**Exploring IRIS Data set with EDA**

**DATA MINING**

LOGISTIC regression and Probit regression

Table of Contents

[Problem and Approach (Simulation Study) 3](#_Toc857927)

[Major Result 3](#_Toc857928)

[Problem and Approach (German Credit data) 5](#_Toc857929)

[Exploratory Data Analysis 5](#_Toc857930)

[Stepwise variable selection using AIC 6](#_Toc857931)

[stepwise variable selection using BIC 6](#_Toc857932)

[Chi-square test for significance of variables 7](#_Toc857933)

[Lasso variable selection 7](#_Toc857934)

[Final model for GLM 7](#_Toc857935)

[Model Evaluation 7](#_Toc857936)

[ROC Plot 8](#_Toc857937)

[Out of sample misclassification rate and AUC score 8](#_Toc857938)

[Asymmetric misclassification rate giving more penalty for false positives 8](#_Toc857939)

[Problem and Approach (Bankruptcy Data) 9](#_Toc857940)

[Major Findings 9](#_Toc857941)

[Conclusion 9](#_Toc857942)

[Exploratory Data Analysis 9](#_Toc857943)

[Model creation using Logistic Regression 12](#_Toc857944)

[Model selection using the BIC criterion 12](#_Toc857945)

[Misclassification for training model: 13](#_Toc857946)

[Performance of the model with testing data: 14](#_Toc857947)

# **Problem and Approach (Simulation Study)**

Usually General Linear Model refers to Classical linear regression models for a continuous response variable given continuous and/or categorical predictors. For general linear model fitting, lm() function is used with identity link. General Linear Model is a special case of Generalized Linear Model (GLIM or GLM). For fitting Generalized Linear Model, glm() function is used. Logistic regression or probit regression is one of the most well-known Generalized Linear Model.

Logit and Probit models can be employed when response variable Y is a binary random variable which takes on only the values of zero and one. In this problem, we considered the case where the probability that y takes on the value zero or one is conditional on two explanatory variables, x1 and x2. The variable x1 was following uniform distribution where minimum value with 0 and maximum value with 1, and it was generated runif() function. Since had value of 1 for odd, and 0 for even, the variable x2 was created using rep() function.

In Part A, we assumed Y was following Binary distribution with probability of success of p where logit() =log( = -1+5.2\*-0.4\*. Consequently, p was calculated as or . After the list of p (p1) was created, y1 was generated using rbiom() function with n was set as 500 and size as 1.

In Part B, we assumed Y was following Binary distribution with probability of success of p where -1+5.2\*-0.4\*. The probability of success, p2, was calculated using pnorm() function. After the list of p2 was created, y2 was created using rbinom() function as above.

# **Major Result**

**Table 1.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Link** | **Estimated coefficient** | | | | |
|  |  |  | **Residual Deviance** | **AIC** |
| Part A.  logit()= -1+5.2\*-0.4\* | **Logit** | -1.40 | 5.44 | -0.04 | 433.6 | 439.6 |
| **Probit** | -0.84 | 3.19 | -0.01 | 431.9 | 437.9 |
| Part B.  -1+5.2\*-0.4\* | **Logit** | -1.70 | 9.43 | -1.01 | 287.4 | 293.4 |
| **Probit** | -0.98 | 5.36 | -0.58 | 286.5 | 292.5 |

1. Part A (Simulation using probability generated Logistic Regression)

In this scenario, the response variable was binary with p, where logit(p)=-1+5.2\*-0.4\* . Generalize linear model (GLM) with both logistic and probit links were performed. With Logit link, the coefficients were much closer to the truth then those with Probit link. However, probit regression has slightly better AIC and residual deviance than those estimated with Logit link.

1. Part B (Simulation using probability generated Probit Regression)

In this scenario, the response variable was binary with p, where -1+5.2\*-0.4\*. Over again generalize linear model (GLM) with both logistic and probit links were executed. This time, the coefficients estimated using probit link were much closer to the truth then those with logit link. Once more, probit regression has slightly better AIC and residual deviance than those estimated with Logit link.

Unexpectedly, Part A had much larger residual deviance and AIC compared to Part B using either logit or probit link. Based on the AIC and residual deviance, it sounds like Probit Regression performs better than logistic regression. Future investigation might be needed to find the reason.

In general, logistic regression is more widely used in many occasion compared to probit regression because Logistic regression has better interpretation than probit. Odds Ratio (OR), which is exponentiation of coefficients, provides information about the strength of association between the binary response and explanatory variable (either categorical or continuous).

