**Generalized Linear Model (GLM) and Logistic Regression**

**Introduction**

The Generalized Linear Model (GLM) is a statistical modeling technique that has become popular since its introduction by John Nelder and Robert Wedderburn in 1972. Unlike linear regression, which assumes that the response variable is continuous and normally distributed, the GLM can model a wider range of response variables, including binary, categorical, and count data with different error distributions. This flexible framework can model the relationship between the response variable and explanatory variables while considering the distribution of the response variable.

The GLM can handle various distributions, such as normal, Poisson, binomial, and gamma distributions, among others, making it an invaluable tool in data analysis across many fields.

**Analysis:**

1. Loading the college data set.

To load the college data set, we have to install the package ISLR In R.

Table

Description automatically generated with medium confidence

After extracting the 'college' dataset, it was found to have a total of 18 columns and 777 rows.

**Exploratory Data Analysis**

To begin, we have created a bar graph to visualize the number of students, out of the 777 observations, enrolled in public and private universities. The graph reveals that the count for public universities is higher than that of private universities, with public universities comprising more than 500 observations and private universities around 200 observation

Chart, bar chart

Description automatically generated

**Descriptive statistics**

The table shows descriptive statistics for 18 variables in the "College" dataset, including information on private/public status, number of applications, acceptance rate, enrollment, top percentage of students, undergraduate and graduate numbers, room and board costs, alumni rate, and more. Each variable has its own mean, standard deviation, minimum, maximum, and range, among other measures. The data suggest that there are differences among colleges with respect to these variables.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Column1** | **vars** | **n** | **mean** | **sd** | **median** | **trimmed** | **mad** | **min** | **max** | **range** | **skew** | **kurtosis** | **se** |
| **Private\*** | 1 | 777 | 1.73 | 0.45 | 2 | 1.78 | 0 | 1 | 2 | 1 | -1.02 | -0.96 | 0.02 |
| **Apps** | 2 | 777 | 3001.64 | 3870.2 | 1558 | 2193.01 | 1463.33 | 81 | 48094 | 48013 | 3.71 | 26.52 | 138.84 |
| **Accept** | 3 | 777 | 2018.8 | 2451.11 | 1110 | 1510.29 | 1008.17 | 72 | 26330 | 26258 | 3.4 | 18.75 | 87.93 |
| **Enroll** | 4 | 777 | 779.97 | 929.18 | 434 | 575.95 | 354.34 | 35 | 6392 | 6357 | 2.68 | 8.74 | 33.33 |
| **Top10perc** | 5 | 777 | 27.56 | 17.64 | 23 | 25.13 | 13.34 | 1 | 96 | 95 | 1.41 | 2.17 | 0.63 |
| **Top25perc** | 6 | 777 | 55.8 | 19.8 | 54 | 55.12 | 20.76 | 9 | 100 | 91 | 0.26 | -0.57 | 0.71 |
| **F.Undergrad** | 7 | 777 | 3699.91 | 4850.42 | 1707 | 2574.88 | 1441.09 | 139 | 31643 | 31504 | 2.6 | 7.61 | 174.01 |
| **P.Undergrad** | 8 | 777 | 855.3 | 1522.43 | 353 | 536.36 | 449.23 | 1 | 21836 | 21835 | 5.67 | 54.52 | 54.62 |
| **Outstate** | 9 | 777 | 10440.67 | 4023.02 | 9990 | 10181.66 | 4121.63 | 2340 | 21700 | 19360 | 0.51 | -0.43 | 144.32 |
| **Room.Board** | 10 | 777 | 4357.53 | 1096.7 | 4200 | 4301.7 | 1005.2 | 1780 | 8124 | 6344 | 0.48 | -0.2 | 39.34 |
| **Books** | 11 | 777 | 549.38 | 165.11 | 500 | 535.22 | 148.26 | 96 | 2340 | 2244 | 3.47 | 28.06 | 5.92 |
| **Personal** | 12 | 777 | 1340.64 | 677.07 | 1200 | 1268.35 | 593.04 | 250 | 6800 | 6550 | 1.74 | 7.04 | 24.29 |
| **PhD** | 13 | 777 | 72.66 | 16.33 | 75 | 73.92 | 17.79 | 8 | 103 | 95 | -0.77 | 0.54 | 0.59 |
| **Terminal** | 14 | 777 | 79.7 | 14.72 | 82 | 81.1 | 14.83 | 24 | 100 | 76 | -0.81 | 0.22 | 0.53 |
| **S.F.Ratio** | 15 | 777 | 14.09 | 3.96 | 13.6 | 13.94 | 3.41 | 2.5 | 39.8 | 37.3 | 0.66 | 2.52 | 0.14 |
| **perc.alumni** | 16 | 777 | 22.74 | 12.39 | 21 | 21.86 | 13.34 | 0 | 64 | 64 | 0.6 | -0.11 | 0.44 |
| **Expend** | 17 | 777 | 9660.17 | 5221.77 | 8377 | 8823.7 | 2730.95 | 3186 | 56233 | 53047 | 3.45 | 18.59 | 187.33 |
| **Grad.Rate** | 18 | 777 | 65.46 | 17.18 | 65 | 65.6 | 17.79 | 10 | 118 | 108 | -0.11 | -0.22 | 0.62 |

