Central News Theorem

Predicting Online News Popularity

Abstract

In this study, we want to investigate what features about articles, such as topic or number of links, best predict online popularity. This issue is relevant because of the phenomenon with fake news spreading across social media, and understanding popularity predictors can mitigate the effects of fake news. To identify what features best predict popularity, we run a logistic regression model on a dataset about articles published by Mashable. Our final model is fitted with the logit link stepwise BIC procedure. The model has an AUC score of .697 and predicts that popular articles are more likely to be published on the weekend, have more keywords and links, be more subjective and opinionated, and be about technology or social media.

1. Background and Significance

Social media permeates our society. By studying how people interact with content shared on social media, we can gain valuable insight into how information spreads, what content is most valuable for a company to advertise on, and how writers and publishers can increase their audiences. This issue has particular relevance because of current events surrounding the proliferation of fake news and election influencing campaigns in the United States, Europe, and Latin America. Previous study of factors leading to the popularity of articles shared on Twitter has established that the content of the article is one of the most important predictors, but is not sufficient on its own to predict the number of shares a given article will receive [1].

We consider factors leading to the probability that an article shared on Mashable.com will become popular using a binomial logistic regression model with particular interest in two research questions:

(1) Are certain categories of predictor variables (i.e. day of the week published or natural language processing) more likely to predict the popularity of an online news article?
(2) Can we use regression analysis to deliver similar results to machine learning techniques?

2. Data

2.1 Data Description

For our study, we used the Online News Popularity dataset from the University of California Irvine Machine Learning Repository [2]. The data is originally from the paper *A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News* [3].

Table 2: List of attributes by category.

Feature	Type (#)
Words	
Number of words in the title	number (1)
Number of words in the article	number (1)
Average word length	number (1)
Rate of non-stop words	ratio (1)
Rate of unique words	ratio (1)
Rate of unique non-stop words	ratio (1)
Links	
Number of links	number (1)
Number of Mashable article links	number (1)
Minimum, average and maximum number of shares of Mashable links	number (3)
Digital Media	
Number of images	number (1)
Number of videos	number (1)
Time	
Day of the week	nominal (1)
Published on a weekend?	bool (1)

Feature	Type (#)
Keywords	
Number of keywords	number (1)
Worst keyword (min./avg./max. shares)	number (3)
Average keyword (min./avg./max. shares)	number (3)
Best keyword (min./avg./max. shares)	number (3)
Article category (Mashable data channel)	nominal (1)
Natural Language Processing	
Closeness to top 5 LDA topics	ratio (5)
Title subjectivity	ratio (1)
Article text subjectivity score and its absolute difference to 0.5	ratio (2)
Title sentiment polarity	ratio (1)
Rate of positive and negative words	ratio (2)
Pos. words rate among non-neutral words	ratio (1)
Neg. words rate among non-neutral words	ratio (1)
Polarity of positive words (min./avg./max.)	ratio (3)
Polarity of negative words (min./avg./max.)	ratio (3)
Article text polarity score and its absolute difference to 0.5	ratio (2)

Target	Type (#)
Number of article Mashable shares	number (1)

Source: Fernandes, Kelwin, Pedro Vinagre, and Paulo Cortez. "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News." Progress in Artificial Intelligence Lecture Notes in Computer Science, 2015, 535-46. doi:10.1007/978-3-319-23485-4_53.

The authors of the paper processed over 39,000 articles from Mashable, published between Jan. 7 2013 and Jan. 7 2015, and extracted 60 summary information and features. The table to the right shows a list of all the features provided in the dataset. Every feature is treated as a potential predictor variable except for the number of article Mashable shares, which is our response of interest. Exploratory graphs are included in Appendix A.

2.2 Data Cleaning

Before starting our model selection process, we removed predictor variables that are either not useful for our analysis, redundant, or hard to interpret from the dataset. While no values are explicitly missing from the dataset, we removed cases that appear to have missing values, such as 0 for length of article. Our trimmed and cleaned dataset has 35 predictor variables and a sample size of roughly 32,000.

