FEBRUARY 2023

ONLINE NEWS POPULARITY

Predict The Number of Shares in Social Networks.

Summary.

Mashable wants to understand what drives people to share their articles on social media.

Findings:

- The dataset is limited to predict if an article will become viral or not - but it can explain what affects the share numbers to an extend.
- Meta keywords are still an important and significant factor when it comes to shareability of an article. The average historical share of the average keyword is the most important feature when it comes to the meta keywords.
- The articles with higher text subjectivity gets more shares - meaning people tend to share articles that contains personal opinion rather than factual information.
- The articles with more external links gets more shares.
- The articles that links to “most shared” internal articles also gets higher number of shares.
- Mondays and weekends are the best times to share an article.
- Entertainment, world and business articles get less shares - the rest of the topics don’t have any effect on the results at all.

Outline

BUSINESS PROBLEM

DATA

METHOD

ITERATIONS

RESULTS

CONCLUSIONS

LIMITATIONS & NEXT STEPS

Business Problem.

Mashable is a global, multi-platform media and entertainment company. With the increase focus of the organisation on the online channels, the marketing team wants to understand what resonates with their readers and what drives them to share the Mashable articles with their own network.

Basically Mashable wants to understand what works for them, what not - and can we predict if an article will become viral.

Data.

The dataset I used came from the online articles published by **www.mashable.com** between **Jan 2013** and **Jan 2015**. It contains 58 features that about the articles, as well as the total number of shares they received.

These features include **metadata** information (title, keywords etc.), **sentiment analysis statistics** (rate of positive words, text subjectivity, title polarity etc.), **SEO features** (number of internal and external shares etc.), and some **other features** around the article (number of **images, videos,** the **day** it's published, the content **channel** it belongs to etc.).

Methods.

Used d**escriptive statistics** and visualizations.

Took an **iterative** and **train-test split approach** to the regression model.

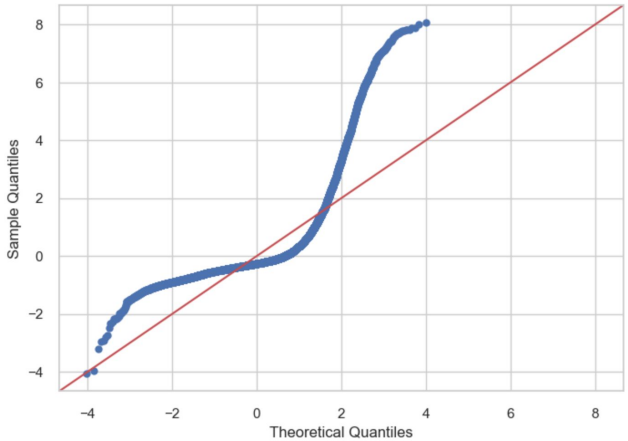
Used both **OLS model** from **statsmodels** and **LinearRegression** from **sklearn**.

Used **plots to visualise the relationships** between the variables and **behaviours of the residuals**.

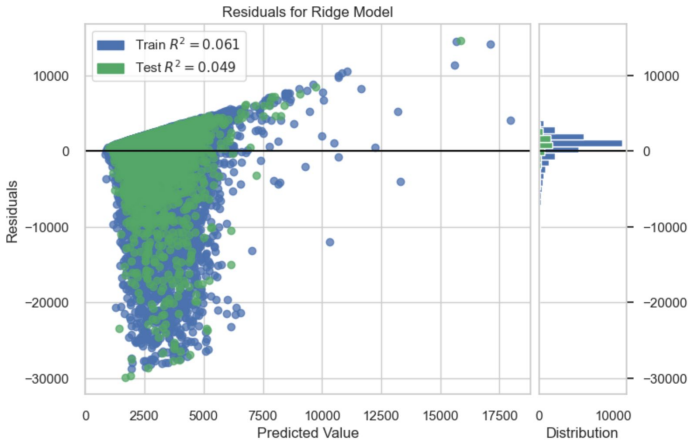
Iterations

Model 1.

The first model didn't use any data transformations but eliminated the outliers of the "Y". It confirmed at least one of the variables is affecting the number of shares significantly. It only explains 6% of the variations in the share number - but it fails to satisfy the primary assumptions for the multilinear regression - and its difference between the train and test datasets are around 25%.



Not linear



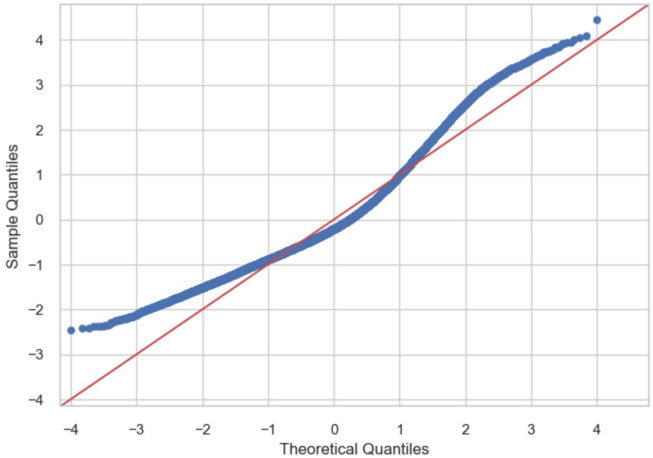
Heteroscedastic

Iterations.

