INTRODUCTION TO DATA MINING

INSTRUCTED BY ASMA SANAM

GROUP MEMBERS
MISHAAL HAJIANI   13050 MUHAMMAD AFFAN SHEIKH 13186

CONTENTS

INTRODUCTION

OVER VIEW:

Mashable is a global, multi-platform media and entertainment company. Powered by its own proprietary technology, Mashable is the go-to source for tech, digital culture and entertainment content for its dedicated and influential audience around the globe.

PROBLEM IN HAND:

The organization (www.mashable.com) wants to evaluate the popularity of an article and the structure of it. The study is going to help in understanding the key factors that help in making an article a hit.

DATA SET INFORMATION:

This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years.

OBJECTIVES OF THE CURRENT STUDY

- The goal is to predict the number of shares in social networks (popularity).
- To determine the specific predators associated to gauge the popularity a news.

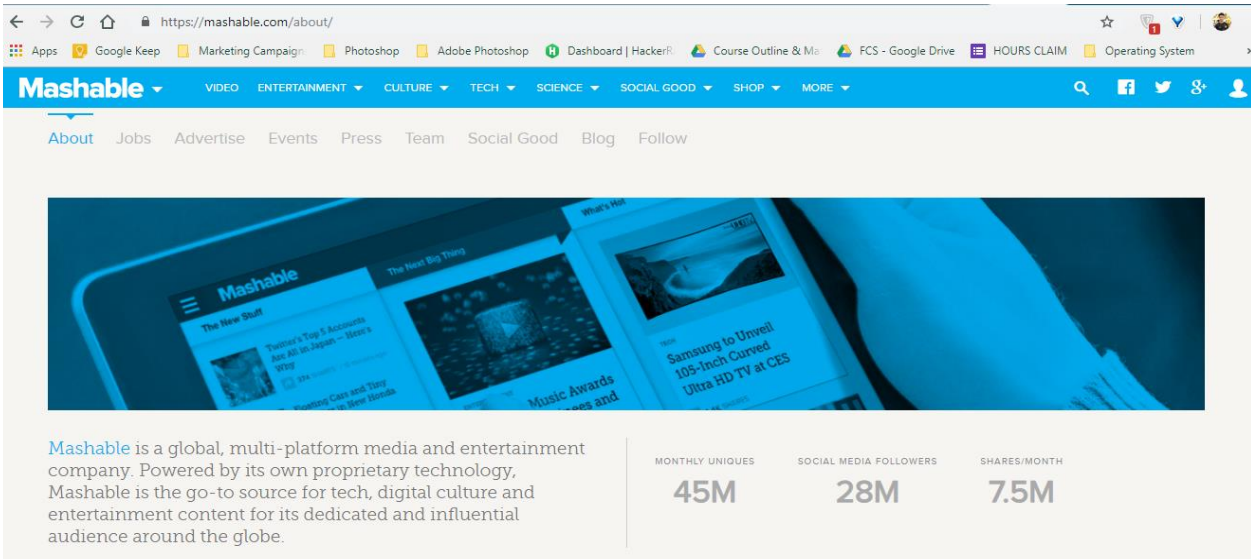
OUTLINE OF THE STUDY

We were assigned a task to develop a statistical report. The topic we have chosen is “Online News Popularity”. This dataset has been picked from an online website named, UCI-Machine Learning Repository. There were all together 61 factors available that were claimed to affect the number of shares of an article.

CONDUCTING DATA ANALYSIS

In order to systematically conduct data mining analysis, **Cross-Industry Standard Process** (CRISP) is used. CRISP is an industry standard process consisting of a sequence of steps that are usually involved in a data mining study.

Step 1-BUSINESS UNDERSTANDING



Mashable is a global, multi-platform media and entertainment company. Powered by its own proprietary technology, Mashable is the go-to source for tech, digital culture and entertainment content for its dedicated and influential audience around the globe.

MONTHLY UNIQUES	SOCIAL MEDIA FOLLOWERS	SHARES/MONTH
45M	28M	7.5M

HOW DOES MASHABLE MAKE MONEY?

Mashable is monetized by offering a variety of advertising formats to its 45M unique monthly readers.

HOW WILL THIS STUDY BOOST MASHABLE'S BUSINESS?

The study on the respective dataset will figure out the factors that make an article go viral on the internet. Being viral means a solid amount of views, comments, likes, and shares. Keeping in mind the business structure of the organization which earns money via advertisements, hence maximum views on an article means greater advertisement display duration which then results in generation of revenue for the firm. This study will state what factors, for example the use of positive or negative words, images or view content, strong title, timestamp of publication, linked content etc., have the optimum involvement in making an article reach to maximum audience.

