Harvard Project Two: Classify Schools

Ferry Edouard

2023-12-05

Introduction

The goal of this project will be exploring the use of tree methods to classify schools as Private or Public based off their features.

Let's start by getting the data which is included in the ISLR library, the College data frame.
A data frame with 777 observations on the following 18 variables.
Private A factor with levels No and Yes indicating private or public university
Apps Number of applications received
Accept Number of applications accepted
Enroll Number of new students enrolled
Top10perc Pct. new students from top 10% of H.S. class
Top25perc Pct. new students from top 25% of H.S. class
F.Undergrad Number of fulltime undergraduates
P.Undergrad Number of parttime undergraduates
Outstate Out-of-state tuition
Room.Board Room and board costs
Books Estimated book costs
Personal Estimated personal spending
PhD Pct. of faculty with Ph.D.'s
Terminal Pct. of faculty with terminal degree
S.F.Ratio Student/faculty ratio
perc.alumni Pct. alumni who donate
Expend Instructional expenditure per student
Grad.Rate Graduation rate

Project goal:

Classify schools as Private or Public based off their features.

For this project we will be exploring the use of tree methods and Random Forest to classify schools as Private or Public based off their features.

Data exploration

```
#Let's start by getting the data which is included in the ISLR library, the College data frame.

library(ISLR)
head(College)
```

```
##                              Private Apps Accept Enroll Top10perc Top25perc
## Abilene Christian University     Yes 1660   1232    721        23        52
## Adelphi University               Yes 2186   1924    512        16        29
## Adrian College                   Yes 1428   1097    336        22        50
## Agnes Scott College              Yes  417    349    137        60        89
## Alaska Pacific University        Yes  193    146     55        16        44
## Albertson College                Yes  587    479    158        38        62
##                              F.Undergrad P.Undergrad Outstate Room.Board Books
## Abilene Christian University        2885         537     7440       3300   450
## Adelphi University                  2683        1227    12280       6450   750
## Adrian College                      1036          99    11250       3750   400
## Agnes Scott College                  510          63    12960       5450   450
## Alaska Pacific University            249         869     7560       4120   800
## Albertson College                    678          41    13500       3335   500
##                              Personal PhD Terminal S.F.Ratio perc.alumni Expend
## Abilene Christian University     2200  70       78      18.1          12   7041
## Adelphi University               1500  29       30      12.2          16  10527
## Adrian College                   1165  53       66      12.9          30   8735
## Agnes Scott College               875  92       97       7.7          37  19016
## Alaska Pacific University        1500  76       72      11.9           2  10922
## Albertson College                 675  67       73       9.4          11   9727
##                              Grad.Rate
## Abilene Christian University        60
## Adelphi University                  56
## Adrian College                      54
## Agnes Scott College                 59
## Alaska Pacific University           15
## Albertson College                   55
```

```
str(College)
```

```
## 'data.frame':    777 obs. of  18 variables:
##  $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Apps       : num  1660 2186 1428 417 193 ...
##  $ Accept     : num  1232 1924 1097 349 146 ...
##  $ Enroll     : num  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : num  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : num  52 29 50 89 44 62 45 68 63 44 ...
##  $ F.Undergrad: num  2885 2683 1036 510 249 ...
##  $ P.Undergrad: num  537 1227 99 63 869 ...
##  $ Outstate   : num  7440 12280 11250 12960 7560 ...
##  $ Room.Board : num  3300 6450 3750 5450 4120 ...
##  $ Books      : num  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : num  2200 1500 1165 875 1500 ...
```

```
##  $ PhD        : num  70 29 53 92 76 67 90 89 79 40 ...
##  $ Terminal   : num  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: num  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : num  7041 10527 8735 19016 10922 ...
##  $ Grad.Rate  : num  60 56 54 59 15 55 63 73 80 52 ...
```

```r
summary(College)
```

```
##  Private        Apps           Accept          Enroll       Top10perc    
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00  
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00  
##            Median : 1558   Median : 1110   Median : 434   Median :23.00  
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56  
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00  
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00  
##    Top25perc      F.Undergrad     P.Undergrad         Outstate    
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340  
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320  
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990  
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441  
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925  
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700  
##    Room.Board       Books           Personal         PhD        
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00  
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00  
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00  
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66  
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00  
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00  
##     Terminal       S.F.Ratio      perc.alumni        Expend     
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186  
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751  
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377  
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660  
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830  
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233  
##    Grad.Rate     
##  Min.   : 10.00  
##  1st Qu.: 53.00  
##  Median : 65.00  
##  Mean   : 65.46  
##  3rd Qu.: 78.00  
##  Max.   :118.00  
```

