kaggle_datadiscovery

Holler Zsuzsa

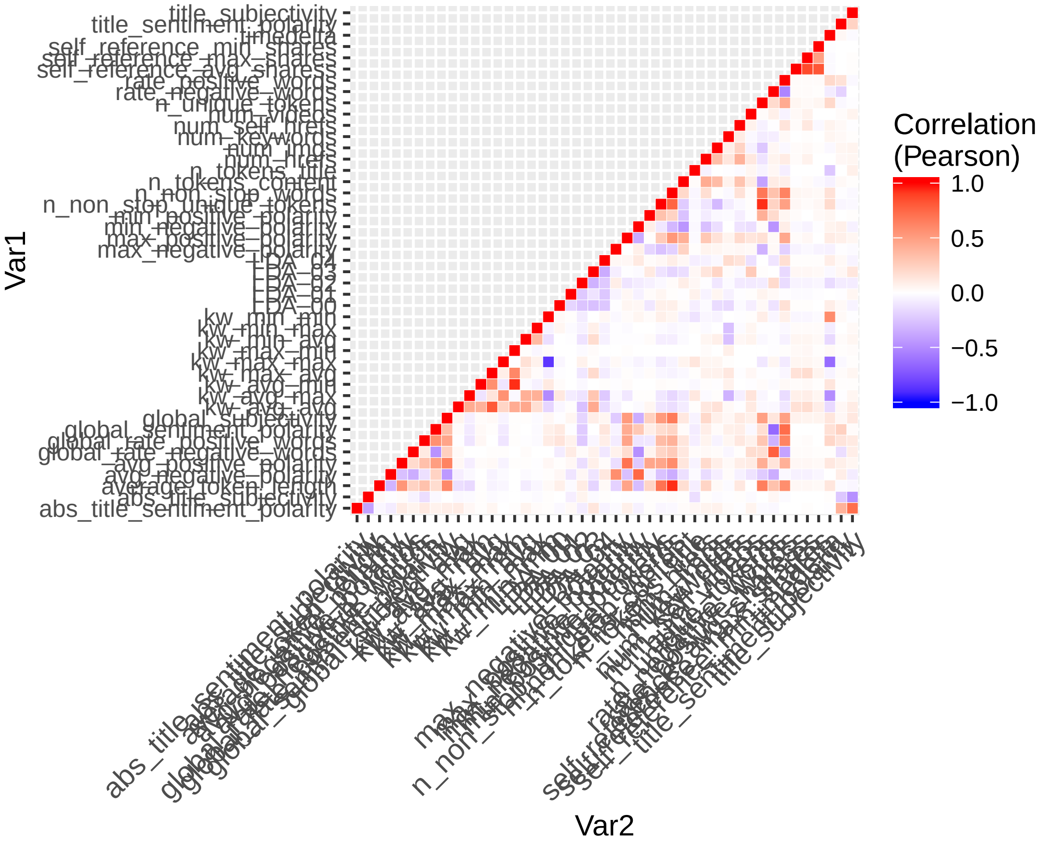
January 30, 2016

Content summarized in 59 features of all articles published from January 7 2013 to January 7 2015 on Mashable. 39,000 articles in total.

Data discovery

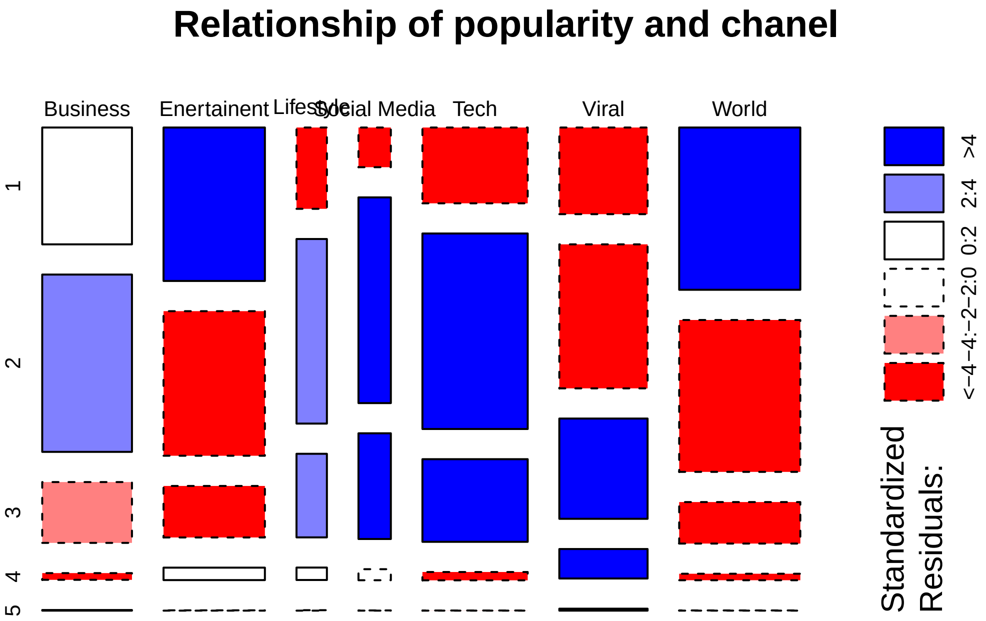
Explore relationship between numeric data

