Predicting Online News Popularity Using Classification Models

Kee-Young Shin, Junyuan Zheng, Manali Phadke

Introduction

In this analysis, we attempted to predict whether or not an online news article would be popular. We used data from Mashable.com, an online social media site, containing various features regarding online articles. In a previous analysis of this dataset, we used a variety of regression techniques, Ordinary Least Square (OLS), Ridge, Lasso, Principal Components Regression (PCR), and Partial Least Squares (PLS), Generalized Additive Model (GAM), as well as Multivariate Adaptive Regression Splines (MARS) to try and predict the number of shares. While we found that GAM and MARS had the best performance among these models, we believed there was still room for improvement. Rather than trying to predict the number of shares, using classification methods to try and predict whether or not an article will be popular may give more accurate and useful models. Additionally, one particular observation seemed to have caused a rather high RMSE using the regression models. We believed classification models could better handle such observations.

Just as with our previous analysis, these classification models could provide insight into the factors that are most important in determining whether an article will be popular, thereby helping authors write more effectively. Publishers could benefit from knowing which factors may potentially help articles be more visible and popular online. Advertisers could use this type of model to predict whether an article will be popular and may be able to make appropriate adjustments. The main goal of our analysis is to find the model that will give the best classification of online news articles into popular vs unpopular.

After determining the best model(s) by comparing classification accuracy, we would like to identify some of the most “important” variables in whichever model(s) is deemed optimal. This indicates which variables are most crucial for classification and is something that those using our recommended model in the future could keep in mind. If our recommended models are complex, we would like to find some way of visualizing or interpreting our model, such as partial dependence plots or individual conditional expectation curves.

Data and Methods

The dataset was collected from Mashable.com, an online social media site, which we in turn obtained from the UCI Machine Learning Repository. There are 39,644 observations, each of which stands for one article published on their website, collected from 2013 to 2015. The dataset contains 45 features regarding keywords, language, links, publishing day, among others (see Table 1).

Pre-processing

All of the data processing procedures were implemented using R statistical software version 3.5.1 (RStudio). Channel types (life style, entertainment, business, social media, technology, and world) and publishing days (Monday through Sunday) are two groups of categorical variables, which were originally coded as a series of dummy variables. We first gathered each of the two groups into one column to have a cleaner format for analysis. In particular, since not every observation belongs to these six types of channels, we created an “other” type for those observations and made this the reference category. In this study, we assumed a binary classification task, where an article was considered to be “popular” if the number of shares was higher than a cut-off value (C), else it was considered to be “unpopular”. We chose C = 1,500 shares, which gave a balanced class distribution in the whole original dataset. Given the computational limitation of our available

devices, we chose to use only 5% (1984 observations) of the original dataset. 70% (1390) and 30% (594) of the observations were split into the training and test datasets, respectively.

The variable “timedelta”, which stands for days between the article publication and the dataset acquisition, was not included in our analysis because we wanted to be able to predict the popularity before the article was published. Three variables, the rate of unique words in the content, rate of non-stop words in the content, and rate of unique non-stop words in the content, were found to have nearly zero variance and were therefore removed from the dataset. Because of collinearity issues that arose in later model fitting, we decided to take out two variables, closeness to LDA topic 4 and publishing days.

Model selection

Training data was used to conduct model selection. In this study, we considered nine classification models: Logistic Regression without regularization or Generalized Linear Model (GLM); Logistic Regression with regularization or Elastic-Net Regularized Generalized Linear Model (GLMN); Linear Discriminant Analysis (LDA); Quadratic Discriminant Analysis (QDA); Classification Tree (CART); Random Forest (RF); Adaptive Boosting or Gradient Boosting Machines (GBM); Support Vector Machine with linear kernel (SVML); and Support Vector Machine with Radial Basis Function kernel (SVMR). The “caret” package in R was used for fitting these models to the training dataset. 10-fold cross validation was performed for tuning the hyperparameters (alpha and lambda for GLMN; complexity parameter for CART; number of predictors at each split and minimum node size for RF; number of trees, interaction depth, and the shrinkage parameter for GBM; budget parameter C for SVML and SVMR; sigma for SVMR). Lastly, we compared accuracy and Cohen’s kappa from the results of cross validation as the criterion for model selection. The models selected were further used to perform prediction on the test dataset. The workflow of this procedure is shown in Figure 1.

Results

Exploratory data analysis

The original data contained no missing values. Plotting of variable distributions did not show a clear separation of the two classes. The correlation coefficient graph of variables shows several correlations (results not shown), and as long as these correlations are not exactly 1 or -1, we decided to leave them in the data. We also looked at different number of clusters using silhouette analysis. The result shows that when trying a number of clusters ranging from one through seven, using two clusters has the highest average silhouette width, which validates our choice of using a binary classification for this data.

Model selection and performance on test data

Figures 2 and 3 show the results of cross validation for the nine fitted models using accuracy and kappa as metrics, respectively. Results using these two performance criteria agree with each other. SVMR has the best performance, while there is an outlier that gives relatively poor accuracy. Since SVMR, RF, and GBM has similar performance that is generally better than the other seven models, we selected these three models to be further tested on the test dataset.

Table 2 shows the performance of the trained SVMR, RF, and GBM models on the test dataset. The three models have similar performance, which are, however, worse than their performance

on the training data. Among the three models, random forest gives the best accuracy and kappa on predicting online news popularity.

Important variables

To further understand the models, we looked at the variable importance for random forest using the "ranger" package in R. The basic idea of the variable importance method is to check whether a variable of interest has a positive effect on the prediction performance. One way is to permute the values of all observations from the out-of-bag (OOB) data for one variable of interest and calculate the accuracy difference from a single tree. Then, the average of importance values from all trees in the random forest gives the permutation importance of that variable. Another way is by calculating the total decrease in node impurities (Gini index) for a variable of interest, averaged over all trees of the ensemble.

