**Logistic Regression**

Logistic Regression is used for Classification problems that is when your DV is a discrete variable.

**Applications of Logistic Regression**

* Risk Scenario: Whether a customer will default or not (1/0)
* Campaign Response: Whether a customer will respond to an offer or not
* Medical Field: Whether a person is diagnosed with a disease or not

The DV has to be a discrete variable whereas the IDV’s can be discrete or continuous

**Logistic Regression Equation**

Unlike the linear Regression Equation where we build a model for “Y” which is the target variable, in logistic Regression we build a model for log(odds ratio)

The derivation is as follows:

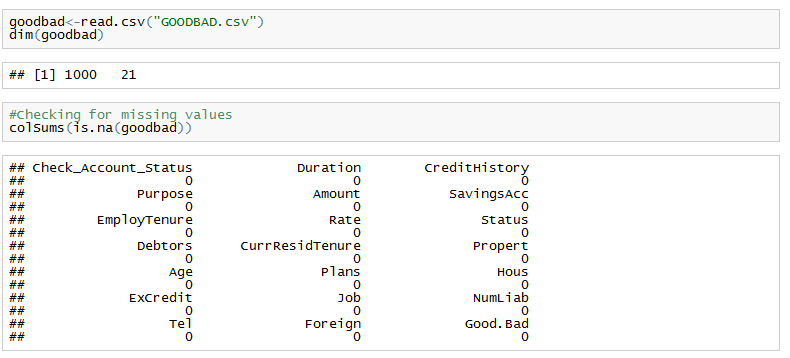
1. The Equation for logistic curve is : p(y)=1/(1+exp(-a-bx))
2. 1-p/p=exp(-a-bx)
3. ln((1-p)/p)=-a-bx
4. Final Equation : ln(p/(1-p))=a+bx

p/(1-p) = Odds Ratio. We build a model to predict the log(odds ratio)

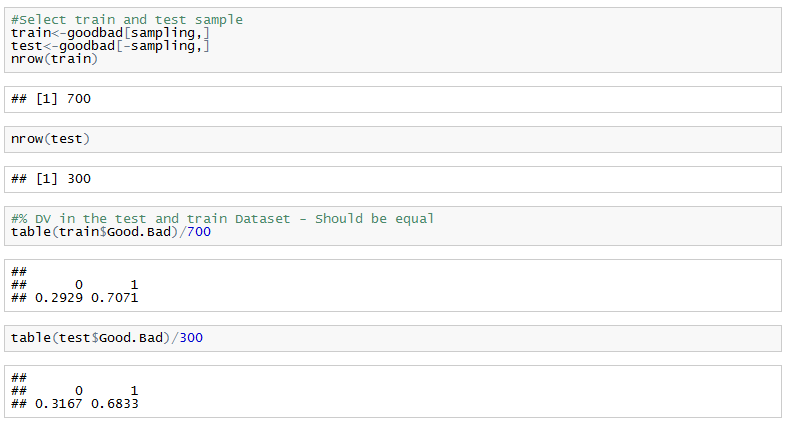
**Logistic Regression in R**

The dataset considered here is GoodBad.csv. It has details about whether a customer will default or not. Our DV is Good.Bad(1=good customer,0=Bad customer)

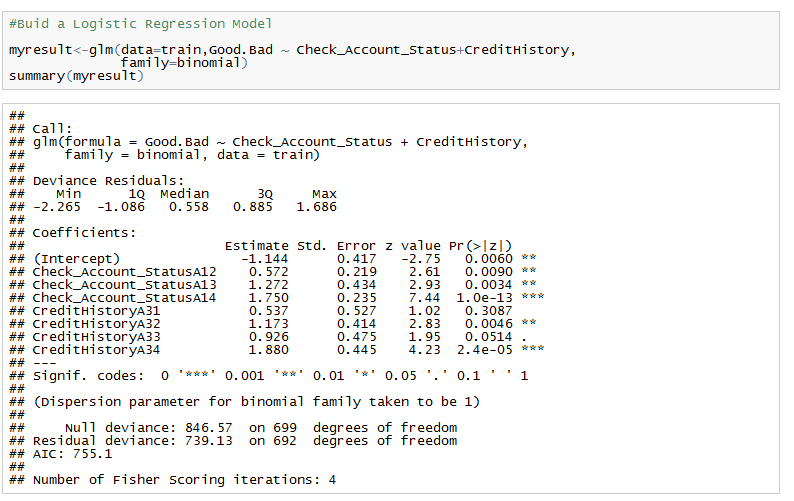
1. Read the data and remove missing values



