College_decision_tree

Nabil Momin

2024-06-10

```
library(corrgram)
library(corrplot)

## corrplot 0.92 loaded

library(caTools)
library(Amelia)

## Loading required package: Rcpp

## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.8.2, built: 2024-04-10)
## ## Copyright (C) 2005-2024 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##

library(ggplot2)
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(rpart)
library(rpart.plot)
library(randomForest)

## randomForest 4.7-1.1

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

##
## Attaching package: 'randomForest'
```

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

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

```
library(ISLR)
```

```
#### College data is inside ISLR
```

```
str(College)
```

```
## 'data.frame':    777 obs. of  18 variables:
##  $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Apps       : num  1660 2186 1428 417 193 ...
##  $ Accept     : num  1232 1924 1097 349 146 ...
##  $ Enroll     : num  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : num  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : num  52 29 50 89 44 62 45 68 63 44 ...
##  $ F.Undergrad: num  2885 2683 1036 510 249 ...
##  $ P.Undergrad: num  537 1227 99 63 869 ...
##  $ Outstate   : num  7440 12280 11250 12960 7560 ...
##  $ Room.Board : num  3300 6450 3750 5450 4120 ...
##  $ Books      : num  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : num  2200 1500 1165 875 1500 ...
##  $ PhD        : num  70 29 53 92 76 67 90 89 79 40 ...
##  $ Terminal   : num  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: num  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : num  7041 10527 8735 19016 10922 ...
##  $ Grad.Rate  : num  60 56 54 59 15 55 63 73 80 52 ...
```

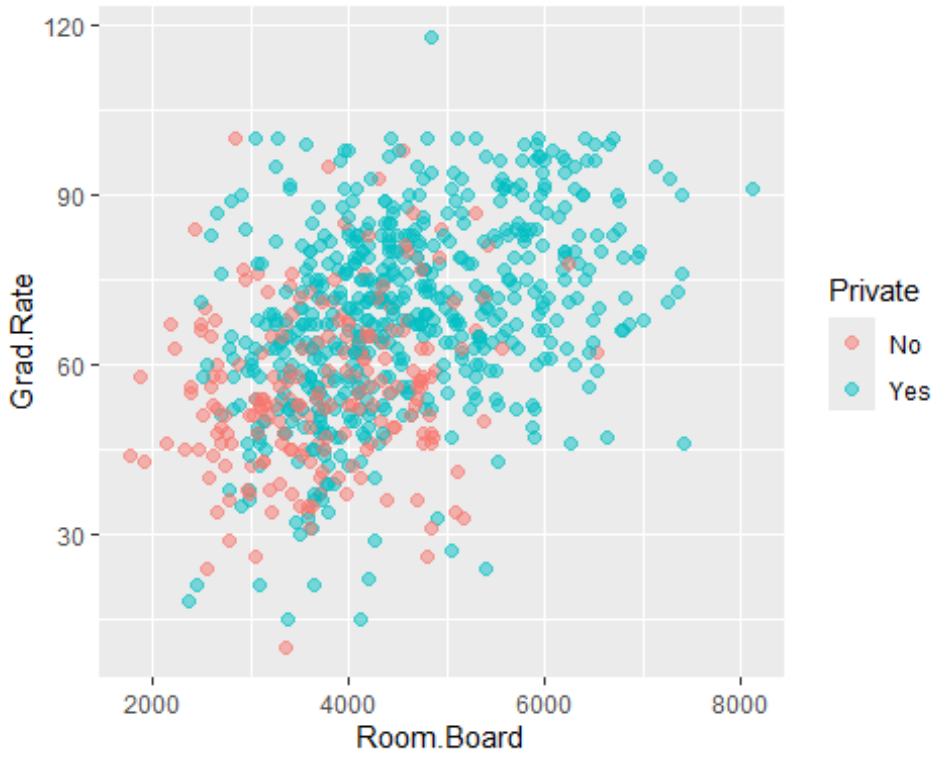
```
View(College)
```

```
#### changing the name for the simplicity
```

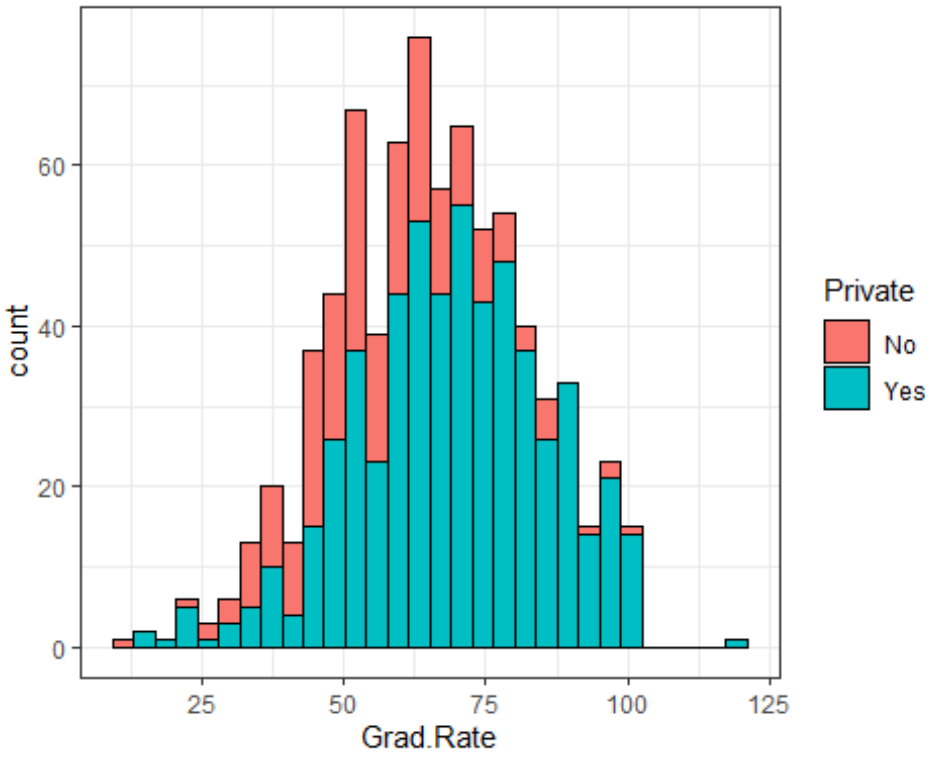
```
df <- College
```

```
#### EDA time
```

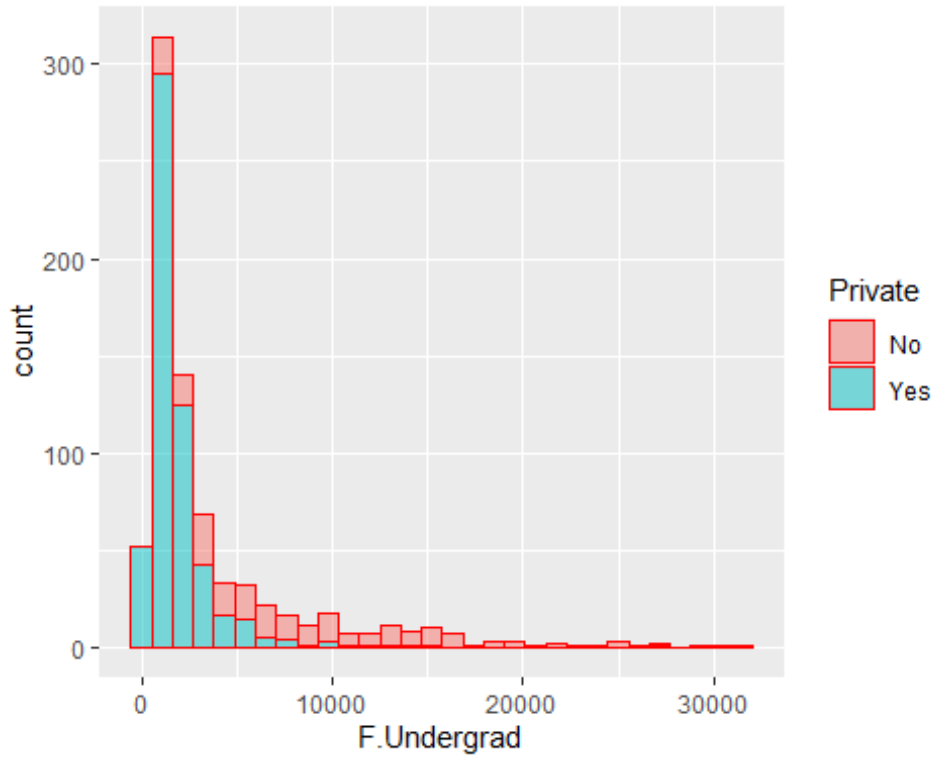
```
ggplot(df,aes(Room.Board, Grad.Rate)) +
geom_point(position=position_jitter(w=1,
h=0),aes(color=Private),alpha=0.5,size=2)
```




```
ggplot(df,aes(Grad.Rate)) + geom_histogram(aes(fill=Private),color='black') +
theme_bw()

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




```
ggplot(df,aes(F.Undergrad)) +
geom_histogram((aes(fill=Private)),color='red',alpha=0.5)

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

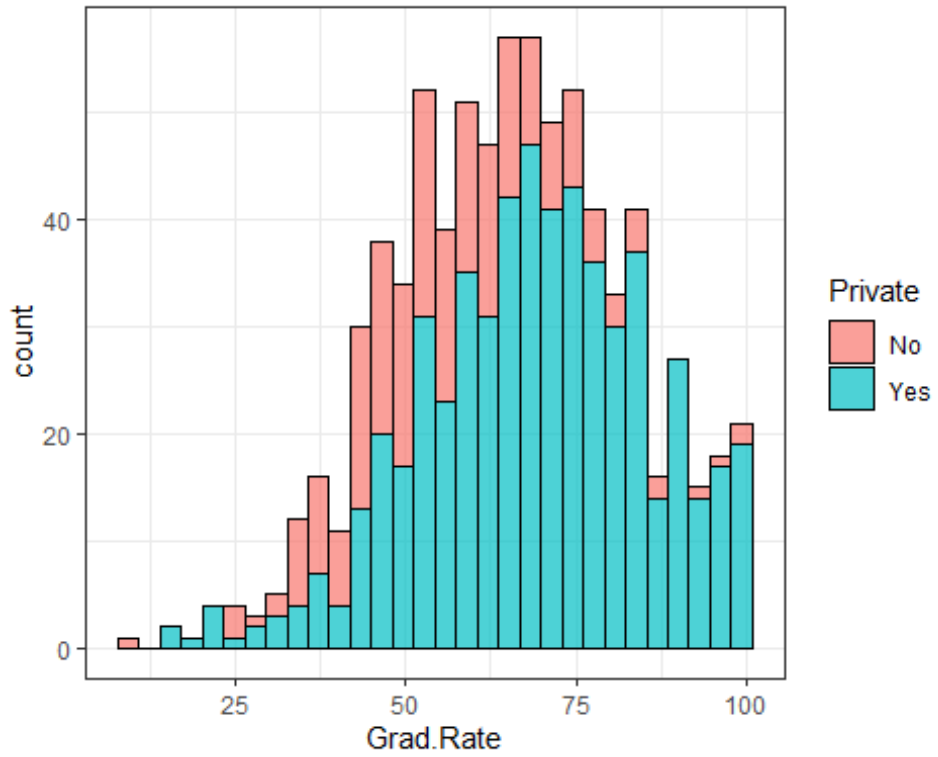



```
#### in the ggplot we see there something off because the graduation rate is
going above 100
#### lets find out which one is it

subset(df,Grad.Rate > 100)

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

#### Getting rid of 118 graduation rate and making it to 100

df['Cazenovia College','Grad.Rate'] <- 100

#### Run subset again just to make sure

subset(df,Grad.Rate > 100)

##  [1] Private     Apps        Accept      Enroll      Top10perc   Top25perc
##  [7] F.Undergrad P.Undergrad Outstate    Room.Board  Books       Personal
```

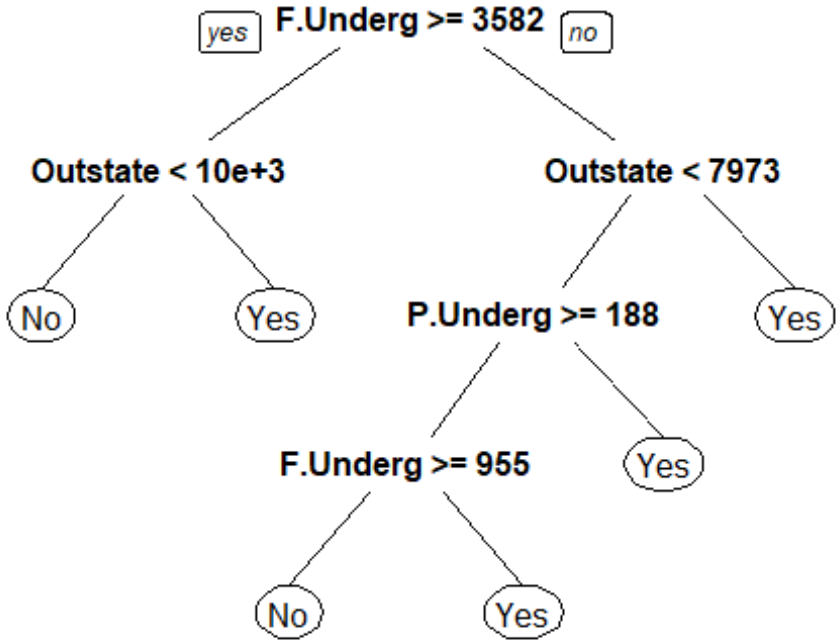
```
## [13] PhD         Terminal    S.F.Ratio   perc.alumni Expend      Grad.Rate
## <0 rows> (or 0-length row.names)
```