Figure 4 shows the variable importance of the fitted random forest. Using the two methods mentioned above, the results both show that the average share of average ranked keywords (kw_avg_avg) plays an important role in predicting online news popularity among all variables in the model. Variable importance from the boosting model shows a similar result (see Figure A1). On the other hand, from the permutation results, the number of keywords (num_keywords) and maximum polarity of negative words (max_negative_polarity) gives even a higher error in the prediction. We could consider removing these two variables in the future for better performance.

Partial dependence plots and individual conditional expectation curves

To further understand how the average number of average keywords affects the prediction of online news popularity, we looked at the partial dependence plots (PDP) for fitted random forest and boosting (Figure A2). The two plots show similar trends of an average change in the probability of predicting popularity while varying the predictor of interest and holding the other predictors constant. In a particular range, as the value of this predictor increases, the average probability of predicting the response variable decreases from above 0.5 to below.

To look more deeply into the change in prediction for each observation, we plotted the individual conditional expectation (ICE) curves for random forest and boosting (Figure 5). Interestingly, when we look at the non-centered ICE curve for random forest, there are two clusters of observations, one of which is generally above 0.5 and the other is below. At any given value of this predictor, random forest is able to distinctly separate observations into two groups. However, the ICE curve for the boosting model does not show a similar pattern. This may reflect why random forest had a better performance on predicting online news popularity.
Since the minimum shares of referenced articles in Mashable (self_reference_min_shares) is another feature shown to be an important variable, we also checked its PDP and ICE curves (Figures A3, A4), which gives the same conclusion as discussed above. The joint effect of these two predictors on averaged possibility of predicting popularity shows two distinct probability regions (see Figure A4).

Discussion

In this study, we used data containing information of online news articles collected from Marshable.com and aimed to predict their popularity. We assumed a binary classification task by cutting the number of shares into "popular" and "unpopular" groups. All of the predictors were collected before the articles were published. We hope this model can give people an idea of how popular an article could be in the future as well as insight into why an article has been popular.

Among the nine classification models we fitted, SVM with radial kernel, random forest, and adaptive boosting had a relatively better performance on the training dataset. It is possible that the truth and the relationships between the features and the response are quite complex, and something that can only be captured by more complex models. However, on the test data, these three models generally gave a poorer performance, with the highest accuracy of 0.6498 provided by random forest. Given this is a binary classification problem, using a random guess will end up with roughly 0.5 accuracy, which makes our prediction model to be mediocre.

Important variables

Measures of variable importance can be a good indicator of which predictors are highly used to make accurate predictions in the model. The variables representing the average shares of average keywords (kw_avg_avg) and the minimum share of referenced articles in Mashable (self_reference_min_shares) were found to be among the top ten most important variables for both the Random Forest (both permutation and impurity) and Boosting models. These variables do make logical sense in their significance in predicting popularity. For example, keywords, which are often utilized by authors and websites to help articles pop up at the top of search engine results pages and draw internet traffic, can be seen as a clear factor of an article's popularity. Similarly for the share of referenced articles in Mashable, the number of links in an article may result in Mashable.com visitors to peruse the article.

Limitations

There are several limitations we faced during this analysis. First, due to computational limitations of our devices, only 5% of the original data was chosen, of which 70% (1390) was used to train the models. Even though we randomly picked this subset, the information may not be enough to capture all of the features in the original dataset. Similarly, some models may not have been tuned as well as possible due to computational limitations. Trying a wider range and more values of tuning parameters may have resulted in better performance. Secondly, the choice of cut-off value for number of shares to get a balanced class distribution is arbitrary and debatable. It may be better to choose a cut-off point that can clearly separate the distribution of number of shares into two groups. Third, we did not include the time between the article publication and the dataset acquisition (timedelta) into the models, even though our previous regression study showed it to be an important variable in predicting the number of shares. A way of offsetting its influence on the number of shares may end up with a more precise popularity distribution. Lastly, as we see from the deciles from the PDP (Figures A2, A3), the distributions of some variables are skewed and the probability of prediction can become very sensitive in a given range of the predictor. We can conduct transformations (e.g., log) on such predictors, while losing some interpretability in exchange.

Figures and Tables:

Table 1. Brief information of the variables in the original dataset.

Variable	Type	Variables	Type
Words Related		**Language Related**	
# of words in the title	numeric	Closeness to LDA topic 0	numeric
# of words in the article	numeric	Closeness to LDA topic 1	numeric
Avg word length	numeric	Closeness to LDA topic 2	numeric
Rate of non-stop words*	numeric	Closeness to LDA topic 3	numeric
Rate of unique words*	numeric	Closeness to LDA topic 4*	numeric
Rate of unique non-stop words*	numeric	Title subjectivity	numeric
Links Related		Article text subjectivity score and its absolute diff. to 0.5	numeric
# of links	numeric	Title sentiment polarity	numeric
# of Mashable article links	numeric	Rate of positive and negative words	numeric
Min # of shares of Mashable links	numeric	Positive words rate among non-neutral words	numeric
Avg # of shares of Mashable links	numeric	Negative words rate among non-neutral words	numeric
Max # of shares of Mashable links	numeric	avg polarity of positive words	numeric
Digital Media Related		Min polarity of positive words	numeric
# of images	numeric	Max polarity of positive words	numeric
# of videos	numeric	Average polarity of negative words	numeric
Time Related		Min polarity of negative words	numeric
Days b/w publication and data acquisition*	numeric	Max polarity of negative words	numeric
Day of the week*	categorical	Article text polarity score and its absolute diff. to 0.5	numeric
Published on a weekend?	categorical		
Keywords Related			
# of keywords	numeric	Max # of shares for the Best keyword	numeric
Min # of shares for the worst keyword	numeric	Min # of shares for keywords in average	numeric
Avg # of shares for the worst keyword	numeric	Average # of shares for keywords in average	numeric
Max # of shares for the worst keyword	numeric	Max # of shares for keywords in average	numeric
Min # of shares for the best keyword	numeric	Article category	categorical
Avg # of shares for the best keyword	numeric		
Target Variable			
# of shares			numeric

