P8106 Data ScienceII Midterm

Yueran Zhang (yz4188)

2023-03-26

1. In this exercise, we will build tree-based models using the College data (see "Col- lege.csv" in Homework 2). The response variable is the out-of-state tuition (Outstate). Partition the dataset into two parts: training data (80%) and test data (20%).

Package Prepare

```
library(tidyverse)
library(ISLR)
library(mlbench)
library(caret)
library(rpart)
library(rpart.plot)
library(party)
library(partykit)
library(pROC)
library(randomForest)
library(ranger)
library(gbm)
library(pdp)
```

Import Dataset

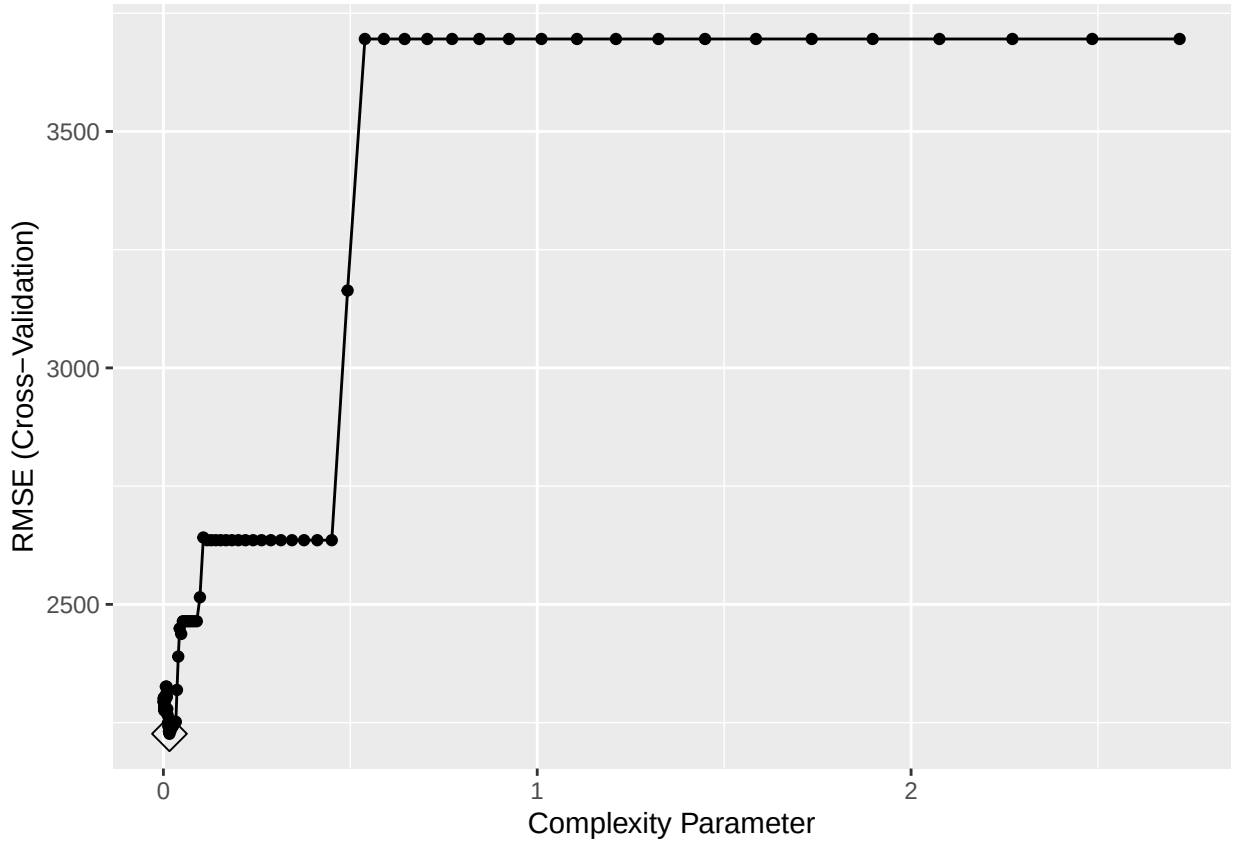
```
set.seed(1234)

# Load dataset + clean data
College = read.csv("/Users/yueranzhang/Desktop/DSII/HW4/DataSet/College.csv")[-1] %>%
janitor::clean_names() %>%
na.omit()

# Data Partition
RowTrain <- createDataPartition(y = College$outstate,
                                p = 0.8,
                                list = FALSE)

training.data <- College[RowTrain,]
```

```r
test.data <- College[-RowTrain,]

# training data
x <- model.matrix(outstate ~. , training.data) [,-1]
y <- training.data$outstate

# test data
x2 <- model.matrix(outstate ~. , test.data)[,-1]
y2 <- test.data$outstate
```