# **Problem and Approach (German Credit data)**

two datasets are provided. the original dataset, in the form provided by Prof. Hofmann, contains categorical/symbolic attributes and is in the file "german.data".   
  
For algorithms that need numerical attributes, Strathclyde University produced the file "german.data-numeric". This file has been edited and several indicator variables added to make it suitable for algorithms which cannot cope with categorical variables. Several attributes that are ordered categorical (such as attribute 17) have been coded as integer. This was the form used by StatLog.

# **Exploratory Data Analysis**

There is a total on 21 attributes in the dataset. Their descriptions and details have been tabulated below:

* Status of existing checking account.
* Duration in month
* Credit history
* Purpose
* Credit amount
* Savings account/bonds
* Present employment since
* Installment rate in percentage of disposable income
* Personal status and sex
* Other debtors / guarantors
* Present residence since
* Property
* Age in years
* Other installment plans
* Housing
* Number of existing credits at this bank
* Job
* Number of people being liable to provide maintenance for
* Telephone
* foreign worker

We take the summary statistics of the dataset, the dataset has a total of 1000 observations with 21 variables, out of which 8 are numerical variables including the response and 13 are categorical variables with various levels

We get the following insights from our EDA of continuous variables:

* From the age variable, we see that the median value for bad records is lesser than that of good records, it might be premature to say young people tend to have bad credit records, but we can safely assume it tends to be riskier.
* The installment\_rate variable has a great deal of difference between the good and bad records, we see that bad records have almost the double median value than good ones.
* The median value and the range of the duration variables appears to be on the higher side of bad records as compared to good records
* For the amount variable, we observe that the amount for bad records is larger in general as compared to good ones
* We further built on this by plotting the density curve along the vertical line for their mean value and find that there is a great deal of difference for the duration as well as amount variable.

We get the following insights from our EDA of categorical variables:

* For chk\_acct we see that, the current status of the checking account matters as the frequency of the response variables is seen to differ from one sub category to another, overall A11 houses more number of bad credit records and A14 the least
* For credit\_his, we observe that proportion of the response variable varies significantly, for categories A30, A31 we see the number of bad credit records are greater.
* For the purpose variable, we observe that the proportion of good and bad credit record varies also overall A44, A45, A410 and A46 seem to include more risky records.
* We also observe these trends in other variables like sex, other\_debtor, saving\_acct, other\_install and foreign. Overall, the trend looks significant in saving\_acct, purpose, credit\_his and chk\_acct as compared to others.

# **Stepwise variable selection using AIC**

From stepwise variable selection method using AIC, the significant variables are:

* chk\_acct
* duration
* credit\_his
* purpose
* amount
* saving\_acct
* installment\_rate
* sex
* other\_debtor
* age
* other\_install
* telephone
* foreign

# **stepwise variable selection using BIC**

From stepwise variable selection method using BIC, the significant variables are:

* chk\_acct
* duration

# **Chi-square test for significance of variables**

Using drop-1 method to check variable importance, we find the significant variables as:

* chk\_acct
* duration
* credit\_his
* purpose
* amount
* saving\_acct
* installment\_rate
* other\_debtor
* other\_install
* foreign

# **Lasso variable selection**

To get variable selection using LASSO, we first create matrix of the dataset.

We fit the LASSO model to our data. From the plot below, we see that as the value of lambda keeps on increasing, the coefficients for the variables tend to 0.

Using cross validation to find perfect lambda value

# **Final model for GLM**

For our final model, we select the final variables:

* chk\_acct
* duration
* credit\_his
* amount
* saving\_acct
* installment\_rate
* other\_install

credit.glm.final <- glm(response ~ chk\_acct + duration + credit\_his + amount + saving\_acct + other\_install + installment\_rate, family = binomial, german\_credit.train)

# **Model Evaluation**

**In-sample misclassification rate**

Keeping cutoff as 0.1667, we calculate the misclassification rate:

**Confusion Matrix**

Checking for the predictions and seeing the False Positive and False negative values from the below confusion matrix:

# **ROC Plot**

ROC Plot for the same is plotted below and the AUC is 0.7896875

# **Out of sample misclassification rate and AUC score**

We get a misclassification rate of 0.395, and AUC of 0.7734524

# Asymmetric misclassification rate giving more penalty for false positives

In cases where we need to penalize the False Negative more than False Positive, we use a 5:1 penalty for misclassification and see an error rate of 0.535

# **Problem and Approach (Bankruptcy Data)**

The problem statement was to explore and analyze bankruptcy dataset using Logistic regression to find the best fit model for bankruptcy prediction.

