Logo

Description automatically generated

**College of Professional Studies**

**Northeastern University San Jose**

**MPS Analytics**

**Course: ALY6015: Intermediate Analytics**

**Assignment:**

Module 3 Assignment - GLM and Logistic Regression

**Submitted on:**

January 29, 2023

**Submitted to:**  **Submitted by:**

Professor: PAROMITA GUHA NIKSHITA RANGANATHAN

# **INTRODUCTION**

**Generalized Linear Models (GLM)** is a framework that can be used to model a variety of data types, such as continuous, binary, count, and categorical data. It is a type of regression analysis that is used to predict a response variable from a set of predictor variables.

GLM is used to model non-linear relationships between variables, as well as to predict outcomes from a variety of data sources. GLMs can also be used to forecast future events using data from the past.

**Advantages of GLM:**

* GLMs can handle non-normal data and non-linear relationships, making them more versatile than traditional linear models.
* GLMs can also be used for model selection and variable selection, which can help identify the most important predictor variables.
* These models can also be regularized to prevent overfitting, which can improve the model's ability to generalize to new data.

**Three main elements of GLM:**

* The response variable's probability distribution is referred to as the random component. The normal, binomial, and Poisson distributions are common examples of probability distributions utilized in GLMs.
* Systematic component refers to the linear predictor, which is a linear combination of the independent variables (predictors) that are used to model the response variable.
* Link functions join the linear predictor with the response variable's mean. The link function is established using the response variable's distribution. Common examples of link functions include identity, logit, and log links.

**Logistic regression** is a type of GLM that is used to model binary outcomes, such as the probability of success or failure in a binary classification problem.

In logistic regression, the response variable is assumed to have a binomial distribution with a probability of success p, and the link function is the logit function, which is defined as log(p/(1-p)). The logit function maps the probability of success to the real line, which allows for a linear relationship between the independent variables (predictors) and the log odds of success.

A table called a confusion matrix is employed to describe how well a classification algorithm performs.

The matrix's rows represent the predicted values, while its columns represent the actual values.

Using a confusion matrix, you may figure out several performance indicators including accuracy, precision, recall, and specificity.

A binary classifier's performance is graphically represented by a **Receiver Operating Characteristic curve**. The true positive rate and false positive rate are plotted against one another in this graph.

A perfect classifier will have a TPR of 1 and a FPR of 0, which corresponds to the top-left corner of the ROC space.

The ability of the classifier to differentiate between the positive and negative classes is measured by the **Area Under the ROC Curve (AUC-ROC).**

**About the dataset:**

The ISLR package in R includes a dataset called "College" which contains information about various colleges and universities in the United States.

**Purpose:**

The purpose of this project is to perform logistic regression modeling in order to predict whether a university is a private or public university. Our focus is also to gain a deeper understanding of how different factors influence the enrollment of students in colleges and universities and to use this understanding to improve predictions about enrollment in a college based on different characteristics of the institutions.

The dataset includes the following features:

|  |  |  |
| --- | --- | --- |
| **No** | **Feature** | **Dictionary** |
| 1 | Private | whether the college is private or public (binary variable) |
| 2 | Apps | total applications received |
| 3 | Accept | total applications accepted |
| 4 | Enroll | total students enrolled |
| 5 | Top10perc | Top 10% - new students from their high school class (in %) |
| 6 | Top25perc | Top 25% - new students from their high school class (in %) |
| 7 | F.Undergrad | the number of undergraduates enrolled full-time |
| 8 | P.Undergrad | the number of undergraduates enrolled part-time |
| 9 | Outstate | the out-of-state tuition |
| 10 | Room.Board | the room and board costs |
| 11 | Books | the estimated book costs |
| 12 | Personal | the estimated personal spending |
| 13 | PhD | faculty with PhD. (in %) |
| 14 | Terminal | faculty with a terminal degree (in %) |
| 15 | S.F.Ratio | student/faculty ratio |
| 16 | perc.alumni | alumni who donate (in %) |
| 17 | Expend | the instructional cost for each student |
| 18 | Grad.Rate | rate of graduating |

*Table 1: Features of the College Data Set with their dictionary*

**ANALYSIS & INTERPRETATION**

**1. Importing the dataset, Descriptive statistics, and EDA**

* Installing and loading the libraries

I have installed and loaded multiple packages in R for this assignment such as ISLR, ellipse, corrgram, RColorBrewer, dplyr, psych, skimr, corrplot, gridExtra etc. These packages provide various functions and tools for data analysis, visualization, and modeling, such as statistical analysis, correlation plot, data visualization and manipulation, imputation of missing values, machine learning, and many more.

