Code

Prediction of Students at Risk on Semester 5

Vanessa Gonzalez

2018-06-24

Open Libraries.

Hide

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

Random Forest model using less variables for semester 5

Hide

```
colnames(dfDataSet4YGImpute) <- c("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")
LessVariablesSet.Sem5 <- dfDataSet4YGImpute
```

Hide

```
LessVariablesSet.Sem5
```

	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
6	Yes	4.000000	3.000000	3.000000	3.000000	2.000000	3.00000
7	Yes	4.000000	4.000000	4.000000	4.000000	4.000000	4.00000
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
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
13	Yes	3.000000	4.000000	4.000000	3.000000	2.000000	3.00000

1-10 of 396 rows | 1-8 of 18 columns Previous **1** 2 3 4 5 6 … 40 Next

Create data partitions and remove variables to leave only classes before 6th semester.

Hide

```
# Creates Data Partitions and removes variables
inTrainingLess.Sem5 <- createDataPartition(LessVariablesSet.Sem5$FourYG, p = 0.80, li
st = FALSE)
LessVariablesSet.Sem5 <- LessVariablesSet[-(2:3)]
LessVariablesSet.Sem5 <- LessVariablesSet[-(3)]
LessVariablesSet.Sem5 <- LessVariablesSet[-(5)]
LessVariablesSet.Sem5 <- LessVariablesSet[-(7)]
LessVariablesSet.Sem5 <- LessVariablesSet[-(11)]
LessVariablesSet.Sem5 <- LessVariablesSet[-(11:12)]
str(LessVariablesSet.Sem5)
```

```
'data.frame':   396 obs. of  10 variables:
 $ FourYG       : Factor w/ 2 levels "No","Yes": 2 1 2 2 2 2 2 2 2 2 ...
 $ Two.CSCI261  : num  4 4 4 3 4 3 3 4 3 4 ...
 $ Two.MATH201  : num  3 3 3 2 4 ...
 $ Three.CSCI262: num  3 3.6 3.67 3 4 ...
 $ Four.CSCI341 : num  2 3.1 3.66 3 4 ...
 $ Four.CSCI358 : num  4 3.1 3.72 2 4 ...
 $ Five.CSCI306 : num  3 3.62 3.74 4 4 ...
 $ Five.CSCI403 : num  4 3.9 4 4 4 ...
 $ Five.MATH332 : num  3 2.56 2.78 3 4 ...
 $ Six.CSCI406  : num  2 2.83 3.16 2 4 ...
```

Hide

```
# Creates Training data Set
trainingLess.Sem5 <- LessVariablesSet.Sem5[inTrainingLess.Sem5, ]
# Creates Testing data Set
testingLess.Sem5 <- LessVariablesSet.Sem5[-inTrainingLess.Sem5, ]
# Data Set with less varialbes
head(LessVariablesSet.Sem5)
```

	Fou… <fctr>	Two.CSCI261 <dbl>	Two.MATH… <dbl>	Three.CSCI262 <dbl>	Four.CSCI341 <dbl>	Four.CSCI358 <dbl>	Five.CSCI3 <dl
1	Yes	4	3.000000	3.000000	2.000000	4.000000	3.0000
2	No	4	3.000000	3.603932	3.100057	3.099580	3.6212
5	Yes	4	3.000000	3.665798	3.664659	3.720739	3.7434
6	Yes	3	2.000000	3.000000	3.000000	2.000000	4.0000
7	Yes	4	4.000000	4.000000	4.000000	4.000000	4.0000
8	Yes	3	3.790084	4.000000	3.517370	3.942883	4.0000