We performed data exploration on the given data set which revealed that dataset had one qualitative response variable and 10 predictors. There were no missing values in the dataset.

The dataset was randomized into training and testing data where training data included 75% of the data points and testing included the remaining 25%. We performed logistic regression using stepwise variable selection method and found out the AIC, BIC values for In-sample data. We also plotted the Area Under the Curve graph (AUC) and computed the area value. We also found out the misclassification rate for the training data. Performing these steps, we found out that the most significant co-variates for our model. We then tested this finalized model using the testing dataset and predicted the response variable using these co-variates in order to verify the above model.

# **Major Findings**

The majority companies in the dataset are bankrupt.

R2, R3, R6, R7, R8, R9 and R10 were the most significant co-variates in the model.

Comparison of In-sample and out of sample dataset resulted in similar AIC, BIC , AUC values which suggest that the finalized model was capable of decent prediction.

After performing the regression, the below model resulted as the final model:

DLRSN ~ R2 + R3 + R6 + R7 + R8 + R9 + R10

# **Conclusion**

The misclassification rate and AUC (Area Under Curve) is close to each other for the training as well as the test dataset. The AUC for training dataset is almost same as the test data while the misclassification rate for the test data is only slightly higher than the training dataset.

Thus, we can say that our finalized model has a performance for both the training and testing dataset. It can predict the bankruptcy status of the companies with around 64% accuracy.

# **Exploratory Data Analysis**

Bankruptcy dataset is a dataset which contains the financial information and the bankruptcy status of the companies for specific years. This dataset is a subset of Man Xu’s project containing the financial information of the companies. It contains 5436 observations with 13 variables. Variable CUSIP contains unique code to identify the company, variable FR determines the fiscal year, where approximately 85% of the records belong to year 1999. Variable DLRSN is the Bankruptcy/Non-Bankruptcy indicator, where 1 stands for Bankruptcy while 0 stands for Non Bankrupt companies. Variables R1 to R10 contain financial information which will be used while building a logistic regression model.

Ten covariates:

R1=Working Capital/Total Asset;

R2=Retained Earning/Total Asset;

R3=Earning Before Interest & Tax/Total Asset;

R4=Market Capital / Total Liability;

R5=SALE/Total Asset;

R6=Total Liability/Total Asset

R7=Current Asset/Current Liability;

R8=Net Income/Total Asset;

R9=LOG(SALE);

R10=LOG(Market Cap)

We performed the exploratory data analysis by first looking at a subset of the data set using head() function. Below is our sample data:

