Solution - Logistic Regression

File: creditset.csv

1. Build Logistic regression model for predicting default10yr on the basis of age, income & loan.
   a. Write logit function with coefficients
   b. Which predictors are good and why?
   c. What is classification accuracy for group '0'?
   d. What is classification accuracy for group '1'?
   e. What is the area under ROC curve?
   f. What is Nagelkerke R square?

Answer a)

$$\begin{aligned}\log it(default10yr) &= logit(odds)\\ &= 9.8166458 + \left((-0.0002367) * Income\right)\\ &+ \left((-0.3465840) \times age\right)\\ &+ \left((0.0017076) \times loan\right)\end{aligned}$$

$$\mathbf{Prob(default10yr)} = \frac{e^{\left(9.8166458+\left((-0.0002367)\times Income\right)+\left((-0.3465840)\times age\right)+\left((0.0017076)\times loan\right)\right)}}{1+\ e^{\left(9.8166458+\left((-0.0002367)\times Income\right)+\left((-0.3465840)\times age\right)+\left((0.0017076)\times loan\right)\right)}}$$

Answer b)

We can say that age, income and loan all are good predictors of our response variable. We have built a model named "model" that gives us the least value of the Residual deviance & AIC value of all the models that can be built using the predictors available.

We need to predict default10yr from the data given. As default10yr is a categorical 1 being defaulter 0 being non -defaulter. We need to use logistic regression for predicting our dependent variable "default10yr.In above we have divided our data set into 2 parts training and testing. We are going to build our model /train our model with training dataset and will test the same model on our testing dataset. training & testing are the subsets of creditset .For building our Logistic Regression model clientid and LTI are not the predictors of default10yr we will not use these variables while building our model . Our predictor variables are age, income and loan .We need to decide and optimize our model in such a way that the accuracy is more and the threshold cut-off probability is such that the model is not dangerous like we predict defaulters as non-defaulters as such an info will be dangerous for bank meaning False-Negative should be minimum but not much accuracy of the model is lost.

Answer c)

Classification Accuracy of group '0' is also known as specificity

The specificity of our training set is **0.960284** meaning **96.02%** the model was able to accurately predict non-defaulter as non-defaulter for our training dataset

Answer d)

Classification Accuracy of group '1' is also known as sensitivity

The sensitivity of our training set is **0.8507463** meaning **85.07%** the model was able to accurately predict defaulter as defaulter for our training dataset

Answer e)

Area under Roc curve is calculated by using the Verification package in R.
The command used for ROC area is

- **roc.area(training$default10yr,res)**

Area under the ROC curve comes out to be **0.9826288** for our training dataset.

The area of the box in which ROC curve is made is considered to have unit area. We need to build a model in order to have maximum area under the ROC Curve.

Answer f)

Nagelkerke R square is obtained in R using the **fmsb** package.

The command for calculating Nagelkerke R square in R is

- Nagelkeker2(model)

The value of Nagelkerke R square comes out for our model equal to **0.7992079** for our training dataset for which the model is trained.

High value of this R square means Logistic Regression model is a goodness of fit of model for predicting the response categorical variable