Group Outliers: Exploring Causes Behind Student Dropout Rates

Khushi Chaudhari, Gabriel Vasquez, Vincent Tang, Nicole Carter, Tiffany Tang

Contents

Libraries

Here are all of the libraries we used throughout the project.

```
# install.packages("car")
# install.packages("caret")
# install.packages("randomForest")
# install.packages("ROCR")
library(tidyverse)
library(gridExtra)
library(car)
library(MASS)
library(caret)
library(randomForest)
library(ROCR)
library(ggplot2)
library(viridis)
```

Project Description

The dataset was downloaded from this link:

https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

```
df <- read.csv("all_data.csv", header = TRUE, sep = ';')
```

Our dataset contains 37 variables, all describing information about students at the time of enrollment, such as academic path, demographics, and social-economic factors. Each instance represents a student and there are 4424 students in this dataset. Our response variable that we will use for a majority of our project is the "Target" variable that describes each student's status at the time of the survey. This is a categorical variable that contains three categories: enrolled, dropout, graduate.

This dataset also contains information on a student while they took a certain course. We will use all this information to explore our research questions.

Research Questions

- What are the biggest reason(s) that cause a student to drop out of college?
- Of the students who dropped out of college, what was their grade avg compared to those who did not drop out of college?
- Is there a relation between student's low grade avg and dropping out of college?
- Do student academic qualifications have a relation to their academic performance?
- Do males or females have a higher drop out rate? What could this possibly relate to?
- Is there a relation between scholarship holder students and their academic performance?
- Does student performance get better or worst with change in age?

Data Cleaning and Manipulation

```
## Rename Variables
rn_df <- df %>%
  rename(Age = Age.at.enrollment,
         Mart.Stat = Marital.status,
         Mom.Occp = Mother.s.occupation,
         Mom.Ed = Mother.s.qualification,
         Dad.Occp = Father.s.occupation,
         Dad.Ed = Father.s.qualification,
         Smstr1.GPA = Curricular.units.1st.sem..grade.,
         Smstr2.GPA = Curricular.units.2nd.sem..grade.,
         Nationality = Nacionality,
         Scholarship = Scholarship.holder,
         Admission.Grade = Admission.grade,
         Attendance = Daytime.evening.attendance.)

## Filter Data to a new data frame. These are the variables we're working with for this project.
fltr_df <- rn_df |>
   dplyr::select(Target, Age,
         Smstr1.GPA, Smstr2.GPA,
         Gender, Mart.Stat,
         Nationality,
```

```
        Mom.Ed, Mom.Occp,
        Dad.Ed, Dad.Occp,
        Course,
        Admission.Grade,
        Scholarship,
        Attendance) |>
  arrange(Age)
head(fltr_df)
```

```
##     Target Age Smstr1.GPA Smstr2.GPA Gender Mart.Stat Nationality Mom.Ed
## 1 Enrolled  17   12.46000   12.46000      0         1           1      3
## 2 Enrolled  17    0.00000    0.00000      1         1           1      3
## 3 Graduate  17   13.00000   12.50000      0         1           1      4
## 4 Graduate  17   13.41625   13.41625      0         1           1     19
## 5 Graduate  17   14.87500   13.42857      1         1           1      1
## 6 Graduate  18   13.30000   14.34500      0         1           1     19
##   Mom.Occp Dad.Ed Dad.Occp Course Admission.Grade Scholarship Attendance
## 1        2     19        3   9500           133.8           0          1
## 2        1     39        3    171           133.5           0          1
## 3        4      1        4   9070           132.0           1          1
## 4        1     19        4   9500           133.3           1          1
## 5        9      1        8   9670           149.5           1          1
## 6        7     38       10   9500           128.4           1          1
```

Creating a new data set, filtering and adding new variables. Converting some numerical variables into categorical variables with new names. Some changes include: Factoring attendance, gender, scholarship. Changing GPA scale from 0-20 to 0-5. Creating a new average gpa category that averages both semesters. Creating a categorical version that holds the letter grades. Create an age group variable. Updating occupation variables.

```
##  [1] pt  ukn br  it  cv  gm  cub sp  rus sm  du  tk  gui col eng lit rom mx  ang
## [20] mz  mld
## 21 Levels: pt gm sp it du eng lit ang cv gui mz sm tk br rom mld mx ukn ... col
```

Write this new data file to a csv file.

```
##   X   Target Age Smstr1.GPA Smstr2.GPA  GPA Gender Mart.Stat Nationality Mom.Ed
## 1 1 Enrolled  17       3.12       3.12 3.12 Female         1          pt      3
## 2 2 Enrolled  17       0.00       0.00 0.00   Male         1          pt      3
## 3 3 Graduate  17       3.25       3.12 3.18 Female         1          pt      4
## 4 4 Graduate  17       3.35       3.35 3.35 Female         1          pt     19
## 5 5 Graduate  17       3.72       3.36 3.54   Male         1          pt      1
## 6 6 Graduate  18       3.33       3.59 3.46 Female         1          pt     19
##   Mom.Occp Dad.Ed Dad.Occp Course Admission.Grade Scholarship Attendance
## 1        2     19        3   9500           133.8          No    Daytime
## 2        1     39        3    171           133.5          No    Daytime
## 3        4      1        4   9070           132.0         Yes    Daytime
## 4        1     19        4   9500           133.3         Yes    Daytime
## 5        9      1        8   9670           149.5         Yes    Daytime
## 6        7     38       10   9500           128.4         Yes    Daytime
##   Age.Group Grade
## 1         1     B
```

```
## 2          1      F
## 3          1      B
## 4          1     B+
## 5          1     B+
## 6          1     B+
```

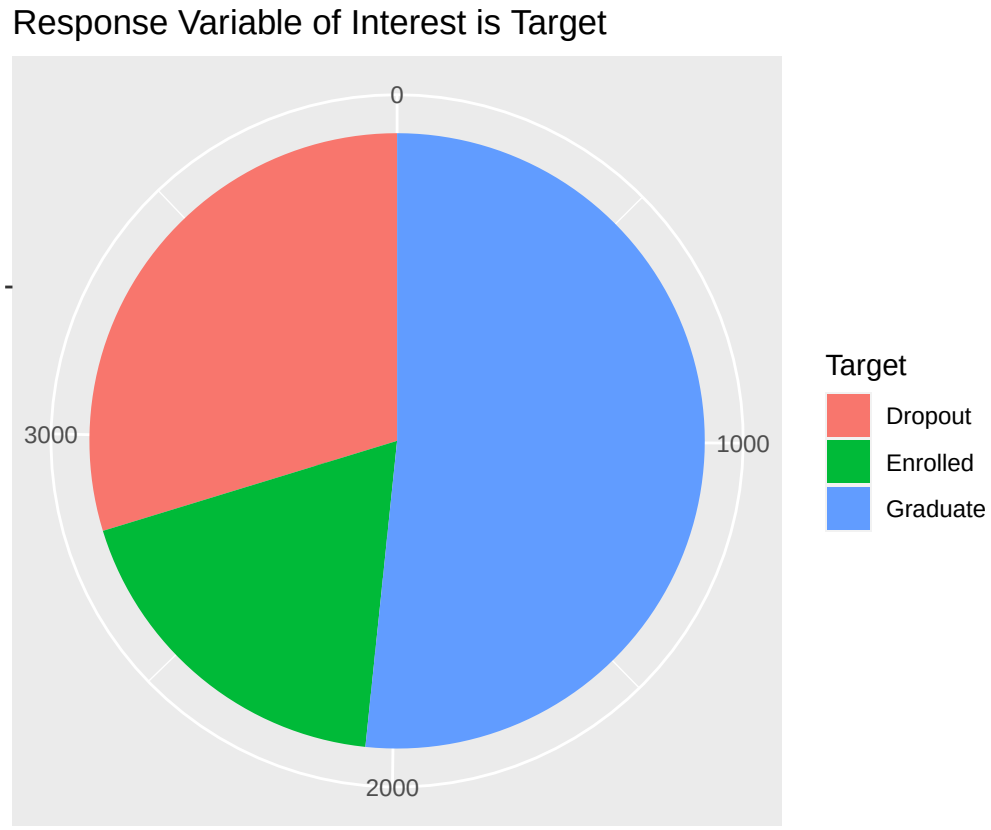
Important Variables we're using:

- Age - The student's Age during their enrollment. Many unique values. (ranges from 17 to 34)
- Age.Group - The students are categorized by age groups. We're only including 4 age groups. Each group ranges 5 years except the first age group, "Young Adults".
- Smstr1.GPA - The student's GPA from their 1st semester. Many unique values. (ranges from 0.0 to 4.72)
- Smstr2.GPA - The student's GPA from their 2nd semester. Many unique values. (ranges from 0.0. to 4.64)
- GPA - The student's overall GPA. This is calculated by averaging the student's semester 1 & 2 GPA. (ranges from 0.0 to 4.57)
- Gender - Male or Female. (2 values)
- Mart.Stat - The student's marital status. (6 values)
- Nationality - The student's ethnicity or national origin. (21 values)
- Mom.Ed - The student's mother's education level. (28 values)
- Mom.Occp - The student's mother's occupation or job. (32 values)
- Dad.Ed - The student's father's education level. (33 values)
- Dad.Occp - The sutdent's father's occupation or job (44 values)
- Scholarship - Yes the student is receiving a scholarship or No the student is not receiving a scholarship. (2 values)

Exploratory Data Analysis and Modeling

We will try to use EDA and different types of modeling to answer our research questions.

