**BANA 7046**

**Homework 3**

Presented By:

Anjali Shalimar, Digvijay Kawale, Meenal Narsinghani

**Index:**

**Simulation Study**

1. Study of Logistic Regression Using the Simulation Study (Logit)
2. Study of Logistic Regression Using the Simulation Study (Probit)

**German Credit Scoring data**

1. Problem Statement and Approach
2. Analysis
3. Results and Conclusion

**Predicting Bankruptcy using Logistic Regression Model**

1. Problem Statement and Approach
2. Exploratory Data Analysis
3. Variable Selection
4. In Sample Prediction
5. Out of Sample Prediction
6. Results and Conclusion

**1. Simulation Study**

**1.1** **Study of Logistic Regression Using the Simulation Study ( Logit )**

A simulation study was conducted to study the Logistic Regression to predict a binary variable. Below are the assumptions of the study:

y|x ~ Binary(p), where p = E(y|x), and   
 logit(pi) = -1.1 + 5\*x1 – 0.4\*x2

Where x1 is a random variable uniformly distributed between 0 and 1 & x2 is a variable which takes value of 1 for odd index and takes value of 0 for even index. The seed set for the generation of random variables is 13437586.

The Binary Response Y is generated using the rbinom plugging the probabilities calculated using the above equation. General linear model is fitted using x1 and x2 as the predictors for the binary response variable Y. The total number of observations used are 500.

After fitting the general linear model (Logit), below results were obtained for the coefficients:

|  |  |  |
| --- | --- | --- |
| Intercept | X1 | X2 |
| -1.2842 | 4.5394 | -0.1186 |

We observed that the estimated coefficients are highly close to the actual parameters of the model. Also, the p-values are significant for the parameter estimate of x1 and not significant for the parameter estimate of x2. The null deviance and residual deviance of the model are 648.68 and 508.53 respectively. The AIC value of the model is 514.53.

This model was used to predict the probabilities of the binary variable Y. By using the cut off of 0.5 used in the above assumption, Y variable was predicted and compared with the actual Y values. Below are its results:

|  |  |  |  |
| --- | --- | --- | --- |
| Y Predicted | | 0 | 1 |
| Y Actual | 0 | 105 | 71 |
| 1 | 55 | 269 |

The area under the curve this model is 0.7134. Which tells us that the model is able to predict the response variable with an accuracy of 71.34%

**1.2 Study of Logistic Regression Using the Simulation Study ( Probit )**

A simulation study was conducted to study the Logistic Regression to predict a binary variable. Below are the assumptions of the study:

y|x ~ Binary(p), where p = E(y|x), and   
 logit(pi) = -1.1 + 5\*x1 – 0.4\*x2

Where x1 is a random variable uniformly distributed between 0 and 1 & x2 is a variable which takes value of 1 for odd index and takes value of 0 for even index. The seed set for the generation of random variables is 13437586.

The Binary Response Y is generated using the rbinom plugging the probabilities calculated using the above equation. General linear model is fitted using x1 and x2 as the predictors for the binary response variable Y. The total number of observations used are 500.

After fitting the general linear model (Probit), below results were obtained for the coefficients:

|  |  |  |
| --- | --- | --- |
| Intercept | X1 | X2 |
| -1.3045 | 4.9743 | -0.1612 |

We observed that the estimated coefficients are highly close to the actual parameters of the model. Also, the p-values are significant for the parameter estimate of x1 and not significant for the parameter estimate of x2. The null deviance and residual deviance of the model are 614.22 and 343.14 respectively. The AIC value of the model is 349.14.

This model was used to predict the probabilities of the binary variable Y. By using the cut off of 0.5 used in the above assumption, Y variable was predicted and compared with the actual Y values. Below are its results:

|  |  |  |  |
| --- | --- | --- | --- |
| Y Predicted | | 0 | 1 |
| Y Actual | 0 | 111 | 41 |
| 1 | 43 | 305 |

The area under the curve this model is 0.8034. Which tells us that the model is able to predict the response variable with an accuracy of 80.34%

**Conclusion:** After comparing both the logit and probit form of the general linear models, both have parameter estimates close to that of the actual parameters. The predictions are better using the probit form of model.

1. **Use German Credit Scoring data**

**2.1 Problem Statement and Approach:**

The German credit scoring data is a dataset provided by Prof. Dr. Hans Hofmann in the file “german.data”. The dataset can be found [here](http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)).

The data set has information about 1000 individuals who have been classified as risky or not. The variable response in the dataset corresponds to the risk label, 1 has been classified as bad and 2 has been classified as good.

The main goal of this analysis is to build a model based on logistic regression technique that will help to identify a customer as good or bad. It is important to note that: It is worse to class a customer as good when they are bad, than it is to class a customer as bad when they are good.

