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 classification for the rest  
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 our focus will be on these  
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, 7-Rocky Mountains, 8-Far West, 9-Outlying Areas)  
# 37-38 - admission rate  
# 39-61 - SAT and ACT Scores  
# 62-99 - percentage of degrees awarded for each field of study  
# 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 expected time to completion)  
# 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  
  
## 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

#show number of rows of filtered research universities  
rows\_usresearchuniv <- nrow(usresearchuniv)  
rows\_usresearchuniv

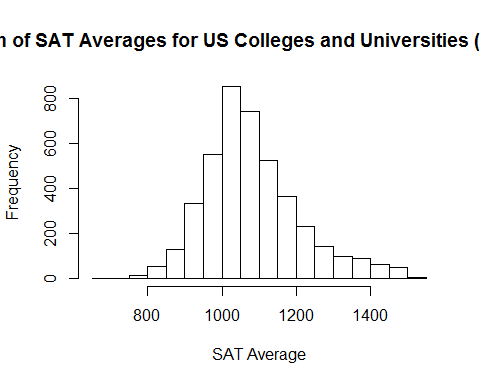
## [1] 815

#percentage of data from filtered research universities to unfiltered  
rows\_usresearchuniv / rows\_usunivaccepted

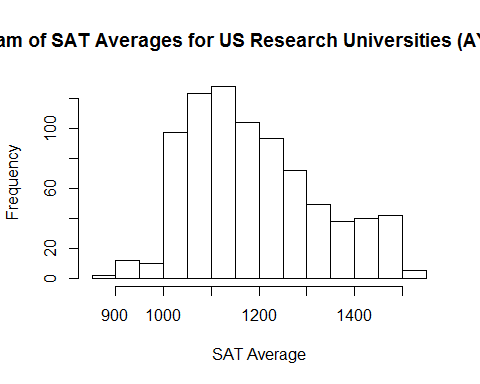
## [1] 0.7062392

# Distributions and Box and Whisker Plots

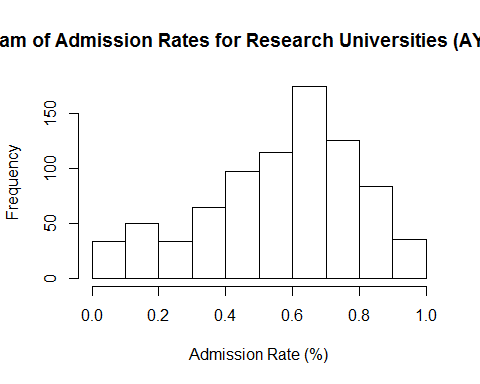
# Histogram of SAT Averages for US Colleges and Universities  
hist(usunivfilter$SAT\_AVG\_ALL, main = "Histogram of SAT Averages for US Colleges and Universities (AY2010-2015)", xlab="SAT Average")



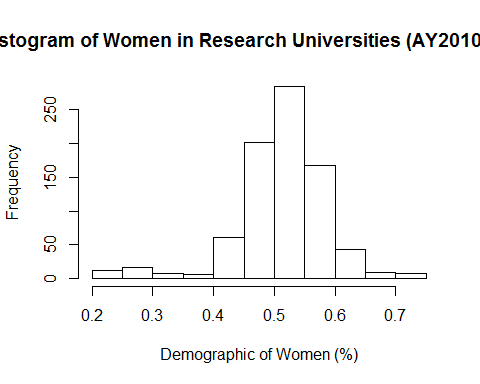
# Histogram of SAT Averages for US Research Universities  
hist(usresearchuniv$SAT\_AVG\_ALL, main = "Histogram of SAT Averages for US Research Universities (AY2010-2015)", xlab="SAT Average")



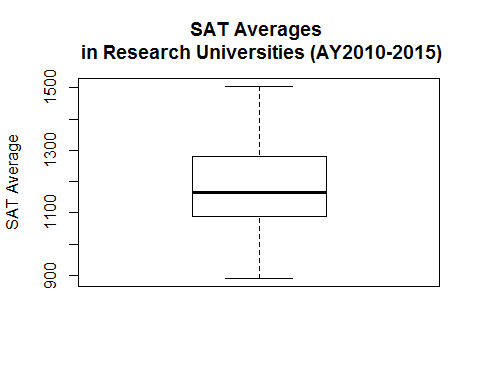
# Histogram of Admission Rates for US Research Universities  
hist(usresearchuniv$ADM\_RATE\_ALL, main = "Histogram of Admission Rates for Research Universities (AY2010-2015)", xlab = "Admission Rate (%)")



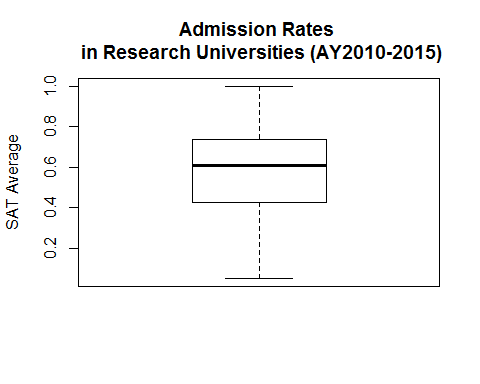
# Histogram of Women in US Research Universities  
hist(usresearchuniv$UGDS\_WOMEN, main = "Histogram of Women in Research Universities (AY2010-2015)", xlab = "Demographic of Women (%)")



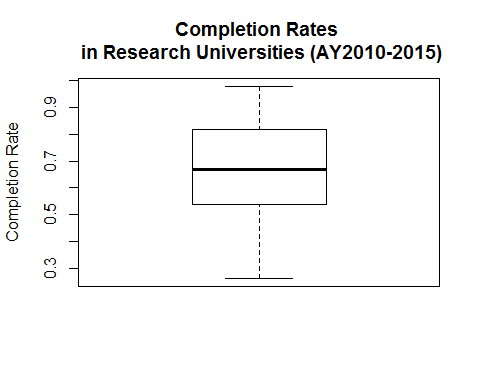
#Boxplot of SAT Average in all US Research Universities  
boxplot(usresearchuniv$SAT\_AVG\_ALL, main = "SAT Averages \n in Research Universities (AY2010-2015)", ylab = "SAT Average")



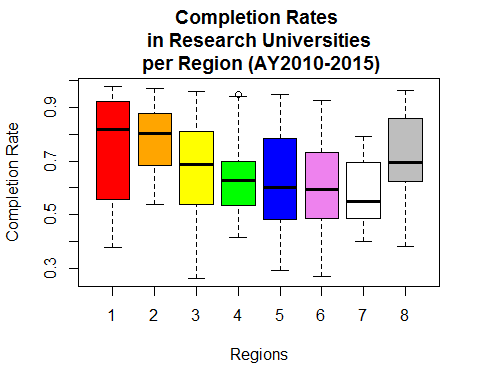
#Boxplot of admission rates in all US Research Universities  
boxplot(usresearchuniv$ADM\_RATE\_ALL, main = "Admission Rates \n in Research Universities (AY2010-2015)", ylab = "SAT Average")



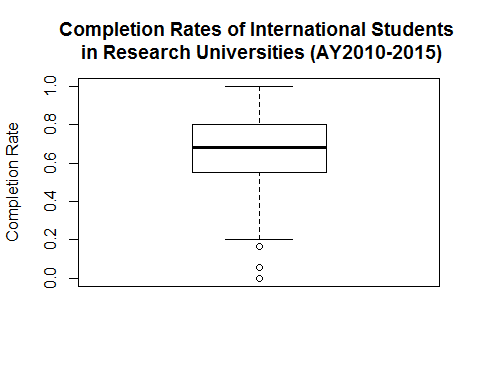
#Boxplot of Completion Rates in all US Research Universities  
boxplot(usresearchuniv$C150\_4, main = "Completion Rates \n in Research Universities (AY2010-2015)", ylab = "Completion Rate")



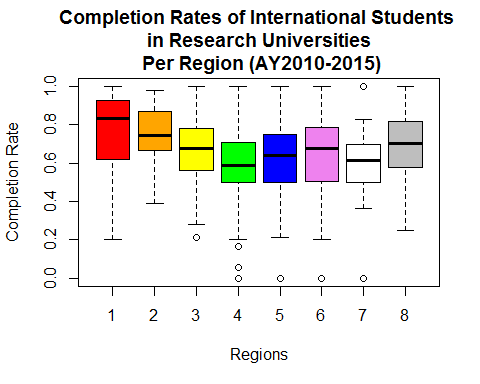
# Boxplot of Completion Rates per Region in US Research Universities  
boxplot(C150\_4 ~ REGION, usresearchuniv, main = "Completion Rates \n in Research Universities \n per Region (AY2010-2015)", col=c("red", "orange", "yellow", "green", "blue", "violet", "white", "gray", "magenta"), ylab = "Completion Rate", xlab = "Regions")



#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 Universities (AY2010-2015)", ylab = "Completion Rate")



# 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 Research Universities \n Per Region (AY2010-2015)", col=c("red", "orange", "yellow", "green", "blue", "violet", "white", "gray", "magenta"), ylab = "Completion Rate", xlab = "Regions")

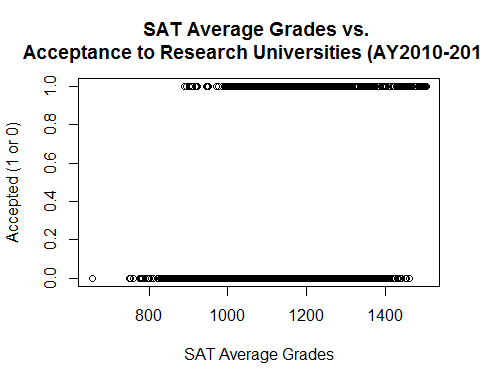


nrow(usresearchuniv[usresearchuniv$C150\_4\_NRA < 0.2,])

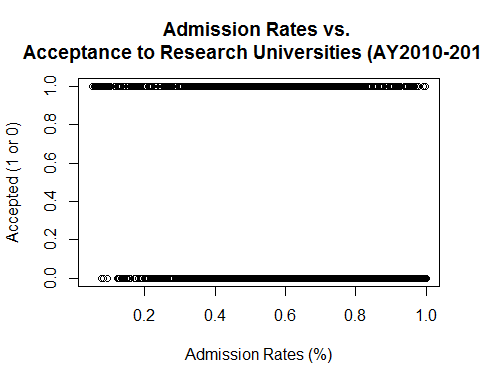
## [1] 9

# Correlations

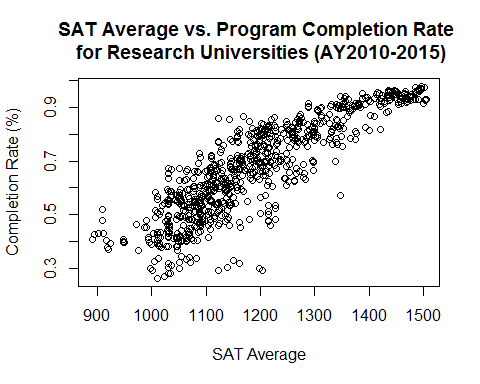
#Correlation between the SAT grades and the acceptance for the research universities  
plot(usunivfilter$SAT\_AVG\_ALL, usunivfilter$ACCEPTED, main="SAT Average Grades vs. \n Acceptance to Research Universities (AY2010-2015)", xlab="SAT Average Grades", ylab="Accepted (1 or 0)")



