US Research University Prediction Model

Philip Gabriel Andrada

October 31, 2016

# Preparation

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

## Warning: package 'randomForest' was built under R version 3.2.5

## randomForest 4.6-12

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

library(caret)

## Warning: package 'caret' was built under R version 3.2.5

## Loading required package: lattice

## Loading required package: ggplot2

##   
## Attaching package: 'ggplot2'

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

library(Boruta)

## Warning: package 'Boruta' was built under R version 3.2.5

## Loading required package: ranger

## Warning: package 'ranger' was built under R version 3.2.5

##   
## Attaching package: 'ranger'

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

library(e1071)

## Warning: package 'e1071' was built under R version 3.2.5

library(ROCR)

## Warning: package 'ROCR' was built under R version 3.2.5

## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.2.5

##   
## Attaching package: 'gplots'

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

library(corrplot)

## Warning: package 'corrplot' was built under R version 3.2.5

library(ggplot2)  
  
#Reading Data Files  
usuniv2010 <- read.csv("C:\\Users\\Philip\\Desktop\\Capstone\\MERGED2010\_11\_PP.csv")  
usuniv2011 <- read.csv("C:\\Users\\Philip\\Desktop\\Capstone\\MERGED2011\_12\_PP.csv")  
usuniv2012 <- read.csv("C:\\Users\\Philip\\Desktop\\Capstone\\MERGED2012\_13\_PP.csv")  
usuniv2013 <- read.csv("C:\\Users\\Philip\\Desktop\\Capstone\\MERGED2013\_14\_PP.csv")  
usuniv2014 <- read.csv("C:\\Users\\Philip\\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)  
  
#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 studnets 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]))  
}

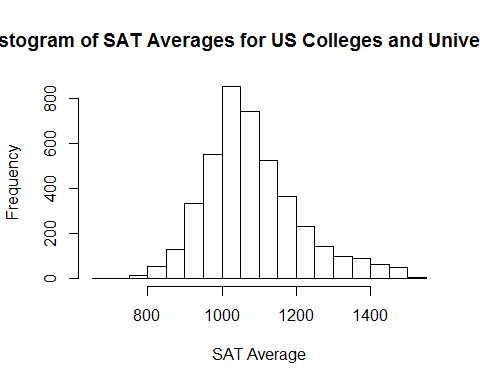
## Warning: NAs introduced by coercion

## 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

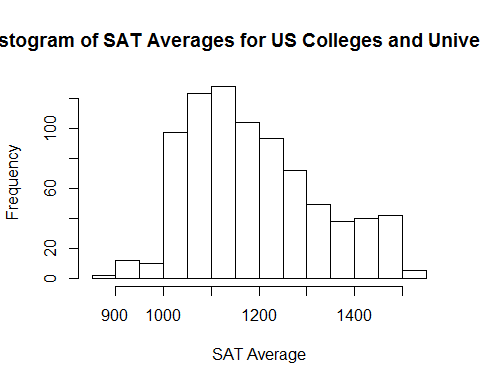
# Clean the results to have all complete   
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150\_4),]  
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150\_4\_ASIAN),]  
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150\_4\_WHITE),]  
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150\_4\_BLACK),]  
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),]

# Distributions and Box and Whisker Plots

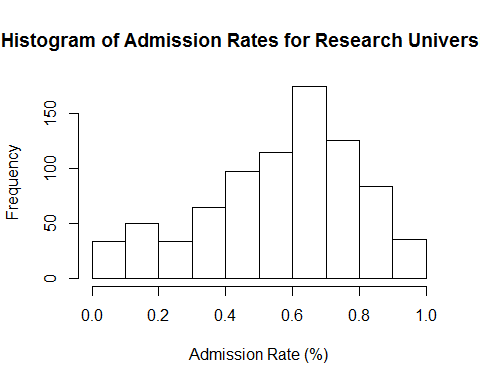
# Histogram of SAT Averages for US Colleges and Universities  
hist(usunivfilter$SAT\_AVG\_ALL, main = "Histogram of SAT Averages for US Colleges and Universities", xlab="SAT Average")