**Table 1: Sample dataset**

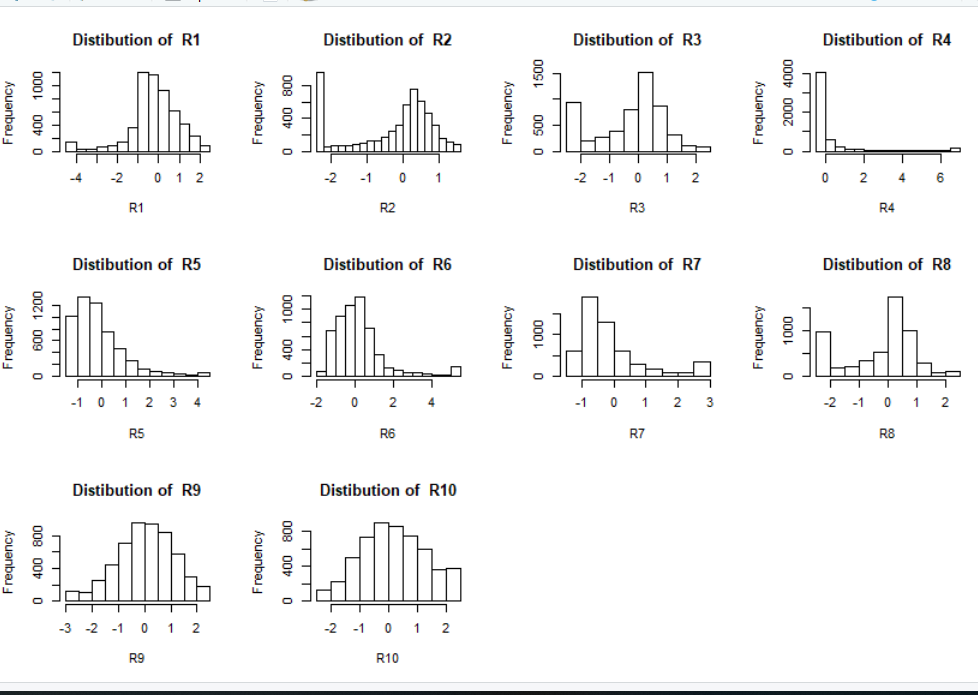
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **FYEAR** | **DLRSN** | **CUSIP** | **R1** | **R2** | **R3** | **R4** | **R5** |
| 1999 | 0 | 36020 | 0.30714 | 0.887006 | 1.647681 | -0.19916 | 1.092964 |
| 1999 | 0 | 36110 | 0.760737 | 0.592493 | 0.453003 | -0.36989 | 0.186154 |
| 1999 | 0 | 37520 | -0.5136 | 0.337615 | 0.299015 | -0.02908 | -0.4326 |
| 1994 | 1 | 78110 | -0.46613 | 0.370747 | 0.496067 | -0.37343 | -0.26742 |
| 1999 | 0 | 00079X10 | 2.023422 | 0.214876 | 0.182595 | 6.69536 | -1.14834 |
| 1999 | 0 | 00086T10 | 0.907499 | 0.38688 | 0.477891 | -0.34716 | 1.407989 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **R6** | **R7** | **R8** | **R9** | **R10** |
| -0.31329 | -0.19679 | 1.206763 | 0.2824709 | 0.158896 |
| 0.039619 | 0.327497 | 0.428418 | 1.1069652 | 0.793443 |
| 0.829993 | -0.70779 | 0.476153 | 2.1791755 | 2.484585 |
| 0.977799 | -0.61098 | 0.45681 | 0.1519511 | 0.047789 |
| -1.50589 | 2.876477 | 0.287375 | -0.986442 | 0.79108 |
| -0.4839 | 0.070259 | 0.527811 | 0.5024659 | -0.16465 |

There are no missing values in the data set.

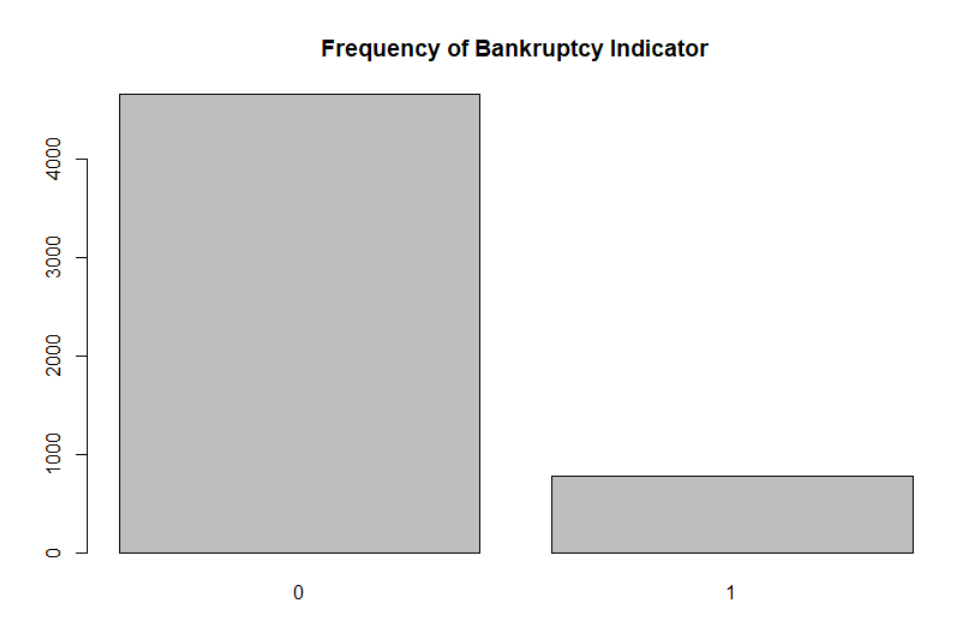
We now check the distributions of all covariates.

**Figure 1: Histogram plot of all co-variates**



From the histogram plot of the distribution of each variable, we can interpret that R9 and R10 follow normal distribution. Nothing much can be inferred about the other variables through the histogram plot.

