Prediction of Online News Popularity

Midterm Report for DATA1030 Fall 2021 at Brown University

Supervised by Dr. Andras Zsom

Project on GitHub

1. Introduction

Over the past couple of decades, with online news platforms increasingly winning out over physical newspapers, media organizations are relying more and more on analytics and machine learning to understand their reader base, moderate interactions and analyze which content is more likely to generate traffic. A study conducted by the Pew Research Center found that the average number of unique monthly visitors to the top 50 US newspaper websites increased from 8.2 million at the end of 2014 to 13.9 million at the end of 2020 [4]. The pandemic has only accelerated this trend, with traditional advertising revenues falling and platforms having the rely on subscriptions and digital advertising to generate revenues.

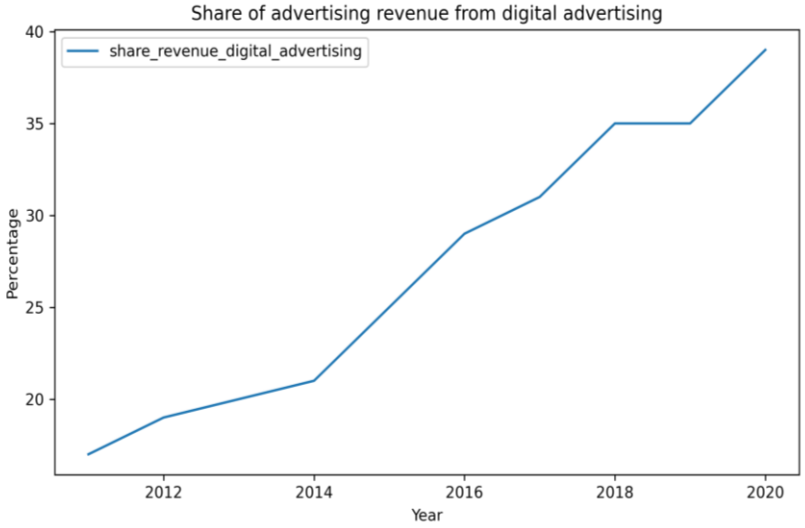

Figure 1: Share of advertising revenue from digital advertising. Data from Pew Research Center [4]

In this context, being able to predict what content is likely to resonate with readers is very important for media organizations, activists and political organizations. Such analysis before publication yields low prediction scores but can be useful as a signal to a content provider if it can indicate where to focus efforts in order to optimize content, or to a platform that publishes such content so that it can focus its moderation efforts on articles which are likely to go viral.

This project explores the performance of various machine learning models in predicting the likelihood of a news article becoming popular, as measured by how often it is shared on social media. The dataset used comes from the UCI Machine Learning Repository, covering a set of articles published on the online platform Mashable over a period of 2 years, from January 7, 2013 to January 7, 2015. The data contain 39,644 records, each corresponding to an article published on Mashable. Each record has 61 attributes - 2 of which are non-predictive, and 1 of which is the target variable (no. of shares). While the number of shares is a continuous variable, in this project a threshold is set, above which an article is labelled 'popular', thus converting the task into a classification problem.

This project uses the paper published by Fernandes et al. [1] as its primary reference, which introduced and published the associated dataset. The authors of the article focus on predicting classifying article popularity before publication, and then make build a system to make recommendations to optimize the likelihood of popularity. They used 47 features and applied various models, eventually settling on a Random Forest classifier with an accuracy of 67% as the best model. The features included fall into one of 6 categories depending on the information they contain and the method of extraction.

Feature Type	Sample Features
Non-descriptive	url, Time since publication
Content	Topic of article, # words/unique words/word length in title/content Keyword shares
Media	# of images/videos
Connections	# of links, # of links to other articles on Mashable, Popularity of referenced artcles
Timeline	Day of week/Weekend
NLP features	LDA topic closeness, Polarity/subjectivity of content/title
Target variable	**Number of social media shares**

Figure 2: Sample of features from each feature type

Further work by Zhang et. al. [3] expands upon this by using PCA for feature selection, but their difference in thresholds and classification (they use low/medium/high popularity tiers) is not comparable to the original paper. Ren et. al. [2] use Mutual Information and Fisher Criterion for feature selection and improve the accuracy of the random forest classifier to 69%. This project attempts to replicate their findings.

2. Exploratory Data Analysis

An initial exploration of the data was done to understand the features, target variable and the kind of information they contain.

