Online News Popularity - Analysis Report

Nagraj Rao, Linhan Cai, Jennifer Hoang

2021-11-27

Our GitHub Repo: **LINK HERE**

Summary

Online articles have become a primary source of news in the digital age. In order to understand factors associated with online news popularity, we examined factors associated with higher shares per day (log transformed) using a multiple linear regression analysis in a dataset containing 36,644 observations (K. Fernandes and Cortez 2015). Our final model, derived using backward model selection, achieved an R-squared score of 0.2132. Additional features that are not among our explanatory variables appear to explain a large portion of variability in the shares per day. Further analysis will be required to better understand the factors which associate with online news popularity.

Introduction

The online news market space has grown rapidly in recent decades, leading to increased competition between traditional news outlets and non-traditional digital news sources. Understanding the factors associated with popularity of news articles online is vital for guiding publishing strategies of news agencies in order for them to remain competitive in the online news space. Here, we assessed factors associated with online news popularity using a public dataset with statistics from originally published on Mashable (www.mashable.com) in 2015 (K. Fernandes and Cortez 2015).

Methods

EDA During EDA on the raw data, we try merging data channels and the weekday columns into one column, then explore data type of the variables in the data set and summary Statistics for each variable. By creating correlation plot and correlation matrix, we try to find out the important features and in the end, we explore bar graph showing how number of shares vary based on topic, how number of shares vary based on day of the week, and try histogram showing how number of shares vary based on day of the week. Finally we made the matrix to pick the features whose coefficient is larger than 0.7. The code used to perform the 2 versions EDA can be found here.

Data Cleaning Upon examining the data during EDA, we observed that the distribution of the response variable Shares was highly right-skewed. Furthermore, we observed that articles had been published at different time points prior to data acquisition, which could confound the number of shares attained per article. To address both of these factors, we transformed the data by creating a Shares per Day features (Shares / Days since Publication), followed by a performing a log transformation of Shares per Day. Lastly, outliers in the log Shares per Day were removed using the Winsorization method, where we defined outliers to be values lower than the 1% percentile and greater than the 99% percentile. Data cleaning was performed using Python (Van Rossum and Drake 2009) and Pandas (team 2020).

Statistical Analysis A Multiple Linear Regression model was used to understand what factors are associated with online news popularity. We estimated six versions of this model using "log_shares_per_day" as our dependent variable until we arrived at a regression where all features were statistically significant at the 95% confidence level. This was compared to both forward and backward selection models using VIF scores, R-Squared, and the time taken to run each model to arrive at the best model, which in our case was backward selection model. Finally, we plot the distribution of residuals to visually assess if it follows a normal distribution.

The R programming language (R Core Team 2019) and the following R packages were used to perform the statistical analysis outlined in this section: broom (Robinson, Hayes, and Couch 2021), car (Fox and Weisberg 2019), docopt (de Jonge 2020), tidyverse (Wickham et al. 2019).

Results and Discussion

EDA Through exploratory data analysis, we determined that some of the features were not informative to answering our question or contained many missing values. We find out the summary statistics for each variable, the features correlation greater than 0.7 and the distributions of shares vary based on day of the week and topics. 2 versions of EDA code can be found here and here.

Bar graph showing how number of shares vary based on topic. Several of the topics were reviewed by plotting the distribution of the shares based on different topic, Except others, "Business" and "Tech" take the largest 2 shares. The shares left "Entertainment,""Lifestyle" and "World" takes almost equal shares, and "Social media" takes the smallest shares.

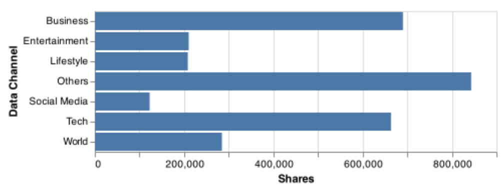

Figure 1: Figure 1. Distribution of Shares Based on Topics

Histogram showing how number of shares vary based on day of the week. Several of the weekdays were reviewed by plotting the distribution of the shares based on different weekdays, Except others, "Wednesday," "Monday" and "Saturday" take the largest 3 shares. The "Tuesday,""Friday" and "Thursday" takes almost equal shares, and "Sunday" takes the smallest shares.