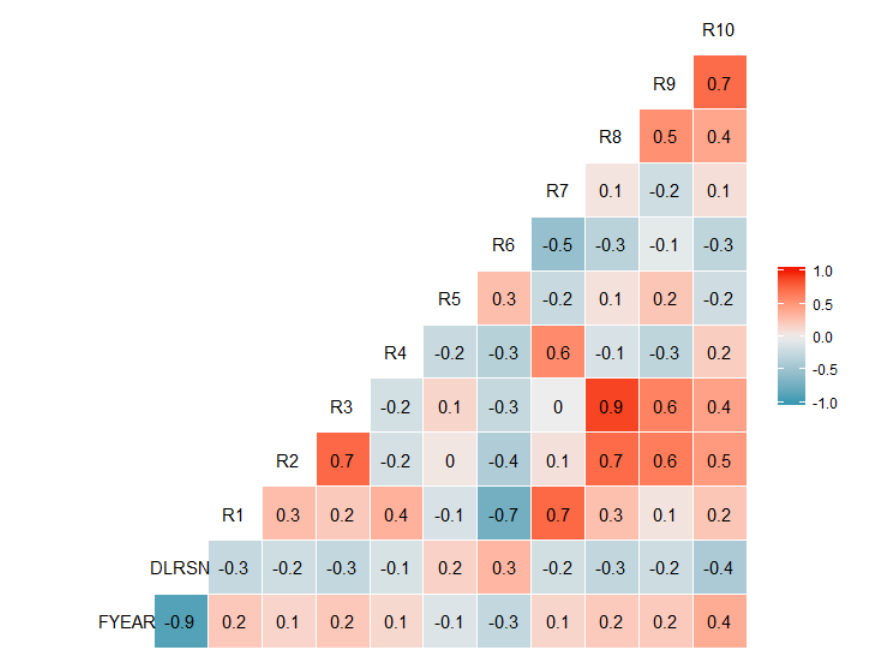
**Figure 2: Frequency distribution of response variable**



The majority companies in the dataset are bankrupt as portrayed in the above graph. The 0 value indicates bankruptcy.

The below plot projects the correlation between the response variable and the co-variates.

**Figure 3: Correlation plot**



It can be seen that R10 is having high correlation with the response variable bankruptcy flag having a correlation of -0.425. R6 is also having a correlation of 0.331 with response variable. There is a high correlation between R8 and R3.

# **Model creation using Logistic Regression**

Logistic regression is suited for response variable being qualitative. The bankruptcy variable is the response variable ion this dataset which is qualitative holding a value of 0 or 1. Hence, we would regress the response variable using the logistic regression in this case.

## Model selection using the BIC criterion

We have created two models in this case. Model 1 contains only the response variable and no co-variates. Model 2 contains all the co-variates.

The best model is selected using the Stepwise method (using BIC criteria)

**Table 2: Summary Statistics using Stepwise**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coefficients** | **Estimate** | **Std.Error** | **z value** | **Pr (>|z|)** |
| (Intercept) | -2.54038 | 0.07913 | -32.106 | <2.00E-16 |
| R10 | -1.57741 | 0.08496 | -18.566 | <2.00E-16 |
| R7 | -0.30274 | 0.07821 | -3.871 | 0.000108 |
| R8 | -0.43076 | 0.09067 | -4.751 | 2.02E-06 |
| R2 | 0.66977 | 0.08374 | 7.998 | 1.26E-15 |
| R6 | 0.18285 | 0.04535 | 4.032 | 5.52E-05 |
| R9 | 0.41526 | 0.08569 | 4.846 | 1.26E-06 |
| R3 | -0.42406 | 0.10451 | -4.057 | 4.96E-05 |

