                     3.924e-04  2.295e-05  17.096  < 2e-16 ***
## self_reference_min_shares      2.959e-06  1.116e-06   2.651  0.00804 **
## self_reference_max_shares     -2.662e-08  5.969e-07  -0.045  0.96443
## self_reference_avg_sharess     1.189e-07  1.517e-06   0.078  0.93755
## weekday_is_monday             -2.400e-01  3.876e-02  -6.193 6.13e-10 ***
## weekday_is_tuesday            -3.289e-01  3.773e-02  -8.716  < 2e-16 ***
## weekday_is_wednesday          -3.418e-01  3.797e-02  -9.003  < 2e-16 ***
## weekday_is_thursday           -3.370e-01  3.810e-02  -8.844  < 2e-16 ***
## weekday_is_friday             -2.433e-01  3.937e-02  -6.180 6.68e-10 ***
## weekday_is_saturday           -1.133e-02  4.772e-02  -0.237  0.81236
## LDA_00                         1.935e-01  6.768e-02   2.859  0.00426 **
## LDA_01                        -1.716e-01  7.585e-02  -2.262  0.02373 *
## LDA_02                        -2.680e-01  6.865e-02  -3.905 9.50e-05 ***
## LDA_03                        -1.595e-01  7.237e-02  -2.204  0.02756 *
## global_subjectivity            2.720e-01  1.258e-01   2.161  0.03069 *
## global_sentiment_polarity      4.864e-02  2.462e-01   0.198  0.84341
## global_rate_positive_words    -2.020e+00  1.075e+00  -1.880  0.06015 .
## global_rate_negative_words     1.136e+00  2.049e+00   0.554  0.57926
## rate_positive_words            5.024e-01  8.664e-01   0.580  0.56205
## rate_negative_words            4.291e-01  8.726e-01   0.492  0.62291
## avg_positive_polarity          3.674e-02  2.010e-01   0.183  0.85501
## min_positive_polarity         -1.375e-01  1.675e-01  -0.821  0.41172
## max_positive_polarity         -4.700e-02  6.459e-02  -0.728  0.46685
## avg_negative_polarity         -2.657e-01  1.863e-01  -1.426  0.15400
## min_negative_polarity         -2.057e-04  6.890e-02  -0.003  0.99762
## max_negative_polarity          1.413e-01  1.550e-01   0.912  0.36197
## title_subjectivity             8.297e-02  4.239e-02   1.958  0.05031 .
## title_sentiment_polarity       1.035e-01  3.856e-02   2.684  0.00728 **
## abs_title_subjectivity         1.382e-01  5.519e-02   2.503  0.01233 *
## abs_title_sentiment_polarity  -1.668e-02  6.034e-02  -0.276  0.78226
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 0.8607 on 9655 degrees of freedom
## Multiple R-squared:  0.1386, Adjusted R-squared:  0.1338
## F-statistic: 28.78 on 54 and 9655 DF,  p-value: < 2.2e-16
```

Table 14-Full AIC model (test data)