3. Methods and Results

3.1 Multiple Linear Regression vs. Binomial Regression

In our initial analysis, we used the response variable number of shares with a multiple linear regression model. The response variable does not follow a normal distribution, so we employed

a Box-Cox transformation and elected to use the natural log transformation on the response variable in our model. Yet even after this transformation and the addition of interaction terms, our best model performed poorly, with an adjusted R^2 value of 0.13. Therefore we created a new binomial response variable, article popularity, by splitting the number of shares variable at the median and assigning the values 1 and 0 to articles above and below the median shares, respectively. We proceeded with our model selection process using our new binomial response variable.

3.2 Model Fitting Procedure

We first separated our data into a training and independent testing set for cross validation to prevent overfitting. We used random sampling to split our cleaned data into ⅔ training and ⅓ testing. Then, with our training set, we identified multicollinearity problems using variance inflation factor (VIF). With a threshold of 8, decided after observing the VIF values and running correlations tests, we eliminated the predictors rate positive words, rate negative words, and published on Thursday from our dataset.

Finally, to find a final model that best predicts popularity that is parsimonious, we compared four different models. Because our response variable is now binary (popular or not popular), models of interest include the (1) logit link using BIC stepwise procedure, (2) logit link using AIC stepwise procedure, (3) c-log-log link using BIC stepwise procedure, and (4) interaction model of the best first order model. We fitted each of these models with the training set.

3.3 Model Evaluation

To determine the performance of each model, we evaluated each with the test set. Number of predictors, AIC score, deviance, AUC, accuracy, sensitivity, and specificity are all possible criteria for evaluating model fit. For our research question, we are most interested in number of predictors, AUC, and accuracy. We want less predictors in our model for easy interpretation, and we want high AUC and accuracy for high prediction power for both popular and not popular articles.

	# Preds	AIC	Deviance	AUC	Accuracy	Sensitivity	Specificity
BIC	**16**	25247	25213	**.696**	**.647**	**.689**	.596
AIC	21	25232	25188	**.696**	**.647**	.686	.601
C Log Log	27	25451	25395	.685	.640	.677	.595
Interaction	25	**25058**	**25006**	**.700**	**.648**	.675	**.616**

Based on the evaluation criteria shown in the chart above, the BIC logistic regression model seems to be our best model. Though the interaction model has slightly higher AUC and accuracy, and by the likelihood ratio test, better represents the data, we select the BIC model because it is more parsimonious.

3.4 Model Diagnostics and Final Model Interpretation

Before finalizing our model, we checked model diagnostics using plots (plots in Appendix B). From the plots, case 129, 8903, and 10473 seem to be outliers. We removed them from our dataset and retrained the model, which increased our AUC to .697 and accuracy to .649.

Our final model includes the predictor variables in the table to the right. Note that each predictor is significant at $\alpha = .01$. The variable n tokens content is the length of the article; num hrefs is the number of links on an article; kw avg min, kw avg max, and kw avg avg are the average shares on all articles containing the worst performing, best performing, and average keyword for a given article; self reference avg shares is the average number of shares on articles linked in a given article; and global subjectivity and title sentiment polarity are natural language processing variables. Subjectivity is rated on a scale from 0 to 1, where 0 is the most objective and 1 is the most subjective. Polarity is on a scale from -1 to 1 where -1 is the most negative, 1 is the most positive, and 0 is neutral.