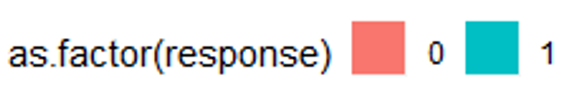
**2.2 Analysis**

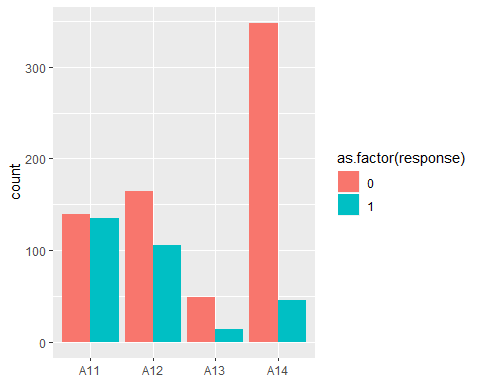
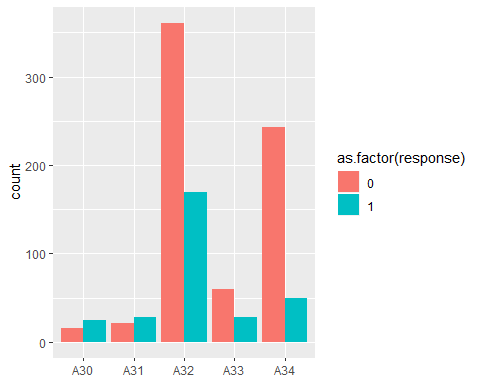
To build a logistic regression-based model that classifies customers as good or bad below steps were followed:

1. Initial exploratory data analysis was performed on the raw dataset
2. The dataset was then split into train and test in 70-30 ratio
3. Best model was chosen by running step AIC and BIC
4. In-sample performance was evaluated by plotting ROC curve and calculating metrics like: AUC and misclassification rate
5. Finally, out-of -sample performance was measured
6. **Exploratory data analysis on raw dataset**
7. **Basic EDA**
8. The dataset contains 1000 instances and 21 attributes, of which there are 13 categorical variables and 8 numerical variables
9. One of the numerical variables “response” actually represents the class (0 = good, 1 = bad)
10. The data is complete – there are no missing values
11. Summary statistics of each attribute shows no signs of abnormality
12. **Univariate Analysis**

***Distributions of the predictor variables were plotted to examine its relation on the “response” variable.***

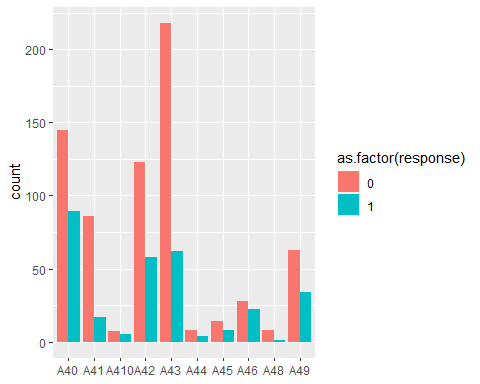
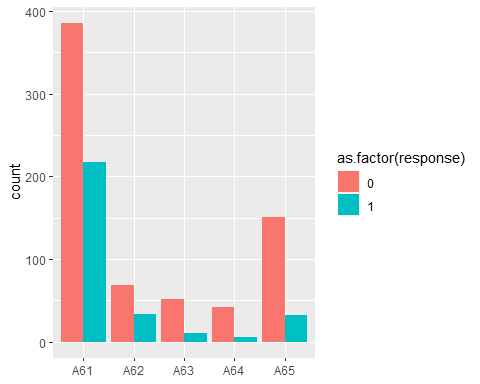
1. **Categorical variables**



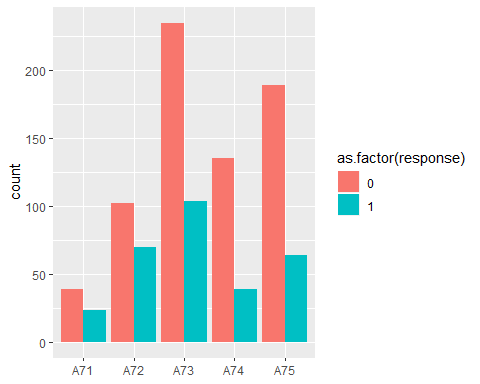
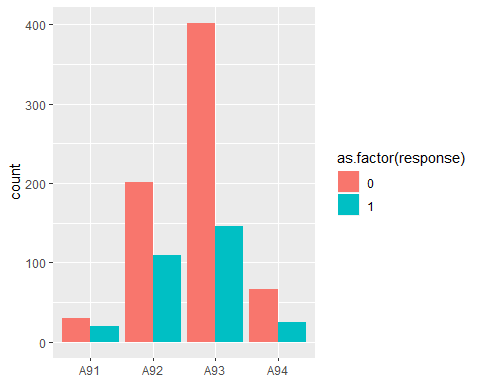
credit\_his