*** variables not included in our model**

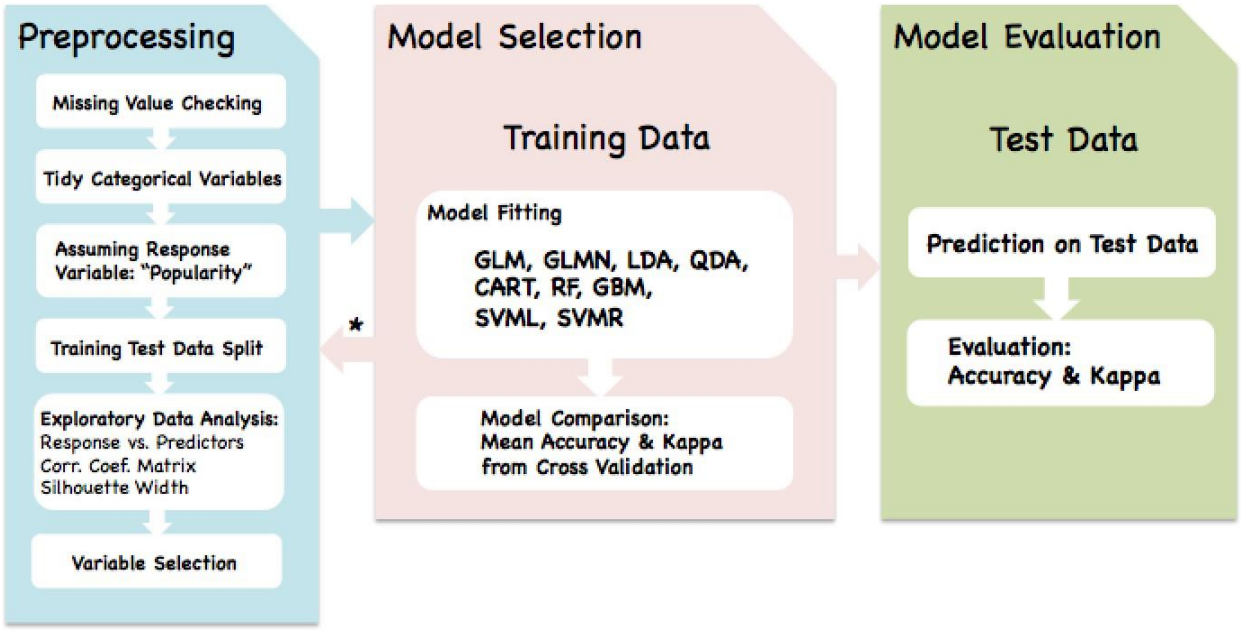

Fig. 1. Workflow diagram of the main procedures.
*Overall performance of candidate models may affect the decisions on variable selection.

