Northeastern University
College *of* Professional Studies

ALY 6015 Intermediate Analytics Module 4 Assignment

Class ALY6015 – Intermediate Analytics

Module 4 Assignment: Regularization

Overview and Rationale

In order to consolidate your theoretical knowledge into technique and skills with practical and applicational value, you will use the glmnet() package in R to implement Ridge and LASSO functions to build linear and logistic models through Ridge and LASSO regression over values of the regularization parameter lambda.

Course Outcomes

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

- Conduct regularization method for models to describe relationships among variables and make useful predictions.

Submission Requirements

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

- Title Page

    o   Your name (as registered in Canvas)

    o   Assignment name

    o   Class number, name and CRN Number

    o   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 regularization models by using Ridge and Lasso (least absolute shrinkage and selection operator). Predict Grad.Rate for all models.

Instructions

1. Split the data into a train and test sets.

Ridge Regression

2. Use the *cv.glmnet* function to estimate the lambda.min and lambda.1se values. Compare and discuss the values.

3. Plot the results from the cv.glmnet function provide an interpretation. What does this plot tell us?

4. Fit a Ridge regression model against the training set and report on the coefficients. Is there anything interesting?

5. Determine the performance of the fit model against the training set by calculating the root mean square error (RMSE). sqrt(mean((actual - predicted)^2))

6. Determine the performance of the fit model against the test set by calculating the root mean square error (RMSE). Is your model overfit?

LASSO

7. Use the *cv.glmnet* function to estimate the lambda.min and lambda.1se values. Compare and discuss the values.

8. Plot the results from the *cv.glmnet* function provide an interpretation. What does this plot tell us?

9. Fit a LASSO regression model against the training set and report on the coefficients. Do any coefficients reduce to zero? If so, which ones?

10. Determine the performance of the fit model against the training set by calculating the root mean square error (RMSE). sqrt(mean((actual - predicted)^2))

11. Determine the performance of the fit model against the test set by calculating the root mean square error (RMSE). Is your model overfit?

Comparison

12. Which model performed better and why? Is that what you expected?

13. Perform stepwise selection and then fit a model.

14. Did this model perform better than or as well as Ridge regression or LASSO?

15. Which method do you prefer and why?

Regularization Assignment Rubric

Criteria	Ratings				Pts
This criterion is linked to a Learning Outcome Introduction	**15 to >13.95 pts** **Above Standard** 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 Standard** Introduction provides a brief and intelligible overview of the goals and methods of the assignment.	**10.5 to >9.0 pts** **Approaching Standard** Introduction provides an overview of the goals and methods of the assignment, but is ambiguous or not concise.	**9 to >0 pts** **Below Standard** Does not introduce project goals, project questions or methods.	15 pts

Regularization Assignment Rubric

Criteria	Ratings				Pts
This criterion is linked to a Learning Outcome Analysis	**25 to >23.25 pts** **Above Standard** 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 Standard** 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 Standard** 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 Standard** 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

ALY 6015 Intermediate Analytics Module 4 Assignment

Regularization Assignment Rubric

Criteria	Ratings				Pts
This criterion is linked to a Learning Outcome Data Visualization	**25 to >23.25 pts** **Above Standard** 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 Standard** 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 Standard** 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 Standard** 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

Regularization Assignment Rubric

Criteria	Ratings				Pts
This criterion is linked to a Learning Outcome Interpretation & Conclusion	**10 to >9.3 pts** **Above Standard** 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 Standard** 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 Standard** 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 Standard** 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

Regularization Assignment Rubric

Criteria	Ratings				Pts
This criterion is linked to a Learning Outcome Report: Writing Mechanics, Title Page, & References	**25 to >23.25 pts** **Above Standard** 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 Standard** 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 Standard** 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 Standard** 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

Total Points: 100