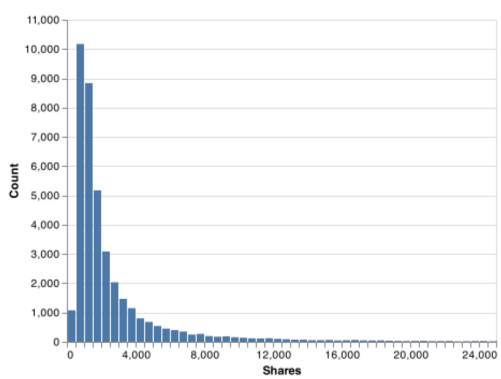

Figure 2: Figure 2. Distribution of Shares

Correlation plot showing the strength, direction, and form of the relationship between 2 features. It shows the kw_avg_avg(Avg. keyword (avg. shares)) has its strongest correlation with shares. The second most correlated feature to shares value is kw_max_max(Best keyword (max. shares)) which quite make sense.

Before model testing, data cleaning was done to address the findings of non-informative features, class imbalance, NAN values. And we calculate the shares_per_day and remove outliers, this code can be found here.

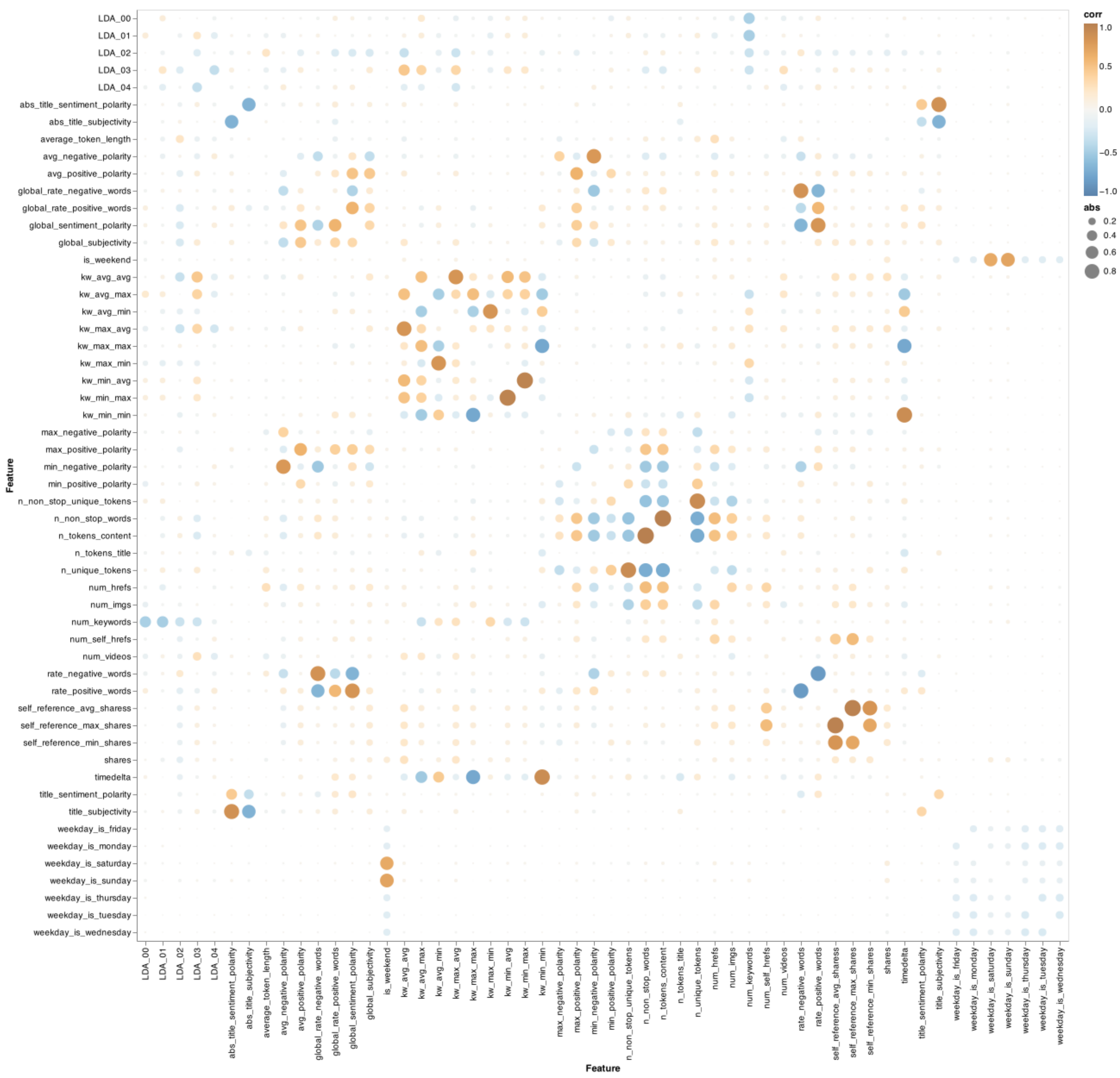

Figure 3: Figure 3. Distribution of Features Correlation Matrix

Statistical Analysis The results of our best model, derived from Backward Model Selection is shown below:

Table 1: Table 1. Backstep Model Results

term	estimate	std.error	statistic	p.value	conf.low	conf.high	is_sig
(Intercept)	1057.9608458	135.1025999	7.830796	0.0000000	793.1563567	1322.7653349	TRUE
n_non_stop_unique_tokens	0.6314715	0.1462972	4.316360	0.0000159	0.3447253	0.9182178	TRUE
global_subjectivity	0.4132123	0.0720999	5.731107	0.0000000	0.2718946	0.5545299	TRUE
data_channel_is_world	0.3451603	0.0308061	11.204293	0.0000000	0.2847796	0.4055410	TRUE
data_channel_is_socmed	0.3106069	0.0310984	9.987885	0.0000000	0.2496533	0.3715605	TRUE
is_weekend	0.2553258	0.0158811	16.077322	0.0000000	0.2241984	0.2864532	TRUE
rate_negative_words	0.2525924	0.0986547	2.560368	0.0104599	0.0592267	0.4459581	TRUE
data_channel_is_tech	0.2396748	0.0308791	7.761706	0.0000000	0.1791509	0.3001987	TRUE
abs_title_subjectivity	0.1550362	0.0333235	4.652463	0.0000033	0.0897214	0.2203511	TRUE
data_channel_is_lifestyle	0.1286847	0.0338225	3.804704	0.0001422	0.0623917	0.1949776	TRUE
title_subjectivity	0.1072980	0.0194179	5.525739	0.0000000	0.0692385	0.1453575	TRUE
title_sentiment_polarity	0.0731971	0.0213573	3.427259	0.0006103	0.0313362	0.1150580	TRUE
