Online News Popularity Data Set Main Assignment for Statistics for Business Analytics I – P.T. (2022-2023)

Panagiotis G. Vaidomarkakis (p2822203)

Using the alldata_onlinenews_45 dataset

Tutor: Ioannis Ntzoufras

06/01/2023

M.Sc. In Business Analytics (Part Time) 2022-2024 at Athens University of Economics and Business (A.U.E.B.)

Table of Contents

Abstract

In our case, data were collected on January 8, 2015, from articles published by Mashable(www.mashable.com). The whole dataset contains 39797 rows. Our random sub-sample training dataset containing 3000 rows of observations and our evaluation/test dataset contains 10000 rows of observations with 61 attributes in each row. 58 of them are predictive attributes, 2 are non-predictive attributes (url, timedelta) and the last of them (shares) is our goal field. Moreover, we have none missing attribute values. After pre-processing the training data, we tested several different prediction models to arrive at the final prediction model which is the following:

$\log(shares) = 1{,}96 - 0{,}188 * n\_unique\_tokens + 0{,}11 * n\_non\_stop\_unique\_tokens + 0{,}001 * num\_hrefs - 0{,}02 * (if\ data\_channel\_is\_lifestyle = yes) - 0{,}04 * (if\ data\_channel\_is\_entertainment = yes) - 0{,}03 * (if\ data\_channel\_is\_bus = yes) - 0{,}0000001 * kw\_\mathrm{max\_max} + 0{,}00005 * \mathrm{kw\_avg\_avg} + 0{,}000001 * \mathrm{self\_reference\_avg\_sharess} - 0.01 * \mathrm{weekday\_is\_wednesday} - 0{,}05 * \mathrm{LDA\_02} + 0{,}34 * \mathrm{global\_rate\_positive\_words} + 0{,}02 * \mathrm{abs\_title\_sentiment\_polarity} - 0{,}000000002 * (\mathrm{kw\_avg\_avg})^2 - 0{,}000000000001 * (\mathrm{self\_reference\_avg\_sharess})^2 + \varepsilon$ where $\varepsilon \sim N(0, 0.1113^2)$

This model managed to have $R^2 = 0.1477$ and Adj. $R^2 = 0.1434$.
In the end of this assignment, we made some test in order to evaluate it.

Introduction

As we said in the Abstract, whole dataset has 39797 rows of observations. We have 58 metrics in order to predict the shares that an article will take. Our training data was a random sub-set of the whole dataset containing 3000 rows. Finally, all the class had a test data of 10000 rows in order to test and evaluate our models. Now, let’s start with our descriptive and exploratory data analysis.

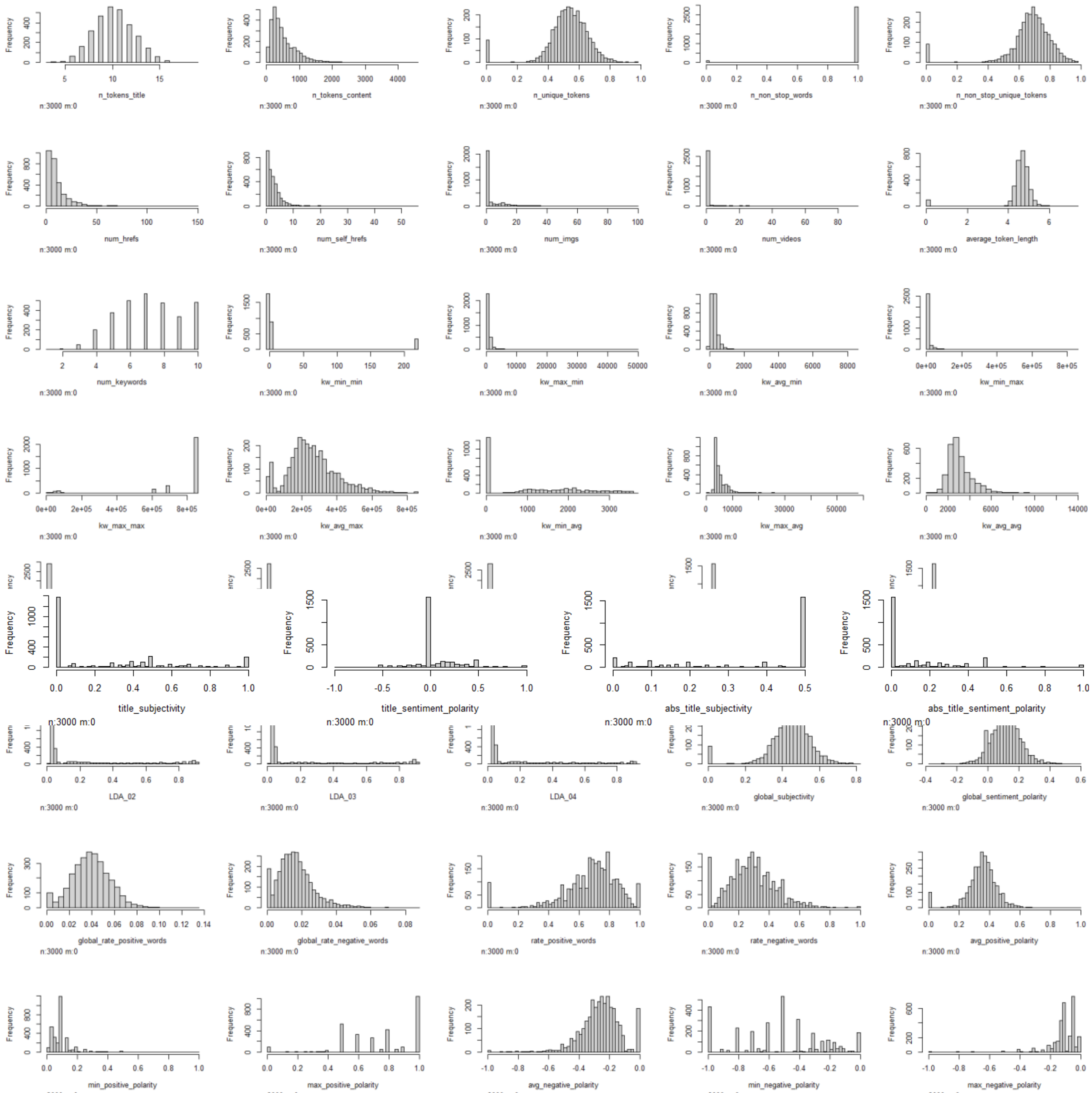
Descriptive analysis and exploratory data analysis