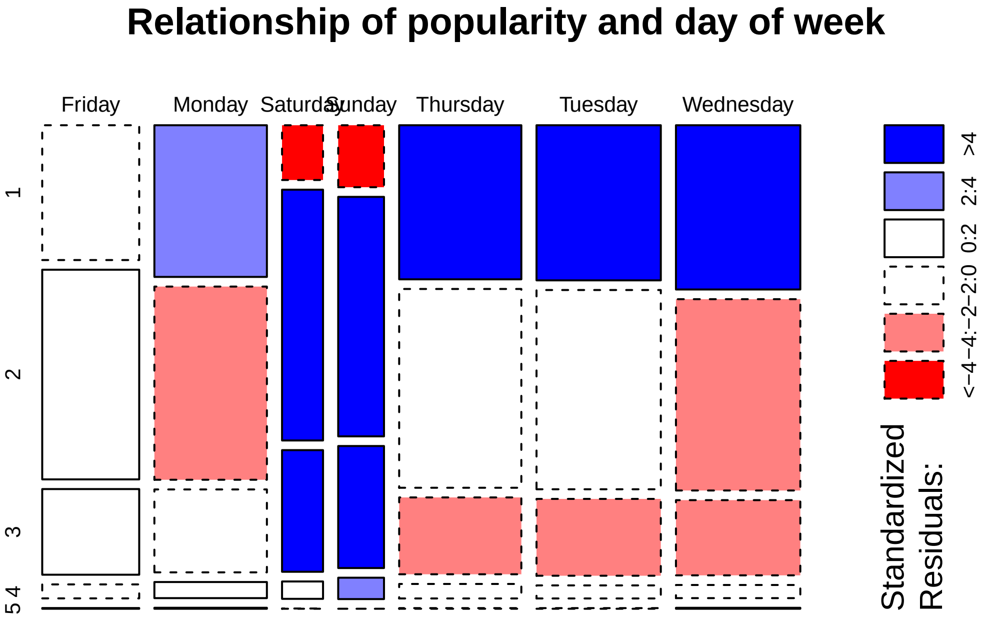
Simple correlations for non-categorical variables visualised. Stronger correlations between variables of the same category. It might go the the PCA section?



Explore relationships between categorical variables (including our y variable.)

Mosaic plot: Shows empirical conditional probabilities of row variable given column variable. Colored by Pearson residuals which shows the departure of each cell from independence. It seems that days are correraletd with the popularity especially significant difference between weekday and weekends. Also, chanel seems to be strongly related to the popularity of articles.