#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 Research Universities (AY2010-2015)", xlab="Admission Rates (%)", ylab="Accepted (1 or 0)")



#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 \n for Research Universities (AY2010-2015)", xlab="SAT Average", ylab="Completion Rate (%)")

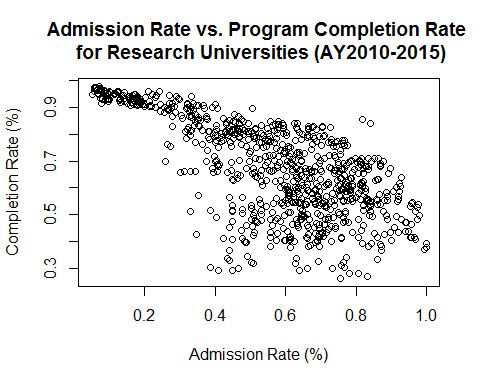


#Correlation coefficient between admission rate and completion rate  
cor(usresearchuniv$SAT\_AVG\_ALL, usresearchuniv$C150\_4, method = "pearson")

## [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 Rate \n for Research Universities (AY2010-2015)", xlab="Admission Rate (%)", ylab="Completion Rate (%)")

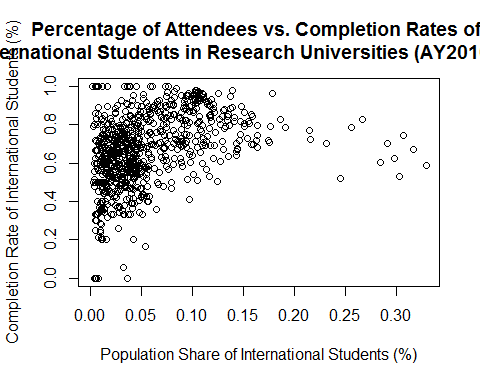


#Correlation coefficient between admission rate and completion rate  
cor(usresearchuniv$ADM\_RATE\_ALL, usresearchuniv$C150\_4, method = "pearson")

## [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 Rates of \n International Students in Research Universities (AY2010-2015)", xlab="Population Share of International Students (%)", ylab="Completion Rate of International Students (%)")

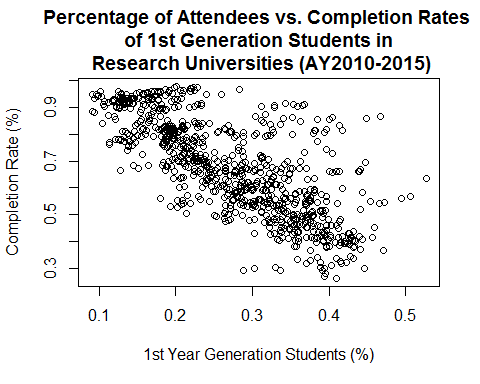


#Correlation coefficient between admission rate and completion rate of international students  
cor(usresearchuniv$UGDS\_NRA, usresearchuniv$C150\_4\_NRA, method = "pearson")

## [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. Completion Rates \n of 1st Generation Students in \n Research Universities (AY2010-2015)", xlab="1st Year Generation Students (%)", ylab="Completion Rate (%)")



#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 based on SAT, average cost, loans, and gender  
formula\_ISAcceptance <- formula(ACCEPTED ~ REGION + ADM\_RATE\_ALL + SAT\_AVG\_ALL + UGDS\_NRA + COSTT4\_A + PCTFLOAN + UGDS\_WOMEN)

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

where

.

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 = 20000, PCTFLOAN = 0.33, UGDS\_WOMEN = .37)  
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 = 25600, PCTFLOAN = 0.57, UGDS\_WOMEN = .55)  
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   
## -2.3543 -0.5475 -0.2959 -0.1118 2.7911   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.454e+01 1.187e+00 -12.249 < 2e-16 \*\*\*  
## REGION 1.572e-01 2.927e-02 5.369 7.91e-08 \*\*\*  
## ADM\_RATE\_ALL 5.969e-01 3.804e-01 1.569 0.116576   
## SAT\_AVG\_ALL 1.430e-02 8.386e-04 17.051 < 2e-16 \*\*\*  
## UGDS\_NRA 7.053e+00 1.328e+00 5.313 1.08e-07 \*\*\*  
## COSTT4\_A -9.023e-05 6.141e-06 -14.693 < 2e-16 \*\*\*  
## PCTFLOAN -9.162e-01 4.885e-01 -1.876 0.060722 .   
## UGDS\_WOMEN -1.794e+00 5.316e-01 -3.375 0.000738 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3173.7 on 3184 degrees of freedom  
## Residual deviance: 2162.1 on 3177 degrees of freedom  
## AIC: 2178.1  
##   
## 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  
## 0 848 102  
## 1 31 81

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

## Accuracy   
## 0.8747646

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

## Neg Pred Value   
## 0.7232143

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

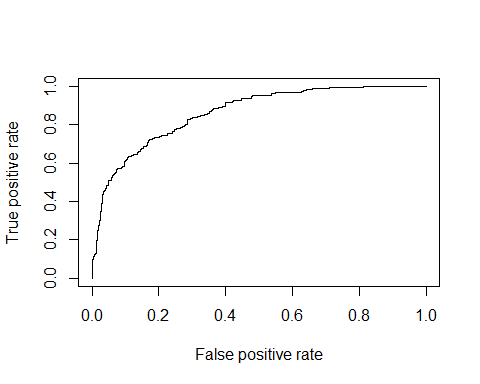
## Specificity   
## 0.442623

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.

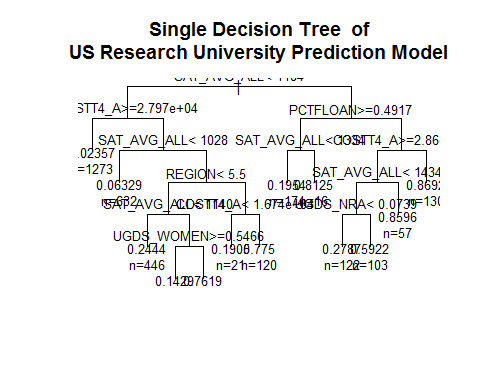
# show the curve on the performance  
perf1 <- performance(pred1, "tpr", "fpr")  
plot(perf1, lty = 1)