Question A

Build a regression tree on the training data to predict the response. Create a plot of the tree.

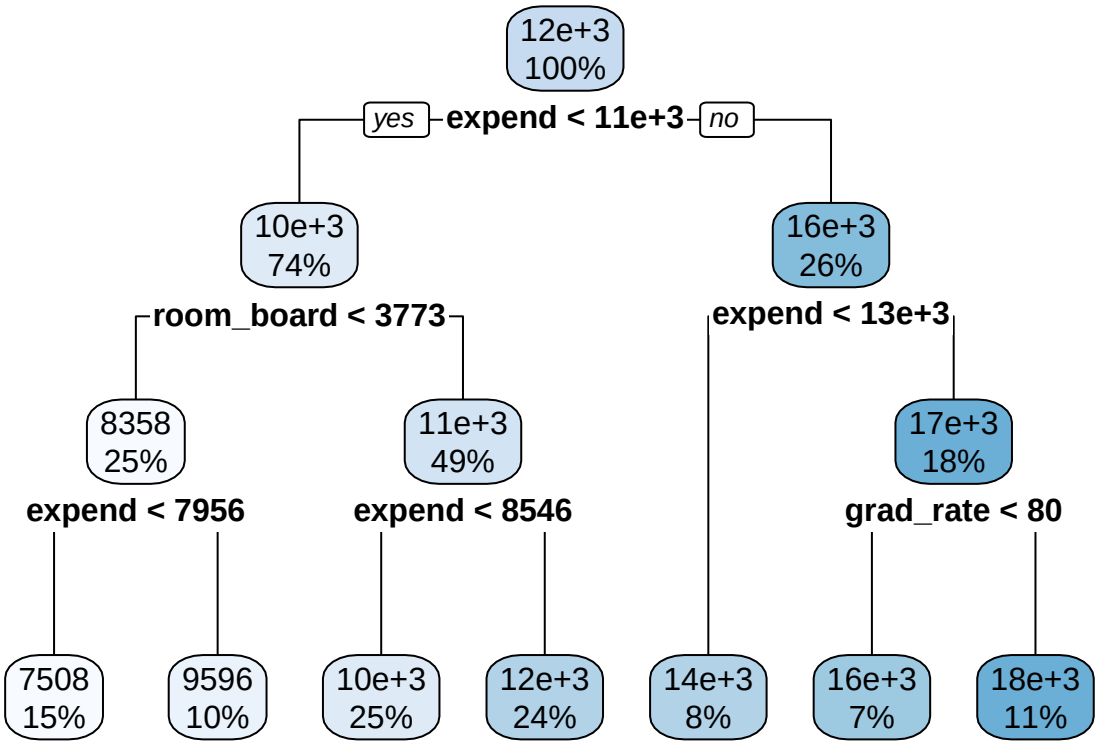
```
ctrl <- trainControl(method = "cv")
set.seed(1234)

# Using Package 'caret' to built the regression tree model

rpart.fit <- train(outstate ~ . ,
                   College[RowTrain,],
                   method = "rpart",
                   tuneGrid = data.frame(cp = exp(seq(-7,1, length = 90))),
                   trControl = ctrl)


# Report the tuning parameter
ggplot(rpart.fit, highlight = TRUE)
```




```
rpart.fit$bestTune
```

```
##           cp
## 33 0.01618621
```

```
# Plot the tree
rpart.plot(rpart.fit$finalModel)
```



