Code ▾

Model with Less Variables

Vanessa Gonzalez

2018-06-24

Random Forest Model using less variables

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/libr
ary")
```

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")
LessVariablesSet <- dfDataSet4YGImpute
# Creates Data Partitions and removes variables
inTrainingLess <- createDataPartition(LessVariablesSet$FourYG, p = 0.80, list = FALSE
)
LessVariablesSet <- LessVariablesSet[-(2:3)]
LessVariablesSet <- LessVariablesSet[-(3)]
LessVariablesSet <- LessVariablesSet[-(5)]
LessVariablesSet <- LessVariablesSet[-(7)]
LessVariablesSet <- LessVariablesSet[-(11)]
# Creates Training data Set
trainingLess <- LessVariablesSet[inTrainingLess, ]
# Creates Testing data Set
testingLess <- LessVariablesSet[-inTrainingLess, ]
# Data Set with less varialbes
head(LessVariablesSet)
```

Fou…	Two.CSCI261	Two.MATH…	Three.CSCI262	Four.CSCI341	Four.CSCI358	Five.CSCI3

	<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<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 12 columns

Hide

```
trainingLess
```

	Fou… <fctr>	Two.CSCI261 <dbl>	Two.MATH… <dbl>	Three.CSCI262 <dbl>	Four.CSCI341 <dbl>	Four.CSCI358 <dbl>	Five.CS(
1	Yes	4.000000	3.000000	3.000000	2.000000	4.000000	3.00
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 12 columns Previous **1** 2 3 4 5 6 ... 32 Next

Hide

```
testingLess
```

	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
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
30	Yes	4.000000	4.000000	4.000000	4.000000	4.000000	4.00
41	Yes	4.000000	2.000000	4.000000	3.000000	2.000000	3.00
43	No	4.000000	2.000000	0.300000	2.266172	1.000000	3.09
53	Yes	4.000000	2.502048	3.775148	3.162417	3.285296	3.46
61	No	0.500000	2.245680	2.873061	1.888522	2.412753	2.94
73	Yes	3.000000	3.000000	4.000000	4.000000	3.000000	4.00
78	No	4.000000	4.000000	4.000000	4.000000	4.000000	4.00

1-10 of 78 rows | 1-8 of 12 columns Previous **1** 2 3 4 5 6 ... 8 Next

Regresion Partition with method “class” for set with less variables

Hide

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

```
n= 318

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

 1) root 318 90 Yes (0.2830189 0.7169811)
   2) Eight.CSCI442< 1.862643 21  4 No (0.8095238 0.1904762) *
   3) Eight.CSCI442>=1.862643 297 73 Yes (0.2457912 0.7542088)
     6) Five.CSCI403< 3.501903 55 25 Yes (0.4545455 0.5454545)
      12) Five.MATH332< 2.071729 29 10 No (0.6551724 0.3448276) *
      13) Five.MATH332>=2.071729 26  6 Yes (0.2307692 0.7692308) *
     7) Five.CSCI403>=3.501903 242 48 Yes (0.1983471 0.8016529)
      14) Four.CSCI358< 3.303554 133 35 Yes (0.2631579 0.7368421)
        28) Four.CSCI358>=3.092642 8  3 No (0.6250000 0.3750000) *
        29) Four.CSCI358< 3.092642 125 30 Yes (0.2400000 0.7600000)
          58) Two.CSCI261< 2.85 9  4 No (0.5555556 0.4444444) *
          59) Two.CSCI261>=2.85 116 25 Yes (0.2155172 0.7844828) *
      15) Four.CSCI358>=3.303554 109 13 Yes (0.1192661 0.8807339) *
```

Hide

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

```


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

Variables actually used in tree construction:
[1] Eight.CSCI442 Five.CSCI403  Five.MATH332  Four.CSCI358  Two.CSCI261

Root node error: 90/318 = 0.28302

n= 318

        CP nsplit rel error  xerror     xstd
1 0.144444      0   1.00000 1.00000 0.089255
2 0.050000      1   0.85556 0.96667 0.088330
3 0.011111      3   0.75556 0.95556 0.088011
4 0.010000      6   0.72222 1.05556 0.090690
```

