Online News Popularity

Objective

The impact that modern social media has had on our daily life, and society in general, is undeniable. The major means that social media influence the public is by creating topics of discussion in the society. In this project, I use a set of predictors to predict if an article would become popular or not. And try to give authors some actionable suggestions to improve the online news popularity. I will use some methods, such as forward stepwise selection, to identify a subset of the predictors related to the response. Then given a set of predictor values (e.g., keywords, number of words in the title, rate of positive words in an article), we could also give some actionable suggestions to improve the popularity of a post.

Data Curation

This dataset summarizes a heterogeneous set of features about articles published by Mashable (www.mashable.com) in a period of two years. The response is the number of shares in social networks. There are 39797 instances and 61 attributes. No missing values. One of the features, url, is URL of the article, which is not useful. So, I delete this feature. This dataset is likely to have multicollinearity problem. For example, there are eight predictors related to weekday. It is possible for collinearity to exist between these predictors even if no pair of predictors has a particularly high correlation. Instead of correlation matrix, I use VIF. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.

So, I calculate VIF values of each predictors and drop the predictor with largest VIF value if it is more than 10. Finally, there are 38463 observations and 51 predictors.

Table 1: Removed Predictors based on VIF values

Redundant Predictors	VIF value
n_non_stop_words	2086.212
weekday_is_monday	Inf
weekday_is_saturday	Inf
LDA_04	981618118
n_unique_tokens	13681.29
rate_negative_words	18.21274
self_reference_avg_sharess	19.14889
kw_max_min	11.30715

Plus, in this dataset there are some observations with very high response. So I delete high leverage observations since high leverage points tend to have a sizable impact on the estimated regression line. And this problem is more pronounced in multiple regression settings with more than two predictors.

Results

First, I use forward stepwise selection and backward stepwise selection.

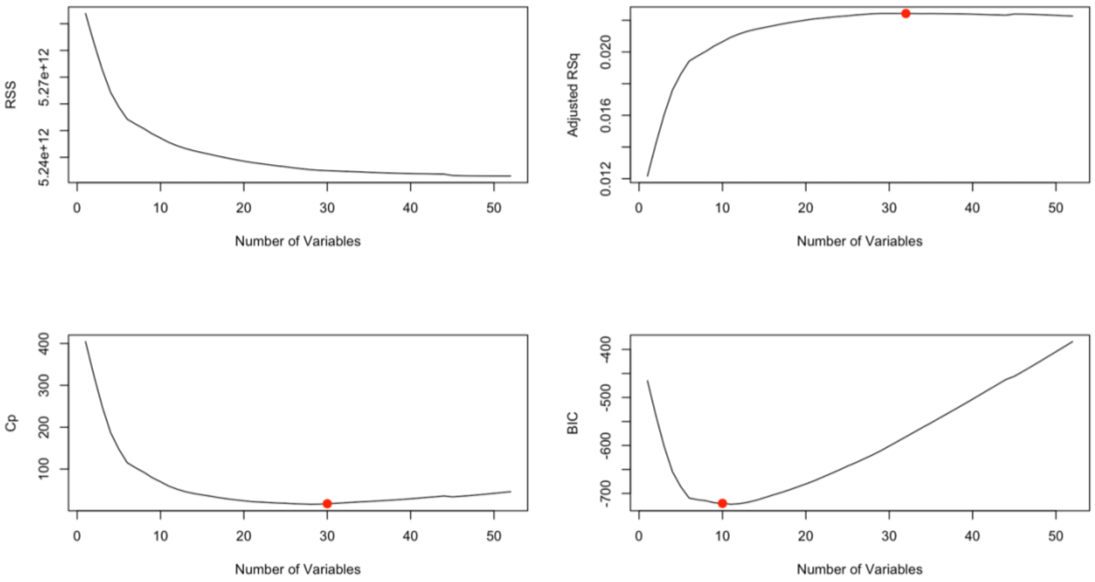

Fig. 1 Feature Selection

BIC: 1 timedelta 2 n_tokens_title 3 num_hrefs 4 data_channel_is_entertainment 5 kw_min_avg 6 kw_max_avg 7 kw_avg_avg 8 self_reference_min_shares 9 global_subjectivity 10 avg_negative_polarity

