MATH 373: Intro to Machine Learning
Case Study 2: Tree Based Methods

by James D. Wilson (University of San Francisco)

Directions: Provide all `R` code and solutions by *knitting* your final `R`Studio file into a single file. You can work in groups of up to 4, and only one person needs to upload the final .pdf to Canvas. *You must work in groups of at least size 2*! Be sure to put all of your group members' names at the top of the document.

1. We will revisit the *Mashable* dataset that you worked on in Homework 3.

   (a) Build classifiers for this data set using the tree-based methods that we've learned in class. In particular, build classifiers on the training data and assess the performance of each of the following methods on the test data (available on Canvas):

      i. Classification tree
      ii. Bagging
      iii. Random Forests
      iv. Boosting

      For each method, consider use cross-validation and grid searches to identify the best tuning parameters (like the depth of the tree(s), the weight for boosting, the number of variables to use in random forests, etc.). The goal here is to build the best classifier that you can for predicting the popularity of a new website.

   (b) What is the MSPE for each of your fitted models? Compare and contrast between models and then compare and contrast these ensemble-based models with the classification models that you fit in Homework 3. What are the advantages and disadvantages of using these ensemble methods?

2. **Bonus**: for the above spam detection task, one can iterate between the creation of features, subsetting of features, and running the classifier. At the end of the day, you want the **best** classification rule to detect spam. The team that comes up with the features that result in the lowest misclassification rate on the full training data will receive +15 points to be spread across and added to the assignments given so far in class. Good luck!