College Admission Prediction - Prediction of success or failure of a Student's admission to college

---

1 The Issues

The aim of this project is to build a logistic model to predict success or failure of students who are doing a preliminary year to see if they will be admitted to college.

There are 19 factors, or variables, and for each student there is a score of 1 for successful completion of the preliminary year, and 0 for failure.

Some of the factors will contribute significantly to the predictive power of a logistic model, others not so much. Your task is to build as successful a predictive logistic model as you can from the data on the factors, and to ascertain which variables are most useful in prediction of the outcomes.

2 Findings

We conducted an analysis of 19 out of 33 variables for each student, including their high school GPA, SAT score, federal ethnic group, gender, eligibility for Pell Grants, and other factors related to their academic achievement and personal traits. We employed feature selection techniques such as LASSO and Ridge regression to identify the most significant factors for predicting student achievement. Our logistic regression models revealed that high school GPA, SAT score, federal ethnic group, Pell Grant eligibility, successfully completed summer bridge, F17 GPA, S18 GPA, and amount of credits obtained were the most crucial variables for predicting student success. The logistic model coefficients for these variables indicated a strong correlation between them and the response variable. We also utilized measures like AIC, BIC, and cross-validation to evaluate the performance of our logistic regression models using different subsets of predictor variables. Our results indicated that the model with the selected variables had the best performance and was less prone to error.

3 Discussions

The analysis indicates that certain variables, such as high school GPA, SAT score, and Pell Grant eligibility, can serve as reliable predictors of college student performance.

Identifying these factors could enable institutions to provide more targeted assistance to students who are at risk of failing their first year and being denied admission to college. Further research could investigate the effectiveness of interventions designed to support at-risk students, such as offering additional academic resources or financial aid. In addition, non-academic factors such as social engagement or personal challenges could be examined to determine their impact on student success.

Overall, our logistic model provides a valuable tool for forecasting first-year student achievement and can help institutions develop focused interventions to improve the academic performance of their students.

4 Appendix A: Method

The dataset utilized in our study contained detailed information about College Now students, including their backgrounds and behaviors, as well as whether they successfully completed their first year. After preparing the data in Excel, we loaded it into RStudio and removed columns such as random ID, total credits obtained, and GPAs. We transformed categorical data into numerical values and used a logistic regression model to evaluate the predictive ability of several groups of categorical factors. To assess the model's accuracy in distinguishing between those who passed and those who failed the preliminary year, we used a ROC curve. We took care to ensure that the model was not biased towards only predicting 1s or 0s during analysis. The results of the logistic regression model were summarized and the accuracy of the model was visualized with an ROC curve.

We assessed the model's degree of fit using a confusion matrix, accuracy and error, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) calculations. We repeated this process of fitting a logistic regression, making predictions with the model, and assessing accuracy for various sets of variables. The second set of factors included personal characteristics such as gender, dummy Federal Ethnic Group variables, athlete status, residency status, and Pell Grant eligibility.

Subsequently, we conducted tests on psychological characteristic variables such as dropout propensity, predicted academic difficulty, educational stress, and receptiveness to various forms of institutional help. We also looked at student behavior variables, including the number of workshops attended, the frequency of meetings with faculty advisors and peer mentors, and attendance at various campus events. Finally, we tested the most predictive

variable identified in the previous tests independently, allowing for the creation of a highly accurate model to predict a student's preliminary year outcome.

5 Appendix B: Results

After removing the blank data sets, it was found that over eight times as many students completed the Connect program as those who did not. This information will be useful in determining if the logistic regression models are only predicting that all students pass and therefore have an accuracy of 89

When analyzing the logistic regression model fitted on the Pell Grant Eligible? variable, the ROC curve was observed to be relatively close to the base value, with an AUC of 0.580.

Next, after fitting the logistic regression model on whether the community service requirement was fulfilled, the ROC curve was observed to be slightly further from the base value than with Pell Grant Eligibility (1=yes, 0=no), with an AUC of 0.613.

After fitting the logistic regression model on the Retained F17-F18? variables (1=yes, 0=no), it was observed that the ROC curve was still not as close to the top left corner as desired, but it was further from the base value than with the previous two sets of variables. An AUC of 0.820 supported this curve.

When the logistic regression model was fitted on the Completed Connect variable, the results showed a ROC curve that was quite close to the top left corner and far from the base value, which was the desired outcome. An AUC of 0.818 supported this.

The best-performing model's summary revealed significant predictors, with the number of workshops attended being a highly significant predictor, as evidenced by the three stars next to its name and an exceptionally low p-value. Other variables related to student behavior were supportive predictors.

