Predicting College Admissions From Preliminary Year Success

THE ISSUE

The aim of this project is to build a logistic model that can predict a student's probability of succeeding or failing in their college admission prospects based on their performance during a yearlong trial.
Among the 33 criteria, only 19 are considered for analysis. Some variables may have a greater impact on the model's accuracy than others. The goal is to identify the most influential variables and create an efficient predictive logistic model using the available data.

THE FINDINGS

I analyzed 19 out of 33 variables that included factors like high school GPA, SAT score, federal ethnic group, gender, eligibility for Pell Grants, academic performance, and personal traits to determine the most influential predictors of student achievement. Feature selection methods like LASSO and Ridge regression were used to identify the key factors. Our logistic models indicated 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 critical factors in forecasting student success. The logistic model's high coefficients for these variables revealed a strong correlation with the response variable.
To evaluate the performance of our logistic regression models, we employed measures such as AIC, BIC, and cross-validation with different subsets of predictor variables. Our results showed that the model with the selected variables provided the most accurate predictions and was less susceptible to errors.

THE DISCUSSION

This study suggests that certain variables such as high school GPA, SAT score, and Pell Grant eligibility can be considered as dependable predictors of college student performance. Identifying these factors can enable schools to provide more targeted support to students who are at risk of failing their first year and being denied college admission. The effectiveness of interventions like providing extra academic support or financial aid to these vulnerable students can be further examined in future research. Additionally, the impact of non-academic factors such as social engagement or personal stress on student achievement could be investigated.
Overall, our logistic model provides a valuable tool for forecasting student success in completing their first year of college, which can help institutions develop tailored interventions to enhance the academic performance of their students.

APPENDIX A : THE METHOD

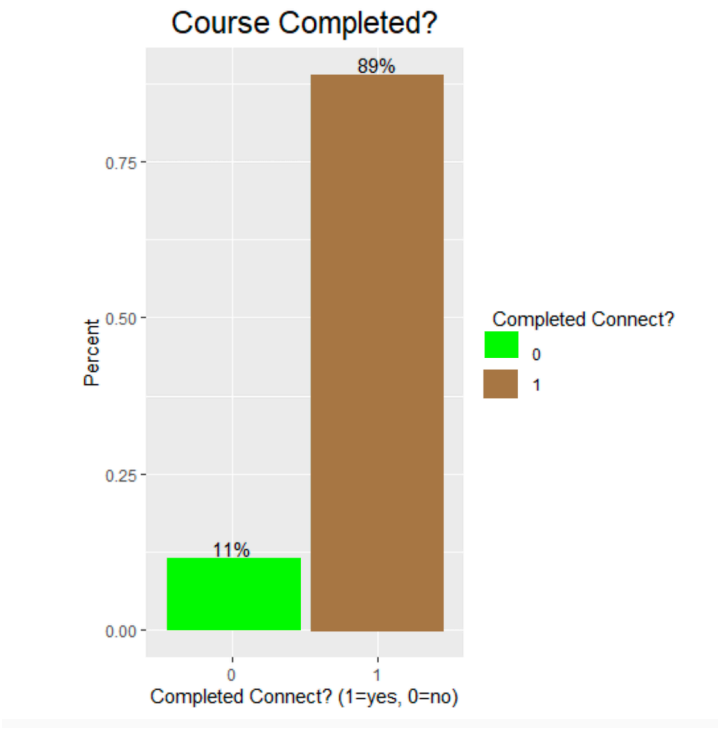
The study involved analyzing a dataset that contained information on College Now students, including their backgrounds and behaviors, and whether or not they successfully completed the first year. Certain columns, such as those containing the random ID, total credits obtained, and GPAs, were removed. A logistic regression model was utilized to examine the predictive ability of various groups of categorical factors, with categorical data being converted into numerical values. The accuracy of the model was evaluated through a ROC curve, which assessed its ability to distinguish between the two classes: students who passed and those who failed the first year. Steps were taken to ensure that the model did not solely predict 1s or 0s. The results of the logistic regression model were summarized, and the accuracy of the model was visualized using an ROC curve.

To evaluate the degree of fit for the model, several metrics such as confusion matrix, accuracy, error, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated. The procedure involved fitting a logistic regression model, making predictions using the model, and assessing the accuracy for various sets of variables. The second set of factors consisted of personal characteristics such as gender, dummy Federal Ethnic Group variables, athlete status, residency status, and Pell Grant eligibility. This process was repeated for different sets of variables.