The below data provides various characteristics of colleges such as their status (private or not), admission details (number of applications, acceptances, and enrollments), student academic achievement (top 10% and 25% of high school class), number of full-time and part-time undergraduates, costs (out-of-state tuition, room and board, books, and personal expenses), faculty qualifications (PhD and terminal degrees), student-to-faculty ratio, alumni donations, expenditure per student, and graduation rate. The data also includes summary statistics such as the minimum, maximum, mean, median, and quartiles for each variable, providing a brief description of the range and distribution of the data.

Table

Description automatically generated

**Plots**

1. This box plot depicts the distribution of room and board fees for both public and private universities. It is apparent that the median room and board fees for public universities are relatively lower, averaging at approximately $3500, compared to private universities, where the average fees are around $4500. The maximum room and board fees for public universities are approximately $7000, while for private universities, the maximum fees are approximately $8000.

Chart, box and whisker chart

Description automatically generated

The graph displays data regarding the number of applicants and enrollments in both public and private universities. The red dots indicate the public universities, while the green dots represent the private ones.

Chart, scatter chart

Description automatically generated

The below histogram shows us the distribution of out of state tuition fees.

Chart, histogram

Description automatically generated

**Split function**

Using the split function, we will partition the college data set into two sections, namely train and test. The train set will encompass 75% of the entire data set, while the test set will account for 25% of the college data set. The train set will be utilized to develop the model, whereas the test set will serve the purpose of evaluating and testing the accuracy of the model.

**Train**

Calendar

Description automatically generated with medium confidence

**Test**

**A picture containing calendar

Description automatically generated**

The data has been split into two subsets for training and testing purposes. The first subset contains 75% of the original data and is used for training the model. The second subset contains 25% of the original data and is used for evaluating the model's performance on unseen data. This technique is commonly known as a 75/25 split or a 3:1 ratio split. By using a portion of the data for testing, we can assess the model's ability to generalize and make predictions on new data. This process helps to prevent overfitting and ensures that the model's performance is not biased towards the training data.