chk\_acct

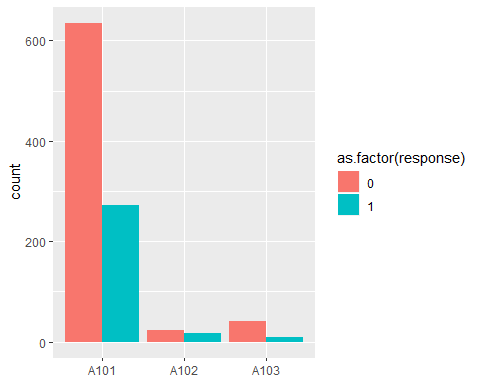
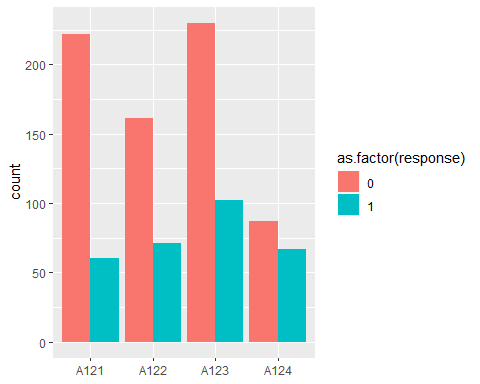
Saving\_acct

purpose

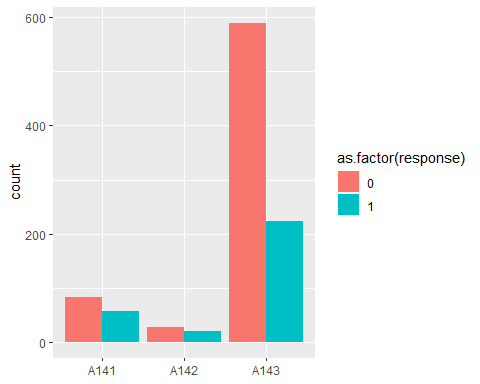
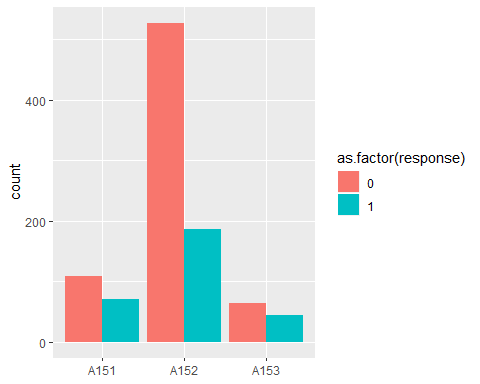
sex

present\_emp

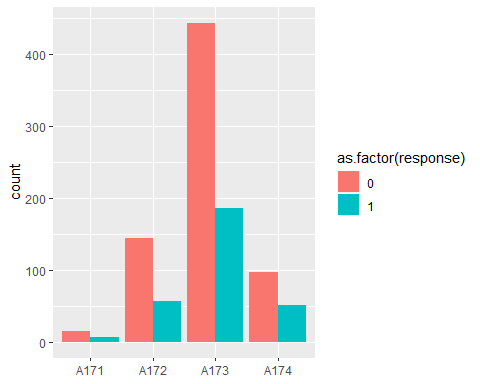
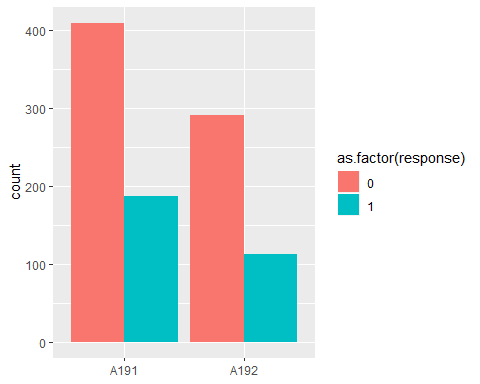
property

