Predicting Virality with Extreme Gradient Boosting on Online News Popularity Data

Ednalyn C. De Dios

D214

September 7, 2023

Western Governors University

Predicting Virality with Extreme Gradient Boosting on Online News Popularity Data

In this digital age that we’re in, understanding the different factors that contribute to the popularity of online news articles is a crucial endeavor for media organizations, marketing professionals, and content creators alike. Boundless amount of data exists and presents a challenge of distilling patterns to uncover hidden insights regarding user behavior. Furthermore, identifying key factors that determine an online news article’s popularity has become the holy grail of many, including data scientists. Machine learning techniques like XGBoost help uncover these hidden insights and translate them into actionable nuggets of information that stakeholders can act upon.

XGBoost, which stands for eXtreme Gradient Boosting, is a powerful and versatile machine learning algorithm that has gained popularity in recent years due to its effectiveness in handling a wide range of predictive tasks. Created by Tianqi Chen, XGBoost is “a scalable tree boosting system” (2023) equipped with both sequential and parallel architectures (Suginoo, 2022). A supervised learning algorithm, gradient boosting predicts “a target variable by combining the estimates of a set of simpler, weaker models” (How XGBoost Works, n.d.).

# A – Research Question

In this study, eXtreme Gradient Boosting or XGBoost is used to analyze the Online News Popularity Data by Fernandes et al. (2015) and predict the popularity of online news articles. The aim is to identify which attributes in the available data are key factors driving the number of social media shares. Stakeholders would like to know this information to optimize their content for “going viral.” This study will examine the different relationship between different attributes of the data in relation to the designation of whether the online news article is popular or not as defined by the number of times the article has been shared on social media channels. Thus, the question can then be summarized as follows: Do certain factors affect the popularity of online news articles?

The null hypothesis of the research question is that certain factors do not have a significant effect on the number of social media shares. For example, an online news article will have the same number of social media shares regardless of how many keywords it is using. The alternative hypothesis is that one or more factors have a significant effect on the number of social media shares. For example, the number of social media shares an online news article will have been greatly dependent on the number of keywords that it uses. If a significant effect is found on the number of social media shares, then it will be worthwhile to optimize online news articles for those factors. Exploratory data analysis will be used to determine whether the null hypothesis can be rejected or not.

# B – Data Collection

This study uses the “Online News Popularity” dataset which is publicly available from the UC Irvine Machine Learning Repository project (Fernandes et al., 2015). The dataset contains statistics on articles published by Mashable.com. The dataset contains 39,797 records and 61 attributes, of which 58 are predictive, two are non-predictive, and one goal field. The dataset is publicly available to the public and licensed under [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/legalcode) (CC BY 4.0). The following table lists the name, type, and description of the attributes that were included in this study.