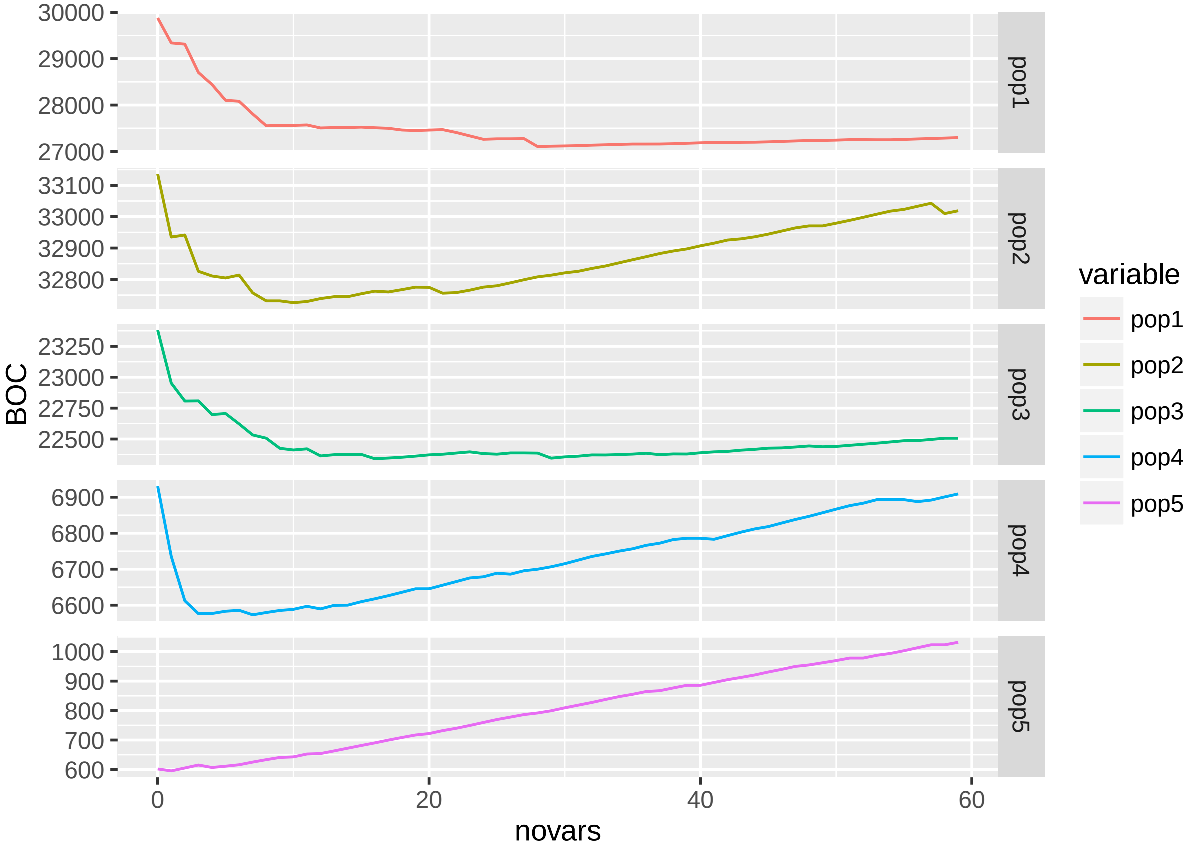

Relationship of popularity and chanel


Relationship of popularity and day of week

Fisher score based logistic regression

For finding the right variables for prediction with logistic regression we applied a simple strategy. We computed fisher scores between popularity and the each of the possible explanatory variables and ranked the variables based on this score. Then starting from the simplest model we added variables by fisher score ranking one-by one. Finally we selected the best performing model based in prediction accuracy based on Bayesian information criterion. This method can be considered as a version of stepwise regression where the models considered are based on a specific variable ranking. The best model obtaine this way had an accuracy arond 49%.

We also tried to fit an ordered logit model to take into account the fact that outcome measure is not a simple categorical but an ordinal variable. Using ordered logit yielded very similar result as the multinomial logit. The best achieved accuracy was around 49%.



