US Research University Prediction Model

Philip Gabriel Andrada

November 15, 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

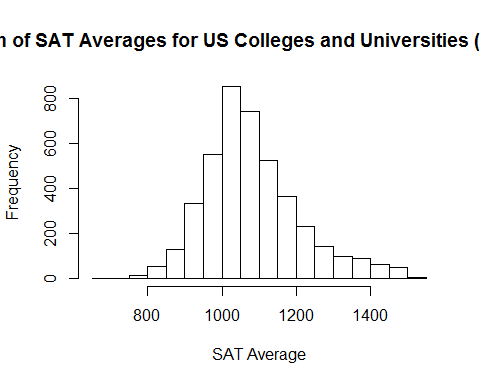
## [1] 815

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

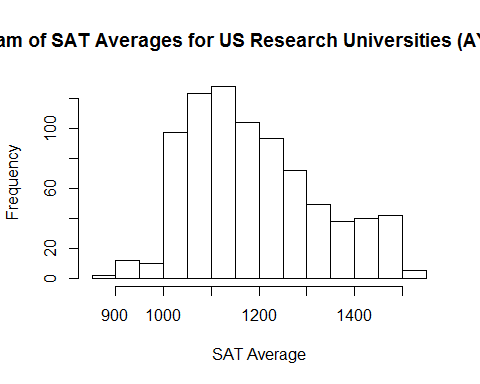
## [1] 0.7062392

# Distributions and Box and Whisker Plots

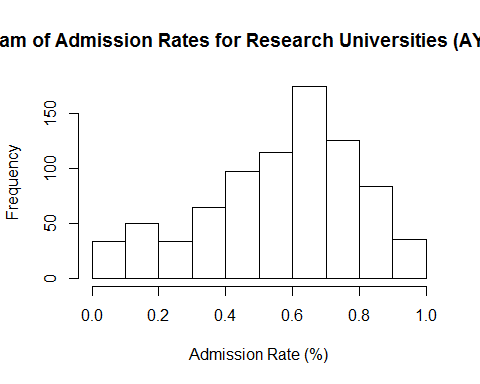
# Histogram of SAT Averages for US Colleges and Universities  
hist(usunivfilter$SAT\_AVG\_ALL, main = "Histogram of SAT Averages for US Colleges and Universities (AY2010-2015)", xlab="SAT Average")



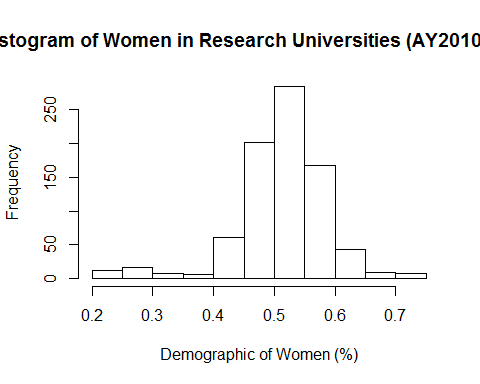
# Histogram of SAT Averages for US Research Universities  
hist(usresearchuniv$SAT\_AVG\_ALL, main = "Histogram of SAT Averages for US Research Universities (AY2010-2015)", xlab="SAT Average")



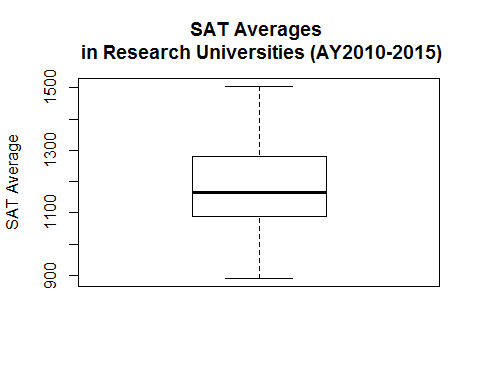
# Histogram of Admission Rates for US Research Universities  
hist(usresearchuniv$ADM\_RATE\_ALL, main = "Histogram of Admission Rates for Research Universities (AY2010-2015)", xlab = "Admission Rate (%)")



