**ALY6015 80472 Intermediate Analytics SEC 04 Spring 2023 CPS**

**Module 3 Assignment — GLM and Logistic Regression REPORT**

Logo

Description automatically generated

**NORTHEASTERN UNIVERSITY**

**College of Professional Studies, Boston, MA, 02215.**

**Submitted by**

Zihan Ma

ma.zihan1@northeastern.edu

**Instructor**

 Prof. Valeriy Shevchenko

**Date**

06/12/2023

**GLM and Logistic Regression**

**Assignment Summary:**

The assignment requires using R to fit a Logistic Regression model on the College dataset to predict whether a university is private or public.

The process involves importing the dataset, conducting exploratory data analysis, splitting the data into training and test sets, fitting a logistic regression model to the training set, creating confusion matrices for both sets and interpreting their results. The assignment also entails calculating metrics such as Accuracy, Precision, Recall, Specificity, ROC curve, and AUC, followed by an in-depth interpretation of these metrics.

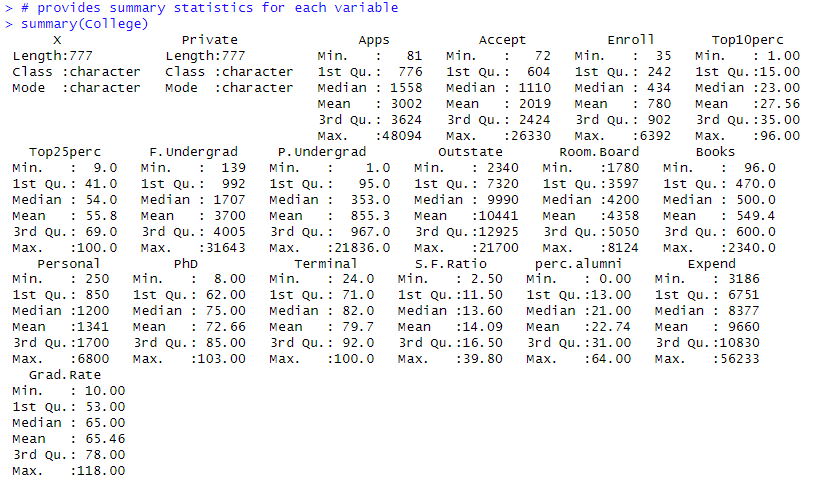
**Process And Report:**

**Process 1: Exploratory Data Analysis**

Import the dataset and perform exploratory data analysis to gain insights into the data. Use descriptive statistics and plots to describe the dataset, including measures of central tendency, dispersion, and any interesting patterns or trends.

