Online News Popularity

Gokulramanan Soundararajan

Abstract

This project is about Online News Popularity. We read many types of news everyday based on our interest and news popularity. However, we never researched about the real factors which makes those news popular and to share more. In this project, mainly focused on factors impacting weekend and weekday’s news and built a classification model to categorize the ranking of news shares. Rank 1 is least shared and 4 is most shared.

Motivation

Main motivation of this project is to analyze the factors which influence the news popularity, and this will help the news publishing firms to improve their business. For example, analyzing the impact of factors like number of images, videos, news released day, news domain, news subjectivity, polarity, can help the news publishing company to change their news covering strategy to satisfy their customers and to attract more customers.

Dataset(s)

Online News Popularity dataset has been taken for this project from USC Machine learning repository. This dataset consists about 40k instances and 60 attributes. Each instance gives the separate news details of 60 attributes. Details are like number of images, videos, news published day, global subjectivity, sentiment polarity, positive subjectivity, negative subjectivity and number of shares of that particular news. More details about all the attributes can be found below.

Dataset path: https://archive.ics.uci.edu/ml/datasets/online+news+popularity

Data Preparation and Cleaning

After aquiring the data from repository basic investigation has been done, like checking the missing values and data types of the attributes. Fortunately, dataset didn’t have any missing values. However, since its really high dimensional data, lot of data exploration was needed to sort out the influencing factors that impact the weekday and weekend’s news.

And also, all of the attributes were in float datatype. To make the conditional check against 1 or 0, some attributes had to be converted into integer.

Research Question(s)

We may think weekend’s news might be more about entertainment and weekday’s news more about technology and business. This exactly is my research question! Comparing the features impact on weekend and weekday’s news. Is our assumption happening really or something different happening? If something is different, news publishing firm may have to take necessary action and analyze the reason.

Second is to build a classification model with most influencing factors. So that, publishing firm can use this model to predict the news popularity for future news.

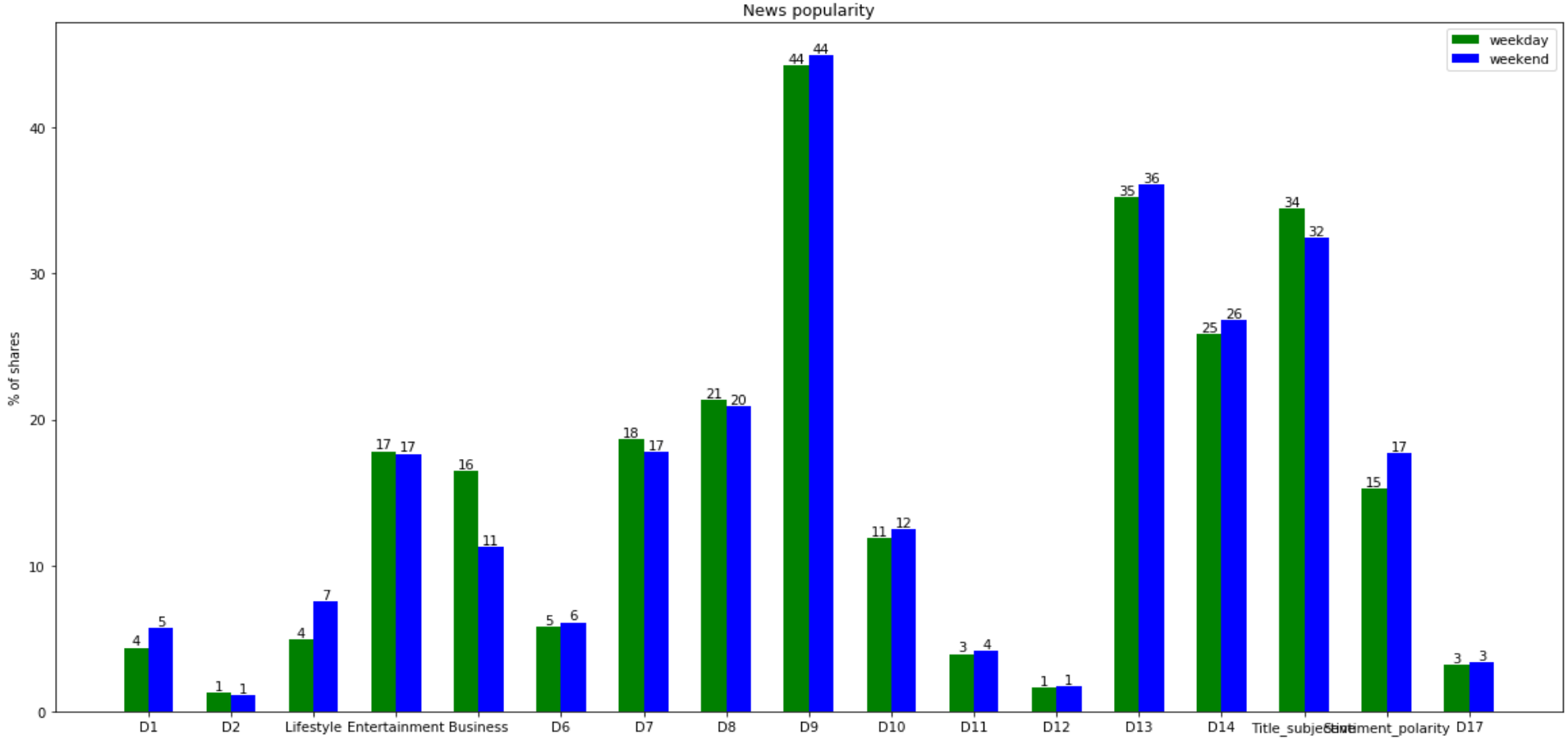
Methods

According to the dataset, target classification attribute(number of news shared) is a continuous variable.