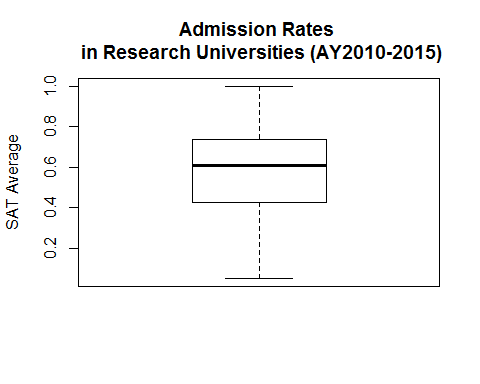
# Histogram of Women in US Research Universities  
hist(usresearchuniv$UGDS\_WOMEN, main = "Histogram of Women in Research Universities (AY2010-2015)", xlab = "Demographic of Women (%)")



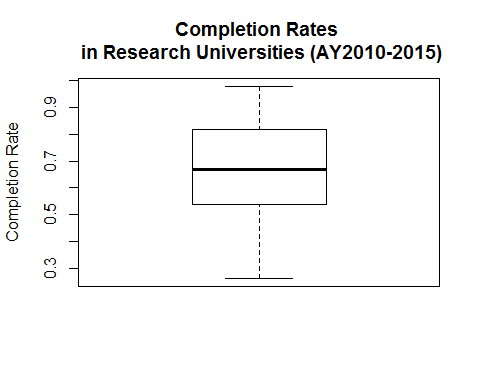
#Boxplot of SAT Average in all US Research Universities  
boxplot(usresearchuniv$SAT\_AVG\_ALL, main = "SAT Averages \n in Research Universities (AY2010-2015)", ylab = "SAT Average")



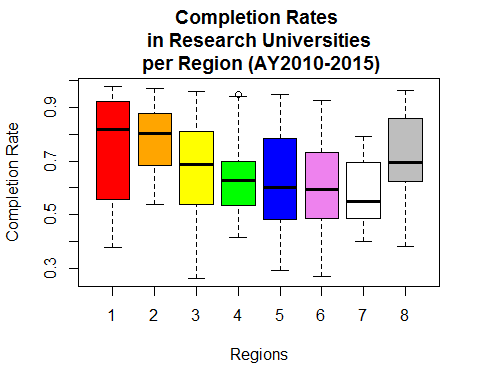
#Boxplot of admission rates in all US Research Universities  
boxplot(usresearchuniv$ADM\_RATE\_ALL, main = "Admission Rates \n in Research Universities (AY2010-2015)", ylab = "SAT Average")



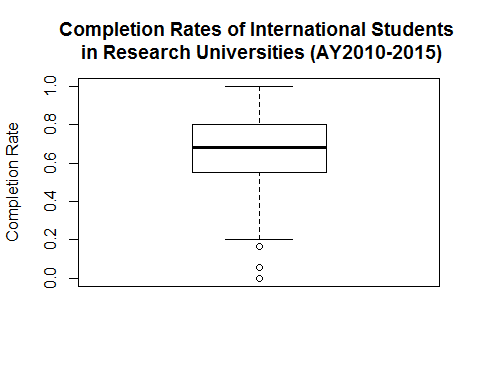
#Boxplot of Completion Rates in all US Research Universities  
boxplot(usresearchuniv$C150\_4, main = "Completion Rates \n in Research Universities (AY2010-2015)", ylab = "Completion Rate")



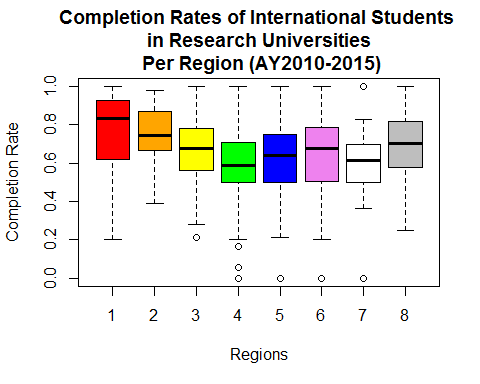
# Boxplot of Completion Rates per Region in US Research Universities  
boxplot(C150\_4 ~ REGION, usresearchuniv, main = "Completion Rates \n in Research Universities \n per Region (AY2010-2015)", col=c("red", "orange", "yellow", "green", "blue", "violet", "white", "gray", "magenta"), ylab = "Completion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