n_tokens_title	0.0577943	0.0026293	21.981160	0.0000000	0.0526408	0.0629477	TRUE
average_token_length	0.0316563	0.0221576	1.428686	0.1531025	-0.0117732	0.0750859	FALSE
num_keywords	0.0144281	0.0033875	4.259282	0.0000206	0.0077886	0.0210676	TRUE
num_hrefs	0.0041843	0.0006199	6.749714	0.0000000	0.0029692	0.0053993	TRUE
kw_avg_avg	0.0003112	0.0000129	24.155547	0.0000000	0.0002859	0.0003364	TRUE
kw_max_min	0.0000589	0.0000046	12.908945	0.0000000	0.0000499	0.0000678	TRUE
n_tokens_content	0.0000328	0.0000201	1.629319	0.1032536	-0.0000067	0.0000722	FALSE
self_reference_min_shares	0.0000024	0.0000007	3.433831	0.0005957	0.0000010	0.0000037	TRUE
kw_avg_max	0.0000015	0.0000001	20.166349	0.0000000	0.0000014	0.0000017	TRUE
kw_max_max	0.0000006	0.0000000	17.769520	0.0000000	0.0000005	0.0000007	TRUE
self_reference_max_shares	0.0000006	0.0000004	1.638230	0.1013818	-0.0000001	0.0000013	FALSE
kw_min_max	-0.0000011	0.0000001	-10.558313	0.0000000	-0.0000013	-0.0000009	TRUE
self_reference_avg_sharess	-0.0000016	0.0000010	-1.660603	0.0968014	-0.0000035	0.0000003	FALSE
kw_min_avg	-0.0000333	0.0000069	-4.831641	0.0000014	-0.0000468	-0.0000198	TRUE
kw_max_avg	-0.0000410	0.0000023	-17.944360	0.0000000	-0.0000454	-0.0000365	TRUE
kw_avg_min	-0.0004220	0.0000280	-15.057736	0.0000000	-0.0004769	-0.0003670	TRUE
num_imgs	-0.0018143	0.0008181	-2.217744	0.0265780	-0.0034178	-0.0002108	TRUE
num_videos	-0.0026074	0.0014325	-1.820119	0.0687486	-0.0054152	0.0002004	FALSE
num_self_hrefs	-0.0210983	0.0016414	-12.854241	0.0000000	-0.0243154	-0.0178812	TRUE
min_negative_polarity	-0.1263717	0.0256066	-4.935115	0.0000008	-0.1765614	-0.0761821	TRUE
max_positive_polarity	-0.1428581	0.0308727	-4.627321	0.0000037	-0.2033694	-0.0823467	TRUE
data_channel_is_bus	-0.1450772	0.0321352	-4.514581	0.0000064	-0.2080631	-0.0820913	TRUE
n_non_stop_words	-0.2661874	0.1354177	-1.965676	0.0493431	-0.5316095	-0.0007653	TRUE