Target Variable



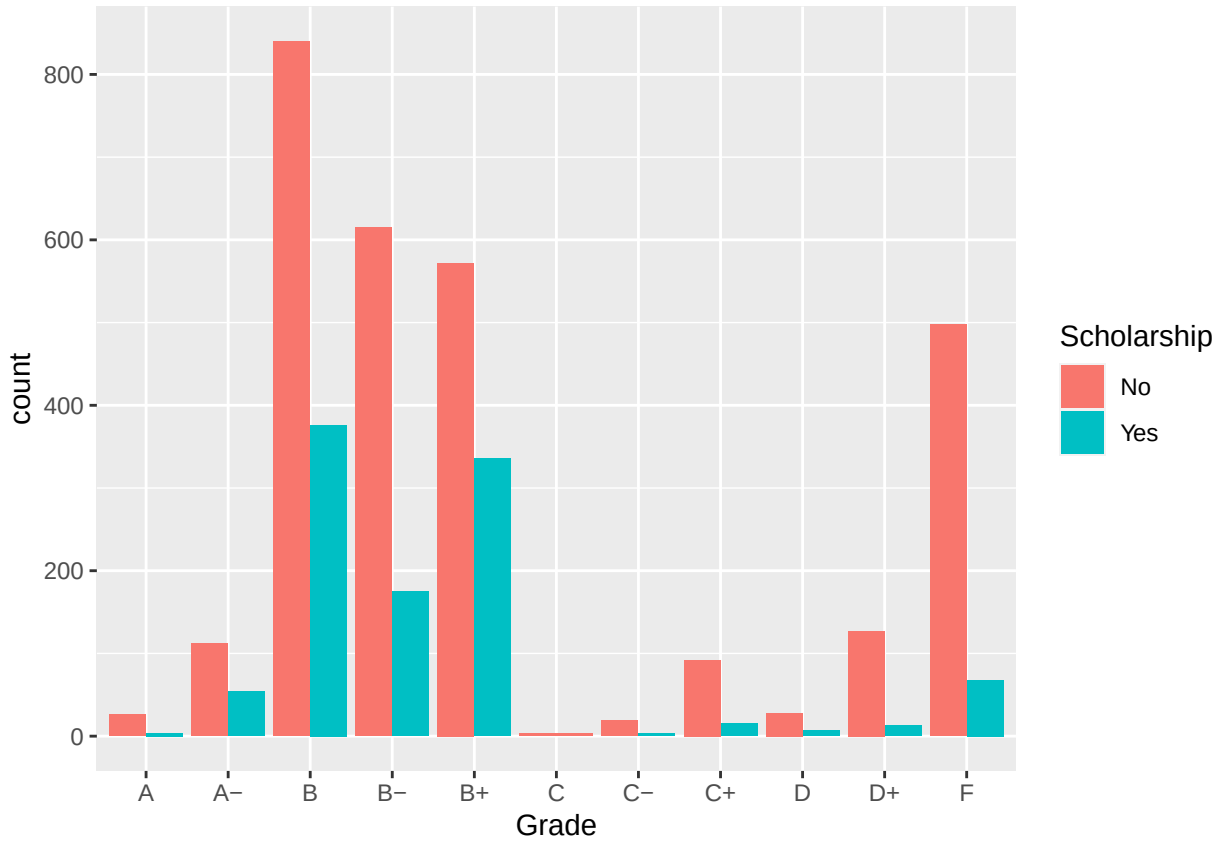
Proportion of Target values in the sample data

```
## # A tibble: 3 x 3
##   Target   Total Ratio
##   <chr>    <int> <dbl>
## 1 Dropout   1184 0.297
## 2 Enrolled   742 0.186
## 3 Graduate  2057 0.516
```

Scholarship Holders and Academic Performance

Is there a relationship between scholarship holder students and their academic performance?

Based on the linear model and our graph, we see that GPA is a significant predictor for whether a student has a scholarship because it has a low p-value. Students with a higher GPA have a higher chance at receiving a



scholarship.

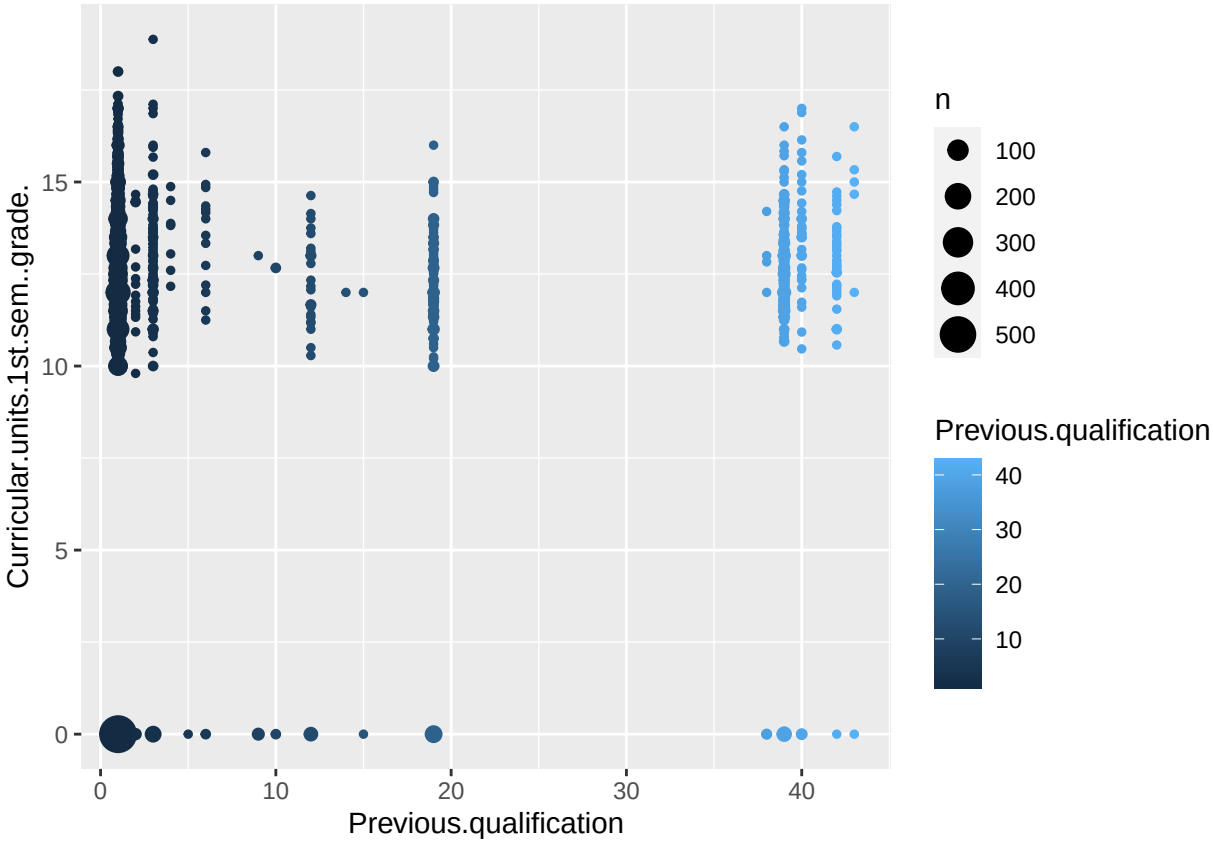
```
##   X   Target Age Smstr1.GPA Smstr2.GPA  GPA Gender Mart.Stat Nationality Mom.Ed
## 1 1 Enrolled  17       3.12       3.12 3.12 Female         1          pt      3
## 2 2 Enrolled  17       0.00       0.00 0.00   Male         1          pt      3
## 3 3 Graduate  17       3.25       3.12 3.18 Female         1          pt      4
## 4 4 Graduate  17       3.35       3.35 3.35 Female         1          pt     19
## 5 5 Graduate  17       3.72       3.36 3.54   Male         1          pt      1
## 6 6 Graduate  18       3.33       3.59 3.46 Female         1          pt     19
##   Mom.Occp Dad.Ed Dad.Occp Course Admission.Grade Scholarship Attendance
## 1        2     19        3   9500           133.8          No    Daytime
## 2        1     39        3    171           133.5          No    Daytime
## 3        4      1        4   9070           132.0         Yes    Daytime
## 4        1     19        4   9500           133.3         Yes    Daytime
## 5        9      1        8   9670           149.5         Yes    Daytime
## 6        7     38       10   9500           128.4         Yes    Daytime
##   Age.Group Grade Scholarship_Int
## 1         1     B               0
## 2         1     F               0
## 3         1     B               1
## 4         1    B+               1
## 5         1    B+               1
## 6         1    B+               1

##
## Call:
## glm(formula = Scholarship_Int ~ GPA, data = readTest)
```

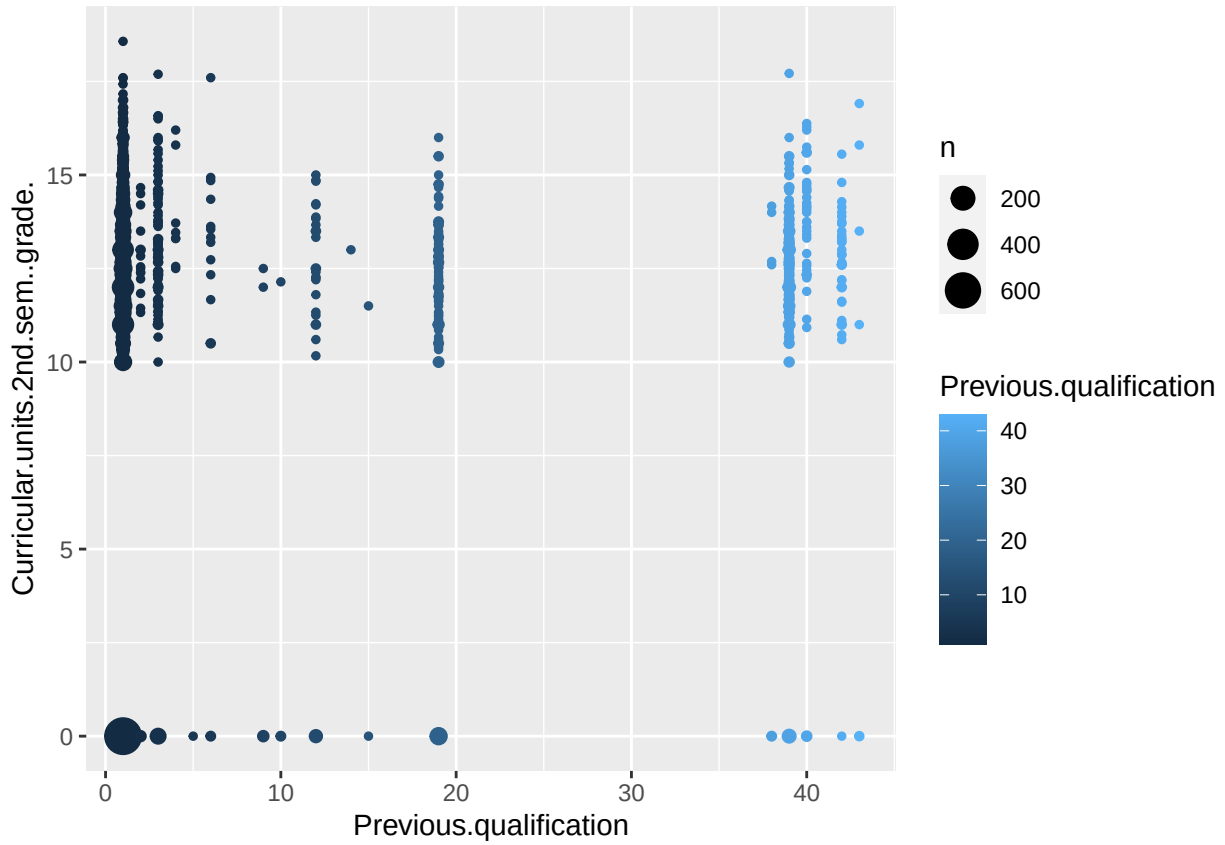
```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3906  -0.2987  -0.2743   0.6669   0.9114
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.088553   0.016965    5.22 1.88e-07 ***
## GPA         0.066097   0.005847   11.30  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1882967)
##
##     Null deviance: 773.67  on 3982  degrees of freedom
## Residual deviance: 749.61  on 3981  degrees of freedom
## AIC: 4656.7
##
## Number of Fisher Scoring iterations: 2
```

Previous Qualifications and Academic Performance

In addition to getting information about the target variable, we also wanted to see if we can make any conclusions about other interesting factors. For this instance, we looked into if there is any relationship between a student's highest achieved level of education with their academic performance in undergraduate studies.