other\_debtor

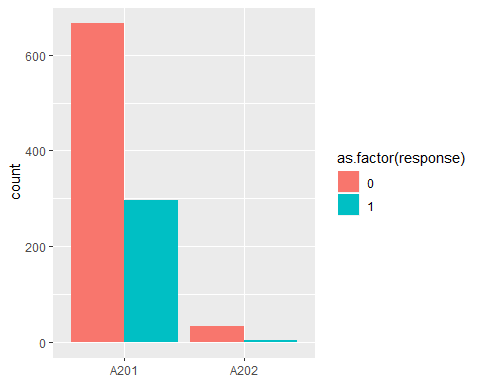
housing

other\_install

job

telephone



foreign

**Some of the key insights from our EDA of categorical variables are:**

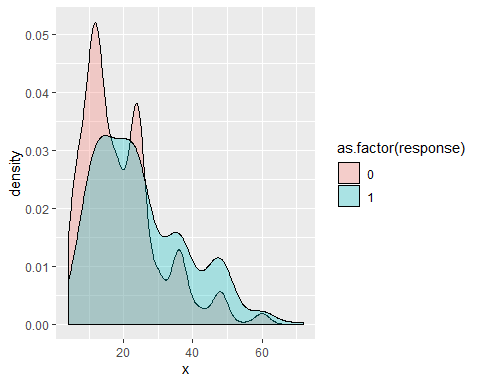
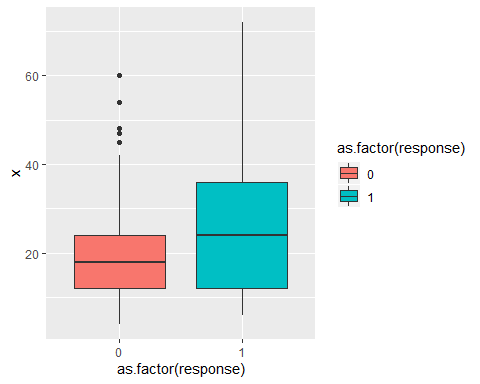
* + chk\_acct of value A11 has the most number of bad credit records, whereas A14 has the least
  + Similar trend is observed in variables purpose, credit\_his where the proportion of good and bad records varies significantly over the range of values
  + Variables like telephone, present\_emp do not have any interesting variation

**Numerical variables**

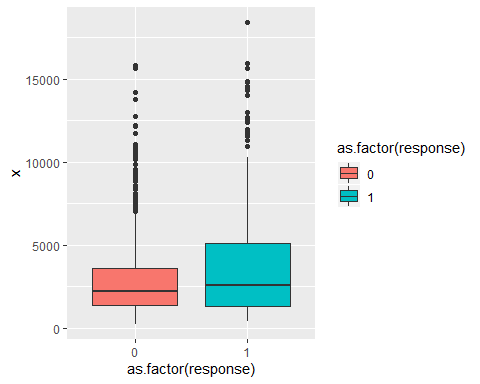
1. duration
2. amount
3. installment.rate
4. present\_resid
5. age
6. n\_credits
7. n\_people

