STUDENT DROPOUT PREDICTION

Thi Diem My Nguyen - School of Business - New Jersey city University - tnguyen9@njcu.edu

Abstract

Dropout remains a persistent challenge within each college. In this research, I present a case study on automatically detecting whether a student is at-risk of dropout at New Jersey City University. I trained several machine learning algorithms in to come up with the best prediction model of student dropout from data on NJCU Student Static Data, NJCU Student Progress Data, and NJCU Financial Aid Data.

Introduction

Teachers and school administrators have striven to reduce dropout for quite some time, but it continues to persist in schools as a problem through the present day. Dropping out of colleges is considered not just a serious educational problem but also a severe social problem, especially in recent decades when technology and societal developments have rendered more and more people without at least a college degree less likely to find a job. It is critical to understand the causes and recognize the signs, in this project I will aim to accurately predict the probability of a student dropping out of a college. I will measure prediction accuracy and analyze aspects of the students' data to recognize the most important factors leading to high dropout rates. Machine learning techniques can effectively facilitate the determination of at-risk students and timely planning for interventions. I will implement several classification algorithms to find the best prediction model.

Data set and features

The data was gathered from New Jersey City University undergraduate students from 2012 to 2017. The data set contains three types of data:

Student Static Data: Static data include demographic and educational background information about each student in the cohort; these data do not change over time. These data are collected through a CSV file, uploaded once for each student. This file contains one record per student, and each student appears in only one static file, corresponding to the year in which he/she first enrolled.

Student Progress Data: Progress/General data reflect your students' academic progression and outcomes over time. These data are CSV files to be uploaded, reflecting each student's activity for each term in each academic year. This file contains one record per student. Multiple cohorts are included in each term file.

Student Financial Aid Data: Financial Aid Data was collected for each student for each academic year, and it is stored in different columns for different years. It contains Financial Aid and other related information such as scholarships, loans, gross income.

The target feature is a 0 or 1 indicating dropout.

The first step was to import and clean the data, in order to determine that there is no information redundancy and blank fields or data that may affect the prediction process.

```
memory.size()

## [1] 44.53

memory.limit()

## [1] 12187

memory.limit(size=500000)

## [1] 5e+05

set.seed(3333)
library(dplyr)
library(Hmisc)
library(ggplot2)
library(MASS)
library(imputeTS)
```

Import Data

```
# Import Student Static Data
getwd()

## [1] "C:/Users/Diem My/Desktop/THI NGUYEN/HOC MY/FALL 2021/Machine Learning
1/Final Project/Code"

setwd("/Users/Diem My/Desktop/THI NGUYEN/HOC MY/FALL 2021/Machine Learning 1/
Final Project/studentdropout/Student Retention Challenge Data/Student Static
Data")
StaticFall2011 <- read.csv("Fall 2011_ST.csv", header = T)
StaticFall2012 <- read.csv("Fall 2012.csv", header = T)
StaticFall2013 <- read.csv("Fall 2013.csv", header = T)
StaticFall2014 <- read.csv("Fall 2014.csv", header = T)
StaticFall2015 <- read.csv("Fall 2015.csv", header = T)
StaticFall2016 <- read.csv("Fall 2016.csv", header = T)
StaticSpring2012 <- read.csv("Spring 2012_ST.csv", header = T)
StaticSpring2013 <- read.csv("Spring 2013.csv", header = T)
StaticSpring2014 <- read.csv("Spring 2014.csv", header = T)
StaticSpring2015 <- read.csv("Spring 2015.csv", header = T)
StaticSpring2016 <- read.csv("Spring 2016.csv", header = T)
StudentStaticData <- rbind(StaticFall2011,StaticFall2012,StaticFall2013,Stati
cFall2014,StaticFall2015,StaticFall2016,StaticSpring2012,StaticSpring2013,Sta
ticSpring2014,StaticSpring2015,StaticSpring2016)

# Remove unused data
```

```
rm(StaticFall2011, StaticFall2012, StaticFall2013, StaticFall2014, StaticFall
2015, StaticFall2016, StaticSpring2012, StaticSpring2013, StaticSpring2014, S
taticSpring2015, StaticSpring2016)
gc()
```

```
##           used  (Mb) gc trigger  (Mb) limit (Mb) max used  (Mb)
## Ncells 2136502 114.2    3980227 212.6         NA  3061712 163.6
## Vcells 3919698  30.0    8388608  64.0     102400  8388307  64.0
```

```
# Import Student Progress Data
getwd()
```

```
## [1] "C:/Users/Diem My/Desktop/THI NGUYEN/HOC MY/FALL 2021/Machine Learning
1/Final Project/studentdropout/Student Retention Challenge Data/Student Stati
c Data"
```

```
setwd("/Users/Diem My/Desktop/THI NGUYEN/HOC MY/FALL 2021/Machine Learning 1/
Final Project/studentdropout/Student Retention Challenge Data/Student Progres
s Data")
ProgressFall2011 <- read.csv("Fall 2011_SP.csv",header = T)
ProgressFall2012 <- read.csv("Fall 2012_SP.csv",header = T)
ProgressFall2013 <- read.csv("Fall 2013_SP.csv",header = T)
ProgressFall2014 <- read.csv("Fall 2014_SP.csv",header = T)
ProgressFall2015 <- read.csv("Fall 2015_SP.csv",header = T)
ProgressFall2016 <- read.csv("Fall 2016_SP.csv",header = T)
ProgressSpring2012 <- read.csv("Spring 2012_SP.csv",header = T)
ProgressSpring2013 <- read.csv("Spring 2013_SP.csv",header = T)
ProgressSpring2014 <- read.csv("Spring 2014_SP.csv",header = T)
ProgressSpring2015 <- read.csv("Spring 2015_SP.csv",header = T)
ProgressSpring2016 <- read.csv("Spring 2016_SP.csv",header = T)
ProgressSpring2017 <- read.csv("Spring 2017_SP.csv",header = T)
ProgressSum2012 <- read.csv("Sum 2012.csv",header = T)
ProgressSum2013 <- read.csv("Sum 2013.csv",header = T)
ProgressSum2014 <- read.csv("Sum 2014.csv",header = T)
ProgressSum2015 <- read.csv("Sum 2015.csv",header = T)
ProgressSum2016 <- read.csv("Sum 2016.csv",header = T)
ProgressSum2017 <- read.csv("Sum 2017.csv",header = T)
```

```
#Create new column AcademicYearID
ProgressFall2011 <- mutate(ProgressFall2011, AcademicYearID = 1)
ProgressSpring2012 <- mutate(ProgressSpring2012, AcademicYearID = 2)
ProgressSum2012 <- mutate(ProgressSum2012, AcademicYearID = 3)
ProgressFall2012 <- mutate(ProgressFall2012, AcademicYearID = 4)
ProgressSpring2013 <- mutate(ProgressSpring2013, AcademicYearID = 5)
ProgressSum2013 <- mutate(ProgressSum2013, AcademicYearID = 6)
ProgressFall2013 <- mutate(ProgressFall2013, AcademicYearID = 7)
ProgressSpring2014 <- mutate(ProgressSpring2014, AcademicYearID = 8)
ProgressSum2014 <- mutate(ProgressSum2014, AcademicYearID = 9)
ProgressFall2014 <- mutate(ProgressFall2014, AcademicYearID = 10)
ProgressSpring2015 <- mutate(ProgressSpring2015, AcademicYearID = 11)
ProgressSum2015 <- mutate(ProgressSum2015, AcademicYearID = 12)
```

```
ProgressFall2015 <- mutate(ProgressFall2015, AcademicYearID = 13)
ProgressSpring2016 <- mutate(ProgressSpring2016, AcademicYearID = 14)
ProgressSum2016 <- mutate(ProgressSum2016, AcademicYearID = 15)
ProgressFall2016 <- mutate(ProgressFall2016, AcademicYearID = 16)
ProgressSpring2017 <- mutate(ProgressSpring2017, AcademicYearID = 17)
ProgressSum2017 <- mutate(ProgressSum2017, AcademicYearID = 18)

StudentProgressData1 <- rbind(ProgressFall2011, ProgressFall2012, ProgressFal
l2013, ProgressFall2014, ProgressFall2015, ProgressFall2016,ProgressSpring201
2,ProgressSpring2013, ProgressSpring2014, ProgressSpring2015, ProgressSpring2
016,ProgressSpring2017, ProgressSum2012, ProgressSum2013, ProgressSum2014, Pr
ogressSum2015, ProgressSum2016, ProgressSum2017)

ProgressData <- StudentProgressData1 %>% group_by(StudentID) %>% top_n(1, Aca
demicYearID)

#Remove unused data
rm(StudentProgressData1)
rm(ProgressFall2011, ProgressFall2012, ProgressFall2013, ProgressFall2014, Pr
ogressFall2015, ProgressFall2016, ProgressSpring2012, ProgressSpring2013, Pro
gressSpring2014, ProgressSpring2015, ProgressSpring2016, ProgressSpring2017,
ProgressSum2012, ProgressSum2013, ProgressSum2014, ProgressSum2015, ProgressS
um2016, ProgressSum2017)
gc()
```

```
##           used  (Mb) gc trigger  (Mb) limit (Mb) max used  (Mb)
## Ncells 2193546 117.2    3980227 212.6         NA  3980227 212.6
## Vcells 4222375  32.3   10146329  77.5     102400  8388569  64.0
```

```
# Import Student Financial Aid Data
getwd()
```

```
## [1] "C:/Users/Diem My/Desktop/THI NGUYEN/HOC MY/FALL 2021/Machine Learning
1/Final Project/studentdropout/Student Retention Challenge Data/Student Progr
ess Data"
```

```
setwd("/Users/Diem My/Desktop/THI NGUYEN/HOC MY/FALL 2021/Machine Learning 1/
Final Project/studentdropout/Student Retention Challenge Data/Student Financi
al Aid Data")
FinancialAid <- read.csv("2011-2017_Cohorts_Financial_Aid_and_Fafsa_Data.csv"
,header = T)

# Import Dropout Train Labels
getwd()
```

```
## [1] "C:/Users/Diem My/Desktop/THI NGUYEN/HOC MY/FALL 2021/Machine Learning
1/Final Project/studentdropout/Student Retention Challenge Data/Student Finan
cial Aid Data"
```

```
setwd("/Users/Diem My/Desktop/THI NGUYEN/HOC MY/FALL 2021/Machine Learning 1/
Final Project/studentdropout")
```

```
TrainLabels <- read.csv("DropoutTrainLabels.csv",header = T)

# Import Test Data
getwd()

## [1] "C:/Users/Diem My/Desktop/THI NGUYEN/HOC MY/FALL 2021/Machine Learning
1/Final Project/studentdropout"

setwd("/Users/Diem My/Desktop/THI NGUYEN/HOC MY/FALL 2021/Machine Learning 1/
Final Project/studentdropout/Student Retention Challenge Data/Test Data")
TestData <- read.csv("TestIDs.csv",header = T)
```