In order to achieve this, first we created plots to evaluate if `previous qualification` has any visual effects on both 1st and 2nd semesters.


Curricular.units.1st.sem..grade.
15
10
5
0
0
10
20
30
40
Previous.qualification
n
100
200
300
400
500
Previous.qualification
40
30
20
10



Based on the above plots, we can make the conclusion that there are no clear patterns for qualification and semester GPA's. An interesting take-away from the plots is that those who had previous qualification of High School have the highest number of students, and their GPA is nearly 0. The largest number of students are represented by characteristics: Previous Qualification = High School, and GPA = 0 (Bottom left points). From this, we can deduce that most of the students in the data just finished high school and are entering undergraduate studies, meaning their undergraduate GPA hasn't been formulated yet.

We can also run statistical analyses to quantitatively prove if a relationship exists.

Correlation Coefficients:

```
## [1] "1st Semester vs. Previous Qualification Coefficient:  -0.000496666070915359"
```

```
## [1] "2nd Semester vs. Previous Qualification Coefficient:  0.000941881532783087"
```

We can identify very small correlation values comparing 1st and 2nd semester GPA to a student's previous qualification. In specific, 1st semester actually has a slight negative correlation to previous qualification, which was expected, since you would think that in the first semester, a student starting out in higher level education would have good grades, compared to those who have already started an undergraduate degree.

Linear Modeling:

In addition to correlation coefficients, we can conduct linear regression models to analyze the significance values.

```
qual.model1 <- lm(Curricular.units.1st.sem..grade. ~ Previous.qualification, data = df)
summary(qual.model1)
```

```
##
## Call:
## lm(formula = Curricular.units.1st.sem..grade. ~ Previous.qualification,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6417   0.3583   1.6441   2.7583   8.2338
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.6418995  0.0798092 133.342   <2e-16 ***
## Previous.qualification -0.0002355  0.0071295  -0.033    0.974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.844 on 4422 degrees of freedom
## Multiple R-squared:  2.467e-07,  Adjusted R-squared:  -0.0002259
## F-statistic: 0.001091 on 1 and 4422 DF,  p-value: 0.9737
```

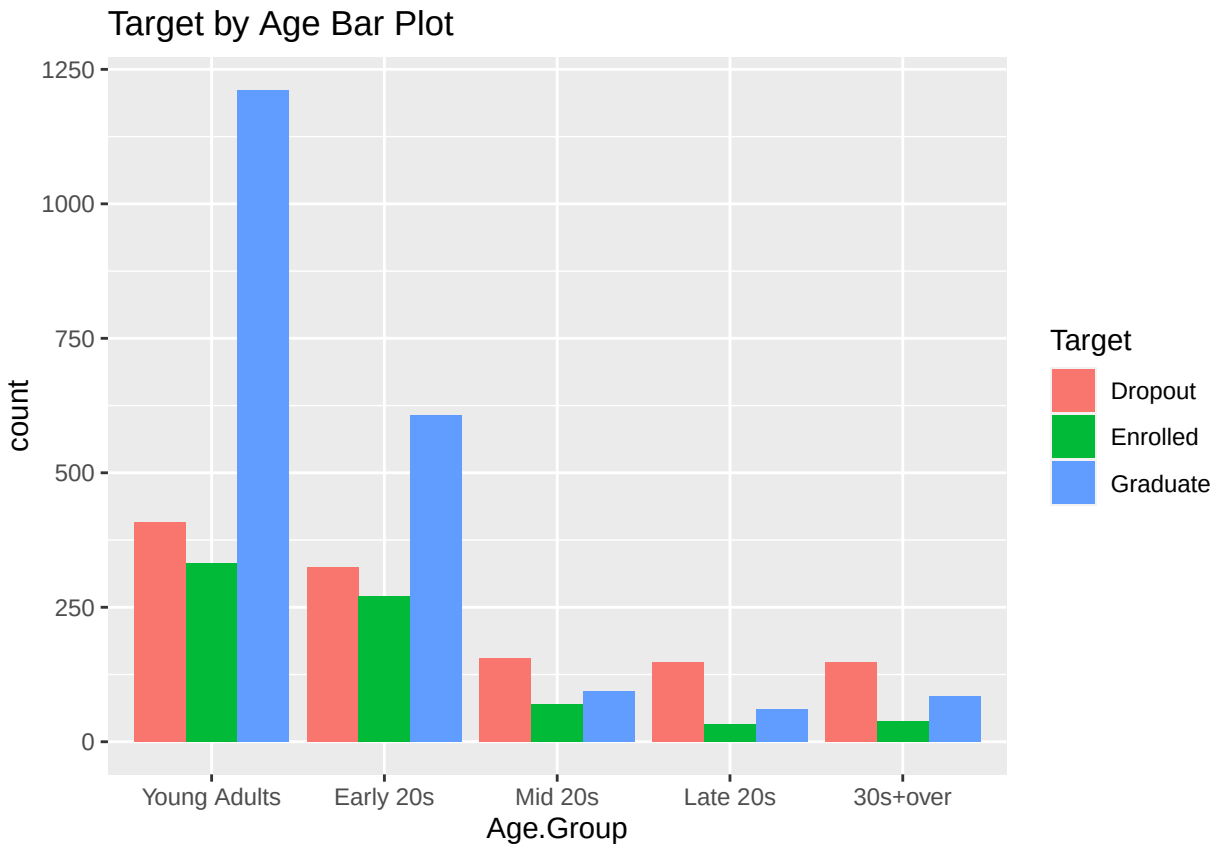
```
qual.model2 <- lm(Curricular.units.1st.sem..grade. ~ Previous.qualification, data = df)
summary(qual.model2)
```

```
##
## Call:
## lm(formula = Curricular.units.1st.sem..grade. ~ Previous.qualification,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6417   0.3583   1.6441   2.7583   8.2338
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.6418995  0.0798092 133.342   <2e-16 ***
## Previous.qualification -0.0002355  0.0071295  -0.033    0.974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.844 on 4422 degrees of freedom
## Multiple R-squared:  2.467e-07,  Adjusted R-squared:  -0.0002259
## F-statistic: 0.001091 on 1 and 4422 DF,  p-value: 0.9737
```

For both models, we can see that the p-values are extremely high, and both R-Squared values are significantly small, leading to the conclusion that there is no significant relationship between a student's previous qualifications and their academic performance in both semesters.

Student Age and Grades vs Dropout Rates

Target by Age Bar Plot



Dropout avg age vs Non-Dropout avg age Table

```
stdnt_avg_age <- fltr_df |>
  dplyr::select(Target, Age) |>
  mutate(Student = ifelse(Target=="Dropout", "Dropout", "Non-Dropout")) |>
  group_by(Student) |>
  summarize(Total = n(), Avg_Age = mean(Age))

table_slide3 <- data.frame(Student = stdnt_avg_age$Student, Avg_Age = stdnt_avg_age$Avg_Age)
table_slide3
```

```
##       Student  Avg_Age
## 1     Dropout 22.88514
## 2 Non-Dropout 20.45731
```

We will perform t-tests to analyze the relationship between age and Target, as well as a few other variables. First we create a new dataset with our variables of interest.

Age vs Target: t-test and ANOVA

```
t_test_age <- t.test(Age ~ Target, data = t_test_df)
print(t_test_age)
```

```
##
##  Welch Two Sample t-test
##
## data:  Age by Target
## t = -15.797, df = 2475.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Graduate and group Dropout is not equa
## 95 percent confidence interval:
##  -4.817320 -3.753387
## sample estimates:
## mean in group Graduate  mean in group Dropout
##               21.78361                26.06897
```

```
anova_results <- aov(Age ~ Target, data = t_test_df)
summary(anova_results)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## Target         1  15880   15880     279 <2e-16 ***
## Residuals   3628 206496      57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
tukey_results <- TukeyHSD(anova_results)
tukey_results
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Age ~ Target, data = t_test_df)
##
## $Target
##                       diff      lwr      upr p adj
## Dropout-Graduate 4.285353 3.782347 4.788359     0
```

