**MODULE: 03**

## **BY: PAYAL SHARMA**

**“GLM and Logistic Regression”**

**Course: ALY 6015 CRN 71547 Intermediate Analytics**

**Instructor: Dr. Paromita Guha (Paro)**

**Winter 2025 – February 04**

**College of Professional Studies, CPS**

**A black background with a black square

Description automatically generated with medium confidence**

MODULE: 03

CONTENT

* Introduction
* Descriptive Analysis, Visualizations
* Splitting Dataset into Train-Test Sets
* GLM Function & Logistic Regression
* Confusion Matrix on Train Set
* Interpretation of Accuracy, Precision, Recall, Specifically
* Confusion Matrox on Test Set
* Plotting & Interpretation on ROC
* Interpretation on AUC
* References
* Appendix

**INTRODUCTION:**

**College Data:**

The College dataset, available in the ISLR (Introduction to Statistical Learning with R) package, is a widely used dataset in statistical learning and data analysis. It contains information about 777 U.S. colleges and universities, providing a comprehensive snapshot of various institutional characteristics. The dataset includes variables such as whether a college is public or private, the number of applications received, the number of students accepted and enrolled, out-of-state tuition costs, graduation rates, and other financial and academic metrics. These features make the dataset particularly useful for exploring relationships between institutional characteristics and outcomes, such as predicting whether a college is private, or public based on its attributes. The dataset is often employed in educational research, statistical modeling, and machine learning tasks, including classification and regression analysis. Its rich set of variables and real-world relevance make it an excellent resource for understanding the factors that influence higher education institutions and their performance.

**Overview of the Assignment:**

In this assignment, we utilized the **College dataset** from the **ISLR** package to build and evaluate a logistic regression model for predicting whether a university is private or public. The assignment began with **Exploratory Data Analysis (EDA)**, where we examined the dataset's structure, summarized its key variables, and visualized relationships using histograms, boxplots, and scatterplots. This step helped us understand the distribution of variables like out-of-state tuition (Outstate), graduation rates (Grad. Rate), and room and board costs (Room. Board), as well as identifying potential predictors for the logistic regression model.

Next, we split the dataset into **training and testing sets** to facilitate model training and evaluation. Using the glm() function in R, we fit a **multivariate** **logistic regression model** to the training data, with Private (a binary outcome variable) as the response and Outstate and Grad. Rate as predictors. We then evaluated the model's performance on the training set by creating a **confusion matrix** and calculating key metrics such as **accuracy**, **precision**, **recall**, and **specificity**. These metrics provided insights into the model's ability to correctly classify universities as private or public, while the confusion matrix highlighted the types of misclassifications (false positives and false negatives) and their potential implications.

To further assess the model's generalizability, we applied it to the **test set**, generated another confusion matrix, and analyzed its performance. Additionally, we plotted the **ROC curve** and calculated the **AUC (Area Under the Curve)** to evaluate the model's discriminatory power. This step aimed to enhance predictive accuracy and provide a more comprehensive understanding of the factors influencing university classification. Overall, the assignment demonstrated the end-to-end process of building, evaluating, and refining a logistic regression model for binary classification tasks.

**Part 1:**

The assignment began with the initial steps of setting up the R environment and loading the \*\*College dataset\*\* from the \*\*ISLR\*\* package. First, the \*\*ISLR\*\* package was installed and loaded using the `install.packages("ISLR")` and `library(ISLR)` commands, respectively. This package contains the dataset, which provides information about 777 U.S. colleges and universities. The dataset was then loaded into the R session using the `data("College")`command. To gain an initial understanding of the dataset, the first few rows were inspected using the `head(College)` function, which displayed a snapshot of the data. The structure of the dataset was examined using `str(College)`, revealing the types of variables (e.g., numeric, factor) and their names. Summary statistics for all variables were generated using `summary(College)`, providing key insights into the central tendencies, variability, and distribution of the data. Additionally, the presence of missing values was checked using `sum(is.na(College))`, which confirmed that the dataset is complete with no missing values. These preliminary steps laid the foundation for further exploratory data analysis (EDA) and modeling by ensuring the dataset was properly loaded, understood, and ready for analysis.