```
## Call:
## lm(formula = shares ~ n_tokens_title + n_tokens_content + n_unique_tokens+
##     n_non_stop_unique_tokens + num_hrefs + num_self_hrefs + num_imgs +
##     average_token_length + data_channel_is_lifestyle +
data_channel_is_entertainment +
##     data_channel_is_bus + data_channel_is_socmed + data_channel_is_tech +
##     kw_min_min + kw_max_min + kw_avg_min + kw_max_max + kw_avg_max +
##     kw_min_avg + kw_max_avg + kw_avg_avg + self_reference_min_shares +
##     weekday_is_monday + weekday_is_tuesday + weekday_is_wednesday +
##     weekday_is_thursday + weekday_is_friday + LDA_00 + LDA_01 +
##     LDA_02 + LDA_03 + global_subjectivity + global_rate_positive_words +
##     avg_negative_polarity + title_subjectivity + title_sentiment_polarity
+
##     abs_title_subjectivity, data = testOnlineNews)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1756 -0.5405 -0.1682  0.3838  5.2190
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    7.285e+00  2.338e-01  31.157  < 2e-16 ***
## n_tokens_title                 9.008e-03  4.338e-03   2.077 0.037865 *
## n_tokens_content               6.922e-05  3.266e-05   2.119 0.034083 *
## n_unique_tokens                4.610e-01  2.699e-01   1.708 0.087595 .
## n_non_stop_unique_tokens      -6.389e-01  2.362e-01  -2.705 0.006844 **
## num_hrefs                      2.256e-03  1.012e-03   2.229 0.025851 *
## num_self_hrefs                -8.689e-03  2.759e-03  -3.150 0.001640 **
## num_imgs                       2.331e-03  1.366e-03   1.706 0.088015 .
## average_token_length          -8.772e-02  3.601e-02  -2.436 0.014856 *
## data_channel_is_lifestyle     -1.014e-01  4.993e-02  -2.031 0.042304 *
## data_channel_is_entertainment -1.531e-01  3.404e-02  -4.496 6.99e-06 ***
## data_channel_is_bus           -1.339e-01  4.764e-02  -2.810 0.004968 **
## data_channel_is_socmed         1.749e-01  4.650e-02   3.763 0.000169 ***
## data_channel_is_tech           1.175e-01  4.281e-02   2.743 0.006093 **
## kw_min_min                     1.080e-03  2.371e-04   4.557 5.25e-06 ***
## kw_max_min                     2.539e-05  7.687e-06   3.303 0.000959 ***
## kw_avg_min                    -2.197e-04  5.499e-05  -3.994 6.53e-05 ***
## kw_max_max                     1.555e-07  8.286e-08   1.877 0.060535 .
## kw_avg_max                    -7.130e-07  1.084e-07  -6.580 4.94e-11 ***
## kw_min_avg                    -7.727e-05  1.116e-05  -6.921 4.75e-12 ***
## kw_max_avg                    -4.705e-05  4.150e-06 -11.339  < 2e-16 ***
## kw_avg_avg                     3.912e-04  2.220e-05  17.620  < 2e-16 ***
## self_reference_min_shares      3.053e-06  4.779e-07   6.388 1.76e-10 ***
## weekday_is_monday             -2.333e-01  3.239e-02  -7.202 6.38e-13 ***
## weekday_is_tuesday            -3.213e-01  3.109e-02 -10.334  < 2e-16 ***
## weekday_is_wednesday          -3.344e-01  3.142e-02 -10.641  < 2e-16 ***
## weekday_is_thursday           -3.301e-01  3.155e-02 -10.461  < 2e-16 ***
## weekday_is_friday             -2.367e-01  3.311e-02  -7.150 9.32e-13 ***
## LDA_00                         2.028e-01  6.729e-02   3.014 0.002588 **
## LDA_01                        -1.489e-01  7.198e-02  -2.069 0.038611 *
## LDA_02                        -2.778e-01  6.504e-02  -4.271 1.96e-05 ***
## LDA_03                        -1.326e-01  6.621e-02  -2.003 0.045208 *
## global_subjectivity            2.843e-01  1.159e-01   2.454 0.014161 *
## global_rate_positive_words    -1.547e+00  6.108e-01  -2.533 0.011317 *
## avg_negative_polarity         -1.890e-01  7.872e-02  -2.401 0.016357 *
## title_subjectivity             7.764e-02  3.280e-02   2.367 0.017935 *
## title_sentiment_polarity       9.433e-02  3.484e-02   2.707 0.006797 **
## abs_title_subjectivity         1.369e-01  5.488e-02   2.495 0.012628 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 0.8602 on 9672 degrees of freedom
## Multiple R-squared:  0.1382, Adjusted R-squared:  0.1349
## F-statistic: 41.92 on 37 and 9672 DF,  p-value: < 2.2e-16
```

Table 15-Full BIC model (test data)