Ensemble Testing

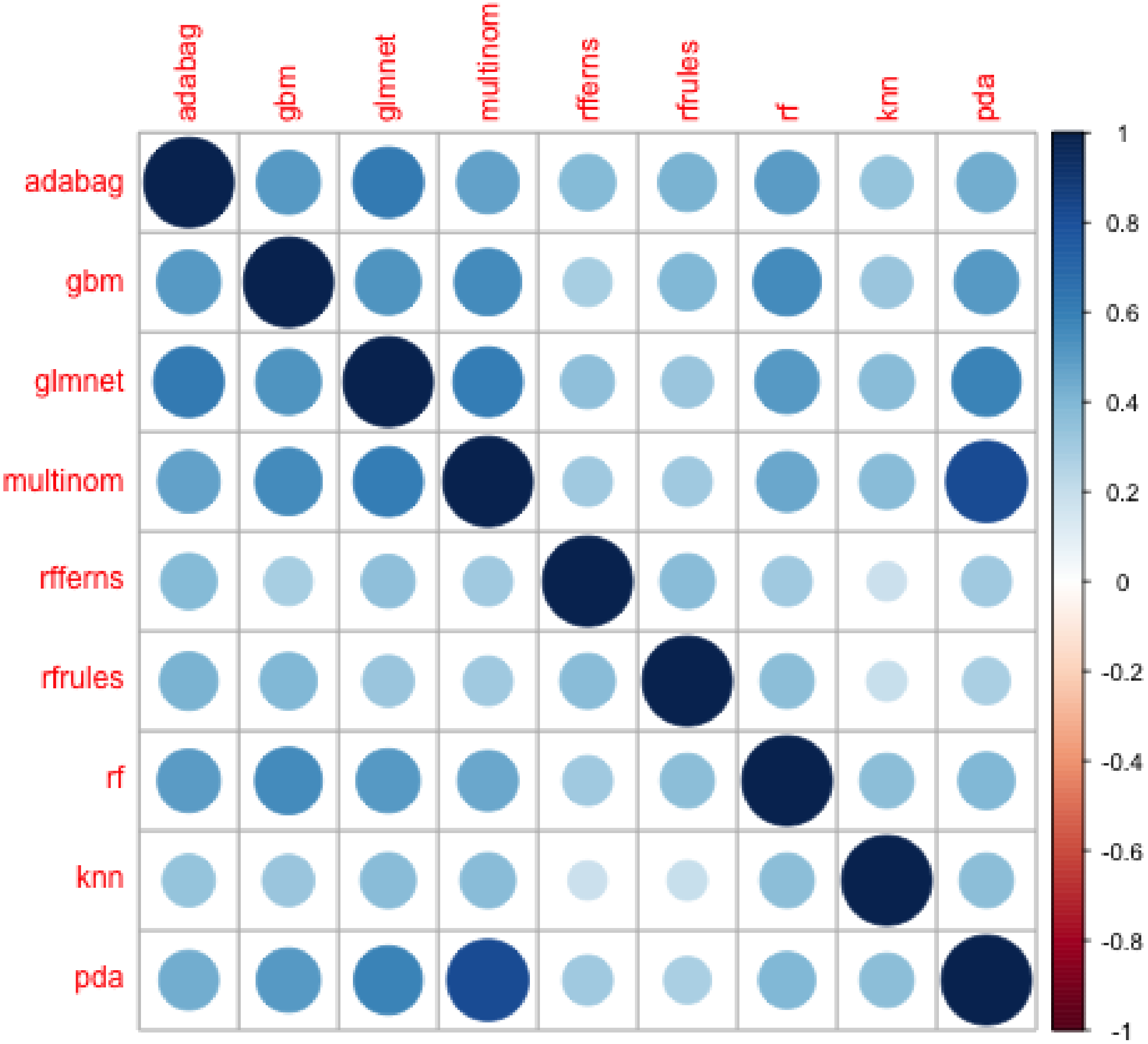
The concept of ensembling has become popular from the intuitive idea that more heads are better than one. It has been shown that making predicions from a collection of models can have greater predictive power than one very good model.

The process to create an ensemble was as follows:

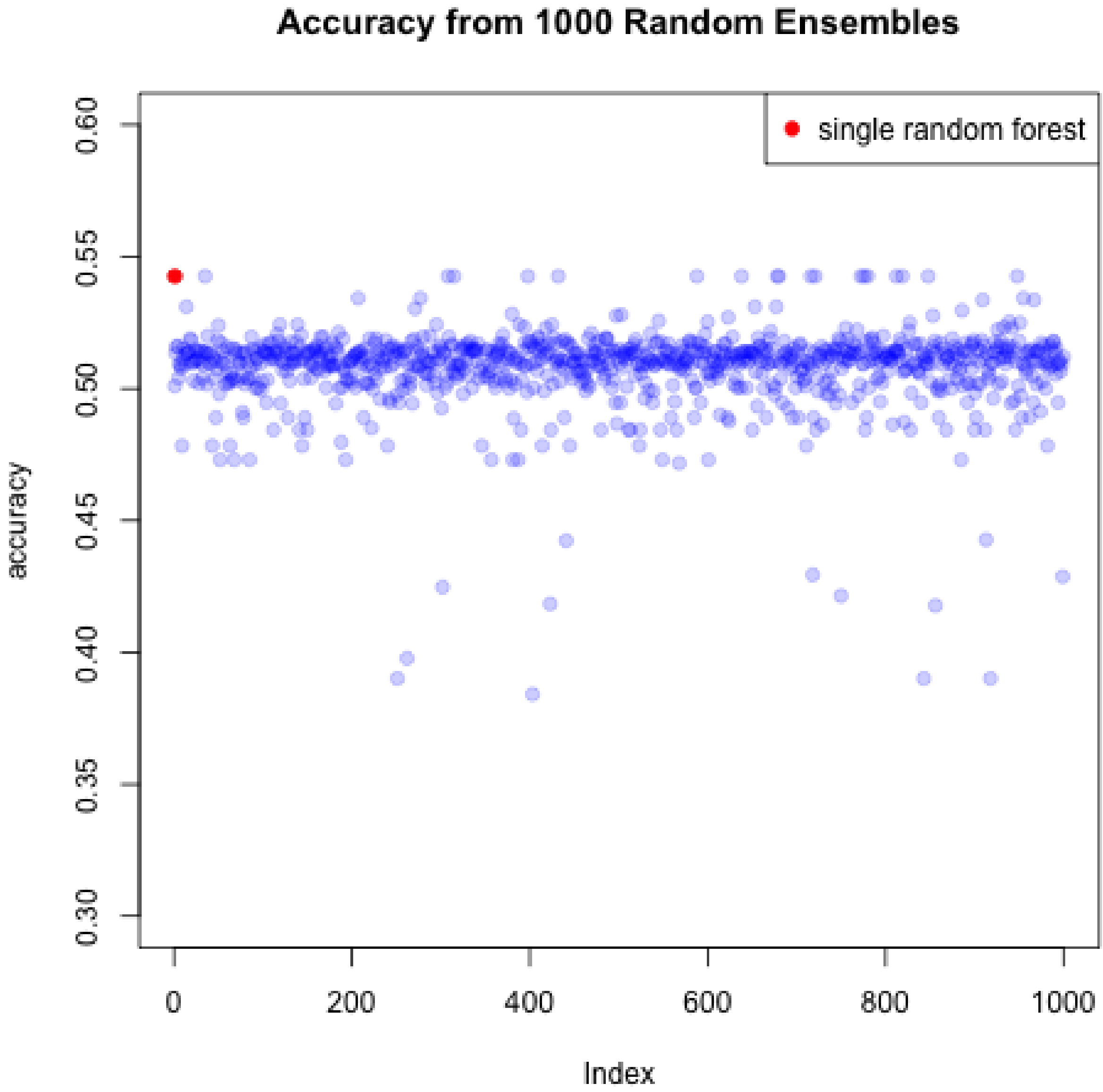
1. Train a collection of models, using caret for tuning, on a smallish data set to iterate faster, on the same training set
2. Evaluate the accuracy of each model on the same test (e.g. validation) data set
3. Evaluate the accuracy of an ensemble of models having low correlation in predictions

