**Module 3 Assignment: GLM and Logistic Regression**

**Sriram Kannepalli**

**Northeastern University: College of Professional Studies**

[**ALY6015.71591.202515**](https://northeastern.instructure.com/courses/196378) **: Intermediate Analytics**

**Professor: Roy Wada**

**November 16th, 2024**

**Introduction**

This report analyzes whether a university is private or public based on its institutional and financial characteristics using logistic regression. The analysis leverages the College dataset from the ISLR library, containing 18 variables such as tuition costs, enrollment numbers, and faculty qualifications.

The primary objective is to build and evaluate a logistic regression model to classify universities as either private or public. Key predictors, including Outstate (out-of-state tuition), F.Undergrad (full-time undergraduate enrollment), and PhD (percentage of faculty with PhDs), are explored for their impact on classification.

**Objectives:**

1. Develop a logistic regression model to classify university types.
2. Refine the model using feature selection techniques to enhance interpretability.
3. Evaluate model performance through metrics such as accuracy, precision, recall, specificity, and AUC.
4. Provide actionable insights supported by visualizations of results.

**Exploratory Data Analysis (EDA)**

**Dataset Overview**

The College dataset contains information about universities' financial and institutional characteristics:

* **Total Observations:** 777
* **Number of Variables:** 18

**Summary Statistics**

A summary of key predictors is shown below:

| **Statistic** | **Outstate Tuition** | **Full-Time Undergrads (F.Undergrad)** | **PhD Faculty (%)** |
| --- | --- | --- | --- |
| **Mean** | 10,441 | 3,700 | 79.70 |
| **Median** | 9,990 | 1,707 | 82.00 |
| **Min** | 2,340 | 139 | 24.00 |
| **Max** | 21,700 | 31,643 | 100.00 |
| **Q1** | 7,320 | 992 | 71.00 |
| **Q3** | 12,925 | 4,005 | 92.00 |
| **Std. Deviation** | 4,023 | 4,850 | 14.72 |

**Key Visualizations and Insights**

1. **Distribution of Outstate Tuition (Figure 1)**  
   **Insight:**
   * Private universities generally have higher tuition fees compared to public universities.
   * Public universities show a narrower range of tuition costs.
2. **Outstate Tuition vs. Undergraduate Enrollment (Figure 2)**  
   **Insight:**
   * Private universities typically cluster around lower enrollment numbers and higher tuition fees.
   * Public universities exhibit higher enrollment numbers with relatively lower tuition.
3. **Correlation Heatmap (Figure 3)**  
   **Insight:**
   * Strong positive correlation between Outstate and Private.
   * Negative correlation between F.Undergrad and Private, indicating public universities tend to have larger undergraduate populations.
   * Moderate positive correlation between Outstate and PhD.

**Logistic Regression Modeling**

**Train-Test Split**

The dataset was divided into:

* **Training Set:** 70% (545 observations)
* **Testing Set:** 30% (232 observations)

**Feature Selection**

Stepwise regression was performed using AIC to identify significant predictors. The final model includes:

* Outstate (Out-of-state tuition)
* F.Undergrad (Full-time undergraduate enrollment)
* PhD (Percentage of faculty with PhDs)

**Model Summary**

**Coefficients:**

| **Predictor** | **Estimate** | **Std. Error** | **z-value** | **p-value** | **Significance** |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | 0.2242 | 1.0974 | 0.2043 | 0.8381 |  |
| Outstate | 0.0009 | 0.0001 | 8.4644 | <0.001 | \*\*\* |
| F.Undergrad | -0.0006 | 0.0001 | -6.721 | <0.001 | \*\*\* |
| PhD | -0.0679 | 0.0183 | -3.715 | <0.001 | \*\*\* |

**Interpretation:**

* **Outstate Tuition:** Higher tuition increases the likelihood of being private.
* **F.Undergrad:** Larger undergraduate populations are associated with public universities.
* **PhD Faculty Percentage:** Higher faculty qualifications are negatively associated with private universities.

**Model Evaluation**

**Training Performance**

| **Metric** | **Value** |
| --- | --- |
| Accuracy | 94.31% |
| Precision | 95.97% |
| Recall | 96.21% |
| Specificity | 89.26% |

**Testing Performance**

| **Metric** | **Value** |
| --- | --- |
| Accuracy | 93.97% |
| Precision | 94.80% |
| Recall | 97.04% |
| Specificity | 85.71% |

**Confusion Matrix for Testing Set (Figure 4)**

| **Prediction** | **Public (0)** | **Private (1)** |
| --- | --- | --- |
| **Public (0)** | 54 | 5 |
| **Private (1)** | 9 | 164 |

**Insights:**

* The model achieves high performance metrics across both training and testing datasets.
* Slightly lower specificity in testing indicates some difficulty in distinguishing public universities.

**ROC Curve and AUC (Figure 5)**

* **AUC (Training):** 0.9799
* **AUC (Testing):** 0.9776

**Insights:**

* The high AUC scores confirm the model’s excellent discriminative power in identifying private and public universities.

