STT 811

In-Class Assignment 14

This problem will use the Weekly dataset, with Direction as the target.

1. Take a look at the data.
   ```
   weekly <- ISLR2::Weekly
   View(weekly)
   ```
2. Split the data into training and test datasets (with a 75/25 split).
   ```
   split_pct <- 0.75
   n <- length(weekly$Direction)*split_pct # train size
   row_samp <- sample(1:length(weekly$Direction), n, replace = FALSE)
   train <- weekly[row_samp,]
   test <- weekly[-row_samp,]
   ```
3. Build a decision tree model based on all the X's. Calculate the accuracy from the confusion matrix.
   ```
   weekly_tree <- tree(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, data = train)
   tree_pred <- predict(weekly_tree, test, type = 'class')
   confusionMatrix(tree_pred, test$Direction)
   ```

4. Now do a bagged decision tree (500 trees). Create the confusion matrix and give the accuracy. How does the accuracy compare to (3)?
   ```
   predicts <- matrix(nrow = length(test$Direction), ncol = 0)
   for(i in 1:500){
     rows <- sample(1:length(train$Direction), length(train$Direction), replace = TRUE)
     samp <- weekly[rows,]
     weekly_trees <- tree(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, data = samp)
     predicts = cbind(predicts, predict(weekly_trees, test)[,1])
   }
   ens <- rowMeans(predicts)

   confusionMatrix(as.factor(ifelse(ens < 0.5, 'Down', 'Up')), reference = test$Direction)
   ```

5. Try a random forest model (500 trees). Use 3 features at a time and set maxnodes = 4. Calculate the confusion matrix and compare the accuracies. Also look at the relative variable importance.
   ```
   weekly_rf <- randomForest(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, data = train, mtry = 3, importance = TRUE, ntree = 500, maxnodes = 4)
   rf_predict <- predict(weekly_rf, test)
   confusionMatrix(rf_predict, test$Direction)
   barplot(weekly_rf$importance[,3])
   ```