

# SurvMeth 640 Assignment 4

Cheng, Chia Wen

2023-03-26

## Setup

```
library(mlbench)
library(foreach)
library(caret)
library(rpart)
library(partykit)
library(xgboost)
```

## Data

In this notebook, we use the Boston Housing data set (again). “This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. It was obtained from the StatLib archive (<http://lib.stat.cmu.edu/datasets/boston>), and has been used extensively throughout the literature to benchmark algorithms.”

Source: <https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html>

```
data(BostonHousing2)
names(BostonHousing2)

## [1] "town"    "tract"   "lon"     "lat"     "medv"    "cmedv"   "crim"
## [8] "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"
## [15] "rad"     "tax"     "ptratio" "b"        "lstat"
```

First, we drop some variables that we will not use in the next sections.

```
BostonHousing2$town <- NULL
BostonHousing2$tract <- NULL
BostonHousing2$cmedv <- NULL
```

Next, we start by splitting the data into a train and test set.

```
set.seed(1293)
train <- sample(1:nrow(BostonHousing2), 0.8*nrow(BostonHousing2))
boston_train <- BostonHousing2[train,]
boston_test <- BostonHousing2[-train,]
```

## 1) Bagging with Trees

a) Build a Bagging model using a foreach loop. Use the maxdepth control option to grow very small trees. These don't have to be stumps, but should not be much larger than a few splits.

```
# In order to get a smaller tree, I set higher thresholds for splitting a node and
# adjust tree depth to be pretty small.
y_tbag_small <- foreach(m = 1:100, .combine = cbind) %do% {
  rows <- sample(nrow(boston_train), replace = T)
  fit <- rpart(medv ~ .,
    data = boston_train[rows, ],
    method = "anova",
    control = rpart.control(cp = 0.01,
      # because I'm not building a really deep tree, I use the default cp value
      maxdepth = 3,
      # max tree depth is set to be 3 for a small tree
      minsplit = 20,
      # minimal obs in a node
      mincriterion = 0.999))
  # 1-p threshold for splitting
  predict(fit, newdata = boston_test)
}
```

##The option minbucket provides the smallest number of observations that are allowed in a terminal node. If a split decision breaks up the data into a node with less than the minbucket, it won't accept it.

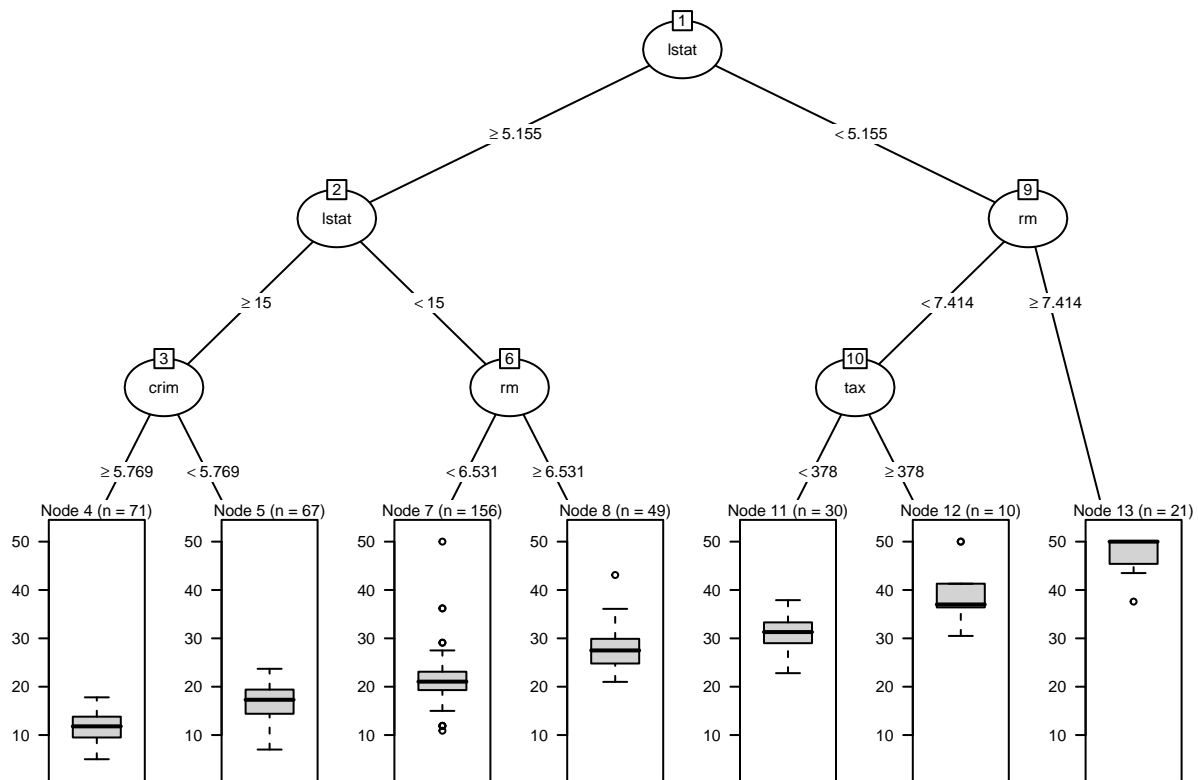
##The minsplit parameter is the smallest number of observations in the parent node that could be split further. The default is 20. If you have less than 20 records in a parent node, it is labeled as a terminal node.