Create a summary for the data set with less variables

Hide

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

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

          CP nsplit rel error    xerror       xstd
1 0.14444444      0 1.0000000 1.0000000 0.08925501
2 0.05000000      1 0.8555556 0.9666667 0.08833027
3 0.01111111      3 0.7555556 0.9555556 0.08801104
4 0.01000000      6 0.7222222 1.0555556 0.09068974

Variable importance
Eight.CSCI442  Four.CSCI358  Five.MATH332  Five.CSCI403   Six.CSCI406 Eight.CSCI400 
Four.CSCI341   Two.CSCI261   Two.MATH201 
           29            16            13            13             7             5 
4             4             4 
 Five.CSCI306 Three.CSCI262 
            2             2 

Node number 1: 318 observations,    complexity param=0.1444444
  predicted class=Yes  expected loss=0.2830189  P(node) =1
    class counts:    90   228
   probabilities: 0.283 0.717 
  left son=2 (21 obs) right son=3 (297 obs)
  Primary splits:
      Eight.CSCI442 < 1.862643 to the left,  improve=12.465940, (0 missing)
      Two.MATH201   < 2.361973 to the left,  improve=10.239170, (0 missing)
      Six.CSCI406   < 2.266458 to the left,  improve=10.222070, (0 missing)
      Five.CSCI403  < 3.419795 to the left,  improve= 9.568449, (0 missing)
      Eight.CSCI400 < 3.215149 to the left,  improve= 9.298103, (0 missing)
  Surrogate splits:
      Four.CSCI358  < 0.85     to the left,  agree=0.940, adj=0.095, (0 split)
      Eight.CSCI400 < 0.85     to the left,  agree=0.937, adj=0.048, (0 split)

Node number 2: 21 observations
  predicted class=No   expected loss=0.1904762  P(node) =0.06603774
    class counts:    17     4
   probabilities: 0.810 0.190 

Node number 3: 297 observations,    complexity param=0.05
  predicted class=Yes  expected loss=0.2457912  P(node) =0.9339623
    class counts:    73   224
   probabilities: 0.246 0.754 
  left son=6 (55 obs) right son=7 (242 obs)
  Primary splits:
      Five.CSCI403  < 3.501903 to the left,  improve=5.883073, (0 missing)
      Four.CSCI358  < 3.386122 to the left,  improve=5.621907, (0 missing)
      Six.CSCI406   < 2.468042 to the left,  improve=4.832347, (0 missing)
      Five.CSCI306  < 3.345445 to the left,  improve=4.796963, (0 missing)
      Eight.CSCI400 < 2.981697 to the left,  improve=4.720930, (0 missing)
  Surrogate splits:
      Eight.CSCI400 < 2.50993  to the left,  agree=0.845, adj=0.164, (0 split)
      Three.CSCI262 < 2.853189 to the left,  agree=0.838, adj=0.127, (0 split)
      Six.CSCI406   < 1.968216 to the left,  agree=0.838, adj=0.127, (0 split)
      Four.CSCI341  < 0.85     to the left,  agree=0.828, adj=0.073, (0 split)
      Five.CSCI306  < 2.15     to the left,  agree=0.825, adj=0.055, (0 split)

Node number 6: 55 observations,    complexity param=0.05
  predicted class=Yes  expected loss=0.4545455  P(node) =0.172956
    class counts:    25    30
   probabilities: 0.455 0.545 
  left son=12 (29 obs) right son=13 (26 obs)
  Primary splits:
      Five.MATH332 < 2.071729 to the left,  improve=4.938510, (0 missing)
      Five.CSCI306 < 3.716578 to the left,  improve=4.446640, (0 missing)
      Four.CSCI358 < 2.344163 to the left,  improve=3.108641, (0 missing)
      Two.MATH201  < 2.780642 to the left,  improve=2.424242, (0 missing)
      Six.CSCI406  < 2.446701 to the left,  improve=2.259286, (0 missing)
  Surrogate splits:
      Six.CSCI406   < 2.133311 to the left,  agree=0.709, adj=0.385, (0 split)
      Four.CSCI341  < 2.502655 to the left,  agree=0.673, adj=0.308, (0 split)
      Four.CSCI358  < 2.344163 to the left,  agree=0.673, adj=0.308, (0 split)
      Two.MATH201   < 2.210063 to the left,  agree=0.618, adj=0.192, (0 split)
      Eight.CSCI442 < 2.534753 to the right, agree=0.618, adj=0.192, (0 split)

Node number 7: 242 observations,    complexity param=0.01111111
  predicted class=Yes  expected loss=0.1983471  P(node) =0.7610063
    class counts:    48   194
   probabilities: 0.198 0.802 
  left son=14 (133 obs) right son=15 (109 obs)
  Primary splits:
      Four.CSCI358  < 3.303554 to the left,  improve=2.480648, (0 missing)
      Eight.CSCI400 < 1.85     to the left,  improve=2.006702, (0 missing)
      Three.CSCI262 < 2.429332 to the left,  improve=1.505686, (0 missing)
      Two.CSCI261   < 2.85     to the left,  improve=1.203605, (0 missing)
      Two.MATH201   < 2.339399 to the left,  improve=1.052872, (0 missing)
  Surrogate splits:
      Five.MATH332  < 3.355349 to the left,  agree=0.727, adj=0.394, (0 split)
      Two.MATH201   < 3.3      to the left,  agree=0.711, adj=0.358, (0 split)
      Five.CSCI306  < 3.72575  to the left,  agree=0.674, adj=0.275, (0 split)
      Six.CSCI406   < 3.064793 to the left,  agree=0.674, adj=0.275, (0 split)
      Eight.CSCI400 < 3.326161 to the left,  agree=0.669, adj=0.266, (0 split)

Node number 12: 29 observations
  predicted class=No   expected loss=0.3448276  P(node) =0.09119497
```

```
    class counts:    19    10
   probabilities: 0.655 0.345 

Node number 13: 26 observations
  predicted class=Yes  expected loss=0.2307692  P(node) =0.08176101
    class counts:     6    20
   probabilities: 0.231 0.769 

Node number 14: 133 observations,    complexity param=0.01111111
  predicted class=Yes  expected loss=0.2631579  P(node) =0.418239
    class counts:    35    98
   probabilities: 0.263 0.737 
  left son=28 (8 obs) right son=29 (125 obs)
  Primary splits:
      Four.CSCI358 < 3.092642 to the right, improve=2.2289470, (0 missing)
      Five.CSCI403 < 3.728962 to the right, improve=1.7944060, (0 missing)
      Two.CSCI261  < 2.85     to the left,  improve=1.6506320, (0 missing)
      Five.MATH332 < 2.36125  to the right, improve=1.2432230, (0 missing)
      Two.MATH201  < 3.15     to the right, improve=0.8300801, (0 missing)

Node number 15: 109 observations
  predicted class=Yes  expected loss=0.1192661  P(node) =0.3427673
    class counts:    13    96
   probabilities: 0.119 0.881 

Node number 28: 8 observations
  predicted class=No   expected loss=0.375  P(node) =0.02515723
    class counts:     5     3
   probabilities: 0.625 0.375 

Node number 29: 125 observations,    complexity param=0.01111111
  predicted class=Yes  expected loss=0.24  P(node) =0.3930818
    class counts:    30    95
   probabilities: 0.240 0.760 
  left son=58 (9 obs) right son=59 (116 obs)
  Primary splits:
      Two.CSCI261   < 2.85     to the left,  improve=1.9314180, (0 missing)
      Five.CSCI403  < 3.801474 to the right, improve=1.4580860, (0 missing)
      Five.MATH332  < 2.415702 to the right, improve=1.0666670, (0 missing)
      Five.CSCI306  < 3.005169 to the left,  improve=0.6272232, (0 missing)
      Eight.CSCI442 < 2.546135 to the right, improve=0.5891068, (0 missing)

Node number 58: 9 observations
  predicted class=No   expected loss=0.4444444  P(node) =0.02830189
    class counts:     5     4
   probabilities: 0.556 0.444 
```

```
Node number 59: 116 observations
  predicted class=Yes  expected loss=0.2155172  P(node) =0.3647799
    class counts:    25    91
   probabilities: 0.216 0.784
```

Prediction

Hide

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

```
     predictionsLess
      No Yes
  No  10  12
  Yes  9  47
```

Confusion Matrix

Hide

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

```
Confusion Matrix and Statistics


predictionsLess No Yes
            No  10   9
            Yes 12  47

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

                  Kappa : 0.3065
 Mcnemar's Test P-Value : 0.6625

            Sensitivity : 0.4545
            Specificity : 0.8393
         Pos Pred Value : 0.5263
         Neg Pred Value : 0.7966
             Prevalence : 0.2821
         Detection Rate : 0.1282
   Detection Prevalence : 0.2436
      Balanced Accuracy : 0.6469

       'Positive' Class : No
```

Hide

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

```
[1] 0.9555556
```

Hide

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

```
3
3
```

Get the cost complecity parameter of the record

Hide

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

```
[1] 0.02777778
```

Hide

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

Hide

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

```
     predictionsLessPrune
      No Yes
  No   7  15
  Yes  4  52
```

Hide

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

```
Confusion Matrix and Statistics


predictionsLessPrune 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
```

Top 10 variables

Hide

```
str(LessVariablesSet)
```

```
'data.frame':	396 obs. of  12 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 ...
 $ Eight.CSCI400: num  3.3 3.58 3.34 3 4 ...
 $ Eight.CSCI442: num  2.3 2.9 3.23 3 4 ...
```

Random Forest model with les variables

Hide

```
FourYG.rf.Less <- randomForest(FourYG ~Two.CSCI261+Two.MATH201+Four.CSCI341+Four.CSCI
358+Five.CSCI306+Five.CSCI403+Five.MATH332+Six.CSCI406+Eight.CSCI400+Eight.CSCI442  ,
data = trainingLess)
FourYG.rf.Less
```

```


Call:
 randomForest(formula = FourYG ~ Two.CSCI261 + Two.MATH201 + Four.CSCI341 +      Four
.CSCI358 + Five.CSCI306 + Five.CSCI403 + Five.MATH332 +      Six.CSCI406 + Eight.CSCI
400 + Eight.CSCI442, data = trainingLess)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 27.36%
Confusion matrix:
    No Yes class.error
No  26  64   0.7111111
Yes 23 205   0.1008772
```

Hide

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

```
FourYG.rf.prediction.Less No Yes
                      No  10   4
                      Yes 12  52
```

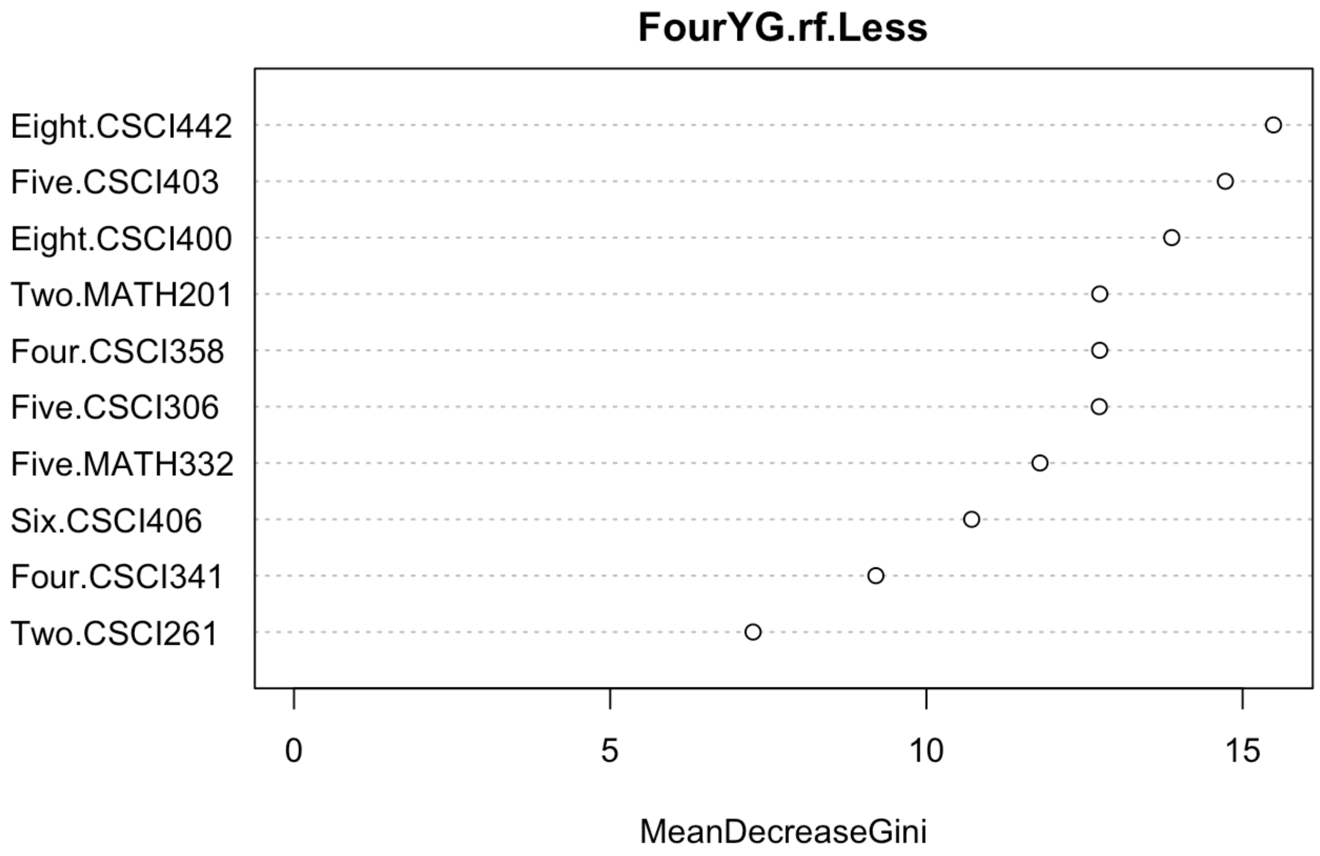
Hide

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

```
              MeanDecreaseGini
Two.CSCI261           7.459634
Two.MATH201          12.278136
Four.CSCI341          9.084376
Four.CSCI358         12.814361
Five.CSCI306         12.923026
Five.CSCI403         14.286177
Five.MATH332         11.681177
Six.CSCI406          10.813423
Eight.CSCI400        13.245431
Eight.CSCI442        16.094796
```

Hide

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



Hide

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

```
Confusion Matrix and Statistics


FourYG.rf.prediction.Less No Yes
                      No  10   4
                      Yes 12  52
                                          
               Accuracy : 0.7949          
                 95% CI : (0.6884, 0.878) 
    No Information Rate : 0.7179          
    P-Value [Acc > NIR] : 0.08014         
                                          
                  Kappa : 0.4307          
 Mcnemar's Test P-Value : 0.08012         
                                          
            Sensitivity : 0.4545          
            Specificity : 0.9286          
         Pos Pred Value : 0.7143          
         Neg Pred Value : 0.8125          
             Prevalence : 0.2821          
         Detection Rate : 0.1282          
   Detection Prevalence : 0.1795          
      Balanced Accuracy : 0.6916          
                                          
       'Positive' Class : No              
```

Logistic Regresion with less variables

Hide

```
# Template code
# Step 1: Build Logit Model on Training Dataset
FourYG.lr.Less <- glm(FourYG ~Two.CSCI261+Two.MATH201+Four.CSCI341+Four.CSCI358+Five.
CSCI306+Five.CSCI403+Five.MATH332+Six.CSCI406+Eight.CSCI400+Eight.CSCI442, family= "b
inomial", data = trainingLess)
FourYG.lr.Less
```

```

Call:  glm(formula = FourYG ~ Two.CSCI261 + Two.MATH201 + Four.CSCI341 + 
    Four.CSCI358 + Five.CSCI306 + Five.CSCI403 + Five.MATH332 + 
    Six.CSCI406 + Eight.CSCI400 + Eight.CSCI442, family = "binomial", 
    data = trainingLess)

Coefficients:
  (Intercept)    Two.CSCI261    Two.MATH201   Four.CSCI341   Four.CSCI358   Five.CSCI
306   Five.CSCI403   Five.MATH332  
     -5.14326        0.18486        0.15821       -0.24165        0.46495        0.35
388        0.47539       -0.11963  
  Six.CSCI406  Eight.CSCI400  Eight.CSCI442  
      0.26569        0.03466        0.26318  

Degrees of Freedom: 317 Total (i.e. Null);  307 Residual
Null Deviance:      378.9 
Residual Deviance: 330.3    AIC: 352.3
```

Hide

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

Variable Importance

Hide

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

	Overall <dbl>
Two.CSCI261	7.459634
Two.MATH201	12.278136
Four.CSCI341	9.084376
Four.CSCI358	12.814361
Five.CSCI306	12.923026
Five.CSCI403	14.286177
Five.MATH332	11.681177
Six.CSCI406	10.813423

Eight.CSCI400	13.245431
Eight.CSCI442	16.094796

1-10 of 10 rows