**GLM**

|  |  |  |
| --- | --- | --- |
| Call: |  |  |
| glm(formula = Private ~ Apps + Enroll + Grad.Rate + PhD | family = "binomial" |  |
| data = train) |  |  |
|  |  |  |
| Deviance Residuals: |  |  |
| Min 1Q Median 3Q Max |  |  |
| -2.9453 -0.0416 0.2759 0.4464 3.2812 |  |  |
|  |  |  |
| Coefficients: |  |  |
| Estimate Std. Error z value Pr(>|z|) |  |  |
| (Intercept) -8.315e-01 7.887e-01 -1.054 0.2918 |  |  |
| Apps 1.507e-05 1.013e-04 0.149 0.8817 |  |  |
| Enroll -3.015e-03 5.310e-04 -5.679 1.36e-08 \*\*\* |  |  |
| Grad.Rate 8.849e-02 1.137e-02 7.785 6.96e-15 \*\*\* |  |  |
| PhD -1.947e-02 1.068e-02 -1.823 0.0683 . |  |  |
| --- |  |  |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |  |  |
|  |  |  |
| (Dispersion parameter for binomial family taken to be 1) |  |  |
|  |  |  |
| Null deviance: 639.40 on 544 degrees of freedom |  |  |
| Residual deviance: 328.74 on 540 degrees of freedom |  |  |
| AIC: 338.74 |  |  |
|  |  |  |
| Number of Fisher Scoring iterations: 6 |  |  |

From the above GLM Model we can determine the following points:

* GLM models use a formula that relates predictor variables to the response variable.
* The provided R code includes four predictor variables, namely Apps, Enroll, Grad.Rate, and PhD, along with an intercept term.
* The coefficients obtained from the GLM model can be used to determine the relationship between each predictor variable and the response variable.
* A one-unit change in Apps corresponds to a 1.507e-05 change in the response variable, a one-unit change in Enroll corresponds to a -3.015e-03 change in the response variable, a one-unit change in Grad.Rate corresponds to an 8.849e-02 change in the response variable, and a one-unit change in PhD corresponds to a -1.947e-02 change in the response variable.
* The p-values for each predictor variable can be used to determine their statistical significance.
* Enroll and Grad.Rate have p-values less than 0.001, indicating that they are statistically significant predictors, while Apps and PhD have p-values much higher than 0.001, suggesting that they are not statistically significant predictors.
* Enroll and Grad.Rate are more important predictors in the model, while Apps and PhD do not have a significant impact on the response variable.

**Confusion matrix for train data set**

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion Matrix and Statistics Reference |  |  |  |
| Prediction No Yes | No Yes |  |  |
|  | 98 | 17 |  |
|  | 51 379 |  |  |
|  | Accuracy . 0.8752 |  |  |
|  | 95% Cl . (0.8445, 0.9018) |  |  |
| No Information Rate |  |  | 0.7266 |
| P-value [Acc > NIR] . < 2.2e-16 Kappa . 0.6619 Mcnemar's Test P-Va1ue . 6.285e-05 Sensitivity . 0.9571 Specificity . 0.6577 Pos Pred Value . 0.8814 |  |  |  |
| Neg Pred Value |  |  | 0.8522 |
| preval ence |  |  | 0.7266 |
| Detection Rate . 0.6954 |  |  |  |
| Detection Prevalence |  |  | 0.789 |
| Balanced Accuracy . 0.8074 'Positive' Class . Yes |  |  |  |

* The provided confusion matrix shows that there are 51 false-positive cases and 17 false-negative cases.
* There are 379 true-positive cases and 98 true-negative cases.
* The number of true-positive and true-negative events is greater than the number of false-positive and false-negative events, indicating that the model is performing reasonably well.
* The accuracy of the model is calculated to be 87.52%, which is about average.
* The precision of the model is calculated to be 0.8814, indicating that the forecast is correct 88.14% of the time for private colleges.
* The sensitivity of the model is calculated to be 0.9571, indicating that the prediction that the college is private is correct 95.71% of the time.
* The specificity of the model is calculated to be 0.6577, indicating that the college will be classified as public or private 65.77% of the time.