# 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 0.13770031 0 1.0000000 1.0009123 0.02682135  
## 2 0.05898713 1 0.8622997 0.8789266 0.02628150  
## 3 0.03636944 3 0.7443254 0.7723719 0.02112462  
## 4 0.03320921 4 0.7079560 0.7678260 0.02229006  
## 5 0.02258599 5 0.6747468 0.7127724 0.02324032  
## 6 0.01717753 6 0.6521608 0.7065240 0.02349796  
## 7 0.01595619 7 0.6349833 0.6784661 0.02383295  
## 8 0.01466381 8 0.6190271 0.6741263 0.02379266  
## 9 0.01205389 9 0.6043633 0.6666893 0.02407650  
## 10 0.01101451 10 0.5923094 0.6586076 0.02448253  
## 11 0.01083813 11 0.5812949 0.6521363 0.02466704  
## 12 0.01000000 12 0.5704567 0.6533475 0.02487291  
##   
## Variable importance  
## SAT\_AVG\_ALL COSTT4\_A PCTFLOAN ADM\_RATE\_ALL UGDS\_WOMEN   
## 36 22 14 9 7   
## REGION UGDS\_NRA   
## 6 6   
##   
## Node number 1: 3185 observations, complexity param=0.1377003  
## mean=0.1984301, MSE=0.1590556   
## left son=2 (2583 obs) right son=3 (602 obs)  
## Primary splits:  
## SAT\_AVG\_ALL < 1183.5 to the left, improve=0.13770030, (0 missing)  
## PCTFLOAN < 0.4936 to the right, improve=0.13201790, (0 missing)  
## UGDS\_WOMEN < 0.55345 to the right, improve=0.08875239, (0 missing)  
## ADM\_RATE\_ALL < 0.3569599 to the right, improve=0.04958056, (0 missing)  
## COSTT4\_A < 27966.5 to the right, improve=0.04363998, (0 missing)  
## Surrogate splits:  
## COSTT4\_A < 51993 to the left, agree=0.881, adj=0.369, (0 split)  
## ADM\_RATE\_ALL < 0.3347178 to the right, agree=0.857, adj=0.246, (0 split)  
## PCTFLOAN < 0.37295 to the right, agree=0.838, adj=0.141, (0 split)  
## UGDS\_NRA < 0.37375 to the left, agree=0.811, adj=0.002, (0 split)  
##   
## Node number 2: 2583 observations, complexity param=0.05898713  
## mean=0.1269841, MSE=0.1108592   
## left son=4 (1273 obs) right son=5 (1310 obs)  
## Primary splits:  
## COSTT4\_A < 27966.5 to the right, improve=0.09375095, (0 missing)  
## PCTFLOAN < 0.6235 to the right, improve=0.07743903, (0 missing)  
## SAT\_AVG\_ALL < 1028.5 to the left, improve=0.05486219, (0 missing)  
## UGDS\_WOMEN < 0.56125 to the right, improve=0.05437147, (0 missing)  
## REGION < 5.5 to the left, improve=0.02433478, (0 missing)  
## Surrogate splits:  
## PCTFLOAN < 0.60705 to the right, agree=0.701, adj=0.393, (0 split)  
## UGDS\_WOMEN < 0.61435 to the right, agree=0.601, adj=0.191, (0 split)  
## REGION < 4.5 to the left, agree=0.576, adj=0.139, (0 split)  
## SAT\_AVG\_ALL < 1032.5 to the right, agree=0.569, adj=0.126, (0 split)  
## UGDS\_NRA < 0.04565 to the right, agree=0.564, adj=0.116, (0 split)  
##   
## Node number 3: 602 observations, complexity param=0.03636944  
## mean=0.5049834, MSE=0.2499752   
## left son=6 (190 obs) right son=7 (412 obs)  
## Primary splits:  
## PCTFLOAN < 0.4917 to the right, improve=0.12243390, (0 missing)  
## COSTT4\_A < 33961 to the right, improve=0.11452710, (0 missing)  
## UGDS\_WOMEN < 0.51265 to the right, improve=0.09162632, (0 missing)  
## UGDS\_NRA < 0.0928 to the left, improve=0.05606864, (0 missing)  
## SAT\_AVG\_ALL < 1434.5 to the left, improve=0.05262837, (0 missing)  
## Surrogate splits:  
## ADM\_RATE\_ALL < 0.6029592 to the right, agree=0.733, adj=0.153, (0 split)  
## UGDS\_WOMEN < 0.2988 to the left, agree=0.726, adj=0.132, (0 split)  
## SAT\_AVG\_ALL < 1197.5 to the left, agree=0.708, adj=0.074, (0 split)  
## UGDS\_NRA < 0.0058 to the left, agree=0.699, adj=0.047, (0 split)  
##   
## Node number 4: 1273 observations  
## mean=0.02356638, MSE=0.023011   
##   
## Node number 5: 1310 observations, complexity param=0.05898713  
## mean=0.2274809, MSE=0.1757333   
## left son=10 (632 obs) right son=11 (678 obs)  
## Primary splits:  
## SAT\_AVG\_ALL < 1028.5 to the left, improve=0.14299650, (0 missing)  
## UGDS\_WOMEN < 0.56125 to the right, improve=0.08192822, (0 missing)  
## COSTT4\_A < 17976 to the left, improve=0.04673986, (0 missing)  
## UGDS\_NRA < 0.02045 to the left, improve=0.04432220, (0 missing)  
## PCTFLOAN < 0.62935 to the right, improve=0.04271969, (0 missing)  
## Surrogate splits:  
## UGDS\_WOMEN < 0.5317 to the right, agree=0.634, adj=0.242, (0 split)  
## COSTT4\_A < 18137.5 to the left, agree=0.601, adj=0.172, (0 split)  
## PCTFLOAN < 0.6755 to the right, agree=0.598, adj=0.166, (0 split)  
## UGDS\_NRA < 0.01805 to the left, agree=0.582, adj=0.133, (0 split)  
## REGION < 1.5 to the left, agree=0.544, adj=0.054, (0 split)  
##   
## Node number 6: 190 observations, complexity param=0.01101451  
## mean=0.2473684, MSE=0.1861773   
## left son=12 (174 obs) right son=13 (16 obs)  
## Primary splits:  
## SAT\_AVG\_ALL < 1334 to the left, improve=0.15774050, (0 missing)  
## UGDS\_WOMEN < 0.52285 to the right, improve=0.15413160, (0 missing)  
## ADM\_RATE\_ALL < 0.3576099 to the right, improve=0.11638180, (0 missing)  
## UGDS\_NRA < 0.02775 to the left, improve=0.09474699, (0 missing)  
## COSTT4\_A < 56352 to the left, improve=0.07857724, (0 missing)  
## Surrogate splits:  
## ADM\_RATE\_ALL < 0.39645 to the right, agree=0.947, adj=0.375, (0 split)  
## COSTT4\_A < 57158.5 to the left, agree=0.926, adj=0.125, (0 split)  
## UGDS\_NRA < 0.1124 to the left, agree=0.921, adj=0.063, (0 split)  
##   
## Node number 7: 412 observations, complexity param=0.02258599  
## mean=0.6237864, MSE=0.2346769   
## left son=14 (282 obs) right son=15 (130 obs)  
## Primary splits:  
## COSTT4\_A < 28598.5 to the right, improve=0.11833960, (0 missing)  
## UGDS\_WOMEN < 0.6111 to the right, improve=0.07135974, (0 missing)  
## ADM\_RATE\_ALL < 0.1285159 to the right, improve=0.04027451, (0 missing)  
## SAT\_AVG\_ALL < 1444 to the left, improve=0.03982250, (0 missing)  
## REGION < 7.5 to the right, improve=0.03574585, (0 missing)  
## Surrogate splits:  
## UGDS\_NRA < 0.0156 to the right, agree=0.772, adj=0.277, (0 split)  
## SAT\_AVG\_ALL < 1215.5 to the right, agree=0.760, adj=0.238, (0 split)  
## ADM\_RATE\_ALL < 0.4401319 to the left, agree=0.745, adj=0.192, (0 split)  
## UGDS\_WOMEN < 0.47165 to the right, agree=0.721, adj=0.115, (0 split)  
## PCTFLOAN < 0.4584 to the left, agree=0.701, adj=0.054, (0 split)  
##   
## Node number 10: 632 observations  
## mean=0.06329114, MSE=0.05928537   
##   
## Node number 11: 678 observations, complexity param=0.03320921  
## mean=0.380531, MSE=0.2357272   
## left son=22 (537 obs) right son=23 (141 obs)  
## Primary splits:  
## REGION < 5.5 to the left, improve=0.10526350, (0 missing)  
## UGDS\_WOMEN < 0.56775 to the right, improve=0.07798958, (0 missing)  
## COSTT4\_A < 16957 to the left, improve=0.06017817, (0 missing)  
## PCTFLOAN < 0.6226 to the right, improve=0.05935574, (0 missing)  
## UGDS\_NRA < 0.00945 to the left, improve=0.05094373, (0 missing)  
## Surrogate splits:  
## ADM\_RATE\_ALL < 0.95345 to the left, agree=0.796, adj=0.021, (0 split)  
## UGDS\_NRA < 0.26405 to the left, agree=0.796, adj=0.021, (0 split)  
##   
## Node number 12: 174 observations  
## mean=0.1954023, MSE=0.1572202   
##   
## Node number 13: 16 observations  
## mean=0.8125, MSE=0.1523438   
##   
## Node number 14: 282 observations, complexity param=0.01717753  
## mean=0.5106383, MSE=0.2498868   
## left son=28 (225 obs) right son=29 (57 obs)  
## Primary splits:  
## SAT\_AVG\_ALL < 1434.5 to the left, improve=0.12348860, (0 missing)  
## UGDS\_NRA < 0.0755 to the left, improve=0.11644730, (0 missing)  
## ADM\_RATE\_ALL < 0.1285159 to the right, improve=0.10593390, (0 missing)  
## UGDS\_WOMEN < 0.51105 to the right, improve=0.10267260, (0 missing)  
## PCTFLOAN < 0.0965 to the right, improve=0.03159341, (0 missing)  
## Surrogate splits:  
## ADM\_RATE\_ALL < 0.1555403 to the right, agree=0.908, adj=0.544, (0 split)  
## PCTFLOAN < 0.1701 to the right, agree=0.855, adj=0.281, (0 split)  
## COSTT4\_A < 61468 to the left, agree=0.816, adj=0.088, (0 split)  
##   
## Node number 15: 130 observations  
## mean=0.8692308, MSE=0.1136686   
##   
## Node number 22: 537 observations, complexity param=0.01595619  
## mean=0.2998138, MSE=0.2099255   
## left son=44 (446 obs) right son=45 (91 obs)  
## Primary splits:  
## SAT\_AVG\_ALL < 1140.5 to the left, improve=0.07170480, (0 missing)  
## UGDS\_WOMEN < 0.56775 to the right, improve=0.06610869, (0 missing)  
## REGION < 2.5 to the left, improve=0.03870125, (0 missing)  
## PCTFLOAN < 0.6226 to the right, improve=0.03787905, (0 missing)  
## COSTT4\_A < 16957 to the left, improve=0.03724255, (0 missing)  
## Surrogate splits:  
## UGDS\_NRA < 0.08035 to the left, agree=0.838, adj=0.044, (0 split)  
##   
## Node number 23: 141 observations, complexity param=0.01205389  
## mean=0.6879433, MSE=0.2146773   
## left son=46 (21 obs) right son=47 (120 obs)  
## Primary splits:  
## COSTT4\_A < 16740.5 to the left, improve=0.20173470, (0 missing)  
## UGDS\_NRA < 0.01085 to the left, improve=0.14846960, (0 missing)  
## SAT\_AVG\_ALL < 1074.5 to the left, improve=0.12913140, (0 missing)  
## ADM\_RATE\_ALL < 0.5745 to the left, improve=0.08404510, (0 missing)  
## UGDS\_WOMEN < 0.53915 to the right, improve=0.07321931, (0 missing)  
## Surrogate splits:  
## PCTFLOAN < 0.338 to the left, agree=0.901, adj=0.333, (0 split)  
## ADM\_RATE\_ALL < 0.4493 to the left, agree=0.872, adj=0.143, (0 split)  
## UGDS\_NRA < 0.2473 to the right, agree=0.872, adj=0.143, (0 split)  
##   
## Node number 28: 225 observations, complexity param=0.01083813  
## mean=0.4222222, MSE=0.2439506   
## left son=56 (122 obs) right son=57 (103 obs)  
## Primary splits:  
## UGDS\_NRA < 0.0739 to the left, improve=0.10002960, (0 missing)  
## PCTFLOAN < 0.2408 to the left, improve=0.07485830, (0 missing)  
## UGDS\_WOMEN < 0.6111 to the right, improve=0.05594406, (0 missing)  
## COSTT4\_A < 52525 to the right, improve=0.03051730, (0 missing)  
## REGION < 5.5 to the left, improve=0.02039915, (0 missing)  
## Surrogate splits:  
## ADM\_RATE\_ALL < 0.4188876 to the right, agree=0.671, adj=0.282, (0 split)  
## SAT\_AVG\_ALL < 1360.5 to the left, agree=0.658, adj=0.252, (0 split)  
## COSTT4\_A < 55702.5 to the left, agree=0.631, adj=0.194, (0 split)  
## REGION < 4.5 to the right, agree=0.613, adj=0.155, (0 split)  
## UGDS\_WOMEN < 0.51265 to the right, agree=0.609, adj=0.146, (0 split)  
##   
## Node number 29: 57 observations  
## mean=0.8596491, MSE=0.1206525   
##   
## Node number 44: 446 observations  
## mean=0.2443946, MSE=0.1846659   
##   
## Node number 45: 91 observations, complexity param=0.01466381  
## mean=0.5714286, MSE=0.244898   
## left son=90 (28 obs) right son=91 (63 obs)  
## Primary splits:  
## UGDS\_WOMEN < 0.5466 to the right, improve=0.3333333, (0 missing)  
## UGDS\_NRA < 0.0091 to the left, improve=0.2256250, (0 missing)  
## PCTFLOAN < 0.49615 to the right, improve=0.1865209, (0 missing)  
## ADM\_RATE\_ALL < 0.6362281 to the left, improve=0.1463415, (0 missing)  
## COSTT4\_A < 18189 to the left, improve=0.1363636, (0 missing)  
## Surrogate splits:  
## COSTT4\_A < 18189 to the left, agree=0.802, adj=0.357, (0 split)  
## UGDS\_NRA < 0.0071 to the left, agree=0.780, adj=0.286, (0 split)  
## PCTFLOAN < 0.5805 to the right, agree=0.747, adj=0.179, (0 split)  
## ADM\_RATE\_ALL < 0.4243039 to the left, agree=0.736, adj=0.143, (0 split)  
##   
## Node number 46: 21 observations  
## mean=0.1904762, MSE=0.154195   
##   
## Node number 47: 120 observations  
## mean=0.775, MSE=0.174375   
##   
## Node number 56: 122 observations  
## mean=0.2786885, MSE=0.2010212   
##   
## Node number 57: 103 observations  
## mean=0.592233, MSE=0.2414931   
##   
## Node number 90: 28 observations  
## mean=0.1428571, MSE=0.122449   
##   
## Node number 91: 63 observations  
## mean=0.7619048, MSE=0.1814059

plot(model\_dtree1, uniform = TRUE, main = "Single Decision Tree of\nUS Research University Prediction Model")  
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] 0.8719397

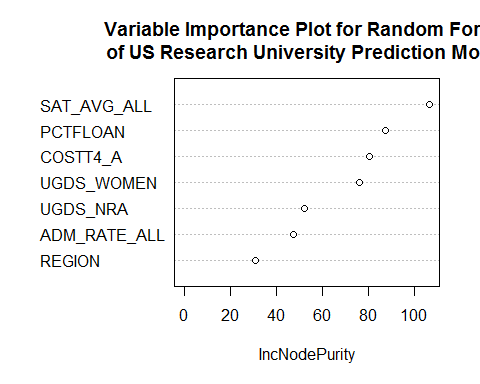
# doing random forest  
model\_forest1 <- randomForest(formula\_ISAcceptance, data = univ\_train)

