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

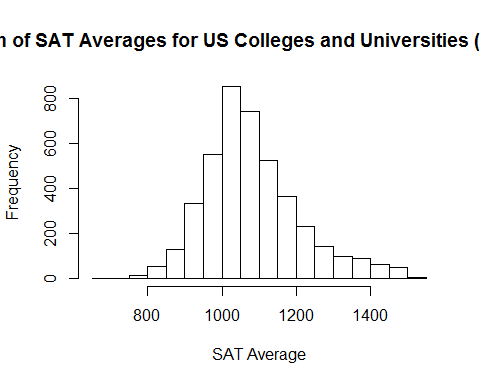
## [1] 815

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

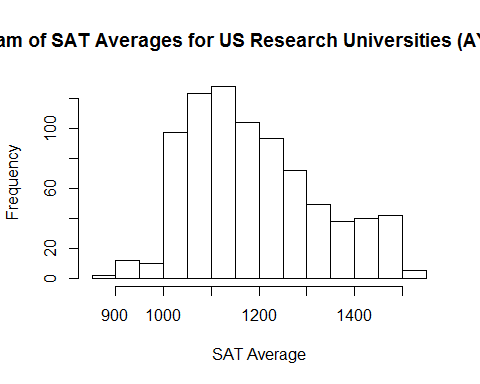
## [1] 0.7062392

# Distributions and Box and Whisker Plots

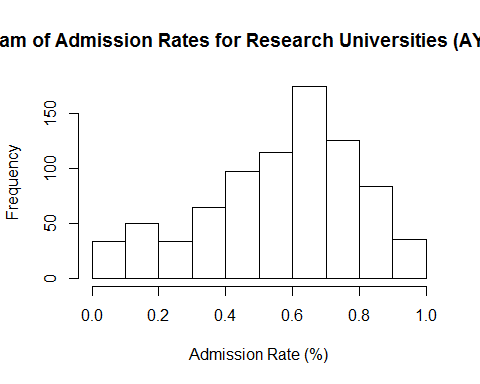
# Histogram of SAT Averages for US Colleges and Universities  
hist(usunivfilter$SAT\_AVG\_ALL, main = "Histogram of SAT Averages for US Colleges and Universities (AY2010-2015)", xlab="SAT Average")



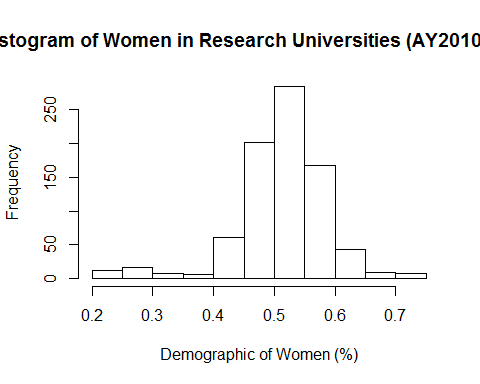
# Histogram of SAT Averages for US Research Universities  
hist(usresearchuniv$SAT\_AVG\_ALL, main = "Histogram of SAT Averages for US Research Universities (AY2010-2015)", xlab="SAT Average")



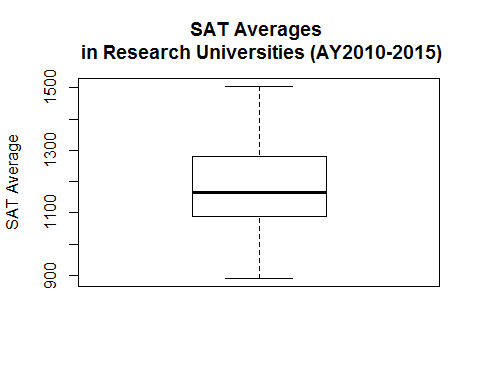
# Histogram of Admission Rates for US Research Universities  
hist(usresearchuniv$ADM\_RATE\_ALL, main = "Histogram of Admission Rates for Research Universities (AY2010-2015)", xlab = "Admission Rate (%)")



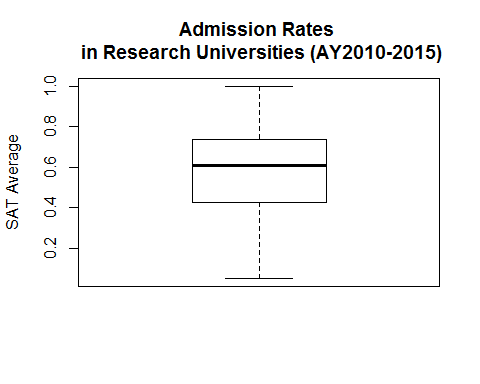
# Histogram of Women in US Research Universities  
hist(usresearchuniv$UGDS\_WOMEN, main = "Histogram of Women in Research Universities (AY2010-2015)", xlab = "Demographic of Women (%)")



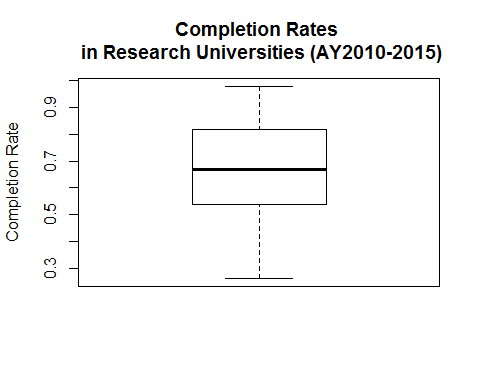
#Boxplot of SAT Average in all US Research Universities  
boxplot(usresearchuniv$SAT\_AVG\_ALL, main = "SAT Averages \n in Research Universities (AY2010-2015)", ylab = "SAT Average")



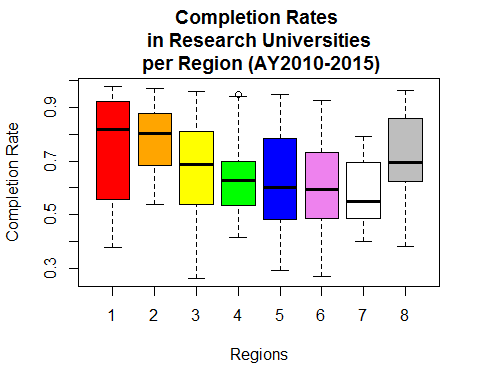
#Boxplot of admission rates in all US Research Universities  
boxplot(usresearchuniv$ADM\_RATE\_ALL, main = "Admission Rates \n in Research Universities (AY2010-2015)", ylab = "SAT Average")



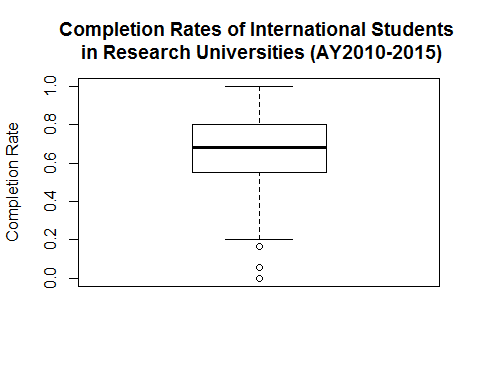
#Boxplot of Completion Rates in all US Research Universities  
boxplot(usresearchuniv$C150\_4, main = "Completion Rates \n in Research Universities (AY2010-2015)", ylab = "Completion Rate")



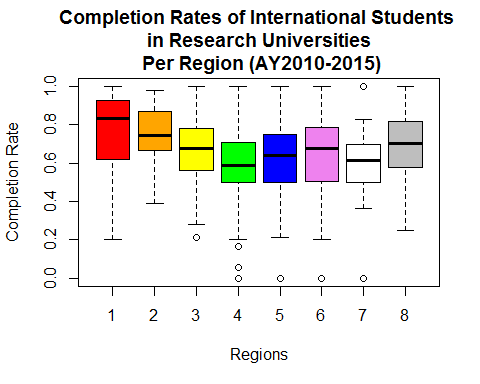
# Boxplot of Completion Rates per Region in US Research Universities  
boxplot(C150\_4 ~ REGION, usresearchuniv, main = "Completion Rates \n in Research Universities \n per Region (AY2010-2015)", col=c("red", "orange", "yellow", "green", "blue", "violet", "white", "gray", "magenta"), ylab = "Completion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

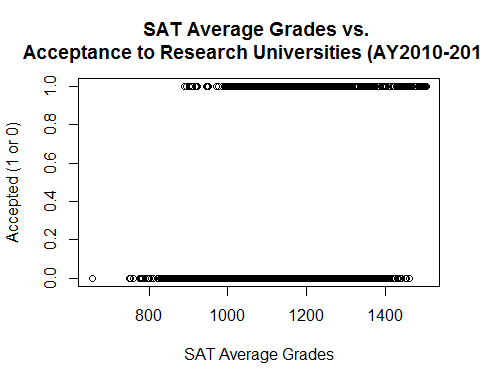


nrow(usresearchuniv[usresearchuniv$C150\_4\_NRA < 0.2,])

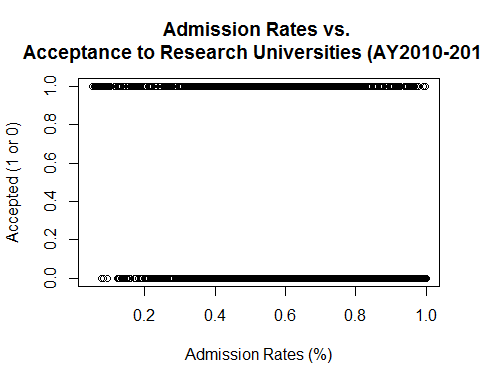
## [1] 9

# Correlations

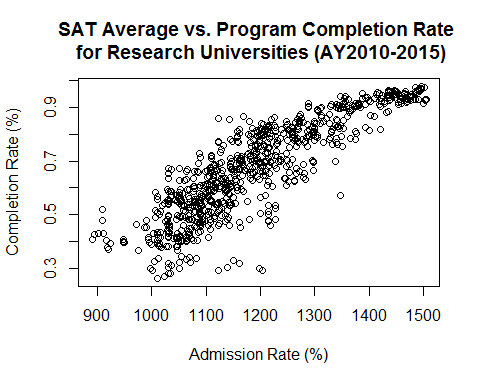
#Correlation between the SAT grades and the acceptance for the research universities  
plot(usunivfilter$SAT\_AVG\_ALL, usunivfilter$ACCEPTED, main="SAT Average Grades vs. \n Acceptance to Research Universities (AY2010-2015)", xlab="SAT Average Grades", ylab="Accepted (1 or 0)")



#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="Admission Rate (%)", ylab="Completion Rate (%)")

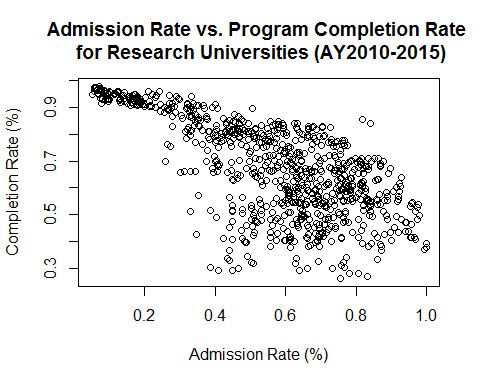


#Correlation coefficient between admission rate and completion rate  
cor(usresearchuniv$SAT\_AVG\_ALL, usresearchuniv$C150\_4, method = "pearson")

