**Assignment 5**

**Author:** Vidhya Venugopal, 8902970

**Course:** PROG8430

**Background**

To determine the factors that predict probability of a tumor diagnosis.

**Data Source**

The dataset contains medical information used in the pre-screening diagnosis of tumors which includes out (tumor present or not present), age, gender, bone density test, bone marrow, spot on Lung, pleura, Liver, brain scan, skin Lesions, stiff neck, Supraclavicular, Axillar& Mediastinum. All columns as a binary variable.

1. Preliminary Data transformation
   1. In the given tumor dataset “Tumor\_21F.Rdata”, all columns are binary variables and there are no factor variables to convert. Considering the summary of data, all columns have the same Min and Max values. Therefore, there is no outliers in this dataset. We are good to proceed with the dataset for further analysis:

A screenshot of a computer screen

Description automatically generated

1. Exploratory Analysis

2.1 Correlation: Numeric correlation matrix has been created, yes, there are co-linear variables which is ‘Out’ and ‘Brain’ & ‘Out’ and ‘Marrow’. These variables have high positive correlation.

A screenshot of a computer

Description automatically generated

2.2 Statistical evidence of one of the significant predictors of tumor is ‘Brain’ which is highly associated with tumor, here p-value<=0.05. Therefore, these variables are not independent:

A computer screen shot of a computer program

Description automatically generated

A computer screen shot of a number

Description automatically generated

Statistical evidence of next column which is the next significant predictor of tumor is ‘Marrow’ which is highly associated with tumor, here p-value<=0.05. Therefore, these variables are not independent:

A computer screen shot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

1. Model Development

3.1 User Model 1 = Full Model – In this model, all the binary variables are utilized

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

User Model 2 = Manual Adjusted Model – In this model, those variables which are non-significant predictors like Bone, Pleura, Liver and Axil because p-value >= 0.05 based on the above model has been dropped manually

A screenshot of a computer

Description automatically generated

A white background with black text

Description automatically generated

3.2 A Criteria Selection Model – In this model, those variables which we used for full Model has been utilized:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

In this above 3 models,

* 1. AIC - Akaike Information Criterion:

User Model – Full Model 1 = 291.8

User Model - Manual Adjusted Mode 2 = 286.7.

A Criteria Selection Model – 285.8

The lower the better, therefore, A Criteria Selection Model is better.

* 1. DEVIANCE - A measure of “goodness of fit” – lower is better.

User Model – Full Model 1:

Null Deviance: 415.89  
Residual Deviance: 263.80

User Model - Manual Adjusted Mode 2

Null Deviance: 415.89  
Residual Deviance: 268.72

A Criteria Selection Model

Null Deviance: 415.89  
Residual Deviance: 265.76

Null Deviance: Indicates how well the dependant variable is predicted using just the intercept.  
Residual Deviance: The deviance with all measures included. Notice the significant reduction from 415 to 263.

Notice also that the fit is just as strong with all variables in Full User Model 1.

* 1. Residual Symmetry

Residual Symmetry section is not populated in the Summary in R Studio

* 1. z-values

Based on the z-values and its corresponding p-value. Skin column is not significant predictor as p-value >=0.05 and all the remaining columns like Age, Sex, Marrow, Lung, Brain, Neck, Supra & Media are all significant predictors as p-value <=0.05.

* 1. Variable Co-Efficients

Coefficients match the correlation matrix and our exploratory analysis above. Brain and Marrow are highly significant except few variables like Bone, Pleura, Liver and Axil. So, these variables are dropped in User Model 2 – Manual Adjusted Model.

1. Model Evaluation
2. User Model 1: Confusion Matrices & Predictions

A computer screen shot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

* + 1. a. Accuracy - 0.78
    2. b. Specificity – 0.8
    3. c. Sensitivity – 0.79
    4. d. Precision – 0.76

User Model 2: Confusion Matrices & Predictions

A computer screen shot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated A screenshot of a computer

Description automatically generated

* + 1. a. Accuracy - 0.79
    2. b. Specificity – 0.8
    3. c. Sensitivity – 0.80
    4. d. Precision – 0.78

Based on the confusion matrices by setting the default predictive level to 50% for “success” and calculating all the predictions, considering the accuracy, specificity, sensitivity, and precision – User Model 2 Manual Adjustment Model is better than User Full Model 1 because in User Model 2, the accuracy, specificity, sensitivity, and precision is higher than User Model 1.

