ASDS 5303 Project 5

Due Friday Nov. 15th, 2024

Answer the following questions. The goal of this lab is to build a logistic regression to classify ‘’ in the ‘Titanic’ dataset.

Note: You need to submit two files for this assignment, a pdf report and the original code file in R. Miss the original code file will have 50% reduction of your score. One day late submission will have 50% reduction. More than one day late submission will have 0 credit.

Download the dataset Titanic.

1. Make appropriate plots between ‘Survival’ and the rest of the columns. By looking at your plots, which predictor do you think would have the strongest power to predict ‘Survival’? Hint: it’s necessary to check the contingency table between the Survival and any categorical variables. You need to figure out the code for making a contingency table yourself.

2. Build a logistic regression, call it fit1, to predict ‘Survival’ with the predictor you choose from the question2. Write down your logistic regression with estimated coefficients and interpret the model estimated coefficients.

3. Split data to get new training (70%) and test (30%) datasets. You need to store these two datasets and use it for Q5. Train your logistic regression Log_fit the training data set and evaluate the accuracy in the test data set. Generate a confusion matrix. Which level (‘Yes’ or ‘No’) has higher prediction accuracy? You need to interpret all results.

4. Repeat 3 with all possible predictors. How much improvement do your new model make compared to the previous model (Log_fit)? What else do you observe in your new model (Write at least one)?

5. Build a LDA model with the same predictors, same training dataset and testing dataset you use in Q3. Train your LDA (LDA_fit) in the training data set and evaluate the accuracy in the test data set. Generate a

confusion matrix. Which level ('Yes' or 'No') has higher prediction accuracy? You need to interpret all results.

6. Compare the result you get in Q3 and Q5. What is your observation? Which model is better? Justify your answer.

Bonus

Run a QDA model for the same training and testing datasets in Q3. Please compare the results with the results from Q3 and Q5. What is your conclusion? Justify it.