## [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 (%)")

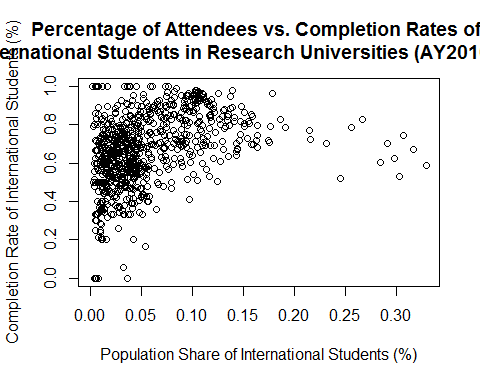


#Correlation coefficient between admission rate and completion rate  
cor(usresearchuniv$ADM\_RATE\_ALL, usresearchuniv$C150\_4, method = "pearson")

## [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 (%)")

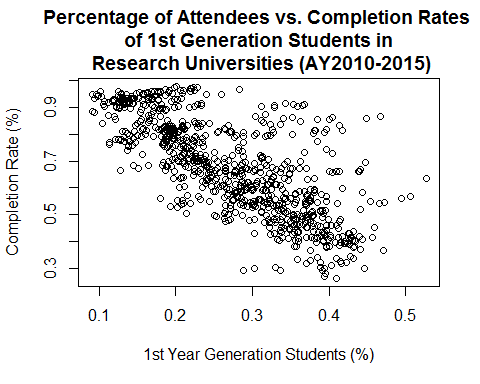


#Correlation coefficient between admission rate and completion rate of international students  
cor(usresearchuniv$UGDS\_NRA, usresearchuniv$C150\_4\_NRA, method = "pearson")

## [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   
## -2.1058 -0.5572 -0.3083 -0.1333 2.7176   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.424e+01 1.167e+00 -12.200 < 2e-16 \*\*\*  
## REGION 1.086e-01 2.852e-02 3.807 0.000141 \*\*\*  
## ADM\_RATE\_ALL 7.686e-01 3.769e-01 2.039 0.041453 \*   
## SAT\_AVG\_ALL 1.402e-02 8.232e-04 17.027 < 2e-16 \*\*\*  
## UGDS\_NRA 6.625e+00 1.345e+00 4.925 8.45e-07 \*\*\*  
## COSTT4\_A -8.553e-05 6.081e-06 -14.067 < 2e-16 \*\*\*  
## PCTFLOAN -6.514e-01 4.833e-01 -1.348 0.177743   
## UGDS\_WOMEN -2.050e+00 5.372e-01 -3.817 0.000135 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3139.9 on 3184 degrees of freedom  
## Residual deviance: 2195.2 on 3177 degrees of freedom  
## AIC: 2211.2  
##   
## Number of Fisher Scoring iterations: 6

# do the first testing with the prediction model  
accepted\_ind <- predict(glm\_ISAcceptance, type="response", newdata = univ\_test)  
pred1 <- prediction(accepted\_ind, univ\_test$ACCEPTED)  
  
# create the confusion matrix and accuracy for this prediction model  
c1 <- confusionMatrix(as.integer(accepted\_ind > 0.5), univ\_test$ACCEPTED)  
c1$table

## Reference  
## Prediction 0 1  
## 0 832 104  
## 1 35 91

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

## Accuracy   
## 0.8691149

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

## Neg Pred Value   
## 0.7222222

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

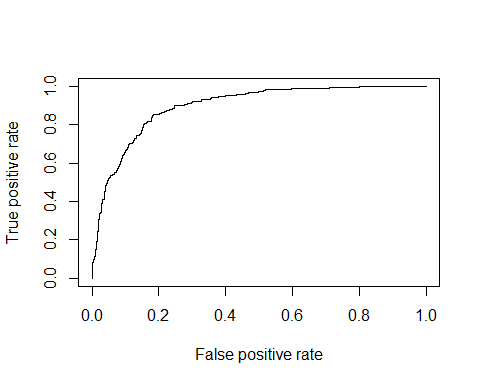
## Specificity   
## 0.4666667

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.

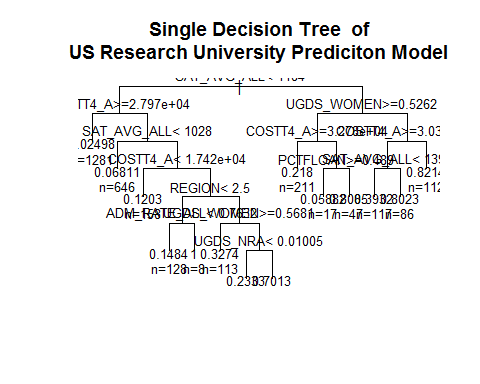
# show the curve on the performance  
perf1 <- performance(pred1, "tpr", "fpr")  
plot(perf1, lty = 1)