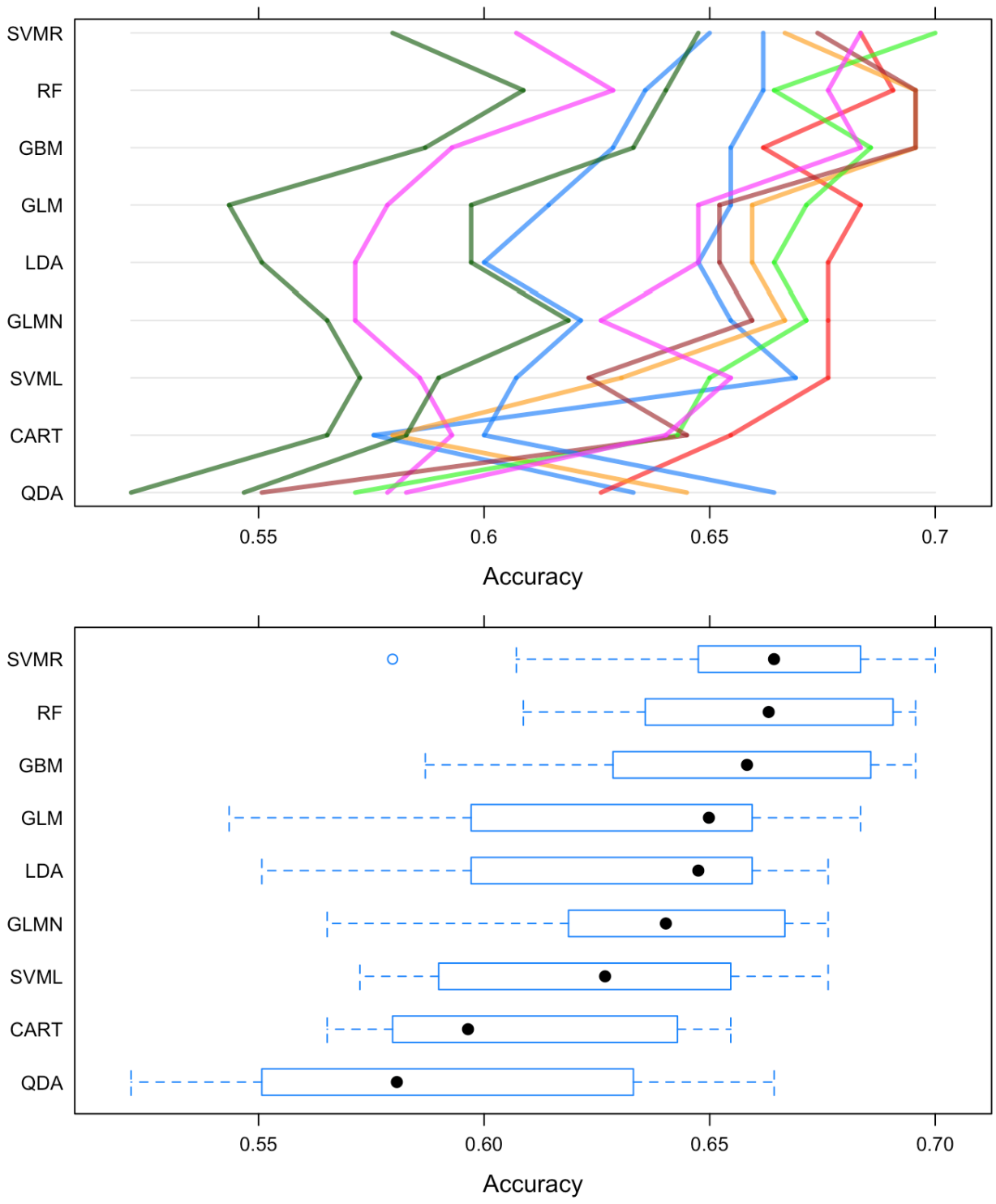

Fig. 2. Parallel plot of accuracy from 10-fold cross validation for different models (upper) and the corresponding boxplot (lower).

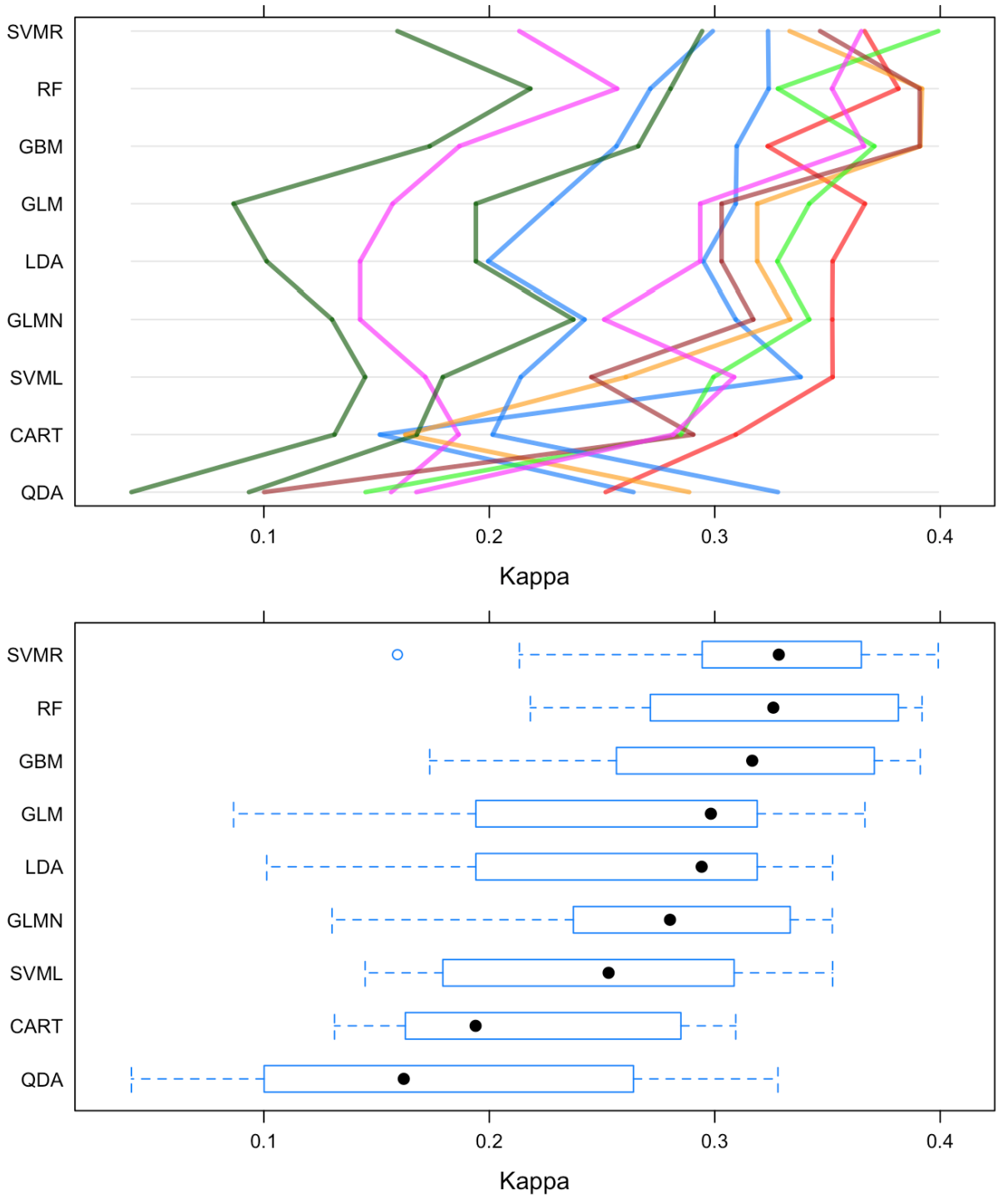

Fig. 3. Parallel plot of Kappa from 10-fold cross validation for different models (upper) and the corresponding boxplot (lower).

Table 2. SVMR, RF, and GBM performance on test data.