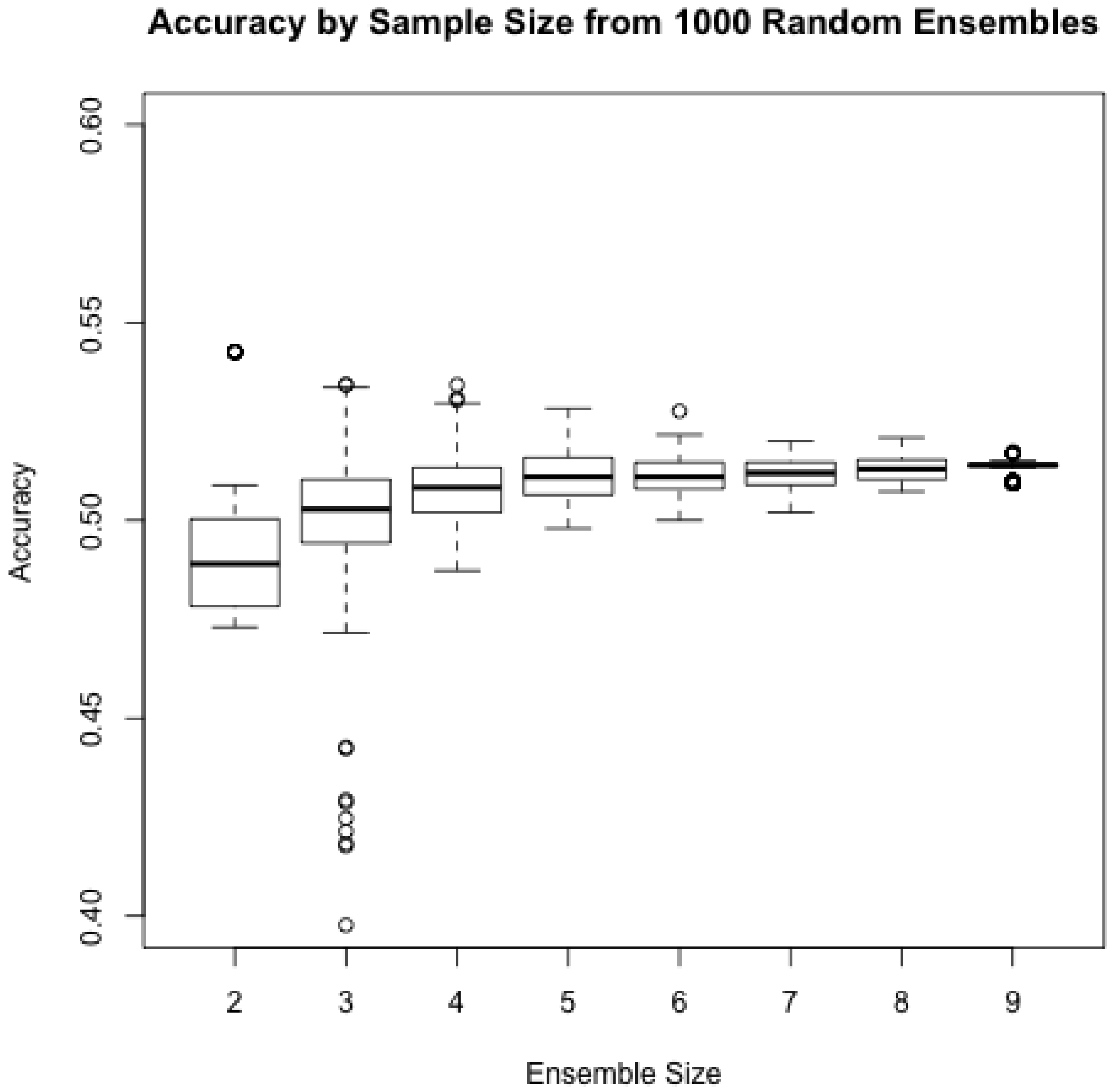
The results of this experiment on the Mashable data set are that while it improved predictive power of nearly all models alone, it did not improve the predictive power of randomForest alone.