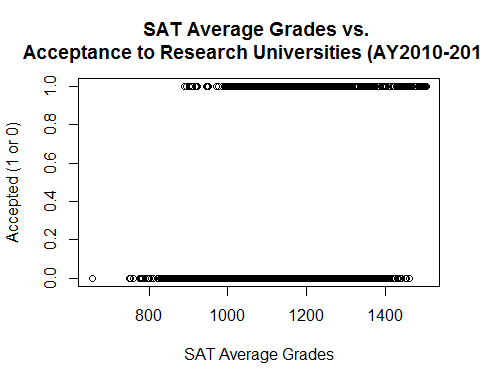


nrow(usresearchuniv[usresearchuniv$C150\_4\_NRA < 0.2,])

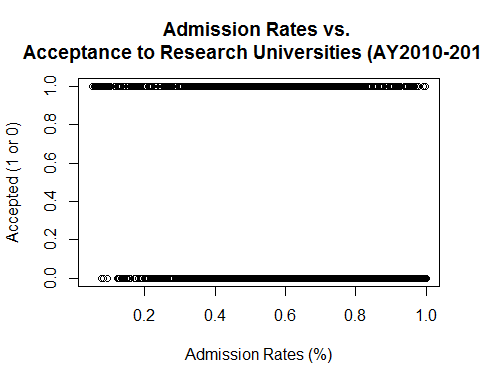
## [1] 9

# Correlations

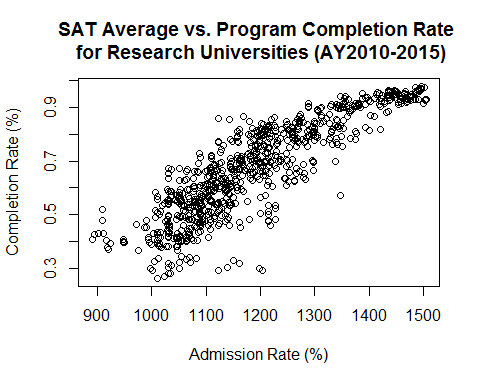
#Correlation between the SAT grades and the acceptance for the research universities  
plot(usunivfilter$SAT\_AVG\_ALL, usunivfilter$ACCEPTED, main="SAT Average Grades vs. \n Acceptance to Research Universities (AY2010-2015)", xlab="SAT Average Grades", ylab="Accepted (1 or 0)")



#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 (%)")

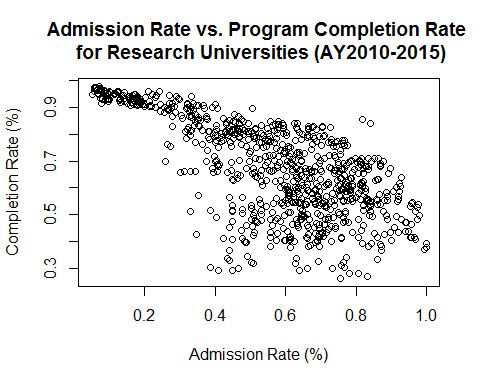


#Correlation coefficient between admission rate and completion rate  
cor(usresearchuniv$SAT\_AVG\_ALL, usresearchuniv$C150\_4, method = "pearson")

## [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 (%)")

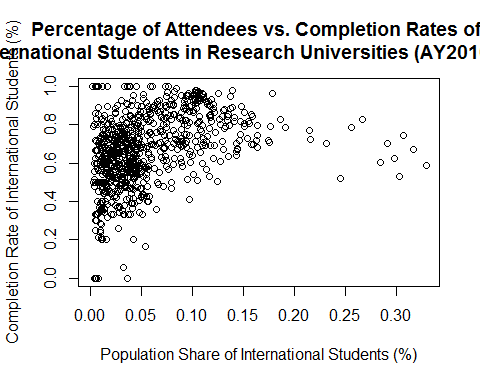


#Correlation coefficient between admission rate and completion rate  
cor(usresearchuniv$ADM\_RATE\_ALL, usresearchuniv$C150\_4, method = "pearson")