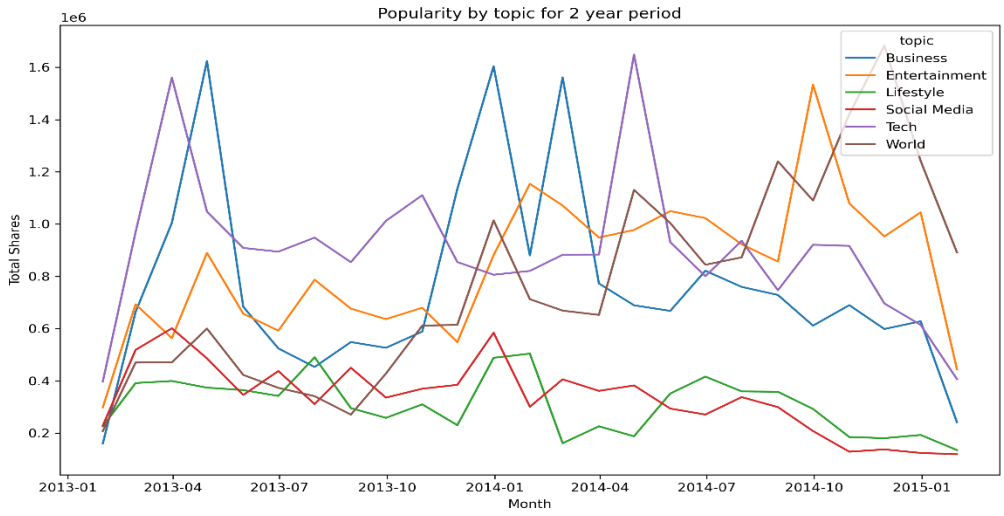

Figure 3: Timeline of topic popularity over each month in the 2-year period

From the data, it can be seen that articles related to the 'World' topic were the most shared whereas articles related to 'Lifestyle' and 'Social Media' were among the least shared overall. It also reflects the general growth trend in the popularity of Mashable articles, although the popularity increase is not proportionally shared across topics. This may reflect the reasons changing preferences/demographics in the Mashable readership. The sharp fall in aggregate number of shares across all topics towards the end likely reflects a lack of sufficient time for articles to start trending before the dataset was pulled.

A major consideration is to convert the target variable from a continuous variable into a target variable by setting a threshold value, above which the article is considered to be popular. The target variable has a long-tailed distribution as highlighted by how far apart the mean and the median are.

Mean	Standard Deviation	Minimum	25th percentile	Median	75th percentile	Maximum
3,395.38	11,626.95	1.00	946.00	1,400.00	2,800.00	843,300.00

Figure 4: Mean, standard deviation and various percentiles of the target variable

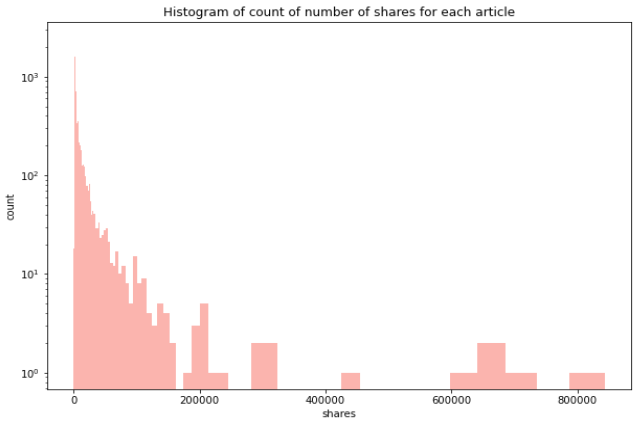

Figure 5: The number of shares has a long-tailed distribution with a small proportion of articles getting a very large number of shares

Converting the problem to a classification problem helps avoid having to make predictions in the long tail where data points are very thinly scattered. A threshold value of 1,400 has been set for the number of shares above which an article is considered to be popular. The threshold has been selected because it creates a balanced split between popular and unpopular articles, and because it facilitates benchmarking against findings from literature.

