Uncovering the Relationship Between Online News Characteristics and Popularity

By Sinclaire Schuetze and Valerie Tseng

Wellesley College

Dataset & Motivation

Online News Popularity Data Set

- Mashables Data from UC Irvine Machine Learning Repository
- Roughly 40,000 observations (articles)
- 61 attributes collected
- Uses 58 predictive attributes to predict 1 target attribute, # of shares
  - Other 2 attributes are URL and id of article
- Predictive attributes include category of news channel, day of week published, keywords, and polarity of content, etc

WHY THIS DATASET?

PREVALENCE OF ONLINE NEWS

- More than 40% of Americans get their online news through FB
- 9 out of 10 Americans get some form of news digitally
- About half of all newspaper and newsletter readers prefer to read their news online
- (From Pew Research)

IMPACT OF ONLINE NEWS

- Misinformation is often associated with online news
- Online news is accompanied by online advertising, and this industry expected to reach $460 billion in revenue by 2024
- Ability to access news online has political and economic impacts
- (From Pew Research)

RQ: What characteristics of news articles lead to greater popularity?

Data Analysis: Model Building & Data Cleaning

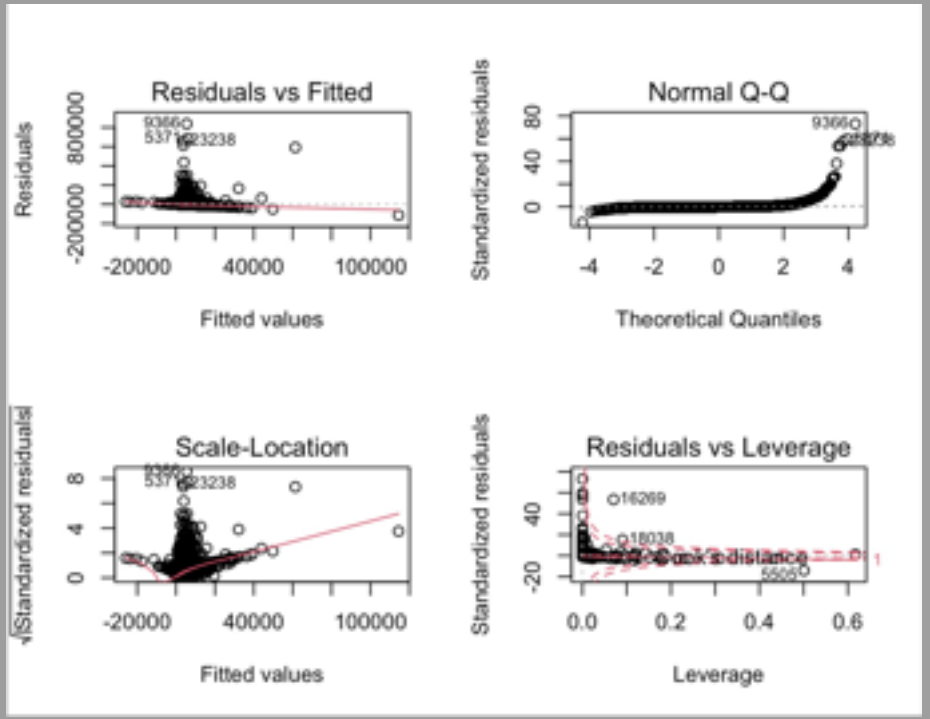
CHECKING MULTICOLLINEARITY

Using Variance Inflation Factor w/ threshold 10, we removed a total of 5 variables.

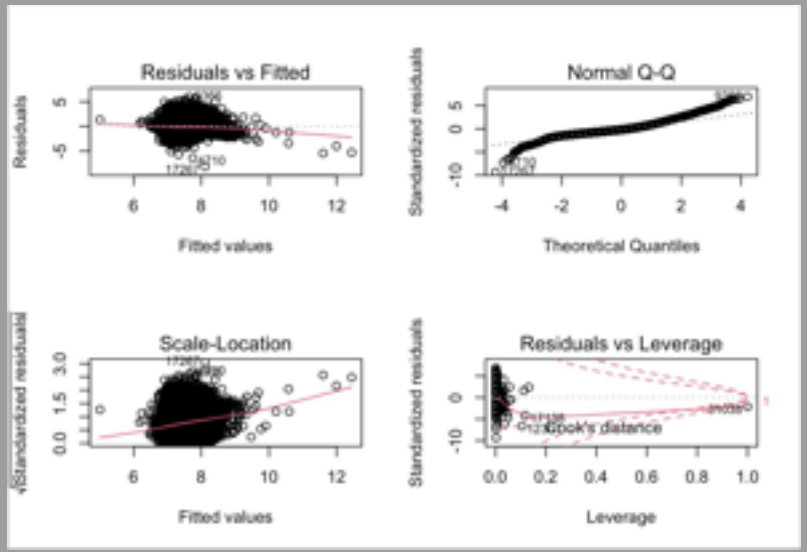
- N_non_stop_unique_tokens (rate of unique non stop words in content)
- N_unique_tokens
- Self_reference_avg_shares (avg share of referenced articles)
- rate_positive_words
- Kw_max_min (best keyword in terms of min shares)

CHECKING ASSUMPTIONS

Linearity, Independence
Constant variance, Normal distribution



Took log of shares and removed outliers



AIC/BIC Model Selection

	AIC Model	BIC Model
Num of Variables	38 Variables	29 Variables
R-Squared	0.1233	0.1226
Adj R-Squared	0.1225	0.1219
CV-score	0.7802742	0.7590034

Conclusion: We chose the BIC Model since it is more parsimonious and the R-squared values and CV scores between the two models are very similar

Final Model

- Decided to not include interaction terms
- Once again removed outliers

R-Squared: 0.1219

Adjusted R-squared: 0.1213

Coefficients:

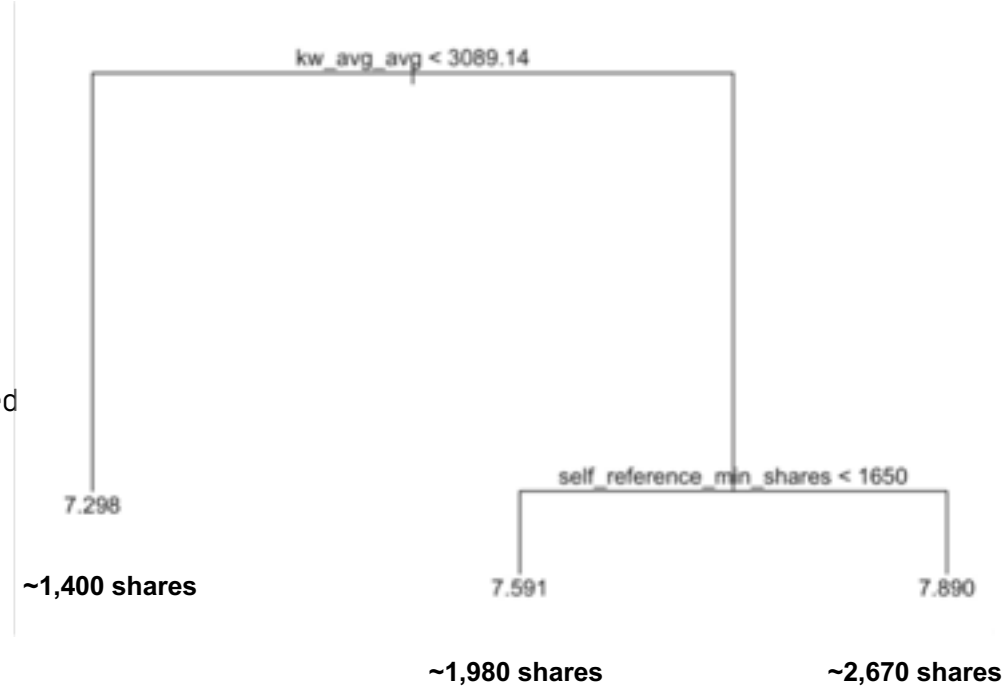
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	6.57924205279	0.05025192050	130.925	< 0.0000000000000002	***
n_tokens_title	0.00773268616	0.00215818674	3.583	0.000340	***
n_tokens_content	0.00004804364	0.00001146752	4.190	0.0000280125982845	***
num_hrefs	0.00419255658	0.00048256665	8.688	< 0.0000000000000002	***
num_self_hrefs	-0.00772391458	0.00132084468	-5.848	0.0000000050229479	***
num_imgs	0.00232235283	0.00059829381	3.882	0.000104	***
average_token_length	-0.05310592191	0.00712265858	-7.456	0.0000000000000911	***
num_keywords	0.01275381907	0.00273659280	4.660	0.0000031650794236	***
data_channel_is_entertainment	-0.19690791954	0.01358269909	-14.497	< 0.0000000000000002	***
data_channel_is_socmed	0.25128557067	0.02033531268	12.357	< 0.0000000000000002	***
data_channel_is_tech	0.16219981375	0.01382189628	11.735	< 0.0000000000000002	***
data_channel_is_world	-0.11638586256	0.01391994099	-8.361	< 0.0000000000000002	***
kw_min_min	0.00081193769	0.00007929771	10.239	< 0.0000000000000002	***
kw_avg_min	-0.00002995029	0.00000879300	-3.406	0.000660	***
kw_min_max	-0.00000036728	0.00000008768	-4.189	0.0000280933311484	***
kw_avg_max	-0.00000021876	0.00000005482	-3.991	0.0000659856831744	***
kw_min_avg	-0.00004673290	0.00000552614	-8.457	< 0.0000000000000002	***
kw_max_avg	-0.00003876279	0.00000170954	-22.674	< 0.0000000000000002	***
kw_avg_avg	0.00032124663	0.00000948613	33.865	< 0.0000000000000002	***
self_reference_min_shares	0.00000171342	0.00000025517	6.715	0.0000000000190803	***
self_reference_max_shares	0.00000052077	0.00000012602	4.132	0.0000359632086949	***
weekday_is_tuesday	-0.06923471824	0.01282710907	-5.398	0.0000006795157O6	***
weekday_is_wednesday	-0.06547673452	0.01280217736	-5.114	0.0000003160368705	***
weekday_is_thursday	-0.05875938388	0.01289377715	-4.557	0.0000051997338208	***
weekday_is_saturday	0.21778140764	0.01936816721	11.244	< 0.0000000000000002	***
weekday_is_sunday	0.21272340526	0.01851891678	11.487	< 0.0000000000000002	***
global_subjectivity	0.42706126099	0.05100724814	8.373	< 0.0000000000000002	***
min_positive_polarity	-0.30649966253	0.06767916391	-4.529	0.0000059516230231	***
title_subjectivity	0.06589698620	0.01596224354	4.128	0.0000366197064862	***
title_sentiment_polarity	0.08026592500	0.01725204313	4.653	0.0000032892044948	***
abs_title_subjectivity	0.14010581422	0.02723282272	5.145	0.0000002691655714	***

Interpretation of Model Output

- Saturday and Sunday (Monday as baseline): 23% increase
  - Tuesday: 7% decrease
- Social media (Business as baseline): 29% increase
  - Entertainment: 18% decrease
- Global subjectivity: 53% increase
- Title subjectivity: 15% increase
- Min polarity of positive words: 26% decrease

Regression Tree

- Regression tree utilized to confirm results of regression model
- Three nodes after pruning
  - Average number of shares of the keywords prior to the date of publication
  - Minimum number of shares of the referenced articles within Mashable
- Both of these variables are also included in the multiple linear regression model



Conclusions

- The importance of global subjectivity indicates that highly subjective or sensationalized news is most popular
- Readers care most about social media related news and least about entertainment news
- Tuesday is the worst day to post, weekends are the best
- Posts with greater min. positive polarity are less popular, so perhaps negative news is more engaging
- If a specific article contains popular keywords or references other popular news articles, then this specific article is expected to be more popular as well

Areas of Improvement

- Determine why R-squared was so low
  - Better predictors, data, etc?
- How do these trends change by country?
- Complete analysis with more widespread news sources (Mashable - limited audience)

THANK YOU!