##Finally, the maxdepth parameter prevents the tree from growing past a certain depth / height. In the example code, I arbitrarily set it to 5. The default is 30 (and anything beyond that, per the help docs, may cause bad results on 32 bit machines).

##The cp value is a stopping parameter. It helps speed up the search for splits because it can identify splits that don't meet this criteria and prune them before going too far. If you take the approach of building really deep trees, the default value of 0.01 might be too restrictive.

b) Plot the last tree of the ensemble to check tree size.

```
party_tree_last <- as.party(fit)
plot(party_tree_last, gp = gpar(fontsize = 6))
```



c) Compare the performance of the last tree in the bagging process with the ensemble. That is, look at the performance of the last tree in the loop and compare it with the performance in the overall averaged bagging model.

```
postResample(y_tbag_small[, 100], boston_test$medv)
```

```
##      RMSE  Rsquared      MAE
## 4.0604494 0.7793784 3.0004123
```

```
##      RMSE  Rsquared      MAE
## 4.0604494 0.7793784 3.0004123
postResample(rowMeans(y_tbag_small), boston_test$medv)
```

```
##      RMSE  Rsquared      MAE
## 3.8048378 0.7976314 2.5544164
```

```
##      RMSE  Rsquared      MAE
## 3.8048378 0.7976314 2.5544164
```

The overall averaged bagging model performs better than the last, individual tree with a smaller RMSE and a greater R-squared.

## 2) Bagging with Bigger Trees

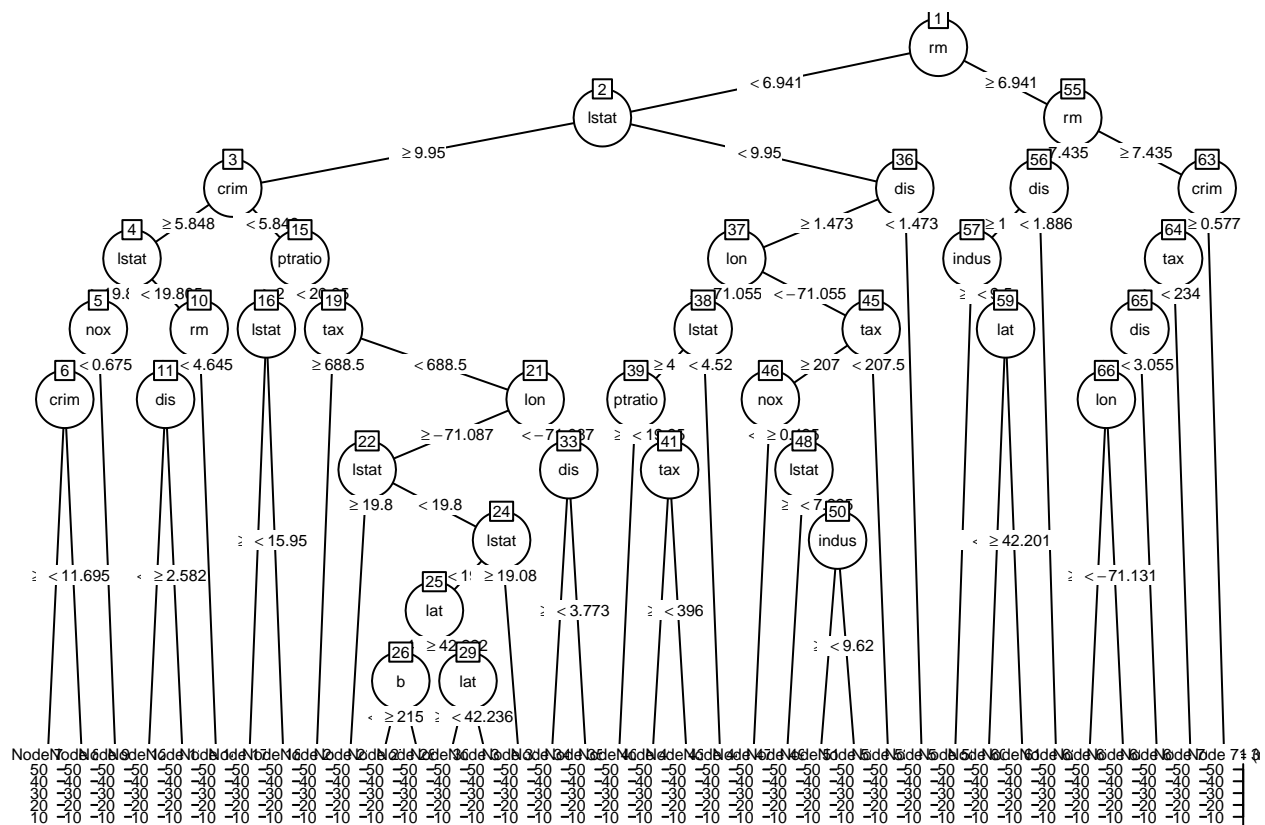
a) In the first loop we've grown small trees. Now, build a new loop and adjust maxdepth such that very large trees are grown as individual pieces of the Bagging model.

```
# In order to get a smaller tree, I set higher thresholds for splitting a node and
# adjust tree depth to be pretty small.
y_tbag_large <- foreach(m = 1:100, .combine = cbind) %do% {
  rows <- sample(nrow(boston_train), replace = T)
  fit <- rpart(medv ~ .,
    data = boston_train[rows, ],
    method = "anova",
    control = rpart.control(cp = 0.001,
      maxdepth = 30, # max tree depth
      minsplit = 10, # minimal obs in a node
      minbucket = 1)) # minimal obs in any terminal node

  predict(fit, newdata = boston_test)
}
```