Model	Accuracy on test data	Kappa on test data
SVM with Radial Kernel (SVMR)	0.6263	0.2521
Random Forest (RF)	0.6498	0.3003
Boosting (GBM)	0.6380	0.2765

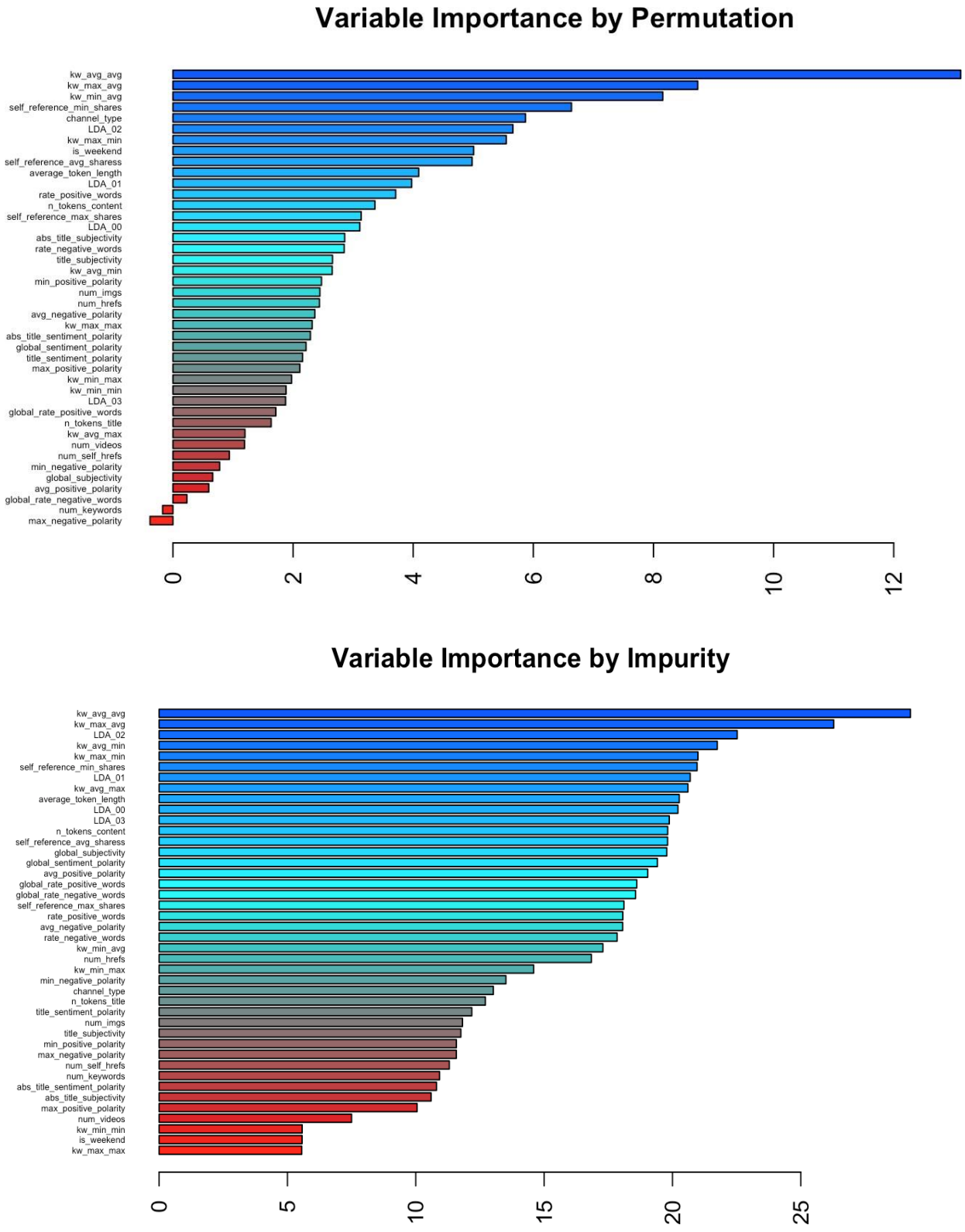

Fig. 4. Variable importance of random forest by the permutation strategy (up) and calculating the impurity (down).

