Online News Shares Popularity

Tim Kurowski, Nicole Root, Ritesh Singla, Abhishek Bhattacharjee

OPIM 5604: Predictive Modeling

Table of Contents

Executive Summary

The declining rate of individuals reading print news articles makes it difficult for businesses to reach their target audience. Unlike print media, the digitally enabled reader can quickly, easily and inexpensively choose amongst many possible sources of information at their fingertips. Given that advertising revenue in the digital space is often driven by reader clicks and other reader generated actions, publishers need to ensure that readers are being driven to their content to gain return on investment. If publishers could know if an article would be popular or not ahead of publication, it would create a strong competitive advantage in capturing an unfair share of advertising revenue.

Work in this space was originally done by K. Fernandes, P. Vinagre and P. Cortez in 2015. This paper presents a predictive modeling approach that narrows the originally propose definition of a "popular" article. Utilizing the "Online News Shares Popularity" dataset on the UCI Machine Learning Repository predictive models were built through logistic regression, lasso and ridge methods, Boosted Tree, Partition, Boosted Forest, and Neural Networks. The initial set of attributes included varying information on the number and depiction of words in the title and content, the data channel platform the article addressed, the weekday the article was published, the subjectivity of the article and more. An alternative subset of data was created with feature engineering and elimination of particular variables. This subset was then processed through the same predictive analytical models with varying parameters. It was then determined that the best predictive model was achieved through Bootstrap Forest without feature engineering which included a total of 46 attributes extracted from the Online News Popularity Dataset.

Business Understanding

In today's digitally enabled society, consumers of news content are increasingly turning away from print media in favor of digital media. According to the Pew Research Center, weekday and weekend newspaper circulation has declined at a CAGR of -8.3% and -7.9% respectively, since 2014 (Trends and Facts on Newspapers: State of the News Media, 2019). At the same time, monthly unique visitors to newspaper websites have increased from 8.23 Million in Q4 2014 to 11.60 Million in Q4 2018 and online advertising revenue has increased from $49.5 Billion in the U.S alone, to $107.5 Billion(Guttmann, A, 2018). Together, the growth in unique users and advertising revenue represents a clear opportunity for content publishers to grow their topline revenue through digital advertising. Given the growth that the market is experiencing, it would seem that the task at hand would be easier today, relative to the past. However, several factors are making capturing the attention of these consumers more difficult: First, the average time spent on newspaper websites has declined from 2.59 minutes to 2.32 minutes over the same period (Trends and Facts on Newspapers: State of the News Media, 2019). Second, consumers are faced with an overwhelming amount of data each day. For example, in a single day 2 million blog posts are written, 294 billion emails are sent and 168 million DVDs worth of information is consumed by internet traffic (Daniells, Kathy, 2012). With both decreasing time spent on websites and increasing saturation of content, the problem publishers are faced with is: "how do publishers select content that is more likely to be viewed as compared to the other content out in the marketplace?" If content is not viewed, advertising revenue will be sub-optimal. Therefore, selecting content more effectively, should in turn allow publishers to capture more than their fair share of advertising revenue.

A common measure of article popularity is the number of shares. Previous work in the space by K. Fernandes, P. Vinagre and P. Cortez, in 2015, defined a "popular" article as one that received more than the median number of shares (1,400) in their dataset of Mashable articles from January 7, 2013 to January 7, 2015. The authors then set out to predict whether an article would become "popular" by looking at the characteristics of the article prior to publication. Given the increased saturation of online media in the years following their study, simply having an article that gets more shares than the median is not enough. Publishers and advertisers need to have content that will truly stand out. The goal of the work behind this paper is to create a predictive model that will classify articles as falling into the top quartile of shares (2,800+). By focusing on the top quartile, as opposed to median, publishers will be able to narrow down the pool of possible online articles to those most likely drive the most viewership and therefore generate the greatest amount of advertising revenue.