## 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 Prediction Model")



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] 0.9397363

# 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: 1425

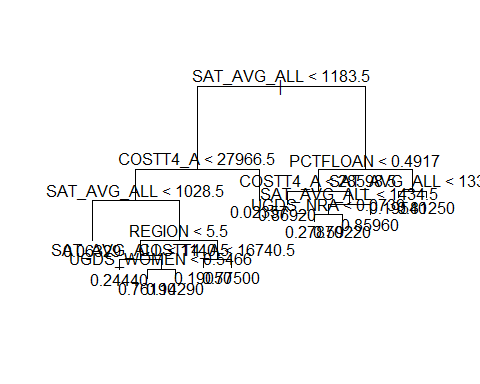
pred\_svm1 <- predict(model\_svm1, newdata = univ\_test)  
accu3 <- abs(pred\_svm1 - univ\_test$ACCEPTED) < 0.5  
frac3 <- sum(accu3)/length(accu3)  
print(frac3)

## [1] 0.8888889

# doing simple tree  
model\_tree1 <- tree(formula\_ISAcceptance, data = univ\_train)  
summary(model\_tree1)

##   
## Regression tree:  
## tree(formula = formula\_ISAcceptance, data = univ\_train)  
## Variables actually used in tree construction:  
## [1] "SAT\_AVG\_ALL" "COSTT4\_A" "REGION" "UGDS\_WOMEN" "PCTFLOAN"   
## [6] "UGDS\_NRA"   
## Number of terminal nodes: 13   
## Residual mean deviance: 0.09111 = 289 / 3172   
## Distribution of residuals:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -0.86920 -0.06329 -0.02357 0.00000 -0.02357 0.97640

plot(model\_tree1, main = "Simple Tree of\nUS Research University Prediction Model")  
text(model\_tree1)



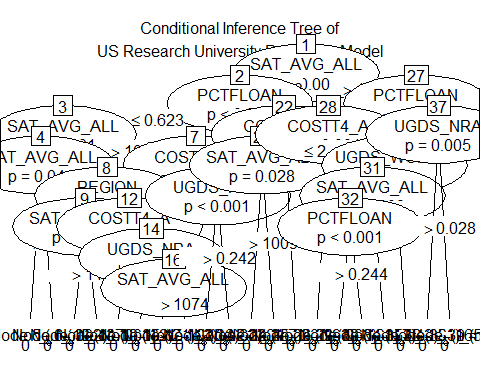
pred\_tree1 <- predict(model\_tree1, newdata = univ\_test)  
accu4 <- abs(pred\_tree1 - univ\_test$ACCEPTED) < 0.5  
frac4 <- sum(accu4)/length(accu4)  
print(frac4)

## [1] 0.8719397

# doing conditional inference tree  
model\_party1 <- ctree(formula\_ISAcceptance, data = univ\_train)  
summary(model\_party1)

## Length Class Mode   
## 1 BinaryTree S4

plot(model\_party1, main = "Conditional Inference Tree of\nUS Research University Prediction Model")



pred\_party1 <- predict(model\_party1, newdata = univ\_test)  
accu5 <- abs(pred\_party1 - univ\_test$ACCEPTED) < 0.5  
frac5 <- sum(accu5)/length(accu5)  
print(frac5)

## [1] 0.8785311

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 Science degrees/majors  
formula\_ISSciAcceptance <- formula(ACCEPTED ~ REGION + ADM\_RATE\_ALL + SAT\_AVG\_ALL + PCIP11 + PCIP12 + PCIP14 + PCIP15 + PCIP24 + PCIP26 + PCIP27 + PCIP40 + PCIP45 + PCIP51 + PCIP52 + UGDS\_NRA + UGDS\_UNKN + COSTT4\_A + PCTFLOAN + UGDS\_WOMEN)  
  
# 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   
## -2.55503 -0.48076 -0.24297 -0.07046 3.08040   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.730e+01 1.436e+00 -12.047 < 2e-16 \*\*\*  
## REGION 1.693e-01 3.180e-02 5.324 1.02e-07 \*\*\*  
## ADM\_RATE\_ALL 8.668e-01 4.214e-01 2.057 0.03971 \*   
## SAT\_AVG\_ALL 1.543e-02 1.020e-03 15.116 < 2e-16 \*\*\*  
## PCIP11 -3.169e-01 2.021e+00 -0.157 0.87536   
## PCIP12 -5.108e+00 2.029e+01 -0.252 0.80120   
## PCIP14 5.395e+00 7.819e-01 6.900 5.20e-12 \*\*\*  
## PCIP15 -3.133e-01 2.417e+00 -0.130 0.89685   
## PCIP24 -6.533e+00 1.293e+00 -5.051 4.39e-07 \*\*\*  
## PCIP26 7.393e+00 1.757e+00 4.207 2.58e-05 \*\*\*  
## PCIP27 -3.232e+01 7.064e+00 -4.575 4.77e-06 \*\*\*  
## PCIP40 -3.401e+01 4.914e+00 -6.921 4.48e-12 \*\*\*  
## PCIP45 7.644e+00 1.210e+00 6.316 2.68e-10 \*\*\*  
## PCIP51 2.011e+00 6.177e-01 3.255 0.00113 \*\*   
## PCIP52 4.257e-01 6.736e-01 0.632 0.52743   
## UGDS\_NRA 9.534e+00 1.496e+00 6.374 1.84e-10 \*\*\*  
## UGDS\_UNKN -2.036e+00 1.636e+00 -1.244 0.21339   
## COSTT4\_A -1.078e-04 7.171e-06 -15.040 < 2e-16 \*\*\*  
## PCTFLOAN -8.620e-01 5.624e-01 -1.533 0.12534   
## UGDS\_WOMEN 6.285e-01 8.105e-01 0.775 0.43807   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3173.7 on 3184 degrees of freedom  
## Residual deviance: 1908.7 on 3165 degrees of freedom  
## AIC: 1948.7  
##   
## 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  
## 0 843 86  
## 1 36 97

c2$overall['Accuracy']

## Accuracy   
## 0.8851224

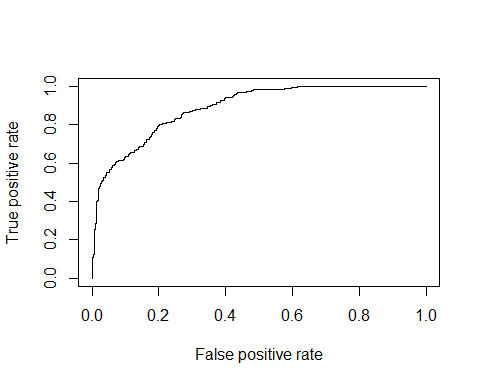
#Precision of the logistic regression model  
c2$byClass['Neg Pred Value']

## Neg Pred Value   
## 0.7293233

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

## Specificity   
## 0.5300546

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



# Now we check on what acceptable ways we could do for regression  
# doing single decision tree  
model\_dtree2 <- rpart(formula\_ISSciAcceptance, method="anova",data = univ\_train)  
summary(model\_dtree2)