# 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 0.13074603 0 1.0000000 1.0005046 0.02734274  
## 2 0.05748337 1 0.8692540 0.8761436 0.02620999  
## 3 0.03562130 3 0.7542872 0.7658771 0.02047583  
## 4 0.02737374 4 0.7186659 0.7310287 0.02172208  
## 5 0.02478727 5 0.6912922 0.7175718 0.02236133  
## 6 0.01618020 6 0.6665049 0.7030683 0.02263949  
## 7 0.01506348 7 0.6503247 0.6809816 0.02330187  
## 8 0.01405262 8 0.6352612 0.6820549 0.02371446  
## 9 0.01300676 9 0.6212086 0.6683377 0.02420925  
## 10 0.01164530 11 0.5951951 0.6629026 0.02441874  
## 11 0.01093514 12 0.5835498 0.6523134 0.02453994  
## 12 0.01000000 13 0.5726147 0.6452961 0.02467096  
##   
## Variable importance  
## SAT\_AVG\_ALL COSTT4\_A PCTFLOAN UGDS\_WOMEN ADM\_RATE\_ALL   
## 31 24 12 11 9   
## UGDS\_NRA REGION   
## 7 7   
##   
## Node number 1: 3185 observations, complexity param=0.130746  
## mean=0.1946625, MSE=0.156769   
## left son=2 (2595 obs) right son=3 (590 obs)  
## Primary splits:  
## SAT\_AVG\_ALL < 1183.5 to the left, improve=0.13074600, (0 missing)  
## PCTFLOAN < 0.49405 to the right, improve=0.11579260, (0 missing)  
## UGDS\_WOMEN < 0.52825 to the right, improve=0.09585995, (0 missing)  
## ADM\_RATE\_ALL < 0.2799627 to the right, improve=0.04760589, (0 missing)  
## COSTT4\_A < 53785 to the left, improve=0.04340306, (0 missing)  
## Surrogate splits:  
## COSTT4\_A < 51565.5 to the left, agree=0.886, adj=0.385, (0 split)  
## ADM\_RATE\_ALL < 0.3750802 to the right, agree=0.860, adj=0.246, (0 split)  
## PCTFLOAN < 0.3729 to the right, agree=0.836, adj=0.114, (0 split)  
## UGDS\_NRA < 0.09575 to the left, agree=0.816, adj=0.007, (0 split)  
##   
## Node number 2: 2595 observations, complexity param=0.05748337  
## mean=0.1263969, MSE=0.1104207   
## left son=4 (1281 obs) right son=5 (1314 obs)  
## Primary splits:  
## COSTT4\_A < 27966.5 to the right, improve=0.09080710, (0 missing)  
## PCTFLOAN < 0.63285 to the right, improve=0.07109482, (0 missing)  
## UGDS\_WOMEN < 0.56585 to the right, improve=0.05170405, (0 missing)  
## SAT\_AVG\_ALL < 1009.5 to the left, improve=0.04978916, (0 missing)  
## REGION < 5.5 to the left, improve=0.01580513, (0 missing)  
## Surrogate splits:  
## PCTFLOAN < 0.61155 to the right, agree=0.707, adj=0.406, (0 split)  
## UGDS\_WOMEN < 0.61435 to the right, agree=0.606, adj=0.201, (0 split)  
## SAT\_AVG\_ALL < 1030.5 to the right, agree=0.580, adj=0.149, (0 split)  
## REGION < 4.5 to the left, agree=0.574, adj=0.137, (0 split)  
## UGDS\_NRA < 0.06015 to the right, agree=0.560, adj=0.108, (0 split)  
##   
## Node number 3: 590 observations, complexity param=0.0356213  
## mean=0.4949153, MSE=0.2499741   
## left son=6 (275 obs) right son=7 (315 obs)  
## Primary splits:  
## UGDS\_WOMEN < 0.52625 to the right, improve=0.12059580, (0 missing)  
## COSTT4\_A < 33250 to the right, improve=0.10602830, (0 missing)  
## PCTFLOAN < 0.4914 to the right, improve=0.09342539, (0 missing)  
## SAT\_AVG\_ALL < 1434.5 to the left, improve=0.06642929, (0 missing)  
## UGDS\_NRA < 0.092 to the left, improve=0.06453814, (0 missing)  
## Surrogate splits:  
## UGDS\_NRA < 0.03885 to the left, agree=0.620, adj=0.185, (0 split)  
## PCTFLOAN < 0.38305 to the right, agree=0.620, adj=0.185, (0 split)  
## ADM\_RATE\_ALL < 0.2551359 to the right, agree=0.615, adj=0.175, (0 split)  
## SAT\_AVG\_ALL < 1370.5 to the left, agree=0.612, adj=0.167, (0 split)  
## REGION < 3.5 to the right, agree=0.592, adj=0.124, (0 split)  
##   
## Node number 4: 1281 observations  
## mean=0.02498048, MSE=0.02435646   
##   
## Node number 5: 1314 observations, complexity param=0.05748337  
## mean=0.2252664, MSE=0.1745214   
## left son=10 (646 obs) right son=11 (668 obs)  
## Primary splits:  
## SAT\_AVG\_ALL < 1028.5 to the left, improve=0.13685580, (0 missing)  
## UGDS\_WOMEN < 0.54175 to the right, improve=0.07528795, (0 missing)  
## COSTT4\_A < 18149.5 to the left, improve=0.06793259, (0 missing)  
## PCTFLOAN < 0.63285 to the right, improve=0.03387021, (0 missing)  
## REGION < 2.5 to the left, improve=0.03027806, (0 missing)  
## Surrogate splits:  
## UGDS\_WOMEN < 0.52995 to the right, agree=0.643, adj=0.274, (0 split)  
## COSTT4\_A < 18137.5 to the left, agree=0.608, adj=0.203, (0 split)  
## PCTFLOAN < 0.60085 to the right, agree=0.587, adj=0.159, (0 split)  
## UGDS\_NRA < 0.01565 to the left, agree=0.578, adj=0.142, (0 split)  
## REGION < 5.5 to the right, agree=0.538, adj=0.060, (0 split)  
##   
## Node number 6: 275 observations, complexity param=0.01506348  
## mean=0.3090909, MSE=0.2135537   
## left son=12 (211 obs) right son=13 (64 obs)  
## Primary splits:  
## COSTT4\_A < 32777 to the right, improve=0.12807220, (0 missing)  
## PCTFLOAN < 0.4914 to the right, improve=0.11377710, (0 missing)  
## ADM\_RATE\_ALL < 0.5614704 to the right, improve=0.06150096, (0 missing)  
## UGDS\_NRA < 0.00665 to the left, improve=0.03390712, (0 missing)  
## SAT\_AVG\_ALL < 1338.5 to the left, improve=0.03315285, (0 missing)  
## Surrogate splits:  
## UGDS\_NRA < 0.0135 to the right, agree=0.811, adj=0.188, (0 split)  
## SAT\_AVG\_ALL < 1194.5 to the right, agree=0.771, adj=0.016, (0 split)  
## PCTFLOAN < 0.25255 to the right, agree=0.771, adj=0.016, (0 split)  
##   
## Node number 7: 315 observations, complexity param=0.01300676  
## mean=0.6571429, MSE=0.2253061   
## left son=14 (203 obs) right son=15 (112 obs)  
## Primary splits:  
## COSTT4\_A < 30317 to the right, improve=0.06609195, (0 missing)  
## UGDS\_NRA < 0.0991 to the left, improve=0.06441611, (0 missing)  
## PCTFLOAN < 0.69545 to the right, improve=0.06284153, (0 missing)  
## UGDS\_WOMEN < 0.22565 to the left, improve=0.06284153, (0 missing)  
## ADM\_RATE\_ALL < 0.1285159 to the right, improve=0.05743563, (0 missing)  
## Surrogate splits:  
## SAT\_AVG\_ALL < 1299 to the right, agree=0.768, adj=0.348, (0 split)  
## ADM\_RATE\_ALL < 0.4401319 to the left, agree=0.762, adj=0.330, (0 split)  
## UGDS\_NRA < 0.02365 to the right, agree=0.714, adj=0.196, (0 split)  
## REGION < 3.5 to the left, agree=0.673, adj=0.080, (0 split)  
##   
## Node number 10: 646 observations  
## mean=0.06811146, MSE=0.06347228   
##   
## Node number 11: 668 observations, complexity param=0.02737374  
## mean=0.3772455, MSE=0.2349313   
## left son=22 (158 obs) right son=23 (510 obs)  
## Primary splits:  
## COSTT4\_A < 17415.5 to the left, improve=0.08709362, (0 missing)  
## REGION < 5.5 to the left, improve=0.07885399, (0 missing)  
## UGDS\_WOMEN < 0.56775 to the right, improve=0.07025798, (0 missing)  
## SAT\_AVG\_ALL < 1106.5 to the left, improve=0.05248366, (0 missing)  
## PCTFLOAN < 0.6226 to the right, improve=0.04525375, (0 missing)  
## Surrogate splits:  
## PCTFLOAN < 0.33755 to the left, agree=0.792, adj=0.120, (0 split)  
## ADM\_RATE\_ALL < 0.3244553 to the left, agree=0.778, adj=0.063, (0 split)  
## SAT\_AVG\_ALL < 1030.5 to the left, agree=0.771, adj=0.032, (0 split)  
## UGDS\_NRA < 0.26405 to the right, agree=0.766, adj=0.013, (0 split)  
##   
## Node number 12: 211 observations  
## mean=0.2180095, MSE=0.1704813   
##   
## Node number 13: 64 observations, complexity param=0.01405262  
## mean=0.609375, MSE=0.2380371   
## left son=26 (17 obs) right son=27 (47 obs)  
## Primary splits:  
## PCTFLOAN < 0.48905 to the right, improve=0.4605770, (0 missing)  
## UGDS\_NRA < 0.00515 to the left, improve=0.4368000, (0 missing)  
## REGION < 4.5 to the left, improve=0.3871780, (0 missing)  
## ADM\_RATE\_ALL < 0.4543585 to the right, improve=0.2413846, (0 missing)  
## UGDS\_WOMEN < 0.5603 to the right, improve=0.1070494, (0 missing)  
## Surrogate splits:  
## UGDS\_NRA < 0.00365 to the left, agree=0.844, adj=0.412, (0 split)  
## REGION < 3.5 to the left, agree=0.828, adj=0.353, (0 split)  
## COSTT4\_A < 17287.5 to the left, agree=0.766, adj=0.118, (0 split)  
## UGDS\_WOMEN < 0.545 to the left, agree=0.766, adj=0.118, (0 split)  
##   
## Node number 14: 203 observations, complexity param=0.01300676  
## mean=0.5665025, MSE=0.2455774   
## left son=28 (117 obs) right son=29 (86 obs)  
## Primary splits:  
## SAT\_AVG\_ALL < 1393.5 to the left, improve=0.16645500, (0 missing)  
## ADM\_RATE\_ALL < 0.1327 to the right, improve=0.15606260, (0 missing)  
## UGDS\_NRA < 0.0647 to the left, improve=0.15317050, (0 missing)  
## UGDS\_WOMEN < 0.41 to the left, improve=0.07404349, (0 missing)  
## PCTFLOAN < 0.5862 to the right, improve=0.05695653, (0 missing)  
## Surrogate splits:  
## ADM\_RATE\_ALL < 0.1714445 to the right, agree=0.882, adj=0.721, (0 split)  
## PCTFLOAN < 0.2691 to the right, agree=0.803, adj=0.535, (0 split)  
## COSTT4\_A < 55839 to the left, agree=0.704, adj=0.302, (0 split)  
## REGION < 1.5 to the right, agree=0.655, adj=0.186, (0 split)  
## UGDS\_NRA < 0.06145 to the left, agree=0.650, adj=0.174, (0 split)  
##   
## Node number 15: 112 observations  
## mean=0.8214286, MSE=0.1466837   
##   
## Node number 22: 158 observations  
## mean=0.1202532, MSE=0.1057923   
##   
## Node number 23: 510 observations, complexity param=0.02478727  
## mean=0.4568627, MSE=0.2481392   
## left son=46 (136 obs) right son=47 (374 obs)  
## Primary splits:  
## REGION < 2.5 to the left, improve=0.09779865, (0 missing)  
## PCTFLOAN < 0.6226 to the right, improve=0.05726940, (0 missing)  
## UGDS\_WOMEN < 0.56725 to the right, improve=0.05163126, (0 missing)  
## ADM\_RATE\_ALL < 0.6801161 to the left, improve=0.05003886, (0 missing)  
## SAT\_AVG\_ALL < 1098.5 to the left, improve=0.04948230, (0 missing)  
## Surrogate splits:  
## PCTFLOAN < 0.6577 to the right, agree=0.798, adj=0.243, (0 split)  
## UGDS\_WOMEN < 0.4316 to the left, agree=0.775, adj=0.154, (0 split)  
## UGDS\_NRA < 0.0036 to the left, agree=0.753, adj=0.074, (0 split)  
## COSTT4\_A < 25660.5 to the right, agree=0.745, adj=0.044, (0 split)  
##   
## Node number 26: 17 observations  
## mean=0.05882353, MSE=0.05536332   
##   
## Node number 27: 47 observations  
## mean=0.8085106, MSE=0.1548212   
##   
## Node number 28: 117 observations  
## mean=0.3931624, MSE=0.2385857   
##   
## Node number 29: 86 observations  
## mean=0.8023256, MSE=0.1585992   
##   
## Node number 46: 136 observations, complexity param=0.01093514  
## mean=0.1985294, MSE=0.1591155   
## left son=92 (128 obs) right son=93 (8 obs)  
## Primary splits:  
## ADM\_RATE\_ALL < 0.763201 to the left, improve=0.2523148, (0 missing)  
## REGION < 1.5 to the right, improve=0.2281456, (0 missing)  
## PCTFLOAN < 0.4787 to the right, improve=0.1479303, (0 missing)  
## SAT\_AVG\_ALL < 1101 to the left, improve=0.1206690, (0 missing)  
## UGDS\_WOMEN < 0.5458 to the right, improve=0.0772241, (0 missing)  
##   
## Node number 47: 374 observations, complexity param=0.0161802  
## mean=0.5508021, MSE=0.2474191   
## left son=94 (113 obs) right son=95 (261 obs)  
## Primary splits:  
## UGDS\_WOMEN < 0.56815 to the right, improve=0.08730689, (0 missing)  
## REGION < 5.5 to the left, improve=0.05461206, (0 missing)  
## UGDS\_NRA < 0.0108 to the left, improve=0.04024279, (0 missing)  
## SAT\_AVG\_ALL < 1076.5 to the left, improve=0.03879262, (0 missing)  
## PCTFLOAN < 0.6226 to the right, improve=0.03792261, (0 missing)  
## Surrogate splits:  
## UGDS\_NRA < 0.00425 to the left, agree=0.717, adj=0.062, (0 split)  
## SAT\_AVG\_ALL < 1029.5 to the left, agree=0.706, adj=0.027, (0 split)  
## PCTFLOAN < 0.75765 to the right, agree=0.703, adj=0.018, (0 split)  
##   
## Node number 92: 128 observations  
## mean=0.1484375, MSE=0.1264038   
##   
## Node number 93: 8 observations  
## mean=1, MSE=0   
##   
## Node number 94: 113 observations  
## mean=0.3274336, MSE=0.2202208   
##   
## Node number 95: 261 observations, complexity param=0.0116453  
## mean=0.6475096, MSE=0.2282409   
## left son=190 (30 obs) right son=191 (231 obs)  
## Primary splits:  
## UGDS\_NRA < 0.01005 to the left, improve=0.09760816, (0 missing)  
## SAT\_AVG\_ALL < 1074.5 to the left, improve=0.06832165, (0 missing)  
## UGDS\_WOMEN < 0.22695 to the left, improve=0.06560559, (0 missing)  
## PCTFLOAN < 0.53295 to the right, improve=0.04899722, (0 missing)  
## COSTT4\_A < 18365 to the left, improve=0.04613861, (0 missing)  
## Surrogate splits:  
## ADM\_RATE\_ALL < 0.9639968 to the right, agree=0.889, adj=0.033, (0 split)  
##   
## Node number 190: 30 observations  
## mean=0.2333333, MSE=0.1788889   
##   
## Node number 191: 231 observations  
## mean=0.7012987, MSE=0.2094788

plot(model\_dtree1, uniform = TRUE, main = "Single Decision Tree of\nUS Research University Prediciton 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] 0.8870056

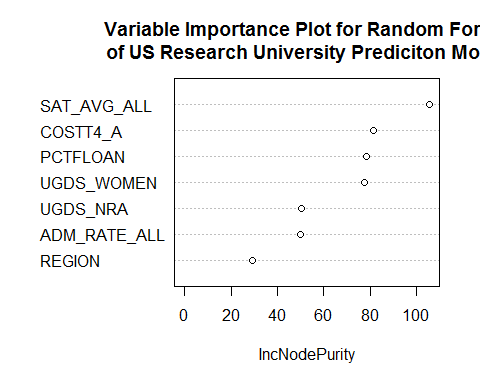
# doing random forest  
model\_forest1 <- randomForest(formula\_ISAcceptance, data = univ\_train)