Below is a plot of the correlations:



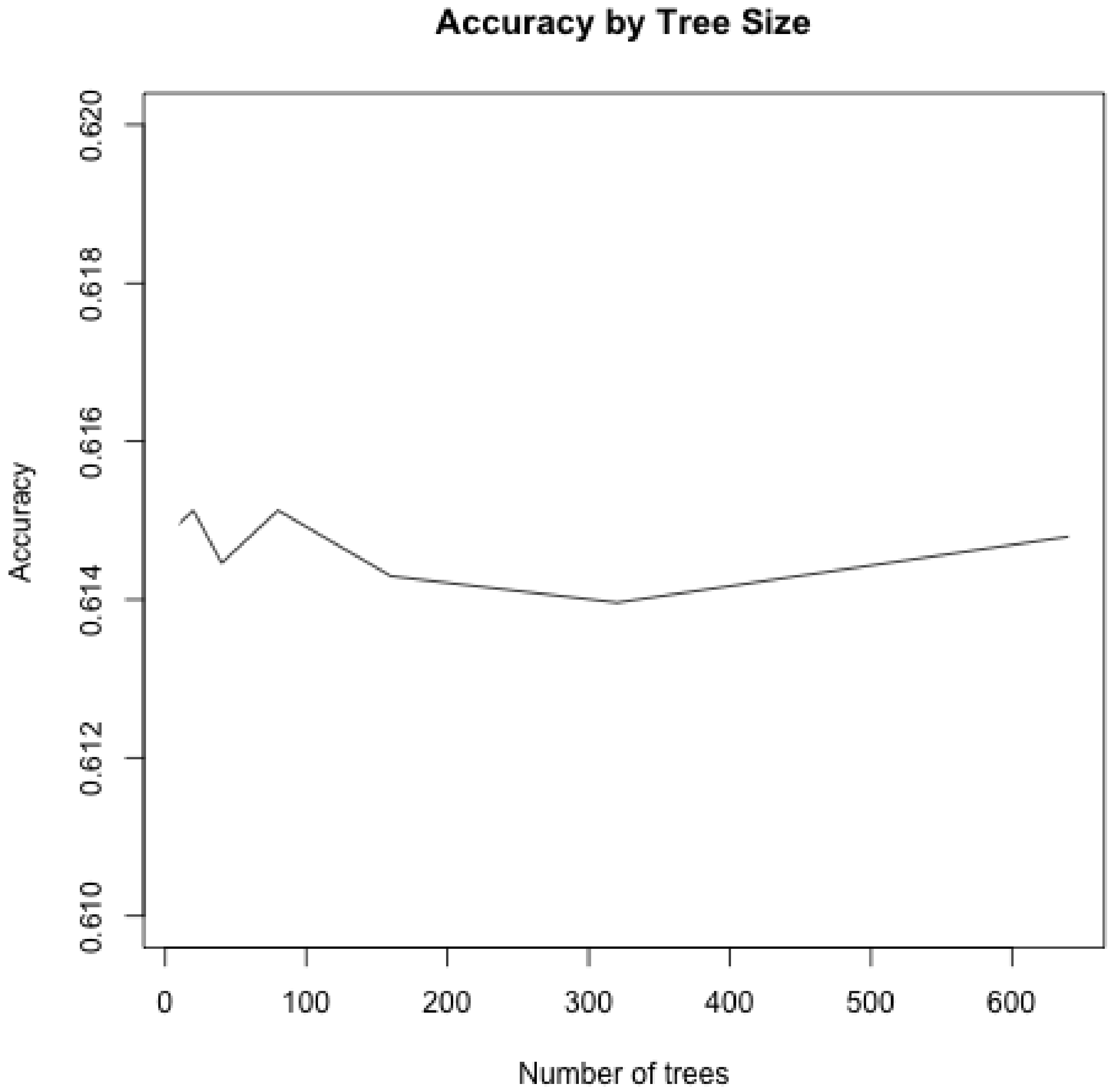
Ensembling predictions from the least correlated models resulted in lower predictive power. In order to affirm this result, we ran 1000 simulations with different random sizes and selections of the model predictions available.