## [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 (%)")

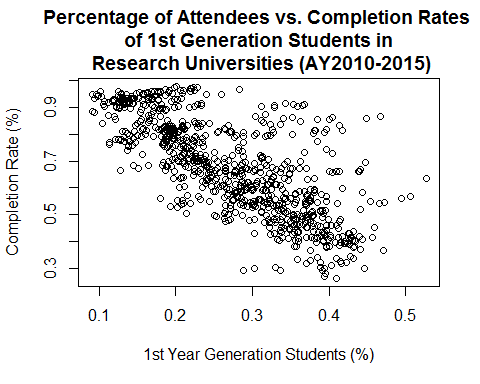


#Correlation coefficient between admission rate and completion rate of international students  
cor(usresearchuniv$UGDS\_NRA, usresearchuniv$C150\_4\_NRA, method = "pearson")

## [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   
## -2.2950 -0.5406 -0.2927 -0.1136 2.8133   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.458e+01 1.186e+00 -12.291 < 2e-16 \*\*\*  
## REGION 1.298e-01 2.956e-02 4.392 1.12e-05 \*\*\*  
## ADM\_RATE\_ALL 6.044e-01 3.807e-01 1.588 0.112   
## SAT\_AVG\_ALL 1.455e-02 8.348e-04 17.435 < 2e-16 \*\*\*  
## UGDS\_NRA 7.328e+00 1.300e+00 5.637 1.73e-08 \*\*\*  
## COSTT4\_A -9.519e-05 6.314e-06 -15.074 < 2e-16 \*\*\*  
## PCTFLOAN -6.484e-01 4.988e-01 -1.300 0.194   
## UGDS\_WOMEN -2.101e+00 5.389e-01 -3.899 9.64e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3154.0 on 3184 degrees of freedom  
## Residual deviance: 2143.3 on 3177 degrees of freedom  
## AIC: 2159.3  
##   
## 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  
## 0 835 116  
## 1 37 74

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

## Accuracy   
## 0.8559322

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

## Neg Pred Value   
## 0.6666667

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

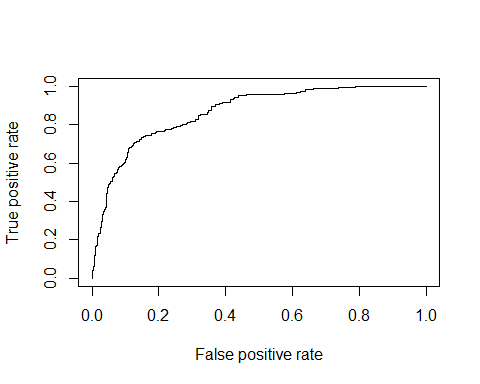
## Specificity   
## 0.3894737

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)  
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] 0.873823

# 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?

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] 0.9416196

# doing support vector machine  
model\_svm1 <- svm(formula\_ISAcceptance, data = univ\_train)  
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] 0.8926554

# doing simple tree  
model\_tree1 <- tree(formula\_ISAcceptance, data = univ\_train)  
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] 0.873823

# doing conditional inference tree  
model\_party1 <- ctree(formula\_ISAcceptance, data = univ\_train)  
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] 0.8775895

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   
## -2.52986 -0.48841 -0.24610 -0.08274 3.02775   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.676e+01 1.414e+00 -11.850 < 2e-16 \*\*\*  
## REGION 1.317e-01 3.169e-02 4.156 3.23e-05 \*\*\*  
## ADM\_RATE\_ALL 9.104e-01 4.173e-01 2.181 0.029154 \*   
## SAT\_AVG\_ALL 1.503e-02 9.999e-04 15.036 < 2e-16 \*\*\*  
## PCIP11 8.879e-01 1.950e+00 0.455 0.648915   
## PCIP12 -2.238e+00 1.981e+01 -0.113 0.910057   
## PCIP14 5.192e+00 7.751e-01 6.698 2.11e-11 \*\*\*  
## PCIP15 -1.392e+00 2.372e+00 -0.587 0.557468   
## PCIP24 -5.818e+00 1.233e+00 -4.718 2.38e-06 \*\*\*  
## PCIP26 5.951e+00 1.726e+00 3.448 0.000565 \*\*\*  
## PCIP27 -2.648e+01 6.823e+00 -3.880 0.000104 \*\*\*  
## PCIP40 -2.699e+01 4.543e+00 -5.942 2.81e-09 \*\*\*  
## PCIP45 7.920e+00 1.199e+00 6.604 3.99e-11 \*\*\*  
## PCIP51 2.085e+00 6.060e-01 3.442 0.000578 \*\*\*  
## PCIP52 4.288e-01 6.588e-01 0.651 0.515132   
## UGDS\_NRA 9.774e+00 1.468e+00 6.659 2.77e-11 \*\*\*  
## UGDS\_UNKN -1.003e+00 1.556e+00 -0.645 0.519216   
## COSTT4\_A -1.119e-04 7.269e-06 -15.401 < 2e-16 \*\*\*  
## PCTFLOAN -4.320e-01 5.679e-01 -0.761 0.446828   
## UGDS\_WOMEN 8.248e-02 7.795e-01 0.106 0.915734   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3154 on 3184 degrees of freedom  
## Residual deviance: 1926 on 3165 degrees of freedom  
## AIC: 1966  
##   
## 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  
## 0 841 99  
## 1 31 91