## 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 Prediciton 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] 0.9519774

# 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: 1401

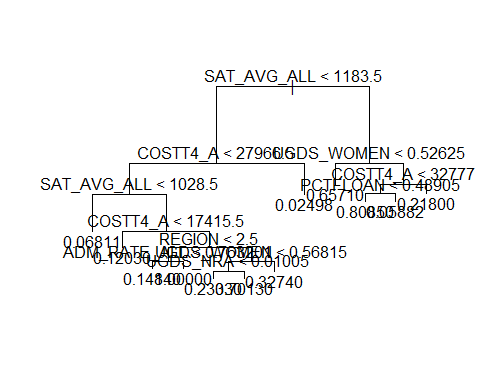
pred\_svm1 <- predict(model\_svm1, newdata = univ\_test)  
accu3 <- abs(pred\_svm1 - univ\_test$ACCEPTED) < 0.5  
frac3 <- sum(accu3)/length(accu3)  
print(frac3)

## [1] 0.8992467

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

##   
## Regression tree:  
## tree(formula = formula\_ISAcceptance, data = univ\_train)  
## Number of terminal nodes: 12   
## Residual mean deviance: 0.0942 = 298.9 / 3173   
## Distribution of residuals:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -0.80850 -0.06811 -0.02498 0.00000 -0.02498 0.97500

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



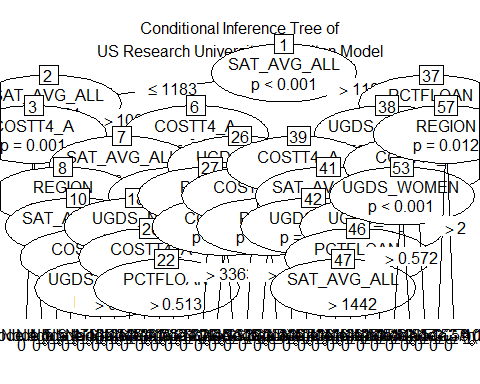
pred\_tree1 <- predict(model\_tree1, newdata = univ\_test)  
accu4 <- abs(pred\_tree1 - univ\_test$ACCEPTED) < 0.5  
frac4 <- sum(accu4)/length(accu4)  
print(frac4)

## [1] 0.8700565

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

## Length Class Mode   
## 1 BinaryTree S4

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



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

## [1] 0.8907721

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   
## -2.62143 -0.47933 -0.24986 -0.08671 3.07458   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.749e+01 1.447e+00 -12.090 < 2e-16 \*\*\*  
## REGION 1.238e-01 3.062e-02 4.041 5.32e-05 \*\*\*  
## ADM\_RATE\_ALL 1.254e+00 4.210e-01 2.978 0.002897 \*\*   
## SAT\_AVG\_ALL 1.498e-02 1.009e-03 14.852 < 2e-16 \*\*\*  
## PCIP11 1.846e+00 1.889e+00 0.977 0.328358   
## PCIP12 3.372e-01 1.909e+01 0.018 0.985905   
## PCIP14 6.789e+00 8.557e-01 7.934 2.13e-15 \*\*\*  
## PCIP15 -2.453e+00 2.157e+00 -1.137 0.255485   
## PCIP24 -5.641e+00 1.272e+00 -4.435 9.22e-06 \*\*\*  
## PCIP26 6.246e+00 1.686e+00 3.704 0.000212 \*\*\*  
## PCIP27 -2.384e+01 6.660e+00 -3.579 0.000345 \*\*\*  
## PCIP40 -3.546e+01 4.833e+00 -7.337 2.18e-13 \*\*\*  
## PCIP45 9.325e+00 1.209e+00 7.715 1.21e-14 \*\*\*  
## PCIP51 2.361e+00 6.054e-01 3.900 9.62e-05 \*\*\*  
## PCIP52 7.748e-01 6.636e-01 1.168 0.243001   
## UGDS\_NRA 8.697e+00 1.515e+00 5.739 9.53e-09 \*\*\*  
## UGDS\_UNKN -1.855e+00 1.566e+00 -1.184 0.236225   
## COSTT4\_A -1.035e-04 7.218e-06 -14.336 < 2e-16 \*\*\*  
## PCTFLOAN -2.620e-01 5.624e-01 -0.466 0.641304   
## UGDS\_WOMEN 3.976e-01 8.230e-01 0.483 0.629018   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3139.9 on 3184 degrees of freedom  
## Residual deviance: 1923.6 on 3165 degrees of freedom  
## AIC: 1963.6  
##   
## 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  
## 0 822 87  
## 1 45 108

c2$overall['Accuracy']

## Accuracy   
## 0.8757062

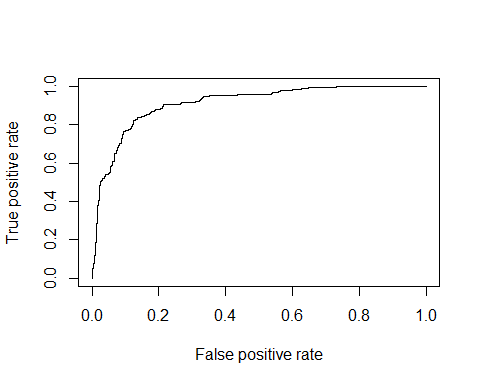
#Precision of the logistic regression model  
c2$byClass['Neg Pred Value']