**Descriptive Analysis:**

The descriptive analysis of the ISLR College dataset provides valuable insights into the characteristics of various colleges and universities. By examining numeric variables such as the number of applications (Apps), out-of-state tuition (Outstate), and graduation rates (Grad.Rate), we observe the central tendencies (mean, median) and variability in data. For example, private colleges generally exhibit higher out-of-state tuition costs than public colleges, as evidenced by boxplots, which also highlight a significant range and potential outliers within private institutions. Histograms reveal the distribution of applications, showing a right-skewed pattern where a few colleges receive an exceptionally high number of applications. Scatterplots further uncover relationships, such as a possible positive correlation between room and board costs and graduation rates, suggesting the impact of institutional resources on student success. These findings lay the groundwork for further statistical and predictive analysis to uncover deeper trends and relationships in the data.

**Visulatisations:**

1. **Histogram of Applications:**

**A graph of applications with numbers and text

Description automatically generated with medium confidence**

**i)Observation:** The histogram shows the distribution of the number of applications received by colleges. It is highly skewed, with most colleges receiving fewer than 10,000 applications. There are very few colleges receiving applications close to or exceeding 50,000.

**ii)Impact:** This indicates that a small number of colleges are significantly more popular or have broader reach compared to others. These outliers could represent prestigious or highly sought-after institutions.

**iii)Justification**: The skewness of the data suggests variability in college application rates, potentially driven by factors such as reputation, location, or program offerings.

**iv)Insight**: This variability indicates that most colleges cater to a regional or niche audience, while a few institutions cater to a national or international audience.

**v)Conclusion:** Understanding this distribution can help identify patterns or trends in student preferences and highlight areas for smaller colleges to improve visibility.

1. **Boxplot of Out-of-State Tuition by Private Status:**

A diagram of a private status

Description automatically generated

**i)Observation**: The boxplot compares out-of-state tuition fees for private and public colleges. Private colleges have a much higher median tuition fee, and a wider range compared to public colleges. Additionally, private colleges exhibit some outliers with exceptionally high tuition fees.

**ii)Impact**: This substantial difference highlights the premium charged by private colleges for out-of-state students, reflecting their often-higher operational costs or perceived value.

**iii)Justification**: Private colleges rely more heavily on tuition for revenue, whereas public colleges receive state funding, allowing them to maintain lower tuition costs.

**Insight**: Students may consider cost a significant factor when deciding between public and private colleges, especially for out-of-state options. This could influence enrollment trends and affordability discussions.

**Conclusion**: This comparison reinforces the financial challenges for students opting for private education and underscores the importance of financial aid and scholarships

1. **Scatterplot: Room & Board vs Graduation Rate**

**A chart of a diagram

Description automatically generated with medium confidence**

**i)Observation**: The scatterplot demonstrates the relationship between room and board costs and graduation rates. While there is no clear linear relationship, colleges with higher room and board costs tend to have graduation rates clustered between 60% and 100%.

**ii)Impact:** Higher room and board costs might indicate institutions with better facilities or resources, which could contribute to higher student retention and graduation rates.

**iii)Justification:** Colleges with better infrastructure and support systems (often indicated by higher room and board costs) might provide a more conducive environment for academic success.

**iv)Insight:** Although room and board costs are not the sole determinant, they could be a proxy for other factors such as institutional quality, which positively influences graduation rates.

**v)Conclusion**: Further analysis might be needed to disentangle the relationship between financial investment and academic outcomes, considering other variables such as student demographics and college funding.