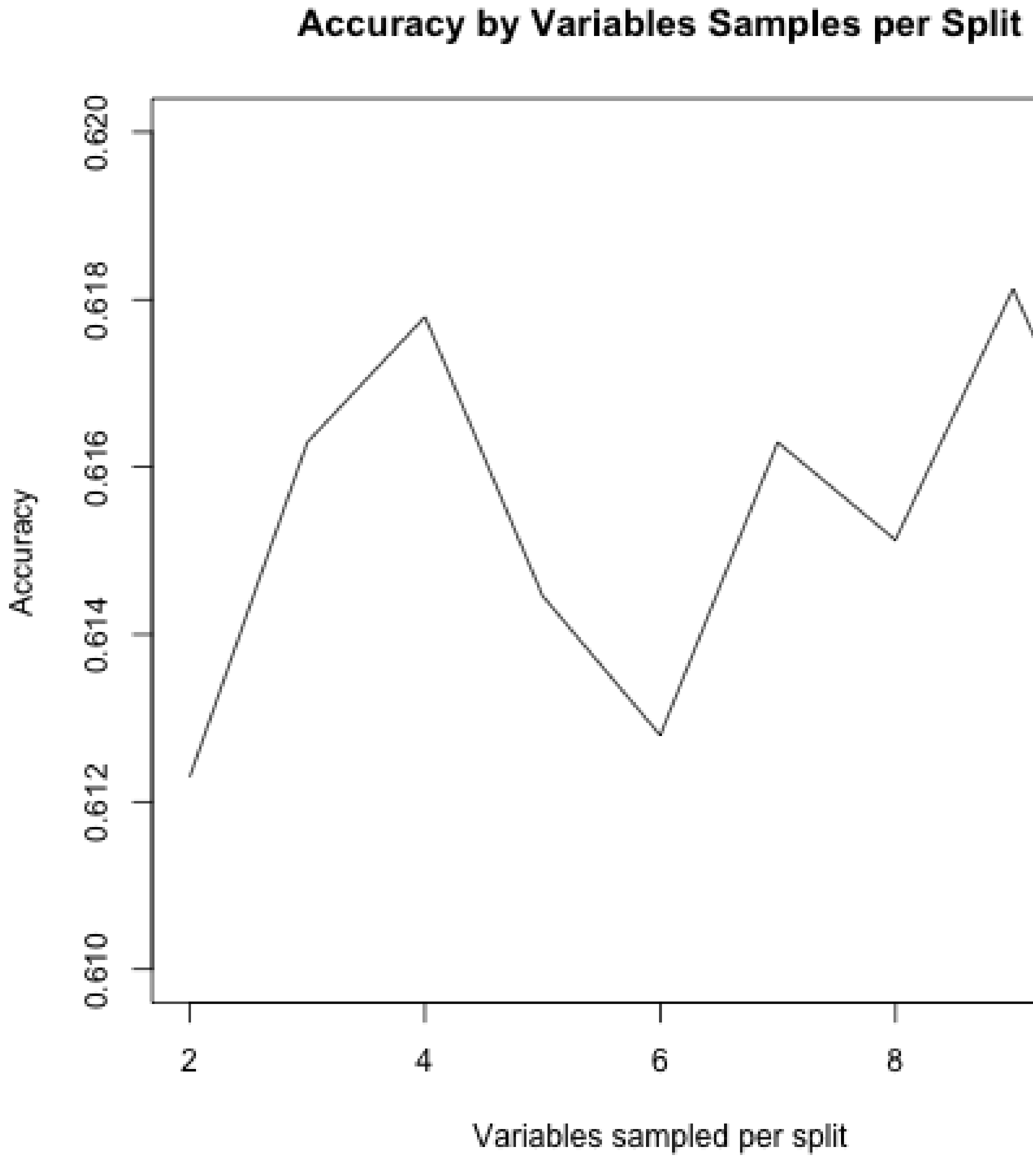

Accuracy from 1000 Random Ensembles
● single random forest
accuracy
0.30
0.35
0.40
0.45
0.50
0.55
0.60
Index
0
200
400
600
800
1000



