Student Name: Ednalyn C. De Dios

Student ID: 001459994

Capstone Project Name: Predicting Virality with Extreme Gradient Boosting on Online News Popularity Data

Project Topic: Predictive Model for Online News Popularity Data

X This project does not involve human subjects research and is exempt from WGU IRB review.

Research Question: Can gradient boosting be constructed based solely on the research data?

Hypothesis: H_0: A gradient boosting model cannot be made from the Online News Popularity dataset.
. H_1: A gradient boosting model can be made from the Online News Popularity dataset.

Context: The contribution of this study to the field of Data Analytics and the MSDA program is to create a predictive model which can predict whether an online article is going to be viral or not. This model can be used to identify important characteristics of a viral online article so that content creators and media companies alike can optimize their content for social media virality. This study will use an Extreme Gradient Boost model to analyze the significance of predictor variables and identify which of them are the best predictors of social media popularity. Gradient boosting is known "for its prediction speed and accuracy, particularly with large and complex datasets" (Saini, 2023).

Data: The data needed to be collected for the research question is publicly available from the UC Irvine Machine Learning Repository (Fernandes et al., 2015). The dataset contains 39,797 records and 61 attributes, of which 58 are predictive, two are non-predictive, and one goal field.

Made available through the UCI Machine Learning Repository, the dataset includes the following predictor variables:

Field	**Type**
url	Categorical
timedelta	Continuous
n_tokens_title	Continuous
n_tokens_content	Continuous
n_unique_tokens	Continuous
n_non_stop_words	Continuous
n_non_stop_unique_tokens	Continuous
num_hrefs	Continuous
num_self_hrefs	Continuous
num_imgs	Continuous
num_videos	Continuous
average_token_length	Continuous
num_keywords	Continuous
data_channel_is_lifestyle	Categorical
data_channel_is_entertainment	Categorical
data_channel_is_bus	Categorical

data_channel_is_socmed	Categorical
data_channel_is_tech	Categorical
data_channel_is_world	Categorical
kw_min_min	Continuous
kw_max_min	Continuous
kw_avg_min	Continuous
kw_min_max	Continuous
kw_max_max	Continuous
kw_avg_max	Continuous
kw_min_avg	Continuous
kw_max_avg	Continuous
kw_avg_avg	Continuous
self_reference_min_shares	Continuous
self_reference_max_shares	Continuous
self_reference_avg_sharess	Continuous
weekday_is_monday	Categorical
weekday_is_tuesday	Categorical
weekday_is_wednesday	Categorical
weekday_is_thursday	Categorical
weekday_is_friday	Categorical
weekday_is_saturday	Categorical
weekday_is_sunday	Categorical
is_weekend	Categorical
LDA_00	Categorical
LDA_01	Categorical
LDA_02	Categorical
LDA_03	Categorical
LDA_04	Categorical
global_subjectivity	Continuous
global_sentiment_polarity	Continuous
global_rate_positive_words	Continuous
global_rate_negative_words	Continuous
rate_positive_words	Continuous
rate_negative_words	Continuous
avg_positive_polarity	Continuous

min_positive_polarity	Continuous
max_positive_polarity	Continuous
avg_negative_polarity	Continuous
min_negative_polarity	Continuous
max_negative_polarity	Continuous
title_subjectivity	Continuous
title_sentiment_polarity	Continuous
abs_title_subjectivity	Continuous
abs_title_sentiment_polarity	Continuous
shares	Continuous

The online articles featured in the dataset were published by Mashable (www.mashable.com) and their content and the rights to reproduce them belongs to them (Fernandes et al., 2015). Hence, this dataset does not share the original content but only some statistics associated with it (Fernandes et al., 2015). The dataset is publicly available to the public and licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

Data Gathering: Treatment of the Data. Data will be downloaded from the publicly available CSV file from archive.ics.uci.ed website which shows a heterogenous set of features about online articles and the number of times these articles were shared through social media for over a period of two years. The data quality is high. There are no missing values and no duplicates in the dataset. Column names will be cleaned up and one misspelling will be corrected. The data mostly contains both quantitative and qualitative variables. Python will be used to clean and wrangle the data. Overall sparsity of the data is ___%.

Data Analytics Tools and Techniques: Design of the Study. 1. Shapiro-Wilk test will be run to determine the normality of the data because it "is the most powerful test when testing for a normal distribution" (Korstanje, 2019). 2. Extreme Gradient Boosting does not require normality but will be run regardless. The XGBoost model will utilize gradient boosting algorithm to build models sequentially. The subsequent models then try to reduce the errors of the previous model by building a new model on the errors of the residuals of preceding model (Saini, 2023). 3. SHAP values will be used to plot the feature importance of the dataset. 4. The dataset will also be split into train and test sets using sklearn's feature selection module.

Justification of Tools/Techniques: Python will be used for the creation of gradient boosting models due to their flexibility of integrating with other platforms and languages (Geeksforgeeks, 2023). Furthermore, Python is currently more popular than R, especially among software developers and data scientists (Luna, 2022). Lastly, Python is free for everyone to use (Luna, 2022) as opposed to SAS.

Project Outcomes: This study will create an Extreme Gradient Boosting model for indicator of virality based on the composition of each online article. The project outcome will be a reusable statistical model. Support for the alternative hypothesis can be found in Uddin (2018) that Gradient Boosting Machine is able to predict the popularity of an online article with decent prediction rate.

Projected Project End Date: 9/20/2023

Sources:

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

Korstanje, Joos (2019). 6 ways to test for a Normal Distribution — which one to use? Retrieved August 29, 2023, from https://towardsdatascience.com/6-ways-to-test-for-a-normal-distribution-which-one-to-use-9dcf47d8fa93

Luna, Javier Canales (2022). Python vs R for Data Science: Which Should You Learn? Retrieved August 29, 2023, from https://www.datacamp.com/blog/python-vs-r-for-data-science-whats-the-difference

Saini, Anshul (2023). Gradient Boosting Algorithm: A Complete Guide for Beginners. Retrieved August 29, 2023, from https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/

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

Course Instructor Signature/Date:

To be filled out by a course mentor:

☒ The research is exempt from an IRB Review.

☐ An IRB approval is in place (provide proof in appendix B).

Course Mentor's Approval Status: Approved

Date: 8/30/2023

Reviewed by: Jsowell

Comments: Click here to enter text.