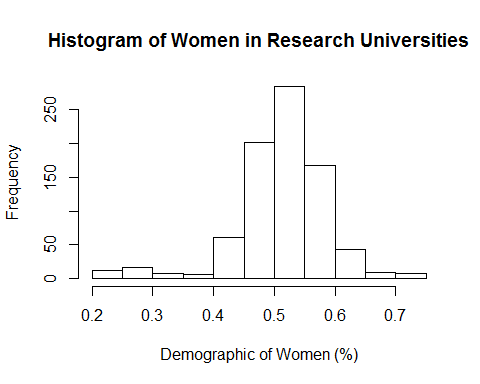
# Histogram of SAT Averages for US Research Universities  
hist(usresearchuniv$SAT\_AVG\_ALL, main = "Histogram of SAT Averages for US Colleges and Universities", xlab="SAT Average")



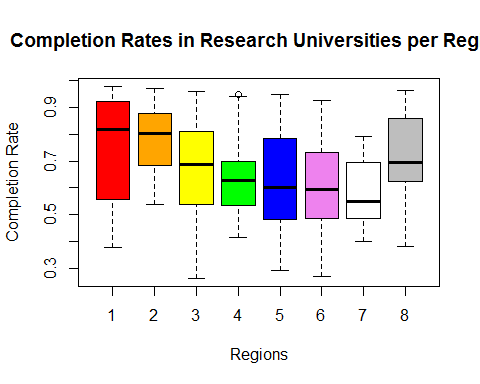
# Histogram of Admission Rates for US Research Universities  
hist(usresearchuniv$ADM\_RATE\_ALL, main = "Histogram of Admission Rates for Research Universities", xlab = "Admission Rate (%)")



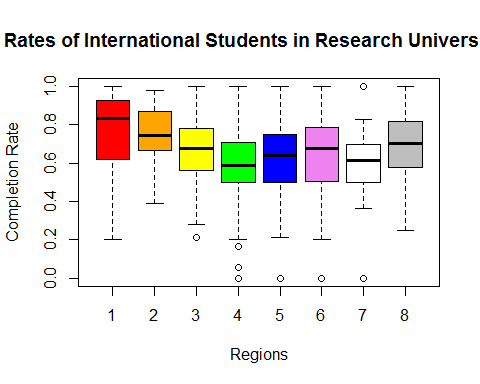
# Histogram of Women in US Research Universitie  
hist(usresearchuniv$UGDS\_WOMEN, main = "Histogram of Women in Research Universities", xlab = "Demographic of Women (%)")



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

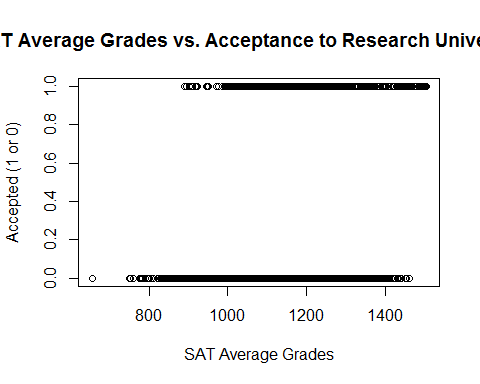


# Boxplot of COmpletion Rates of International Students per Region in US Research Universities  
boxplot(C150\_4\_NRA ~ REGION, usresearchuniv, main = "Completion Rates of International Students in Research Universities Per Region", col=c("red", "orange", "yellow", "green", "blue", "violet", "white",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