6 rows | 1-8 of 10 columns

Hide

```
trainingLess.Sem5
```

	Fou… <fctr>	Two.CSCI261 <dbl>	Two.MATH… <dbl>	Three.CSCI262 <dbl>	Four.CSCI341 <dbl>	Four.CSCI358 <dbl>	Five.CS
2	No	4.000000	3.000000	3.603932	3.100057	3.099580	3.62
5	Yes	4.000000	3.000000	3.665798	3.664659	3.720739	3.74
7	Yes	4.000000	4.000000	4.000000	4.000000	4.000000	4.00
8	Yes	3.000000	3.790084	4.000000	3.517370	3.942883	4.00
10	Yes	3.000000	1.000000	2.984141	3.000000	2.166370	3.05
11	Yes	4.000000	3.000000	4.000000	4.000000	4.000000	4.00
12	Yes	3.000000	2.000000	4.000000	3.000000	4.000000	4.00
13	Yes	4.000000	2.000000	3.000000	3.000000	3.000000	4.00
14	No	3.000000	2.000000	2.000000	2.000000	4.000000	3.00
16	Yes	4.000000	3.000000	4.000000	4.000000	4.000000	4.00

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

Hide

```
testingLess.Sem5
```

	Fou…	Two.CSCI261	Two.MATH…	Three.CSCI262	Four.CSCI341	Four.CSCI358	Five.CS
	<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
1	Yes	4.000000	3.000000	3.000000	2.000000	4.000000	3.00
6	Yes	3.000000	2.000000	3.000000	3.000000	2.000000	4.00
18	Yes	4.000000	1.000000	4.000000	3.000000	1.000000	3.00
25	No	4.000000	2.000000	4.000000	2.000000	2.000000	2.00
26	Yes	4.000000	4.000000	4.000000	4.000000	4.000000	4.00
27	Yes	3.744411	3.672929	3.798099	3.724800	3.419065	3.68
29	Yes	3.000000	3.000000	3.000000	4.000000	2.000000	1.00
34	Yes	4.000000	2.000000	3.903218	3.559483	3.271073	3.58
42	Yes	4.000000	3.000000	3.272966	2.678475	3.055259	3.37
44	Yes	3.000000	4.000000	3.000000	4.000000	4.000000	4.00

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

Regresion Partition with method “class” for less variables for 5th semester courses.

Hide

```
FourYG.rp.Less.Sem5 = rpart(FourYG ~ ., data=trainingLess.Sem5, method = "class")
FourYG.rp.Less.Sem5
```

```
n= 318

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

 1) root 318 90 Yes (0.28301887 0.71698113)
   2) Six.CSCI406< 2.835387 113 52 Yes (0.46017699 0.53982301)
     4) Five.CSCI403< 3.570705 50 18 No (0.64000000 0.36000000)
       8) Five.MATH332< 1.984246 14  1 No (0.92857143 0.07142857) *
       9) Five.MATH332>=1.984246 36 17 No (0.52777778 0.47222222)
        18) Five.CSCI306< 3.572585 29 11 No (0.62068966 0.37931034)
          36) Four.CSCI341>=1.998648 22  6 No (0.72727273 0.27272727) *
          37) Four.CSCI341< 1.998648 7  2 Yes (0.28571429 0.71428571) *
        19) Five.CSCI306>=3.572585 7  1 Yes (0.14285714 0.85714286) *
     5) Five.CSCI403>=3.570705 63 20 Yes (0.31746032 0.68253968)
      10) Five.MATH332>=2.338313 29 13 Yes (0.44827586 0.55172414)
        20) Four.CSCI341< 3.328436 20  8 No (0.60000000 0.40000000)
          40) Five.CSCI306< 3.097392 7  1 No (0.85714286 0.14285714) *
          41) Five.CSCI306>=3.097392 13  6 Yes (0.46153846 0.53846154) *
        21) Four.CSCI341>=3.328436 9  1 Yes (0.11111111 0.88888889) *
      11) Five.MATH332< 2.338313 34  7 Yes (0.20588235 0.79411765) *
   3) Six.CSCI406>=2.835387 205 38 Yes (0.18536585 0.81463415)
     6) Three.CSCI262< 2.85 11  5 No (0.54545455 0.45454545) *
     7) Three.CSCI262>=2.85 194 32 Yes (0.16494845 0.83505155) *
```

Hide

```
printcp(FourYG.rp.Less.Sem5)
```

```
Classification tree:
rpart(formula = FourYG ~ ., data = trainingLess.Sem5, method = "class")

Variables actually used in tree construction:
[1] Five.CSCI306  Five.CSCI403  Five.MATH332  Four.CSCI341  Six.CSCI406   Three.CSCI2
62

Root node error: 90/318 = 0.28302

n= 318

        CP nsplit rel error  xerror     xstd
1 0.077778      0   1.00000 1.00000 0.089255
2 0.027778      2   0.84444 0.94444 0.087686
3 0.022222      5   0.75556 0.97778 0.088644
4 0.011111      7   0.71111 0.96667 0.088330
5 0.010000      9   0.68889 0.94444 0.087686
```

Summary for regression partition.