**Confusion matrix for test data set**

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion Matrix and Statistics Reference |  |  |  |
| Prediction No | No Yes |  |  |
|  | 45 | 12 |  |
| Yes | 18 157 |  |  |
|  | Accuracy . 0.8707 |  |  |
|  | 95% Cl . (0.8206, o. 911) |  |  |
| No Information Rate |  |  | 0.7284 |
| P-value [Acc > NIR] |  |  | 1. 326e-07 |
| Kappa . 0.6631 |  |  |  |
| Mcnemar's Test P-Va1ue |  |  | 0. 3613 |
| Sensitivity . 0.9290 Specificity . 0.7143 |  |  |  |
| Pos Pred Value |  |  | 0.8971 |
| Neg Pred Value |  |  | 0.7895 |
| Prevalence |  |  | 0.7284 |
| Detection Rate |  |  | 0.6767 |
| Detection Preval ence |  |  | 0.7543 |
| Bal anced Accuracy . 0.8216 'Positive' Class : Yes |  |  |  |

* The provided confusion matrix indicates that there are 18 false-positive cases and 12 false-negative cases.
* There are 157 true-positive cases and 45 true-negative cases.
* The number of true-positive and true-negative events is greater than the number of false-positive and false-negative events, indicating that the model is performing reasonably well.
* The accuracy of the model is calculated to be 87.07%, which is fairly average.
* The precision of the model is calculated to be 0.8917, indicating that the forecast is correct 89.17% of the time for private colleges.
* The sensitivity of the model is calculated to be 0.9290, indicating that the prediction that the college is private is correct 92.90% of the time.
* The specificity of the model is calculated to be 0.7143, indicating that the college will be classified as public or not private 71.43% of the time.

**Confusion Matrix for Best Model Test data**

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion Matrix and Statistics Reference |  |  |  |
| Predi ction | No Yes |  |  |
| No | 53 | 11 |  |
| Yes | 10 158 |  |  |
|  | Accuracy |  | 0.9095 |
|  | 95% Cl |  | (0.865, 0.9431) |
| No Information Rate |  |  | 0.7284 |
| P-value [Acc > NIRJ |  |  | 5.89E-12 |
| Kappa . 0.7723 |  |  |  |
| Mcnemar's Test P-Va1ue |  |  | 1 |
| Sensitivity |  |  | 0.9349 |
| Speci fi ci ty |  |  | 0.8413 |
| Pos Pred Value |  |  | 0.9405 |
| Neg pred Value |  |  | 0.8281 |
| Preval ence |  |  | 0.7284 |
| Detection Rate |  |  | 0.681 |
| Detection preval ence . 0.7241 Bal anced Accuracy . 0.8881 'positive' Class . Yes |  |  |  |

The confusion matrix shows the classification model's test data predictions. The accuracy is 0.9095 and the kappa score is 0.7723, indicating substantial agreement. The model has a sensitivity of 0.9349 and specificity of 0.8413, with a balanced accuracy of 0.8881. The positive predictive value is 0.9405 and negative predictive value is 0.8281.

**ROC Curve Plots**

1. **Roc Curve**

The ROC diagram indicates that the model's performance is suboptimal.

One reason for this is that the curve on the diagram does not touch the Y-axis and runs straight across the top.

The non-diagonal shape of the curve suggests that the model is only partially correct in its predictions.

Despite its suboptimal performance, the curve does show some improvement over random guessing, as it is closer to the y-axis.

To enhance the model's accuracy further, one possible solution is to provide it with more data.

Chart, line chart

Description automatically generated

1. **ROC Curve Test data**

Chart, line chart

Description automatically generated

The ROC diagram shows that the model is not optimal because the curve does not intersect the Y-axis and runs straight across the top.

Since the curve is not a diagonal line, the model is only partially correct.

However, the performance of the model is better because the curve is initially closer to the y-axis but moves closer to the diagonal line later.

To improve the model's performance, additional data can be provided.

1. **ROC Curve Best Model Test Data**

Chart, line chart

Description automatically generated

The ROC diagram shows that the model is optimal.

The curve is generated after calculating the confusion matrix for the best model for test data.

The best model roc curve is therefore generated in green.

