CSC 240 Data Mining

Final Project Proposal

04/04/2021

Sarah Zaman, szaman

Syeda Sarah Shahrin, sshahrin

Project Proposal: Online News Popularity

We propose to analyze the UCI Online News Popularity Dataset (link: https://www.kaggle.com/thehapyone/uci-online-news-popularity-data-set) with the goal to predict whether an article will have good or bad popularity. The data set uses the number of shares as a measure of popularity. As netizens in today's day and age, we believe it is important to understand the trends behind what makes a news article successful and engaging.

Both of us spent the majority of our high school lives debating and, thus, keeping track of the political, economical, and social issues that are happening around the world. Online articles were a significant source of matter for arguments and opinions we learned to form over the years. We spent most of our weekends in high school researching legit and popular news articles regarding the motion/subject we were assigned to for a debate. We are very enthusiastic about this project because it will allow us to see what kind of content, tone, and wording is likely to get the most traction and perhaps pose questions about perceptions.

Dataset Description and Project Question

Our dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of 2 years. There are 61 attributes in this dataset of which, 58 are predictive attributes, 2 are non-predictive and 1 is the goal field (no. of shares). Our question is -

Can we predict the popularity of an online news article by analyzing the dataset?

Goal of Analysis

Our analysis of this data set produces impactful work in predicting the virality of articles and opens the possibility to further explore how wording can generate more engagement or discourse for the same content.

Technical Approach

Data Preprocessing

We will be using Jupyter Notebooks to analyze the provided .csv file. We will clean up the dataset for use in our model by replacing all missing data values with NaN to indicate the irrelevant attributes.

Data Visualization

First, we will visualize the data using visualization tools in python (such as scatterplots, box plots, histograms, pie charts, etc).

Model Selection

We plan on using 3 algorithms on this dataset and compare the results. (This is subject to change as we progress with our project).

1. Naive K-means: Also known as the k-means clustering algorithm, this model is our choice for a basic yet fast comparison. Using k-means will allow us to cluster certain attributes together and see how the data behaves with respect to our chosen attributes, allowing us to see patterns between popularity and the clustered attributes.

2. Logistic Regression Model: As we know, the Logistic Regression model is a classification algorithm used to attain binary outcomes based on data.

3. Random Forest: Random Forest is a classification and regression algorithm that grows multiple trees which are then merged together in order to attain a more accurate prediction. They use decision trees that draw a pathway to an outcome based on binary decisions inputted.

We believe the combination of these classification and regression models, we can successfully analyze both the categorical and numerical data from the dataset and successfully predict the outcome.

Team Composition & Role of Team Members

Team Member 1: Sarah Zaman

NetID: szaman

Undergraduate

Major: BA in Computer Science

Team Member 2: Syeda Sarah Shahrin

NetID: sshahrin

Undergraduate

Major: BS in Computer Science

We will equally perform all the steps provided in the technical approach section and deliver the final presentation as well.