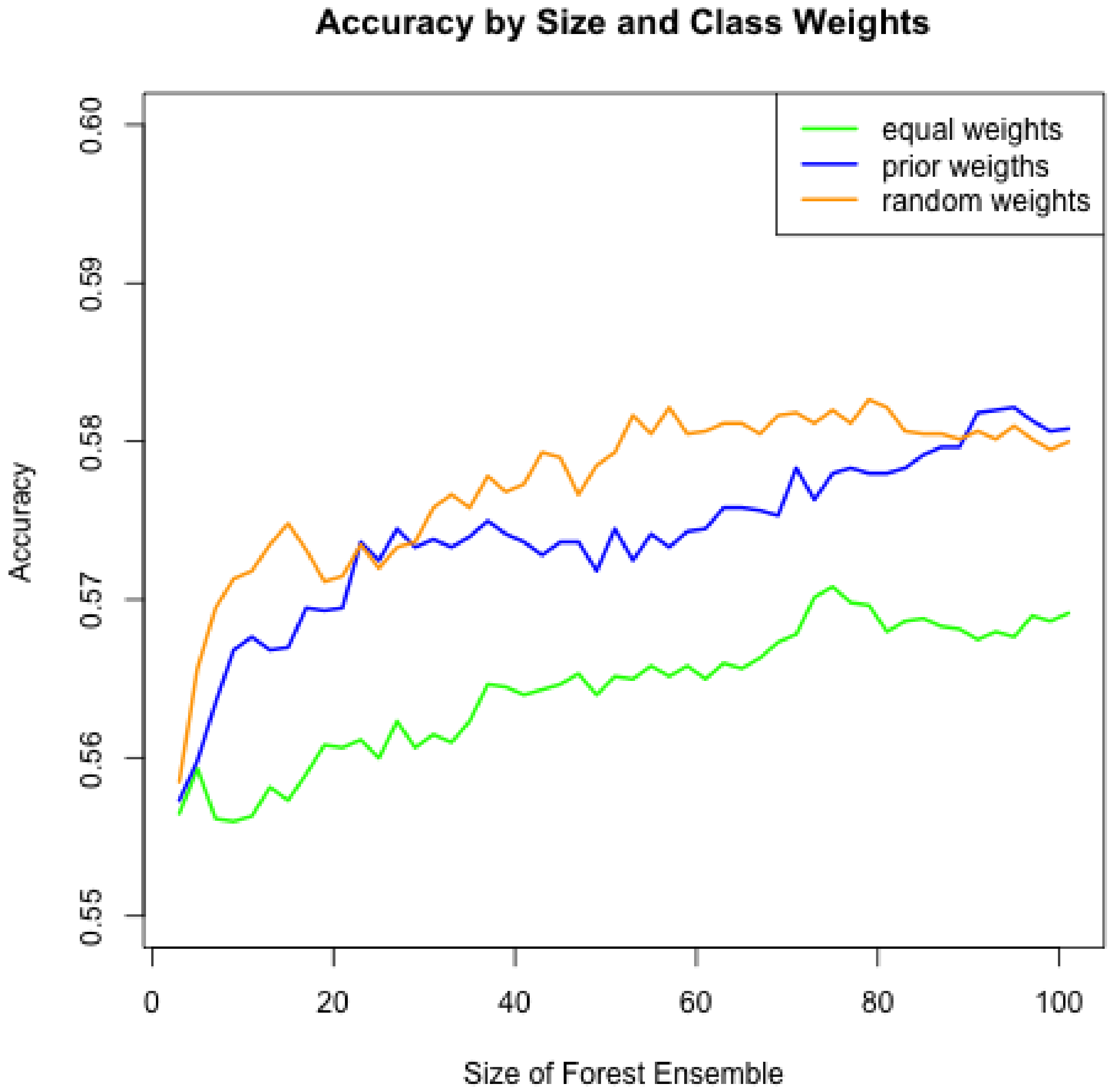
Ensembling Random Forests

Since ensembling different models was unsuccessful, we doubled down on random forest. Most parameters were tested for optimal tuning and it was found the default tuning (e.g. nodesize, mtry, ntrees) were hard to improve upon.


Accuracy by Tree Size
Accuracy
0.610
0.612
0.614
0.616
0.618
0.620
0
100
200
300
400
500
600
Number of trees



I was particularly interested in the effect of tuning classwt parameter: Could it be ensembled randomly to produce greater predictive power?


Accuracy by Size and Class Weights
equal weights
prior weigths
random weights
Accuracy
0.55
0.56
0.57
0.58
0.59
0.60
0
20
40
60
80
100
Size of Forest Ensemble