```
## Call:
## lm(formula = shares ~ n_non_stop_unique_tokens +
data_channel_is_entertainment +
##     data_channel_is_socmed + data_channel_is_tech + kw_min_min +
##     kw_max_min + kw_avg_min + kw_avg_max + kw_min_avg + kw_max_avg +
##     kw_avg_avg + self_reference_min_shares + weekday_is_monday +
##     weekday_is_tuesday + weekday_is_wednesday + weekday_is_thursday +
##     weekday_is_friday + LDA_00 + LDA_02 + global_subjectivity,
##     data = testOnlineNews)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2145 -0.5387 -0.1687  0.3785  5.2998
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   7.217e+00  9.519e-02  75.817  < 2e-16 ***
## n_non_stop_unique_tokens     -5.787e-01  8.804e-02  -6.573 5.18e-11 ***
## data_channel_is_entertainment -1.377e-01  2.753e-02  -5.002 5.78e-07 ***
## data_channel_is_socmed        2.258e-01  3.954e-02   5.709 1.17e-08 ***
## data_channel_is_tech          1.806e-01  2.820e-02   6.405 1.57e-10 ***
## kw_min_min                    7.246e-04  1.521e-04   4.764 1.92e-06 ***
## kw_max_min                    2.939e-05  7.631e-06   3.852 0.000118 ***
## kw_avg_min                   -2.527e-04  5.446e-05  -4.641 3.52e-06 ***
## kw_avg_max                   -6.772e-07  9.869e-08  -6.861 7.24e-12 ***
## kw_min_avg                   -8.508e-05  1.085e-05  -7.842 4.91e-15 ***
## kw_max_avg                   -4.948e-05  4.033e-06 -12.270  < 2e-16 ***
## kw_avg_avg                    4.120e-04  2.066e-05  19.944  < 2e-16 ***
## self_reference_min_shares     3.155e-06  4.782e-07   6.598 4.40e-11 ***
## weekday_is_monday            -2.396e-01  3.239e-02  -7.396 1.52e-13 ***
## weekday_is_tuesday           -3.256e-01  3.108e-02 -10.476  < 2e-16 ***
## weekday_is_wednesday         -3.434e-01  3.138e-02 -10.945  < 2e-16 ***
## weekday_is_thursday          -3.374e-01  3.154e-02 -10.698  < 2e-16 ***
## weekday_is_friday            -2.380e-01  3.310e-02  -7.191 6.91e-13 ***
## LDA_00                        1.396e-01  4.262e-02   3.275 0.001060 **
## LDA_02                       -1.465e-01  4.366e-02  -3.355 0.000797 ***
## global_subjectivity           3.781e-01  1.058e-01   3.573 0.000354 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 0.8627 on 9689 degrees of freedom
## Multiple R-squared:  0.1317, Adjusted R-squared:  0.1299
## F-statistic: 73.46 on 20 and 9689 DF, p-value: < 2.2e-16
```

8. APPENDIX II (R-CODE)

# QUESTION 1: DATA EXPLORATION

```
 # Import Data

onlineNews <- read.csv2("alldata_onlinenews_33.csv", header = TRUE)

# Data pre-processing

        # Cleaning
```

```
     trimws(names(onlineNews), which = c("both", "left", "right")
     if (sum(onlineNews$n_unique_tokens)==0) {
     onlineNews$n_unique_tokens<-NULL
      }
     if (sum(onlineNews$n_non_stop_words)==0) {
     onlineNews$n_non_stop_words<-NULL
       }

     if (sum(onlineNews$n_non_stop_unique_tokens)==0) {
      onlineNews$n_non_stop_unique_tokens<-NULL
     }

      # Preparing
              # Duplicating days of week
              OnlineNews$is_weekend = NULL
               # URL and timedelta columns are meta-data and cannot be treated as features
               onlineNews$url<-NULL
              onlineNews$timedelta <- NULL
              # The X column and the articles with no words included have no meaning
               onlineNews$X<-NULL
               onlineNews<-onlineNews[!onlineNews$n_tokens_content==0,]
       # Checking for null/NaN/inf values
           for (i in 1:ncol(onlineNews))
           {
            na <- is.na(onlineNews[,i])
            inf <- is.infinite(onlineNews[,i])
            nan <- is.nan(onlineNews[,i])
           }
           any(na)
           any(nan)
           any(inf)

 str(onlineNews)
# Factor variables
onlineNews$weekday_is_monday <- factor(onlineNews$weekday_is_monday)
```