Methodology – The SEMMA Process

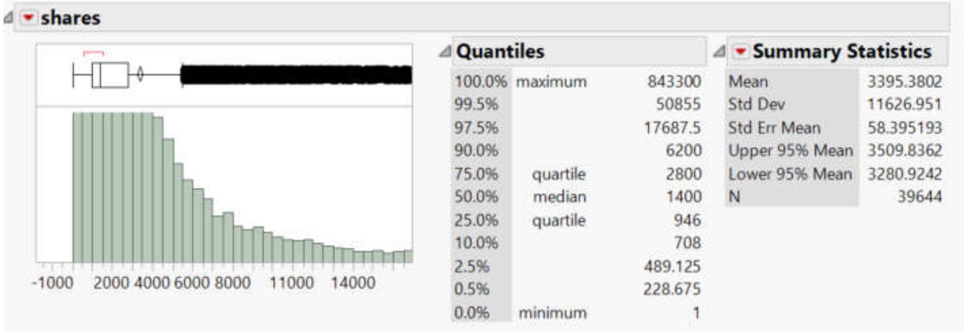
Sample

Sampling datasets to use for predictive modeling analysis, requires detailed research to gather domain knowledge of various attributes. After researching, it was determined that the Online News Shares Popularity dataset had potential within its variables, with some modifications. This dataset initially had 61 attributes (58 predictive attributes, 2 non-predictive, 1 goal field "Number of Shares"). It was clear that the business problem was in line with predicting the popularity of an online news article based on the number of shares.

With the dataset in the beginning stages, the next step was to split the data 50/50 between validation and training sets. This was done by using the seed 888 to keep reproducibility in the event on needing to recreate this split. Before moving to the baseline model, it was imperative to check the dimensionality of the dataset with the formula: $n>5(p+2)$. It is also important to note that both the training and validation sets have 19,822 observations with 58 predictive attributes. Where $n = 19822$ and $p = 58$, then $5(p+2) = 300$. Concluding that this data has appropriate dimensionality to use in a predictive model set.

Explore & Modify

The data exploration and modification process focused on understanding the distribution of the predictors, relationships amongst predictors, identifying missing values, reviewing outliers and determining the cut-off for the definition of the "Popular" Yes/No target variable in this study. For the latter focus of exploration, the goal was to determine a cut-off for what is popular or not to use as the response variable for predicting popularity of online news share articles. The approach included SEMMA "modify" step of deriving a new categorical variable: PopularYN to use as our target variable. This variable was created by finding the fourth quartile of all shares through a box-plot analysis of the distribution of the variable “shares”.



Quantiles

100.0%	maximum	843300
99.5%		50855
97.5%		17687.5
90.0%		6200
75.0%	quartile	2800
50.0%	median	1400
25.0%	quartile	946
10.0%		708
2.5%		489.125
0.5%		228.675
0.0%	minimum	1

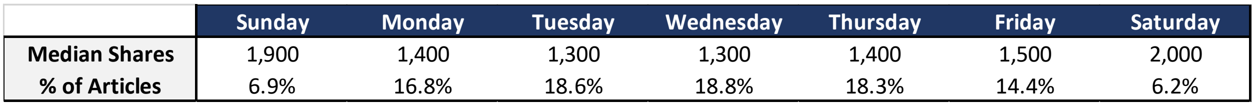
Summary Statistics

Mean	3395.3802
Std Dev	11626.951
Std Err Mean	58.395193
Upper 95% Mean	3509.8362
Lower 95% Mean	3280.9242
N	39644

As shown in the image above, the 75th percentile of shares is 2,800. When defining the target variable PopularYN, 2,800 was used to populate a 1, indicating, “Yes” the news article is popular. If the PopularYN variable populates a 0, then that will indicate that the number of shares for the news article was less than 2800 and is determined to not be popular.

The next process was to remove potential target leakage variables before creating the initial baseline test. Amongst these attributes, “timedelta” was removed, as this would not be available at the time of modeling. “timedelta” is defined as the “days between the article publication and the dataset acquisition.” The 5 LDA variables were also removed as this was the original author’s version of a Principal Component Analysis. The goal with this removal was to remove the influence of the prior author’s analysis from this study. It was determined that the data set has no missing values by performing JMPs Missing Values Screening capability.