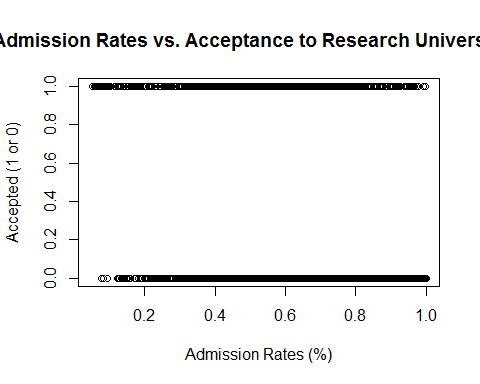


# Correlations

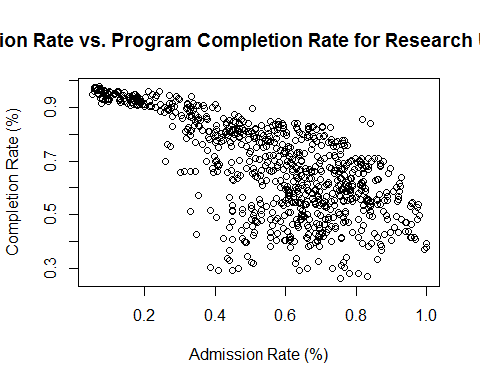
#Correlation between the SAT grades and the acceptance for the research universities  
plot(usunivfilter$SAT\_AVG\_ALL, usunivfilter$ACCEPTED, main="SAT Average Grades vs. Acceptance to Research Universities", xlab="SAT Average Grades", ylab="Accepted (1 or 0)")



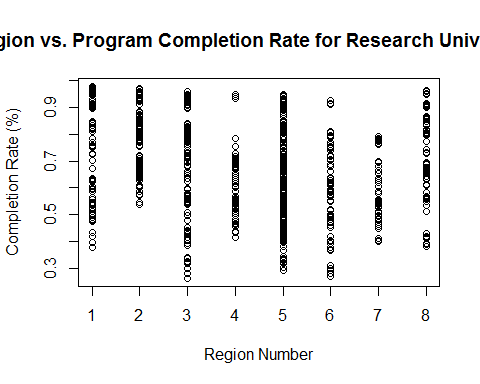
#Correlation between the admission rates and the acceptance for the research universities  
plot(usunivfilter$ADM\_RATE\_ALL, usunivfilter$ACCEPTED, main="Admission Rates vs. Acceptance to Research Universities", xlab="Admission Rates (%)", ylab="Accepted (1 or 0)")



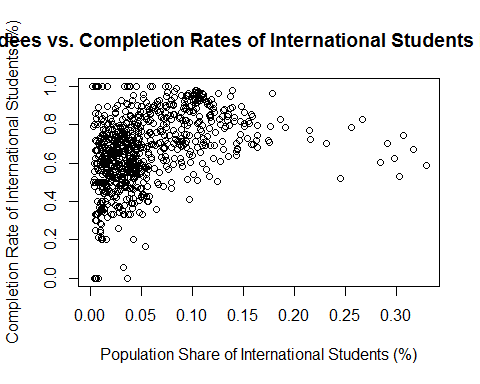
#Correlation between admission rate for research universities and program completion rate  
plot(usresearchuniv$ADM\_RATE\_ALL, usresearchuniv$C150\_4, main="Admission Rate vs. Program Completion Rate for Research Universities", xlab="Admission Rate (%)", ylab="Completion Rate (%)")



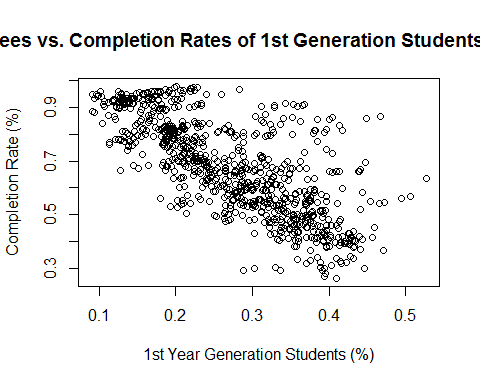
#Correlation between admission rate for research universities and program completion rate  
plot(usresearchuniv$REGION, usresearchuniv$C150\_4, main="Region vs. Program Completion Rate for Research Universities", xlab="Region Number", ylab="Completion Rate (%)")



#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 International Students in Research Universities", xlab="Population Share of International Students (%)", ylab="Completion Rate of International Students (%)")



#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 of 1st Generation Students in Research Universities", xlab="1st Year Generation Students (%)", ylab="Completion Rate (%)")