Step 2- DATA UNDERSTANDING

We are using R-statistical software to interpret the dataset.

ATTRIBUTE INFORMATION:

Number of Attributes: 61 (58 predictive attributes, 2 non-predictive, 1 goal field)

Attribute Information:
0. url: URL of the article (non-predictive)
1. timedelta: Days between the article publication and the dataset acquisition (non-predictive)
2. n_tokens_title: Number of words in the title
3. n_tokens_content: Number of words in the content
4. n_unique_tokens: Rate of unique words in the content
5. n_non_stop_words: Rate of non-stop words in the content
6. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
7. num_hrefs: Number of links
8. num_self_hrefs: Number of links to other articles published by Mashable
9. num_imgs: Number of images
10. num_videos: Number of videos
11. average_token_length: Average length of the words in the content
12. num_keywords: Number of keywords in the metadata
13. data_channel_is_lifestyle: Is data channel 'Lifestyle'?
14. data_channel_is_entertainment: Is data channel 'Entertainment'?
15. data_channel_is_bus: Is data channel 'Business'?
16. data_channel_is_socmed: Is data channel 'Social Media'?
17. data_channel_is_tech: Is data channel 'Tech'?
18. data_channel_is_world: Is data channel 'World'?
19. kw_min_min: Worst keyword (min. shares)
20. kw_max_min: Worst keyword (max. shares)
21. kw_avg_min: Worst keyword (avg. shares)
22. kw_min_max: Best keyword (min. shares)
23. kw_max_max: Best keyword (max. shares)
24. kw_avg_max: Best keyword (avg. shares)
25. kw_min_avg: Avg. keyword (min. shares)
26. kw_max_avg: Avg. keyword (max. shares)
27. kw_avg_avg: Avg. keyword (avg. shares)
28. self_reference_min_shares: Min. shares of referenced articles in Mashable
29. self_reference_max_shares: Max. shares of referenced articles in Mashable
30. self_reference_avg_sharess: Avg. shares of referenced articles in Mashable
31. weekday_is_monday: Was the article published on a Monday?
32. weekday_is_tuesday: Was the article published on a Tuesday?
33. weekday_is_wednesday: Was the article published on a Wednesday?
34. weekday_is_thursday: Was the article published on a Thursday?
35. weekday_is_friday: Was the article published on a Friday?
36. weekday_is_saturday: Was the article published on a Saturday?
37. weekday_is_sunday: Was the article published on a Sunday?
38. is_weekend: Was the article published on the weekend?
39. LDA_00: Closeness to LDA topic 0
40. LDA_01: Closeness to LDA topic 1
41. LDA_02: Closeness to LDA topic 2

42. LDA_03: Closeness to LDA topic 3
43. LDA_04: Closeness to LDA topic 4
44. global_subjectivity: Text subjectivity
45. global_sentiment_polarity: Text sentiment polarity
46. global_rate_positive_words: Rate of positive words in the content
47. global_rate_negative_words: Rate of negative words in the content
48. rate_positive_words: Rate of positive words among non-neutral tokens
49. rate_negative_words: Rate of negative words among non-neutral tokens
50. avg_positive_polarity: Avg. polarity of positive words
51. min_positive_polarity: Min. polarity of positive words
52. max_positive_polarity: Max. polarity of positive words
53. avg_negative_polarity: Avg. polarity of negative words
54. min_negative_polarity: Min. polarity of negative words
55. max_negative_polarity: Max. polarity of negative words
56. title_subjectivity: Title subjectivity
57. title_sentiment_polarity: Title polarity
58. abs_title_subjectivity: Absolute subjectivity level
59. abs_title_sentiment_polarity: Absolute polarity level
60. shares: Number of shares (target)