Cp: 1 timedelta 2 n_tokens_title 3 n_tokens_content 4 num_hrefs 5 num_self_hrefs 6 average_token_length 7 data_channel_is_lifestyle 8 data_channel_is_entertainment 9 data_channel_is_bus 10 data_channel_is_tech 11 kw_min_max 12 kw_min_avg 13 kw_max_avg 14 kw_avg_avg 15 self_reference_min_shares 16 self_reference_max_shares 17 weekday_is_tuesday 18 weekday_is_wednesday 19 weekday_is_thursday 20 weekday_is_friday 21 LDA_02 22 global_subjectivity 23 global_rate_positive_words 24 min_positive_polarity 25 avg_negative_polarity 26 abs_title_subjectivity 27 abs_title_sentiment_polarity

News feeds are pre-tagged with category labels describing the content. So design a t-density score to represent a prior distribution on the popularity of categories.

$$t - density = \frac{Number\ of\ Shares}{Number\ of\ Links}$$

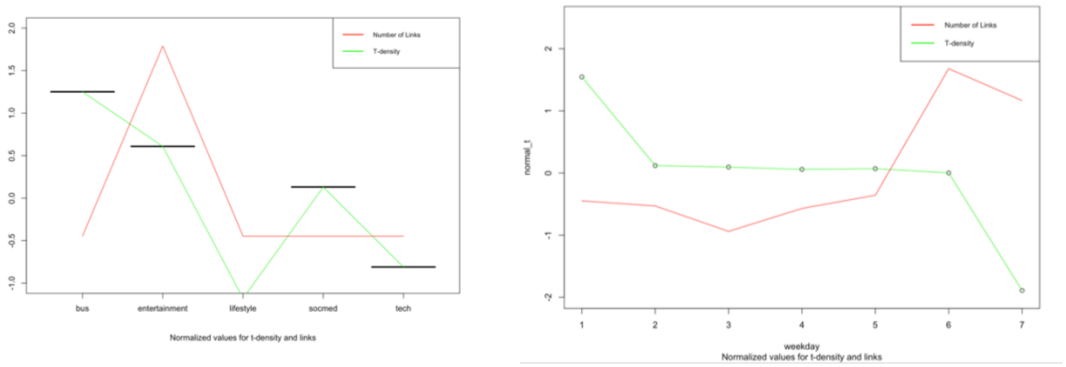

Fig. 2 t-density

We can see that news related to Entertainment receives more shares on average and thus has a prominent presence in our dataset. Plus, categories (such as Social Media) with low number of published links but higher rates of t-density may have a niche following and loyal readers. Similarly, we can see that news posted on Monday may receive more shares on average. However, news posted on Saturday and Sunday may have less shares although they may have more links. Plus, we can see the distribution of number of words in the title, showing that the number of words is centered at 10.

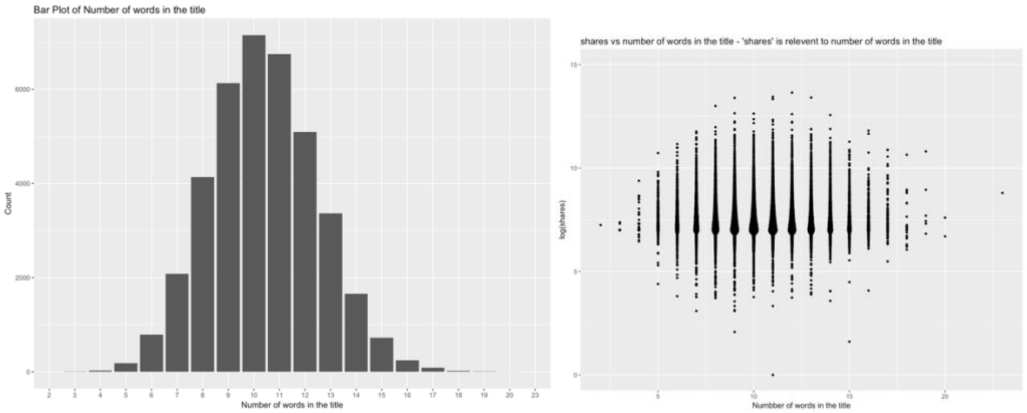

Fig. 3 number of word in the title

1) Classification

To predict if an article will be popular or not, I fit the model using training data and then examine how well it predicts the held out data. I create training data corresponding to the observations in the first year. Then I used Logistic Regression, LDA, QDA, RF, SVM and KNN. The best obtained result (78.47%) is 14 percentage points higher than QDA. While not perfect, an interesting discrimination level, higher than 75%, was achieved.