```
ggplot(df,aes(Grad.Rate)) +
geom_histogram(aes(fill=Private),color='black',alpha=0.7) + theme_bw()
```

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




```
#### I think our data is all clean to undergo model transformation
#### train and test data

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

train <- subset(df,sample == TRUE)

test <- subset(df,sample == FALSE)

#### Making the model now for decision tree

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

prp(tree.model)
```




```
#### Prediction
tree.predict <- predict(tree.model,test)

print(tree.predict)
```

```
##                                             No        Yes
## Adelphi University                   0.006289308 0.99371069
## Albion College                       0.006289308 0.99371069
## Allegheny College                    0.006289308 0.99371069
## Alverno College                      0.006289308 0.99371069
## Anderson University                  0.006289308 0.99371069
## Angelo State University              0.963636364 0.03636364
## Antioch University                   0.006289308 0.99371069
## Arizona State University Main campus 0.963636364 0.03636364
## Assumption College                   0.006289308 0.99371069
## Augsburg College                     0.006289308 0.99371069
## Averett College                      0.006289308 0.99371069
## Bard College                         0.006289308 0.99371069
## Beaver College                       0.006289308 0.99371069
## Bellarmine College                   0.006289308 0.99371069
## Belmont University                   0.777777778 0.22222222
## Bennington College                   0.006289308 0.99371069
## Bethel College KS                    0.006289308 0.99371069
## Bloomsburg Univ. of Pennsylvania     0.963636364 0.03636364
## Bradley University                   0.176470588 0.82352941
## Brenau University                    0.006289308 0.99371069
## Brewton-Parker College               0.777777778 0.22222222
```

```
## Briar Cliff College                              0.006289308 0.99371069
## Bryn Mawr College                                0.006289308 0.99371069
## Cabrini College                                  0.006289308 0.99371069
## Campbell University                              0.777777778 0.22222222
## Campbellsville College                           0.142857143 0.85714286
## Canisius College                                 0.006289308 0.99371069
## Capital University                               0.006289308 0.99371069
## Carnegie Mellon University                       0.176470588 0.82352941
## Cazenovia College                                0.006289308 0.99371069
## Central College                                  0.006289308 0.99371069
## Central Missouri State University                0.963636364 0.03636364
## College of Santa Fe                              0.006289308 0.99371069
## College of St. Scholastica                       0.006289308 0.99371069
## Columbia College                                 0.006289308 0.99371069
## Concordia Lutheran College                       0.142857143 0.85714286
## Creighton