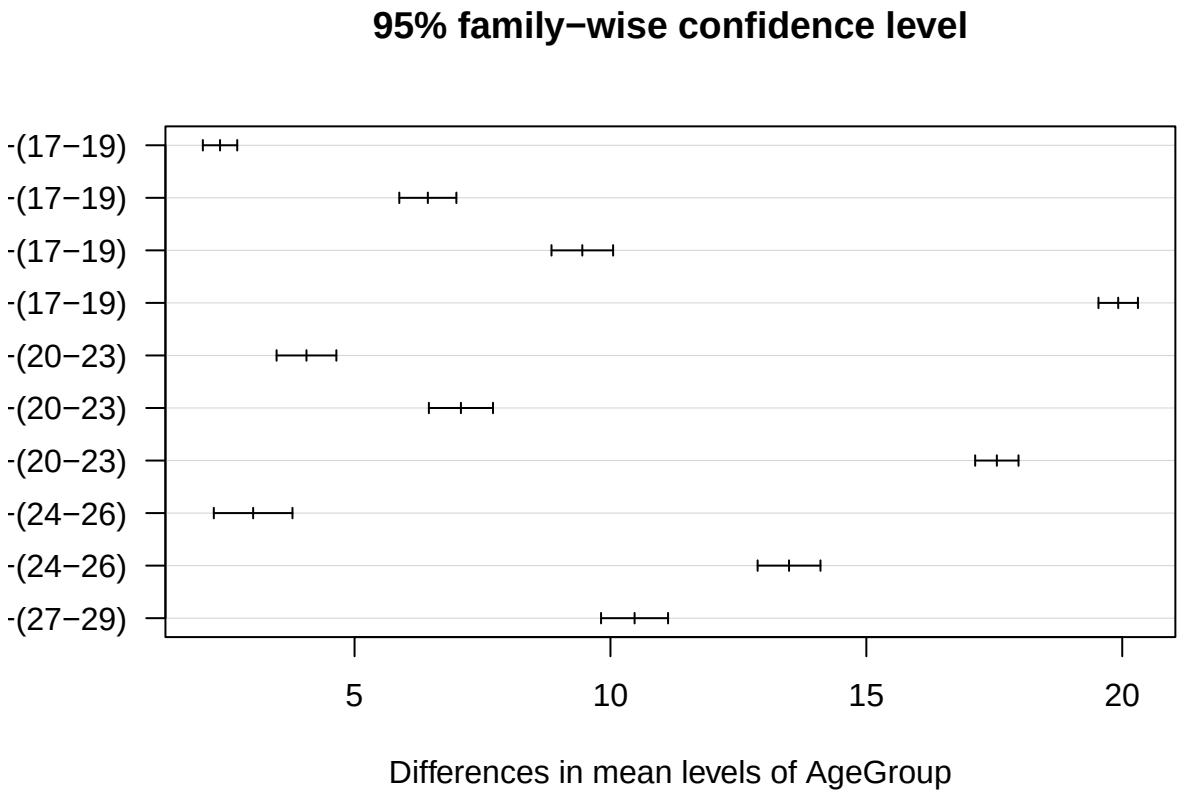
```
dropout_df <- t_test_df %>%
  mutate(AgeGroup = cut(Age, breaks = c(17, 20, 24, 27, 30, Inf), right = FALSE,
                        labels = c("(17-19)", "(20-23)", "(24-26)", "(27-29)", "(30 Over)")))
anova_age_group <- aov(Age ~ AgeGroup, data = dropout_df)
summary(anova_age_group)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## AgeGroup       4 189860   47465    5292 <2e-16 ***
## Residuals   3625  32516       9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
tukey_age_group <- TukeyHSD(anova_age_group)
print(tukey_age_group)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Age ~ AgeGroup, data = dropout_df)
##
## $AgeGroup
##                          diff       lwr       upr p adj
## (20-23)-(17-19)     2.371470  2.035461  2.707478     0
## (24-26)-(17-19)     6.431867  5.874538  6.989196     0
## (27-29)-(17-19)     9.450167  8.848152 10.052183     0
## (30 Over)-(17-19) 19.921569 19.535820 20.307318     0
## (24-26)-(20-23)     4.060397  3.476372  4.644422     0
## (27-29)-(20-23)     7.078698  6.451886  7.705509     0
## (30 Over)-(20-23) 17.550099 17.126694 17.973504     0
## (27-29)-(24-26)     3.018300  2.249791  3.786809     0
## (30 Over)-(24-26) 13.489702 12.875710 14.103693     0
## (30 Over)-(27-29) 10.471402  9.816579 11.126225     0
```

```r
plot(tukey_age_group, las = 1)
```



```r
average_age_dropout <- mean(dropout_df$Age, na.rm = TRUE)
average_age_dropout
```

```
## [1] 23.46116
```

There is a highly significant difference in the age of students who graduate and those who drop out. The average age of dropouts is significantly higher than that of graduates, suggesting that older students are more likely to drop out.

Anova: Since Pvalue is less than 0.05, it indicates that there is a significant difference in the ages between the groups. We see a average age difference between dropouts and graduates is approximately 4.29 years. With a 95% CI, the difference in means ranges from about 3.78 to 4.79 years.

Admission Grade vs Target

```
##
##  Welch Two Sample t-test
##
## data:  Admission.Grade by Target
## t = 7.6565, df = 2869.4, p-value = 2.593e-14
## alternative hypothesis: true difference in means between group Graduate and group Dropout is not equa
## 95 percent confidence interval:
##  2.851440 4.814693
## sample estimates:
## mean in group Graduate  mean in group Dropout
##               128.7944               124.9614
```

There is a highly significant difference in admission grades between graduates and dropouts. Students who graduate have significantly higher admission grades compared to those who drop out, indicating that initial academic performance is a strong predictor of graduation.

Semester 1 and 2 GPA vs Target

```
##
##  Welch Two Sample t-test
##
## data:  Smstr1.GPA by Target
## t = 31.691, df = 1790.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Graduate and group Dropout is not equa
## 95 percent confidence interval:
##  5.053606 5.720392
## sample estimates:
## mean in group Graduate  mean in group Dropout
##              12.643655               7.256656
```

There is a highly significant difference in first semester GPAs between graduates and dropouts. Students who graduate have significantly higher GPAs in their first semester than those who drop out, suggesting that early academic performance is crucial for student retention.

```
##
##  Welch Two Sample t-test
```

```
##
## data:  Smstr2.GPA by Target
## t = 39.504, df = 1776.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Graduate and group Dropout is not equa
## 95 percent confidence interval:
##  6.460434 7.135440
## sample estimates:
## mean in group Graduate  mean in group Dropout
##               12.697276                5.899339
```

There is a highly significant difference in second semester GPAs between graduates and dropouts. Students who graduate have significantly higher GPAs in their second semester than those who drop out, reinforcing the importance of sustained academic performance for successful completion of studies.

Biggest Reasons Students Drop Out

We will first create a new data set that we need specifically for our models. We will only consider dropouts and graduates for the sake of the model and the question we are trying to answer (what factors lead to a student dropping out). We want to analyze what predictors most influence our response variable, based on our research questions.

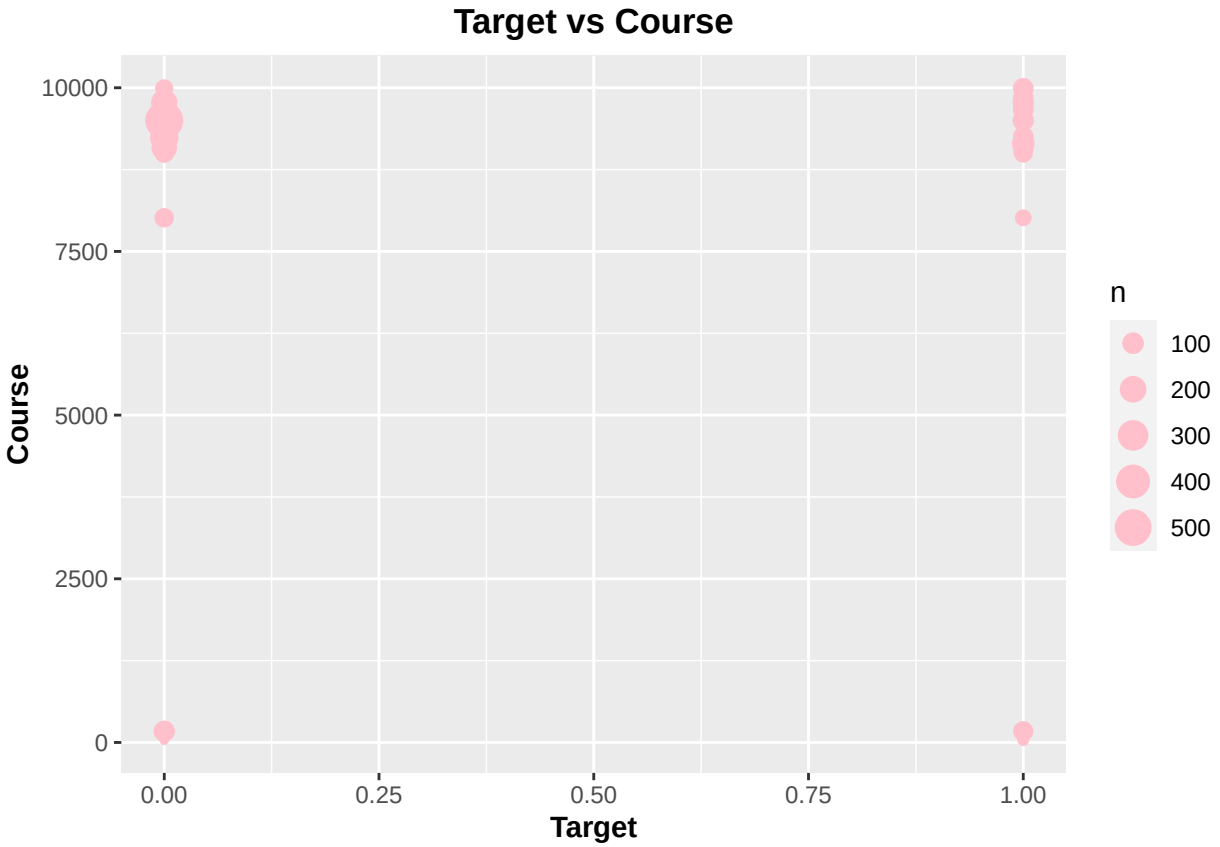
```
mod_df <- fltr_df %>%
  filter(Target == "Dropout" | Target == "Graduate") %>%
  dplyr::select(Target, Age, GPA, Gender, Mom.Ed, Mom.Occp, Dad.Ed, Dad.Occp,
                Course, Admission.Grade, Scholarship, Attendance)
mod_df$Target <- ifelse(mod_df$Target == "Dropout", 1, 0)
```

Let's take a look at the distributions of each of our variables of interest against our Target variable. We are selecting about 9 out of our 18 variables to use in two models: logistic regression and random forest.

We won't need to look at the distributions for variables we explored earlier. Based on our previous modeling and EDA, age and scholarship are significant variables that we want to include in our preliminary model. Based on previous t-tests, we know that admission grade and GPA are also significant variables in determining whether students drop out.