Further investigation of the dataset revealed several correlated variables. These variables were dummy variables created for the seven days of the week and a weekend yes/no indicator. The decision was made to utilize the “is_weekend” predictor. The decision was made in part due to similarities that exist amongst the Saturday and Sunday as compared to the weekdays. As is shown in the chart below, the median shares are similar between Saturday and Sunday and are also similar amongst the five weekdays. Additionally, the percent of Articles in the dataset shows similarities in the same nature as shares does.



	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Median Shares	1,900	1,400	1,300	1,300	1,400	1,500	2,000
% of Articles	6.9%	16.8%	18.6%	18.8%	18.3%	14.4%	6.2%

The data set also had several right skewed predictors. These included "n_tokens_content", "num_hrefs", "num_self_hrefs", "num_imgs", "num_videos", "kw_max_min", "kw_min_max", "kw_max_avg", "kw_avg_avg", "self_reference_min_shares", "self_reference_max_shares", "self_reference_avg_shares". New columns were added to the dataset representing the log + 1 of each of the aforementioned predictors. Log+1 was used instead of log because predictors including Num_Self_hrefs had values of 0. As noted previously in this paper, each model type was run twice; once using the original version of the predictors and once using the log + 1 transformed version of the predictors.

The final aspect of feature engineering performed in the Explore and Modify steps was combining the "Data_Channel" predictors into like channels. Data_Channel_is_Lifestyle and Data_Channel_is_Entertainment were combined into Channel_LE, where if Data_Channel_is_Lifestyle = 1, then code Channel_LE with 1. Data_Channel_is_Entertainment = 1, then code Channel_LE = 1. If both of them = 0, then code Channel_LE to = 0. Data_Channel_is_World and Data_Channel_is_Business were combined into Channel_WB where if Data_Channel_is_World = 1, then coded Channel_WB with 1. Data_Channel_is_Business = 1, then code Channel_WB = 1. If both of them = 0, then code Channel_WB to = 0. No other Channel groupings were made.

Model & Assess

The modeling for this study can be approached from one of two general ways. The choice used in this study treats the target variable as categorical (Popular = Yes OR Popular = No). An alternative method would be to use Number of Shares as a continuous target variable. To predict the categorical target, seven classification methods were employed. Models were assessed with two goals in mind: Primary Goal = Achieve the highest AUC, Secondary Goal = In the event of an AUC tie, select the simplest model amongst contending models. Each of the seven classification methods were re-run with the feature engineered variables described in the prior section. Ultimately the choice of a final model came down to two predictive analytical techniques: Neural Network Classifier with a validation AUC of 0.6947 and Bootstrap Forest with the same AUC. Bootstrap Forest was selected as the final model because in following Occam's Razor, bootstrap forests are simpler than neural networks. The sections below highlight the process and parameters by which each model was built. The focus below is on the version of each model without feature engineering, as each feature engineered model resulted in a lower AUC relative to its non-feature engineered peer. Additionally, the original authors used Linear Discriminant Analysis to reduce the dimensionality of the dataset. In order to create distance between this analysis and the original author's analysis, the 5 LDA columns were removed when running all of the models below.

Logistic Regression:

Logistic regression was the first model chosen in our modeling process as it is the most explainable and straight forward. The model was first run with the 53 remaining predictors. During the initial run, it was discovered that the dummy 7 predictors titled "weekday_is_monday

/ tuesday / … / sunday" and the predictor "is_weekend" were biased. The cause for this is that "is_weekend" is a derived predictor from the 7 "weekday_is_" predictors. A choice was made to continue the modeling process with only the "is_weekend" predictor, thus the other 7 predictors were removed. This approach was followed in the 6 other classifiers that were explored for this study.

The logistic modeling process continued by removing predictors 1 by 1 based on the highest, insignificant p-value, rerunning the model between each removal. Ultimately 25 predictors were removed. The final logistic model yielded an AUC on validation of 0.6842. Digging a bit into the odds ratios, there were several observations that logically made sense. For example, "is_weekend" had an odds ratio of 1.621522, suggesting that news articles published on the weekend are 1.621522 times more likely to be popular than those that are not, i.e 62% higher chances to be popular when published on a weekend. This aligns with an earlier observation during the exploratory phase the research that the median and 3rd quartile of shares for weekend articles was 1,900 and 3,600 vs. 1,400 and 2,600 for weekday articles.

Ridge:

We ran RIDGE to understand if the variables that we excluded following the traditional p-value approach in our Logistic regression model have coefficients closer to zero using a K-Fold validation of 5 folds. In each of the insignificant variables, the coefficient was close to 0(three zeros left to decimal) bolstering our belief on our variable selection. The AUC is in this case was 0.6905 in case of Training Data Set and 0.6807 in Validation Data Set which is consistent with no sign of overfitting. The same RIDGE was run using our feature engineered variables and in this case our AUC was significantly on the lower side-0.5555 for Training and 0.5661 for Validation splits.

Lasso:

We ran LASSO taking 46 predictors using a K-Fold of 5 validation and compared those variables for which we had coefficients close to zero in case of RIDGE and p-value greater than 0.05(at 5% significance level) in case of normal LOGIT. We had 4 variables with coefficients shrunk to zero and the rest were very close to zero compared to LOGIT and RIDGE insignificant variables. AUC in this case was 0.6914 in case of Training set and 0.6782 in case of Validation set.

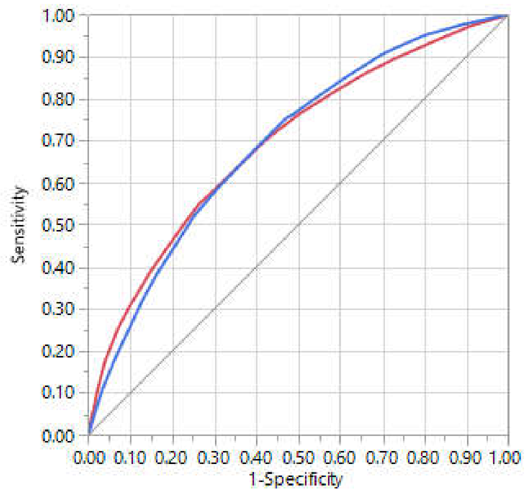
As was observed in case of RIDGE, when run on feature engineered variables LASSO gave a substantial lower AUC 0.5458 for Training and 0.5496 for Validation. Note that for either of the cases there was no sign of overfitting which suggests that for models based on feature engineered variables our model would have higher bias and lower variance.

Bootstrap Forest – "BEST" Model:

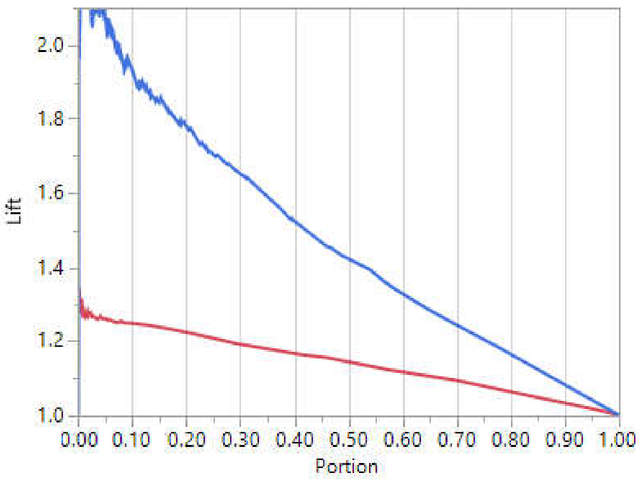
Bootstrap Forest was run initially with the default parameters in JMP. The initial run yielded a strong overfitting scenario whereby the Training AUC registered at 0.8574 compared to a Validation AUC of 0.7098. In order to alleviate the overfitting issue, the hyperparameters of the model were further tuned. Specifically, 15 tuning scenarios were tested by adjusting combinations of the Number of Trees in the Forest, the Minimum Splits Per Tree, the Maximum Splits Per Tree and the Minimum Split Size. Reducing the Maximum Splits Per Tree had the most substantial impact on reducing overfitting. On the surface, this is logical in that shallower trees should improve generalization of an overfit model by adding bias and thus reducing variance. Our final hyperparameters for Bootstrap Forest are shown below.