## Call:  
## rpart(formula = formula\_ISSciAcceptance, data = univ\_train, method = "anova")  
## n= 3185   
##   
## CP nsplit rel error xerror xstd  
## 1 0.33035437 0 1.0000000 1.0009005 0.02682139  
## 2 0.07005047 1 0.6696456 0.6743438 0.02196207  
## 3 0.04076907 2 0.5995952 0.6149427 0.02381785  
## 4 0.02639615 3 0.5588261 0.5826456 0.02425302  
## 5 0.01670652 4 0.5324299 0.5713389 0.02469309  
## 6 0.01568867 5 0.5157234 0.5606989 0.02462141  
## 7 0.01471420 6 0.5000347 0.5590773 0.02480022  
## 8 0.01275077 7 0.4853205 0.5560829 0.02545241  
## 9 0.01246383 10 0.4470682 0.5447612 0.02549885  
## 10 0.01000000 11 0.4346044 0.5196665 0.02517616  
##   
## Variable importance  
## PCIP14 SAT\_AVG\_ALL PCTFLOAN UGDS\_WOMEN PCIP45   
## 35 14 8 8 7   
## ADM\_RATE\_ALL PCIP26 COSTT4\_A UGDS\_NRA UGDS\_UNKN   
## 5 5 4 3 2   
## REGION PCIP52 PCIP40 PCIP27 PCIP15   
## 2 2 2 1 1   
## PCIP51 PCIP11   
## 1 1   
##   
## Node number 1: 3185 observations, complexity param=0.3303544  
## mean=0.1984301, MSE=0.1590556   
## left son=2 (2286 obs) right son=3 (899 obs)  
## Primary splits:  
## PCIP14 < 0.0269 to the left, improve=0.33035440, (0 missing)  
## SAT\_AVG\_ALL < 1183.5 to the left, improve=0.13770030, (0 missing)  
## PCTFLOAN < 0.4936 to the right, improve=0.13201790, (0 missing)  
## UGDS\_WOMEN < 0.55345 to the right, improve=0.08875239, (0 missing)  
## PCIP45 < 0.04825 to the left, improve=0.06992954, (0 missing)  
## Surrogate splits:  
## UGDS\_WOMEN < 0.51585 to the right, agree=0.778, adj=0.214, (0 split)  
## SAT\_AVG\_ALL < 1182.5 to the left, agree=0.747, adj=0.103, (0 split)  
## ADM\_RATE\_ALL < 0.1685543 to the right, agree=0.729, adj=0.039, (0 split)  
## PCTFLOAN < 0.38235 to the right, agree=0.726, adj=0.028, (0 split)  
## PCIP11 < 0.1152 to the left, agree=0.724, adj=0.021, (0 split)  
##   
## Node number 2: 2286 observations, complexity param=0.01275077  
## mean=0.05468066, MSE=0.05169069   
## left son=4 (2049 obs) right son=5 (237 obs)  
## Primary splits:  
## SAT\_AVG\_ALL < 1194.5 to the left, improve=0.04089789, (0 missing)  
## PCIP45 < 0.3399 to the left, improve=0.03823503, (0 missing)  
## PCTFLOAN < 0.62345 to the right, improve=0.03506829, (0 missing)  
## COSTT4\_A < 59882.5 to the left, improve=0.02332492, (0 missing)  
## PCIP14 < 0.00415 to the left, improve=0.02131276, (0 missing)  
## Surrogate splits:  
## PCIP45 < 0.21745 to the left, agree=0.935, adj=0.376, (0 split)  
## COSTT4\_A < 53312.5 to the left, agree=0.935, adj=0.371, (0 split)  
## ADM\_RATE\_ALL < 0.2968292 to the right, agree=0.920, adj=0.224, (0 split)  
## PCIP40 < 0.0546 to the left, agree=0.906, adj=0.097, (0 split)  
## PCIP52 < 0.0034 to the right, agree=0.906, adj=0.093, (0 split)  
##   
## Node number 3: 899 observations, complexity param=0.07005047  
## mean=0.56396, MSE=0.2459091   
## left son=6 (415 obs) right son=7 (484 obs)  
## Primary splits:  
## PCTFLOAN < 0.51625 to the right, improve=0.16052220, (0 missing)  
## PCIP45 < 0.0317 to the left, improve=0.14435210, (0 missing)  
## PCIP26 < 0.02435 to the left, improve=0.10418290, (0 missing)  
## SAT\_AVG\_ALL < 1120.5 to the left, improve=0.09747897, (0 missing)  
## PCIP40 < 0.00695 to the left, improve=0.08437388, (0 missing)  
## Surrogate splits:  
## PCIP45 < 0.03885 to the left, agree=0.703, adj=0.357, (0 split)  
## SAT\_AVG\_ALL < 1121.5 to the left, agree=0.697, adj=0.345, (0 split)  
## ADM\_RATE\_ALL < 0.6110151 to the right, agree=0.689, adj=0.325, (0 split)  
## REGION < 4.5 to the left, agree=0.647, adj=0.236, (0 split)  
## PCIP26 < 0.04755 to the left, agree=0.643, adj=0.227, (0 split)  
##   
## Node number 4: 2049 observations  
## mean=0.03904344, MSE=0.03751905   
##   
## Node number 5: 237 observations, complexity param=0.01275077  
## mean=0.1898734, MSE=0.1538215   
## left son=10 (154 obs) right son=11 (83 obs)  
## Primary splits:  
## PCIP52 < 0.13035 to the left, improve=0.18828560, (0 missing)  
## PCIP51 < 0.0362 to the left, improve=0.15723430, (0 missing)  
## PCIP40 < 0.0278 to the right, improve=0.10925960, (0 missing)  
## PCIP45 < 0.3399 to the left, improve=0.09138216, (0 missing)  
## PCIP14 < 0.0021 to the left, improve=0.08455997, (0 missing)  
## Surrogate splits:  
## PCIP51 < 0.01485 to the left, agree=0.806, adj=0.446, (0 split)  
## PCIP45 < 0.1779 to the right, agree=0.797, adj=0.422, (0 split)  
## PCIP40 < 0.02165 to the right, agree=0.764, adj=0.325, (0 split)  
## SAT\_AVG\_ALL < 1253.5 to the right, agree=0.743, adj=0.265, (0 split)  
## PCIP27 < 0.01325 to the right, agree=0.743, adj=0.265, (0 split)  
##   
## Node number 6: 415 observations, complexity param=0.02639615  
## mean=0.3493976, MSE=0.2273189   
## left son=12 (201 obs) right son=13 (214 obs)  
## Primary splits:  
## COSTT4\_A < 27256.5 to the right, improve=0.14174750, (0 missing)  
## PCIP45 < 0.0317 to the left, improve=0.13234170, (0 missing)  
## PCTFLOAN < 0.67375 to the right, improve=0.10063070, (0 missing)  
## PCIP26 < 0.01525 to the left, improve=0.08377179, (0 missing)  
## PCIP40 < 0.00505 to the left, improve=0.06536961, (0 missing)  
## Surrogate splits:  
## SAT\_AVG\_ALL < 1109.5 to the right, agree=0.687, adj=0.353, (0 split)  
## UGDS\_NRA < 0.05535 to the right, agree=0.667, adj=0.313, (0 split)  
## UGDS\_UNKN < 0.0384 to the right, agree=0.646, adj=0.269, (0 split)  
## PCIP15 < 0.00535 to the left, agree=0.641, adj=0.259, (0 split)  
## PCIP26 < 0.07135 to the right, agree=0.619, adj=0.214, (0 split)  
##   
## Node number 7: 484 observations, complexity param=0.04076907  
## mean=0.7479339, MSE=0.1885288   
## left son=14 (64 obs) right son=15 (420 obs)  
## Primary splits:  
## SAT\_AVG\_ALL < 1066 to the left, improve=0.22634260, (0 missing)  
## PCIP26 < 0.02705 to the left, improve=0.10567940, (0 missing)  
## PCIP24 < 0.01125 to the right, improve=0.09710353, (0 missing)  
## UGDS\_UNKN < 0.00035 to the left, improve=0.08838507, (0 missing)  
## PCIP40 < 0.00695 to the left, improve=0.07166574, (0 missing)  
## Surrogate splits:  
## PCIP26 < 0.02535 to the left, agree=0.880, adj=0.094, (0 split)  
## ADM\_RATE\_ALL < 0.98265 to the right, agree=0.874, adj=0.047, (0 split)  
## PCIP12 < 0.005 to the right, agree=0.874, adj=0.047, (0 split)  
## PCIP51 < 0.2316 to the right, agree=0.874, adj=0.047, (0 split)  
## COSTT4\_A < 11984.5 to the left, agree=0.874, adj=0.047, (0 split)  
##   
## Node number 10: 154 observations  
## mean=0.06493506, MSE=0.0607185   
##   
## Node number 11: 83 observations, complexity param=0.01275077  
## mean=0.4216867, MSE=0.243867   
## left son=22 (60 obs) right son=23 (23 obs)  
## Primary splits:  
## PCIP45 < 0.156 to the left, improve=0.3795048, (0 missing)  
## PCTFLOAN < 0.41885 to the right, improve=0.3265804, (0 missing)  
## SAT\_AVG\_ALL < 1330 to the left, improve=0.3025210, (0 missing)  
## PCIP14 < 6e-04 to the left, improve=0.1607641, (0 missing)  
## PCIP24 < 0.02695 to the left, improve=0.1488283, (0 missing)  
## Surrogate splits:  
## SAT\_AVG\_ALL < 1250.5 to the left, agree=0.916, adj=0.696, (0 split)  
## ADM\_RATE\_ALL < 0.363659 to the right, agree=0.843, adj=0.435, (0 split)  
## COSTT4\_A < 52679 to the left, agree=0.795, adj=0.261, (0 split)  
## PCIP24 < 0.07665 to the left, agree=0.771, adj=0.174, (0 split)  
## PCIP52 < 0.13335 to the right, agree=0.747, adj=0.087, (0 split)  
##   
## Node number 12: 201 observations, complexity param=0.0147142  
## mean=0.1641791, MSE=0.1372243   
## left son=24 (173 obs) right son=25 (28 obs)  
## Primary splits:  
## PCIP45 < 0.10325 to the left, improve=0.2702513, (0 missing)  
## COSTT4\_A < 51693.5 to the left, improve=0.2494925, (0 missing)  
## ADM\_RATE\_ALL < 0.5764284 to the right, improve=0.2002029, (0 missing)  
## PCIP26 < 0.12095 to the left, improve=0.1900521, (0 missing)  
## PCTFLOAN < 0.5842 to the right, improve=0.1672563, (0 missing)  
## Surrogate splits:  
## PCIP26 < 0.1206 to the left, agree=0.905, adj=0.321, (0 split)  
## PCIP27 < 0.0431 to the left, agree=0.876, adj=0.107, (0 split)  
## COSTT4\_A < 57292 to the left, agree=0.871, adj=0.071, (0 split)  
## ADM\_RATE\_ALL < 0.3535498 to the right, agree=0.866, adj=0.036, (0 split)  
## SAT\_AVG\_ALL < 1350.5 to the left, agree=0.866, adj=0.036, (0 split)  
##   
## Node number 13: 214 observations, complexity param=0.01670652  
## mean=0.5233645, MSE=0.2494541   
## left son=26 (27 obs) right son=27 (187 obs)  
## Primary splits:  
## PCIP26 < 0.014 to the left, improve=0.1585404, (0 missing)  
## PCIP45 < 0.0151 to the left, improve=0.1433680, (0 missing)  
## UGDS\_NRA < 0.01335 to the left, improve=0.1410271, (0 missing)  
## PCIP40 < 0.00195 to the left, improve=0.1368955, (0 missing)  
## PCIP27 < 0.0021 to the left, improve=0.1174942, (0 missing)  
## Surrogate splits:  
## PCIP40 < 0.00195 to the left, agree=0.944, adj=0.556, (0 split)  
## PCIP45 < 0.0082 to the left, agree=0.925, adj=0.407, (0 split)  
## PCIP27 < 0.00045 to the left, agree=0.921, adj=0.370, (0 split)  
## UGDS\_WOMEN < 0.22525 to the left, agree=0.897, adj=0.185, (0 split)  
## PCIP15 < 0.092 to the right, agree=0.893, adj=0.148, (0 split)  
##   
## Node number 14: 64 observations  
## mean=0.21875, MSE=0.1708984   
##   
## Node number 15: 420 observations, complexity param=0.01246383  
## mean=0.8285714, MSE=0.1420408   
## left son=30 (9 obs) right son=31 (411 obs)  
## Primary splits:  
## UGDS\_UNKN < 0.00035 to the left, improve=0.10583940, (0 missing)  
## PCIP52 < 0.30305 to the right, improve=0.07731458, (0 missing)  
## PCIP51 < 0.00545 to the left, improve=0.06388152, (0 missing)  
## PCIP11 < 0.00305 to the left, improve=0.03910111, (0 missing)  
## PCIP40 < 0.00635 to the left, improve=0.03800287, (0 missing)  
## Surrogate splits:  
## UGDS\_WOMEN < 0.10785 to the left, agree=0.981, adj=0.111, (0 split)  
##   
## Node number 22: 60 observations  
## mean=0.2333333, MSE=0.1788889   
##   
## Node number 23: 23 observations  
## mean=0.9130435, MSE=0.07939509   
##   
## Node number 24: 173 observations  
## mean=0.0867052, MSE=0.07918741   
##   
## Node number 25: 28 observations  
## mean=0.6428571, MSE=0.2295918   
##   
## Node number 26: 27 observations  
## mean=0, MSE=0   
##   
## Node number 27: 187 observations, complexity param=0.01568867  
## mean=0.5989305, MSE=0.2402128   
## left son=54 (44 obs) right son=55 (143 obs)  
## Primary splits:  
## UGDS\_NRA < 0.01335 to the left, improve=0.17693220, (0 missing)  
## COSTT4\_A < 18126 to the left, improve=0.12741300, (0 missing)  
## SAT\_AVG\_ALL < 990.5 to the left, improve=0.09657143, (0 missing)  
## PCTFLOAN < 0.6469 to the right, improve=0.09339239, (0 missing)  
## PCIP52 < 0.10585 to the right, improve=0.06622578, (0 missing)  
## Surrogate splits:  
## SAT\_AVG\_ALL < 848.5 to the left, agree=0.791, adj=0.114, (0 split)  
## UGDS\_WOMEN < 0.59305 to the right, agree=0.791, adj=0.114, (0 split)  
## ADM\_RATE\_ALL < 0.4384291 to the left, agree=0.786, adj=0.091, (0 split)  
## PCIP27 < 0.0231 to the right, agree=0.775, adj=0.045, (0 split)  
## COSTT4\_A < 15122.5 to the left, agree=0.770, adj=0.023, (0 split)  
##   
## Node number 30: 9 observations  
## mean=0, MSE=0   
##   
## Node number 31: 411 observations  
## mean=0.8467153, MSE=0.1297885   
##   
## Node number 54: 44 observations  
## mean=0.2272727, MSE=0.1756198   
##   
## Node number 55: 143 observations  
## mean=0.7132867, MSE=0.2045088

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] 0.9086629

# 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] 0.9613936

# 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: 1712

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] 0.9161959

# 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" "PCTFLOAN" "SAT\_AVG\_ALL" "UGDS\_UNKN" "COSTT4\_A"   
## [6] "PCIP26" "UGDS\_NRA" "PCIP45"   
## Number of terminal nodes: 9   
## Residual mean deviance: 0.07542 = 239.5 / 3176   
## Distribution of residuals:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -0.84670 -0.05468 -0.05468 0.00000 -0.05468 0.94530

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] 0.9048964

# 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] 0.8954802

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

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

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

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

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

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

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

# First, we will create another copy of the dataset  
usunivnoccbasic <- usunivfilter  
  
# Next, we will change those that have "NA" to 0, since there is no data in it  
usunivnoccbasic[usunivnoccbasic == "NA"] <- 0  
  
# Next, we will choose rows that have complete cases  
usunivnoccbasic <- usunivnoccbasic[complete.cases(usunivnoccbasic),]  
  
# Now that we have the cleansed dataset, we will implement Boruta  
boruta.train <- Boruta(ACCEPTED ~ .-CCBASIC2, data=usunivnoccbasic)  
print(boruta.train)

## Boruta performed 99 iterations in 26.68167 secs.  
## 61 attributes confirmed important: ADM\_RATE, ADM\_RATE\_ALL,  
## C150\_4, C150\_4\_2MOR, C150\_4\_AIAN and 56 more.  
## 7 attributes confirmed unimportant: C150\_4\_NHPI, PCIP12, PCIP25,  
## PCIP29, PCIP46 and 2 more.  
## 2 tentative attributes left: PCIP10, PCIP22.

