Predicting 6 Year Graduation at NJCU

FINAL R PROJECT FINC614

Julian Garcia

December 6, 2017

A

```
library(readr)
library(dplyr)

##
## Attaching package: 'dplyr'

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

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

library(ggplot2)
library(rJava)
library(randomForest)

## Warning: package 'randomForest' was built under R version 3.4.3

## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

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

library(FSelector)
library(caret)

## Loading required package: lattice
```

```
library(e1071)
library(rattle)

## Rattle: A free graphical interface for data science with R.
## Version 5.1.0 Copyright (c) 2006-2017 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.

##
## Attaching package: 'rattle'

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

library(rpart.plot)

## Loading required package: rpart

library(RColorBrewer)
library(klaR)

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

library(ROCR)

## Loading required package: gplots

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

library(pROC)

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.4.3

library(labeling)
library(reshape2)
FALL2010 <- read_excel("C:/Users/jgarcia/Desktop/school stuff/FALL2010FTFTPRO
JECTFINAL10.xlsx")
```

B Data Description

The dataset I chose is student data dealing with 2 cohorts from the years 2010 and 2011 that I built from snapshots we capture within the office on Institiutional Effectiveness here at NJCU and live financial aid information stored on the schools system. I loaded a variety of pre-college characteristics as follows: Birth-Date (BirthDate), High School Percentile (hspctl), Math SAT, Critical Reading SAT and the combined total SAT Scores(msat, crsat, and totnusat). In addition I also included entering perfomance based and financial aid characteristics from snapshot records and from the FAFSA (Free Application for Federal Student Aid) they are as follows: Credits enrolled (creden), entering term (Term), Ethnicity(Nues), Basic Skills test scores (BASICENGL, BASICREAD, BASICARIT and BASICELAG), Declaration of Major (PlannedorDeclared), Transfer Credits (TotalTrnsfr), Test Credits (TotalTest), student/parent income (AdjustedGrossIncome and ParentAdjustedGrossIncome), financial need (need), housing status (Housing) grants (grantactual), scholarships (scholactual), loans (loanactual), work-study (workstudyactual). Also included are end of term and end of year one and two characteristics such as: cumulative GPA after first term (CUMGPA), Credits passed after first term (TermPassed), total credits passed/transfered/tested (TotalUnits), End of first Year credits and GPA (TotalPassedEOY and CUMGPAEOY), End of second Year credits and GPA (TotalPassedEOYT and CUMGPAEOYT) and 1st year retention (retained).

C Questions

What characteristics impact retention?

What kinds of correlations exist between pre-college characteristics and student performance?

What factors can help predict a 6 year graduate?

D

a

```
View(FALL2010)
sum(is.na(FALL2010))
```

```
## [1] 1154
```

D

b

Removed Nulls

Replaced negative values from need and parent adjusted gross income wth 0's

```
FALL2010 <- na.omit(FALL2010)
FALL2010$ParentAdjustedGrossIncome <- ifelse(FALL2010$ParentAdjustedGrossInco
me< 0, 0, FALL2010$ParentAdjustedGrossIncome)
FALL2010$need <- ifelse(FALL2010$need< 0, 0, FALL2010$need)
```

D

C

```
summary(FALL2010)
```

```
##     emplid              Term               creden          hspctl
##  Length:1007        Length:1007        Min.   :10.00   Min.   :0.0000
##  Class :character   Class :character   1st Qu.:13.00   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :14.00   Median :0.0000
##                                        Mean   :14.22   Mean   :0.3054
##                                        3rd Qu.:15.00   3rd Qu.:0.6550
##                                        Max.   :19.00   Max.   :1.0000
##       msat           crsat           Nues              totnusat
##  Min.   :20.00   Min.   :20.00   Length:1007        Min.   : 49.0
##  1st Qu.:41.00   1st Qu.:39.00   Class :character   1st Qu.: 82.0
##  Median :45.00   Median :43.00   Mode  :character   Median : 88.0
##  Mean   :45.54   Mean   :43.76                      Mean   : 89.3
##  3rd Qu.:50.00   3rd Qu.:48.00                      3rd Qu.: 96.0
##  Max.   :72.00   Max.   :76.00                      Max.   :130.0
##   sixyrGRAD          BASICENGL       BASICREAD        BASICARIT
##  Length:1007        Min.   :0.000   Min.   :  0.00   Min.   :  0.00
##  Class :character   1st Qu.:2.000   1st Qu.: 62.00   1st Qu.: 43.00
##  Mode  :character   Median :2.000   Median : 74.00   Median : 69.00
##                     Mean   :2.521   Mean   : 73.37   Mean   : 67.57
##                     3rd Qu.:3.000   3rd Qu.: 87.00   3rd Qu.: 91.00
##                     Max.   :4.000   Max.   :149.00   Max.   :167.00
##    BASICELAG      PlannedorDeclared    TermPassed  TotalTrnsfr
##  Min.   :  0.00   Length:1007        Min.   : 0   Min.   : 0.0000
##  1st Qu.:  0.00   Class :character   1st Qu.:10   1st Qu.: 0.0000
##  Median : 35.00   Mode  :character   Median :13   Median : 0.0000
##  Mean   : 38.66                      Mean   :12   Mean   : 0.3188
##  3rd Qu.: 74.00                      3rd Qu.:15   3rd Qu.: 0.0000
##  Max.   :161.00                      Max.   :19   Max.   :66.0000
##    TotalTest         TotalUnits        CUMGPA
##  Min.   : 0.0000   Min.   : 0.00   Min.   :0.000
```

```
##  1st Qu.: 0.0000   1st Qu.:11.00   1st Qu.:2.430
##  Median : 0.0000   Median :13.00   Median :3.000
##  Mean   : 0.2393   Mean   :13.05   Mean   :2.785
##  3rd Qu.: 0.0000   3rd Qu.:15.50   3rd Qu.:3.440
##  Max.   :12.0000   Max.   :84.00   Max.   :4.000
##    BirthDate                   AdjustedGrossIncome FatherHighestGradeLevel
##  Min.   :1984-07-30 00:00:00   Min.   :     0      Length:1007
##  1st Qu.:1992-02-21 12:00:00   1st Qu.:     0      Class :character
##  Median :1992-09-10 00:00:00   Median :     0      Mode  :character
##  Mean   :1992-08-12 13:23:39   Mean   :  1818
##  3rd Qu.:1993-05-01 00:00:00   3rd Qu.:     0
##  Max.   :1995-01-07 00:00:00   Max.   :234461
##  MotherHighestGradeLevel ParentAdjustedGrossIncome      need
##  Length:1007             Min.   :     0            Min.   :    0
##  Class :character        1st Qu.: 15180            1st Qu.:14498
##  Mode  :character        Median : 31019            Median :18286
##                          Mean   : 45074            Mean   :17857
##                          3rd Qu.: 60331            3rd Qu.:25962
##                          Max.   :298966            Max.   :35940
##    retained           Housing           grantactual     scholactual
##  Length:1007        Length:1007        Min.   :    0   Min.   :    0.0
##  Class :character   Class :character   1st Qu.:    0   1st Qu.:    0.0
##  Mode  :character   Mode  :character   Median : 7068   Median :    0.0
##                                        Mean   : 6488   Mean   :  782.3
##                                        3rd Qu.:11412   3rd Qu.:    0.0
##                                        Max.   :14349   Max.   :27631.9
##    loanactual    workstudyactual  TotalPassedEOY  TotalPassedEOYT
##  Min.   :    0   Min.   :   0.0   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:    0   1st Qu.:   0.0   1st Qu.:19.00   1st Qu.:25.00
##  Median : 1750   Median :   0.0   Median :25.00   Median :45.00
##  Mean   : 3096   Mean   : 155.3   Mean   :23.18   Mean   :40.55
##  3rd Qu.: 5500   3rd Qu.:   0.0   3rd Qu.:29.00   3rd Qu.:56.00
##  Max.   :23500   Max.   :3000.0   Max.   :41.00   Max.   :80.00
##    CUMGPAEOY       CUMGPAEOYT
##  Min.   :0.000   Min.   :0.000
##  1st Qu.:2.270   1st Qu.:2.190
##  Median :2.870   Median :2.760
##  Mean   :2.661   Mean   :2.576
##  3rd Qu.:3.350   3rd Qu.:3.250
##  Max.   :4.000   Max.   :4.000
```

```
FALL2010 %>% count(sixyrGRAD)
```

```
## # A tibble: 2 x 2
##   sixyrGRAD     n
##       <chr> <int>
## 1        no   657
## 2       yes   350
```

```
FALL2010 %>% count(retained)
```

```
## # A tibble: 2 x 2
##   retained     n
##      <chr> <int>
## 1       no   263
## 2      yes   744
```

