US Research University Prediction Model

Philip Gabriel Andrada

November 18, 2016

Preparation

```
# loading necessary libraries
library(rpart)
library(randomForest)
```

```
## randomForest 4.6-12

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

```
library(tree)
library(party)
```

```
## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich
```

```
library(caret)
```

```
## Loading required package: lattice

## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'

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

```
library(Boruta)
```

```
## Loading required package: ranger

##
## Attaching package: 'ranger'

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

```
library(e1071)
library(ROCR)
```

```
## Loading required package: gplots

##
## Attaching package: 'gplots'

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

```
library(corrplot)
library(ggplot2)
```

```
#Reading Data Files
usuniv2010 <- read.csv("C:\\Users\\pandrada\\Desktop\\Capstone\\MERGED2010_11_PP.csv")
usuniv2011 <- read.csv("C:\\Users\\pandrada\\Desktop\\Capstone\\MERGED2011_12_PP.csv")
usuniv2012 <- read.csv("C:\\Users\\pandrada\\Desktop\\Capstone\\MERGED2012_13_PP.csv")
usuniv2013 <- read.csv("C:\\Users\\pandrada\\Desktop\\Capstone\\MERGED2013_14_PP.csv")
usuniv2014 <- read.csv("C:\\Users\\pandrada\\Desktop\\Capstone\\MERGED2014_15_PP.csv")

#Binding All Data Files into One Data Frame
usuniv <- rbind(usuniv2010,usuniv2011,usuniv2012,usuniv2013,usuniv2014)
```

```
## Warning in `[<-.factor`(`*tmp*`, ri, value = c(100200L, 105200L,
## 2503400L, : invalid factor level, NA generated

## Warning in `[<-.factor`(`*tmp*`, ri, value = c(100200L, 105200L,
## 2503400L, : invalid factor level, NA generated

## Warning in `[<-.factor`(`*tmp*`, ri, value = c(100200L, 105200L,
```

```
## 2503400L, : invalid factor level, NA generated

## Warning in `[<-.factor`(`*tmp*`, ri, value = c(100200L, 105200L,
## 2503400L, : invalid factor level, NA generated

## Warning in `[<-.factor`(`*tmp*`, ri, value = c(100200L, 105200L,
## 2503400L, : invalid factor level, NA generated

## Warning in `[<-.factor`(`*tmp*`, ri, value = c(100200L, 105200L,
## 2503400L, : invalid factor level, NA generated
```

```
#Since there are some incomplete Carnegie Classifications, we use usuniv2014 as basis for the classific
usuniv$CCBASIC2 <- usuniv2014$CCBASIC[match(usuniv$OPEID6,usuniv2014$OPEID6)]

#added the ACCEPTED column for those that are research universities (CCBASIC2 is equal to 15 or 16), as
usuniv$ACCEPTED <- ifelse(usuniv$CCBASIC2 %in% c(15,16), 1, 0)

#number of rows in the usuniv data frame
rows_usuniv <- nrow(usuniv)
rows_usuniv
```

```
## [1] 38389
```

```
#number of columns that are in the usuniv data frame
ncol(usuniv)
```

```
## [1] 1745
```

```
#number of rows that are research universities in the data frame before cleansing
rows_usunivaccepted <- nrow(usuniv[usuniv$ACCEPTED == 1,])
rows_usunivaccepted
```

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

```
#grab a head of research universities to see if we got the correct ones
head(usuniv[usuniv$ACCEPTED == 1,c(4,1744:1745)], 30)
```

```
##                                                INSTNM CCBASIC2
## 2                  University of Alabama at Birmingham       15
## 4                  University of Alabama in Huntsville       16
## 6                            The University of Alabama       16
## 10                                 Auburn University       16
## 50                       University of South Alabama       16
## 61                   University of Alaska Fairbanks       16
## 82                   Arizona State University-Tempe       15
## 84                            University of Arizona       15
## 113                     Northern Arizona University       16
## 144                          University of Arkansas       15
## 237                California Institute of Technology       15
## 254                 University of California-Berkeley       15
## 255                    University of California-Davis       15
```

```
## 256                       University of California-Irvine       15
## 257                   University of California-Los Angeles       15
## 258                     University of California-Riverside       15
## 259                     University of California-San Diego       15
## 261                 University of California-Santa Barbara       15
## 262                    University of California-Santa Cruz       15
## 294                          Claremont Graduate University       16
## 518                             San Diego State University       16
## 567                        University of Southern California       15
## 604 University of Colorado Denver/Anschutz Medical Campus       16
## 607                           University of Colorado Boulder       15
## 614                                Colorado School of Mines       16
## 616                 Colorado State University-Fort Collins       15
## 627                                    University of Denver       16
## 644                         University of Northern Colorado       16
## 675                               University of Connecticut       15
## 720                                           Yale University       15
##     ACCEPTED
## 2          1
## 4          1
## 6          1
## 10         1
## 50         1
## 61         1
## 82         1
## 84         1
## 113        1
## 144        1
## 237        1
## 254        1
## 255        1
## 256        1
## 257        1
## 258        1
## 259        1
## 261        1
## 262        1
## 294        1
## 518        1
## 567        1
## 604        1
## 607        1
## 614        1
## 616        1
## 627        1
## 644        1
## 675        1
## 720        1
```

```
#Create a vector with the columns that is needed from the study
# 19 - institution region (1-New England, 2-Mid East, 3-Great Lakes, 4-Plains, 5-Southeast, 6-Southwest
# 37-38 - admission rate
# 39-61 - SAT and ACT Scores
# 62-99 - percentage of degrees awarded for each field of study
```

```r
# 293-299 - total share of enrollment for different ethnicities
# 300 - total share of enrollment that are non-resident aliens (i.e. international students)
# 301 - total share of enrollment that have unknown race
# 314 - share of undergraduate, degree-/certificate-seeking students who are part-time
# 377 - average cost of attendance in an academic year institution
# 379 - in-state tuition and fees
# 380 - out-of-state tuition and fees
# 387 - completion rate of first-time, full-time students at four-year institutions with 150% of expect
# 397-403 - completion rate for first-time, full-time students for different ethnicities
# 404 - completion rate for first-time, full-time students for non-resident aliens
# 405 - completion rate for first-time, full-time students that have unknown race
# 429 - retention rate for first-time, full time students at four-year institutions
# 438 - percent of all federal undergraduate students receiving a federal student loan
# 1412 - percentage of first-generation students
# 1740-1741 - total share of enrollment per gender
# 1745 - acceptance flag
col_select <- c(19,37:38,61:99,293:301,314,377,379:380,387,397:405,429,438,1412,1740:1741, 1744, 1745)

# Create a new data frame with the columns that will be filtered out
usunivfilter <- usuniv[,col_select]

# Change the factor columns to numeric for faster processing
for (i in 1:ncol(usunivfilter)){
  usunivfilter[,i] <- as.numeric(as.character(usunivfilter[,i]))
}
```

```
## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

```
## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

```
## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

```
## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

```
# Clean the results to have all complete
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150_4),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150_4_ASIAN),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150_4_WHITE),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150_4_BLACK),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150_4_NRA),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$ADM_RATE_ALL),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$SAT_AVG_ALL),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$UGDS_ASIAN),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$UGDS_WHITE),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$UGDS_BLACK),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$UGDS_NRA),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$UGDS_WOMEN),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$UGDS_MEN),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$COSTT4_A),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP11),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP12),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP14),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP15),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP24),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP26),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP27),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP40),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP45),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP51),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP52),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCTFLOAN),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PPTUG_EF),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$RET_FT4),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PAR_ED_PCT_1STGEN),]

#We will create another data frame for the research universities only
usresearchuniv <- usunivfilter[usunivfilter$CCBASIC2 %in% c(15,16),]

#show number of rows in the filtered usuniv
rows_usunivfilter <- nrow(usunivfilter)
rows_usunivfilter
```

```
## [1] 4247
```

```
#percentage of data from filtered to unfiltered
rows_usunivfilter / rows_usuniv
```

```
## [1] 0.1106306
```

```r
#show number of rows of filtered research universities
rows_usresearchuniv <- nrow(usresearchuniv)
rows_usresearchuniv
```

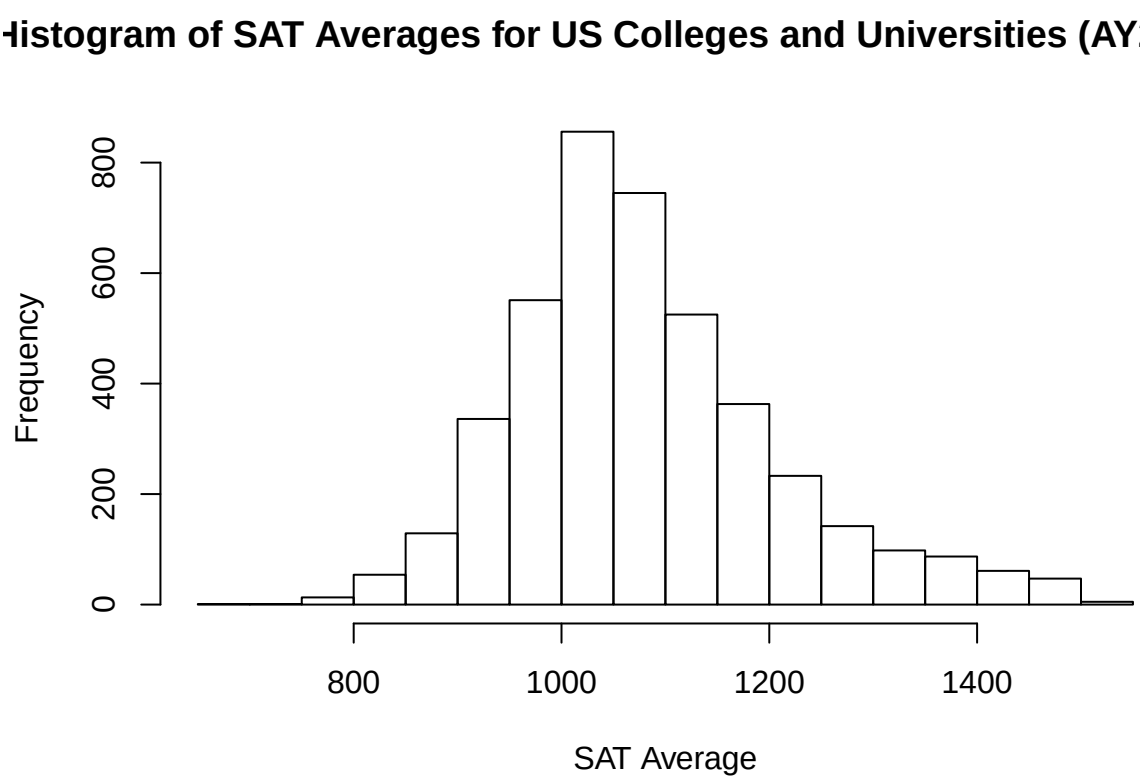
```
## [1] 815
```

```r
#percentage of data from filtered research universities to unfiltered
rows_usresearchuniv / rows_usunivaccepted
```

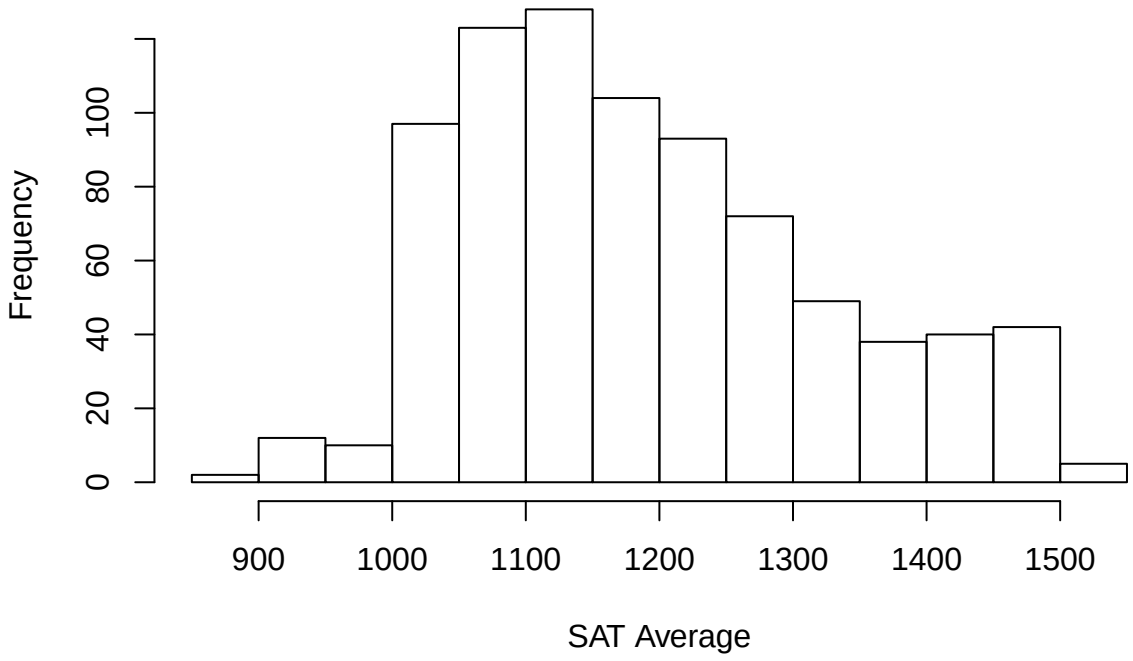
```
## [1] 0.7062392
```

Distributions and Box and Whisker Plots

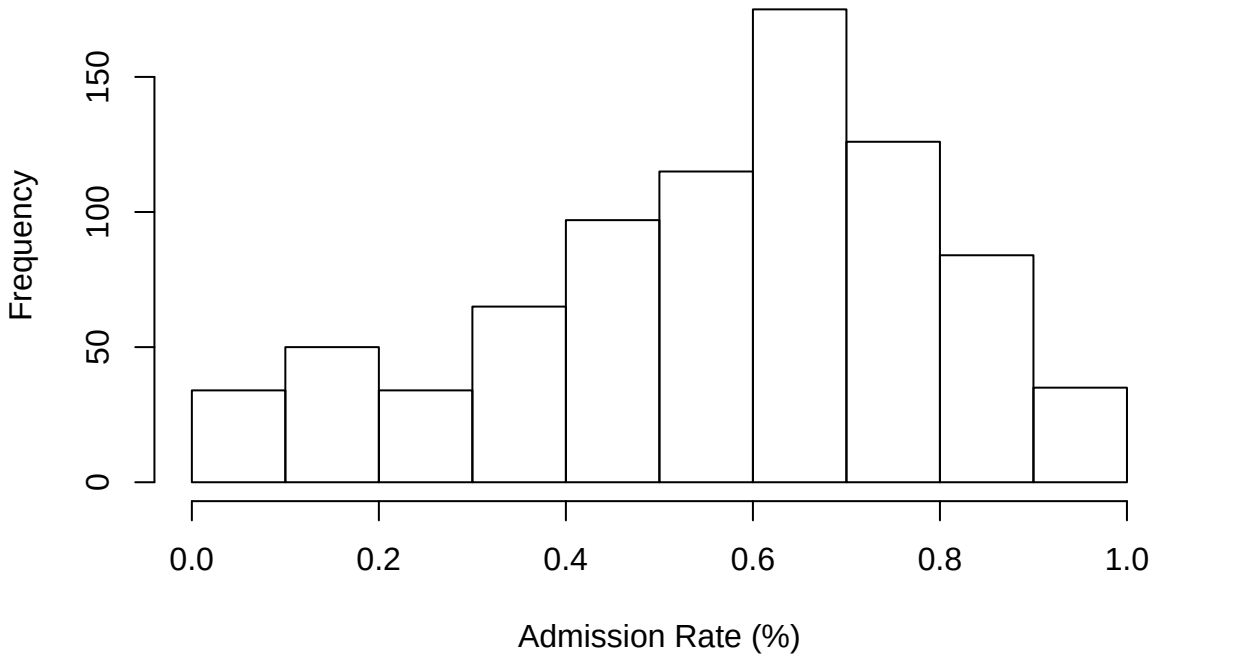
```r
# Histogram of SAT Averages for US Colleges and Universities
hist(usunivfilter$SAT_AVG_ALL, main = "Histogram of SAT Averages for US Colleges and Universities (AY20
```



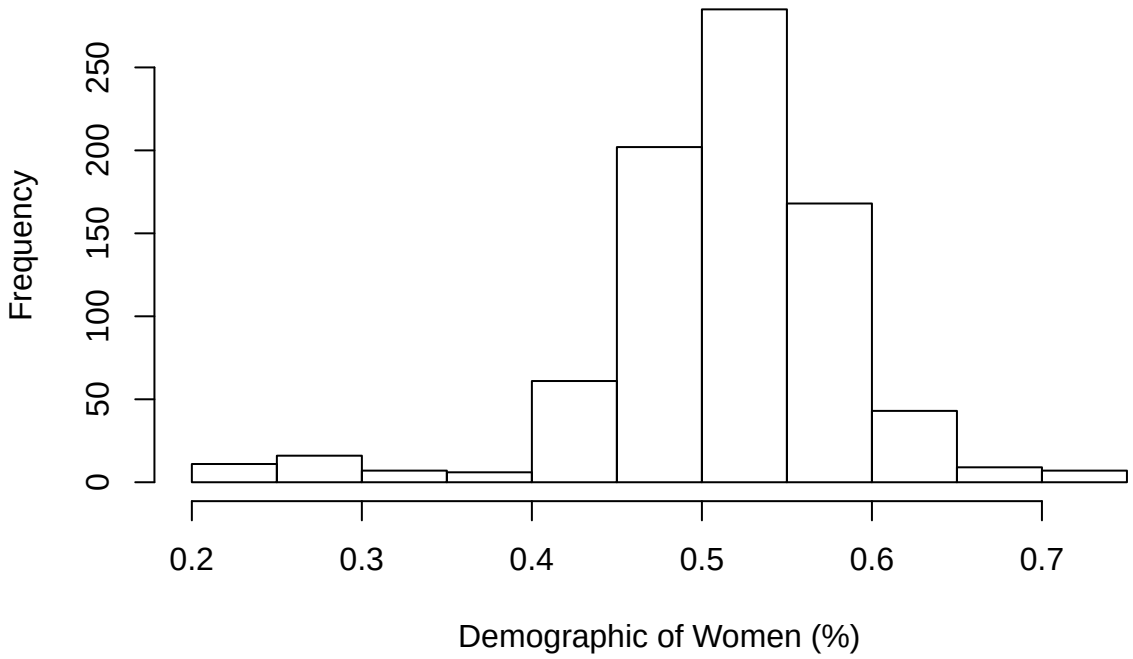

```r
# Histogram of SAT Averages for US Research Universities
hist(usresearchuniv$SAT_AVG_ALL, main = "Histogram of SAT Averages for US Research Universities (AY2010
```



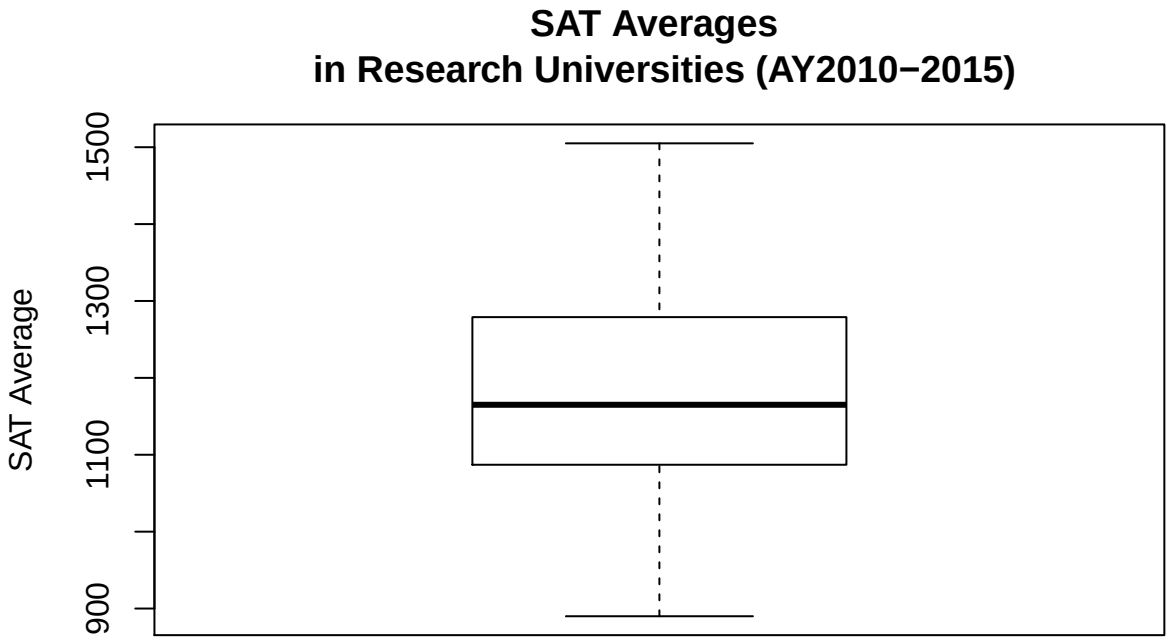

```r
# Histogram of Admission Rates for US Research Universities
hist(usresearchuniv$ADM_RATE_ALL, main = "Histogram of Admission Rates for Research Universities (AY201
```



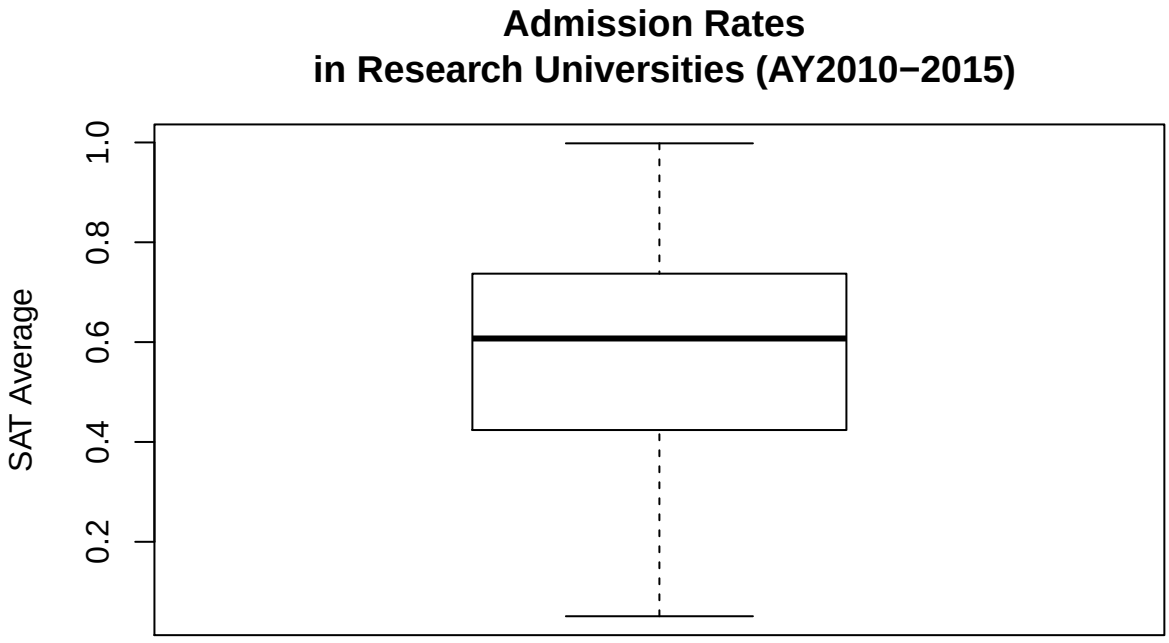

```r
# Histogram of Women in US Research Universities
hist(usresearchuniv$UGDS_WOMEN, main = "Histogram of Women in Research Universities (AY2010-2015)", xlab
```



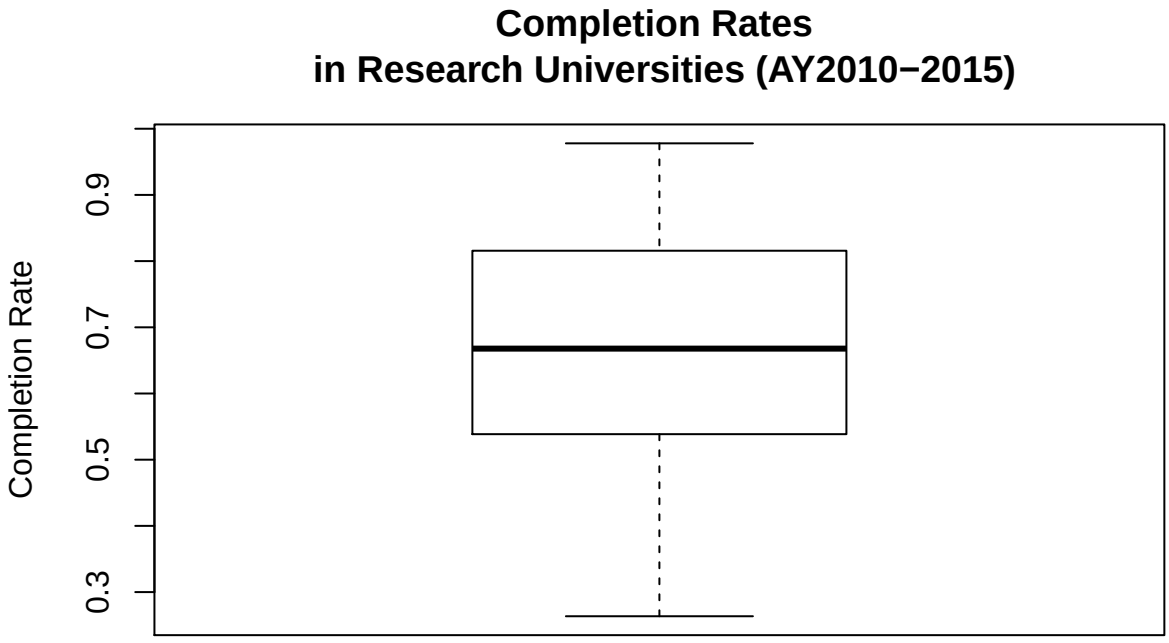

```
#Boxplot of SAT Average in all US Research Universities
boxplot(usresearchuniv$SAT_AVG_ALL, main = "SAT Averages \n in Research Universities (AY2010-2015)", yla
```



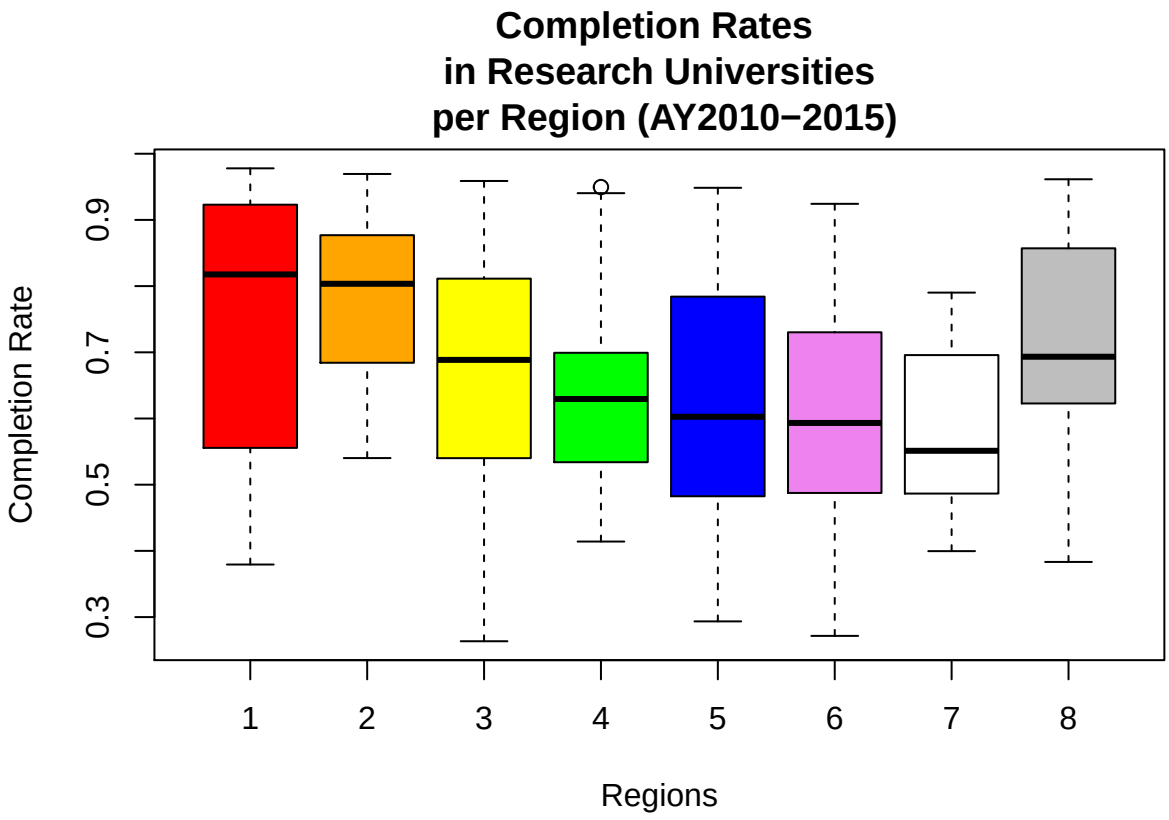

```
#Boxplot of admission rates in all US Research Universities
boxplot(usresearchuniv$ADM_RATE_ALL, main = "Admission Rates \n in Research Universities (AY2010-2015)"
```



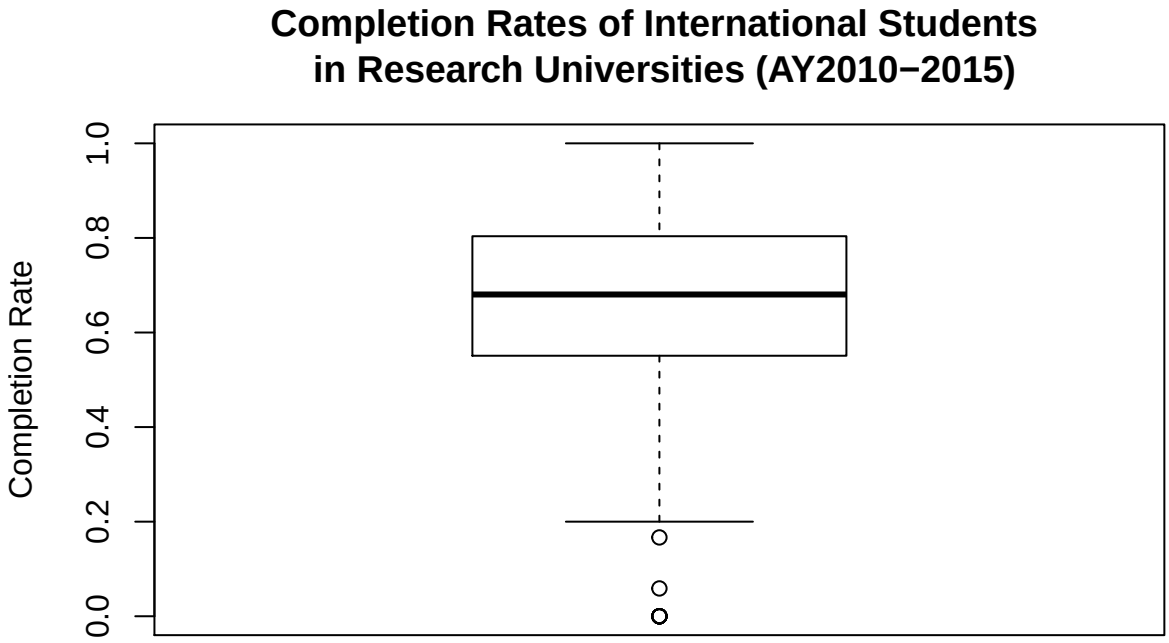

```
#Boxplot of Completion Rates in all US Research Universities
boxplot(usresearchuniv$C150_4, main = "Completion Rates \n in Research Universities (AY2010-2015)", ylab
```



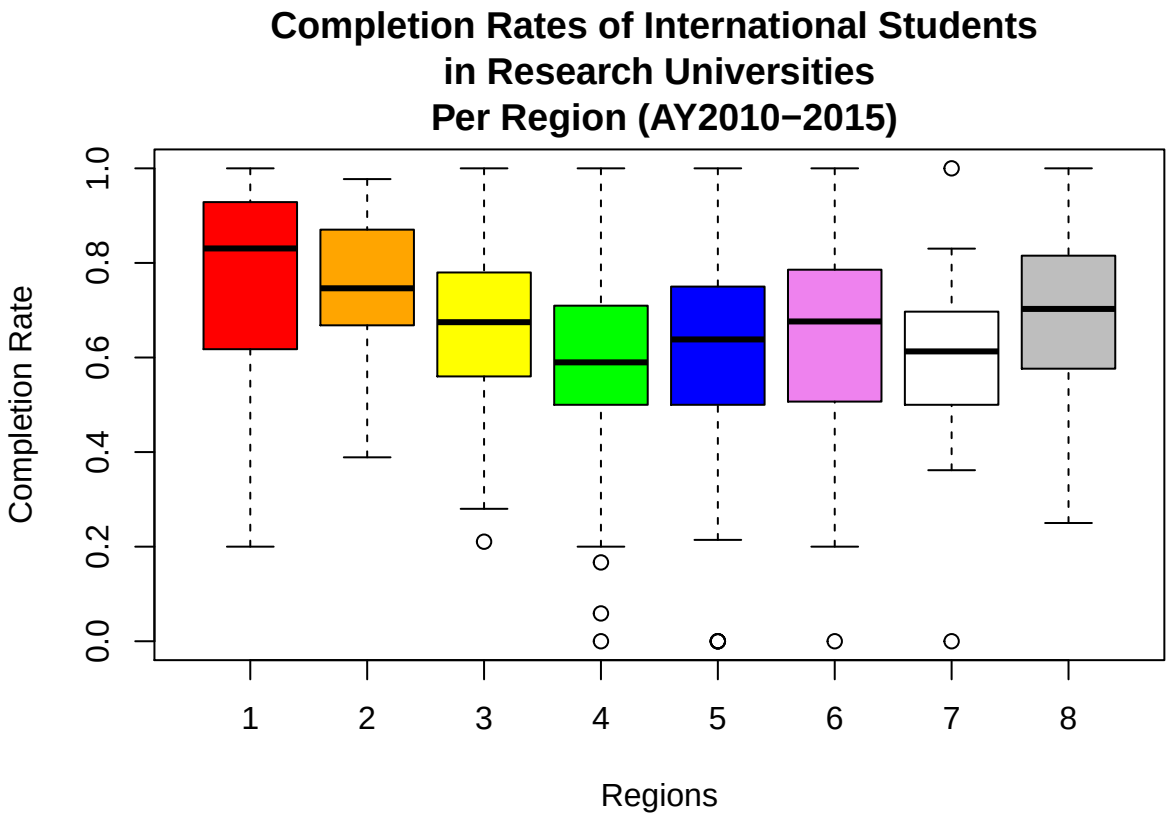

```r
# Boxplot of Completion Rates per Region in US Research Universities
boxplot(C150_4 ~ REGION, usresearchuniv, main = "Completion Rates \n in Research Universities \n per Reg
```




```
#Boxplot of Completion Rates of International Students in all US Research Universities
boxplot(usresearchuniv$C150_4_NRA, main = "Completion Rates of International Students \n in Research Uni
```




```r
# Boxplot of Completion Rates of International Students per Region in US Research Universities
boxplot(C150_4_NRA ~ REGION, usresearchuniv, main = "Completion Rates of International Students \n in R
```

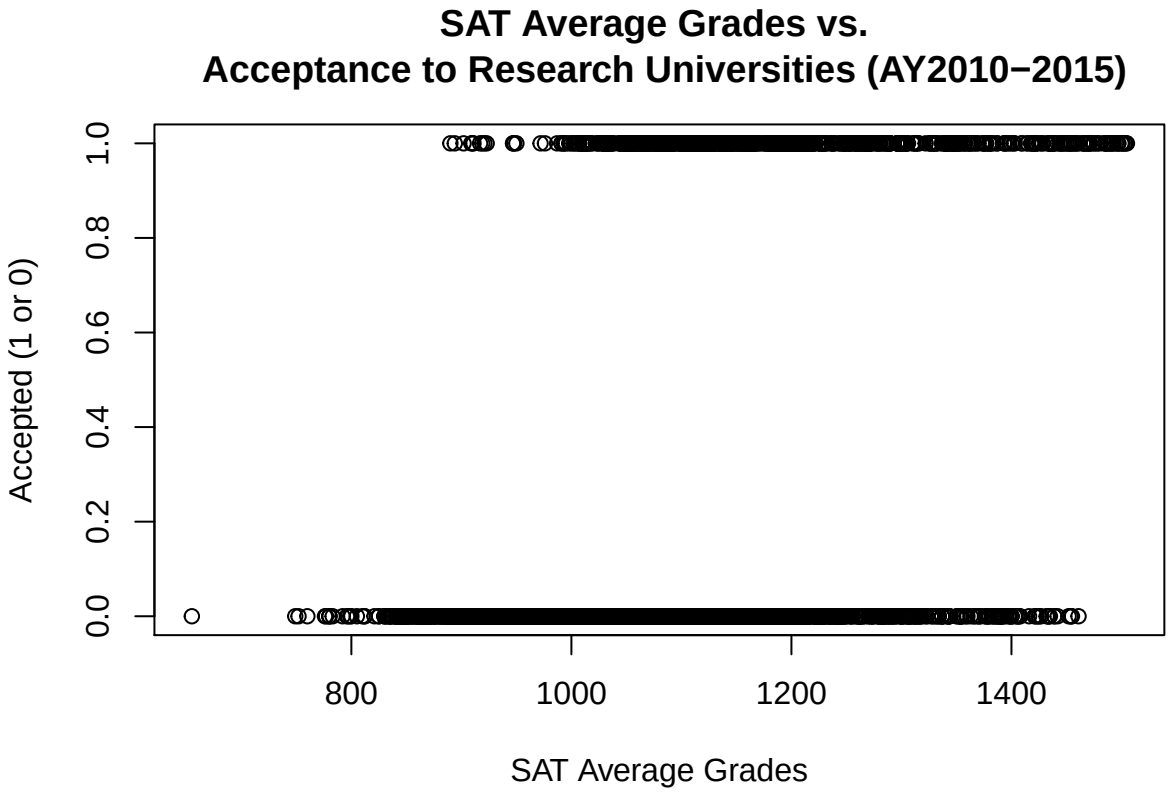



```
nrow(usresearchuniv[usresearchuniv$C150_4_NRA < 0.2,])
```

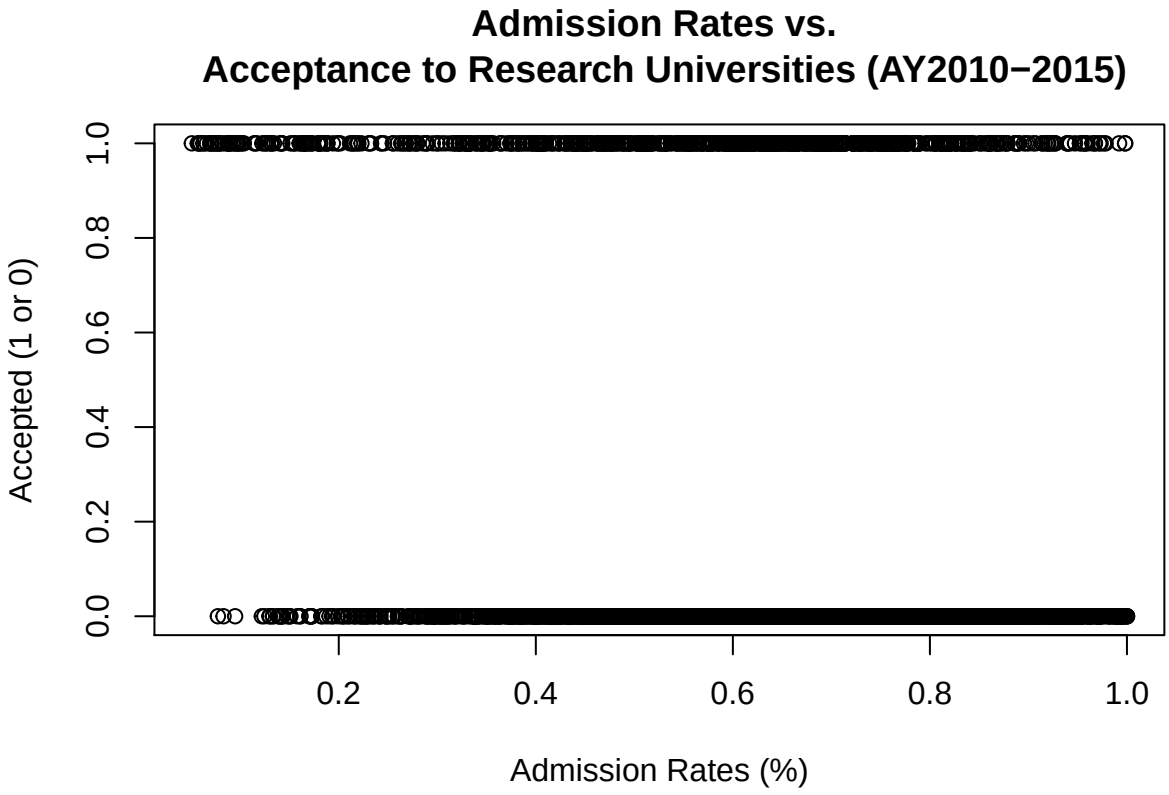
```
## [1] 9
```

Correlations

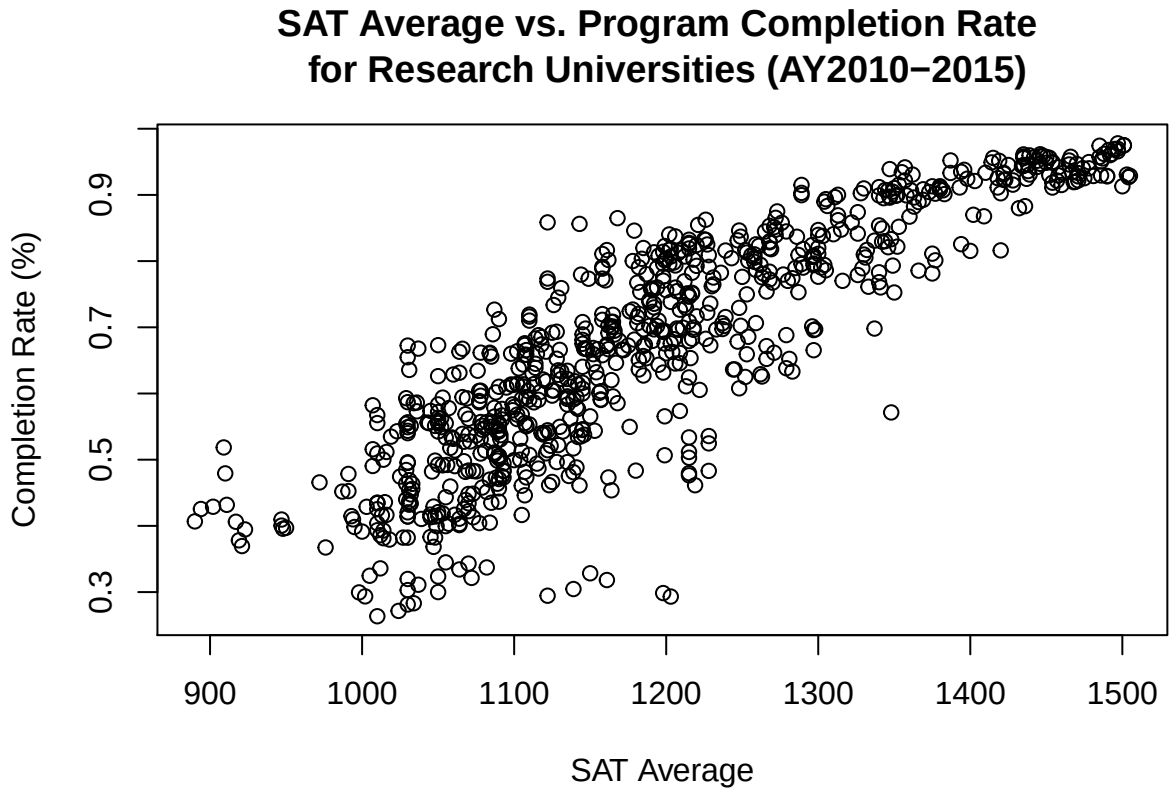
```
#Correlation between the SAT grades and the acceptance for the research universities
plot(usunivfilter$SAT_AVG_ALL, usunivfilter$ACCEPTED, main="SAT Average Grades vs. \n Acceptance to Res
```




```
#Correlation between the admission rates and the acceptance for the research universities
plot(usunivfilter$ADM_RATE_ALL, usunivfilter$ACCEPTED, main="Admission Rates vs. \n Acceptance to Resea
```




```
#Correlation between admission rate for research universities and program completion rate
plot(usresearchuniv$SAT_AVG_ALL, usresearchuniv$C150_4, main="SAT Average vs. Program Completion Rate \
```

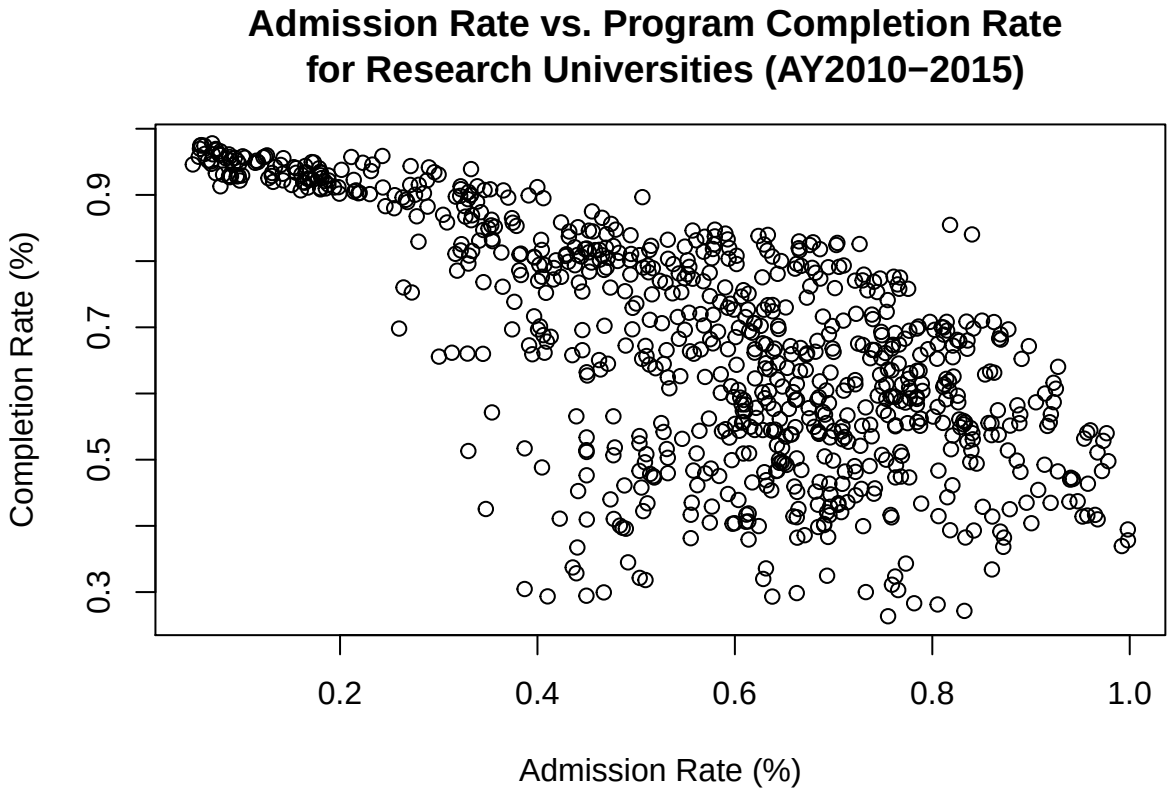



```
#Correlation coefficient between admission rate and completion rate
cor(usresearchuniv$SAT_AVG_ALL, usresearchuniv$C150_4, method = "pearson")
```

```
## [1] 0.8702261
```

This means that there is a strong positive correlation between the SAT average scores and the completion rate for all students.

```
#Correlation between admission rate for research universities and program completion rate
plot(usresearchuniv$ADM_RATE_ALL, usresearchuniv$C150_4, main="Admission Rate vs. Program Completion Ra
```

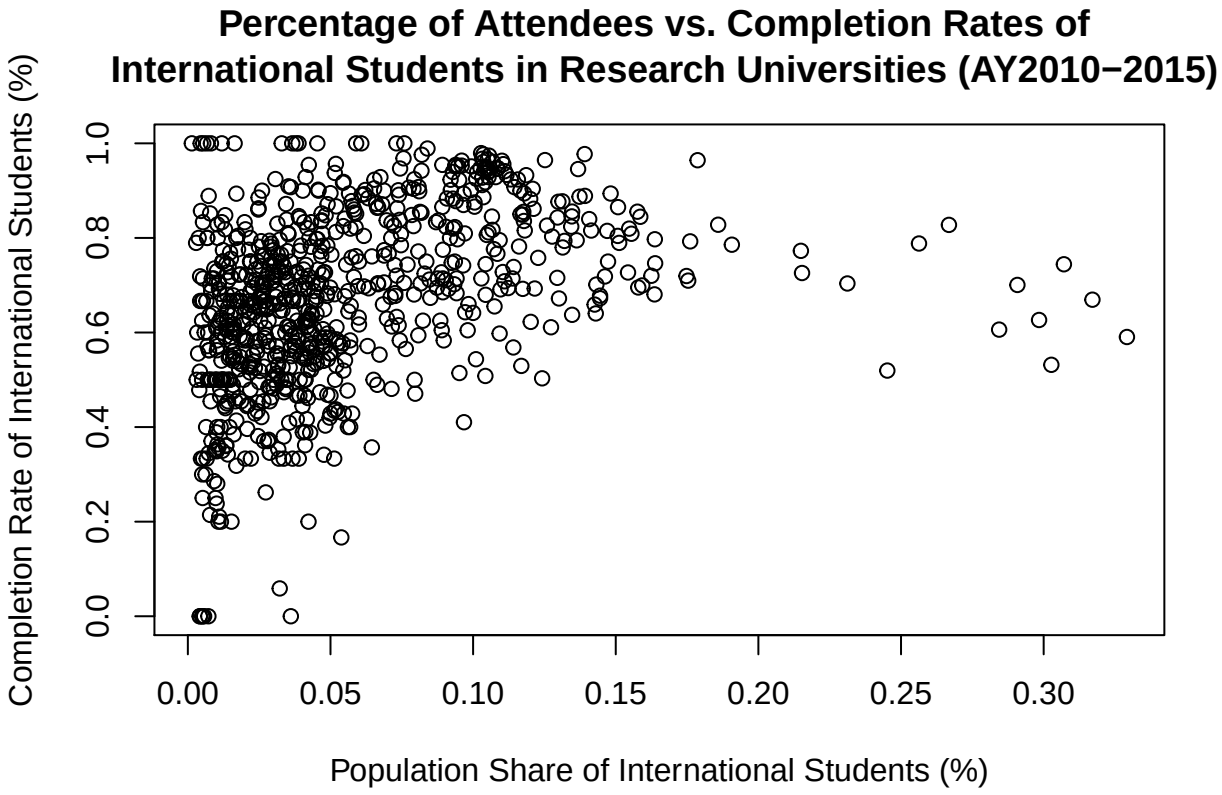



```
#Correlation coefficient between admission rate and completion rate
cor(usresearchuniv$ADM_RATE_ALL, usresearchuniv$C150_4, method = "pearson")
```

```
## [1] -0.6825525
```

This means that there is a strong negative correlation between the admission rates and the completion rates for the research universities.

```
#Correlation between attendees and completion rate of non-resident aliens (International Students)
plot(usresearchuniv$UGDS_NRA, usresearchuniv$C150_4_NRA, main="Percentage of Attendees vs. Completion Ra
```

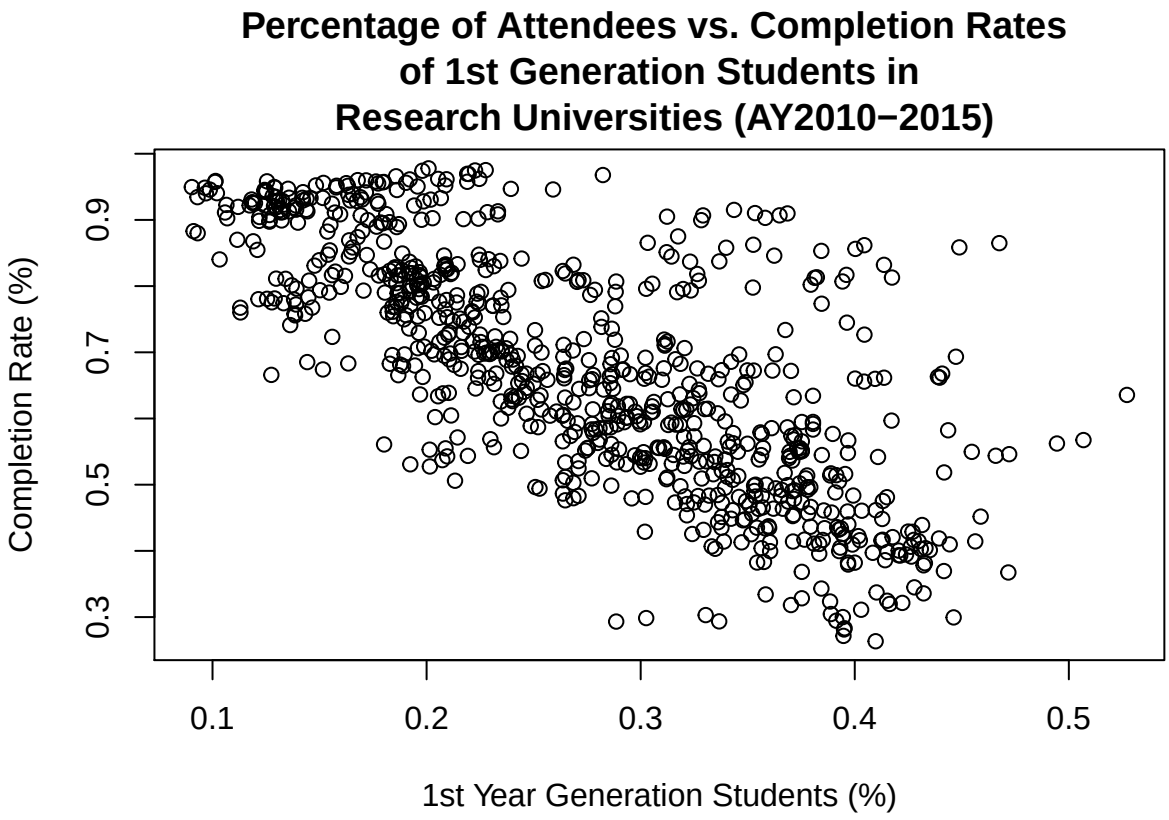



```
#Correlation coefficient between admission rate and completion rate of international students
cor(usresearchuniv$UGDS_NRA, usresearchuniv$C150_4_NRA, method = "pearson")
```

```
## [1] 0.370641
```

This means that there is a weak positive correlation between international student population and their completion rate.

```
#Correlation between attendees and completion rate of 1st Generation students in Research Universities
plot(usresearchuniv$PAR_ED_PCT_1STGEN, usresearchuniv$C150_4, main="Percentage of Attendees vs. Complet
```




```
#Correlation coefficient between admission rate and completion rate of 1st Generation students
cor(usresearchuniv$PAR_ED_PCT_1STGEN, usresearchuniv$C150_4, method = "pearson")
```

```
## [1] -0.7419477
```

This means that there is a strong negative correlation between 1st generation students and completion rates in research universities.

U.S. Research University Acceptance Model

In this report section, we are going to create a formula on getting an acceptance to a US Research University based on the College Scorecard statistics. We will try different methods of regression, and find the best regression technique from the following sources.

We will also consider another formula based on an international student taking up science degree/major.

```
# create a training and test model using a 75%/25% from the data set
rm_train <- sample(nrow(usunivfilter), floor(nrow(usunivfilter)*0.75))
univ_train <- usunivfilter[rm_train,]
univ_test <- usunivfilter[-rm_train,]

# create a generic formula for the US research university acceptance model for International Students b
formula_ISAcceptance <- formula(ACCEPTED ~ REGION + ADM_RATE_ALL + SAT_AVG_ALL + UGDS_NRA + COSTT4_A + I
```

We will do a generalized logistic regression formula.

```
# create a logistic regression
fit1 <- glm(formula_ISAcceptance, data = usunivfilter, family  = binomial())
summary(fit1)
```

```
##
## Call:
## glm(formula = formula_ISAcceptance, family = binomial(), data = usunivfilter)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2091  -0.5400  -0.2922  -0.1192   2.7993
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.478e+01  1.029e+00 -14.362  < 2e-16 ***
## REGION        1.246e-01  2.550e-02   4.886 1.03e-06 ***
## ADM_RATE_ALL  7.036e-01  3.297e-01   2.134   0.0328 *
## SAT_AVG_ALL   1.462e-02  7.312e-04  19.999  < 2e-16 ***
## UGDS_NRA      6.637e+00  1.147e+00   5.784 7.28e-09 ***
## COSTT4_A     -9.181e-05  5.441e-06 -16.872  < 2e-16 ***
## PCTFLOAN     -7.486e-01  4.247e-01  -1.763   0.0779 .
## UGDS_WOMEN   -1.995e+00  4.619e-01  -4.318 1.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4153.3  on 4246  degrees of freedom
## Residual deviance: 2838.4  on 4239  degrees of freedom
## AIC: 2854.4
##
## Number of Fisher Scoring iterations: 6
```

Based on the logistic regression, the formula will be

$$\frac{1}{1+e^{-x}}$$

where

$x = -14.8 + 0.125REGION + 0.704ADM\_RATE\_ALL + 0.0146SAT\_AVG\_ALL + 6.64UGDS\_NRA - 0.0000918COSTT4$

.

We will test this regression with some data types.

```
# this will not accept the person because of the SAT average
df_accept <- data.frame(REGION = 5, SAT_AVG_ALL = 900, ADM_RATE_ALL = .55, UGDS_NRA=.010, COSTT4_A = 20
predict(fit1, type = "response", newdata = df_accept)
```

```
##          1
## 0.03356807
```

```
# this will accept because of the SAT average and the cost
df_accept2 <- data.frame(REGION = 3, SAT_AVG_ALL = 1350, ADM_RATE_ALL = .35, UGDS_NRA=.25, COSTT4_A = 25
predict(fit1, type = "response", newdata = df_accept2)
```

```
##         1
## 0.9667774
```

Now, we will do some testing of performance with the logistic regression. Since we have split the dataset into training and testing set, we will see how the performance will be done.

```
# do a logistic regression model based on this
glm_ISAcceptance <- glm(formula_ISAcceptance, data = univ_train, family = binomial())
summary(glm_ISAcceptance)
```

```
##
## Call:
## glm(formula = formula_ISAcceptance, family = binomial(), data = univ_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1968  -0.5272  -0.2762  -0.1079   2.8655
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.543e+01  1.223e+00 -12.613  < 2e-16 ***
## REGION        1.335e-01  2.973e-02   4.491 7.10e-06 ***
## ADM_RATE_ALL  5.981e-01  3.866e-01   1.547    0.122
## SAT_AVG_ALL   1.543e-02  8.755e-04  17.622  < 2e-16 ***
## UGDS_NRA      6.534e+00  1.304e+00   5.010 5.44e-07 ***
## COSTT4_A     -9.760e-05  6.423e-06 -15.196  < 2e-16 ***
## PCTFLOAN     -5.260e-01  4.978e-01  -1.057    0.291
## UGDS_WOMEN   -2.374e+00  5.248e-01  -4.524 6.06e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3125.6  on 3184  degrees of freedom
## Residual deviance: 2085.6  on 3177  degrees of freedom
## AIC: 2101.6
##
## Number of Fisher Scoring iterations: 6
```

```
# do the first testing with the prediction model
accepted_ind <- predict(glm_ISAcceptance, type="response", newdata = univ_test)
pred1 <- prediction(accepted_ind, univ_test$ACCEPTED)

# create the confusion matrix and accuracy for this prediction model
c1 <- confusionMatrix(as.integer(accepted_ind > 0.5), univ_test$ACCEPTED)
c1$table
```

```
##           Reference
```

```
## Prediction   0   1
##          0 829 116
##          1  33  84
```

```
#Accuracy of the logistic regression model
c1$overall['Accuracy']
```

```
##  Accuracy
## 0.8596987
```

```
#Precision of the logistic regression model
c1$byClass['Neg Pred Value']
```

```
## Neg Pred Value
##      0.7179487
```

```
#Recall of the logistic regression model
c1$byClass['Specificity']
```

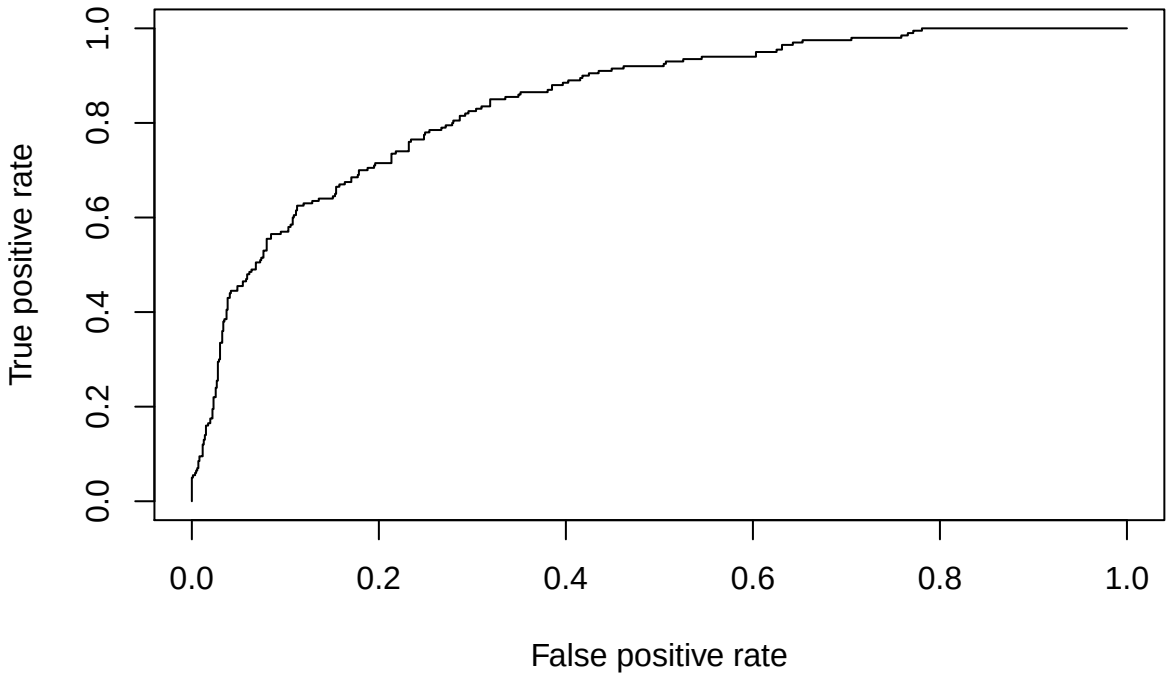
```
## Specificity
##        0.42
```

Accuracy shows the correct value. But in precision and recall, it is using "Neg Pred Value" and "Specificity" respectively. It should have been "Pos Pred Value" and "Sensitivity", as defined before. However, I manually calculated for the precision and recall for these values, and they are displayed correctly as it should be.

Precision: TP / (FP + TP) Recall: TP / (FN + TP)

As I show the precision and recall, it would be done the same thing, and verified manually that these are the correct percentages.

```
# show the curve on the performance
perf1 <- performance(pred1, "tpr", "fpr")
plot(perf1, lty = 1)
```




```
# Now we check on what acceptable ways we could do for regression
# doing single decision tree
model_dtree1 <- rpart(formula_ISAcceptance, method="anova",data = univ_train)
summary(model_dtree1)
```

```
## Call:
## rpart(formula = formula_ISAcceptance, data = univ_train, method = "anova")
##   n= 3185
##
##            CP nsplit rel error    xerror       xstd
## 1  0.15114670      0 1.0000000 1.0002558 0.02756113
## 2  0.05840615      1 0.8488533 0.8611606 0.02661874
## 3  0.03985339      3 0.7320410 0.7669612 0.02140972
## 4  0.03527832      4 0.6921876 0.7467538 0.02274813
## 5  0.02844596      5 0.6569093 0.7035155 0.02319756
## 6  0.02393626      6 0.6284633 0.6903009 0.02346670
## 7  0.01706800      7 0.6045271 0.6670968 0.02423084
## 8  0.01126752      8 0.5874591 0.6409908 0.02400885
## 9  0.01006366      9 0.5761916 0.6437463 0.02476801
## 10 0.01000000     10 0.5661279 0.6391691 0.02490182
##
## Variable importance
##  SAT_AVG_ALL     COSTT4_A     PCTFLOAN ADM_RATE_ALL   UGDS_WOMEN
##           36           25           10           10            9
##     UGDS_NRA       REGION
##            6            4
```

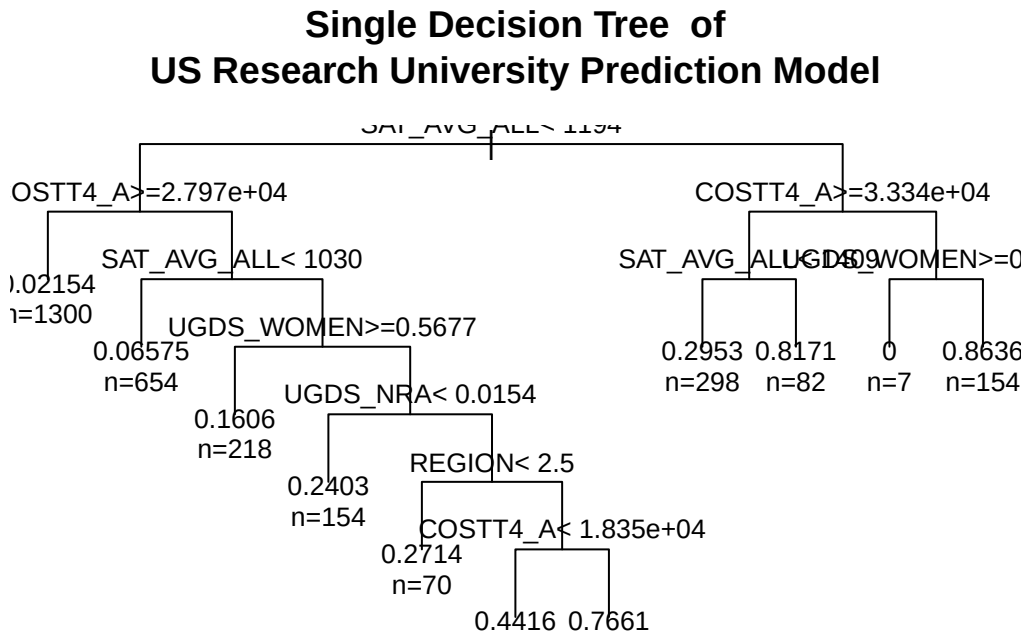
```
##
## Node number 1: 3185 observations,    complexity param=0.1511467
##   mean=0.1930926, MSE=0.1558079
##   left son=2 (2644 obs) right son=3 (541 obs)
##   Primary splits:
##       SAT_AVG_ALL  < 1194.5    to the left,  improve=0.15114670, (0 missing)
##       PCTFLOAN     < 0.49355   to the right, improve=0.12118250, (0 missing)
##       UGDS_WOMEN   < 0.52825   to the right, improve=0.09759761, (0 missing)
##       ADM_RATE_ALL < 0.20205   to the right, improve=0.05320142, (0 missing)
##       COSTT4_A     < 27966.5   to the right, improve=0.04026824, (0 missing)
##   Surrogate splits:
##       COSTT4_A     < 51237     to the left,  agree=0.896, adj=0.388, (0 split)
##       ADM_RATE_ALL < 0.3545465 to the right, agree=0.879, adj=0.287, (0 split)
##       PCTFLOAN     < 0.37295   to the right, agree=0.852, adj=0.128, (0 split)
##
## Node number 2: 2644 observations,    complexity param=0.05840615
##   mean=0.1236762, MSE=0.1083804
##   left son=4 (1300 obs) right son=5 (1344 obs)
##   Primary splits:
##       COSTT4_A    < 27966.5   to the right, improve=0.09310352, (0 missing)
##       PCTFLOAN    < 0.61485   to the right, improve=0.07318501, (0 missing)
##       UGDS_WOMEN  < 0.56775   to the right, improve=0.06186774, (0 missing)
##       SAT_AVG_ALL < 1028.5    to the left,  improve=0.04770672, (0 missing)
##       UGDS_NRA    < 0.02325   to the left,  improve=0.01398625, (0 missing)
##   Surrogate splits:
##       PCTFLOAN    < 0.62145   to the right, agree=0.702, adj=0.395, (0 split)
##       UGDS_WOMEN  < 0.62705   to the right, agree=0.601, adj=0.189, (0 split)
##       SAT_AVG_ALL < 1078.5    to the right, agree=0.579, adj=0.144, (0 split)
##       REGION      < 4.5       to the left,  agree=0.576, adj=0.138, (0 split)
##       UGDS_NRA    < 0.05695   to the right, agree=0.561, adj=0.106, (0 split)
##
## Node number 3: 541 observations,    complexity param=0.03985339
##   mean=0.5323475, MSE=0.2489536
##   left son=6 (380 obs) right son=7 (161 obs)
##   Primary splits:
##       COSTT4_A     < 33336     to the right, improve=0.14684130, (0 missing)
##       UGDS_WOMEN   < 0.5257    to the right, improve=0.11658930, (0 missing)
##       PCTFLOAN     < 0.48925   to the right, improve=0.07359364, (0 missing)
##       SAT_AVG_ALL  < 1443      to the left,  improve=0.06716186, (0 missing)
##       ADM_RATE_ALL < 0.1327    to the right, improve=0.05307765, (0 missing)
##   Surrogate splits:
##       UGDS_NRA     < 0.0156    to the right, agree=0.760, adj=0.193, (0 split)
##       SAT_AVG_ALL  < 1215.5    to the right, agree=0.719, adj=0.056, (0 split)
##       ADM_RATE_ALL < 0.6967171 to the left,  agree=0.708, adj=0.019, (0 split)
##
## Node number 4: 1300 observations
##   mean=0.02153846, MSE=0.02107456
##
## Node number 5: 1344 observations,    complexity param=0.05840615
##   mean=0.2224702, MSE=0.1729772
##   left son=10 (654 obs) right son=11 (690 obs)
##   Primary splits:
##       SAT_AVG_ALL < 1029.5    to the left,  improve=0.13458420, (0 missing)
##       UGDS_WOMEN  < 0.56565   to the right, improve=0.08742330, (0 missing)
```

```
##       COSTT4_A    < 17415.5   to the left,  improve=0.04975748, (0 missing)
##       PCTFLOAN    < 0.63165   to the right, improve=0.04002908, (0 missing)
##       UGDS_NRA    < 0.02325   to the left,  improve=0.03573591, (0 missing)
##   Surrogate splits:
##       UGDS_WOMEN < 0.5614    to the right, agree=0.638, adj=0.257, (0 split)
##       COSTT4_A   < 18512     to the left,  agree=0.609, adj=0.196, (0 split)
##       PCTFLOAN   < 0.65795   to the right, agree=0.578, adj=0.133, (0 split)
##       UGDS_NRA   < 0.01805   to the left,  agree=0.566, adj=0.109, (0 split)
##       REGION     < 4.5       to the right, agree=0.554, adj=0.084, (0 split)
##
## Node number 6: 380 observations,    complexity param=0.03527832
##   mean=0.4078947, MSE=0.2415166
##   left son=12 (298 obs) right son=13 (82 obs)
##   Primary splits:
##       SAT_AVG_ALL  < 1409      to the left,  improve=0.19075500, (0 missing)
##       ADM_RATE_ALL < 0.1327    to the right, improve=0.13948320, (0 missing)
##       UGDS_WOMEN   < 0.51245   to the right, improve=0.12944720, (0 missing)
##       UGDS_NRA     < 0.0737    to the left,  improve=0.11169260, (0 missing)
##       PCTFLOAN     < 0.41885   to the right, improve=0.06200991, (0 missing)
##   Surrogate splits:
##       ADM_RATE_ALL < 0.1709001 to the right, agree=0.942, adj=0.732, (0 split)
##       PCTFLOAN     < 0.2691    to the right, agree=0.879, adj=0.439, (0 split)
##       COSTT4_A     < 61397     to the left,  agree=0.808, adj=0.110, (0 split)
##
## Node number 7: 161 observations,    complexity param=0.01006366
##   mean=0.826087, MSE=0.1436673
##   left son=14 (7 obs) right son=15 (154 obs)
##   Primary splits:
##       UGDS_WOMEN < 0.5842    to the right, improve=0.21590910, (0 missing)
##       UGDS_NRA   < 0.0047    to the left,  improve=0.15028660, (0 missing)
##       COSTT4_A   < 19931.5   to the left,  improve=0.10486060, (0 missing)
##       REGION     < 4.5       to the left,  improve=0.10311430, (0 missing)
##       PCTFLOAN   < 0.4762    to the right, improve=0.07542965, (0 missing)
##
## Node number 10: 654 observations
##   mean=0.06574924, MSE=0.06142627
##
## Node number 11: 690 observations,    complexity param=0.02844596
##   mean=0.3710145, MSE=0.2333627
##   left son=22 (218 obs) right son=23 (472 obs)
##   Primary splits:
##       UGDS_WOMEN  < 0.56775   to the right, improve=0.08766753, (0 missing)
##       REGION      < 5.5       to the left,  improve=0.06760415, (0 missing)
##       PCTFLOAN    < 0.6226    to the right, improve=0.06424421, (0 missing)
##       COSTT4_A    < 17415.5   to the left,  improve=0.06214370, (0 missing)
##       SAT_AVG_ALL < 1089.5    to the left,  improve=0.05098514, (0 missing)
##   Surrogate splits:
##       COSTT4_A     < 14401.5   to the left,  agree=0.709, adj=0.078, (0 split)
##       SAT_AVG_ALL  < 1042.5    to the left,  agree=0.699, adj=0.046, (0 split)
##       UGDS_NRA     < 0.00225   to the left,  agree=0.696, adj=0.037, (0 split)
##       ADM_RATE_ALL < 0.29805   to the left,  agree=0.691, adj=0.023, (0 split)
##       PCTFLOAN     < 0.7931    to the right, agree=0.688, adj=0.014, (0 split)
##
## Node number 12: 298 observations
```

```
##   mean=0.295302, MSE=0.2080987
##
## Node number 13: 82 observations
##   mean=0.8170732, MSE=0.1494646
##
## Node number 14: 7 observations
##   mean=0, MSE=0
##
## Node number 15: 154 observations
##   mean=0.8636364, MSE=0.1177686
##
## Node number 22: 218 observations
##   mean=0.1605505, MSE=0.134774
##
## Node number 23: 472 observations,    complexity param=0.02393626
##   mean=0.4682203, MSE=0.2489901
##   left son=46 (154 obs) right son=47 (318 obs)
##   Primary splits:
##       UGDS_NRA    < 0.0154    to the left,  improve=0.10107210, (0 missing)
##       PCTFLOAN    < 0.61115   to the right, improve=0.08576401, (0 missing)
##       REGION      < 2.5       to the left,  improve=0.07438918, (0 missing)
##       UGDS_WOMEN  < 0.43185   to the left,  improve=0.06368816, (0 missing)
##       SAT_AVG_ALL < 1089.5    to the left,  improve=0.06130002, (0 missing)
##   Surrogate splits:
##       PCTFLOAN     < 0.64065   to the right, agree=0.703, adj=0.091, (0 split)
##       UGDS_WOMEN   < 0.1588    to the left,  agree=0.699, adj=0.078, (0 split)
##       REGION       < 1.5       to the left,  agree=0.693, adj=0.058, (0 split)
##       ADM_RATE_ALL < 0.9623977 to the right, agree=0.676, adj=0.006, (0 split)
##
## Node number 46: 154 observations
##   mean=0.2402597, MSE=0.182535
##
## Node number 47: 318 observations,    complexity param=0.017068
##   mean=0.5786164, MSE=0.2438195
##   left son=94 (70 obs) right son=95 (248 obs)
##   Primary splits:
##       REGION      < 2.5       to the left,  improve=0.10924100, (0 missing)
##       PCTFLOAN    < 0.60465   to the right, improve=0.07698848, (0 missing)
##       SAT_AVG_ALL < 1089.5    to the left,  improve=0.07048827, (0 missing)
##       UGDS_WOMEN  < 0.4311    to the left,  improve=0.06337372, (0 missing)
##       COSTT4_A    < 17552     to the left,  improve=0.04645596, (0 missing)
##   Surrogate splits:
##       PCTFLOAN    < 0.6576    to the right, agree=0.852, adj=0.329, (0 split)
##       COSTT4_A    < 24369     to the right, agree=0.805, adj=0.114, (0 split)
##       UGDS_WOMEN  < 0.42585   to the left,  agree=0.796, adj=0.071, (0 split)
##       SAT_AVG_ALL < 1188      to the right, agree=0.792, adj=0.057, (0 split)
##       UGDS_NRA    < 0.1451    to the right, agree=0.783, adj=0.014, (0 split)
##
## Node number 94: 70 observations
##   mean=0.2714286, MSE=0.1977551
##
## Node number 95: 248 observations,    complexity param=0.01126752
##   mean=0.6653226, MSE=0.2226684
##   left son=190 (77 obs) right son=191 (171 obs)
```

```
##   Primary splits:
##       COSTT4_A    < 18349     to the left,  improve=0.10125510, (0 missing)
##       SAT_AVG_ALL < 1074.5    to the left,  improve=0.09146780, (0 missing)
##       PCTFLOAN    < 0.31295   to the left,  improve=0.07486011, (0 missing)
##       UGDS_WOMEN  < 0.42955   to the left,  improve=0.05678224, (0 missing)
##       REGION      < 4.5       to the left,  improve=0.02863475, (0 missing)
##   Surrogate splits:
##       ADM_RATE_ALL < 0.9326655 to the right, agree=0.722, adj=0.104, (0 split)
##       PCTFLOAN     < 0.34845   to the left,  agree=0.710, adj=0.065, (0 split)
##       UGDS_WOMEN   < 0.19025   to the left,  agree=0.702, adj=0.039, (0 split)
##       UGDS_NRA     < 0.2577    to the right, agree=0.698, adj=0.026, (0 split)
##
## Node number 190: 77 observations
##   mean=0.4415584, MSE=0.2465846
##
## Node number 191: 171 observations
##   mean=0.7660819, MSE=0.1792004
```

```
plot(model_dtree1, uniform = TRUE, main = "Single Decision Tree  of\nUS Research University Prediction M
text(model_dtree1, use.n = TRUE, cex = .8)
```




```
pred_dtree1 <- predict(model_dtree1, newdata = univ_test)
accu1 <- abs(pred_dtree1 - univ_test$ACCEPTED) < 0.5
frac1 <- sum(accu1)/length(accu1)
print(frac1)
```

```
## [1] 0.8625235
```

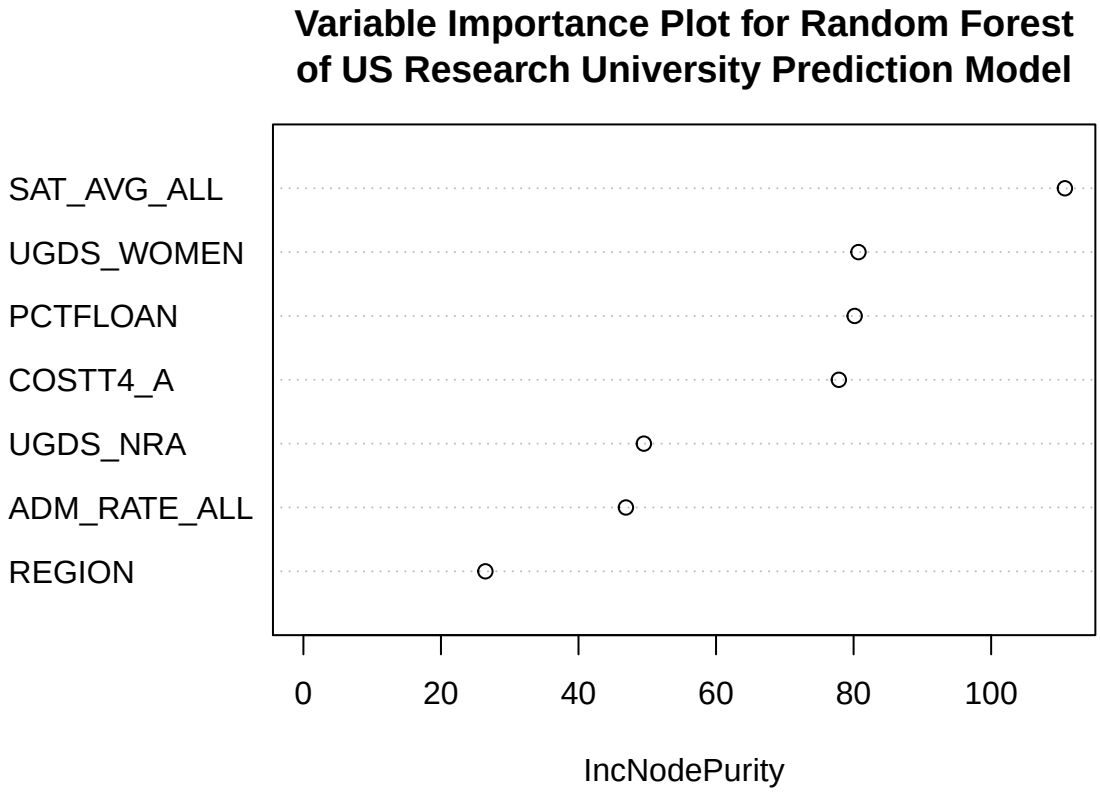
```
# doing random forest
model_forest1 <- randomForest(formula_ISAcceptance, data = univ_train)
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values. Are you sure you want to do regression?
```

```
summary(model_forest1)
```

```
##                 Length Class  Mode
## call               3   -none- call
## type               1   -none- character
## predicted       3185   -none- numeric
## mse              500   -none- numeric
## rsq              500   -none- numeric
## oob.times       3185   -none- numeric
## importance         7   -none- numeric
## importanceSD       0   -none- NULL
## localImportance    0   -none- NULL
## proximity          0   -none- NULL
## ntree              1   -none- numeric
## mtry               1   -none- numeric
## forest            11   -none- list
## coefs              0   -none- NULL
## y               3185   -none- numeric
## test               0   -none- NULL
## inbag              0   -none- NULL
## terms              3   terms  call
```

```
varImpPlot(model_forest1, main = "Variable Importance Plot for Random Forest\nof US Research University
```




```
pred_forest1 <- predict(model_forest1, newdata = univ_test)
accu2 <- abs(pred_forest1 - univ_test$ACCEPTED) < 0.5
frac2 <- sum(accu2)/length(accu2)
print(frac2)
```

```
## [1] 0.9303202
```

```
# doing support vector machine
model_svm1 <- svm(formula_ISAcceptance, data = univ_train)
summary(model_svm1)
```

```
##
## Call:
## svm(formula = formula_ISAcceptance, data = univ_train)
##
##
## Parameters:
##    SVM-Type:  eps-regression
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  0.1428571
##     epsilon:  0.1
##
##
## Number of Support Vectors:  1419
```

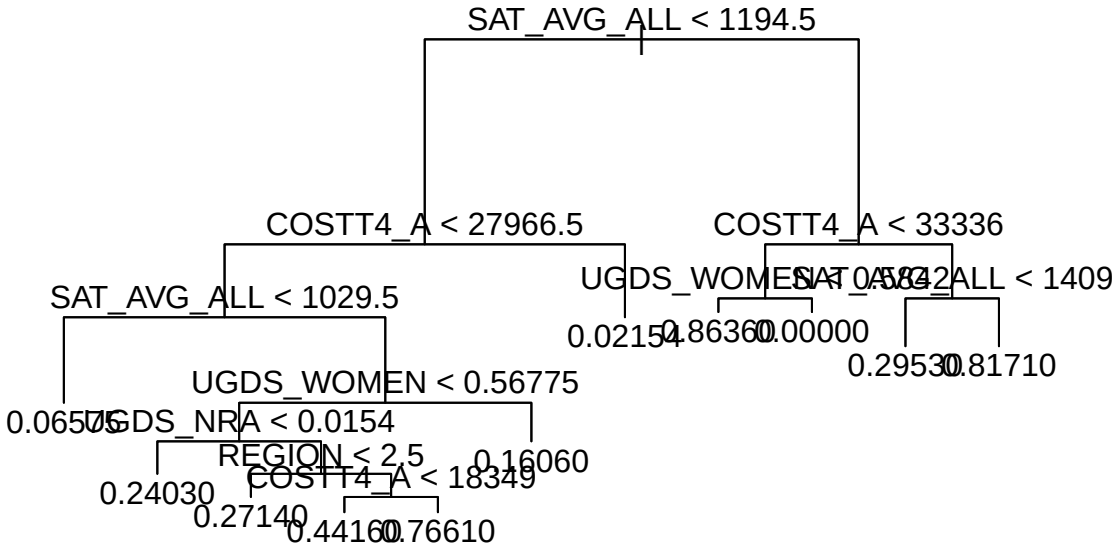
```r
pred_svm1 <- predict(model_svm1, newdata = univ_test)
accu3 <- abs(pred_svm1 - univ_test$ACCEPTED) < 0.5
frac3 <- sum(accu3)/length(accu3)
print(frac3)
```

```
## [1] 0.8832392
```

```r
# doing simple tree
model_tree1 <- tree(formula_ISAcceptance, data = univ_train)
summary(model_tree1)
```

```
##
## Regression tree:
## tree(formula = formula_ISAcceptance, data = univ_train)
## Variables actually used in tree construction:
## [1] "SAT_AVG_ALL" "COSTT4_A"    "UGDS_WOMEN"  "UGDS_NRA"    "REGION"
## Number of terminal nodes:  11
## Residual mean deviance:  0.08851 = 280.9 / 3174
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.86360 -0.06575 -0.02154  0.00000 -0.02154  0.97850
```

```r
plot(model_tree1, main = "Simple Tree of\nUS Research University Prediction Model")
text(model_tree1)
```



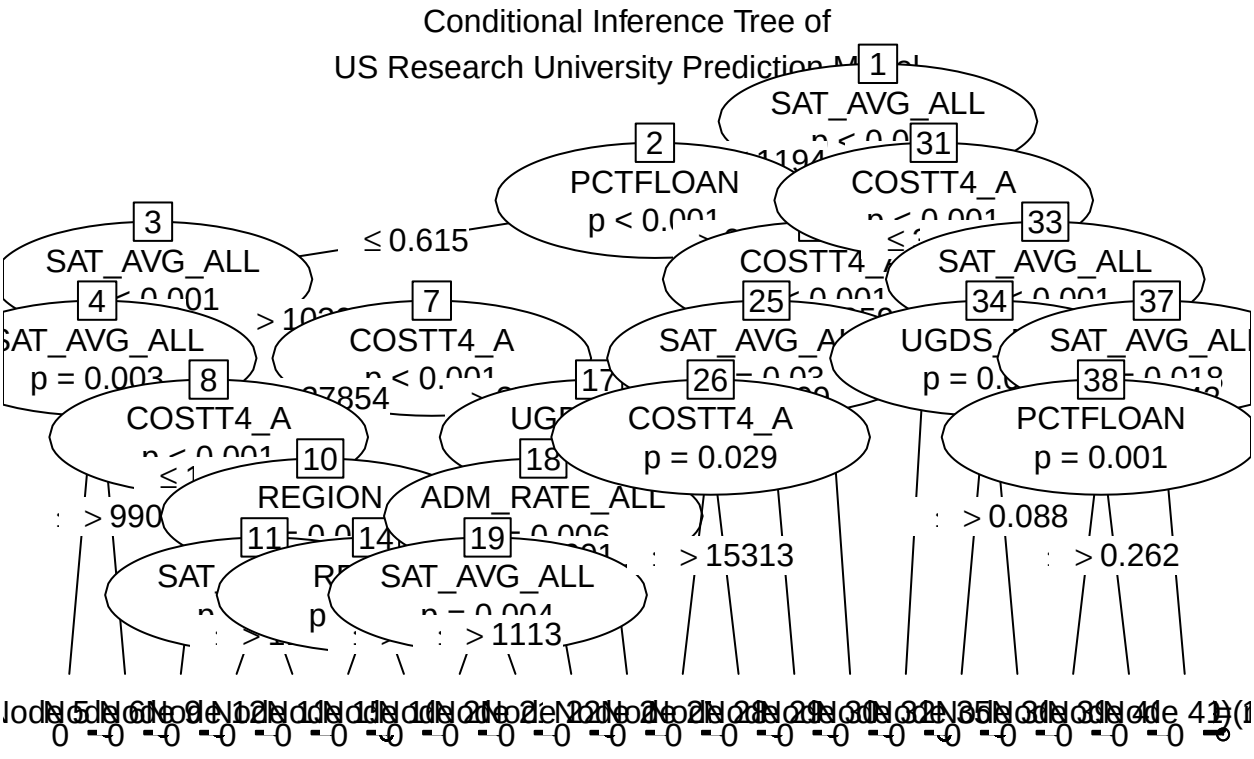
```
pred_tree1 <- predict(model_tree1, newdata = univ_test)
accu4 <- abs(pred_tree1 - univ_test$ACCEPTED) < 0.5
frac4 <- sum(accu4)/length(accu4)
print(frac4)
```

```
## [1] 0.8625235
```

```
# doing conditional inference tree
model_party1 <- ctree(formula_ISAcceptance, data = univ_train)
summary(model_party1)
```

```
##      Length      Class       Mode
##           1 BinaryTree         S4
```

```
plot(model_party1, main = "Conditional Inference Tree of\nUS Research University Prediction Model")
```




```
pred_party1 <- predict(model_party1, newdata = univ_test)
accu5 <- abs(pred_party1 - univ_test$ACCEPTED) < 0.5
frac5 <- sum(accu5)/length(accu5)
print(frac5)
```

```
## [1] 0.8709981
```

Based on the run, random forest is the best regression method to use in this model.

Next, another formula is created. This is an acceptance model for an international student that wants to take up Science degree/major

```
# create a formula for the US research university acceptance model for International Students taking up
formula_ISSciAcceptance <- formula(ACCEPTED ~ REGION + ADM_RATE_ALL + SAT_AVG_ALL + PCIP11 + PCIP12 + PC

# do a logistic regression model based on the formula created
glm_ISSciAcceptance <- glm(formula_ISSciAcceptance, data=univ_train,family=binomial())
summary(glm_ISSciAcceptance)
```

```
##
## Call:
## glm(formula = formula_ISSciAcceptance, family = binomial(), data = univ_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.59349  -0.46374  -0.22302  -0.06606   3.02515
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.869e+01  1.491e+00 -12.530  < 2e-16 ***
## REGION        1.470e-01  3.222e-02   4.564 5.01e-06 ***
## ADM_RATE_ALL  1.154e+00  4.341e-01   2.659 0.007831 **
## SAT_AVG_ALL   1.662e-02  1.077e-03  15.428  < 2e-16 ***
## PCIP11        1.644e+00  2.360e+00   0.697 0.486003
## PCIP12        1.347e+00  1.836e+01   0.073 0.941519
## PCIP14        5.987e+00  8.093e-01   7.398 1.39e-13 ***
## PCIP15        2.637e-02  2.264e+00   0.012 0.990706
## PCIP24       -6.865e+00  1.414e+00  -4.854 1.21e-06 ***
## PCIP26        7.250e+00  1.844e+00   3.932 8.42e-05 ***
## PCIP27       -2.405e+01  7.164e+00  -3.357 0.000788 ***
## PCIP40       -4.076e+01  5.041e+00  -8.086 6.15e-16 ***
## PCIP45        9.107e+00  1.262e+00   7.216 5.36e-13 ***
## PCIP51        2.094e+00  6.507e-01   3.218 0.001289 **
## PCIP52        1.037e+00  6.644e-01   1.562 0.118383
## UGDS_NRA      8.756e+00  1.490e+00   5.875 4.22e-09 ***
## UGDS_UNKN    -1.351e+00  1.575e+00  -0.858 0.391020
## COSTT4_A     -1.156e-04  7.541e-06 -15.326  < 2e-16 ***
## PCTFLOAN     -5.723e-01  5.794e-01  -0.988 0.323257
## UGDS_WOMEN    8.945e-02  7.771e-01   0.115 0.908356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3125.6  on 3184  degrees of freedom
## Residual deviance: 1813.8  on 3165  degrees of freedom
## AIC: 1853.8
##
## Number of Fisher Scoring iterations: 6
```

```r
# do the testing with the prediction model
accepted_ind2 <- predict(glm_ISSciAcceptance, type="response", newdata = univ_test)
pred2 <- prediction(accepted_ind2, univ_test$ACCEPTED)

# prepare confusion matrix and accuracy to see the scores
c2 <- confusionMatrix(as.integer(accepted_ind2 > 0.5), univ_test$ACCEPTED)
c2$table
```

```
##            Reference
## Prediction   0   1
##          0 814 100
##          1  48 100
```

```r
c2$overall['Accuracy']
```

```
##  Accuracy
## 0.8606403
```

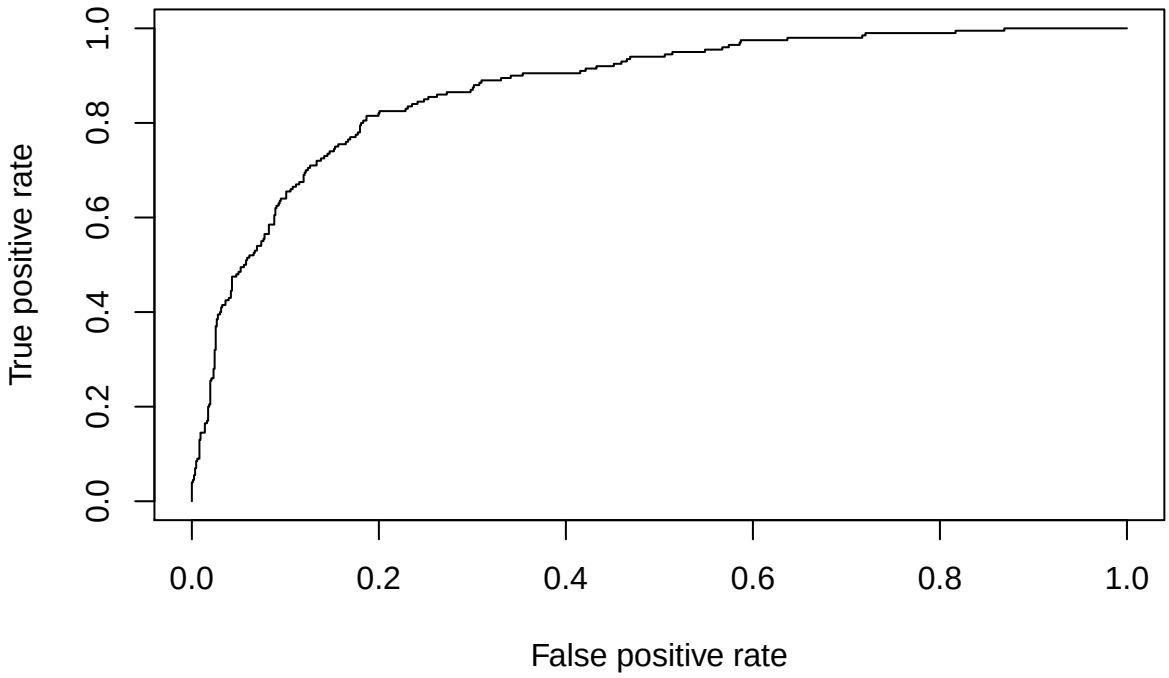
```r
#Precision of the logistic regression model
c2$byClass['Neg Pred Value']
```

```
## Neg Pred Value
##      0.6756757
```

```r
#Recall of the logistic regression model
c2$byClass['Specificity']
```

```
## Specificity
##         0.5
```

```r
# show the curve on the performance
perf2 <- performance(pred2,"tpr","fpr")
plot(perf2, lty = 1)
```




```
# Now we check on what acceptable ways we could do for regression
# doing single decision tree
model_dtree2 <- rpart(formula_ISSciAcceptance, method="anova",data = univ_train)
summary(model_dtree2)
```

```
## Call:
## rpart(formula = formula_ISSciAcceptance, data = univ_train, method = "anova")
##   n= 3185
##
##            CP nsplit rel error    xerror       xstd
## 1  0.32389751      0 1.0000000 1.0004844 0.02756746
## 2  0.06911055      1 0.6761025 0.6775320 0.02211370
## 3  0.04443134      2 0.6069919 0.6234205 0.02411208
## 4  0.03265113      3 0.5625606 0.5842627 0.02446834
## 5  0.01912956      4 0.5299095 0.5599525 0.02486940
## 6  0.01349312      5 0.5107799 0.5465998 0.02530315
## 7  0.01302300      6 0.4972868 0.5429956 0.02522273
## 8  0.01157068      7 0.4842638 0.5427670 0.02537074
## 9  0.01142749      8 0.4726931 0.5338634 0.02508416
## 10 0.01052931     11 0.4376201 0.5320077 0.02513873
## 11 0.01000000     12 0.4270908 0.5232169 0.02506642
##
## Variable importance
##       PCIP14  SAT_AVG_ALL     PCTFLOAN   UGDS_WOMEN       PCIP45
##           33           14            8            8            7
##     COSTT4_A ADM_RATE_ALL       PCIP26       PCIP51       REGION
```

```
##           6           6           3           2           2
##    UGDS_UNKN       PCIP24       PCIP52       PCIP11       PCIP27
##           2           1           1           1           1
##       PCIP40       PCIP15     UGDS_NRA
##           1           1           1
##
## Node number 1: 3185 observations,    complexity param=0.3238975
##   mean=0.1930926, MSE=0.1558079
##   left son=2 (2298 obs) right son=3 (887 obs)
##   Primary splits:
##       PCIP14      < 0.0269    to the left,  improve=0.32389750, (0 missing)
##       SAT_AVG_ALL < 1194.5    to the left,  improve=0.15114670, (0 missing)
##       PCTFLOAN    < 0.49355   to the right, improve=0.12118250, (0 missing)
##       UGDS_WOMEN  < 0.52825   to the right, improve=0.09759761, (0 missing)
##       PCIP45      < 0.04825   to the left,  improve=0.07446954, (0 missing)
##   Surrogate splits:
##       UGDS_WOMEN   < 0.52185   to the right, agree=0.784, adj=0.223, (0 split)
##       SAT_AVG_ALL  < 1178.5    to the left,  agree=0.751, adj=0.106, (0 split)
##       ADM_RATE_ALL < 0.1685543 to the right, agree=0.733, adj=0.043, (0 split)
##       PCIP11       < 0.0787    to the left,  agree=0.728, adj=0.024, (0 split)
##       COSTT4_A     < 60088     to the left,  agree=0.724, adj=0.010, (0 split)
##
## Node number 2: 2298 observations,    complexity param=0.013023
##   mean=0.0535248, MSE=0.0506599
##   left son=4 (2287 obs) right son=5 (11 obs)
##   Primary splits:
##       PCIP45      < 0.3399    to the left,  improve=0.05551311, (0 missing)
##       SAT_AVG_ALL < 1194.5    to the left,  improve=0.05178677, (0 missing)
##       PCTFLOAN    < 0.61455   to the right, improve=0.03271923, (0 missing)
##       PCIP14      < 0.00465   to the left,  improve=0.02897509, (0 missing)
##       COSTT4_A    < 53532     to the left,  improve=0.02002876, (0 missing)
##   Surrogate splits:
##       SAT_AVG_ALL < 1461      to the left,  agree=0.996, adj=0.182, (0 split)
##       COSTT4_A    < 62014.5   to the left,  agree=0.996, adj=0.091, (0 split)
##
## Node number 3: 887 observations,    complexity param=0.06911055
##   mean=0.5546787, MSE=0.2470102
##   left son=6 (411 obs) right son=7 (476 obs)
##   Primary splits:
##       PCTFLOAN    < 0.51445   to the right, improve=0.1565325, (0 missing)
##       PCIP45      < 0.0324    to the left,  improve=0.1444779, (0 missing)
##       PCIP26      < 0.02745   to the left,  improve=0.1130469, (0 missing)
##       SAT_AVG_ALL < 1197.5    to the left,  improve=0.1118886, (0 missing)
##       PCIP40      < 0.00495   to the left,  improve=0.0896419, (0 missing)
##   Surrogate splits:
##       PCIP45       < 0.03925   to the left,  agree=0.703, adj=0.360, (0 split)
##       SAT_AVG_ALL  < 1121.5    to the left,  agree=0.692, adj=0.336, (0 split)
##       ADM_RATE_ALL < 0.60835   to the right, agree=0.691, adj=0.333, (0 split)
##       PCIP26       < 0.04765   to the left,  agree=0.641, adj=0.226, (0 split)
##       REGION       < 4.5       to the left,  agree=0.634, adj=0.209, (0 split)
##
## Node number 4: 2287 observations,    complexity param=0.01142749
##   mean=0.04984696, MSE=0.04736224
##   left son=8 (2064 obs) right son=9 (223 obs)
```

```
##   Primary splits:
##       SAT_AVG_ALL < 1194.5    to the left,  improve=0.03566706, (0 missing)
##       PCIP14      < 0.00465   to the left,  improve=0.03361250, (0 missing)
##       PCIP45      < 0.04985   to the left,  improve=0.02991969, (0 missing)
##       PCTFLOAN    < 0.61455   to the right, improve=0.02903460, (0 missing)
##       COSTT4_A    < 26319     to the right, improve=0.01925006, (0 missing)
##   Surrogate splits:
##       COSTT4_A     < 53312.5   to the left,  agree=0.938, adj=0.368, (0 split)
##       PCIP45       < 0.2007    to the left,  agree=0.934, adj=0.327, (0 split)
##       ADM_RATE_ALL < 0.3310702 to the right, agree=0.923, adj=0.211, (0 split)
##       PCIP40       < 0.06815   to the left,  agree=0.913, adj=0.108, (0 split)
##       PCIP52       < 0.0028    to the right, agree=0.909, adj=0.067, (0 split)
##
## Node number 5: 11 observations
##   mean=0.8181818, MSE=0.1487603
##
## Node number 6: 411 observations,    complexity param=0.03265113
##   mean=0.3430657, MSE=0.2253716
##   left son=12 (215 obs) right son=13 (196 obs)
##   Primary splits:
##       COSTT4_A < 26509.5   to the right, improve=0.17492660, (0 missing)
##       PCIP45   < 0.0324    to the left,  improve=0.12950830, (0 missing)
##       PCTFLOAN < 0.64055   to the right, improve=0.10333630, (0 missing)
##       PCIP26   < 0.0198    to the left,  improve=0.08252529, (0 missing)
##       PCIP40   < 0.0034    to the left,  improve=0.07570640, (0 missing)
##   Surrogate splits:
##       SAT_AVG_ALL  < 1080.5    to the right, agree=0.703, adj=0.378, (0 split)
##       PCIP15       < 0.0053    to the left,  agree=0.645, adj=0.255, (0 split)
##       UGDS_NRA     < 0.01875   to the right, agree=0.642, adj=0.250, (0 split)
##       UGDS_UNKN    < 0.04425   to the right, agree=0.640, adj=0.245, (0 split)
##       ADM_RATE_ALL < 0.828083  to the left,  agree=0.637, adj=0.240, (0 split)
##
## Node number 7: 476 observations,    complexity param=0.04443134
##   mean=0.737395, MSE=0.1936436
##   left son=14 (86 obs) right son=15 (390 obs)
##   Primary splits:
##       SAT_AVG_ALL < 1075.5    to the left,  improve=0.23920930, (0 missing)
##       PCIP26      < 0.02865   to the left,  improve=0.12458880, (0 missing)
##       PCIP24      < 0.02745   to the right, improve=0.10268900, (0 missing)
##       COSTT4_A    < 19673.5   to the left,  improve=0.09578348, (0 missing)
##       PCIP14      < 0.05145   to the left,  improve=0.07971109, (0 missing)
##   Surrogate splits:
##       PCIP26     < 0.0255    to the left,  agree=0.845, adj=0.140, (0 split)
##       PCIP24     < 0.1415    to the right, agree=0.842, adj=0.128, (0 split)
##       COSTT4_A   < 16881.5   to the left,  agree=0.840, adj=0.116, (0 split)
##       UGDS_WOMEN < 0.6106    to the right, agree=0.832, adj=0.070, (0 split)
##       PCIP51     < 0.2316    to the right, agree=0.830, adj=0.058, (0 split)
##
## Node number 8: 2064 observations
##   mean=0.03633721, MSE=0.03501682
##
## Node number 9: 223 observations,    complexity param=0.01142749
##   mean=0.1748879, MSE=0.1443021
##   left son=18 (185 obs) right son=19 (38 obs)
```

```
##   Primary splits:
##       PCIP51   < 0.03005   to the left,  improve=0.23239660, (0 missing)
##       PCIP52   < 0.13035   to the left,  improve=0.22262050, (0 missing)
##       PCIP14   < 6e-04     to the left,  improve=0.14101150, (0 missing)
##       PCIP40   < 0.02755   to the right, improve=0.12033630, (0 missing)
##       COSTT4_A < 28751.5   to the right, improve=0.09170619, (0 missing)
##   Surrogate splits:
##       SAT_AVG_ALL < 1196      to the right, agree=0.843, adj=0.079, (0 split)
##
## Node number 12: 215 observations,    complexity param=0.01349312
##   mean=0.1534884, MSE=0.1299297
##   left son=24 (190 obs) right son=25 (25 obs)
##   Primary splits:
##       COSTT4_A     < 51693.5   to the left,  improve=0.2396980, (0 missing)
##       ADM_RATE_ALL < 0.5755735 to the right, improve=0.2209638, (0 missing)
##       PCIP45       < 0.11255   to the left,  improve=0.2194046, (0 missing)
##       SAT_AVG_ALL  < 1337      to the left,  improve=0.1818506, (0 missing)
##       PCIP26       < 0.118     to the left,  improve=0.1347889, (0 missing)
##   Surrogate splits:
##       SAT_AVG_ALL  < 1264.5    to the left,  agree=0.935, adj=0.44, (0 split)
##       ADM_RATE_ALL < 0.4055649 to the right, agree=0.898, adj=0.12, (0 split)
##       PCIP45       < 0.23265   to the left,  agree=0.893, adj=0.08, (0 split)
##       PCIP27       < 0.0431    to the left,  agree=0.888, adj=0.04, (0 split)
##       PCIP52       < 0.01985   to the right, agree=0.888, adj=0.04, (0 split)
##
## Node number 13: 196 observations,    complexity param=0.01912956
##   mean=0.5510204, MSE=0.2473969
##   left son=26 (30 obs) right son=27 (166 obs)
##   Primary splits:
##       PCIP45 < 0.01385   to the left,  improve=0.1957733, (0 missing)
##       PCIP51 < 3e-04     to the left,  improve=0.1732439, (0 missing)
##       PCIP26 < 0.01465   to the left,  improve=0.1712474, (0 missing)
##       PCIP40 < 0.00195   to the left,  improve=0.1306275, (0 missing)
##       PCIP27 < 0.00215   to the left,  improve=0.1153012, (0 missing)
##   Surrogate splits:
##       PCIP27 < 0.00045   to the left,  agree=0.929, adj=0.533, (0 split)
##       PCIP26 < 0.01465   to the left,  agree=0.918, adj=0.467, (0 split)
##       PCIP40 < 0.00195   to the left,  agree=0.908, adj=0.400, (0 split)
##       PCIP51 < 3e-04     to the left,  agree=0.893, adj=0.300, (0 split)
##       PCIP11 < 0.00135   to the left,  agree=0.878, adj=0.200, (0 split)
##
## Node number 14: 86 observations
##   mean=0.2790698, MSE=0.2011898
##
## Node number 15: 390 observations,    complexity param=0.01157068
##   mean=0.8384615, MSE=0.1354438
##   left son=30 (8 obs) right son=31 (382 obs)
##   Primary splits:
##       UGDS_UNKN  < 0.00035   to the left,  improve=0.10870110, (0 missing)
##       PCIP52     < 0.3127    to the right, improve=0.09486510, (0 missing)
##       UGDS_WOMEN < 0.27015   to the left,  improve=0.05980446, (0 missing)
##       PCIP51     < 0.0054    to the left,  improve=0.05269858, (0 missing)
##       UGDS_NRA   < 0.09835   to the left,  improve=0.04145392, (0 missing)
##   Surrogate splits:
```

```
##       UGDS_WOMEN < 0.10785   to the left,  agree=0.982, adj=0.125, (0 split)
##
## Node number 18: 185 observations
##   mean=0.09189189, MSE=0.08344777
##
## Node number 19: 38 observations,    complexity param=0.01142749
##   mean=0.5789474, MSE=0.2437673
##   left son=38 (20 obs) right son=39 (18 obs)
##   Primary splits:
##       PCTFLOAN     < 0.42265   to the right, improve=0.6545455, (0 missing)
##       ADM_RATE_ALL < 0.5723214 to the right, improve=0.5887446, (0 missing)
##       PCIP24       < 0.00035   to the left,  improve=0.4090909, (0 missing)
##       PCIP40       < 0.044     to the right, improve=0.3666667, (0 missing)
##       PCIP45       < 0.0723    to the left,  improve=0.3360795, (0 missing)
##   Surrogate splits:
##       ADM_RATE_ALL < 0.5723214 to the right, agree=0.868, adj=0.722, (0 split)
##       REGION       < 4.5       to the left,  agree=0.842, adj=0.667, (0 split)
##       PCIP24       < 0.00035   to the left,  agree=0.816, adj=0.611, (0 split)
##       SAT_AVG_ALL  < 1288      to the left,  agree=0.763, adj=0.500, (0 split)
##       PCIP45       < 0.13095   to the left,  agree=0.711, adj=0.389, (0 split)
##
## Node number 24: 190 observations
##   mean=0.08947368, MSE=0.08146814
##
## Node number 25: 25 observations
##   mean=0.64, MSE=0.2304
##
## Node number 26: 30 observations
##   mean=0.03333333, MSE=0.03222222
##
## Node number 27: 166 observations
##   mean=0.6445783, MSE=0.2290971
##
## Node number 30: 8 observations
##   mean=0, MSE=0
##
## Node number 31: 382 observations,    complexity param=0.01052931
##   mean=0.8560209, MSE=0.1232491
##   left son=62 (7 obs) right son=63 (375 obs)
##   Primary splits:
##       PCIP52     < 0.3127    to the right, improve=0.11098180, (0 missing)
##       UGDS_WOMEN < 0.27015   to the left,  improve=0.04535685, (0 missing)
##       PCIP51     < 0.0026    to the left,  improve=0.04307904, (0 missing)
##       PCIP26     < 0.02355   to the left,  improve=0.04015713, (0 missing)
##       UGDS_NRA   < 0.09835   to the left,  improve=0.03570491, (0 missing)
##
## Node number 38: 20 observations
##   mean=0.2, MSE=0.16
##
## Node number 39: 18 observations
##   mean=1, MSE=0
##
## Node number 62: 7 observations
##   mean=0, MSE=0
```

```
##
## Node number 63: 375 observations
##   mean=0.872, MSE=0.111616
```

```
pred_dtree2 <- predict(model_dtree2, newdata = univ_test)
accu6 <- abs(pred_dtree2 - univ_test$ACCEPTED) < 0.5
frac6 <- sum(accu6)/length(accu6)
print(frac6)
```

```
## [1] 0.9001883
```

```
# doing random forest
model_forest2 <- randomForest(formula_ISSciAcceptance, data = univ_train)
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values. Are you sure you want to do regression?
```

```
summary(model_forest2)
```

```
##                 Length Class  Mode
## call               3   -none- call
## type               1   -none- character
## predicted       3185   -none- numeric
## mse              500   -none- numeric
## rsq              500   -none- numeric
## oob.times       3185   -none- numeric
## importance        19   -none- numeric
## importanceSD       0   -none- NULL
## localImportance    0   -none- NULL
## proximity          0   -none- NULL
## ntree              1   -none- numeric
## mtry               1   -none- numeric
## forest            11   -none- list
## coefs              0   -none- NULL
## y               3185   -none- numeric
## test               0   -none- NULL
## inbag              0   -none- NULL
## terms              3   terms  call
```

```
pred_forest2 <- predict(model_forest2, newdata = univ_test)
accu7 <- abs(pred_forest2 - univ_test$ACCEPTED) < 0.5
frac7 <- sum(accu7)/length(accu7)
print(frac7)
```

```
## [1] 0.9519774
```

```
# doing support vector machine
model_svm2 <- svm(formula_ISSciAcceptance, data = univ_train)
summary(model_svm2)
```

```
##
## Call:
## svm(formula = formula_ISSciAcceptance, data = univ_train)
##
##
## Parameters:
##    SVM-Type:  eps-regression
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  0.05263158
##     epsilon:  0.1
##
##
## Number of Support Vectors:  1660
```

```
pred_svm2 <- predict(model_svm2, newdata = univ_test)
accu8 <- abs(pred_svm2 - univ_test$ACCEPTED) < 0.5
frac8 <- sum(accu8)/length(accu8)
print(frac8)
```

```
## [1] 0.9039548
```

```
# doing simple tree
model_tree2 <- tree(formula_ISSciAcceptance, data = univ_train)
summary(model_tree2)
```

```
##
## Regression tree:
## tree(formula = formula_ISSciAcceptance, data = univ_train)
## Variables actually used in tree construction:
## [1] "PCIP14"      "PCIP45"      "PCTFLOAN"    "SAT_AVG_ALL" "UGDS_UNKN"
## [6] "PCIP52"      "COSTT4_A"
## Number of terminal nodes:  10
## Residual mean deviance:  0.07224 = 229.3 / 3175
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.87200 -0.04985 -0.04985  0.00000 -0.04985  0.96670
```

```
pred_tree2 <- predict(model_tree2, newdata = univ_test)
accu9 <- abs(pred_tree2 - univ_test$ACCEPTED) < 0.5
frac9 <- sum(accu9)/length(accu9)
print(frac9)
```

```
## [1] 0.8964218
```

```
# doing conditional inference tree
model_party2 <- ctree(formula_ISSciAcceptance, data = univ_train)
summary(model_party2)
```

```
##     Length      Class       Mode
##          1 BinaryTree         S4
```

```
pred_party2 <- predict(model_party2, newdata = univ_test)
accu10 <- abs(pred_party2 - univ_test$ACCEPTED) < 0.5
frac10 <- sum(accu10)/length(accu10)
print(frac10)
```

```
## [1] 0.886064
```

Based on this, random forest is the best regression method to use.

In this project, I have selected a couple of variables that we could use in this model. However, we could use more than a few variables to get the optimal result.

With this in mind, feature selection is very essential, especially with datasets that have many variables for model selection. Although in this report, we have 1745 variables, and deduced it to 72 variables, we have to check which variables will be very useful in doing our research model.

In this portion, we will consider all variables, and use Boruta and RFE to use what variables we could use for doing a better outcome of the model.

Boruta is a package created was written by Miron B. Kursa and Witold R. Rudnicki to use an all relevant feature selection wrapper algorithm. According to their description, it "finds relevant features by comparing original attributes' importance with importance achievable at random, estimated using their permuted copies". (Source: https://cran.r-project.org/web/packages/Boruta/Boruta.pdf)

The Recursive Feature Elimination, or RFE, is a function in R's Caret package that uses the random forest algorithm to evaluate the attributes needed to be able to get an optimal result in the data that we have. (Source: http://machinelearningmastery.com/feature-selection-with-the-caret-r-package/)

Now, we will be doing some feature eliminations using Boruta and RFE.

```
# First, we will create another copy of the dataset
usunivnoccbasic <- usunivfilter

# Next, we will change those that have "NA" to 0, since there is no data in it
usunivnoccbasic[usunivnoccbasic == "NA"] <- 0

# Next, we will choose rows that have complete cases
usunivnoccbasic <- usunivnoccbasic[complete.cases(usunivnoccbasic),]

# Now that we have the cleansed dataset, we will implement Boruta
boruta.train <- Boruta(ACCEPTED ~ .-CCBASIC2, data=usunivnoccbasic)
print(boruta.train)
```

```
## Boruta performed 99 iterations in 30.57906 secs.
##  61 attributes confirmed important: ADM_RATE, ADM_RATE_ALL,
## C150_4, C150_4_2MOR, C150_4_AIAN and 56 more.
##  7 attributes confirmed unimportant: C150_4_NHPI, PCIP12, PCIP25,
## PCIP29, PCIP46 and 2 more.
##  2 tentative attributes left: PCIP10, PCIP22.
```

```
getSelectedAttributes(boruta.train)
```

```
##  [1] "REGION"           "ADM_RATE"         "ADM_RATE_ALL"
##  [4] "SAT_AVG_ALL"      "PCIP01"           "PCIP03"
```

```
##  [7] "PCIP04"            "PCIP05"            "PCIP09"
## [10] "PCIP11"            "PCIP13"            "PCIP14"
## [13] "PCIP15"            "PCIP16"            "PCIP19"
## [16] "PCIP23"            "PCIP24"            "PCIP26"
## [19] "PCIP27"            "PCIP30"            "PCIP31"
## [22] "PCIP38"            "PCIP39"            "PCIP40"
## [25] "PCIP41"            "PCIP42"            "PCIP43"
## [28] "PCIP44"            "PCIP45"            "PCIP49"
## [31] "PCIP50"            "PCIP51"            "PCIP52"
## [34] "PCIP54"            "UGDS_WHITE"        "UGDS_BLACK"
## [37] "UGDS_HISP"         "UGDS_ASIAN"        "UGDS_AIAN"
## [40] "UGDS_NHPI"         "UGDS_2MOR"         "UGDS_NRA"
## [43] "UGDS_UNKN"         "PPTUG_EF"          "COSTT4_A"
## [46] "TUITIONFEE_IN"     "TUITIONFEE_OUT"    "C150_4"
## [49] "C150_4_WHITE"      "C150_4_BLACK"      "C150_4_HISP"
## [52] "C150_4_ASIAN"      "C150_4_AIAN"       "C150_4_2MOR"
## [55] "C150_4_NRA"        "C150_4_UNKN"       "RET_FT4"
## [58] "PCTFLOAN"          "PAR_ED_PCT_1STGEN" "UGDS_MEN"
## [61] "UGDS_WOMEN"
```

```
# We will print the stats of the variables that would be accepted or not
stats <- attStats(boruta.train)
print(stats)
```

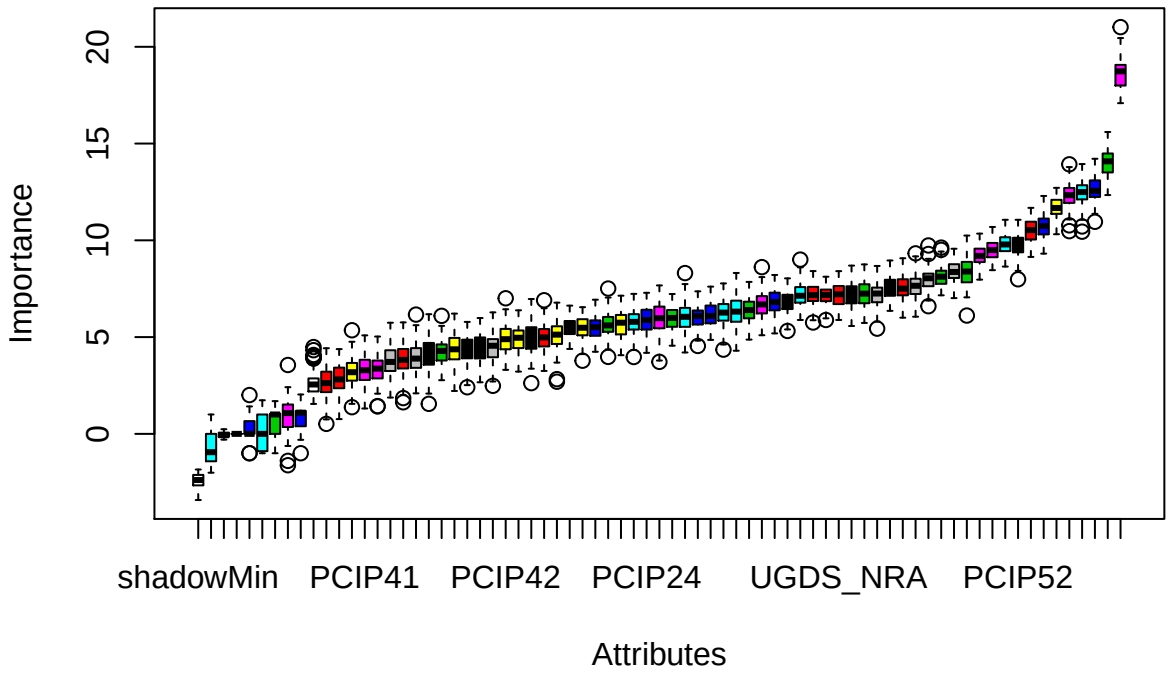
```
##                    meanImp  medianImp     minImp    maxImp   normHits
## REGION            5.5081725  5.4627199  4.3879973  6.625689 1.00000000
## ADM_RATE          7.2095002  7.1720367  5.7533737  8.411783 1.00000000
## ADM_RATE_ALL      7.2589052  7.2465149  5.7321808  8.756515 1.00000000
## SAT_AVG_ALL      12.6677731 12.5585593 10.9573350 14.213991 1.00000000
## PCIP01            6.2315129  6.2668661  4.3527331  7.774242 1.00000000
## PCIP03            6.6808669  6.6885526  5.0966550  8.615427 1.00000000
## PCIP04           11.6940475 11.6710569 10.3164257 12.712270 1.00000000
## PCIP05            8.3704024  8.3664562  7.0140846  9.564879 1.00000000
## PCIP09            4.9416377  4.9709496  2.6223312  6.967504 1.00000000
## PCIP10            2.8479120  2.8132397  0.7623389  4.379303 0.59595960
## PCIP11            6.4139389  6.3977043  4.8702094  7.852816 1.00000000
## PCIP12            0.7449681  1.0949366 -1.0010015  2.035296 0.00000000
## PCIP13            6.0183391  6.0488807  4.2069486  8.308735 1.00000000
## PCIP14           18.6592916 18.7315569 17.0877118 21.016000 1.00000000
## PCIP15            4.9167030  4.9654877  3.2171267  6.383403 0.96969697
## PCIP16            7.6616507  7.6526487  6.0536753  9.325951 1.00000000
## PCIP19            7.5038144  7.4496841  6.3554507  8.958478 1.00000000
## PCIP22            2.6291588  2.6224867  0.5169543  4.425715 0.51515152
## PCIP23            8.3742813  8.3880246  6.1130030 10.245796 1.00000000
## PCIP24            5.8692848  5.8832731  4.1868368  7.288873 1.00000000
## PCIP25           -0.7162428 -0.9466502 -2.0072601  1.001002 0.00000000
## PCIP26            5.9780349  5.9764247  3.7214850  7.669611 1.00000000
## PCIP27            5.0949145  5.1264566  2.6948283  6.779980 0.98989899
## PCIP29            0.0000000  0.0000000  0.0000000  0.000000 0.00000000
## PCIP30            4.1082751  4.0659795  1.5517149  6.181563 0.92929293
## PCIP31            4.9398011  4.9813503  3.2513531  6.904634 1.00000000
## PCIP38            4.2348597  4.2863171  2.7734359  6.098650 0.94949495
## PCIP39            5.4731635  5.5102502  4.2421279  6.932346 1.00000000
## PCIP40            5.8207331  5.7874198  3.9710213  7.233373 1.00000000
```

```
## PCIP41             3.2914394  3.2865787  1.3114749  5.089456 0.76767677
## PCIP42             4.9192337  4.8922045  3.3085043  7.006398 0.98989899
## PCIP43             7.1785507  7.2518902  5.4369174  8.676564 1.00000000
## PCIP44             4.4448209  4.5038920  2.6660761  5.985451 0.97979798
## PCIP45             7.5724495  7.5043748  6.0032489  9.073648 1.00000000
## PCIP46             0.5942746  1.0010015 -1.0010015  1.691816 0.00000000
## PCIP47             0.2618315  0.0000000 -1.0010015  2.005977 0.00000000
## PCIP48             0.2171161  0.0000000 -1.0010015  1.737100 0.00000000
## PCIP49             3.3092357  3.3715370  1.4194199  5.028358 0.79797980
## PCIP50             5.6666487  5.7443429  4.0657938  7.140368 1.00000000
## PCIP51             3.9321210  3.8892114  2.0940462  6.162679 0.90909091
## PCIP52             9.7553187  9.8197560  7.9866688 11.061929 1.00000000
## PCIP54             3.8347579  3.8243966  1.6311327  5.760333 0.89898990
## UGDS_WHITE         8.1449747  8.1210944  7.1591712  9.624241 1.00000000
## UGDS_BLACK        10.7133608 10.7414476  9.3122711 12.293652 1.00000000
## UGDS_HISP          6.3348615  6.3163742  4.2961123  8.313597 1.00000000
## UGDS_ASIAN         9.2114054  9.1951579  7.9696756 10.347958 1.00000000
## UGDS_AIAN          4.3463187  4.3690990  2.2161830  6.213952 0.97979798
## UGDS_NHPI          3.8084254  3.7173187  1.8767980  5.742722 0.90909091
## UGDS_2MOR          4.3777558  4.4300297  2.4078699  5.784833 0.95959596
## UGDS_NRA           7.1847905  7.2125180  5.8863381  8.407740 1.00000000
## UGDS_UNKN          5.9413167  5.9892452  4.5555932  7.226263 1.00000000
## PPTUG_EF           6.7961820  6.8091423  5.1921094  8.210218 1.00000000
## COSTT4_A           9.7983991  9.7869611  8.6439791 11.060685 1.00000000
## TUITIONFEE_IN      9.5129484  9.4955091  8.4578019 10.690246 1.00000000
## TUITIONFEE_OUT     5.4907310  5.4817032  3.7807516  6.545610 1.00000000
## C150_4             8.0034858  8.0396025  6.5959422  9.734829 1.00000000
## C150_4_WHITE       6.7808288  6.8387931  5.3180947  8.036055 1.00000000
## C150_4_BLACK       7.1342903  7.1766900  5.8845291  8.119342 1.00000000
## C150_4_HISP        5.6420301  5.6040174  3.9775933  7.508769 1.00000000
## C150_4_ASIAN       6.1559994  6.0925102  4.8555044  7.608205 1.00000000
## C150_4_AIAN        7.1764284  7.1351994  5.8846373  9.003808 1.00000000
## C150_4_NHPI        0.8901589  1.0732144 -1.6206667  3.558408 0.02020202
## C150_4_2MOR        3.1636871  3.1851329  1.3785923  5.353744 0.74747475
## C150_4_NRA         4.4333786  4.5516215  2.4822177  6.276687 0.96969697
## C150_4_UNKN        7.2340428  7.2176670  5.5727722  8.694206 1.00000000
## RET_FT4           10.4839983 10.5279623  9.1432848 11.678124 1.00000000
## PCTFLOAN          14.0460059 14.0802416 12.3387896 15.598037 1.00000000
## PAR_ED_PCT_1STGEN  6.0380151  6.0880363  4.5425596  7.364983 1.00000000
## UGDS_MEN          12.4614180 12.4937905 10.4473895 13.940432 1.00000000
## UGDS_WOMEN        12.3260541 12.3318389 10.4745491 13.927399 1.00000000
##                    decision
## REGION            Confirmed
## ADM_RATE          Confirmed
## ADM_RATE_ALL      Confirmed
## SAT_AVG_ALL       Confirmed
## PCIP01            Confirmed
## PCIP03            Confirmed
## PCIP04            Confirmed
## PCIP05            Confirmed
## PCIP09            Confirmed
## PCIP10            Tentative
## PCIP11            Confirmed
## PCIP12             Rejected
```

```
## PCIP13           Confirmed
## PCIP14           Confirmed
## PCIP15           Confirmed
## PCIP16           Confirmed
## PCIP19           Confirmed
## PCIP22           Tentative
## PCIP23           Confirmed
## PCIP24           Confirmed
## PCIP25            Rejected
## PCIP26           Confirmed
## PCIP27           Confirmed
## PCIP29            Rejected
## PCIP30           Confirmed
## PCIP31           Confirmed
## PCIP38           Confirmed
## PCIP39           Confirmed
## PCIP40           Confirmed
## PCIP41           Confirmed
## PCIP42           Confirmed
## PCIP43           Confirmed
## PCIP44           Confirmed
## PCIP45           Confirmed
## PCIP46            Rejected
## PCIP47            Rejected
## PCIP48            Rejected
## PCIP49           Confirmed
## PCIP50           Confirmed
## PCIP51           Confirmed
## PCIP52           Confirmed
## PCIP54           Confirmed
## UGDS_WHITE       Confirmed
## UGDS_BLACK       Confirmed
## UGDS_HISP        Confirmed
## UGDS_ASIAN       Confirmed
## UGDS_AIAN        Confirmed
## UGDS_NHPI        Confirmed
## UGDS_2MOR        Confirmed
## UGDS_NRA         Confirmed
## UGDS_UNKN        Confirmed
## PPTUG_EF         Confirmed
## COSTT4_A         Confirmed
## TUITIONFEE_IN    Confirmed
## TUITIONFEE_OUT   Confirmed
## C150_4           Confirmed
## C150_4_WHITE     Confirmed
## C150_4_BLACK     Confirmed
## C150_4_HISP      Confirmed
## C150_4_ASIAN     Confirmed
## C150_4_AIAN      Confirmed
## C150_4_NHPI       Rejected
## C150_4_2MOR      Confirmed
## C150_4_NRA       Confirmed
## C150_4_UNKN      Confirmed
## RET_FT4          Confirmed
```

```
## PCTFLOAN          Confirmed
## PAR_ED_PCT_1STGEN Confirmed
## UGDS_MEN          Confirmed
## UGDS_WOMEN        Confirmed
```

```
# We will plot on the number of variables and its importance for Boruta
plot(boruta.train, type = c("g","o"), cex = 1.0, col = 1:70)
```




```
#Now, let us try RFE
rfe_control <- rfeControl(functions=rfFuncs, method="cv", number = 10)
rfe.train <- rfe(usunivnoccbasic[,1:70], usunivnoccbasic[,72], sizes = 1:70, rfeControl = rfe_control)
```

```
##
## Attaching package: 'plyr'
```

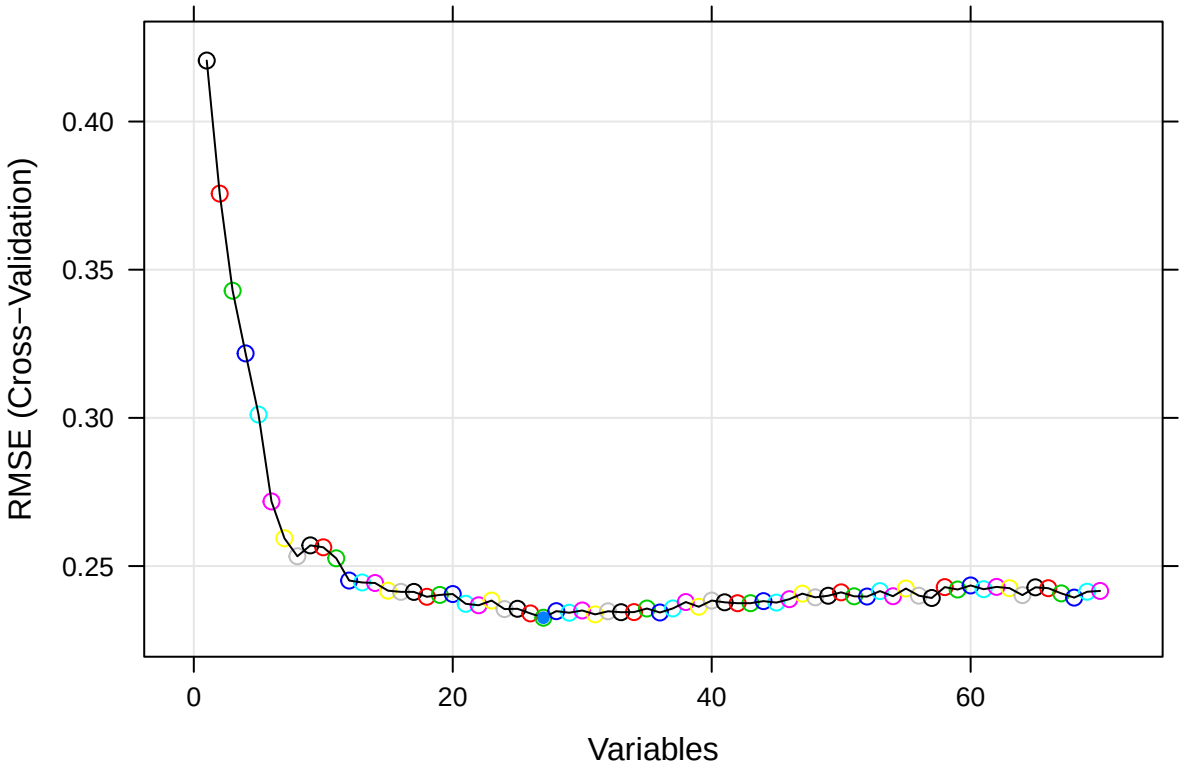
```
## The following object is masked from 'package:modeltools':
##
##     empty
```

```
predictors(rfe.train)
```

```
##  [1] "PCIP14"        "PCTFLOAN"      "PCIP04"        "SAT_AVG_ALL"
##  [5] "PCIP52"        "UGDS_BLACK"    "UGDS_MEN"      "PCIP45"
##  [9] "UGDS_WOMEN"    "PCIP43"        "COSTT4_A"      "PCIP23"
```

```
## [13] "RET_FT4"       "UGDS_HISP"     "TUITIONFEE_IN" "C150_4_AIAN"
## [17] "PCIP39"        "UGDS_ASIAN"    "PCIP16"        "UGDS_WHITE"
## [21] "PCIP19"        "C150_4"        "PCIP05"        "UGDS_NRA"
## [25] "PCIP26"        "PCIP03"        "PCIP24"
```

```
# We will plot on the number of variables and its importance for RFE
plot(rfe.train, type = c("g","o"), cex = 1.0, col = 1:70)
```



Based on these runs, RFE determines fewer variables needed for the prediction model than Boruta. There would be some cases that the Boruta package could be used, depending on the number of variables.

US Research University Completion Rate Prediction Model

```
rm_train2 <- sample(nrow(usresearchuniv), floor(nrow(usresearchuniv)*0.75))
univ_train2 <- usresearchuniv[rm_train2,]
univ_test2 <- usresearchuniv[-rm_train2,]

formula_completionrate <- formula(C150_4_NRA ~ REGION + ADM_RATE_ALL + UGDS_NRA + PPTUG_EF + COSTT4_A +
```

We will do a generalized multivariate linear regression formula.

```r
# create a logistic regression
fit2 <- lm(formula_completionrate, data = usresearchuniv)
summary(fit2)
```

```
##
## Call:
## lm(formula = formula_completionrate, data = usresearchuniv)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62640 -0.05949  0.00907  0.07396  0.51024
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.323e-01  3.881e-02  24.021  < 2e-16 ***
## REGION            -2.791e-03  2.847e-03  -0.980  0.32728
## ADM_RATE_ALL      -1.472e-01  3.336e-02  -4.412 1.16e-05 ***
## UGDS_NRA           2.210e-01  1.274e-01   1.735  0.08314 .
## PPTUG_EF          -3.508e-01  7.451e-02  -4.708 2.94e-06 ***
## COSTT4_A           1.588e-06  5.358e-07   2.965  0.00312 **
## PCTFLOAN          -3.614e-01  5.114e-02  -7.068 3.41e-12 ***
## PAR_ED_PCT_1STGEN -9.581e-02  8.656e-02  -1.107  0.26865
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1408 on 807 degrees of freedom
## Multiple R-squared:  0.4242, Adjusted R-squared:  0.4192
## F-statistic: 84.94 on 7 and 807 DF,  p-value: < 2.2e-16
```

Based on the regression, the formula will be

$C150\_4\_NRA = 0.932 - 0.00279REGION - 0.147ADM\_RATE\_ALL + 0.021UGDS\_NRA - 0.351PPTUG\_EF + 0.00000$

.

We will test this regression with some data types.

```r
# for Ivy League schools with high admission rates for all and international students
df_accept3 <- data.frame(REGION = 1, ADM_RATE_ALL = .55, UGDS_NRA=.25, PPTUG_EF = 0.07, COSTT4_A = 5000
predict(fit2, newdata = df_accept3)
```

```
##         1
## 0.7757938
```

```r
# for Ivy League schools with less admission rates, but have high shares of students doing part-time
df_accept4 <- data.frame(REGION = 1, ADM_RATE_ALL = .05, UGDS_NRA=.05, PPTUG_EF = 0.46, COSTT4_A = 5000
predict(fit2, newdata = df_accept4)
```

```
##        1
## 0.612912
```

Now, we will do some testing of performance with the logistic regression. Since we have split the dataset into training and testing set, we will see how the performance will be done.

```r
# using multivariate linear regression to calculate the completion rate for international students
lm_NRAcompletion <- lm(formula_completionrate, data = univ_train2)
summary(lm_NRAcompletion)
```

```
##
## Call:
## lm(formula = formula_completionrate, data = univ_train2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58903 -0.05884  0.01043  0.07397  0.47438
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.635e-01  4.283e-02  22.497  < 2e-16 ***
## REGION            -4.079e-03  3.085e-03  -1.322   0.1867
## ADM_RATE_ALL      -1.526e-01  3.659e-02  -4.169 3.51e-05 ***
## UGDS_NRA           1.522e-01  1.364e-01   1.116   0.2648
## PPTUG_EF          -3.609e-01  8.151e-02  -4.427 1.13e-05 ***
## COSTT4_A           1.359e-06  5.864e-07   2.317   0.0208 *
## PCTFLOAN          -3.900e-01  5.585e-02  -6.983 7.66e-12 ***
## PAR_ED_PCT_1STGEN -6.653e-02  9.321e-02  -0.714   0.4757
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1338 on 603 degrees of freedom
## Multiple R-squared:  0.4353, Adjusted R-squared:  0.4287
## F-statistic: 66.39 on 7 and 603 DF,  p-value: < 2.2e-16
```

```r
# do the testing with the prediction model
accepted_ind3 <- predict(lm_NRAcompletion, interval="prediction", newdata = univ_test2)

# Checking on PRED(25)
errors <- accepted_ind3[,"fit"] - univ_test2$C150_4_NRA
rel_change <- abs(errors) / univ_test2$C150_4_NRA
table(rel_change<0.25)["TRUE"] / nrow(univ_test2)
```

```
##      TRUE
## 0.7647059
```

```r
# Now we check on what acceptable ways we could do for regression
# Doing single decision tree
model_dtree3 <- rpart(formula_completionrate, method="anova",data = univ_train2)
summary(model_dtree3)
```

```
## Call:
## rpart(formula = formula_completionrate, data = univ_train2, method = "anova")
##   n= 611
##
##           CP nsplit rel error    xerror       xstd
## 1 0.27649341      0 1.0000000 1.0028935 0.06369181
## 2 0.07062335      1 0.7235066 0.7469230 0.05634590
```