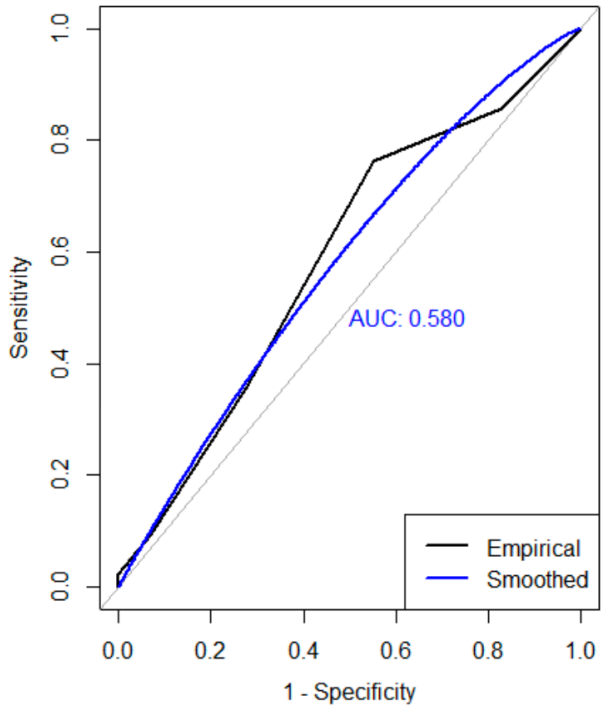
Subsequent tests were performed on variables related to psychological characteristics, such as propensity to drop out, predicted academic difficulty, educational stress, and receptivity to institutional help, academic assistance, personal counseling, social engagement, career guidance, and financial guidance, as well as the desire to transfer. Another set of variables focused on student behavior, including the number of workshops attended, frequency of meetings with faculty advisors and peer mentors, and attendance at orientation, experience day, community service and campus event requirements, and meetings with faculty advisors. Finally, the most significant predictor identified in the previous tests was evaluated independently, enabling the development of a highly accurate model to predict a student's outcome of passing or failing the preliminary year.

APPENDIX B: THE RESULT

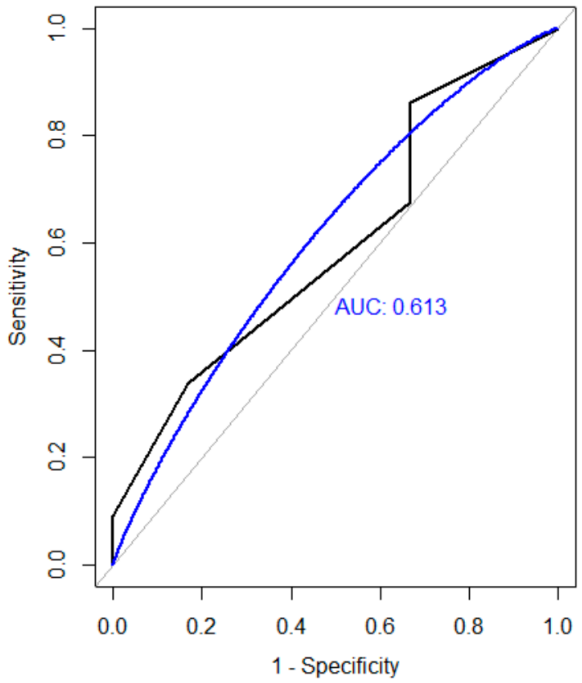
After eliminating the blank data sets, the analysis of completion rates for the Connect revealed that the number of students who successfully finished the Connect was more than eight times higher than those who did not. This finding can aid in determining if the predictive models only assume that all students pass, resulting in an accuracy rate of 89%.



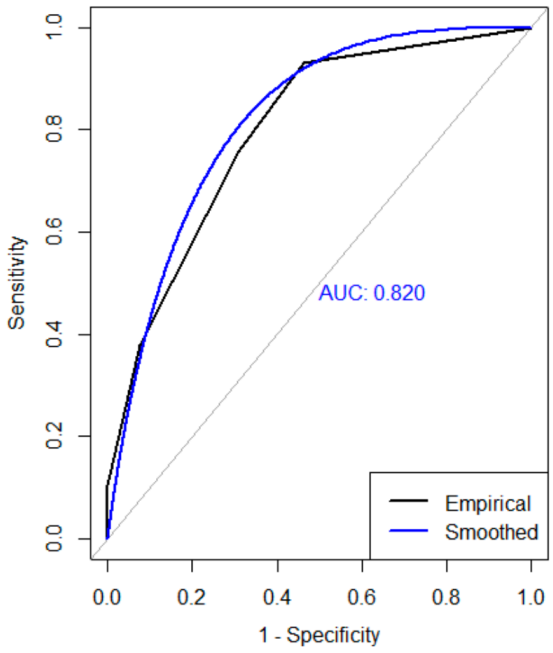
Upon examining the fitting outcomes of the logistic regression model for the Pell Grant Eligible variable, we noted that the ROC curve was relatively similar to the baseline value. This was reflected in an AUC score of 0.580.