# U.S. Research University Acceptance Model

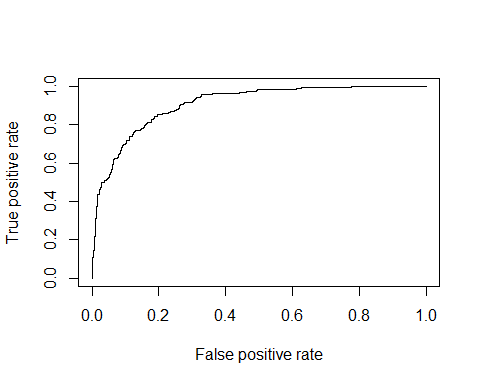
# create a training and test model using a 75%/25% from the data set   
rm\_train <- sample(nrow(usunivfilter), floor(nrow(usunivfilter)\*0.75))  
univ\_train <- usunivfilter[rm\_train,]  
univ\_test <- usunivfilter[-rm\_train,]  
  
# create a formula for the US research university acceptance model for International Students taking up Science degrees/majors  
test\_formulagen <- 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  
model\_glm <- glm(test\_formulagen, data=univ\_train,family=binomial())  
summary(model\_glm)

##   
## Call:  
## glm(formula = test\_formulagen, family = binomial(), data = univ\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.55632 -0.48638 -0.24229 -0.07769 3.06783   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.773e+01 1.442e+00 -12.303 < 2e-16 \*\*\*  
## REGION 1.391e-01 3.115e-02 4.466 7.97e-06 \*\*\*  
## ADM\_RATE\_ALL 1.364e+00 4.197e-01 3.250 0.001152 \*\*   
## SAT\_AVG\_ALL 1.575e-02 1.018e-03 15.459 < 2e-16 \*\*\*  
## PCIP11 9.323e-01 1.915e+00 0.487 0.626315   
## PCIP12 -5.023e-01 2.009e+01 -0.025 0.980057   
## PCIP14 4.931e+00 7.568e-01 6.516 7.22e-11 \*\*\*  
## PCIP15 6.008e-01 2.234e+00 0.269 0.787933   
## PCIP24 -5.244e+00 1.200e+00 -4.372 1.23e-05 \*\*\*  
## PCIP26 7.049e+00 1.760e+00 4.005 6.20e-05 \*\*\*  
## PCIP27 -2.577e+01 6.851e+00 -3.761 0.000169 \*\*\*  
## PCIP40 -3.526e+01 4.821e+00 -7.313 2.62e-13 \*\*\*  
## PCIP45 8.232e+00 1.209e+00 6.810 9.78e-12 \*\*\*  
## PCIP51 1.954e+00 5.931e-01 3.295 0.000983 \*\*\*  
## PCIP52 5.294e-01 6.632e-01 0.798 0.424729   
## UGDS\_NRA 9.166e+00 1.531e+00 5.985 2.16e-09 \*\*\*  
## UGDS\_UNKN -1.892e+00 1.593e+00 -1.188 0.234951   
## COSTT4\_A -1.077e-04 7.288e-06 -14.784 < 2e-16 \*\*\*  
## PCTFLOAN -7.882e-01 5.565e-01 -1.416 0.156712   
## UGDS\_WOMEN 1.181e-01 7.831e-01 0.151 0.880084   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3122.7 on 3184 degrees of freedom  
## Residual deviance: 1911.3 on 3165 degrees of freedom  
## AIC: 1951.3  
##   
## Number of Fisher Scoring iterations: 6

# do the testing with the prediction model  
univ\_test$scores <- predict(model\_glm, type="response", newdata = univ\_test)  
pred <- prediction(univ\_test$scores, univ\_test$ACCEPTED)  
  
# prepare confusion matrix to see the scores  
c <- confusionMatrix(as.integer(univ\_test$scores > 0.5), univ\_test$ACCEPTED)  
c$table

