Predicting Virality with Extreme Gradient Boosting on Online News Popularity Data

Ednalyn C. De Dios

D214

September 16, 2023

Western Governors University

Executive Summary

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. Machine learning techniques like XGBoost help uncover these hidden insights and translate them into actionable nuggets of information that stakeholders can act upon.

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 construct a model with more than 65% accuracy and an AUC score of above 60%. Thus, the question can then be summarized as follows: Can gradient boosting be constructed based solely on the research data?

The null hypothesis of the research question is that gradient boosting cannot be made from the Online News Popularity dataset. The alternative hypothesis is that an optimized gradient boosting model can be made from the Online News Popularity dataset.

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.

Python 3.9.9 and Jupyter Notebook 7.0.2 was used as the interactive development environment. The following steps were taken to prepare the dataset:

1. Correcting the column names
2. Creating the target variables based on the 'shares' using a threshold of 1400

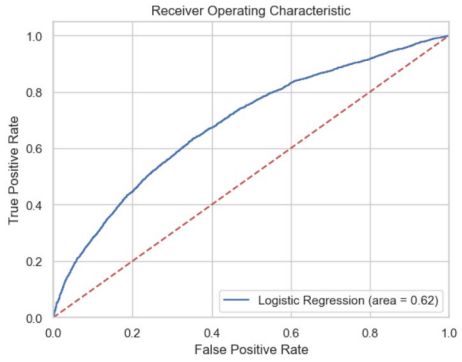
3. Removing outliers

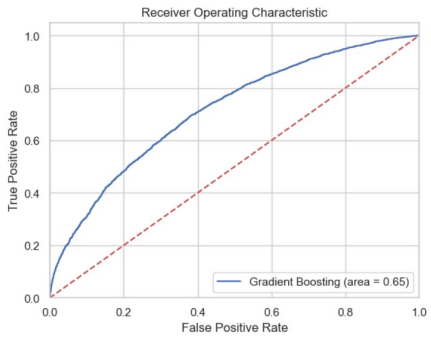
Exploratory Data Analysis revealed several interesting findings. 1) The weekend is a slow news cycle and articles published were more likely to be popular than those published during the week. 2) Tech, Business, and Entertainment topics dominate the articles. 3) LDA Topic #2 shows more unpopular articles than popular ones. Additionally, T-tests revealed more significant differences between the popular and unpopular groups than insignificant ones.