Step 3- DATA PREPARATION

We did few sample t-tests to check the significance that each factor holds and finalized 35 factors that we think are more impactful.

```
Coefficients:
                              Estimate  Std. Error  t value  Pr(>|t|)
n_tokens_title               7.182e+01  2.797e+01    2.568 0.010233 *
n_tokens_content             5.364e-01  1.912e-01    2.806 0.005017 **
n_unique_tokens              2.045e+03  8.488e+02    2.409 0.016004 *
num_hrefs                    2.968e+01  6.492e+00    4.572 4.84e-06 ***
num_self_hrefs              -6.353e+01  1.730e+01   -3.672 0.000241 ***
num_imgs                     1.426e+01  8.251e+00    1.728 0.083916 .
average_token_length        -5.092e+02  1.348e+02   -3.776 0.000160 ***
num_keywords                 1.139e+02  3.266e+01    3.488 0.000488 ***
data_channel_is_lifestyle   -1.104e+03  3.886e+02   -2.841 0.004499 **
data_channel_is_entertainment -1.456e+03  2.431e+02   -5.989 2.13e-09 ***
data_channel_is_bus         -1.128e+03  3.784e+02   -2.981 0.002873 **
data_channel_is_socmed      -7.576e+02  3.645e+02   -2.078 0.037674 *
data_channel_is_tech        -7.935e+02  3.668e+02   -2.163 0.030533 *
data_channel_is_world       -7.546e+02  3.688e+02   -2.046 0.040742 *
kw_avg_min                   2.926e-01  1.130e-01    2.591 0.009583 **
kw_max_avg                  -1.145e-01  1.954e-02   -5.860 4.66e-09 ***
kw_avg_avg                   1.004e+00  9.467e-02   10.602  < 2e-16 ***
self_reference_avg_sharess   2.021e-02  2.442e-03    8.278  < 2e-16 ***
weekday_is_monday            2.786e+02  2.626e+02    1.061 0.288591
weekday_is_tuesday          -2.543e+02  2.587e+02   -0.983 0.325652
weekday_is_wednesday        -1.010e+02  2.587e+02   -0.390 0.286278
weekday_is_thursday         -2.682e+02  2.593e+02   -1.034 0.301024
weekday_is_friday           -2.342e+02  2.686e+02    -0.872 0.0383315  .
weekday_is_saturday          3.727e+02  3.206e+02    1.163 0.0244999.
LDA_00                       1.429e+06  5.949e+05    2.403 0.016285 *
LDA_01                       1.429e+06  5.949e+05    2.401 0.016337 *
LDA_02                       1.428e+06  5.949e+05    2.400 0.016401 *
LDA_03                       1.429e+06  5.949e+05    2.402 0.016292 *
LDA_04                       1.429e+06  5.949e+05    2.402 0.016326 *
global_subjectivity          2.649e+03  8.001e+02    3.311 0.000929 ***
global_rate_positive_words  -7.757e+03  4.051e+03   -1.915 0.055530 .
```

```
global_rate_negative_words    -3.750e+03  5.942e+03  -0.631 0.0527926.
avg_positive_polarity         -1.350e+03  7.567e+02  -1.784 0.074477 .
avg_negative_polarity         -1.647e+03  5.402e+02  -3.048 0.002306 **
title_sentiment_polarity       3.134e+02  2.247e+02   1.395 0.163039
---
Signif. codes:  0 ‘***’ 0.001 ‘**’
0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Step-4 MODELLING

TYPE OF REGRESSION MODEL: Observational (values of x’s(attributes) are uncontrolled

Modeling a response:

a) **Multiple Regression**: A Model Relating E(y) for Qualitative Independent Variables

E(y)= β0+β1n_tokens_title+ β2n_tokens_content+ β3n_unique_tokens+ β4num_hrefs+ β5num_self_hrefs+ β6num_imgs+ β7average_token_length+ β8num_keywords+ β9data_channel_is_lifestyle+ β10data_channel_is_entertainment+ β11data_channel_is_bus+ β12data_channel_is_socmed+ β13data_channel_is_tech+ β14data_channel_is_world+ β15kw_avg_min+ β16kw_max_avg+ β17kw_avg_avg+ β18self_reference_avg_shares+ β19weekday_is_monday+ β20weekday_is_tuesday+ β21weekday_is_wednesday+ β22weekday_is_thursday+ β23weekday_is_friday+ β24weekday_is_saturday+ β25LDA_00+ β26LDA_01+ β27LDA_02+ β28LDA_03+ β329LDA_04+ β30global_subjectivity+ β31global_rate_positive_words+ β32global_rate_negative_words+ β33avg_positive_polarity+ β34avg_negative_polarity+ β35title_sentiment_polarity+ ε

Where:

y=number of shares of an article (shares ranging from 1 to 843300)

Dummy variables

data_channel_is_lifestyle	data_channel_is_entertainment	data_channel_is_bus	data_channel_is_socmed	data_channel_is_tech	data_channel_is_world	weekday_is_monday
1-If Yes 0-If No	1-If Yes 0-If No	1-If Yes 0-If No	1-If Yes 0-If No	1-If Yes 0-If No	1-If Yes 0-If No	1-If yes 0-If not
Base case: No	Base case: No	Base case: No	Base case: No	Base case: No	Base case: No	Base Case: No

weekday_is_tuesday	weekday_is_wednesday	weekday_is_Thursday	weekday_is_friday	weekday_is_Saturday
1-If Yes 0-If No	1-If Yes 0-If No	1-If Yes 0-If No	1-If Yes 0-If No	1-If Yes 0-If No
Base case: Good	Base case: No	Base case: No	Base case: No	Base case: No

MINITAB FOR MULTIPLE REGRESSION

