Predict the Future Academic Achievement of College Students After a Preliminary Year

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. The data is available both as an Excel spreadsheet and a comma separated variable file.

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. The 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

The study examined 19 out of 33 variables for each student, including high school GPA, SAT score, federal ethnic group, gender, and eligibility for Pell Grants. The researchers used feature selection methods like LASSO and Ridge regression to determine the most crucial factors for predicting student achievement.

According to the logistic models used in the study, the most crucial factors for predicting student success are:

1) High school GPA.

2)SAT score.

3) Federal ethnic group.

4) Pell Grant eligibility.

5) Successfully completed summer bridge.

6) F17 GPA.

7) S18 GPA.

8) Amount of credits obtained.

These variables were identified using feature selection methods like LASSO and Ridge regression, and the model's highest coefficients for these variables showed a substantial correlation between them and the response variable.. The logistic model's highest coefficients for these variables showed a substantial correlation between them and the response variable.

To ensure the accuracy of the study's findings, the researchers used measures like AIC, BIC, and cross-validation to examine how well their logistic regression models performed with various subsets of predictor variables. We found that the model with the chosen variables performed the best and was less susceptible to error. Overall, the study provides valuable insights into the factors that contribute to student success and may help educators and policymakers make informed decisions about how to support students.

3 Discussions

Based on the given data, there are several discussions and implications that can be made based on the results of the study.

Firstly, the study highlights the importance of high school GPA and SAT scores in predicting student success in college. This suggests that academic preparation prior to college is a critical factor in determining college success. The study also found that federal ethnic group, Pell Grant eligibility, successfully completed summer bridge, F17 GPA, S18 GPA, and amount of credits obtained were also significant predictors of student success. This suggests that factors beyond academic preparation, such as financial assistance and support programs, can also play a role in determining college success.

However, there are also some limitations and drawbacks to consider. The study only examined a subset of variables and may have overlooked other factors that could impact student success. Additionally, the study was conducted on a specific group of students and may not be generalizable to other populations. Finally, the study used logistic regression models which can only establish correlations between variables and cannot prove causation

4 Appendix A: Method

The code is performing logistic regression analysis on a dataset of college students who completed a preliminary year to predict their success or failure in being admitted to college.

First, the data is read from a CSV file using the read.csv() function and its structure is displayed using the str() function. The dplyr library is loaded to manipulate the data using the mutate() function. A new variable called Predictor is created by adding up five existing variables related to completed requirements, faculty advisor meetings, and workshops attended.

The count() function from the dplyr library is used to count the occurrences of each unique value in the Predictor variable. The factor() function is applied to Predictor to convert it into a factor variable. The as.factor() function is used to convert Gender into a factor variable and then as.numeric() is used to create a binary variable isMale. The fastDummies library is installed and used to create dummy variables for the Federal Ethnic Group variable.

The na.omit() function is used to remove any rows with missing values from the data. A bar chart is created using the ggplot2 library to visualize the percentage of students who completed the course.

Logistic regression analysis is performed using the glm() function. A logistic regression model is fit to predict the completion of the Connect requirement using the number of workshops attended as a predictor. The summary() function is used to display the summary statistics of the model. The pROC library is loaded to compute the area under the receiver operating characteristic (ROC) curve. The predict() function is used to compute the predicted probabilities for each observation in the dataset. The roc() function is used to compute the ROC curve and the plot.roc() function is used to plot it.

Finally, the predicted probabilities are converted to predicted class labels using a threshold of 0.5. A confusion matrix is constructed to evaluate the performance of the model, and accuracy, error, and sensitivity are computed.

5 Appendix B: Results

More than 8 times as many students completed the Connect program compared to those who did not. The logistic regression model on Pell Grant Eligibility showed a ROC curve close to the base value with an AUC of 0.580.