University                             0.006289308 0.99371069
## Culver-Stockton College                          0.006289308 0.99371069
## D'Youville College                               0.006289308 0.99371069
## Dana College                                     0.006289308 0.99371069
## Dartmouth College                                0.176470588 0.82352941
## Delta State University                           0.777777778 0.22222222
## DePauw University                                0.006289308 0.99371069
## Doane College                                    0.006289308 0.99371069
## Duke University                                  0.176470588 0.82352941
## East Carolina University                         0.963636364 0.03636364
## Eastern Connecticut State University             0.777777778 0.22222222
## Eastern Illinois University                      0.963636364 0.03636364
## Eckerd College                                   0.006289308 0.99371069
## Elms College                                     0.006289308 0.99371069
## Emory & Henry College                            0.006289308 0.99371069
## Evergreen State College                          0.777777778 0.22222222
## Florida State University                         0.963636364 0.03636364
## Fort Lewis College                               0.963636364 0.03636364
## Francis Marion University                        0.777777778 0.22222222
## Gardner Webb University                          0.006289308 0.99371069
## Geneva College                                   0.006289308 0.99371069
## Georgia Institute of Technology                  0.963636364 0.03636364
## Georgian Court College                           0.006289308 0.99371069
## Goucher College                                  0.006289308 0.99371069
## Green Mountain College                           0.006289308 0.99371069
## Grinnell College                                 0.006289308 0.99371069
## Guilford College                                 0.006289308 0.99371069
## Hampton University                               0.963636364 0.03636364
## Hanover College                                  0.006289308 0.99371069
## Hartwick College                                 0.006289308 0.99371069
## Harvard University                               0.176470588 0.82352941
## Hastings College                                 0.006289308 0.99371069
## Hobart and William Smith Colleges                0.006289308 0.99371069
## Houghton College                                 0.006289308 0.99371069
## Huntington College                               0.006289308 0.99371069
```

```
## Illinois Institute of Technology              0.006289308 0.99371069
## Illinois State University                     0.963636364 0.03636364
## Illinois Wesleyan University                  0.006289308 0.99371069
## Immaculata College                            0.006289308 0.99371069
## Indiana State University                      0.963636364 0.03636364
## Iowa State University                         0.963636364 0.03636364
## Ithaca College                                0.176470588 0.82352941
## Jamestown College                             0.963636364 0.03636364
## Judson College                                0.006289308 0.99371069
## Juniata College                               0.006289308 0.99371069
## Kentucky Wesleyan College                     0.006289308 0.99371069
## King's College                                0.006289308 0.99371069
## La Roche College                              0.006289308 0.99371069
## La Salle University                           0.006289308 0.99371069
## Lamar University                              0.963636364 0.03636364
## Le Moyne College                              0.006289308 0.99371069
## LeTourneau University                         0.006289308 0.99371069
## Livingstone College                           0.142857143 0.85714286
## Lock Haven University of Pennsylvania         0.777777778 0.22222222
## Louisiana College                             0.142857143 0.85714286
## Loyola College                                0.006289308 0.99371069
## Loyola University                             0.006289308 0.99371069
## Lynchburg College                             0.006289308 0.99371069
## Lyndon State College                          0.142857143 0.85714286
## Macalester College                            0.006289308 0.99371069
## MacMurray College                             0.006289308 0.99371069
## Manhattan College                             0.006289308 0.99371069
## Marian College of Fond du Lac                 0.006289308 0.99371069
## Maryville College                             0.006289308 0.99371069
## McMurry University                            0.090909091 0.90909091
## McPherson College                             0.142857143 0.85714286
## Mercer University                             0.006289308 0.99371069
## Mesa State College                            0.777777778 0.22222222
## Michigan State University                     0.176470588 0.82352941
## MidAmerica Nazarene College                   0.142857143 0.85714286
## Montreat-Anderson College                     0.006289308 0.99371069
## Moravian College                              0.006289308 0.99371069
## Morehouse College                             0.142857143 0.85714286
## Morningside College                           0.006289308 0.99371069
## Morris College                                0.142857143 0.85714286
## Mount Vernon Nazarene College                 0.142857143 0.85714286
## Nazareth College of Rochester                 0.006289308 0.99371069
## North Carolina A. & T. State University       0.963636364 0.03636364
## North Carolina State University at Raleigh    0.963636364 0.03636364
## North Central College                         0.006289308 0.99371069
## Northwest Missouri State University           0.963636364 0.03636364
## Northwest Nazarene College                    0.006289308 0.99371069
## Northwestern College                          0.006289308 0.99371069
## Northwestern University                       0.176470588 0.82352941
## Norwich University                            0.006289308 0.99371069
```

```