First, we insert our data into R-studio. After that, we need to remove the id of the observation, the url and the timedelta which is the time from 08/01/2015 until the time the article was published. After that, we also remove the is_weekend attribute because we already have is_saturday and is_sunday and it will be an overlap. Then, we find the categorical variables and identify them as factor variables with 2 possible outcomes (1=”Yes” and 0=”No”). Additionally, we separate the numerical variables from factor variables in order to have different visuals for each variable.

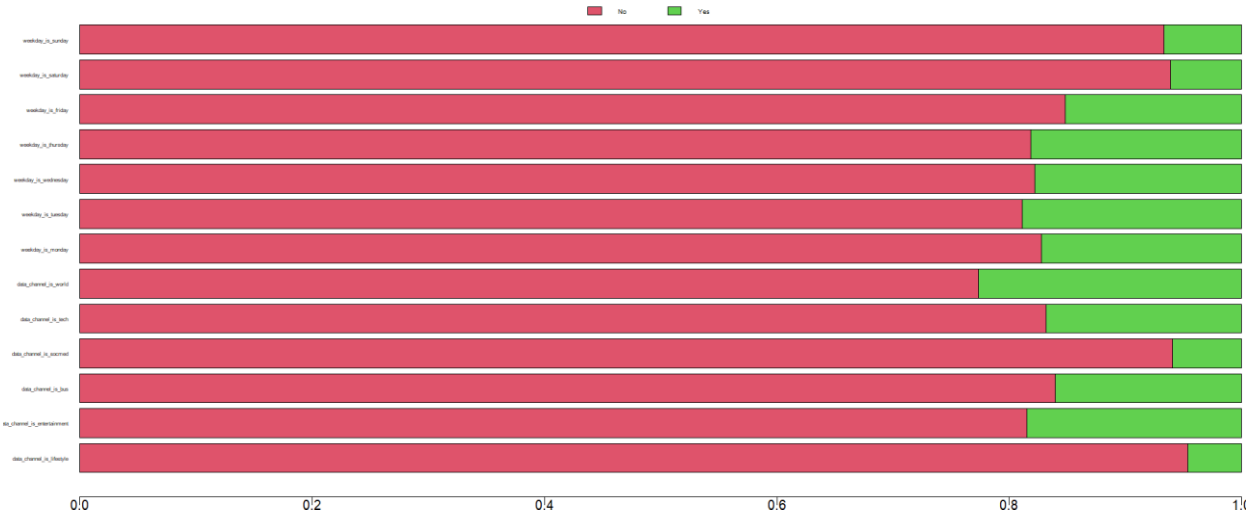
```
> summary(training_dataset)
 n_tokens_title n_tokens_content n_unique_tokens  n_non_stop_words n_non_stop_unique_tokens   num_hrefs      num_self_hrefs      num_imgs
 Min.   : 3.0   Min.   :   0.0   Min.   :0.0000   Min.   :0.000    Min.   :0.0000             Min.   :  0.00   Min.   : 0.000   Min.   :  0.000
 1st Qu.: 9.0   1st Qu.: 238.0   1st Qu.:0.4711   1st Qu.:1.000    1st Qu.:0.6253             1st Qu.:  4.00   1st Qu.: 1.000   1st Qu.:  1.000
 Median :10.0   Median : 401.5   Median :0.5393   Median :1.000    Median :0.6884             Median :  8.00   Median : 2.000   Median :  1.000
 Mean   :10.4   Mean   : 537.2   Mean   :0.5310   Mean   :0.969    Mean   :0.6713             Mean   : 10.97   Mean   : 3.262   Mean   :  4.594
 3rd Qu.:12.0   3rd Qu.: 707.0   3rd Qu.:0.6119   3rd Qu.:1.000    3rd Qu.:0.7538             3rd Qu.: 14.00   3rd Qu.: 4.000   3rd Qu.:  4.000
 Max.   :19.0   Max.   :4514.0   Max.   :0.9730   Max.   :1.000    Max.   :0.9706             Max.   :150.00   Max.   :56.000   Max.   :100.000
   num_videos     average_token_length  num_keywords     data_channel_is_lifestyle data_channel_is_entertainment data_channel_is_bus
 Min.   : 0.000   Min.   :0.000        Min.   : 1.000   No :2862                  No :2445                      No :2519
 1st Qu.: 0.000   1st Qu.:4.495        1st Qu.: 6.000   Yes: 138                  Yes: 555                      Yes: 481
 Median : 0.000   Median :4.675        Median : 7.000
 Mean   : 1.258   Mean   :4.554        Mean   : 7.165
 3rd Qu.: 1.000   3rd Qu.:4.859        3rd Qu.: 9.000
 Max.   :91.000   Max.   :7.218        Max.   :10.000
 data_channel_is_socmed data_channel_is_tech data_channel_is_world   kw_min_min       kw_max_min       kw_avg_min        kw_min_max
 No :2822               No :2495             No :2320              Min.   : -1.0   Min.   :    0   Min.   :  -1.0   Min.   :     0
 Yes: 178               Yes: 505             Yes: 680              1st Qu.: -1.0   1st Qu.:  438   1st Qu.: 135.9   1st Qu.:     0
                                                                   Median : -1.0   Median :  642   Median : 232.0   Median :  1600
                                                                   Mean   : 25.4   Mean   : 1014   Mean   : 289.5   Mean   : 14471
                                                                   3rd Qu.:  4.0   3rd Qu.: 1000   3rd Qu.: 351.3   3rd Qu.:  8300
                                                                   Max.   :217.0   Max.   :50000   Max.   :8494.3   Max.   :843300
   kw_max_max       kw_avg_max       kw_min_avg      kw_max_avg       kw_avg_avg     self_reference_min_shares self_reference_max_shares
 Min.   :     0   Min.   :     0   Min.   :   0   Min.   :    0   Min.   :    0   Min.   :     0            Min.   :     0
 1st Qu.:843300   1st Qu.:177064   1st Qu.:   0   1st Qu.: 3537   1st Qu.: 2373   1st Qu.:   654            1st Qu.:  1100
 Median :843300   Median :246475   Median :1067   Median : 4307   Median : 2846   Median :  1200            Median :  2800
 Mean   :754973   Mean   :262090   Mean   :1147   Mean   : 5381   Mean   : 3106   Mean   :  4546            Mean   : 11203
 3rd Qu.:843300   3rd Qu.:330463   3rd Qu.:2087   3rd Qu.: 5943   3rd Qu.: 3591   3rd Qu.:  2700            3rd Qu.:  7500