term	estimate	std.error	statistic	p.value	conf.low	conf.high	is_sig
global_rate_positive_words	-1.5283499	0.5910915	-2.585640	0.0097234	-2.6869040	-0.3697957	TRUE
n_unique_tokens	-1.6969142	0.1690463	-10.038164	0.0000000	-2.0282491	-1.3655793	TRUE
global_rate_negative_words	-4.3456218	1.1788783	-3.686234	0.0002279	-6.6562530	-2.0349907	TRUE
LDA_00	-1057.8152585	135.0944797	-7.830189	0.0000000	-1322.6038318	-793.0266852	TRUE
LDA_04	-1057.9194454	135.0922015	-7.831092	0.0000000	-1322.7035535	-793.1353373	TRUE
LDA_01	-1057.9347375	135.0925029	-7.831188	0.0000000	-1322.7194363	-793.1500388	TRUE
LDA_03	-1058.0458221	135.0924559	-7.832013	0.0000000	-1322.8304287	-793.2612156	TRUE
LDA_02	-1058.1926285	135.0952899	-7.832935	0.0000000	-1322.9827898	-793.4024673	TRUE

Table 2: Table 2. Backstep Model Model Performance

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.2059672	0.2051079	1.044721	239.6737	0	42	-56803.93	113695.9	114072.8	42355.58	38807	38850

Overall, our model has an R-Squared of 0.2051. This seems like a low R-Squared, particularly given the large number of features included in the model and their statistical significance at alpha = 0.05. This indicates that other variables that are not currently included in the model explain a large portion of the variability in our data. There is not much we can do about this problem, beyond including some interaction variables to assess if there are any interaction effects.

Finally, we plot a distribution of the residuals, which looks normally distributed, one of the assumptions of a linear regression.

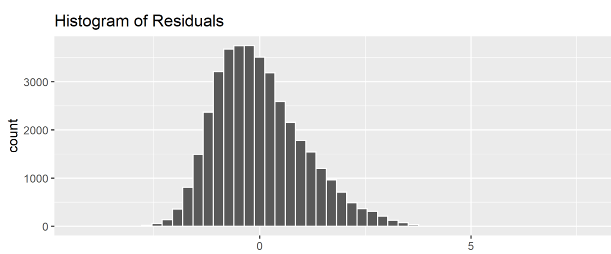

Figure 4: Figure 4. Histogram of Residuals

As next steps, we need to consider if interaction terms can help improve model performance, and perform rigorous statistical tests for the remaining assumptions of a multiple linear regression model – heteroscedasticity and normality of residuals.

References

de Jonge, Edwin. 2020. *Docopt: Command-Line Interface Specification Language*. https://CRAN.R-project.org/package=docopt.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

K. Fernandes, P. Vinagre, and P. Cortez. 2015. *A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News*. Coimbra, Portugal: Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence. https://archive-beta.ics.uci.edu/ml/datasets/online+news+popularity.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2021. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

team, The pandas development. 2020. *Pandas-Dev/Pandas: Pandas* (version latest). Zenodo. https://doi.org/10.5281/zenodo.3509134.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.