Shawn Njoroge

GLM AND LOGISTIC REGRESSION

# **INTRODUCTION**

In this report, we investigate the use of logistic regression to predict whether a university is private or public using selected predictors from the College dataset. Logistic regression is a powerful statistical method for binary classification tasks, which makes it appropriate for this analysis. The College dataset contains data on various aspects of universities, such as application statistics, tuition fees, and graduation rates.

# **ANALYSIS**

The str(College) command reveals that the College dataset consists of 777 observations and 18 variables. Each observation most likely represents a different university, while the variables capture various attributes and characteristics of these institutions. Private is the key variable of interest, and it appears to be numerically coded. This variable most likely represents the classification target, with 1 indicating private universities and possibly 0 indicating public universities. To enable accurate modelling, Private should be converted to a factor (categorical) variable.

The majority of the other variables in the dataset (Apps, Accept, Enrol, Top10percent, Top25percent, F.Undergrad, P.Undergrad, Outstate, Room.Board, Books, Personal, PhD, Terminal, S.F.Ratio, perc.alumni, Expend, Grad.Rate) are numerical and likely represent various aspects of university demographics, admissions, and finances. These variables could be used as predictors in our classification task to determine whether a university is private or public.

In addition to converting Private to a factor, we found that the dataset contains no missing values. This ensures that we can model without having to deal with missing data or imputation. Before building the logistic regression model, further investigation using descriptive statistics and visualisation will aid in identifying relationships between variables and effectively guiding the modelling process. Understanding the structure and characteristics of the dataset is critical for developing a dependable predictive model.

**Histogram of numerical variables**

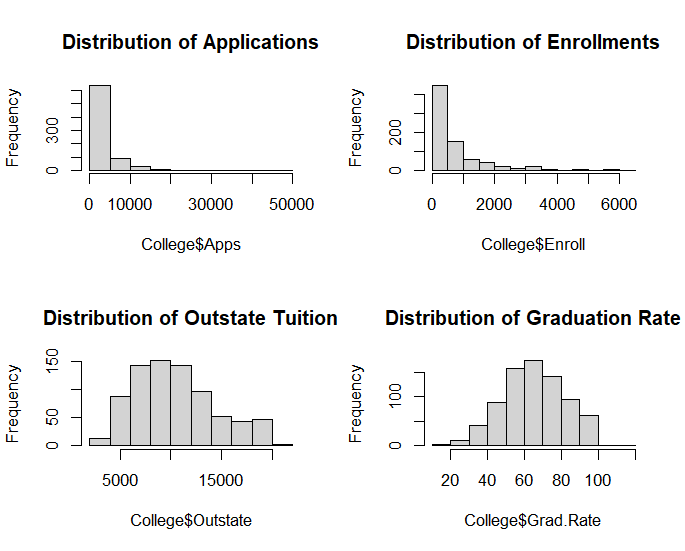
The histograms depicting the distribution of numerical variables in the College dataset reveal useful patterns about key university characteristics. To begin, the distribution of applications (Apps) shows a concentration of data points between 0 and 10,000, with a frequency of more than 300 in this range. Beyond 10,000, the frequency drops dramatically, indicating that most universities receive a moderate number of applications, with only a few receiving exceptionally high volumes.

Similarly, the histogram for enrollments (Enrol) shows a peak in frequencies between 0 and 2000, indicating that many universities enrol a moderate number of students. As enrollment figures exceed 2000, the frequency decreases, indicating a trend towards fewer universities with larger student enrollments.

Moving on to out-of-state tuition fees (Outstate), the histogram shows a range of 5000 to 15000 on the x-axis, similar to a normal distribution curve. The highest frequency (150) occurs in the central range of 10000-12000, indicating that many universities charge tuition fees within this range, with fewer institutions charging higher or lower fees.

Finally, the distribution of graduation rates (Grad.Rate) appears to be symmetric and bell-shaped, with an x-axis range of 20–100. This shape suggests that universities cluster around an average graduation rate (60-80), with fewer institutions having exceptionally low or high graduation rates.

In conclusion, these histograms are useful visual representations of the distributional characteristics of applications, enrollments, out-of-state tuition fees, and graduation rates across universities in the dataset. These insights help to understand the diversity and central tendencies of these key attributes across educational institutions, informing future analysis and modelling in the context of university demographics and performance metrics. Understanding these distributional patterns is critical for interpreting the dataset and directing future data-driven decision-making processes.