A screenshot of a computer

Description automatically generated with medium confidence

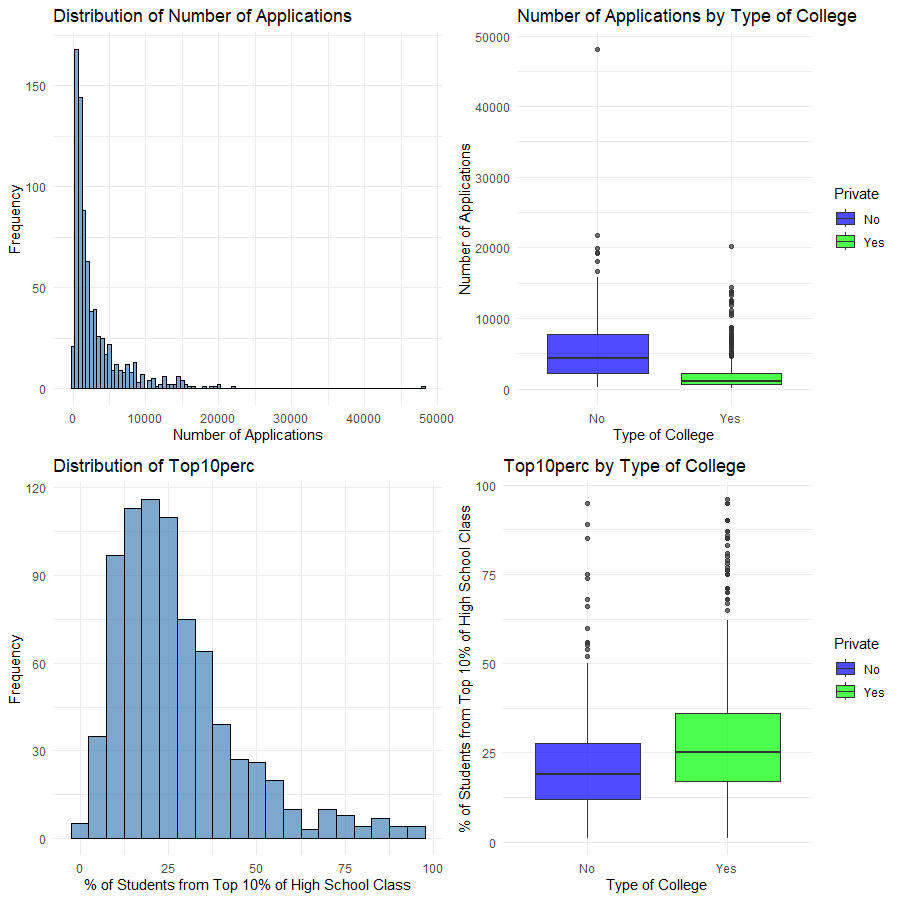


From the summary of the data set, we know that:

The dataset contains diverse information on 777 U.S. colleges, including aspects of enrollment, academic achievement, faculty characteristics, and costs. There is significant variability in the number of applications received and accepted by colleges, the academic standing of enrolled students, and the associated costs. The dataset paints a detailed picture of the higher education landscape in the U.S., with insights into institutional differences in the student body, faculty qualifications, and financial requirements.

Assume the 'Private' variable (indicating whether the college is private or not) is the target variable. We should then select those features that show different behaviors when the college is private or not.

In this case, I decided to choose variables `$ Apps` and ` $ Top10perc ` to show this difference.



The four graphs collectively provide insights into the college application landscape and the academic competitiveness of U.S. colleges.

The first histogram shows the distribution of the number of applications across all colleges, indicating a right-skewed distribution, with most colleges receiving fewer than 10,000 applications.

The second boxplot contrasts the number of applications between private and public colleges. Public colleges tend to receive a higher number of applications than Private colleges but also exhibit greater variability.

The third histogram visualizes the percentage of enrolled students who come from the top 10% of their high school classes. The distribution is somewhat uniform between 10% to 40%, then gradually decreases.

Finally, the fourth boxplot compares the Top10perc between private and public colleges. Private colleges generally have a higher percentage of students from the top 10% of their high school classes, suggesting that they are more selective or academically rigorous.

**Process 2: Train-Test Split**

Split the dataset into a training set and a test set. Refer to the "Feature\_Selection\_R.pdf" document for information on how to perform the split. Ensure that the splitting preserves the distribution of the target variable.

A picture containing text, font, screenshot, line

Description automatically generated

After the split, we have 2 data sets. In this case, we're allocating 80% of the data to the training set.

**Process 3: Logistic Regression Model Training**

Use the glm() function from the 'stats' package in R to fit a logistic regression model to the training set. Select at least two predictors to include in the model. Provide a brief rationale for the predictor selection.

A screenshot of a computer

Description automatically generated with medium confidence

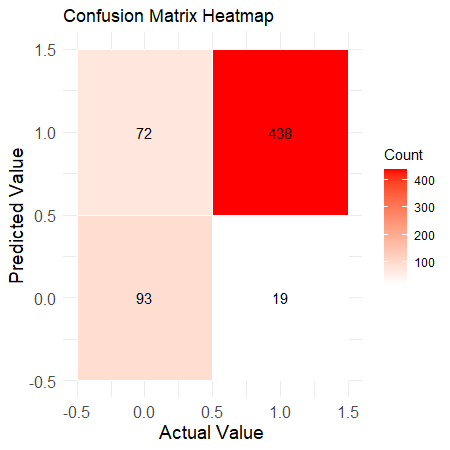
From the summary of the model, we can observe that:

1. Both Apps and Top10perc have large z values, indicating they are significant predictors.
2. The p-values for both Apps and Top10perc are less than 0.001, indicating strong evidence that the relationship observed in the data did not occur by chance alone.
3. The model has a residual deviance of 470.67, which is considerably smaller than the null deviance of 719.65 (a model with no predictors), suggesting that adding the predictors Apps and Top10perc to the model improved the fit.