**Conclusion**

**Key Findings**

1. **Outstate Tuition** strongly predicts university type, with higher tuition increasing the likelihood of being private.
2. **Undergraduate Enrollment** is inversely related to private classification.
3. **PhD Faculty Percentage** also negatively predicts private university classification.

**Model Performance**

* **Training Accuracy:** 94.31%
* **Testing Accuracy:** 93.97%
* The model demonstrates robust predictive power and reliability.

**Recommendations**

1. **Practical Applications:**
   * Use this model for funding allocation or policy differentiation between public and private universities.
2. **Continuous Monitoring:**
   * Update the model with new data to maintain accuracy over time.
3. **Stakeholder Engagement:**
   * Educate stakeholders about possible misclassifications, particularly false negatives.

**Appendix: Visuals**

* **Figure 1:** Distribution of Outstate Tuition.

A graph with a blue and red squares

Description automatically generated with medium confidence

* **Figure 2:** Scatter Plot of Outstate Tuition vs. Undergraduate Enrollment.

A graph of a graph with red and blue dots

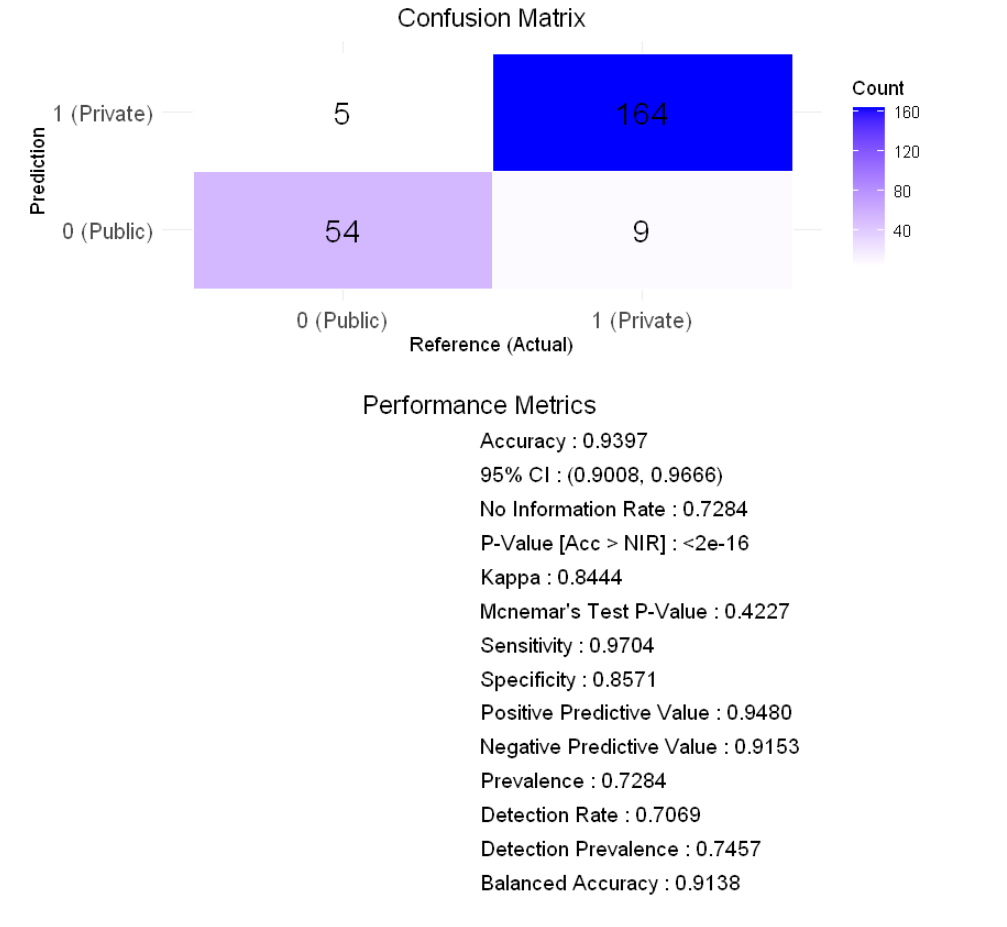
Description automatically generated

* **Figure 3:** Correlation Heatmap of Key Predictors.

A diagram of heatmap

Description automatically generated

* **Figure 4:** Confusion Matrix for Testing Set.



* **Figure 5:** ROC Curve for Training and Testing.

