Predicting the Popularity of Online News Articles

Jeremy McCormick

12/14/2021

## Introduction

Online news is a competitive area where many different sites compete for the public’s limited attention. Predicting the popularity of a news story before publication could be useful for publishers and authors as a decision support system for helping them maximize the impact of their work. Such a system could be used for predictive or prescriptive decision support, where small changes to content before publication, such as modifications to the title wording or sentiment, might increase the article’s popularity.

## Goals

In this study, we will attempt to develop a statistical model to predict whether an online news story will be popular by evaluating pre-publication information. This can be used to evaluate a story’s potential popularity based on its metadata. We will also identify which variables are most important in this classification, in order to draw attention to areas that can be focused upon by publishers and authors to increase readership.

## Data Preparation

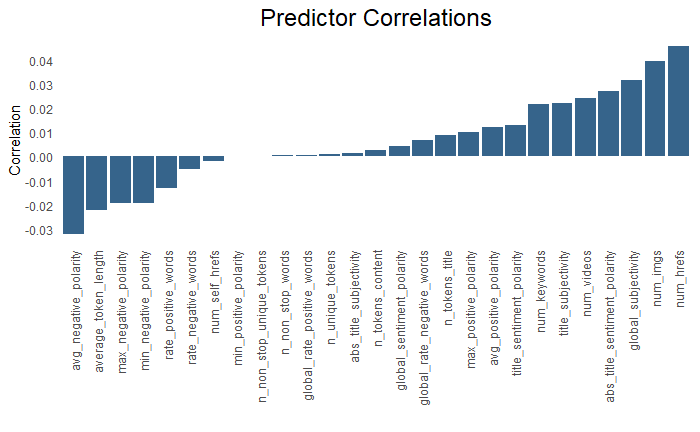
The data used is the Online News Popularity Data Set from the UCI Machine Learning Repository [<https://archive.ics.uci.edu/ml/datasets/online+news+popularity>]. This is comprised of all the articles published on the Mashable platform during a two year period from 2013 to 2015. There are 61 variables total, with two non-predictive columns, and a target variable, which is the number of shares, for a total of 58 potential predictors. We discard all of the predictors derived from post-publication analysis, as the goal of this study is pre-publication assessment. After this initial variable selection, we are left with 27 numeric predictors.

The *day\_of\_week* and *channel* categorical variables have already been one-hot encoded. We re-encoded these as factors in order to reduce the number of variables and simplify the analysis. Our statistical models will automatically one hot encode these during tuning and evaluation.

Some initial investigation revealed that directly predicting the number of shares as a linear output was difficult, so the target was transformed into a binary variable. Those articles with a number of shares above the median were given a classification of *popular* while those below were considered *unpopular*. This division allows us to have an even number of articles in each classification group, which should improve the model performance.

## Data Exploration

We can calculate Pearson’s correlation coefficient in order to examine the relationship between each predictor and the original shares variable, which is the number of shares the article received on the Mashable platform. The following shows these correlations between shares and each predictor, from least to most correlated.

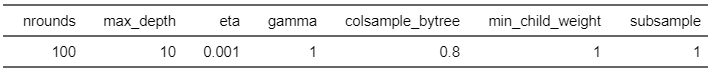


Predictors show weak correlation with target

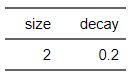
This shows that the linear correlations between the predictors and shares are quite weak, presenting challenges for prediction. Given that there are no clear linear relationships between any of the predictors and shares, accurately predicting popularity will likely be challenging.

## Model Tuning and Selection

Given the lack of linear correlation in our data, we choose two supervised learning models which can theoretically model non-linear relationships. The first is a Neural Net (NN) with a single hidden layer, and the other is a Boosted Decision Tree (BDT). We tuned each model on a subset of the data using a parameter grid with CV5. The BDT model was tested with depth from one to 10 and eta values (learning rate) across several different orders of magnitude. The NN was tuned with the hidden layer size from 3 to 10 and decay from 0.1 to 0.5. Since the dataset is fairly large, and these models are both computationally intensive to train, we used a randomly sampled validation set for the initial tuning. The tuning was performed on this subset of the data using the *caret* library with 5-fold cross validation. The data for the BDT was left un-scaled, but the data was centered and scaled for the NN.