Hide

```
summary(FourYG.rp.Less.Sem5)
```

```
Call:
rpart(formula = FourYG ~ ., data = trainingLess.Sem5, method = "class")
  n= 318

          CP nsplit rel error    xerror       xstd
1 0.07777778      0 1.0000000 1.0000000 0.08925501
2 0.02777778      2 0.8444444 0.9444444 0.08768622
3 0.02222222      5 0.7555556 0.9777778 0.08864397
4 0.01111111      7 0.7111111 0.9666667 0.08833027
5 0.01000000      9 0.6888889 0.9444444 0.08768622

Variable importance
  Six.CSCI406  Five.CSCI306  Four.CSCI341  Five.MATH332 Three.CSCI262   Two.MATH201
Five.CSCI403  Four.CSCI358   Two.CSCI261
          17            16            15            15            13             9
9             3             3

Node number 1: 318 observations,    complexity param=0.07777778
  predicted class=Yes  expected loss=0.2830189  P(node) =1
    class counts:    90   228
   probabilities: 0.283 0.717
```

```
  left son=2 (113 obs) right son=3 (205 obs)
  Primary splits:
      Six.CSCI406   < 2.835387 to the left,  improve=11.002820, (0 missing)
      Five.CSCI403  < 3.419795 to the left,  improve= 9.874517, (0 missing)
      Five.CSCI306  < 3.716578 to the left,  improve= 9.635865, (0 missing)
      Two.MATH201   < 2.769009 to the left,  improve= 8.119167, (0 missing)
      Three.CSCI262 < 2.773857 to the left,  improve= 7.093058, (0 missing)
  Surrogate splits:
      Five.MATH332  < 2.684672 to the left,  agree=0.774, adj=0.363, (0 split)
      Five.CSCI306  < 3.537257 to the left,  agree=0.764, adj=0.336, (0 split)
      Three.CSCI262 < 3.471227 to the left,  agree=0.761, adj=0.327, (0 split)
      Four.CSCI341  < 2.961475 to the left,  agree=0.761, adj=0.327, (0 split)
      Two.MATH201   < 2.849651 to the left,  agree=0.748, adj=0.292, (0 split)

Node number 2: 113 observations,    complexity param=0.07777778
  predicted class=Yes  expected loss=0.460177  P(node) =0.3553459
    class counts:    52    61
   probabilities: 0.460 0.540
  left son=4 (50 obs) right son=5 (63 obs)
  Primary splits:
      Five.CSCI403 < 3.570705 to the left,  improve=5.800006, (0 missing)
      Five.CSCI306 < 3.85     to the left,  improve=5.356559, (0 missing)
      Two.CSCI261  < 3.255126 to the left,  improve=3.927214, (0 missing)
      Five.MATH332 < 1.984246 to the left,  improve=2.684050, (0 missing)
      Four.CSCI358 < 1.15     to the left,  improve=1.738741, (0 missing)
  Surrogate splits:
      Three.CSCI262 < 2.853189 to the left,  agree=0.664, adj=0.24, (0 split)
      Two.CSCI261   < 2.15     to the left,  agree=0.637, adj=0.18, (0 split)
      Four.CSCI341  < 2.464343 to the left,  agree=0.637, adj=0.18, (0 split)
      Five.CSCI306  < 2.15     to the left,  agree=0.611, adj=0.12, (0 split)
      Five.MATH332  < 1.984246 to the left,  agree=0.611, adj=0.12, (0 split)

Node number 3: 205 observations,    complexity param=0.01111111
  predicted class=Yes  expected loss=0.1853659  P(node) =0.6446541
    class counts:    38   167
   probabilities: 0.185 0.815
  left son=6 (11 obs) right son=7 (194 obs)
  Primary splits:
      Three.CSCI262 < 2.85     to the left,  improve=3.0143510, (0 missing)
      Four.CSCI358  < 3.304964 to the left,  improve=1.9284230, (0 missing)
      Two.MATH201   < 2.067636 to the left,  improve=1.1313070, (0 missing)
      Five.CSCI306  < 3.047521 to the left,  improve=0.6972385, (0 missing)