Get the Data

```r
#Call the ISLR library and check the head of College (a built-in data frame with ISLR,
#use data() to check this.) Then reassign College to a dataframe called df

data("College")
df <- College
head(df)
```

```
##                              Private Apps Accept Enroll Top10perc Top25perc
## Abilene Christian University     Yes 1660   1232    721        23        52
## Adelphi University               Yes 2186   1924    512        16        29
## Adrian College                   Yes 1428   1097    336        22        50
## Agnes Scott College              Yes  417    349    137        60        89
## Alaska Pacific University        Yes  193    146     55        16        44
## Albertson College                Yes  587    479    158        38        62
##                              F.Undergrad P.Undergrad Outstate Room.Board Books
## Abilene Christian University        2885         537     7440       3300   450
## Adelphi University                  2683        1227    12280       6450   750
## Adrian College                      1036          99    11250       3750   400
## Agnes Scott College                  510          63    12960       5450   450
## Alaska Pacific University            249         869     7560       4120   800
## Albertson College                    678          41    13500       3335   500
##                              Personal PhD Terminal S.F.Ratio perc.alumni Expend
## Abilene Christian University     2200  70       78      18.1          12   7041
## Adelphi University               1500  29       30      12.2          16  10527
## Adrian College                   1165  53       66      12.9          30   8735
## Agnes Scott College               875  92       97       7.7          37  19016
## Alaska Pacific University        1500  76       72      11.9           2  10922
## Albertson College                 675  67       73       9.4          11   9727
##                              Grad.Rate
## Abilene Christian University        60
## Adelphi University                  56
## Adrian College                      54
## Agnes Scott College                 59
## Alaska Pacific University           15
## Albertson College                   55
```

EDA

```r
#Let's explore the data!

#Create a scatterplot of Grad.Rate versus Room.Board, colored by the
#Private column.

#Call of
library(car)
```

```
## Loading required package: carData
```

```r
library(carData)
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:randomForest':
##
##     margin
```