Visualizing the distribution of these variables against the response variable

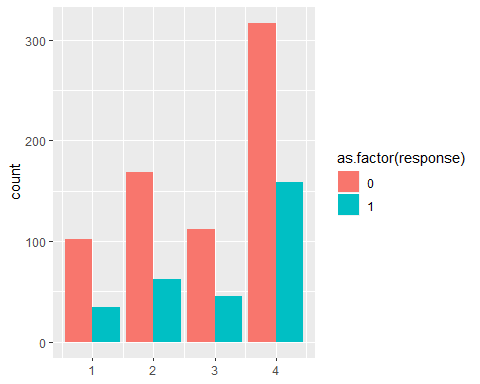
**1. Duration**

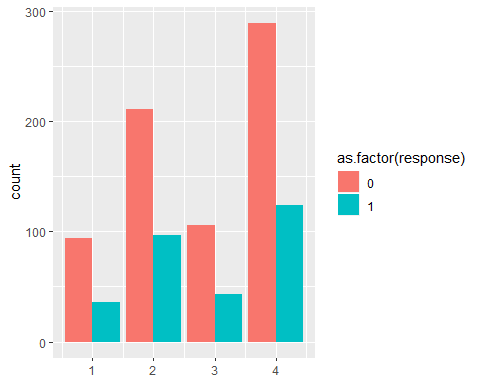
**2. Amount**



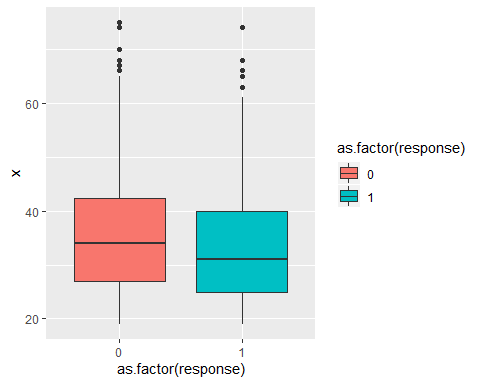
**3. Installment Rate**



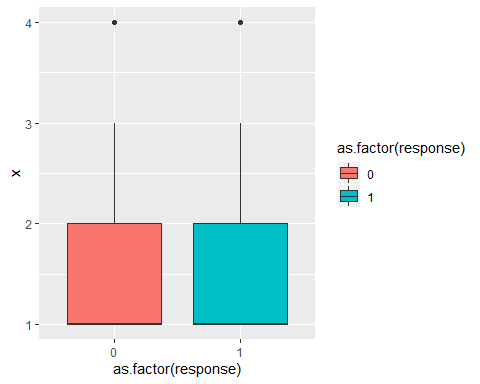
**4. Residence Since**



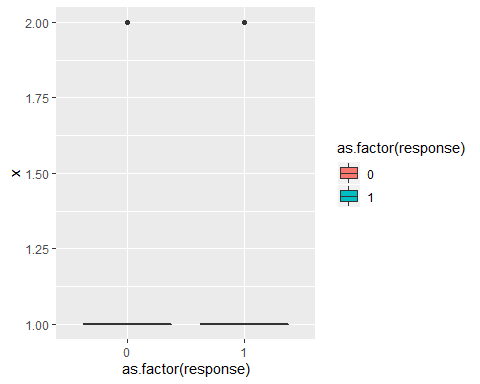
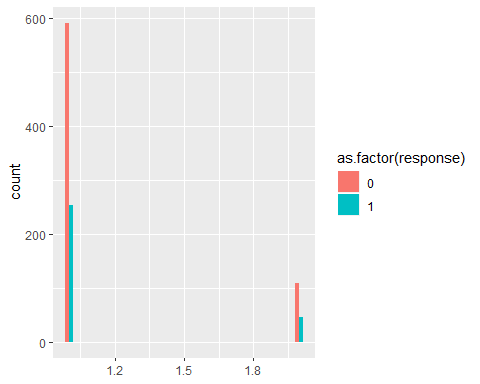
**5. Age**



**6. Credits**



**7. n\_people**

*Some of the insights from EDA of continuous variables are:*

* + 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
  + The value of “amount” variable for bad records is higher than that for the good records
  + The median value of “age” for bad records is lesser than that of good records, which indicate that young people tend to be more risky

1. **Train/Test split**

Randomly split the data to training (70%) and testing (30%) datasets:

1. **Find a best model for Credit Scoring data using logistic regression with AIC and BIC**

***Training the model with all the predictor variables***

credit.glm0 <- glm(response~., family=binomial, data=credit.train)

**Extracting certain criteria from the zero-model**

**Deviance = 577.13**

**AIC = 675.13**

**BIC = 898.13**

***Variable selection with stepwise approach***

**AIC based model**

glm(formula = response ~ chk\_acct + duration + credit\_his + purpose + amount + saving\_acct + installment\_rate + sex + other\_debtor + other\_install + housing + foreign, family = binomial, data = credit.train)

Significant predictors –

1. chk\_acct

2. duration

3. credit\_his

4. purpose

5. amount

6. saving\_acct

7. installment\_rate

8. sex

9. other\_debtor

10. other\_install

11. housing

12. foreign

**Model metrics**

**## Deviance = 594.54**

**## AIC = 660.54**

**## BIC = 810.73**

***BIC based model***

glm(formula = response ~ chk\_acct + duration + other\_debtor + foreign, family = binomial, data = credit.train)

**Significant predictors are –**

**1. chk\_acct**

**2. duration**

**3. other\_debtor**

**4. foreign**

***Model metrics***

**## Deviance = 710.11**

**## AIC = 722.11**

**## BIC = 749.42**

***Given the AIC of the model based on AIC stepwise approach is smaller. We will select the AIC based model.***

***Final model***

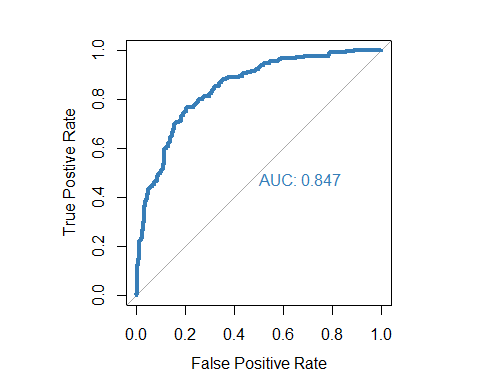
credit.model <- glm(formula = response ~ chk\_acct + duration + credit\_his + purpose +   
 amount + saving\_acct + installment\_rate + sex + other\_debtor +   
 other\_install + housing + foreign, family = binomial, data = credit.train)  
  
  
summary(credit.model)