Exploratory Data Analysis - EDA

Student Static Data

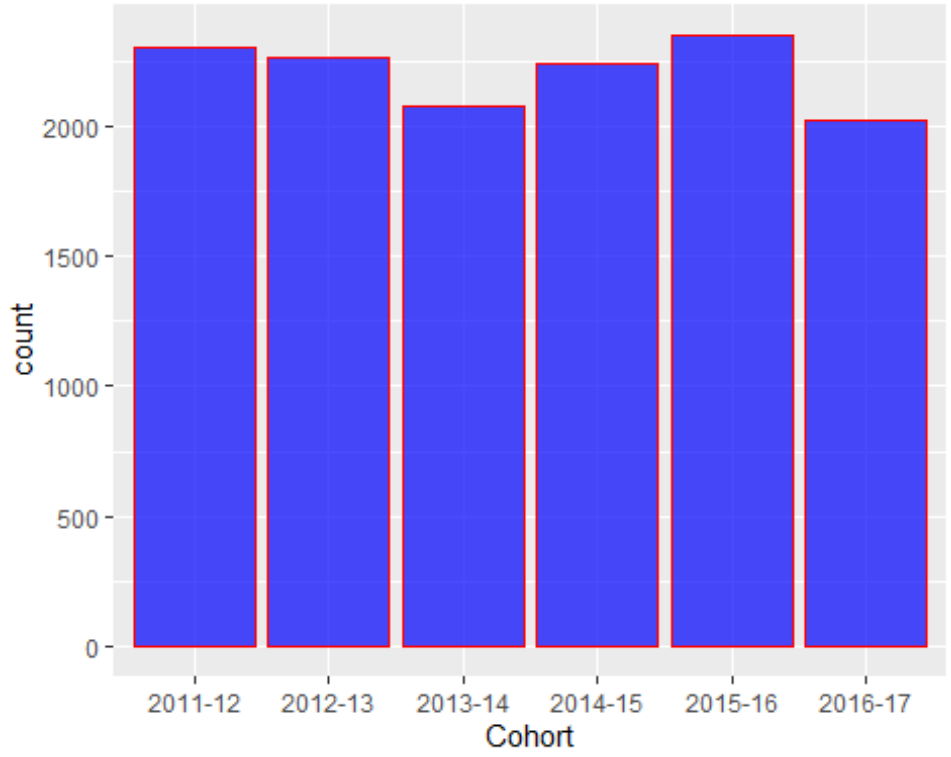
Basic descriptive statistics of the variables in the Student Static Data

```
summary(StudentStaticData)

##    StudentID         Cohort            CohortTerm     Campus
##  Min.   : 20932   Length:13261       Min.   :1.000   Mode:logical
##  1st Qu.:305254   Class :character   1st Qu.:1.000   NA's:13261
##  Median :321478   Mode  :character   Median :1.000
##  Mean   :316151                      Mean   :1.391
##  3rd Qu.:343511                      3rd Qu.:1.000
##  Max.   :359783                      Max.   :3.000
##
##    Address1           Address2             City              State
##  Length:13261       Length:13261       Length:13261       Length:13261
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##       Zip        RegistrationDate       Gender        BirthYear
##  Min.   :  747   Min.   :20110111   Min.   :1.000   Min.   :1945
##  1st Qu.: 7060   1st Qu.:20120710   1st Qu.:1.000   1st Qu.:1986
##  Median : 7304   Median :20140121   Median :2.000   Median :1992
##  Mean   : 7790   Mean   :20136109   Mean   :1.596   Mean   :1989
##  3rd Qu.: 7307   3rd Qu.:20150624   3rd Qu.:2.000   3rd Qu.:1995
##  Max.   :98118   Max.   :20160912   Max.   :2.000   Max.   :2000
##  NA's   :134
##    BirthMonth        Hispanic       AmericanIndian         Asian
##  Min.   : 1.000   Min.   :-1.0000   Min.   :-1.00000   Min.   :-1.00000
##  1st Qu.: 4.000   1st Qu.: 0.0000   1st Qu.: 0.00000   1st Qu.: 0.00000
##  Median : 7.000   Median : 0.0000   Median : 0.00000   Median : 0.00000
```

