**Logo

Description automatically generatedIntermediate Analytics**

**Module 3 GLM and Logistic Regression**

**ALY6015, Spring 2022**

**Week-3**

**Professor: Roy Wada**

**Submitted by: Abhinav Jain**

**Date: 05/01/2022**

Introduction

**Background**

In this paper, we used the college dataset from the ISLR library to analyze and forecast the results of the college, whether it is private or public. To get the findings, used R Language, and then execute both (GLM and logistics regression) models to forecast the outcome of the College dataset.

**GLM():** This is the model used when the linear regression model fails to predict a variable in a specific model. However, GLM consists of a random component that is an exponential family of probability distributions, a systemic component that works as a linear predictor , and a link function that helps in generalizing linear regression. R's glm() function is utilized in this implementation.

**Logistics Regression model**: In this method, the outcome of the input variable is determined by modeling the likelihood of the discrete outcome. Typically, logistics regression produces a binary result, which can have two values, such as 0/1, and yes/no. Even during logistics regression, there will be a single binary dependent variable that will be translated to “0 or 1” form, and also an independent variable that might be binary or continuous with real values.

On the other hand, while performing the regression modeling techniques in R, the college dataset has University name, private with yes and no which will be the dependent variable and depending on the correlation between the variable we have decided independent variable which will be discussed at later stage of the report while performing the regression modeling.

**Implementations**

Importing the college dataset in R from the ISLR library will enable insights to execute the logistics regression model, whether the institution is public or private, based on a model of one dependent and one or more independent variables using the logistics regression model. Furthermore, studying the dataset in R using the dplyr libraries describe () function will help in understanding statistics and cleaning the dataset as needed. To evaluate the dataset, we will create several graphs and gain a basic knowledge of the data.

While implementing the regression model, divide the dataset into test data and train data. Using the feature selection strategy, partition the data and use the glm() function to fit the model by keeping at least two predictors using the logistics regression model. A confusion matrix was utilized to examine the results, which helps in discovering the misclassification of false positives or false negatives. More findings like Accuracy, precision, recall, and specificity, plotting of ROC curve and calculating the AUC for the linear regression model.

**Task 1**: In this task Imported the dataset which has a dimension of 777 rows and 18 columns

Table 1: Descriptive Analysis

A picture containing text, receipt

Description automatically generated

Plots1: Created Scatter plot which show the private university and with blue dots whereas public university as red dots with respect to the application accepted in the private university.

**Chart1: Scatterplot Applications accepted in Private University**

Chart, scatter chart

Description automatically generated

Plot 2: In this below graph purple bar show the number of application received in private university in the given data

Chart2: Number of Applications in Private University

Chart, bar chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart

Description automatically generated

**Chart 4: Boxplot**

**Task 2**: In this task created a Training and Test dataset, whereas the test dataset of college has 234 observations with 18 variables and the training dataset of college has 543 observations with the same 18 variables. Moreover, split the data into 70:30 ratio to implementation of the logistics regression model.

**Training dataset descriptive analysis table**

**Table2: Train dataset**

**A picture containing text, receipt

Description automatically generated**

**Test dataset descriptive analysis Table**

**Table3: Test dataset**

A screenshot of a computer

Description automatically generated with low confidence

**Task3**: In this task fitting the logistics regression model by using the glm(), by using two predictor variables.

**Model 1:**

Text

Description automatically generated

**Model 2:**

Text, letter

Description automatically generated

**Accuracy Level of Model 2**: Model 2 has 93.58 percent accuracy to predict the value from the new variable which will give the best result whether the university is private or public.



**Precision level:**

A picture containing logo

Description automatically generated

**Recall:**

A picture containing logo

Description automatically generated

**Specificity:**



**ROC curve**

In this finding, the below curve indicates that the value of predicting the new variable will show the success and failure rate of the model. In our implementation of the model, the success rate would be 93% as per the variables we have selected with reference to the significance level of the variable. This graph clearly shows the clinical sensitivity and specificity for the most possible out would be positive with 95 % of specificity.

**Chart 5**: ROC Curve

Chart

Description automatically generated

**AUC:** This finding suggests that the model we have implemented will gives the 98% possibility to define the best outcome between the range of 0.5 to 1.

A picture containing text

Description automatically generated

**Conclusion:**

In this report, we have implemented the logistics regression model. Firstly, Importing the college dataset in R using the ISLR package will allow insights to run the logistics regression model, whether the school is public or private, using a model of one dependent and one or more independent variables. In addition, reviewing the dataset in R with the dplyr library description () function will aid in understanding statistics and cleaning the dataset as needed. To assess the dataset, we will generate numerous graphs and obtain a basic understanding of the data. Divide the dataset into test and train data while executing the regression model. Partition the data using the feature selection method and use the glm() function to fit the model by maintaining at least two predictors using the logistics regression. By implementing the current model we can get the 93 % accuracy of results if new observations will be added to the dataset to analyze whether the university is public or private. Whereas, the

**References:**

[0] <https://github.com/nguyen-toan/ISLR/blob/master/dataset/College.csv>

[0.1]https://rdrr.io/cran/ISLR/man/College.html#heading-3

[1] <https://towardsdatascience.com/linear-regression-or-generalized-linear-model-1636e29803d0>

[2] <https://en.wikipedia.org/wiki/Generalized_linear_model>