## Reference  
## Prediction 0 1  
## 0 814 91  
## 1 47 110

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



# Now we check on what acceptable ways we could do for regression  
#doing single decision tree  
model\_tree <- rpart(test\_formulagen, method="anova",data = univ\_train)  
pred\_tree <- predict(model\_tree, newdata = univ\_test)  
accu = abs(pred\_tree - univ\_test$ACCEPTED) < 0.25  
frac = sum(accu)/length(accu)  
print(frac)

## [1] 0.8766478

#doing random forest  
model\_forest <- randomForest(test\_formulagen, 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\_forest <- predict(model\_forest, newdata = univ\_test)  
accu2 <- abs(pred\_forest - univ\_test$ACCEPTED) < 0.25  
frac2 <- sum(accu2)/length(accu2)  
print(frac2)

## [1] 0.8719397

#doing support vector machine  
model\_svm <- svm(test\_formulagen, data = univ\_train)  
pred\_svm <- predict(model\_svm, newdata = univ\_test)  
accu3 <- abs(pred\_svm - univ\_test$ACCEPTED) < 0.25  
frac3 <- sum(accu3)/length(accu3)  
print(frac3)

## [1] 0.834275

# We will consider all variables, and use Boruta to use what variables we could use for doing a better outcome of the moded  
  