      Four.CSCI341  < 2.833078 to the left,  improve=0.6531152, (0 missing)

Node number 4: 50 observations,    complexity param=0.02777778
  predicted class=No   expected loss=0.36  P(node) =0.1572327
    class counts:    32    18
```

```
   probabilities: 0.640 0.360
  left son=8 (14 obs) right son=9 (36 obs)
  Primary splits:
      Five.MATH332 < 1.984246 to the left,  improve=3.2384130, (0 missing)
      Two.CSCI261  < 3.255126 to the left,  improve=2.9498230, (0 missing)
      Five.CSCI306 < 3.572585 to the left,  improve=2.8971430, (0 missing)
      Two.MATH201  < 2.85     to the left,  improve=0.9700699, (0 missing)
      Four.CSCI341 < 1.998648 to the right, improve=0.9700699, (0 missing)
  Surrogate splits:
      Five.CSCI306 < 2.965983 to the left,  agree=0.78, adj=0.214, (0 split)
      Two.CSCI261  < 1.25     to the left,  agree=0.76, adj=0.143, (0 split)
      Four.CSCI358 < 0.85     to the left,  agree=0.76, adj=0.143, (0 split)

Node number 5: 63 observations,    complexity param=0.02222222
  predicted class=Yes  expected loss=0.3174603  P(node) =0.1981132
    class counts:    20    43
   probabilities: 0.317 0.683
  left son=10 (29 obs) right son=11 (34 obs)
  Primary splits:
      Five.MATH332 < 2.338313 to the right, improve=1.8391130, (0 missing)
      Four.CSCI341 < 3.428408 to the left,  improve=1.0975060, (0 missing)
      Five.CSCI306 < 3.85     to the left,  improve=0.8941799, (0 missing)
      Two.MATH201  < 2.339399 to the left,  improve=0.6220741, (0 missing)
      Two.CSCI261  < 3.15     to the left,  improve=0.5676627, (0 missing)
  Surrogate splits:
      Two.MATH201   < 2.339399 to the right, agree=0.683, adj=0.310, (0 split)
      Three.CSCI262 < 3.166968 to the right, agree=0.667, adj=0.276, (0 split)
      Four.CSCI358  < 2.233185 to the right, agree=0.651, adj=0.241, (0 split)
      Two.CSCI261   < 3.383106 to the right, agree=0.635, adj=0.207, (0 split)
      Five.CSCI306  < 3.189194 to the right, agree=0.635, adj=0.207, (0 split)

Node number 6: 11 observations
  predicted class=No   expected loss=0.4545455  P(node) =0.03459119
    class counts:     6     5
   probabilities: 0.545 0.455

Node number 7: 194 observations
  predicted class=Yes  expected loss=0.1649485  P(node) =0.6100629
    class counts:    32   162
   probabilities: 0.165 0.835

Node number 8: 14 observations
  predicted class=No   expected loss=0.07142857  P(node) =0.04402516
    class counts:    13     1
   probabilities: 0.929 0.071

Node number 9: 36 observations,    complexity param=0.02777778
```

```
  predicted class=No   expected loss=0.4722222  P(node) =0.1132075
    class counts:    19    17
   probabilities: 0.528 0.472
  left son=18 (29 obs) right son=19 (7 obs)
  Primary splits:
      Five.CSCI306 < 3.572585 to the left,  improve=2.574986, (0 missing)
      Two.CSCI261  < 3.255126 to the left,  improve=2.173016, (0 missing)
      Six.CSCI406  < 2.056729 to the right, improve=1.500000, (0 missing)
      Five.MATH332 < 2.223379 to the right, improve=1.469444, (0 missing)
      Four.CSCI341 < 2.133086 to the right, improve=1.344444, (0 missing)
  Surrogate splits:
      Two.MATH201 < 3.15     to the left,  agree=0.861, adj=0.286, (0 split)

Node number 10: 29 observations,    complexity param=0.02222222
  predicted class=Yes  expected loss=0.4482759  P(node) =0.09119497
    class counts:    13    16
   probabilities: 0.448 0.552
  left son=20 (20 obs) right son=21 (9 obs)
  Primary splits:
      Four.CSCI341 < 3.328436 to the left,  improve=2.967050, (0 missing)
      Two.MATH201  < 2.339399 to the left,  improve=1.934301, (0 missing)
      Four.CSCI358 < 3.269341 to the left,  improve=1.721451, (0 missing)
      Five.CSCI306 < 3.097392 to the left,  improve=1.244828, (0 missing)
      Five.MATH332 < 2.823027 to the left,  improve=1.244828, (0 missing)
  Surrogate splits:
      Two.MATH201  < 3.65     to the left,  agree=0.828, adj=0.444, (0 split)
      Four.CSCI358 < 3.448719 to the left,  agree=0.828, adj=0.444, (0 split)
      Five.CSCI306 < 3.85     to the left,  agree=0.828, adj=0.444, (0 split)

Node number 11: 34 observations
  predicted class=Yes  expected loss=0.2058824  P(node) =0.1069182
    class counts:     7    27
   probabilities: 0.206 0.794

Node number 18: 29 observations,    complexity param=0.02777778
  predicted class=No   expected loss=0.3793103  P(node) =0.09119497
    class counts:    18    11
   probabilities: 0.621 0.379
  left son=36 (22 obs) right son=37 (7 obs)
  Primary splits:
      Four.CSCI341  < 1.998648 to the right, improve=2.0707570, (0 missing)
      Two.CSCI261   < 3.455126 to the left,  improve=1.4742200, (0 missing)
      Five.MATH332  < 2.223379 to the right, improve=1.0397880, (0 missing)
      Four.CSCI358  < 2.15     to the left,  improve=1.0317960, (0 missing)
      Three.CSCI262 < 1.85     to the right, improve=0.6811464, (0 missing)
  Surrogate splits:
      Two.CSCI261 < 2.65     to the right, agree=0.793, adj=0.143, (0 split)
```

```
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 20: 20 observations,    complexity param=0.01111111
  predicted class=No   expected loss=0.4  P(node) =0.06289308
    class counts:    12     8
   probabilities: 0.600 0.400
  left son=40 (7 obs) right son=41 (13 obs)
  Primary splits:
      Five.CSCI306  < 3.097392 to the left,  improve=1.4241760, (0 missing)
      Three.CSCI262 < 3.361978 to the right, improve=0.6000000, (0 missing)
      Two.MATH201   < 2.339399 to the left,  improve=0.4000000, (0 missing)
      Four.CSCI341  < 2.687203 to the right, improve=0.4000000, (0 missing)
      Two.CSCI261   < 3.579058 to the right, improve=0.2666667, (0 missing)
  Surrogate splits:
      Four.CSCI341 < 2.687203 to the right, agree=0.75, adj=0.286, (0 split)
      Two.MATH201  < 2.15     to the left,  agree=0.70, adj=0.143, (0 split)
      Six.CSCI406  < 1.5      to the left,  agree=0.70, adj=0.143, (0 split)

Node number 21: 9 observations
  predicted class=Yes  expected loss=0.1111111  P(node) =0.02830189
    class counts:     1     8
   probabilities: 0.111 0.889

Node number 36: 22 observations
  predicted class=No   expected loss=0.2727273  P(node) =0.06918239
    class counts:    16     6
   probabilities: 0.727 0.273

Node number 37: 7 observations
  predicted class=Yes  expected loss=0.2857143  P(node) =0.02201258
    class counts:     2     5
   probabilities: 0.286 0.714

Node number 40: 7 observations
  predicted class=No   expected loss=0.1428571  P(node) =0.02201258
    class counts:     6     1
   probabilities: 0.857 0.143

Node number 41: 13 observations
  predicted class=Yes  expected loss=0.4615385  P(node) =0.0408805
    class counts:     6     7
   probabilities: 0.462 0.538
```