```
onlineNews$weekday_is_wednesday <- factor(onlineNews$weekday_is_wednesday)
onlineNews$weekday_is_thursday <- factor(onlineNews$weekday_is_thursday)
onlineNews$weekday_is_friday <- factor(onlineNews$weekday_is_friday)
onlineNews$weekday_is_tuesday <- factor(onlineNews$weekday_is_tuesday)
onlineNews$weekday_is_saturday <- factor(onlineNews$weekday_is_saturday)
onlineNews$weekday_is_sunday <- factor(onlineNews$weekday_is_sunday)
onlineNews$data_channel_is_lifestyle <- factor(onlineNews$data_channel_is_lifestyle)
onlineNews$data_channel_is_entertainment <- factor(onlineNews$data_channel_is_entertainment)
onlineNews$data_channel_is_bus <- factor(onlineNews$data_channel_is_bus)
onlineNews$data_channel_is_socmed <- factor(onlineNews$data_channel_is_socmed)
onlineNews$data_channel_is_tech <- factor(onlineNews$data_channel_is_tech)
onlineNews$data_channel_is_world <- factor(onlineNews$data_channel_is_world
summary(onlineNews)
    # Create numeric variable dataset
    library(psych)
    index <- sapply(onlineNews, class) !='factor'
    newsNumeric <- onlineNews[,index]
    round(t(describe(newsNumeric <- onlineNews[,index])),2)
    n <- nrow(newsNumeric)
   # Create factor variable dataset
   newsFactors <- onlineNews[,!index]
 # Visualization of numeric variables
   par(mfrow=c(2,4))
   for (i in c(1,2,3,4,5,6,8,9))
  {
   h1 <- hist(newsNumeric[,i], main=names(newsNumeric)[i], border='pink', col='purple')
  }
  par(mfrow=c(2,4))
  for (i in c(24,25,26,27,28,41,42,45))
  {
   h2 <- hist(newsNumeric[,i], main=names(newsNumeric)[i], border='pink', col='purple')
  }
  par(mfrow=c(2,3))
  for (i in c(29,30,31,32,33,34))
  {
```

```
                  h3 <- hist(newsNumeric[,i], main=names(newsNumeric)[i], border='pink', col='purple')

                 }

                # Visualization of factor variables

                barplot(sapply(newsFactors[,c(1:6)],table)/n, names.arg = c("Lfst", "Entrt", "Bus", "Soc", "Tech",
           "World"), horiz=T, las=1, col=2:3, ylim=c(0,9), cex.names=1.3)

                legend('topleft', fil=2:3, legend=c('No','Yes'), ncol=2, bty='n',cex=1.5)

               barplot(sapply(newsFactors[,c(7:13)],table)/n, names.arg = c("Mon", "Tue", "Wed", "Thu", "Fri",
           "Sat", "Sun"), horiz=T, las=1, col=2:3, ylim=c(0,11), cex.names=1.3)
           legend('topleft', fil=2:3, legend=c('No','Yes'), ncol=2, bty='n',cex=1.5)


# Data exploration of shares

        # Outliers for shares exist

        outlier_values_shares <- boxplot.stats(newsNumeric[,45])$out  # outlier values

        print("Outliers of shares")

        print(length(outlier_values_shares))

        boxplot(newsNumeric[,45], main=names(newsNumeric)[45], boxwex=0.1)

        # Distribution of shares

        onlineNews$shares <- log(onlineNews$shares)

        newsNumeric$shares<- log(newsNumeric$shares)

        hist(onlineNews$shares, freq = FALSE, col="blue", main = "Distribution of log_shares", xlab =
      "Log_shares")

        curve(dnorm(x, mean = mean(onlineNews$shares), sd = sd(onlineNews$shares)), col = 2, lty = 1, lwd = 2,
      add = TRUE)