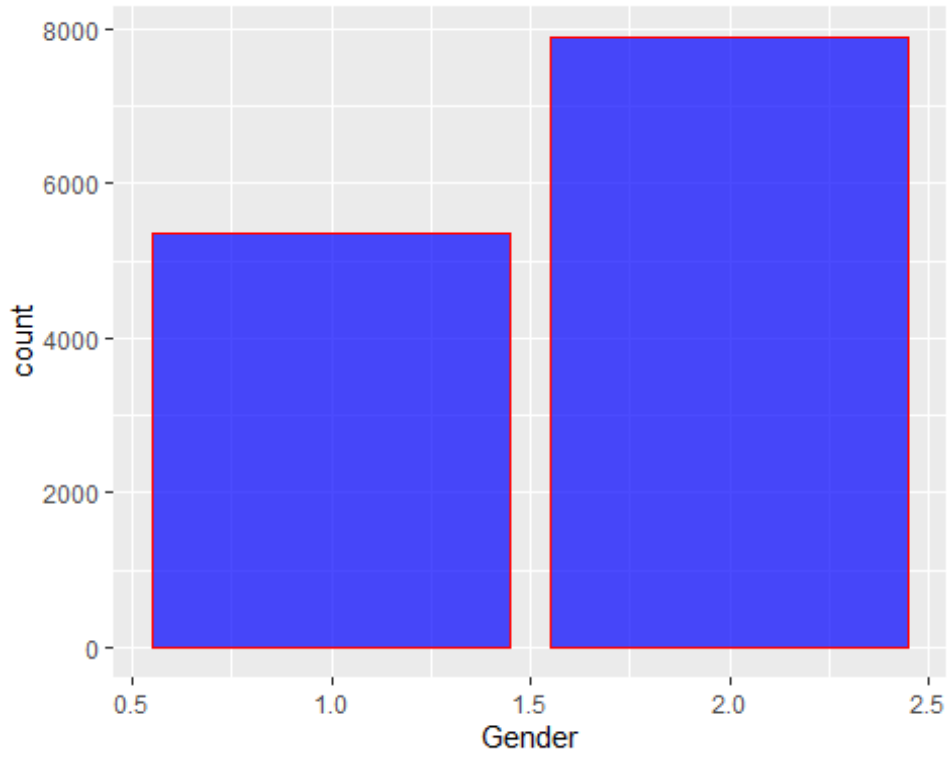
```
##  Mean   : 6.581   Mean   : 0.2568   Mean   :-0.06742   Mean   : 0.01848
##  3rd Qu.:10.000   3rd Qu.: 1.0000   3rd Qu.: 0.00000   3rd Qu.: 0.00000
##  Max.   :12.000   Max.   : 1.0000   Max.   : 1.00000   Max.   : 1.00000
##
##      Black         NativeHawaiian         White        TwoOrMoreRace
##  Min.   :-1.0000   Min.   :-1.00000   Min.   :-1.000   Min.   :-1.00000
##  1st Qu.: 0.0000   1st Qu.: 0.00000   1st Qu.: 0.000   1st Qu.: 0.00000
##  Median : 0.0000   Median : 0.00000   Median : 0.000   Median : 0.00000
##  Mean   : 0.1447   Mean   :-0.06757   Mean   : 0.183   Mean   :-0.05181
##  3rd Qu.: 0.0000   3rd Qu.: 0.00000   3rd Qu.: 1.000   3rd Qu.: 0.00000
##  Max.   : 1.0000   Max.   : 1.00000   Max.   : 1.000   Max.   : 1.00000
##
##      HSDip            HSDipYr         HSGPAUnwtd         HSGPAWtd     Firs
tGen
##  Min.   :-1.0000   Min.   :  -1.0   Min.   :-1.0000   Min.   :-1   Min.
:-1
##  1st Qu.: 1.0000   1st Qu.:  -1.0   1st Qu.:-1.0000   1st Qu.:-1   1st Qu.
:-1
##  Median : 1.0000   Median :  -1.0   Median :-1.0000   Median :-1   Median
:-1
##  Mean   : 0.9643   Mean   : 557.8   Mean   : 0.1624   Mean   :-1   Mean
:-1
##  3rd Qu.: 1.0000   3rd Qu.:2010.0   3rd Qu.: 2.4000   3rd Qu.:-1   3rd Qu.
:-1
##  Max.   : 4.0000   Max.   :2016.0   Max.   : 4.0000   Max.   :-1   Max.
:-1
##
##  DualHSSummerEnroll EnrollmentStatus NumColCredAttemptTransfer
##  Min.   :0          Min.   :1.000    Min.   : -2.00
##  1st Qu.:0          1st Qu.:1.000    1st Qu.: -2.00
##  Median :0          Median :2.000    Median : 14.00
##  Mean   :0          Mean   :1.589    Mean   : 36.97
##  3rd Qu.:0          3rd Qu.:2.000    3rd Qu.: 73.00
##  Max.   :0          Max.   :2.000    Max.   :150.00
##
##  NumColCredAcceptTransfer CumLoanAtEntry      HighDeg       MathPlacement
##  Min.   :-2.00            Min.   :-2.000   Min.   :0.0000   Min.   :-1.000
0
##  1st Qu.:-2.00            1st Qu.:-2.000   1st Qu.:0.0000   1st Qu.: 0.000
0
##  Median :22.00            Median :-1.000   Median :0.0000   Median : 0.000
0
##  Mean   :31.77            Mean   :-1.411   Mean   :0.5849   Mean   : 0.279
3
##  3rd Qu.:66.00            3rd Qu.:-1.000   3rd Qu.:2.0000   3rd Qu.: 1.000
0
##  Max.   :96.00            Max.   :-1.000   Max.   :4.0000   Max.   : 1.000
0
##
##   EngPlacement     GatewayMathStatus GatewayEnglishStatus
```

```
##  Min.   :-1.0000   Min.   :0.0000    Min.   :0.0000
##  1st Qu.: 0.0000   1st Qu.:0.0000    1st Qu.:0.0000
##  Median : 0.0000   Median :0.0000    Median :0.0000
##  Mean   : 0.1869   Mean   :0.1197    Mean   :0.1902
##  3rd Qu.: 0.0000   3rd Qu.:0.0000    3rd Qu.:0.0000
##  Max.   : 1.0000   Max.   :1.0000    Max.   :1.0000
##
```

```
#Distribution of Cohort
bar1 <- ggplot(data=StudentStaticData, aes(x=Cohort)) + geom_bar(color="red",
fill=rgb(0,0,1,0.7))
bar1
```




```
#Distribution of Gender, most students were female
bar2 <- ggplot(data=StudentStaticData, aes(x=Gender)) + geom_bar(color="red",
fill=rgb(0,0,1,0.7))
bar2
```



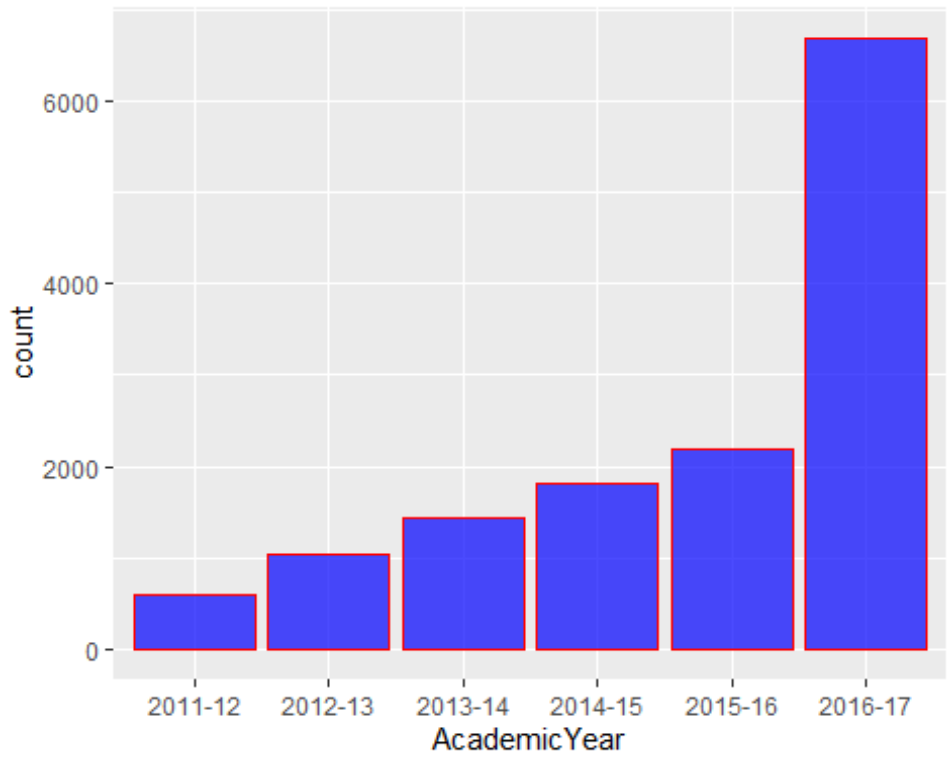
Student Progress Data

Basic descriptive statistics of the variables in the Student Progress Data

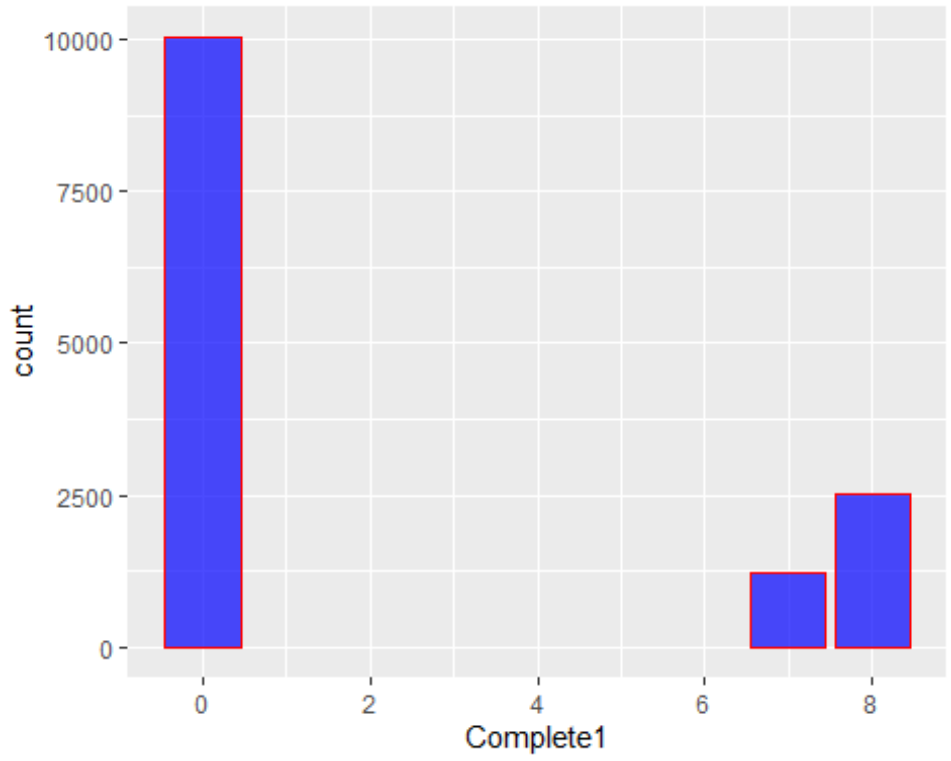
```
summary(ProgressData)

##    StudentID         Cohort            CohortTerm        Term
##  Min.   : 20932   Length:13767       Min.   :1.00   Min.   :1.000
##  1st Qu.:305676   Class :character   1st Qu.:1.00   1st Qu.:3.000
##  Median :322282   Mode  :character   Median :1.00   Median :3.000
##  Mean   :317090                      Mean   :1.45   Mean   :3.011
##  3rd Qu.:344785                      3rd Qu.:1.00   3rd Qu.:3.000
##  Max.   :364184                      Max.   :3.00   Max.   :6.000
##  AcademicYear       CompleteDevMath  CompleteDevEnglish     Major1
##  Length:13767       Min.   :-2.000   Min.   :-2.000     Min.   :-1.00
##  Class :character   1st Qu.:-2.000   1st Qu.:-2.000     1st Qu.:26.01
##  Mode  :character   Median :-2.000   Median :-2.000     Median :43.04
##                     Mean   :-1.256   Mean   :-1.414     Mean   :38.33
##                     3rd Qu.: 0.000   3rd Qu.:-1.000     3rd Qu.:51.38
##                     Max.   : 1.000   Max.   : 1.000     Max.   :54.01
##      Major2           Complete1       Complete2  CompleteCIP1    CompleteC
IP2
##  Min.   :-1.00000   Min.   :0.000   Min.   :0   Min.   :-2.00   Min.   :-2
##  1st Qu.:-1.00000   1st Qu.:0.000   1st Qu.:0   1st Qu.:-2.00   1st Qu.:-2
##  Median :-1.00000   Median :0.000   Median :0   Median :-2.00   Median :-2
##  Mean   : 0.02398   Mean   :2.081   Mean   :0   Mean   :10.52   Mean   :-2
```

```
##  3rd Qu.:-1.00000   3rd Qu.:7.000   3rd Qu.:0   3rd Qu.:23.01   3rd Qu.:-2
##  Max.   :54.01010   Max.   :8.000   Max.   :0   Max.   :54.01   Max.   :-2
##  TransferIntent DegreeTypeSought    TermGPA          CumGPA
##  Min.   :-1     Min.   :6        Min.   :0.000   Min.   :0.000
##  1st Qu.:-1     1st Qu.:6        1st Qu.:1.725   1st Qu.:2.300
##  Median :-1     Median :6        Median :3.080   Median :3.070
##  Mean   :-1     Mean   :6        Mean   :2.592   Mean   :2.778
##  3rd Qu.:-1     3rd Qu.:6        3rd Qu.:3.700   3rd Qu.:3.580
##  Max.   :-1     Max.   :6        Max.   :4.000   Max.   :4.000
##  AcademicYearID
##  Min.   : 1.00
##  1st Qu.:10.00
##  Median :15.00
##  Mean   :13.17
##  3rd Qu.:17.00
##  Max.   :18.00
```

```
#Distribution of Academic Year, most students were in the year 2016-2017
bar3 <- ggplot(data=ProgressData, aes(x=AcademicYear)) + geom_bar(color="red"
, fill=rgb(0,0,1,0.7))
bar3
```




```
#Distribution of Complete1 (Highest award received by the student during the
current term), most value = 0 mean that no award was conferred.
bar4 <- ggplot(data=ProgressData, aes(x=Complete1)) + geom_bar(color="red", f
ill=rgb(0,0,1,0.7))
bar4
```



