Data Science Project: Digital Media Consumption & Sharing

Published on the UC Irvine Machine Learning Repository, the Online News Popularity dataset provides a sample of online news articles' attributes and metadata (words in the title, number of 'positive' words in the title, day of the week that the article is published, etc) and popularity (as indicated by the total number of times that the article is 'shared' on social media). The articles were collected from Mashable over a period of two years between 2013 and 2015. The dataset is cited below.

> 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.

Features

The dataset has sixty features and one response (number of shares). Two of its features (`url`, the article's url, and `timedelta`, the time between the article being published and the original authors pulling it into their dataset) are non-predictive.

Conveniently, the dataset authors have already engineered a number of useful features to aid analysis. These features include `n_tokens_title` (the total number of words in the title of the article), `n_tokens_content` (the total number of words in the body of the article), `n_non_stop_words` (the 'rate' of non-stop words in the article text), `num_videos` (the number of videos in the article), channel data (which channels—for example, tech, world, business, lifestyle, etc—was the article published under), time features such as `is_weekend` (was the article published on the weekend), and dozens of other features.

For a complete list and explanation of the dataset's figures and response, as described by the dataset authors themselves, see the Appendix.

Response

The response of the dataset is `shares`, which denotes the number of times that the article was shared on social media at the time that the authors pulled the article into their dataset.

Samples

The dataset contains 39,644 samples. There is no missing data.

Appendix: Detailed Feature Descriptions

Note: the following information is pulled from the dataset description in the UC Irvine Machine Learning Repository (see citation above).

Statements marked with an asterisk `*` are calculated by the authors and not objective. I therefore might omit these features from my analysis.

Statements marked with a plus sign + are not fully explained by the authors' description file, and therefore likely will be omitted from my analysis.

0. `url`: URL of the article
1. `timedelta`: Days between the article publication and the dataset acquisition
2. `n_tokens_title`: Number of words in the title
3. `n_tokens_content`: Number of words in the content
4. `n_unique_tokens`: Rate of unique words in the content
5. `n_non_stop_words`: Rate of non-stop words in the content
6. `n_non_stop_unique_tokens`: Rate of unique non-stop words in the content
7. `num_hrefs`: Number of links
8. `num_self_hrefs`: Number of links to other articles published by Mashable
9. `num_imgs`: Number of images
10. `num_videos`: Number of videos
11. `average_token_length`: Average length of the words in the content
12. `num_keywords`: Number of keywords in the metadata
13. `data_channel_is_lifestyle`: Is data channel 'Lifestyle'?
14. `data_channel_is_entertainment`: Is data channel 'Entertainment'?
15. `data_channel_is_bus`: Is data channel 'Business'?
16. `data_channel_is_socmed`: Is data channel 'Social Media'?
17. `data_channel_is_tech`: Is data channel 'Tech'?
18. `data_channel_is_world`: Is data channel 'World'?
19. `kw_min_min`: Worst keyword (min. shares)
20. `kw_max_min`: Worst keyword (max. shares)
21. `kw_avg_min`: Worst keyword (avg. shares)
22. `kw_min_max`: Best keyword (min. shares)
23. `kw_max_max`: Best keyword (max. shares)
24. `kw_avg_max`: Best keyword (avg. shares)
25. `kw_min_avg`: Avg. keyword (min. shares)
26. `kw_max_avg`: Avg. keyword (max. shares)
27. `kw_avg_avg`: Avg. keyword (avg. shares)
28. `self_reference_min_shares`: Min. shares of referenced articles in Mashable
29. `self_reference_max_shares`: Max. shares of referenced articles in Mashable
30. `self_reference_avg_sharess`: Avg. shares of referenced articles in Mashable
31. `weekday_is_monday`: Was the article published on a Monday?
32. `weekday_is_tuesday`: Was the article published on a Tuesday?
33. `weekday_is_wednesday`: Was the article published on a Wednesday?
34. `weekday_is_thursday`: Was the article published on a Thursday?
35. `weekday_is_friday`: Was the article published on a Friday?
36. `weekday_is_saturday`: Was the article published on a Saturday?
37. `weekday_is_sunday`: Was the article published on a Sunday?
38. `is_weekend`: Was the article published on the weekend?
39. `LDA_00`: Closeness to LDA topic 0 +
40. `LDA_01`: Closeness to LDA topic 1 +
41. `LDA_02`: Closeness to LDA topic 2 +
42. `LDA_03`: Closeness to LDA topic 3 +
43. `LDA_04`: Closeness to LDA topic 4 +

44. `global_subjectivity`: Text subjectivity + *
45. `global_sentiment_polarity`: Text sentiment polarity + *
46. `global_rate_positive_words`: Rate of positive words in the content + *
47. `global_rate_negative_words`: Rate of negative words in the content + *
48. `rate_positive_words`: Rate of positive words among non-neutral tokens + *
49. `rate_negative_words`: Rate of negative words among non-neutral tokens + *
50. `avg_positive_polarity`: Avg. polarity of positive words + *
51. `min_positive_polarity`: Min. polarity of positive words + *
52. `max_positive_polarity`: Max. polarity of positive words + *
53. `avg_negative_polarity`: Avg. polarity of negative words + *
54. `min_negative_polarity`: Min. polarity of negative words + *
55. `max_negative_polarity`: Max. polarity of negative words + *
56. `title_subjectivity`: Title subjectivity + *
57. `title_sentiment_polarity`: Title polarity + *
58. `abs_title_subjectivity`: Absolute subjectivity level + *
59. `abs_title_sentiment_polarity`: Absolute polarity level + *
60. `shares`: Number of shares (target)