Logo

Description automatically generated

ALY6015 Module 3 Project – R Practice

**Authored by: Prateek Singh**

Course: ALY6015 – Intermediate Analytics

Submission Date: 02/09/2022

**Instructor’s name: Jiyoung Yun**

# Introduction

In this exercise we are going to practically implement the technique and skills about the generalized linear models to fit the Logistic Regression Model. GLM is traditional linear regression models with continuous response and continuous or categorical predictor variables.

Fitting any model, it turns out to be very crucial to divide the original data set into training and testing data set. We did the same here in this assignment as well. The training data set is employed to train the model and then the test data set is applied to test the same model. This ensures the model fitness is done accurately and appropriately and it will work correctly for any new data coming in. A common approach is to divide the data set in 70:30 ratio, where 70% of the data set is used to train the model and the rest 30% is for testing the trained model.

# Analysis

The College data set to be modelled in this assignment is the inbuilt R data set from the ISLR package and contains information about large number of colleges in US. Upon investigating the data set further, one strange and ambiguous data is recovered regarding the percentage of faculty holding PhD degrees. There are total of 777 colleges listed in the data set with two classifications as whether they are Private or Public. Out of these 777 colleges, only Texas A&M University at Galveston has the percentage of faculties more than 100%, exact numbers are 103%, which is impossible. So, either the data captured is wrong or there was some mistake at the university end while making the calculation.

Graphical user interface

Description automatically generated with medium confidence

Fig. 1: University having maximum PhD Staff

Texas A&M Univ. at College Station has the maximum number of enrollments and can be considered as the most popular and yet affordable university amongst students. When compared with top 6 colleges students prefer to enroll in, the total expenses calculated comes to $11,286 for Texas A&M Univ. at College Station and is less than other assessed universities.

Graphical user interface

Description automatically generated with medium confidence

Fig. 2: Highest Enrolled University

We got interested to capture the least enrolled university after this and noticed, Capitol College with 90% acceptance rate is having lowest number of enrollments.

Graphical user interface

Description automatically generated with low confidence

Fig. 3: Least Enrolled University

We compared the Private vs Public University Applications in the below histogram plot and can see a considerable difference in the applications received by both the categories. Private Universities are receiving more applications than public universities.

Chart

Description automatically generated

Fig. 4: Private vs Public University Application Received

On similar ground, I checked if there is any similarity in the acceptance rate with the applications received and which category is accepting more students than other. It is evident that students are getting accepted more in private universities hence the inclination towards the private university can be explained.

Chart

Description automatically generated

Fig. 5: Private vs Public University Acceptance Rate

Above analysis explains the reason of more enrollments in the Private Universities. The claim is supported by below histogram plot as well.

Chart

Description automatically generated

Fig. 6: Private vs Public Enrollment Status

Moving ahead, it was very interesting to see the faculty available for each undergraduate students whether part time or full time. Below scatter plot graph tells us about two things in general. One is there are higher number of Full-time undergraduate students in the data set. This shows people are more interested in full time studies rather than part time and they consider full time to be more beneficial for their future. Second highlighting point to be noticed here is, averagely the maximum student faculty ratio range in most of the colleges lies between 10 to 18. Which means normally there is 1 Faculty assigned for every 10-18 students in US Colleges.

Chart, scatter chart

Description automatically generated

Fig. 7: Student & Faculty Ratio

Upon comparing Out of State Tuition Fees amongst all the colleges, we found that the maximum colleges fees lie in between $5000 to $15000. However, there are some extreme outliers too going above $20000 and below to $5000. To note here, it wasn’t possible to display names of all the 777 colleges hence the display has been adjusted likewise.

Chart

Description automatically generated

Fig. 8: Out of State Tuition Fees Distribution

Post analyzing the data set, we move on to fitting the model. Before doing that, we must see which variables are to be used in the model. We need an optimal model with not many features as it can become challenging. We run response versus all the predictors in the model and then select the best predictor variable for our model. Because we are running the logistic linear model, we have specified the family as ‘Binomial’. From the output, we can confidently state that except from Full Time Undergraduate, Out of state Tuition Fees and Alumni Percentage there is no other significant variables present in the model which can be used as the predictor variable. So, we went ahead to remove all the other variables and keep only the significant predictor variables to fit the model. We compared our significant predictors against each other using trial and test methodology to predict which amongst them form the best fit pair, yet I could only conclude that using all the three predictors together is the best fit as it yields the maximum difference between Null Deviance and the Residual Deviance. Also, the AIC value 216.2 is the lowest amongst other tested predictor variable permutations and thus supports the logic of the assertion.

Text

Description automatically generated

Fig. 9: Best Fit Logistic Regression Model

The coefficient indicates that the odds of Private increase by a factor of 0.999 for each one unit increase in full time undergrad students maintaining other variables constant. The odds of Private increase by a factor of 1.001 for each one unit increase in out of state tuition fees maintaining other variables constant. Finally, the odds of Private increase by a factor of 1.050 for each one unit increase in alumni percentage maintaining other variables constant.

Graphical user interface, text

Description automatically generated

Fig. 10: Coefficient of Logistic Regression Model

We predicted the probability of college being private or public. We assumed that any value equals to 0.5 (which is a typical threshold) or above is going to be considered as private and values below 0.5 will be denoted for public. Upon running this predictive model, I got six college details (due to head function) and all of them were categorised as Private.

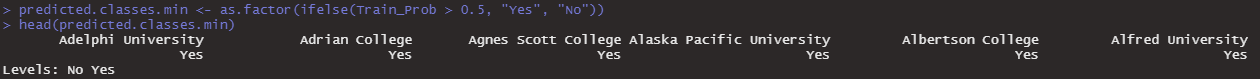


Fig. 11: Predicted Train Model

Since we have the predicted values and the actual values, we can simply compare those in a confusion matrix. We specify our true positives and true negatives very carefully, which means when we predict something it did happen exactly like the prediction. In our model, our true positive value comes out to be 381, true negative to be 128, false positive to be 21 and false negative to be 15. Our model shows 93.39% accuracy, which is very good, and the true positive and the true negative count is higher than the false positive and false negative count confirming the accuracy of our predictions. The value of prevalence also is higher around 73% stating that the number of Private colleges in our data set for which we are training the model for is more than the number of Public colleges. There are total of 565 Private colleges in our data set and remaining 212 are Public listings, hence the model is very good at predicting Private colleges.

Text

Description automatically generated

Fig. 12: Confusion Matrix for Training Model

We know for a fact that it is much more dangerous predicting something which is not going to happen, and it happens eventually because then we are never prepared for the same. Nevertheless, if we predict something to happen and it never happens is comparatively less dangerous. Based on this contention, we can state that False Negative is more risky than False Positive. Based on these True Positive, True Negative, False Positive and False Negative values, we can calculate the accuracy, precision, recall and specificity of the train model. Accuracy tells us how accurate the model is in predicting the outcomes. Precision minimizes the False Positives in the model. Recall gives the proportion of True Positive values and used to minimize the False Negatives. Specificity is the ability of the model being trained to identify the True Negatives. The values for each metrics are presented in the below figure.

Text

Description automatically generated

Fig. 13: Train Model Statistics

Now we need to create a confusion matrix for Test data set, so that we can test the trained model precisely. Similar to what we did for the training data set, we can repeat the process just using test data set this time. We can see the accuracy of Test model, 93.53% is very alike to Training Model and hence there is not a likelihood of overfitting the model. The TP and TN values are again much greater than FP and FN making the model dependable.

Text

Description automatically generated

Fig. 14: Confusion Matrix for Testing Model

Post the modelling and confusion matrix analysis for both the models – train and test, is done we are going to plot receiver operator characteristic curve. We can see the vertical line is going straight up almost until the top left corner and then becoming horizontal. This shows that our model is adept of performing flawless predictions under the specified threshold of 0.5.

Chart

Description automatically generated

Fig. 15: Test Model ROC Curve

And finally, Area under the ROC Curve is calculated to be 0.977.



Fig. 16: Area under the curve

# Conclusion

We can safely conclude the model we trained and tested is accurate and will yield precise results. Our model is 93-94% accurate and qualifies other conditions of False Positive and False Negative being less than True Positive and True Negative. Also, Yes values are greater than No values in our data set, all setting up the model to be a good fit for “Yes” prediction.

# Reference

1. 6.1 - Introduction to Glms. PennState: Statistics Online Courses. (n.d.). Retrieved February 9, 2022, from <https://online.stat.psu.edu/stat504/lesson/6/6.1>
2. brettljausn. (2017, October 18). How to represent categorical variable vs continuous variable using ggplot? Stack Overflow. Retrieved February 9, 2022, from <https://stackoverflow.com/questions/46802891/how-to-represent-categorical-variable-vs-continuous-variable-using-ggplot>
3. Bryan, thelatemail, & Irenabyss. (2013, July 11). *Exact axis ticks and labels in R lattice xyplot*. Stack Overflow. Retrieved February 9, 2022, from <https://stackoverflow.com/questions/17584248/exact-axis-ticks-and-labels-in-r-lattice-xyplot>
4. Elizabeth, A., Peterson, M., Turek, D., & Valagam, D. (2016, September 23). Creating a scatterplot with one X and multiple y values. Stack Overflow. Retrieved February 9, 2022, from <https://stackoverflow.com/questions/39646383/creating-a-scatterplot-with-one-x-and-multiple-y-values>
5. Heeter, T., Dey, S., & Li, Z. (2016, October 8). How to plot a one column data frame with ggplot? Stack Overflow. Retrieved February 9, 2022, from <https://stackoverflow.com/questions/39936690/how-to-plot-a-one-column-data-frame-with-ggplot>
6. Narula, M. (2020, December 30). Create stacked histogram in R. Delft Stack. Retrieved February 9, 2022, from <https://www.delftstack.com/howto/r/stacked-histogram-in-r/>