In summary, both Apps and Top10perc appear to be significant predictors of whether a college is private or not, according to this logistic regression model.

**Process 4: Confusion Matrix and Train Set Results**

Create a confusion matrix based on the predictions of the logistic regression model on the train set. Report the results and interpret the confusion matrix.

  
A screenshot of a computer code

Description automatically generated with low confidence

* True Positives (TP): The model correctly predicted 438 times that the university is private when it is indeed private.
* False Positives (FP): The model incorrectly predicted 72 times that the university is private when it is actually public.
* True Negatives (TN): The model correctly predicted 93 times that the university is public when it is indeed public.
* False Negatives (FN): The model incorrectly predicted 19 times that the university is public when it is actually private.
* Accuracy: About 85.4% of the total predictions made by the model are correct.
* Precision: The precision is about 85.9%. High precision indicates a low false positive rate.
* Recall: The recall is about 95.8%. High recall indicates the model is good at catching positive cases (here, identifying private universities).
* Specificity: The specificity is about 56.4%. The model seems to be less efficient at correctly identifying public universities compared to private ones, as indicated by the lower specificity compared to recall.

This model performs well in identifying private universities but less so for public ones.

From the graph "Number of Applications by Type of College" on p1, we can see that private universities have a stricter distribution than public ones, which could cause this problem.

**Process 5: Impact of Misclassifications**

Discuss which misclassifications are more damaging for the analysis, False Positives (Type I Error) or False Negatives (Type II Error)? Provide a brief explanation.

For this dataset, we are trying to predict whether a college is private or public based on the number of applications and the percentage of accepted students in the top 10% of their high school class.

False Positives, in this context, means that the model predicted a college to be private when it is actually public. This could result in students erroneously expecting higher tuition fees, smaller class sizes, or different admission standards than what the public college might actually offer.

False Negatives, on the other hand, mean that the model predicted a college to be public when it is, in fact, private. This could lead to students being unprepared for potentially higher academic standards or different educational environments than they would find at a public university. It could also lead to unexpected financial burdens if students budget for a public university's tuition, but the university is actually private with higher tuition.

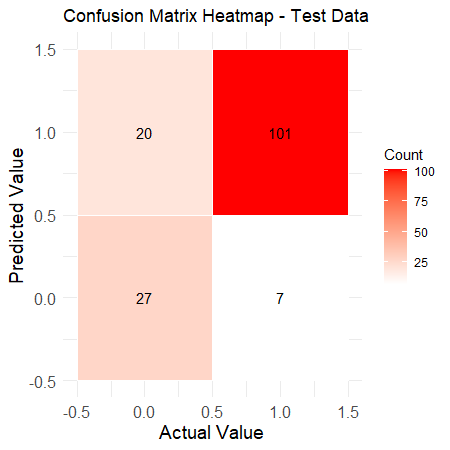
Which error is more damaging depends on the stakeholders' priorities:

* If the primary concern is ensuring that students aren't scared off by the prospect of high tuition fees at private colleges, then False Positives might be considered more damaging.
* If the main focus is to prepare students for the potentially higher academic standards at private colleges, then False Negatives could be considered more harmful.

In this situation, the cost of both types of errors can be high, but which one is more damaging will depend on the particular context and the goals of the stakeholders involved in making or using these predictions.

**Process 6: Confusion Matrix and Test Set Results**

Create a confusion matrix based on the predictions of the logistic regression model on the test set. Report the results and analyze the performance of the model on unseen data.