Student Financial Aid Data

Basic descriptive statistics of the variables in the Student Financial Aid Data

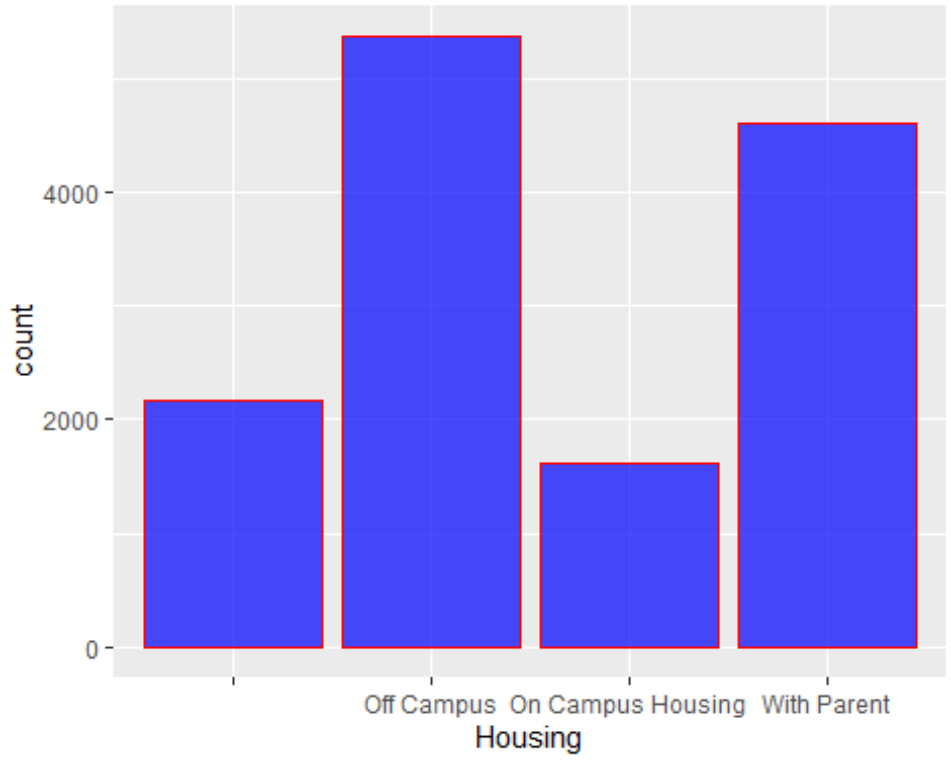
```
summary(FinancialAid)
```

```
##    StudentID         cohort            cohortterm    MaritalStatus
##  Min.   : 20932   Length:13769       Min.   :1.000   Length:13769
##  1st Qu.:305677   Class :character   1st Qu.:1.000   Class :character
##  Median :322283   Mode  :character   Median :1.000   Mode  :character
##  Mean   :317095                      Mean   :1.451
##  3rd Qu.:344790                      3rd Qu.:1.000
##  Max.   :364184                      Max.   :3.000
##
##  AdjustedGrossIncome ParentAdjustedGrossIncome FathersHighestGradeLevel
##  Min.   : -24326     Min.   :-62979            Length:13769
##  1st Qu.:      0     1st Qu.:     0            Class :character
##  Median :   2637     Median : 12372            Mode  :character
##  Mean   :  13125     Mean   : 28102
##  3rd Qu.:  16323     3rd Qu.: 38587
##  Max.   :2576425     Max.   :657631
##  NA's   :2154        NA's   :2154
##  MotherHighestGradeLevel   Housing            X2012Loan     X2012Scholarsh
ip
##  Length:13769           Length:13769       Min.   :  337   Min.   :  283
##  Class :character       Class :character   1st Qu.: 3500   1st Qu.: 2000
```