 Max.   :843300   Max.   :843300   Max.   :3613   Max.   :57513   Max.   :13595   Max.   :663600            Max.   :843300
 self_reference_avg_sharess weekday_is_monday weekday_is_tuesday weekday_is_wednesday weekday_is_thursday weekday_is_friday weekday_is_saturday
 Min.   :     0.0           No :2484          No :2434           No :2466             No :2455            No :2545          No :2817
 1st Qu.:   993.8           Yes: 516          Yes: 566           Yes: 534             Yes: 545            Yes: 455          Yes: 183
 Median :  2200.0
 Mean   :  7001.5

3rd Qu.:  5000.0
Max.   :663600.0
weekday_is_sunday     LDA_00            LDA_01            LDA_02            LDA_03            LDA_04          global_subjectivity
No :2799          Min.   :0.01884   Min.   :0.01819   Min.   :0.01819   Min.   :0.01820   Min.   :0.01829   Min.   :0.0000
Yes: 201          1st Qu.:0.02512   1st Qu.:0.02501   1st Qu.:0.02857   1st Qu.:0.02857   1st Qu.:0.02857   1st Qu.:0.3936
                  Median :0.03347   Median :0.03335   Median :0.04005   Median :0.04000   Median :0.04001   Median :0.4534
                  Mean   :0.18723   Mean   :0.14647   Mean   :0.22534   Mean   :0.22299   Mean   :0.21797   Mean   :0.4428
                  3rd Qu.:0.24635   3rd Qu.:0.16816   3rd Qu.:0.36315   3rd Qu.:0.37508   3rd Qu.:0.36442   3rd Qu.:0.5073
                  Max.   :0.92000   Max.   :0.91985   Max.   :0.92000   Max.   :0.91998   Max.   :0.92653   Max.   :0.8069
global_sentiment_polarity global_rate_positive_words global_rate_negative_words rate_positive_words rate_negative_words avg_positive_polarity
Min.   :-0.37500          Min.   :0.00000            Min.   :0.000000            Min.   :0.0000      Min.   :0.0000      Min.   :0.0000
1st Qu.: 0.05563          1st Qu.:0.02822            1st Qu.:0.009644            1st Qu.:0.6000      1st Qu.:0.1892      1st Qu.:0.3061
Median : 0.11868          Median :0.03893            Median :0.015316            Median :0.7083      Median :0.2838      Median :0.3585
Mean   : 0.11718          Mean   :0.03941            Mean   :0.016776            Mean   :0.6786      Mean   :0.2904      Mean   :0.3528
3rd Qu.: 0.17495          3rd Qu.:0.05014            3rd Qu.:0.021807            3rd Qu.:0.8000      3rd Qu.:0.3846      3rd Qu.:0.4085
Max.   : 0.60000          Max.   :0.13223            Max.   :0.086168            Max.   :1.0000      Max.   :1.0000      Max.   :1.0000
min_positive_polarity max_positive_polarity avg_negative_polarity min_negative_polarity max_negative_polarity title_subjectivity
Min.   :0.00000       Min.   :0.0000        Min.   :-1.0000       Min.   :-1.0000       Min.   :-1.0000       Min.   :0.0000
1st Qu.:0.05000       1st Qu.:0.6000        1st Qu.:-0.3254       1st Qu.:-0.7000       1st Qu.:-0.1250       1st Qu.:0.0000
Median :0.10000       Median :0.8000        Median :-0.2517       Median :-0.5000       Median :-0.1000       Median :0.1250
Mean   :0.09773       Mean   :0.7468        Mean   :-0.2591       Mean   :-0.5182       Mean   :-0.1087       Mean   :0.2861
3rd Qu.:0.10000       3rd Qu.:1.0000        3rd Qu.:-0.1847       3rd Qu.:-0.3000       3rd Qu.:-0.0500       3rd Qu.:0.5000
Max.   :1.00000       Max.   :1.0000        Max.   : 0.0000       Max.   : 0.0000       Max.   : 0.0000       Max.   :1.0000
title_sentiment_polarity abs_title_subjectivity abs_title_sentiment_polarity     shares
Min.   :-1.00000         Min.   :0.0000         Min.   :0.000                Min.   :    42
1st Qu.: 0.00000         1st Qu.:0.1500         1st Qu.:0.000                1st Qu.:   931
Median : 0.00000         Median :0.5000         Median :0.000                Median :  1400
Mean   : 0.07496         Mean   :0.3421         Mean   :0.157                Mean   :  3424
3rd Qu.: 0.16000         3rd Qu.:0.5000         3rd Qu.:0.250                3rd Qu.:  2800
Max.   : 1.00000         Max.   :0.5000         Max.   :1.000                Max.   :843300
```