**Part 02**

**Splitting Dataset into Train-Test Set**

The dataset College is being split into training and testing datasets to prepare for model building and evaluation. Specifically:

**Setting the Seed:**

* + **set.seed(123)** ensures that the random sampling is reproducible. This means that every time the code runs, the same random indices will be generated, resulting in consistent splits of the dataset**.**

**Sampling:**

* + The sample\_size is calculated as 70% of the total number of rows in the College dataset (floor(0.7 \* nrow(College))), adhering to the commonly used 70-30 train-test split ratio.
  + train\_indices are randomly selected indices corresponding to 70% of the rows, which are used to create the train\_data.
  + The remaining 30% of the rows are assigned to test\_data

**Why is this step important?**

Splitting the dataset into training and testing subsets is a crucial step in supervised learning for the following reasons:

**Prevent Overfitting:**

* + If we train and test a model on the same dataset, it might perform well during training but fail to generalize on unseen data. This phenomenon is called overfitting.
  + By creating separate datasets for training and testing, we ensure that the model is evaluated on data it has not seen during training, which provides a realistic estimate of its performance.

**Model Evaluation:**

* + The training dataset is used to build and fit the model, allowing it to learn patterns from the data.
  + The testing dataset is reserved for validating the model's predictive capabilities, ensuring that the model generalizes well to new, unseen data.

**Reproducibility:**

* + By using set.seed(), we ensure that the random sampling process is consistent. This is essential for reproducibility in analytics and machine learning, where different data splits could lead to variations in model performance.

The overall structure of the data remains intact, but the split ensures a balance between having enough data for training and sufficient data for testing.

**Part 03**

**Why We Use the glm() Function and Multivariate Logistic Regression**

The glm() function in R (Generalized Linear Model) is a versatile tool for modeling relationships between a dependent variable and one or more independent variables. Specifically, we use the logistic regression model when the dependent variable is binary or categorical, as is the case with the Private variable in the ISLR College dataset (indicating whether a college is private or public).

**What is Multivariate Logistic Regression?**

A multivariate logistic regression model extends simple logistic regression by incorporating multiple predictor variables. It estimates the relationship between these predictors and a binary outcome by modeling the log-odds of the outcome. In our case:

* The dependent variable (Private) is binary: either "Yes" (Private) or "No" (Public).
* The independent variables (Outstate and Grad.Rate) are continuous predictors of the binary outcome.

The multivariate logistic regression equation is:



Where:

A black and white text

Description automatically generated

**Why Use glm() for Logistic Regression?**

The glm() function is specifically designed for Generalized Linear Models, allowing flexibility to handle:

1. **Non-normal data distributions**: Logistic regression assumes a binomial distribution for the dependent variable, which is well-suited for binary outcomes.
2. **Link functions**: Logistic regression uses the **logit link function** to transform probabilities into log-odds, enabling linear modeling in the transformed space.

**Key Observations from the Model Output**

**1. Coefficients Interpretation:**

* **Intercept (-4.828)**: This represents the log-odds of a college being private when both Outstate and Grad.Rate are 0. However, this interpretation is not practical in this context since tuition (Outstate) and graduation rate (Grad.Rate) are not zero. The intercept mainly centers the model.
* **Outstate (0.0006207)**: A positive and significant coefficient indicates that as the out-of-state tuition increases, the odds of a college being private increase. The low value is due to the scaling of the variable (tuition is in thousands of dollars).
* **Grad.Rate (0.005315)**: The coefficient is positive but **not statistically significant** (p-value = 0.514). This means there is no strong evidence that the graduation rate contributes significantly to predicting whether a college is private or public.

1. **Significance of Predictors:**

* **Outstate Tuition (Significant):** The very low p-value (<2e-16) indicates that Outstate is a highly significant predictor of whether a college is private.
* **Graduation Rate (Not Significant):** The p-value (0.514) suggests no significant relationship between Grad.Rate and the college's private/public status in this model.