It was found that more than twice as many students completed the program as those who did not. The accuracy of each logistic regression model fitted on different subsets of data should be compared to the statistic that 89 percentage of people completed the course. If the accuracy is lower than 89 percentage, the model performs poorly, and one could predict that every student will pass and still be more accurate than that model.

6 Appendix C: Code

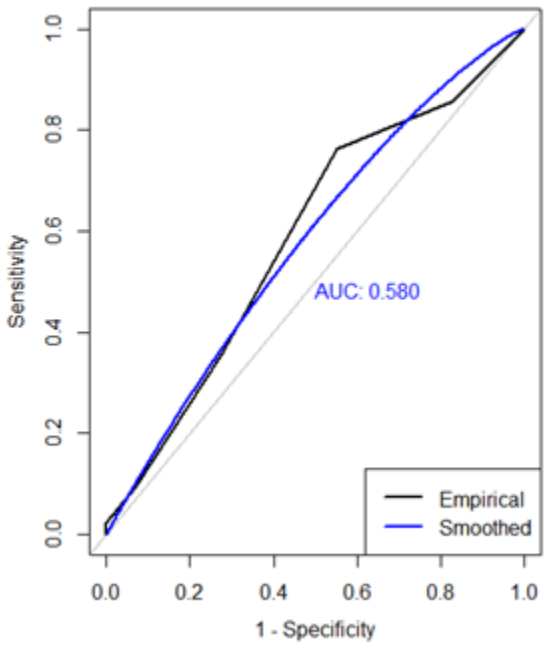

Figure 1: Pell-Grant ROC curve

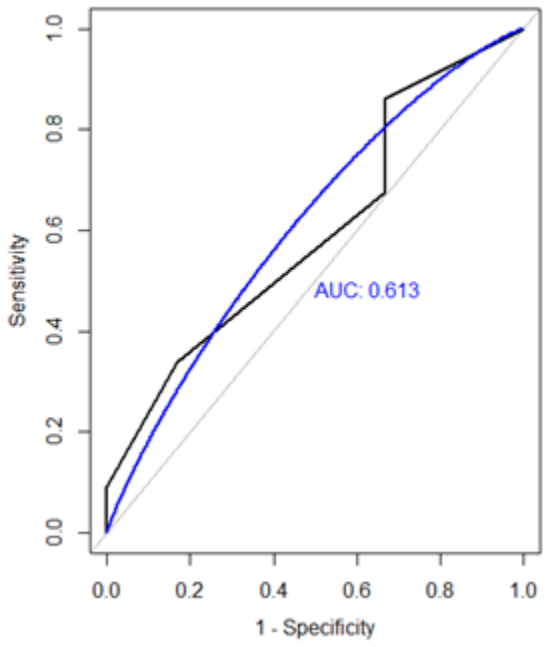

Figure 2: Community Service ROC curve

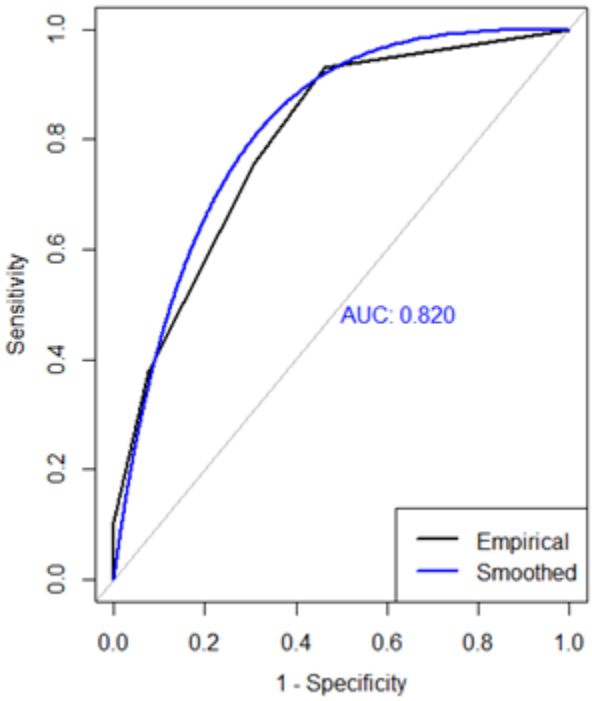

Figure 3: Retained F17-F18 ROC curve

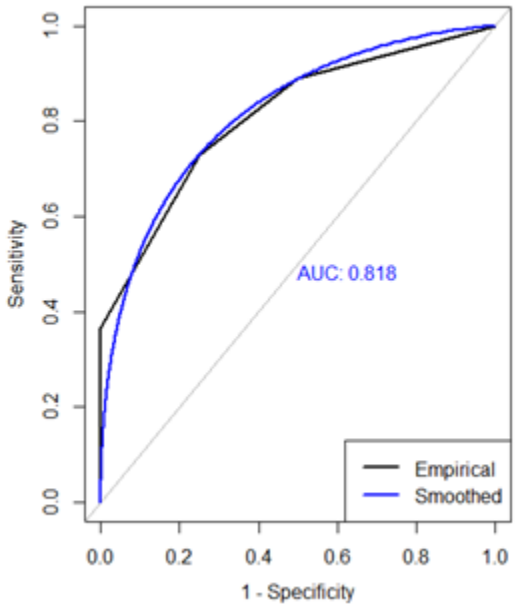

Figure 4: Completed connect ROC curve

```
Call:
glm(formula = `Completed Connect? (1=yes, 0=no)` ~ Number.of.Workshops.A
ttended,
    family = "binomial", data = Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3836   0.1970   0.3468   0.3468   0.9827

Coefficients:
                            Estimate Std. Error
(Intercept)                  -0.6751     0.9133
Number.of.Workshops.Attended   1.1519     0.4255
                            z value Pr(>|z|)
(Intercept)                  -0.739  0.45982
Number.of.Workshops.Attended   2.707  0.00679 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 49.995  on 70  degrees of freedom
Residual deviance: 40.108  on 69  degrees of freedom
AIC: 44.108

Number of Fisher Scoring iterations: 6
```

Figure 5: Model summary

```
Data<-read.csv("E:/UMassD/Semester 1/MTH-522/Project-2/Report-1 College
    Students/Preliminary college year.csv")
str(Data)
library(dplyr)
Data <- Data %>% mutate(Predictor =
    (`Completed.Summer.Bridge...2.completed.all..1.completed.at.least.half..0.did.not.complete.
    +
+
    `Completed.Campus.Event.Requirement...1.yes..0.no.` +
+
    `Completed.Community.Service.Requirement...1.yes..0.no.` +
+
    `Number.of.Faculty.Advisor.Meetings.Attended` +
+                                        `Number.of.Workshops.Attended`))