c2$overall['Accuracy']

## Accuracy   
## 0.8775895

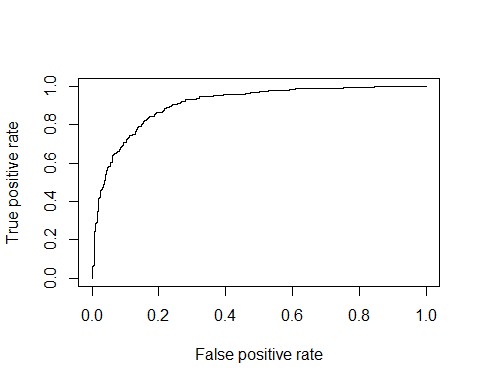
#Precision of the logistic regression model  
c2$byClass['Neg Pred Value']

## Neg Pred Value   
## 0.7459016

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

## Specificity   
## 0.4789474

# 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)  
pred\_dtree2 <- predict(model\_dtree2, newdata = univ\_test)  
accu6 <- abs(pred\_dtree2 - univ\_test$ACCEPTED) < 0.5  
frac6 <- sum(accu6)/length(accu6)  
print(frac6)

## [1] 0.9067797

# doing random forest  
model\_forest2 <- randomForest(formula\_ISSciAcceptance, data = univ\_train)

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

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] 0.9670433

# doing support vector machine  
model\_svm2 <- svm(formula\_ISSciAcceptance, data = univ\_train)  
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] 0.9227872

# doing simple tree  
model\_tree2 <- tree(formula\_ISSciAcceptance, data = univ\_train)  
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] 0.9067797

# doing conditional inference tree  
model\_party2 <- ctree(formula\_ISSciAcceptance, data = univ\_train)  
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] 0.9048964

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

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

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

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

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

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

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

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

## Boruta performed 99 iterations in 26.39764 secs.  
## 60 attributes confirmed important: ADM\_RATE, ADM\_RATE\_ALL,  
## C150\_4, C150\_4\_AIAN, C150\_4\_ASIAN and 55 more.  
## 7 attributes confirmed unimportant: C150\_4\_NHPI, PCIP12, PCIP25,  
## PCIP29, PCIP46 and 2 more.  
## 3 tentative attributes left: C150\_4\_2MOR, PCIP10, PCIP22.

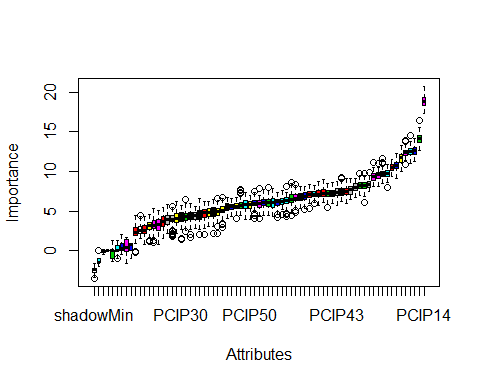
getSelectedAttributes(boruta.train)