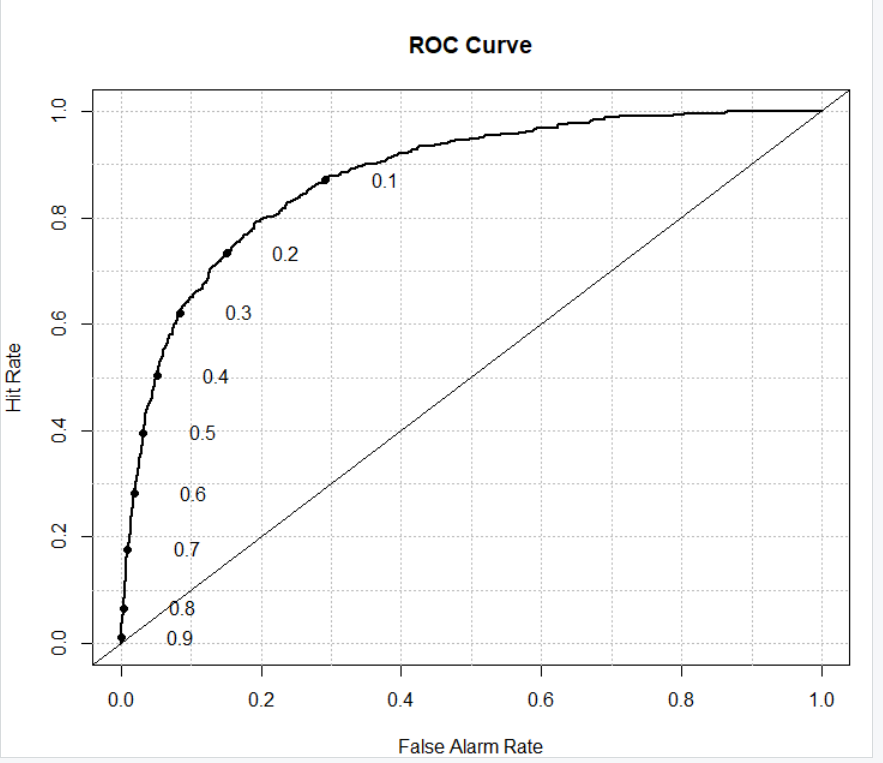
The model with BIC criterion has 7 predictors namely R2, R3, R6, R7, R8, R9 and R10.

Resulting model:

DLRSN ~ R2 + R3 + R6 + R7 + R8 + R9 + R10

Using the above model, we proceed to plot the ROC curve to get the area under the curve (AUC). The ROC curve is plot of Sensitivity (True Positives) vs Specificity (False Positives). Higher AUC implies better model performance.

**Figure 4: ROC curve of training dataset**



**Table 3: Area Under the curve**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Area** | **p.value** | **binorm.area** |
| Model 1 | 0.87685 | 1.70E-188 | NA |

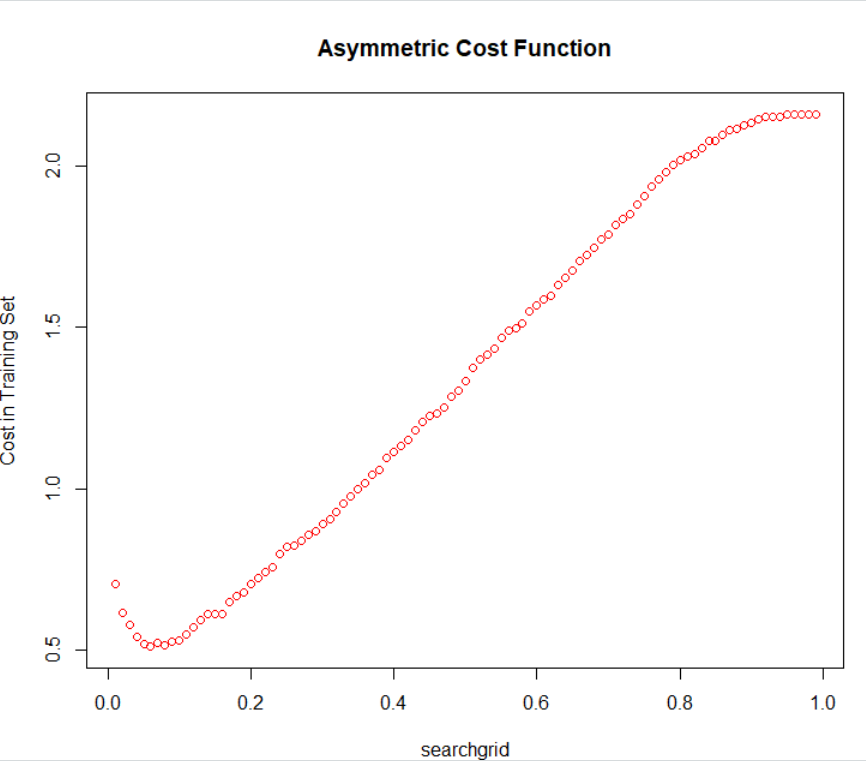
The above figure shows the ROC Curve for the training Dataset and the Area under the Curve (AUC) obtained is 0.88

# **Misclassification for training model:**

For computing the misclassification rate, first we find out the threshold for categorizing the bankruptcy status. We plot a graph of cost of training data vs threshold value using the cost function. We choose the value with lowest cost as the threshold to categorize the bankruptcy status. The ratio of weightage for False Positive and false negative is 1:15. Hence, classifying the Non-bankrupt companies as bankrupt may be a serious problem than classifying the bankrupt as Non-bankrupt. So, we try to reduce the false negatives here.