Table 2: Comparison of models (best values in **bold**)

Model	Accuracy
Logistic Regression	0.7076331
Linear Discriminant Analysis (LDA)	0.7072536
Quadratic Discriminant Analysis (QDA)	0.6837255
Random Forest (RF)	**0.7876316**
Support Vector Machine (SVM)	0.71466035
K-Nearest Neighbors (KNN)	0.6996097

And kw_avg_avg, LDA_02, kw_max_avg, LDA_01 are the most important predictors in this model.

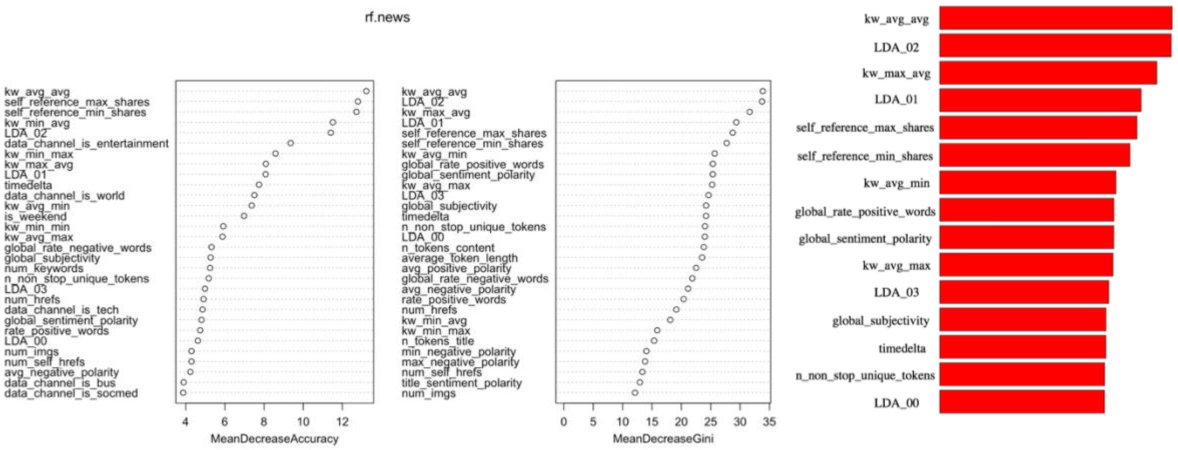

Fig. 4 Variable importance plots for the *news* data.

2) Prediction

I used KNN, GAM, RF, Ridge Regression, LASSO, PCA and PLS. The best obtained result (test MSE=126437588) is almost 100 percentage points lower than PLS. While the test MSE for RF is still large, this is a good result compared with the other models.

Table 3: Comparison of models (best values in **bold**)

Model	Test MSE
K-Nearest Neighbors (KNN)	203394969
Smoothing spline (GAM)	156557987
Random forest (RF)	**126437588**
Ridge Regression	206494352
LASSO	201407596
Principal Components Regression (PCA)	158072532
Partial Least Squares (PLS)	255595567

3) Inference: give authors some actionable suggestions to improve popularity

I perform LASSO to shrink some coefficient estimates toward zero since it could yield a sparse model. This model is simpler and more interpretable for me to give suggestions. I also use random forest and plot variable importance to identify which predictors are important for popularity.

1. Increase the number of words in the title.
2. Increase the number of links in the article.
3. Let the article be more subjective.
4. Decrease the average polarity of negative words.
5. The number of words in the title should be from 7 to 13.
6. Post news on Monday.
7. Write the article related to entertainment.