## Neg Pred Value   
## 0.7058824

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

## Specificity   
## 0.5538462

# 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 0.33104225 0 1.0000000 1.0006221 0.02734596  
## 2 0.05601820 1 0.6689578 0.6834344 0.02300713  
## 3 0.03259472 2 0.6129396 0.6427926 0.02497454  
## 4 0.02074622 3 0.5803448 0.6197855 0.02536444  
## 5 0.01817888 4 0.5595986 0.5999983 0.02575770  
## 6 0.01767021 5 0.5414197 0.5936382 0.02580687  
## 7 0.01293240 6 0.5237495 0.5782633 0.02559146  
## 8 0.01204567 7 0.5108171 0.5626835 0.02534925  
## 9 0.01099359 9 0.4867258 0.5535836 0.02521501  
## 10 0.01096512 10 0.4757322 0.5375351 0.02488643  
## 11 0.01071942 11 0.4647671 0.5363307 0.02490256  
## 12 0.01000000 14 0.4304237 0.5315796 0.02483723  
##   
## Variable importance  
## PCIP14 PCTFLOAN SAT\_AVG\_ALL UGDS\_WOMEN PCIP45   
## 42 10 9 8 8   
## ADM\_RATE\_ALL COSTT4\_A PCIP51 REGION PCIP24   
## 4 3 2 2 2   
## UGDS\_UNKN PCIP26 PCIP15 PCIP52 PCIP27   
## 2 1 1 1 1   
## PCIP11 UGDS\_NRA PCIP40   
## 1 1 1   
##   
## Node number 1: 3185 observations, complexity param=0.3310422  
## mean=0.1946625, MSE=0.156769   
## left son=2 (2377 obs) right son=3 (808 obs)  
## Primary splits:  
## PCIP14 < 0.03595 to the left, improve=0.33104220, (0 missing)  
## SAT\_AVG\_ALL < 1183.5 to the left, improve=0.13074600, (0 missing)  
## PCTFLOAN < 0.49405 to the right, improve=0.11579260, (0 missing)  
## UGDS\_WOMEN < 0.52825 to the right, improve=0.09585995, (0 missing)  
## PCIP45 < 0.04945 to the left, improve=0.07096971, (0 missing)  
## Surrogate splits:  
## UGDS\_WOMEN < 0.5158 to the right, agree=0.791, adj=0.176, (0 split)  
## SAT\_AVG\_ALL < 1248.5 to the left, agree=0.759, adj=0.051, (0 split)  
## ADM\_RATE\_ALL < 0.1285159 to the right, agree=0.755, adj=0.035, (0 split)  
## PCIP11 < 0.1109 to the left, agree=0.752, adj=0.024, (0 split)  
## PCIP15 < 0.2006 to the left, agree=0.749, adj=0.009, (0 split)  
##   
## Node number 2: 2377 observations, complexity param=0.01817888  
## mean=0.06184266, MSE=0.05801814   
## left son=4 (2361 obs) right son=5 (16 obs)  
## Primary splits:  
## PCIP45 < 0.3384 to the left, improve=0.06581784, (0 missing)  
## SAT\_AVG\_ALL < 1194.5 to the left, improve=0.04958316, (0 missing)  
## PCIP14 < 0.0202 to the left, improve=0.04455203, (0 missing)  
## PCTFLOAN < 0.62345 to the right, improve=0.03275809, (0 missing)  
## COSTT4\_A < 53515 to the left, improve=0.02585318, (0 missing)  
## Surrogate splits:  
## SAT\_AVG\_ALL < 1456.5 to the left, agree=0.995, adj=0.250, (0 split)  
## ADM\_RATE\_ALL < 0.09145359 to the right, agree=0.995, adj=0.187, (0 split)  
## COSTT4\_A < 62038.5 to the left, agree=0.994, adj=0.125, (0 split)  
##   
## Node number 3: 808 observations, complexity param=0.0560182  
## mean=0.585396, MSE=0.2427075   
## left son=6 (210 obs) right son=7 (598 obs)  
## Primary splits:  
## PCTFLOAN < 0.58815 to the right, improve=0.14262780, (0 missing)  
## PCIP45 < 0.03155 to the left, improve=0.12973290, (0 missing)  
## PCIP26 < 0.0244 to the left, improve=0.11207500, (0 missing)  
## SAT\_AVG\_ALL < 1118.5 to the left, improve=0.10477250, (0 missing)  
## PCIP40 < 0.00505 to the left, improve=0.08421334, (0 missing)  
## Surrogate splits:  
## SAT\_AVG\_ALL < 903.5 to the left, agree=0.757, adj=0.067, (0 split)  
## PCIP15 < 0.0496 to the right, agree=0.756, adj=0.062, (0 split)  
## PCIP45 < 0.0358 to the left, agree=0.756, adj=0.062, (0 split)  
## UGDS\_WOMEN < 0.6216 to the right, agree=0.755, adj=0.057, (0 split)  
## PCIP27 < 0.00335 to the left, agree=0.754, adj=0.052, (0 split)  
##   
## Node number 4: 2361 observations, complexity param=0.01204567  
## mean=0.05675561, MSE=0.05353441   
## left son=8 (2231 obs) right son=9 (130 obs)  
## Primary splits:  
## PCIP14 < 0.0202 to the left, improve=0.04564549, (0 missing)  
## SAT\_AVG\_ALL < 1194.5 to the left, improve=0.03050129, (0 missing)  
## PCIP45 < 0.04985 to the left, improve=0.02975223, (0 missing)  
## PCTFLOAN < 0.62345 to the right, improve=0.02823914, (0 missing)  
## COSTT4\_A < 27651.5 to the right, improve=0.01739769, (0 missing)  
##   
## Node number 5: 16 observations  
## mean=0.8125, MSE=0.1523438   
##   
## Node number 6: 210 observations, complexity param=0.02074622  
## mean=0.2714286, MSE=0.1977551   
## left son=12 (115 obs) right son=13 (95 obs)  
## Primary splits:  
## COSTT4\_A < 26509.5 to the right, improve=0.24943740, (0 missing)  
## PCIP45 < 0.02765 to the left, improve=0.12366920, (0 missing)  
## PCTFLOAN < 0.67325 to the right, improve=0.10541330, (0 missing)  
## PCIP26 < 0.02475 to the left, improve=0.07967811, (0 missing)  
## PCIP40 < 0.00375 to the left, improve=0.07707911, (0 missing)  
## Surrogate splits:  
## UGDS\_NRA < 0.01845 to the right, agree=0.714, adj=0.368, (0 split)  
## SAT\_AVG\_ALL < 1050.5 to the right, agree=0.676, adj=0.284, (0 split)  
## PCIP45 < 0.059 to the left, agree=0.676, adj=0.284, (0 split)  
## PCTFLOAN < 0.6645 to the right, agree=0.667, adj=0.263, (0 split)  
## ADM\_RATE\_ALL < 0.81835 to the left, agree=0.657, adj=0.242, (0 split)  
##   
## Node number 7: 598 observations, complexity param=0.03259472  
## mean=0.6956522, MSE=0.2117202   
## left son=14 (127 obs) right son=15 (471 obs)  
## Primary splits:  
## SAT\_AVG\_ALL < 1074.5 to the left, improve=0.12854440, (0 missing)  
## PCIP26 < 0.02375 to the left, improve=0.11458840, (0 missing)  
## PCIP45 < 0.01515 to the left, improve=0.09332801, (0 missing)  
## PCIP40 < 0.00695 to the left, improve=0.07604117, (0 missing)  
## PCIP52 < 0.3013 to the right, improve=0.06895389, (0 missing)  
## Surrogate splits:  
## PCIP24 < 0.0714 to the right, agree=0.813, adj=0.118, (0 split)  
## COSTT4\_A < 16973.5 to the left, agree=0.808, adj=0.094, (0 split)  
## PCIP26 < 0.02835 to the left, agree=0.806, adj=0.087, (0 split)  
## UGDS\_WOMEN < 0.60405 to the right, agree=0.799, adj=0.055, (0 split)  
## PCIP40 < 0.0016 to the left, agree=0.798, adj=0.047, (0 split)  
##   
## Node number 8: 2231 observations, complexity param=0.01071942  
## mean=0.04482295, MSE=0.04281385   
## left son=16 (2002 obs) right son=17 (229 obs)  
## Primary splits:  
## SAT\_AVG\_ALL < 1194.5 to the left, improve=0.03117160, (0 missing)  
## PCTFLOAN < 0.62805 to the right, improve=0.02786754, (0 missing)  
## PCIP45 < 0.04985 to the left, improve=0.01864373, (0 missing)  
## PCIP14 < 0.0041 to the left, improve=0.01470923, (0 missing)  
## COSTT4\_A < 26319 to the right, improve=0.01417084, (0 missing)  
## Surrogate splits:  
## COSTT4\_A < 52398.5 to the left, agree=0.934, adj=0.358, (0 split)  
## PCIP45 < 0.2007 to the left, agree=0.929, adj=0.306, (0 split)  
## ADM\_RATE\_ALL < 0.3308226 to the right, agree=0.915, adj=0.170, (0 split)  
## PCIP40 < 0.05475 to the left, agree=0.909, adj=0.118, (0 split)  
## PCIP52 < 0.0037 to the right, agree=0.902, adj=0.048, (0 split)  
##   
## Node number 9: 130 observations, complexity param=0.01204567  
## mean=0.2615385, MSE=0.1931361   
## left son=18 (66 obs) right son=19 (64 obs)  
## Primary splits:  
## PCIP45 < 0.0629 to the left, improve=0.24931330, (0 missing)  
## ADM\_RATE\_ALL < 0.6912 to the right, improve=0.13464450, (0 missing)  
## PCIP40 < 0.0061 to the left, improve=0.11832890, (0 missing)  
## PCIP26 < 0.0284 to the left, improve=0.08854167, (0 missing)  
## SAT\_AVG\_ALL < 1204.5 to the left, improve=0.08642129, (0 missing)  
## Surrogate splits:  
## PCIP26 < 0.03305 to the left, agree=0.677, adj=0.344, (0 split)  
## REGION < 5.5 to the left, agree=0.662, adj=0.312, (0 split)  
## PCIP51 < 0.0998 to the right, agree=0.662, adj=0.312, (0 split)  
## ADM\_RATE\_ALL < 0.6757158 to the right, agree=0.654, adj=0.297, (0 split)  
## PCIP40 < 0.0061 to the left, agree=0.646, adj=0.281, (0 split)  
##   
## Node number 12: 115 observations  
## mean=0.06956522, MSE=0.0647259   
##   
## Node number 13: 95 observations, complexity param=0.0129324  
## mean=0.5157895, MSE=0.2497507   
## left son=26 (60 obs) right son=27 (35 obs)  
## Primary splits:  
## UGDS\_UNKN < 0.02445 to the left, improve=0.2721564, (0 missing)  
## PCIP45 < 0.0273 to the left, improve=0.2669363, (0 missing)  
## PCIP51 < 0.02455 to the left, improve=0.1841116, (0 missing)  
## PCIP40 < 0.00195 to the left, improve=0.1394928, (0 missing)  
## UGDS\_NRA < 0.016 to the left, improve=0.1361299, (0 missing)  
## Surrogate splits:  
## REGION < 1.5 to the right, agree=0.789, adj=0.429, (0 split)  
## PCIP45 < 0.08705 to the left, agree=0.716, adj=0.229, (0 split)  
## PCIP24 < 0.02875 to the left, agree=0.705, adj=0.200, (0 split)  
## SAT\_AVG\_ALL < 1114 to the left, agree=0.684, adj=0.143, (0 split)  
## PCIP27 < 0.0093 to the left, agree=0.674, adj=0.114, (0 split)  
##   
## Node number 14: 127 observations, complexity param=0.01096512  
## mean=0.3779528, MSE=0.2351045   
## left son=28 (109 obs) right son=29 (18 obs)  
## Primary splits:  
## PCTFLOAN < 0.5548 to the left, improve=0.1833658, (0 missing)  
## PCIP45 < 0.1261 to the left, improve=0.1612094, (0 missing)  
## PCIP15 < 0.00205 to the right, improve=0.1350381, (0 missing)  
## PCIP40 < 0.00635 to the left, improve=0.1343720, (0 missing)  
## UGDS\_WOMEN < 0.46005 to the left, improve=0.1261968, (0 missing)  
## Surrogate splits:  
## PCIP45 < 0.183 to the left, agree=0.874, adj=0.111, (0 split)  
##   
## Node number 15: 471 observations, complexity param=0.01767021  
## mean=0.7813163, MSE=0.1708611   
## left son=30 (36 obs) right son=31 (435 obs)  
## Primary splits:  
## PCIP45 < 0.01385 to the left, improve=0.10963450, (0 missing)  
## PCIP52 < 0.3308 to the right, improve=0.09340694, (0 missing)  
## PCTFLOAN < 0.51945 to the right, improve=0.08377489, (0 missing)  
## PCIP26 < 0.01505 to the left, improve=0.08017503, (0 missing)  
## UGDS\_WOMEN < 0.2026 to the left, improve=0.07750142, (0 missing)  
## Surrogate splits:  
## UGDS\_WOMEN < 0.37715 to the left, agree=0.955, adj=0.417, (0 split)  
## PCIP26 < 0.01505 to the left, agree=0.947, adj=0.306, (0 split)  
## PCIP15 < 0.0707 to the right, agree=0.945, adj=0.278, (0 split)  
## PCIP27 < 0.00085 to the left, agree=0.943, adj=0.250, (0 split)  
## PCIP14 < 0.3224 to the right, agree=0.938, adj=0.194, (0 split)  
##   
## Node number 16: 2002 observations  
## mean=0.03246753, MSE=0.03141339   
##   
## Node number 17: 229 observations, complexity param=0.01071942  
## mean=0.1528384, MSE=0.1294788   
## left son=34 (186 obs) right son=35 (43 obs)  
## Primary splits:  
## PCIP51 < 0.03005 to the left, improve=0.26060720, (0 missing)  
## PCIP52 < 0.13255 to the left, improve=0.23344770, (0 missing)  
## PCIP14 < 0.0021 to the left, improve=0.14032420, (0 missing)  
## PCIP40 < 0.02975 to the right, improve=0.12212590, (0 missing)  
## PCIP24 < 0.02935 to the left, improve=0.09131809, (0 missing)  
## Surrogate splits:  
## PCIP15 < 2e-04 to the left, agree=0.825, adj=0.070, (0 split)  
## SAT\_AVG\_ALL < 1196 to the right, agree=0.821, adj=0.047, (0 split)  
## PCTFLOAN < 0.60215 to the left, agree=0.817, adj=0.023, (0 split)  
##   
## Node number 18: 66 observations  
## mean=0.04545455, MSE=0.04338843   
##   
## Node number 19: 64 observations  
## mean=0.484375, MSE=0.2497559   
##   
## Node number 26: 60 observations  
## mean=0.3166667, MSE=0.2163889   
##   
## Node number 27: 35 observations  
## mean=0.8571429, MSE=0.122449   
##   
## Node number 28: 109 observations  
## mean=0.293578, MSE=0.20739   
##   
## Node number 29: 18 observations  
## mean=0.8888889, MSE=0.09876543   
##   
## Node number 30: 36 observations  
## mean=0.3055556, MSE=0.2121914   
##   
## Node number 31: 435 observations, complexity param=0.01099359  
## mean=0.8206897, MSE=0.1471581   
## left son=62 (8 obs) right son=63 (427 obs)  
## Primary splits:  
## PCIP52 < 0.3308 to the right, improve=0.08575032, (0 missing)  
## UGDS\_UNKN < 0.001 to the left, improve=0.06161818, (0 missing)  
## COSTT4\_A < 25314 to the right, improve=0.05040126, (0 missing)  
## SAT\_AVG\_ALL < 1321.5 to the left, improve=0.04143869, (0 missing)  
## UGDS\_NRA < 0.09835 to the left, improve=0.03634280, (0 missing)  
##   
## Node number 34: 186 observations  
## mean=0.06451613, MSE=0.0603538   
##   
## Node number 35: 43 observations, complexity param=0.01071942  
## mean=0.5348837, MSE=0.2487831   
## left son=70 (23 obs) right son=71 (20 obs)  
## Primary splits:  
## PCTFLOAN < 0.44025 to the right, improve=0.6023110, (0 missing)  
## PCIP40 < 0.044 to the right, improve=0.4451613, (0 missing)  
## ADM\_RATE\_ALL < 0.5665704 to the right, improve=0.3861161, (0 missing)  
## PCIP24 < 0.0216 to the left, improve=0.2989130, (0 missing)  
## SAT\_AVG\_ALL < 1328.5 to the left, improve=0.2301790, (0 missing)  
## Surrogate splits:  
## ADM\_RATE\_ALL < 0.4278319 to the right, agree=0.814, adj=0.60, (0 split)  
## PCIP24 < 0.0216 to the left, agree=0.791, adj=0.55, (0 split)  
## REGION < 4.5 to the left, agree=0.767, adj=0.50, (0 split)  
## SAT\_AVG\_ALL < 1299.5 to the left, agree=0.767, adj=0.50, (0 split)  
## PCIP14 < 0.00135 to the left, agree=0.744, adj=0.45, (0 split)  
##   
## Node number 62: 8 observations  
## mean=0, MSE=0   
##   
## Node number 63: 427 observations  
## mean=0.8360656, MSE=0.1370599   
##   
## Node number 70: 23 observations  
## mean=0.173913, MSE=0.1436673   
##   
## Node number 71: 20 observations  
## mean=0.95, MSE=0.0475