```
Coefficients:
                                 Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                     -1.429e+06  5.950e+05  -2.401 0.016339 *
n_tokens_title                   7.182e+01  2.797e+01   2.568 0.010233 *
n_tokens_content                 5.364e-01  1.912e-01   2.806 0.005017 **
n_unique_tokens                  2.045e+03  8.488e+02   2.409 0.016004 *
num_hrefs                        2.968e+01  6.492e+00   4.572 4.84e-06 ***
num_self_hrefs                  -6.353e+01  1.730e+01  -3.672 0.000241 ***
num_imgs                         1.426e+01  8.251e+00   1.728 0.083916 .
average_token_length            -5.092e+02  1.348e+02  -3.776 0.000160 ***
num_keywords                     1.139e+02  3.266e+01   3.488 0.000488 ***
data_channel_is_lifestyleyes     -1.104e+03  3.886e+02  -2.841 0.004499 **
data_channel_is_entertainmentyes 1.456e+03  2.431e+02  -5.989 2.13e-09 ***
data_channel_is_busyes           1.128e+03  3.784e+02  -2.981 0.002873 **
data_channel_is_socmedyes        7.576e+02  3.645e+02  -2.078 0.037674 *
data_channel_is_techyes          7.935e+02  3.668e+02  -2.163 0.030533 *
data_channel_is_worldyes         -7.546e+02  3.688e+02  -2.046 0.040742 *
kw_avg_min                       2.926e-01  1.130e-01   2.591 0.009583 **
kw_max_avg                      -1.145e-01  1.954e-02  -5.860 4.66e-09 ***
kw_avg_avg                       1.004e+00  9.467e-02  10.602  < 2e-16 ***
self_reference_avg_shares        2.021e-02  2.442e-03   8.278  < 2e-16 ***
weekday_is_mondayYes             2.786e+02  2.626e+02   1.061 0.288591
weekday_is_tuesdayYes           -2.543e+02  2.587e+02  -0.983 0.325652
weekday_is_wednesdayYes         -1.010e+02  2.587e+02  -0.390 0.696278
weekday_is_thursdayYes          -2.682e+02  2.593e+02  -1.034 0.301024
weekday_is_fridayYes            -2.342e+02  2.686e+02  -0.872 0.383315
weekday_is_saturdayYes           3.727e+02  3.206e+02   1.163 0.244999
```

```
LDA_00                         1.429e+06  5.949e+05   2.403 0.016285 *
LDA_01                         1.429e+06  5.949e+05   2.401 0.016337 *
LDA_02                         1.428e+06  5.949e+05   2.400 0.016401 *
LDA_03                         1.429e+06  5.949e+05   2.402 0.016292 *
LDA_04                         1.429e+06  5.949e+05   2.402 0.016326 *
global_subjectivity            2.649e+03  8.001e+02   3.311 0.000929 ***
global_rate_positive_words    -7.757e+03  4.051e+03  -1.915 0.055530 .
global_rate_negative_words    -3.750e+03  5.942e+03  -0.631 0.527926
avg_positive_polarity         -1.350e+03  7.567e+02  -1.784 0.074477 .
avg_negative_polarity         -1.647e+03  5.402e+02  -3.048 0.002306 **
title_sentiment_polarity       3.134e+02  2.247e+02   1.395 0.163039
---
Signif. codes:  0 ‘***’ 0.001 ‘**’
0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

COEFFICIENT INTERPRETATIONS

Coefficient of n_tokens_title: Keeping all other variables fixed, when n_tokens_title is increased by 1 word, the share is increased by 7.182e+01 on average.

Coefficient of n_tokens_content: Keeping all other variables fixed, when n_tokens_content is increased by 1 word, the share is increased by 5.364e-01 on average

Coefficient of n_unique_tokens: Keeping all other variables fixed, when n_unique_content is increased by 1 unique word, the share is increased by 2.045e+03 on average

Coefficient of num_hrefs: Keeping all other variables fixed, when num_hrefs is increased by 1 link, the share is increased by 2.968e+01 on average

Coeffocient of num_self_hrefs: Keeping all other variables fixed, when num_self_hrefs is increased by 1 link, the share is decreased by 6.353e+01 on average

Coefficient of num_imgs: Keeping all other variables fixed, when num_images is increased by 1 image, the share is increased by 1.426e+01on average

Coefficient of average_token_length: Keeping all other variables fixed, when average_token_length is increased by 1 word, the share is decreased by 5.092e+02 on average.