1. Split the data into training and test data samples



1. Build a logistic model



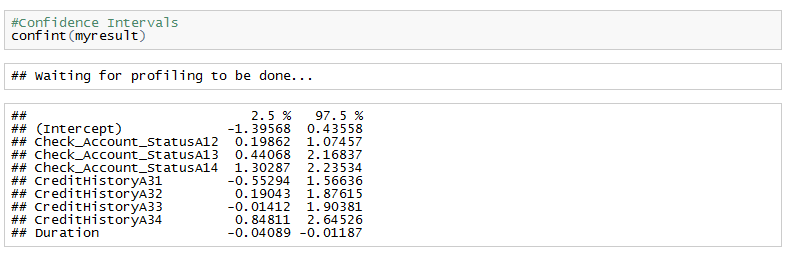
Here the DV=Good.Bad (whether a customer is a good or a bad customer

Following parameters we get from the summary of “myresult”

* **Estimate** :
  + Tells us by how the IDV is influencing the DV(Good.Bad)
  + For Example : When account status increases by 1 unit then the log of odds increases by 57% .This is because we are predicting the log of odds ratio and not the DV directly
* **Pr** : This is the p value. Tells us whether the variable is significant or not.
* **Deviances**
  + **Null deviance**: Gives the model Error when No predictor variables (IDV’s ) are involved
  + **Residual deviance**: The model Error when predictors are taken into account
  + Hence always Residual deviance< Null deviance so that we can say that the IDv’s have contributed in predicting the DV well
* **AIC** :It’s the model performance metric. Lower the value,the better the model is and also it’s a less complex model (with optimum no of variables/right number of variables)

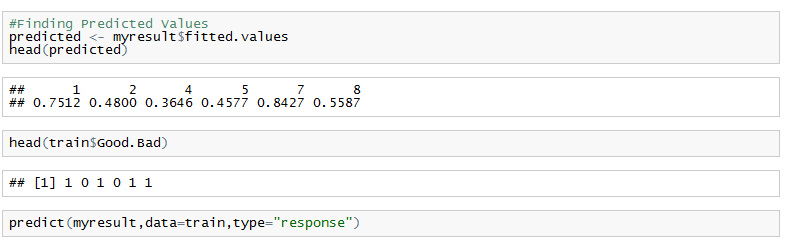
1. **Checking confidence intervals** :

* This tells us how confident we are about our predictions wrt to the IDV’s.
* The narrower the confidence intervals are the more accurate the predictions will be.
* For eg: we see that CreditHistoryA31 ranges between -0.5 to 1.5 which is huge confidence interval. So next time we run a model we could get a different value.
* So a good model should have a narrow Confidence Interval



1. **Predictions**

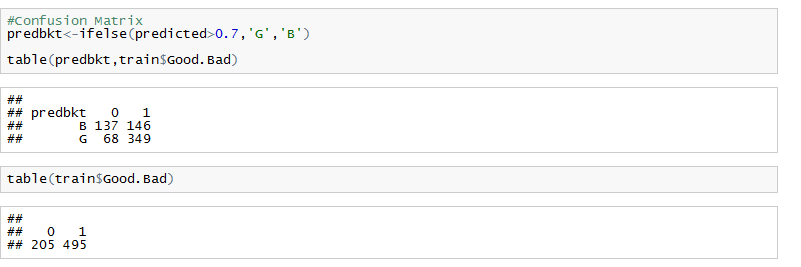
* We can predict values using either the “fitted.values” or predict function with type=”response”.
* The output we get is predicted probabilities and not discrete values(0/1).
* For Eg if the actual value of Good.Bad=1 and we get predicted output as 0.8 , then this can be interpreted as : We are 80% confident that Good.Bad is 1



1. **Confusion Matrix**

This is used for validating the model. It gives us a matrix of Actual Vs Predicted values.