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] 0.8898305

# 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] 0.9595104

# 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: 1664

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] 0.9096045

# doing simple tree  
model\_tree2 <- tree(formula\_ISSciAcceptance, data = univ\_train)  
summary(model\_tree2)

##   
## Regression tree:  
## tree(formula = formula\_ISSciAcceptance, data = univ\_train)  
## Variables actually used in tree construction:  
## [1] "PCIP14" "PCIP45" "PCTFLOAN" "SAT\_AVG\_ALL" "PCIP52"   
## [6] "COSTT4\_A" "UGDS\_UNKN"   
## Number of terminal nodes: 12   
## Residual mean deviance: 0.07314 = 232.1 / 3173   
## Distribution of residuals:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -0.88890 -0.04482 -0.04482 0.00000 -0.04482 0.95520

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] 0.8879473

# 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] 0.8907721

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 27.276 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.

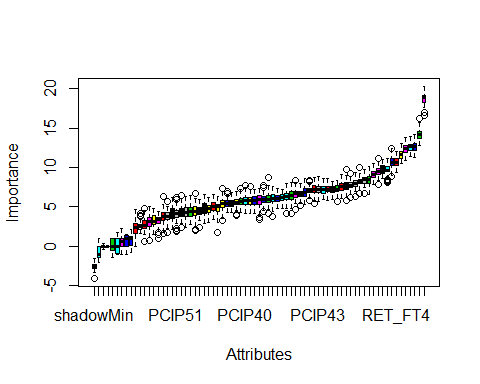
getSelectedAttributes(boruta.train)

## [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 5.51126524 5.54894123 4.24049025 6.793117 1.00000000  
## ADM\_RATE 7.20767885 7.18608658 5.38800550 8.793640 1.00000000  
## ADM\_RATE\_ALL 7.27670715 7.26197089 5.71127710 8.379075 1.00000000  
## SAT\_AVG\_ALL 12.65261967 12.67101025 11.27073872 14.241541 1.00000000  
## PCIP01 6.11980159 6.14763973 4.78083092 7.169711 1.00000000  
## PCIP03 6.64873968 6.68469457 4.66129047 8.382289 1.00000000  
## PCIP04 11.63916670 11.65104991 10.55331657 12.998818 1.00000000  
## PCIP05 8.38676963 8.42916009 6.76209687 9.698037 1.00000000  
## PCIP09 4.78162000 4.73877462 2.89840281 6.639328 0.97979798  
## PCIP10 2.79165086 2.61746224 0.56453678 4.818660 0.55555556  
## PCIP11 6.46511163 6.45699567 4.23206338 8.081567 1.00000000  
## PCIP12 0.75206510 1.02989108 -0.77122933 2.180724 0.01010101  
## PCIP13 6.04509861 6.03864857 3.80706483 7.652644 1.00000000  
## PCIP14 18.75413049 18.83421223 16.57956759 20.300596 1.00000000  
## PCIP15 4.76333045 4.77777635 3.02783354 6.250375 0.98989899  
## PCIP16 7.73519219 7.75840082 6.24217714 9.296801 1.00000000  
## PCIP19 7.63671569 7.63620761 5.79041391 9.728163 1.00000000  
## PCIP22 2.30116386 2.35584166 -0.02574986 4.649051 0.38383838  
## PCIP23 8.40649440 8.49479434 6.56633908 10.149158 1.00000000  
## PCIP24 5.88537837 5.85427953 4.21001897 7.238844 1.00000000  
## PCIP25 -0.85757646 -1.00100150 -2.00694617 1.001002 0.00000000  
## PCIP26 5.84159749 5.90262843 3.38419983 7.338262 1.00000000  
## PCIP27 5.34552877 5.40275248 3.28864181 7.301105 1.00000000  
## PCIP29 0.00000000 0.00000000 0.00000000 0.000000 0.00000000  
## PCIP30 4.16719678 4.11298375 2.14757380 6.211766 0.95959596  
## PCIP31 5.02458870 5.02808154 3.37143359 6.579676 1.00000000  
## PCIP38 4.35268360 4.37350865 2.71432209 6.405423 0.97979798  
## PCIP39 5.46459782 5.48860225 4.29164684 6.925625 1.00000000  
## PCIP40 5.73896419 5.75978719 4.23557736 7.670370 1.00000000  
## PCIP41 3.09267963 3.16290853 0.78254671 5.095272 0.70707071  
## PCIP42 4.91218422 5.08692048 1.77335547 6.316996 0.98989899  
## PCIP43 7.24806676 7.20204980 5.72724852 8.872772 1.00000000  
## PCIP44 4.50198984 4.52807023 2.42026992 5.936823 0.98989899  
## PCIP45 7.59354952 7.57888549 6.07964826 9.146177 1.00000000  
## PCIP46 0.03791968 0.00000000 -1.41704486 1.001002 0.00000000  
## PCIP47 0.67292323 1.00100150 -1.00100150 1.564722 0.00000000  
## PCIP48 0.15998342 0.04336047 -1.09302410 1.726766 0.00000000  
## PCIP49 3.23969342 3.35443989 1.05480145 4.450157 0.81818182  
## PCIP50 5.82559879 5.81851387 4.11842749 7.625256 1.00000000  
## PCIP51 4.15949914 4.15232553 1.93515545 6.154640 0.91919192  
## PCIP52 9.73915717 9.65484516 8.60139203 11.030027 1.00000000  
## PCIP54 3.77448626 3.80656199 1.58262688 6.314206 0.91919192  
## UGDS\_WHITE 8.21293113 8.26739441 6.73470979 10.003768 1.00000000  
## UGDS\_BLACK 10.67575608 10.67727259 8.85848665 12.472461 1.00000000  
## UGDS\_HISP 6.31566879 6.29192199 4.17496472 7.977846 1.00000000  
## UGDS\_ASIAN 9.17239948 9.17804474 7.89659489 10.481348 1.00000000  
## UGDS\_AIAN 4.39770980 4.40014806 2.03092397 6.731413 0.97979798  
## UGDS\_NHPI 3.80261613 3.81594645 1.78098961 5.690864 0.91919192  
## UGDS\_2MOR 4.35079910 4.28834026 2.37308382 6.448532 0.97979798  
## UGDS\_NRA 7.24799082 7.21904293 5.82669815 8.599984 1.00000000  
## UGDS\_UNKN 6.05730116 6.00219237 4.15359043 8.686824 1.00000000  
## PPTUG\_EF 6.82445030 6.88507227 5.29047804 8.657967 1.00000000  
## COSTT4\_A 9.74268521 9.66782895 8.11327098 11.123805 1.00000000  
## TUITIONFEE\_IN 9.48884129 9.48917467 7.89806733 11.154066 1.00000000  
## TUITIONFEE\_OUT 5.58367676 5.62293850 3.96807510 6.956440 1.00000000  
## C150\_4 7.99532368 8.00393184 6.36126521 9.428266 1.00000000  
## C150\_4\_WHITE 6.70197515 6.69906130 5.19937866 7.926092 1.00000000  
## C150\_4\_BLACK 7.13926029 7.11175602 5.80029068 8.329869 1.00000000  
## C150\_4\_HISP 5.76121225 5.68857668 4.28113133 7.383579 1.00000000  
## C150\_4\_ASIAN 6.08917430 6.10545401 4.81850996 7.397430 1.00000000  
## C150\_4\_AIAN 7.21644119 7.21220843 5.98174624 8.671905 1.00000000  
## C150\_4\_NHPI 0.59867175 0.67873184 -1.03211729 2.279194 0.01010101  
## C150\_4\_2MOR 3.29718864 3.21474151 1.56321827 4.766082 0.78787879  
## C150\_4\_NRA 4.33775940 4.36242619 2.40113695 6.326319 0.93939394  
## C150\_4\_UNKN 7.21184689 7.24786052 5.92940285 8.269922 1.00000000  
## RET\_FT4 10.66182193 10.69181216 9.42792054 11.976761 1.00000000  
## PCTFLOAN 14.17277782 14.17062583 12.83940898 16.171268 1.00000000  
## PAR\_ED\_PCT\_1STGEN 5.91165812 5.94088169 4.46042241 7.754593 1.00000000  
## UGDS\_MEN 12.59144867 12.64813755 11.40981387 13.880126 1.00000000  
## UGDS\_WOMEN 12.40092179 12.42303284 10.91048576 13.776179 1.00000000  
## decision  
## REGION Confirmed  
## ADM\_RATE Confirmed  
## ADM\_RATE\_ALL Confirmed  
## SAT\_AVG\_ALL Confirmed  
## PCIP01 Confirmed  
## PCIP03 Confirmed  
## PCIP04 Confirmed  
## PCIP05 Confirmed  
## PCIP09 Confirmed  
## PCIP10 Tentative  
## PCIP11 Confirmed  
## PCIP12 Rejected  
## PCIP13 Confirmed  
## PCIP14 Confirmed  
## PCIP15 Confirmed  
## PCIP16 Confirmed  
## PCIP19 Confirmed  
## PCIP22 Tentative  
## PCIP23 Confirmed  
## PCIP24 Confirmed  
## PCIP25 Rejected  
## PCIP26 Confirmed  
## PCIP27 Confirmed  
## PCIP29 Rejected  
## PCIP30 Confirmed  
## PCIP31 Confirmed  
## PCIP38 Confirmed  
## PCIP39 Confirmed  
## PCIP40 Confirmed  
## PCIP41 Confirmed  
## PCIP42 Confirmed  
## PCIP43 Confirmed  
## PCIP44 Confirmed  
## PCIP45 Confirmed  
## PCIP46 Rejected  
## PCIP47 Rejected  
## PCIP48 Rejected  
## PCIP49 Confirmed  
## PCIP50 Confirmed  
## PCIP51 Confirmed  
## PCIP52 Confirmed  
## PCIP54 Confirmed  
## UGDS\_WHITE Confirmed  
## UGDS\_BLACK Confirmed  
## UGDS\_HISP Confirmed  
## UGDS\_ASIAN Confirmed  
## UGDS\_AIAN Confirmed  
## UGDS\_NHPI Confirmed  
## UGDS\_2MOR Confirmed  
## UGDS\_NRA Confirmed  
## UGDS\_UNKN Confirmed  
## PPTUG\_EF Confirmed  
## COSTT4\_A Confirmed  
## TUITIONFEE\_IN Confirmed  
## TUITIONFEE\_OUT Confirmed  
## C150\_4 Confirmed  
## C150\_4\_WHITE Confirmed  
## C150\_4\_BLACK Confirmed  
## C150\_4\_HISP Confirmed  
## C150\_4\_ASIAN Confirmed  
## C150\_4\_AIAN Confirmed  
## C150\_4\_NHPI Rejected  
## C150\_4\_2MOR Confirmed  
## C150\_4\_NRA Confirmed  
## C150\_4\_UNKN Confirmed  
## RET\_FT4 Confirmed  
## PCTFLOAN Confirmed  
## PAR\_ED\_PCT\_1STGEN Confirmed  
## UGDS\_MEN Confirmed  
## UGDS\_WOMEN Confirmed

