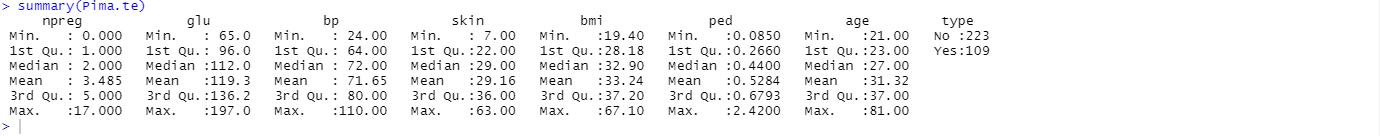
Pima.te

The object is to “Find the best predictions for diabetes on the Pima.te data set”.

The Pima.te in in library MASS, it contains 8 rows, and 322 records.

Text, letter

Description automatically generatedWe can see that in the data and finds out that the TYPE and other attributes did not show any arrangement so we can use 257 records as the training set and the other 65 as the testing set (80/20). And checking the baseline accuracy.

A picture containing company name

Description automatically generatedFor easy interpretation and calculation of the odds ratios, we are introducing the dummy variables to the table.

Now, we have the ideal data which is about to analyze.

Graphical user interface, text

Description automatically generated with medium confidence1.GLM

Using type as the objective, which is a dummy variable using 0 and 1 indicate the individual is a diabetes or not. Firstly, we use all the variables to do the logistic linear regression and can see in the summary that some attributes--- bp, skin, age are not related to the regression of having diabetes or not.

Chart, box and whisker chart

Description automatically generatedTable

Description automatically generatedAnd we could do boxplot to show the relationship between type and attributes.

Chart, box and whisker chart

Description automatically generated

So, instead of doing step by step trying, we can use stepwise function to look for the ideal regression model.

Text

Description automatically generatedBy forward

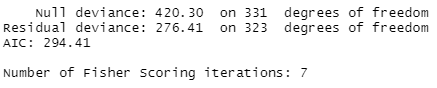
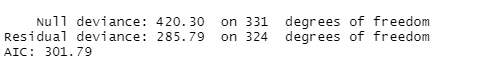
Showing plot for forward\_1

Chart

Description automatically generated

and we can see clearly in the Residual vs Fitted, there is a pattern, which indicates that the relationship between the TYPE and variables is not linear.

The result of glm1 and forwards is showed as followed, the dropping AIC and Residual deviance and degree of freedom showed the improvement of the model



Graphical user interface, text, application

Description automatically generatedA screenshot of a computer

Description automatically generated with medium confidenceText

Description automatically generatedNow we are working on the test subset using it to predict the value of the type and compared it with the actual value.

After we done the glm(), the output is great, and after we done the prediction we got 80.59% accuracy when we set the Threshold = 0.46, and we can see in the ROC curve , it is a moderate model.

Chart, histogram

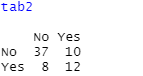
Description automatically generated2. LDA

After, we set the group we could just using lda() function to do the analyze

Graphical user interface, text, application, email

Description automatically generated

Tab2 accuracy is 0.7313



Tab3 accuracy is 0.7313

Tab4 accuracy is 0.7761

3.QDA

Text

Description automatically generated with medium confidenceKind of the same process of LDA, so in my point of view, it would be saving some time and effort to use the formula in LDA. We got

A picture containing calendar

Description automatically generated

The table is correctly predicted 52 out of 67 which is 0.7761194 accurate.

4.K-NN

Graphical user interface, text, application, email

Description automatically generatedFor this one we should, regroup the testing set and training set, along with 2 vectors of the actual type of both. The set should be also the same size of the training set above. Using cbind() to make a set of the variables we are going to be used.

The tables and the accuracies are

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated