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
```

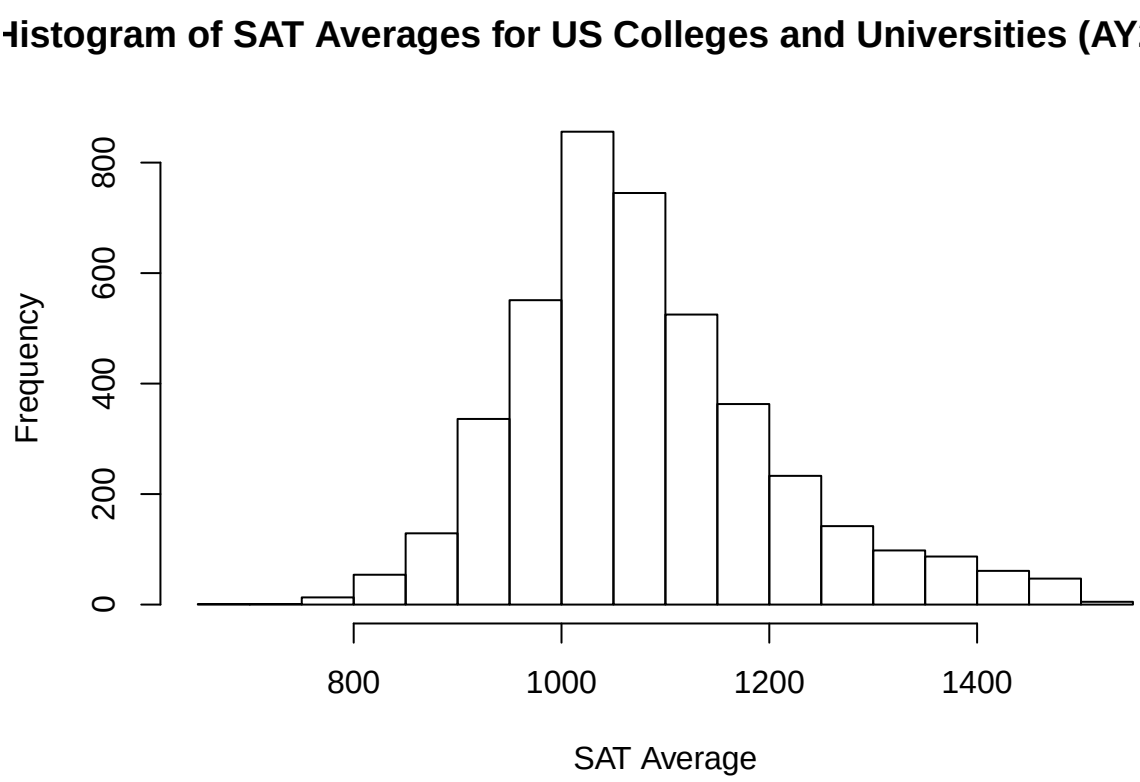
```
## [1] 815
```

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

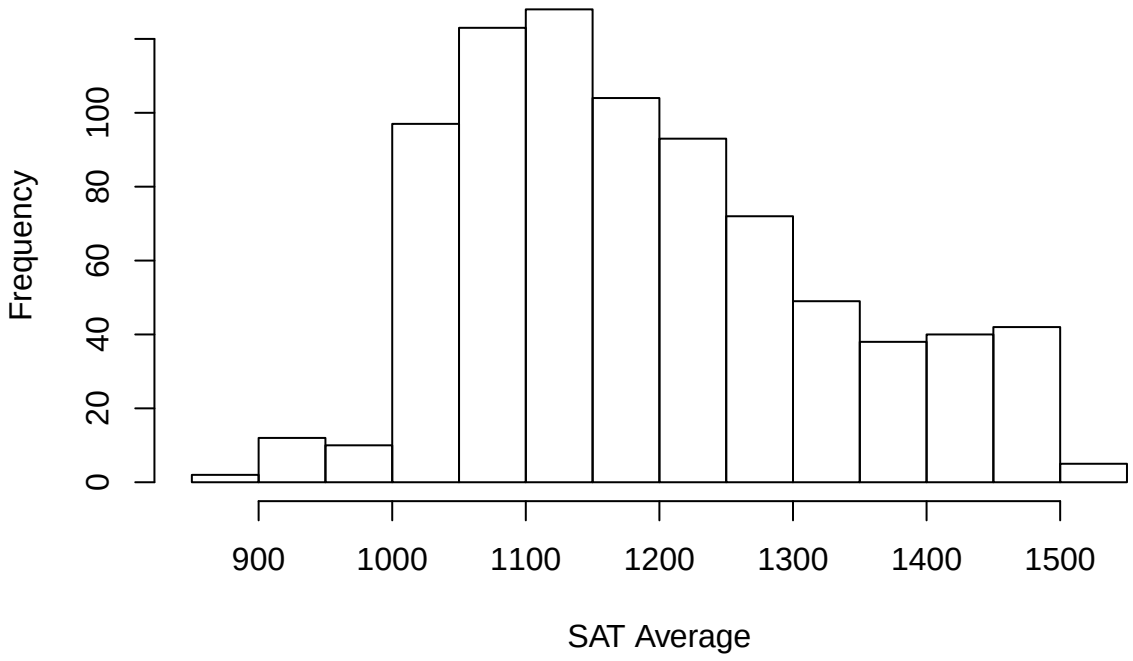
```
## [1] 0.7062392
```

Distributions and Box and Whisker Plots

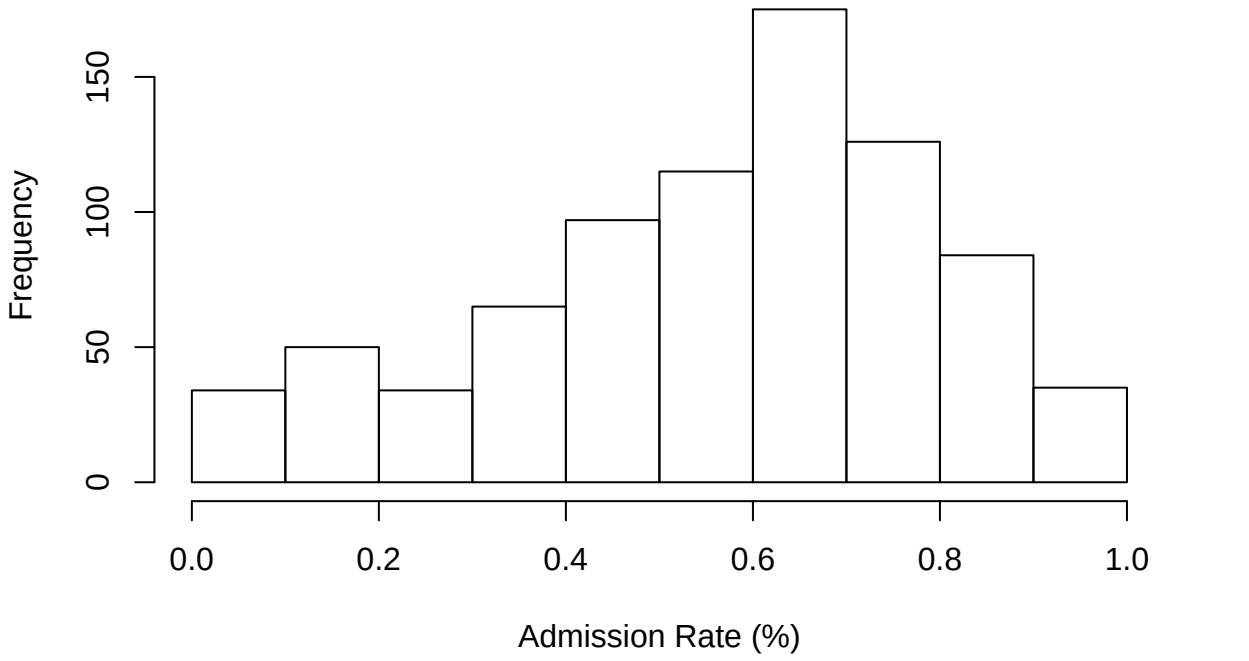
```r
# Histogram of SAT Averages for US Colleges and Universities
hist(usunivfilter$SAT_AVG_ALL, main = "Histogram of SAT Averages for US Colleges and Universities (AY20
```



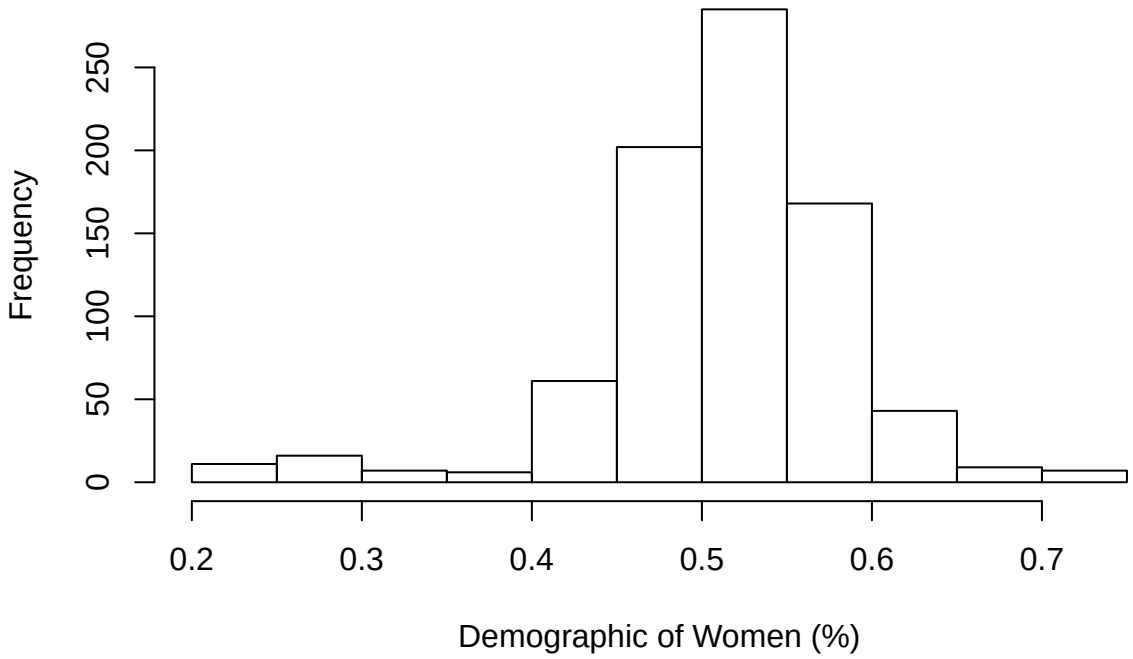

```r
# Histogram of SAT Averages for US Research Universities
hist(usresearchuniv$SAT_AVG_ALL, main = "Histogram of SAT Averages for US Research Universities (AY2010
```



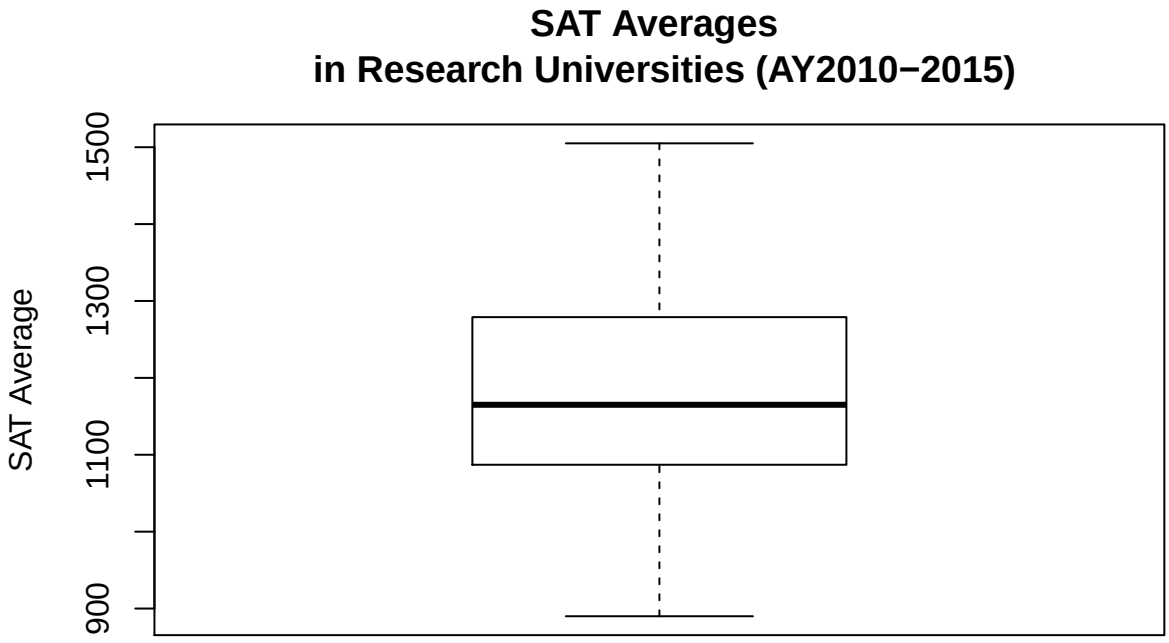

```r
# Histogram of Admission Rates for US Research Universities
hist(usresearchuniv$ADM_RATE_ALL, main = "Histogram of Admission Rates for Research Universities (AY201
```



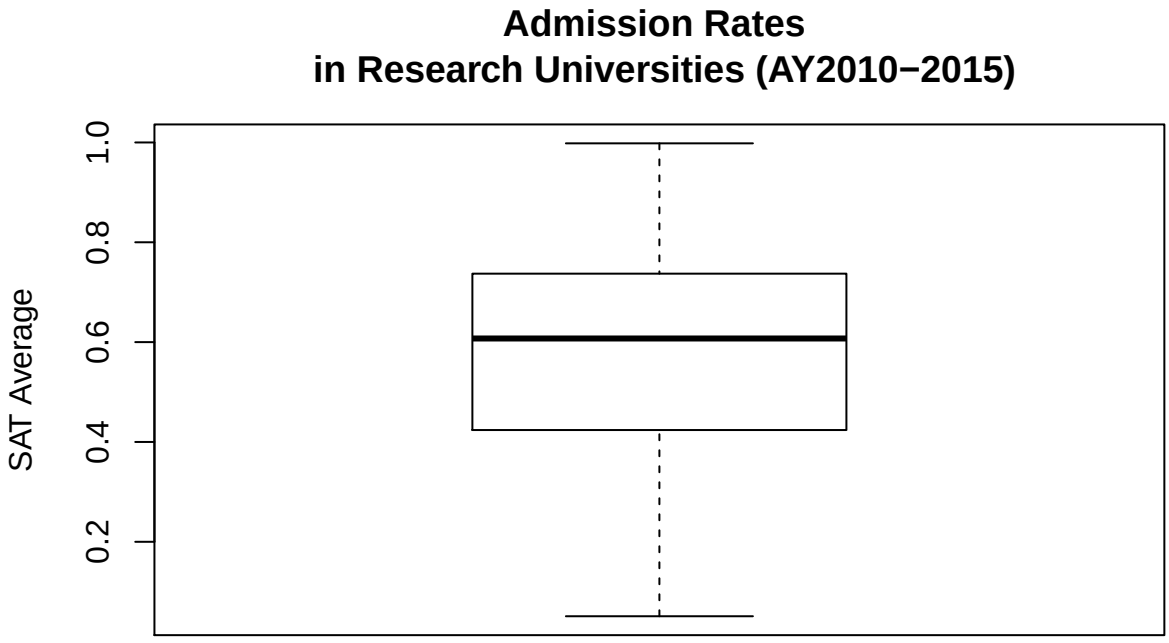

```r
# Histogram of Women in US Research Universities
hist(usresearchuniv$UGDS_WOMEN, main = "Histogram of Women in Research Universities (AY2010-2015)", xlab
```



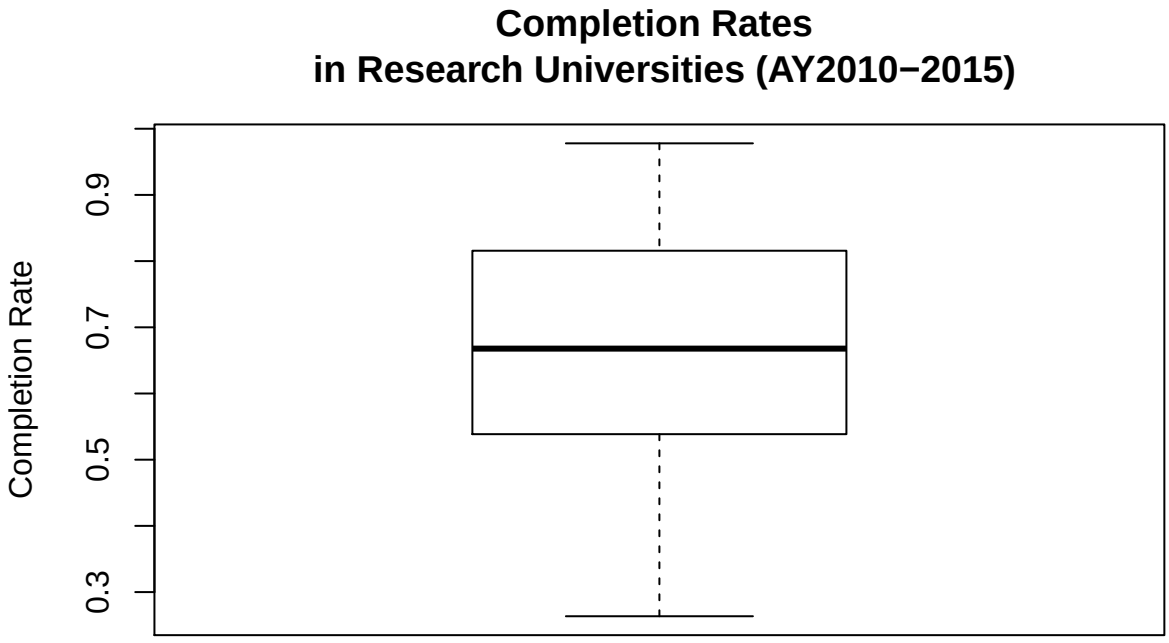

```
#Boxplot of SAT Average in all US Research Universities
boxplot(usresearchuniv$SAT_AVG_ALL, main = "SAT Averages \n in Research Universities (AY2010-2015)", yla
```



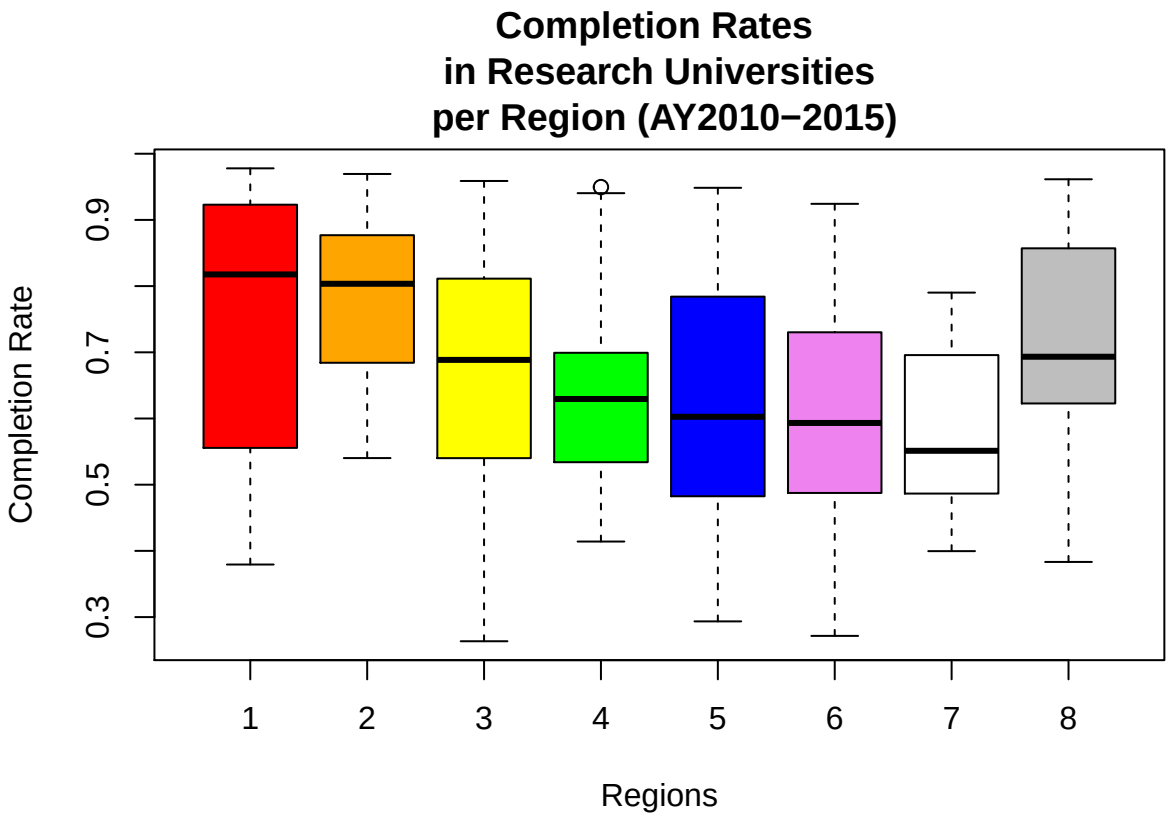

```
#Boxplot of admission rates in all US Research Universities
boxplot(usresearchuniv$ADM_RATE_ALL, main = "Admission Rates \n in Research Universities (AY2010-2015)"
```



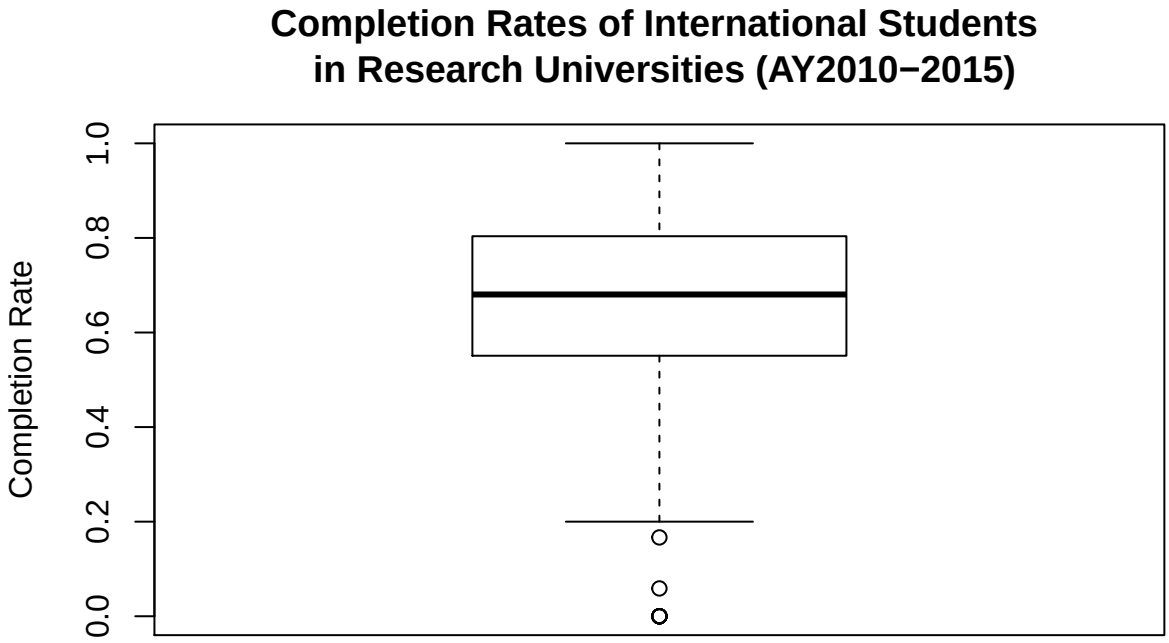

```
#Boxplot of Completion Rates in all US Research Universities
boxplot(usresearchuniv$C150_4, main = "Completion Rates \n in Research Universities (AY2010-2015)", ylab
```



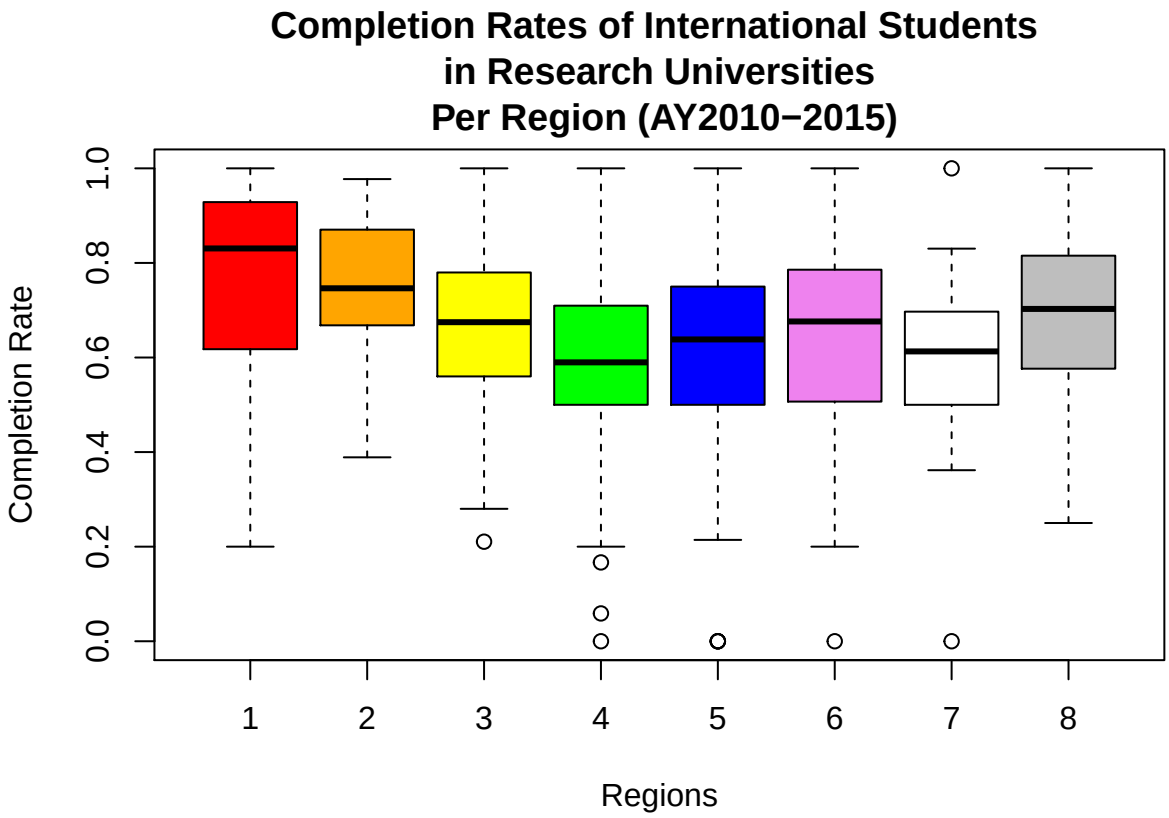

```r
# Boxplot of Completion Rates per Region in US Research Universities
boxplot(C150_4 ~ REGION, usresearchuniv, main = "Completion Rates \n in Research Universities \n per Reg
```




```
#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
```

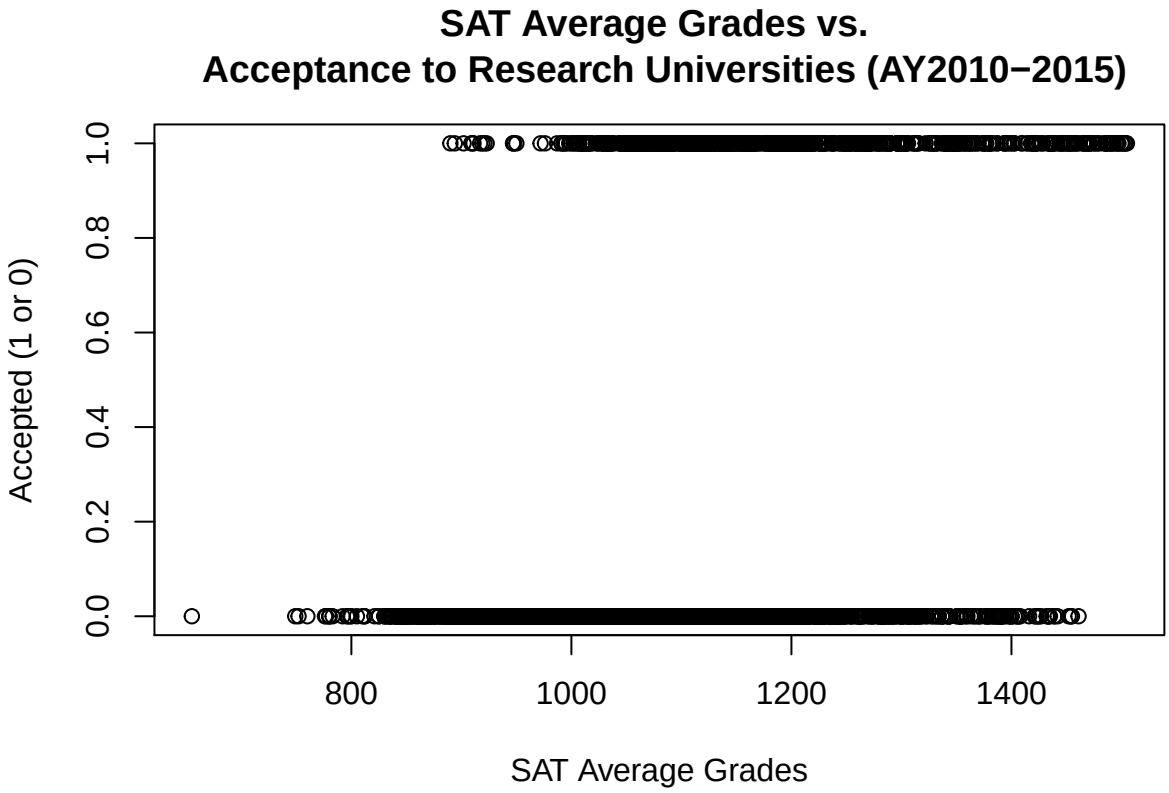



```
nrow(usresearchuniv[usresearchuniv$C150_4_NRA < 0.2,])
```

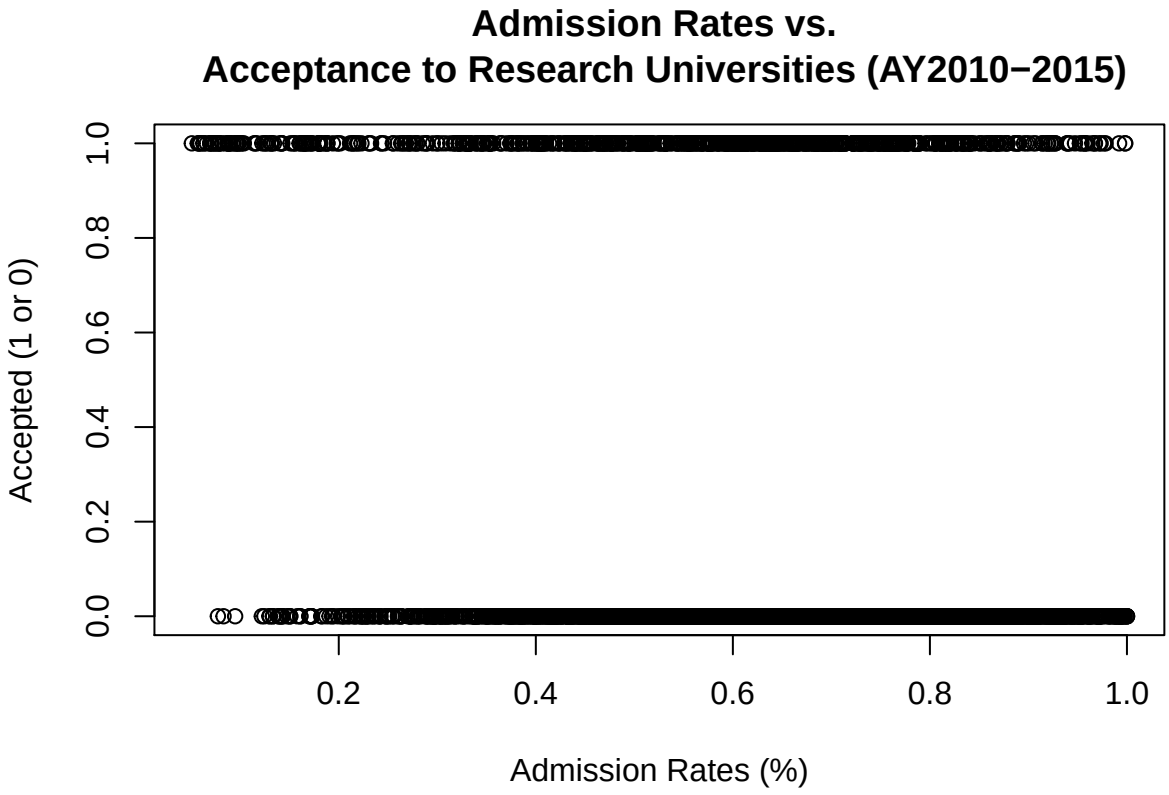
```
## [1] 9
```

Correlations

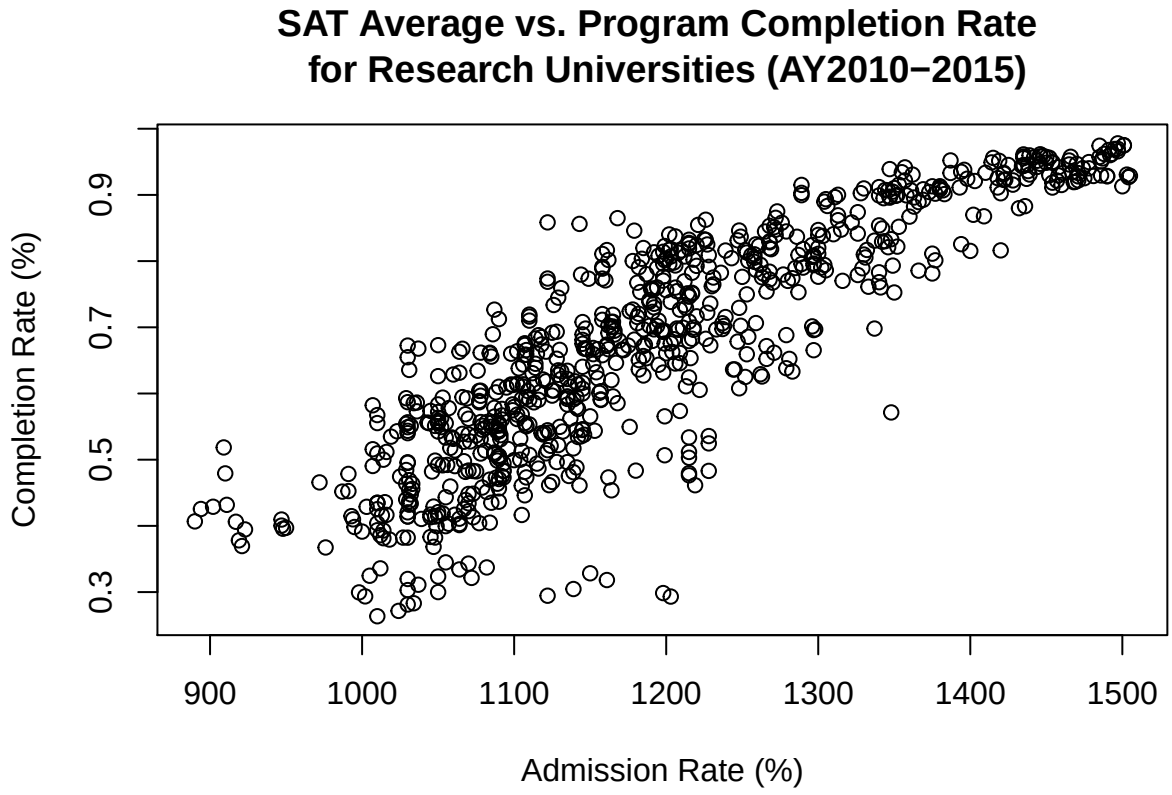
```
#Correlation between the SAT grades and the acceptance for the research universities
plot(usunivfilter$SAT_AVG_ALL, usunivfilter$ACCEPTED, main="SAT Average Grades vs. \n Acceptance to Res
```




```
#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 \
```

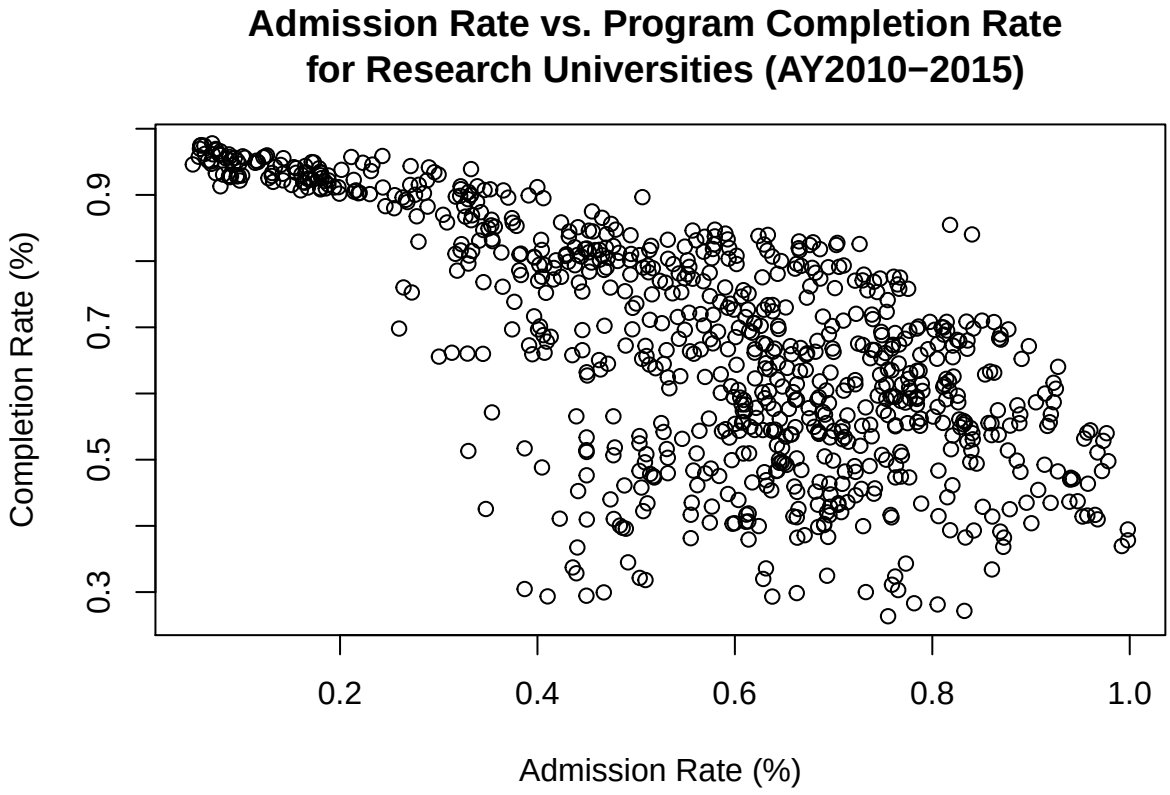



```
#Correlation coefficient between admission rate and completion rate
cor(usresearchuniv$SAT_AVG_ALL, usresearchuniv$C150_4, method = "pearson")
```

```
## [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
```

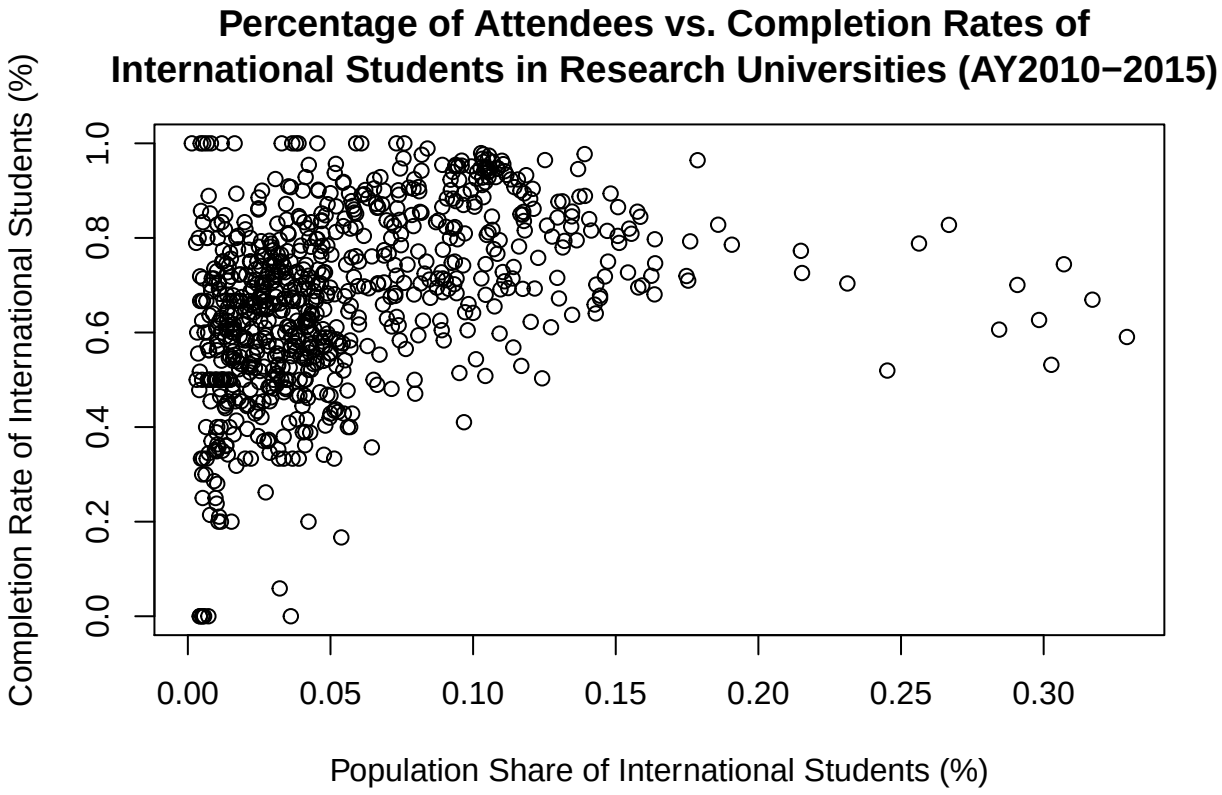



```
#Correlation coefficient between admission rate and completion rate
cor(usresearchuniv$ADM_RATE_ALL, usresearchuniv$C150_4, method = "pearson")
```

```
## [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
```

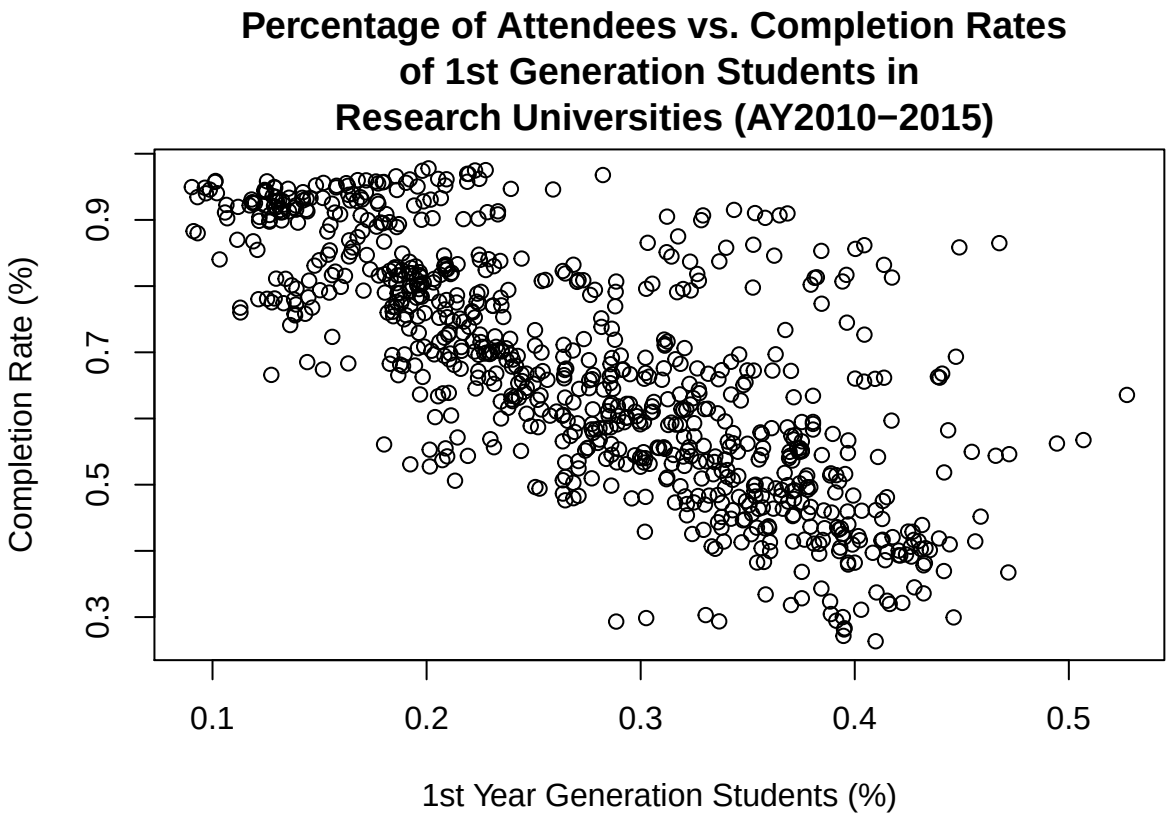



```
#Correlation coefficient between admission rate and completion rate of international students
cor(usresearchuniv$UGDS_NRA, usresearchuniv$C150_4_NRA, method = "pearson")
```

```
## [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
## -2.3718  -0.5290  -0.2902  -0.1226   2.7023
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.510e+01  1.200e+00 -12.581  < 2e-16 ***
## REGION        1.295e-01  2.983e-02   4.342 1.41e-05 ***
## ADM_RATE_ALL  6.051e-01  3.841e-01   1.576 0.115131
## SAT_AVG_ALL   1.475e-02  8.550e-04  17.255  < 2e-16 ***
## UGDS_NRA      7.984e+00  1.367e+00   5.840 5.21e-09 ***
## COSTT4_A     -9.639e-05  6.459e-06 -14.924  < 2e-16 ***
## PCTFLOAN     -4.276e-01  4.957e-01  -0.863 0.388400
## UGDS_WOMEN   -1.805e+00  5.333e-01  -3.385 0.000712 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3052.8  on 3184  degrees of freedom
## Residual deviance: 2090.2  on 3177  degrees of freedom
## AIC: 2106.2
##
## 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
##          0 798 124
##          1  39 101
```

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

```
## Accuracy
## 0.846516
```

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

```
## Neg Pred Value
##      0.7214286
```

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

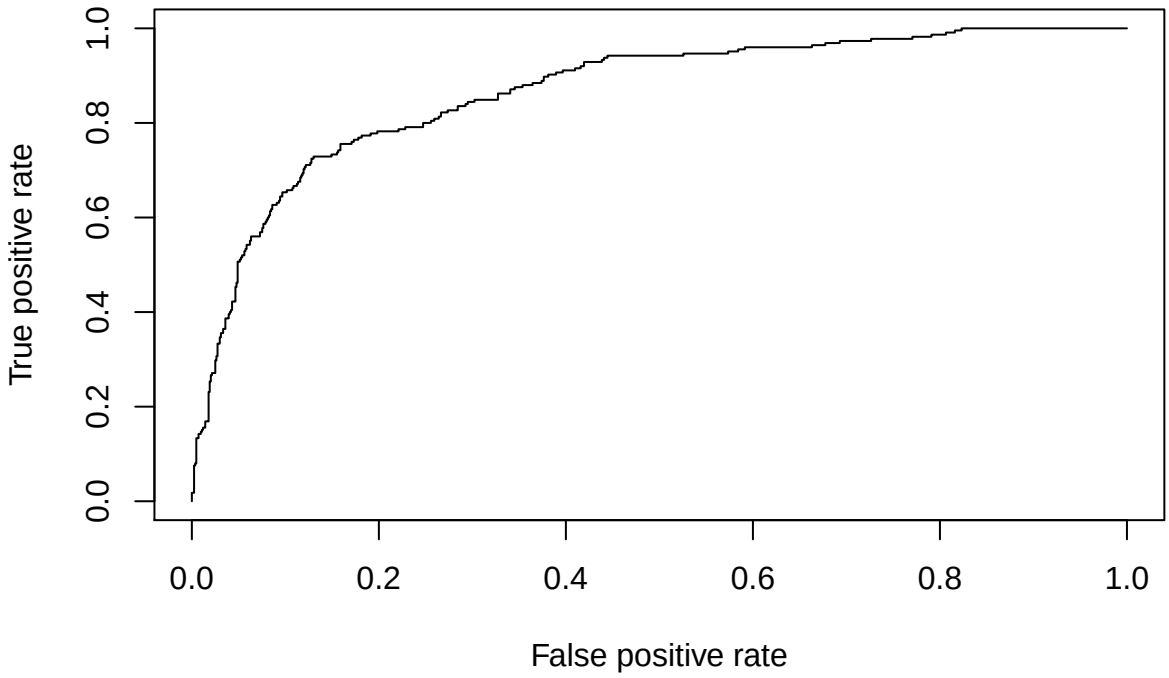
```
## Specificity
##   0.4488889
```

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  0.12607669      0 1.0000000 1.0006518 0.02873066
## 2  0.06105988      1 0.8739233 0.8799653 0.02721397
## 3  0.03619644      3 0.7518035 0.7675499 0.02097609
## 4  0.02473128      5 0.6794107 0.7071576 0.02283742
## 5  0.01693606      6 0.6546794 0.6784317 0.02375405
## 6  0.01531058      8 0.6208073 0.6685045 0.02425859
## 7  0.01499247      9 0.6054967 0.6577531 0.02429686
## 8  0.01309608     10 0.5905042 0.6494780 0.02452321
## 9  0.01233887     11 0.5774081 0.6365913 0.02494318
## 10 0.01213212     12 0.5650692 0.6330536 0.02504019
## 11 0.01184526     13 0.5529371 0.6305995 0.02507268
## 12 0.01000000     14 0.5410919 0.6264104 0.02524221
##
## Variable importance
##  SAT_AVG_ALL     COSTT4_A     PCTFLOAN ADM_RATE_ALL     UGDS_NRA
##           31           22           14           12            7
```

```
##   UGDS_WOMEN       REGION
##            7            7
##
## Node number 1: 3185 observations,    complexity param=0.1260767
##   mean=0.1852433, MSE=0.1509282
##   left son=2 (2605 obs) right son=3 (580 obs)
##   Primary splits:
##       SAT_AVG_ALL  < 1183.5    to the left,  improve=0.12607670, (0 missing)
##       PCTFLOAN     < 0.49355   to the right, improve=0.11604830, (0 missing)
##       UGDS_WOMEN   < 0.52825   to the right, improve=0.08505936, (0 missing)
##       ADM_RATE_ALL < 0.3570456 to the right, improve=0.05307651, (0 missing)
##       UGDS_NRA     < 0.03055   to the left,  improve=0.03973025, (0 missing)
##   Surrogate splits:
##       COSTT4_A     < 51565.5   to the left,  agree=0.889, adj=0.390, (0 split)
##       ADM_RATE_ALL < 0.3680017 to the right, agree=0.868, adj=0.278, (0 split)
##       PCTFLOAN     < 0.37295   to the right, agree=0.835, adj=0.097, (0 split)
##       UGDS_NRA     < 0.09575   to the left,  agree=0.819, adj=0.009, (0 split)
##
## Node number 2: 2605 observations,    complexity param=0.06105988
##   mean=0.1201536, MSE=0.1057167
##   left son=4 (1280 obs) right son=5 (1325 obs)
##   Primary splits:
##       COSTT4_A    < 27966.5   to the right, improve=0.09108962, (0 missing)
##       PCTFLOAN    < 0.6147    to the right, improve=0.06861081, (0 missing)
##       SAT_AVG_ALL < 1028.5    to the left,  improve=0.05304131, (0 missing)
##       UGDS_WOMEN  < 0.56775   to the right, improve=0.04921139, (0 missing)
##       REGION      < 4.5       to the left,  improve=0.01829512, (0 missing)
##   Surrogate splits:
##       PCTFLOAN    < 0.60705   to the right, agree=0.702, adj=0.393, (0 split)
##       UGDS_WOMEN  < 0.6047    to the right, agree=0.612, adj=0.210, (0 split)
##       SAT_AVG_ALL < 1030.5    to the right, agree=0.581, adj=0.148, (0 split)
##       REGION      < 4.5       to the left,  agree=0.577, adj=0.140, (0 split)
##       UGDS_NRA    < 0.05475   to the right, agree=0.565, adj=0.115, (0 split)
##
## Node number 3: 580 observations,    complexity param=0.03619644
##   mean=0.4775862, MSE=0.2494976
##   left son=6 (408 obs) right son=7 (172 obs)
##   Primary splits:
##       COSTT4_A    < 33250     to the right, improve=0.11491320, (0 missing)
##       UGDS_WOMEN  < 0.51655   to the right, improve=0.10137280, (0 missing)
##       PCTFLOAN    < 0.4917    to the right, improve=0.09885532, (0 missing)
##       SAT_AVG_ALL < 1434.5    to the left,  improve=0.06975453, (0 missing)
##       UGDS_NRA    < 0.08615   to the left,  improve=0.06302371, (0 missing)
##   Surrogate splits:
##       UGDS_NRA     < 0.0151    to the right, agree=0.743, adj=0.134, (0 split)
##       SAT_AVG_ALL  < 1194.5    to the right, agree=0.709, adj=0.017, (0 split)
##       ADM_RATE_ALL < 0.8409027 to the left,  agree=0.707, adj=0.012, (0 split)
##       PCTFLOAN     < 0.785     to the left,  agree=0.707, adj=0.012, (0 split)
##
## Node number 4: 1280 observations
##   mean=0.0203125, MSE=0.0198999
##
## Node number 5: 1325 observations,    complexity param=0.06105988
##   mean=0.2166038, MSE=0.1696866
```

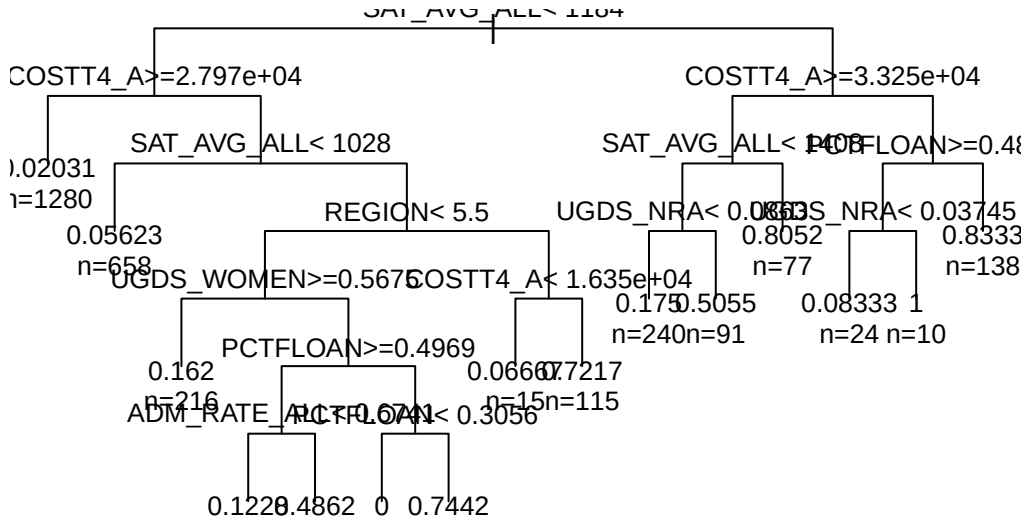
```
##   left son=10 (658 obs) right son=11 (667 obs)
##   Primary splits:
##       SAT_AVG_ALL < 1028.5    to the left,  improve=0.14952500, (0 missing)
##       UGDS_WOMEN  < 0.56515   to the right, improve=0.07526070, (0 missing)
##       COSTT4_A    < 18143.5   to the left,  improve=0.05103363, (0 missing)
##       REGION      < 2.5       to the left,  improve=0.03733577, (0 missing)
##       PCTFLOAN    < 0.6147    to the right, improve=0.03568280, (0 missing)
##   Surrogate splits:
##       UGDS_WOMEN   < 0.5299    to the right, agree=0.634, adj=0.263, (0 split)
##       COSTT4_A     < 18133.5   to the left,  agree=0.605, adj=0.204, (0 split)
##       PCTFLOAN     < 0.59805   to the right, agree=0.589, adj=0.172, (0 split)
##       UGDS_NRA     < 0.01805   to the left,  agree=0.580, adj=0.153, (0 split)
##       ADM_RATE_ALL < 0.86195   to the right, agree=0.533, adj=0.059, (0 split)
##
## Node number 6: 408 observations,    complexity param=0.03619644
##   mean=0.3676471, MSE=0.2324827
##   left son=12 (331 obs) right son=13 (77 obs)
##   Primary splits:
##       SAT_AVG_ALL  < 1408.5    to the left,  improve=0.19156800, (0 missing)
##       UGDS_NRA     < 0.08645   to the left,  improve=0.14960590, (0 missing)
##       ADM_RATE_ALL < 0.1321699 to the right, improve=0.13848910, (0 missing)
##       UGDS_WOMEN   < 0.51545   to the right, improve=0.13713820, (0 missing)
##       COSTT4_A     < 53793.5   to the left,  improve=0.06012946, (0 missing)
##   Surrogate splits:
##       ADM_RATE_ALL < 0.1935801 to the right, agree=0.939, adj=0.675, (0 split)
##       PCTFLOAN     < 0.2691    to the right, agree=0.887, adj=0.403, (0 split)
##       COSTT4_A     < 61468     to the left,  agree=0.833, adj=0.117, (0 split)
##
## Node number 7: 172 observations,    complexity param=0.01309608
##   mean=0.7383721, MSE=0.1931787
##   left son=14 (34 obs) right son=15 (138 obs)
##   Primary splits:
##       PCTFLOAN     < 0.48925   to the right, improve=0.18946700, (0 missing)
##       UGDS_NRA     < 0.00425   to the left,  improve=0.12689560, (0 missing)
##       REGION       < 4.5       to the left,  improve=0.11792620, (0 missing)
##       ADM_RATE_ALL < 0.5545675 to the right, improve=0.06970379, (0 missing)
##       SAT_AVG_ALL  < 1294      to the left,  improve=0.06877416, (0 missing)
##   Surrogate splits:
##       UGDS_WOMEN   < 0.2563    to the left,  agree=0.843, adj=0.206, (0 split)
##       UGDS_NRA     < 0.00105   to the left,  agree=0.831, adj=0.147, (0 split)
##       ADM_RATE_ALL < 0.8500731 to the right, agree=0.820, adj=0.088, (0 split)
##
## Node number 10: 658 observations
##   mean=0.056231, MSE=0.05306908
##
## Node number 11: 667 observations,    complexity param=0.02473128
##   mean=0.3748126, MSE=0.2343281
##   left son=22 (537 obs) right son=23 (130 obs)
##   Primary splits:
##       REGION      < 5.5       to the left,  improve=0.07606350, (0 missing)
##       UGDS_WOMEN  < 0.5641    to the right, improve=0.07346194, (0 missing)
##       COSTT4_A    < 17552     to the left,  improve=0.06323340, (0 missing)
##       PCTFLOAN    < 0.62945   to the right, improve=0.04756623, (0 missing)
##       SAT_AVG_ALL < 1106.5    to the left,  improve=0.04060213, (0 missing)
```

```
##   Surrogate splits:
##       ADM_RATE_ALL < 0.9707706 to the left,  agree=0.813, adj=0.038, (0 split)
##       UGDS_NRA     < 0.2649    to the left,  agree=0.808, adj=0.015, (0 split)
##
## Node number 12: 331 observations,    complexity param=0.01499247
##   mean=0.265861, MSE=0.1951789
##   left son=24 (240 obs) right son=25 (91 obs)
##   Primary splits:
##       UGDS_NRA     < 0.0863    to the left,  improve=0.11155580, (0 missing)
##       ADM_RATE_ALL < 0.5756857 to the right, improve=0.03697151, (0 missing)
##       UGDS_WOMEN   < 0.51395   to the right, improve=0.03417994, (0 missing)
##       REGION       < 7.5       to the right, improve=0.03021531, (0 missing)
##       SAT_AVG_ALL  < 1335.5    to the left,  improve=0.02651130, (0 missing)
##   Surrogate splits:
##       ADM_RATE_ALL < 0.19435   to the right, agree=0.758, adj=0.121, (0 split)
##       REGION       < 1.5       to the right, agree=0.752, adj=0.099, (0 split)
##       UGDS_WOMEN   < 0.6347    to the left,  agree=0.746, adj=0.077, (0 split)
##       SAT_AVG_ALL  < 1376.5    to the left,  agree=0.740, adj=0.055, (0 split)
##       COSTT4_A     < 59755     to the left,  agree=0.737, adj=0.044, (0 split)
##
## Node number 13: 77 observations
##   mean=0.8051948, MSE=0.1568561
##
## Node number 14: 34 observations,    complexity param=0.01233887
##   mean=0.3529412, MSE=0.2283737
##   left son=28 (24 obs) right son=29 (10 obs)
##   Primary splits:
##       UGDS_NRA     < 0.03745   to the left,  improve=0.76388890, (0 missing)
##       UGDS_WOMEN   < 0.4983    to the right, improve=0.57408010, (0 missing)
##       COSTT4_A     < 20645.5   to the left,  improve=0.22727270, (0 missing)
##       ADM_RATE_ALL < 0.6729404 to the left,  improve=0.21976010, (0 missing)
##       SAT_AVG_ALL  < 1205      to the left,  improve=0.05892857, (0 missing)
##   Surrogate splits:
##       UGDS_WOMEN   < 0.4437    to the right, agree=0.882, adj=0.6, (0 split)
##       REGION       < 3.5       to the left,  agree=0.824, adj=0.4, (0 split)
##       ADM_RATE_ALL < 0.6729404 to the left,  agree=0.824, adj=0.4, (0 split)
##
## Node number 15: 138 observations
##   mean=0.8333333, MSE=0.1388889
##
## Node number 22: 537 observations,    complexity param=0.01693606
##   mean=0.3091248, MSE=0.2135666
##   left son=44 (216 obs) right son=45 (321 obs)
##   Primary splits:
##       UGDS_WOMEN  < 0.56755   to the right, improve=0.06816606, (0 missing)
##       SAT_AVG_ALL < 1138      to the left,  improve=0.06573282, (0 missing)
##       REGION      < 2.5       to the left,  improve=0.04574707, (0 missing)
##       COSTT4_A    < 17552     to the left,  improve=0.04558109, (0 missing)
##       PCTFLOAN    < 0.4948    to the right, improve=0.04174317, (0 missing)
##   Surrogate splits:
##       SAT_AVG_ALL  < 1049.5    to the left,  agree=0.665, adj=0.167, (0 split)
##       COSTT4_A     < 17300     to the left,  agree=0.633, adj=0.088, (0 split)
##       UGDS_NRA     < 0.0044    to the left,  agree=0.616, adj=0.046, (0 split)
##       ADM_RATE_ALL < 0.3281553 to the left,  agree=0.607, adj=0.023, (0 split)
```

```
##       PCTFLOAN     < 0.76005   to the right, agree=0.601, adj=0.009, (0 split)
##
## Node number 23: 130 observations,    complexity param=0.01184526
##   mean=0.6461538, MSE=0.2286391
##   left son=46 (15 obs) right son=47 (115 obs)
##   Primary splits:
##       COSTT4_A     < 16350     to the left,  improve=0.19157140, (0 missing)
##       SAT_AVG_ALL  < 1074.5    to the left,  improve=0.12525880, (0 missing)
##       UGDS_NRA     < 0.00995   to the left,  improve=0.12480000, (0 missing)
##       PCTFLOAN     < 0.48965   to the left,  improve=0.08873114, (0 missing)
##       ADM_RATE_ALL < 0.5964316 to the left,  improve=0.08170838, (0 missing)
##   Surrogate splits:
##       PCTFLOAN     < 0.3033    to the left,  agree=0.923, adj=0.333, (0 split)
##       ADM_RATE_ALL < 0.28635   to the left,  agree=0.900, adj=0.133, (0 split)
##       UGDS_NRA     < 0.24815   to the right, agree=0.900, adj=0.133, (0 split)
##
## Node number 24: 240 observations
##   mean=0.175, MSE=0.144375
##
## Node number 25: 91 observations
##   mean=0.5054945, MSE=0.2499698
##
## Node number 28: 24 observations
##   mean=0.08333333, MSE=0.07638889
##
## Node number 29: 10 observations
##   mean=1, MSE=0
##
## Node number 44: 216 observations
##   mean=0.162037, MSE=0.135781
##
## Node number 45: 321 observations,    complexity param=0.01693606
##   mean=0.4080997, MSE=0.2415543
##   left son=90 (223 obs) right son=91 (98 obs)
##   Primary splits:
##       PCTFLOAN    < 0.49685   to the right, improve=0.10916970, (0 missing)
##       REGION      < 2.5       to the left,  improve=0.10126560, (0 missing)
##       SAT_AVG_ALL < 1118.5    to the left,  improve=0.07803735, (0 missing)
##       UGDS_WOMEN  < 0.4316    to the left,  improve=0.07408316, (0 missing)
##       UGDS_NRA    < 0.00985   to the left,  improve=0.06036711, (0 missing)
##   Surrogate splits:
##       SAT_AVG_ALL  < 1125.5    to the left,  agree=0.769, adj=0.245, (0 split)
##       REGION       < 4.5       to the left,  agree=0.763, adj=0.224, (0 split)
##       ADM_RATE_ALL < 0.38755   to the right, agree=0.710, adj=0.051, (0 split)
##       COSTT4_A     < 15671.5   to the right, agree=0.707, adj=0.041, (0 split)
##       UGDS_WOMEN   < 0.56565   to the left,  agree=0.698, adj=0.010, (0 split)
##
## Node number 46: 15 observations
##   mean=0.06666667, MSE=0.06222222
##
## Node number 47: 115 observations
##   mean=0.7217391, MSE=0.2008318
##
## Node number 90: 223 observations,    complexity param=0.01531058
```

```
##   mean=0.3004484, MSE=0.2101792
##   left son=180 (114 obs) right son=181 (109 obs)
##   Primary splits:
##       ADM_RATE_ALL < 0.6741    to the left,  improve=0.15702800, (0 missing)
##       REGION       < 1.5       to the right, improve=0.08462717, (0 missing)
##       UGDS_NRA     < 0.02315   to the left,  improve=0.07137413, (0 missing)
##       UGDS_WOMEN   < 0.43185   to the left,  improve=0.06788575, (0 missing)
##       SAT_AVG_ALL  < 1031.5    to the right, improve=0.02525419, (0 missing)
##   Surrogate splits:
##       REGION      < 2.5       to the left,  agree=0.807, adj=0.606, (0 split)
##       SAT_AVG_ALL < 1066.5    to the right, agree=0.673, adj=0.330, (0 split)
##       COSTT4_A    < 18867.5   to the right, agree=0.646, adj=0.275, (0 split)
##       UGDS_NRA    < 0.01625   to the left,  agree=0.623, adj=0.229, (0 split)
##       PCTFLOAN    < 0.65225   to the right, agree=0.610, adj=0.202, (0 split)
##
## Node number 91: 98 observations,    complexity param=0.01213212
##   mean=0.6530612, MSE=0.2265723
##   left son=182 (12 obs) right son=183 (86 obs)
##   Primary splits:
##       PCTFLOAN     < 0.3056    to the left,  improve=0.2626539, (0 missing)
##       COSTT4_A     < 19026.5   to the left,  improve=0.2102419, (0 missing)
##       SAT_AVG_ALL  < 1143.5    to the left,  improve=0.2007555, (0 missing)
##       UGDS_NRA     < 0.0089    to the left,  improve=0.1673203, (0 missing)
##       ADM_RATE_ALL < 0.4014088 to the left,  improve=0.1093980, (0 missing)
##   Surrogate splits:
##       ADM_RATE_ALL < 0.383785  to the left,  agree=0.939, adj=0.500, (0 split)
##       REGION       < 2.5       to the left,  agree=0.918, adj=0.333, (0 split)
##       COSTT4_A     < 15069.5   to the left,  agree=0.908, adj=0.250, (0 split)
##
## Node number 180: 114 observations
##   mean=0.122807, MSE=0.1077255
##
## Node number 181: 109 observations
##   mean=0.4862385, MSE=0.2498106
##
## Node number 182: 12 observations
##   mean=0, MSE=0
##
## Node number 183: 86 observations
##   mean=0.744186, MSE=0.1903732
```

```
plot(model_dtree1, uniform = TRUE, main = "Single Decision Tree  of\nUS Research University Prediciton 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] 0.8681733
```

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

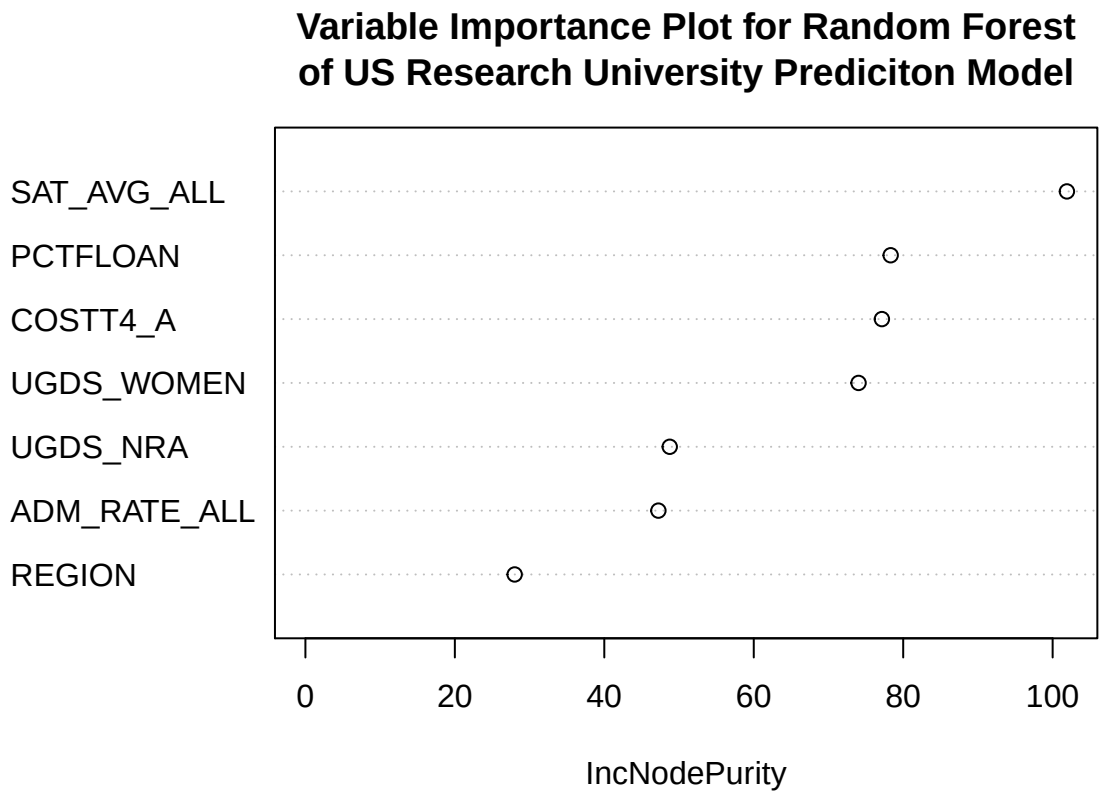
```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values. Are you sure you want to do regression?
```

```
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] 0.933145
```

```
# 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:  1364
```

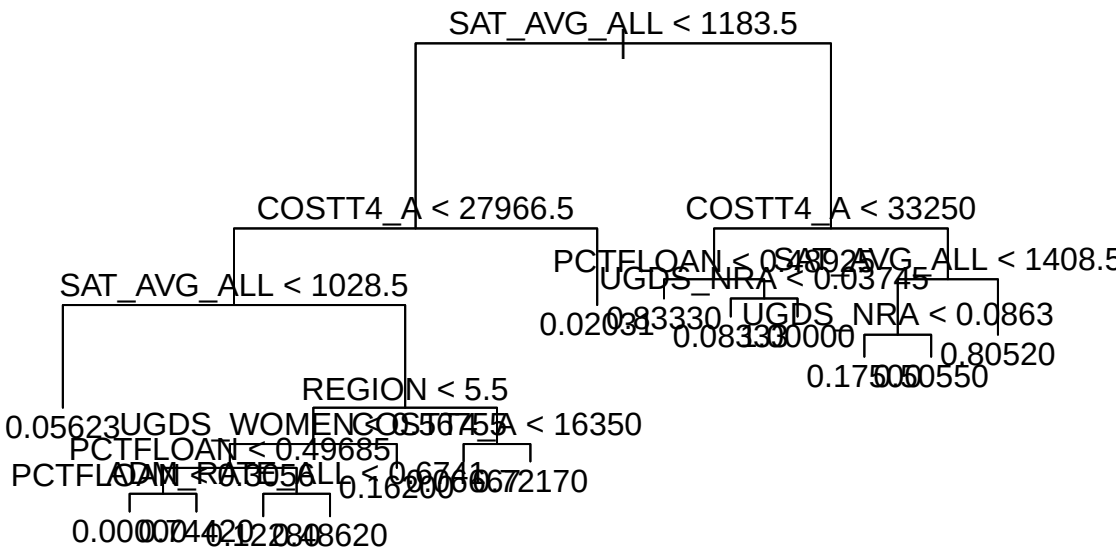
```
pred_svm1 <- predict(model_svm1, newdata = univ_test)
accu3 <- abs(pred_svm1 - univ_test$ACCEPTED) < 0.5
frac3 <- sum(accu3)/length(accu3)
print(frac3)
```

```
## [1] 0.8757062
```

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

```
##
## Regression tree:
## tree(formula = formula_ISAcceptance, data = univ_train)
## Number of terminal nodes:  15
## Residual mean deviance:  0.08205 = 260.1 / 3170
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.83330 -0.05623 -0.02031  0.00000 -0.02031  0.97970
```

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



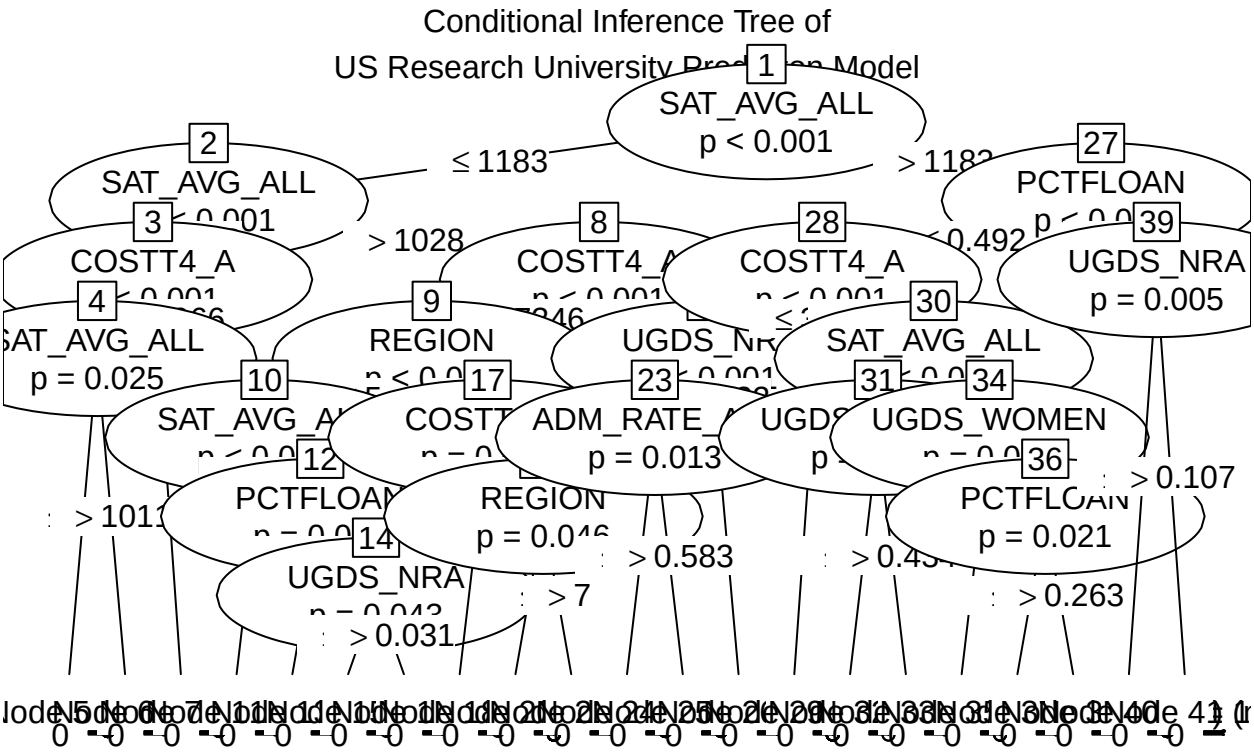

```
pred_tree1 <- predict(model_tree1, newdata = univ_test)
accu4 <- abs(pred_tree1 - univ_test$ACCEPTED) < 0.5
frac4 <- sum(accu4)/length(accu4)
print(frac4)
```

```
## [1] 0.8681733
```

```
# 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 Prediciton 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] 0.8596987
```

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
```

```
## -2.51711  -0.46763  -0.23106  -0.08032   3.11252
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.819e+01  1.447e+00 -12.568  < 2e-16 ***
## REGION        1.524e-01  3.261e-02   4.675 2.94e-06 ***
## ADM_RATE_ALL  9.423e-01  4.280e-01   2.202 0.027669 *
## SAT_AVG_ALL   1.609e-02  1.036e-03  15.533  < 2e-16 ***
## PCIP11        1.034e+00  2.243e+00   0.461 0.644759
## PCIP12       -5.412e+00  2.383e+01  -0.227 0.820352
## PCIP14        5.284e+00  7.825e-01   6.753 1.45e-11 ***
## PCIP15       -4.636e-01  2.229e+00  -0.208 0.835214
## PCIP24       -6.050e+00  1.242e+00  -4.873 1.10e-06 ***
## PCIP26        7.704e+00  1.824e+00   4.224 2.40e-05 ***
## PCIP27       -3.056e+01  7.269e+00  -4.204 2.62e-05 ***
## PCIP40       -3.465e+01  4.962e+00  -6.984 2.88e-12 ***
## PCIP45        8.301e+00  1.206e+00   6.881 5.93e-12 ***
## PCIP51        2.280e+00  6.128e-01   3.720 0.000199 ***
## PCIP52        4.144e-01  6.540e-01   0.634 0.526276
## UGDS_NRA      1.082e+01  1.572e+00   6.883 5.88e-12 ***
## UGDS_UNKN    -1.644e+00  1.599e+00  -1.028 0.303975
## COSTT4_A     -1.160e-04  7.616e-06 -15.235  < 2e-16 ***
## PCTFLOAN     -1.818e-01  5.749e-01  -0.316 0.751759
## UGDS_WOMEN    2.478e-01  8.024e-01   0.309 0.757435
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3052.8  on 3184  degrees of freedom
## Residual deviance: 1839.4  on 3165  degrees of freedom
## AIC: 1879.4
##
## Number of Fisher Scoring iterations: 6
```

```
# 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
##          0 796 109
##          1  41 116
```

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

```
##  Accuracy
## 0.8587571
```

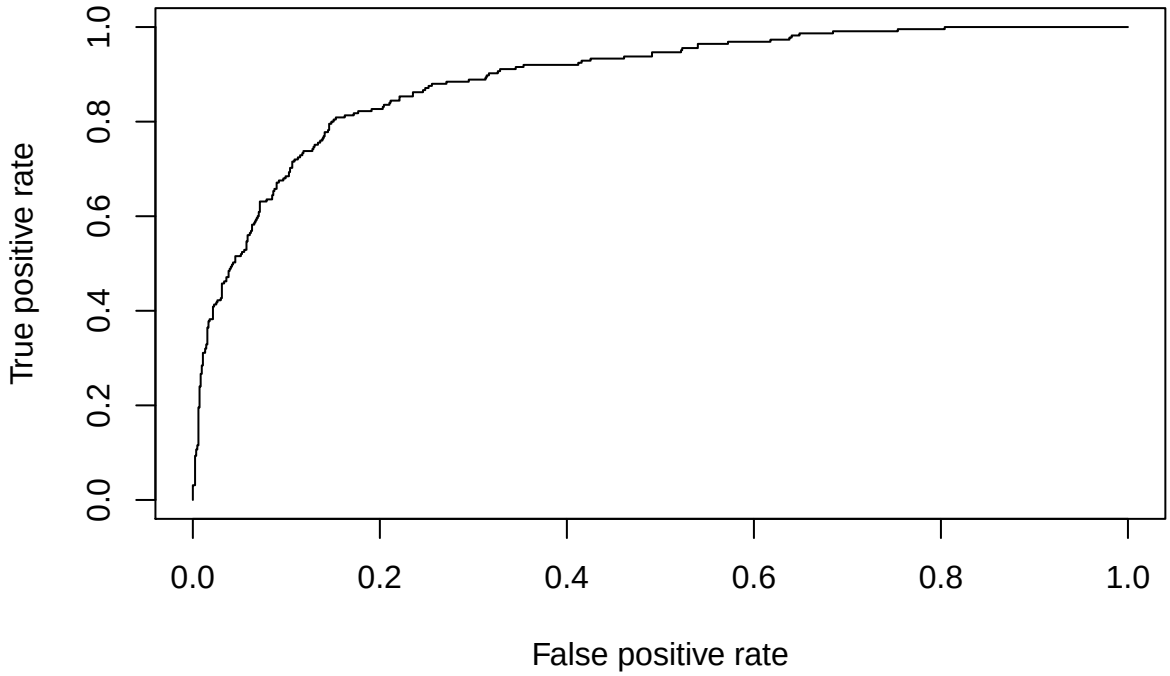
```r
#Precision of the logistic regression model
c2$byClass['Neg Pred Value']
```

```
## Neg Pred Value
##      0.7388535
```

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

```
## Specificity
##   0.5155556
```

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




```r
# 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  0.30665795      0 1.0000000 1.0004619 0.02872522
## 2  0.06769369      1 0.6933421 0.6945412 0.02281588
## 3  0.04258518      2 0.6256484 0.6292733 0.02457547
## 4  0.03049514      3 0.5830632 0.5963441 0.02521387
## 5  0.01366054      4 0.5525680 0.5745269 0.02552859
## 6  0.01343097      5 0.5389075 0.5682624 0.02588794
## 7  0.01312960      6 0.5254765 0.5656665 0.02592561
## 8  0.01143779      7 0.5123469 0.5567499 0.02609446
## 9  0.01109708      8 0.5009092 0.5546416 0.02619277
## 10 0.01064777      9 0.4898121 0.5476650 0.02616162
## 11 0.01043254     11 0.4685165 0.5416316 0.02602606
## 12 0.01000000     12 0.4580840 0.5395116 0.02592798
##
## Variable importance
##       PCIP14  SAT_AVG_ALL     PCTFLOAN   UGDS_WOMEN ADM_RATE_ALL
##           36           14           10            8            7
##       PCIP45     COSTT4_A       PCIP26       PCIP51       REGION
##            6            5            3            2            2
##    UGDS_UNKN       PCIP52       PCIP11       PCIP40       PCIP27
##            1            1            1            1            1
##     UGDS_NRA       PCIP15
##            1            1
##
## Node number 1: 3185 observations,    complexity param=0.3066579
##   mean=0.1852433, MSE=0.1509282
##   left son=2 (2320 obs) right son=3 (865 obs)
##   Primary splits:
##       PCIP14      < 0.0269    to the left,  improve=0.30665790, (0 missing)
##       SAT_AVG_ALL < 1183.5    to the left,  improve=0.12607670, (0 missing)
##       PCTFLOAN    < 0.49355   to the right, improve=0.11604830, (0 missing)
##       UGDS_WOMEN  < 0.52825   to the right, improve=0.08505936, (0 missing)
##       PCIP40      < 0.00615   to the left,  improve=0.05588648, (0 missing)
##   Surrogate splits:
##       UGDS_WOMEN   < 0.5158    to the right, agree=0.782, adj=0.199, (0 split)
##       SAT_AVG_ALL  < 1174.5    to the left,  agree=0.745, adj=0.062, (0 split)
##       ADM_RATE_ALL < 0.20305   to the right, agree=0.740, adj=0.042, (0 split)
##       PCIP11       < 0.0661    to the left,  agree=0.733, adj=0.018, (0 split)
##       COSTT4_A     < 56914     to the left,  agree=0.732, adj=0.014, (0 split)
##
## Node number 2: 2320 observations,    complexity param=0.01343097
##   mean=0.05387931, MSE=0.05097633
##   left son=4 (2309 obs) right son=5 (11 obs)
##   Primary splits:
##       PCIP45      < 0.3441    to the left,  improve=0.05459222, (0 missing)
##       SAT_AVG_ALL < 1194.5    to the left,  improve=0.04768078, (0 missing)
##       PCTFLOAN    < 0.62345   to the right, improve=0.03200701, (0 missing)
##       COSTT4_A    < 53519.5   to the left,  improve=0.02689840, (0 missing)
##       PCIP14      < 0.00585   to the left,  improve=0.01839136, (0 missing)
##   Surrogate splits:
##       SAT_AVG_ALL < 1456.5    to the left,  agree=0.996, adj=0.182, (0 split)
##       COSTT4_A    < 62161.5   to the left,  agree=0.996, adj=0.091, (0 split)
```

```
##
## Node number 3: 865 observations,    complexity param=0.06769369
##   mean=0.5375723, MSE=0.2485883
##   left son=6 (405 obs) right son=7 (460 obs)
##   Primary splits:
##       PCTFLOAN    < 0.51465   to the right, improve=0.15133220, (0 missing)
##       PCIP45      < 0.03155   to the left,  improve=0.13021730, (0 missing)
##       PCIP26      < 0.02435   to the left,  improve=0.10523000, (0 missing)
##       SAT_AVG_ALL < 1120.5    to the left,  improve=0.08978442, (0 missing)
##       PCIP40      < 0.00695   to the left,  improve=0.08562038, (0 missing)
##   Surrogate splits:
##       PCIP45       < 0.03875   to the left,  agree=0.702, adj=0.363, (0 split)
##       SAT_AVG_ALL  < 1120.5    to the left,  agree=0.691, adj=0.341, (0 split)
##       ADM_RATE_ALL < 0.60835   to the right, agree=0.680, adj=0.316, (0 split)
##       REGION       < 4.5       to the left,  agree=0.637, adj=0.225, (0 split)
##       PCIP26       < 0.04765   to the left,  agree=0.635, adj=0.220, (0 split)
##
## Node number 4: 2309 observations,    complexity param=0.01064777
##   mean=0.0502382, MSE=0.04771432
##   left son=8 (2077 obs) right son=9 (232 obs)
##   Primary splits:
##       SAT_AVG_ALL < 1194.5    to the left,  improve=0.03252178, (0 missing)
##       PCTFLOAN    < 0.62345   to the right, improve=0.02863470, (0 missing)
##       PCIP14      < 0.00415   to the left,  improve=0.02138492, (0 missing)
##       PCIP45      < 0.07155   to the left,  improve=0.01977433, (0 missing)
##       COSTT4_A    < 26322.5   to the right, improve=0.01576139, (0 missing)
##   Surrogate splits:
##       COSTT4_A     < 52398.5   to the left,  agree=0.935, adj=0.349, (0 split)
##       PCIP45       < 0.19985   to the left,  agree=0.929, adj=0.293, (0 split)
##       ADM_RATE_ALL < 0.3507965 to the right, agree=0.922, adj=0.228, (0 split)
##       PCIP40       < 0.0546    to the left,  agree=0.913, adj=0.134, (0 split)
##       PCIP27       < 0.0451    to the left,  agree=0.905, adj=0.052, (0 split)
##
## Node number 5: 11 observations
##   mean=0.8181818, MSE=0.1487603
##
## Node number 6: 405 observations,    complexity param=0.03049514
##   mean=0.3308642, MSE=0.2213931
##   left son=12 (213 obs) right son=13 (192 obs)
##   Primary splits:
##       COSTT4_A < 26481.5   to the right, improve=0.16349010, (0 missing)
##       PCIP45   < 0.0154    to the left,  improve=0.12859600, (0 missing)
##       PCTFLOAN < 0.67325   to the right, improve=0.08363907, (0 missing)
##       PCIP26   < 0.02005   to the left,  improve=0.07425912, (0 missing)
##       PCIP40   < 0.00505   to the left,  improve=0.07328137, (0 missing)
##   Surrogate splits:
##       SAT_AVG_ALL  < 1080.5    to the right, agree=0.691, adj=0.349, (0 split)
##       UGDS_NRA     < 0.05185   to the right, agree=0.649, adj=0.260, (0 split)
##       ADM_RATE_ALL < 0.7622236 to the left,  agree=0.637, adj=0.234, (0 split)
##       PCTFLOAN     < 0.6056    to the right, agree=0.630, adj=0.219, (0 split)
##       PCIP15       < 0.00515   to the left,  agree=0.622, adj=0.203, (0 split)
##
## Node number 7: 460 observations,    complexity param=0.04258518
##   mean=0.7195652, MSE=0.2017911
```

```
##   left son=14 (69 obs) right son=15 (391 obs)
##   Primary splits:
##       SAT_AVG_ALL < 1065.5    to the left,  improve=0.22053550, (0 missing)
##       PCIP26      < 0.03355   to the left,  improve=0.11088820, (0 missing)
##       PCIP24      < 0.01125   to the right, improve=0.09910066, (0 missing)
##       COSTT4_A    < 18908.5   to the left,  improve=0.08512946, (0 missing)
##       UGDS_UNKN   < 0.00085   to the left,  improve=0.06872924, (0 missing)
##   Surrogate splits:
##       COSTT4_A   < 15777     to the left,  agree=0.870, adj=0.130, (0 split)
##       PCIP26     < 0.02535   to the left,  agree=0.867, adj=0.116, (0 split)
##       PCIP24     < 0.0772    to the right, agree=0.861, adj=0.072, (0 split)
##       UGDS_WOMEN < 0.58055   to the right, agree=0.859, adj=0.058, (0 split)
##       PCIP12     < 0.005     to the right, agree=0.857, adj=0.043, (0 split)
##
## Node number 8: 2077 observations
##   mean=0.0370727, MSE=0.03569832
##
## Node number 9: 232 observations,    complexity param=0.01064777
##   mean=0.1681034, MSE=0.1398447
##   left son=18 (208 obs) right son=19 (24 obs)
##   Primary splits:
##       PCIP51   < 0.0698    to the left,  improve=0.20508900, (0 missing)
##       PCIP52   < 0.13765   to the left,  improve=0.15445720, (0 missing)
##       PCIP14   < 6e-04     to the left,  improve=0.12147500, (0 missing)
##       PCIP40   < 0.02975   to the right, improve=0.09552187, (0 missing)
##       COSTT4_A < 28481     to the right, improve=0.09466655, (0 missing)
##   Surrogate splits:
##       PCIP11 < 0.06685   to the left,  agree=0.901, adj=0.042, (0 split)
##
## Node number 12: 213 observations,    complexity param=0.01143779
##   mean=0.1502347, MSE=0.1276643
##   left son=24 (159 obs) right son=25 (54 obs)
##   Primary splits:
##       ADM_RATE_ALL < 0.5837101 to the right, improve=0.2021962, (0 missing)
##       PCIP45       < 0.11255   to the left,  improve=0.1752086, (0 missing)
##       COSTT4_A     < 52496.5   to the left,  improve=0.1632919, (0 missing)
##       SAT_AVG_ALL  < 1335.5    to the left,  improve=0.1166243, (0 missing)
##       PCTFLOAN     < 0.5842    to the right, improve=0.0923882, (0 missing)
##   Surrogate splits:
##       SAT_AVG_ALL < 1308.5    to the left,  agree=0.808, adj=0.241, (0 split)
##       COSTT4_A    < 52887     to the left,  agree=0.793, adj=0.185, (0 split)
##       PCIP40      < 0.0532    to the left,  agree=0.770, adj=0.093, (0 split)
##       PCIP45      < 0.18875   to the left,  agree=0.765, adj=0.074, (0 split)
##       PCIP14      < 0.03105   to the right, agree=0.756, adj=0.037, (0 split)
##
## Node number 13: 192 observations,    complexity param=0.01366054
##   mean=0.53125, MSE=0.2490234
##   left son=26 (27 obs) right son=27 (165 obs)
##   Primary splits:
##       PCIP45     < 0.01385   to the left,  improve=0.1373429, (0 missing)
##       PCIP51     < 3e-04     to the left,  improve=0.1325490, (0 missing)
##       PCIP26     < 0.014     to the left,  improve=0.1172414, (0 missing)
##       UGDS_WOMEN < 0.22175   to the left,  improve=0.1030303, (0 missing)
##       COSTT4_A   < 18195     to the left,  improve=0.1024003, (0 missing)
```

```
##   Surrogate splits:
##       PCIP27 < 0.00045   to the left,  agree=0.938, adj=0.556, (0 split)
##       PCIP26 < 0.014     to the left,  agree=0.922, adj=0.444, (0 split)
##       PCIP40 < 0.00195   to the left,  agree=0.922, adj=0.444, (0 split)
##       PCIP51 < 3e-04     to the left,  agree=0.911, adj=0.370, (0 split)
##       PCIP11 < 0.00135   to the left,  agree=0.901, adj=0.296, (0 split)
##
## Node number 14: 69 observations
##   mean=0.2173913, MSE=0.1701323
##
## Node number 15: 391 observations,    complexity param=0.01109708
##   mean=0.8081841, MSE=0.1550225
##   left son=30 (8 obs) right son=31 (383 obs)
##   Primary splits:
##       UGDS_UNKN  < 0.00085   to the left,  improve=0.08800696, (0 missing)
##       PCIP52     < 0.30305   to the right, improve=0.07748562, (0 missing)
##       PCIP51     < 0.0054    to the left,  improve=0.06686897, (0 missing)
##       UGDS_WOMEN < 0.26465   to the left,  improve=0.04372425, (0 missing)
##       UGDS_NRA   < 0.08615   to the left,  improve=0.03851656, (0 missing)
##   Surrogate splits:
##       UGDS_WOMEN < 0.14275   to the left,  agree=0.982, adj=0.125, (0 split)
##
## Node number 18: 208 observations
##   mean=0.1105769, MSE=0.09834967
##
## Node number 19: 24 observations
##   mean=0.6666667, MSE=0.2222222
##
## Node number 24: 159 observations
##   mean=0.05660377, MSE=0.05339979
##
## Node number 25: 54 observations
##   mean=0.4259259, MSE=0.244513
##
## Node number 26: 27 observations
##   mean=0.07407407, MSE=0.06858711
##
## Node number 27: 165 observations,    complexity param=0.0131296
##   mean=0.6060606, MSE=0.2387511
##   left son=54 (26 obs) right son=55 (139 obs)
##   Primary splits:
##       SAT_AVG_ALL  < 990.5     to the left,  improve=0.16021460, (0 missing)
##       COSTT4_A     < 18123.5   to the left,  improve=0.12622380, (0 missing)
##       UGDS_NRA     < 0.01285   to the left,  improve=0.11613540, (0 missing)
##       PCTFLOAN     < 0.7306    to the right, improve=0.07879614, (0 missing)
##       ADM_RATE_ALL < 0.4384291 to the left,  improve=0.06815969, (0 missing)
##   Surrogate splits:
##       PCTFLOAN     < 0.7441    to the right, agree=0.933, adj=0.577, (0 split)
##       ADM_RATE_ALL < 0.429423  to the left,  agree=0.879, adj=0.231, (0 split)
##       PCIP26       < 0.1014    to the right, agree=0.855, adj=0.077, (0 split)
##       UGDS_WOMEN   < 0.62      to the right, agree=0.855, adj=0.077, (0 split)
##
## Node number 30: 8 observations
##   mean=0, MSE=0
```

```
##
## Node number 31: 383 observations,    complexity param=0.01043254
##   mean=0.8250653, MSE=0.1443326
##   left son=62 (17 obs) right son=63 (366 obs)
##   Primary splits:
##       PCIP52    < 0.30305   to the right, improve=0.09072080, (0 missing)
##       PCIP51    < 0.0026    to the left,  improve=0.05966030, (0 missing)
##       UGDS_UNKN < 0.1177    to the right, improve=0.03752225, (0 missing)
##       UGDS_NRA  < 0.07935   to the left,  improve=0.03458145, (0 missing)
##       PCIP26    < 0.0238    to the left,  improve=0.03356125, (0 missing)
##
## Node number 54: 26 observations
##   mean=0.1538462, MSE=0.1301775
##
## Node number 55: 139 observations
##   mean=0.6906475, MSE=0.2136535
##
## Node number 62: 17 observations
##   mean=0.2941176, MSE=0.2076125
##
## Node number 63: 366 observations
##   mean=0.8497268, MSE=0.1276912
```

```
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] 0.9067797
```

```
# 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] 0.9642185
```

```
# 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:  1636
```

```
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] 0.9171375
```

```
# 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"
## [5] "UGDS_UNKN"    "PCIP52"       "COSTT4_A"     "ADM_RATE_ALL"
## Number of terminal nodes:  11
## Residual mean deviance:  0.0726 = 230.4 / 3174
```

```
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.84970 -0.05024 -0.05024  0.00000 -0.05024  0.94980
```

```
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] 0.9096045
```

```
# 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] 0.9020716
```

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
```

```r
# 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 26.59466 secs.
##  60 attributes confirmed important: ADM_RATE, ADM_RATE_ALL,
## C150_4, C150_4_AIAN, C150_4_ASIAN and 55 more.
##  7 attributes confirmed unimportant: C150_4_NHPI, PCIP12, PCIP25,
## PCIP29, PCIP46 and 2 more.
##  3 tentative attributes left: C150_4_2MOR, PCIP10, PCIP22.
```

```r
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_NRA"
## [55] "C150_4_UNKN"       "RET_FT4"           "PCTFLOAN"
## [58] "PAR_ED_PCT_1STGEN" "UGDS_MEN"          "UGDS_WOMEN"
```

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

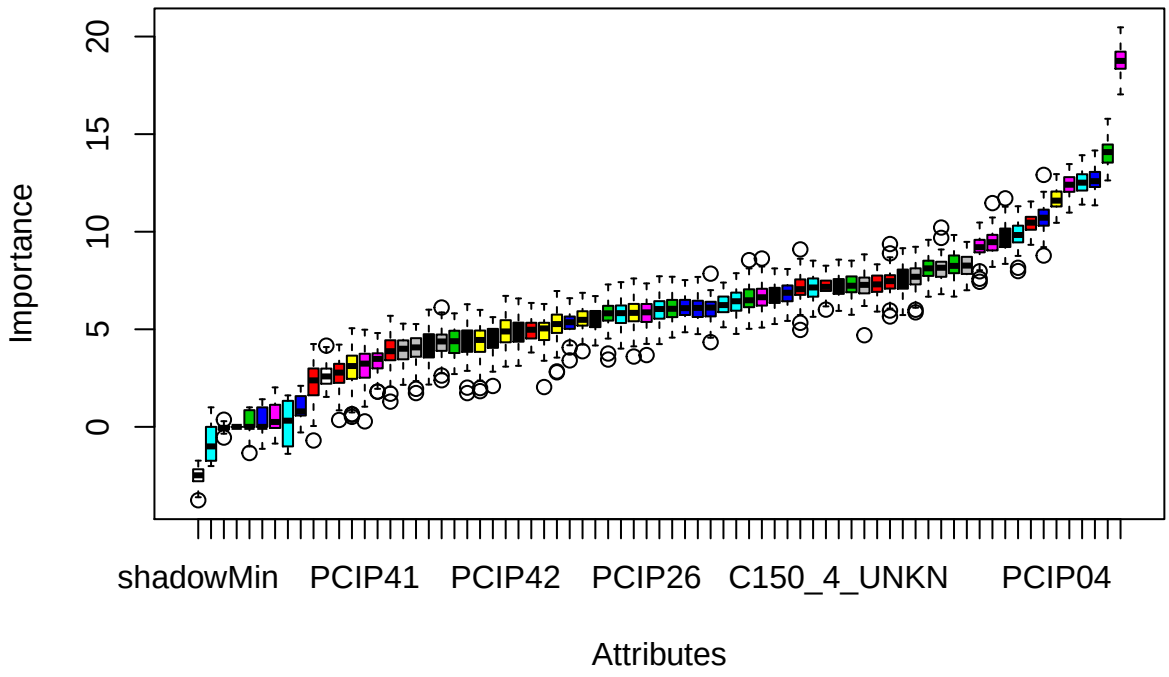
```
##                    meanImp  medianImp     minImp    maxImp  normHits
## REGION            5.5411953  5.4975196  4.1702870  6.702311 1.0000000
## ADM_RATE          7.3051074  7.3006426  5.9077095  8.328965 1.0000000
## ADM_RATE_ALL      7.2649879  7.2323929  5.7433352  8.518129 1.0000000
## SAT_AVG_ALL      12.6640120 12.5903155 11.3520121 14.161979 1.0000000
## PCIP01            6.2587815  6.2412386  5.1025928  7.394145 1.0000000
## PCIP03            6.6396503  6.6427899  5.0875318  8.626703 1.0000000
## PCIP04           11.6640658 11.5935032 10.4562788 12.951170 1.0000000
## PCIP05            8.2848424  8.2651669  7.0003142  9.479125 1.0000000
## PCIP09            4.8579386  4.9138689  3.1287016  6.580582 1.0000000
## PCIP10            2.7294977  2.7810113  0.3526242  4.214184 0.5858586
```

```
## PCIP11             6.6031990  6.4868313  5.0245742  8.543592 1.0000000
## PCIP12             0.9036968  0.7929059 -0.2843330  2.103476 0.0000000
## PCIP13             5.9868587  6.0385532  4.2401922  7.717525 1.0000000
## PCIP14            18.7707396 18.7532057 17.0361673 20.471370 1.0000000
## PCIP15             4.9152381  5.0456932  2.0411208  6.298222 0.9898990
## PCIP16             7.6457891  7.6912255  5.8677886  9.227438 1.0000000
## PCIP19             7.5179182  7.5454738  5.7298423  9.155925 1.0000000
## PCIP22             2.3339623  2.3824931 -0.6975178  4.246354 0.4141414
## PCIP23             8.3015679  8.2488921  6.6762439  9.833747 1.0000000
## PCIP24             6.0504761  6.0940834  4.7524544  7.684477 1.0000000
## PCIP25            -0.9382558 -1.0010015 -2.0075638  1.001002 0.0000000
## PCIP26             5.8226892  5.8690856  3.6785019  7.351316 1.0000000
## PCIP27             5.2127242  5.2684641  2.7965131  6.962251 1.0000000
## PCIP29             0.0000000  0.0000000  0.0000000  0.000000 0.0000000
## PCIP30             4.1857154  4.2365848  2.1685592  6.005528 0.9393939
## PCIP31             4.9989776  4.9776136  3.8115516  6.351910 1.0000000
## PCIP38             4.3448255  4.3892773  2.7061331  5.817189 0.9595960
## PCIP39             5.3396014  5.3620328  3.4059937  6.593988 1.0000000
## PCIP40             5.7980404  5.8280568  4.0112510  7.415592 1.0000000
## PCIP41             3.1910638  3.2397970  0.2816409  4.979572 0.7272727
## PCIP42             4.8814116  4.8894008  3.0869807  6.712171 0.9696970
## PCIP43             7.2508487  7.2771567  4.6952133  8.839186 1.0000000
## PCIP44             4.5252346  4.6263945  2.0948523  5.623487 0.9696970
## PCIP45             7.4619293  7.4660933  5.6556099  9.368665 1.0000000
## PCIP46             0.1022306  0.0000000 -1.3398513  1.001002 0.0000000
## PCIP47             0.2566136  0.0000000 -1.1257649  1.416832 0.0000000
## PCIP48             0.2116250  0.3170444 -1.3814144  1.612699 0.0000000
## PCIP49             3.3579129  3.4833980  1.8041762  4.809261 0.7474747
## PCIP50             5.8206652  5.8422201  3.6037018  7.604553 1.0000000
## PCIP51             3.9854026  4.0043244  2.1536258  5.289989 0.9191919
## PCIP52             9.7489714  9.8179765  8.3526025 11.711160 1.0000000
## PCIP54             3.8989724  3.8767975  1.2935035  5.692184 0.9292929
## UGDS_WHITE         8.1499779  8.1224943  6.6701766  9.585892 1.0000000
## UGDS_BLACK        10.7405537 10.7172410  8.7778332 12.911723 1.0000000
## UGDS_HISP          6.3890335  6.4371713  4.7625086  7.875194 1.0000000
## UGDS_ASIAN         9.2414199  9.2074002  7.4354700 10.469291 1.0000000
## UGDS_AIAN          4.3339945  4.4541683  1.8306193  5.995413 0.9494949
## UGDS_NHPI          3.9926305  4.0722658  1.7313182  5.268734 0.9191919
## UGDS_2MOR          4.3518092  4.4026115  1.7300445  6.284687 0.9292929
## UGDS_NRA           7.1087232  7.0558150  4.9675576  9.091865 1.0000000
## UGDS_UNKN          6.0863406  6.0510971  4.5739495  7.693875 1.0000000
## PPTUG_EF           6.8555662  6.8718372  5.4290769  8.090365 1.0000000
## COSTT4_A           9.8721817  9.8312883  7.9819382 11.305353 1.0000000
## TUITIONFEE_IN      9.4541022  9.4648532  8.1975936 11.466391 1.0000000
## TUITIONFEE_OUT     5.5289182  5.4857357  3.8760500  6.869499 1.0000000
## C150_4             8.0848696  8.1528886  6.7998644 10.213070 1.0000000
## C150_4_WHITE       6.7395075  6.7097316  5.2688363  8.119994 1.0000000
## C150_4_BLACK       7.2293737  7.1722255  6.0069691  8.258410 1.0000000
## C150_4_HISP        5.7829461  5.8203560  3.4499579  7.296527 1.0000000
## C150_4_ASIAN       6.1281138  6.0840814  4.8491531  7.523248 1.0000000
## C150_4_AIAN        7.1324135  7.1421341  5.6401391  8.519306 1.0000000
## C150_4_NHPI        0.4364940  0.2518653 -0.8548426  2.023303 0.0000000
## C150_4_2MOR        3.0251448  3.1169638  0.5143681  5.054858 0.6363636
## C150_4_NRA         4.3071019  4.3756831  2.3893522  6.116222 0.9797980
```

```
## C150_4_UNKN        7.1772133  7.2184048  5.9432228  8.570165 1.0000000
## RET_FT4           10.4612537 10.4633196  9.3288358 11.549600 1.0000000
## PCTFLOAN          14.0347227 14.0800947 12.6268176 15.786100 1.0000000
## PAR_ED_PCT_1STGEN  5.9962383  6.1115126  4.3402909  7.851685 1.0000000
## UGDS_MEN          12.5262334 12.5142443 11.3893612 13.919321 1.0000000
## UGDS_WOMEN        12.4041657 12.4069207 10.9796363 13.466496 1.0000000
##                    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        Tentative
## C150_4_NRA         Confirmed
## C150_4_UNKN        Confirmed
## RET_FT4            Confirmed
## PCTFLOAN           Confirmed
## PAR_ED_PCT_1STGEN Confirmed
## UGDS_MEN           Confirmed
## UGDS_WOMEN         Confirmed
```

```r
# 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)
```

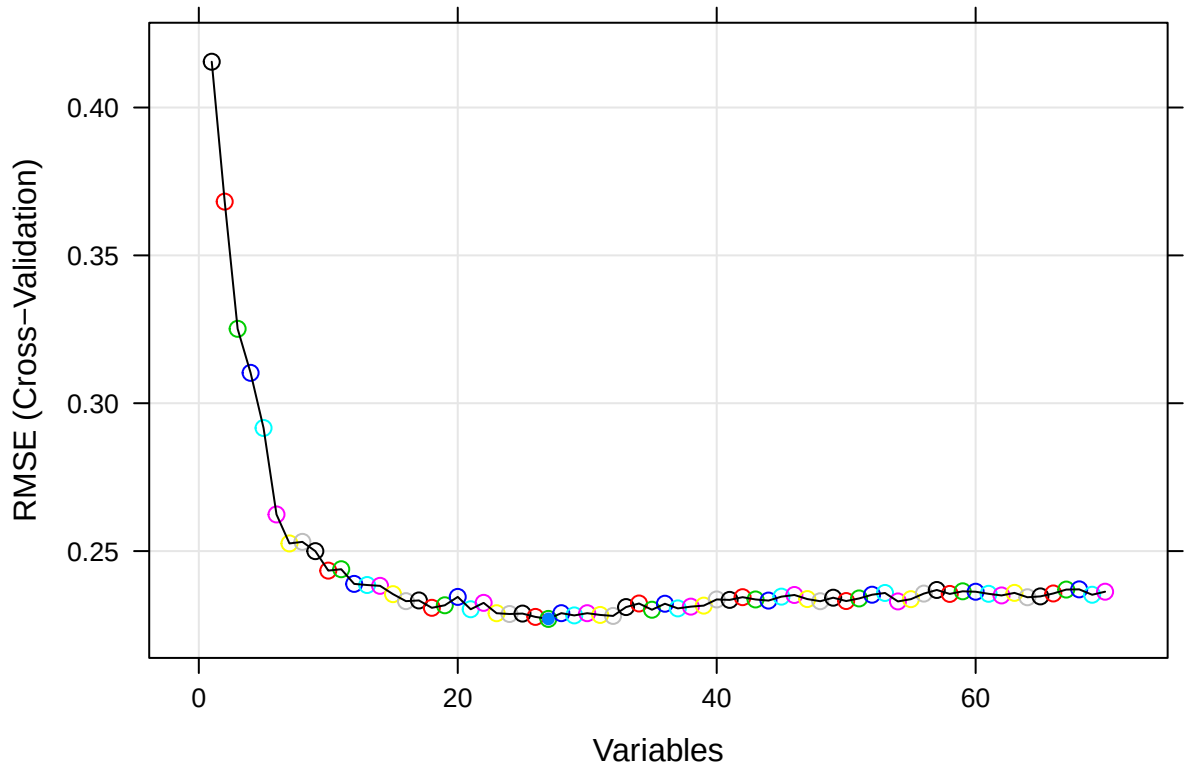
```
##
## Attaching package: 'plyr'
```

```
## 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] "TUITIONFEE_IN" "C150_4_AIAN"   "RET_FT4"       "UGDS_HISP"
## [17] "PCIP39"        "UGDS_ASIAN"    "PCIP16"        "UGDS_WHITE"
## [21] "PCIP19"        "UGDS_NRA"      "C150_4"        "PPTUG_EF"
## [25] "PCIP26"        "PCIP05"        "PCIP50"
```

```
# 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.

```
# 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.

```
# 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 = 50000
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
## -0.60482 -0.06064  0.01281  0.07491  0.51763
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.354e-01  4.472e-02  20.915  < 2e-16 ***
## REGION            -3.636e-03  3.293e-03  -1.104 0.269915
## ADM_RATE_ALL      -1.457e-01  3.833e-02  -3.801 0.000159 ***
## UGDS_NRA           2.169e-01  1.464e-01   1.482 0.138847
## PPTUG_EF          -3.280e-01  8.483e-02  -3.867 0.000122 ***
## COSTT4_A           1.672e-06  6.186e-07   2.702 0.007081 **
## PCTFLOAN          -4.127e-01  5.845e-02  -7.062 4.55e-12 ***
## PAR_ED_PCT_1STGEN -4.416e-02  9.948e-02  -0.444 0.657267
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1395 on 603 degrees of freedom
## Multiple R-squared:  0.4333, Adjusted R-squared:  0.4267
## F-statistic: 65.86 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
## 0.7892157
```

```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  0.29401850      0 1.0000000 1.0029639 0.06480981
## 2  0.06331002      1 0.7059815 0.7234215 0.05722964
## 3  0.02995005      2 0.6426715 0.6877923 0.05468780
## 4  0.02804232      3 0.6127214 0.6986102 0.05674321
## 5  0.02101282      4 0.5846791 0.7061201 0.05829297
## 6  0.01809964      5 0.5636663 0.6932180 0.05744853
## 7  0.01678080      7 0.5274670 0.6859515 0.05717466
## 8  0.01299550      8 0.5106862 0.6895656 0.05560876
## 9  0.01092446      9 0.4976907 0.6762075 0.05340993
## 10 0.01000000     10 0.4867662 0.6761857 0.05259241
##
## Variable importance
##      ADM_RATE_ALL          PPTUG_EF          COSTT4_A          PCTFLOAN
##                29                17                17                15
## PAR_ED_PCT_1STGEN          UGDS_NRA            REGION
##                11                10                 1
##
## Node number 1: 611 observations,    complexity param=0.2940185
##   mean=0.6611142, MSE=0.03387809
##   left son=2 (506 obs) right son=3 (105 obs)
##   Primary splits:
##       ADM_RATE_ALL < 0.3363178 to the right, improve=0.2940185, (0 missing)
##       COSTT4_A     < 51980     to the left,  improve=0.2470990, (0 missing)
##       PPTUG_EF     < 0.06485   to the right, improve=0.2290053, (0 missing)
##       UGDS_NRA     < 0.0577    to the left,  improve=0.2237433, (0 missing)
##       PCTFLOAN     < 0.42385   to the right, improve=0.2010692, (0 missing)
##   Surrogate splits:
##       PCTFLOAN          < 0.31645   to the right, agree=0.902, adj=0.429, (0 split)
##       COSTT4_A          < 55204.5   to the left,  agree=0.900, adj=0.419, (0 split)
##       PPTUG_EF          < 0.014     to the right, agree=0.897, adj=0.400, (0 split)
##       PAR_ED_PCT_1STGEN < 0.1801479 to the right, agree=0.885, adj=0.333, (0 split)
##
## Node number 2: 506 observations,    complexity param=0.06331002
##   mean=0.6156504, MSE=0.02728552
##   left son=4 (384 obs) right son=5 (122 obs)
##   Primary splits:
##       UGDS_NRA          < 0.0579    to the left,  improve=0.09491828, (0 missing)
##       PPTUG_EF          < 0.08135   to the right, improve=0.09199408, (0 missing)
##       PCTFLOAN          < 0.50685   to the right, improve=0.08236296, (0 missing)
##       PAR_ED_PCT_1STGEN < 0.347892  to the right, improve=0.06248149, (0 missing)
##       COSTT4_A          < 25725.5   to the left,  improve=0.05463281, (0 missing)
##   Surrogate splits:
##       COSTT4_A          < 49140.5   to the left,  agree=0.826, adj=0.279, (0 split)
```

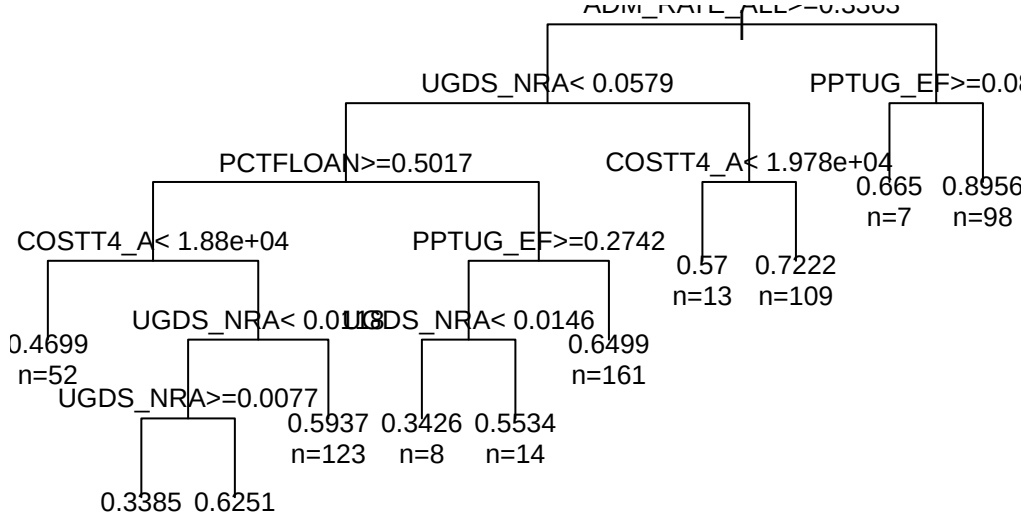
```
##       ADM_RATE_ALL      < 0.4165    to the right, agree=0.792, adj=0.139, (0 split)
##       PPTUG_EF          < 0.05175   to the right, agree=0.792, adj=0.139, (0 split)
##       PAR_ED_PCT_1STGEN < 0.1743518 to the right, agree=0.773, adj=0.057, (0 split)
##
## Node number 3: 105 observations,    complexity param=0.0167808
##   mean=0.8802067, MSE=0.00768572
##   left son=6 (7 obs) right son=7 (98 obs)
##   Primary splits:
##       PPTUG_EF          < 0.0896    to the right, improve=0.4304263, (0 missing)
##       COSTT4_A          < 23711.5   to the left,  improve=0.3874953, (0 missing)
##       PAR_ED_PCT_1STGEN < 0.3446017 to the right, improve=0.3717428, (0 missing)
##       ADM_RATE_ALL      < 0.2721    to the right, improve=0.3364179, (0 missing)
##       UGDS_NRA          < 0.04595   to the left,  improve=0.2743237, (0 missing)
##   Surrogate splits:
##       COSTT4_A          < 20751.5   to the left,  agree=0.962, adj=0.429, (0 split)
##       PAR_ED_PCT_1STGEN < 0.3843536 to the right, agree=0.952, adj=0.286, (0 split)
##       UGDS_NRA          < 0.03465   to the left,  agree=0.943, adj=0.143, (0 split)
##
## Node number 4: 384 observations,    complexity param=0.02995005
##   mean=0.5869654, MSE=0.02851568
##   left son=8 (201 obs) right son=9 (183 obs)
##   Primary splits:
##       PCTFLOAN          < 0.50175   to the right, improve=0.05661646, (0 missing)
##       PPTUG_EF          < 0.06485   to the right, improve=0.05176600, (0 missing)
##       PAR_ED_PCT_1STGEN < 0.347892  to the right, improve=0.03572350, (0 missing)
##       UGDS_NRA          < 0.01185   to the left,  improve=0.02988927, (0 missing)
##       COSTT4_A          < 19063     to the left,  improve=0.02646810, (0 missing)
##   Surrogate splits:
##       REGION            < 4.5       to the left,  agree=0.633, adj=0.230, (0 split)
##       ADM_RATE_ALL      < 0.5964643 to the right, agree=0.617, adj=0.197, (0 split)
##       PAR_ED_PCT_1STGEN < 0.2998108 to the right, agree=0.612, adj=0.186, (0 split)
##       PPTUG_EF          < 0.0782    to the right, agree=0.581, adj=0.120, (0 split)
##       COSTT4_A          < 17240     to the right, agree=0.560, adj=0.077, (0 split)
##
## Node number 5: 122 observations,    complexity param=0.0129955
##   mean=0.7059377, MSE=0.01267183
##   left son=10 (13 obs) right son=11 (109 obs)
##   Primary splits:
##       COSTT4_A          < 19778     to the left,  improve=0.17400190, (0 missing)
##       PPTUG_EF          < 0.1849    to the right, improve=0.17098310, (0 missing)
##       ADM_RATE_ALL      < 0.6030879 to the right, improve=0.14510170, (0 missing)
##       PAR_ED_PCT_1STGEN < 0.2089352 to the right, improve=0.09457956, (0 missing)
##       PCTFLOAN          < 0.571     to the right, improve=0.09271540, (0 missing)
##   Surrogate splits:
##       PPTUG_EF     < 0.23325   to the right, agree=0.951, adj=0.538, (0 split)
##       ADM_RATE_ALL < 0.80155   to the right, agree=0.918, adj=0.231, (0 split)
##
## Node number 6: 7 observations
##   mean=0.665, MSE=0.02443381
##
## Node number 7: 98 observations
##   mean=0.8955786, MSE=0.002944996
##
## Node number 8: 201 observations,    complexity param=0.02101282
```

```
##   mean=0.5486264, MSE=0.03068981
##   left son=16 (52 obs) right son=17 (149 obs)
##   Primary splits:
##       COSTT4_A     < 18804     to the left,  improve=0.07051058, (0 missing)
##       UGDS_NRA     < 0.0118    to the left,  improve=0.06755005, (0 missing)
##       PPTUG_EF     < 0.0654    to the right, improve=0.03835758, (0 missing)
##       ADM_RATE_ALL < 0.662844  to the right, improve=0.03101977, (0 missing)
##       REGION       < 3.5       to the right, improve=0.02656306, (0 missing)
##   Surrogate splits:
##       ADM_RATE_ALL < 0.9163613 to the right, agree=0.766, adj=0.096, (0 split)
##       PCTFLOAN     < 0.7501    to the right, agree=0.756, adj=0.058, (0 split)
##
## Node number 9: 183 observations,    complexity param=0.02804232
##   mean=0.6290754, MSE=0.02273999
##   left son=18 (22 obs) right son=19 (161 obs)
##   Primary splits:
##       PPTUG_EF          < 0.27425   to the right, improve=0.13948660, (0 missing)
##       COSTT4_A          < 25306.5   to the left,  improve=0.05445775, (0 missing)
##       PAR_ED_PCT_1STGEN < 0.3425707 to the right, improve=0.05129478, (0 missing)
##       UGDS_NRA          < 0.00665   to the left,  improve=0.05064729, (0 missing)
##       PCTFLOAN          < 0.4236    to the right, improve=0.03529774, (0 missing)
##   Surrogate splits:
##       ADM_RATE_ALL < 0.9851702 to the right, agree=0.891, adj=0.091, (0 split)
##
## Node number 10: 13 observations
##   mean=0.5699692, MSE=0.009797479
##
## Node number 11: 109 observations
##   mean=0.7221541, MSE=0.01054675
##
## Node number 16: 52 observations
##   mean=0.4698827, MSE=0.02574005
##
## Node number 17: 149 observations,    complexity param=0.01809964
##   mean=0.5761074, MSE=0.02949808
##   left son=34 (26 obs) right son=35 (123 obs)
##   Primary splits:
##       UGDS_NRA          < 0.0118    to the left,  improve=0.04967755, (0 missing)
##       PCTFLOAN          < 0.6998    to the right, improve=0.03344117, (0 missing)
##       PAR_ED_PCT_1STGEN < 0.3871971 to the right, improve=0.03083482, (0 missing)
##       ADM_RATE_ALL      < 0.681     to the right, improve=0.02847557, (0 missing)
##       PPTUG_EF          < 0.0813    to the right, improve=0.01938223, (0 missing)
##   Surrogate splits:
##       PCTFLOAN < 0.671     to the right, agree=0.852, adj=0.154, (0 split)
##
## Node number 18: 22 observations,    complexity param=0.01092446
##   mean=0.4767182, MSE=0.03119188
##   left son=36 (8 obs) right son=37 (14 obs)
##   Primary splits:
##       UGDS_NRA          < 0.0146    to the left,  improve=0.3295306, (0 missing)
##       PAR_ED_PCT_1STGEN < 0.3822091 to the left,  improve=0.3024247, (0 missing)
##       PPTUG_EF          < 0.3452    to the left,  improve=0.1897162, (0 missing)
##       COSTT4_A          < 18345     to the left,  improve=0.1516320, (0 missing)
##       PCTFLOAN          < 0.4521    to the right, improve=0.1323283, (0 missing)
```

```
##   Surrogate splits:
##       ADM_RATE_ALL      < 0.5006937 to the left,  agree=0.864, adj=0.625, (0 split)
##       PAR_ED_PCT_1STGEN < 0.3767416 to the left,  agree=0.818, adj=0.500, (0 split)
##       PPTUG_EF          < 0.31555   to the left,  agree=0.773, adj=0.375, (0 split)
##
## Node number 19: 161 observations
##   mean=0.6498944, MSE=0.01797972
##
## Node number 34: 26 observations,    complexity param=0.01809964
##   mean=0.4928462, MSE=0.0456708
##   left son=68 (12 obs) right son=69 (14 obs)
##   Primary splits:
##       UGDS_NRA     < 0.0077    to the right, improve=0.44714990, (0 missing)
##       COSTT4_A     < 19786     to the right, improve=0.18094420, (0 missing)
##       PCTFLOAN     < 0.60195   to the right, improve=0.16297720, (0 missing)
##       ADM_RATE_ALL < 0.7566865 to the right, improve=0.07691589, (0 missing)
##       REGION       < 5.5       to the left,  improve=0.05741993, (0 missing)
##   Surrogate splits:
##       PPTUG_EF          < 0.11415   to the right, agree=0.692, adj=0.333, (0 split)
##       COSTT4_A          < 19923.5   to the right, agree=0.692, adj=0.333, (0 split)
##       ADM_RATE_ALL      < 0.7566865 to the right, agree=0.654, adj=0.250, (0 split)
##       PCTFLOAN          < 0.60625   to the right, agree=0.654, adj=0.250, (0 split)
##       PAR_ED_PCT_1STGEN < 0.3162031 to the right, agree=0.654, adj=0.250, (0 split)
##
## Node number 35: 123 observations
##   mean=0.5937073, MSE=0.0243043
##
## Node number 36: 8 observations
##   mean=0.3426, MSE=0.04993862
##
## Node number 37: 14 observations
##   mean=0.5533571, MSE=0.004327248
##
## Node number 68: 12 observations
##   mean=0.3384917, MSE=0.02289754
##
## Node number 69: 14 observations
##   mean=0.62515, MSE=0.02726474
```

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

Single Decision Tree of
US Research University Completion Rate Prediction Model




```
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)
```

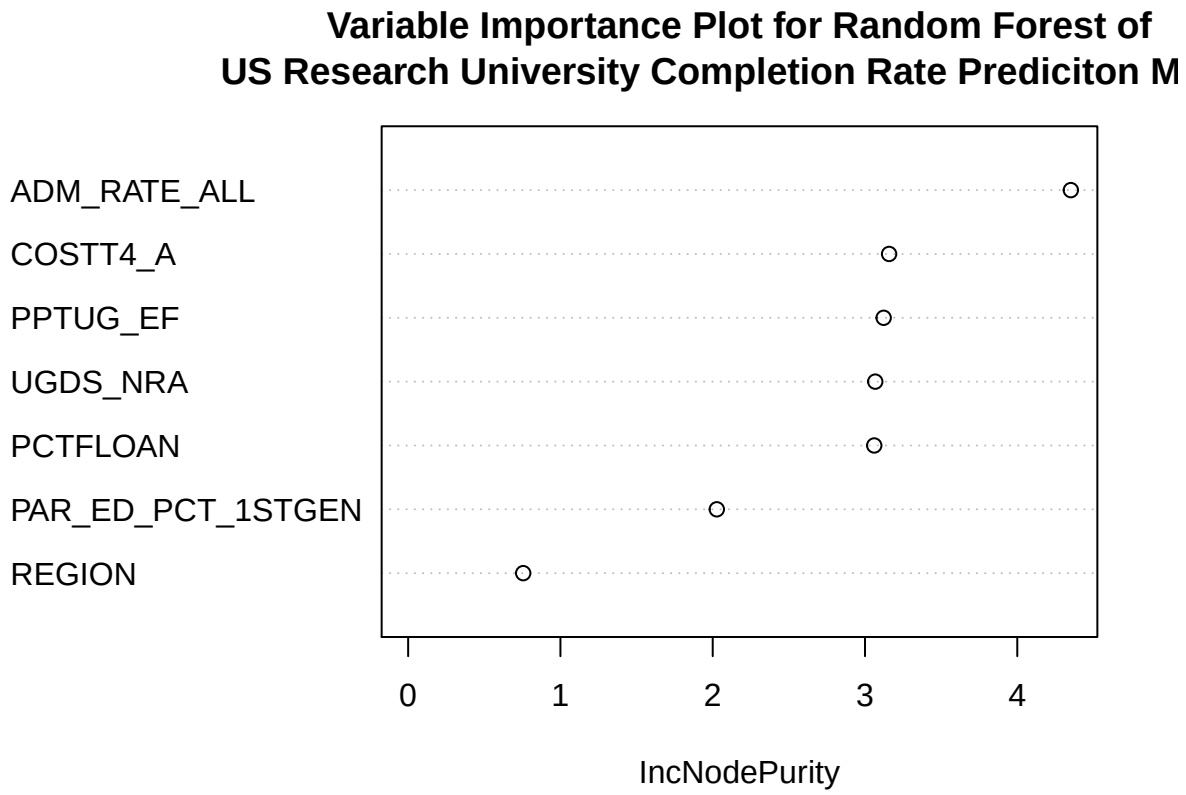
```
## [1] 0.9117647
```

```
# 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
```

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




```
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] 0.9215686
```

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

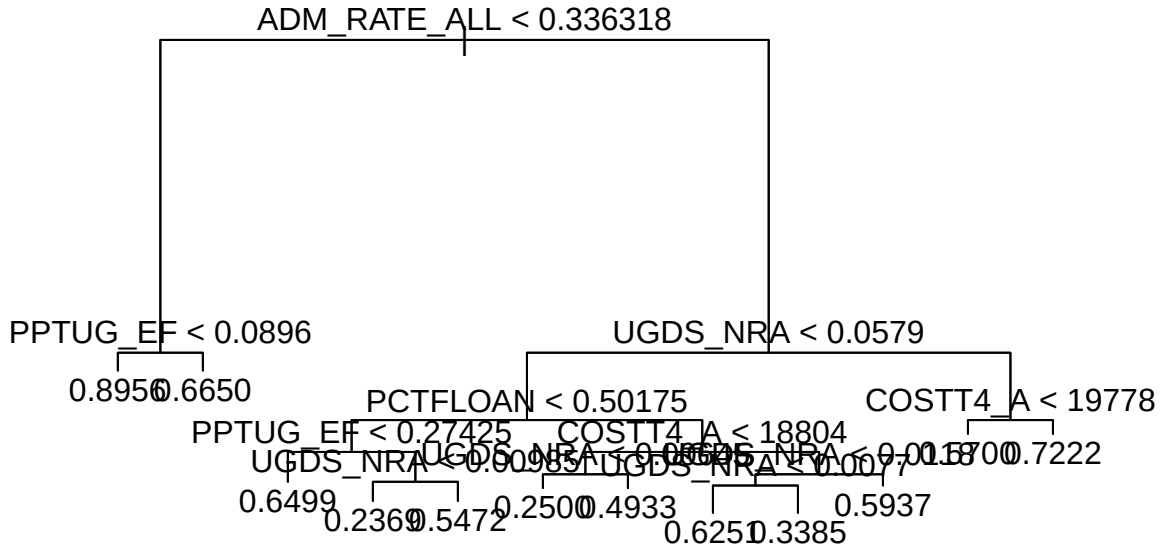
```
##
## Call:
## svm(formula = formula_completionrate, data = univ_train2)
##
##
```

```
## Parameters:
##    SVM-Type:  eps-regression
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  0.1428571
##     epsilon:  0.1
##
##
## Number of Support Vectors:  513
```

```
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] 0.9117647
```

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




```
pred_tree3 <- predict(model_tree3, newdata = univ_test2)
accu14 <- abs(pred_tree3 - univ_test2$C150_4_NRA) < 0.25
```

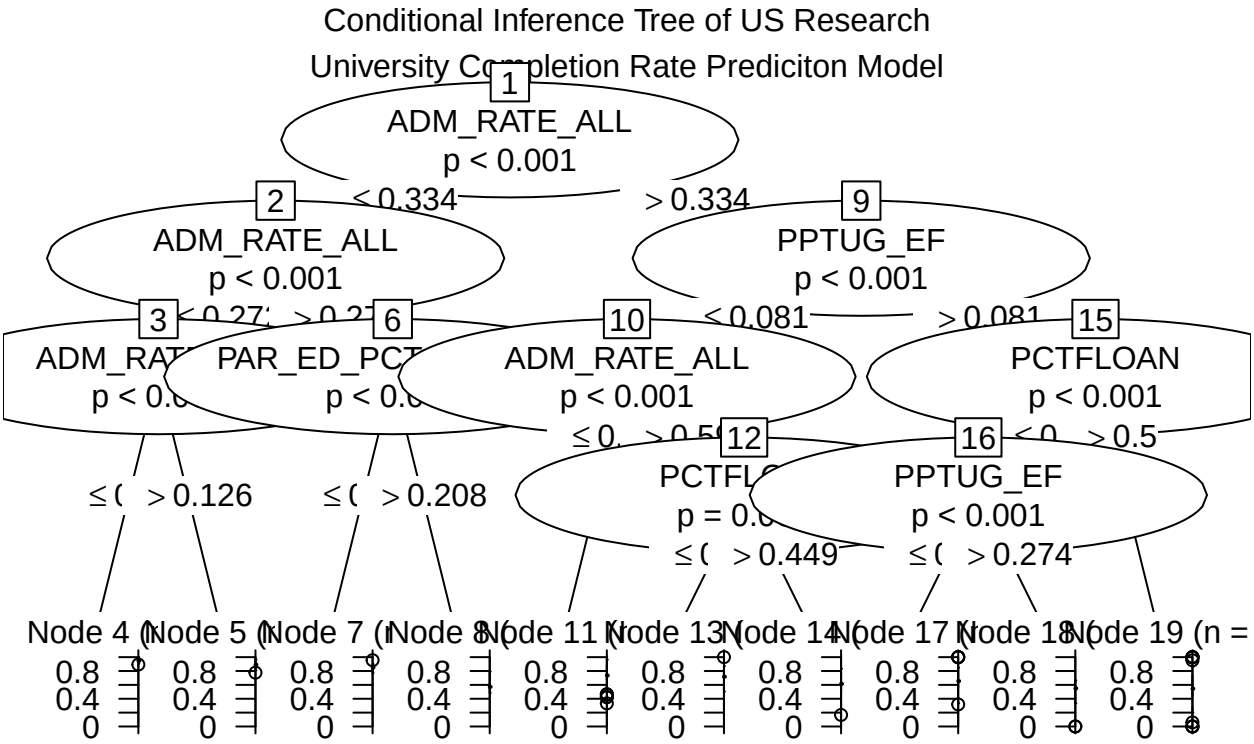
```
frac14 <- sum(accu14)/length(accu14)
print(frac14)
```

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

```
# 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] 0.8970588
```

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