Random Forest Quiz

NAMES:

1. **Enough Data?**

   As an analyst at a car company, you are attempting to build a machine learning model that will determine whether or not a driver is too impaired to drive. Your training set contains 20 attributes describing driver state and behavior (swerviness, blink rate etc.) Along with a class label describing whether or not a collision occurred during the trip. Your training set consists of 2000 collisions and 2000 non-collisions. After training a random forest, your test set accuracy is 89%. Your supervisor asks whether you could get better results with a larger training set. How might determine the answer to her question using only the labeled data that you currently have?

2. **Random Forests**

   (a) Bagging is an ensemble method that involves training multiple base classifiers on different subsets of the original training data, then allowing those classifiers to vote. How is the Random Forest algorithm different from just applying bagging to our standard decision tree algorithm? Based on your answer, do you think that the Random Forest improvement is likely be helpful for one-dimensional training data?

   (b) Random forests generally perform better than individual decision trees. Can you imagine a situation where a simple decision tree may be preferable to a random forest?

   (c) Our textbook suggests that we can avoid creating an explicit validation set by using out-of-bag samples to evaluate a random forest classifier (or any classifier constructed using bagging). Some online sources suggest that this approach might slightly underestimate the generalization performance of our classifier. Why would this be the case?

(d) Random forests are typically composed of fully-expanded decision trees. In contrast, when using a simple decision tree, we are likely to use prune the final tree or to carefully tune the tree depth as a hyperparameter. How do you explain the difference?