4.2 User Model 1: ROC Curve and Calculate the AUC

A computer code with red and blue text

Description automatically generated

A graph of a curve

Description automatically generated with medium confidence

Red Diagonal Line represents a random selection. AUC = 0.88, Theoretical Maximum = 1

User Model 2: ROC Curve and Calculate the AUC

A screen shot of a computer

Description automatically generated

A graph with red lines

Description automatically generated

Red Diagonal Line represents a random selection. AUC = 0.876, Theoretical Maximum = 1

1. Final Recommendation

5.1 Based on our preceding analysis, though the AUC of the User model 1 & 2 are more are less same when we are rounding to 2 digits. Therefore, by considering the accuracy, specificity, sensitivity, and precision User Model 2 is will be selected to do the tumor predictions.

PART B

1. Logistic Regression – Stepwise
   1. A Criteria Selection Model – Stepwise Model Result matches exactly with the above 3.1 results

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A white background with black text

Description automatically generated

* 1. Summarizing the results in a confusion matrix, the test for independence of all factors is significant as p-value <= 0.05.

A screenshot of a computer program

Description automatically generated

A black text on a white background

Description automatically generated

1.3 Calculated the time (in seconds) it took to fit the model.

A screen shot of a computer code

Description automatically generated

1. Naïve-Bayes Classification

2.1 Transformed the variables as necessary for N-B classification.

A screenshot of a computer code

Description automatically generated

2.2 Used all the variables in the dataset to fit a Naïve-Bayesian classification model.

A computer screen shot of a computer code

Description automatically generated

2.3 Summarize the results in a Confusion Matrix

A close-up of a computer screen

Description automatically generated

2.4 Calculated the time (in seconds) it took to fit the model

A computer screen shot of a program

Description automatically generated

1. Linear Discriminant Analysis
   1. Transform the variables as necessary for LDA classification.

A screen shot of a computer code

Description automatically generated

3.2 Use all the variables in the dataset to fit an LDA classification model.

A close-up of a computer code

Description automatically generated

3.3 Summarize the results in a Confusion Matrix

A computer screen shot of a computer error

Description automatically generated

3.4 Calculated the time (in seconds) it took to fit the model.

A close-up of a code

Description automatically generated

1. Compare All Three Classifiers

A Criteria Selection Model

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

Naïve-Bayes Classification

A screenshot of a computer program

Description automatically generated

A screenshot of a computer code

Description automatically generated

Linear Discriminant Analysis

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

* 1. Which classifier is most accurate? - When comparing all the three accuracy Linear Discriminant Analysis model has high accuracy.
  2. Which classifier is most suitable when processing speed is most important? – When comparing all the 3 models processing speed, Linear Discriminant Analysis model has taken very less time in processing. Therefore, Linear Discriminant Analysis model is most suitable.
  3. Which classifier minimizes Type 1 error? – When comparing all the 3 False Positive values, Stepwise Model has the lowest value FP = 29
  4. Which classifier minimizes Type 2 error? – When comparing all the 3 False Negative values, Naïve Bayes Model has the lowest value FN = 34
  5. Which classifier is best overall? – When comparing all the parameters of the 3 models. Linear Discriminant Analysis seems to best overall.

Stepwise Model Naïve Bayes Linear Discriminant Analysis

A screenshot of a computer code

Description automatically generatedA screenshot of a computer code

Description automatically generatedA screenshot of a computer code

Description automatically generated

* 1. How do these classifiers compare to the best model you built in Part 1? - When comparing the Linear Discriminant Analysis and User Model 2 from Part 1 based on the accuracy, precision, sensitivity, and specificity the User Model 2 is the best model for tumor predictions.

User Model 2 from PART 1

A screenshot of a computer code

Description automatically generated A screenshot of a computer

Description automatically generated

Linear Discriminant Analysis

A screenshot of a computer code

Description automatically generated