Let's take a look at the rest of the variables we can use. We will use variables that either have two categories or they are numerical continuous/discrete. We will not use nominal discrete variables for classification modeling because we cannot order them in a reasonable format and would like to explore the variables that we can.

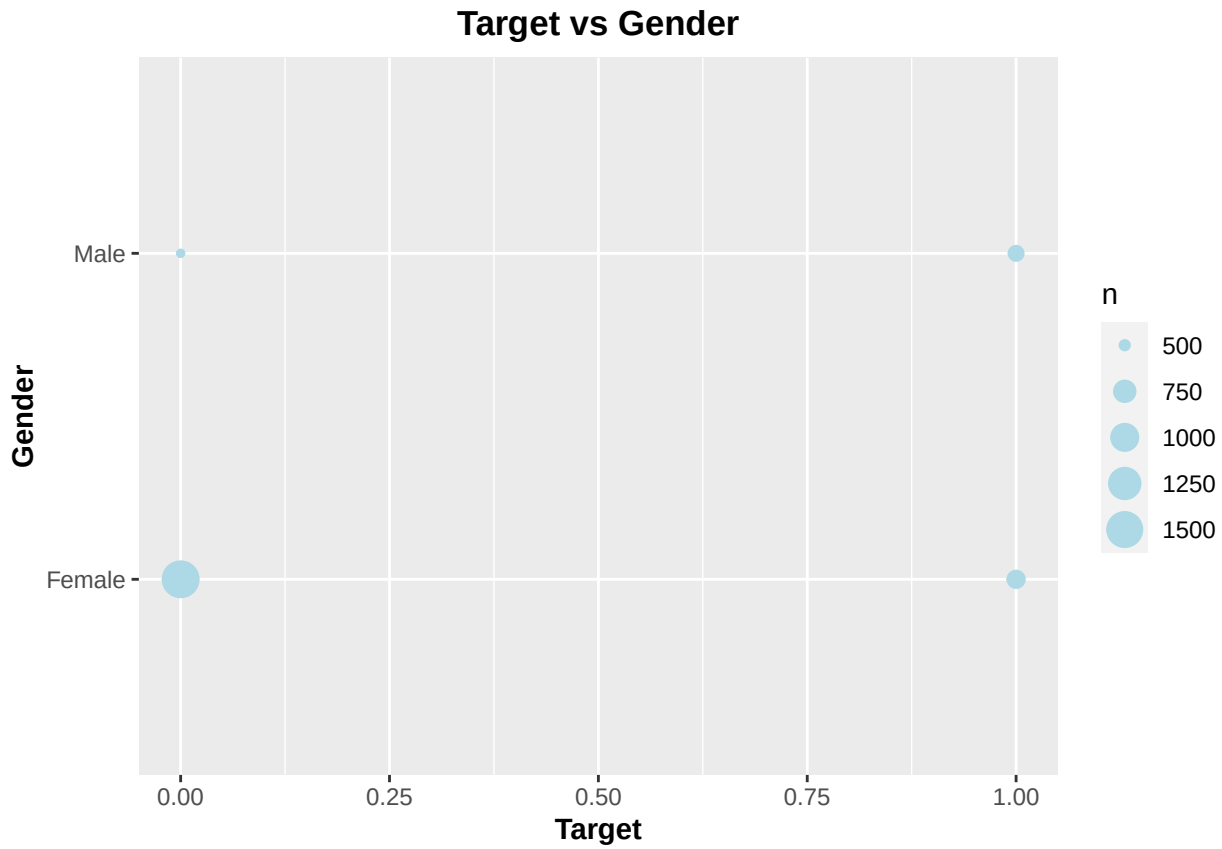
```
ggplot(mod_df) +
  geom_count(aes(x = Target, y = Course), color = "pink")+
  labs(title = "Target vs Course")+
  scale_color_viridis_c(option = "plasma")+
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        axis.title = element_text(face = "bold"),
        legend.position = "right")
```




```
print(paste("Correlation Coefficient between Target and Course: ", cor(mod_df$Course, as.numeric(mod_df$
```

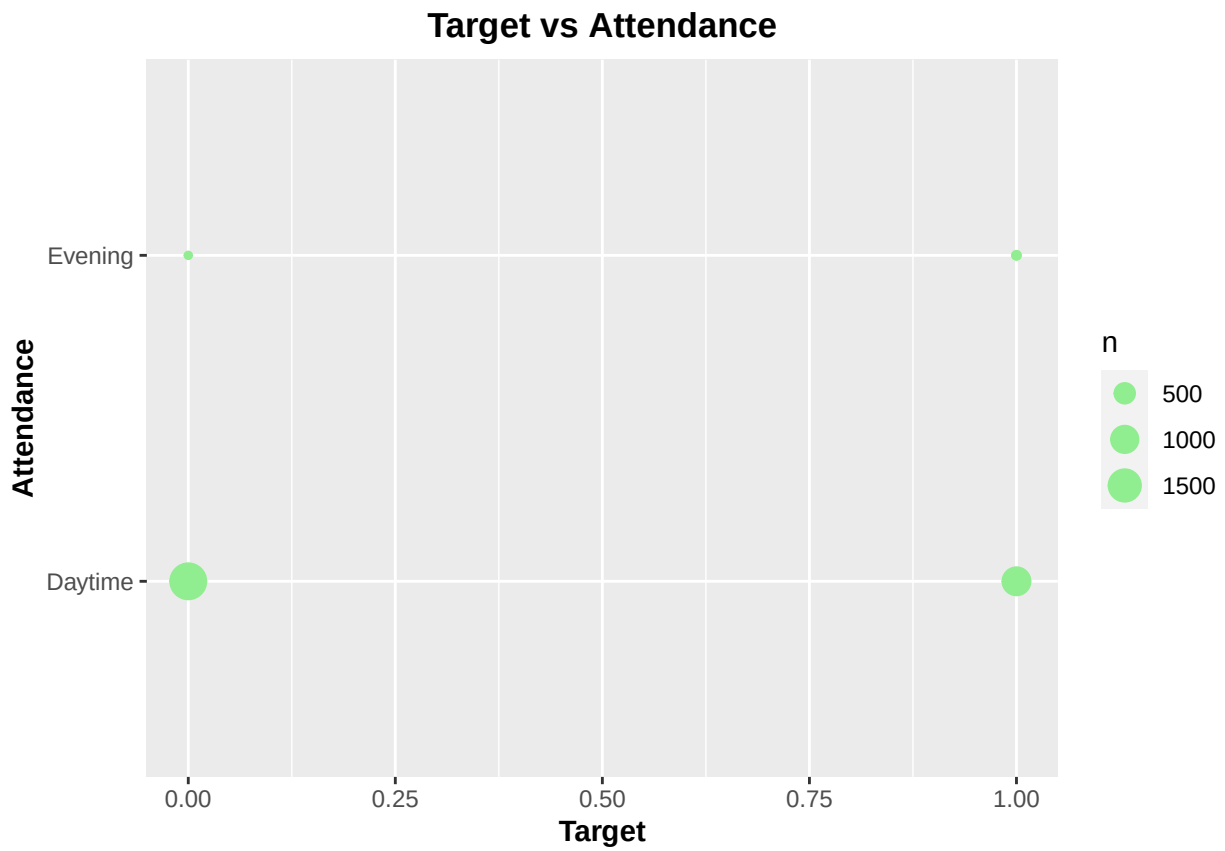
```
## [1] "Correlation Coefficient between Target and Course:  -0.0477309685199663"
```

Based on this graph, we don't see any visible variation between the different courses on the target variable. The distributions looks very similar. Based on our correlation coefficient, there is very little correlation between the two variables because the number is close to 0.



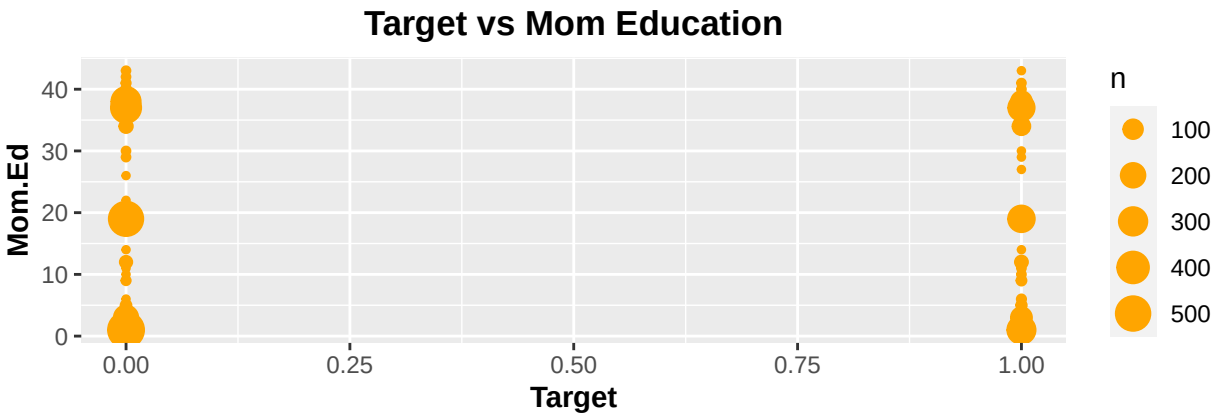

```
## [1] "Correlation Coefficient between Target and Gender:  0.245590892375208"
```

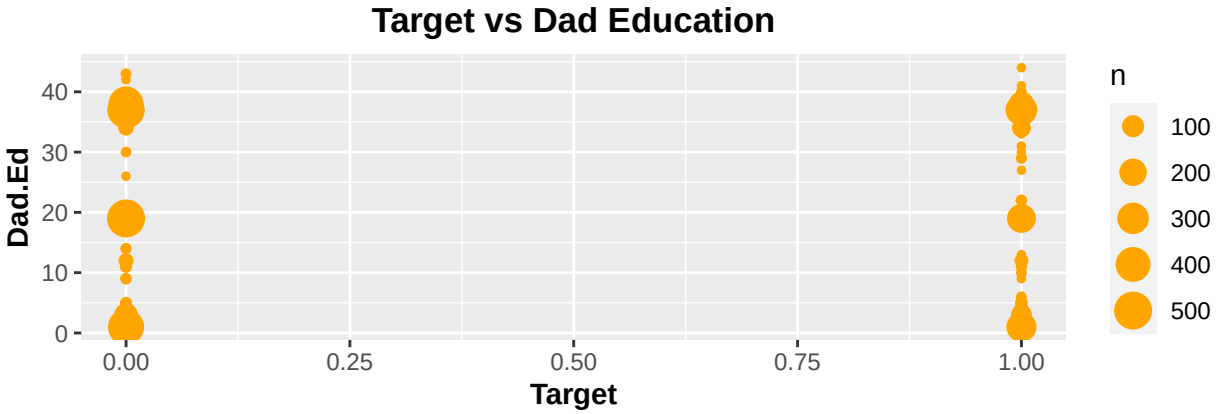
Based on the graph, there seems to be some correlation between gender and Target. There are more females that are more likely to graduate. Based on the correlation coefficient, there is some positive correlation between the two variables.




```
## [1] "Correlation Coefficient between Target and Attendance:  0.0853815296292931"
```

Based on this graph, we don't see any visible variation between the attendance and target. The distributions looks very similar. Based on our correlation coefficient, there is very little correlation between the two variables because the number is close to 0.