E

Grouped SAT Scores, GPA's and Credit's passed to analyze retention and graduation based on performance and it is evident as you move up in the group it is more likely to be retained and to graduate within 6 years. Also created binomial fields based on predefined performance standards for gpa and credits earned in hopes to impove the logistic regression with some additional binomial chracteristics.

```
FALL2010 <- mutate(FALL2010, SATGROUP = as.numeric(cut(FALL2010$totnusat,5)))
FALL2010 <- mutate(FALL2010, GPAGROUP = as.numeric(cut(FALL2010$CUMGPA,4)))
FALL2010 <- mutate(FALL2010, GPAEOYGROUP = as.numeric( cut(FALL2010$CUMGPAEOY
,4)))
FALL2010 <- mutate(FALL2010, GPAEOYTGROUP = as.numeric( cut(FALL2010$CUMGPAEO
YT,4)))
FALL2010 <- mutate(FALL2010, TermPassedGROUP = as.numeric( cut(FALL2010$TermP
assed,6)))
FALL2010 <- mutate(FALL2010, TotalPassedEOYGROUP = as.numeric( cut(FALL2010$T
otalPassedEOY,6)))
FALL2010 <- mutate(FALL2010, TotalPassedEOYTGROUP = as.numeric( cut(FALL2010$
TotalPassedEOYT,6)))
FALL2010 <- mutate(FALL2010, EOYoverthirty = if_else( TotalPassedEOY>29, "yes
", "no" ))
FALL2010 <- mutate(FALL2010, EOYToversixty = if_else( TotalPassedEOYT>59, "ye
s", "no" ))
FALL2010 <- mutate(FALL2010, EOYgpagood = if_else( CUMGPAEOY>2.8, "yes", "no"
))
FALL2010 <- mutate(FALL2010, EOYTgpagood = if_else( CUMGPAEOYT>2.8, "yes", "n
o" ))

FALL2010 %>% count(GPAGROUP, sixyrGRAD)
```

```
## # A tibble: 8 x 3
##   GPAGROUP sixyrGRAD     n
##      <dbl>     <chr> <int>
## 1        1        no    70
## 2        1       yes     2
## 3        2        no    80
## 4        2       yes     7
## 5        3        no   264
## 6        3       yes    89
```

```
## 7        4        no   243
## 8        4       yes   252
```

```
FALL2010 %>% count(GPAGROUP, retained)
```

```
## # A tibble: 8 x 3
##   GPAGROUP retained     n
##      <dbl>    <chr> <int>
## 1        1       no    51
## 2        1      yes    21
## 3        2       no    40
## 4        2      yes    47
## 5        3       no    96
## 6        3      yes   257
## 7        4       no    76
## 8        4      yes   419
```

```
FALL2010 %>% count(GPAEOYGROUP, sixyrGRAD)
```

```
## # A tibble: 7 x 3
##   GPAEOYGROUP sixyrGRAD     n
##         <dbl>     <chr> <int>
## 1           1        no    80
## 2           2        no   108
## 3           2       yes     5
## 4           3        no   276
## 5           3       yes   108
## 6           4        no   193
## 7           4       yes   237
```

```
FALL2010 %>% count(GPAEOYGROUP, retained)
```

```
## # A tibble: 8 x 3
##   GPAEOYGROUP retained     n
##         <dbl>    <chr> <int>
## 1           1       no    67
## 2           1      yes    13
## 3           2       no    53
## 4           2      yes    60
## 5           3       no    88
## 6           3      yes   296
## 7           4       no    55
## 8           4      yes   375
```

```
FALL2010 %>% count(GPAEOYTGROUP, sixyrGRAD)
```

```
## # A tibble: 7 x 3
##   GPAEOYTGROUP sixyrGRAD     n
##          <dbl>     <chr> <int>
## 1            1        no    83
## 2            2        no   132
## 3            2       yes     4
```

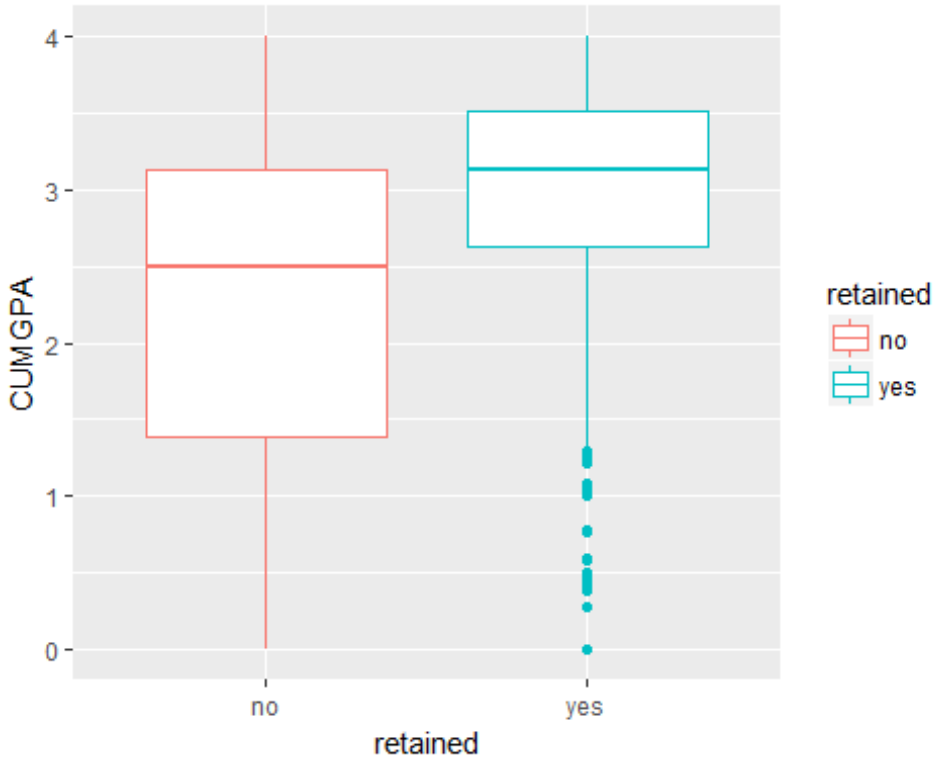
```
## 4              3         no   295
## 5              3        yes   113
## 6              4         no   147
## 7              4        yes   233
```

D

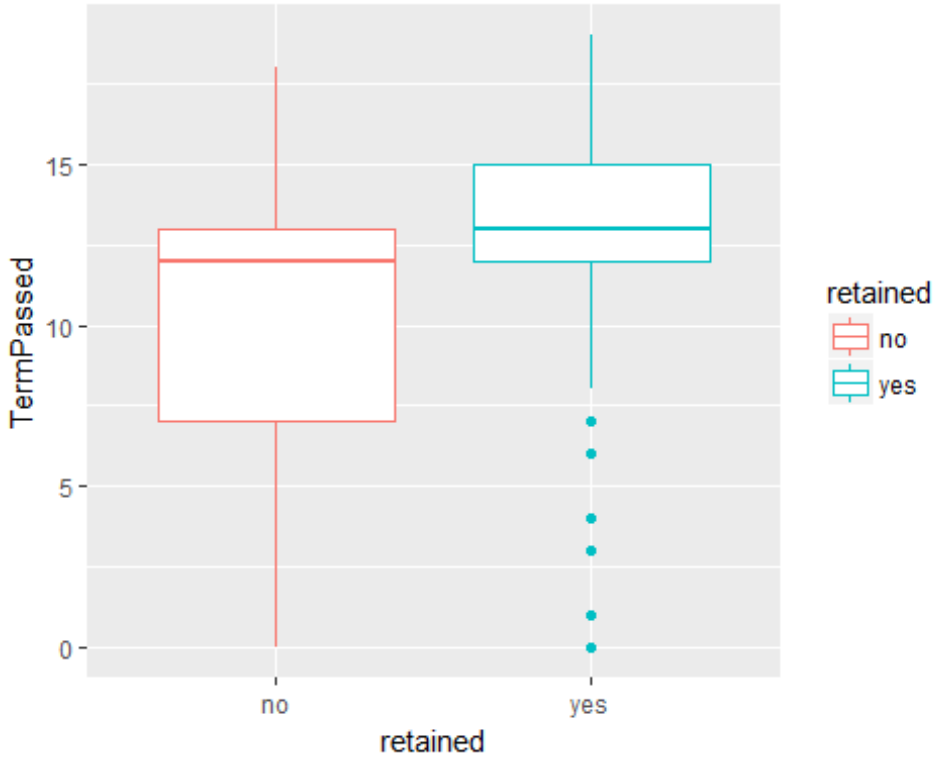
d-1

From these box plots we can see the different impacts certain pre college characteristics have on retention in comparison to first semester performance characteristics. It is evident that neither math or critical reading sat scores provide any distinction between the groups that are retained or graduated where as students with higher gpas and more term credits passed after their first semester tend to be retained and/or graduate within six years. The strongest distinction amongst students who are retained is Term Credits passed.

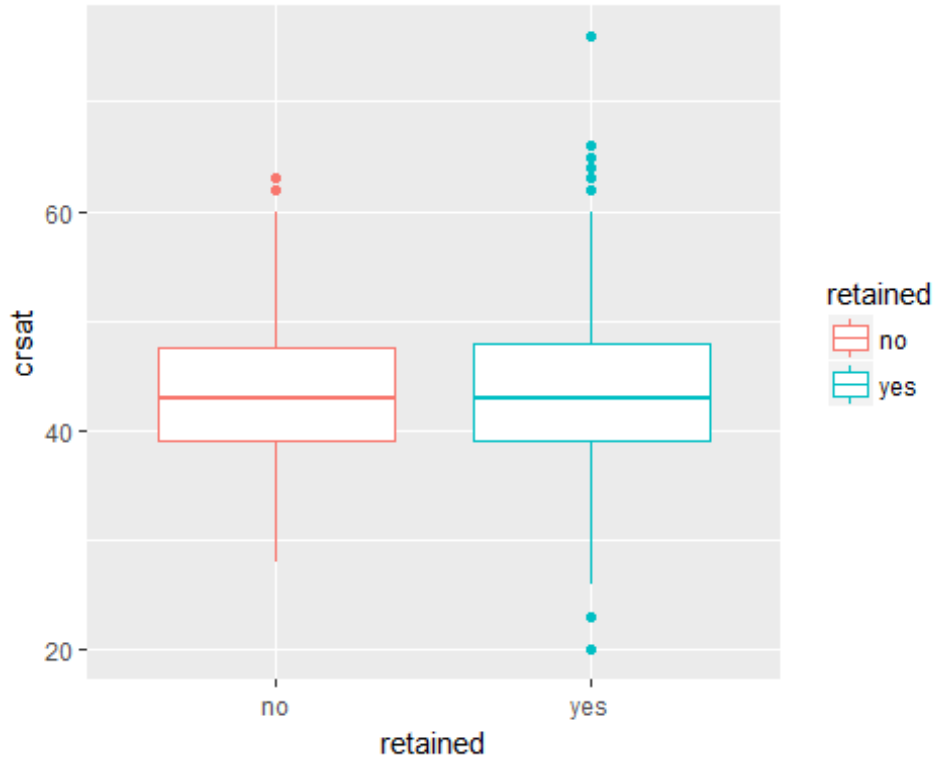
```
ggplot(FALL2010, aes(retained, CUMGPA, color = retained)) + geom_boxplot()
```



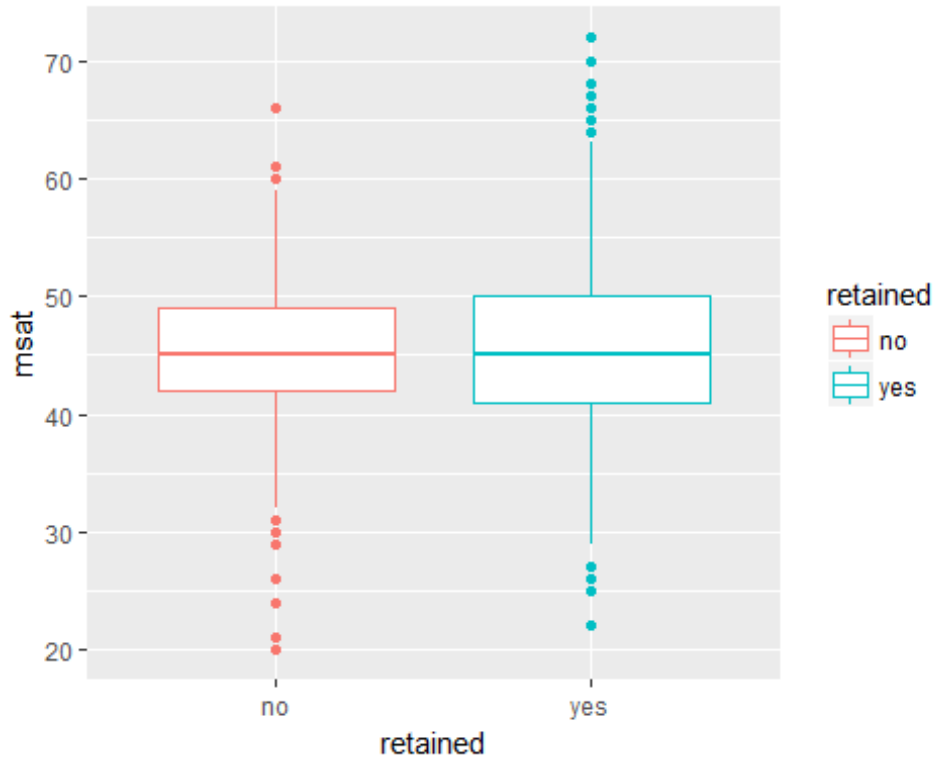

```
ggplot(FALL2010, aes(retained, TermPassed, color = retained)) + geom_boxplot(
)
```




```
ggplot(FALL2010, aes(retained, crsat, color = retained)) + geom_boxplot()
```




```
ggplot(FALL2010, aes(retained, msat, color = retained)) + geom_boxplot()
```

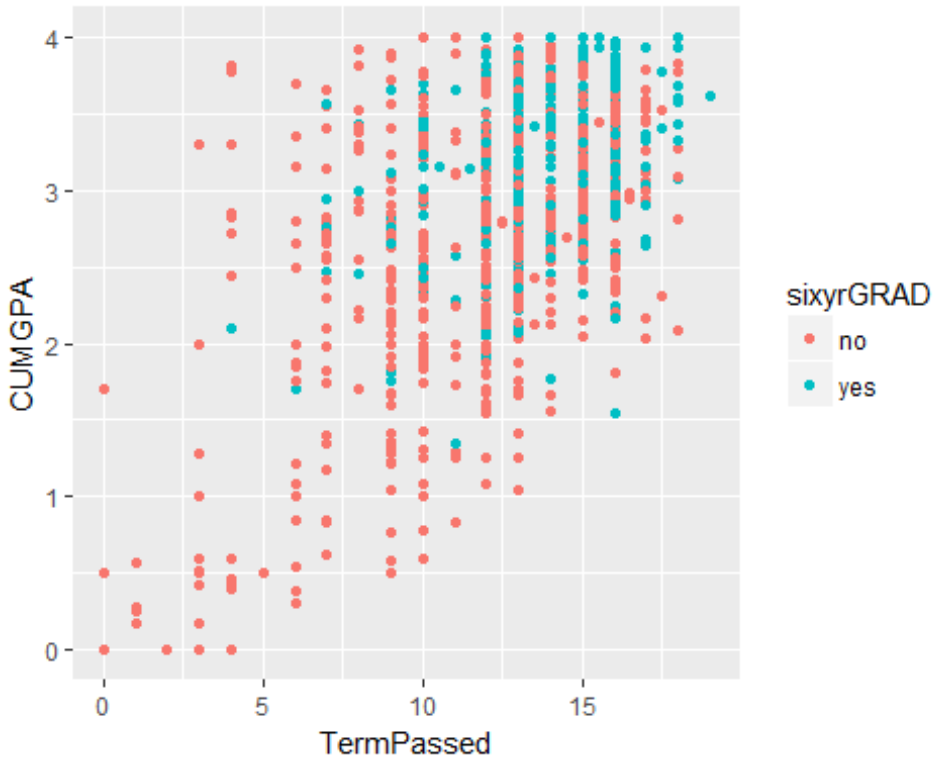


D

d-2

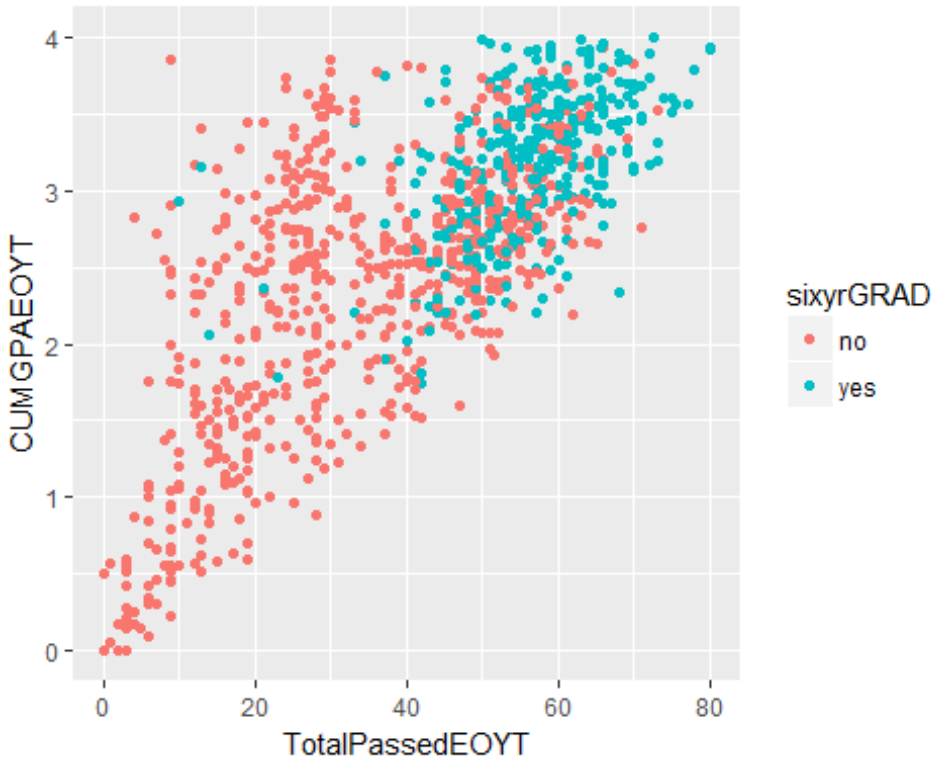
This scatter plot shows some grouping of succesful 6 year graduates in the upper right corner where students have the greatest combination of cumulative gpa and first term credits passed. I produced three scatter plots showing cumulative GPA and total credits earned at 3 different points in the students career and it is evident that these groups begin to cluster even more as they progress into their academic careers.

```
ggplot(data=FALL2010, aes(TermPassed, CUMGPA, color=sixyrGRAD)) + geom_point(
)
```




```
ggplot(data=FALL2010, aes(TotalPassedEOY, CUMGPAEOY, color=sixyrGRAD)) + geom
_point()
```

```
ggplot(data=FALL2010, aes(TotalPassedEOYT, CUMGPAEOYT, color=sixyrGRAD)) + ge
om_point()
```

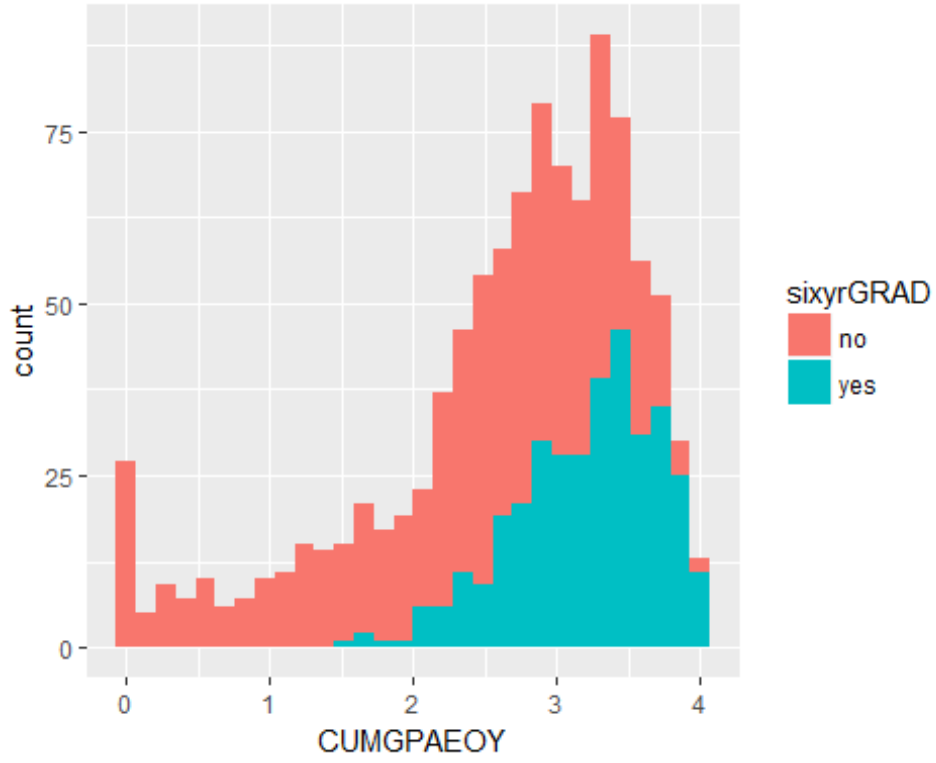


D

d-3

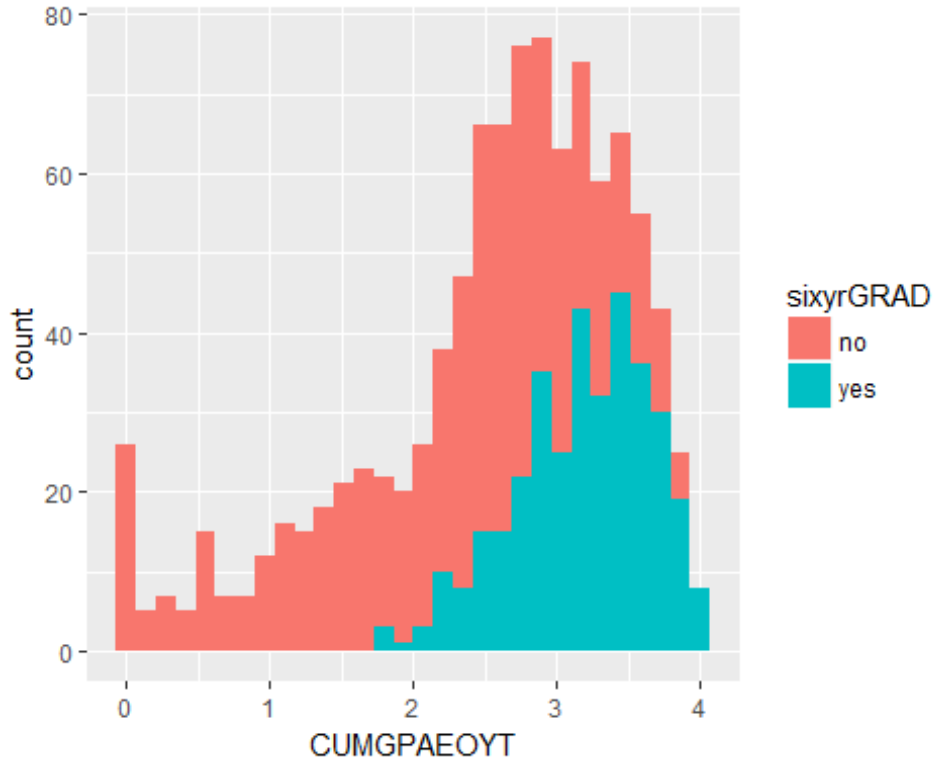
From this histogram we can tell again that as cumulative gpa goes up so do the amount of students who graduate in six years in the second histogram we can see the same goes for total credits earned by the end of the second year and there seems to be a real sweet spot for students who stay on target and accumulate 60 credits by the end of their second year. Further proving that target credit accumulation is the biggest impact on student graduation.

```
ggplot(data=FALL2010, aes(CUMGPAEOY, fill = sixyrGRAD)) + geom_histogram()

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

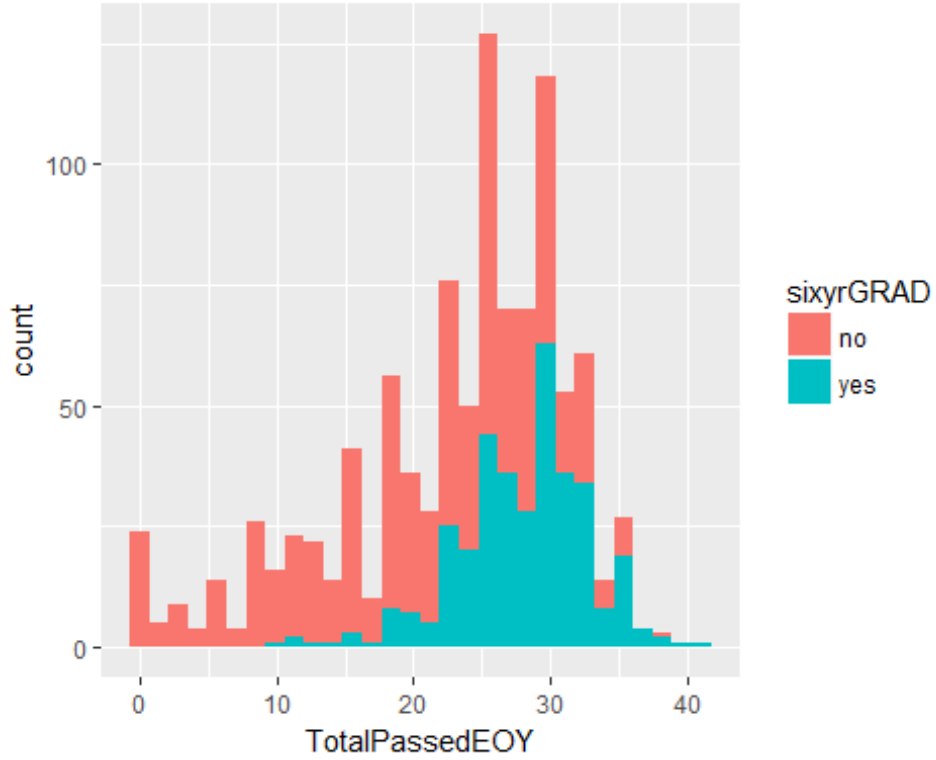



```
ggplot(data=FALL2010, aes(CUMGPAEOYT, fill = sixyrGRAD)) + geom_histogram()
```

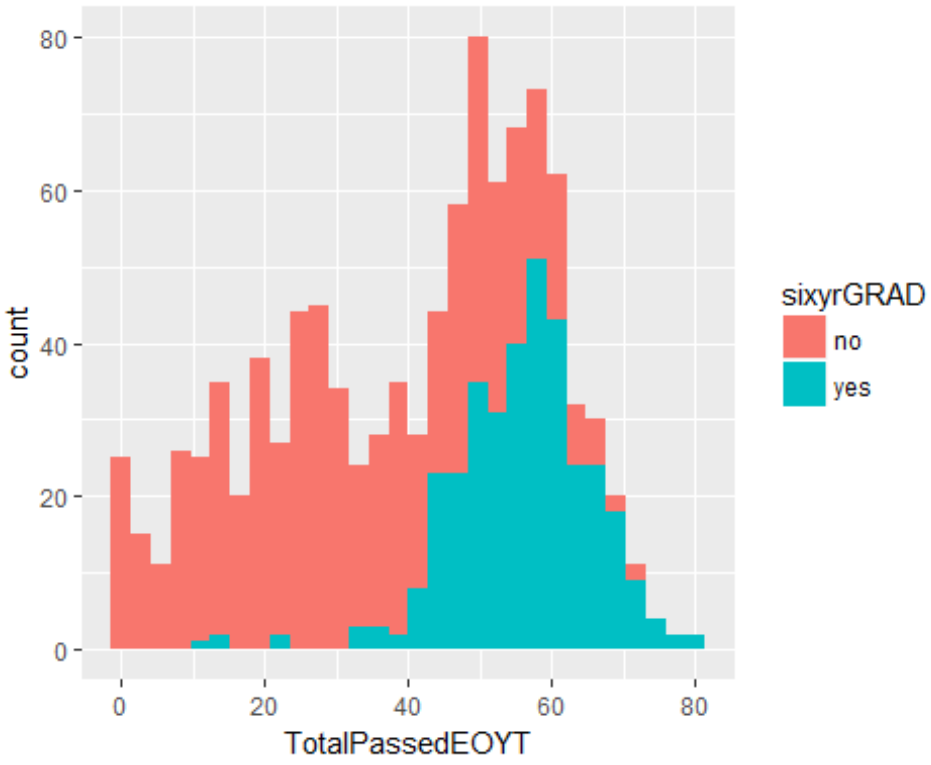
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(data=FALL2010, aes(TotalPassedEOY, fill = sixyrGRAD)) + geom_histogram
()

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




```
ggplot(data=FALL2010, aes(TotalPassedEOYT, fill = sixyrGRAD)) + geom_histogra
m()

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


80
60
40
20
0
count
0
20
40
60
80
TotalPassedEOYT
sixyrGRAD
no
yes

F

I plan to implement 2 different classification models on my data as they are appropriate for this kind of study. Classification will predict whether a student does or does not graduate in 6 years and I plan to use a decision tree because it does well handling a variety of continuous variables. I will also use a logistic regression as it handles binomial variables better and I want to see whether some of my mutated fields will improve the use of that method over the decision tree.

G

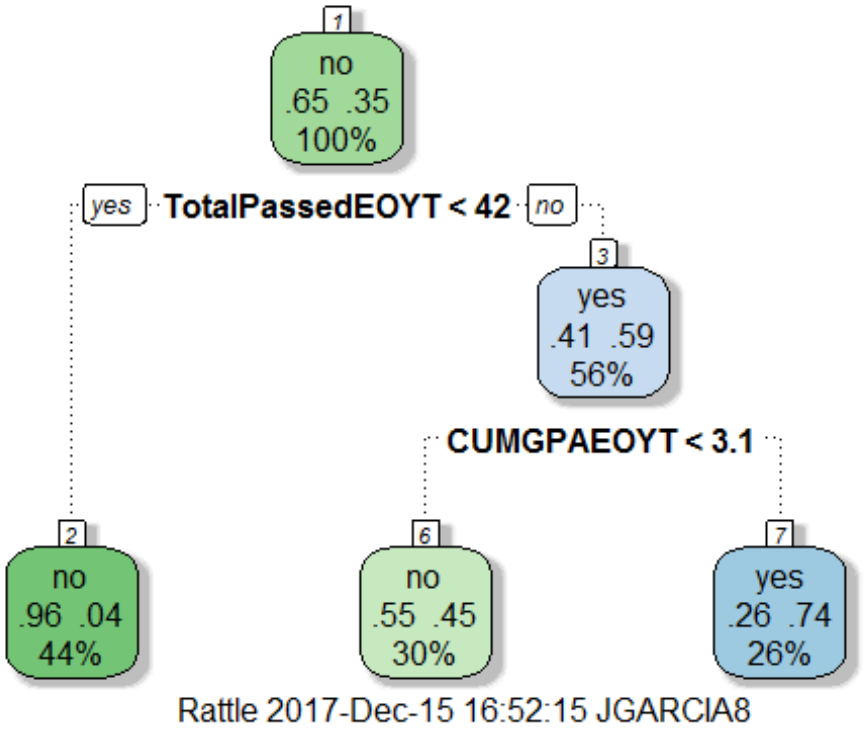
In order to avoid overfitting I removed fields that did not add to the models in addition to using a 10 fold cross validation.

H-k

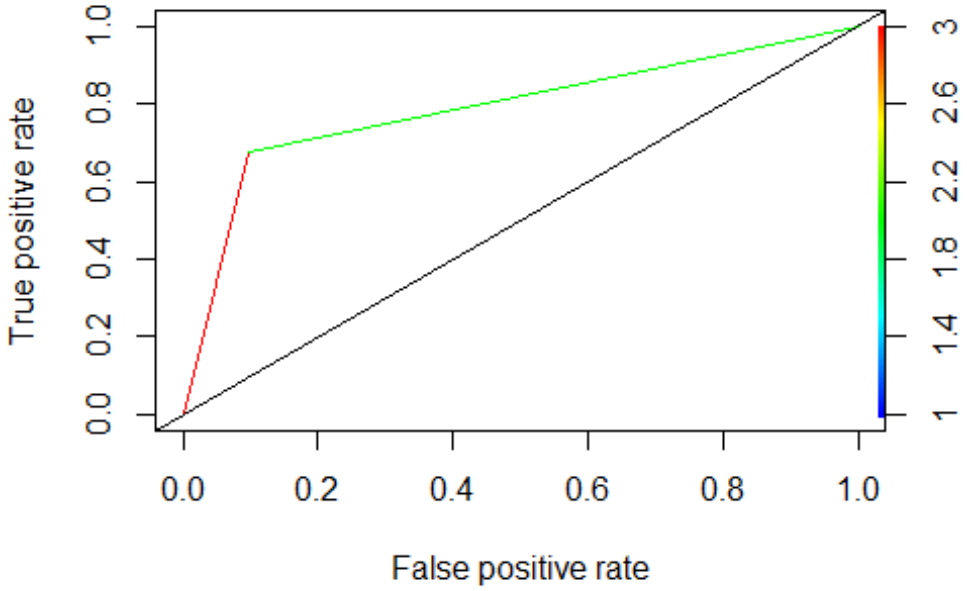
```
FALL2010$sixyrGRAD <- as.factor(FALL2010$sixyrGRAD)
set.seed(1234567)
intrain <- createDataPartition(FALL2010$sixyrGRAD,p=0.70,list=FALSE)
train <- FALL2010[intrain,]
test <- FALL2010[-intrain,]
nrow(test)

## [1] 302

trctrl <- trainControl(method="cv")
modFit <- train(sixyrGRAD ~ ., method='rpart', data=train[,-1], trControl = t
rctrl)
decisiontreemodel <- modFit$finalModel
predictionstm <- predict(modFit, newdata = test)
confusionMatrix(predictionstm,test$sixyrGRAD)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  no yes
##        no  178  34
##        yes  19  71
##
##                Accuracy : 0.8245
##                  95% CI : (0.7768, 0.8657)
```

```
##     No Information Rate : 0.6523
##     P-Value [Acc > NIR] : 2.518e-11
##
##                   Kappa : 0.5997
##  Mcnemar's Test P-Value : 0.05447
##
##             Sensitivity : 0.9036
##             Specificity : 0.6762
##          Pos Pred Value : 0.8396
##          Neg Pred Value : 0.7889
##              Prevalence : 0.6523
##          Detection Rate : 0.5894
##    Detection Prevalence : 0.7020
##       Balanced Accuracy : 0.7899
##
##        'Positive' Class : no
##
```

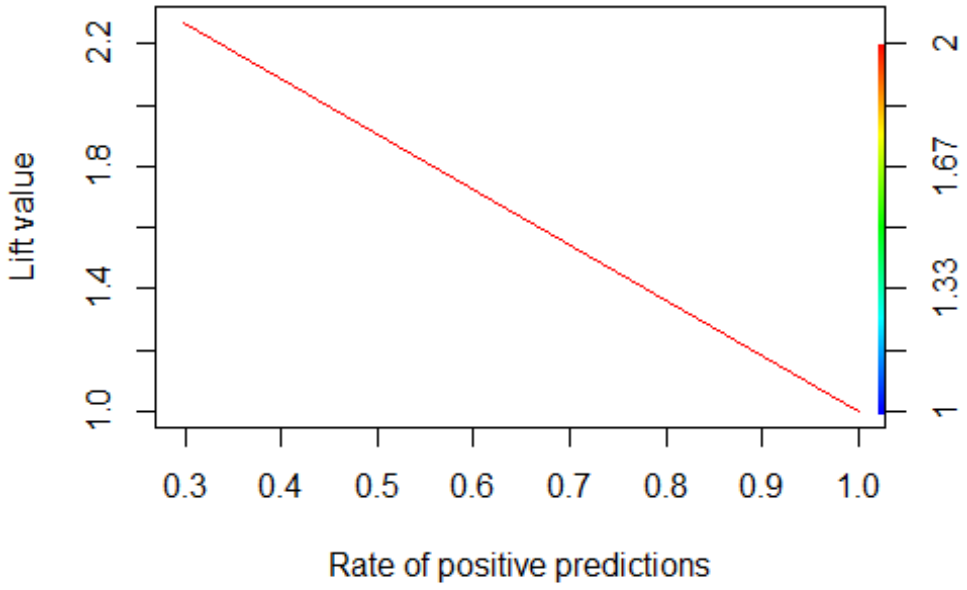
```
fancyRpartPlot(decisiontreemodel)
```



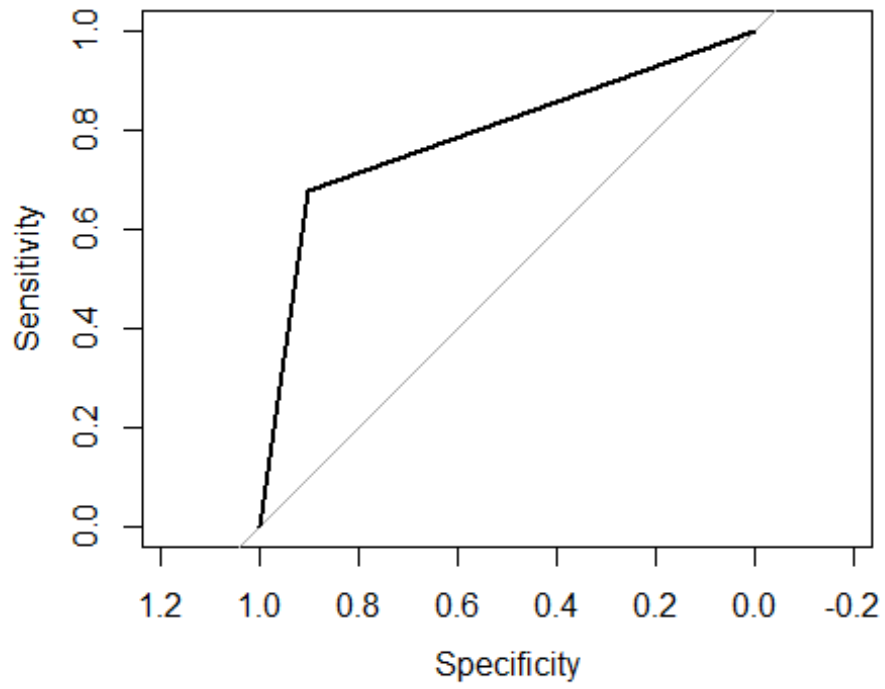

```
ROCRpredtm <- prediction(as.numeric(predictionstm), as.numeric(test$sixyrGRAD
))
ROCRperftm <- performance(ROCRpredtm, 'tpr','fpr')
plot(ROCRperftm, colorize = TRUE, text.adj = c(-0.2,1.7))
abline(0, 1)
```




```r
#Lift Curve
ROCRlifttm <- performance(ROCRpredtm, 'lift','rpp')
plot(ROCRlifttm, colorize = TRUE, text.adj = c(-0.2,1.7))
```



```
roccurvetm <- roc(test$sixyrGRAD ~ as.numeric(predictionstm))
roccurvetm$auc

## Area under the curve: 0.7899

roccurvetm$sensitivities

## [1] 1.0000000 0.6761905 0.0000000

roccurvetm$specificities

## [1] 0.0000000 0.9035533 1.0000000

plot(roccurvetm)
```




```
#logistic regression
FALL2010$sixyrGRAD <- as.factor(FALL2010$sixyrGRAD)
set.seed(12345678)
intrain1 <- createDataPartition(FALL2010$sixyrGRAD,p=0.70,list=FALSE)
train1 <- FALL2010[intrain1,]
test1 <- FALL2010[-intrain1,]
trctrl1 <- trainControl(method = "cv")
modFit1 <- train(retained ~ ., method='glm',trControl=trctrl1,data=train1[,-1
],
                 family=binomial(link='logit'))

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
logitmodel <- modFit1$finalModel
summary(logitmodel)
```

```
##
## Call:
```

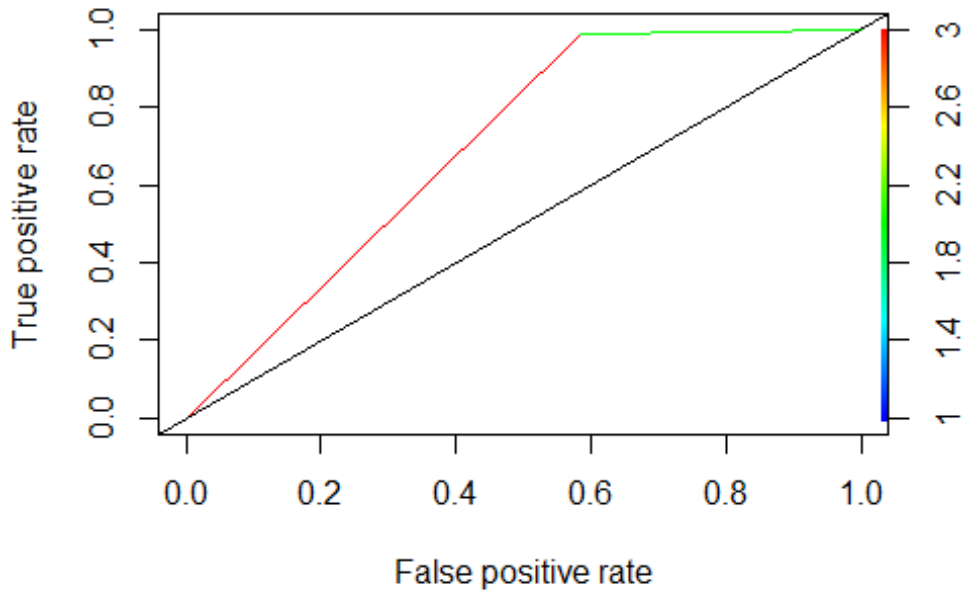
```
## NULL
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.87637  -0.01795   0.00182   0.01713   2.67589
##
## Coefficients: (1 not defined because of singularities)
##                                      Estimate Std. Error z value
## (Intercept)                         -1.105e+01  9.602e+00  -1.151
## `TermFall 2011`                      2.960e-02  7.331e-01   0.040
## creden                              -1.423e-01  1.952e-01  -0.729
## hspctl                              -4.439e-01  8.260e-01  -0.537
## msat                                 3.604e-02  7.727e-02   0.466
## crsat                                1.000e-01  7.895e-02   1.267
## Nues3                               -3.590e+00  3.350e+01  -0.107
## Nues4                                2.870e+00  1.355e+00   2.118
## Nues5                                5.162e-01  6.927e-01   0.745
## Nues7                                3.913e-01  7.627e-01   0.513
## Nues8                                3.061e+00  1.500e+00   2.041
## Nues9                               -8.451e-04  9.929e-01  -0.001
## totnusat                                   NA         NA      NA
## sixyrGRADyes                        -1.822e+00  1.143e+00  -1.593
## BASICENGL                           -4.625e-02  3.560e-01  -0.130
## BASICREAD                           -3.961e-03  2.089e-02  -0.190
## BASICARIT                            7.968e-03  1.746e-02   0.456
## BASICELAG                           -3.044e-03  1.193e-02  -0.255
## PlannedorDeclaredPlanned             1.875e+00  2.723e+00   0.689
## TermPassed                          -2.108e-01  3.310e-01  -0.637
## TotalTrnsfr                          8.121e-03  3.005e-01   0.027
## TotalTest                           -1.248e-01  3.169e-01  -0.394
## TotalUnits                           2.032e-01  2.484e-01   0.818
## CUMGPA                               1.173e+00  1.048e+00   1.119
## BirthDate                            1.092e-08  1.251e-08   0.873
## AdjustedGrossIncome                 -1.668e-05  3.517e-05  -0.474
## `FatherHighestGradeLevelHigh School`    4.218e-01  7.246e-01   0.582
## `FatherHighestGradeLevelMiddle School`  3.810e-01  1.097e+00   0.347
## FatherHighestGradeLevelUnknown          5.110e-01  9.639e-01   0.530
## `MotherHighestGradeLevelHigh School`   -1.611e-01  6.695e-01  -0.241
## `MotherHighestGradeLevelMiddle School` -6.514e-01  1.208e+00  -0.539
## MotherHighestGradeLevelUnknown         -1.792e+00  1.115e+00  -1.607
## ParentAdjustedGrossIncome              -3.383e-06  1.359e-05  -0.249
## need                                   -1.168e-04  6.830e-05  -1.709
## `HousingOn Campus Housing`              6.979e-01  9.908e-01   0.704
## Housingunknown                          5.721e-03  9.942e-01   0.006
## `HousingWith Parent`                    3.428e-01  9.828e-01   0.349
## grantactual                             1.598e-04  9.198e-05   1.737
## scholactual                            -2.286e-04  1.953e-04  -1.171
## loanactual                             -5.460e-05  8.346e-05  -0.654
## workstudyactual                        -3.081e-04  4.731e-04  -0.651
## TotalPassedEOY                         -4.433e-01  1.693e-01  -2.618
```

```
## TotalPassedEOYT                              6.609e-01  1.200e-01   5.505
## CUMGPAEOY                                    4.345e+00  1.863e+00   2.332
## CUMGPAEOYT                                  -7.495e+00  2.095e+00  -3.577
## SATGROUP                                    -7.467e-01  9.171e-01  -0.814
## GPAGROUP                                    -1.358e+00  1.069e+00  -1.271
## GPAEOYGROUP                                  1.145e+00  1.261e+00   0.907
## GPAEOYTGROUP                                -2.860e-01  1.393e+00  -0.205
## TermPassedGROUP                             -4.976e-02  8.116e-01  -0.061
## TotalPassedEOYGROUP                          1.282e-01  9.067e-01   0.141
## TotalPassedEOYTGROUP                        -8.487e-01  9.787e-01  -0.867
## EOYoverthirtyyes                            -1.576e+00  1.550e+00  -1.017
## EOYToversixtyyes                             1.029e+01  1.184e+03   0.009
## EOYgpagoodyes                               -1.029e+00  1.299e+00  -0.792
## EOYTgpagoodyes                               1.996e+00  1.330e+00   1.500
##                                             Pr(>|z|)
## (Intercept)                                 0.249735
## `TermFall 2011`                             0.967795
## creden                                      0.466228
## hspctl                                      0.590958
## msat                                        0.640891
## crsat                                       0.205223
## Nues3                                       0.914646
## Nues4                                       0.034152 *
## Nues5                                       0.456182
## Nues7                                       0.607922
## Nues8                                       0.041243 *
## Nues9                                       0.999321
## totnusat                                          NA
## sixyrGRADyes                                0.111077
## BASICENGL                                   0.896625
## BASICREAD                                   0.849637
## BASICARIT                                   0.648216
## BASICELAG                                   0.798609
## PlannedorDeclaredPlanned                    0.491089
## TermPassed                                  0.524117
## TotalTrnsfr                                 0.978441
## TotalTest                                   0.693870
## TotalUnits                                  0.413240
## CUMGPA                                      0.263213
## BirthDate                                   0.382618
## AdjustedGrossIncome                         0.635429
## `FatherHighestGradeLevelHigh School`        0.560502
## `FatherHighestGradeLevelMiddle School`      0.728304
## FatherHighestGradeLevelUnknown              0.596041
## `MotherHighestGradeLevelHigh School`        0.809791
## `MotherHighestGradeLevelMiddle School`      0.589658
## MotherHighestGradeLevelUnknown              0.107970
## ParentAdjustedGrossIncome                   0.803439
## need                                        0.087389 .
## `HousingOn Campus Housing`                  0.481211
```

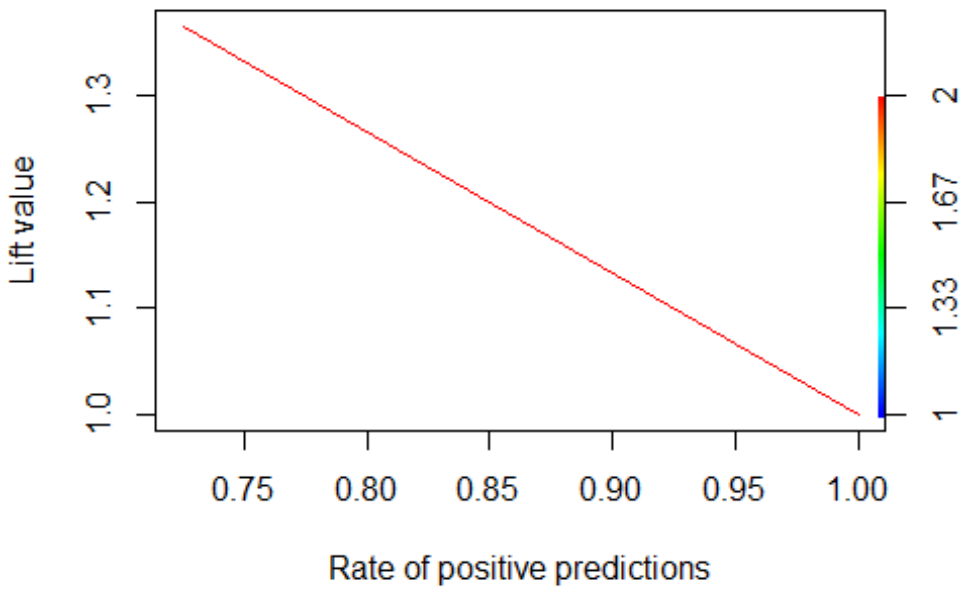
```
## Housingunknown                          0.995408
## `HousingWith Parent`                    0.727257
## grantactual                             0.082377 .
## scholactual                             0.241637
## loanactual                              0.512998
## workstudyactual                         0.514883
## TotalPassedEOY                          0.008832 **
## TotalPassedEOYT                         3.68e-08 ***
## CUMGPAEOY                               0.019694 *
## CUMGPAEOYT                              0.000348 ***
## SATGROUP                                0.415529
## GPAGROUP                                0.203799
## GPAEOYGROUP                             0.364222
## GPAEOYTGROUP                            0.837355
## TermPassedGROUP                         0.951111
## TotalPassedEOYGROUP                     0.887565
## TotalPassedEOYTGROUP                    0.385836
## EOYoverthirtyyes                        0.309000
## EOYToversixtyyes                        0.993068
## EOYgpagoodyes                           0.428226
## EOYTgpagoodyes                          0.133509
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 798.88  on 704  degrees of freedom
## Residual deviance: 140.23  on 650  degrees of freedom
## AIC: 250.23
##
## Number of Fisher Scoring iterations: 19

predictionslr <- predict(modFit1, newdata = test1)

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

confusionMatrix(predictionslr,test1$sixyrGRAD)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  no yes
##        no   82   1
##        yes 115 104
##
##                Accuracy : 0.6159
##                  95% CI : (0.5585, 0.671)
##     No Information Rate : 0.6523
##     P-Value [Acc > NIR] : 0.9169
##
```

```
##                   Kappa : 0.3245
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.4162
##             Specificity : 0.9905
##          Pos Pred Value : 0.9880
##          Neg Pred Value : 0.4749
##              Prevalence : 0.6523
##          Detection Rate : 0.2715
##    Detection Prevalence : 0.2748
##       Balanced Accuracy : 0.7034
##
##        'Positive' Class : no
##
```

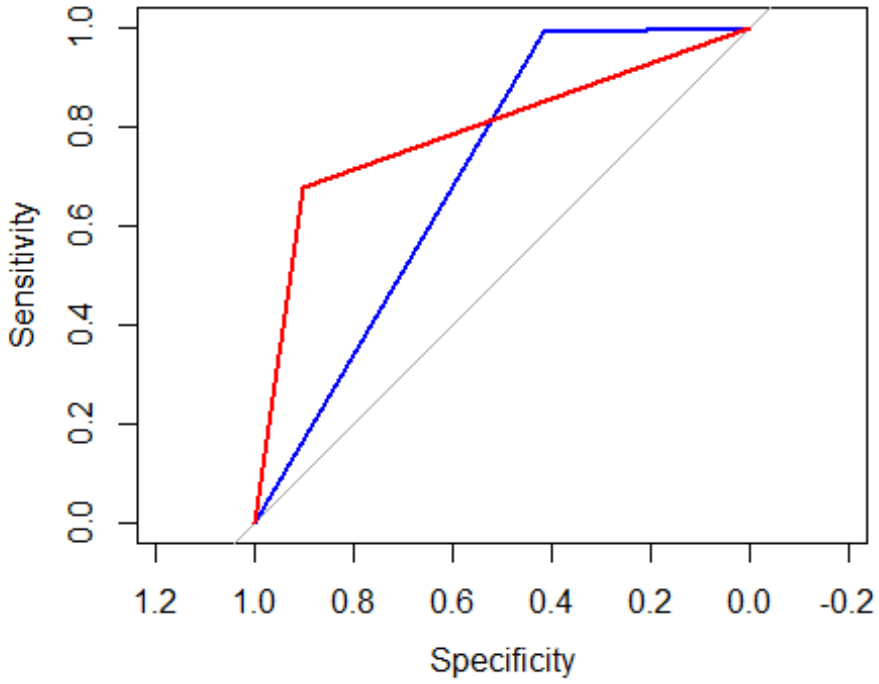
```
#ROC Curve
ROCRpred <- prediction(as.numeric(predictionslr), as.numeric(test1$sixyrGRAD)
)
ROCRperf <- performance(ROCRpred, 'tpr','fpr')
plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7))
abline(0, 1)
```




```
#Lift Curve
ROCRlift <- performance(ROCRpred, 'lift','rpp')
plot(ROCRlift, colorize = TRUE, text.adj = c(-0.2,1.7))
```




```
roccurve <- roc(test1$sixyrGRAD ~ as.numeric(predictionslr))
roccurve$auc

## Area under the curve: 0.7034

roccurve$sensitivities

## [1] 1.0000000 0.9904762 0.0000000

roccurve$specificities

## [1] 0.0000000 0.4162437 1.0000000

plot(roccurve, col = 'blue')
plot(roccurvetm, col = 'red', add = TRUE)
```



From my results the decision tree out performs the logistic regression in accuracy in addition to its rates of true positives this is due to the structure of data and the trees ability to classify continuous variables. From my EDA it was apparent that the biggest impact on student graduation were the GPA and Total Credits earned which are both continuous variables.