b) Confirm that these trees are larger by plotting the last tree.

```
party_tree_last.1 <- as.party(fit)
plot(party_tree_last.1, gp = gpar(fontsize = 6))
```



c) Show how this ensemble model performs.

```
# I look at the performance of the last tree in the loop and compare it with the performance in the
# overall averaged bagging model.
postResample(y_tbag_large[, 100], boston_test$medv)
```

```
##      RMSE  Rsquared      MAE
## 4.9430781 0.6939056 2.8733794
```

```
## RMSE Rsquared MAE 4.9430781 0.6939056 2.8733794
postResample(rowMeans(y_tbag_large), boston_test$medv)
```

```
##      RMSE  Rsquared      MAE
## 3.2541248 0.8540099 2.2197547
```

```
## RMSE Rsquared MAE 3.2541248 0.8540099 2.2197547
```

d) In summary, which setting of `maxdepth` did you expect to work better? Why?

I expect the ensemble model along with the greater ‘maxdepth,’ which is 30 in my case, to work the best among models I’ve built above because the R-squared is the highest and the RMSE is the lowest. Across the smaller trees, which ‘maxdepth’ is 3 in my case, the ensemble model also works better than any individual models.

### 3) Building a Boosting Model with XGBoost

a) Now let’s try using a boosting model using trees as the base learner. Here, we will use the XGBoost model. First, set up the `trainControl` parameters.

```
ctrl <- trainControl(method = "cv",
                     number = 5,
                     summaryFunction = defaultSummary,
                     verboseIter = TRUE,
                     classProbs = TRUE)
```

b) Next, set up the tuning parameters by creating a grid of parameters to try.

```
grid <- expand.grid(max_depth = c(1, 3, 5),
                  nrounds = c(500, 1000, 1500),
                  eta = c(0.05, 0.01, 0.005),
                  min_child_weight = 10,
                  subsample = 0.7,
                  gamma = 0,
                  colsample_bytree = 1)
```

c) Using CV to tune, fit an XGBoost model.

```
set.seed(1293)
xgb <- train(medv ~ .,
            data = boston_train,
            method = "xgbTree",
```