Above, we can see a summary of our data.

Below, you will see histograms about every variable and bar plots for categorical variables.



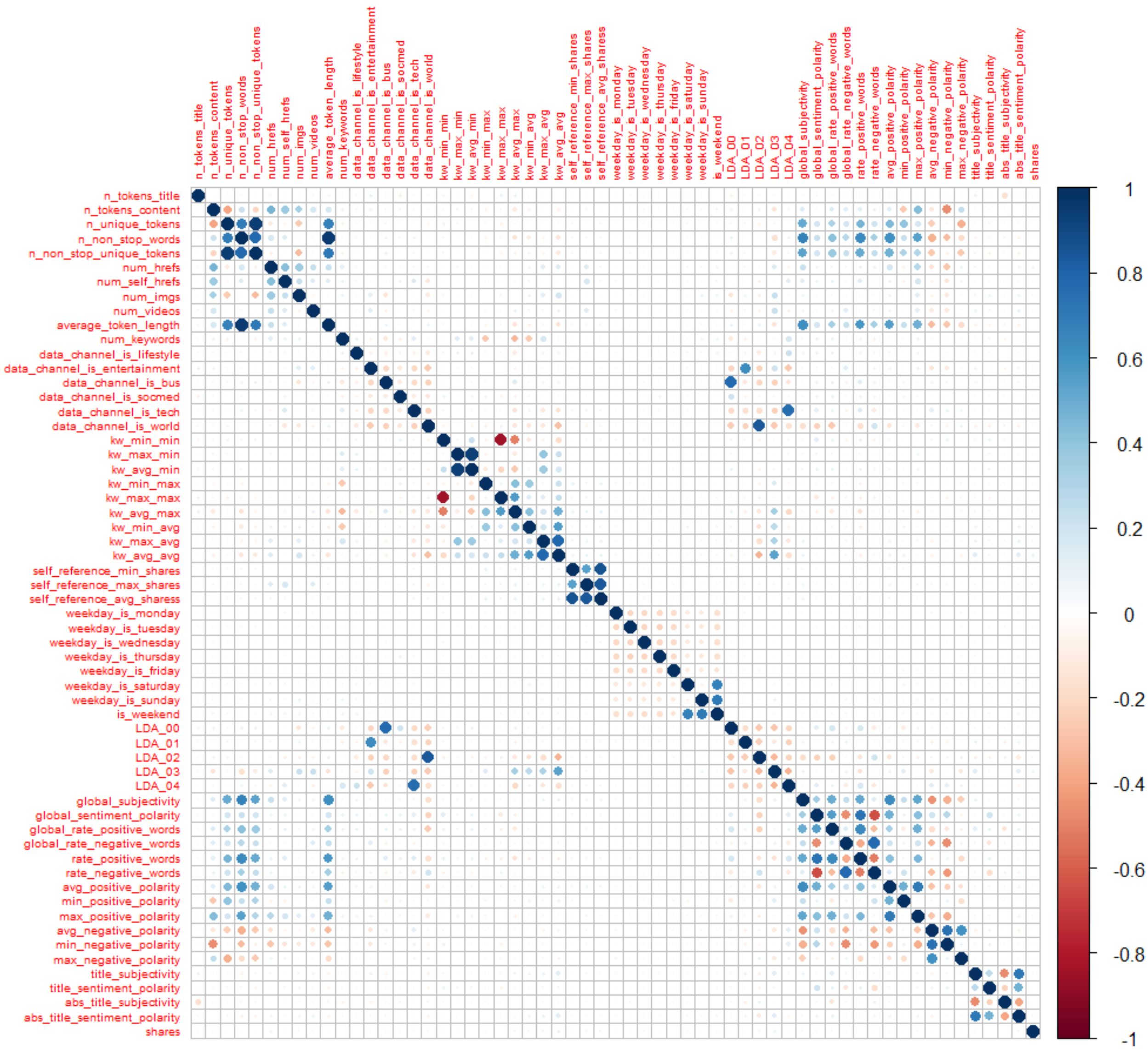
As wee can observe, most of the variables seems normal distributed.