# 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,doTrace = 2)

## 1. run of importance source...

## 2. run of importance source...

## 3. run of importance source...

## 4. run of importance source...

## 5. run of importance source...

## 6. run of importance source...

## 7. run of importance source...

## 8. run of importance source...

## 9. run of importance source...

## 10. run of importance source...

## 11. run of importance source...

## 12. run of importance source...

## 13. run of importance source...

## Confirmed 48 attributes: ADM\_RATE, ADM\_RATE\_ALL, C150\_4, C150\_4\_AIAN, C150\_4\_ASIAN and 43 more.

## Rejected 5 attributes: C150\_4\_NHPI, PCIP25, PCIP29, PCIP47, PCIP48.

## 14. run of importance source...

## 15. run of importance source...

## 16. run of importance source...

## 17. run of importance source...

## Confirmed 5 attributes: C150\_4\_NRA, PCIP13, PCIP51, UGDS\_2MOR, UGDS\_NHPI.

## Rejected 2 attributes: PCIP12, PCIP46.

## 18. run of importance source...

## 19. run of importance source...

## 20. run of importance source...

## 21. run of importance source...

## Confirmed 3 attributes: PCIP30, PCIP38, PCIP54.

## 22. run of importance source...

## 23. run of importance source...

## 24. run of importance source...

## 25. run of importance source...

## 26. run of importance source...

## 27. run of importance source...

## 28. run of importance source...

## Confirmed 2 attributes: PCIP09, UGDS\_AIAN.

## 29. run of importance source...

## 30. run of importance source...

## 31. run of importance source...

## 32. run of importance source...

## 33. run of importance source...

## 34. run of importance source...

## 35. run of importance source...

## 36. run of importance source...

## 37. run of importance source...

## 38. run of importance source...

## 39. run of importance source...

## 40. run of importance source...

## Confirmed 1 attributes: PCIP49.

## 41. run of importance source...

## 42. run of importance source...

## 43. run of importance source...

## 44. run of importance source...

## 45. run of importance source...

## 46. run of importance source...

## 47. run of importance source...

## 48. run of importance source...

## 49. run of importance source...

## 50. run of importance source...

## 51. run of importance source...

## 52. run of importance source...

## 53. run of importance source...

## Confirmed 1 attributes: C150\_4\_2MOR.

## 54. run of importance source...

## 55. run of importance source...

## 56. run of importance source...

## 57. run of importance source...

## 58. run of importance source...

## 59. run of importance source...

## 60. run of importance source...

## 61. run of importance source...

## 62. run of importance source...

## 63. run of importance source...

## 64. run of importance source...

## 65. run of importance source...

## 66. run of importance source...

## 67. run of importance source...

## 68. run of importance source...

## 69. run of importance source...

## Confirmed 1 attributes: PCIP41.

## 70. run of importance source...

## 71. run of importance source...

## 72. run of importance source...

## 73. run of importance source...

## 74. run of importance source...

## 75. run of importance source...

## 76. run of importance source...

## 77. run of importance source...

## 78. run of importance source...

## 79. run of importance source...

## 80. run of importance source...

## 81. run of importance source...

## 82. run of importance source...

## 83. run of importance source...

## 84. run of importance source...

## 85. run of importance source...

## 86. run of importance source...

## 87. run of importance source...

## 88. run of importance source...

## 89. run of importance source...

## 90. run of importance source...

## 91. run of importance source...

## 92. run of importance source...

## 93. run of importance source...

## 94. run of importance source...

## 95. run of importance source...

## 96. run of importance source...

## 97. run of importance source...

## 98. run of importance source...

## 99. run of importance source...

print(boruta.train)