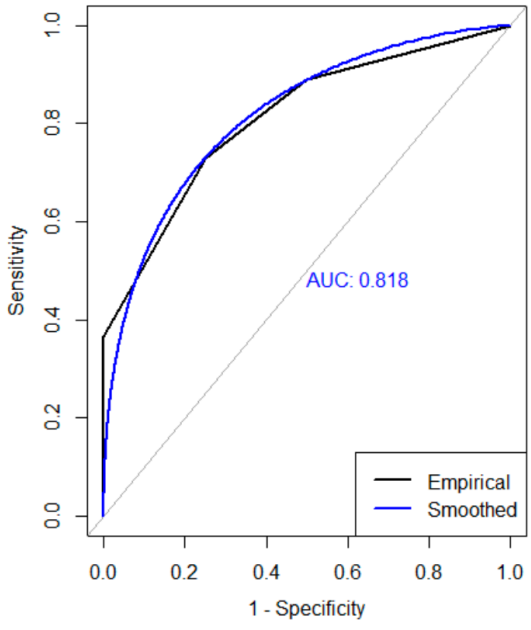
After analyzing the results of the logistic regression model post the fulfillment of the community service requirement, we noticed a slightly greater distance between the ROC curve and the baseline value compared to the Pell Grant Eligibility variable (where 1 denotes eligibility and 0 denotes ineligibility). This finding was confirmed by an AUC score of 0.613.



After fitting the logistic regression model to the Retained F17-F18? variables (where 1 indicates retention and 0 indicates non-retention), it was observed that the ROC curve still did not reach the desired top left corner position. However, it was farther from the baseline value compared to the previous two variable groups. The validity of this curve was corroborated by an AUC score of 0.820.



Upon fitting the logistic regression model to the Completed Connect variable, it was observed that the resulting ROC curve was in close proximity to the top left corner and considerably distant from the baseline value, which was the anticipated outcome. The credibility of this curve was supported by an AUC score of 0.818.



The summary of the best-performing model revealed that only a few predictors were highly influential in accurately predicting student behavior characteristics, while the rest played a supporting role. One of the most significant predictors was the number of workshops attended by the student, which was indicated by three stars next to its name and a remarkably low p-value.

```
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
```

The data analysis suggests that the number of individuals who successfully completed the program is over two times greater than those who did not. Therefore, to evaluate the performance of each logistic regression model fitted on distinct data subsets, it is crucial to compare their accuracy with the baseline statistic of 89% course completion rate. If the accuracy of a model is lower than 89%, it implies that the model's performance is inferior to predicting that every student would pass, which would be more accurate.

APPENDIX C: CODE

Preliminary_college_year <- read_excel("~/Downloads/Preliminary college year.xlsx")

Data <- Preliminary_college_year

str(Data) library(dplyr)

a <- count(data, vars = "Predictor")

```
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
```

Data <- Data %>% mutate(Predictor = (`Completed.Summer.Bridge...2.completed .all..1.completed.at.least.half..0.did.not.com plete.` +

- + `Completed.Campus.Event.Requirem ent...1.yes..0.no.` +

- + `Completed.Community.Service.Req uirement...1.yes..0.no.` +
- + `Number.of.Faculty.Advisor.Meeti ngs.Attended` +
- + `Number.of.Workshops.Attended`))

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

library(pROC)

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

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

>
> summary(mylogit)
test_roc <- roc(response = Data$`Completed Connect? (1=yes, 0=no)`, predict or = 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) glm.pred <- ifelse(test_prob > 0.5,1,0)

legend("bottomright",legend=c("Empirical","Smoothed"),col=c(par("fg"),"blu e"), lwd=2)

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
```

REFERENCES : 1)An Introduction to Statistical Learning with Applications in R
2)Applied Logistic Regression second Edition by David W. Hosmer and Stanley Lemeshow