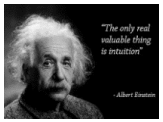
Bootstrap Aggregation (Bagging)



Bagging Models: Intuition

- Generally, **tree models** have **high variance** and cross-validation error, which **Bagging** aims to **reduce**
- In general, the **bootstrap** approach to reducing variance is based on a **simple concept**

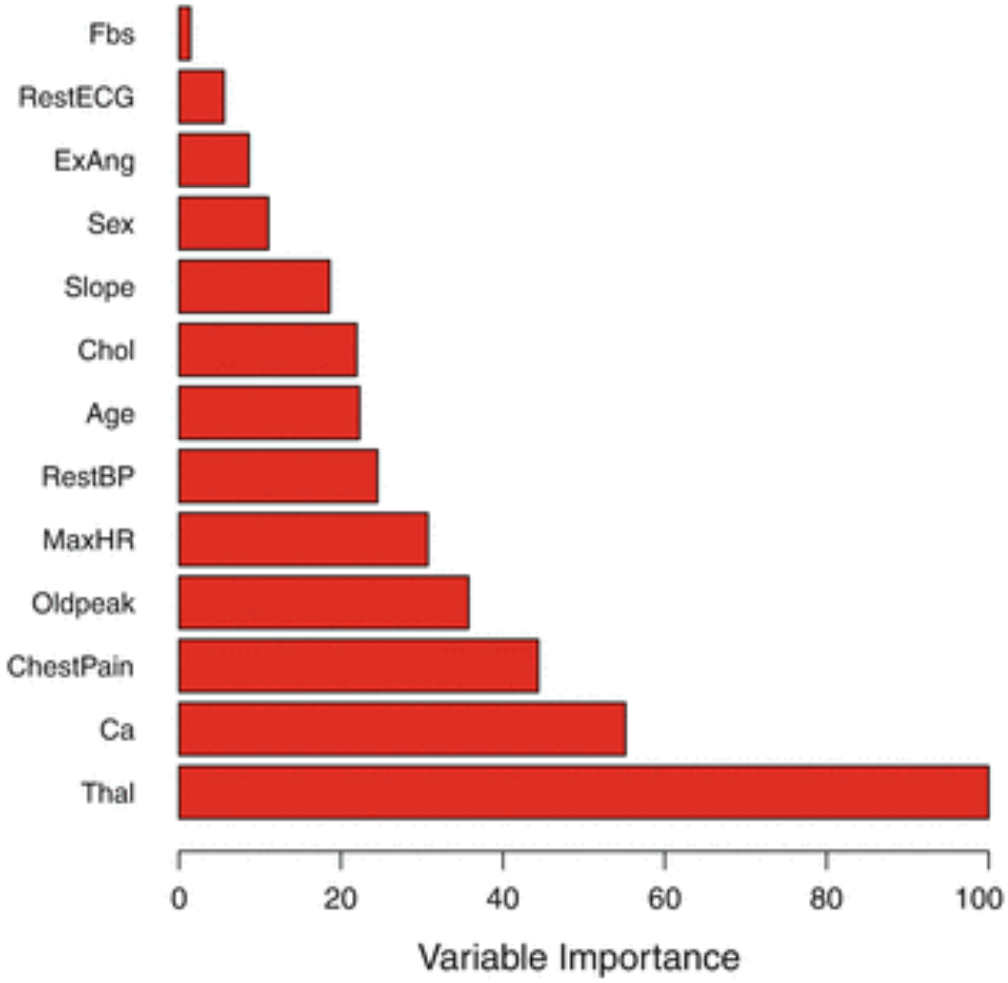
  → If you have a data set with ***n*** **observations** and high variance, you can produce a new data set with ***n*** **random samples** (with replacement) of the same data set and using the respective **averages**, rather than the **actual** observations.
- This applies to many statistical methods but it is particularly **useful** for **trees** because we can generate as **many samples** as we need
- This works because the **variance** of **sample means** is always **lower** than the **actual variance** of the raw data.
- **Bagging** applies this principle to decision trees by **fitting many training sets** from the same data set, building **separate** tree **models** for each, and averaging the resulting **predictions** across all models.
- Hence the name: **bootstrap + aggregation !!**

Out-of-Bag (OOB) Estimation

- Because bootstrap uses a sample from the data to do the estimation each time, the **remaining observations** not selected in the bootstrap sample can be use for **cross-validation** without further sampling.
- The observations not selected in the bootstrap sample are called **"out-of-bag"** observations, which can be used to compute the **MSE**, **SSE**, **classification error**, or other statistics.
- As the bootstrap **sample increases**, the OOB approach becomes equivalent with Leave-One-Out cross validation (**LOOCV**) discussed earlier.
- As with other tree models, the results are **difficult** to **interpret**, but this is more problematic with bagging because of the multiple samples and averaging results across models.
- However, one can display the **"importance"** of each variable by analyzing the **proportion** each **variable** contributes to reducing **MSE** or **error rate** across all samples.

Variable Importance Illustration

Source: Heart data in ISLR package – see textbook



R Tips

`randomForest()`{`randomForest`} → Function used to fit various random forest models (please note the cap F). Bootstrap Aggregation (Bagging) is a special case of Random Forest

```
bag.fit=randomForest(y~x1+x2+etc.,data=dataName,
                     mtry=13,importance=TRUE)
```
→ `mtry=13` tells Random Forest to use 13 predictors; if p is the full set of predictors in the model, then `mtry=p` generates a Bagging model; by default this method computes 500 trees.

```
bag.fit.25=randomForest(y~x1+x2+etc.,data=dataName,
                        mtry=13,importance=TRUE,
                        ntree=25)
```
→ Use the `ntree=25` attribute to change the default number of trees to generate (to 25, for example)

KOGOD SCHOOL
of
BUSINESS