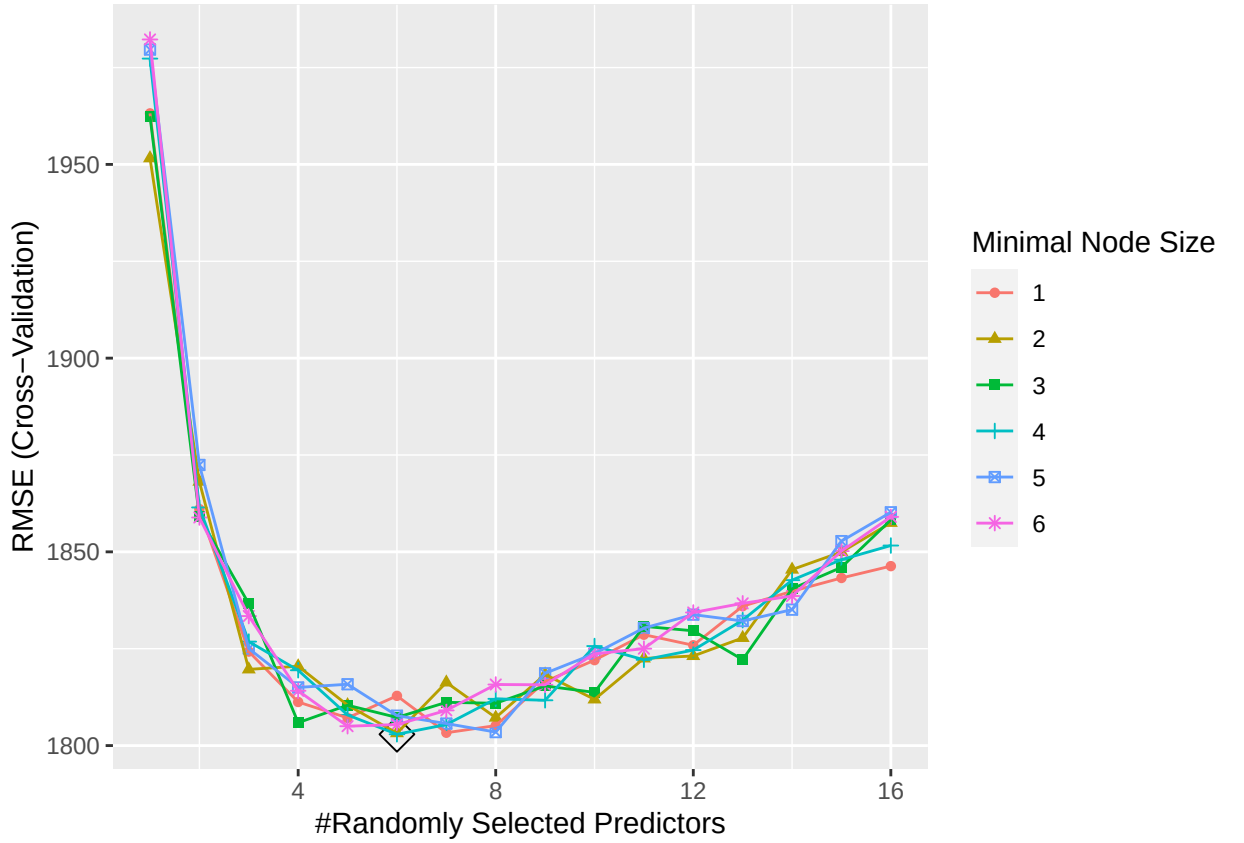
* From the final model, we can report the best tuning parameter `cp` is 0.0161862120750658.