* Importing the dataset

In this step, we load the "College" data set and attach it to the R environment. The "College" data set is a built-in data set in the ISLR package.

* Understanding the College dataset
* The str() function will return the structure of the data frame and data types of each variable and summary() function will provide summary statistics such as mean, median, minimum, and maximum for each variable.
* 'College' is a data frame containing 777 observations of 18 variables.

A screenshot of a computer

Description automatically generated with low confidence

**A screenshot of a computer

Description automatically generated with low confidence**

**Figure 1 – str() and summary()**

* Descriptive statistics

To get a quick overview of the dataset's key features, the skim function is used. This function provides a summary of the variable types, the number of missing values, mean, standard deviation, percentiles, and a histogram of the variable distribution.

* It shows that the dataset has a one-factor variable (Private) and 17 numeric variables.

Text, table

Description automatically generated with medium confidence

**Figure 2 – skim()**

Next, I wanted to generate descriptive statistics for a given dataset, so I used describe function.

* The mean value of the variable "Apps" is 3001.64, the median value is 1558, and the minimum and maximum values are 81 and 48094 respectively.
* On the other side, "Accept" has a mean value of 2018.80 and a median value of 1110. 72 and 26330 are the min and max values.

A positive skew indicates that more of the tail of the distribution is on the positive side of the mean (i.e., the right tail is longer). A high kurtosis indicates that the distribution has a sharper peak and heavier tails than a normal distribution, while a low kurtosis indicates that the distribution has a flatter peak and lighter tails.

* Variables like "Apps", "Accept", and "Enroll" have a positive skew and high kurtosis.

A picture containing text

Description automatically generated

**Figure 3 – describe()**

I wanted to understand two groups- Private Yes and Private No. So, I used describeBy to generate summary statistics for different groups within a dataset.

* There are 212 universities that are public and 565 universities that are private.
* It is observed that the mean number of applications received, accepted, and student enrollments for public institutions is more than that of private institutions.
* Public institutions typically have larger enrollment numbers and may receive more applications due to their lower tuition costs and often more extensive outreach efforts compared to private institutions.
* Additionally, public institutions may have more open admissions policies, which could also contribute to higher application and enrollment numbers.

**Text

Description automatically generated**

**A picture containing table

Description automatically generated**

**Figure 4 – describeBy()**

* Check missingness

To check for null or <NA> values in a college dataset, you can use the is.na() function or the is.null() function in R. There are no missing values in the dataset.

**A picture containing logo

Description automatically generated**

**Shape

Description automatically generated**

**Figure 5 – NA and null values**

* Duplicate rows

The college dataset does not contain any duplicate rows.

* Exploratory Data Analysis

**Scatterplots:**

The number of applications received and accepted for both public and private universities are positively correlated. This may be because of a variety of factors such as the reputation and prestige of the university, the quality of the programs offered, and the availability of financial aid.