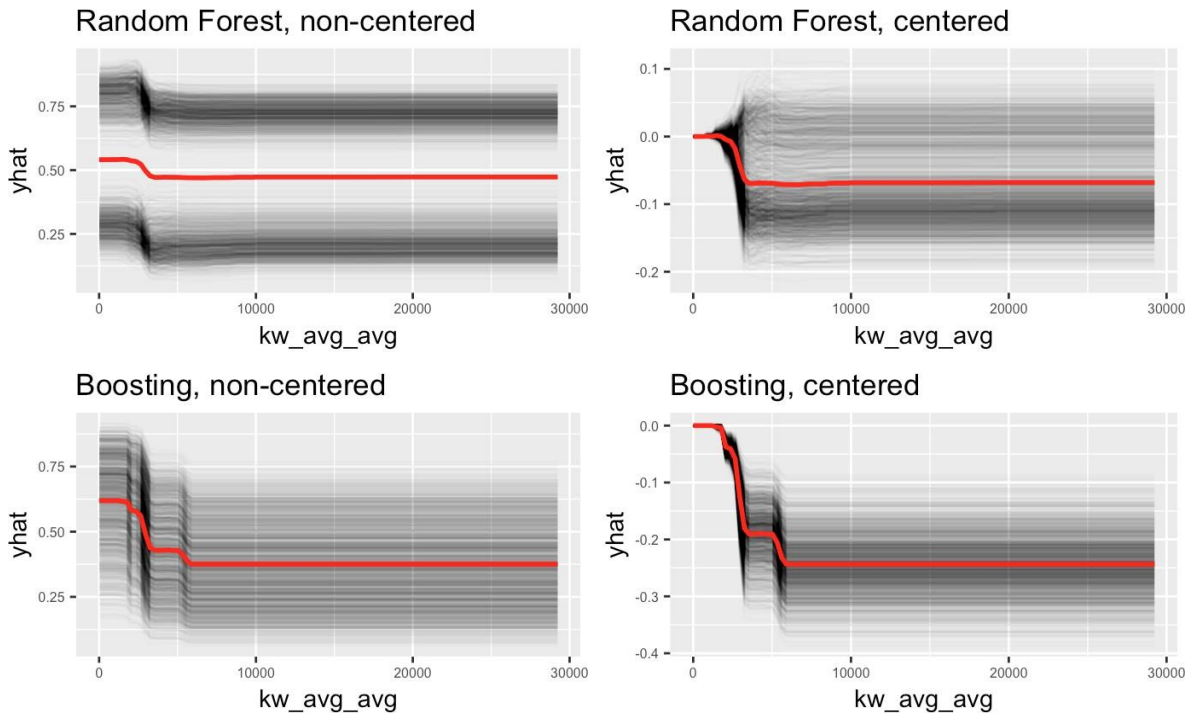

Fig. 5. Individual Conditional Expectation (ICE) curve of average number of shares of keywords in average, on random forest (up) and boosting (down).

Appendix:

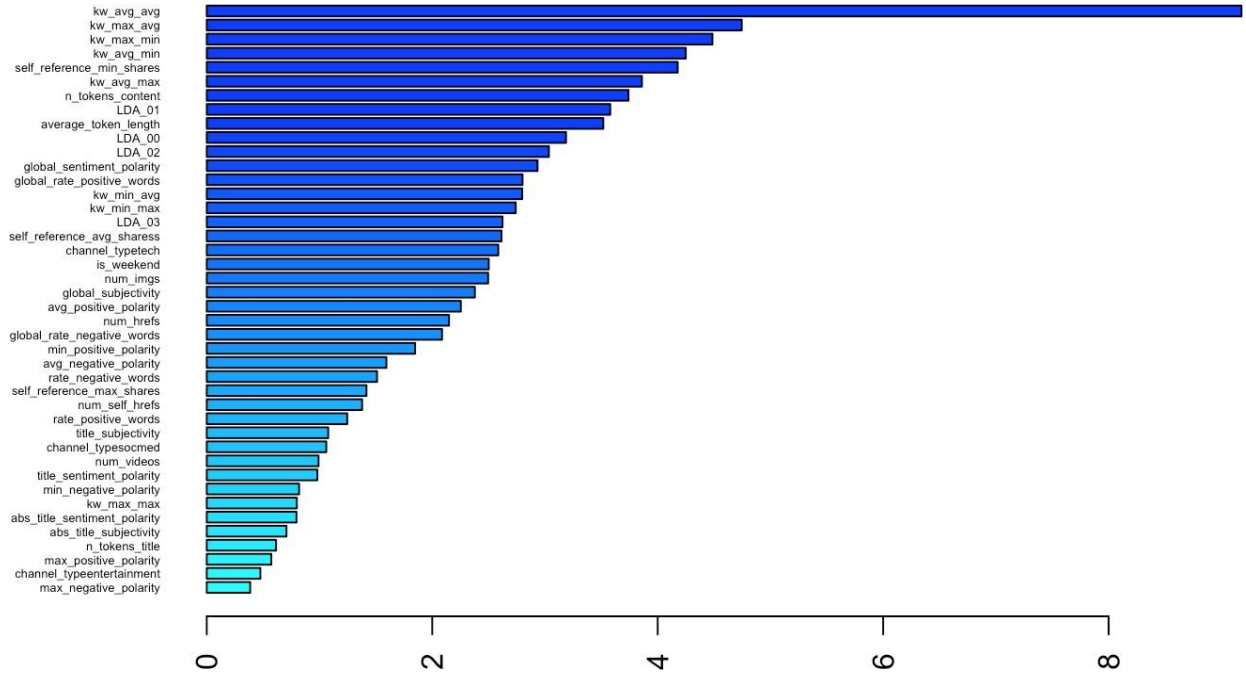

Fig. A1. Variable importance of boosting.