Prediction for regretion partition for less courses before Semester 5.

Hide

```
predictionsLess.Sem5 = predict(FourYG.rp.Less.Sem5, testingLess.Sem5, type="class")
table(testingLess.Sem5$FourYG, predictionsLess.Sem5)
```

```
     predictionsLess.Sem5
      No Yes
  No   8  14
  Yes  7  49
```

Confusion Matrix

Hide

```
library(caret)
confusionMatrix(table(predictionsLess.Sem5, testingLess.Sem5$FourYG))
```

```
Confusion Matrix and Statistics


predictionsLess.Sem5 No Yes
                 No   8   7
                 Yes 14  49

               Accuracy : 0.7308
                 95% CI : (0.6184, 0.825)
    No Information Rate : 0.7179
    P-Value [Acc > NIR] : 0.4572

                  Kappa : 0.2642
 Mcnemar's Test P-Value : 0.1904

            Sensitivity : 0.3636
            Specificity : 0.8750
         Pos Pred Value : 0.5333
         Neg Pred Value : 0.7778
             Prevalence : 0.2821
         Detection Rate : 0.1026
   Detection Prevalence : 0.1923
      Balanced Accuracy : 0.6193

       'Positive' Class : No
```

Hide

```
min(FourYG.rp.Less.Sem5$cptable[,"xerror"])
```

```
[1] 0.9444444
```

Hide

```
which.min(FourYG.rp.Less.Sem5$cptable[,"xerror"])
```

```
2
2
```

Get the cost complecity parameter of the record

Hide

```
FourYG.cp.Less.Sem5 = FourYG.rp.Less.Sem5$cptable[1,"CP"]
FourYG.cp.Less.Sem5
```

```
[1] 0.07777778
```

Prune tree.

Hide

```
prune.tree.Less.Sem5 = prune(FourYG.rp.Less.Sem5, cp = FourYG.cp.Less.Sem5)
predictionsLessPrune.Sem5 = predict(prune.tree.Less.Sem5, testingLess.Sem5, type="cla
ss")
table(testingLess.Sem5$FourYG, predictionsLessPrune.Sem5)
```

```
     predictionsLessPrune.Sem5
      No Yes
  No   0  22
  Yes  0  56
```

Confusion Matrix

Hide

```
confusionMatrix(table(predictionsLessPrune.Sem5, testingLess.Sem5$FourYG))
```

```
Confusion Matrix and Statistics


predictionsLessPrune.Sem5 No Yes
                       No   0   0
                       Yes 22  56

               Accuracy : 0.7179
                 95% CI : (0.6047, 0.8141)
    No Information Rate : 0.7179
    P-Value [Acc > NIR] : 0.5572

                  Kappa : 0
 Mcnemar's Test P-Value : 7.562e-06

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.7179
             Prevalence : 0.2821
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : No
```

Top 10 variables

Hide

```
str(LessVariablesSet.Sem5)
```

```
'data.frame':   396 obs. of  10 variables:
 $ FourYG       : Factor w/ 2 levels "No","Yes": 2 1 2 2 2 2 2 2 2 2 ...
 $ Two.CSCI261  : num  4 4 4 3 4 3 3 4 3 4 ...
 $ Two.MATH201  : num  3 3 3 2 4 ...
 $ Three.CSCI262: num  3 3.6 3.67 3 4 ...
 $ Four.CSCI341 : num  2 3.1 3.66 3 4 ...
 $ Four.CSCI358 : num  4 3.1 3.72 2 4 ...
 $ Five.CSCI306 : num  3 3.62 3.74 4 4 ...
 $ Five.CSCI403 : num  4 3.9 4 4 4 ...
 $ Five.MATH332 : num  3 2.56 2.78 3 4 ...
 $ Six.CSCI406  : num  2 2.83 3.16 2 4 ...
```

Random Forest method with less variables on 5th semester.