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

# 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 5.5269389 5.510555613 4.0672209 6.746214 1.0000000  
## ADM\_RATE 7.1805443 7.125331145 5.7834063 8.767147 1.0000000  
## ADM\_RATE\_ALL 7.2377345 7.261082300 5.4339045 8.562596 1.0000000  
## SAT\_AVG\_ALL 12.6189537 12.704846238 11.3115566 14.152091 1.0000000  
## PCIP01 6.1934711 6.242209734 4.6319094 7.571458 1.0000000  
## PCIP03 6.7075429 6.669541140 4.9097406 8.358227 1.0000000  
## PCIP04 11.7340949 11.731580500 10.0667085 13.321824 1.0000000  
## PCIP05 8.4228505 8.431654565 7.5372230 9.883665 1.0000000  
## PCIP09 4.8645798 4.887578352 2.2561943 6.465213 0.9898990  
## PCIP10 2.6390290 2.640956575 0.8307419 4.285870 0.4949495  
## PCIP11 6.5926515 6.517371661 4.4176297 8.612306 1.0000000  
## PCIP12 0.5212074 0.282394290 -0.4370785 2.028440 0.0000000  
## PCIP13 6.0114780 6.107352732 4.3635068 7.871953 1.0000000  
## PCIP14 18.8383279 18.832251137 17.3481116 20.743977 1.0000000  
## PCIP15 4.8231243 4.855207631 3.0857547 6.295388 0.9797980  
## PCIP16 7.6237443 7.739206780 6.4644383 9.168934 1.0000000  
## PCIP19 7.5068607 7.446903041 6.2090005 9.311529 1.0000000  
## PCIP22 2.4852663 2.392042939 -0.1594531 4.437363 0.4545455  
## PCIP23 8.2566877 8.247985838 6.1238757 9.730715 1.0000000  
## PCIP24 5.9993766 6.015963131 4.1276886 7.466202 0.9898990  
## PCIP25 -1.2933294 -1.415782155 -2.0075835 0.000000 0.0000000  
## PCIP26 5.9746237 6.020256429 4.0692449 7.830207 1.0000000  
## PCIP27 5.1414173 5.099130356 2.9425979 6.759870 0.9797980  
## PCIP29 0.0000000 0.000000000 0.0000000 0.000000 0.0000000  
## PCIP30 4.2554778 4.306483207 1.4361826 5.746038 0.9595960  
## PCIP31 4.7823011 4.816080446 2.1053710 6.189775 0.9797980  
## PCIP38 4.2780595 4.330124636 2.4576521 6.501061 0.9595960  
## PCIP39 5.5282093 5.617760403 4.3097042 6.671240 1.0000000  
## PCIP40 5.8324209 5.808155334 4.0111389 7.343603 0.9898990  
## PCIP41 3.2795455 3.296516235 0.8477233 5.133030 0.7171717  
## PCIP42 4.9012972 4.916078976 2.2630570 6.606747 0.9797980  
## PCIP43 7.3663528 7.418684057 5.7155402 8.776034 1.0000000  
## PCIP44 4.4962901 4.563129023 2.1350496 6.655892 0.9393939  
## PCIP45 7.5127420 7.472107622 6.0213162 9.146786 1.0000000  
## PCIP46 -0.2356178 0.000000000 -1.3662007 1.338010 0.0000000  
## PCIP47 0.3642691 0.842034898 -1.0010015 1.001002 0.0000000  
## PCIP48 0.1190946 0.004221514 -1.0010015 1.320408 0.0000000  
## PCIP49 3.2584921 3.228649371 1.1130735 5.335159 0.7575758  
## PCIP50 5.8024633 5.813891846 3.9369749 6.999662 0.9898990  
## PCIP51 3.9993394 3.941730047 2.1804918 5.907579 0.9191919  
## PCIP52 9.7363004 9.677719526 8.5360151 11.613814 1.0000000  
## PCIP54 3.6787062 3.772669880 1.5165488 5.508369 0.8282828  
## UGDS\_WHITE 8.2392186 8.216904747 6.9608331 9.827593 1.0000000  
## UGDS\_BLACK 10.7988571 10.832958797 9.1835939 12.276183 1.0000000  
## UGDS\_HISP 6.2999137 6.308632034 4.3129830 8.168312 1.0000000  
## UGDS\_ASIAN 9.3210507 9.368840237 8.1266744 11.199288 1.0000000  
## UGDS\_AIAN 4.1792034 4.088007905 2.2511322 6.286778 0.9292929  
## UGDS\_NHPI 3.9769698 4.024303231 1.7795875 5.579967 0.8989899  
## UGDS\_2MOR 4.3597001 