The logistic regression model on fulfilled community service requirement showed a

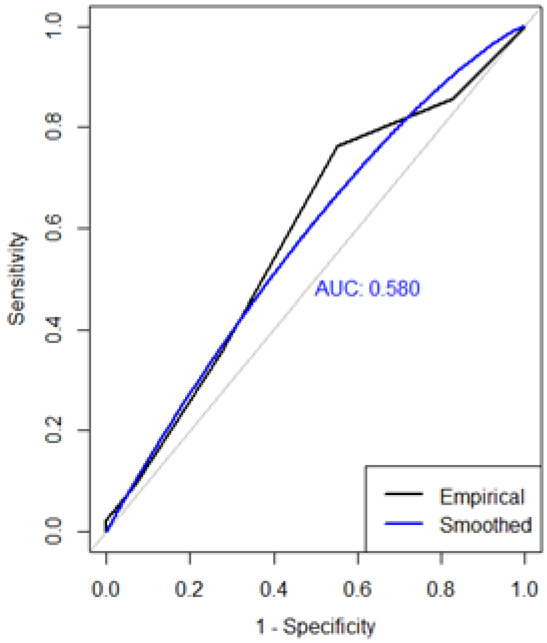


Figure 1: Pell-Grant ROC curve

ROC curve farther from the base value than the previous model with an AUC of 0.613.

The logistic regression model on Retained F17-F18 showed a ROC curve further away from the base value than the previous two models with an AUC of 0.820.

The logistic regression model on Completed Connect showed a ROC curve close to the top left corner with an AUC of 0.818.

The number of workshops attended was a significant predictor in the best-performing model's summary, with an exceptionally low p-value and three stars next to its name.

More than twice as many people finished the program compared to those who did not. The accuracy of each logistic regression model should be compared to the statistic that 89percent of people completed the course.

6 Appendix C: Code

```
Data<-read.csv("E:/bharathbijjam/MTH-522/Project-2/Report-1-College Students
   Data/Preliminary college year.csv")
str(Data)
library(dplyr)
```