Bosted Decision Tree parameters after tuning

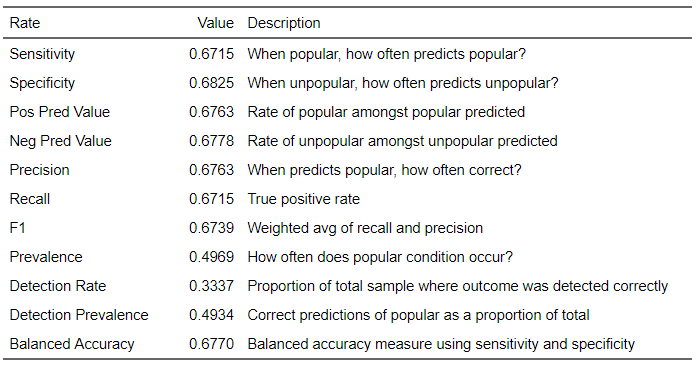


Neural Net parameters after tuning

After tuning, the two models were compared against each other for classification accuracy using five-fold double cross-validation with the entire dataset. The results of this comparison showed that the BDT outperformed the NN for all but one of the folds, but the accuracy was only slightly better. The overall accuracy from the double cross-validation was 0.636 with a corresponding error rate of 0.363. This is a lower performance level than we would have liked to achieve but but still indicates that the model has some predictive power.

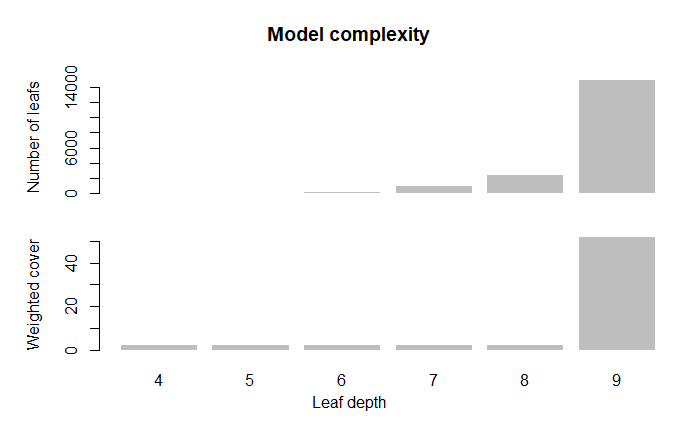
## Results

The BDT model was then used to predict popularity on the entire original dataset, with the following results.

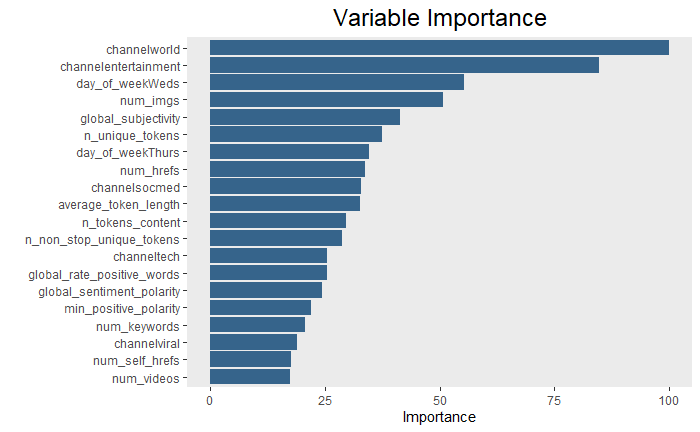


Summary table for final model results

This shows an increase of around 4 percent for sensitivity compared with the CV5 results, which is a welcome improvement. The final XGB model arrived at an optimal tree depth of 9 for most paths, based on the initial tuning, with a minimum of depth 4. Number of leaves was around 14000, indicating that this is an extremely complex model. The relatively low accuracy from the final, tuned model suggests that the correlations between the predictors and the target are weak. Since BDTs are typically one of the best performing models for classification, we conclude that the performance was probably not contingent on the model itself but the very weak associations with popularity.



The variable importance based on the final XGB model shows which predictors were most heavily used in the popularity classification model.



The variable importance calculation shows interesting associations between characteristics of the news stories and their popularity. For instance, the world and entertainment channels seem to be most popular; publishing on a Wednesday may have some positive effect on popularity; and the number of images in the story may increase its shares. The length of the story also seems to have been important in the classification. These associations can be used for evaluating potential stories and making changes in order to increase their number of shares. Versions before and after these changes can be compared using the models to assess their popularity. Even though our classification accuracy is sub-optimal, the model still does better than random guessing and has identified which variables in the original data are most related to whether an article becomes popular or not.

# Next Steps

Additional tuning could be performed of the model by varying the classification threshold, which was fixed at 0.5 in this analysis, in order to arrive at the best accuracy. The popularity classification could also be changed to use thresholds based on an upper quartile, rather than just the median. The number of samples would probably need to be balanced in this case between the two classifications for the models in order to be optimally tuned with equivalently sized groups. Finally, the classification could also be changed from a binary to a multi-class prediction on popularity level (low, medium or high, for instance instead of just popular or unpopular).