## Oakland University                                 0.963636364 0.03636364
## Oberlin College                                    0.006289308 0.99371069
## Oglethorpe University                              0.006289308 0.99371069
## Ohio Northern University                           0.006289308 0.99371069
## Ohio Wesleyan University                           0.006289308 0.99371069
## Oklahoma Baptist University                        0.777777778 0.22222222
## Oklahoma Christian University                      0.777777778 0.22222222
## Oklahoma State University                          0.963636364 0.03636364
## Ouachita Baptist University                        0.142857143 0.85714286
## Pacific Lutheran University                        0.006289308 0.99371069
## Pennsylvania State Univ. Main Campus               0.176470588 0.82352941
## Pitzer College                                     0.006289308 0.99371069
## Point Loma Nazarene College                        0.006289308 0.99371069
## Prairie View A. and M. University                  0.963636364 0.03636364
## Presbyterian College                               0.006289308 0.99371069
## Princeton University                               0.176470588 0.82352941
## Providence College                                 0.176470588 0.82352941
## Quincy University                                  0.006289308 0.99371069
## Regis College                                      0.006289308 0.99371069
## Rhodes College                                     0.006289308 0.99371069
## Rider University                                   0.006289308 0.99371069
## Rivier College                                     0.006289308 0.99371069
## Rockhurst College                                  0.006289308 0.99371069
## Rocky Mountain College                             0.006289308 0.99371069
## Rowan College of New Jersey                        0.963636364 0.03636364
## Rutgers State University at Camden                 0.777777778 0.22222222
## Sacred Heart University                            0.006289308 0.99371069
## Saint Francis College                              0.006289308 0.99371069
## Saint Joseph College                               0.006289308 0.99371069
## Saint Mary-of-the-Woods College                    0.006289308 0.99371069
## Saint Michael's College                            0.006289308 0.99371069
## Saint Olaf College                                 0.006289308 0.99371069
## Saint Peter's College                              0.006289308 0.99371069
## Saint Vincent College                              0.006289308 0.99371069
## Salem College                                      0.006289308 0.99371069
## San Diego State University                         0.963636364 0.03636364
## Santa Clara University                             0.176470588 0.82352941
## Sarah Lawrence College                             0.006289308 0.99371069
## Scripps College                                    0.006289308 0.99371069
## Seattle Pacific University                         0.006289308 0.99371069
## Sioux Falls College                                0.006289308 0.99371069
## Smith College                                      0.006289308 0.99371069
## Southeast Missouri State University                0.963636364 0.03636364
## Southeastern Oklahoma State Univ.                  