HW 11: Code Competition Austin Funcheon & Viraj Rane

* Looking at the dataset, it varies from 0 to 1, for most of the predictors. As such, there is no scaling to be applied to the data.
* With 710 variables, number of observations are fairly close to balanced data, which poses a challenge for many methods. Trimming proved mildly effective, but overall didn’t improve AUC score all that much, with a modest 5% of predictors trimmed.
* Also, reviewing the data, it is highly unbalanced.
  + 0:452 1:3548
  + As such, dataset for training set will need to be balanced to generate an appropriate model, in this case, we have selected undersampling down to the lower class.
* Given the type of data, classification, and the type of variables, a word presence count, it conceptually identified several methods to explore:
  + Naïve Bayes
  + Random Forest
  + ANN
  + SVM RBF

Reviewing methods, RF seemed to take an early lead with the methods, with a 57.32% AUC. By trimming the bottom 5% of predictors based upon the RF MeanDecreaseAccuracy, accuracy was improved with the test data set.

For Final Validation set,

Summary:

|  |  |
| --- | --- |
| Method | AUC Result |
| NB | 52.4% |
| RF | 57.32% |
| SVM RBF | 44.6% (error) |
| ANN | 51.4% |
| Final RF Trimmed | 57.76% |