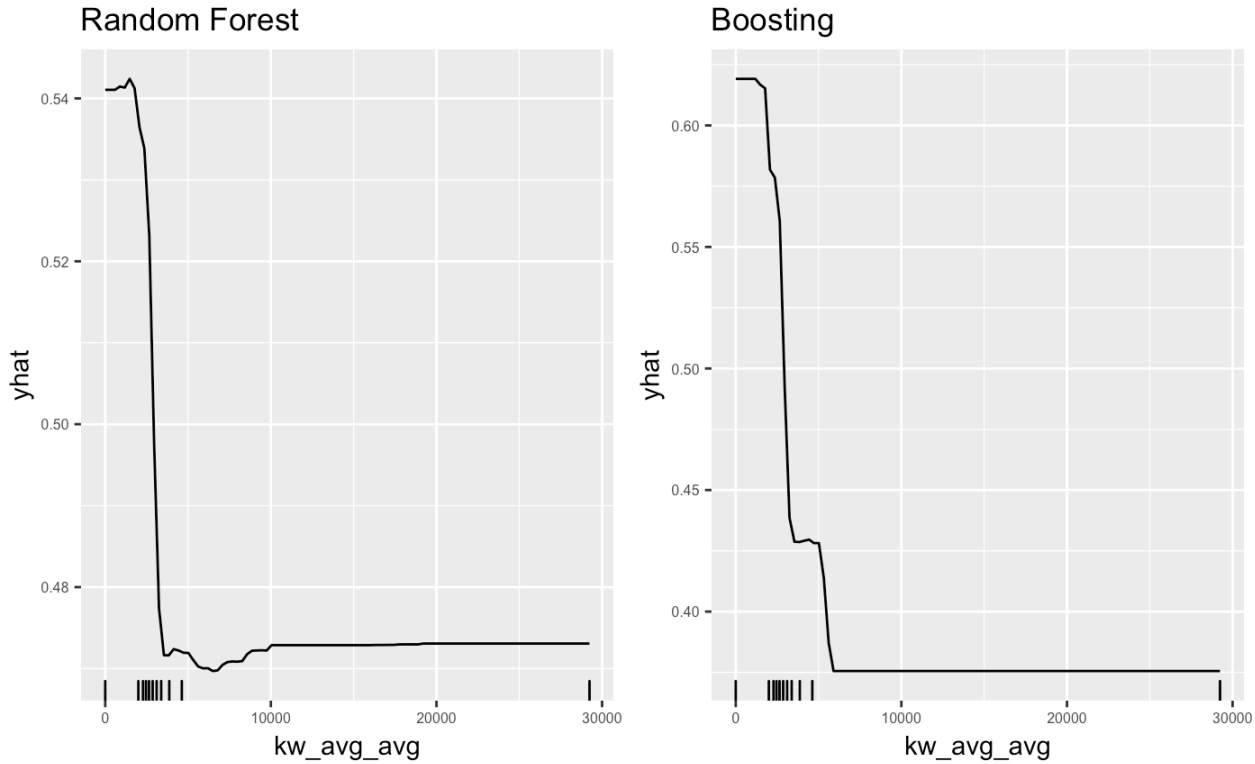

Fig. A2. Partial Dependence Plots (PDP) of average number of shares of keywords in average, on random forest (left) and boosting (right).

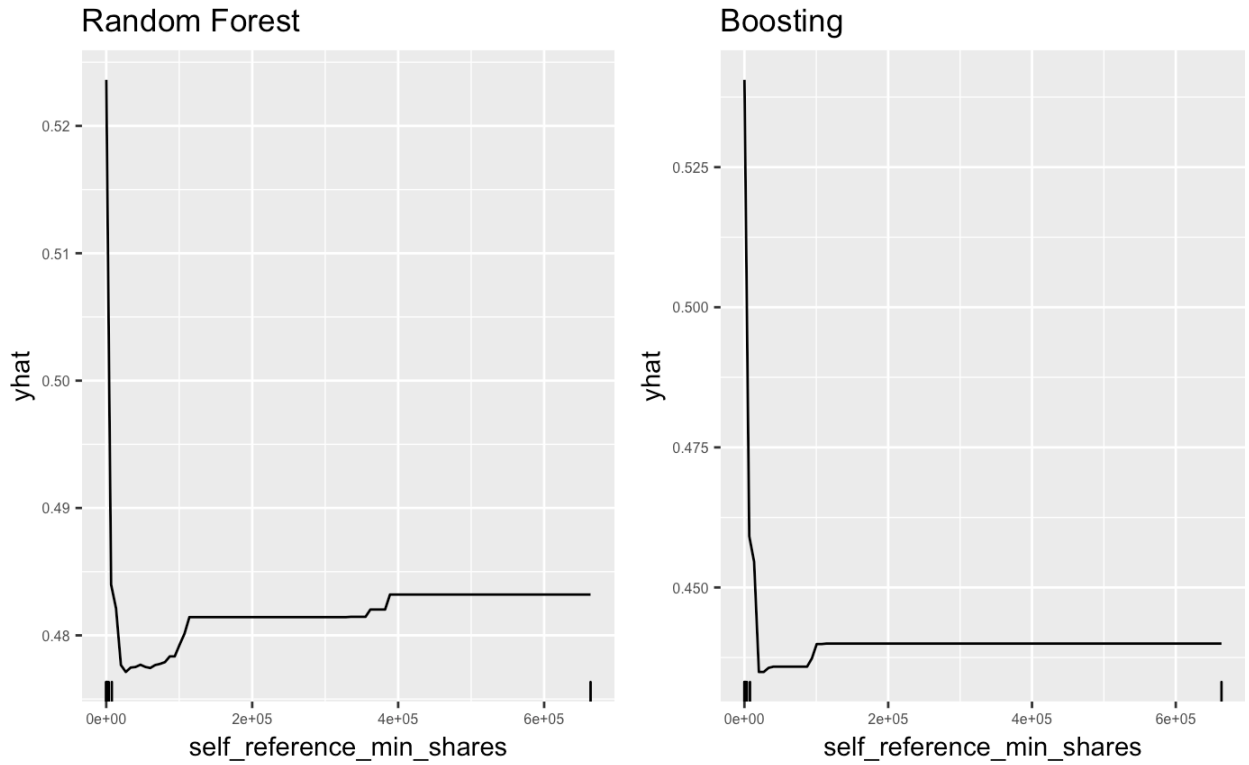

Fig. A3. Partial Dependence Plots (PDP) of minimum shares of referenced articles in Mashable, on random forest (left) and boosting (right).