4.445176325 1.7226476 5.968450 0.9292929  
## UGDS\_NRA 7.2819443 7.226026402 6.0958708 8.978152 1.0000000  
## UGDS\_UNKN 6.0588011 6.086472476 4.5318681 7.984769 1.0000000  
## PPTUG\_EF 6.8565582 6.867977636 5.1872759 8.140597 1.0000000  
## COSTT4\_A 9.8030040 9.821931175 7.9542785 10.919788 1.0000000  
## TUITIONFEE\_IN 9.5582799 9.594688856 8.1410799 11.015160 1.0000000  
## TUITIONFEE\_OUT 5.5833043 5.636463182 4.1511166 6.721850 0.9898990  
## C150\_4 8.0146861 7.985240237 6.7403966 9.119211 1.0000000  
## C150\_4\_WHITE 6.7761819 6.714555013 5.3713135 8.181660 1.0000000  
## C150\_4\_BLACK 7.1105654 7.067731216 5.3516642 8.413031 1.0000000  
## C150\_4\_HISP 5.7536856 5.684441898 4.5219874 7.711917 1.0000000  
## C150\_4\_ASIAN 6.0744306 6.175628413 4.2278531 7.334815 1.0000000  
## C150\_4\_AIAN 7.1303455 7.093564557 6.0147092 8.278065 1.0000000  
## C150\_4\_NHPI 0.5528640 0.380959581 -1.5157394 1.750305 0.0000000  
## C150\_4\_2MOR 3.1104781 3.204083198 1.1793716 4.562546 0.6565657  
## C150\_4\_NRA 4.3856526 4.448286780 2.8879907 5.643754 0.9595960  
## C150\_4\_UNKN 7.2387720 7.272178907 6.0394090 8.669905 1.0000000  
## RET\_FT4 10.5303281 10.482001399 9.4138850 11.577750 1.0000000  
## PCTFLOAN 14.1602827 14.147957098 12.6907470 16.474380 1.0000000  
## PAR\_ED\_PCT\_1STGEN 6.1332351 6.070090945 4.4955547 7.706977 1.0000000  
## UGDS\_MEN 12.5384243 12.576847281 11.0714416 14.524958 1.0000000  
## UGDS\_WOMEN 12.4066099 12.378606810 10.9340984 13.896772 1.0000000  
## decision  
## REGION Confirmed  
## ADM\_RATE Confirmed  
## ADM\_RATE\_ALL Confirmed  
## SAT\_AVG\_ALL Confirmed  
## PCIP01 Confirmed  
## PCIP03 Confirmed  
## PCIP04 Confirmed  
## PCIP05 Confirmed  
## PCIP09 Confirmed  
## PCIP10 Tentative  
## PCIP11 Confirmed  
## PCIP12 Rejected  
## PCIP13 Confirmed  
## PCIP14 Confirmed  
## PCIP15 Confirmed  
## PCIP16 Confirmed  
## PCIP19 Confirmed  
## PCIP22 Tentative  
## PCIP23 Confirmed  
## PCIP24 Confirmed  
## PCIP25 Rejected  
## PCIP26 Confirmed  
## PCIP27 Confirmed  
## PCIP29 Rejected  
## PCIP30 Confirmed  
## PCIP31 Confirmed  
## PCIP38 Confirmed  
## PCIP39 Confirmed  
## PCIP40 Confirmed  
## PCIP41 Confirmed  
## PCIP42 Confirmed  
## PCIP43 Confirmed  
## PCIP44 Confirmed  
## PCIP45 Confirmed  
## PCIP46 Rejected  
## PCIP47 Rejected  
## PCIP48 Rejected  
## PCIP49 Confirmed  
## PCIP50 Confirmed  
## PCIP51 Confirmed  
## PCIP52 Confirmed  
## PCIP54 Confirmed  
## UGDS\_WHITE Confirmed  
## UGDS\_BLACK Confirmed  
## UGDS\_HISP Confirmed  
## UGDS\_ASIAN Confirmed  
## UGDS\_AIAN Confirmed  
## UGDS\_NHPI Confirmed  
## UGDS\_2MOR Confirmed  
## UGDS\_NRA Confirmed  
## UGDS\_UNKN Confirmed  
## PPTUG\_EF Confirmed  
## COSTT4\_A Confirmed  
## TUITIONFEE\_IN Confirmed  
## TUITIONFEE\_OUT Confirmed  
## C150\_4 Confirmed  
## C150\_4\_WHITE Confirmed  
## C150\_4\_BLACK Confirmed  
## C150\_4\_HISP Confirmed  
## C150\_4\_ASIAN Confirmed  
## C150\_4\_AIAN Confirmed  
## C150\_4\_NHPI Rejected  
## C150\_4\_2MOR Tentative  
## C150\_4\_NRA Confirmed  
## C150\_4\_UNKN Confirmed  
## RET\_FT4 Confirmed  
## PCTFLOAN Confirmed  
## PAR\_ED\_PCT\_1STGEN Confirmed  
## UGDS\_MEN Confirmed  
## UGDS\_WOMEN Confirmed