Name	*Number of Trees in the Forest*	Number of Terms Sampled Per Split	Bootstrap Sample Rate	*Minimum Splits Per Tree*	*Maximum Splits Per Tree*	*Minimum Split Size*	Training AUC	Validation AUC	Actual Number of Trees
BSF_11	*500*	11	1	*10*	*25*	*50*	0.7183	0.6947	46

Receiver Operating Characteristic on Validation Data



Lift Curve on Validation Data



Partition Decision Tree:

When running the model with the featured splits, the partition founded the best number of tree splits to be 21 with an AUC for the validation data of 0.6679. The featured engineered peer had an AUC of 0.6673.

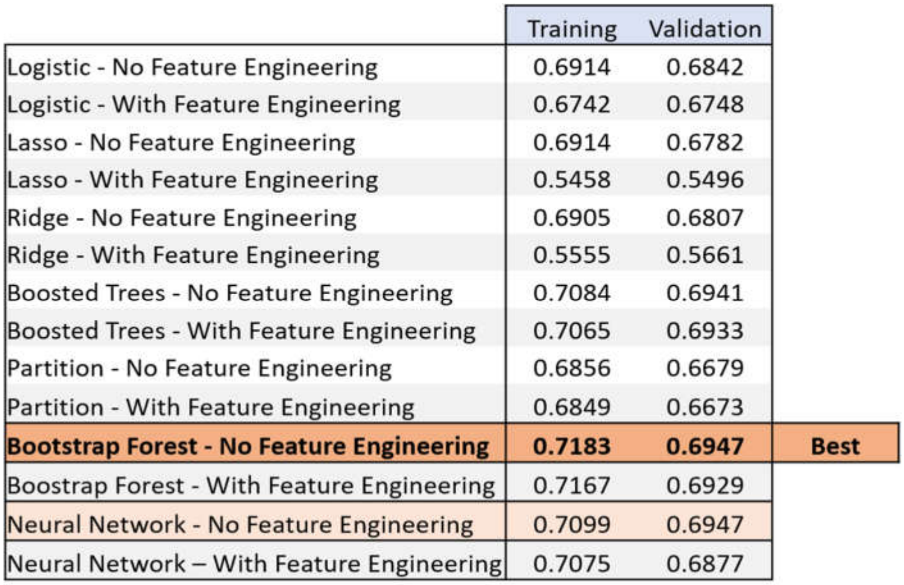
Boosted Trees:

A series of Boosted Tree models were run. The best model was fund with 50 Layers and 3 Splits Per Tree. This yielded an AUC on Validation of 0.6941. The feature engineered peer yielded a Validation AUC of 0.6933. The Boosted Trees without feature engineering was a contender for the “Best” model, however, Random Forest provided slightly better results.

Neural Network:

The final modeling technique we explored was a neural network with 1 hidden layer. Results of the Neural Network without feature engineering produced a Test AUC of 0.7099 and a Validation AUC of 0.6947. Compared to our “best model” choice of Random Forest, Validation AUC matched exactly to the fourth decimal point. The neural network yielded closer AUCs between Test and Validation, thus exhibiting less variance. However, the delta between Test and Validation in the Bootstrap Forest was only modestly larger.

Our final AUC values for all model runs can be seen below



	Training	Validation	
Logistic - No Feature Engineering	0.6914	0.6842	
Logistic - With Feature Engineering	0.6742	0.6748	
Lasso - No Feature Engineering	0.6914	0.6782	
Lasso - With Feature Engineering	0.5458	0.5496	
Ridge - No Feature Engineering	0.6905	0.6807	
Ridge - With Feature Engineering	0.5555	0.5661	
Boosted Trees - No Feature Engineering	0.7084	0.6941	
Boosted Trees - With Feature Engineering	0.7065	0.6933	
Partition - No Feature Engineering	0.6856	0.6679	
Partition - With Feature Engineering	0.6849	0.6673	
Bootstrap Forest - No Feature Engineering	**0.7183**	**0.6947**	**Best**
Boostrap Forest - With Feature Engineering	0.7167	0.6929	
Neural Network - No Feature Engineering	0.7099	0.6947	
Neural Network – With Feature Engineering	0.7075	0.6877	

Evaluation

Our results should be evaluated under two frameworks; One focused on number of shares, the other focused on advertising revenue.