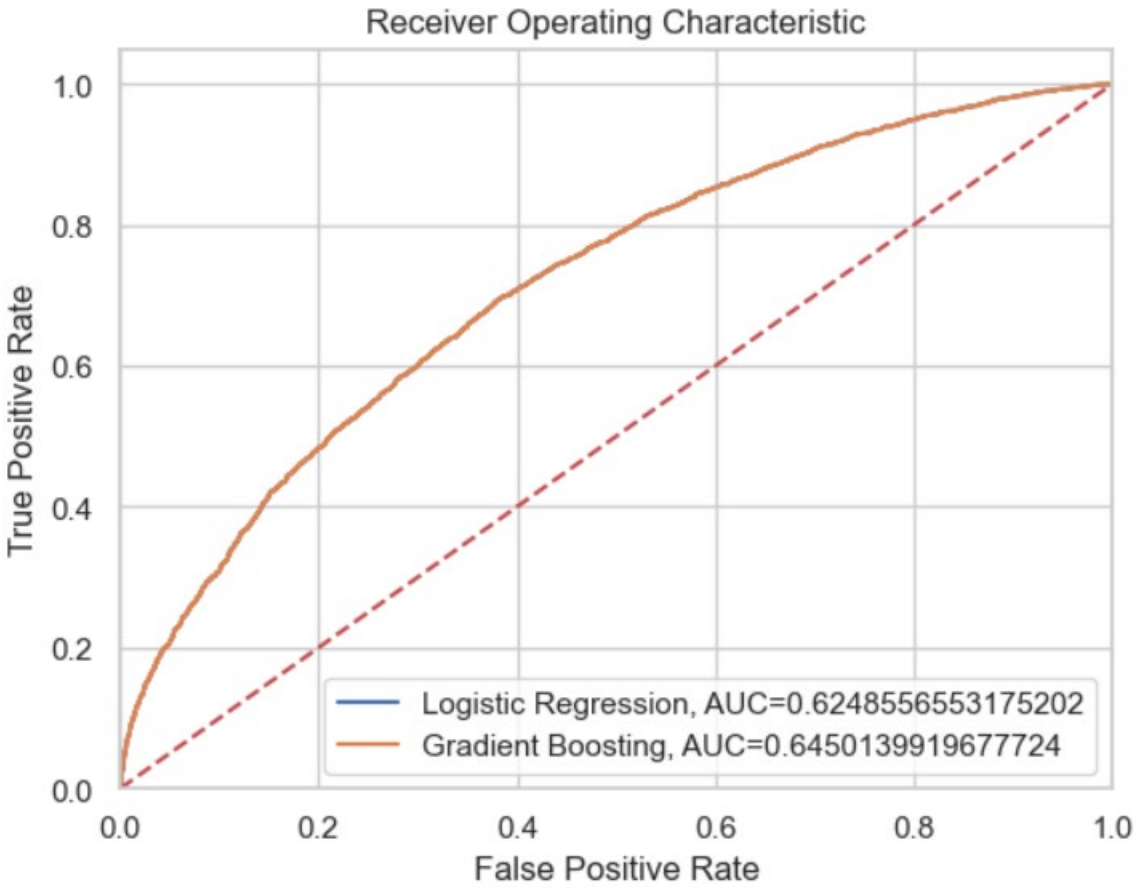
These are the steps involved in the modeling part of the analysis:

1. Splitting the dataset into training and test sets
2. Building logistic regression models for reference
3. Building XGBoost classifier models
4. Extracting feature importance based on the best XGBoost model

After splitting the data, we built an optimized logistic regression model with an accuracy of 0.65 and an AUC score of 0.62. Then we compared that with the best performing XGBoost model which had an accuracy of 0.66 and AUC score of 0.65.