## Boruta performed 99 iterations in 2.333855 mins.  
## 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.

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

## meanImp medianImp minImp maxImp  
## REGION 5.59590070 5.60881747 3.68306598 7.409247855  
## ADM\_RATE 7.15868996 7.13700959 5.42674038 8.837163535  
## ADM\_RATE\_ALL 7.17227824 7.11433019 5.32249735 8.489047750  
## SAT\_AVG\_ALL 12.57399314 12.61157714 10.89845324 14.149015930  
## PCIP01 6.25362862 6.34434033 4.70926472 7.514339576  
## PCIP03 6.55907358 6.51706164 4.89603487 7.818602020  
## PCIP04 11.71423762 11.74086693 10.03202778 13.354837722  
## PCIP05 8.29875631 8.29965953 7.25335135 9.571185634  
## PCIP09 4.84375860 4.73274441 2.12695595 6.646705476  
## PCIP10 2.76226544 2.80836503 0.27023966 4.347018187  
## PCIP11 6.47956196 6.55439709 4.09052046 8.401998667  
## PCIP12 1.08123736 1.28594455 -0.46549930 2.257870104  
## PCIP13 5.97623869 5.98676191 3.53553362 7.568568876  
## PCIP14 18.69054105 18.63470959 16.68296336 20.450613312  
## PCIP15 4.90223126 4.83999149 2.50984985 6.360767747  
## PCIP16 7.61184430 7.59031460 5.70784235 9.495639939  
## PCIP19 7.54822419 7.55671161 6.27076592 8.812650303  
## PCIP22 2.49226865 2.54506470 -0.08690792 4.347883257  
## PCIP23 8.30753362 8.35561507 6.56384595 9.950867250  
## PCIP24 5.86613099 5.84343698 4.38979238 7.657641087  
## PCIP25 -1.19168137 -1.41375504 -2.38065294 0.008222033  
## PCIP26 6.02243787 5.92160083 4.79754330 7.920466023  
## PCIP27 5.26020685 5.28411704 2.57741118 6.868432832  
## PCIP29 0.23100035 0.00000000 0.00000000 1.001001503  
## PCIP30 4.13273744 4.31131753 1.61177302 6.192867280  
## PCIP31 4.85590697 4.86782522 2.91215341 6.238148650  
## PCIP38 4.17430616 4.22263547 2.22664295 5.507767123  
## PCIP39 5.49827812 5.53979735 4.27129025 6.694960897  
## PCIP40 5.77887869 5.79752294 4.02192622 7.149522948  
## PCIP41 3.26377559 3.28207813 1.28732164 5.049338843  
## PCIP42 4.88944534 4.89193788 3.27970161 6.443029785  
## PCIP43 7.10307336 7.11631041 5.53904673 9.103272001  
## PCIP44 4.58967230 4.69588030 2.37602308 6.193636905  
## PCIP45 7.54275016 7.54071207 5.75214960 9.424697577  
## PCIP46 0.47892762 1.00100150 -1.33259295 1.970045670  
## PCIP47 0.36898964 0.00000000 -1.00100150 1.416319437  
## PCIP48 0.05743157 0.01553987 -1.37749264 1.001001503  
## PCIP49 3.39128637 3.41131174 1.72950228 5.131308982  
## PCIP50 5.77483001 5.85781450 3.55092395 7.634362253  
## PCIP51 3.99827643 4.10968984 1.44266813 5.882954718  
## PCIP52 9.77454683 9.73259896 8.70082596 11.318409719  
## PCIP54 3.75659660 3.87538743 1.46463582 5.801084644  
## UGDS\_WHITE 8.18831612 8.14971388 7.05572632 9.762520201  
## UGDS\_BLACK 10.77070034 10.89687742 8.52946003 12.138459415  
## UGDS\_HISP 6.25365921 6.25207818 4.15019289 8.359714041  
## UGDS\_ASIAN 9.20778005 9.21588322 8.10425546 10.128801362  
## UGDS\_AIAN 4.12794934 4.03577717 2.34387250 6.358553562  
## UGDS\_NHPI 3.90120812 3.89707764 2.65714282 5.298123781  
## UGDS\_2MOR 4.36451701 4.39638530 1.75804865 6.355576895  
## UGDS\_NRA 7.19510623 7.19217953 5.54931236 8.660363461  
## UGDS\_UNKN 6.02221859 6.04596923 4.30396461 7.664937749  
## PPTUG\_EF 6.77757745 6.81807806 5.24974169 8.280212766  
## COSTT4\_A 9.82573931 9.85097217 8.17515530 11.439334414  
## TUITIONFEE\_IN 9.46186315 9.46164286 8.36952429 10.956464390  
## TUITIONFEE\_OUT 5.58192889 5.63615231 3.64901040 7.040279060  
## C150\_4 7.97559808 7.95076338 6.94533074 9.334247968  
## C150\_4\_WHITE 6.75759590 6.71612408 5.22863485 7.967125766  
## C150\_4\_BLACK 7.19890718 7.17676128 6.13538901 8.247794358  
## C150\_4\_HISP 5.80407250 5.90877651 4.26516875 7.099246974  
## C150\_4\_ASIAN 6.00687459 5.99027922 4.72088843 7.356160558  
## C150\_4\_AIAN 7.13889322 7.13595803 5.84028552 8.273495699  
## C150\_4\_NHPI 1.09763716 1.30922991 -0.25464310 2.623949114  
## C150\_4\_2MOR 3.15890064 3.20276104 1.62356393 4.785896179  
## C150\_4\_NRA 4.40872320 4.42527935 2.94222053 5.907551847  
## C150\_4\_UNKN 7.19592324 7.17998521 5.56319782 8.448857592  
## RET\_FT4 10.59646439 10.62586400 9.29534464 11.993002254  
## PCTFLOAN 14.01886843 14.10580679 12.54004727 15.554679044  
## PAR\_ED\_PCT\_1STGEN 6.03354311 6.15236284 4.80459372 7.513489454  
## UGDS\_MEN 12.47748622 12.43175383 10.89531092 14.220308926  
## UGDS\_WOMEN 12.35052125 12.36484564 11.26898659 13.582459380  
## normHits decision  
## REGION 1.00000000 Confirmed  
## ADM\_RATE 1.00000000 Confirmed  
## ADM\_RATE\_ALL 1.00000000 Confirmed  
## SAT\_AVG\_ALL 1.00000000 Confirmed  
## PCIP01 1.00000000 Confirmed  
## PCIP03 1.00000000 Confirmed  
## PCIP04 1.00000000 Confirmed  
## PCIP05 1.00000000 Confirmed  
## PCIP09 0.92929293 Confirmed  
## PCIP10 0.55555556 Tentative  
## PCIP11 1.00000000 Confirmed  
## PCIP12 0.01010101 Rejected  
## PCIP13 0.98989899 Confirmed  
## PCIP14 1.00000000 Confirmed  
## PCIP15 0.97979798 Confirmed  
## PCIP16 1.00000000 Confirmed  
## PCIP19 1.00000000 Confirmed  
## PCIP22 0.45454545 Tentative  
## PCIP23 1.00000000 Confirmed  
## PCIP24 1.00000000 Confirmed  
## PCIP25 0.00000000 Rejected  
## PCIP26 1.00000000 Confirmed  
## PCIP27 0.98989899 Confirmed  
## PCIP29 0.00000000 Rejected  
## PCIP30 0.89898990 Confirmed  
## PCIP31 0.96969697 Confirmed  
## PCIP38 0.93939394 Confirmed  
## PCIP39 0.98989899 Confirmed  
## PCIP40 1.00000000 Confirmed  
## PCIP41 0.71717172 Confirmed  
## PCIP42 0.98989899 Confirmed  
## PCIP43 1.00000000 Confirmed  
## PCIP44 0.96969697 Confirmed  
## PCIP45 1.00000000 Confirmed  
## PCIP46 0.01010101 Rejected  
## PCIP47 0.00000000 Rejected  
## PCIP48 0.00000000 Rejected  
## PCIP49 0.78787879 Confirmed  
## PCIP50 1.00000000 Confirmed  
## PCIP51 0.87878788 Confirmed  
## PCIP52 1.00000000 Confirmed  
## PCIP54 0.83838384 Confirmed  
## UGDS\_WHITE 1.00000000 Confirmed  
## UGDS\_BLACK 1.00000000 Confirmed  
## UGDS\_HISP 1.00000000 Confirmed  
## UGDS\_ASIAN 1.00000000 Confirmed  
## UGDS\_AIAN 0.90909091 Confirmed  
## UGDS\_NHPI 0.87878788 Confirmed  
## UGDS\_2MOR 0.93939394 Confirmed  
## UGDS\_NRA 1.00000000 Confirmed  
## UGDS\_UNKN 1.00000000 Confirmed  
## PPTUG\_EF 1.00000000 Confirmed  
## COSTT4\_A 1.00000000 Confirmed  
## TUITIONFEE\_IN 1.00000000 Confirmed  
## TUITIONFEE\_OUT 0.98989899 Confirmed  
## C150\_4 1.00000000 Confirmed  
## C150\_4\_WHITE 1.00000000 Confirmed  
## C150\_4\_BLACK 1.00000000 Confirmed  
## C150\_4\_HISP 1.00000000 Confirmed  
## C150\_4\_ASIAN 1.00000000 Confirmed  
## C150\_4\_AIAN 1.00000000 Confirmed  
## C150\_4\_NHPI 0.00000000 Rejected  
## C150\_4\_2MOR 0.70707071 Confirmed  
## C150\_4\_NRA 0.95959596 Confirmed  
## C150\_4\_UNKN 1.00000000 Confirmed  
## RET\_FT4 1.00000000 Confirmed  
## PCTFLOAN 1.00000000 Confirmed  
## PAR\_ED\_PCT\_1STGEN 1.00000000 Confirmed  
## UGDS\_MEN 1.00000000 Confirmed  
## UGDS\_WOMEN 1.00000000 Confirmed