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



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

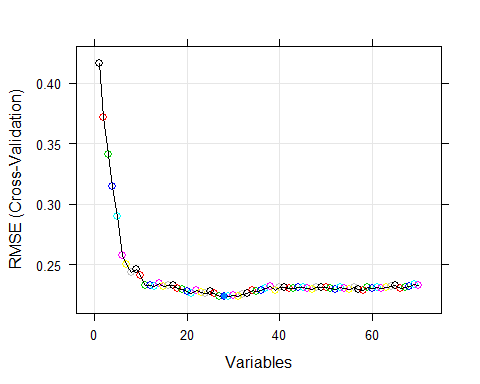
##   
## Attaching package: 'plyr'

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

predictors(rfe.train)

## [1] "PCIP14" "PCTFLOAN" "PCIP04" "SAT\_AVG\_ALL"   
## [5] "PCIP52" "UGDS\_BLACK" "UGDS\_MEN" "PCIP45"   
## [9] "UGDS\_WOMEN" "PCIP43" "PCIP23" "COSTT4\_A"   
## [13] "RET\_FT4" "UGDS\_HISP" "TUITIONFEE\_IN" "C150\_4\_AIAN"   
## [17] "PCIP16" "UGDS\_ASIAN" "PCIP39" "UGDS\_WHITE"   
## [21] "UGDS\_NRA" "C150\_4" "PCIP19" "PCIP24"   
## [25] "PCIP26" "PCIP05" "PCIP03" "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   
## -0.61816 -0.06369 0.01059 0.07862 0.51163   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.307e-01 4.695e-02 19.825 < 2e-16 \*\*\*  
## REGION -4.975e-03 3.377e-03 -1.473 0.141180   
## ADM\_RATE\_ALL -1.378e-01 3.969e-02 -3.473 0.000552 \*\*\*  
## UGDS\_NRA 2.617e-01 1.518e-01 1.725 0.085115 .   
## PPTUG\_EF -3.589e-01 8.865e-02 -4.049 5.83e-05 \*\*\*  
## COSTT4\_A 1.571e-06 6.397e-07 2.456 0.014335 \*   
## PCTFLOAN -3.508e-01 6.082e-02 -5.769 1.28e-08 \*\*\*  
## PAR\_ED\_PCT\_1STGEN -1.084e-01 1.016e-01 -1.067 0.286279   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1451 on 603 degrees of freedom  
## Multiple R-squared: 0.4089, Adjusted R-squared: 0.402   
## F-statistic: 59.59 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   
## 0.8039216

# 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)  
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] 0.9068627

# Doing random forest  
model\_forest3 <- randomForest(formula\_completionrate, data = univ\_train2)  
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] 0.9264706

# Doing support vector machine  
model\_svm3 <- svm(formula\_completionrate, data = univ\_train2)  
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] 0.9264706

# doing simple tree  
model\_tree3 <- tree(formula\_completionrate, data = univ\_train2)  
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] 0.9019608

# doing conditional inference tree  
model\_party3 <- ctree(formula\_completionrate, data = univ\_train2)  
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] 0.9313725

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