Lesson 13: Cheating Works Even on Randomized Data

We can push the example from our previous lesson to the extreme. We will randomize the classification data. That is, we will take the column with the class values and randomly permute it. We will use the Randomize widget to do this.

Later, we will do classification on this data set. We expect really low classification accuracy on randomized data set. Then, we will select five features that are most associated with the class. Even though we have randomly permuted the classes, there have to be some features that are weakly correlated with the class. Simply because we have tens of thousands of features, and we have only a few samples. There are enough features that some of them correlate with class simply by chance. Finally, we will score a random forest on a randomized data set with selected features.

Compare the scores reported by cross-validation on different data sets in this pipeline. Why is the accuracy in the final one rather high? Would adding more "most informative features" improve or degrade the cross-validated performance on a randomized data set?

Instead of selecting five most informative features, you can reduce this number even further. Say, to two most informative features. What happens? Why does accuracy raise after this change?