# 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 ~ REGION + ADM\_RATE\_ALL + UGDS\_NRA + PPTUG\_EF + COSTT4\_A + PCTFLOAN + PAR\_ED\_PCT\_1STGEN)  
  
model\_tree2 <- rpart(formula\_completionrate, method="anova",data = univ\_train2)  
pred\_tree2 <- predict(model\_tree2, newdata = univ\_test2)  
accu4 = abs(pred\_tree2 - univ\_test2$C150\_4\_NRA) < 0.25  
frac4 = sum(accu4)/length(accu4)  
print(frac4)

## [1] 0.8823529

model\_forest2 <- randomForest(formula\_completionrate, data = univ\_train2)  
pred\_forest2 <- predict(model\_forest2, newdata = univ\_test2)  
accu5 <- abs(pred\_forest2 - univ\_test2$ACCEPTED) < 0.25  
frac5 <- sum(accu5)/length(accu5)  
print(frac5)

## [1] 0.3921569

model\_svm2 <- svm(formula\_completionrate, data = univ\_train2)  
pred\_svm2 <- predict(model\_svm2, newdata = univ\_test2)  
accu6 <- abs(pred\_svm2 - univ\_test2$ACCEPTED) < 0.25  
frac6 <- sum(accu6)/length(accu6)  
print(frac6)

## [1] 0.3823529