Above we can see the bar plots which show that we have in general 80% of “No” in every variable.

Pairwise comparisons

In the end, we have to check if we have any correlation between variable share and any other variable, so a corrplot is what we need in order to check it. If there is a negative correlation, then, in the row of share, we will see a red dot and if there is a positive correlation, we will see a blue dot. That’s why, we can observe blue dots along the main diagonal of this table because a variable has 100% positive correlation with itself. Below, you can see the matrix:



As we can see, no attribute has any correlation related to shares so we are ready to start making prediction models. There are some correlations between some variables which we can try later to minimize.

Predictive or Descriptive models

Our first model is the full model which contains all the variables in order to predict shares. Our R^2 was 0,026, adj. R^2 was 0,008 and $\varepsilon \sim N\ (0{,}16590^2)$. This indicates that our model cannot predict very well shares. After that, we tried to make a LASSO model in order to keep only the significant variables and the use a stepwise procedure but in the end, R^2 was 0,016, adj. R^2 was

0,013 and ε ~ N (0,16540^2) from the LASSO so we didn't go with this procedure. We need to make it better, so we skip LASSO model and we made only a stepwise procedure. After that, our R^2 was 0,021, adj. R^2 was 0,015 and ε ~ N (0,16530^2) which is still not great, but it is better than LASSO and stepwise. After that, we tried to remove the intercept in order to have better presentation and our R^2 was 0,018 so we skipped that model because intercept was significant.

All of the above methods didn't really worked out so we need to attempt non-linear transformations. We transform the data from the last linear model in order to use logarithmic procedure. In shares variable, adding 1 share make little to no difference so we add 1 to this, just to avoid -inf as an answer after the transformation. This had a result of 0,1401 R^2, 0,1346 adj. R^2 and ε ~ N (0, $(e^{0,11192})^2$=1,251) which means that the error is about 25,1% of the predictive value. This is a far better model from the previous ones so we keep this one and we try to make modifications in it.

After the logarithmic model, we try again a stepwise procedure in order to have better prediction. After the stepmodel2, R^2 was 0,1393, adj. R^2 was 0,1353 and ε ~ N (0, $(e^{0,11192})^2$=1,251). We decided to keep this one because of the slightly better adj. R^2. In this almost final model, all the variables are significant. But we don't stop here because we can try polynomial models on top of it.

After many trials, we managed to have our final model which is the above plus (kw_avg_avg)^2 + (self_reference_avg_sharess)^2 minus the self_reference_min_shares. R^2 was 0,1477, adj. R^2 was 0,1434 and ε ~ N (0, $(e^{0,1113})^2$=1.253). So, we have a slightly larger error, but we have almost 1% better predictive power.

So, our final model is:

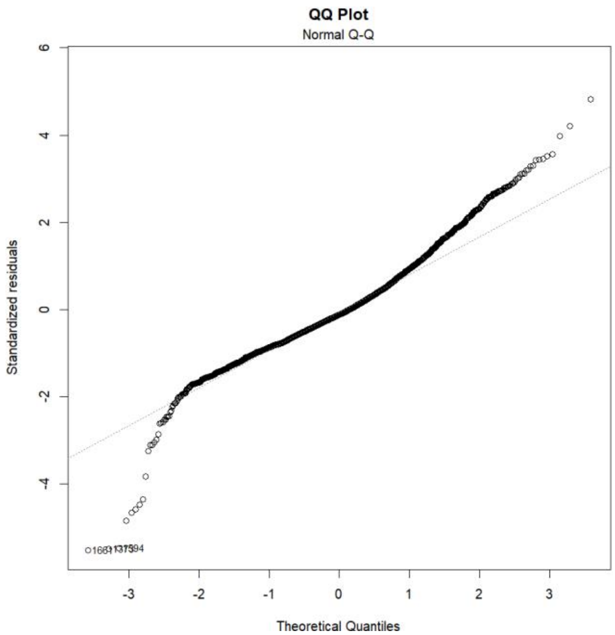
$\log(shares) = 1{,}96 - 0{,}188 * n\_unique\_tokens + 0{,}11 * n\_non\_stop\_unique\_tokens + 0{,}001 * num\_hrefs - 0{,}02 * (if\ data\_channel\_is\_lifestyle = yes) - 0{,}04 * (if\ data\_channel\_is\_entertainment = yes) - 0{,}03 * (if\ data\_channel\_is\_bus = yes) - 0{,}0000001 * kw\_\mathrm{max\_max} + 0{,}00005 * \mathrm{kw\_avg\_avg} + 0{,}000001 * \mathrm{self\_reference\_avg\_sharess} - 0.01 * \mathrm{weekday\_is\_wednesday} - 0{,}05 * \mathrm{LDA\_02} + 0{,}34 * \mathrm{global\_rate\_positive\_words} + 0{,}02 * \mathrm{abs\_title\_sentiment\_polarity} - 0{,}000000002 * (\mathrm{kw\_avg\_avg})^2 - 0{,}000000000001 * (\mathrm{self\_reference\_avg\_sharess})^2 + \varepsilon$