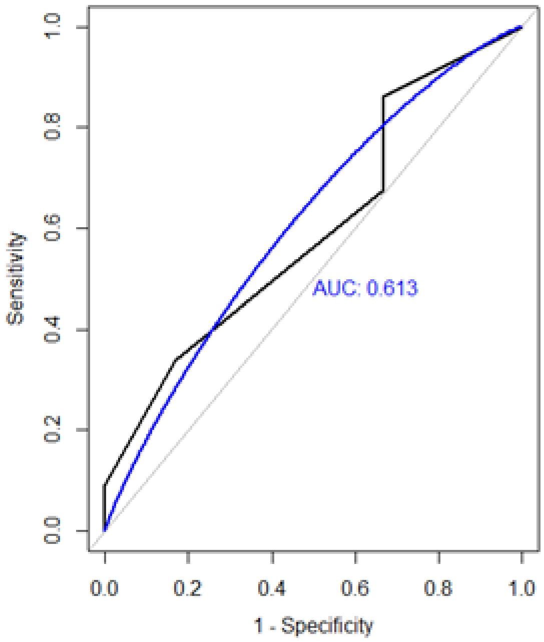


Figure 2: Community Service ROC curve

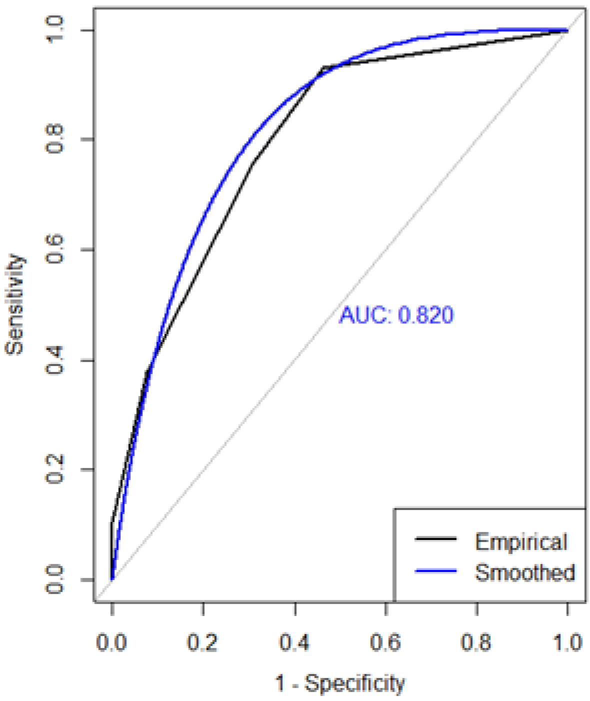


Figure 3: Retained F17-F18 ROC curve

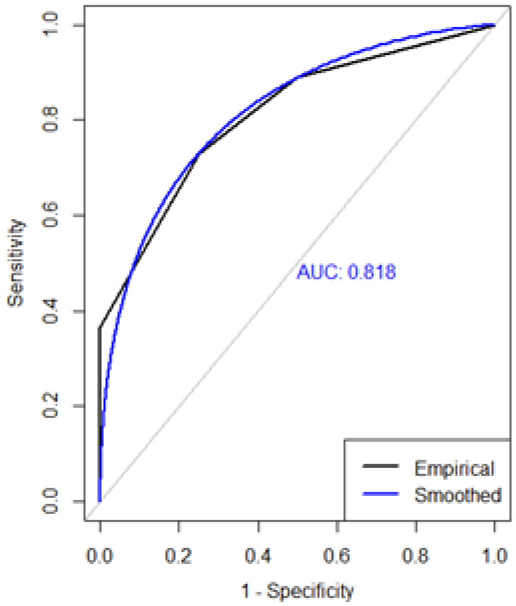
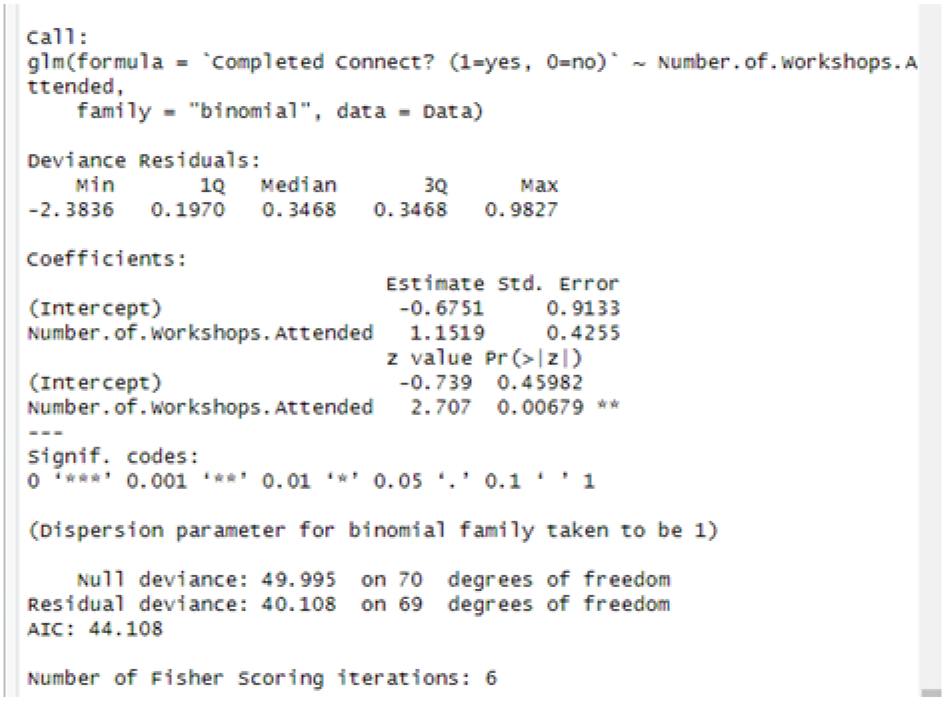


Figure 4: Completed connect ROC curve



```
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: Entire model summary

```
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
```