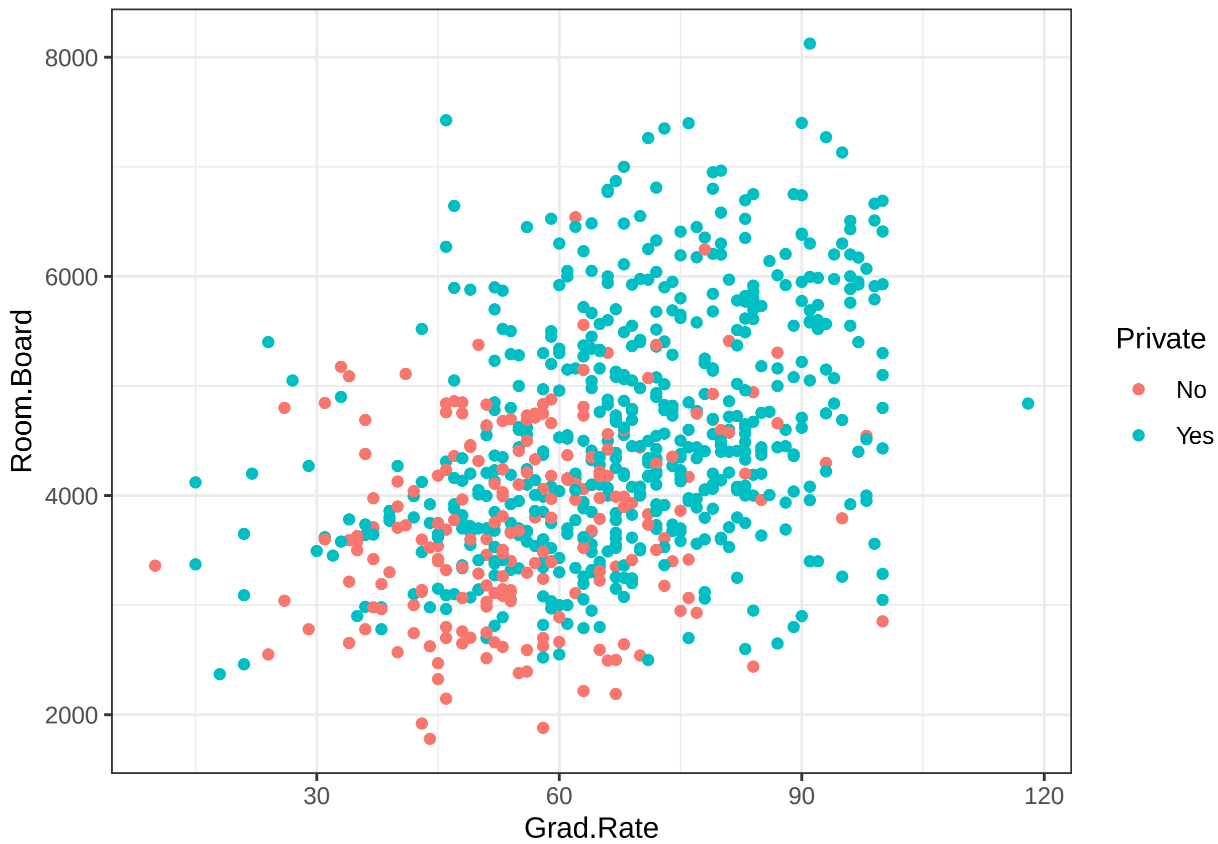
```r
set.seed(101)

head(df)
```

```
##                              Private Apps Accept Enroll Top10perc Top25perc
## Abilene Christian University     Yes 1660   1232    721        23        52
## Adelphi University               Yes 2186   1924    512        16        29
## Adrian College                   Yes 1428   1097    336        22        50
## Agnes Scott College              Yes  417    349    137        60        89
## Alaska Pacific University        Yes  193    146     55        16        44
## Albertson College                Yes  587    479    158        38        62
##                              F.Undergrad P.Undergrad Outstate Room.Board Books
## Abilene Christian University        2885         537     7440       3300   450
## Adelphi University                  2683        1227    12280       6450   750
## Adrian College                      1036          99    11250       3750   400
## Agnes Scott College                  510          63    12960       5450   450
## Alaska Pacific University            249         869     7560       4120   800
## Albertson College                    678          41    13500       3335   500
##                              Personal PhD Terminal S.F.Ratio perc.alumni Expend
## Abilene Christian University     2200  70       78      18.1          12   7041
## Adelphi University               1500  29       30      12.2          16  10527
## Adrian College                   1165  53       66      12.9          30   8735
## Agnes Scott College               875  92       97       7.7          37  19016
## Alaska Pacific University        1500  76       72      11.9           2  10922
## Albertson College                 675  67       73       9.4          11   9727
##                              Grad.Rate
## Abilene Christian University        60
## Adelphi University                  56
```

```
## Adrian College                      54
## Agnes Scott College                 59
## Alaska Pacific University           15
## Albertson College                   55
```

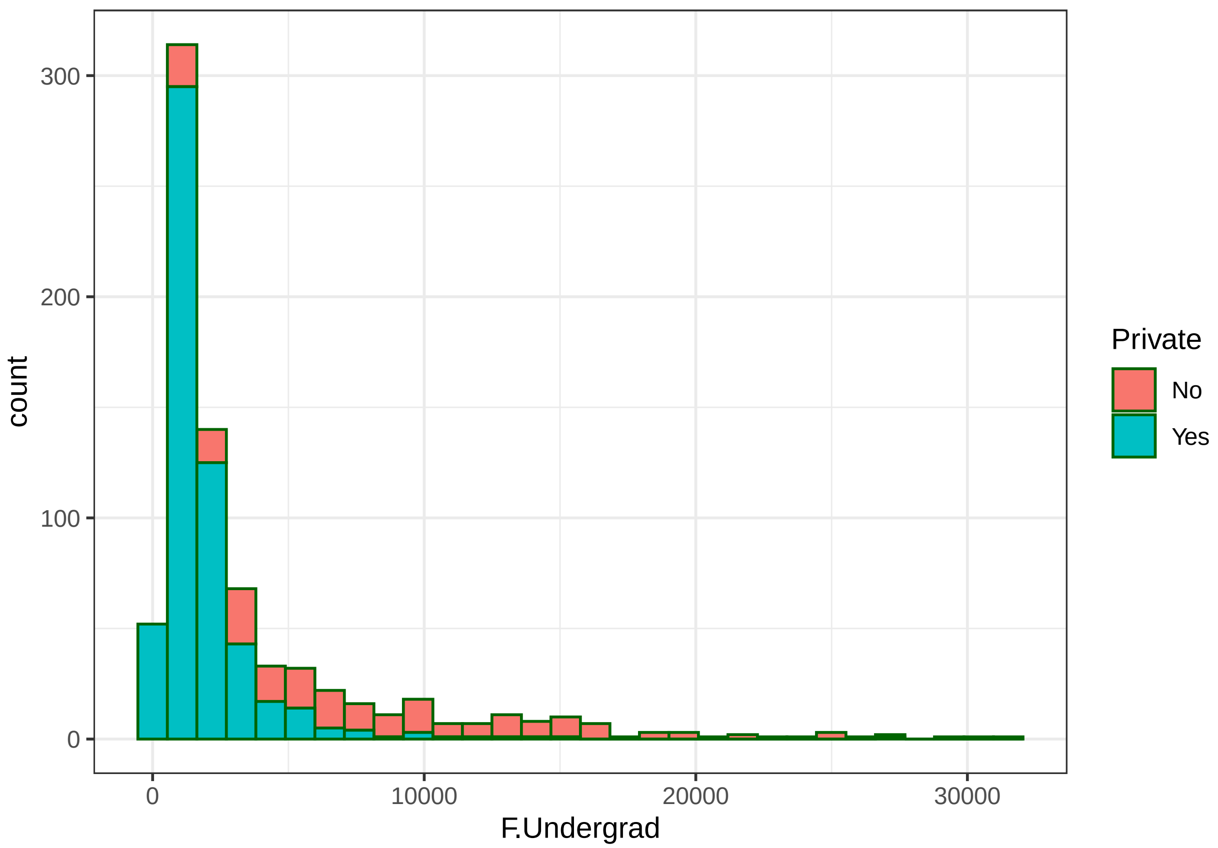
```r
pl <- ggplot(df, aes(Grad.Rate, Room.Board)) + geom_point(aes(color = Private)) + theme_bw()
print(pl)
```



Create a histogram of full time undergrad students, color by Private.

```r
pl2 <- ggplot(df, aes(F.Undergrad)) + geom_histogram(aes(fill = Private), color = "dark green") + theme_
print(pl2)
```

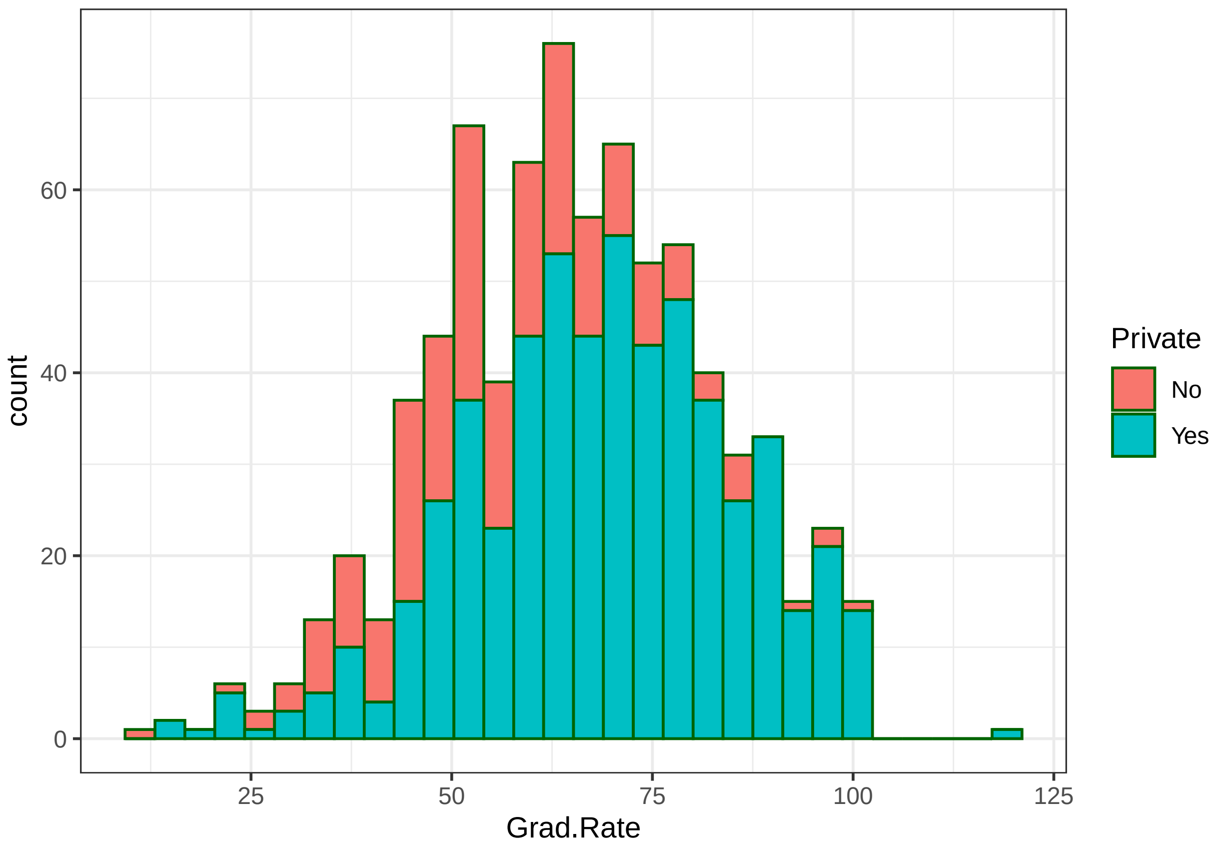
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Create a histogram of Grad.Rate colored by Private. We should see something odd here.

```
ggplot(df, aes(Grad.Rate)) + geom_histogram(aes(fill = Private), color = "dark green") + theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



What college had a Graduation Rate of above 100% ?

```r
subset(df, Grad.Rate > 100)
```

```
##                   Private Apps Accept Enroll Top10perc Top25perc F.Undergrad
## Cazenovia College     Yes 3847   3433    527         9        35        1010
##                   P.Undergrad Outstate Room.Board Books Personal PhD Terminal
## Cazenovia College          12     9384       4840   600      500  22       47
##                   S.F.Ratio perc.alumni Expend Grad.Rate
## Cazenovia College      14.3          20   7697       118
```

```r
#Change that college's grad rate to 100%
df["Cazenovia College", "Grad.Rate"] <- 100
```

Train Test Split

Now let's split the data into training and testing sets 70/30. Use the caTools library to do this.

```
library(caTools)
set.seed(101)

sample <- sample.split(df$Private, SplitRatio = 0.7)

#Training
train = subset(df, sample = T)
```

```
## Warning: In subset.data.frame(df, sample = T) :
##  extra argument 'sample' will be disregarded
```

```
#Testing
test = subset(df, sample = F)
```

```
## Warning: In subset.data.frame(df, sample = F) :
##  extra argument 'sample' will be disregarded
```

```
#Decision Tree

#Use the rpart library to build a decision tree to predict whether
#or not a school is Private. Remember to only build your tree off
#the training data.