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



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

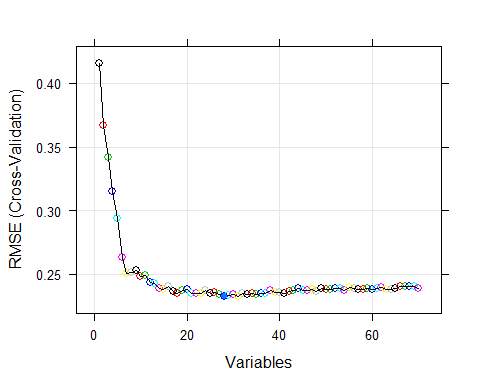
##   
## Attaching package: 'plyr'

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

predictors(rfe.train)

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

# 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   
## -0.62863 -0.05922 0.01180 0.07523 0.49376   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.011e-01 4.708e-02 19.142 < 2e-16 \*\*\*  
## REGION -2.592e-03 3.349e-03 -0.774 0.43938   
## ADM\_RATE\_ALL -1.081e-01 3.894e-02 -2.776 0.00568 \*\*   
## UGDS\_NRA 2.053e-01 1.563e-01 1.313 0.18968   
## PPTUG\_EF -4.491e-01 8.893e-02 -5.050 5.86e-07 \*\*\*  
## COSTT4\_A 1.978e-06 6.404e-07 3.088 0.00210 \*\*   
## PCTFLOAN -3.611e-01 6.087e-02 -5.932 5.04e-09 \*\*\*  
## PAR\_ED\_PCT\_1STGEN -7.244e-02 1.043e-01 -0.695 0.48751   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1441 on 603 degrees of freedom  
## Multiple R-squared: 0.4218, Adjusted R-squared: 0.4151   
## F-statistic: 62.84 on 7 and 603 DF, p-value: < 2.2e-16

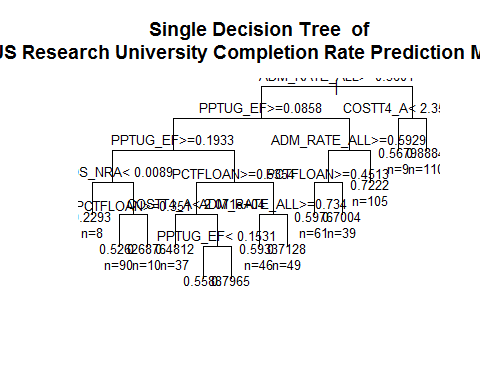
# do the testing with the prediction model  
accepted\_ind3 <- predict(lm\_NRAcompletion, interval="prediction", newdata = univ\_test2)  
  
# Checking on PRED(25)  
errors <- accepted\_ind3[,"fit"] - univ\_test2$C150\_4\_NRA  
rel\_change <- abs(errors) / univ\_test2$C150\_4\_NRA  
table(rel\_change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

