Northeastern University
College *of* Professional Studies

Logistic Regression

Overview and Rationale

In order to consolidate your theoretical knowledge into technique and skills with practical and applicational value, you will use the *glm()* function in R to fit a Logistic Regression model to perform classification.

Course Outcomes

This assignment is directly linked to the following key learning outcomes from the course syllabus:

CLO10: Extend general linear regression to allow for a response variable to have an error distribution model other than a normal distribution.

CLO11: Extend general linear regression by allowing the linear model to be related to the response variable via a link function.

CLO12: Model the relationship between predictor variables and a categorical response variable.

Assignment Summary

Use the College dataset from the ISLR library to build a logistic regression model to predict whether a university is private or public.

1. Import the dataset and perform Exploratory Data Analysis by using descriptive statistics and plots to describe the dataset.
2. Split the data into a train and test set – refer to the Feature_Selection_R.pdf document for information on how to split a dataset.
3. Use the *glm()* function in the 'stats' package to fit a logistic regression model to the training set using at least two predictors.
4. Create a confusion matrix and report the results of your model for the train set. Interpret and discuss the confusion matrix. Which misclassifications are more damaging for the analysis, False Positives or False Negatives?
5. Report and interpret metrics for Accuracy, Precision, Recall, and Specificity.
6. Create a confusion matrix and report the results of your model for the test set.
7. Plot and interpret the ROC curve.
8. Calculate and interpret the AUC.

Report

Refer to the attached rubric for more details on the report. The report should contain a well written cover/title page, introduction, body, conclusion, and references. It must follow APA format and have at least 1000 words (excluding title page and references page. All R code used for your report should be included in an appendix at the end of the report.

Graphs, figures, charts, and tables are very useful visual effects to communicate your results and impress your readers. However, such items should not be included in the report unless they are well described and interpreted. Please use subtitles to make your assignment more reader friendly

as well.

Format & Guidelines

The report should follow the following format:

(i) Title page

(ii) Introduction

(iii) Analysis

(iv) Conclusion/Interpretations

(v) References

Northeastern University
College *of* Professional Studies

Assignment Rubric

Category	Meets Standards	Approaching Standards	Below Standards
Introduction	Introduction provides a brief and intelligible overview of the goals and methods of the assignment.	Introduction provides an overview of the goals and methods of the assignment, but is ambiguous or not concise.	Does not introduce project goals, project questions or methods.
Analysis	Provides all R code and the outputs. Includes interpretation of the output, graphs, figures, charts, and tables and the significance of the results in the analysis.	Provides R codes and outputs, but the R code does not match the outputs or is missing some code or outputs. Includes limited interpretations, charts, and tables and the significance of the results in the analysis.	Does not provide R code or its outputs or minimal R code is provided. Includes few interpretations, charts, or tables. Does not identify the significance of the results in the analysis.
Data Visualizations	Data visualizations are appropriate for the level and type of analysis. Graphs, figures and tables communicate insights and significance to the reader.	Data visualization are useful for the level and type of analysis, but graphs, figures and tables do not clearly communicate significance of the results to the reader.	Data visualization are used minimally or not at all. If graphs, figures and tables are used, it is unclear what they are intended to communicate or why.
Interpretation & Conclusions	The conclusion summarizes and makes sense of the results, making good points that reflect clear understanding of the assignment material.	The conclusion summarizes and makes sense of the results, making good points that reflect a basic understanding of the assignment material.	The conclusion does not summarize or attempt to make sense of the results. Conclusions do not reflect an understanding or reflect a misunderstanding of the material.
Report: Writing Mechanics, Title Page, & References	There are no noticeable errors in grammar, spelling, and punctuation; and completely correct usage of title page, citations, and references. The report contains approximately of 1000 words.	There are very few errors in grammar, spelling, and punctuation; and completely correct usage of title page, citations, and references. The report contains approximately 1000 words.	There are more than five errors in grammar, spelling, and punctuation; or the usage of title page, citations, and references are incomplete; or the report contains far less than 1000 words.