Training

Hide

```
FourYG.rf.Less.Sem5 <- randomForest(FourYG ~Two.CSCI261+Two.MATH201+Four.CSCI341+Four
.CSCI358+Five.CSCI306+Five.CSCI403+Five.MATH332+Six.CSCI406 , data = trainingLess.Sem
5)
FourYG.rf.Less.Sem5
```

```
Call:
 randomForest(formula = FourYG ~ Two.CSCI261 + Two.MATH201 + Four.CSCI341 +      Four
.CSCI358 + Five.CSCI306 + Five.CSCI403 + Five.MATH332 +      Six.CSCI406, data = trai
ningLess.Sem5)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 29.25%
Confusion matrix:
    No Yes class.error
No  22  68   0.7555556
Yes 25 203   0.1096491
```

Prediction

Hide

```
FourYG.rf.prediction.Less.Sem5 <- predict(FourYG.rf.Less.Sem5, testingLess.Sem5)
table(FourYG.rf.prediction.Less.Sem5, testingLess.Sem5$FourYG)
```

```
FourYG.rf.prediction.Less.Sem5 No Yes
                            No  10   5
                            Yes 12  51
```

Importance of variables.

Hide

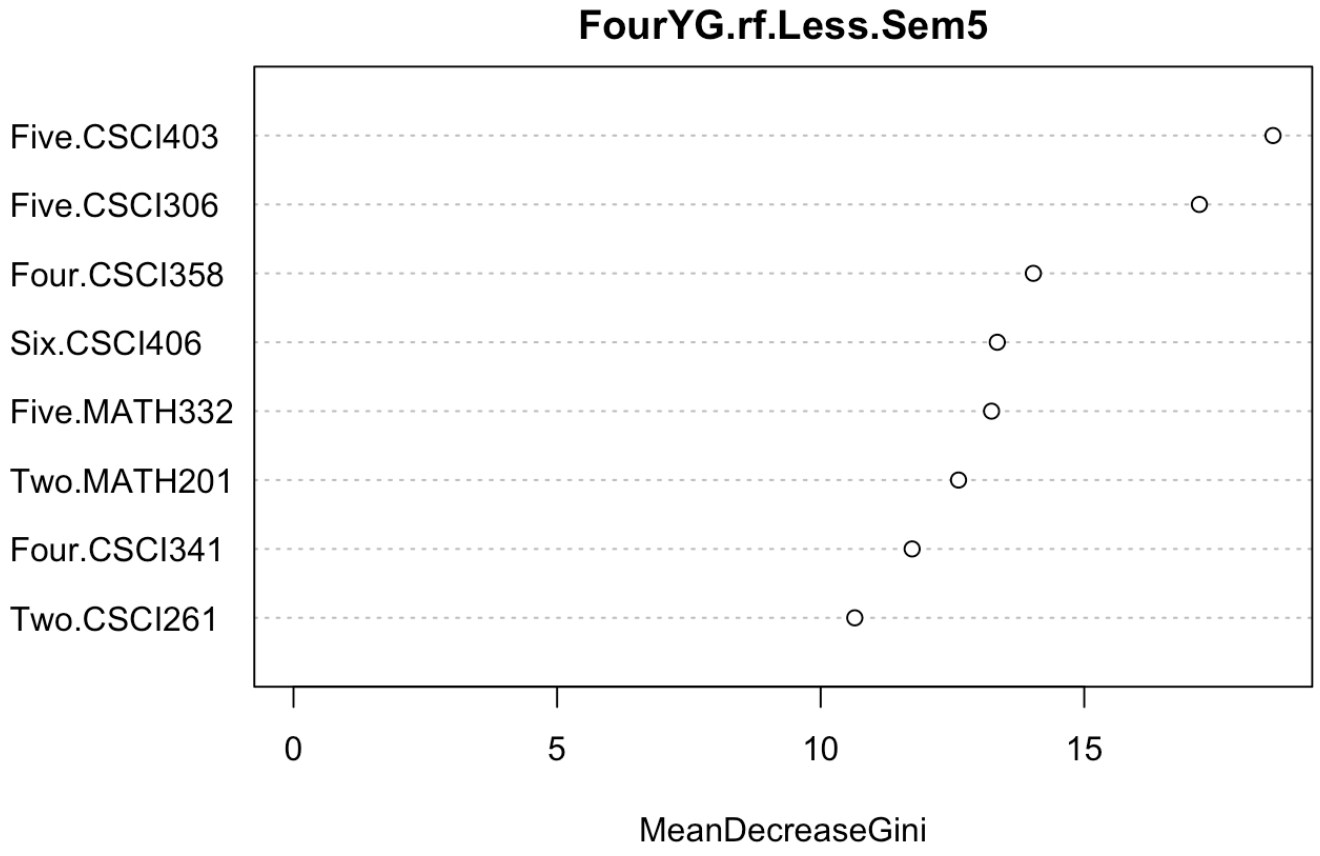
```
importance(FourYG.rf.Less.Sem5)
```

```
             MeanDecreaseGini
Two.CSCI261          10.64683
Two.MATH201          12.61523
Four.CSCI341         11.73470
Four.CSCI358         14.03414
Five.CSCI306         17.18344
Five.CSCI403         18.58088
Five.MATH332         13.24096
Six.CSCI406          13.35002
```