A graph of a training and testing

Description automatically generated

The R code used in this report is provided here for reproducibility:

|  |
| --- |
| # Load required libraries  library(ISLR)  library(ggplot2)  library(pROC)  library(caret)  library(kableExtra)  library(MASS)  library(corrplot)  # Load and prepare the dataset  data("College")  College$Private <- as.factor(ifelse(College$Private == "Yes", 1, 0))  # EDA: Descriptive Statistics  summary(College)  cor\_matrix <- cor(College[sapply(College, is.numeric)])  corrplot(cor\_matrix, method = "color", tl.cex = 0.8, addCoef.col = "black", number.cex = 0.7)  # Visualizations for EDA  ggplot(College, aes(x = as.factor(Private), y = Outstate, fill = as.factor(Private))) +  geom\_boxplot() +  labs(title = "Outstate Tuition by College Type", x = "College Type (0 = Public, 1 = Private)", y = "Outstate Tuition") +  theme\_minimal()  ggplot(College, aes(x = F.Undergrad, y = Outstate, color = as.factor(Private))) +  geom\_point(alpha = 0.6) +  labs(title = "Outstate Tuition vs. Full-Time Undergraduates", x = "Full-Time Undergraduates", y = "Outstate Tuition") +  theme\_minimal()  # Train-test split  set.seed(123)  train\_index <- createDataPartition(College$Private, p = 0.7, list = FALSE)  train\_data <- College[train\_index, ]  test\_data <- College[-train\_index, ]  # Stepwise Regression for Feature Selection  full\_model <- glm(Private ~ ., data = train\_data, family = binomial)  stepwise\_model <- stepAIC(full\_model, direction = "both")  # Final Model (based on stepwise results)  final\_model <- glm(Private ~ Outstate + F.Undergrad + PhD, data = train\_data, family = binomial)  # Display model coefficients  coefficients\_df <- as.data.frame(summary(final\_model)$coefficients)  colnames(coefficients\_df) <- c("Estimate", "Std. Error", "z value", "P value")  kable(coefficients\_df, format = "html", caption = "Logistic Regression Coefficients") %>%  kable\_styling(bootstrap\_options = c("striped", "hover", "condensed"), full\_width = FALSE)  # Predictions for training and test sets  train\_probs <- predict(final\_model, train\_data, type = "response")  train\_preds <- ifelse(train\_probs > 0.5, 1, 0)  test\_probs <- predict(final\_model, test\_data, type = "response")  test\_preds <- ifelse(test\_probs > 0.5, 1, 0)  # Confusion Matrices  conf\_matrix\_train <- confusionMatrix(as.factor(train\_preds), as.factor(train\_data$Private), positive = "1")  conf\_matrix\_test <- confusionMatrix(as.factor(test\_preds), as.factor(test\_data$Private), positive = "1")  # Combine performance metrics into a table  metrics <- data.frame(  Dataset = c("Training", "Training", "Training", "Training", "Testing", "Testing", "Testing", "Testing"),  Metric = c("Accuracy", "Precision", "Recall", "Specificity", "Accuracy", "Precision", "Recall", "Specificity"),  Value = c(  conf\_matrix\_train$overall["Accuracy"],  conf\_matrix\_train$byClass["Precision"],  conf\_matrix\_train$byClass["Recall"],  conf\_matrix\_train$byClass["Specificity"],  conf\_matrix\_test$overall["Accuracy"],  conf\_matrix\_test$byClass["Precision"],  conf\_matrix\_test$byClass["Recall"],  conf\_matrix\_test$byClass["Specificity"]  )  )  kable(metrics, format = "html", caption = "Training and Test Performance Metrics") %>%  kable\_styling(bootstrap\_options = c("striped", "hover", "condensed"), full\_width = FALSE)  # ROC Curve and AUC  roc\_train <- roc(train\_data$Private, train\_probs)  roc\_test <- roc(test\_data$Private, test\_probs)  auc\_train <- auc(roc\_train)  auc\_test <- auc(roc\_test)  # Display AUC values  auc\_table <- data.frame(Dataset = c("Training", "Testing"), AUC = c(auc\_train, auc\_test))  kable(auc\_table, format = "html", caption = "AUC for Training and Test Sets") %>%  kable\_styling(bootstrap\_options = c("striped", "hover", "condensed"), full\_width = FALSE)  # Plot ROC curves  ggplot() +  geom\_line(data = data.frame(fpr = 1 - roc\_train$specificities, tpr = roc\_train$sensitivities), aes(x = fpr, y = tpr), color = "blue", size = 1) +  geom\_line(data = data.frame(fpr = 1 - roc\_test$specificities, tpr = roc\_test$sensitivities), aes(x = fpr, y = tpr), color = "red", size = 1) +  geom\_abline(linetype = "dashed", color = "black") +  labs(title = "ROC Curve (Blue: Training, Red: Testing)", x = "False Positive Rate", y = "True Positive Rate") +  theme\_minimal() |

**References**

DataCamp. (n.d.). *Logistic Regression in R Tutorial*. Retrieved November 16, 2024, from <https://www.datacamp.com/tutorial/logistic-regression-R>

Statology. (2020, October 28). *How to Perform Logistic Regression in R (Step-by-Step)*. Retrieved November 16, 2024, from <https://www.statology.org/logistic-regression-in-r/>

R-bloggers. (2015, August 24). *Evaluating Logistic Regression Models*. Retrieved November 16, 2024, from <https://www.r-bloggers.com/2015/08/evaluating-logistic-regression-models/>

Machine Learning Plus. (n.d.). *Logistic Regression – A Complete Tutorial With Examples in R*. Retrieved November 15, 2024, from <https://www.machinelearningplus.com/machine-learning/logistic-regression-tutorial-examples-r/>

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Retrieved November 16, 2024, from <https://www.R-project.org/>