##   
## Call:  
## glm(formula = response ~ chk\_acct + duration + credit\_his + purpose +   
## amount + saving\_acct + installment\_rate + sex + other\_debtor +   
## other\_install + housing + foreign, family = binomial, data = credit.train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4835 -0.6741 -0.3490 0.6155 2.7575   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 7.925e-01 8.243e-01 0.961 0.33635   
## chk\_acctA12 -3.158e-01 2.616e-01 -1.207 0.22733   
## chk\_acctA13 -8.144e-01 4.437e-01 -1.836 0.06641 .   
## chk\_acctA14 -2.025e+00 2.908e-01 -6.966 3.26e-12 \*\*\*  
## duration 2.414e-02 1.063e-02 2.271 0.02317 \*   
## credit\_hisA31 -5.302e-01 6.216e-01 -0.853 0.39364   
## credit\_hisA32 -7.826e-01 4.891e-01 -1.600 0.10958   
## credit\_hisA33 -1.559e+00 5.783e-01 -2.696 0.00701 \*\*   
## credit\_hisA34 -1.594e+00 5.167e-01 -3.085 0.00204 \*\*   
## purposeA41 -1.800e+00 4.483e-01 -4.016 5.92e-05 \*\*\*  
## purposeA410 -6.692e-01 8.594e-01 -0.779 0.43620   
## purposeA42 -7.657e-01 3.149e-01 -2.431 0.01504 \*   
## purposeA43 -9.912e-01 3.048e-01 -3.253 0.00114 \*\*   
## purposeA44 -5.263e-01 8.398e-01 -0.627 0.53083   
## purposeA45 -2.803e-01 6.842e-01 -0.410 0.68206   
## purposeA46 8.091e-01 4.927e-01 1.642 0.10051   
## purposeA48 -1.619e+00 1.242e+00 -1.303 0.19266   
## purposeA49 -6.717e-01 4.028e-01 -1.668 0.09535 .   
## amount 1.632e-04 5.028e-05 3.244 0.00118 \*\*   
## saving\_acctA62 -5.794e-01 3.466e-01 -1.672 0.09452 .   
## saving\_acctA63 -8.590e-01 5.245e-01 -1.638 0.10149   
## saving\_acctA64 -1.090e+00 6.153e-01 -1.771 0.07651 .   
## saving\_acctA65 -1.255e+00 3.198e-01 -3.924 8.71e-05 \*\*\*  
## installment\_rate 3.264e-01 1.055e-01 3.094 0.00197 \*\*   
## sexA92 -3.615e-01 4.795e-01 -0.754 0.45087   
## sexA93 -9.988e-01 4.689e-01 -2.130 0.03316 \*   
## sexA94 -6.185e-01 5.706e-01 -1.084 0.27833   
## other\_debtorA102 6.160e-01 4.400e-01 1.400 0.16144   
## other\_debtorA103 -5.852e-01 4.967e-01 -1.178 0.23871   
## other\_installA142 8.446e-01 4.920e-01 1.717 0.08607 .   
## other\_installA143 -2.335e-02 2.990e-01 -0.078 0.93776   
## housingA152 -3.963e-01 2.729e-01 -1.452 0.14646   
## housingA153 -4.439e-01 4.056e-01 -1.094 0.27379   
## foreignA202 -1.137e+00 6.915e-01 -1.645 0.10005   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 850.06 on 699 degrees of freedom  
## Residual deviance: 602.82 on 666 degrees of freedom  
## AIC: 670.82  
##   
## Number of Fisher Scoring iterations: 5

**## Deviance = 602.82**

**## AIC = 670.82**

**## BIC = 825.56**

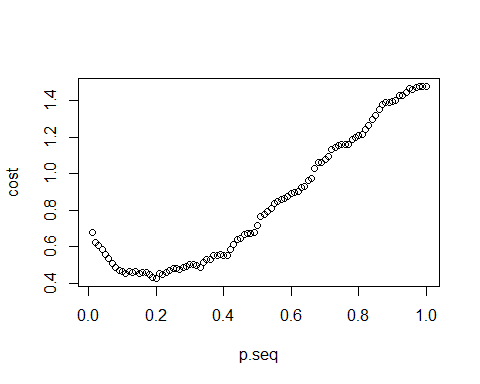
1. **Draw ROC curve, report the AUC, and present the misclassification rate table of your final model.** 

**Misclassification rate table**

**Naive Choice of Cut-off probability**

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| True | **0** | **1** |
| 0 | **375** | **118** |
| 1 | **46** | **161** |

**Optimal cut-off Probability using Cost Function**



**optimal.pcut**

[1] 0.2

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| True | **0** | **1** |
| 0 | **315** | **178** |
| 1 | **24** | **183** |
|  |  |  |

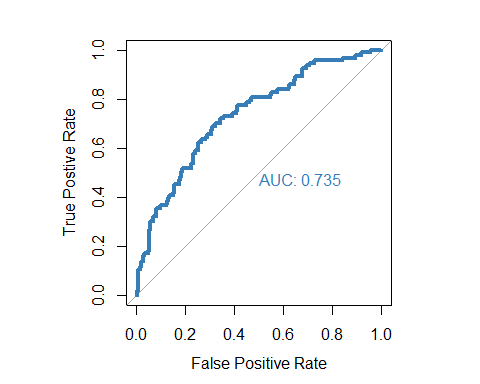
## Misclassification Rate = 0.2885714

## False Positive Rate = 0.3610548  
## False Negative Rate = 0.115942  
## Cost = 0.4257143

1. **Test the out-of-sample performance. Using final logistic linear model built from (i) on the 70% of original data, test with the remaining 30% testing data**

**Out of sample Prediction**

**ROC Curve and AUC**



|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| True | **0** | **1** |
| 0 | **119** | **88** |
| 1 | **21** | **72** |
|  |  |  |

## Misclassification Rate = 0.3633333

## False Positive Rate = 0.4251208

## False Negative Rate = 0.2258065

## Cost = 0.6433333

* 1. **Results and Conclusion**

The model based on step AIC best suited the data. A total of 12 predictor variables significantly influenced the value of the “response” variable.

|  |  |  |
| --- | --- | --- |
|  | AUC value | Misclassification Rate |
| In-sample | **0.847** | **28.85** |
| Out-of-sample | **0.735** | **36.33** |

1. **Predicting Bankruptcy using Logistic Regression Model**

**3.1 Problem Statement and Approach**

The objective of the model is to identify the probability of bankruptcy using the method of logistic regression. The response variable ‘DLRSN’ is the bankruptcy indicator (a value of 1 corresponds to bankruptcy).