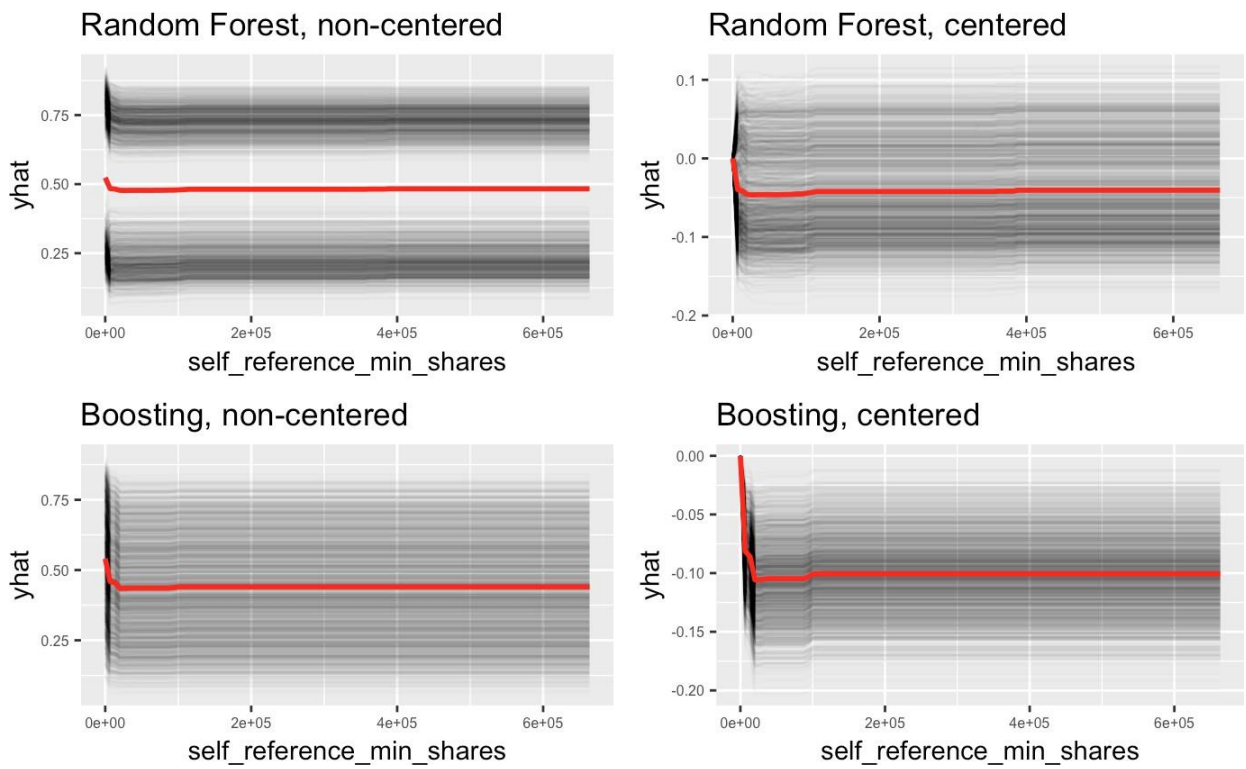

Fig. A4. Individual Conditional Expectation (ICE) curve of minimum shares of referenced articles in Mashable, on random forest (up) and boosting (down).

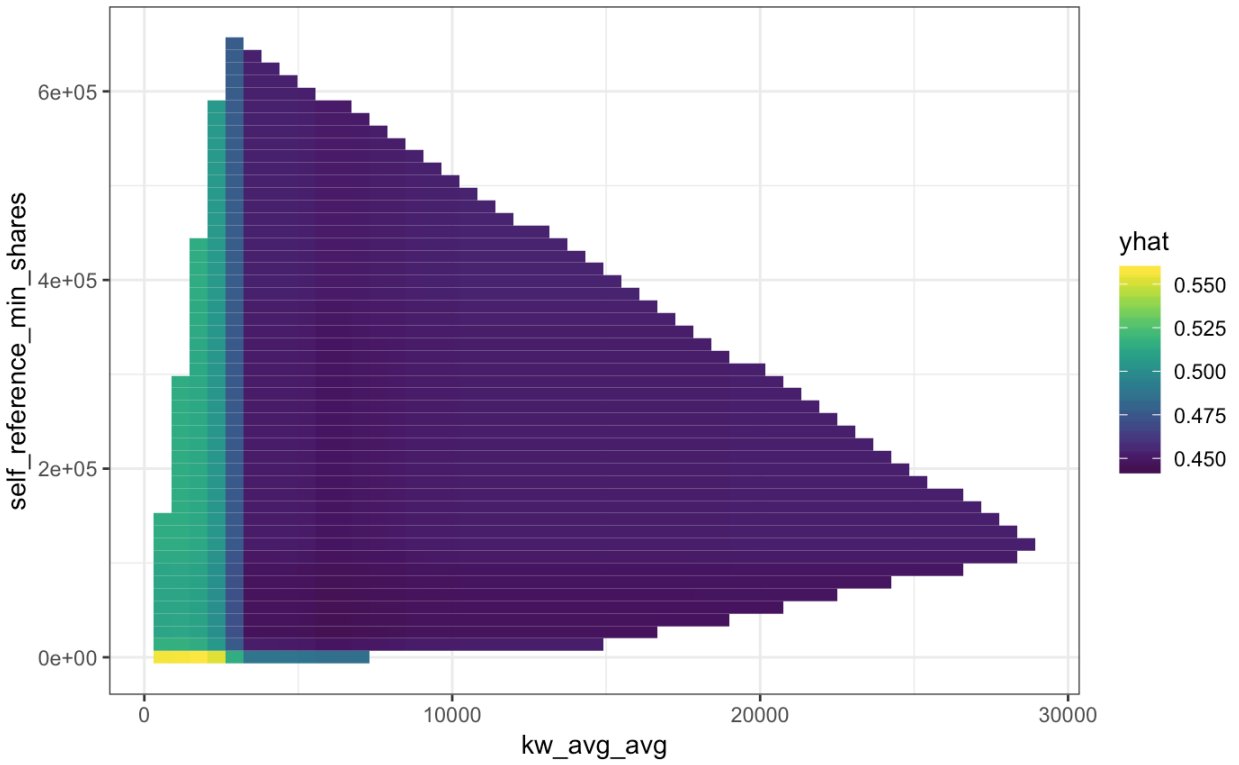

Fig. A5. Partial dependence of popularity on average number of shares of keywords in average and minimum shares of referenced articles in Mashable.