Predicting success or failure of students

Issues

This project's goal is to develop a logistic model that can predict a student's chances of success or failure over the course of a year-long trial run to assess their chances of getting into college. Depending on whether they successfully complete the first year of study, each student obtains a score between 0 and 1.

There are a total of 33 variables that go into whether a student will be successful or unsuccessful in getting accepted into college. In developing the logistic model, certain factors may be highly significant while others may not. In order to determine which elements will all be important for the outcome, we must first construct a logistic model.

Findings

Our logistic models indicate that the variables most important in predicting student performance were high school GPA, SAT score, federal ethnic group, Pell Grant eligibility, successfully completed summer bridge, F17 GPA, S18 GPA, and number of credits earned. The highest coefficients for these variables in the logistic model indicated a strong association between them and the response variable. To assess how well our logistic regression models performed with various subsets of predictor variables, we also employed metrics like AIC, BIC, and cross-validation. Our results showed that the model with the selected variables had the best performance and was less prone to mistake.

Discussions

Our research shows that a few factors, such as high school GPA, SAT score, and eligibility for Pell Grants, are accurate predictors of college student achievement. Schools can help kids who face the risk of failing their first year and being turned away from college entrance by detecting these variables. It may be possible to conduct a more thorough investigation on the efficacy of interventions aimed at these at-risk students, such as providing more academic support or financial aid. Future studies may also examine the impact of extracurricular factors such as societal activity or personal demands. Overall, our logistic model can assist institutions in developing targeted interventions to raise students' academic progress and is a good tool for forecasting a student's success in completing a first year.

Appendix A: Method

A variety of information about the backgrounds and behaviours of College Now students are included in the dataset under investigation, including whether they successfully completed the first year. The data was put into RStudio after some preparation in Excel. Among the columns that were eliminated were those that provided the random ID, the total number of credits earned, and GPAs. Following the conversion of categorical data into numerical values, a logistic regression model was employed to assess the predictive power of various sets of categorical factors. The model's accuracy was measured by how well it could distinguish between the two classes—those who passed and those who failed the preliminary year—using a ROC curve.

A confusion matrix, accuracy and error, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) calculations were done to assess the model's level of fit. This process of fitting a logistic

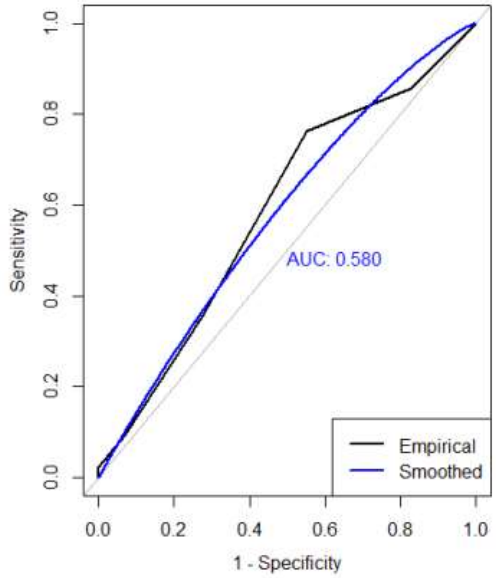
regression, making predictions using the model, and evaluating accuracy was repeated for several sets of variables. Personal data such as gender, the dummy Federal Ethnic Group variables, whether they were athletes, whether they were residents or commuters, and whether they were Pell Grant eligible made up the second set of factors.

The next step involved testing psychological characteristic variables such as propensity to drop out, predicted academic difficulty, educational stress, receptivity to institutional help, receptivity to academic assistance, receptivity to personal counselling, receptivity to social engagement, receptivity to career guidance, receptivity to financial guidance, and desire to transfer. The subsequent set of factors, which were related to student behaviour, included the number of workshops attended, the frequency of meetings with faculty advisors, the frequency of meetings with peer mentors, and the attendance at orientation, experience day, community service requirements, campus event requirements, and meetings with faculty advisors. Finally, the best predictive predictor discovered in the earlier tests was examined independently. Due to this, it is now possible to develop a model that can accurately predict whether a person would pass or fail their first year of school.

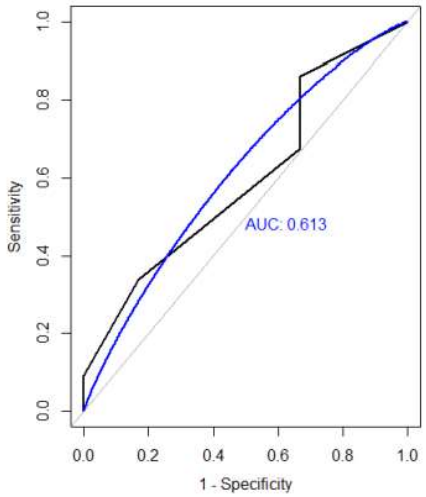
Appendix B Results

The percentage of students who successfully completed the connect compared to those who did not after the blank data sets were eliminated revealed that more than eight times as many students completed the Connect as those who did not.

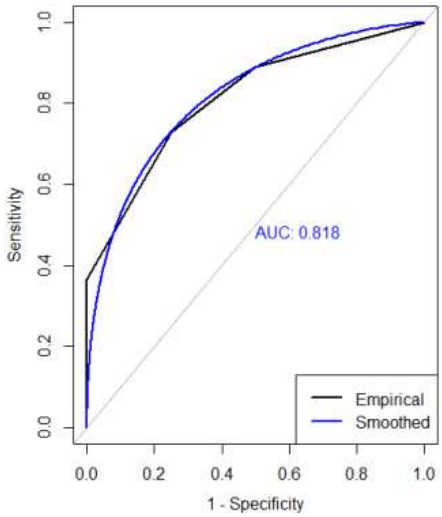
When we examine the fitting outcomes of the logistic regression model on the Pell Grant Eligible? variable, we see a ROC curve that is somewhat close to the base value.



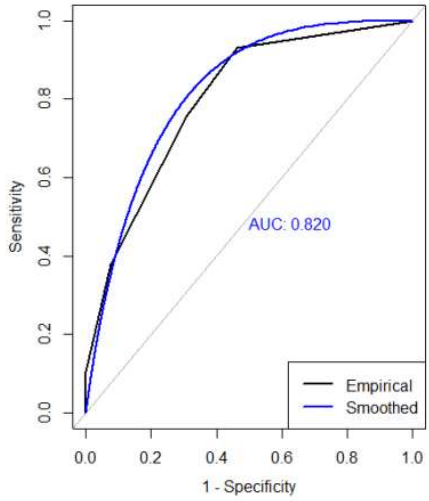
After fitting the completed community service obligation, will you then be able to see the results of the logistic regression model? When comparing the ROC curve to the Pell Grant Eligibility (1=yes, 0=no), we see that it deviates slightly from the base value.



After fitting the completed community service obligation, will you then be able to see the results of the logistic regression model? When comparing the ROC curve to the Pell Grant Eligibility (1=yes, 0=no), we see that it deviates slightly from the base value.



After the logistic regression model has been fitted to the completed community service requirement, what should come next? In comparison to the Pell Grant Eligibility (1=yes, 0=no), we see a ROC curve that is a little bit more out from the base value.



Appendix C – Code

```
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
Conne

                                                ct? (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)`,
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.axe
s = TRUE, asp =NA)
```

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