 # Pairs of shares and other numerical variables

 pairs(newsNumeric[,c(45,1,2,3,5,6,8,9)])

 pairs(newsNumeric[,c(45,24,25,26,27,28)])

 pairs(newsNumeric[,c(45,29,30,31,32,33,43,42)])

 # Shares on each factor variable

 par(mfrow=c(2,3))

 for(j in 1:6){

   boxplot(newsNumeric[,1]~newsFactors[,j], xlab=names(newsFactors)[j], ylab='Shares',cex.lab=1.5)

   abline(lm(newsNumeric[,1]~newsFactors[,j]),col=2)

 }

 par(mfrow=c(2,4))

 for(j in 7:13){

   boxplot(newsNumeric[,1]~newsFactors[,j], xlab=names(newsFactors)[j], ylab='Shares',cex.lab=1.5)
```

```
    abline(lm(newsNumeric[,1]~newsFactors[,j]),col=2)
  }
  # Correlation visualization of shares and calculations
  col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))
  library(corrplot)
  corrplot(cor(newsNumeric[,c(45,1,2,3,5,6,8,9)]), method= "color", col = col(200),
        type = "upper", order = "hclust", number.cex = .7,
        addCoef.col = "black", # Add coefficient of correlation
        tl.col = "black", tl.srt = 90, # Text label color and rotation
        # Combine with significance
        sig.level = 0.05, insig = "blank",
        # hide correlation coefficient on the principal diagonal
        diag = FALSE)
  corrplot(cor(newsNumeric[,c(45,24,25,26,27,28)]), method= "color", col = col(200),
        type = "upper", order = "hclust", number.cex = .7,
        addCoef.col = "black", # Add coefficient of correlation
        tl.col = "black", tl.srt = 90, # Text label color and rotation
        # Combine with significance
        sig.level = 0.05, insig = "blank",
        # hide correlation coefficient on the principal diagonal
        diag = FALSE)
  corrplot(cor(newsNumeric[,c(45,29,30,31,32,33,43,42)]), method= "color", col = col(200),
        type = "upper", order = "hclust", number.cex = .7,
        addCoef.col = "black", # Add coefficient of correlation
        tl.col = "black", tl.srt = 90, # Text label color and rotation
        # Combine with significance
        sig.level = 0.05, insig = "blank",
        # hide correlation coefficient on the principal diagonal
        diag = FALSE)

 # QUESTION 2: BEST PREDICTION MODEL
# Initial regression model
fit_full <- lm(shares ~ ., data = onlineNews)
summary(fit_full)
      # Remove the variables that have multicolinearity
```

```
        onlineNews$weekday_is_sunday<-NULL
        onlineNews$LDA_04<-NULL
        onlineNews$n_non_stop_words <-NULL
# Second regression model
fit_full2 <- lm(shares ~ ., data = onlineNews)
summary(fit_full2)
# Final regression model
 # AIC method
 selected_aic_model <- step(fit_full2, direction='both')
 summary(selected_aic_model)
 length(coef(selected_aic_model))
 # BIC method
 nRegres = length(resid(fit_full))
 selected_bic_model = step(fit_full2, direction = "both", k = log(nRegres))
 summary(selected_bic_model)
 length(coef(selected_bic_model))


# QUESTION 3: REGRESSION HYPOTHESIS
# Residual Analysis and Diagnostics
  par(mfrow=c(2,4))
  # Normality of residuals
   # Visualization
   plot(selected_bic_model, which=2, col=c("red"))  # Q-Q Plot
   hist(selected_bic_model$residuals,probability = T, main = 'Residuals')
   x0<- seq (min(selected_bic_model$residuals), max(selected_bic_model$residuals), length.out = 100)
   y0 <- dnorm(x0, mean(selected_bic_model$residuals), sd(selected_bic_model$residuals))
   lines(x0,y0, col=2,lty=2)
   # Testing normality
   install.packages("nortest")
   library(nortest)
   lillie.test(selected_bic_model$res)
   shapiro.test(selected_bic_model$res)
# Data linearity
   # Visualization
   plot(selected_bic_model, which=1, col=c("blue")) # Residuals vs Fitted Plot
```