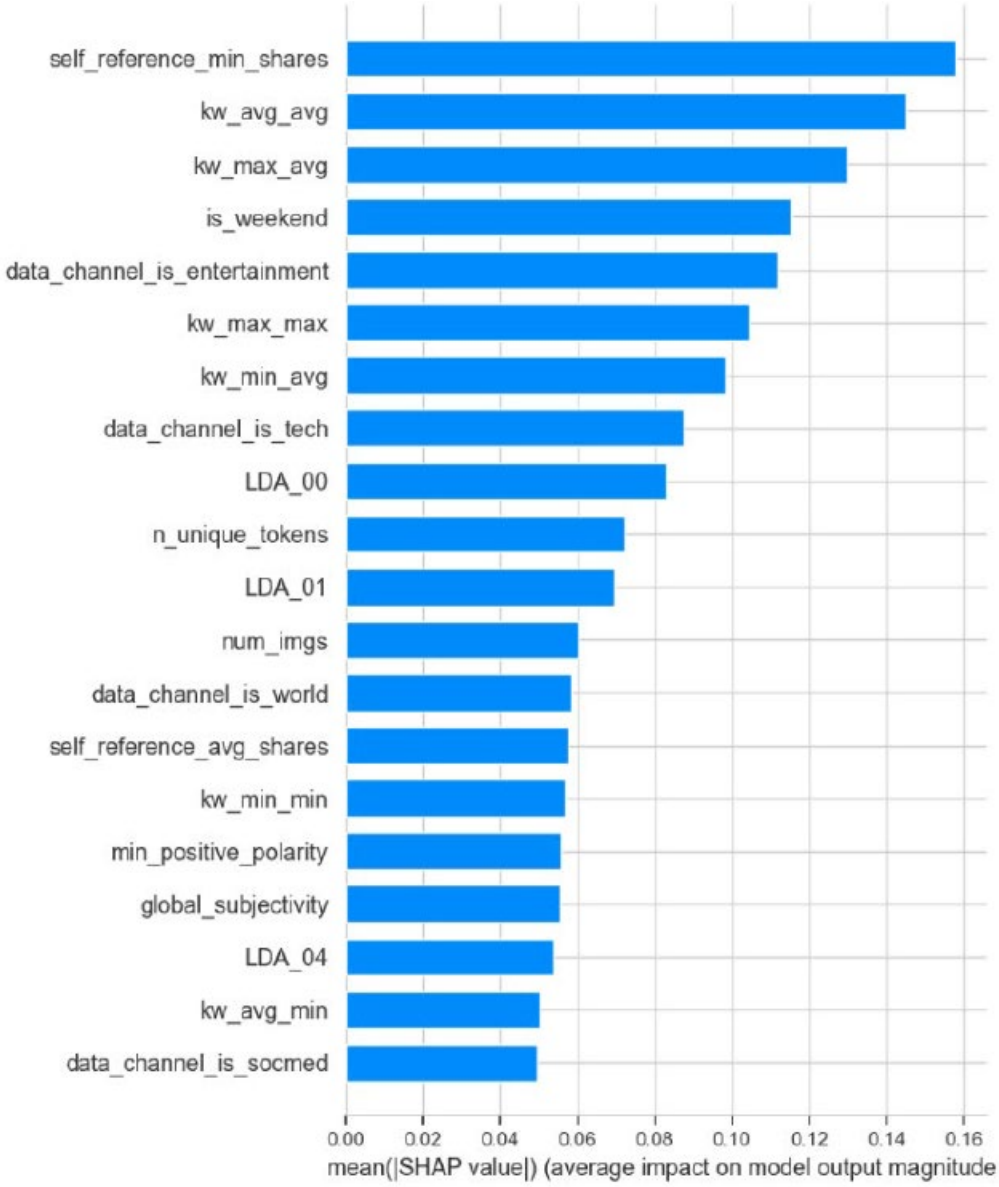






Although the accuracy is not stellar, the final XGBoost model can generalize well for new data.

Feature importances were also extracted using mean SHAP values. Self-reference_min_shares, kw_avg_avg, kw_max_avg, is_weekend, and data_channel_ius_entertainment were deemed as the most important feature of the final model.



One limitation of this study is that the original dataset only included articles from one website (Mashable.com). The prevalence of popular articles in tech, business, and entertainment reflects the niched demographic of Mashable's distribution. A sample containing articles from all sorts of publications would make a better dataset that could generalize better in predicting previously unseen articles.

One recommendation that can be made is to pay particular attention to the kind of articles that are published during the weekend. Even though the number of articles published on the weekend is less than those published during the week, the study shows that articles published during the weekend are more likely to be popular than not. This phenomenon surely warrants more investigation to determine the reason why.

For future studies, the author recommends the following: 1) use of XGBoost regression to predict the number of social media shares instead of using a threshold value, and 2) clustering the articles based on channels or topic. Initially converting the number of social media shares into the categorical number of 0 or 1 presented the possibility of information loss. Predicting the number of social media shares could possibly yield better results. In addition, leveraging clustering algorithms could also yield more insights about the different segments within the publication's reader base.

The findings of this study can benefit organizations, marketing professionals, and content creators. By building an XGBoost model using their own dataset, they can extract the feature importances and figure out which features are important. Based on these features, the writers or copywriters can then model their articles to have follow the optimum characteristics of the features. For example, they can target how many keywords to use for SEO, how many self-referral links to create, how many words they should aim for, and when to publish their content. While the aforementioned are not hard and fast rules, they can serve as guidelines for writing viral content.

F – Sources

- Fernandes, Kelwin, Vinagre, Pedro, Cortez, Paulo, and Sernadela, Pedro. (2015). Online News Popularity. UCI Machine Learning Repository. https://doi.org/10.24432/C5NS3V