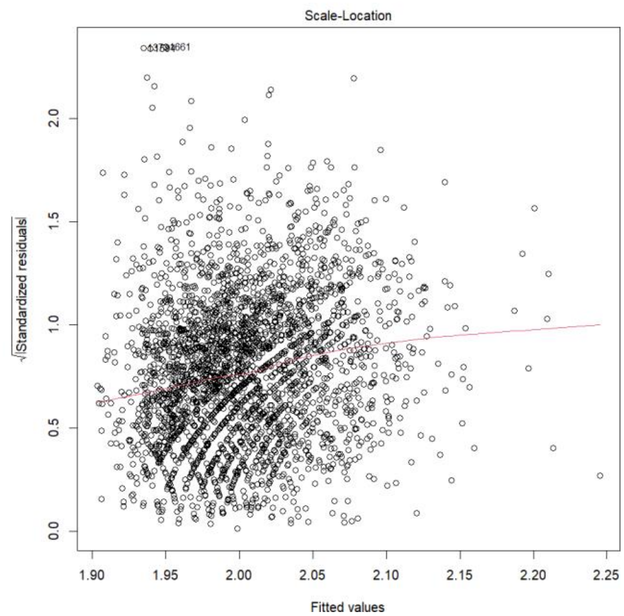
where ε ~ N (0,$(e^{0,1113})^2$=1.253) (about ± 25,3% error from each predictive value)

We need to make some evaluation tests in order to keep this model.

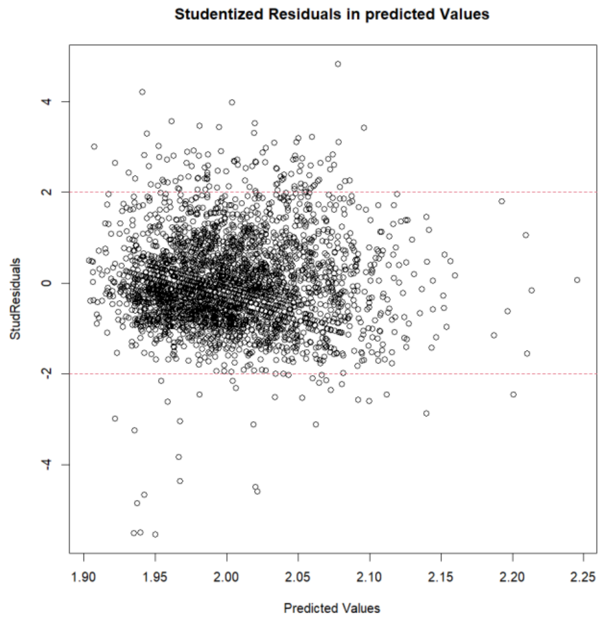
```
        Shapiro-Wilk normality test

data:  rstandard(model)
W = 0.96578, p-value < 2.2e-16

[1] "NCV Test"
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 29.16826, Df = 1, p = 6.6357e-08
```

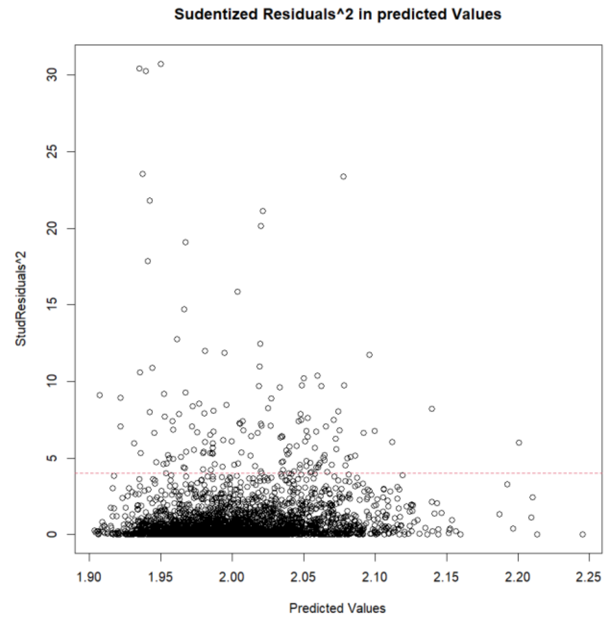
As we can see, we reject normality of errors (SW $p=2.2*10^{-16}<0.05$)