Coefficient of num_keywords: Keeping all other variables fixed, when num_keywords increased by 1 word, the share is increased 1.139e+02 on average.

Coefficient of data_channel_is_lifestyle: Keeping all other variables fixed, if an article is about lifestyle the shares will be 1.104e+03 less than the shares of an article which is not about lifestyle.

Coefficient of data_channel_is_entertainment: Keeping all other variables fixed, if an article is about entertainment the shares will be 1.456e+03 more than the shares of an article which is not about entertainment

Coefficient of data_channel_is_bus: Keeping all other variables fixed, if an article is about business the shares will be 1.128e+03 more than the shares of an article which is not about business

Coefficient of data_channel_is_socmed: Keeping all other variables fixed, if an article is about social media the shares will be 7.576e+02 more than the shares of an article which is not about social media

Coefficient of data_channel_is_tech: Keeping all other variables fixed, if an article is about technology the shares will be 7.935e+02 more than the shares of an article which is not about technology

Coefficient of data_channel_is_world: Keeping all other variables fixed, if an article is about world affairs the shares will be `-7.546e+02 less` than the shares of an article which is not about world affairs

MODEL SUMMARY

Residual standard error: `11510 on 39608 degrees of freedom`	
Multiple R-squared: 0.6627	Adjusted R-squared: `0.6041`

The value R^2 = .5527 is highlighted on the printout. This implies that using the independent variables, the model explains 55.27% of the total sample variation in shares, y. Thus, $R^2$is a sample statistic that tells how well the model fits the data and thereby represents a measure of the usefulness of the entire model.

ANOVA

F-statistic: `24.6 on 35 and 39608 DF,  p-value: < 2.2e-16`

TESTING THE UTILITY OF A MODEL: THE ANALYSIS OF VARIANCE F-TEST

Null hypothesis: β1 = β2 = β3 =···= β35 = 0

Alternate hypotheses: At least one of the coefficients is non-zero

The test statistic used to test this hypothesis is an F statistic, the statistical software calculates the F statistic):

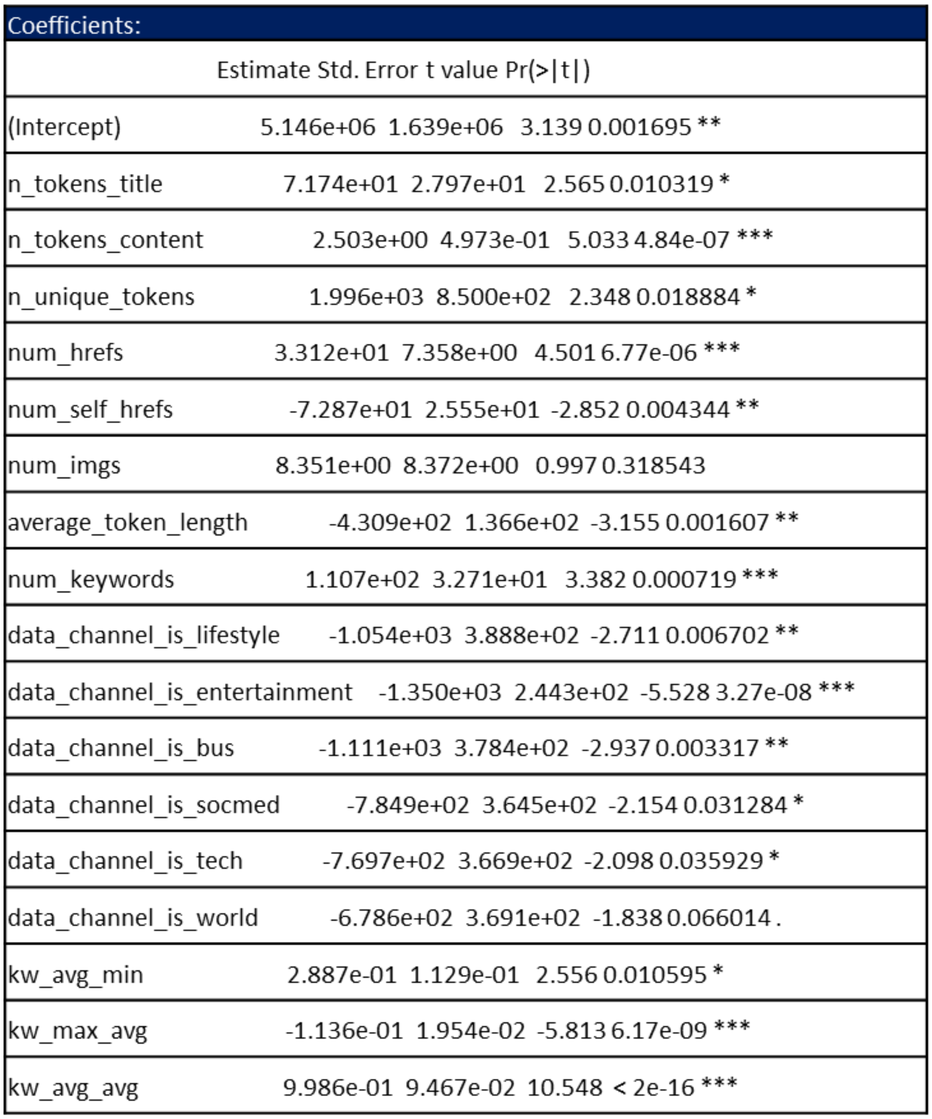
Test statistic: F = (SSyy − SSE)/k /SSE/ [n − (k + 1)] = 24.6

