MSDS 6372 Project 3

Team members:

Chang, Kevin | Hill, Brandon | Le, Vinh

Popularity of Online News

**Introduction**

With the Internet, there is a vast amount of news information from different online sources. Whether it is from social media network, news website, or even online videos such as YouTube, reading and sharing news have become the center of people’s entertainment lives. Therefore, it would be helpful if we could accurately predict the popularity of news and understand the need for specific content depending on our audience before publication. The purpose of this paper is to attempt to find a suitable model that will predict the popularity of online news. We will employ logistic regression methods to confirm that, based on number of shares, computer based prediction models can assist in identifying the popularity of online news.

**Descriptive Statistics**

Our dataset is provided from the UC Irvine Machine Learning Repository. The dataset summarizes a heterogeneous set of features from 39,644 articles that was published by Mashable in a period of two years. From the 59 non-response variables, we decided to focus only on the variables that have the most impact of an article being popular or not. These variables are:

|  |  |
| --- | --- |
| **Variables Names** | **Description** |
| *pop (response variable)* | *Popularity based off number of times the article was shared*  *0=Low (0-946), 1=Medium (946-1400), 2=High (1400-2800), 3=Very High (>=2800)* |
| timedelta | Duration since the article was posted |
| n\_tokens\_content | The number of words in the article title |
| num\_imgs | The number of images in the article |
| num\_videos | The number of videos in the article |
| weekday | The day of the week the article was posted  0=Monday, 1=Tuesday, 2=Wednesday, 3=Thursday, 4=Friday, 5=Saturday, 6=Sunday |
| data\_channel | The categorical topic the article is classified as  0=Lifestyle, 1=Entertainment, 2=Business, 3=Social Media, 4=Technology, 5=World Channel, 6=Other |

Table 1. Variable Descriptions

Investigating the impact of the subset of variables chosen, we checked the correlation between the variables. As shown in Figure 1 below, we found correlations between the day of the week, number of images, and the data channel utilized (.10356, .07849, and .05396 respectively). This correlates to some known elements that release of news on a weekday or weekend, number of graphics to depict the news, and which channel the news is released on all definitely make an impact to the overall popularity

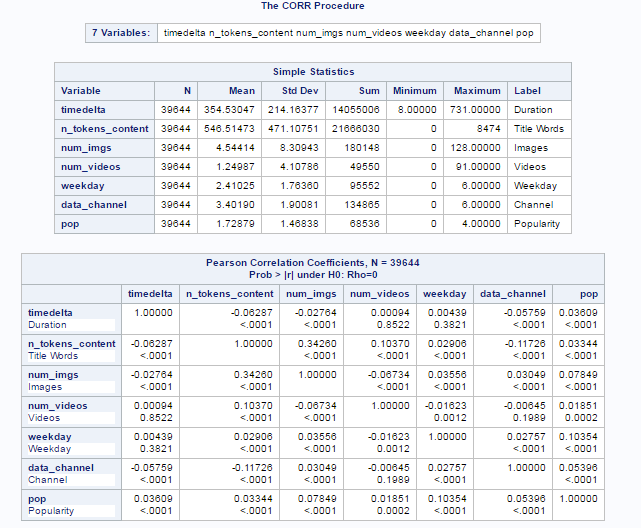


Figure 1. Pearson’s Correlation Table

**Popularity of News**

The response variables for our different models will tell us whether the news is popular or not. This is just for online news and does not include any other types of news. The popularity variable itself is categorical in 4 levels ranging from 0-3 based on the number of shares of the article.

*timedelta* – The time duration since the article was posted. There are no confounding variables to this as it is used as a metric of time and is purely independent.

*n\_tokens\_content* – The number of words in the article title. There are no confounding variables to this as the length of an article title can be manipulated to be longer or shorter based on intentions.

*num\_imgs* – The number of images in the article. There are no confounding variables to this as the number of images depends on the article and its intentions.

*num\_videos* – The number of videos in the article. This variable is largely dependent on the content itself, as if there are no possible videos to attach to an article, this would not have an impact.