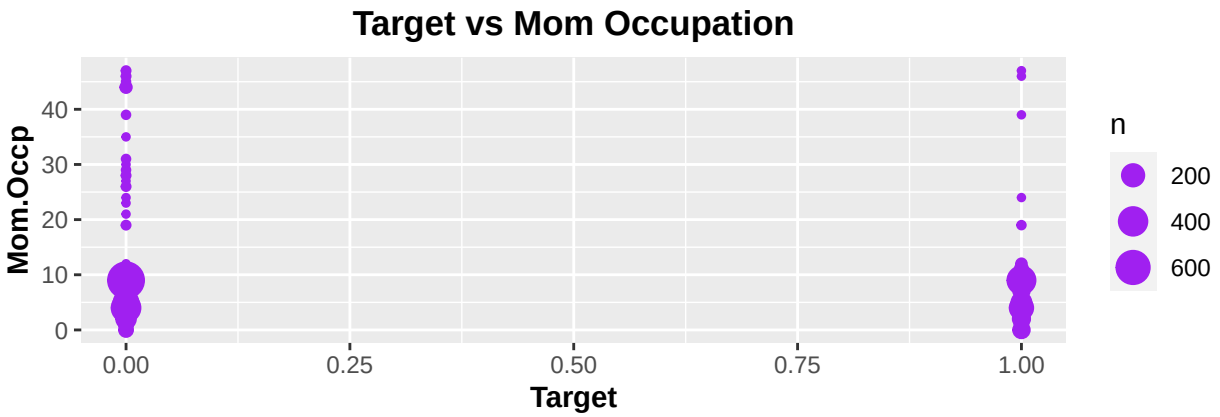


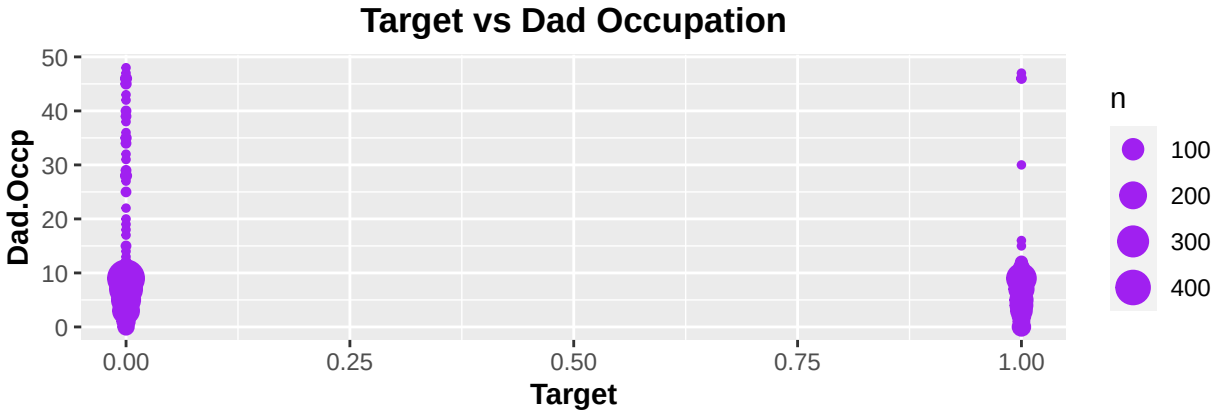



```
## [1] "Correlation Coefficient between Target and Mom Education:  0.0164539088819473"
```

```
## [1] "Correlation Coefficient between Target and Dad Education:  -0.0192256820313096"
```

Based on both of these graphs, we don't see any visible variation between the parents' education and the target variable. The distributions looks very similar. Based on our correlation coefficients for both of these variables, there is very little correlation between the two variables because the numbers are close to 0.






```
## [1] "Correlation Coefficient between Target and Mom Occupation:  -0.0757384106838734"
```

```
## [1] "Correlation Coefficient between Target and Dad Occupation:  -0.0743177144558512"
```

Based on these graphs, we can see that there are slightly different distributions between the variables for each of the parent's occupations. But the correlation coefficients are very low. For the sake of being safe because the distributions are visibly different, we will keep these variables in our models.

Multiple Logistic Regression Model

This preliminary model has all of our variables. It is not a very good model just based on the fact that some of our predictors have very high p-values. Such high p-values indicate that the variables are not statistically significant, so we want a model that only includes variables that are statistically significant.

```
mod_log <- glm(Target ~ Age + Gender + Mom.Occp + Dad.Occp
               + Admission.Grade + Scholarship + GPA,
               mod_df, family = binomial)
summary(mod_log)
```

```
##
## Call:
## glm(formula = Target ~ Age + Gender + Mom.Occp + Dad.Occp + Admission.Grade +
##     Scholarship + GPA, family = binomial, data = mod_df)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7309  -0.6735  -0.3763   0.4305   2.7429
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.786303   0.552256   3.235  0.00122 **
## Age              0.113737   0.011574   9.827  < 2e-16 ***
## GenderMale       0.528260   0.100372   5.263 1.42e-07 ***
## Mom.Occp        -0.023515   0.015776  -1.491  0.13608
## Dad.Occp        -0.012463   0.017145  -0.727  0.46729
## Admission.Grade -0.021134   0.003542  -5.967 2.41e-09 ***
## ScholarshipNo    1.251889   0.129395   9.675  < 2e-16 ***
## GPA             -1.147126   0.051200 -22.405  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4254.9  on 3240  degrees of freedom
## Residual deviance: 2737.3  on 3233  degrees of freedom
## AIC: 2753.3
##
## Number of Fisher Scoring iterations: 5
```

Let's try to select the best predictors using forwards and backwards selection.

Backwards selection

We will first use backwards elimination to remove predictors until the AIC value is stable and as low as possible. The AIC value is a measure for accuracy of the model, another metric used to measure how good the model is. We can use the step function to perform backwards elimination.

```
## Start:  AIC=2753.31
## Target ~ Age + Gender + Mom.Occp + Dad.Occp + Admission.Grade +
##     Scholarship + GPA
##
##                   Df Deviance    AIC
## - Dad.Occp         1   2737.8 2751.8
## <none>                 2737.3 2753.3
## - Mom.Occp         1   2739.6 2753.6
## - Gender           1   2764.7 2778.7
## - Admission.Grade  1   2774.1 2788.1
## - Age              1   2836.3 2850.3
## - Scholarship      1   2844.3 2858.3
## - GPA              1   3564.6 3578.6
##
## Step:  AIC=2751.84
## Target ~ Age + Gender + Mom.Occp + Admission.Grade + Scholarship +
##     GPA
##
##                   Df Deviance    AIC
## <none>                 2737.8 2751.8
```

```
## - Mom.Occp         1   2746.8 2758.8
## - Gender           1   2765.3 2777.3
## - Admission.Grade  1   2774.5 2786.5
## - Age              1   2837.7 2849.7
## - Scholarship      1   2845.2 2857.2
## - GPA              1   3564.8 3576.8

##
## Call:
## glm(formula = Target ~ Age + Gender + Mom.Occp + Admission.Grade +
##     Scholarship + GPA, family = binomial, data = mod_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7590  -0.6768  -0.3772   0.4294   2.7257
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.74116    0.54835   3.175  0.00150 **
## Age              0.11418    0.01156   9.875  < 2e-16 ***
## GenderMale       0.52893    0.10036   5.271 1.36e-07 ***
## Mom.Occp        -0.03166    0.01115  -2.840  0.00451 **
## Admission.Grade -0.02111    0.00354  -5.963 2.47e-09 ***
## ScholarshipNo    1.25333    0.12935   9.690  < 2e-16 ***
## GPA             -1.14699    0.05121 -22.398  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4254.9  on 3240  degrees of freedom
## Residual deviance: 2737.8  on 3234  degrees of freedom
## AIC: 2751.8
##
## Number of Fisher Scoring iterations: 5
```

After performing the backwards elimination, we are left with a model with the following predictors:

Age, Gender, Mom.Occp, Admission.Grade, Scholarship, GPA

Forwards selection

We will use forwards elimination to add predictors starting with an empty model until the AIC value is stable and as low as possible. We want to use a forwards selection method to confirm that we chose the most significant predictors from backwards selection.

```
## Start:  AIC=4256.9
## Target ~ 1
##
##                   Df Deviance    AIC
## + GPA              1   3109.4 3113.4
## + Scholarship      1   3918.6 3922.6
## + Age              1   3947.5 3951.5
```

```
## + Gender           1   4062.6 4066.6
## + Admission.Grade  1   4193.5 4197.5
## + Mom.Occp         1   4234.3 4238.3
## + Dad.Occp         1   4235.2 4239.2
## <none>                 4254.9 4256.9
##
## Step:  AIC=3113.43
## Target ~ GPA
##
##                   Df Deviance    AIC
## + Scholarship      1   2924.3 2930.3
## + Age              1   2941.7 2947.7
## + Gender           1   3045.9 3051.9
## + Admission.Grade  1   3065.8 3071.8
## + Dad.Occp         1   3094.1 3100.1
## + Mom.Occp         1   3095.2 3101.2
## <none>                 3109.4 3113.4
##
## Step:  AIC=2930.33
## Target ~ GPA + Scholarship
##
##                   Df Deviance    AIC
## + Age              1   2809.8 2817.8
## + Admission.Grade  1   2882.9 2890.9
## + Gender           1   2885.7 2893.7
## + Dad.Occp         1   2917.8 2925.8
## + Mom.Occp         1   2918.9 2926.9
## <none>                 2924.3 2930.3
##
## Step:  AIC=2817.79
## Target ~ GPA + Scholarship + Age
##
##                   Df Deviance    AIC
## + Admission.Grade  1   2774.7 2784.7
## + Gender           1   2782.9 2792.9
## + Mom.Occp         1   2801.0 2811.0
## + Dad.Occp         1   2802.6 2812.6
## <none>                 2809.8 2817.8
##
## Step:  AIC=2784.65
## Target ~ GPA + Scholarship + Age + Admission.Grade
##
##            Df Deviance    AIC
## + Gender    1   2746.8 2758.8
## + Mom.Occp  1   2765.3 2777.3
## + Dad.Occp  1   2767.0 2779.0
## <none>          2774.7 2784.7
##
## Step:  AIC=2758.8
## Target ~ GPA + Scholarship + Age + Admission.Grade + Gender
##
##            Df Deviance    AIC
## + Mom.Occp  1   2737.8 2751.8
## + Dad.Occp  1   2739.6 2753.6
```

```
## <none>          2746.8 2758.8
##
## Step:  AIC=2751.84
## Target ~ GPA + Scholarship + Age + Admission.Grade + Gender +
##     Mom.Occp
##
##            Df Deviance    AIC
## <none>          2737.8 2751.8
## + Dad.Occp  1   2737.3 2753.3

##
## Call:
## glm(formula = Target ~ GPA + Scholarship + Age + Admission.Grade +
##     Gender + Mom.Occp, family = binomial, data = mod_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7590  -0.6768  -0.3772   0.4294   2.7257
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.74116    0.54835   3.175  0.00150 **
## GPA             -1.14699    0.05121 -22.398  < 2e-16 ***
## ScholarshipNo    1.25333    0.12935   9.690  < 2e-16 ***
## Age              0.11418    0.01156   9.875  < 2e-16 ***
## Admission.Grade -0.02111    0.00354  -5.963 2.47e-09 ***
## GenderMale       0.52893    0.10036   5.271 1.36e-07 ***
## Mom.Occp        -0.03166    0.01115  -2.840  0.00451 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4254.9  on 3240  degrees of freedom
## Residual deviance: 2737.8  on 3234  degrees of freedom
## AIC: 2751.8
##
## Number of Fisher Scoring iterations: 5
```

After performing the forwards elimination, we are left with a model with the following predictors:

Age, Gender, Mom.Occp, Admission.Grade, Scholarship, GPA

These are the same predictors that we got from our backwards elimination. We will now perform diagnostics on this model to understand where it has flaws.

Final multiple logistic regression model

Let's take a look at the predictors and their significance based on the model summary.

```
final_log_mod <- glm(Target ~ Age + Gender + Mom.Occp
               + Admission.Grade + Scholarship + GPA,
               mod_df, family = binomial)
summary(final_log_mod)
```

```
##
## Call:
## glm(formula = Target ~ Age + Gender + Mom.Occp + Admission.Grade +
##     Scholarship + GPA, family = binomial, data = mod_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7590  -0.6768  -0.3772   0.4294   2.7257
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.74116    0.54835   3.175  0.00150 **
## Age              0.11418    0.01156   9.875  < 2e-16 ***
## GenderMale       0.52893    0.10036   5.271 1.36e-07 ***
## Mom.Occp        -0.03166    0.01115  -2.840  0.00451 **
## Admission.Grade -0.02111    0.00354  -5.963 2.47e-09 ***
## ScholarshipNo    1.25333    0.12935   9.690  < 2e-16 ***
## GPA             -1.14699    0.05121 -22.398  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4254.9  on 3240  degrees of freedom
## Residual deviance: 2737.8  on 3234  degrees of freedom
## AIC: 2751.8
##
## Number of Fisher Scoring iterations: 5
```

Our AIC is 2751. Our p-values indicate that all of our selected predictors are significant because they are less than 0.05. The coefficients indicate:

- A higher GPA gives a lower chance of a student dropping out.
- A student without a scholarship gives a lower chance of a student dropping out.
- A higher age gives a higher chance of a student dropping out.
- A higher admission rate gives a lower chance of a student dropping out.
- Male students are more likely to drop out.
- Certain mother's occupations effect whether a student drops out.

Model analysis and Diagnostics

Let's look at some model diagnostics and analyze our final model.

AIC and BIC

Let's first look at the AIC and BIC scores of our model.

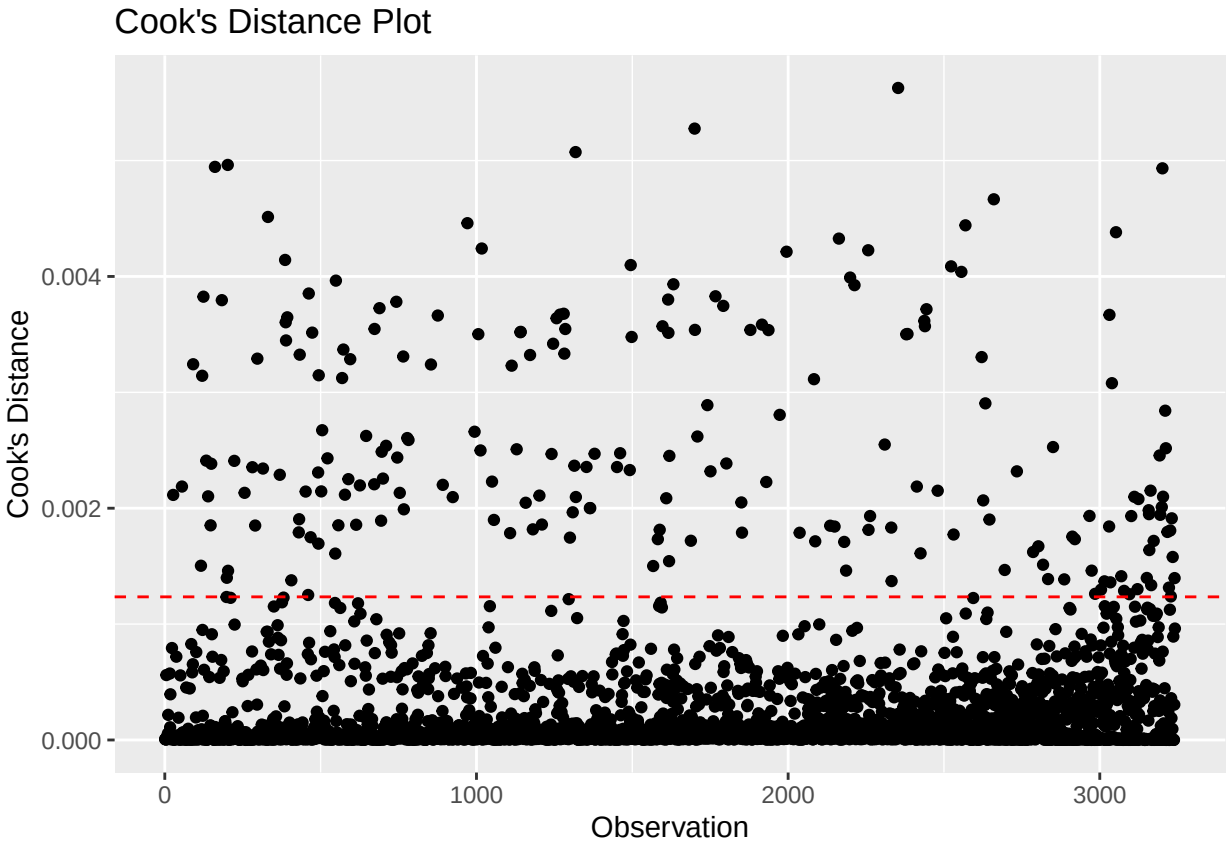
```
## [1] "AIC:  2751.83560771897"
```

```
## [1] "BIC: 2794.42106814096"
```

Cook's plot

Let's take a look at Cook's plot. We want to take a look at how many outliers significantly impact our response variable and our model.

```
## [1] "Proportion of points above the Cook's Distance:  0.0672631903733416"
```



Based on this plot, there are quite a few points that are above our Cook's distance line. But, when we look at the proportion of the number of influential points to the total number of points, we can see that it's only about 0.067, or 6.7% of the total points. This means that our model does not actually have too many influential points and our model is reliable.

Multicollinearity

We want to analyze multicollinearity because we want to make sure that none of our predictors are highly correlated with each other. If they are, we would need to either remove or combine them as their numbers give very similar information about the response variable.

```
##             Age          Gender        Mom.Occp Admission.Grade     Scholarship
##        1.025603        1.015018        1.012477        1.008172        1.031668
##             GPA
##        1.014523
```

After using VIF to print VIF values, we can see that all of our predictors are around 1-2, which are much less than 5. We can conclude that our predictors do not exhibit multicollinearity.

Confusion Matrix

Next, we want to look at a confusion matrix using our model to look at some metrics to measure our model's performance.

```
## [1] "Confusion Matrix for Training Data:"
```

```
##       Predicted
## Actual    0    1
##      0 1526  123
##      1  369  575
```

```
## [1] "Confusion Matrix for Test Data:"
```

```
##       Predicted
## Actual   0   1
##      0 375  33
##      1  82 158
```

```
## Accuracy:  0.8225309
```

```
## Precision:  0.8272251
```

```
## Recall:  0.6583333
```

```
## Specificity:  0.9191176
```

```
## F1 Score:  0.7331787
```

We have very high accuracy and precision. We do have a lower recall of 65% but our specificity is 92%. Overall, out model performs very well, which we can analyze using the F1 value. The F1 value balances our precision and recall and tells us whether the model balances both well. Based on our F1 value of 0.73, we conclude that our model has reasonably good precision and recall.

Comparing the training and test data matrices, we can clearly see that our model performs better on the test data than the training data. This is a good sign that our model is doing very well and has successfully learned the underlying patterns in the training data. This means that our model is likely to perform well when given unseen or new data.

K-Fold Cross Validation

We want to analyze how well our model performs using k-fold cross validation.

```
## Generalized Linear Model
##
## 3241 samples
##    6 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2917, 2917, 2917, 2917, 2916, 2917, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.8120955  0.5717786
```

A kappa value of approximately 0.57 suggests a moderate level of agreement beyond chance. Our classification model performs very well, achieving an accuracy of around 81% with a moderate level of agreement between predicted and actual classifications.

