Northeastern University
College of Professional Studies

ALY 6015 Intermediate Analytics Module 3 Assignment

Class ALY6015 – Intermediate Analytics

Module 3 Assignment: GLM and 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:

- Use "R" effectively to process, analyze and depict data
- Develop more advanced models to interpret data
- Use advanced generalized linear methods to answer strategic and operational questions
- Prepare complex dataset for analysis
- Use multivariable and logistic regression method to improve predictive outcomes

Submission Requirements

1. Complete paperwork in MS Word format (.docx) must include:

- Title Page
    - Your name (as registered in Canvas)
    - Assignment name
    - Class number, name and CRN Number
    - Your contact information (NEU email)
- Assignment summary section.(Explain assignment summary, plans, goals, dataset).
- Each step of the research with supporting screenshots, charts, results generated by R code.
- Explain each screen shot from the data standpoint.
- Each output generated by R code must be present and explained in the paperwork.
- Each output, chart, table, screenshot shown in the paperwork must have corresponding R code that generates it.
- Final conclusions section. (Explain if goals were achieved as expected or not, summary of you findings about analyzed data).

- References (optional)

2. Complete R code file meet the following criteria:

- Submitted in R script (.r file format). Only .r file format will be accepted.
- Each line of code must be commented. (Explain why do you execute this line of code, not what the command does).
- Code must be runnable on any computer. Any errors in executing R code will results in significant points deduction. (*Follow the guidelines provided in the class and user R code examples provided in Canvas*)

3. Submit dataset(s) used in the research.

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.

Instructions

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.
5. Which misclassifications are more damaging for the analysis, False Positives or False Negatives?
6. Report and interpret metrics for Accuracy, Precision, Recall, and Specificity.
7. Create a confusion matrix and report the results of your model for the test set.
8. Plot and interpret the ROC curve from the data standpoint.
9. Calculate and interpret the AUC from the data standpoint.

GLM and Logistic Regression Assignment Rubric

Criteria	Ratings				Pts
This criterion is linked to a Learning Outcome Introduction	**15 to >13.95 pts** **Above Standards** Clearly and briefly introduces the goals of the project, the question that needs to be answered and the methods used in the analysis. The goals, questions and methods outlined are consistent with one another.	**13.95 to >10.5 pts** **Meets Standards** Introduction provides a brief and intelligible overview of the goals and methods of the assignment.	**10.5 to >9.0 pts** **Approaching Standards** Introduction provides an overview of the goals and methods of the assignment, but is ambiguous or not concise.	**9 to >0 pts** **Below Standards** Does not introduce project goals, project questions or methods.	15 pts

GLM and Logistic Regression Assignment Rubric

Criteria	Ratings				Pts
This criterion is linked to a Learning Outcome Analysis	**25 to >23.25 pts** **Above Standards** Incorporates R code and the outputs. Provides detailed analysis of the output focusing on significance results. Uses visualizations to make major points.	**23.25 to >17.5 pts** **Meets Standards** 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.	**17.5 to >15.0 pts** **Approaching Standards** 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.	**15 to >0 pts** **Below Standards** 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.	25 pts

GLM and Logistic Regression Assignment Rubric

Criteria	Ratings				Pts
This criterion is linked to a Learning Outcome Data Visualizations	**25 to >23.25 pts** **Above Standards** Data visualizations are appropriate for the level and type of analysis. Uses graphs, figures, charts, and tables to increase visual effects of the main points being made based on the results.	**23.25 to >17.5 pts** **Meets Standards** Data visualizations are appropriate for the level and type of analysis. Graphs, figures and tables communicate insights and significance to the reader.	**17.5 to >15.0 pts** **Approaching Standards** Data visualization are useful for the level and type of analysis, but graphs, figures and tables do not clearly communicate the significance of the results to the reader.	**15 to >0 pts** **Below Standards** 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.	25 pts

GLM and Logistic Regression Assignment Rubric

Criteria	Ratings				Pts
This criterion is linked to a Learning Outcome Interpretation & Conclusions	**25 to >23.25 pts** **Above Standards** Wraps up the findings in a conclusion that provides an answer to the question(s) posed in the introduction. Makes specific recommendations based on the data presented.	**23.25 to >17.5 pts** **Meets Standards** The conclusion summarizes and makes sense of the results, making good points that reflect clear understanding of the assignment material.	**17.5 to >15.0 pts** **Approaching Standards** The conclusion summarizes and makes sense of the results, making good points that reflect a basic understanding of the assignment material.	**15 to >0 pts** **Below Standards** 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.	25 pts

GLM and Logistic Regression Assignment Rubric

Criteria	Ratings				Pts
This criterion is linked to a Learning Outcome Report: Writing Mechanics, Title Page, & References	**10 to >9.3 pts** **Above Standards** There are no noticeable errors in grammar, spelling, and punctuation; and completely correct usage of title page, citations, and references. The report contains approximately 1,000 words.	**9.3 to >7.0 pts** **Meets Standards** There are no noticeable errors in grammar, spelling, and punctuation; and completely correct usage of title page, citations, and references. The report contains approximately 1,000 words.	**7 to >6.0 pts** **Approaching Standards** There are very few errors in grammar, spelling, and punctuation; and completely correct usage of title page, citations, and references. The report contains approximately 1,000 words.	**6 to >0 pts** **Below Standards** 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 1,000 words.	10 pts
Total Points: 100					