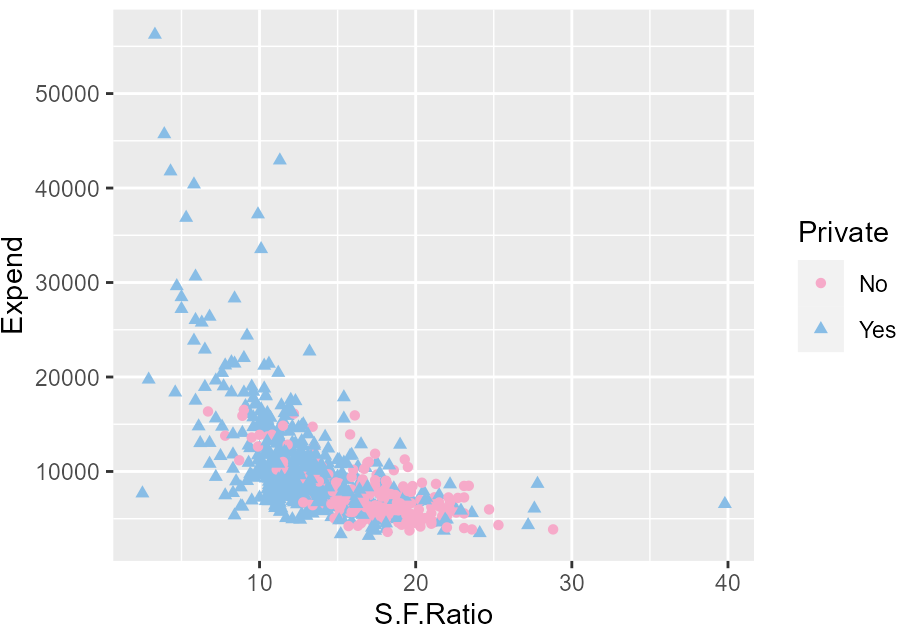
Chart, scatter chart

Description automatically generated

**Figure 6 – Scatterplot A (Number of applications received vs accepted)**

The relationship between the student-faculty ratio and the instructional expenditure per student for private and public universities is likely to be inversely correlated. This could be due to several factors, such as the availability of funding, the size of the university, and the number of students enrolled.

In general, private universities tend to have smaller student-faculty ratios and higher instructional expenditure per student, while public universities tend to have larger student-faculty ratios and lower instructional expenditure per student. This may be because private universities have more resources to devote to smaller class sizes and more individualized instruction, while public universities may have more students to serve and less funding to allocate to instructional expenses.

****

**Figure 7 – Scatterplot B (Instructional expenditure per student vs student-faculty ratio)**

**Boxplots:**

Both Outstate tuition and Graduation rate is more in private universities.

The higher graduation rate in private institutions compared to public institutions is possibly due to factors like more personalized attention from professors and staff. Additionally, private institutions may have more selective admissions processes, meaning that the student body is generally more academically prepared and motivated to succeed.

Out-of-state tuition, in particular, is often higher at private institutions because these institutions do not receive the same level of funding from state governments as public institutions. This means that private institutions have to charge more for tuition to make up for the lack of funding, which can make attendance more expensive for students from out of state.

**Chart, box and whisker chart

Description automatically generated**

**Figure 8 – Boxplots of Out-of-state tuition and Graduate rate**

**Barplot:**

The below graph shows the enrollment numbers for the top 10 universities, with the highest enrollment ranging from around 4500 to 6400.

Chart, bar chart

Description automatically generated

**Figure 9 – Barplot (Number of enrollments vs name of the university)**

**Histograms:**

The histogram for Apps, Accept, Enroll, Expend, and Top10perc has right-skewed distribution whereas Top25perc and Outstate have a multimodal distribution. It appears that the graduation rate's distribution, as depicted in the histogram, is very similar to a normal distribution.

Chart, histogram

Description automatically generated

**Figure 10 – Histograms**

**Correlation:**

* Enroll has a very strong positive relationship with F.Undergrad (Correlation Coefficient – 0.96)
* Apps has a very strong positive relationship with Accept (Correlation Coefficient – 0.94)
* Accept has a very strong positive relationship with Enroll (Correlation Coefficient – 0.91)

A picture containing background pattern

Description automatically generated

A picture containing table

Description automatically generated

**Figure 11 – Correlograms**

**2. Splitting the data into a train and test set**

An essential step in the machine learning process is dividing the data into train and test sets. The intention is to train a model using the training set, and then use the test set to assess the model's performance.

The data is typically split into two parts, with a percentage of the data (such as 70%) used for training, and the remaining percentage (such as 30%) used for testing. This allows the model to be trained on a large portion of the data and then evaluated on a smaller portion of unseen data to see how well it performs on new data.

**3. Logistic regression model using at least two predictors**

The glm (Generalized Linear Model) function is a function in the 'stats' package in R that is used to fit a variety of different types of models, including linear regression, logistic regression, and other types. To use this function, you would need to specify the dependent variable (in this case, the Private variable) and at least two predictors.

Using every predictor present in the train data set, we will fit a logistic regression model for Model 1 in this case.