Before that we need to choose a cutoff value to convert our predicted probabilities back to 0’s and 1’s.Here we have chosen cutoff as “0.7”

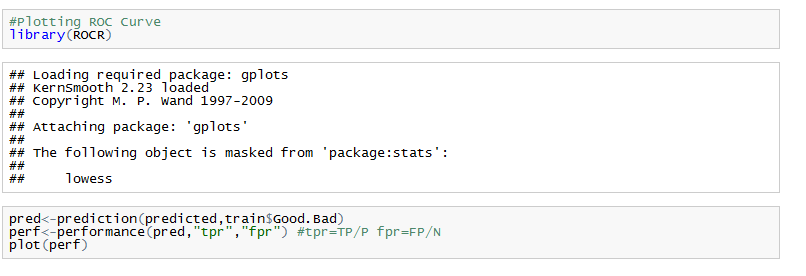


Looking at the confusion matrix, the four metrics we get are :

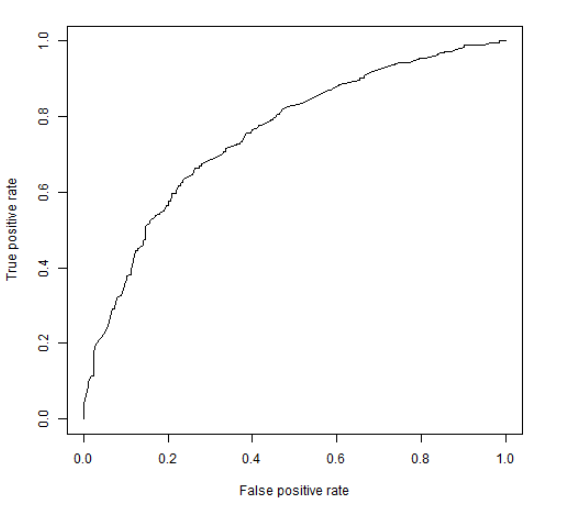
* True Positives: The model has accurately predicted the 1’s (G=1)
* False Positives/Alarms: The actual values are 0’s but the model has predicted them as G
* True Negatives: The model has accurately predicted the 0’s (B=0)
* False Negatives :The actual values are 1’s but the model has predicted them as B

We must always try and maximise our True Positives and True Negatives

1. **ROCR Curve**

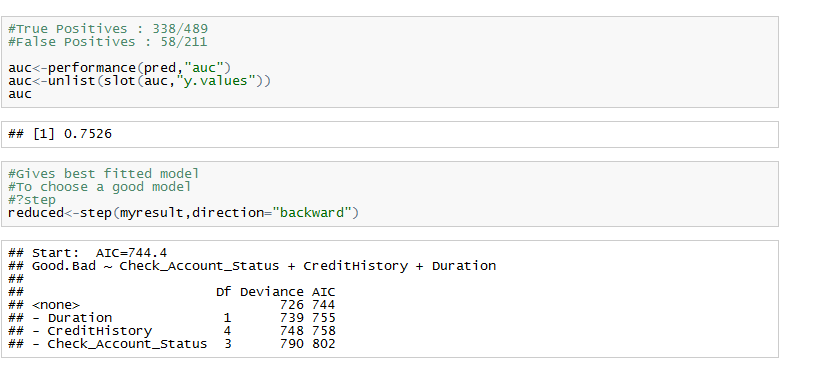


* ROCR curve means Receiver Operator Characteristics Curve
* Before plotting it, we need to convert the predicted values to object of type “prediction”
* These values are then fed into the function “performance” .The “performance” function in R directly helps us find values such as “fpr” (False Positive Rate”),tpr(True Positive Rate”)
* Check *?performance* to know more details about other parameters it gives.
* The ROCR Plot is a plot between FPR and TPR
  + The “performance” function, for all cut off values -> calculates the FPR and TPR and throws the result, which is then plotted.
  + We should also pick up a higher TPR and a lower FPR. That will be a good model.



1. **AUC value**

* This is a model performance metric. It is the area under the curve.
* This is obtained from the “performance” function where you can choose the option “AUC”
* In the ROCR the higher the value of TPR, the greater will the area under the curve (it is obvious if you observe the graph).
* Generally a good logistic model has an AUC >70%



1. **To choose the best model**

As shown above,Use function called as step. IF you mention direction =”backward”, it will throw all the variables and the respective AIC values. Then from this you can choose variables with the least AIC value and re run your model