**Calculating AUC**

|  |
| --- |
| **"AUC for train data: 0. 933174022100197"** |
| **"AUC for test data: 0. 893303277918662"** |
| **"AUC for the best mode/ for test data: 0.964684887761811"** |

The analysis calculates the AUC for a binary classification model using both the train and test datasets.

The AUC score for the train data is found to be 0.9332, which is higher than average and suggests that the model can make reasonably accurate predictions on the training data.

The AUC score for the test data is found to be 0.8933, indicating that the model is capable of making reasonable predictions on new, unseen data.

The AUC score for the best model on the test data is calculated using a function called auc().

The AUC score for the best model on the test data is found to be 0.9647, which is significantly higher than the other models, indicating that the best model performs much better than the other models.

Overall, the analysis provides insights into the performance of the binary classification model and its ability to make accurate predictions on both the training and test data.

**Conclusion:**

In conclusion, the Generalized Linear Model (GLM) is a powerful statistical modeling technique that can handle a wide range of response variables with different error distributions. This flexibility makes it an invaluable tool in data analysis across many fields. By analyzing the "College" dataset, we can see that there are differences among colleges with respect to various variables, including private/public status, number of applications, acceptance rate, enrollment, undergraduate and graduate numbers, room and board costs, alumni rate, and more. The insights gained from this analysis can be useful for making informed decisions about college admissions, funding, and other related matters.

**References:**

* Zach. (2022, December 19). How to create a confusion matrix in R (step-by-step). Retrieved January 29, 2023, from <https://www.statology.org/confusion-matrix-in-r/>
* Northeastern University. Retrieved from <https://northeastern.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=391e9cc3-cfbe-430e-a2bb-ac94003dfcd3&start=0>

**Appendix:**#Avi Milan Jani

#ALY6015 - Module 3 - GLM & Logistic Regression

#Installing packages

install.packages("ISLR")

install.packages("pROC")

install.packages("caret")

install.packages("corrplot")

install.packages("dplyr")

install.packages("ggplot2")

install.packages("correlation")

#Loading libraries

library(ISLR)

library(pROC)

library(caret)

library(corrplot)

library(dplyr)

library(ggplot2)

library(correlation)

#loading the Data set

data("College")

View(College)

# EDA

summary(College)

psych::describe(College)

str(College)

head(College)

tail(College)

#Number of Public and private universities

library(dplyr)

library(ggplot2)

College %>%

count(Private) %>%

ggplot(aes(x = Private, y = n, fill = Private)) +

geom\_bar(stat = "identity", width = 0.5) +

labs(title = "Number of Public and Private Universities",

x = "", y = "Count") +

scale\_fill\_manual(values = c("#2671B2", "#4F82C4")) +

scale\_x\_discrete(labels = c("Private", "Public")) +

theme(plot.title = element\_text(size = 16, face = "bold"),

axis.text = element\_text(size = 12),

axis.title = element\_text(size = 14),

legend.position = "none")

# Scatter plot of Accept vs Enroll

library(ggplot2)

ggplot(College, aes(x = Accept, y = Enroll, color = Private)) +

geom\_point(size = 3) +

geom\_smooth(method = "lm", se = FALSE, color = "black") +

labs(title = "Acceptance vs Enrollment by Private/Public University",

x = "Number of Applications", y = "Number of Enrollments") +

theme(plot.title = element\_text(size = 16, face = "bold"),

axis.text = element\_text(size = 12),

axis.title = element\_text(size = 14)) +

annotate("text", x = 40000, y = 5000, label = "Private universities", size = 4, color = "#2671B2") +

annotate("text", x = 40000, y = 3000, label = "Public universities", size = 4, color = "#4F82C4")

# Histogram of Outstate

hist(College$Outstate, breaks=30, col="#4F82C4", xlab="Out-of-State Tuition",

main="Distribution of Out-of-State Tuition Fees")