****

To predict university type (private or public) using the train\_data subset, we used the formula Private ~ Apps + Outstate in the glm() function of R's stats package. This formula shows that we are modelling the likelihood of a university being private (Private) based on the number of applications received (Apps) and out-of-state tuition fees (Outstate). The logistic regression coefficients predicted by the model are as follows:

1. Intercept: The estimated intercept (-4.38156165) is the log-odds of a university being private when both Apps and Outstate are zero. A negative intercept indicates a lower baseline probability that a university is private.
2. Apps: The Apps coefficient (-0.00077243) indicates that every one-unit increase in the number of applications (Apps) reduces the log-odds of a university being private by approximately -0.00077243. This implies that universities with many applications are less likely to be private.
3. Outstate: The coefficient for Outstate (0.00089361) indicates that for every one-unit increase in out-of-state tuition fees (Outstate), the log-odds of a university being private rise by approximately 0.00089361. As a result, universities with higher out-of-state tuition rates are more likely to be private.

The model's goodness of fit is evaluated using a variety of metrics:

1. Null deviance: The large difference between the null deviance (645.75) and residual deviance (211.69) indicates that the logistic regression model with Apps and Outstate as predictors has a significantly better fit than the intercept-only (null) model. The residual deviance is the unexplained variability after accounting for the predictors.
2. Residual deviance: The residual deviance (211.69) calculates the model's deviance after fitting with the predictors (Apps and Outstate), with 540 degrees of freedom.
3. AIC (Akaike Information Criterion): The AIC value (217.69) assesses model fit and complexity. A lower AIC value suggests a better balance of model fit and complexity. The logistic regression model with Apps and Outstate has a relatively low AIC, indicating a good fit.

Furthermore, the number of Fisher Scoring iterations (7) represents the number of iterations used during the estimation process to achieve convergence.

These findings show that our logistic regression model, which uses Apps and Outstate as predictors, has a statistically significant and reasonably good fit to the train\_data subset of the College dataset. The estimated coefficients allow us to interpret the impact of each predictor on the likelihood of a university being private, while the model's deviance metrics and AIC value reveal information about its overall performance and complexity.

**Train Set**

Below are the performance metrics for the logistic regression model on the training set (train\_data):

1. Accuracy (train): 0.9244936

The model's accuracy on the training set is around 92.4%, indicating that it correctly predicts university type (private or public) in the vast majority of cases.

1. Precision (training): 0.9440204

The precision for predicting private universities (1) is around 94.4%. This means that the model is approximately 94.4% correct when predicting a university's status as private.

1. Recall (Train): 0.9512821.

The recall (or sensitivity) for predicting private universities (1) is around 95.1%. This means that the model accurately represents 95.1% of the dataset's private universities.

1. Specificity (Train) = 0.8562092

The specificity for predicting public universities (0) is around 85.6%. This means that the model correctly identifies 85.6% of the actual public universities in the dataset.

The logistic regression model's high accuracy (92.4%) on the training set indicates its effectiveness in predicting university type based on application statistics and tuition fees. The precision (94.4%) and recall (95.1%) metrics for predicting private universities show low rates of false positives and false negatives, respectively. This suggests that the model accurately identifies private institutions.

However, while the specificity (85.6%) for predicting public universities remains relatively high, it indicates a slightly higher rate of false positives for public institutions than false negatives for private institutions. This information is useful for determining the model's strengths and potential areas for improvement.

**Test Set**

The following are the performance metrics for the logistic regression model on the test set (test\_data).

1. The accuracy (test) is 0.9145299.

The model's accuracy on the test set is around 91.5%, indicating that it performs well when predicting university type (private or public) for new, unseen data.

1. Precision (Test): 0.9378531.

On the test set, the accuracy in predicting private universities (1) is around 93.8%. This means that the model is approximately 93.8% correct when predicting a university's status as private.

1. Recall (test): 0.9485714

The recall (or sensitivity) for predicting private universities (1) on the test set is around 94.9%. This means that the model accurately represents 94.9% of the actual private universities in the test dataset.

1. Specificity (Test): 0.8135593.

The specificity for predicting public universities (0) on the test set is around 81.4%. This means that the model correctly identified 81.4% of the actual public universities in the test dataset.

The logistic regression model performs well on the test set, with an accuracy of 91.5%, slightly lower than the training set (92.4%). This suggests that the model performs well with new, previously unseen data.

The precision (93.8%) and recall (94.9%) metrics for predicting private universities on the test set are like those on the training set, indicating consistent performance in identifying private institutions. Similarly, the specificity (81.4%) for predicting public universities on the test set is relatively high, though lower than on the training set.