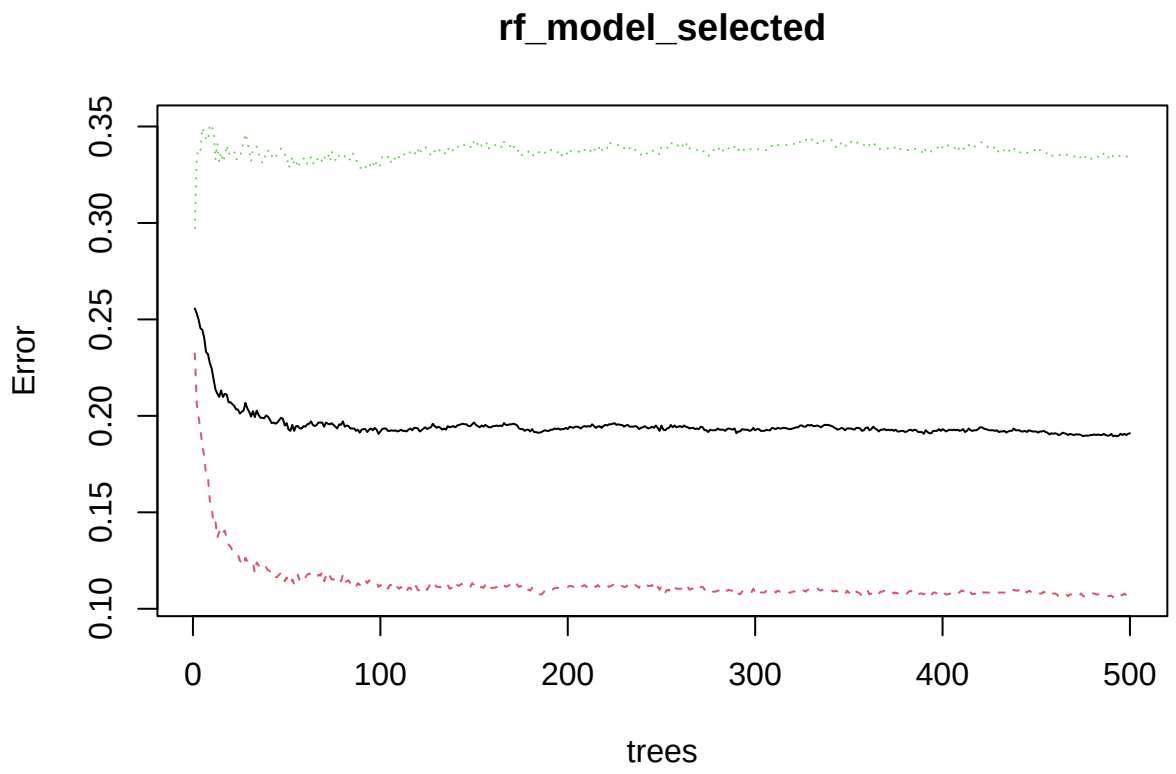
Random Forest Model

Let's create a random forest model using all of our predictors next. We will use this model to compare it to our multiple logistic regression model.

```
set.seed(12345)
rf_model <- randomForest(Target ~ Age + Gender + Mom.Occp + Dad.Occp +
                           Admission.Grade + Scholarship + GPA, data = mod_df)

var_importance <- as.data.frame(importance(rf_model))
selected_predictors <- rownames(var_importance)[var_importance$MeanDecreaseGini > 100]

rf_model_selected <- randomForest(Target ~ .,
                                  data = subset(mod_df, select = c("Target", selected_predictors)))

rf_model_selected
```

```
##
## Call:
##  randomForest(formula = Target ~ ., data = subset(mod_df, select = c("Target",      selected_predicto
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 19.1%
## Confusion matrix:
##      0   1 class.error
## 0 1836 221   0.1074380
## 1  398 786   0.3361486
```

```
plot(rf_model_selected)
```




```
print(selected_predictors)
```

```
## [1] "Age"             "Dad.Occp"        "Admission.Grade" "GPA"
```

The first thing to note about this model is that after pruning the model using a very high value of importance, it chose the following variables: Age, Admission.Grade, GPA, and father's occupation. This is a smaller and different subset of variables. This could indicate that these subset of variables predicts the target more accurately. We will compare the analytics of this model to our first one to see which one is better.

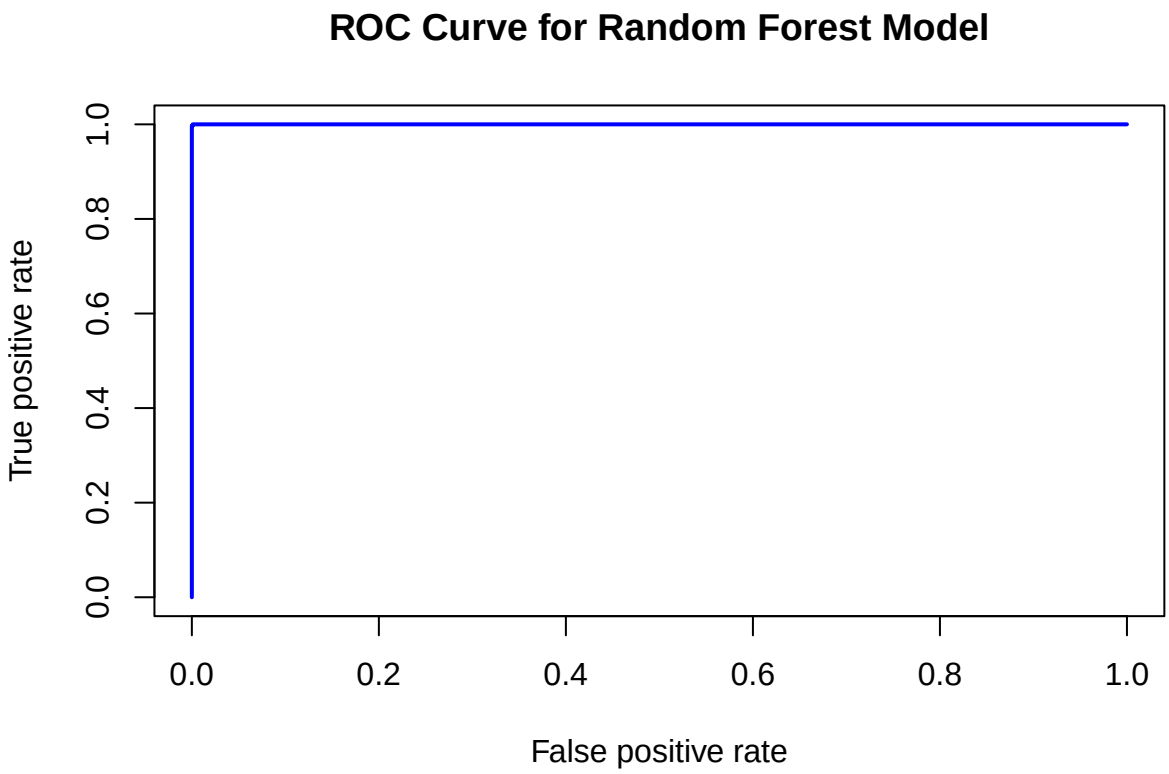
We can see the error rate for all the variables and the highest error rate is low and the overall error rate of the model is 19.28%. This is a fairly low error rate and the model performs well for the most part. We will take a look at a confusion matrix using training and test data later to confirm how well the model performs.

Model analysis and Diagnostics

Let's now analyze this forest and compare it to our multiple logistic regression model.

ROC Curve

We will first look at the ROC for our random forest.



The ROC curve allows us to visualize the sensitivity of our model. Looking at our curve, it is very close to the top left of the graph. This means that the sensitivity of the model is very high and it has a lower false positive rate. Overall, our model performs very well for our response variable.

Confusion Matrix

Let's now create a confusion matrix and calculating the different metrics.

```
## [1] "Confusion Matrix for Training Data:"
```

```
##       Predicted
## Actual    0    1
##      0 1643    3
##      1    0  948
```

```
## [1] "Confusion Matrix for Test Data:"
```

```
##       Predicted
## Actual   0   1
##      0 374  37
##      1  85 151
```

```
## Accuracy:  0.8114374
```

```
## Precision:  0.8031915
```

```
## Recall:  0.6398305
```

```
## Specificity:  0.9099757
```

```
## F1 Score:  0.7122642
```

Similar to the multiple logistic regression model, our random forest model has high accuracy and precision. Furthermore, we have a lower recall of 64% but our specificity is 91%. Overall, out model performs very well, which we can analyze using the F1 value. Based on our F1 value of 0.71, we conclude that our model has reasonably good precision and recall.

An interesting note is that our F1 values for both models are very similar, which means that both model are quite reliable with a fairly high agreement between recall and precision.

Comparing the confusion matrices for the training and test data, we can see that the model performs far worse on the test data compared to the training data, which it performed nearly perfectly on. This is a sign that our model may be over fitting based on the data and may not be a very reliable model. This is because the model is likely not going to be able to perform well when given unseen data so it is not as reliable.

K-Fold Cross Validation

Now, we will use k-fold-cross-validation to train our model and test it using the random forest models.

```
## note: only 2 unique complexity parameters in default grid. Truncating the grid to 2 .
```

```
## Random Forest
##
## 3241 samples
##    3 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2917, 2917, 2917, 2918, 2917, 2918, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##   2     0.7932880  0.5417715
##   3     0.7840277  0.5226441
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```

Based on a 10-fold cross validation, we got an accuracy of 79% and a kappa value of 0.54. Our model performs very well.

The kappa value indicates that our model has a moderate level of agreement between the predicted and actual classes as well.

Based on all of these analyses and comparison of the two models, we can conclude that our random forest model is better than our multiple logistic regression model.

After the model analysis, we concluded that our logistic regression model was a more reliable model. But, based on the fact that both models performed significantly well in a variety of tests, we concluded that Age, Admission Grade, and GPA are the variables that seem to have the most influence on whether a student drops out or not. These were the common predictors between both models after pruning.

Contributions

Based on the Table of Contents

- Khushi: Biggest Reasons Students Drop Out
- Gabe: Data Cleaning and Manipulation, Target Variable, Student Age and Grades vs Dropout Rates
- Tiffany: Research Questions, Scholarship Holders and Academic Performance
- Vincent: Student Age and Grades vs Dropout Rates
- Nicole: Previous Qualifications and Academic Performance