Question B

Perform random forest on the training data. Report the variable importance and the test error.

```
set.seed(1234)

# Using 'caret' package to perform random forest

rf.grid <- expand.grid(mtry = 1:16,
                       splitrule = "variance",
                       min.node.size = 1:6)

rf.fit <- train(outstate ~ . ,
                College[RowTrain,],
                method = "ranger",
                tuneGrid = rf.grid,
                trControl = ctrl)
```

```
# Report the tuning parameter
ggplot(rf.fit, highlight = TRUE)
```

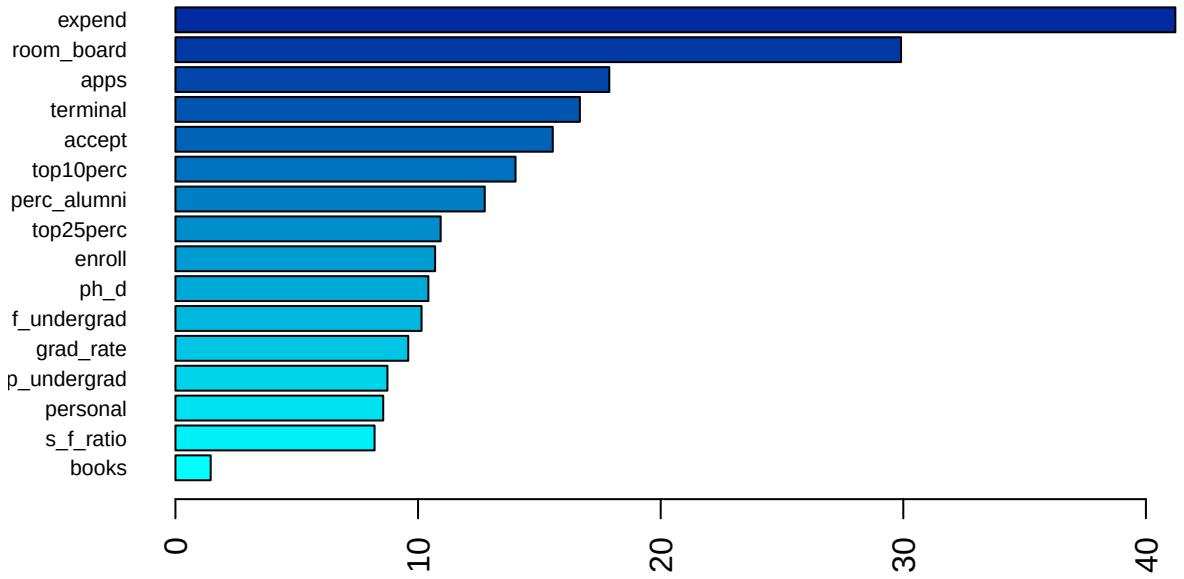



```
rf.fit$bestTune
```

```
##    mtry splitrule min.node.size
## 34    6  variance             4
```

- From the above output, ww can know that the best tuning parameters selected via CV are `mtry = 6`, `splitrule = variance` and `min.node.size = 4`.

```
set.seed(1234)
# We can extract the variable importance from the fitted models.

rf2.final.per <- ranger(outstate ~ . ,
                        College[RowTrain,],
                         mtry = rf.fit$bestTune[[1]],
                         splitrule = "variance",
                         min.node.size = rf.fit$bestTune[[3]],
                         importance = "permutation",
                         scale.permutation.importance = TRUE)
barplot(sort(ranger::importance(rf2.final.per), decreasing = FALSE),
        las = 2, horiz = TRUE, cex.names = 0.7,
        col = colorRampPalette(colors = c("cyan","darkblue"))(19))
```



* We see that the variables `expend(Instructional expenditure per student)`, `Room_Board(Room and board costs)` and `apps(Number of applications received)` are the top 3 from the variable importance.

```
pred.rf <- predict(rf.fit, newdata = College[-RowTrain,])
test.error <- RMSE(pred.rf, College$outstate[-RowTrain])
test.error
```

```
## [1] 1565.881
```

- The test error is 1565.880541.