Below is the plot of asymmetric cost function:

**Figure 5: Asymmetric cost function**



We find the cutoff probability which is 0.06 here.

Then, we calculate the misclassification rate of the training dataset using this threshold value.

**Table 4: Misclassification table**

|  |  |  |
| --- | --- | --- |
|  | **TRUE** | |
| **PRED** | 0 | 1 |
| 0 | 2095 | 46 |
| 1 | 1395 | 541 |

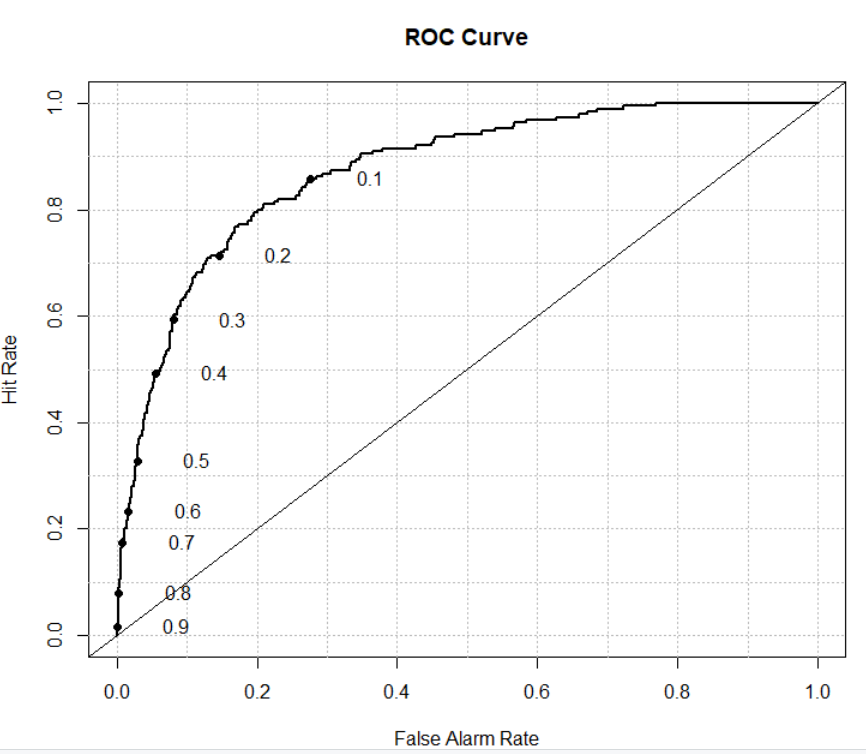
**Mean:0.35**Hence the misclassification rate obtained is 0.35 which means our model is 65% accurate in predicting the bankruptcy status.

# **Performance of the model with testing data:**

Let’s check the performance of the model on the testing data. Now we use the 25% of the data sampled initially as the test dataset. To test the performance of the final model above, we will predict the bankruptcy status for the testing data. We will check the asymmetric classification rate and area under the curve for this testing dataset.

Let us plot the ROC Curve for test dataset.

**Figure 6: ROC curve of test dataset**



**Table 5: Area under the curve**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Area** | **p.value** | **binorm.area** |
| Model 1 | 0.87304 | 2.65E-61 | NA |

The Area Under the Curve (AUC) for the test dataset is 0.87304.

We now calculate the asymmetric classification rate for the test data set. For this we will use the same cut-off probability as 0.06 as we found in the previous step.

**Table 6: Misclassification table**

|  |  |  |
| --- | --- | --- |
|  | **TRUE** | |
| **PRED** | 0 | 1 |
| 0 | 573 | 14 |
| 1 | 371 | 130 |

**Mean:** 0.3605592Hence the misclassification rate obtained is 0.3605592

which means our model is 64% accurate in predicting the bankruptcy status for the observations in testing dataset.