# Appendix

# ALY6015 Module 3 R Practice: Singh Prateek ------------------------------------------------

#----------------- Author: Prateek Singh

#----------------- Submission Date: 9th Feb, 2022

#----------------- Tutor: Jiyoung Yun

# Step: Installing New Libraries ------------------------------------------------

install.packages("ISLR")

install.packages("caret")

install.packages("lattice")

# Step: Importing Libraries ------------------------------------------------

library(ISLR)

library(dplyr)

library(psych)

library(caret)

library(ggplot2)

library(lattice)

library(pROC)

# Step: Load the inbuilt Dataset and do the EDA ------------------------------------------------

data("College")

str(College)

describe(College)

summary(College)

head(arrange(College, Enroll)) # Lowest Enrollment in the colleges

head(arrange(College, desc(Enroll))) # Highest Enrolling colleges

head(arrange(College, desc(PhD))) # To investigate the percentage going above 100% for faculty

ggplot(data = College, aes(x = Apps,fill = Private)) +

geom\_histogram() +

labs(title = "Private vs Public University Applications")

ggplot(data = College, aes(x = Accept,fill = Private)) +

geom\_histogram() +

labs(title = "Private vs Public University Acceptance Rate")

ggplot(data = College, aes(x = Enroll,fill = Private)) +

geom\_histogram() +

labs(title = "Private vs Public University Enrollment")

xyplot(F.Undergrad + P.Undergrad ~ S.F.Ratio, College, auto.key = TRUE,

ylab = "FullTime & PartTime Undergrad Students",

xlab = "Student Faculty Ratio",

xlim=c(0,45),

scales = list(

x = list(

at=seq(0,45,3)

)))

ggplot(College, aes(x = rownames(College), Outstate)) +

geom\_bar(stat='identity') +

labs(title = "Out of State Tuition Fees Distribution") +

theme(axis.text.x = element\_text(angle = 90)) +

scale\_x\_discrete(name = "Colleges", guide = guide\_axis(check.overlap = TRUE))

# Step: Split the data into a train and test set ------------------------------------------------

set.seed(3456)

Index\_Train <- createDataPartition(College$Private, p = 0.7, list = FALSE, times = 1)

Col\_train <- College[Index\_Train,]

Col\_test <- College[-Index\_Train,]

# Step: Fitting Logistic Regression Model to a training set ------------------------------------------------

model1 <- glm(Private ~., data = Col\_train , family = binomial(link = "logit"))

summary(model1)

model2 <- glm(Private ~ F.Undergrad + Outstate + perc.alumni, data = Col\_train , family = binomial(link = "logit"))

summary(model2)

coef(model2)

exp(coef(model2))

# Train set predictions

Train\_Prob <- predict(model2, newdata = Col\_train, type = "response")

predicted.classes.min <- as.factor(ifelse(Train\_Prob >= 0.5, "Yes", "No"))

# Step: Creating a confusion matrix for Train data set ------------------------------------------------

confusionMatrix(predicted.classes.min, Col\_train$Private, positive = 'Yes')

# Step: Calculating Accuracy, Precision, Recall, and Specificity ------------------------------------------------

TP <- 381

TN <- 128

FP <- 21

FN <- 15

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Accuracy

Precison = TP/(FP+TP)

Precison

Recall = TP/(TP+FN)

Recall

Specificity = TN/(TN+FP)

Specificity

# Step: Creating a confusion matrix for the test set ------------------------------------------------

# Test set predictions

Test\_Prob <- predict(model2, newdata = Col\_test, type = "response")

predicted.classes.min <- as.factor(ifelse(Test\_Prob >= 0.5, "Yes", "No"))

confusionMatrix(predicted.classes.min, Col\_test$Private, positive = 'Yes')

# Step: Plotting ROC Curve ------------------------------------------------

Test\_ROC <- roc(Col\_test$Private, Test\_Prob)

plot(Test\_ROC, col = "cyan4", ylab = "Sensitivity - TP Rate", xlab = "Specificity - FP Rate")

# Step: Calculating the Area under the ROC Curve ------------------------------------------------

Test\_AUC <- auc(Test\_ROC)

Test\_AUC