1. **Model Fit Statistics:**

* **Null Deviance (645.75):** This represents the deviance of the model with no predictors, serving as a baseline.
* **Residual Deviance (394.78):** This indicates the deviance of the model with predictors (Outstate and Grad.Rate). A lower residual deviance compared to null deviance shows improvement in the model.
* **AIC (400.78):** The Akaike Information Criterion measures model quality. Lower values indicate better fit but also penalize complexity. This value is used to compare multiple models if needed.

The logistic regression model shows that **Outstate tuition is a strong predictor of whether a college is private**, while **Graduation Rate does not significantly contribute**. This highlights the economic factor (tuition fees) as a key determinant in distinguishing between private and public colleges.

**Part 04**

**Confusion Matrix On Train\_Set: Explanation, Use, and Analysis**

**What is a Confusion Matrix?**

A **confusion matrix** is a summary table used to evaluate the performance of a classification model. It compares the actual outcomes (true values) with the predicted outcomes from the model. The matrix provides counts of:

1. **True Positives (TP):** Cases correctly classified as positive.
2. **True Negatives (TN):** Cases correctly classified as negative.
3. **False Positives (FP):** Cases incorrectly classified as positive.
4. **False Negatives (FN):** Cases incorrectly classified as negative.

**Why Are We Using a Confusion Matrix in the Assignment?**

In our logistic regression analysis of the ISLR College dataset, the confusion matrix helps us assess how well the model predicts whether a college is private (Private = "Yes") or public (Private = "No") based on the predictors Outstate and Grad.Rate. Specifically, it:

1. **Validates Model Performance:** Helps determine whether the model generalizes well on the training set.
2. **Highlights Prediction Errors:** Identifies the number of false positives (misclassifying a public college as private) and false negatives (misclassifying a private college as public).
3. **Informs Model Refinement:** If performance metrics indicate room for improvement, we can adjust thresholds, add features, or try alternative models.

**Insights:**

A screenshot of a math test

Description automatically generated

**Purpose of Printing Metrics**

* **Accuracy:** Indicates the overall reliability of the model by summarizing all correct predictions.
* **Precision:** Helps in understanding how "precise" the model is when predicting a specific class (e.g., Private).
* **Recalling:** Measures of how well the model captures all instances of the positive class.
* **Specificity:** Ensures that the model doesn't misclassify too many negative cases as positive, which is critical for balanced evaluation.

**By printing these metrics, we can:**

1. Understand the strengths and weaknesses of the model's predictions.
2. Compare these metrics across different models or training/test datasets to select the most effective one.
3. Detect class imbalance issues if one metric (e.g., Precision or Recall) is significantly lower than others.

**Part 05**

**Confusion Matrix on Test\_Set**

**Purpose of Evaluating the Model on the Test Set**

After building and testing the logistic regression model on the training set, it is crucial to evaluate its performance on a separate test set. This process determines how well the model generalizes to unseen data and ensures that it is not overfitted to the training data.

**Impacts and Takeaways**

**Strong at Identifying Private Colleges:**

* The model excels at detecting private colleges (90.3% recall), making it suitable for applications where identifying private institutions is a priority.

**Struggles with Public Colleges:**

* The low specificity (57.6%) suggests the model confuses many public colleges as private.
* A higher false positive rate (25 misclassified public colleges) means the model may need additional predictors to distinguish between public and private colleges better.

**Potential Model Improvements:**

* Feature Engineering: Adding other factors like Room & Board costs or faculty-student ratio may help improve classification.
* Adjusting the Decision Threshold: Instead of using 0.5 as the probability cutoff, a higher threshold (e.g., 0.55 or 0.6) might reduce false positives.
* Alternative Models: Decision trees or ensemble methods like Random Forest could potentially improve classification.
* Overall, the model performs well, achieving an accuracy of 82.1%.
* It is highly effective at classifying private colleges but has difficulty with public colleges, misclassifying 25 as private.
* Further tuning, additional predictors, or different classification models could improve public-private distinction.

This confusion matrix provides **valuable insights into the strengths and weaknesses of the logistic regression model**, guiding further optimization strategies for better classification accuracy.