A picture containing text, newspaper, receipt

Description automatically generated

**Figure 12 – Coefficients for Model 1**

The null deviance in both models represents the deviance of a null model (i.e., a model that only includes the intercept) on the training dataset.

The residual deviance represents the deviance of the model with predictors included, in the training dataset.

A statistical model's relative quality is evaluated by the AIC (Akaike information criterion), where lower values signify a better model.

**Null deviance: 639.40 on 544 degrees of freedom**

**Residual deviance: 168.25 on 527 degrees of freedom**

**AIC: 204.25**

When a variable's p-value is less than 0.05, it indicates that the association between the variable and the dependent variable in the model is statistically significant. These variables include Apps, Enroll, F.Undergrad, Outstate, Terminal, perc.alumni

For Model 2, we will fit a logistic regression model to the training dataset using only the significant predictors identified in Model 1 (p-value less than 0.05).

Text

Description automatically generated

**Figure 13 – Coefficients for Model 2**

**Null deviance: 639.40 on 544 degrees of freedom**

**Residual deviance: 186.56 on 538 degrees of freedom**

**AIC: 200.56**

Based on the AIC values, it appears that Model 2 with 200.56 is the better model.

**4. Confusion matrix for the train set**

A screenshot of a computer

Description automatically generated with medium confidence

**Figure 14 – Confusion Matrix (Training Data)**

A false positive occurs when a test result incorrectly indicates that a particular condition or attribute is present. On the other hand, a false negative occurs when a test result incorrectly indicates that a particular condition or attribute is not present.

**True Negative - 129**

**False Negative – 18 (Type II Error)**

**True Positive - 378**

**False Positive - 20 (Type I Error)**

**Which is more damaging for the analysis, False Positives or False Negatives?**

Both false positives and false negatives can be damaging to analysis. However, which one is more damaging would depend on the specific use case and objectives of the analysis.

If the goal is to identify the private universities for targeted funding, then a false positive (predicting a public university to be private) would be more damaging as resources would be wasted on a university that is not private. Whereas, if the goal is to identify the public universities for governmental regulations and monitoring, then a false negative (predicting a private university to be public) would be more damaging as a private university may not be subjected to the same regulations and monitoring as public universities.

It's essential to consider the impact of misclassification on the decision-making process before determining which type of misclassification is more damaging

**5. Accuracy, Precision, Recall, and Specificity.**

Accuracy, precision, recall, and specificity are all commonly used metrics to evaluate the performance of a binary classification model.

* Accuracy is the proportion of correct predictions made by the model (both true positive and true negatives) out of the total number of predictions. It can be calculated by (TP+TN)/(TP+TN+FP+FN) where TP is the no. of true positives, TN is the no. of true negatives, FP is the no. of false positives and FN is the no. of false negatives.
* Precision is the proportion of true positive predictions out of the total number of positive predictions made by the model. It can be calculated by TP/(TP+FP).
* Recall (also known as sensitivity or true positive rate) is the proportion of true positive predictions out of the total number of actual positive cases. It can be calculated by TP/(TP+FN).
* Specificity (also known as true negative rate) is the proportion of true negative predictions out of the total number of actual negative cases. It can be calculated by TN/(TN+FP).

From the table, it can be seen that the overall accuracy of the model is 0.9303, which means that the model correctly predicted the class (private or public) for 93.03% of the instances. The 95% CI (confidence interval) of the accuracy is (0.9056, 0.9502) which means that we are 95% confident that the true accuracy of the model is between 90.56% and 95.02%.

Sensitivity of 0.9545 is relatively high level of sensitivity, indicating that the model is able to detect a large proportion of actual positive cases. The model has a high level of specificity (0.8658), indicating that it can accurately identify a large proportion of negative cases.