To normalize the distribution of very large shares and very small shares, number of shares attribute has been converted to log value. Which gives the smooth distribution.

Second to avoid the model overfitting of classification, log of news share has been categorized into 4 ranks. 1 – least shared news, 2 – below average shared news, 3 – above average shared news and 4 – most popular news.

Influence of factors on News



Finding - 1

- As we expect, news categories make certain level of impact based on weekend news and weekday news
- First finding,11% of the weekend news are fall under business category, on the other hand 16% of the weekday’s news fall under business category. So, business news are more popular in weekdays than weekend.
- So, news publishing firms may focus about business more on weekdays and less on weekend to attract more customers

Finding - 2

- However, another assumption, entertainment news are more popular in weekend than weekdays, gave another finding
- As per the analysis, entertainment has same level of reach in both weekdays and weekend which is 17%.
- So, news publishing firms should always give equal importance to entertainment news in both weekdays and weekend

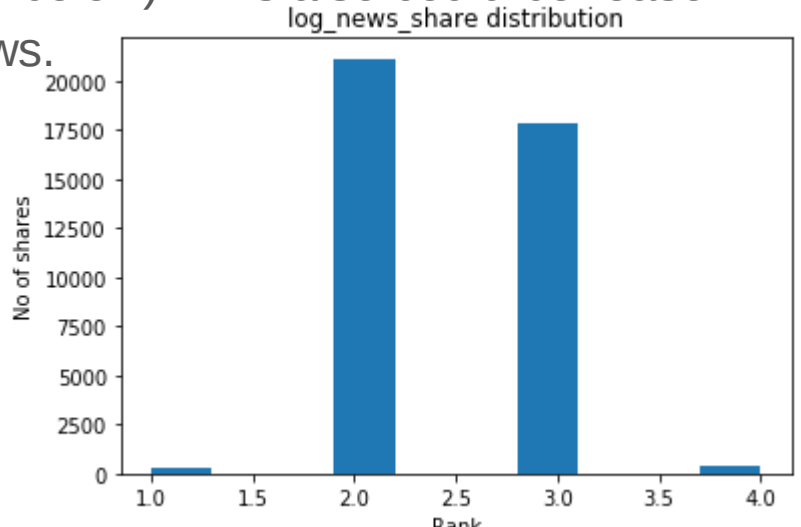
Finding - 3

- And also, another category gives the more importance to weekend; lifestyle.
- As per the analysis, lifestyle news are more popular in weekend than weekday as we expect. 7% of the weekends news are lifestyle, whereas only 4% of the weekdays news are lifestyle.
- So, people are interested about entertainment on weekend and weekday as well. However, they are interested in lifestyle more in weekend than weekday. News publishing firms should consider this factor.

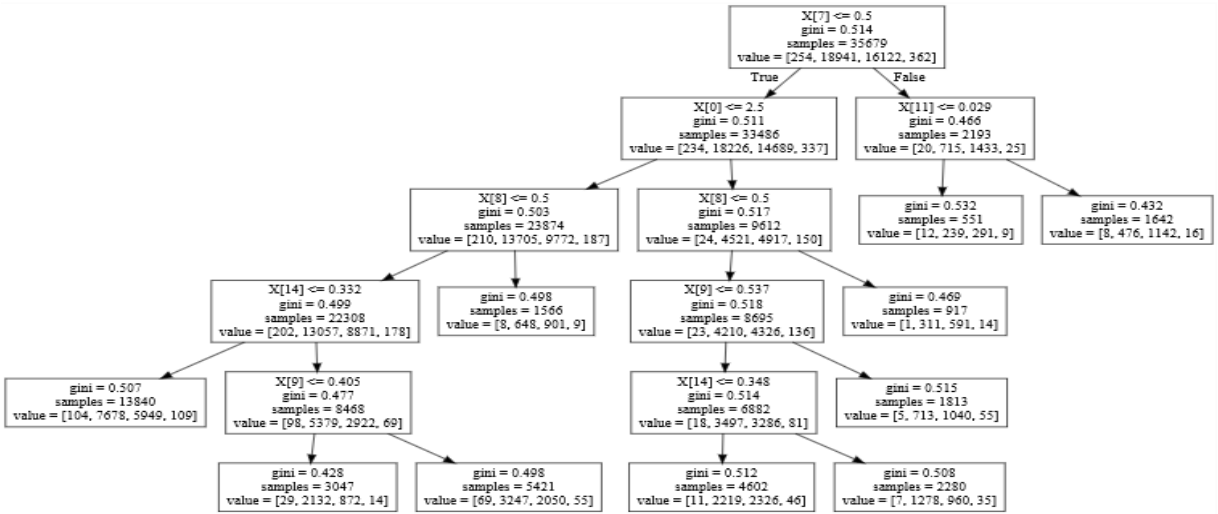
Limitations

Classification model has given around 60% accuracy, because of high dimensionality(high number of features) which lead to model overfitting.

Since the least shared and most share news count are very less. Training data are very less to train the model(distribution is show below). This also could be reason for model underfitting for rank 1 and rank 4 news.



Classifier of News shares rank



Conclusions

According to the pattern and findings, News publishing firms may focus entertainment news both on weekday and weekend equally. Business news less on weekend. Lifestyle news more on weekend. And other categories are equally popular in weekend and weekdays.

And also, firms may use the classification training model to predict the news popularity.