**Part 06**

**Plotting and Interpreting the ROC Curve**

**What is an ROC Curve?**

The ROC (Receiver Operating Characteristic) curve is a graphical representation used to evaluate the performance of a binary classifier system. It plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity) at various threshold settings.

**Interpretation of the ROC Curve:**

1. **Sensitivity (True Positive Rate)**: This is the proportion of actual positives correctly identified by the model. In your plot, sensitivity ranges from 0.0 to 1.0.
2. **Specificity (True Negative Rate)**: This is the proportion of actual negatives correctly identified by the model. The False Positive Rate (1 - Specificity) is plotted on the x-axis.

A graph of a curve

Description automatically generated

**Insights from the ROC Curve:**

* **Area Under the Curve (AUC)**: The AUC provides a single measure of the model's performance. An AUC of 1.0 indicates a perfect classifier, while an AUC of 0.5 suggests no discriminative ability (equivalent to random guessing).
* **Curve Shape**: The closer the curve follows the left-hand border and then the top border of the ROC space, the better the model's performance. A curve that is closer to the 45-degree diagonal indicates a less effective model.

**Justification:**

* **Model Performance**: If the ROC curve is in the upper left corner, it indicates high sensitivity and specificity, meaning the model is effective at distinguishing between the positive and negative classes.
* **Threshold Selection**: The ROC curve helps in selecting the optimal threshold for classification, balancing sensitivity and specificity based on the specific requirements of the application.

**Part 7**

**AUC**

1. **auc\_value <- auc(roc\_curve):**
   * auc(): This function calculates the Area Under the Curve (AUC) for a given ROC curve. The AUC is a single scalar value that summarizes the performance of a binary classifier across all possible classification thresholds.
   * roc\_curve: This is an object that contains the data for the ROC curve, typically generated by a function like roc() from the pROC package or similar.
   * auc\_value: This variable stores the computed AUC value.
2. **cat("AUC:", auc\_value, "\n"):**
   * cat(): This function is used to concatenate and print strings and values. It is often used for displaying output in R.
   * "AUC:": This is a string label that will be printed before the AUC value to indicate what the value represents.
   * auc\_value: This is the variable containing the AUC value calculated in the previous step.
   * "\n": This is a newline character, which ensures that the output is followed by a line break, making it easier to read if there are subsequent outputs.
3. **AUC: 0.8707022:**
   * This is the output produced by the cat() function. It shows that the AUC value for the ROC curve is approximately 0.8707.
   * AUC Interpretation: An AUC value of 0.8707 indicates that the model has good discriminatory power. Here's how to interpret theAUC value:
     + 0.5: No discrimination (equivalent to random guessing).
     + 0.7 - 0.8: Acceptable discrimination.
     + 0.8 - 0.9: Excellent discrimination.
     + > 0.9: Outstanding discrimination.

In summary, calculates the AUC for an ROC curve and prints it out, indicating that the model has excellent discriminatory ability with an AUC of approximately 0.8707. This value helps in assessing the overall performance of the classification model.

**Part 8**

**Comparison**

The \*\*ROC Curve\*\* (Receiver Operating Characteristic Curve) is a graphical representation used to evaluate the performance of a classification model. It plots the \*\*True Positive Rate (TPR)\*\* against the \*\*False Positive Rate (FPR)\*\* at various threshold settings. The area under the ROC curve (AUC) is a measure of the model's ability to distinguish between classes. A higher AUC indicates better model performance.

**Key Observations from the ROC Curves:**

1. \*\*Training Set ROC Curve\*\*:

- This curve represents the model's performance on the training dataset.

- If the curve is close to the top-left corner, it indicates high TPR and low FPR, meaning the model performs well on the training data.

- However, a curve that is too close to the top-left corner might indicate overfitting, where the model performs exceptionally well on the training data but poorly on unseen data.

2. \*\*Testing Set ROC Curve\*\*:

- This curve represents the model's performance on the testing dataset.