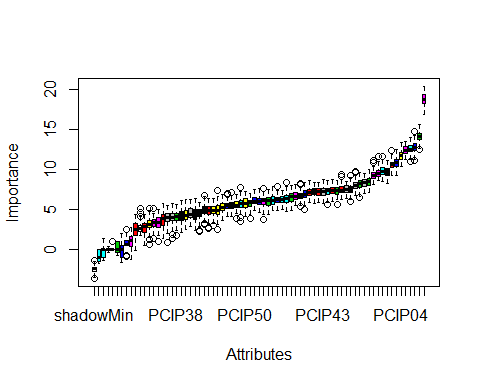
getSelectedAttributes(boruta.train)

## [1] "REGION" "ADM\_RATE" "ADM\_RATE\_ALL"   
## [4] "SAT\_AVG\_ALL" "PCIP01" "PCIP03"   
## [7] "PCIP04" "PCIP05" "PCIP09"   
## [10] "PCIP11" "PCIP13" "PCIP14"   
## [13] "PCIP15" "PCIP16" "PCIP19"   
## [16] "PCIP23" "PCIP24" "PCIP26"   
## [19] "PCIP27" "PCIP30" "PCIP31"   
## [22] "PCIP38" "PCIP39" "PCIP40"   
## [25] "PCIP41" "PCIP42" "PCIP43"   
## [28] "PCIP44" "PCIP45" "PCIP49"   
## [31] "PCIP50" "PCIP51" "PCIP52"   
## [34] "PCIP54" "UGDS\_WHITE" "UGDS\_BLACK"   
## [37] "UGDS\_HISP" "UGDS\_ASIAN" "UGDS\_AIAN"   
## [40] "UGDS\_NHPI" "UGDS\_2MOR" "UGDS\_NRA"   
## [43] "UGDS\_UNKN" "PPTUG\_EF" "COSTT4\_A"   
## [46] "TUITIONFEE\_IN" "TUITIONFEE\_OUT" "C150\_4"   
## [49] "C150\_4\_WHITE" "C150\_4\_BLACK" "C150\_4\_HISP"   
## [52] "C150\_4\_ASIAN" "C150\_4\_AIAN" "C150\_4\_2MOR"   
## [55] "C150\_4\_NRA" "C150\_4\_UNKN" "RET\_FT4"   
## [58] "PCTFLOAN" "PAR\_ED\_PCT\_1STGEN" "UGDS\_MEN"   
## [61] "UGDS\_WOMEN"

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

## meanImp medianImp minImp maxImp normHits  
## REGION 5.47942183 5.46964504 4.1538281 7.080243 1.0000000  
## ADM\_RATE 7.31370115 7.34158553 5.9575645 8.682065 1.0000000  
## ADM\_RATE\_ALL 7.29667769 7.36472452 5.5650597 8.638561 1.0000000  
## SAT\_AVG\_ALL 12.74858260 12.75941490 11.0197997 14.722248 1.0000000  
## PCIP01 6.24194849 6.24876094 5.0268305 7.333695 1.0000000  
## PCIP03 6.66038177 6.59482682 5.4311020 8.120056 1.0000000  
## PCIP04 11.64650663 11.70326513 10.3178158 13.307337 1.0000000  
## PCIP05 8.28031024 8.29715303 6.9668829 9.476272 1.0000000  
## PCIP09 4.91687253 4.86669918 3.1063624 6.724992 0.9797980  
## PCIP10 2.73999285 2.84768123 0.4895816 4.284176 0.5050505  
## PCIP11 6.45137409 6.53982999 4.8711187 7.901310 1.0000000  
## PCIP12 0.75845752 0.59856872 -0.8721671 2.488382 0.0000000  
## PCIP13 6.11519317 6.11411803 4.7869445 7.893389 1.0000000  
## PCIP14 18.73317178 18.68075768 16.8249828 20.328698 1.0000000  
## PCIP15 4.85185558 4.88803727 2.9356288 6.362803 0.9898990  
## PCIP16 7.47240366 7.40903862 6.2892436 8.514377 1.0000000  
## PCIP19 7.49171282 7.49920449 5.9058008 9.169181 1.0000000  
## PCIP22 2.49348993 2.48176602 -0.1219153 4.328714 0.4646465  
## PCIP23 8.39699279 8.44609400 7.0658809 9.981093 1.0000000  
## PCIP24 6.07078209 5.93648751 4.8040451 7.502615 1.0000000  
## PCIP25 -0.96033155 -1.00100150 -1.7669632 0.000000 0.0000000  
## PCIP26 5.92940617 6.02905237 3.7186663 7.584382 1.0000000  
## PCIP27 5.22607164 5.23572987 3.6350918 6.783261 1.0000000  
## PCIP29 0.07700012 0.00000000 0.0000000 1.001002 0.0000000  
## PCIP30 4.10305534 4.13425628 2.1344666 6.138143 0.9292929  
## PCIP31 4.82295316 4.86692930 2.6560936 6.372045 0.9797980  
## PCIP38 4.03613871 4.04288711 1.6693988 5.601888 0.8989899  
## PCIP39 5.41976047 5.46210178 4.0462837 6.940330 1.0000000  
## PCIP40 5.66828885 5.66936510 3.5274178 7.679365 0.9898990  
## PCIP41 3.32852534 3.36416953 0.9441927 4.996688 0.7777778  
## PCIP42 4.92969784 4.90325766 2.5045039 7.384272 0.9696970  
## PCIP43 7.19382012 7.21369603 5.6892709 8.674257 1.0000000  
## PCIP44 4.45797698 4.52366020 2.2446995 5.886658 0.9595960  
## PCIP45 7.47898935 7.39611418 6.2256162 9.272526 1.0000000  
## PCIP46 0.22271741 0.00000000 -0.7162887 1.001002 0.0000000  
## PCIP47 -0.02151351 0.00000000 -1.6851208 2.003904 0.0000000  
## PCIP48 -0.41019956 -0.08594479 -1.4161652 1.370378 0.0000000  
## PCIP49 3.31210226 3.34713255 1.3104724 5.098957 0.7979798  
## PCIP50 5.79574691 5.76840684 4.0842360 7.265149 1.0000000  
## PCIP51 3.97425561 4.01896718 1.3068290 5.704495 0.8888889  
## PCIP52 9.67118638 9.76461492 8.1855220 11.128491 1.0000000  
## PCIP54 3.71374303 3.84637944 1.3691949 5.692630 0.8484848  
## UGDS\_WHITE 8.12278627 8.15721708 6.4076573 9.630323 1.0000000  
## UGDS\_BLACK 10.75118118 10.82386925 9.2218304 12.423843 1.0000000  
## UGDS\_HISP 6.30921622 6.32028993 4.6807852 8.360271 1.0000000  
## UGDS\_ASIAN 9.21663069 9.19726455 8.0639288 11.056395 1.0000000  
## UGDS\_AIAN 4.23741493 4.22810074 2.4186109 6.188673 0.9090909  
## UGDS\_NHPI 3.87426884 3.96664860 0.8456296 5.429106 0.8686869  
## UGDS\_2MOR 4.37219984 4.41870384 2.3091198 6.306687 0.9191919  
## UGDS\_NRA 7.17864570 7.20600444 5.8936414 8.385783 1.0000000  
## UGDS\_UNKN 5.95275179 6.06144769 4.4047298 7.479594 1.0000000  
## PPTUG\_EF 6.87355654 6.84414364 5.0239237 8.397178 1.0000000  
## COSTT4\_A 9.78605544 9.76179784 8.4512643 11.557316 1.0000000  
## TUITIONFEE\_IN 9.54898067 9.46226601 8.5673501 11.547049 1.0000000  
## TUITIONFEE\_OUT 5.51913790 5.50586781 3.8486614 6.778249 1.0000000  
## C150\_4 7.97692954 7.94227403 6.6535409 9.746049 1.0000000  
## C150\_4\_WHITE 6.65247095 6.62579174 5.3362076 8.189016 1.0000000  
## C150\_4\_BLACK 7.13650584 7.13086776 5.8523399 8.534632 1.0000000  
## C150\_4\_HISP 5.72122472 5.78829323 3.8476927 6.985596 1.0000000  
## C150\_4\_ASIAN 6.15856437 6.20928976 4.9016454 7.790729 1.0000000  
## C150\_4\_AIAN 7.17889144 7.27726176 5.6441681 8.528892 1.0000000  
## C150\_4\_NHPI 0.90701674 1.05753667 -1.2221456 2.674707 0.0000000  
## C150\_4\_2MOR 3.11715100 3.12167299 0.6290531 5.037113 0.7272727  
## C150\_4\_NRA 4.47833158 4.38740670 2.8558954 6.170639 0.9494949  
## C150\_4\_UNKN 7.17101075 7.18121570 5.7195723 8.342651 1.0000000  
## RET\_FT4 10.56012433 10.57214175 9.2307913 12.332062 1.0000000  
## PCTFLOAN 14.10103965 14.08782834 12.5127879 15.543151 1.0000000  
## PAR\_ED\_PCT\_1STGEN 5.97956710 5.94885749 4.6266454 7.559058 1.0000000  
## UGDS\_MEN 12.55337373 12.40386949 10.9114268 13.986566 1.0000000  
## UGDS\_WOMEN 12.37698907 12.38440489 11.2109533 13.430464 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 Confirmed  
## C150\_4\_NRA Confirmed  
## C150\_4\_UNKN Confirmed  
## RET\_FT4 Confirmed  
## PCTFLOAN Confirmed  
## PAR\_ED\_PCT\_1STGEN Confirmed  
## UGDS\_MEN Confirmed  
## UGDS\_WOMEN Confirmed

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



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

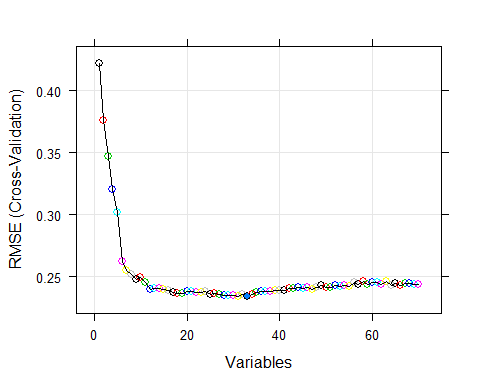
##   
## Attaching package: 'plyr'

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

predictors(rfe.train)

## [1] "PCIP14" "PCTFLOAN" "PCIP04" "SAT\_AVG\_ALL"   
## [5] "PCIP52" "UGDS\_BLACK" "UGDS\_MEN" "PCIP45"   
## [9] "UGDS\_WOMEN" "PCIP43" "COSTT4\_A" "PCIP23"   
## [13] "TUITIONFEE\_IN" "RET\_FT4" "UGDS\_HISP" "UGDS\_ASIAN"   
## [17] "PCIP39" "C150\_4\_AIAN" "PCIP16" "UGDS\_WHITE"   
## [21] "C150\_4" "UGDS\_NRA" "PCIP19" "PCIP24"   
## [25] "PCIP05" "PCIP03" "PPTUG\_EF" "PCIP26"   
## [29] "PCIP50" "PCIP09" "PCIP11" "UGDS\_UNKN"   
## [33] "PCIP41"

# 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 + PCTFLOAN + PAR\_ED\_PCT\_1STGEN)

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

.