Assumptions:

- Views of articles and advertising revenue increase directly proportional to number of shares
- The maintenance cost of the model is a fixed cost per year over the life of the model
- Model would be deployed only if marginal gain in revenue over baseline exceeds the fixed cost of maintaining the model.

Framework 1 – Number of Shares

The goal of Framework 1 is to evaluate whether or not our model leads to publishers selecting a greater concentration of articles that have greater than 2,800 shares. We propose setting up a test by which publishers select three sample groups of articles to publish. Sample 1 would include articles that our classifier ranks as having a probability in the top two deciles as defined by the lift chart. For reference, our lift at Decile 2 is approximately 1.78. Sample 2 would consist of articles selected from the lower 4 deciles. Sample 3 would be a selection based on the publisher's previous criteria for article selection. Success would be defined as the model selecting a higher concentration of popular articles than the previous method employed by the publisher. Comparisons to the lower 4 deciles would be made for reference.

Framework 2 – Advertising Revenue

During the Framework 1 test, we would also capture and compare the advertising revenue generated by each of the 3 samples of articles. For the model to be economical, Sample 1 (top 2 deciles) would have generated advertising revenue that is both in excess of Sample 2 as well as projected to in the long-term cover the cost of maintaining the model. Per our assumptions, financial success can be defined as the following statement being true:

$$(Annual\ Revenue_{Model} - Annual\ Revenue_{NoModel}) > Annual\ Model\ Cost)$$

Deployment

In the marketplace, the model would sit as a pre-screening tool that scored articles prior to reaching the department in charge of deciding if an article gets published or not.

- **Step 1** - Editors submit final copy of article to repository each day
- **Step 2** - Model would run against the articles for that day and assign a probability of being in the "Popular" class.
- **Step 3** - The articles would be ranked most likely to least likely and sent to the team responsible for making publishing decisions.

Though the model would drive additional revenue into the business, there are some important ethical considerations to be aware of. The most concerning ethical consideration is that editors and writers discover what drives the model to determine if an article will become popular or not. If discovered, editors and writers may adjust their writing style and article structure in such a way that it is more likely to be ranked high by the model. Additionally, weekend articles stood out as having a higher likelihood to be chosen. If known, writers and editors may withhold important stories from submission until the weekend. In this case, important information may not get out to the public in a timely manner. This could be mitigated in a number of ways. First, the details of the model must be kept from the writers and editors. Second, publishers should be given the full list of ranked articles, not just those most likely to be popular. This would give the team responsible for selection an opportunity to balance the content of an article and the potential popularity when selecting articles. Additionally, the compensation of all parties should not be tied exclusively to popularity and/or advertising revenue.

Appendix

References

Daniells, Kathy. "Infographic: 24 Hours on the Internet." Infographic: 24 Hours on the Internet, Digital Buzz Blog, 13 Mar. 2012, http://www.digitalbuzzblog.com/infographic-24-hours-on-the-internet/.

Fernandes, Kelwin, et al. "Online News Popularity Data Set." UCI Machine Learning Repository: Online News Popularity Data Set, UCI, 8 Jan. 2015, https://archive.ics.uci.edu/ml/datasets/Online News Popularity.

Guttmann, A. "U.S. Online Advertising Revenue 2018." Statista, Statista, 9 Aug. 2019, https://www.statista.com/statistics/183816/us-online-advertising-revenue-since-2000/.

Meyer, Robinson. "How Many Stories Do Newspapers Publish Per Day?" *The Atlantic*, Atlantic Media Company, 26 May 2016, https://www.theatlantic.com/technology/archive/2016/05/how-many-stories-do-newspapers-publish-per-day/483845/.

"Trends and Facts on Newspapers: State of the News Media." Pew Research Center's Journalism Project, Pew Research Center, 9 July 2019, https://www.journalism.org/fact-sheet/newspapers/.

Documentation of Models and Supporting Materials – See Zip File