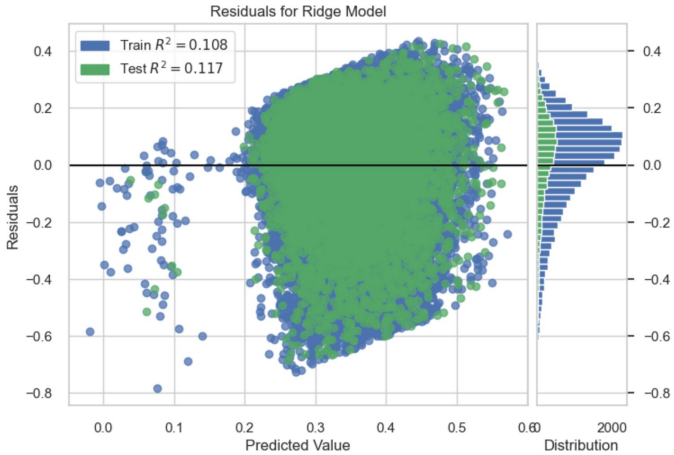
Model 0	Model 1	Model 2	Model 3	Model 4
Baseline model - uses the mean of the observed the values of Y.	First model - eliminated the outlierss The adj. R2 is 6% Residuals are not linear and are heteroscedastic.	Removed the irrelevant features. The adj. R2 is 5.9% No significance in the residual behaviour,	Performed log transformation. The adj. R2 is 10.8% The residuals are behaving more linear and are homoscedastic.	Normalised the data using min-max scaler. Value between -1 and 1 The adj. R2 is 10.8% Final model.

Model 4.

The last model have more linearity between the variables, it satisfies primary assumptions - the difference between train and test splits is below 1% - and the model is a better fit which means it's more reliable. However, it only explains 10.8% of the variations in the share numbers



More linear



Very close to being homoscedastic

Results

Let’s Address The Elephant In The Room...

The **R2 score** is way **too low** to be considered as a good predictive tool. I managed to increased the score by 80% from the first iteration and created a reliable model - but it can only explain 10.8% of the observed data - due to the limitations in the dataset.

On the other hand, it statistically proved that some features have significant effect on the share numbers. That’s why, I used the model more in an explanatory way.

kw_avg_avg	0.75
global_subjectivity	0.08
self_reference_min_shares	0.08
num_external_hrefs	0.08
is_weekend	0.06
num_videos	0.05
title_sentiment_polarity	0.04
num_imgs	0.04
abs_title_subjectivity	0.02
weekday_is_monday	0.01
title_subjectivity	0.01
global_sentiment_polarity	-0.02
data_channel_is_bus	-0.03
data_channel_is_lifestyle	-0.03
num_self_hrefs	-0.06
data_channel_is_world	-0.07
data_channel_is_entertainment	-0.08
average_token_length	-0.15
kw_avg_max	-0.33

The top feature (the average ranked keyword)

The most significant positive features

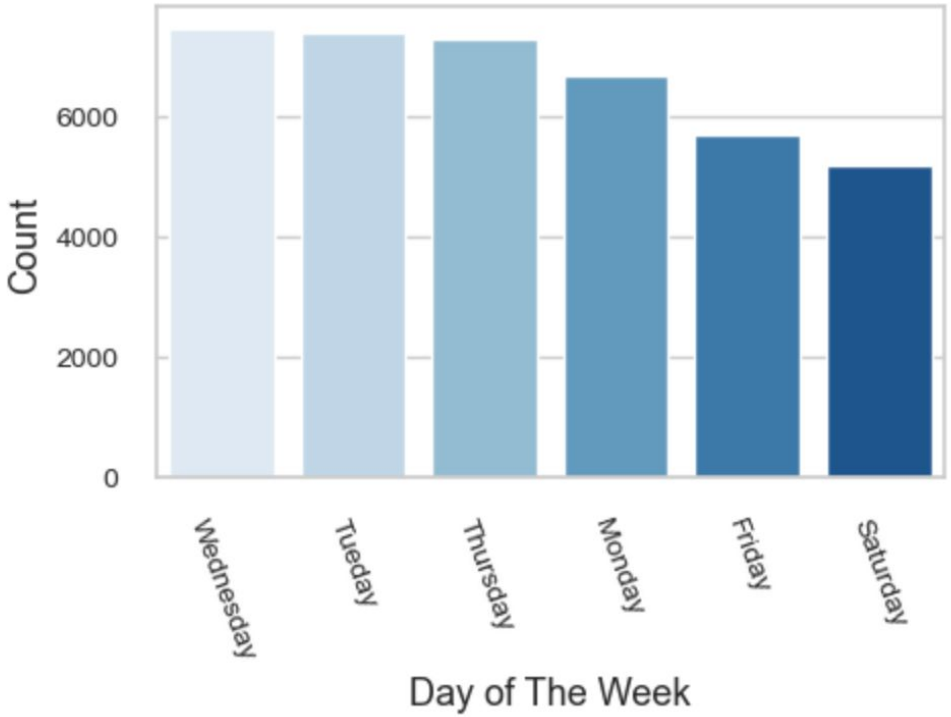
The most significant negative features

Top Features.