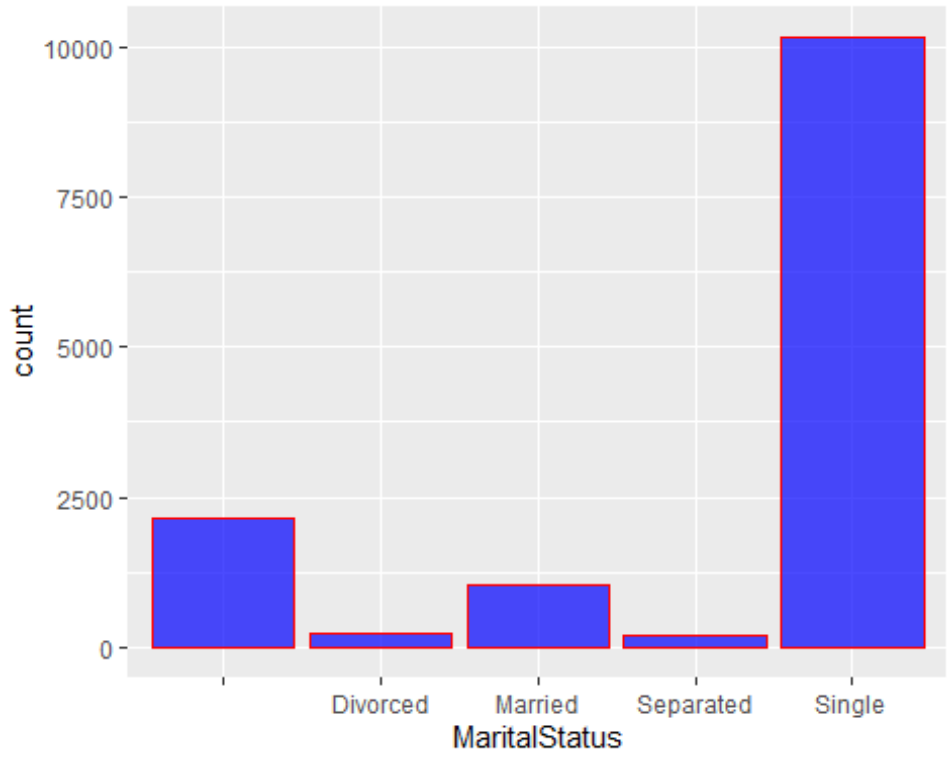
```
##  Mode  :character        Mode  :character   Median : 5500   Median : 4000
##                                            Mean   : 7169   Mean   : 5225
##                                            3rd Qu.: 9500   3rd Qu.: 6000
##                                            Max.   :55626   Max.   :27632
##                                            NA's   :12532   NA's   :13598
##  X2012Work_Study   X2012Grant         X2013Loan     X2013Scholarship
##  Min.   : 200    Min.   :   79.09   Min.   :  103   Min.   :   23
##  1st Qu.:1700    1st Qu.: 3368.25   1st Qu.: 3500   1st Qu.: 2000
##  Median :2000    Median : 5794.00   Median : 5500   Median : 3549
##  Mean   :1873    Mean   : 6660.93   Mean   : 7156   Mean   : 4793
##  3rd Qu.:2121    3rd Qu.:10714.00   3rd Qu.: 9500   3rd Qu.: 6409
##  Max.   :3000    Max.   :13263.00   Max.   :50555   Max.   :28737
##  NA's   :13666   NA's   :12415      NA's   :11582   NA's   :13459
##  X2013Work_Study   X2013Grant      X2014Loan     X2014Scholarship
##  Min.   :  25    Min.   :  162   Min.   :  128   Min.   :  100
##  1st Qu.:2000    1st Qu.: 3683   1st Qu.: 3783   1st Qu.: 2000
##  Median :2000    Median : 6089   Median : 6250   Median : 4000
##  Mean   :2084    Mean   : 7094   Mean   : 7280   Mean   : 4999
##  3rd Qu.:2200    3rd Qu.:11040   3rd Qu.:10500   3rd Qu.: 6000
##  Max.   :4000    Max.   :13790   Max.   :49845   Max.   :38851
##  NA's   :13590   NA's   :11450   NA's   :11028   NA's   :13353
##  X2014Work_Study   X2014Grant         X2015Loan     X2015Scholarship
##  Min.   :  70    Min.   :   97.24   Min.   :   25   Min.   :  200
##  1st Qu.:2000    1st Qu.: 3528.00   1st Qu.: 4162   1st Qu.: 2000
##  Median :2000    Median : 6245.00   Median : 6250   Median : 4000
##  Mean   :1933    Mean   : 7208.11   Mean   : 7241   Mean   : 4755
##  3rd Qu.:2000    3rd Qu.:11725.89   3rd Qu.:10500   3rd Qu.: 5730
##  Max.   :3300    Max.   :14001.00   Max.   :47824   Max.   :30478
##  NA's   :13526   NA's   :10840      NA's   :10718   NA's   :13174
##  X2015Work_Study   X2015Grant      X2016Loan     X2016Scholarship
##  Min.   :  10    Min.   :  209   Min.   :  103   Min.   :   28.3
##  1st Qu.:2000    1st Qu.: 3880   1st Qu.: 4500   1st Qu.: 2000.0
##  Median :2000    Median : 6358   Median : 6420   Median : 4000.0
##  Mean   :2127    Mean   : 7370   Mean   : 7625   Mean   : 4897.3
##  3rd Qu.:2800    3rd Qu.:11592   3rd Qu.:10500   3rd Qu.: 6000.0
##  Max.   :4600    Max.   :19038   Max.   :52880   Max.   :31265.5
##  NA's   :13520   NA's   :10365   NA's   :10594   NA's   :13084
##  X2016Work_Study   X2016Grant         X2017Loan     X2017Scholarship
##  Min.   :  75    Min.   :    9.69   Min.   :  103   Min.   :  100
##  1st Qu.:2000    1st Qu.: 3963.25   1st Qu.: 5354   1st Qu.: 2000
##  Median :2000    Median : 6428.00   Median : 6500   Median : 4000
##  Mean   :2036    Mean   : 7458.96   Mean   : 8256   Mean   : 5024
##  3rd Qu.:2000    3rd Qu.:11717.50   3rd Qu.:11812   3rd Qu.: 6906
##  Max.   :4000    Max.   :18505.00   Max.   :60118   Max.   :33848
##  NA's   :13497   NA's   :10075      NA's   :10445   NA's   :12784
##  X2017Work_Study   X2017Grant
##  Min.   :  45    Min.   :    0.1
##  1st Qu.:1500    1st Qu.: 4261.0
##  Median :2000    Median : 7305.0
##  Mean   :1929    Mean   : 7794.2
```

```
##  3rd Qu.:2000    3rd Qu.:12173.0
##  Max.   :3000    Max.   :19823.0
##  NA's   :13402   NA's   :9732
```

```
#Distribution of Housing, most students were living out of campus
bar5 <- ggplot(data=FinancialAid, aes(x=Housing)) + geom_bar(color="red", fil
l=rgb(0,0,1,0.7))
bar5
```




```
#Distribution of Mairital Status, most students were single
bar6 <- ggplot(data=FinancialAid, aes(x=MaritalStatus)) + geom_bar(color="red
", fill=rgb(0,0,1,0.7))
bar6
```



Data Cleaning

Data cleaning for Student Static Data

```
#Remove address columns because we won't use them for training and testing da
ta: Address1, Address2,    City,   State,  Zip,    RegistrationDate
#Remove columns because of missing most values: Campus, HSDipYr HSGPAWtd, Fir
stGen, DualHSSummerEnroll, CumLoanAtEntry,
StudentStatic <- StudentStaticData[-c(4, 5, 6, 7, 8, 9, 10, 22, 24, 25, 26, 3
0)]
#Replace value = -1 in Hispanic,    AmericanIndian, Asian,  Black,  NativeHaw
aiian, White,  TwoOrMoreRace = 0
StudentStatic["Hispanic"][StudentStatic["Hispanic"] == -1] <- 0
StudentStatic["AmericanIndian"][StudentStatic["AmericanIndian"] == -1] <- 0
StudentStatic["Asian"][StudentStatic["Asian"] == -1] <- 0
StudentStatic["Black"][StudentStatic["Black"] == -1] <- 0
StudentStatic["NativeHawaiian"][StudentStatic["NativeHawaiian"] == -1] <- 0
StudentStatic["White"][StudentStatic["White"] == -1] <- 0
StudentStatic["TwoOrMoreRace"][StudentStatic["TwoOrMoreRace"] == -1] <- 0
#Replace value = -1 in HSDip = 1 because all students completed high school b
efore applying for college
StudentStatic["HSDip"][StudentStatic["HSDip"] == -1] <- 0
#Replace values = -1 in HSGPAUnwtd = mean
StudentStatic["HSGPAUnwtd"][StudentStatic["HSGPAUnwtd"] == -1] <- mean(Studen
```

```
tStatic$HSGPAUnwtd>0)
#Replace missing values = -1, -2 in NumColCredAttemptTransfer = 0
StudentStatic["NumColCredAttemptTransfer"][StudentStatic["NumColCredAttemptTr
ansfer"] == -1] <- 0
StudentStatic["NumColCredAttemptTransfer"][StudentStatic["NumColCredAttemptTr
ansfer"] == -2] <- 0
#Replace missing values = -1, -2 in NumColCredAcceptTransfer = 0
StudentStatic["NumColCredAcceptTransfer"][StudentStatic["NumColCredAcceptTran
sfer"] == -1] <- 0
StudentStatic["NumColCredAcceptTransfer"][StudentStatic["NumColCredAcceptTran
sfer"] == -2] <- 0
#Replace missing values = -1 in MathPlacement column by majority value = 0
StudentStatic["MathPlacement"][StudentStatic["MathPlacement"] == -1] <- 0
#Replace missing values = -1 in EngPlacement column by majority value = 0
StudentStatic["EngPlacement"][StudentStatic["EngPlacement"] == -1] <- 0
```

Data cleaning for Student Progress Data

```
# Data cleaning for Student Progress Data
#Remove columns because missing data: Complete2,CompleteCIP2,TransferIntent,D
egreeTypeSought,AcademicYearID
Progress <- ProgressData[-c(11, 13, 14, 15, 18)]
#Replace missing values = -1, -2 in CompleteDevMath = 0
Progress["CompleteDevMath"][Progress["CompleteDevMath"] == -1] <- 0
Progress["CompleteDevMath"][Progress["CompleteDevMath"] == -2] <- 0
#Replace missing values = -1, -2 in CompleteDevEnglish = 0
Progress["CompleteDevEnglish"][Progress["CompleteDevEnglish"] == -1] <- 0
Progress["CompleteDevEnglish"][Progress["CompleteDevEnglish"] == -2] <- 0
#Replace missing values = -1 in Major1 = 0
Progress["Major1"][Progress["Major1"] == -1] <- 0
#Replace missing values = -1 in Major2 = 0
Progress["Major2"][Progress["Major2"] == -1] <- 0
#Replace missing values = -2 in CompleteCIP1 = 0
Progress["CompleteCIP1"][Progress["CompleteCIP1"] == -2] <- 0
```

Data cleaning for Financial Aid Data

```
# Data cleaning for Financial Aid Data
# Most of students are single, so fill the empty values of Marital Status col
umn with Single.
FinancialAid["MaritalStatus"][FinancialAid["MaritalStatus"] == ""] <- "Single
"
# Most of students live Off campus, so fill the empty values of Housing colum
n with Off Campus.
FinancialAid["Housing"][FinancialAid["Housing"] == ""] <- "Off Campus"
# Fill the empty values of parent's Highest Grade level with 'Unknown'.
FinancialAid["FathersHighestGradeLevel"][FinancialAid["FathersHighestGradeLev
el"] == ""] <- "Unknown"
FinancialAid["MotherHighestGradeLevel"][FinancialAid["MotherHighestGradeLevel
```

```
"] == ""] <- "Unknown"
# Replace all other missing values by 0
FinancialAid <- na_replace(FinancialAid, 0)
```