Plot of importance of Variabels.

Hide

```
varImpPlot(FourYG.rf.Less.Sem5)
```



Confusion Matrix

Hide

```
confusionMatrix(table(FourYG.rf.prediction.Less.Sem5, testingLess.Sem5$FourYG))
```

```
Confusion Matrix and Statistics


FourYG.rf.prediction.Less.Sem5 No Yes
                            No  10   5
                            Yes 12  51

               Accuracy : 0.7821
                 95% CI : (0.6741, 0.8676)
    No Information Rate : 0.7179
    P-Value [Acc > NIR] : 0.1274

                  Kappa : 0.4043
 Mcnemar's Test P-Value : 0.1456

            Sensitivity : 0.4545
            Specificity : 0.9107
         Pos Pred Value : 0.6667
         Neg Pred Value : 0.8095
             Prevalence : 0.2821
         Detection Rate : 0.1282
   Detection Prevalence : 0.1923
      Balanced Accuracy : 0.6826

       'Positive' Class : No
```

Logistic Regression method for variable importance

Hide

```
# Template code
# Step 1: Build Logit Model on Training Dataset
FourYG.lr.Less.Sem5 <- glm(FourYG ~Two.CSCI261+Two.MATH201+Four.CSCI341+Four.CSCI358+
Five.CSCI306+Five.CSCI403+Five.MATH332+Six.CSCI406, family= "binomial", data = traini
ngLess.Sem5)
FourYG.lr.Less.Sem5
```

```
Call:  glm(formula = FourYG ~ Two.CSCI261 + Two.MATH201 + Four.CSCI341 +
    Four.CSCI358 + Five.CSCI306 + Five.CSCI403 + Five.MATH332 +
    Six.CSCI406, family = "binomial", data = trainingLess.Sem5)

Coefficients:
 (Intercept)   Two.CSCI261   Two.MATH201  Four.CSCI341  Four.CSCI358  Five.CSCI306  F
ive.CSCI403  Five.MATH332   Six.CSCI406
    -5.78757       0.33388       0.06878      -0.16941       0.36688       0.70112
0.62001      -0.16263       0.18636

Degrees of Freedom: 317 Total (i.e. Null);  309 Residual
Null Deviance:      378.9
Residual Deviance: 331.7    AIC: 349.7
```

Hide

```
# Step 2: Predict Y on Test Dataset
predicted.lr.Less.Sem5 <- predict(FourYG.lr.Less.Sem5, testingLess.Sem5, type="respon
se")
```

Variable Importance

Hide

```
gbmImp.Less.Sem5 <- varImp(FourYG.rf.Less.Sem5, scale = FALSE)
gbmImp.Less.Sem5
```

	Overall <dbl>
Two.CSCI261	10.64683
Two.MATH201	12.61523
Four.CSCI341	11.73470
Four.CSCI358	14.03414
Five.CSCI306	17.18344
Five.CSCI403	18.58088
Five.MATH332	13.24096
Six.CSCI406	13.35002

8 rows