*weekday* – The day of the week the article was posted. There are no confounding variables to this as the day of posting is largely related to the driving factor for the article. If it is new news that needs to be reported as a current event, the article will post regardless of what day of the week it is.

*data\_channel* – The categorical top the article is classified as. There are no confounding variables to this as this is a descriptive classification of the article.

**Initial Modeling Analysis**

Looking at the shares univariate results from SAS in figure 2, we see the quantities of different quantiles for the data response.

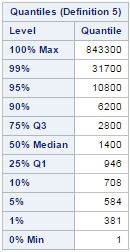


Figure 2. Shares Univariable Results

After running the shares univariate calculation, we want to verify that shares were broken into equal categories based off the quartiles using PROC FREQ in SAS, as seen below in Figure 3.

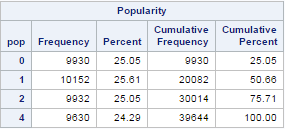


Figure 3. PROC FREQ frequency table for Popularity Variable

Each makes up roughly a quarter of the dataset. Based on the data provided above, we conducted a Random Sampling of the data in order to reduce the sample size to 500 articles as seen in Figure 4 below.

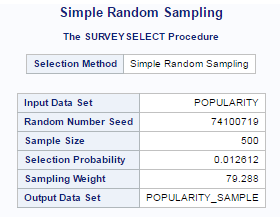


Figure 4. Random sampling of 500 articles

Re-evaluating the distribution of shares by popularity level, we see that a similar grouping exists in relation to the entire population.

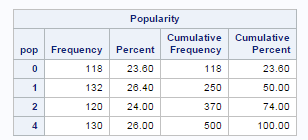
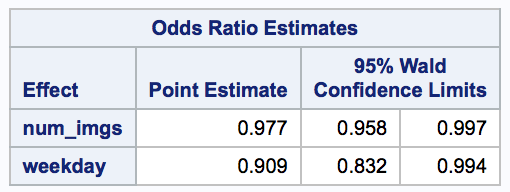


Figure 5. PROC FREQ verification

**Model Generation**

From the initial model, we looked at the factors in multiple perspectives. Based on stepwise variable selection, we found that the significant factors chosen included the number of images and the day of week factor. In the media space, these factors are synonymous with elements that drive shares. The number of images is significant to the visual appeal of an article, and the day of the week is significant in the web traffic of sharing articles on the web. The odds ratio estimates below in Figure 6 show that the number of images has an almost 100% likelihood of increasing the popularity class of an article. The day of the week also explains that about 90% likelihood of increasing the popularity class of an article.

  
Figure 6. Odds Ratio Estimates

An additional variable based on basic understanding of the industry, we also evaluated the time delta of measurement to determine its significance when evaluating the popularity of shares of an article. As seen in Figure 7 below, we see that time delta has a 100% likelihood of increasing the popularity class of an article.

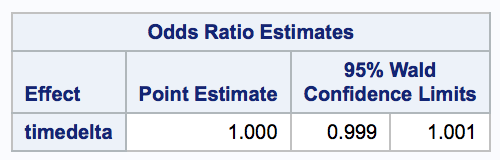


Figure 7. Odds Ratio Estimate for timedelta

Based on this analyses above, we end up failing to include time delta in our final model. The final results of the model are seen below in Figure 8.

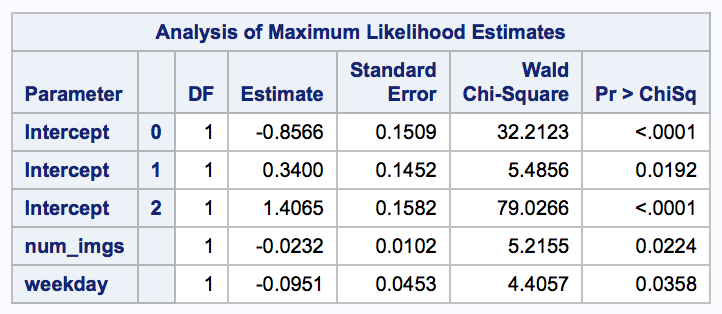


Figure 8. Estimates of Final Model

With the results seen above, we come to our final model shown below for the logistic regression of predictors for article share popularity.

