Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)
Department of Computer Science and Engineering (Data Science)

Report on Mini Project

Machine Learning -I (DJ19DSC402)

AY: 2022-23

ONLINE NEWS POPULARITY

NAME: FARIN I. KHAN

SAP ID: 60009210140

Guided By -

PROF. PREETAM VERNEKAR

CHAPTER 1: INTRODUCTION

With the growth of social media and the internet, the popularity of online news articles has assumed a greater significance. Because it directly affects readers and revenue, the popularity of news articles is a crucial factor for news websites. This project is an attempt to build a model for predicting the popularity of online news articles using a number of criteria, including the article's length, subject, and date of publication. The model was trained and tested using machine learning algorithms on a dataset of 7,000 news articles from 'Mashable'.

CHAPTER 2: DATA DESCRIPTION

The 'OnlineNewsPopularity' dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years. The goal is to predict the number of shares in social networks (popularity).

Acquisition date: January 8, 2015

Attribute Information: No. of Attributes: 61 (58 predictive attributes, 1 non-predictive, 1 goal field)

1. url: URL of the article (non-predictive)
2. timedelta: Days between the article publication and the dataset acquisition
3. n_tokens_title: Number of words in the title
4. n_tokens_content: Number of words in the content
5. n_unique_tokens: Rate of unique words in the content
6. n_non_stop_words: Rate of non-stop words in the content
7. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
8. num_hrefs: Number of links
9. num_self_hrefs: Number of links to other articles published by Mashable
10. num_imgs: Number of images
11. num_videos: Number of videos

12. average_token_length: Average length of the words in the content

13. 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)

CHAPTER 3: DATA ANALYSIS

Data analysis show the nature of data. The data has outliers but they cannot be removed as pertaining to the context, they could be novelties, meaning there might be some special cases, some trending news. This could also be seasonal i.e., depended on the particular days of the year or month; hence, date is calculated and new columns are added for the same. We observe that though the outliers don't seem to have a pattern, other ratings of popularity are seen in clusters. Hence, there must be some dependence of the date with the popularity.

To simplify the data and aid in better model building, in an attempt to reduce redundancy, columns having high correlation have been dropped.

The plots between features show that it may not be the right choice to build regression model on the data. Hence, bins with ratings of popularity are created to aid in building a classification model which would also be a better way to comment on popularity measure.

Further analyzing the data, A plot on average number of words for articles across levels of popularity does not show a lot of dependence. Countplot against days of week show week days to be more popular than weekends. Comparing the domain of news, world news has the highest popularity followed by entertainment.

CHAPTER 4: DATA MODELLING

From the data visualization and analysis, we see that the data is not linearly separable hence logistic regression is not used which would otherwise have been good for the outliers. Decision tree classifier gave an accuracy of 30.96% using the kfold cross validation method. Two of the effective ensemble models – Random Forest and XGBoost was trained in an attempt to improve the accuracy also considering the complexity and size of the data. Random forest model gave and accuracy of 43.56% and XGBoost an accuracy of 43.53%. Model was build using SVM kernel next to test if it could perform better as it is known for it's robust characteristics and ability to deal with noise or any other data impurities. Observation is done on rbf kernel and polynomial kernel. The accuracy was 36.74% and 37.43% respectively.

As we can see the Ensemble models have performed the best, performing grid search cv might help in finding the better parameters for them. Hence, GridSearchCV is done on random forest has a lesser computational time than XGBoost. The best accuracy seen is 44.32%.

CHAPTER 4: CONCLUSION

This project aimed to build a model for predicting the popularity of online news articles using various criteria. With the growth of social media and the internet, the popularity of news articles has assumed a greater significance for news websites. 'OnlineNewsPopularity' dataset, which condenses a diverse range of features about articles published by Mashable, was used in the project. The dataset includes 61 attributes, such as the word count in the title and body, the percentage of unique words, the number of links, images, and videos, the topic of the article, and the publication date. Data outliers, which cannot be eliminated because they may be novelties or special cases, were revealed by the data analysis. Both the model's training and testing used machine learning algorithms. The final model that performed the best was the Random Forest model with the parameters - {'max_depth': None, 'min_samples_split': 10, 'n_estimators': 200}. The model will assist news websites in predicting the popularity of their articles and in optimizing their content in light of those factors.