```
    # Testing linearity
    library(car)
    residualPlots(selected_bic_model, plot=F, type = "rstudent")
# Homoscedasticity of residuals variance
    # Visualization
    plot(selected_bic_model, which=3, col=c("blue"))  # Scale-Location Plot
    studentResiduals <- rstudent(selected_bic_model)
    yhat <- fitted(selected_bic_model)
    plot(yhat, studentResiduals, main = "Residuals",col=c("blue"))
    abline(h=c(-2,2), col=2, lty=2)
    plot(yhat, studentResiduals^2, main = "Residuals",col=c("blue"))
    abline(h=4, col=2, lty=2)
    # Testing homoscedasticity
    qyhat.quantiles <- cut(yhat, breaks=quantile(yhat,probs = seq(0,1,0.25), dig.lab=6))
    library(car)
    leveneTest(rstudent(selected_bic_model)~qyhat.quantiles)
    ncvTest(selected_bic_model)
# Outliers and high levarage points
    # Visualization
    plot(selected_bic_model, which=4, col=c("blue"))
    plot(selected_aic_model, which=5, col=c("blue"))
    leveragePlots(selected_bic_model,col="blue")
# Multicollinearity
    library(car)
    round(vif(selected_bic_model),2)


# Fixing problems of assumptions
 # (1) Polynominals
 polynominal <- lm(shares ~ poly(n_tokens_title,5, raw=TRUE)
                 + poly(n_tokens_content,5, raw=TRUE)
                 + poly(num_hrefs, 5, raw=TRUE)
                 + poly(kw_min_min,5, raw=TRUE)
                 + poly(kw_min_avg,5, raw=TRUE)
                 + poly(kw_max_avg,5, raw=TRUE)
                 + poly(kw_avg_avg,5, raw=TRUE)
```

```
            + poly(LDA_00,5, raw=TRUE)
            + poly(global_subjectivity,5, raw=TRUE)
            + poly(avg_negative_polarity,5, raw=TRUE)
            + poly(min_negative_polarity,5, raw=TRUE)
            + data_channel_is_entertainment
            + data_channel_is_tech
            + weekday_is_tuesday
            + weekday_is_wednesday
            + weekday_is_thursday,
            data = onlineNews)
summary(polynominal)
length(coef(polynominal))



# AIC method
polyn_selected_aic_model <- step(polynominal, direction='both')
summary(polyn_selected_aic_model)
length(coef(polyn_selected_aic_model))
# BIC method
polyn_nRegres = length(resid(polynominal))
poly_selected_bic_model = step(polynominal, direction = "both", k = log(polyn_nRegres))
summary(poly_selected_bic_model)
length(coef(poly_selected_bic_model))
 # (2) Logarithms
 logarithms <- lm(shares ~ log(n_tokens_title)
            + log(n_tokens_content)
            + num_hrefs
            + kw_min_min
            + kw_min_avg
            + kw_max_avg
            + kw_avg_avg
            + log(LDA_00)
            + global_subjectivity
            + avg_negative_polarity
            + min_negative_polarity
```

```
                + data_channel_is_entertainment
                + data_channel_is_tech
                + weekday_is_tuesday
                + weekday_is_wednesday
                + weekday_is_thursday,
                data = onlineNews)
summary(logarithms)
length(coef(logarithms))
# AIC method
log_selected_aic_model <- step(logarithms, direction='both')
summary(log_selected_aic_model)
length(coef(log_selected_aic_model))
# BIC method
log_nRegres = length(resid(logarithms))
log_selected_bic_model = step(logarithms, direction = "both", k = log(log_nRegres))
summary(log_selected_bic_model)
length(coef(log_selected_bic_model))
```

# QUESTION 4: 10-FOLD CROSS VALIDATION

