Online News Popularity

Data Description

- Kaggle : Predicting_shares
- resource : Mashable (A global media and entertainment company)
- Observations : 39797 (61 attributes)

Data Description

Article type

- data_channel_is_lifestyle: Is data channel 'Lifestyle'?
- data_channel_is_entertainment: Is data channel 'Entertainment'?
- data_channel_is_bus: Is data channel 'Business'?
- data_channel_is_socmed: Is data channel 'Social Media'?
- data_channel_is_tech: Is data channel 'Tech'?
- data_channel_is_world: Is data channel 'World'?

Publication day

- weekday_is_monday: Was the article published on a Monday?
- weekday_is_tuesday: Was the article published on a Tuesday?
- weekday_is_wednesday: Was the article published on a Wednesday?
- weekday_is_thursday: Was the article published on a Thursday?
- weekday_is_friday: Was the article published on a Friday?
- weekday_is_saturday: Was the article published on a Saturday?
- weekday_is_sunday: Was the article published on a Sunday?
- is_weekend: Was the article published on the weekend?

Data Description

Polarity

- avg_positive_polarity: Avg. polarity of positive words
- min_positive_polarity: Min. polarity of positive words
- max_positive_polarity: Max. polarity of positive words
- avg_negative_polarity: Avg. polarity of negative words
- min_negative_polarity: Min. polarity of negative words
- max_negative_polarity: Max. polarity of negative words
- title_subjectivity: Title subjectivity
- title_sentiment_polarity: Title polarity
- abs_title_subjectivity: Absolute subjectivity level
- abs_title_sentiment_polarity: Absolute polarity level

Data Description

Content/title

- n_tokens_title: Number of words in the title
- n_tokens_content: Number of words in the content
- n_unique_tokens: Rate of unique words in the content
- num_hrefs: Number of links
- num_self_hrefs: Number of links to other articles published by Mashable
- num_imgs: Number of images
- num_videos: Number of videos
- average_token_length: Average length of the words in the content
- num_keywords: Number of keywords in the metadata

Thesis

A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News

- Purpose : How to predict the populaity of online news and optimize it
- result : 1000 articles improve 15% popularity by the improvement
- What we want to do?

Statistics

The distribution is flat from Monday to Friday.
On weekends, the frequency is lower.

Research question 1:
The treatment effect of the day of the week.

Variable	Obs	Mean	Std. Dev.	Min	Max
Monday	39,644	.1680204	.3738891	0	1
Tuesday	39,644	.186409	.3894413	0	1
Wednesday	39,644	.1875441	.3903526	0	1
Thurersday	39,644	.1833064	.3869224	0	1
Friday	39,644	.1438049	.3508962	0	1
Saturday	39,644	.0618757	.2409327	0	1
Sunday	39,644	.0690395	.2535244	0	1
is weekend	39,644	.1309151	.3373118	0	1

Statistics

We divide the data into two categories:

- Work
- Entertainment (Note in yellow)

Research question 2:
We investigate the interaction between the day of the week and the above two categories.

Variable	Obs	Mean	Std. Dev.	Min	Max
data_channel_is_lifestyle	39,644	.0529462	.223929	0	1
data_channel_is_entertainment	39,644	.1780093	.3825254	0	1
data_channel_is_bus	39,644	.1578549	.3646095	0	1
data_channel_is_socmed	39,644	.0585965	.2348709	0	1
data_channel_is_tech	39,644	.1852992	.388545	0	1
data_channel_is_world	39,644	.2125668	.4091288	0	1

Statistics

Research question 3:
Are these variables important?
How to improve shares using these variables?

Research question 4:
Is negativity better?

Variable	Obs	Mean	Std. Dev.	Min	Max
# of title	39,644	10.39875	2.114037	2	23
# of content	39,644	546.5147	471.1075	0	8474
# of imgs	39,644	4.544143	8.309434	0	128
# of videos	39,644	1.249874	4.107855	0	91
# of keywords	39,644	7.223767	1.90913	1	10

Model

01

Lasso

Using Lasso to investigate the coefficients of variables

02

Random Forest

Using Random Forest to study the improtance of the variables

Prelimary result in Lasso

The variables with large coefficient (in absolute value):

- Number of words in the title
- Average length of the words in the content
- Data channel type
- The day of week
- Subjectivity
- Positive polarity
- Negative polarity

Reference

K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.

Thank You!