Merge Static Data, Progress Data, Fiancial Data

```
StaticProgressData <- merge(x=StudentStatic,y=Progress,by="StudentID")
FinancialStaticProgressData <- merge(x=StaticProgressData,y=FinancialAid, by=
"StudentID")
# Merge FinancailStaticProgressData with TrainLabels Data
StaticProgressData_Train <- merge(x=FinancialStaticProgressData,y=TrainLabels
,by="StudentID")
DataTrain <- StaticProgressData_Train[-c(2, 3, 4, 24, 25)]
DataTrain$Dropout <- as.factor(DataTrain$Dropout)
head(DataTrain)
```

```
##   StudentID BirthYear BirthMonth Hispanic AmericanIndian Asian Black
## 1     20932      1971          4        0              0     0     1
## 2     21868      1980          8        0              0     0     0
## 3     21943      1982          7        1              0     0     0
## 4     22163      1982          4        0              0     0     1
## 5     22672      1969          3        0              0     0     1
## 6     23538      1981          6        0              0     1     0
##   NativeHawaiian White TwoOrMoreRace HSDip HSGPAUnwtd EnrollmentStatus
## 1              0     0             0     1  0.2973381                2
## 2              0     1             0     1  0.2973381                2
## 3              0     0             0     1  0.2973381                2
## 4              0     0             0     1  0.2973381                2
## 5              0     0             0     1  0.2973381                2
## 6              0     0             0     1  0.2973381                2
##   NumColCredAttemptTransfer NumColCredAcceptTransfer HighDeg MathPlacement
## 1                        81                       65       0             0
## 2                        71                       66       0             0
## 3                        81                       81       0             0
## 4                        91                       81       0             0
## 5                         0                       96       0             0
## 6                         0                       79       2             0
##   EngPlacement GatewayMathStatus GatewayEnglishStatus Term AcademicYear
## 1            0                 0                    0    1      2014-15
## 2            0                 0                    0    6      2016-17
## 3            0                 0                    0    3      2012-13
## 4            0                 0                    0    3      2016-17
## 5            0                 0                    0    1      2016-17
## 6            0                 0                    0    6      2014-15
##   CompleteDevMath CompleteDevEnglish  Major1 Major2 Complete1 CompleteCIP1
## 1               0                  0  0.0000      0         0       0.0000
## 2               0                  0 23.0101      0         7      23.0101
## 3               0                  0 26.0101      0         0       0.0000
## 4               0                  0 52.0201      0         0       0.0000
## 5               0                  0 52.0801      0         0       0.0000
```

```
## 6               0                  0 51.3801      0         8      51.3801
##   TermGPA CumGPA  cohort cohortterm MaritalStatus AdjustedGrossIncome
## 1    0.00   0.00 2014-15          1       Married               52555
## 2    4.00   3.82 2014-15          1        Single               30600
## 3    0.00   0.00 2012-13          1        Single               27879
## 4    4.00   3.30 2013-14          3        Single               26794
## 5    1.85   3.21 2013-14          1        Single                   0
## 6    3.70   3.73 2013-14          3        Single               28376
##   ParentAdjustedGrossIncome FathersHighestGradeLevel MotherHighestGradeLev
el
## 1                         0                 Unknown                 Unkno
wn
## 2                         0             High School             High Scho
ol
## 3                         0                 Unknown             High Scho
ol
## 4                         0                 Unknown                 Colle
ge
## 5                         0                 Unknown                 Unkno
wn
## 6                         0                 College             High Scho
ol
##      Housing X2012Loan X2012Scholarship X2012Work_Study X2012Grant X2013Lo
an
## 1 Off Campus         0                0              0          0
0
## 2 Off Campus         0                0              0          0
0
## 3 Off Campus         0                0              0          0      49
98
## 4 Off Campus         0                0              0          0
0
## 5 Off Campus         0                0              0          0
0
## 6 Off Campus         0                0              0          0
0
##   X2013Scholarship X2013Work_Study X2013Grant X2014Loan X2014Scholarship
## 1                0               0          0         0                0
## 2                0               0          0         0                0
## 3                0               0          0         0                0
## 4                0               0          0      1650                0
## 5                0               0          0         0                0
## 6                0               0          0         0                0
##   X2014Work_Study X2014Grant X2015Loan X2015Scholarship X2015Work_Study
## 1               0          0         0                0               0
## 2               0          0      7500                0               0
## 3               0          0         0                0               0
## 4               0       1411      2300              250               0
## 5               0          0         0                0               0
## 6               0          0         0                0               0
```

```
##   X2015Grant X2016Loan X2016Scholarship X2016Work_Study X2016Grant X2017Lo
an
## 1          0         0                0               0          0
0
## 2       4260      5500                0               0       2888     125
00
## 3          0         0                0               0          0
0
## 4       3582      5079             1000               0       3610      35
00
## 5          0         0                0               0          0      62
50
## 6          0         0                0               0          0
0
##   X2017Scholarship X2017Work_Study X2017Grant Dropout
## 1                0               0          0       1
## 2                0               0          0       0
## 3                0               0          0       1
## 4             3500               0       3635       0
## 5                0               0       2181       1
## 6                0               0          0       0

rm(StaticProgressData_Train)
gc()

##           used  (Mb) gc trigger  (Mb) limit (Mb) max used  (Mb)
## Ncells 2320446 124.0    3980227 212.6         NA  3980227 212.6
## Vcells 7382153  56.4   12255594  93.6     102400 10145318  77.5
```

