Online News Popularity Prediction

Milestone: Project Planning

Group 5

Sravya Paramata (paramata.s@northeastern.edu)

Shreya Karakata (karakata.s@northeastern.edu)

Amoolya Shettar (shettar.a@northeastern.edu)

Amogha Shettar (shettar.am@northeastern.edu)

Submission Date: 10/06/2022

Problem Setting:

In the modern age, the age of science and technology, the traditional newspapers and magazines have been taken over by the internet thereby making it crucial for the channels to predict the popularity of an article on the web. Since a major portion of a company's budget is allocated to its marketing, an efficient way to get the popularity is by building a model that can reveal the best available options to the user in understanding the features that affect the output and predict the score of a given article/blog. The principle of this project is to use data mining techniques to analyze the data and build multiple machine learning models to predict the popularity of a news article.

Problem Definition:

Number of views/clicks, number of likes/dislikes, number of re-shares are few of the many ways to estimate the popularity of a particular paper. In this project we will be considering shares to estimate the popularity of an online new article.

The assumptions we will be making as part of this project are,

- There is no relation between popularity of the article with the article itself and the current events that are happening in the world.
- The quality of the article and the affinity between the reader and the article is ignored.

The following items will be inspected and addressed as part of this project:

- What are the contributing factors/attributes that make an article popular?
- What are the predictors that affect the shares of an article?
- What are the best models to predict and classify a particular article as the best article?

Data Description:

The dataset is obtained from the Mashable Internet news platform. It contains 61 attributes and 39797. In the 61 attributes, there are 58 predictive, 2 non-predictive (URL and time), and 1 target variable - number of shares. The goal field (target attribute) is the shares column. The 58 predictive variables constitute both numerical and categorical fields. They give information about the article published like: length of title (N), length of article (N), count of unique words, stop words, average token length, has images attached (C), has video attached (C), type of channel the article was published in(C), day of the week the article was published on (C).

Since the articles had the textual data, the sentiment analysis was already performed on the data. The sentiment polarity scores for the articles and titles along with positive and negative word count data was generated and is already part of the predictors. There are few other derived parameters along with these major predictors which we will be discussing in detail later on.

Problem Planning:

We will be building machine learning models to analyze various attributes of news and predict the popularity of an article. With 61 attributes as predictors, we assume to be using dimension reduction and feature selection on the data in order to get the most relevant predictors.

Supervised machine learning models will then be used for regression and classification. Since the number of shares is a continuous numerical variable, we can build regression models. For classification, the target variable 'shares' will be used to perform binary classification based on a predefined threshold, thereby classifying whether an article is popular or not.