Pop = -0.8566 + 0.3400 + 1.4065 + -0.0232*x1* + -0.0951*x2*

where x1 = number of images and x2 = the day of the week the article is published

**Conclusion**

Article popularity is a major driver in a number of media elements that drive businesses today. The ability to predict the sharing popularity of a news article could drive strategies around advertising and revenue based on this popularity. The most significant factors driving article popularity revolve around the number of visual images and the day of the week on which the article is published. We actually find that the number of images can be detrimental to the sharing popularity of a news article.

**Appendix**

**Final SAS Code for Reference**

data popularity\_v1;

infile

'/folders/myshortcuts/SMU/MSDS-6372/Project3/OnlineNewsPopularity.csv'

dlm=',' firstobs=2;

input url & $61. timedelta 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

num\_videos average\_token\_length num\_keywords data\_channel\_is\_lifestyle

data\_channel\_is\_entertainment data\_channel\_is\_bus data\_channel\_is\_socmed

data\_channel\_is\_tech data\_channel\_is\_world kw\_min\_min kw\_max\_min kw\_avg\_min

kw\_min\_max kw\_max\_max kw\_avg\_max kw\_min\_avg kw\_max\_avg kw\_avg\_avg

self\_reference\_min\_shares self\_reference\_max\_shares

self\_reference\_avg\_sharess weekday\_is\_monday weekday\_is\_tuesday

weekday\_is\_wednesday weekday\_is\_thursday weekday\_is\_friday

weekday\_is\_saturday weekday\_is\_sunday is\_weekend LDA\_00 LDA\_01 LDA\_02 LDA\_03

LDA\_04 global\_subjectivity global\_sentiment\_polarity

global\_rate\_positive\_words global\_rate\_negative\_words rate\_positive\_words

rate\_negative\_words avg\_positive\_polarity min\_positive\_polarity

max\_positive\_polarity avg\_negative\_polarity min\_negative\_polarity

max\_negative\_polarity title\_subjectivity title\_sentiment\_polarity

abs\_title\_subjectivity abs\_title\_sentiment\_polarity shares;

KEEP timedelta n\_tokens\_content num\_imgs num\_videos weekday\_is\_monday

weekday\_is\_tuesday weekday\_is\_wednesday weekday\_is\_thursday weekday\_is\_friday

weekday\_is\_saturday weekday\_is\_sunday is\_weekend shares

data\_channel\_is\_lifestyle data\_channel\_is\_entertainment data\_channel\_is\_bus

data\_channel\_is\_socmed data\_channel\_is\_tech data\_channel\_is\_world;

run;

/\* Group Weekday categorical variables into single categorical variable, "weekday". Drop single day variables.

0 = Monday

1 = Tuesday

2 = Wednesday

3 = Thursday

4 = Friday

5 = Saturday

6 = Sunday

\*/

data popularity\_v2;

set popularity\_v1;

if weekday\_is\_monday=1 then

weekday=0;

else if weekday\_is\_tuesday=1 then

weekday=1;

else if weekday\_is\_wednesday=1 then

weekday=2;

else if weekday\_is\_thursday=1 then

weekday=3;

else if weekday\_is\_friday=1 then

weekday=4;

else if weekday\_is\_saturday=1 then

weekday=5;

else if weekday\_is\_sunday=1 then

weekday=6;

KEEP timedelta n\_tokens\_content num\_imgs num\_videos weekday shares

data\_channel\_is\_lifestyle data\_channel\_is\_entertainment data\_channel\_is\_bus

data\_channel\_is\_socmed data\_channel\_is\_tech data\_channel\_is\_world;

run;

/\* Group Channel categorical variables into a single categorical variable, "data\_channel". Drop various channel variables. (Created "Other" for records that were not categorized.)

