**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**: H0: A gradient boosting model cannot be made from the Online News Popularity dataset. . H1: 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](http://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](https://creativecommons.org/licenses/by/4.0/legalcode) (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: Select one

Date: Click here to enter a date.

Reviewed by:

Comments: Click here to enter text.