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| url | Categorical | URL of the article |
| timedelta | Continuous | Days between the article publication and the dataset acquisition |
| n\_tokens\_title | Continuous | Number of words in the title |
| n\_tokens\_content | Continuous | Number of words in the content |
| n\_unique\_tokens | Continuous | Rate of unique words in the content |
| n\_non\_stop\_words | Continuous | Rate of non-stop words in the content |
| n\_non\_stop\_unique\_tokens | Continuous | Rate of unique non-stop words in the  content |
| num\_hrefs | Continuous | Number of links |
| num\_self\_hrefs | Continuous | Number of links to other articles published by Mashable |
| num\_imgs | Continuous | Number of images |
| num\_videos | Continuous | Number of videos |
| average\_token\_length | Continuous | Average length of the words in the content |
| num\_keywords | Continuous | Number of keywords in the metadata |
| data\_channel\_is\_lifestyle | Categorical | Is data channel 'Lifestyle'? |
| data\_channel\_is\_entertainment | Categorical | Is data channel 'Entertainment'? |
| data\_channel\_is\_bus | Categorical | Is data channel 'Business'? |
| data\_channel\_is\_socmed | Categorical | Is data channel 'Social Media'? |
| data\_channel\_is\_tech | Categorical | Is data channel 'Tech'? |
| data\_channel\_is\_world | Categorical | Is data channel 'World'? |
| kw\_min\_min | Continuous | Worst keyword (min. shares) |
| kw\_max\_min | Continuous | Worst keyword (max. shares) |
| kw\_avg\_min | Continuous | Worst keyword (avg. shares) |
| kw\_min\_max | Continuous | Best keyword (min. shares) |
| kw\_max\_max | Continuous | Best keyword (max. shares) |
| kw\_avg\_max | Continuous | Best keyword (avg. shares) |
| kw\_min\_avg | Continuous | Avg. keyword (min. shares) |
| kw\_max\_avg | Continuous | Avg. keyword (max. shares) |
| kw\_avg\_avg | Continuous | Avg. keyword (avg. shares) |
| self\_reference\_min\_shares | Continuous | Min. shares of referenced articles in Mashable |
| self\_reference\_max\_shares | Continuous | Max. shares of referenced articles in Mashable |
| self\_reference\_avg\_sharess | Continuous | Avg. shares of referenced articles in Mashable |
| weekday\_is\_monday | Categorical | Was the article published on a Monday? |
| weekday\_is\_tuesday | Categorical | Was the article published on a Tuesday? |
| weekday\_is\_wednesday | Categorical | Was the article published on a Wednesday? |
| weekday\_is\_thursday | Categorical | Was the article published on a Thursday? |
| weekday\_is\_friday | Categorical | Was the article published on a Friday? |
| weekday\_is\_saturday | Categorical | Was the article published on a Saturday? |
| weekday\_is\_sunday | Categorical | Was the article published on a Sunday? |
| is\_weekend | Categorical | Was the article published on the weekend? |
| LDA\_00 | Categorical | Closeness to LDA topic 0 |
| LDA\_01 | Categorical | Closeness to LDA topic 1 |
| LDA\_02 | Categorical | Closeness to LDA topic 2 |
| LDA\_03 | Categorical | Closeness to LDA topic 3 |
| LDA\_04 | Categorical | Closeness to LDA topic 4 |
| global\_subjectivity | Continuous | Text subjectivity |
| global\_sentiment\_polarity | Continuous | Text sentiment polarity |
| global\_rate\_positive\_words | Continuous | Rate of positive words in the content |
| global\_rate\_negative\_words | Continuous | Rate of negative words in the content |
| rate\_positive\_words | Continuous | Rate of positive words among non-neutral tokens |
| rate\_negative\_words | Continuous | Rate of negative words among non-neutral tokens |
| avg\_positive\_polarity | Continuous | Avg. polarity of positive words |
| min\_positive\_polarity | Continuous | Min. polarity of positive words |
| max\_positive\_polarity | Continuous | Max. polarity of positive words |
| avg\_negative\_polarity | Continuous | Avg. polarity of negative words |
| min\_negative\_polarity | Continuous | Min. polarity of negative words |
| max\_negative\_polarity | Continuous | Max. polarity of negative words |
| title\_subjectivity | Continuous | Title subjectivity |
| title\_sentiment\_polarity | Continuous | Title polarity |
| abs\_title\_subjectivity | Continuous | Absolute subjectivity level |
| abs\_title\_sentiment\_polarity | Continuous | Absolute polarity level |
| shares | Continuous | Number of shares (target) |

There are several advantages with the chosen dataset. There are no missing values, the class distribution is relatively balanced, and the categorical variables have already been encoded into numerical form. However, one observed disadvantage is the sheer size of the dataset. One trial run of the experiment took almost three hours to run hyperparameter tuning. The challenge of long running time was overcome by setting the tree\_method to ‘gpu\_hist’ and limiting the range of the parameter search space.



C – Data Extraction and Preparation

The next phase of the analysis is data preparation. The tool used is Python 3.9.9 and Jupyter Notebook 7.0.2 was used as the interactive development environment. Python was chosen for its versatility and easy-to-learn syntax. Jupyter Notebook was chosen for its markdown capability. One disadvantage of Python is performance while Jupyter makes good code versioning very difficult (Mueller, 2018).