Conclusion: Since p-value< level of significance=0.05 we will reject null hypothesis and conclude that at least one of coefficient is non-zero. (conclusion based on p-value given in the table)

MODEL 2:

b) Interaction Model: An interaction model with qualitative predictors

E(y)= β0+β1n_tokens_title+ β2n_tokens_content+ β3n_unique_tokens+ β4num_hrefs+ β5num_self_hrefs+ β6num_imgs+ β7average_token_length+ β8num_keywords+ β9data_channel_is_lifestyle+ β*10*data_channel_is_entertainment+ β11data_channel_is_bus+ β12data_channel_is_socmed+ β13data_channel_is_tech+ β14data_channel_is_world+ β15kw_avg_min+ β16kw_max_avg+ β17kw_avg_avg+ β*18*self_reference_avg_shares+ β19weekday_is_monday+ β20weekday_is_tuesday+ β21weekday_is_wednesday+ β22weekday_is_thursday+ β23weekday_is_friday+ β24weekday_is_saturday+ β*25*LDA_00+ β26LDA_01+ β27LDA_02+ β28LDA_03+ β329LDA_04+ β30global_subjectivity+ β31global_rate_positive_words+ β32global_rate_negative_words+ β33avg_positive_polarity+ _polarity+ β36num_hrefs*num_self_hrefs+ β37n_tokens_content*n_unique_tokens + ε



Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5.146e+06	1.639e+06	3.139	0.001695	**
n_tokens_title	7.174e+01	2.797e+01	2.565	0.010319	*
n_tokens_content	2.503e+00	4.973e-01	5.033	4.84e-07	***
n_unique_tokens	1.996e+03	8.500e+02	2.348	0.018884	*
num_hrefs	3.312e+01	7.358e+00	4.501	6.77e-06	***
num_self_hrefs	-7.287e+01	2.555e+01	-2.852	0.004344	**
num_imgs	8.351e+00	8.372e+00	0.997	0.318543	
average_token_length	-4.309e+02	1.366e+02	-3.155	0.001607	**
num_keywords	1.107e+02	3.271e+01	3.382	0.000719	***
data_channel_is_lifestyle	-1.054e+03	3.888e+02	-2.711	0.006702	**
data_channel_is_entertainment	-1.350e+03	2.443e+02	-5.528	3.27e-08	***
data_channel_is_bus	-1.111e+03	3.784e+02	-2.937	0.003317	**
data_channel_is_socmed	-7.849e+02	3.645e+02	-2.154	0.031284	*
data_channel_is_tech	-7.697e+02	3.669e+02	-2.098	0.035929	*
data_channel_is_world	-6.786e+02	3.691e+02	-1.838	0.066014	.
kw_avg_min	2.887e-01	1.129e-01	2.556	0.010595	*
kw_max_avg	-1.136e-01	1.954e-02	-5.813	6.17e-09	***
kw_avg_avg	9.986e-01	9.467e-02	10.548	< 2e-16	***

