* GBM
  + What is boosting -
  + What is the goal of boosting from a model bias / variance perspective
  + Describe some differences between random forest and GBM (parallel – sequential, RF reduces variance, GBM has reduces bias)
  + Describe some similarities between random forest and GBM (both are ensemble techiques, )
  + Describe some characteristics of individual trees in a GBM forest
  + Describe the GBM training process
  + What is one method of preventing correlation between GBM trees
  + How does the learning rate in GBM relate to the number of trees in the forest
  + How does tree depth relate to model accuracy in GBM
* Bootstrap Sampling / Bagging
  + Describe how Bootstrap sampling works. How is sampling performed. How many observations are contained in a bootstrap sample as compared to the original data set we are sampling.

100% of the observation

* + What is bagging and how does it relate to bootstrap sampling?

Collection of bootstrap Samples

* + Describe how bagging works in relation to random forest
  + Is it possible to have duplicate observations in a bootstrap sample

Yes, it’s possible to have

* + What is an out of bag sample

Samples that did not make it into the bootstrap sample

* + Approximately how many out of bag samples exist per bootstrap sample.

33% or 1/3 of population

* + Describe how out of bag samples are useful in training models

For testing

* Random Forest
  + What are some differences between random forest and regular decision trees
  + How is training data generated for each tree in the forest
  + What is unique about the way random forest performs splits that helps minimize correlation between trees.  
    Column Sampling
  + Provide a brief summary on how random forest prevents high correlation between trees in the forest.

Randomly sampling the columns – reduce high variance

* + Why would are trees in random forest tend to be correlated.

Bootstrap sampling, observations will be shared between trees

* + Why correlation between trees is bad?

High variance

* + How does the number of trees in the forest relate to model bias / variance

Low Variance

* + How does tree depth relate to model bias / variance

Low Bias, and High Variance

* + How is entropy and gini index used in the training process
  + What is information gain and how does it relate to relate to the training process
  + How is feature importance determined in a random forest model  
    Based on information gain

Number of columns -> square root of (num of features) for each tree in the forest.