0.777777778 0.22222222
## Southern California College                        0.006289308 0.99371069
## Southern Illinois University at Edwardsville       0.963636364 0.03636364
## Southwestern University                            0.006289308 0.99371069
## St. Bonaventure University                         0.006289308 0.99371069
## Stevens Institute of Technology                    0.006289308 0.99371069
## Stockton College of New Jersey                     0.963636364 0.03636364
```

```
## Stonehill College                              0.006289308 0.99371069
## SUNY at Binghamton                             0.963636364 0.03636364
## SUNY College at Cortland                       0.963636364 0.03636364
## SUNY College at Geneseo                        0.963636364 0.03636364
## Sweet Briar College                            0.006289308 0.99371069
## Texas Christian University                     0.963636364 0.03636364
## Texas Lutheran College                         0.090909091 0.90909091
## Texas Southern University                      0.963636364 0.03636364
## Texas Wesleyan University                      0.777777778 0.22222222
## Thiel College                                  0.006289308 0.99371069
## Tiffin University                              0.090909091 0.90909091
## Univ. of Wisconsin at OshKosh                  0.963636364 0.03636364
## University of California at Berkeley           0.176470588 0.82352941
## University of Central Florida                  0.963636364 0.03636364
## University of Dayton                           0.176470588 0.82352941
## University of Delaware                         0.963636364 0.03636364
## University of Denver                           0.006289308 0.99371069
## University of Kansas                           0.963636364 0.03636364
## University of Maine at Presque Isle            0.777777778 0.22222222
## University of Miami                            0.176470588 0.82352941
## University of Michigan at Ann Arbor            0.176470588 0.82352941
## University of Minnesota at Morris              0.006289308 0.99371069
## University of Missouri at Saint Louis          0.963636364 0.03636364
## University of Montevallo                       0.777777778 0.22222222
## University of New England                      0.006289308 0.99371069
## University of New Hampshire                    0.176470588 0.82352941
## University of North Carolina at Chapel Hill    0.963636364 0.03636364
## University of North Carolina at Charlotte      0.963636364 0.03636364
## University of North Carolina at Greensboro     0.963636364 0.03636364
## University of North Texas                      0.963636364 0.03636364
## University of Notre Dame                       0.176470588 0.82352941
## University of Oregon                           0.176470588 0.82352941
## University of Pittsburgh-Main Campus           0.176470588 0.82352941
## University of Rhode Island                     0.963636364 0.03636364
## University of Scranton                         0.176470588 0.82352941
## University of Southern California              0.176470588 0.82352941
## University of the Pacific                      0.006289308 0.99371069