self_reference_avg_sharess 2.022e-02 2.441e-03 8.285 < 2e-16 ***
weekday_is_monday 2.921e+02 2.625e+02 1.113 0.265894
weekday_is_tuesday -2.459e+02 2.587e+02 -0.951 0.341809
weekday_is_wednesday -8.789e+01 2.587e+02 -0.340 0.734001
weekday_is_thursday -2.616e+02 2.592e+02 -1.009 0.312956
weekday_is_friday -2.282e+02 2.686e+02 -0.849 0.395624
weekday_is_saturday 3.889e+02 3.205e+02 1.213 0.225047
LDA_00 -5.145e+06 1.639e+06 -3.139 0.001697 **
LDA_01 -5.146e+06 1.639e+06 -3.139 0.001694 **
LDA_02 -5.147e+06 1.639e+06 -3.140 0.001691 **
LDA_03 -5.145e+06 1.639e+06 -3.139 0.001696 **
LDA_04 -5.146e+06 1.639e+06 -3.139 0.001695 **
global_subjectivity 2.608e+03 7.999e+02 3.260 0.001116 **
global_rate_positive_words -7.131e+03 4.053e+03 -1.759 0.078501 .
global_rate_negative_words -1.678e+03 5.963e+03 -0.281 0.778424
avg_positive_polarity -1.266e+03 7.569e+02 -1.673 0.094426 .
avg_negative_polarity -1.828e+03 5.417e+02 -3.374 0.000742 ***
title_sentiment_polarity 3.159e+02 2.247e+02 1.406 0.159685
num_hrefs:num_self_hrefs 4.030e-01 5.617e-01 0.717 0.0473090*
n_tokens_content:n_unique_tokens -5.945e+00 1.388e+00 -4.282 1.86e-05 ***

ANALYSIS OF VARIANCE

```
Residual standard error: 11510 on 39606 degrees of freedom
Multiple R-squared:  0.6674, Adjusted R-squared:  0.6011
F-statistic: 23.8 on 37 and 39606 DF, p-value: < 2.2e-16
```

TESTING THE OVERALL UTILITY OF MODEL USING THE GLOBAL F-TEST AT α = .05

The global F-test is used to test the null hypothesis

Null hypothesis: $\beta1 = \beta2 = \beta3 = \cdots = \beta37 = 0$

Alternate hypotheses: At least one of the coefficients is non-zero

The test statistic and p-value of the test (highlighted on the MINITAB printout) are F = 23.8 and p = 2.2e-16, respectively. Since α = .05 exceeds the p-value, there is sufficient evidence to conclude that the model fit is a statistically useful predictor of shares, y. Reject null hypothesis.

Step 5- EVALUATION

After observing both the models:

When we looked at the adjusted R-squared value for both the regression models, they are relatively close, but since the interaction test is insignificant we may declare model A to be more suitable to predict the shares/ popularity of an article

CONCLUSION

Selected Model- Model A:

E(y)= β0+β1n_tokens_title+ β2n_tokens_content+ β3n_unique_tokens+ β4num_hrefs+ β5num_self_hrefs+ β6num_imgs+ β7average_token_length+ β8num_keywords+ β9data_channel_is_lifestyle+ β*10*data_channel_is_entertainment+ β11data_channel_is_bus+ β12data_channel_is_socmed+ β13data_channel_is_tech+ β14data_channel_is_world+ β15kw_avg_min+ β16kw_max_avg+ β17kw_avg_avg+ β*18*self_reference_avg_shares+ β19weekday_is_monday+ β20weekday_is_tuesday+ β21weekday_is_wednesday+ β22weekday_is_thursday+ β23weekday_is_friday+ β24weekday_is_saturday+ β*25*LDA_00+ β26LDA_01+ β27LDA_02+ β28LDA_03+ β329LDA_04+ β30global_subjectivity+ β31global_rate_positive_words+ β32global_rate_negative_words+ β33avg_positive_polarity+ β34avg_negative_polarity+ β35title_sentiment_polarity+ ε

CROSS VALIDATING A MODEL A:

To generate training and testing data for cross validating online news popularity model:
70% of the data is trained using model A. The rest 30% is referred as test data. Next, the model is build using test data to predict shares.

Correlation table of actual and predicted for cross validating sample

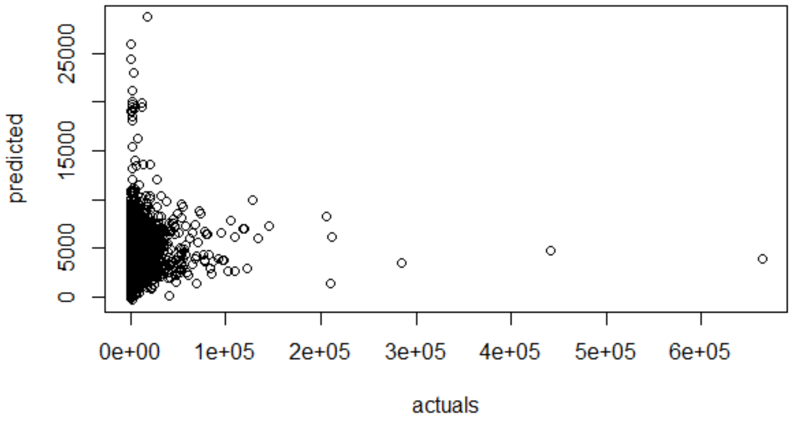
	actuals	predicted
actuals	1.0000000	0.6580839
predicted	0.6580839	1.0000000

We can say that the model selected i.e. **MODEL A**, is 65.8% good in giving accurate results.