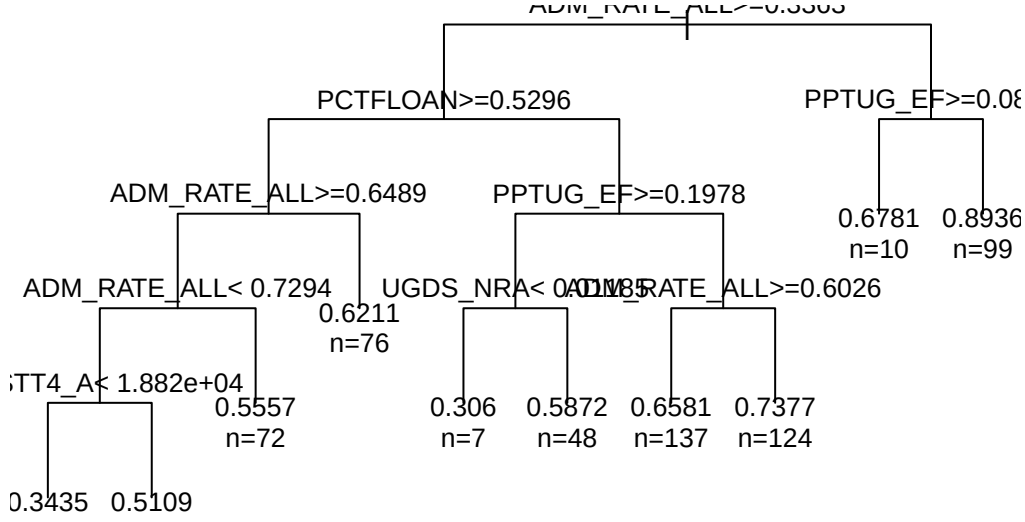
```
## 3 0.04960622      2 0.6528832 0.7097115 0.05755015
## 4 0.02528588      3 0.6032770 0.6929204 0.05539394
## 5 0.02243677      4 0.5779911 0.6753507 0.05329303
## 6 0.02208896      5 0.5555544 0.6649698 0.05236172
## 7 0.02157032      6 0.5334654 0.6649698 0.05236172
## 8 0.01138938      7 0.5118951 0.6192243 0.04968555
## 9 0.01000000      9 0.4891163 0.6125933 0.04596624
##
## Variable importance
##      ADM_RATE_ALL          PPTUG_EF          PCTFLOAN          COSTT4_A
##                33                20                17                15
## PAR_ED_PCT_1STGEN          UGDS_NRA            REGION
##                10                 4                 1
##
## Node number 1: 611 observations,    complexity param=0.2764934
##   mean=0.6742969, MSE=0.03126865
##   left son=2 (502 obs) right son=3 (109 obs)
##   Primary splits:
##       ADM_RATE_ALL < 0.3363178 to the right, improve=0.2764934, (0 missing)
##       COSTT4_A     < 51980     to the left,  improve=0.2392747, (0 missing)
##       PPTUG_EF     < 0.05795   to the right, improve=0.2062040, (0 missing)
##       UGDS_NRA     < 0.05435   to the left,  improve=0.1915478, (0 missing)
##       PCTFLOAN     < 0.3564    to the right, improve=0.1842236, (0 missing)
##   Surrogate splits:
##       COSTT4_A          < 52427     to the left,  agree=0.897, adj=0.422, (0 split)
##       PPTUG_EF          < 0.01375   to the right, agree=0.894, adj=0.404, (0 split)
##       PCTFLOAN          < 0.3108    to the right, agree=0.884, adj=0.349, (0 split)
##       PAR_ED_PCT_1STGEN < 0.1795873 to the right, agree=0.872, adj=0.284, (0 split)
##
## Node number 2: 502 observations,    complexity param=0.07062335
##   mean=0.6309699, MSE=0.02569606
##   left son=4 (186 obs) right son=5 (316 obs)
##   Primary splits:
##       PCTFLOAN     < 0.52965   to the right, improve=0.10459920, (0 missing)
##       PPTUG_EF     < 0.08135   to the right, improve=0.09209080, (0 missing)
##       UGDS_NRA     < 0.057     to the left,  improve=0.07690063, (0 missing)
##       ADM_RATE_ALL < 0.6555841 to the right, improve=0.06698155, (0 missing)
##       COSTT4_A     < 20653     to the left,  improve=0.06628840, (0 missing)
##   Surrogate splits:
##       UGDS_NRA          < 0.00725   to the left,  agree=0.643, adj=0.038, (0 split)
##       REGION            < 3.5       to the left,  agree=0.639, adj=0.027, (0 split)
##       ADM_RATE_ALL      < 0.8229921 to the right, agree=0.639, adj=0.027, (0 split)
##       PAR_ED_PCT_1STGEN < 0.3278438 to the right, agree=0.639, adj=0.027, (0 split)
##       PPTUG_EF          < 0.001     to the left,  agree=0.635, adj=0.016, (0 split)
##
## Node number 3: 109 observations,    complexity param=0.02208896
##   mean=0.8738394, MSE=0.00847039
##   left son=6 (10 obs) right son=7 (99 obs)
##   Primary splits:
##       PPTUG_EF          < 0.0897    to the right, improve=0.4570835, (0 missing)
##       COSTT4_A          < 23707.5   to the left,  improve=0.3702758, (0 missing)
##       ADM_RATE_ALL      < 0.2576    to the right, improve=0.3391687, (0 missing)
##       UGDS_NRA          < 0.0438    to the left,  improve=0.3167517, (0 missing)
##       PAR_ED_PCT_1STGEN < 0.3373642 to the right, improve=0.3006233, (0 missing)
```

```
##   Surrogate splits:
##       COSTT4_A          < 20751.5   to the left,  agree=0.936, adj=0.3, (0 split)
##       PAR_ED_PCT_1STGEN < 0.3843536 to the right, agree=0.927, adj=0.2, (0 split)
##       UGDS_NRA          < 0.03465   to the left,  agree=0.917, adj=0.1, (0 split)
##
## Node number 4: 186 observations,    complexity param=0.02243677
##   mean=0.5633952, MSE=0.02493515
##   left son=8 (110 obs) right son=9 (76 obs)
##   Primary splits:
##       ADM_RATE_ALL < 0.64885   to the right, improve=0.09242418, (0 missing)
##       COSTT4_A     < 20667     to the left,  improve=0.07785026, (0 missing)
##       PPTUG_EF     < 0.08085   to the right, improve=0.05581419, (0 missing)
##       REGION       < 3.5       to the right, improve=0.05420677, (0 missing)
##       UGDS_NRA     < 0.0158    to the left,  improve=0.05021006, (0 missing)
##   Surrogate splits:
##       COSTT4_A          < 26724     to the left,  agree=0.694, adj=0.250, (0 split)
##       PPTUG_EF          < 0.0654    to the right, agree=0.677, adj=0.211, (0 split)
##       PAR_ED_PCT_1STGEN < 0.2157511 to the right, agree=0.629, adj=0.092, (0 split)
##       UGDS_NRA          < 0.0768    to the left,  agree=0.624, adj=0.079, (0 split)
##       REGION            < 2.5       to the right, agree=0.618, adj=0.066, (0 split)
##
## Node number 5: 316 observations,    complexity param=0.04960622
##   mean=0.6707449, MSE=0.0218741
##   left son=10 (55 obs) right son=11 (261 obs)
##   Primary splits:
##       PPTUG_EF          < 0.1978    to the right, improve=0.13711000, (0 missing)
##       COSTT4_A          < 24729.5   to the left,  improve=0.07150608, (0 missing)
##       PAR_ED_PCT_1STGEN < 0.3463965 to the right, improve=0.07079260, (0 missing)
##       UGDS_NRA          < 0.05725   to the left,  improve=0.05865867, (0 missing)
##       REGION            < 3.5       to the right, improve=0.04541987, (0 missing)
##   Surrogate splits:
##       PAR_ED_PCT_1STGEN < 0.3848453 to the right, agree=0.854, adj=0.164, (0 split)
##       COSTT4_A          < 16035     to the left,  agree=0.848, adj=0.127, (0 split)
##       ADM_RATE_ALL      < 0.93205   to the right, agree=0.839, adj=0.073, (0 split)
##
## Node number 6: 10 observations
##   mean=0.67806, MSE=0.01823085
##
## Node number 7: 99 observations
##   mean=0.8936152, MSE=0.003221731
##
## Node number 8: 110 observations,    complexity param=0.01138938
##   mean=0.5234918, MSE=0.02330008
##   left son=16 (38 obs) right son=17 (72 obs)
##   Primary splits:
##       ADM_RATE_ALL < 0.7294199 to the left,  improve=0.08436483, (0 missing)
##       PCTFLOAN     < 0.70995   to the right, improve=0.07851020, (0 missing)
##       COSTT4_A     < 20052     to the left,  improve=0.07811049, (0 missing)
##       REGION       < 3.5       to the right, improve=0.03879852, (0 missing)
##       UGDS_NRA     < 0.012     to the left,  improve=0.03793614, (0 missing)
##   Surrogate splits:
##       PCTFLOAN          < 0.6804    to the right, agree=0.709, adj=0.158, (0 split)
##       UGDS_NRA          < 0.01115   to the left,  agree=0.691, adj=0.105, (0 split)
##       PAR_ED_PCT_1STGEN < 0.4303247 to the right, agree=0.691, adj=0.105, (0 split)
```

```
##       PPTUG_EF          < 0.23515   to the right, agree=0.682, adj=0.079, (0 split)
##       REGION            < 6.5       to the right, agree=0.673, adj=0.053, (0 split)
##
## Node number 9: 76 observations
##   mean=0.62115, MSE=0.02166148
##
## Node number 10: 55 observations,    complexity param=0.02528588
##   mean=0.5514455, MSE=0.0326402
##   left son=20 (7 obs) right son=21 (48 obs)
##   Primary splits:
##       UGDS_NRA          < 0.01185   to the left,  improve=0.26909950, (0 missing)
##       PCTFLOAN          < 0.3722    to the right, improve=0.17247590, (0 missing)
##       PAR_ED_PCT_1STGEN < 0.3256895 to the right, improve=0.13074400, (0 missing)
##       PPTUG_EF          < 0.2773    to the right, improve=0.09900907, (0 missing)
##       REGION            < 6.5       to the left,  improve=0.08207734, (0 missing)
##
## Node number 11: 261 observations,    complexity param=0.02157032
##   mean=0.6958847, MSE=0.01597422
##   left son=22 (137 obs) right son=23 (124 obs)
##   Primary splits:
##       ADM_RATE_ALL      < 0.6026    to the right, improve=0.09884318, (0 missing)
##       COSTT4_A          < 23813.5   to the left,  improve=0.05481396, (0 missing)
##       PAR_ED_PCT_1STGEN < 0.1735899 to the right, improve=0.05383448, (0 missing)
##       UGDS_NRA          < 0.04425   to the left,  improve=0.05274882, (0 missing)
##       PPTUG_EF          < 0.065     to the right, improve=0.04826549, (0 missing)
##   Surrogate splits:
##       PPTUG_EF          < 0.065     to the right, agree=0.686, adj=0.339, (0 split)
##       COSTT4_A          < 28918.5   to the left,  agree=0.659, adj=0.282, (0 split)
##       REGION            < 2.5       to the right, agree=0.628, adj=0.218, (0 split)
##       UGDS_NRA          < 0.06035   to the left,  agree=0.621, adj=0.202, (0 split)
##       PAR_ED_PCT_1STGEN < 0.2086112 to the right, agree=0.605, adj=0.169, (0 split)
##
## Node number 16: 38 observations,    complexity param=0.01138938
##   mean=0.4624632, MSE=0.03119489
##   left son=32 (11 obs) right son=33 (27 obs)
##   Primary splits:
##       COSTT4_A          < 18823     to the left,  improve=0.18471610, (0 missing)
##       REGION            < 2.5       to the right, improve=0.07404089, (0 missing)
##       PAR_ED_PCT_1STGEN < 0.3386463 to the left,  improve=0.06080481, (0 missing)
##       PCTFLOAN          < 0.5814    to the right, improve=0.05599123, (0 missing)
##       PPTUG_EF          < 0.073     to the right, improve=0.04803588, (0 missing)
##   Surrogate splits:
##       PAR_ED_PCT_1STGEN < 0.3072623 to the left,  agree=0.816, adj=0.364, (0 split)
##       PCTFLOAN          < 0.76655   to the right, agree=0.763, adj=0.182, (0 split)
##
## Node number 17: 72 observations
##   mean=0.5557014, MSE=0.0161302
##
## Node number 20: 7 observations
##   mean=0.3060286, MSE=0.04588194
##
## Node number 21: 48 observations
##   mean=0.5872354, MSE=0.02064473
##
```

```
## Node number 22: 137 observations
##   mean=0.658081, MSE=0.01378832
##
## Node number 23: 124 observations
##   mean=0.7376516, MSE=0.01506587
##
## Node number 32: 11 observations
##   mean=0.3435364, MSE=0.02048353
##
## Node number 33: 27 observations
##   mean=0.5109148, MSE=0.02744902
```

```
plot(model_dtree3, uniform = TRUE, main = "Single Decision Tree  of\nUS Research University Completion R
text(model_dtree3, use.n = TRUE, cex = .8)
```



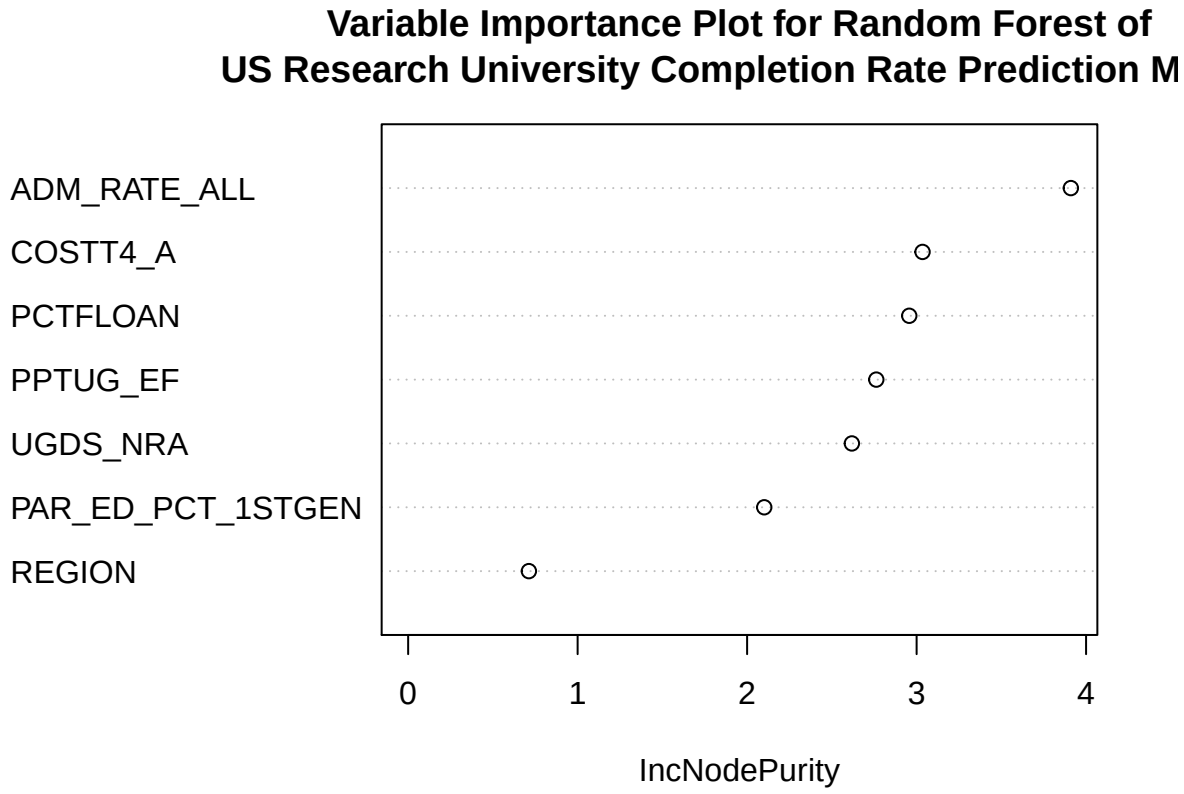

```
pred_dtree3 <- predict(model_dtree3, newdata = univ_test2)
accu11 <- abs(pred_dtree3 - univ_test2$C150_4_NRA) < 0.25
frac11 <- sum(accu11)/length(accu11)
print(frac11)
```

```
## [1] 0.872549
```

```
# Doing random forest
model_forest3 <- randomForest(formula_completionrate, data = univ_train2)
summary(model_forest3)
```

```
##                 Length Class  Mode
## call               3   -none- call
## type               1   -none- character
## predicted        611   -none- numeric
## mse              500   -none- numeric
## rsq              500   -none- numeric
## oob.times        611   -none- numeric
## importance         7   -none- numeric
## importanceSD       0   -none- NULL
## localImportance    0   -none- NULL
## proximity          0   -none- NULL
## ntree              1   -none- numeric
## mtry               1   -none- numeric
## forest            11   -none- list
## coefs              0   -none- NULL
## y                611   -none- numeric
## test               0   -none- NULL
## inbag              0   -none- NULL
## terms              3   terms  call
```

```r
varImpPlot(model_forest3, main = "Variable Importance Plot for Random Forest of\nUS Research University
```




```r
pred_forest3 <- predict(model_forest3, newdata = univ_test2)
accu12 <- abs(pred_forest3 - univ_test2$C150_4_NRA) < 0.25
frac12 <- sum(accu12)/length(accu12)
print(frac12)
```

```
## [1] 0.9117647
```

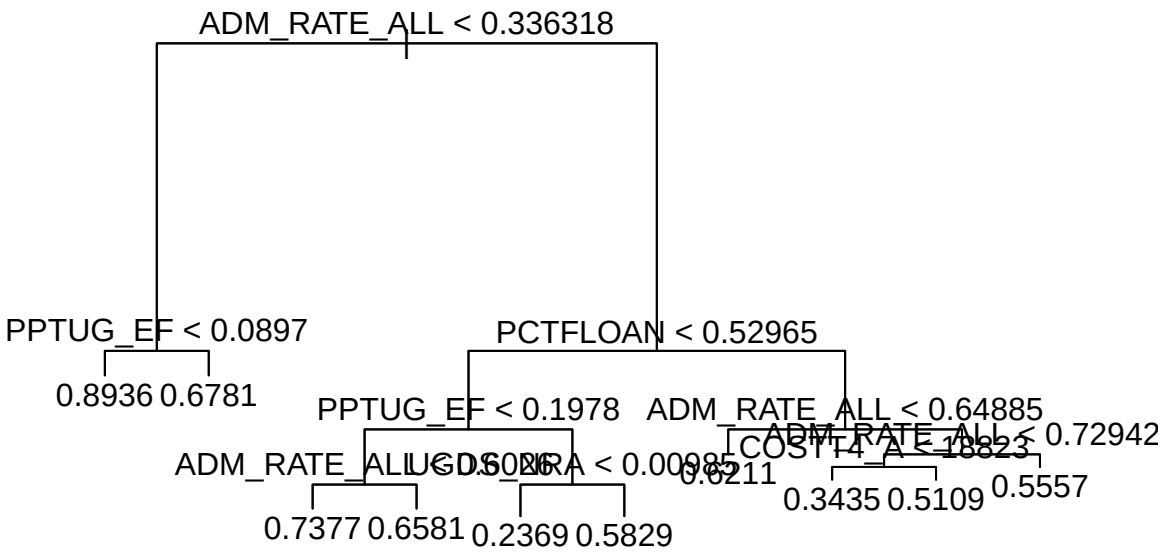
```
# Doing support vector machine
model_svm3 <- svm(formula_completionrate, data = univ_train2)
summary(model_svm3)
```

```
##
## Call:
## svm(formula = formula_completionrate, data = univ_train2)
##
##
## Parameters:
##    SVM-Type:  eps-regression
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  0.1428571
##     epsilon:  0.1
##
##
## Number of Support Vectors:  511
```

```
pred_svm3 <- predict(model_svm3, newdata = univ_test2)
accu13 <- abs(pred_svm3 - univ_test2$C150_4_NRA) < 0.25
frac13 <- sum(accu13)/length(accu13)
print(frac13)
```

```
## [1] 0.9019608
```

```
# doing simple tree
model_tree3 <- tree(formula_completionrate, data = univ_train2)
plot(model_tree3, main = "Simple Tree of US Research\nUniversity Completion Rate Prediction Model")
text(model_tree3)
```



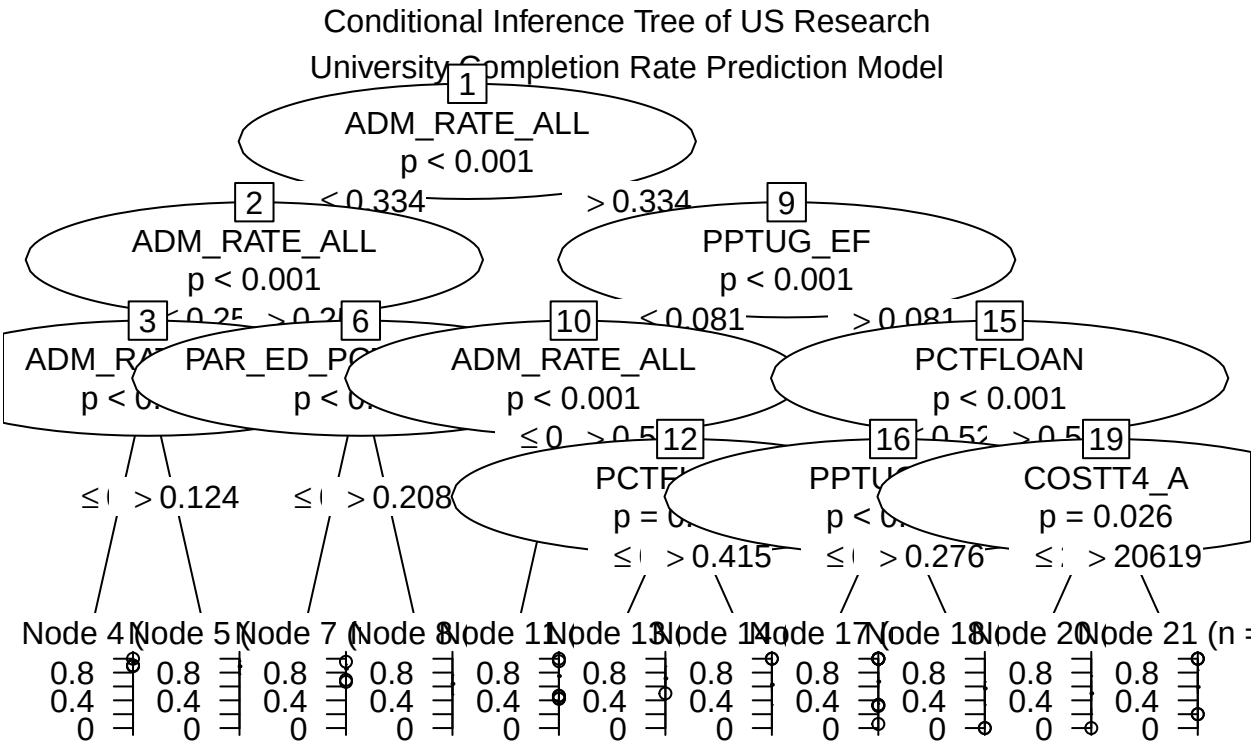

```
pred_tree3 <- predict(model_tree3, newdata = univ_test2)
accu14 <- abs(pred_tree3 - univ_test2$C150_4_NRA) < 0.25
frac14 <- sum(accu14)/length(accu14)
print(frac14)
```

```
## [1] 0.8872549
```

```
# doing conditional inference tree
model_party3 <- ctree(formula_completionrate, data = univ_train2)
summary(model_party3)
```

```
##      Length      Class       Mode
##           1 BinaryTree         S4
```

```
plot(model_party3, main = "Conditional Inference Tree of US Research\nUniversity Completion Rate Predic
```




```
pred_party3 <- predict(model_party3, newdata = univ_test2)
accu15 <- abs(pred_party3 - univ_test2$C150_4_NRA) < 0.25
frac15 <- sum(accu15)/length(accu15)
print(frac15)
```

```
## [1] 0.872549
```

From the regressions that we have run, the random forest is the best regression model to use for determining completion rates for international students.