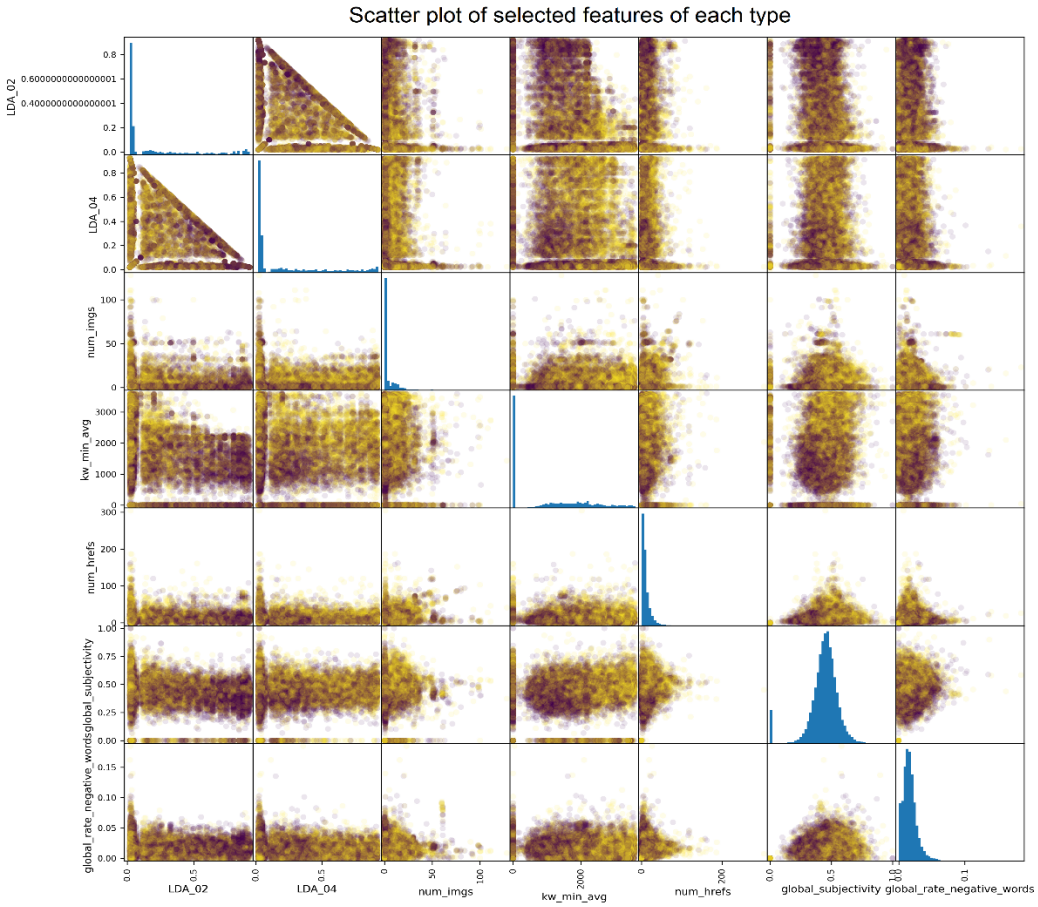

Figure 6: Scatter matrix of selected predictive features colored by class

A scatter matrix of the features reveals that several features also have long tailed distributions. While none of them individually stand out as predictive of popularity, a scatter matrix of different types of features shows some promise for combinations of features of different types in explaining popularity.

3. Methods

3.1 Data Splitting

The dataset consists of records of online articles published by Mashable. Each row corresponds to a single article and there are a total of 39,644 articles. The target variable is 'popular' - a binary variable where 0 corresponds to not-popular and 1 corresponds to popular. This variable is derived from the 'shares' variable present in the original dataset. Shares is the number of shares for a given article, and we set a threshold = 1400, above which an article is considered popular.

The dataset was generated in [1] by collecting information about all articles published on Mashable – no group structure was induced by article selection or linkages between articles that is mentioned in the original paper. Therefore, the data can be considered to be independent and identically distributed (IID). However, given the long-tailed nature of the distribution of the shares of each article, it is important to make sure that the train, validation and test sets all contain articles that exhibit a wide range of popularity, so that the models generalize well. To achieve this, groups have been created which map each article to the popularity quantile that they fall into, so that the bottom 10% of articles fall into one such group, the articles with popularity quantiles between 10% and 20% fall into another group and so on. This group information is used when splitting the dataset and is discarded thereafter, so that it is not used while training the model. A split of 80:10:10 has been used, which gives us 31715, 3964, 3965 articles for the train, validation and test sets respectively. When the model is trained, k-fold cross validation will be used to optimize the model hyperparameters.

3.2 Data Preprocessing

On the continuous features, the MinMaxEncoder is used for features relating to polarity, Latent Dirichlet Allocation (LDA) features, no. of tokens in title, average token length and number of keywords. None of the other continuous features are well bounded and exhibit fat tailed distributions. Hence, they are encoded with a StandardScaler.

On the categorical features (is_weekend, day_of_week and topic), a OneHotEncoder has been used since the features do not exhibit inherent order which will affect the prediction task. Eg. a OneHot encoder has been used on the topic feature since it contains the topic of the article and doesn't exhibit order.

There are no missing values in the data.

4. References

[1] Kelwin Fernandes, Pedro Vinagre, and Paulo 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.

[2] Yuchao Zhang and Kun Lin *"Predicting and Evaluating the Online News Popularity based on Random Forest"*, 2021 J. Phys.: Conf. Ser. 1994 012040.

[3] He Ren, Quan Yang, Stanford University, *"Predicting and Evaluating the Popularity of Online News"*, Machine Learning Project Work Report, 2015, pp. 1-5.

[4] Michael Barthel, Kirsten Worden Newspapers Fact Sheet, Pew Research Center
https://www.pewresearch.org/journalism/fact-sheet/newspapers/

GitHub repository: https://github.com/spacegoat1/news_article_popularity