PROG8430 – Data Analysis, Modeling and Algorithms

Assignment 4

Classification

DUE BEFORE 10PM AUGUST 11, 2021

1. Submission Guidelines

All assignments must be submitted via the econestoga course website before the due date in to the assignment folder.
You may make multiple submissions, but only the most current submission will be graded.

SUBMISSIONS

In the Assignment 4 Folder submit:

1. Your R Code
2. Your report in Word

All variables in your code must abide by the naming convention [variable_name]_[intials]. For example, a variable I create for State would be State_DM.

You may only use the 'R' packages discussed and demonstrated in class:

1. **pROC**
2. **MASS**
3. **klaR**

THIS IS AN INDIVIDUAL ASSIGNMENT. UNAUTHORIZED COLLABORATION IS AN ACADEMIC OFFENSE. Please see the Conestoga College Academic Integrity Policy for details.

2. Grading

This assignment will be marked out of 30 and is worth 12.5% of your total grade in the course.
Late assignments will receive a 20% penalty.
Assignments received after start of class the day after due will receive a mark of 0.

3. Data

Each student will be using one dataset:

Tumor_21S.Rdata

4. Background

The dataset contains medical information used in the pre-screening diagnosis of tumors.

Your task is to use logistic regression to determine the factors that predict probability of a tumor diagnosis.

You will then be using two other classification techniques and will compare all three of them.

Your work should follow the format of the sample report used previously.

5. Assignment Tasks

Nbr	Description	Marks
1	Preliminary Data Preparation 1. As demonstrated in class and conducted in previous assignments (MLR), make sure that the data is free from outliers or unnecessary data.	1
2	Exploratory Analysis 1. Correlations: Create numeric correlations (as demonstrated) and comment on what you see. Are there co-linear variables? 2. Identify the two most significant predictors of tumors and provide statistical evidence (in addition to the correlation coefficients) that suggest they are associated with tumors (Think of the contingency tables we did in class).	1 2
3	Model Development As demonstrated in class, create three models. 1. A forward selection model. 2. Two additional models using variables that you select based on the above output (recall lecture slides on variable selection). We will refer to these models as "User Model 1" and "User Model 2". Make sure you mention why you chose the variables you did. For each model, interpret and comment on the main measures we discussed in class: 1. Fisher's Scoring Iteration (does it converge?) 2. AIC 3. Deviance 4. Residual symmetry 5. z-values 6. Variable Co-Efficients	1 2
4	Model Evaluation 1. For User Model 1 and User Model 2, create and evaluate the confusion matrix. Set the default predictive level to 50% for "success". Based on the confusion matrix, calculate and comment on: a. Accuracy b. Specificity c. Sensitivity	2

	d. Precision 2. For each of the two models, create the ROC curve and calculate the AUC. Comment on how you interpret each of them.	2
5	Final Recommendation 1. Based on your preceding analysis, recommend which model should be selected and explain why.	1
PART B		
1	Logistic Regression – Stepwise 1. As above, use the forward option in the glm function to fit the model 2. Summarize the results in a Confusion Matrix . 3. As demonstrated in class, calculate the time (in seconds) it took to fit the model and include this in your summary.	 1 1 1
2	Naïve-Bayes Classification 1. As demonstrated in class, transform the variables as necessary for N-B classification. 2. Use all the variables in the dataset to fit a Naïve-Bayesian classification model. 3. Summarize the results in a Confusion Matrix. 4. As demonstrated in class, calculate the time (in seconds) it took to fit the model and include this in your summary.	 1 1 1 1
3	Linear Discriminant Analysis 1. As demonstrated in class, transform the variables as necessary for LDA classification. 2. Use all the variables in the dataset to fit an LDA classification model. 3. Summarize the results in a Confusion Matrix. 4. As demonstrated in class, calculate the time (in seconds) it took to fit the model and include this in your summary.	 1 1 1 1
4	Compare All Three Classifiers For all questions below please provide evidence. 1. Which classifier is most accurate? (provide evidence) 2. Which classifier is most suitable when processing speed is most important? 3. Which classifier minimizes Type 1 errors? 4. Which classifier minimizes Type 2 errors? 5. Which classifier is best overall? 6. How do these classifiers compare to the best model you built in Part 1?	 4
5	Professionalism and Clarity	3

APPENDIX ONE: DATA DICTIONARY

Name	Description
Out	Tumor is present=1, Is not present=0
Age	Older =1, Younger=0
Sex	Male=1, Female=0
Bone	Bone Density Test: Good=0, Bad=1
Marrow	Bone Marrow: Good=0, Bad=1
Lung	Spot on Lung: Yes=1, No=0
Pleura	Pleura: Yes=1, No=0
Liver	Spot on Liver: Yes=1, No=0
Brain	Brain Scan: Yes=1, No=0
Skin	Lesions: Yes=1, No=0
Neck	Stiff Neck? Yes=1, No=0
Supra	Supraclavicular: Yes=0, No=1
Axil	Axillar: Yes=0, No=1
Media	Mediastinum: Yes=1, No=0