DATA SCIENCE: ONLINE NEWS POPULARITY

Abstract: This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years (Mashable's primary focus is on technology, lifestyle and entertainment news. It is a global, multi-platform media and entertainment company).

Internship Task: Please carry out an Exploratory Data Analysis and create compelling story based on the given dataset; also predict the number of shares in social networks (popularity).

Attribute Information:

1. url: URL of the article (non-predictive)
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. weekday_is_monday: Was the article published on a Monday?
20. weekday_is_tuesday: Was the article published on a Tuesday?
21. weekday_is_wednesday: Was the article published on a Wednesday?
22. weekday_is_thursday: Was the article published on a Thursday?

23. weekday_is_friday: Was the article published on a Friday?
24. weekday_is_saturday: Was the article published on a Saturday?
25. weekday_is_sunday: Was the article published on a Sunday?
26. is_weekend: Was the article published on the weekend?
27. LDA_00: Closeness to LDA topic 0
28. LDA_01: Closeness to LDA topic 1
29. LDA_02: Closeness to LDA topic 2
30. LDA_03: Closeness to LDA topic 3
31. LDA_04: Closeness to LDA topic 4
32. global_sentiment_polarity: Text sentiment polarity
33. global_rate_positive_words: Rate of positive words in the content
34. global_rate_negative_words: Rate of negative words in the content
35. avg_positive_polarity: Avg. polarity of positive words
36. avg_negative_polarity: Avg. polarity of negative words
37. title_sentiment_polarity: Title polarity
38. shares: Number of shares (TARGET)

Citation: 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. http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity