Regression Classification Trees and Random Forest

Code

Vanessa Gonzalez

2018-06-25

Open libraries

Hide

```
library("mlbench")
library("dplyr")
library("caret")
library("randomForest")
library("lattice")
library("ggplot2")
library("rpart")
library("e1071")
library("caret", lib.loc="/Library/Frameworks/R.framework/Versions/3.4/Resources/libr
ary")
```

Partitions Creation

Hide

```
head(dfDataSet4YGImpute)
```

	4...	1_CSCI1...	1_MATH...	2_CSCI2...	2_MATH...	2_MATH...	3_CSCI2...	3_MATH...	4_CSCI3.
	<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl
1	Yes	4.000000	3	4	2	3.000000	3.000000	4	2.00000
2	No	3.625546	3	4	2	3.000000	3.603932	3	3.10005
5	Yes	3.776819	3	4	3	3.000000	3.665798	3	3.66465
6	Yes	4.000000	3	3	3	2.000000	3.000000	4	3.00000
7	Yes	4.000000	4	4	4	4.000000	4.000000	4	4.00000
8	Yes	4.000000	3	3	4	3.790084	4.000000	4	3.51737

6 rows | 1-10 of 18 columns

The RandomForest library does not accept numbers or “_" in columng titles so the column titles need to be modified.

Hide

```
colnames(dfDataSet4YGImpute) <- c("FourYG", "One.CSCI101","One.MATH111","Two.CSCI261"
,"Two.MATH112","Two.MATH201","Three.CSCI262","Three.MATH213","Four.CSCI341","Four.CSC
I358","Four.MATH225","Five.CSCI306","Five.CSCI403","Five.MATH332","Six.CSCI406","Seve
n.CSCI370","Eight.CSCI400","Eight.CSCI442")
head(dfDataSet4YGImpute)
```

	Fou... <fctr>	One.CSCI101 <dbl>	One.MATH... <dbl>	Two.CSCI261 <dbl>	Two.MATH... <dbl>	Two.MATH... <dbl>	Three.CSCI262 <dbl>
1	Yes	4.000000	3	4	2	3.000000	3.000000
2	No	3.625546	3	4	2	3.000000	3.603932
5	Yes	3.776819	3	4	3	3.000000	3.665798
6	Yes	4.000000	3	3	3	2.000000	3.000000
7	Yes	4.000000	4	4	4	4.000000	4.000000
8	Yes	4.000000	3	3	4	3.790084	4.000000

6 rows | 1-9 of 18 columns

Parrtitions created with 80% of data for training and 20% of data for testing.

Hide

```
inTraining <- createDataPartition(dfDataSet4YGImpute$FourYG, p = 0.80, list = FALSE)
training <- dfDataSet4YGImpute[inTraining, ]
testing <- dfDataSet4YGImpute[-inTraining, ]
```

Hide

```
training
```

	Fou... <fctr>	One.CSCI101 <dbl>	One.MATH... <dbl>	Two.CSCI261 <dbl>	Two.MATH... <dbl>	Two.MATH... <dbl>	Three.CSCI26 <dbl
1	Yes	4.000000	3.000000	4.000000	2.000000	3.000000	3.00000
2	No	3.625546	3.000000	4.000000	2.000000	3.000000	3.60393
5	Yes	3.776819	3.000000	4.000000	3.000000	3.000000	3.66579

7	Yes	4.000000	4.000000	4.000000	4.000000	4.000000	4.00000
11	Yes	4.000000	3.000000	4.000000	3.000000	3.000000	4.00000
12	Yes	4.000000	2.000000	3.000000	3.000000	2.000000	4.00000
14	No	1.000000	3.000000	3.000000	3.000000	2.000000	2.00000
18	Yes	4.000000	3.000000	4.000000	3.000000	1.000000	4.00000
20	Yes	4.000000	3.000000	4.000000	2.000000	3.014598	3.00000
21	Yes	4.000000	3.000000	3.000000	2.000000	1.000000	3.00000

1-10 of 318 rows | 1-8 of 18 columns Previous **1** 2 3 4 5 6 ... 32 Next

Hide

```
testing
```

	Fou… <fctr>	One.CSCI101 <dbl>	One.MATH… <dbl>	Two.CSCI261 <dbl>	Two.MATH… <dbl>	Two.MATH… <dbl>	Three.CSCI26 <dbl
6	Yes	4.000000	3.000000	3.000000	3.000000	2.000000	3.00000
8	Yes	4.000000	3.000000	3.000000	4.000000	3.790084	4.00000
10	Yes	3.328175	3.000000	3.000000	2.000000	1.000000	2.98414
13	Yes	3.000000	4.000000	4.000000	3.000000	2.000000	3.00000
16	Yes	3.889459	3.000000	4.000000	3.000000	3.000000	4.00000
30	Yes	4.000000	4.000000	4.000000	4.000000	4.000000	4.00000
44	Yes	4.000000	3.000000	3.000000	3.000000	4.000000	3.00000
52	No	3.000000	4.000000	3.000000	3.000000	2.000000	2.00000
57	No	3.000000	4.000000	4.000000	4.000000	3.000000	4.00000
65	Yes	3.642363	3.000000	3.455580	3.000000	2.000000	3.66243

1-10 of 78 rows | 1-8 of 18 columns Previous **1** 2 3 4 5 6 ... 8 Next

Regresion Partition with method “class”.

Hide

```
FourYG.rp = rpart(FourYG ~ ., data=training, method = "class")
FourYG.rp
```

```
n= 318

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 318 90 Yes (0.2830189 0.7169811)
   2) Five.CSCI403< 3.419795 65 28 No (0.5692308 0.4307692)
     4) Five.CSCI306< 3.572585 49 14 No (0.7142857 0.2857143)
       8) Four.CSCI358< 2.705482 26  3 No (0.8846154 0.1153846) *
       9) Four.CSCI358>=2.705482 23 11 No (0.5217391 0.4782609)
        18) Eight.CSCI442>=2.169579 16  5 No (0.6875000 0.3125000) *
        19) Eight.CSCI442< 2.169579 7  1 Yes (0.1428571 0.8571429) *
     5) Five.CSCI306>=3.572585 16  2 Yes (0.1250000 0.8750000) *
   3) Five.CSCI403>=3.419795 253 53 Yes (0.2094862 0.7905138)
     6) One.CSCI101< 2.65 8  2 No (0.7500000 0.2500000) *
     7) One.CSCI101>=2.65 245 47 Yes (0.1918367 0.8081633)
      14) Eight.CSCI400< 1.85 8  3 No (0.6250000 0.3750000) *
      15) Eight.CSCI400>=1.85 237 42 Yes (0.1772152 0.8227848)
        30) Six.CSCI406< 1.818216 7  3 No (0.5714286 0.4285714) *
        31) Six.CSCI406>=1.818216 230 38 Yes (0.1652174 0.8347826) *
```

Display CP table for Fitted Rpart Object. The main variables used in this classification tree were: CSCI206, CSCI403, CSCI341, and CSCI406.

Hide

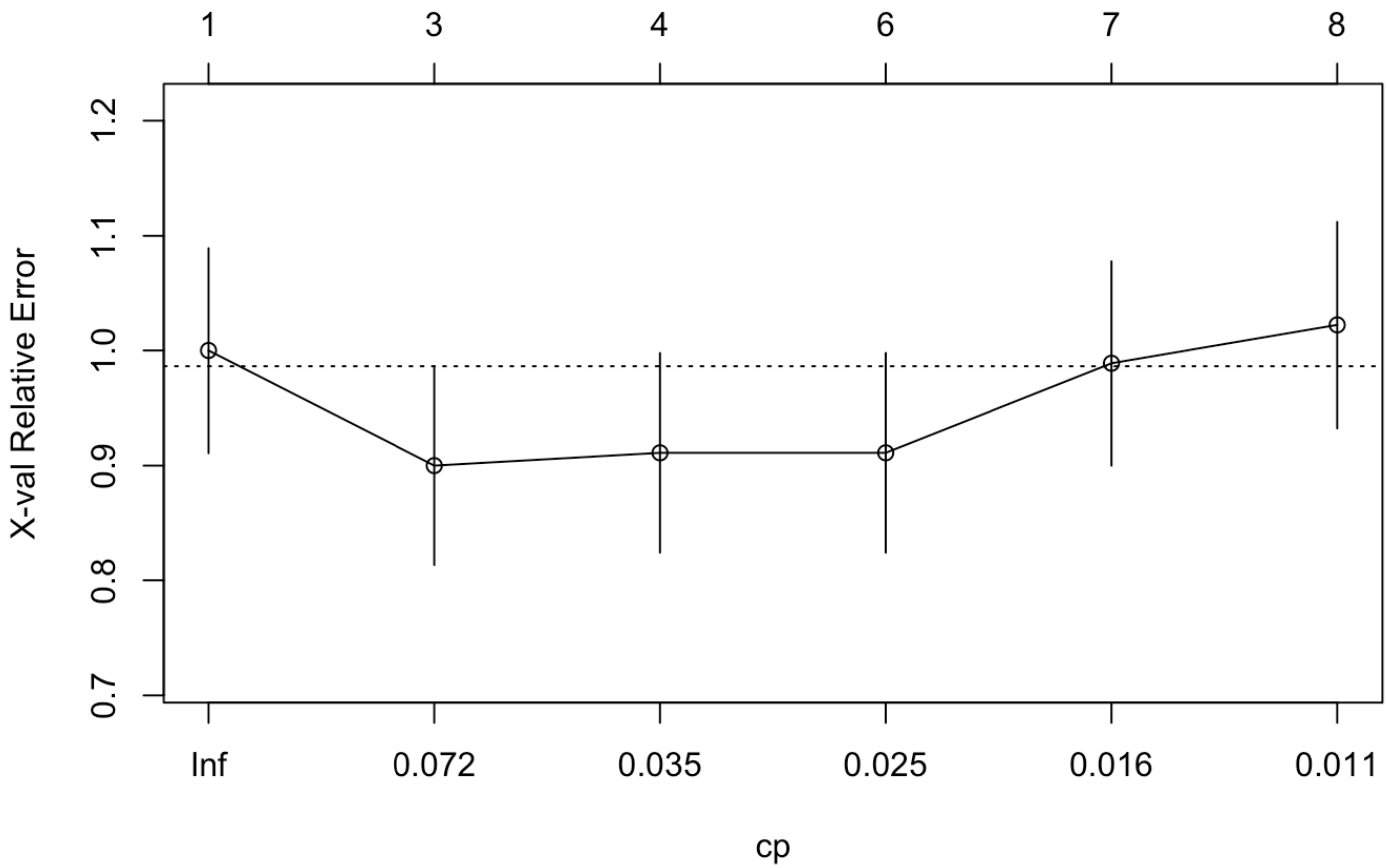
```
printcp(FourYG.rp)
```

```
Classification tree:
rpart(formula = FourYG ~ ., data = training, method = "class")

Variables actually used in tree construction:
[1] Eight.CSCI400 Eight.CSCI442 Five.CSCI306  Five.CSCI403  Four.CSCI358  One.CSCI101
Six.CSCI406

Root node error: 90/318 = 0.28302

n= 318

        CP nsplit rel error  xerror     xstd
1 0.116667      0   1.00000 1.00000 0.089255
2 0.044444      2   0.76667 0.90000 0.086330
3 0.027778      3   0.72222 0.91111 0.086678
4 0.022222      5   0.66667 0.91111 0.086678
5 0.011111      6   0.64444 0.98889 0.088952
6 0.010000      7   0.63333 1.02222 0.089845
```

Plot CP.

Hide

```
plotcp(FourYG.rp)
```



To look at the importance of variables in the regression partition with method “class”.

Hide

```
summary(FourYG.rp)
```

```
Call:
rpart(formula = FourYG ~ ., data = training, method = "class")
  n= 318

          CP nsplit rel error    xerror       xstd
1 0.11666667      0 1.0000000 1.0000000 0.08925501
2 0.04444444      2 0.7666667 0.9000000 0.08632978
3 0.02777778      3 0.7222222 0.9111111 0.08667760
4 0.02222222      5 0.6666667 0.9111111 0.08667760
5 0.01111111      6 0.6444444 0.9888889 0.08895220
6 0.01000000      7 0.6333333 1.0222222 0.08984462

Variable importance
 Five.CSCI403  Five.CSCI306 Eight.CSCI400   One.CSCI101   Six.CSCI406 Eight.CSCI442
Four.CSCI358   Two.MATH201   Two.MATH112
           22            17            13            12            10             9
```

```
5              3             3 
Three.CSCI262   One.MATH111 Three.MATH213 
            2             1             1 

Node number 1: 318 observations,    complexity param=0.1166667
  predicted class=Yes  expected loss=0.2830189  P(node) =1
    class counts:    90   228
   probabilities: 0.283 0.717 
  left son=2 (65 obs) right son=3 (253 obs)
  Primary splits:
      Five.CSCI403  < 3.419795 to the left,  improve=13.385210, (0 missing)
      Five.CSCI306  < 3.325026 to the left,  improve=11.971780, (0 missing)
      Four.CSCI358  < 2.696418 to the left,  improve=11.258070, (0 missing)
      Six.CSCI406   < 2.266458 to the left,  improve=10.967200, (0 missing)
      Eight.CSCI400 < 3.008728 to the left,  improve= 9.909563, (0 missing)
  Surrogate splits:
      Eight.CSCI400 < 2.387419 to the left,  agree=0.836, adj=0.200, (0 split)
      Eight.CSCI442 < 2.089747 to the left,  agree=0.827, adj=0.154, (0 split)
      Six.CSCI406   < 1.911649 to the left,  agree=0.814, adj=0.092, (0 split)
      Five.CSCI306  < 2.15     to the left,  agree=0.811, adj=0.077, (0 split)
      Two.MATH201   < 1.35     to the left,  agree=0.808, adj=0.062, (0 split)

Node number 2: 65 observations,    complexity param=0.1166667
  predicted class=No   expected loss=0.4307692  P(node) =0.2044025
    class counts:    37    28
   probabilities: 0.569 0.431 
  left son=4 (49 obs) right son=5 (16 obs)
  Primary splits:
      Five.CSCI306 < 3.572585 to the left,  improve=8.376923, (0 missing)
      Four.CSCI358 < 2.705482 to the left,  improve=5.934066, (0 missing)
      Six.CSCI406  < 2.363306 to the left,  improve=4.191934, (0 missing)
      Two.MATH201  < 2.780642 to the left,  improve=3.452681, (0 missing)
      Five.MATH332 < 1.945752 to the left,  improve=3.450256, (0 missing)
  Surrogate splits:
      Eight.CSCI400 < 3.380197 to the left,  agree=0.800, adj=0.188, (0 split)
      Two.MATH112   < 3.5      to the left,  agree=0.785, adj=0.125, (0 split)
      Six.CSCI406   < 3.5      to the left,  agree=0.785, adj=0.125, (0 split)
      Eight.CSCI442 < 3.504269 to the left,  agree=0.769, adj=0.063, (0 split)

Node number 3: 253 observations,    complexity param=0.04444444
  predicted class=Yes  expected loss=0.2094862  P(node) =0.7955975
    class counts:    53   200
   probabilities: 0.209 0.791 
  left son=6 (8 obs) right son=7 (245 obs)
  Primary splits:
      One.CSCI101   < 2.65     to the left,  improve=4.827119, (0 missing)
      Eight.CSCI400 < 1.85     to the left,  improve=3.901024, (0 missing)
```

```
      Eight.CSCI442 < 3.233423 to the left,  improve=2.556588, (0 missing)
      Three.CSCI262 < 2.573857 to the left,  improve=2.484895, (0 missing)
      Four.CSCI358  < 2.344163 to the left,  improve=2.296267, (0 missing)

Node number 4: 49 observations,    complexity param=0.02777778
  predicted class=No   expected loss=0.2857143  P(node) =0.1540881
    class counts:    35    14
   probabilities: 0.714 0.286 
  left son=8 (26 obs) right son=9 (23 obs)
  Primary splits:
      Four.CSCI358 < 2.705482 to the left,  improve=3.214047, (0 missing)
      Five.MATH332 < 1.945752 to the left,  improve=2.594595, (0 missing)
      Six.CSCI406  < 2.363306 to the left,  improve=2.260684, (0 missing)
      Five.CSCI306 < 2.968379 to the left,  improve=2.074510, (0 missing)
      Two.CSCI261  < 3.455126 to the left,  improve=1.333333, (0 missing)
  Surrogate splits:
      One.CSCI101   < 3.297335 to the left,  agree=0.776, adj=0.522, (0 split)
      Three.CSCI262 < 3.136483 to the left,  agree=0.714, adj=0.391, (0 split)
      Six.CSCI406   < 2.220706 to the left,  agree=0.714, adj=0.391, (0 split)
      Two.MATH201   < 2.295413 to the left,  agree=0.694, adj=0.348, (0 split)
      Five.CSCI306  < 2.968379 to the left,  agree=0.694, adj=0.348, (0 split)

Node number 5: 16 observations
  predicted class=Yes  expected loss=0.125  P(node) =0.05031447
    class counts:     2    14
   probabilities: 0.125 0.875 

Node number 6: 8 observations
  predicted class=No   expected loss=0.25  P(node) =0.02515723
    class counts:     6     2
   probabilities: 0.750 0.250 

Node number 7: 245 observations,    complexity param=0.02222222
  predicted class=Yes  expected loss=0.1918367  P(node) =0.7704403
    class counts:    47   198
   probabilities: 0.192 0.808 
  left son=14 (8 obs) right son=15 (237 obs)
  Primary splits:
      Eight.CSCI400 < 1.85     to the left,  improve=3.103423, (0 missing)
      Four.CSCI358  < 2.344163 to the left,  improve=2.207347, (0 missing)
      Six.CSCI406   < 1.818216 to the left,  improve=2.076591, (0 missing)
      Three.CSCI262 < 2.573857 to the left,  improve=2.066146, (0 missing)
      Eight.CSCI442 < 3.233423 to the left,  improve=1.937271, (0 missing)

Node number 8: 26 observations
  predicted class=No   expected loss=0.1153846  P(node) =0.08176101
    class counts:    23     3
```

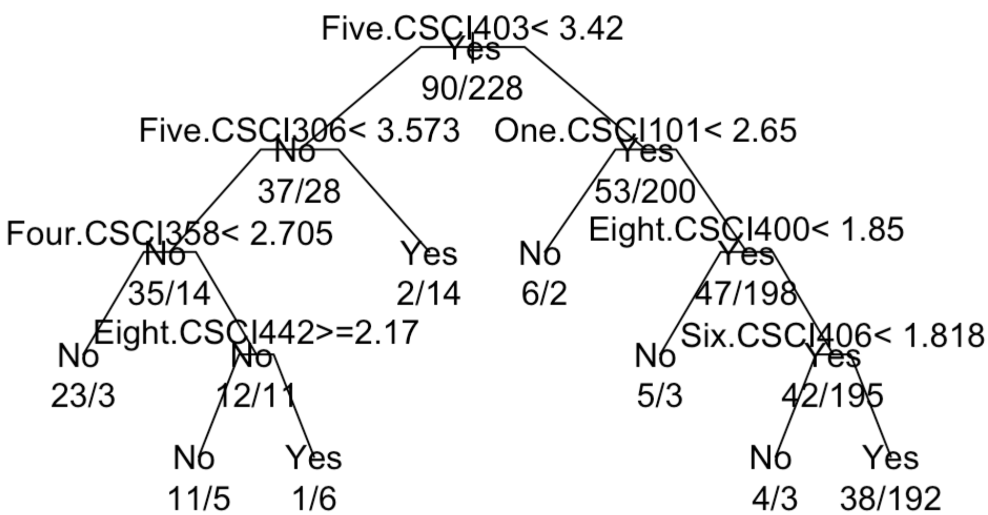
```
   probabilities: 0.885 0.115 

Node number 9: 23 observations,    complexity param=0.02777778
  predicted class=No   expected loss=0.4782609  P(node) =0.07232704
    class counts:    12    11
   probabilities: 0.522 0.478 
  left son=18 (16 obs) right son=19 (7 obs)
  Primary splits:
      Eight.CSCI442 < 2.169579 to the right, improve=2.888975, (0 missing)
      Eight.CSCI400 < 2.831697 to the right, improve=2.739799, (0 missing)
      Five.MATH332  < 2.080122 to the left,  improve=2.739799, (0 missing)
      Four.MATH225  < 2.5      to the right, improve=1.124415, (0 missing)
      Three.CSCI262 < 3.274875 to the right, improve=1.054018, (0 missing)
  Surrogate splits:
      One.CSCI101   < 2.5      to the right, agree=0.783, adj=0.286, (0 split)
      One.MATH111   < 2.5      to the right, agree=0.783, adj=0.286, (0 split)
      Two.MATH112   < 2.5      to the right, agree=0.783, adj=0.286, (0 split)
      Three.MATH213 < 2.5      to the right, agree=0.783, adj=0.286, (0 split)
      Eight.CSCI400 < 1.35     to the right, agree=0.783, adj=0.286, (0 split)

Node number 14: 8 observations
  predicted class=No   expected loss=0.375  P(node) =0.02515723
    class counts:     5     3
   probabilities: 0.625 0.375 

Node number 15: 237 observations,    complexity param=0.01111111
  predicted class=Yes  expected loss=0.1772152  P(node) =0.745283
    class counts:    42   195
   probabilities: 0.177 0.823 
  left son=30 (7 obs) right son=31 (230 obs)
  Primary splits:
      Six.CSCI406   < 1.818216 to the left,  improve=2.2418740, (0 missing)
      Four.CSCI358  < 2.344163 to the left,  improve=1.7871840, (0 missing)
      Eight.CSCI442 < 3.233423 to the left,  improve=1.5535340, (0 missing)
      Three.CSCI262 < 2.423857 to the left,  improve=1.4494800, (0 missing)
      Eight.CSCI400 < 2.648349 to the right, improve=0.9345519, (0 missing)

Node number 18: 16 observations
  predicted class=No   expected loss=0.3125  P(node) =0.05031447
    class counts:    11     5
   probabilities: 0.688 0.312 

Node number 19: 7 observations
  predicted class=Yes  expected loss=0.1428571  P(node) =0.02201258
    class counts:     1     6
   probabilities: 0.143 0.857 
```

```
Node number 30: 7 observations
  predicted class=No   expected loss=0.4285714  P(node) =0.02201258
    class counts:     4     3
   probabilities: 0.571 0.429

Node number 31: 230 observations
  predicted class=Yes  expected loss=0.1652174  P(node) =0.7232704
    class counts:    38   192
   probabilities: 0.165 0.835
```

Plot of classification Tree

Hide

```
plot(FourYG.rp, uniform=TRUE, branch=.3, margin=0.2)
text(FourYG.rp, all=TRUE, use.n = TRUE)
```



Utilizing the regression classification tree model the results for the testing set are below.

Hide

```
predictions = predict(FourYG.rp, testing, type="class")
table(testing$FourYG, predictions)
```

```
     predictions
      No Yes
  No   7  15
  Yes  6  50
```

A confusion matrix is created to compare prediction results with testing results.

Hide

```
library(caret)
confusionMatrix(table(predictions, testing$FourYG))
```

```
Confusion Matrix and Statistics


predictions No Yes
        No   7   6
        Yes 15  50

               Accuracy : 0.7308
                 95% CI : (0.6184, 0.825)
    No Information Rate : 0.7179
    P-Value [Acc > NIR] : 0.45719

                  Kappa : 0.241
 Mcnemar's Test P-Value : 0.08086

            Sensitivity : 0.31818
            Specificity : 0.89286
         Pos Pred Value : 0.53846
         Neg Pred Value : 0.76923
             Prevalence : 0.28205
         Detection Rate : 0.08974
   Detection Prevalence : 0.16667
      Balanced Accuracy : 0.60552

       'Positive' Class : No
```

Hide

```
min(FourYG.rp$cptable[,"xerror"])
```

```
[1] 0.9
```

Hide

```
which.min(FourYG.rp$cptable[,"xerror"])
```

```
3
3
```

Prune the tree to increase accuracy. Get the cost complecity parameter of the record

Hide

```
FourYG.cp = FourYG.rp$cptable[3,"CP"]
FourYG.cp
```

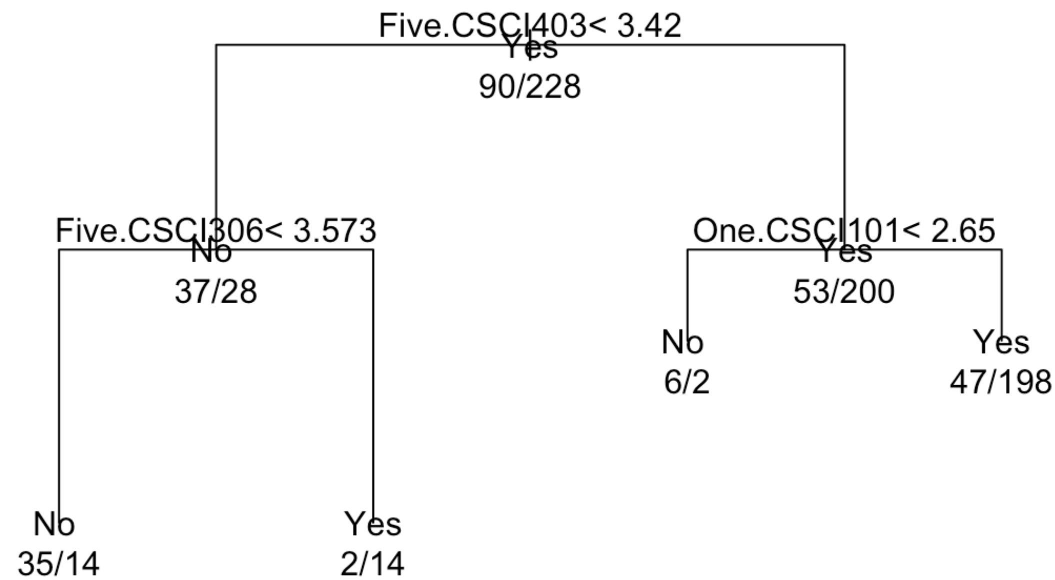
```
[1] 0.02777778
```

Hide

```
prune.tree = prune(FourYG.rp, cp= FourYG.cp)
```

Hide

```
plot(prune.tree, margin= 0.1)
text(prune.tree, all=TRUE , use.n=TRUE)
```



Prune tree.

Hide

```
prune.tree = prune(FourYG.rp, cp = FourYG.cp)
predictions.prune = predict(prune.tree, testing, type="class")
table(testing$FourYG, predictions.prune)
```

```
     predictions.prune
      No Yes
  No   9  13
  Yes  6  50
```

Confusion matrix for punned tree.

Hide

```
confusionMatrix(table(predictions.prune, testing$FourYG))
```

```
Confusion Matrix and Statistics


predictions.prune No Yes
              No   9   6
              Yes 13  50

               Accuracy : 0.7564
                 95% CI : (0.646, 0.8465)
    No Information Rate : 0.7179
    P-Value [Acc > NIR] : 0.2685

                  Kappa : 0.3342
 Mcnemar's Test P-Value : 0.1687

            Sensitivity : 0.4091
            Specificity : 0.8929
         Pos Pred Value : 0.6000
         Neg Pred Value : 0.7937
             Prevalence : 0.2821
         Detection Rate : 0.1154
   Detection Prevalence : 0.1923
      Balanced Accuracy : 0.6510

       'Positive' Class : No
```

Random Forest Method

Hide

```
FourYG.rf <- randomForest(FourYG ~One.CSCI101+One.MATH111+Two.CSCI261+Two.MATH112+Two
.MATH201+Three.CSCI262+Three.MATH213+Four.CSCI341+Four.CSCI358+Four.MATH225+Five.CSCI
306+Five.CSCI403+Five.MATH332+Six.CSCI406+Seven.CSCI370+Eight.CSCI400+Eight.CSCI442
, data = training)
FourYG.rf
```

```
Call:
 randomForest(formula = FourYG ~ One.CSCI101 + One.MATH111 + Two.CSCI261 +      Two.M
ATH112 + Two.MATH201 + Three.CSCI262 + Three.MATH213 +      Four.CSCI341 + Four.CSCI3
58 + Four.MATH225 + Five.CSCI306 +      Five.CSCI403 + Five.MATH332 + Six.CSCI406 + S
even.CSCI370 +      Eight.CSCI400 + Eight.CSCI442, data = training)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 26.73%
Confusion matrix:
    No Yes class.error
No  26  64  0.71111111
Yes 21 207  0.09210526
```

Hide

```
FourYG.rf.prediction <- predict(FourYG.rf, testing)
table(FourYG.rf.prediction, testing$FourYG)
```

```
FourYG.rf.prediction No Yes
                 No   7   4
                 Yes 15  52
```

To determine variable imortance.

Hide

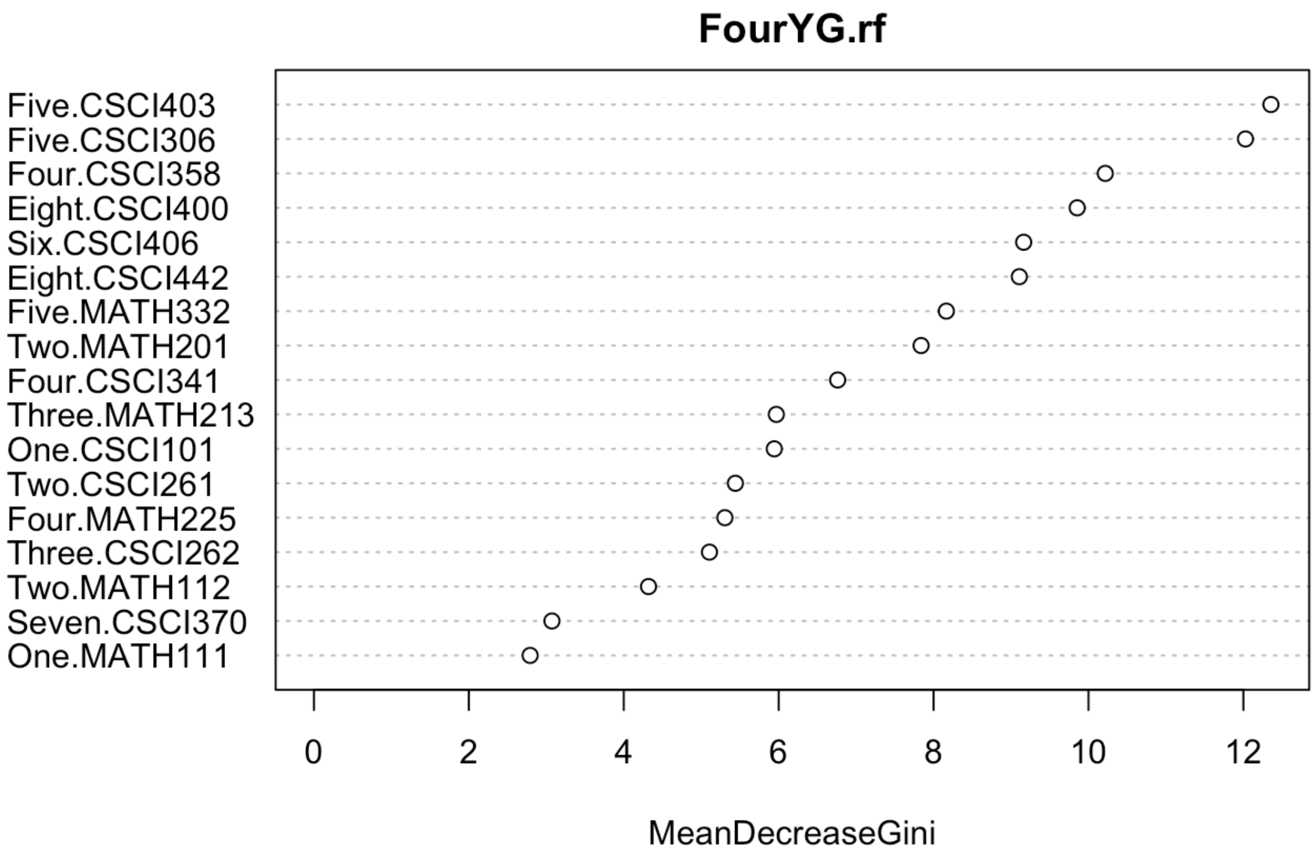
```
importance(FourYG.rf)
```

```
               MeanDecreaseGini
One.CSCI101            5.943245
One.MATH111            2.791131
Two.CSCI261            5.441233
Two.MATH112            4.319703
Two.MATH201            7.838733
Three.CSCI262          5.106540
Three.MATH213          5.969680
Four.CSCI341           6.761883
Four.CSCI358          10.216689
Four.MATH225           5.305591
Five.CSCI306          12.026363
Five.CSCI403          12.355417
Five.MATH332           8.164668
Six.CSCI406            9.163340
Seven.CSCI370          3.072915
Eight.CSCI400          9.854125
Eight.CSCI442          9.107240
```

Plot for variable importance

Hide

```
varImpPlot(FourYG.rf)
```



Hide

```
confusionMatrix(table(FourYG.rf.prediction, testing$FourYG))
```

```
Confusion Matrix and Statistics


FourYG.rf.prediction No Yes
                 No   7   4
                 Yes 15  52

               Accuracy : 0.7564
                 95% CI : (0.646, 0.8465)
    No Information Rate : 0.7179
    P-Value [Acc > NIR] : 0.26848

                  Kappa : 0.2909
 Mcnemar's Test P-Value : 0.02178

            Sensitivity : 0.31818
            Specificity : 0.92857
         Pos Pred Value : 0.63636
         Neg Pred Value : 0.77612
             Prevalence : 0.28205
         Detection Rate : 0.08974
   Detection Prevalence : 0.14103
      Balanced Accuracy : 0.62338

       'Positive' Class : No
```

Logistic Regression Method for Variable Importance

A different method was tried to confirm the variable importance obtained through Random Forest. The results were very similar.

Hide

```
# Template code
# Step 1: Build Logit Model on Training Dataset
FourYG.lr <- glm(FourYG ~One.CSCI101+One.MATH111+Two.CSCI261+Two.MATH112+Two.MATH201+
Three.CSCI262+Three.MATH213+Four.CSCI341+Four.CSCI358+Four.MATH225+Five.CSCI306+Five.
CSCI403+Five.MATH332+Six.CSCI406+Seven.CSCI370+Eight.CSCI400+Eight.CSCI442, family= "
binomial", data = training)
FourYG.lr
```

```
Call:  glm(formula = FourYG ~ One.CSCI101 + One.MATH111 + Two.CSCI261 + 
    Two.MATH112 + Two.MATH201 + Three.CSCI262 + Three.MATH213 + 
    Four.CSCI341 + Four.CSCI358 + Four.MATH225 + Five.CSCI306 + 
    Five.CSCI403 + Five.MATH332 + Six.CSCI406 + Seven.CSCI370 + 
    Eight.CSCI400 + Eight.CSCI442, family = "binomial", data = training)

Coefficients:
  (Intercept)    One.CSCI101    One.MATH111    Two.CSCI261    Two.MATH112    Two.MATH
201  Three.CSCI262  Three.MATH213  
     -7.58580        0.12995        0.19853        0.15016       -0.04938       -0.12
892       -0.06610        0.08899  
 Four.CSCI341   Four.CSCI358   Four.MATH225   Five.CSCI306   Five.CSCI403   Five.MATH
332    Six.CSCI406  Seven.CSCI370  
     -0.07123        0.45633       -0.22594        0.55716        0.62781        0.09
829        0.23332        0.36814  
Eight.CSCI400  Eight.CSCI442  
      0.08548        0.02620  

Degrees of Freedom: 317 Total (i.e. Null);  300 Residual
Null Deviance:      378.9 
Residual Deviance: 321.3    AIC: 357.3
```

Hide

```
# Step 2: Predict Y on Test Dataset
predictedY <- predict(FourYG.lr, testing, type="response")
```

Check prediction

Hide

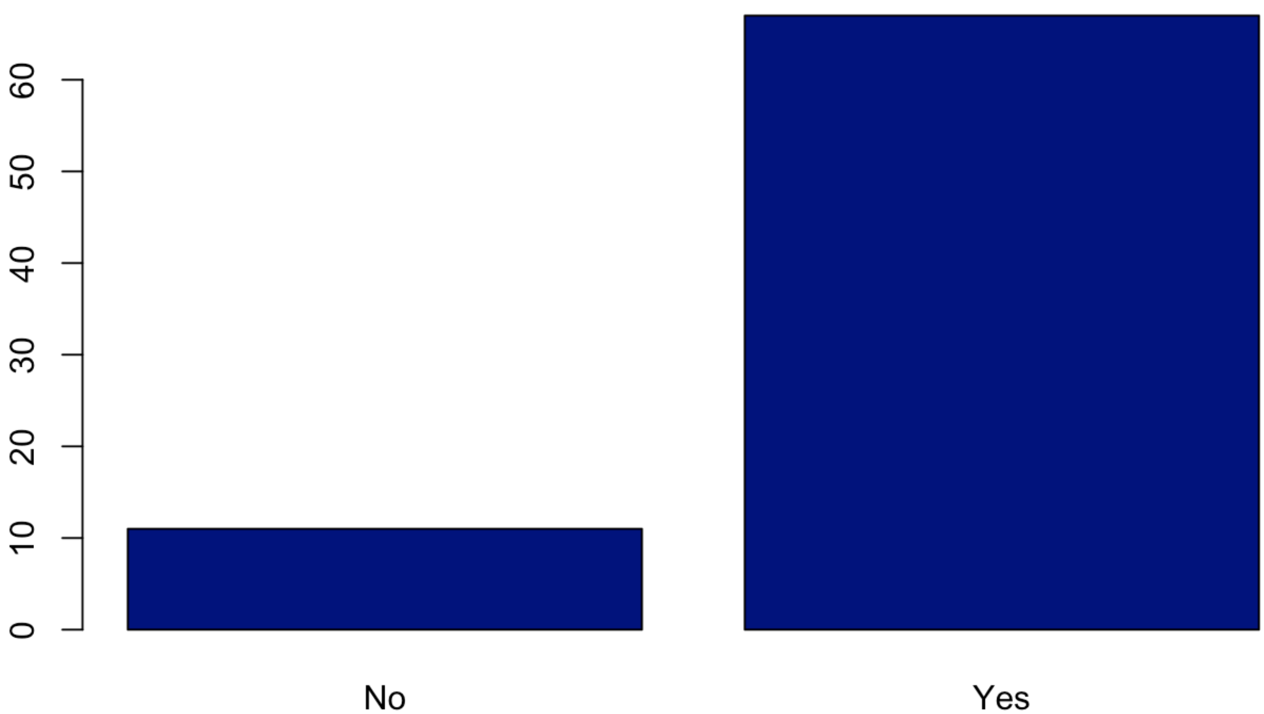
```
predictedY.rf <- predict(FourYG.rf, testing, type="response")
predictedY.rf
```

```
  6   8  10  13  16  30  44  52  57  65  70  74  76  81  83  87  91 102 103 113 118 1
23 130 139 148 167 185 200 202 206 209 213 219
Yes Yes Yes Yes Yes Yes Yes  No Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes  No  No Y
es Yes Yes Yes Yes Yes Yes Yes Yes Yes  No  No
226 236 254 256 265 269 273 276 277 280 288 312 315 322 326 330 343 356 363 365 367 3
69 378 383 391 392 397 402 403 417 423 429 432
Yes Yes Yes Yes Yes Yes Yes  No Yes Yes  No Yes Yes Yes  No Yes Yes Yes Yes Yes Yes
No Yes  No Yes Yes Yes Yes Yes Yes Yes Yes Yes
435 436 440 454 457 459 473 478 496 499 510 526
Yes Yes  No Yes Yes Yes Yes Yes Yes Yes Yes Yes
Levels: No Yes
```

Plot prediction

Hide

```
plot(predictedY.rf, col = "navy blue")
```



For a list of importance of variables

Hide

```
gbmImp <- varImp(FourYG.rf, scale = FALSE)
gbmImp
```

	Overall <dbl>
One.CSCI101	5.943245
One.MATH111	2.791131
Two.CSCI261	5.441233
Two.MATH112	4.319703
Two.MATH201	7.838733
Three.CSCI262	5.106540
Three.MATH213	5.969680
Four.CSCI341	6.761883
Four.CSCI358	10.216689
Four.MATH225	5.305591

1-10 of 17 rows Previous **1** 2 Next