- A curve that is close to the top-left corner indicates good generalization to unseen data.

- If the testing curve is significantly lower than the training curve, it may indicate overfitting.

**Comparison of Performance:**

- \*\*AUC (Area Under the Curve) \*\*:

- The AUC for both the training and testing sets should be compared. A higher AUC indicates better performance.

- If the AUC for the training set is much higher than the testing set, the model may be overfitting.

- If the AUC values for both sets are similar and high, the model generalizes well.

- \*\*Shape of the Curves\*\*:

- A curve that rises steeply towards the top-left corner indicates a model with high sensitivity and specificity.

- If the testing curve follows a similar shape to the training curve, it suggests consistent performance across both datasets.

- \*\*Overfitting Check\*\*:

- If the training curve is significantly higher than the testing curve, the model may be overfitting the training data.

- If both curves are close to each other and near the top-left corner, the model is likely well-balanced and generalizes well.

**Practical Implications:**

**-** \*\*Model Selection\*\*: If the testing ROC curve has a high AUC and is close to the training curve, the model is likely a good choice for deployment.

- \*\*Model Improvement\*\*: If there is a significant gap between the training and testing curves, consider techniques like regularization, cross-validation, or feature selection to reduce overfitting.

**Conclusion:**

To compare the performance of the two ROC curves (training and testing), focus on the AUC values and the proximity of the curves to the top-left corner. A model with high and similar AUC values for both training and testing sets, with curves close to the top-left corner, is considered to perform well and generalize effectively. If the testing curve is significantly lower, it may indicate overfitting, and further model tuning is required.

**REFERENCES:**

[**https://www.geo.fu-berlin.de/en/v/soga-r/Basics-of-statistics/Logistic-Regression/Logistic-Regression-in-R---An-Example**](https://www.geo.fu-berlin.de/en/v/soga-r/Basics-of-statistics/Logistic-Regression/Logistic-Regression-in-R---An-Example) **(Logistic Regression using R by Freie Universitat Berlin)**

[**https://stats.oarc.ucla.edu/r/dae/logit-regression/**](https://stats.oarc.ucla.edu/r/dae/logit-regression/) **(Logit Regression, glm function by UCLA** Statistical Methods and Data Analytics)

**APPENDIX:**

**## MODULE :03**

**# STEP 1: Install and Load ISLR Package**

**install.packages("ISLR")**

**library(ISLR)**

**# STEP 2: Load the dataset**

**data("College")**

**# STEP 3: Verifying the dataset**

**head(College) #View frist few rows of dataset**

**str(College) #Check Structure of dataset**

**summary(College) #To get summary of dataset( Descriptive Statistic)**

**sum(is.na(College)) #Check for missing values**

**#STEP 4: EDA --> Descriptive Statistics and Plots**

**#Descriptive Statistics for numeric variables**

**summary(College)**

**#Histogram for numeric variables**

**hist(College$Apps, main = "Histogram of Applications", xlab = "Applications", col = "pink")**

**#Boxplot for Out-of-State Tuition by Private Status**

**boxplot(Outstate ~ Private, data = College, main = " Out-of-State Tuition by Private Status", xlab = "Private", ylab = "Out-of-State Tuition", col = "peachpuff")**

**#Scatterplot: Room & Board vs Graduation Rate**

**plot(College$Room.Board, College$Grad.Rate, main = "Room & Board vs Graduation Rate", xlab = "Room & Board", ylab = "Graduation Rate", col = "orange")**

**##STEP 5: Split the Data into Train and Test Sets**

**set.seed(123) # Set seeds for reproducibility**

**#Split the data**

**sample\_size <- floor(0.7 \* nrow(College))**

**train\_indices <- sample(seq\_len(nrow(College)), size = sample\_size)**

**train\_data <- College[train\_indices, ]**

**test\_data <- College[-train\_indices,]**

**##STEP 6: Fit a multivariate Logistic Regression Model**

**# NOTE: We will use glm() function to fit a logistic regression model on the training set. Outstate and Grad.Rate as predictors.**