0 = Lifestyle Channel

1 = Entertainment Channel

2 = Business Channel

3 = Social Media Channel

4 = Technology Channel

5 = World Channel

6 = Other (No category)

\*/

data popularity\_v3;

set popularity\_v2;

if data\_channel\_is\_lifestyle=1 then

data\_channel=0;

else if data\_channel\_is\_entertainment=1 then

data\_channel=1;

else if data\_channel\_is\_bus=1 then

data\_channel=2;

else if data\_channel\_is\_socmed=1 then

data\_channel=3;

else if data\_channel\_is\_tech=1 then

data\_channel=4;

else if data\_channel\_is\_world=1 then

data\_channel=5;

else

data\_channel=6;

drop data\_channel\_is\_lifestyle data\_channel\_is\_entertainment

data\_channel\_is\_bus data\_channel\_is\_socmed data\_channel\_is\_tech

data\_channel\_is\_world;

run;

/\* Reviewing Response Variable to determine Quality Level Grouping into Category Variable \*/

proc univariate data=popularity\_v3 plot;

var shares;

histogram / normal;

run;

/\* Create Categorical Variable that classifies the Shares records into 4 categories, Popularity Levels, based off Quartiles of shares variable

Q1 = 946

Q2 = 1400

Q3 = 2800

Q4 = > 2800

Popularity Level

0 = Low

1 = Medium

2 = High

3 = Very High

\*/

data popularity;

set popularity\_v3;

if (shares < 947) then

pop=0;

else if ((shares > 946) and (shares < 1401)) then

pop=1;

else if ((shares > 1400) and (shares < 2801)) then

pop=2;

else if (shares < 2800) then

pop=3;

else

pop=4;

drop shares;

label timedelta="Duration"

n\_tokens\_content="Title Words"

num\_imgs="Images"

num\_videos="Videos"

pop="Popularity"

weekday="Weekday"

data\_channel="Channel";

run;

/\* Verify that Shares was broken into 4 equal categories based off the quartiles, using proc freq. Each category is roughly a 4th of the population. \*/

proc freq data=popularity;

run;

proc print data=popularity(obs=300);

run;

/\* Check Correlation of Entire Dataset\*/

proc corr data=popularity;

var timedelta n\_tokens\_content num\_imgs num\_videos weekday data\_channel pop;

run;

/\* Capture a random sample of Dataset for Efficient Computing purposes. n=200\*/

title2 'Simple Random Sampling';

proc surveyselect data=popularity method=srs n=500 out=popularity\_sample seed=74100719;

run;

proc freq data=popularity\_sample;

run;

title 'Weekday Scatterplots';

proc sgscatter data=popularity\_sample;

matrix timedelta n\_tokens\_content num\_imgs num\_videos weekday data\_channel

/ diagonal=(histogram) group=pop;

run;

proc logistic data=popularity\_sample;

class weekday;

model pop = timedelta n\_tokens\_content num\_imgs num\_videos weekday data\_channel

data\_channel /risklimits;

run;

proc logistic data=popularity\_sample;

class weekday;

model pop = weekday /risklimits;

run;

proc logistic data=popularity\_sample;

model pop = timedelta n\_tokens\_content num\_imgs num\_videos weekday data\_channel / selection=stepwise sle=.05 sls=.05;

run;

ods graphics on;

proc logistic data=popularity\_sample plots=EFFECT plots=ROC;

model pop = weekday / outroc = rocout;

output out=estimated predicted=estprob l=lower95 u=upper95;

roc 'Weekday' weekday;

run;

proc logistic data=popularity\_sample plots=EFFECT plots=ROC;

model pop = timedelta / outroc = rocout;

output out=estimated predicted=estprob l=lower95 u=upper95;

roc 'Time Delta' timedelta;

run;

proc logistic data=popularity\_sample plots=EFFECT plots=ROC;

model pop = num\_imgs / outroc = rocout;

output out=estimated predicted=estprob l=lower95 u=upper95;

roc '# Images' num\_imgs;

run;