**6. Confusion matrix for the test set.**

A screenshot of a computer

Description automatically generated with low confidence

**Figure 15 – Confusion Matrix (Test Data)**

Results of confusion matrix of test data:

* The confusion matrix shows the number of true positive (165), false positive (10), false negative (4), and true negative (53) predictions made by the model.
* The accuracy is 0.9397, meaning that the model correctly predicts the class for 93.97% of instances.
* The 95% Confidence Interval (CI) for the accuracy is (0.9008, 0.9666), which means that if the model was retrained multiple times with different training sets, the accuracy would fall within this interval in 95% of cases.
* The Precision of the model is 0.9429, meaning that 94.29% of the positive predictions made by the model are correct.
* Sensitivity is the true positive rate or the proportion of positive instances that are correctly classified as positive. The sensitivity of the model is 0.9763, meaning that it correctly identifies positive instances as positive 97.63% of the time.
* Specificity is the true negative rate or the proportion of negative instances that are correctly classified as negative. The specificity of the model is 0.8413, meaning that it correctly identifies negative instances as negative 84.13% of the time.

**7. ROC curve.**

Chart, histogram

Description automatically generated

**Figure 16 – ROC Curve**

In the figure, the curve is closer to the top-left corner of the plot, which means the accuracy of the model is high. This means that the model has a high true positive rate (sensitivity) and a low false positive rate (specificity).

**8. AUC**

Logo

Description automatically generated with low confidence

**Figure 17 – AUC for ROC Curve**

An AUC of 0.97 for a ROC curve is considered to be very good. It means that the model has a high ability to distinguish between positive and negative cases. An AUC of 0.97 indicates that the model has a high true positive rate and a low false positive rate, which is desirable for many applications.

**CONCLUSION**

Through this assignment, we were able to better comprehend the steps involved in logistic regression, including data preparation, model fitting, and evaluation, as well as the interpretation of the results. The model was able to accurately predict whether a university is private or public with an accuracy of 93% on the training and testing data set. The analyses' findings can be applied to higher education decision-making.

The key findings are :

* The dataset was imported, and exploratory data analysis was performed using descriptive statistics and plots to get an understanding of the data.
* The data was split into a training and test set in a 70-30 ratio to evaluate the accuracy of the model.
* The stats package's glm() function was used to fit the model (logistic regression).
* Metrics, confusion matrix, the ROC curve, and the AUC were used to measure the model's accuracy. The performance indicators such as Precision, Recall, and Specificity are also high and significant.
* The model showed good accuracy in both the training and test sets, with no signs of underfitting or overfitting, and the Type I and Type II errors were low.

**REFERENCES**

Bluman, A. G. (2018). Elementary Statistics, 10th ed. McGraw Hill.

Kabacoff, R. I. (2011). R in action: Data analysis and graphics with R. Manning Publications Co.

Alice, M. (2020, July 5). How to perform a Logistic Regression in R. R-bloggers. https://www.r-bloggers.com/2015/09/how-to-perform-a-logistic-regression-in-r/

Home - RDocumentation. (n.d.). https://www.rdocumentation.org/

**APPENDIX: CODE**

#---------------------- Week\_3\_Module\_3 R Script ----------------------#

print("Author : Nikshita Ranganathan")

print("Module 3 Assignment - GLM and Logistic Regression")

print("Course Name - ALY6015: Intermediate Analytics")

# Installing and loading the packages

library(ISLR)

library(ellipse)

library(corrgram)

library(RColorBrewer)

library(dplyr)

library(psych)

library(skimr)

library(corrplot)

library(gridExtra)

library(ggpubr)

library(MASS)

library(caret)

library(hrbrthemes)

library(grid)

library(pROC)

library(ochRe)

library(Amelia)

library(mlbench)

library(wesanderson)

# Load the data

attach(College)

College

# Descriptive statistics

str(College)

summary(College)

glimpse(College)

headTail(College)

dim(College)

skim(College)

describe(College)

describeBy(College,Private)

# Checking Missing data

missmap(College, col=c("blue", "red"), legend=FALSE)

sum(is.na(College))

sum(is.null(College))

# Checking for duplicated rows and removing them

duplicated(College)

anyDuplicated(College)

# Scatterplot

qplot(x=Apps,y=Accept,color=Private,shape=Private)+scale\_shape\_manual(values = c(16, 17)) +scale\_color\_manual(values = c("#F6AAC9", "#88BDE6"))

qplot(x=S.F.Ratio,y=Expend,color=Private,shape=Private)+scale\_shape\_manual(values = c(16, 17)) +scale\_color\_manual(values = c("#F6AAC9", "#88BDE6"))