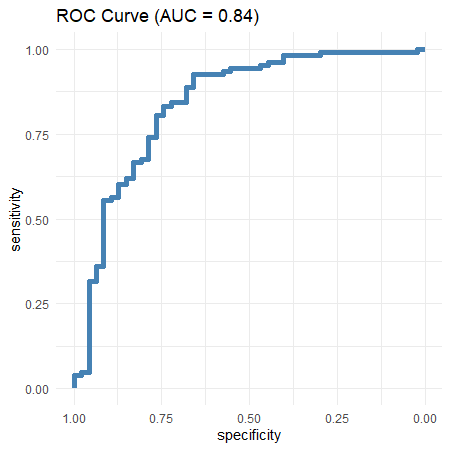


* True Positives (TP): The model identified more colleges as private correctly on the training data than on the test data (438 vs. 101).
* False Positives (FP): The model incorrectly identified more colleges as private on the training data than on the test data (72 vs. 20). This suggests that the model may be overfitting to the training data and mistaking some public colleges as private.
* True Negatives (TN): The model correctly identified more colleges as public on the training data than on the test data (93 vs. 27).
* False Negatives (FN): The model incorrectly identified more colleges as public on the training data than on the test data (19 vs. 7).
* Accuracy: The model's accuracy is higher on the training data than on the test data (85.37% vs 82.58%). This suggests that the model may be overfitting to the training data.
* Precision: The model's precision is higher on the training data than on the test data (85.88% vs 83.47%). This suggests that the model is more accurate in its predictions of private colleges on the training data.
* Recall: The model's recall is higher on the training data than on the test data (95.84% vs 93.52%). This suggests that the model is better at identifying private colleges on the training data.
* Specificity: The model's specificity is lower on the training data than on the test data (56.36% vs 57.45%). This suggests that the model is slightly better at identifying public colleges on the test data, but this difference is quite small.

The results show the model seems to perform slightly better on the training data, which is expected as the model was trained on this data. However, the differences in performance are not very large, suggesting that the model generalizes reasonably well to unseen data. Overfitting might be occurring given that performance on the training data is somewhat better, but the degree of overfitting appears to be limited.

**Process 7: Receiver Operating Characteristic (ROC) Curve And Area Under the Curve (AUC)**

Plot and interpret the ROC curve based on the data standpoint. Discuss the trade-off between true positive rate and false positive rate and its implications for the model's classification performance.  
Calculate and interpret the AUC score from the data standpoint. Discuss the discriminatory power of the logistic regression model based on the AUC value.



Based on the ROC curve graph, the model demonstrates relatively strong performance. The Area Under the Curve (AUC) score is 0.84, which means that the model correctly classifies positive and negative examples 84% of the time when choosing a random threshold. Generally, an AUC score of 0.5 denotes a model with random predictive power, while a score of 1.0 denotes a perfect model. Thus, a score of 0.84 is quite good, suggesting that the model's predictive power is notably above chance.

Looking at the data points on the ROC curve, we see that the True Positive Rate (TPR, or sensitivity) remains very high (at one or close to 1) even as the False Positive Rate (FPR, or 1-specificity) increases up to about 0.02. This means that, for a wide range of decision thresholds, the model is capable of identifying the most positive examples (i.e., correctly predicting 'Private' universities) without generating many false positives.

As the FPR increases beyond 0.02, the TPR starts to decline gradually. This indicates that the model begins to generate more false positives (incorrectly predicting 'Private' when the university is actually 'Public') for higher sensitivity. There is a trade-off between the TPR and FPR: to increase the TPR, the model needs to predict more examples as 'Private', which in turn leads to more false positives and, thus, a higher FPR.

Towards the end of the curve (FPR close to 1), the TPR drops more sharply, which implies the model is misclassifying more 'Public' universities as 'Private'. This end of the ROC curve is associated with low decision thresholds, where the model tends to predict most examples as 'Private'.

In terms of implications for the model's classification performance, the ROC curve suggests that the model can achieve a high sensitivity (up to about 0.99) with a relatively low FPR (around 0.30). Beyond this point, increasing sensitivity leads to larger increases in the FPR, which might not be desirable depending on the specific context and costs associated with false positives versus false negatives.

Thus, this model would be more useful in scenarios where high sensitivity is required, and some increase in the FPR can be tolerated. However, for applications where a low FPR is crucial, a lower decision threshold would need to be chosen, which would reduce the model's sensitivity.

**Summary:**

In this study, we analyzed U.S. college's data using logistic regression, focusing on predicting if a college is public or private. Our exploratory analysis revealed public colleges receive more applications, but private colleges have more top-achieving students.

The model demonstrated promising results with an accuracy of 85.4% on training data but was more efficient at classifying private colleges. Misclassifications, both false positives, and negatives, highlighted the importance of accuracy in this context.

Despite the trade-off between sensitivity and specificity reflected in the ROC curve, the model showcased good discriminatory power (AUC = 0.84), suggesting its utility in decision-making and scope for future improvements.