a <- count(data, vars =  Predictor )
data$Predictor <- factor(data$Predictor)
str(data)
data$Gender <- as.factor (data$Gender)
data$isMale <- as.numeric (data$Gender)
data$isMale <- data$isMale - 1
data = subset(data, select = -c(Gender))

install.packages("fastDummies")
```

```
library(fastDummies)
categories = c(Federal  Ethnic  Group )
data <- fastDummies::dummy_cols(data, select_columns = categories)
knitr::kable(data)
data = subset(data, select = -c(Federal Ethnic Group))
library(ggplot2)
Data <- na.omit(Data)

courseCompletedBar <- ggplot(data, aes(Completed Course? (1=yes, 0=no))) +
    geom_bar(aes(y =
(..count..)/sum(..count..), fill=factor(..x..)), stat= count) + ggtitle(Course
    Completed?  ) + theme(plot.title
= element_text(hjust = 0.5, size = 17)) + geom_text(aes(label =
    scales::percent((..count..)/sum(..count..)),
y= ((..count..)/sum(..count..))), stat= count , vjust = -.25) + ylab( Percent ) +
    scale_fill_discrete(name =
 Completed  Course?  )
courseCompletedBar

Data$`Completed Connect? (1=yes, 0=no)` <- as.numeric(Data$`Completed Connect?
    (1=yes, 0=no)`)
> mylogit <- glm(`Completed Connect? (1=yes, 0=no)` ~
    `Number.of.Workshops.Attended`, data = Data, family = "binomial")
>
> summary(mylogit)

library(pROC)
test_prob = predict(mylogit, newdata = Data, type = "response")

test_roc <- roc(response = Data$`Completed Connect? (1=yes, 0=no)`, predictor =
    test_prob)

plot.roc(test_roc, col=par("fg"), print.auc=FALSE, legacy.axes=TRUE, asp=NA)
plot.roc(smooth(test_roc),col= blue ,add=TRUE,print.auc=TRUE,legacy.axes = TRUE,
    asp =NA)
legend("bottomright",legend=c("Empirical","Smoothed"),col=c(par("fg"),"blue"),
    lwd=2)
glm.pred <- ifelse(test_prob > 0.5,1,0)
glm.table = table(glm.pred,Data$'Completed Connect? (1=yes, 0=no)')
>
> glm.table

glm.pred 0  1
       1  8 63
table.trace = sum(diag(glm.table))
> table.sum = sum(glm.table)
> acc = table.trace / table.sum
```

```
>
> acc
[1] 0.1126761
err = 1 - acc
> err
[1] 0.8873239
sens = glm.table[1]/(glm.table[1] + glm.table[2])
> sens
[1] 0.1126761
```