## TRUE   
## 0.7892157

# 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 0.26806025 0 1.0000000 1.0013136 0.06615839  
## 2 0.06421962 1 0.7319398 0.7715956 0.05877302  
## 3 0.03946564 2 0.6677201 0.7429201 0.05862012  
## 4 0.02843285 3 0.6282545 0.7202865 0.05727949  
## 5 0.02263047 5 0.5713888 0.7053863 0.05795586  
## 6 0.01689051 6 0.5487583 0.6878506 0.05687233  
## 7 0.01674966 7 0.5318678 0.6714101 0.05457408  
## 8 0.01564732 9 0.4983685 0.6677493 0.05351844  
## 9 0.01162417 10 0.4827212 0.6496380 0.05203788  
## 10 0.01082797 11 0.4710970 0.6690445 0.05546656  
## 11 0.01000000 12 0.4602691 0.6724556 0.05510621  
##   
## Variable importance  
## ADM\_RATE\_ALL PPTUG\_EF COSTT4\_A PAR\_ED\_PCT\_1STGEN   
## 26 19 18 14   
## PCTFLOAN UGDS\_NRA REGION   
## 13 8 2   
##   
## Node number 1: 611 observations, complexity param=0.2680602  
## mean=0.6659948, MSE=0.0354359   
## left son=2 (492 obs) right son=3 (119 obs)  
## Primary splits:  
## ADM\_RATE\_ALL < 0.3601335 to the right, improve=0.2680602, (0 missing)  
## PPTUG\_EF < 0.0607 to the right, improve=0.2517726, (0 missing)  
## COSTT4\_A < 49172 to the left, improve=0.2396818, (0 missing)  
## PAR\_ED\_PCT\_1STGEN < 0.2093101 to the right, improve=0.2063361, (0 missing)  
## UGDS\_NRA < 0.0577 to the left, improve=0.2021123, (0 missing)  
## Surrogate splits:  
## COSTT4\_A < 53939 to the left, agree=0.902, adj=0.496, (0 split)  
## PPTUG\_EF < 0.01375 to the right, agree=0.889, adj=0.429, (0 split)  
## PAR\_ED\_PCT\_1STGEN < 0.1802756 to the right, agree=0.882, adj=0.395, (0 split)  
## PCTFLOAN < 0.311 to the right, agree=0.879, adj=0.378, (0 split)  
## UGDS\_NRA < 0.09355 to the left, agree=0.815, adj=0.050, (0 split)  
##   
## Node number 2: 492 observations, complexity param=0.06421962  
## mean=0.6180624, MSE=0.02933659   
## left son=4 (287 obs) right son=5 (205 obs)  
## Primary splits:  
## PPTUG\_EF < 0.0858 to the right, improve=0.09633357, (0 missing)  
## COSTT4\_A < 20346.5 to the left, improve=0.07798452, (0 missing)  
## UGDS\_NRA < 0.0579 to the left, improve=0.06981951, (0 missing)  
## PAR\_ED\_PCT\_1STGEN < 0.2090995 to the right, improve=0.06499428, (0 missing)  
## PCTFLOAN < 0.5341 to the right, improve=0.06007770, (0 missing)  
## Surrogate splits:  
## PAR\_ED\_PCT\_1STGEN < 0.2794355 to the right, agree=0.776, adj=0.463, (0 split)  
## COSTT4\_A < 23813.5 to the left, agree=0.746, adj=0.390, (0 split)  
## ADM\_RATE\_ALL < 0.5834978 to the right, agree=0.689, adj=0.254, (0 split)  
## UGDS\_NRA < 0.0562 to the left, agree=0.681, adj=0.234, (0 split)  
## PCTFLOAN < 0.49 to the right, agree=0.608, adj=0.059, (0 split)  
##   
## Node number 3: 119 observations, complexity param=0.03946564  
## mean=0.8641689, MSE=0.01188125   
## left son=6 (9 obs) right son=7 (110 obs)  
## Primary splits:  
## COSTT4\_A < 23579.5 to the left, improve=0.6043585, (0 missing)  
## PPTUG\_EF < 0.0897 to the right, improve=0.4024565, (0 missing)  
## PAR\_ED\_PCT\_1STGEN < 0.3615154 to the right, improve=0.3555563, (0 missing)  
## ADM\_RATE\_ALL < 0.2721 to the right, improve=0.3294809, (0 missing)  
## UGDS\_NRA < 0.0491 to the left, improve=0.3028274, (0 missing)  
## Surrogate splits:  
## UGDS\_NRA < 0.02795 to the left, agree=0.958, adj=0.444, (0 split)  
## PAR\_ED\_PCT\_1STGEN < 0.3795346 to the right, agree=0.958, adj=0.444, (0 split)  
## PPTUG\_EF < 0.0897 to the right, agree=0.941, adj=0.222, (0 split)  
##   
## Node number 4: 287 observations, complexity param=0.02843285  
## mean=0.5731331, MSE=0.031027   
## left son=8 (108 obs) right son=9 (179 obs)  
## Primary splits:  
## PPTUG\_EF < 0.19335 to the right, improve=0.05676054, (0 missing)  
## ADM\_RATE\_ALL < 0.4408 to the left, improve=0.04945094, (0 missing)  
## UGDS\_NRA < 0.01155 to the left, improve=0.04933555, (0 missing)  
## PAR\_ED\_PCT\_1STGEN < 0.3460932 to the right, improve=0.04226952, (0 missing)  
## COSTT4\_A < 19814.5 to the left, improve=0.03964047, (0 missing)  
## Surrogate splits:  
## PAR\_ED\_PCT\_1STGEN < 0.3807375 to the right, agree=0.774, adj=0.398, (0 split)  
## ADM\_RATE\_ALL < 0.5844877 to the left, agree=0.697, adj=0.194, (0 split)  
## COSTT4\_A < 16900 to the left, agree=0.662, adj=0.102, (0 split)  
## PCTFLOAN < 0.33185 to the left, agree=0.634, adj=0.028, (0 split)  
##   
## Node number 5: 205 observations, complexity param=0.01689051  
## mean=0.6809634, MSE=0.02018738   
## left son=10 (100 obs) right son=11 (105 obs)  
## Primary splits:  
## ADM\_RATE\_ALL < 0.5929458 to the right, improve=0.08836771, (0 missing)  
## PCTFLOAN < 0.4147 to the right, improve=0.07206288, (0 missing)  
## UGDS\_NRA < 0.04235 to the left, improve=0.05624424, (0 missing)  
## PPTUG\_EF < 0.0607 to the right, improve=0.05070259, (0 missing)  
## PAR\_ED\_PCT\_1STGEN < 0.1876551 to the right, improve=0.04933058, (0 missing)  
## Surrogate splits:  
## UGDS\_NRA < 0.0389 to the left, agree=0.683, adj=0.35, (0 split)  
## COSTT4\_A < 28918.5 to the left, agree=0.673, adj=0.33, (0 split)  
## PPTUG\_EF < 0.05325 to the right, agree=0.659, adj=0.30, (0 split)  
## PAR\_ED\_PCT\_1STGEN < 0.2082203 to the right, agree=0.605, adj=0.19, (0 split)  
## PCTFLOAN < 0.48375 to the right, agree=0.585, adj=0.15, (0 split)  
##   
## Node number 6: 9 observations  
## mean=0.5679222, MSE=0.01699388   
##   
## Node number 7: 110 observations  
## mean=0.8884073, MSE=0.003694912   
##   
## Node number 8: 108 observations, complexity param=0.02843285  
## mean=0.5191065, MSE=0.02981413   
## left son=16 (8 obs) right son=17 (100 obs)  
## Primary splits:  
## UGDS\_NRA < 0.0089 to the left, improve=0.22540260, (0 missing)  
## PCTFLOAN < 0.351 to the right, improve=0.09710919, (0 missing)  
## ADM\_RATE\_ALL < 0.4497 to the left, improve=0.09697429, (0 missing)  
## PAR\_ED\_PCT\_1STGEN < 0.3425707 to the right, improve=0.07995585, (0 missing)  
## REGION < 6.5 to the left, improve=0.06081527, (0 missing)  
##   
## Node number 9: 179 observations, complexity param=0.02263047  
## mean=0.6057302, MSE=0.02893511   
## left son=18 (84 obs) right son=19 (95 obs)  
## Primary splits:  
## PCTFLOAN < 0.53545 to the right, improve=0.09460192, (0 missing)  
## UGDS\_NRA < 0.02905 to the left, improve=0.05425596, (0 missing)  
## PPTUG\_EF < 0.18295 to the left, improve=0.05121529, (0 missing)  
## ADM\_RATE\_ALL < 0.67635 to the right, improve=0.04467173, (0 missing)  
## COSTT4\_A < 19189 to the left, improve=0.02715591, (0 missing)  
## Surrogate splits:  
## REGION < 5.5 to the left, agree=0.654, adj=0.262, (0 split)  
## UGDS\_NRA < 0.0275 to the left, agree=0.637, adj=0.226, (0 split)  
## PAR\_ED\_PCT\_1STGEN < 0.3352895 to the right, agree=0.603, adj=0.155, (0 split)  
## COSTT4\_A < 21992 to the right, agree=0.575, adj=0.095, (0 split)  
## ADM\_RATE\_ALL < 0.8898 to the right, agree=0.547, adj=0.036, (0 split)  
##   
## Node number 10: 100 observations, complexity param=0.01162417  
## mean=0.637684, MSE=0.02321523   
## left son=20 (61 obs) right son=21 (39 obs)  
## Primary splits:  
## PCTFLOAN < 0.4513 to the right, improve=0.10841110, (0 missing)  
## PPTUG\_EF < 0.05775 to the right, improve=0.05460045, (0 missing)  
## PAR\_ED\_PCT\_1STGEN < 0.209449 to the right, improve=0.05176826, (0 missing)  
## COSTT4\_A < 20348 to the left, improve=0.03594446, (0 missing)  
## UGDS\_NRA < 0.01495 to the left, improve=0.02767774, (0 missing)  
## Surrogate splits:  
## UGDS\_NRA < 0.05695 to the left, agree=0.75, adj=0.359, (0 split)  
## REGION < 5.5 to the left, agree=0.65, adj=0.103, (0 split)  
## COSTT4\_A < 49384.5 to the left, agree=0.64, adj=0.077, (0 split)  
## PAR\_ED\_PCT\_1STGEN < 0.1318897 to the right, agree=0.64, adj=0.077, (0 split)  
## ADM\_RATE\_ALL < 0.596237 to the right, agree=0.63, adj=0.051, (0 split)  
##   
## Node number 11: 105 observations  
## mean=0.7221819, MSE=0.01382083   
##   
## Node number 16: 8 observations  
## mean=0.229275, MSE=0.03988229   
##   
## Node number 17: 100 observations, complexity param=0.01082797  
## mean=0.542293, MSE=0.02175088   
## left son=34 (90 obs) right son=35 (10 obs)  
## Primary splits:  
## PCTFLOAN < 0.351 to the right, improve=0.10778410, (0 missing)  
## PPTUG\_EF < 0.2085 to the left, improve=0.10617690, (0 missing)  
## PAR\_ED\_PCT\_1STGEN < 0.432151 to the right, improve=0.08220064, (0 missing)  
## REGION < 6.5 to the left, improve=0.05454218, (0 missing)  
## ADM\_RATE\_ALL < 0.524178 to the left, improve=0.03864783, (0 missing)  
##   
## Node number 18: 84 observations, complexity param=0.01674966  
## mean=0.5500905, MSE=0.03551997   
## left son=36 (37 obs) right son=37 (47 obs)  
## Primary splits:  
## COSTT4\_A < 20712.5 to the left, improve=0.10529960, (0 missing)  
## UGDS\_NRA < 0.01145 to the left, improve=0.09959870, (0 missing)  
## PPTUG\_EF < 0.123 to the left, improve=0.07488862, (0 missing)  
## PAR\_ED\_PCT\_1STGEN < 0.2848234 to the left, improve=0.04365379, (0 missing)  
## ADM\_RATE\_ALL < 0.593539 to the right, improve=0.03834943, (0 missing)  
## Surrogate splits:  
## REGION < 3.5 to the right, agree=0.833, adj=0.622, (0 split)  
## UGDS\_NRA < 0.0194 to the left, agree=0.726, adj=0.378, (0 split)  
## ADM\_RATE\_ALL < 0.7089247 to the left, agree=0.631, adj=0.162, (0 split)  
## PPTUG\_EF < 0.15765 to the right, agree=0.631, adj=0.162, (0 split)  
## PCTFLOAN < 0.6003 to the left, agree=0.619, adj=0.135, (0 split)  
##   
## Node number 19: 95 observations, complexity param=0.01564732  
## mean=0.6549274, MSE=0.01795503   
## left son=38 (46 obs) right son=39 (49 obs)  
## Primary splits:  
## ADM\_RATE\_ALL < 0.7339946 to the right, improve=0.19861630, (0 missing)  
## PPTUG\_EF < 0.18295 to the left, improve=0.11541890, (0 missing)  
## PCTFLOAN < 0.5232 to the left, improve=0.05288903, (0 missing)  
## PAR\_ED\_PCT\_1STGEN < 0.3153789 to the left, improve=0.04250287, (0 missing)  
## COSTT4\_A < 21304.5 to the right, improve=0.04250232, (0 missing)  
## Surrogate splits:  
## REGION < 5.5 to the right, agree=0.726, adj=0.435, (0 split)  
## PAR\_ED\_PCT\_1STGEN < 0.2940859 to the right, agree=0.632, adj=0.239, (0 split)  
## PPTUG\_EF < 0.15445 to the right, agree=0.611, adj=0.196, (0 split)  
## PCTFLOAN < 0.47775 to the right, agree=0.611, adj=0.196, (0 split)  
## COSTT4\_A < 17666 to the right, agree=0.579, adj=0.130, (0 split)  
##   
## Node number 20: 61 observations  
## mean=0.5975705, MSE=0.02735313   
##   
## Node number 21: 39 observations  
## mean=0.7004256, MSE=0.01028981   
##   
## Node number 34: 90 observations  
## mean=0.5261533, MSE=0.02091026   
##   
## Node number 35: 10 observations  
## mean=0.68755, MSE=0.005872475   
##   
## Node number 36: 37 observations  
## mean=0.4811622, MSE=0.02958379   
##   
## Node number 37: 47 observations, complexity param=0.01674966  
## mean=0.6043532, MSE=0.03350846   
## left son=74 (38 obs) right son=75 (9 obs)  
## Primary splits:  
## PPTUG\_EF < 0.15305 to the left, improve=0.26104860, (0 missing)  
## UGDS\_NRA < 0.029 to the left, improve=0.17027260, (0 missing)  
## REGION < 5.5 to the left, improve=0.14275790, (0 missing)  
## COSTT4\_A < 22840 to the right, improve=0.09692777, (0 missing)  
## PAR\_ED\_PCT\_1STGEN < 0.2972428 to the left, improve=0.07068106, (0 missing)  
## Surrogate splits:  
## ADM\_RATE\_ALL < 0.9046185 to the left, agree=0.851, adj=0.222, (0 split)  
## REGION < 5.5 to the left, agree=0.830, adj=0.111, (0 split)  
## PAR\_ED\_PCT\_1STGEN < 0.37795 to the left, agree=0.830, adj=0.111, (0 split)  
##   
## Node number 38: 46 observations  
## mean=0.5932935, MSE=0.01235694   
##   
## Node number 39: 49 observations  
## mean=0.7127878, MSE=0.01629639   
##   
## Node number 74: 38 observations  
## mean=0.5588368, MSE=0.02411135   
##   
## Node number 75: 9 observations  
## mean=0.7965333, MSE=0.0275046

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)



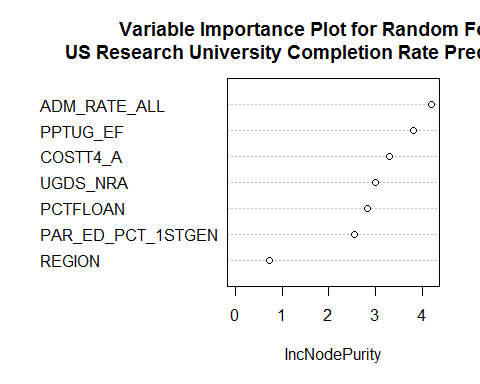
pred\_dtree3 <- predict(model\_dtree3, newdata = univ\_test2)  
accu11 <- abs(pred\_dtree3 - univ\_test2$C150\_4\_NRA) < 0.25  
frac11 <- sum(accu11)/length(accu11)  
print(frac11)

## [1] 0.9313725

# 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 Prediciton 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] 0.9460784

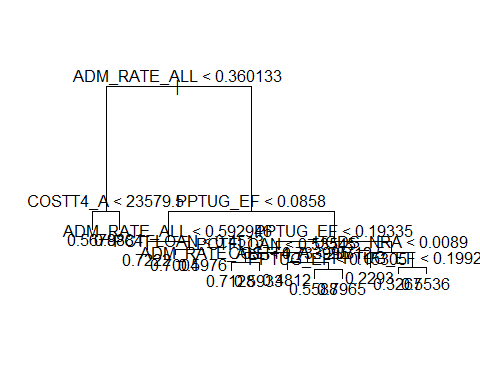
# Doing support vector machine  
model\_svm3 <- svm(formula\_completionrate, data = univ\_train2)  
summary(model\_svm3)

##   
## 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: 502

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 Prediciton Model")  
text(model\_tree3)



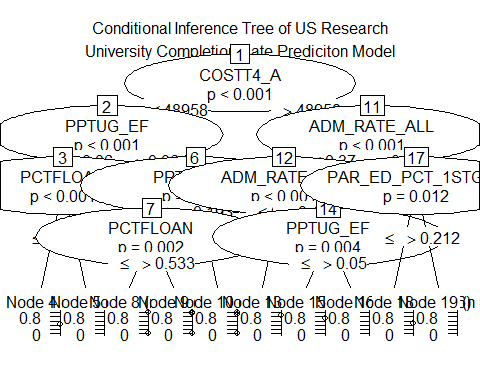
pred\_tree3 <- predict(model\_tree3, newdata = univ\_test2)  
accu14 <- abs(pred\_tree3 - univ\_test2$C150\_4\_NRA) < 0.25  
frac14 <- sum(accu14)/length(accu14)  
print(frac14)

## [1] 0.9166667

# 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 Prediciton 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] 0.9362745

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