Data Science 316 A1 Project

Sebastien Meniere, Jaunré de Swardt

April 4, 2024

1 Introduction and Problem Statement

Online news popularity prediction is a well-developed research field, with the aim to model the underlying mechanisms influencing how digital media is distributed through online social networks and, to make predictions about news article popularity prior to publication.

Existing literature measures popularity through likes, shares, and views. Reliably predicting the popularity of a news article before it is published is of great importance to the news agencies / providers. Allowing for data driven article optimization for improved popularity, competitive advantages over other publishers, higher user engagement with articles, all of which could lead to greater industry success.

Furthermore, an acute understanding of what news will be popular is valuable to many other sectors, including, consumer markets, political affair, marketing, and entertainment. This is because popular / trending news can have a strong influence on the public's opinions, interests, and decision making.

The difficulty in creating a robust predictive model in this context, is that there are many unknown, and unmeasured variables in the physical world that will influence which articles become popular. These could be, current political affairs, fashion trends, consumer fads. These factors are difficult to incorporate into training data sets, as they are time / period specific, and the information captured might not generalize to future events. Some of these confounding variables are not knowable before publication.

The news affects how people act, and the way in which people act makes the news. This rapid, and constant feedback loop is increasingly difficult to model as people consume more media.

In this project we will be investigating and evaluating the predictive classification models presented in the research of Fernande et. Al. We will also attempt to improve upon the model's classification performance, and attempt to construct a high preforming and highly interpretable model.

2 Data Description

We acquired the data set from the UCI machine learning repository. The authors collected 39644 articles published by the reputable news organisation Mashable, over a two year time period. Each article was processed as to summarise some important characteristic. In total there are 61 features, that stem from 6 categories.

Category	Feature examples
WORD	Number of words in article / title, Average word length, Rate of unique / non-stop words
Links	Number of links, Number of Mashable article links
Digital Media	Number of images / images
Time	Day of the week, Published on weekend
Keywords	Number of keywords, Article category
NLP	Title subjectivity, Article text subjectivity, Title sentiment polarity
Target	Number of shares

All of the variables fall into two types, number and ratio. Where the ratio variables are generally continuous between 0 and 1. And number type variables are either floats or integers. There are no missing values in the data set. The target variable is an integer type, ranging from zero to eight hundred thousand. This means that it will have to be partitioned into bins such that we can use classification models for prediction. We will discuss the partitioning threshold and it's implications later.

We must also point out that there are some variables that will be omitted from model training process as they do contain any useful information, for example to url to the article page is non-informative.

We did observe several outliers, these were observations where the number of shares was far larger than any reasonable amount of standard deviations from the mean.

3 The Current Approach

df