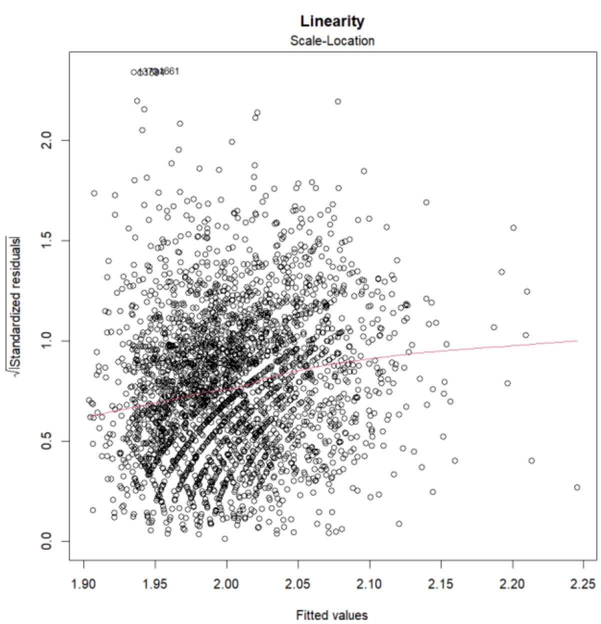


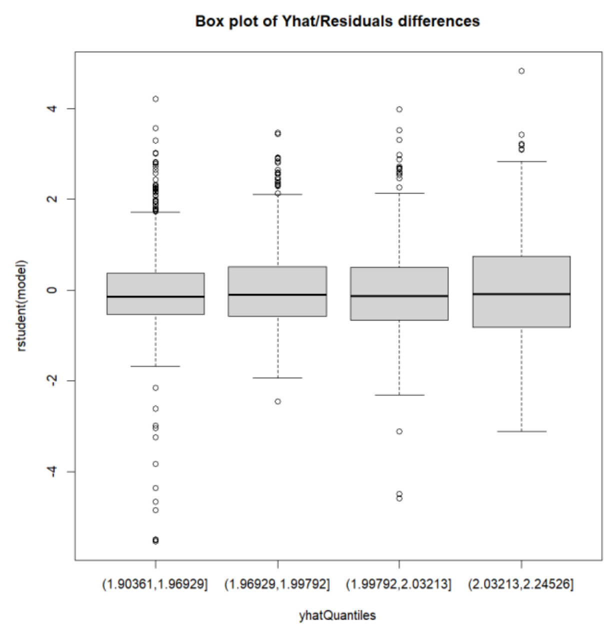


We can see from the above that we don’t have normality of errors. Let’s check for constant variance (NCV $p=6.6357*10^{-8}<0.05$). NCV rejection shows to us that the error variance is not constant, but changes based on predictors but in our case, we reject H_0 hypothesis.



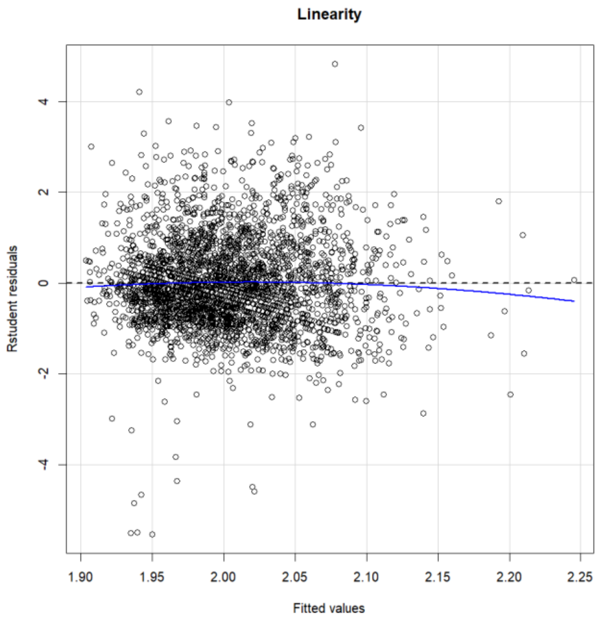


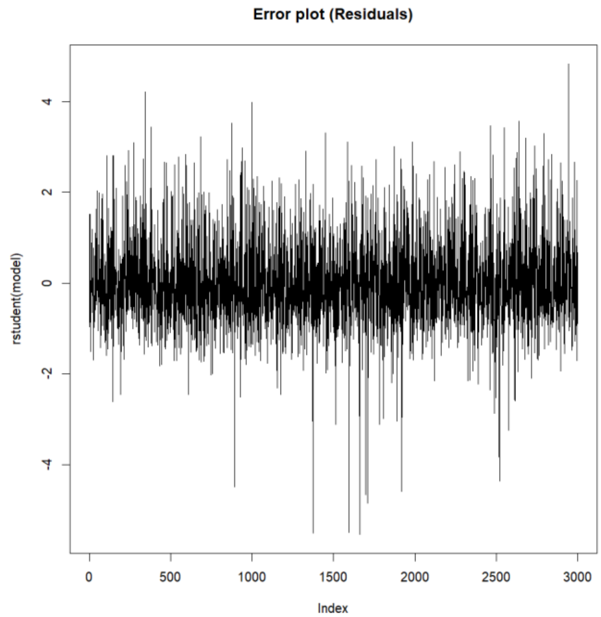




From all the above, we can see that the errors are not constant (first boxplot has many spread errors).

Next, we will check for Linearity and in the end, we will check about independency of errors.