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 = 50000, PCTFLOAN = 0.40, PAR\_ED\_PCT\_1STGEN = .40)  
predict(fit2, newdata = df\_accept3)

## 1   
## 0.7757938

# 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, PCTFLOAN = 0.58, PAR\_ED\_PCT\_1STGEN = .30)  
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.

# 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   
## -0.60336 -0.06269 0.00467 0.07538 0.48399   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.983e-01 4.623e-02 19.429 < 2e-16 \*\*\*  
## REGION -1.896e-03 3.377e-03 -0.561 0.574753   
## ADM\_RATE\_ALL -1.241e-01 3.913e-02 -3.170 0.001599 \*\*   
## UGDS\_NRA 2.155e-01 1.579e-01 1.365 0.172770   
## PPTUG\_EF -3.384e-01 8.887e-02 -3.807 0.000155 \*\*\*  
## COSTT4\_A 2.022e-06 6.364e-07 3.177 0.001562 \*\*   
## PCTFLOAN -3.634e-01 6.034e-02 -6.023 2.97e-09 \*\*\*  
## PAR\_ED\_PCT\_1STGEN -8.624e-02 1.011e-01 -0.853 0.393794   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1437 on 603 degrees of freedom  
## Multiple R-squared: 0.4192, Adjusted R-squared: 0.4125   
## F-statistic: 62.18 on 7 and 603 DF, p-value: < 2.2e-16

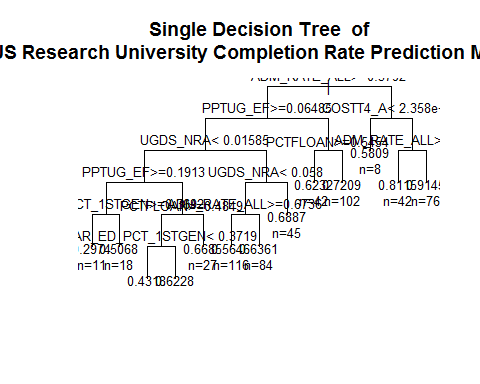
# do the testing with the prediction model  
accepted\_ind3 <- predict(lm\_NRAcompletion, interval="prediction", newdata = univ\_test2)  
  
# Checking on PRED(25)  
errors <- accepted\_ind3[,"fit"] - univ\_test2$C150\_4\_NRA  
rel\_change <- abs(errors) / univ\_test2$C150\_4\_NRA  
table(rel\_change<0.25)["TRUE"] / nrow(univ\_test2)