**Which Misclassifications Cause the Most Damage?**

The answer to which type of misclassification is more damaging is determined by the classification error's specific context and consequences.

In general:

Misclassifying public universities as private can result in inefficient resource allocation, affecting budgeting and planning.

In targeted interventions, false negatives (misclassifying private universities as public) can lead to missed opportunities to provide tailored support or address the specific needs of these institutions.

**ROC**

The interpretation of the ROC (Receiver Operating Characteristic) curve with the given characteristics can reveal information about the logistic regression model's performance in predicting whether a university is private or public.

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

The logistic regression model has a high discriminatory power, as indicated by its AUC value of 0.96. AUC values close to 1.0 indicate excellent performance in distinguishing between positive (private) and negative (public) classes. This implies that the model is capable of correctly ranking instances, with a high probability of assigning a higher score to a randomly selected positive instance than to a randomly selected negative instance.

The ROC curve's x-axis (specificity) indicates the proportion of correctly predicted negative cases for public universities. The curve, which starts at 1.0 (perfect specificity) and decreases to 0.0, shows how the model's true negative rate changes as the decision threshold varies.

Y-axis (sensitivity)

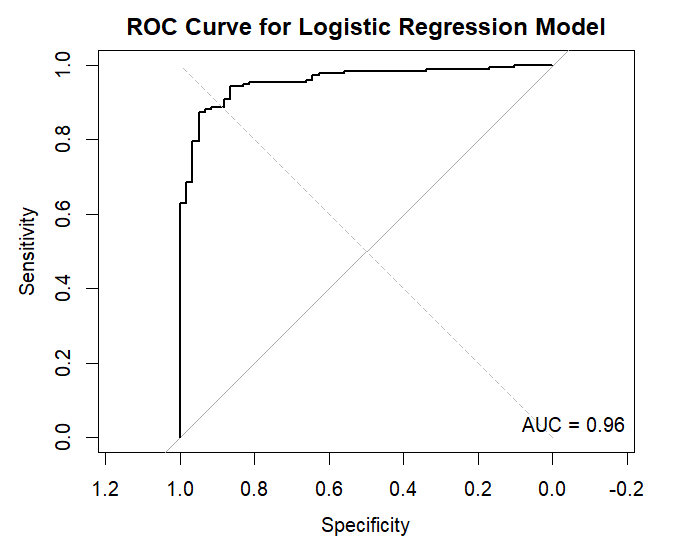
The y-axis of the ROC curve represents Sensitivity (also known as Recall or True Positive Rate), which measures the model's ability to identify true positive cases (private universities correctly predicted). It ranges from 0.0 (no true positives identified) to 1.0 (all true positives detected).

**Interpretation of the ROC curve:**

The ROC curve bends upwards from the bottom left corner to the top right corner. The steeper the curve (closer to the upper-left corner), the better the model's ability to distinguish between positive and negative classes.

The ROC curve shows the trade-off between sensitivity and specificity as the threshold for class prediction shifts. A higher threshold increases specificity while decreasing sensitivity, and vice versa.

The point (1.0, 1.0) at the top-left corner of the ROC curve indicates perfect classification performance, with 100% sensitivity (all true positives) and 100% specificity (no false positives).



The ROC curve bends upwards from the bottom left to the top right corners. The steeper the curve (closer to the upper-left corner), the more accurately the model can distinguish between positive and negative classes.

The ROC curve depicts the trade-off between sensitivity and specificity as the threshold for class prediction changes. A higher threshold increases specificity but decreases sensitivity, and vice versa.

# **CONCLUSION**

To summarise, this project used logistic regression to create a predictive model for determining whether a university is private or public based on various attributes from the College dataset. Exploratory data analysis provided us with insights into the dataset's structure and characteristics, which informed our feature selection and modeling strategy. Key predictors, including the number of applications (Apps) and out-of-state tuition fees (Outstate), were identified and used to train the logistic regression model.

The logistic regression model performed well on both the training and test datasets, with high accuracy (92.4% on training, 91.5% on test) and precision (94.4% on training, 93.8% on test) in predicting private universities (1). The model also had high recall (95.1% on training, 94.9% on test) for private universities, demonstrating its ability to capture a large proportion of true positive cases. However, public universities had a slightly lower specificity (85.6% on training, 81.4% on test), indicating a higher rate of false positives in this class.