# Boxplots

y<-qplot(x=Private,y=Outstate,fill=Private,geom='boxplot')+guides(fill=FALSE)+scale\_fill\_manual(values= wes\_palette("Chevalier1", n = 2))

z<-qplot(x=Private,y=Grad.Rate,fill=Private,geom='boxplot')+guides(fill=FALSE)+scale\_fill\_manual(values= wes\_palette("Chevalier1", n = 2))

grid.arrange(y,z,nrow=1)

# Barplot

df1<-College %>% arrange(desc(Enroll))

df1<-head(df1,10)

University\_name<-row.names(df1)

ggplot(df1, aes(x=University\_name, y=Enroll,fill=University\_name)) + geom\_bar(stat = "identity")+coord\_flip()+scale\_fill\_brewer(palette="Set3")+ggtitle("Universities with Top 10 Enrollments")+labs(x = "University Name", y = "Number of Enrollments",fill = "University Name")

# Histograms

hist1<-ggplot(College,aes(x=Apps)) +geom\_histogram(fill= "#00A9FF",colour = "black")

hist2<-ggplot(College,aes(x=Accept)) +geom\_histogram(fill= "#A54657",colour = "black")

hist3<-ggplot(College,aes(x=Enroll)) +geom\_histogram(fill= "#97ce4c",colour = "black")

hist4<-ggplot(College,aes(x=Top10perc)) +geom\_histogram(fill= "#00C19A",colour = "black")

hist5<-ggplot(College,aes(x=Top25perc)) +geom\_histogram(fill= "#F39B7FFF",colour = "black")

hist6<-ggplot(College,aes(x=Outstate)) +geom\_histogram(fill= "#938dd2",colour = "black")

hist7<-ggplot(College,aes(x=Expend)) +geom\_histogram(fill= "#F8766D",colour = "black")

hist8<-ggplot(College,aes(x=Grad.Rate)) +geom\_histogram(fill= "#00BFC4",colour = "black")

grid.arrange(hist1,hist2,hist3,hist4,hist5,hist6,hist7,hist8,top=textGrob("Histograms of Variables"))

# Correlation Matrix

corr <- select\_if(College, is.numeric)

cormatrix<-round(cor(corr,method = "pearson"),digits=2)

corrgram(corr, order = TRUE, upper.panel = panel.pie,lower.panel=panel.shade,

main = "Correlogram",)

corrgram(corr, lower.panel=panel.pts, upper.panel=panel.conf,

diag.panel=panel.density)

corr<-round(cor(corr,method = "pearson"),digits=2)

plotcorr(corr, col = colorRampPalette(c("#E08214", "white", "#8073AC"))(10), mar=c(0,0.1,0,0))

# Split data into train and test sets

datasplit <-createDataPartition(College$Private,p=0.7,list=FALSE)

train<-College[datasplit,]

test<-College[-datasplit,]

head(train)

head(test)

# Fit a logistic regression model

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

summary(model1)

# Display regression coefficients (log-odds)

coef(model1)

# Display regression coefficients (odds)

exp(coef(model1))

# Model 2 after variable selection

model2<-glm(Private~Apps+Enroll+F.Undergrad+Outstate+Terminal+perc.alumni,data=train,family=binomial(link="logit"))

summary(model2)

# Display regression coefficients (log-odds)

coef(model2)

# Display regression coefficients (odds)

exp(coef(model2))

# Train set predictions

probabilities.train <- predict(model2, newdata = train, type = "response")

predicted.classes.train <- as.factor(ifelse(probabilities.train >= 0.5, "Yes", "No"))

# Model Accuracy

cm\_train<-confusionMatrix(predicted.classes.train,train$Private,positive='Yes')

cm\_train

# Test set predictions

probabilities.test <- predict(model2, newdata = test, type = "response")

predicted.classes.test <- as.factor(ifelse(probabilities.test >= 0.5, "Yes", "No"))

# Model Accuracy

cm\_test<-confusionMatrix(predicted.classes.test,test$Private,positive='Yes')

cm\_test

# Plot the Receiver operator characteristic curve

ROC1<- roc(test$Private, probabilities.test)

plot(ROC1, col = "blue", ylab = "Sensitivity - TP Rate", xlab = "Sensitivity - FP Rate")

# Plot the Area Under the ROC Curve

auc <- auc(ROC1)

auc