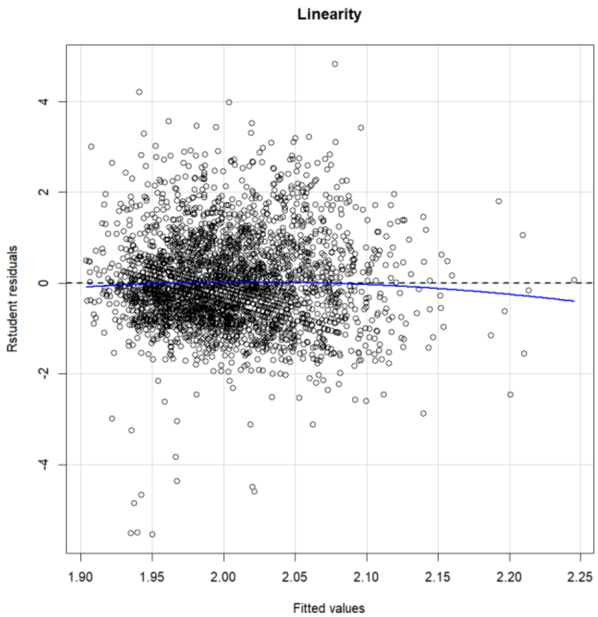


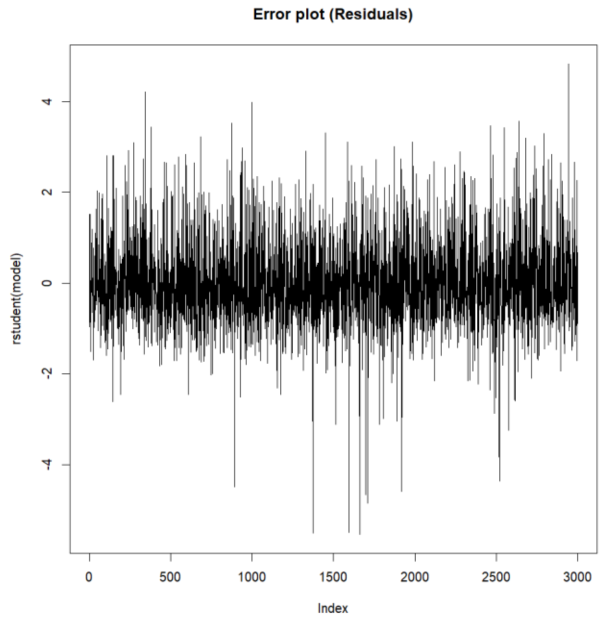


From the above and from (runs.test p=0,6091>0.05), we don't reject linearity. About independency of errors, we have the below results:

```
        Approximate runs test

data:  model$res
Runs = 1515, p-value = 0.6091
alternative hypothesis: two.sided

[1] "DW test"

        Durbin-Watson test

data:  model
DW = 2.0535, p-value = 0.9285
alternative hypothesis: true autocorrelation is greater than 0
```

(DW p=0,9285>0,05), so we don't reject independency.





Error plot seems fine though so we will keep this model.

Further Analysis

We need to do 10-fold validation in order to evaluate our model. Below you can see the results in training data:

```
> model$resample
        RMSE  Rsquared        MAE Resample
1  0.1189527 0.1159618 0.08643597   Fold01
2  0.1086600 0.1359145 0.08430627   Fold02
3  0.1083114 0.1269846 0.08091533   Fold03
4  0.1140376 0.1078250 0.08339175   Fold04
5  0.1234766 0.1312695 0.08718464   Fold05
6  0.1130057 0.1304441 0.08610332   Fold06
7  0.1079110 0.2249911 0.08304458   Fold07
8  0.1090557 0.1337477 0.08250810   Fold08
9  0.1036521 0.1329969 0.08225201   Fold09
10 0.1083966 0.1637771 0.08418736   Fold10
```

We also need our test data so we import them and do the same procedure. Below you will find the results in test data:

```
> modelvalidation$resample
        RMSE   Rsquared        MAE Resample
1  0.1102980 0.09104564 0.08339984   Fold01
2  0.1109974 0.08521975 0.08406980   Fold02
3  0.1112349 0.11136145 0.08389769   Fold03
4  0.1105052 0.09139405 0.08501599   Fold04
5  0.1128130 0.08839445 0.08375240   Fold05
6  0.1173226 0.13909585 0.08798165   Fold06
7  0.1167361 0.09680143 0.08802689   Fold07
8  0.1104957 0.09490521 0.08300843   Fold08
9  0.1169051 0.11809557 0.08711818   Fold09
10 0.1169881 0.12814433 0.08864601   Fold10
```

As we can see, there aren't many changes which lead us that our model predicts about the same for training data and test data.

```
> predict(quadModel,newdata=baseDataExpotest, interval = 'confidence')
          fit       lwr      upr
1     1.9585208  1.9471392 1.969902
2     1.9600699  1.9492905 1.970849
3     2.0135146  1.9942089 2.032820
4     1.9742778  1.9635078 1.985048
5     1.9637013  1.9449333 1.982469
6     1.9629317  1.9518563 1.974007
7     1.9665154  1.9547590 1.978272
8     2.0340169  2.0250630 2.042971
9     1.9965463  1.9731500 2.019943
10    2.0466599  2.0363244 2.056995
```

We also use predict fucction to see some predictions on test data (with lowest and highest error).

Conclusions

Our conclusions about these articles are:

If an article is for lifestyle, entertainment and bus, then it will have less shares.

If an article will be published on Wednesday, then it will have less shares than in other days.

If an article has many links, it will have more shares.

If it has average keywords, then it will have more shares.

If an article has average shares of referenced articles in Mashable, then it will have more shares.

The most important thing to increase shares is having high rate of positive words in the content.(1% higher rate of positive words leads to ($e^{0,04512}$)=1.046 which means 4,6% more shares). The same logic goes for all the above attributes.

Reference:

K. Fernandes, P. Vinagre and P. Cortez.(2015). A Proactive Intelligent Decision Support Systemfor Predicting the Popularity of Online News.Proceedings of the 17th EPIA 2015- PortugueseConference on Artificial Intelligence, September, Coimbra, Portugal

Source:

Kelwin Fernandes-INESC TEC, Porto, Portugal/Universidade do Porto, Portugal.

Pedro Vinagre-ALGORITMI Research Centre, Universidade do Minho, Portugal

Paulo Cortez-ALGORITMIResearch Centre, Universidade do Minho, Portugal

Pedro Sernadela-Universidade de Aveiro