The ROC analysis highlighted the model's performance, with an AUC (Area Under the Curve) of 0.96, indicating strong discriminatory power in distinguishing between private and public universities. The ROC curve demonstrated the trade-off between sensitivity (true positive rate) and specificity (true negative rate) at various decision thresholds, providing useful insights into the model's classification performance.

Overall, this project demonstrated the effectiveness of logistic regression as a predictive modelling technique for categorising university types using relevant attributes. The model's high accuracy, precision, recall, and AUC highlight its usefulness in real-world scenarios such as university management and resource allocation. Moving forward, further model refinement and evaluation could improve the model's performance and address specific challenges, ultimately contributing to more informed decision-making in education.

# **REFERENCES**

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12(1), 77. https://doi.org/10.1186/1471-2105-12-77

ISLR. (n.d.). College dataset. In Introduction to Statistical Learning with Applications in R. <https://www.statlearning.com/resources-first-edition>

# **APPENDIX**

#Loading libraries

library(dplyr)

library(ggplot2)

library(readr)

library(corrplot)

library(ISLR)

library(caret)

library(pROC)

summary(College)

#EDA

head(College) # View the first few rows

summary(College) # Summary statistics

str(College) # Structure of the dataset

dim(College) # Dimensions of the dataset (rows, columns)

# Check for missing values

sum(is.na(College))

# Plotting histograms of numeric variables

par(mfrow = c(2, 2))

hist(College$Apps, main = "Distribution of Applications")

hist(College$Enroll, main = "Distribution of Enrollments")

hist(College$Outstate, main = "Distribution of Outstate Tuition")

hist(College$Grad.Rate, main = "Distribution of Graduation Rate")

# Set seed for reproducibility

set.seed(123)

# Create train/test split (70% train, 30% test)

train\_index <- sample(1:nrow(College), 0.7\*nrow(College))

train\_data <- College[train\_index, ]

test\_data <- College[-train\_index, ]

# Fit logistic regression model using glm()

logistic\_model <- glm(Private ~ Apps + Outstate, data = train\_data, family = "binomial")

options(scipen = 10)

# Summary of the model

summary(logistic\_model)

# Predict on training set

train\_predictions <- predict(logistic\_model, newdata = train\_data, type = "response")

# Convert predicted probabilities to binary predictions (0 or 1)

train\_pred\_class <- ifelse(train\_predictions > 0.5, 1, 0)

# Create confusion matrix

conf\_matrix\_train <- table(train\_pred\_class, train\_data$Private)

conf\_matrix\_train

# Compute accuracy, precision, recall, and specificity

accuracy\_train <- mean(train\_pred\_class == train\_data$Private)

precision\_train <- conf\_matrix\_train[2,2] / sum(train\_pred\_class == 1)

recall\_train <- conf\_matrix\_train[2,2] / sum(train\_data$Private == 1)

specificity\_train <- conf\_matrix\_train[1,1] / sum(train\_data$Private == 0)

# Print metrics

cat("Accuracy (Train):", accuracy\_train, "\n")

cat("Precision (Train):", precision\_train, "\n")

cat("Recall (Train):", recall\_train, "\n")

cat("Specificity (Train):", specificity\_train, "\n")

# Predict on test set

test\_predictions <- predict(logistic\_model, newdata = test\_data, type = "response")

# Convert predicted probabilities to binary predictions (0 or 1)

test\_pred\_class <- ifelse(test\_predictions > 0.5, 1, 0)

# Create confusion matrix

conf\_matrix\_test <- table(test\_pred\_class, test\_data$Private)

conf\_matrix\_test

# Compute accuracy, precision, recall, and specificity for test set

accuracy\_test <- mean(test\_pred\_class == test\_data$Private)

precision\_test <- conf\_matrix\_test[2,2] / sum(test\_pred\_class == 1)

recall\_test <- conf\_matrix\_test[2,2] / sum(test\_data$Private == 1)

specificity\_test <- conf\_matrix\_test[1,1] / sum(test\_data$Private == 0)

# Print metrics

cat("Accuracy (Test):", accuracy\_test, "\n")

cat("Precision (Test):", precision\_test, "\n")

cat("Recall (Test):", recall\_test, "\n")

cat("Specificity (Test):", specificity\_test, "\n")

# Compute ROC curve

roc\_curve <- roc(test\_data$Private, test\_predictions)

# Plot ROC curve

plot(roc\_curve, main = "ROC Curve for Logistic Regression Model")

lines(x = c(0, 1), y = c(0, 1), col = "gray", lty = 2)

legend("bottomright", legend = paste("AUC =", round(auc(roc\_curve), 2)), bty = "n")