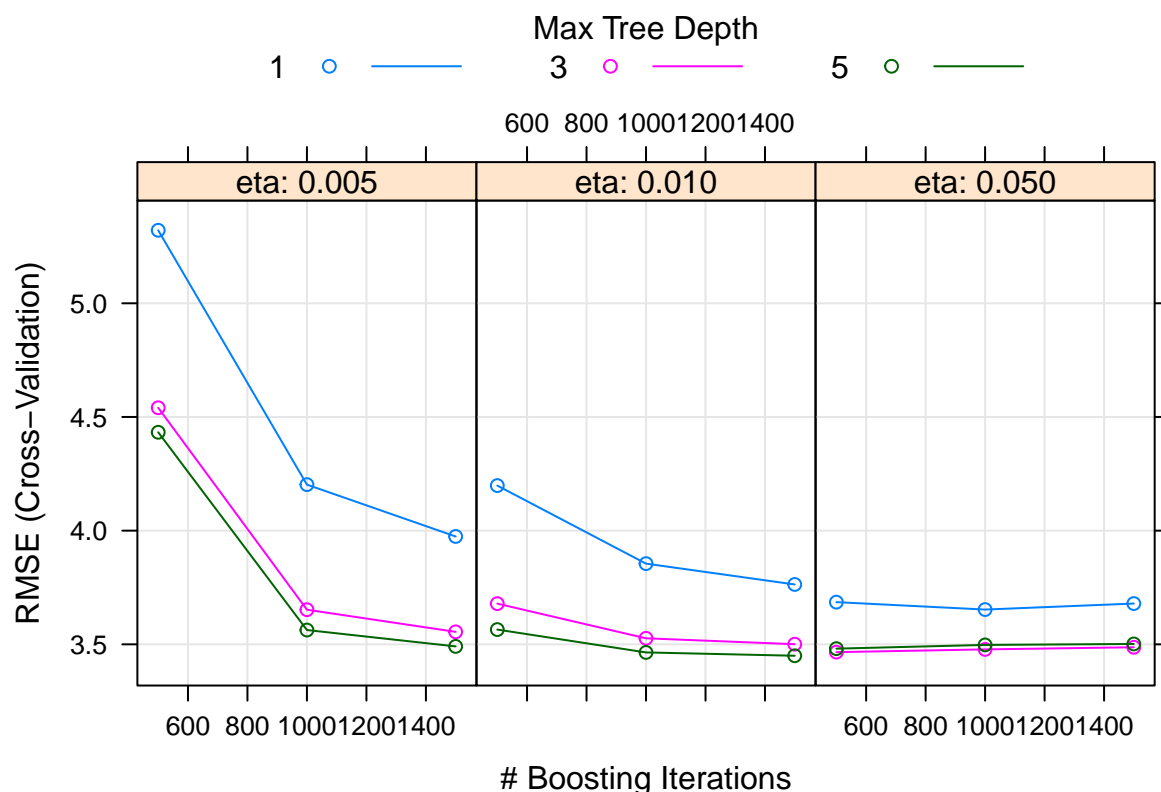


```

## + Fold5: eta=0.010, max_depth=1, gamma=0, colsample_bytree=1, min_child_weight=10, subsample=0.7, nr
## [14:01:53] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instea
## [14:01:53] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instea
## - Fold5: eta=0.010, max_depth=1, gamma=0, colsample_bytree=1, min_child_weight=10, subsample=0.7, nr
## + Fold5: eta=0.010, max_depth=3, gamma=0, colsample_bytree=1, min_child_weight=10, subsample=0.7, nr
## [14:01:54] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instea
## [14:01:54] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instea
## - Fold5: eta=0.010, max_depth=3, gamma=0, colsample_bytree=1, min_child_weight=10, subsample=0.7, nr
## + Fold5: eta=0.010, max_depth=5, gamma=0, colsample_bytree=1, min_child_weight=10, subsample=0.7, nr
## [14:01:56] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instea
## [14:01:56] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instea
## - Fold5: eta=0.010, max_depth=5, gamma=0, colsample_bytree=1, min_child_weight=10, subsample=0.7, nr
## + Fold5: eta=0.050, max_depth=1, gamma=0, colsample_bytree=1, min_child_weight=10, subsample=0.7, nr
## [14:01:57] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instea
## [14:01:57] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instea
## - Fold5: eta=0.050, max_depth=1, gamma=0, colsample_bytree=1, min_child_weight=10, subsample=0.7, nr
## + Fold5: eta=0.050, max_depth=3, gamma=0, colsample_bytree=1, min_child_weight=10, subsample=0.7, nr
## [14:01:58] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instea
## [14:01:58] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instea
## - Fold5: eta=0.050, max_depth=3, gamma=0, colsample_bytree=1, min_child_weight=10, subsample=0.7, nr
## + Fold5: eta=0.050, max_depth=5, gamma=0, colsample_bytree=1, min_child_weight=10, subsample=0.7, nr
## [14:02:01] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instea
## [14:02:01] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instea
## - Fold5: eta=0.050, max_depth=5, gamma=0, colsample_bytree=1, min_child_weight=10, subsample=0.7, nr
## Aggregating results
## Selecting tuning parameters
## Fitting nrounds = 1500, max_depth = 5, eta = 0.01, gamma = 0, colsample_bytree = 1, min_child_weight

plot(xgb)

```



# ROC is for categorical variables

d) Compare the performance of the boosting model with the models run previously in this assignment. How does it compare?