Methodology and Results

The data set was split in 75% train and 25% test, training the models using grid search and cross-validation on the training set and evaluating them on the test set.

```
library(caret)
intrain <- createDataPartition(DataTrain$Dropout,p=0.75,list = FALSE)
head(intrain)

##      Resample1
## [1,]         1
## [2,]         2
## [3,]         3
## [4,]         4
## [5,]         5
## [6,]         7

train1 <- DataTrain[intrain,]
head(train1)

##   StudentID BirthYear BirthMonth Hispanic AmericanIndian Asian Black
## 1     20932      1971          4        0              0     0     1
```

```
## 2     21868      1980          8        0              0     0     0
## 3     21943      1982          7        1              0     0     0
## 4     22163      1982          4        0              0     0     1
## 5     22672      1969          3        0              0     0     1
## 7     23548      1981         12        0              0     0     0
##   NativeHawaiian White TwoOrMoreRace HSDip HSGPAUnwtd EnrollmentStatus
## 1              0     0             0     1  0.2973381                2
## 2              0     1             0     1  0.2973381                2
## 3              0     0             0     1  0.2973381                2
## 4              0     0             0     1  0.2973381                2
## 5              0     0             0     1  0.2973381                2
## 7              0     1             0     1  0.2973381                2
##   NumColCredAttemptTransfer NumColCredAcceptTransfer HighDeg MathPlacement
## 1                        81                       65       0             0
## 2                        71                       66       0             0
## 3                        81                       81       0             0
## 4                        91                       81       0             0
## 5                         0                       96       0             0
## 7                        80                       49       0             0
##   EngPlacement GatewayMathStatus GatewayEnglishStatus Term AcademicYear
## 1            0                 0                    0    1      2014-15
## 2            0                 0                    0    6      2016-17
## 3            0                 0                    0    3      2012-13
## 4            0                 0                    0    3      2016-17
## 5            0                 0                    0    1      2016-17
## 7            0                 0                    0    3      2014-15
##   CompleteDevMath CompleteDevEnglish  Major1 Major2 Complete1 CompleteCIP1
## 1               0                  0  0.0000      0         0       0.0000
## 2               0                  0 23.0101      0         7      23.0101
## 3               0                  0 26.0101      0         0       0.0000
## 4               0                  0 52.0201      0         0       0.0000
## 5               0                  0 52.0801      0         0       0.0000
## 7               0                  0 52.0201      0         8      52.0201
##   TermGPA CumGPA  cohort cohortterm MaritalStatus AdjustedGrossIncome
## 1    0.00   0.00 2014-15          1       Married               52555
## 2    4.00   3.82 2014-15          1        Single               30600
## 3    0.00   0.00 2012-13          1        Single               27879
## 4    4.00   3.30 2013-14          3        Single               26794
## 5    1.85   3.21 2013-14          1        Single                   0
## 7    3.50   3.03 2012-13          1        Single               34962
##   ParentAdjustedGrossIncome FathersHighestGradeLevel MotherHighestGradeLev
el
## 1                         0                 Unknown                 Unkno
wn
## 2                         0             High School             High Scho
ol
## 3                         0                 Unknown             High Scho
ol
## 4                         0                 Unknown                 Colle
ge
```

```
## 5                          0                  Unknown                  Unkno
wn
## 7                          0              High School              High Scho
ol
##      Housing X2012Loan X2012Scholarship X2012Work_Study X2012Grant X2013Lo
an
## 1 Off Campus         0                0               0          0
0
## 2 Off Campus         0                0               0          0
0
## 3 Off Campus         0                0               0          0      49
98
## 4 Off Campus         0                0               0          0
0
## 5 Off Campus         0                0               0          0
0
## 7 Off Campus         0                0               0          0      75
00
##   X2013Scholarship X2013Work_Study X2013Grant X2014Loan X2014Scholarship
## 1                0               0          0         0                0
## 2                0               0          0         0                0
## 3                0               0          0         0                0
## 4                0               0          0      1650                0
## 5                0               0          0         0                0
## 7                0               0       5300     10500                0
##   X2014Work_Study X2014Grant X2015Loan X2015Scholarship X2015Work_Study
## 1               0          0         0                0               0
## 2               0          0      7500                0               0
## 3               0          0         0                0               0
## 4               0       1411      2300              250               0
## 5               0          0         0                0               0
## 7               0       5495     10500                0               0
##   X2015Grant X2016Loan X2016Scholarship X2016Work_Study X2016Grant X2017Lo
an
## 1          0         0                0               0          0
0
## 2       4260      5500                0               0       2888     125
00
## 3          0         0                0               0          0
0
## 4       3582      5079             1000               0       3610      35
00
## 5          0         0                0               0          0      62
50
## 7       3885         0                0               0          0
0
##   X2017Scholarship X2017Work_Study X2017Grant Dropout
## 1                0               0          0       1
## 2                0               0          0       0
## 3                0               0          0       1
```

```
## 4             3500               0       3635       0
## 5                0               0       2181       1
## 7                0               0          0       0

test1 <- DataTrain[-intrain,]
head(test1)

##    StudentID BirthYear BirthMonth Hispanic AmericanIndian Asian Black
## 6      23538      1981          6        0              0     1     0
## 9      23897      1982          5        1              0     0     0
## 11     26047      1990          9        0              0     0     1
## 17     27743      1992         12        1              0     0     0
## 19     28117      1992         10        0              0     0     1
## 21     28567      1982          4        0              0     0     1
##    NativeHawaiian White TwoOrMoreRace HSDip HSGPAUnwtd EnrollmentStatus
## 6               0     0             0     1  0.2973381                2
## 9               0     0             0     1  0.2973381                2
## 11              0     0             0     1  0.2973381                2
## 17              0     0             0     1  0.2973381                2
## 19              0     0             0     1  2.0000000                1
## 21              0     0             0     1  0.2973381                2
##    NumColCredAttemptTransfer NumColCredAcceptTransfer HighDeg MathPlacemen
t
## 6                          0                       79       2
0
## 9                         93                       66       2
0
## 11                        65                       66       0
0
## 17                       107                       66       2
0
## 19                         0                        0       0
1
## 21                       120                       96       3
0
##    EngPlacement GatewayMathStatus GatewayEnglishStatus Term AcademicYear
## 6             0                 0                    0    6      2014-15
## 9             0                 0                    0    3      2015-16
## 11            0                 0                    0    3      2016-17
## 17            0                 0                    0    3      2015-16
## 19            1                 0                    0    1      2015-16
## 21            0                 0                    0    1      2011-12
##    CompleteDevMath CompleteDevEnglish  Major1 Major2 Complete1 CompleteCIP
1
## 6               0                 0 51.3801      0         8      51.380
1
## 9               0                 0 50.0701      0         7      50.070
1
## 11              0                 0 42.0101      0         7      42.010
1
```

```
## 17               0                   0 42.0101      0         7      42.010
1
## 19               1                   0  9.0101      0         0       0.000
0
## 21               0                   0 52.0301      0         0       0.000
0
##    TermGPA CumGPA  cohort cohortterm MaritalStatus AdjustedGrossIncome
## 6     3.70   3.73 2013-14          3        Single               28376
## 9     3.81   3.84 2014-15          1        Single               14626
## 11    2.33   2.67 2014-15          1        Single               12685
## 17    2.68   2.56 2014-15          1        Single                   0
## 19    0.00   0.00 2015-16          1        Single                   0
## 21    1.54   1.54 2011-12          1        Single               25094
##    ParentAdjustedGrossIncome FathersHighestGradeLevel MotherHighestGradeLe
vel
## 6                          0                 College             High Sch
ool
## 9                          0             High School             High Sch
ool
## 11                         0                 College             High Sch
ool
## 17                         0             High School                 Coll
ege
## 19                    105950                 Unknown             High Sch
ool
## 21                         0           Middle School           Middle Sch
ool
##        Housing X2012Loan X2012Scholarship X2012Work_Study X2012Grant X2013
Loan
## 6   Off Campus         0                0              0          0
0
## 9  With Parent         0                0              0          0
0
## 11  Off Campus         0                0              0          0
0
## 17 With Parent         0                0              0          0
0
## 19  Off Campus         0                0              0          0
0
## 21  Off Campus      3750                0              0          0
0
##    X2013Scholarship X2013Work_Study X2013Grant X2014Loan X2014Scholarship
## 6                 0               0          0         0                0
## 9                 0               0          0         0                0
## 11                0               0          0         0                0
## 17                0               0          0         0                0
## 19                0               0          0         0                0
## 21                0               0          0         0                0
##    X2014Work_Study X2014Grant X2015Loan X2015Scholarship X2015Work_Study
## 6                0          0         0                0                0
```

```
## 9                0          0      4000                0               0
## 11               0          0      5500                0               0
## 17               0          0         0                0               0
## 19               0          0         0                0               0
## 21               0          0         0                0               0
##    X2015Grant X2016Loan X2016Scholarship X2016Work_Study X2016Grant X2017L
oan
## 6           0        0                0                0          0
0
## 9        6908     5500                0                0       5225
0
## 11       5280     5500                0                0       9443     12
500
## 17      12116        0                0                0       5775
0
## 19          0        0                0                0          0
0
## 21          0        0                0                0          0
0
##    X2017Scholarship X2017Work_Study X2017Grant Dropout
## 6                 0                0          0       0
## 9                 0                0          0       0
## 11                0                0      10691       0
## 17                0                0          0       0
## 19                0                0          0       1
## 21                0                0          0       1
```

```
#Create cross validation
trctrl <- trainControl(method = "cv", number = 5)
```

Fit the classification tree model

```
model1 <- train(Dropout ~., data = train1, method = "rpart", trControl=trctrl
)
predictions1 <- predict(model1, newdata = test1)
confusionMatrix(predictions1,test1$Dropout)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1845   93
##          1   36 1090
##
##                Accuracy : 0.9579
##                  95% CI : (0.9502, 0.9647)
##     No Information Rate : 0.6139
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9104
```

```
##
##  Mcnemar's Test P-Value : 8.201e-07
##
##             Sensitivity : 0.9809
##             Specificity : 0.9214
##          Pos Pred Value : 0.9520
##          Neg Pred Value : 0.9680
##              Prevalence : 0.6139
##          Detection Rate : 0.6022
##    Detection Prevalence : 0.6325
##       Balanced Accuracy : 0.9511
##
##        'Positive' Class : 0
##
```

```
bagImp1 <- varImp(model1, scale=TRUE)
bagImp1
```

```
## rpart variable importance
##
##   only 20 most important variables shown (out of 80)
##
##                                          Overall
## CompleteCIP1                            100.0000
## Complete1                                99.9018
## CumGPA                                   38.4670
## AcademicYear2016-17                      36.8324
## X2017Grant                               21.0350
## TermGPA                                  20.0160
## EnrollmentStatus                         11.6191
## BirthYear                                 3.9111
## StudentID                                 2.9725
## cohort2016-17                             2.6334
## X2012Grant                                2.3654
## cohort2015-16                             2.0686
## X2016Loan                                 1.7867
## X2016Scholarship                          1.7374
## X2013Grant                                1.6631
## ParentAdjustedGrossIncome                 1.1870
## X2016Grant                                1.1677
## X2012Loan                                 0.5082
## X2017Scholarship                          0.4335
## `FathersHighestGradeLevelMiddle School`   0.0000
```

Fit the Logistic Regression Model

```
model2 <- train(Dropout ~., data = train1, method = "glm", trControl=trctrl)
predictions2 <- predict(model2, newdata = test1)
confusionMatrix(predictions2,test1$Dropout)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1832   90
##          1   49 1093
##
##                Accuracy : 0.9546
##                  95% CI : (0.9467, 0.9617)
##     No Information Rate : 0.6139
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9037
##
##  Mcnemar's Test P-Value : 0.0006919
##
##             Sensitivity : 0.9740
##             Specificity : 0.9239
##          Pos Pred Value : 0.9532
##          Neg Pred Value : 0.9571
##              Prevalence : 0.6139
##          Detection Rate : 0.5979
##    Detection Prevalence : 0.6273
##       Balanced Accuracy : 0.9489
##
##        'Positive' Class : 0
##
```

```
bagImp2 <- varImp(model2, scale=TRUE)
bagImp2
```

```
## glm variable importance
##
##   only 20 most important variables shown (out of 77)
##
##                              Overall
## Complete1                     100.00
## `cohort2015-16`                87.61
## X2016Grant                     61.81
## `cohort2014-15`                47.16
## `AcademicYear2016-17`          39.74
## CumGPA                         37.36
## ParentAdjustedGrossIncome      35.13
## X2016Loan                      34.63
## X2016Scholarship               27.97
## EnrollmentStatus               27.67
## X2017Grant                     27.12
## CompleteDevMath                26.51
## `AcademicYear2015-16`          25.48
## `cohort2013-14`                24.26
```

```
## HSGPAUnwtd                    23.29
## X2015Loan                     22.48
## `HousingOn Campus Housing`    21.19
## MathPlacement                 21.10
## Term                          20.87
## X2012Work_Study               19.57
```

Fit the Bagging Model

```
model3 <- train(Dropout ~., data = train1, method = "treebag", trControl=trctrl)
predictions3 <- predict(model3, newdata = test1)
confusionMatrix(predictions3,test1$Dropout)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1846   78
##          1   35 1105
##
##                Accuracy : 0.9631
##                  95% CI : (0.9558, 0.9695)
##     No Information Rate : 0.6139
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9217
##
##  Mcnemar's Test P-Value : 7.782e-05
##
##             Sensitivity : 0.9814
##             Specificity : 0.9341
##          Pos Pred Value : 0.9595
##          Neg Pred Value : 0.9693
##              Prevalence : 0.6139
##          Detection Rate : 0.6025
##    Detection Prevalence : 0.6279
##       Balanced Accuracy : 0.9577
##
##        'Positive' Class : 0
##
```

```
bagImp3 <- varImp(model3, scale=TRUE)
bagImp3
```

```
## treebag variable importance
##
##   only 20 most important variables shown (out of 93)
##
##                            Overall
```

```
## CompleteCIP1               100.000
## Complete1                   99.579
## CumGPA                      39.351
## AcademicYear2016-17         35.413
## TermGPA                     28.644
## X2017Grant                  22.798
## EnrollmentStatus            12.036
## StudentID                    8.478
## BirthYear                    7.320
## BirthMonth                   3.808
## X2016Grant                   3.672
## X2016Loan                    3.619
## Major1                       3.504
## NumColCredAttemptTransfer    3.416
## ParentAdjustedGrossIncome    3.283
## cohort2016-17                3.148
## cohort2015-16                3.024
## X2013Grant                   2.991
## X2012Grant                   2.965
## NumColCredAcceptTransfer     2.923
```

Fit the SVM Radial Model

```
model4 <- train(Dropout ~., data = train1, method = "svmRadial", trControl=tr
ctrl)
predictions4 <- predict(model4, newdata = test1)
confusionMatrix(predictions4,test1$Dropout)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1860  120
##          1   21 1063
##
##                Accuracy : 0.954
##                  95% CI : (0.946, 0.9611)
##     No Information Rate : 0.6139
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9014
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9888
##             Specificity : 0.8986
##          Pos Pred Value : 0.9394
##          Neg Pred Value : 0.9806
##              Prevalence : 0.6139
##          Detection Rate : 0.6070
```

```
##    Detection Prevalence : 0.6462
##       Balanced Accuracy : 0.9437
##
##        'Positive' Class : 0
##
```

Stacking using Random Forest

```
# Construct data frame with predictions
library(caret)
predDF <- data.frame(predictions1, predictions2, predictions3, predictions4,
class = test1$Dropout)
predDF$class <- as.factor(predDF$class)
#Combine models using random forest
combModFit.rf <- train(class ~ ., method = "rf", data = predDF, distribution
= 'multinomial')
combPred.rf <- predict(combModFit.rf, predDF)
confusionMatrix(combPred.rf, predDF$class)$overall[1]
```

```
##  Accuracy
## 0.9631201
```

Compare the accuracy of each model

The performance of the classifiers is assessed using the standard measure of accuracy.

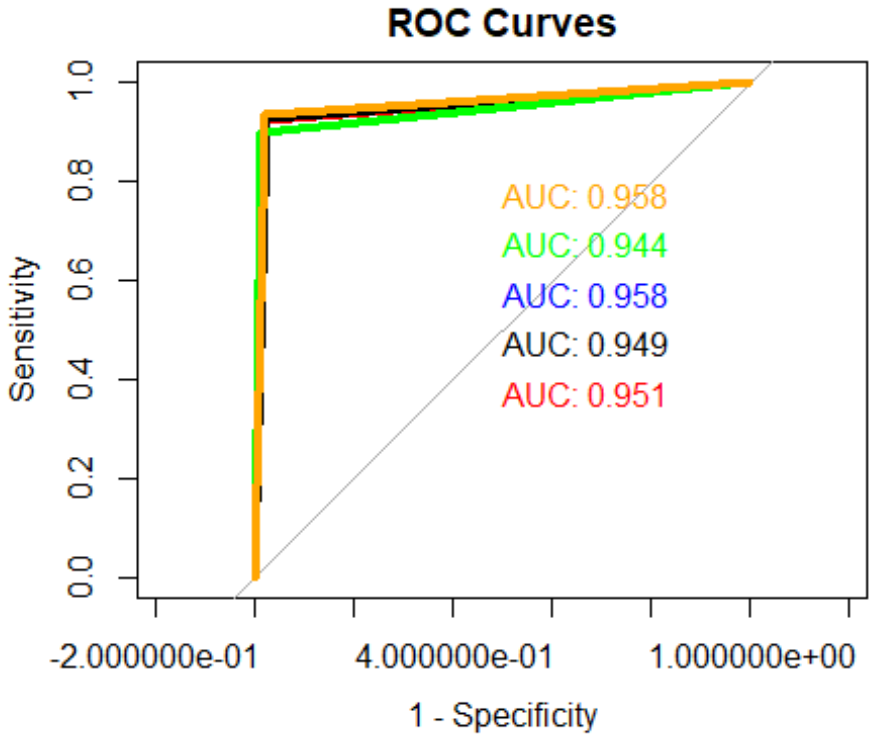
Model	Accuracy Score
Classification Tree	95.79%
Logistic Regression	95.46%
Bagging	96.31%
SVM Radial	95.4 %
Stacking with Random Forest	96.31%

Bagging and Stacking model have the higher accuracy score than the others.

#ROC Curve

```
library(pROC)
# ROC Curve
roccurve1 <- roc(test1$Dropout ~ as.numeric(predictions1))
roccurve2 <- roc(test1$Dropout ~ as.numeric(predictions2))
roccurve3 <- roc(test1$Dropout ~ as.numeric(predictions3))
roccurve4 <- roc(test1$Dropout ~ as.numeric(predictions4))
roccurve <- roc(predDF$class ~ as.numeric(combPred.rf))
roccurve$auc
```

```
## Area under the curve: 0.9577

roccurve$sensitivities

## [1] 1.0000000 0.9340659 0.0000000

roccurve$specificities

## [1] 0.0000000 0.9813929 1.0000000

plot(roccurve1, print.auc = TRUE, col = "red",print.auc.y = .4, lwd =4,legacy
.axes=TRUE,main="ROC Curves")
plot(roccurve2, print.auc = TRUE, col = "black", print.auc.y = .5, add = TRUE
, lwd =4,legacy.axes=TRUE,main="ROC Curves")
plot(roccurve3, print.auc = TRUE, col = "blue", print.auc.y = .6, add = TRUE,
lwd =4,legacy.axes=TRUE,main="ROC Curves")
plot(roccurve4, print.auc = TRUE, col = "green", print.auc.y = .7, add = TRUE
, lwd =4,legacy.axes=TRUE,main="ROC Curves")
plot(roccurve, print.auc = TRUE, col = "orange", print.auc.y = .8, add = TRUE
,lwd =4,legacy.axes=TRUE,main="ROC Curves")
```



The plot presents the ROC curves for the fine binary classifiers used in this study. The bagging model and Stacking model using Random Forest performed the same AUC and better than the Classification Tree, Logistic Regression and SVM.

#Results on TESTIDs data: Kaggle challenge

```
DatatestIDs <- merge(x = TestData, y = FinancialStaticProgressData,by = "Stud
entID")
predictions1 <- predict(model1, newdata = DatatestIDs)
predictions2 <- predict(model2, newdata = DatatestIDs)
predictions3 <- predict(model3, newdata = DatatestIDs)
predictions4 <- predict(model4, newdata = DatatestIDs)

test_predDF <- data.frame(predictions1, predictions2, predictions3, predictio
ns4)
test_combPred.rf <- predict(combModFit.rf,newdata = test_predDF)
submitfile <- data.frame(DatatestIDs$StudentID, test_combPred.rf)
colnames(submitfile) <- c("StudentID", "Dropout")

getwd()
```

```
## [1] "C:/Users/Diem My/Desktop/THI NGUYEN/HOC MY/FALL 2021/Machine Learning
1/Final Project/Code"
```

```
write.csv(submitfile,file = 'SubmissionFile9.csv')
```

Conclusion and Future Works

Conclusion

By this project, I have presented many machine learning models to predict New Jersey City student dropout. We see that these models achieve high predictive power, combining values of AUC ROC for decision-making with capable of achieving with accuracy score of over 96% in its predictions. The result was that the Bagging and Stacking with Random Forest performed the best, followed by the Classification Tree, Logistic Regression and SVM.

Limitation

Some improvements that can be made to the experiment include a more advanced solution dealing with missing values rather than replacing missing values to 0 or the majority value. For great quality to be achieved, this means there should be no missing or wrong data points in the dataset, as well as consistent and useable formatting of the data.

Developing such a model demands analytics and coding skills. These two skills, even if required, are not enough: having subject-matter experts providing input on the industry practices and interpreting results and data is crucial to success.

Future works

In this study, I limited our scope to New Jersey City students, but the same models developed for this purpose could be used for colleges, given that the models are trained and supplied with the appropriate data. Consequently, the relevant factors we have identified as impact for predicting dropout students in these models are relevant for any other college's students.