kw_avg_avg	0.75
global_subjectivity	0.08
self_reference_min_shares	0.08
num_external_hrefs	0.08

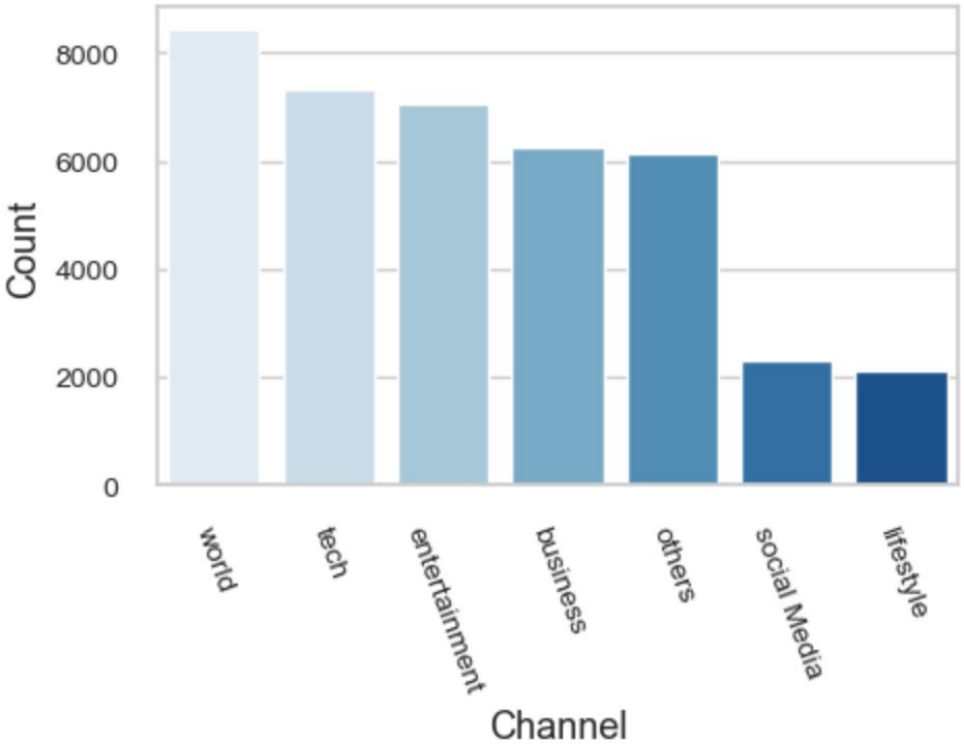
- The **higher the avg. position** of the avg. keyword is by far the most feature.
- The articles and titles with **higher text subjectivity** gets more shares - meaning people **tend to share articles that contains personal opinion** rather than factual information.
- The articles with **more external links** gets more shares.
- The articles that **links to “most shared” internal articles** also gets higher number of shares.

Day of The Week.



Mondays and weekends are the best times to share an article. However, the least number of articles are published that day.

The Article Category/Channel.



Entertainment, world and business articles get less shares - the rest of the topics don't have any effect on the results at all.

Bottom Features.

num_self_hrefs	-0.06
data_channel_is_world	-0.07
data_channel_is_entertainment	-0.08
average_token_length	-0.15
kw_avg_max	-0.33

- **Higher ranked keywords** get less shares.
- The **longer titles** get less share,
- The articles with **more internal links** gets less shares - I recommend focusing on the external links.

Conclusions.

The model should be used to identify the significant features affecting the total number of shares.

Recommendations:

- Sticking to average keyword tags are better than choosing high ranked keywords is a better strategy - perhaps high ranked keywords have more competitive environment.
- Opinion pieces that are strongly worded gets more shares. I recommend having opinion pieces more often.
- The shorter titles are better than the longer ones.
- I recommend having more external links than the internal links.
- Articles published on Mondays and Weekends have higher chance to become viral.
- Finally, I recommend publishing less articles in entertainment, business and world channels.

Limitations & Next Steps.

The biggest limitation and the challenge of this study was the dataset itself. It does a good job to identify some of the important features but is limited to fully explain the dynamics behind the Mashable's readers' article sharing behaviour.

Looking at the features in the dataset, it lacks external factors and the group behaviours of individuals.

We know that the relevancy of the articles to the current topic and trends, and whether or not someone influential shares the article impacts its total shares. For the next studies, the dataset needs new features to reflects these.

Thank you!

Hazal Aydin

h.aydinhazal@gmail.com

https://www.linkedin.com/in/hazalaydin/

https://github.com/hazal-aydin