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following object is masked from 'package:randomForest':
##
##     combine

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(explore)

library(rpart)
library(rpart.plot)

tree <- rpart(Private ~ . , method='class', data= train)
prp(tree)



#Use predict() to predict the Private label on the test data.

tree.prediction <- predict(tree, test)



#Check the Head of the predicted values. We should notice that we
#actually have two columns with the probabilities.

head(tree.prediction)
```

```
##                                      No       Yes
## Abilene Christian University 0.35714286 0.6428571
## Adelphi University           0.00462963 0.9953704
## Adrian College               0.00462963 0.9953704
## Agnes Scott College          0.00462963 0.9953704
## Alaska Pacific University    0.08823529 0.9117647
## Albertson College            0.00462963 0.9953704
```

```r
#Turn these two columns into one column to match the original
#Yes/No Label for a Private column.

tree.prediction <- as.data.frame(tree.prediction)

unity <- function(x){
  if (x >= 0.5){
    return("Yes")

  }else{
    return("No")
  }

}

tree.prediction$Private <- sapply(tree.prediction$Yes, unity)
head(tree.prediction)
```

```
##                                      No       Yes Private
## Abilene Christian University 0.35714286 0.6428571     Yes
## Adelphi University           0.00462963 0.9953704     Yes
## Adrian College               0.00462963 0.9953704     Yes
## Agnes Scott College          0.00462963 0.9953704     Yes
```

```
## Alaska Pacific University    0.08823529 0.9117647     Yes
## Albertson College            0.00462963 0.9953704     Yes
```

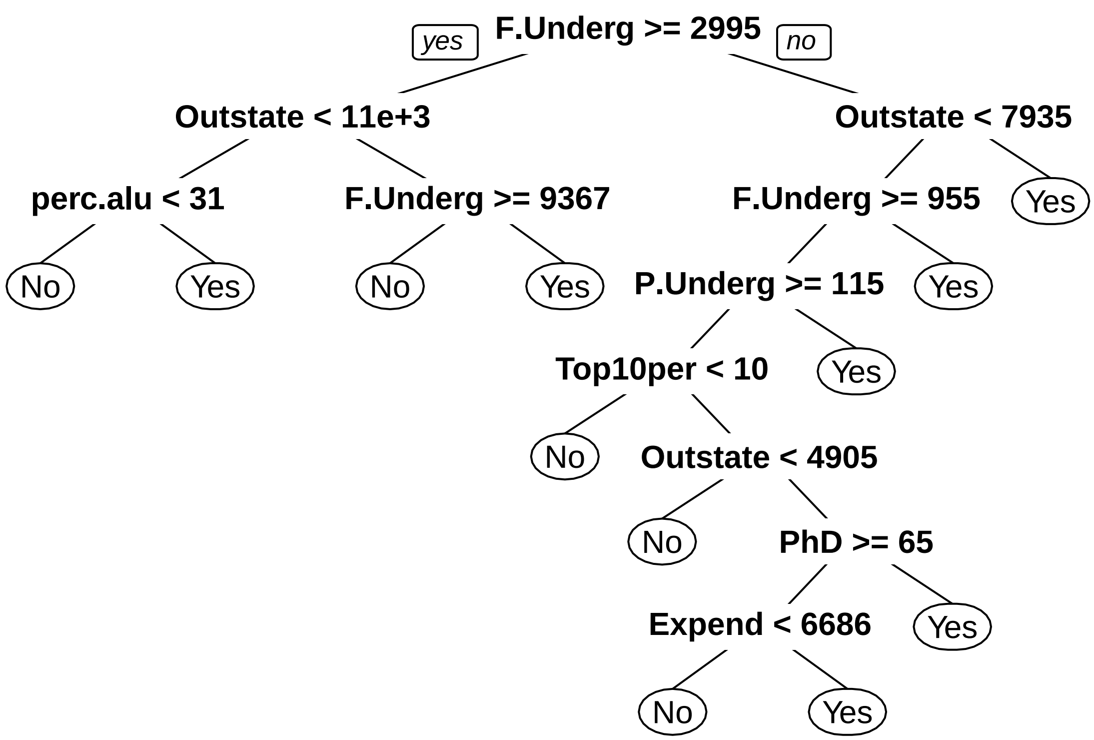
```r
#Now let's use table() to create a confusion matrix of the tree model.

table(tree.prediction$Private, test$Private)
```

```
##
##          No Yes
##   No    195  14
##   Yes    17 551
```

```r
#Use the rpart.plot library and the prp() function to plot out
#the tree model.

prp(tree)
```



Random Forest

```r
#Now let's build out a random forest model!