**logistic\_model <- glm(Private ~ Outstate + Grad.Rate, data = train\_data, family = binomial)**

**summary(logistic\_model) #To summarize the model**

**##STEP 6 : Create a confusion Matrix for the Train set**

**#Predict probabilites on the train set**

**train\_data$predicted\_prob <- predict(logistic\_model, type = "response")**

**#Convert probabilities to binary predictions (0 = Public, 1 = Private)**

**train\_data$predicted\_class <- ifelse(train\_data$predicted\_prob > 0.5, "Yes", "No")**

**#Create confusion matrix**

**confusion\_matrix\_train <- table(Actual = train\_data$Private, Predicted = train\_data$predicted\_class)**

**print(confusion\_matrix\_train)**

**##STEP 7: Extract values from confusion matrix**

**TP <- confusion\_matrix\_train["Yes", "Yes"]**

**TN <- confusion\_matrix\_train["No","No"]**

**FP <-confusion\_matrix\_train["No","Yes"]**

**FN <- confusion\_matrix\_train["Yes","No"]**

**# Calculate metrices**

**accuracy <- (TP + TN) / (TP + TN + FP + FN)**

**precision <- TP / (TP + FP)**

**recall <- TP / (TP + FN)**

**specifically <- TN / (TN + FP)**

**#print metrices**

**cat("Accuracy:", accuracy, "\n")**

**cat("Precision:", precision, "\n")**

**cat("Recall:", recall, "\n")**

**cat("Specifically:", specifically, "\n")**

**##STEP 8: Evaluate the Model on the Test Set**

**#Predict probabilities on the test set**

**test\_data$predicted\_prob <- predict(logistic\_model, newdata = test\_data, type = "response")**

**# Convert probabilities to binary predictions**

**test\_data$predicted\_class <- ifelse(test\_data$predicted\_prob > 0.5, "Yes", "No")**

**# Create confusion matrix**

**Confusion\_matrix\_test <- table(Actual = test\_data$Private, Predicted = test\_data$predicted\_class)**

**print(Confusion\_matrix\_test)**

**##STEP 9: Plot and Interpret the ROC Curve**

**install.packages # install and load the pROC package**

**library(pROC)**

**#Create ROC curve**

**roc\_curve <- roc(test\_data$Private, test\_data$predicted\_prob,**

**levels = c("No", "Yes"), # Define levels**

**direction = "<") # Define direction**

**plot(roc\_curve, main = "ROC Curve", col = "brown")**

**##STEP 10: Calculate and Interpret AUC(Area under the curve)**

**auc\_value <- auc(roc\_curve)**

**cat("AUC:", auc\_value, "\n")**

**# Create ROC Curve for Training Set**

**roc\_curve\_train <- roc(train\_data$Private, train\_data$predicted\_prob,**

**levels = c("No", "Yes"),**

**direction = "<")**

**plot(roc\_curve\_train, main = "ROC Curve - Training/Testing Set", col = "blue")**

**print(roc\_curve\_train)**

**# Create ROC Curve for Test Set**

**roc\_curve\_test <- roc(test\_data$Private, test\_data$predicted\_prob,**

**levels = c("No", "Yes"),**

**direction = "<")**

**plot(roc\_curve\_test, add = TRUE, col = "red")**

**print(roc\_curve\_test)**

**# Add legend to differentiate training and test ROC curves**

**legend("bottomright", legend = c("Training Set", "Test Set"), col = c("blue", "red"), lwd = 2)**

**# Calculate and Compare AUC values**

**auc\_train <- auc(roc\_curve\_train)**

**auc\_test <- auc(roc\_curve\_test)**

**# Print AUC values**

**cat("AUC for Training Set:", auc\_train, "\n")**

**cat("AUC for Test Set:", auc\_test, "\n")**

**\*End of Module 3\***