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.604e+00  1.175e-01 -13.647  < 2e-16 ***
n_tokens_content               1.850e-04  3.770e-05   4.907 9.24e-07 ***
num_hrefs                      9.383e-03  1.567e-03   5.987 2.14e-09 ***
num_keywords                   4.780e-02  8.957e-03   5.336 9.48e-08 ***
data_channel_is_entertainment -6.680e-01  4.376e-02 -15.264  < 2e-16 ***
data_channel_is_socmed         1.011e+00  7.654e-02  13.209  < 2e-16 ***
data_channel_is_tech           4.604e-01  4.544e-02  10.132  < 2e-16 ***
data_channel_is_world         -5.041e-01  4.609e-02 -10.938  < 2e-16 ***
kw_avg_min                    -1.908e-04  3.597e-05  -5.306 1.12e-07 ***
kw_avg_max                    -7.566e-07  1.547e-07  -4.892 9.96e-07 ***
kw_avg_avg                     3.061e-04  1.978e-05  15.473  < 2e-16 ***
self_reference_avg_sharess     7.999e-06  1.194e-06   6.697 2.13e-11 ***
weekday_is_friday              2.501e-01  4.318e-02   5.792 6.97e-09 ***
weekday_is_saturday            9.687e-01  7.050e-02  13.740  < 2e-16 ***
weekday_is_sunday              8.044e-01  6.375e-02  12.617  < 2e-16 ***
global_subjectivity            7.778e-01  1.797e-01   4.329 1.50e-05 ***
title_sentiment_polarity       2.903e-01  5.792e-02   5.012 5.38e-07 ***
```

4. Conclusion

4.1 Discussion

In response to our first research question, our final model contains a diverse group of predictor variables including content type, natural language processing (NLP) variables, and publication details, suggesting that no given category of predictor has more significance than another. This suggests that the number of shares on an article depends on a network of predictor variables rather than one in particular.

In response to our second research question, we compared the results of our regression model to Fernandes, Vinagre, and Cortez's machine learning models. Their best performing model was generated by a random forest, which is a non-parametric process, meaning they did not have to account for the skewed distribution of the response variable. The accuracy and AUC for the random forest model were 0.67 and 0.73, respectively, compared with our 0.65 accuracy and 0.70 AUC. This suggests that our regression model is closely comparable with the random forest method, without necessitating the computing power to complete a random forest on a dataset of this size.

4.2 Further Considerations

To take this study further, we would first try to do a more thorough cleaning and investigation of possible null values in the NLP category of variables. For NLP variables, the 0s may either represent nulls or be intentional. Also, because our data originates from a machine learning paper, we would look into the advantages and disadvantages of machine learning versus regression. Lastly, an interesting extension would be to observe how our model performs for non-Mashable articles. Would the predictors be the same? For example, if we looked into New York Times articles, would a different topic like world be more predictive? It would also be interesting to investigate whether other factors, such as current events or trends, help predict popularity.

5. References

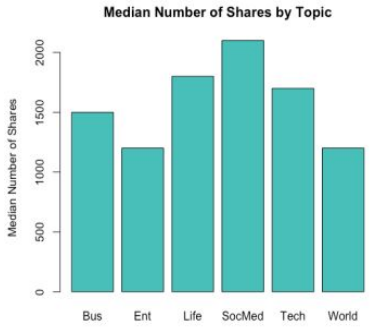
[1] Roja Bandari, Sitaram Asur, and Bernardo A. Huberman. “The Pulse of News in Social Media: Forecasting Popularity.” *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, n.d.

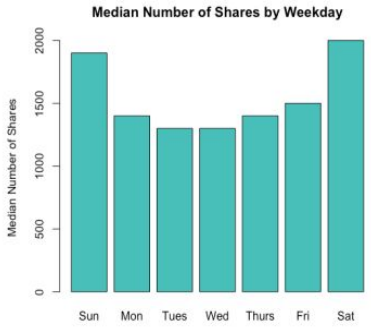
[2] "UCI Machine Learning Repository: Online News Popularity Data Set." UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set. http://archive.ics.uci.edu/ml/datasets/Online News Popularity.

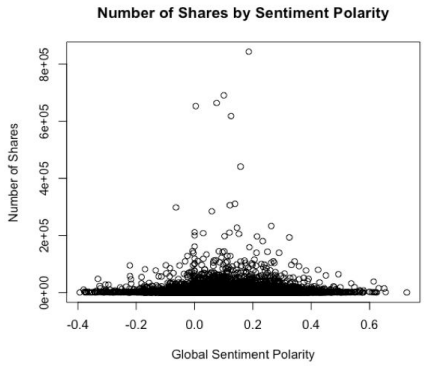
[3] Fernandes, Kelwin, Pedro Vinagre, and Paulo Cortez. "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News." Progress in Artificial Intelligence Lecture Notes in Computer Science, 2015, 535-46. doi:10.1007/978-3-319-23485-4_53.

6. Appendix

A. Data Exploration Graphs







B. Model Diagnostic Plots