## TRUE   
## 0.8431373

# 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 0.27403158 0 1.0000000 1.0033301 0.06218336  
## 2 0.05485508 1 0.7259684 0.7680901 0.05641990  
## 3 0.03080516 2 0.6711133 0.7252923 0.05512253  
## 4 0.03025883 3 0.6403082 0.7311668 0.05690279  
## 5 0.02155523 4 0.6100493 0.7130296 0.05481675  
## 6 0.01517275 6 0.5669389 0.7115720 0.05487648  
## 7 0.01396383 7 0.5517661 0.6970487 0.05176186  
## 8 0.01339279 8 0.5378023 0.6987959 0.05198648  
## 9 0.01324165 9 0.5244095 0.7048866 0.05321058  
## 10 0.01187002 10 0.5111679 0.6973300 0.05289773  
## 11 0.01163708 11 0.4992978 0.6973884 0.05290986  
## 12 0.01000000 12 0.4876608 0.6875797 0.05256865  
##   
## Variable importance  
## ADM\_RATE\_ALL COSTT4\_A PPTUG\_EF PAR\_ED\_PCT\_1STGEN   
## 28 17 17 15   
## PCTFLOAN UGDS\_NRA   
## 14 9   
##   
## Node number 1: 611 observations, complexity param=0.2740316  
## mean=0.6665936, MSE=0.03507882   
## left son=2 (485 obs) right son=3 (126 obs)  
## Primary splits:  
## ADM\_RATE\_ALL < 0.3791867 to the right, improve=0.2740316, (0 missing)  
## COSTT4\_A < 47799 to the left, improve=0.2470153, (0 missing)  
## PPTUG\_EF < 0.05865 to the right, improve=0.2303289, (0 missing)  
## UGDS\_NRA < 0.05725 to the left, improve=0.2212987, (0 missing)  
## PCTFLOAN < 0.34615 to the right, improve=0.1989469, (0 missing)  
## Surrogate splits:  
## COSTT4\_A < 53857.5 to the left, agree=0.897, adj=0.500, (0 split)  
## PPTUG\_EF < 0.01375 to the right, agree=0.879, adj=0.413, (0 split)  
## PAR\_ED\_PCT\_1STGEN < 0.1801479 to the right, agree=0.872, adj=0.381, (0 split)  
## PCTFLOAN < 0.311 to the right, agree=0.869, adj=0.365, (0 split)  
## UGDS\_NRA < 0.0927 to the left, agree=0.822, adj=0.135, (0 split)  
##   
## Node number 2: 485 observations, complexity param=0.05485508  
## mean=0.6166204, MSE=0.02861418   
## left son=4 (341 obs) right son=5 (144 obs)  
## Primary splits:  
## PPTUG\_EF < 0.06485 to the right, improve=0.08471886, (0 missing)  
## UGDS\_NRA < 0.05695 to the left, improve=0.06904941, (0 missing)  
## PCTFLOAN < 0.5341 to the right, improve=0.06783895, (0 missing)  
## COSTT4\_A < 20337.5 to the left, improve=0.06011140, (0 missing)  
## PAR\_ED\_PCT\_1STGEN < 0.2098901 to the right, improve=0.05240259, (0 missing)  
## Surrogate splits:  
## COSTT4\_A < 28286.5 to the left, agree=0.819, adj=0.389, (0 split)  
## PAR\_ED\_PCT\_1STGEN < 0.2232916 to the right, agree=0.814, adj=0.375, (0 split)  
## ADM\_RATE\_ALL < 0.4488 to the right, agree=0.730, adj=0.090, (0 split)  
## UGDS\_NRA < 0.05655 to the left, agree=0.730, adj=0.090, (0 split)  
## PCTFLOAN < 0.31785 to the right, agree=0.713, adj=0.035, (0 split)  
##   
## Node number 3: 126 observations, complexity param=0.03080516  
## mean=0.8589508, MSE=0.01334858   
## left son=6 (8 obs) right son=7 (118 obs)  
## Primary splits:  
## COSTT4\_A < 23579.5 to the left, improve=0.3925582, (0 missing)  
## ADM\_RATE\_ALL < 0.2576 to the right, improve=0.3543141, (0 missing)  
## UGDS\_NRA < 0.05075 to the left, improve=0.2991775, (0 missing)  
## PPTUG\_EF < 0.0897 to the right, improve=0.2866781, (0 missing)  
## PCTFLOAN < 0.36115 to the right, improve=0.2500567, (0 missing)  
## Surrogate splits:  
## PAR\_ED\_PCT\_1STGEN < 0.3843536 to the right, agree=0.96, adj=0.375, (0 split)  
##   
## Node number 4: 341 observations, complexity param=0.03025883  
## mean=0.5846252, MSE=0.03131276   
## left son=8 (96 obs) right son=9 (245 obs)  
## Primary splits:  
## UGDS\_NRA < 0.01585 to the left, improve=0.06073830, (0 missing)  
## PCTFLOAN < 0.50175 to the right, improve=0.04614655, (0 missing)  
## COSTT4\_A < 19063 to the left, improve=0.03112851, (0 missing)  
## PPTUG\_EF < 0.1874 to the right, improve=0.03054918, (0 missing)  
## REGION < 5.5 to the left, improve=0.02560307, (0 missing)  
## Surrogate splits:  
## PCTFLOAN < 0.6687 to the right, agree=0.751, adj=0.115, (0 split)  
## COSTT4\_A < 12195.5 to the left, agree=0.721, adj=0.010, (0 split)  
##   
## Node number 5: 144 observations, complexity param=0.01324165  
## mean=0.6923868, MSE=0.01405908   
## left son=10 (42 obs) right son=11 (102 obs)  
## Primary splits:  
## PCTFLOAN < 0.54535 to the right, improve=0.14018740, (0 missing)  
## ADM\_RATE\_ALL < 0.5475 to the right, improve=0.11962090, (0 missing)  
## PAR\_ED\_PCT\_1STGEN < 0.4266404 to the right, improve=0.07585182, (0 missing)  
## UGDS\_NRA < 0.04185 to the left, improve=0.05172373, (0 missing)  
## PPTUG\_EF < 0.0366 to the right, improve=0.03606381, (0 missing)  
## Surrogate splits:  
## PPTUG\_EF < 0.0129 to the left, agree=0.743, adj=0.119, (0 split)  
## PAR\_ED\_PCT\_1STGEN < 0.489486 to the right, agree=0.722, adj=0.048, (0 split)  
##   
## Node number 6: 8 observations  
## mean=0.5809375, MSE=0.02467567   
##   
## Node number 7: 118 observations, complexity param=0.01339279  
## mean=0.8777992, MSE=0.006985287   
## left son=14 (42 obs) right son=15 (76 obs)  
## Primary splits:  
## ADM\_RATE\_ALL < 0.2491359 to the right, improve=0.3482499, (0 missing)  
## PCTFLOAN < 0.36975 to the right, improve=0.3193976, (0 missing)  
## PPTUG\_EF < 0.0069 to the right, improve=0.1094391, (0 missing)  
## UGDS\_NRA < 0.14025 to the right, improve=0.1037470, (0 missing)  
## COSTT4\_A < 52241 to the left, improve=0.1026357, (0 missing)  
## Surrogate splits:  
## PCTFLOAN < 0.34615 to the right, agree=0.839, adj=0.548, (0 split)  
## COSTT4\_A < 30814 to the left, agree=0.746, adj=0.286, (0 split)  
## UGDS\_NRA < 0.05695 to the left, agree=0.729, adj=0.238, (0 split)  
## PPTUG\_EF < 0.13365 to the right, agree=0.686, adj=0.119, (0 split)  
##   
## Node number 8: 96 observations, complexity param=0.02155523  
## mean=0.5149562, MSE=0.04569073   
## left son=16 (29 obs) right son=17 (67 obs)  
## Primary splits:  
## PPTUG\_EF < 0.1913 to the right, improve=0.07268514, (0 missing)  
## PCTFLOAN < 0.48485 to the right, improve=0.07246463, (0 missing)  
## ADM\_RATE\_ALL < 0.449895 to the left, improve=0.05839661, (0 missing)  
## COSTT4\_A < 17518 to the right, improve=0.05047381, (0 missing)  
## REGION < 5.5 to the left, improve=0.04941225, (0 missing)  
## Surrogate splits:  
## ADM\_RATE\_ALL < 0.4754317 to the left, agree=0.781, adj=0.276, (0 split)  
## UGDS\_NRA < 0.01475 to the right, agree=0.729, adj=0.103, (0 split)  
## PAR\_ED\_PCT\_1STGEN < 0.3807375 to the right, agree=0.719, adj=0.069, (0 split)  
##   
## Node number 9: 245 observations, complexity param=0.01517275  
## mean=0.6119241, MSE=0.02303183   
## left son=18 (200 obs) right son=19 (45 obs)  
## Primary splits:  
## UGDS\_NRA < 0.058 to the left, improve=0.05763098, (0 missing)  
## COSTT4\_A < 19810.5 to the left, improve=0.05525262, (0 missing)  
## ADM\_RATE\_ALL < 0.6801 to the right, improve=0.05139144, (0 missing)  
## PCTFLOAN < 0.4178 to the right, improve=0.04317412, (0 missing)  
## PAR\_ED\_PCT\_1STGEN < 0.3781181 to the right, improve=0.02574886, (0 missing)  
## Surrogate splits:  
## COSTT4\_A < 44982.5 to the left, agree=0.861, adj=0.244, (0 split)  
## ADM\_RATE\_ALL < 0.401225 to the right, agree=0.820, adj=0.022, (0 split)  
##   
## Node number 10: 42 observations  
## mean=0.6232024, MSE=0.01474872   
##   
## Node number 11: 102 observations  
## mean=0.7208745, MSE=0.01099266   
##   
## Node number 14: 42 observations  
## mean=0.8114524, MSE=0.009096337   
##   
## Node number 15: 76 observations  
## mean=0.9144645, MSE=0.002041683   
##   
## Node number 16: 29 observations, complexity param=0.01396383  
## mean=0.4273621, MSE=0.03212559   
## left son=32 (11 obs) right son=33 (18 obs)  
## Primary splits:  
## PAR\_ED\_PCT\_1STGEN < 0.3691543 to the right, improve=0.3212488, (0 missing)  
## UGDS\_NRA < 0.00935 to the left, improve=0.3126479, (0 missing)  
## ADM\_RATE\_ALL < 0.45645 to the left, improve=0.2568820, (0 missing)  
## PPTUG\_EF < 0.2873 to the right, improve=0.1183324, (0 missing)  
## COSTT4\_A < 18019 to the left, improve=0.1158147, (0 missing)  
## Surrogate splits:  
## UGDS\_NRA < 0.00935 to the left, agree=0.828, adj=0.545, (0 split)  
## ADM\_RATE\_ALL < 0.45005 to the left, agree=0.759, adj=0.364, (0 split)  
## PPTUG\_EF < 0.29755 to the right, agree=0.759, adj=0.364, (0 split)  
## PCTFLOAN < 0.6062 to the right, agree=0.724, adj=0.273, (0 split)  
## COSTT4\_A < 20715 to the right, agree=0.655, adj=0.091, (0 split)  
##   
## Node number 17: 67 observations, complexity param=0.02155523  
## mean=0.5528701, MSE=0.0468037   
## left son=34 (40 obs) right son=35 (27 obs)  
## Primary splits:  
## PCTFLOAN < 0.48485 to the right, improve=0.19298570, (0 missing)  
## ADM\_RATE\_ALL < 0.601 to the right, improve=0.10606310, (0 missing)  
## COSTT4\_A < 18166.5 to the right, improve=0.08868176, (0 missing)  
## PPTUG\_EF < 0.14585 to the left, improve=0.08427680, (0 missing)  
## PAR\_ED\_PCT\_1STGEN < 0.375729 to the left, improve=0.04699657, (0 missing)  
## Surrogate splits:  
## PPTUG\_EF < 0.14585 to the left, agree=0.701, adj=0.259, (0 split)  
## PAR\_ED\_PCT\_1STGEN < 0.2178337 to the right, agree=0.687, adj=0.222, (0 split)  
## REGION < 7.5 to the left, agree=0.657, adj=0.148, (0 split)  
## COSTT4\_A < 14028.5 to the right, agree=0.642, adj=0.111, (0 split)  
## ADM\_RATE\_ALL < 0.5859 to the right, agree=0.627, adj=0.074, (0 split)  
##   
## Node number 18: 200 observations, complexity param=0.01163708  
## mean=0.5946425, MSE=0.02272172   
## left son=36 (116 obs) right son=37 (84 obs)  
## Primary splits:  
## ADM\_RATE\_ALL < 0.6735546 to the right, improve=0.05488566, (0 missing)  
## COSTT4\_A < 19063 to the left, improve=0.03455355, (0 missing)  
## PCTFLOAN < 0.4183 to the right, improve=0.03264477, (0 missing)  
## PAR\_ED\_PCT\_1STGEN < 0.432151 to the right, improve=0.02897695, (0 missing)  
## UGDS\_NRA < 0.0524 to the right, improve=0.02430589, (0 missing)  
## Surrogate splits:  
## PPTUG\_EF < 0.18235 to the left, agree=0.615, adj=0.083, (0 split)  
## PCTFLOAN < 0.4241 to the right, agree=0.615, adj=0.083, (0 split)  
## REGION < 2.5 to the right, agree=0.610, adj=0.071, (0 split)  
## COSTT4\_A < 26346 to the left, agree=0.605, adj=0.060, (0 split)  
## PAR\_ED\_PCT\_1STGEN < 0.4315034 to the left, agree=0.605, adj=0.060, (0 split)  
##   
## Node number 19: 45 observations  
## mean=0.6887311, MSE=0.01718344   
##   
## Node number 32: 11 observations  
## mean=0.2974091, MSE=0.02764022   
##   
## Node number 33: 18 observations  
## mean=0.5067778, MSE=0.01823949   
##   
## Node number 34: 40 observations, complexity param=0.01187002  
## mean=0.4747875, MSE=0.04531037   
## left son=68 (31 obs) right son=69 (9 obs)  
## Primary splits:  
## PAR\_ED\_PCT\_1STGEN < 0.3719407 to the left, improve=0.14037180, (0 missing)  
## UGDS\_NRA < 0.00555 to the right, improve=0.12027280, (0 missing)  
## COSTT4\_A < 18166.5 to the right, improve=0.08172658, (0 missing)  
## PCTFLOAN < 0.6943 to the left, improve=0.05328135, (0 missing)  
## ADM\_RATE\_ALL < 0.6121318 to the right, improve=0.05171366, (0 missing)  
## Surrogate splits:  
## ADM\_RATE\_ALL < 0.5740898 to the right, agree=0.9, adj=0.556, (0 split)  
## PPTUG\_EF < 0.1762 to the left, agree=0.8, adj=0.111, (0 split)  
## PCTFLOAN < 0.74725 to the left, agree=0.8, adj=0.111, (0 split)  
##   
## Node number 35: 27 observations  
## mean=0.6685481, MSE=0.0266022   
##   
## Node number 36: 116 observations  
## mean=0.5645914, MSE=0.01753124   
##   
## Node number 37: 84 observations  
## mean=0.6361417, MSE=0.02692024   
##   
## Node number 68: 31 observations  
## mean=0.4318161, MSE=0.04020261   
##   
## Node number 69: 9 observations  
## mean=0.6228, MSE=0.03463577

plot(model\_dtree3, uniform = TRUE, main = "Single Decision Tree of\nUS Research University Completion Rate Prediction Model")  
text(model\_dtree3, use.n = TRUE, cex = .8)



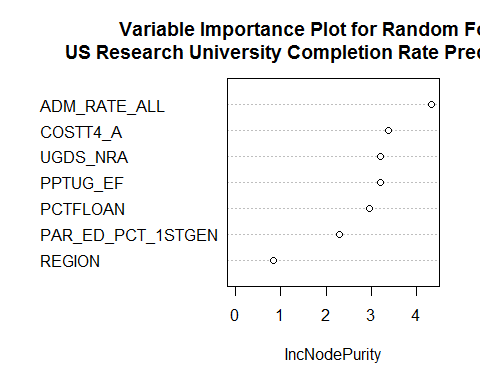
pred\_dtree3 <- predict(model\_dtree3, newdata = univ\_test2)  
accu11 <- abs(pred\_dtree3 - univ\_test2$C150\_4\_NRA) < 0.25  
frac11 <- sum(accu11)/length(accu11)  
print(frac11)

## [1] 0.9215686

# 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 Completion Rate Prediction Model")



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] 0.9411765

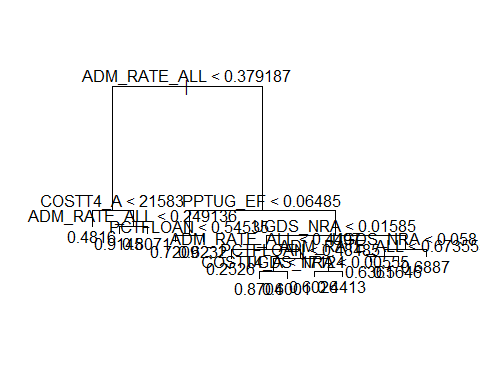
# Doing support vector machine  
model\_svm3 <- svm(formula\_completionrate, data = univ\_train2)  
summary(model\_svm3)

##   
## 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: 507

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] 0.9362745

# doing simple tree  
model\_tree3 <- tree(formula\_completionrate, data = univ\_train2)  
plot(model\_tree3, main = "Simple Tree of US Research\nUniversity Completion Rate Prediction Model")  
text(model\_tree3)



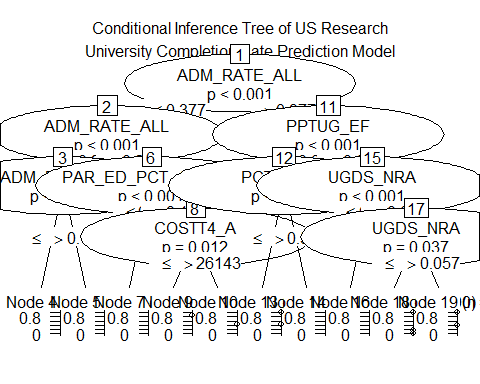
pred\_tree3 <- predict(model\_tree3, newdata = univ\_test2)  
accu14 <- abs(pred\_tree3 - univ\_test2$C150\_4\_NRA) < 0.25  
frac14 <- sum(accu14)/length(accu14)  
print(frac14)

## [1] 0.9068627

# 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 Prediction Model")



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] 0.9117647

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