ONLINE NEWS POPULARITY

Introduction

Popular news can help the social media company attract more people. So they can make more profits. So if one company can predict whether news or articles will be popular before them are published, they will definitely try their best to get the information.

In this project, I am analyzing articles popularity using target variable - shares which indicated number of times that article is shared. This can help online news companies who can check popularity of article before publishing it. Also, it can help businesses which rely heavily on social networks to grow and flourish

Data Collection

I am using Online News Popularity dataset from UCI Machine Learning Repository to predict Online News Popularity. The dataset consists of 39644 records and 61 attributes 58 are predictive attributes and 1 is target variable which is Number of shares.

Data Description

Attribute Information:

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

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)

Data Preprocessing

1. Handling Missing Values : The UCI machine learning repository which hosts the current data set has indicated that the data set has no missing values
2. Further, while checking the dataset for 0 values, I found 0 values in rate\_positive\_words, rate\_negative \_words, and average\_token\_length. Removed all these values while doing data clean up
3. Outliers : From the histograms , boxplots and the pairplots ,it is clear that the following attributes contain outliers

n\_unique\_tokens,n\_non\_stop\_words,n\_non\_stop\_unique\_tokens ,

n\_tokens\_title , num\_hrefs ,num\_self\_hrefs

I have deleted those observations from the data.

1. For the Outliers in the target variable I have used the standard deviation method .

5. From the available 60 attribute , we can drop

URL - Since every URL is unique for each column ,and

timedelta: Days between the article publication and the dataset acquisition

Finally after cleaning the data , dataset contains 37895 rows × 58 columns .