Due: Monday Dec 11 at 11:59pm. Hard deadline! Solutions to be posted shortly after deadline.

Instructions: Upload the well organized output (html or pdf) from an Rmarkdown (or equivalent) document to Canvas.

1. On Canvas in the assignment area, there is a data set named `datasalaries.csv` that contains information on data science and STEM salaries — this is a cleaned up version of the data from `https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries`. Since the `tree` function will skip character vectors, when you use 'read.csv' (or an equivalent) please ensure that you set 'stringsAsFactors=TRUE', or otherwise coerce your character string predictors to factors after the fact.

   (a) Create a regression tree for a data science/STEM salaries given the remainder of the variables in the data set. Provide the tree, including labels — using the command `text(treename, pretty=0)` will provide a (somewhat) more understandable split labelling for the questions that follow.

   (b) Based **only** on the tree outputted in the previous question, which companies included in this data set would you prefer to work for? Why?

   (c) Using `set.seed(51341)`, perform 10-fold cross-validation using `cv.tree`. Plot the resulting object. How many terminal nodes does cross-validation suggest?

   (d) Prune your original tree. Give the predicted salary for a self-identified Asian female with a PhD working at Google with 10 years of experience and 10 years at that company. Use the `predict()` function to do this, but PLEASE double-check with your tree diagram and brain. My warning is to pay careful attention to how the character vectors are factored, and note that you will have to setup the entry as a 'data.frame'. You will likely find this finicky...but it is good practice for real life data science messiness.

   (e) Use the following commands to setup a training and testing set:

   ```
   set.seed(763)
   dsindex <- sample(1:nrow(datasalaries), 4000)
   dstrain <- datasalaries[dsindex, ]
   dstest <- datasalaries[-dsindex, ]
   ```

   Now fit a tree to the training set, prune via 10-fold CV, and once again give the predicted salary for the individual from part (d) via the predict function.

   (f) Provide the estimated MSE of the model in part (e) — that is, calculate the MSE of the test set. Is the MSE of the test set close to the expected MSE from the 10-fold CV from question (c)?

2. Here I'll run you through some code that could seem aggravating/confusing at first. Pause and consider what you're asking the computer to do for each of these estimates of the MSE. Which estimate is more believable as a long-run estimate of the MSE? Why?

   ```
   > library(randomForest)
   > library(gclus)
   > data(body)
   ```

```
> set.seed(02139)
> bodrun <- randomForest(Weight~Height+Gender, data=body)
> bodrun

Call:
 randomForest(formula = Weight ~ Height + Gender, data = body)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 82.15451
                    % Var explained: 53.78

> #FYI this matches the MSE from the printout
> bodrun$mse[500]

[1] 82.15451

> #FYI this also matches that printout
> sum((body$Weight-predict(bodrun))^2)/length(body$Weight)

[1] 82.15451

> #but this is different!
> sum((body$Weight-predict(bodrun, newdata=body))^2)/length(body$Weight)

[1] 74.8239
```

3. On Canvas (in the assignment area), you will find a data set (insurance.csv) on individuals' health insurance charges in the US along with some demographic information. Note that the data includes both categorical and numeric measures. Provide a thorough regression analysis attempting to predict the 'charges' variable using the remainder of the predictors in the data set. At minimum, trees, boosting, linear models, random forests, and lasso should be used...with appropriate diagnostics, sensible training/testing split, cross-validation, etc. Which model is most likely to provide the lowest MSE in the long-run? Which model would you choose if you were consulting with an insurance company on this data set? If they don't match, explain why.