After reading the data into a dataframe, the next step is to check for any missing values.A computer code with text

Description automatically generated

A computer code with red text

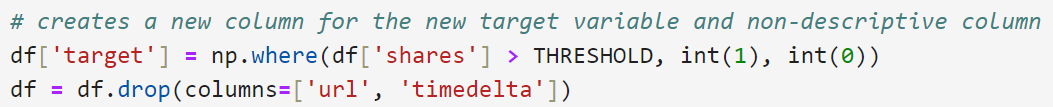
Description automatically generatedNext is correcting the column names.

Then, duplicates are dropped if any. There weren’t any as shown by df.shape before and after the operation.

A screenshot of a computer

Description automatically generated

The target variable was created by applying a condition on the ‘shares’ variable and designating 1 or 0 depending on the THRESHOLD value. The threshold value for this notebook is 1400.



Finally, the cleaned and prepared dataset is exported using pandas’ to\_csv() method.

D – Analysis

Once the dataset is prepared, the next step in the analysis is to conduct EDA or explanatory data analysis. The high-levels steps are:

1. Get familiar with the data
2. Review class distribution
3. Get summary statistics
4. Remove outliers
5. Compare the interactions of the variables

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generatedAs shown in the graph below, the dataset has a relatively balanced class.

A screen shot of a graph

Description automatically generated Next is to conduct a visual inspection of the boxplot to look for outliers in the dataset.

A screenshot of a computer code

Description automatically generated After confirming the existence of outliers visually, the interquartile range or IQR was calculated so that the records outside the lower and upper bound could be removed.A graph with a blue rectangle and black lines

Description automatically generated

A screenshot of a graph

Description automatically generated Next is the plotting of histograms for the continuous variables. As shown, the continuous variables are not distributed normally.

A computer code with black text

Description automatically generated with medium confidence The correlation heatmap shows a few interesting relationship between several variables.

A red and white squares with black text

Description automatically generated with medium confidence

According to the graph below, the weekend is a slow news cycle for both popular and unpopular articles.

A graph of blue and orange bars

Description automatically generated

A graph of blue and orange bars

Description automatically generated Evident in the graph below, Tech, Business, and Entertainment dominate both popular and unpopular articles. Volume wise, the difference between the count of popular and unpopular articles in the World and Entertainment is noteworthy.

Similarly, LDA Topic #2 shows the same imbalance between popular and unpopular articles.

A graph of a bar graph

Description automatically generated with medium confidence

Conducting an T-test revealed more significant differences between the popular and unpopular groups than insignificant ones. In this case, n\_non\_stop\_words, kw\_min\_max, kw\_avg\_max, min\_negative\_polarity, max\_negative polarity, and abs\_title\_subjectivity have insignificant differences in samples.

A screenshot of a computer

Description automatically generated

Confirming the visual inspection earlier, Shapiro test indicate that none of the variables are distributed normally.

A screenshot of a computer program

Description automatically generated



The exploration of the dataset involved both visual exploration and statistical testing. An advantage of visualization is the ease it provides the reader to grasp the characteristics of the data that is being inspected. In addition, the output of the statistical testing eliminates guesswork by supplying a statistic and p-values. One disadvantage is that visual inspection can only go so far. It does not provide a value up front without extensive coding of matplotlib methods.

These are the steps involved in the modeling part of the analysis:

1. Splitting the dataset into training and test sets
2. Building logistic regression models for reference
3. Building XGBoost classifier models
4. A screenshot of a computer program

   Description automatically generatedExtracting feature importance based on the best XGBoost model

After splitting the data and building the initial logistic regression model, GridSearchCV was utilized to determine the ideal parameters that maximize the accuracy of the logistic regression model.

A screenshot of a computer program

Description automatically generated

In the following code, the final logistic regression model is built using the best parameters given by grid search. The accuracy of the final logistic regression model is 0.65, a value that is relatively close to 0.64 which is the accuracy of the training set. This means that the model generalizes well.

A screenshot of a computer program

Description automatically generated

Below, the area-under-the-curve (AUC) score was calculated using the test set and the receiver operating characteristic ROC was plotted. This graph will be used later to compare against the final XGBoost model.