# Boxplot of Room.Board vs Private

boxplot(College$Room.Board ~ College$Private, xlab="Private", ylab="Room and Board Fees",

main="Room and Board Fees by Private/Public University", col=c("#2671B2", "#4F82C4"),

names=c("Public", "Private"))

#Split the data into a train and test set

library(caret)

set.seed(123)

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

train <- College[trainIndex,]

View(train)

test <- College[-trainIndex,]

View(test)

##GLM Model

glm\_model <- glm(Private ~ Apps + Enroll + Grad.Rate + PhD, family = "binomial", data = train)

View(glm\_model)

summary(glm\_model)

#Creating Confusion Matrix

##Log-odds coefficient

coef(glm\_model)

#Odds coefficient

exp(coef(glm\_model))

#Making predictions

train\_pred <- predict(glm\_model, newdata = train, type = "response")

train\_pred <- as.factor(ifelse(train\_pred >=0.5, "Yes","No"))

head(train\_pred)

#Accuracy of Model for the train set

confusionMatrix(train\_pred, train$Private, positive = 'Yes')

#Confusion Matrix for Test data

#Predictions

test\_pred <- predict(glm\_model, newdata = test, type = "response")

test\_pred

## GLM Model

library(stats)

College <- glm(Private ~ Apps + Enroll + Grad.Rate + PhD,

family = "binomial", data = train)

View(College)

summary(College)

## Creating Confusion Matrix

## Log-odds coefficient

coef(College)

## Odds coefficient

exp(coef(College))

## Making predictions

ptrain <- predict(College, newdata = train, type = "response")

pl <- as.factor(ifelse(ptrain >= 0.5, "Yes", "No"))

head(pl)

## Accuracy of Model

confusionMatrix(pl, train$Private, positive = "Yes")

## Confusion Matrix for Test data

## Predictions

predicttest <- predict(College, newdata = test, type = "response")

pt <- as.factor(ifelse(predicttest >= 0.5, "Yes", "No"))

head(pt)

## Accuracy of model

confusionMatrix(pt, test$Private, positive = "Yes")

## ROC

library(pROC)

## For train data set

ROC <- roc(train$Private, ptrain)

plot(ROC, main = "ROC Curve",

xlab = "False Positive Rate", ylab = "True Positive Rate", col = "red")

## For test data set

ROC\_test <- roc(test$Private, predicttest)

plot(ROC\_test, main = "ROC Curve Test Data",

xlab = "False Positive Rate", ylab = "True Positive Rate", col = "blue")

## Calculating AUC

AUC\_train <- auc(ROC)

AUC\_test <- auc(ROC\_test)

print(paste0("AUC for train data: ", AUC\_train))

print(paste0("AUC for test data: ", AUC\_test))

## AIC and BIC Values

print(paste0("AIC Value: ", AIC(College)))

print(paste0("BIC Value: ", BIC(College)))

## Stepwise Selection

## Stepwise logistic regression

library(MASS)

glm\_fit <- stepAIC(College, direction = "both", trace = FALSE)

## Best model

glm\_best <- glm(Private ~ Accept + Enroll + Outstate, data = train, family = "binomial")

summary(glm\_best)

## Confusion Matrix for Test data for the best model

## Predictions

predicttest\_best <- predict(glm\_best, newdata = test, type = "response")

pt\_best <- as.factor(ifelse(predicttest\_best >= 0.5, "Yes", "No"))

head(pt\_best)

## Accuracy of model

confusionMatrix(pt\_best, test$Private, positive = "Yes")

## ROC for the best model

## For test data set

ROC\_test\_best <- roc(test$Private, predicttest\_best)

plot(ROC\_test\_best, main = "ROC Curve Best Model Test Data",

xlab = "False Positive Rate", ylab = "True Positive Rate", col = "green")

## Calculating AUC

AUC\_test\_best <- auc(ROC\_test\_best)

print(paste0("AUC for the best model for test data: ", AUC\_test\_best))