Using the XGBoost model across this sample, the RMSE, which is between 3.4 and 3.5, will be the smallest when the maximum tree depth is 5, eta is 0.01, and when the number of boosting iterations is more than 1400. The second smallest RMSE, which is also between 3.4 and 3.5, takes place when the maximum tree depth is 5, eta is 0.01, and when the number of boosting iterations is approximately 1000.

Compare to the previous models I've created, the trees that are built using the extreme gradient boosting are shallower than the bigger trees I built. However, the extreme gradient boosting has significantly more trees than the previous models. In addition, with 3 as the tree depth, XGBoost usually performs better than the ensemble model run previously in this assignment. Exceptions happen when eta is 0.005 and the number of boosting iterations is smaller than 1000. Nonetheless, among all the models run in this assignment, deeper trees have more outstanding performance.

#### 4) Comparing Models with caretList

a) Use caretList to run a Bagging model, a Random Forest model, and an XGBoost model using the same CV splits with 5-fold CV. Plot the performance by RMSE. How do the models compare?



[illegible]





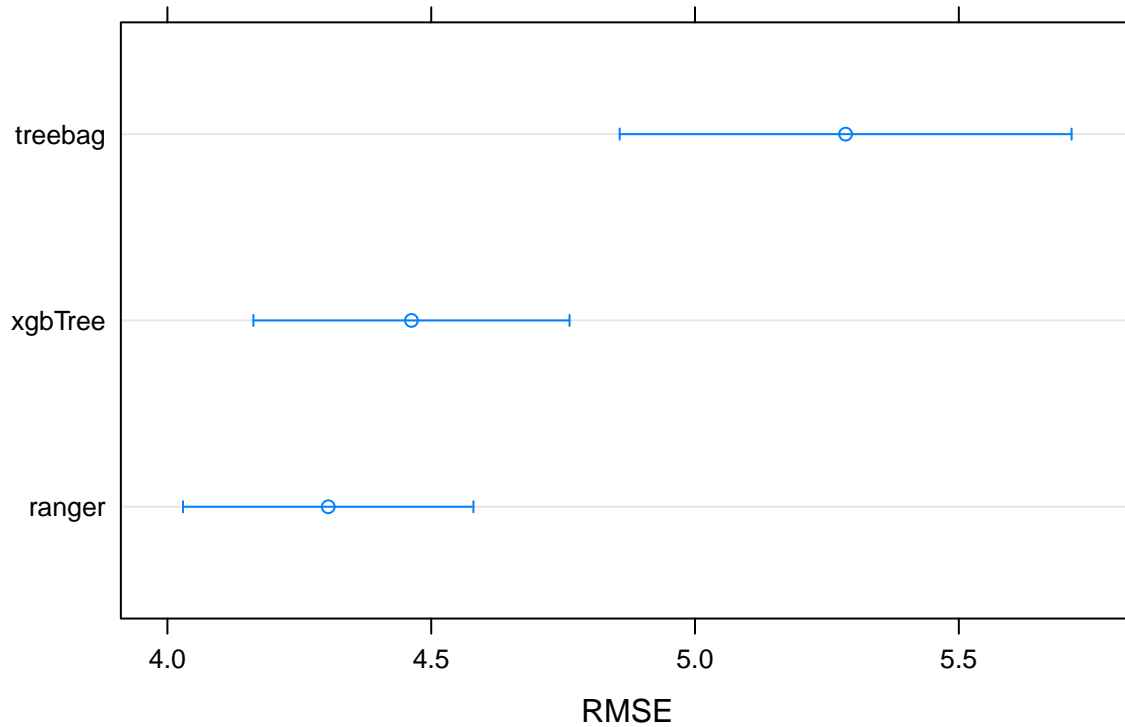


[illegible]



[illegible]

```
dotplot(resamples(model_list), metric = 'RMSE')
```



**Confidence Level: 0.95**

*Hint: You can use `treebag`, `ranger`, and `xgbTree` for the models.*

It turns out that the Random Forest model has the lowest RMSE, and the Bagging model has the greatest. To conclude, across this sample, the Random Forest model has a performance better than both the XGBoost model and the Bagging model, while the XGBoost model has a still better performance than the Bagging model.