A screenshot of a computer program

Description automatically generated

A graph of a logistic regression

Description automatically generated

The ROC curve is larger than the unskilled (0.50) line which signifies that the model is slightly a little bit better at predicting an event than flipping a standard coin.

The initial XGBoost model that was trained had an accuracy of 0.90 on the training set and 0.64 on the test set. The initial model did not generalize well and it well perhaps overfitted.

A screenshot of a computer

Description automatically generated

To remedy overfitting, we created a pipeline that utilized grid search again to determine the best parameter for XGBoost.

A screenshot of a computer program

Description automatically generated

This time around, the accuracy is 0.69 on the train set and 0.67 on the test set. Although the accuracy is not stellar, the model generalized well and did not overfit.

A screenshot of a computer

Description automatically generated

The best parameters were used in the final XGBoost model.

A screenshot of a computer

Description automatically generated

The final XGBoost model sports an accuracy of 0.73 on the train set and 0.66 on the test set. The AUC was calculated and ROC plotted.

A graph of a positive rate

Description automatically generated with medium confidence

When the best logistic regression model and the best XGBoost were plotted together, there was no indication that the two models were different at all. However, the AUC says otherwise. Nevertheless, XGBoost saw an improvement over logistic regression by one point in accuracy and two points in AUC score.

A graph with a red line

Description automatically generated

Now, feature importance will be extracted to determine the factors that contribute to an online news article’s popularity. For this purpose, the author decided to use the SHAP package to extract feature importance because those ranked by XGBoost were inconsistent. In “Interpretable Machine Learning with XGBoost,” Lundberg found that “feature importance orderings are very different for each of the three options provided by XGBoost” (2018). Therefore, SHAP method will be used to extract feature importance.

A screenshot of a computer

Description automatically generated

Above, we use the shap package to extract feature importance. Below, the shap package helps in extracting the most important features of our dataset. In relation to the final XGBoost model, the most important features are:

1. self\_reference\_min\_shares
2. kw\_avg\_avg
3. kw\_max\_avg
4. is\_weekend
5. data\_channel\_is\_entertainment

A graph with blue and black text

Description automatically generated

In the graph below, every article has one dot on each row. The x position of the dot is the impact of that feature on the model’s prediction for the article, and the color of the dot represents the value of that feature for the article. Dots that do not fit on the row pile up to show density (Lundberg, 2018).

A screenshot of a graph

Description automatically generated

E – Data Summary and Implications

The research question of this study is “Do certain factors affect the popularity of online news articles?” Based on the analysis, it is clear that certain factors affect the popularity of online news articles. Some variables seem to have a stronger influence on the target variable when it comes to statistical significance.

# C – Sources

* How XGBoost Works. (n.d.). Amazon Sagemaker. Retrieved August 31, 2023, from https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html
* Mueller, Alex. (2018, March 24). 5 reasons why jupyter notebooks suck. Towards Data Science. Retrieved September 1, 2023, from https://towardsdatascience.com/5-reasons-why-jupyter-notebooks-suck-4dc201e27086
* Chen, Tianqi. (n.d.). Tianqi Chen. Retrieved August 31, 2023, from https://tqchen.com.
* Fernandes,Kelwin, Vinagre,Pedro, Cortez,Paulo, and Sernadela,Pedro. (2015). Online News Popularity. UCI Machine Learning Repository. https://doi.org/10.24432/C5NS3V
* Lundberg, Scott. (2018, April 17). Interpretable Machine Learning with XGBoost. Towards Data Science. Retrieved from https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27.
* Suginoo, Michio. (2022, October 20). XGBoost: its Genealogy, its Architectural Features, and its Innovation. Towards Data Science. Retrieved August 31, 2023, from https://towardsdatascience.com/xgboost-its-genealogy-its-architectural-features-and-its-innovation-bf32b15b45d2.
* Uddin, Md. Taufeeq (2018). Predicting the Popularity of Online News from Content Metadata. Retrieved August 29, 2023, from https://github.com/krishnakartik1/onlineNewsPopularity/blob/master/Paper2/Predicting%20the%20Popularity%20of%20Online%20News%20from%20Content%20Metadata.pdf