```
testOnlineNews <- read.csv2("OnlineNewsPopularity_test.csv", header = TRUE)
# Data pre-processing
  # Cleaning
  trimws(names(testOnlineNews), which = c("both", "left", "right"))

  if (sum(testOnlineNews$n_unique_tokens)==0) {
    testOnlineNews$n_unique_tokens<-NULL
  }

  if (sum(testOnlineNews$n_non_stop_words)==0) {
    testOnlineNews$n_non_stop_words<-NULL
  }

  if (sum(testOnlineNews$n_non_stop_unique_tokens)==0) {
    testOnlineNews$n_non_stop_unique_tokens<-NULL
  }
```

```
# Preparing
   # Duplicating days of week
   testOnlineNews$is_weekend = NULL
   # URL and timedelta columns are are meta-data and cannot be treated as features
   testOnlineNews$url<-NULL
   testOnlineNews$timedelta <- NULL
   # The X column and the articles with no words included
   testOnlineNews$X<-NULL
   testOnlineNews<-testOnlineNews[!testOnlineNews$n_tokens_content==0,]
   # Checking for null/NaN values
   for (i in 1:ncol(testOnlineNews))
   {
     na1 <- is.na(testOnlineNews[,i])
     inf1 <- is.infinite(testOnlineNews[,i])
     nan1 <- is.nan(testOnlineNews[,i])
   }
   any(na1)
   any(nan1)
   any(inf1)

   # Converting shares to log_shares
   testOnlineNews$shares <- log(testOnlineNews$shares)
  str(testOnlineNews)
  wilcox.test(onlineNews[,55],testOnlineNews[,55])


  # Full model
  library("caret")
  model_full <- train(
  shares ~., testOnlineNews,
  method = "lm", trControl = trainControl(
  method = "cv",
  number = 10,
  verboseIter = TRUE))


  summary(model_full)
```

```
     # Remove columns with multicollinearity problem
      testOnlineNews$n_non_stop_words <-NULL
      testOnlineNews$weekday_is_sunday <- NULL
      testOnlineNews$LDA_04 <- NULL
 library("caret")
 model_full_new <- train(
 shares ~., testOnlineNews,
 method = "lm", trControl = trainControl(
 method = "cv",
  number = 10,
 verboseIter = TRUE))
 summary(model_full_new)

 # Full BIC model
 library("caret")
 model_bic_full <- train(
shares ~ n_tokens_title + n_tokens_content + num_hrefs + data_channel_is_entertainment
+ data_channel_is_tech + kw_min_min + kw_min_avg + kw_max_avg + kw_avg_avg + weekday_is_tuesday
+ weekday_is_wednesday + weekday_is_thursday + LDA_00 + global_subjectivity + avg_negative_polarity
+ min_negative_polarity, testOnlineNews,
  method = "lm", trControl = trainControl(
  method = "cv",
  number = 10,
  verboseIter = TRUE))
 summary(model_bic_full)

 # Full AIC model
 library("caret")
 model_aic_full <- train(
shares ~ n_tokens_title + n_tokens_content + num_hrefs + data_channel_is_entertainment
+ data_channel_is_tech + data_channel_is_socmed + data_channel_is_tech + kw_min_min
+ kw_min_max + kw_min_avg + kw_max_avg + kw_avg_avg + self_reference_min_shares
+ weekday_is_monday + weekday_is_tuesday + weekday_is_wednesday + weekday_is_thursday
+ weekday_is_friday + LDA_00 + global_subjectivity + min_positive_polarity + avg_negative_polarity
+ min_negative_polarity + title_sentiment_polarity + global_rate_positive_words, testOnlineNews,
  method = "lm", trControl = trainControl(
  method = "cv",
  number = 10,
  verboseIter = TRUE))
 summary(model_aic_full)

 # AIC polynomial model
 library("caret")
 model_aic_polyn <- train(
 shares ~ poly(n_tokens_title, 5, raw = TRUE) + poly(num_hrefs, 5, raw = TRUE)
   + data_channel_is_tech + weekday_is_tuesday + weekday_is_wednesday
   + poly(kw_min_min, 5, raw = TRUE) + poly(kw_min_avg, 5, raw = TRUE)
   + poly(kw_max_avg, 5, raw = TRUE) + poly(kw_avg_avg, 5, raw = TRUE) + poly(LDA_00, 5, raw = TRUE)
   + poly(global_subjectivity, 5, raw = TRUE)
   + data_channel_is_entertainment
   + data_channel_is_tech
   + weekday_is_tuesday
   + weekday_is_wednesday
   + weekday_is_thursday, testOnlineNews,
     method = "lm", trControl = trainControl(
     method = "cv",
     number = 10,
     verboseIter = TRUE))
 summary(model_aic_polyn)

 # BIC polynomial model
 library("caret")
 model_bic_polyn <- train(
shares ~ poly(kw_min_avg, 5, raw = TRUE) + data_channel_is_entertainment + data_channel_is_tech
+ weekday_is_tuesday + weekday_is_wednesday + weekday_is_thursday, testOnlineNews,
  method = "lm", trControl = trainControl(
  method = "cv",
  number = 10,
  verboseIter = TRUE))
 summary(model_bic_polyn)
```

```
# QUESTION 5: TEST DATA FILE
 # Selecting model
    # Initial regression model
    fit_full_3 <- lm(shares ~ ., data = testOnlineNews)
    summary(fit_full_3)
    # Final regression model
    # AIC method
    test_selected_aic_model <- step(fit_full_3, direction='both')
    summary(test_selected_aic_model)
    length(test_selected_aic_model$coefficients)

    # BIC method
    test_nRegres = length(resid(fit_full_3))
    test_selected_bic_model = step(fit_full_3, direction = "both", k = log(test_nRegres))
    summary(test_selected_bic_model)
    length(test_selected_bic_model$coefficients)
    # 10-fold cross validation of full model
    library("caret")
    model_full_test <- train(
   shares ~., onlineNews,
   method = "lm", trControl = trainControl(
     method = "cv",
     number = 10,
     verboseIter = TRUE))
    summary(model_full_test)
    # 10-fold cross validation of AIC model
    library("caret")
    model_full_aic_test_2 <- train(
   shares ~ n_tokens_title + n_tokens_content + n_unique_tokens + n_non_stop_unique_tokens + num_hrefs
   + num_self_hrefs + num_imgs + average_token_length + data_channel_is_lifestyle
   + data_channel_is_entertainment + data_channel_is_bus + data_channel_is_socmed + data_channel_is_tech
   + kw_min_min + kw_max_min + kw_avg_min + kw_max_max + kw_avg_max + kw_min_avg + kw_max_avg
   + kw_avg_avg + self_reference_min_shares
   + weekday_is_monday + weekday_is_tuesday + weekday_is_wednesday + weekday_is_thursday
```

```
  + weekday_is_friday + LDA_00 + LDA_01 + LDA_02 + LDA_03
  + global_subjectivity + global_rate_positive_words + avg_negative_polarity
  + title_subjectivity + title_sentiment_polarity + abs_title_subjectivity, onlineNews,
  method = "lm", trControl = trainControl(
 method = "cv",
 number = 10,
 verboseIter = TRUE))
 summary(model_full_aic_test_2)
# 10-fold cross validation of BIC model
library("caret")
model_bic_full_test_2 <- train(
shares ~ n_non_stop_unique_tokens +
  + data_channel_is_entertainment + data_channel_is_socmed + data_channel_is_tech
  + kw_min_min + kw_max_min + kw_avg_min + kw_avg_max + kw_min_avg + kw_max_avg + kw_avg_avg
  + self_reference_min_shares
  + weekday_is_monday + weekday_is_tuesday + weekday_is_wednesday + weekday_is_thursday
  + weekday_is_friday + LDA_00 +
  + global_subjectivity, onlineNews,
  method = "lm", trControl = trainControl(
    method = "cv",
    number = 10,
    verboseIter = TRUE))
   summary(model_bic_full_test_2)
```