#Call the randomForest package library

library(randomForest)



#Now use randomForest() to build out a model to predict Private
#class. Add importance=TRUE as a parameter in the model.
#(Use help(randomForest) to find out what this does.


model <- randomForest(Private ~ .,   data = train, importance = T)
print(model) # view results
```

```
##
## Call:
##  randomForest(formula = Private ~ ., data = train, importance = T)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 6.05%
## Confusion matrix:
##      No Yes class.error
## No  183  29  0.13679245
## Yes  18 547  0.03185841
```

```r
importance(model) # importance of each predictor
```

```
##                    No        Yes MeanDecreaseAccuracy MeanDecreaseGini
## Apps         9.034161 10.0481973           13.9560396        12.037964
## Accept      10.275079 10.2035503           14.6311592        18.383314
## Enroll      13.738844 15.4805852           20.0473507        32.007931
## Top10perc    8.422479  6.3309716           10.2012964         5.839418
## Top25perc    6.815916  7.8765842           10.8976170         5.218735
## F.Undergrad 28.234686 24.5077686           36.4090879        60.311072
## P.Undergrad 17.009457  9.6050630           19.7839475        23.452528
## Outstate    37.076161 33.1300906           45.5805811        63.717629
## Room.Board  11.718784 18.0363376           21.0432009        16.885770
## Books        1.944080 -0.7662283            0.9563453         2.733010
## Personal     3.949495  3.2066984            5.1190105         4.204704
## PhD          9.164732 10.1710671           13.6789546         6.885597
## Terminal     5.490625 10.9494548           12.4408339         5.304365
## S.F.Ratio   14.471752  8.0742058           16.6404388        19.993068
## perc.alumni 17.188324  4.9908065           16.6018721         9.113808
## Expend      13.488598 12.3195344           16.5521631        13.177318
## Grad.Rate   10.930332  7.7852865           12.5690049         8.684059
```

```r
#What was the model's confusion matrix on its own training set?
#Use model$confusion.

model$confusion
```

```
##      No Yes class.error
## No  183  29  0.13679245
## Yes  18 547  0.03185841
```

```r
#
# Confusion matrix:
#   No Yes class.error
# No  183  29  0.13679245
# Yes  18 547  0.03185841
# > model$confusion
# No Yes class.error
# No  183  29  0.13679245
# Yes  18 547  0.03185841
# >



#Grab the feature importance with model$importance.
model$importance
```

```
##                      No           Yes MeanDecreaseAccuracy MeanDecreaseGini
## Apps        0.026400519  0.0132810757         0.0167738117        12.037964
## Accept      0.036728848  0.0139517989         0.0201423818        18.383314
## Enroll      0.051260145  0.0333490697         0.0382285046        32.007931
## Top10perc   0.008525153  0.0035623266         0.0048685548         5.839418
## Top25perc   0.005854879  0.0038389543         0.0043593354         5.218735
## F.Undergrad 0.164482590  0.0621680611         0.0901789617        60.311072
## P.Undergrad 0.053091887  0.0073666472         0.0198098231        23.452528
## Outstate    0.162510583  0.0657032835         0.0920753821        63.717629
## Room.Board  0.020449840  0.0151832826         0.0165841658        16.885770
## Books       0.001120627 -0.0001493092         0.0002075755         2.733010
## Personal    0.002716087  0.0009497650         0.0014177517         4.204704
## PhD         0.012249642  0.0065743826         0.0081273010         6.885597
## Terminal    0.006032989  0.0068166915         0.0066042199         5.304365
## S.F.Ratio   0.032858213  0.0052031651         0.0127746714        19.993068
## perc.alumni 0.034714988  0.0026478245         0.0113764480         9.113808
## Expend      0.022916743  0.0134608296         0.0160353445        13.177318
## Grad.Rate   0.013940550  0.0052197427         0.0075682611         8.684059
```

Predictions

```r
#Now use your random forest model to predict on the test set!

p <- predict(model, test)

table(p, test$Private)
```

```
##
## p      No Yes
##   No  212   0
##   Yes   0 565
```

Conclusion

This is the end of this project, classify schools as Private or | Public based off their features.. Performance wise, it should have been better if it wasn't just a single tree, how much better depends on whether we are emasuring recall, precision, or accuracy as the most important measure of the model.