## University of Utah                             0.963636364 0.03636364
## University of West Florida                     0.963636364 0.03636364
## University of Wisconsin-Superior               0.777777778 0.22222222
## University of Wisconsin at Madison             0.963636364 0.03636364
## University of Wisconsin at Milwaukee           0.963636364 0.03636364
## Valley City State University                   0.090909091 0.90909091
## Vanderbilt University                          0.176470588 0.82352941
## Villanova University                           0.176470588 0.82352941
## Virginia Wesleyan College                      0.006289308 0.99371069
## Wartburg College                               0.006289308 0.99371069
## Washington College                             0.006289308 0.99371069
## Western State College of Colorado              0.142857143 0.85714286
## Westminster College MO                         0.006289308 0.99371069
```

```
## Wheaton College IL                          0.006289308 0.99371069
## Whitman College                             0.006289308 0.99371069
## Wilkes University                           0.006289308 0.99371069
## Willamette University                       0.006289308 0.99371069
## William Jewell College                      0.006289308 0.99371069
## William Woods University                    0.006289308 0.99371069
## Williams College                            0.006289308 0.99371069
## Wilson College                              0.006289308 0.99371069
## Wofford College                             0.006289308 0.99371069
## Worcester State College                     0.777777778 0.22222222
## Yale University                             0.176470588 0.82352941
## York College of Pennsylvania                0.777777778 0.22222222
```

```
#### Lets make a column next to Yes which will say YES if its above 0.5

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

View(tree.predict)

predict.column <- function(x){
  if (x >= 0.5){
    return ('YES')
  }else {
    return('NO')
  }
}

tree.predict$Predict <- sapply(tree.predict$Yes,predict.column)

str(tree.predict)
```

```
## 'data.frame':    233 obs. of  3 variables:
##  $ No     : num  0.00629 0.00629 0.00629 0.00629 0.00629 ...
##  $ Yes    : num  0.994 0.994 0.994 0.994 0.994 ...
##  $ Predict: chr  "YES" "YES" "YES" "YES" ...
```

```
print(head(tree.predict))
```

```
##                                  No        Yes Predict
## Adelphi University      0.006289308 0.99371069     YES
## Albion College          0.006289308 0.99371069     YES
## Allegheny College       0.006289308 0.99371069     YES
## Alverno College         0.006289308 0.99371069     YES
## Anderson University     0.006289308 0.99371069     YES
## Angelo State University 0.963636364 0.03636364      NO
```

```
#### Now make the confusion matrix with test and model in mind

table(tree.predict$Predict,test$Private)
```

```
##
##        No Yes
##   NO   53  11
##   YES  11 158
```

```
#### calculate the accuracy of our model
#### accuracy = true positives + true negatives / total predictions

accuracy <- (162+52) / (162+52+7+12)

print(accuracy)
```

```
## [1] 0.9184549
```