Random Forest Exercise 2

Part A: Understanding Variable Importance in Random Forest

1. CART uses a point scoring system to determine variable importance. If an X variable is chosen as the best split, points are credited to that X variable. If this split is higher up the tree, there are more points. If an X variable is chosen as a Surrogate, it earns points too. Since Random Forest is an ensemble of 500 CARTs, can we use the same idea to get variable importance in Random Forest? Explain.

2. Produce and explain the Variable Importance chart from the Heart.csv analysis in the provided Rscript.

Part B: Understanding Missing Values in Random Forest

3. A quick way to handle missing values in Random Forest is to overwrite the parameter na.action to be either na.omit or na.roughfix. Can the Surrogate idea in CART be used to handle missing values in Random Forest? Explain.

4. There is another way to handle missing value in Random Forest. The rfImpute() function. Read the documentation for this function within RStudio or online at http://math.furman.edu/~dcs/courses/math47/R/library/randomForest/html/rfImpute.html. Explain how this works and why this may be better than na.roughfix.

Part C: Random Forest for Continuous Y.

5. Use Random Forest to predict resale flat price [Dataset: resale-flat-prices-2019.csv]. Modify the codes in flatprice-RF.R to add Random Forest and compare against MARS.
   a. Set.seed(2) and do a 70-30 train-test split.
   b. What is the MARS (degree 2) testset error?
   c. What is the Random Forest OOB error and testset error?
   d. Which model performed the better?
   e. Which variables are important?