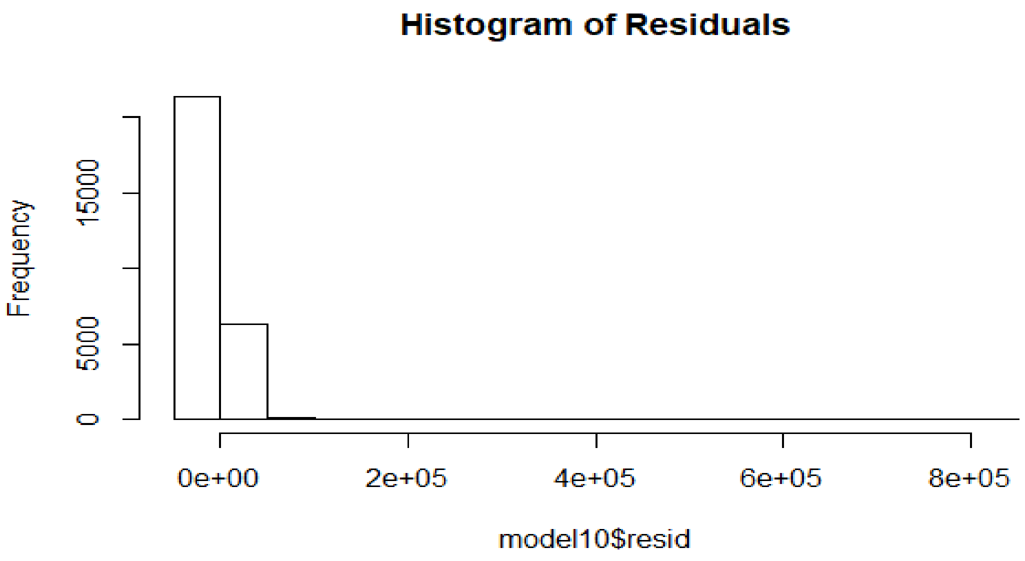
Step-6 DEPOLYMENT

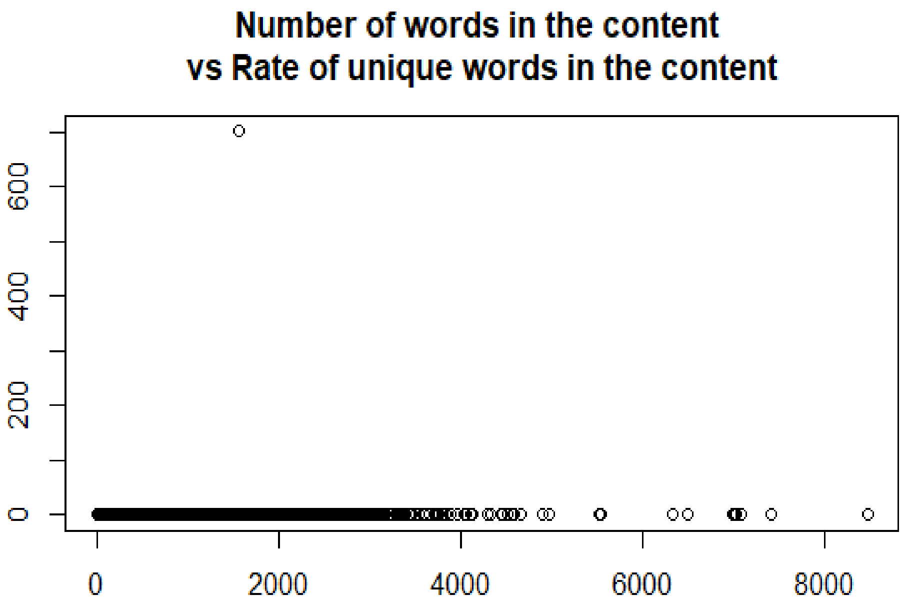
The model is finally ready for deployment

DATA REPRESENTATION USING MODEL A



Scatter plot for predicted shares vs actual shares


Histogram of Residuals
Frequency
0
5000
15000
0e+00
2e+05
4e+05
6e+05
8e+05
model10$resid


Number of words in the content
vs Rate of unique words in the content
0
200
400
600
0
2000
4000
6000
8000

REFERENCES

1. Pedro Vinagre (pedro.vinagre.sousa â€™@â€™ gmail.com) - ALGORITMI Research Centre, Universidade do Minho, Portugal

2. Paulo Cortez - ALGORITMI Research Centre, Universidade do Minho, Portugal
Pedro Sernadela - Universidade de Aveiro

3. K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.