The following variables are available in the data set, of which the most significant risk factors are identified using AIC and BIC approach. The model giving the lowest BIC value (in the training data set) is taken as the model of best fit and the prediction accuracy is obtained on the test data set (30% of the given data set).

**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**=CurrentAsset/Current Liability

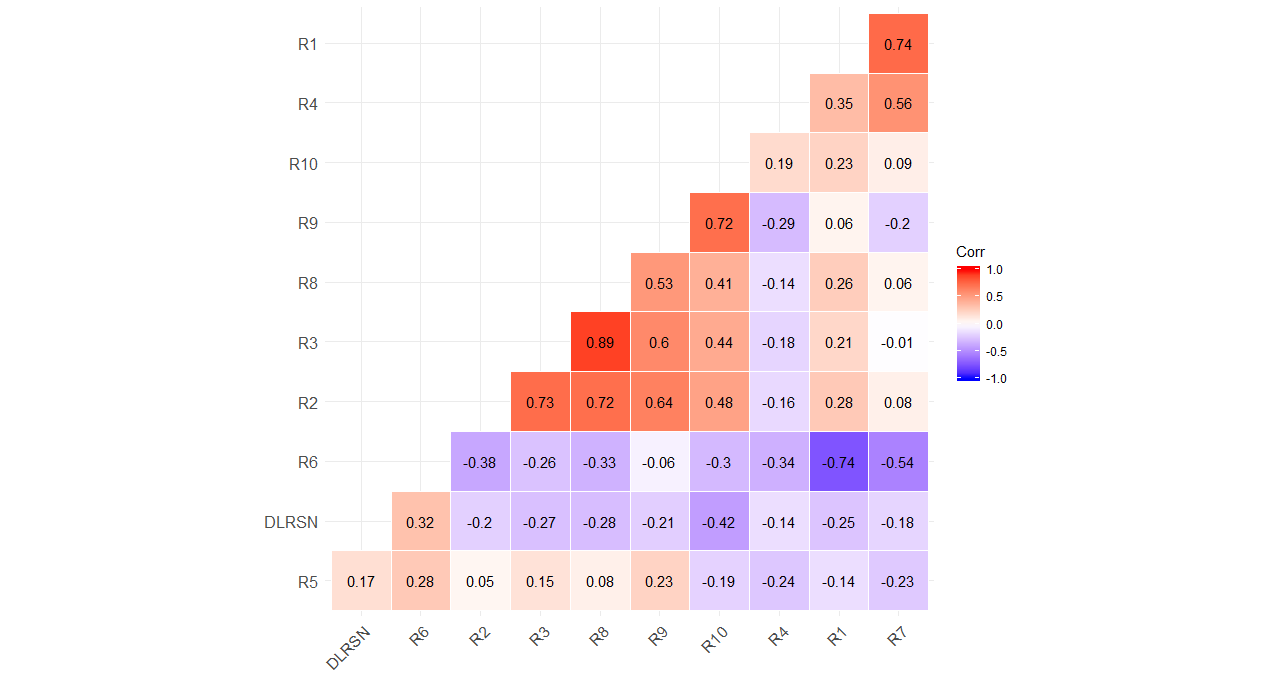
**R8**=Net Income/Total Asset

**R9**=LOG(SALE)

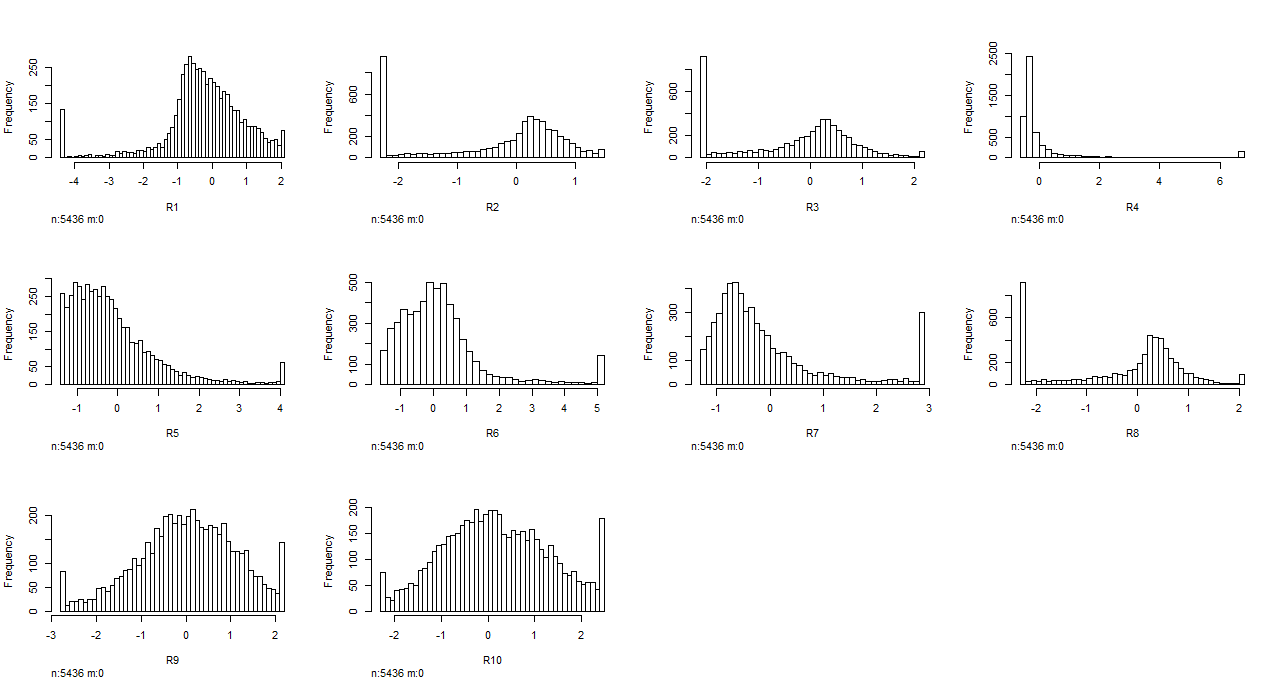
**R10**=LOG(Market Cap)

**3.2 Exploratory Data Analysis**

***Correlation Matrix***



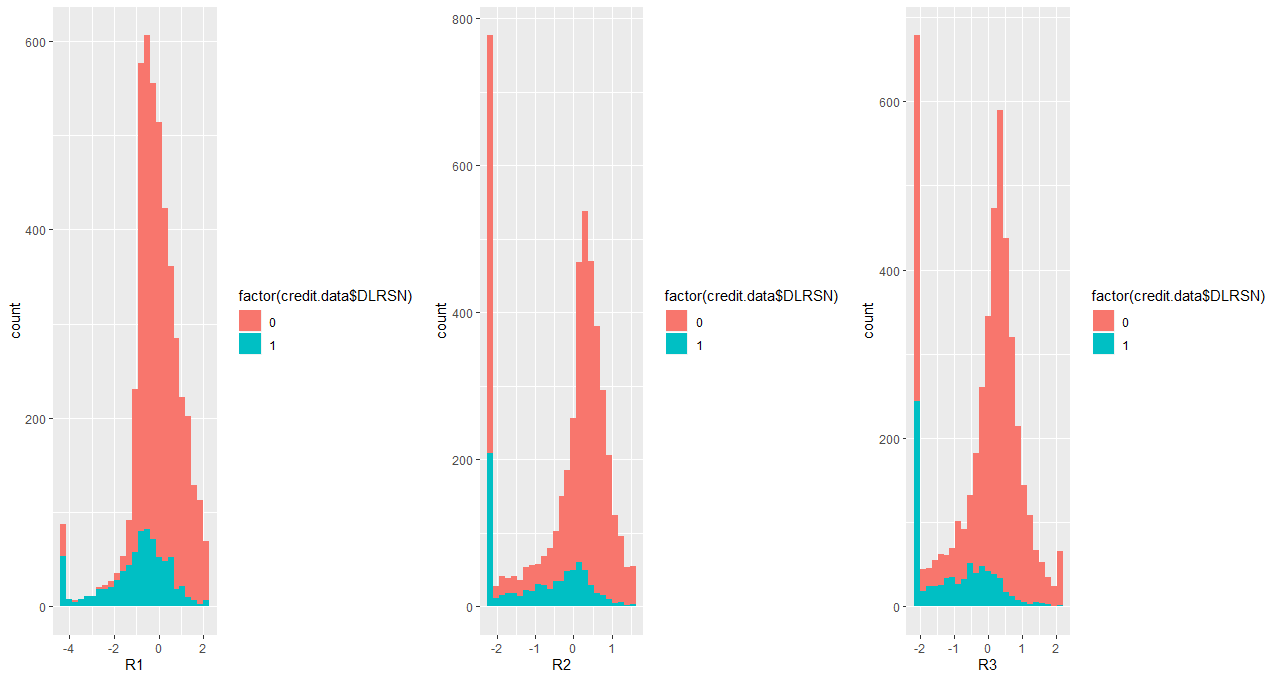
***Understanding the distribution of Numeric Variables***



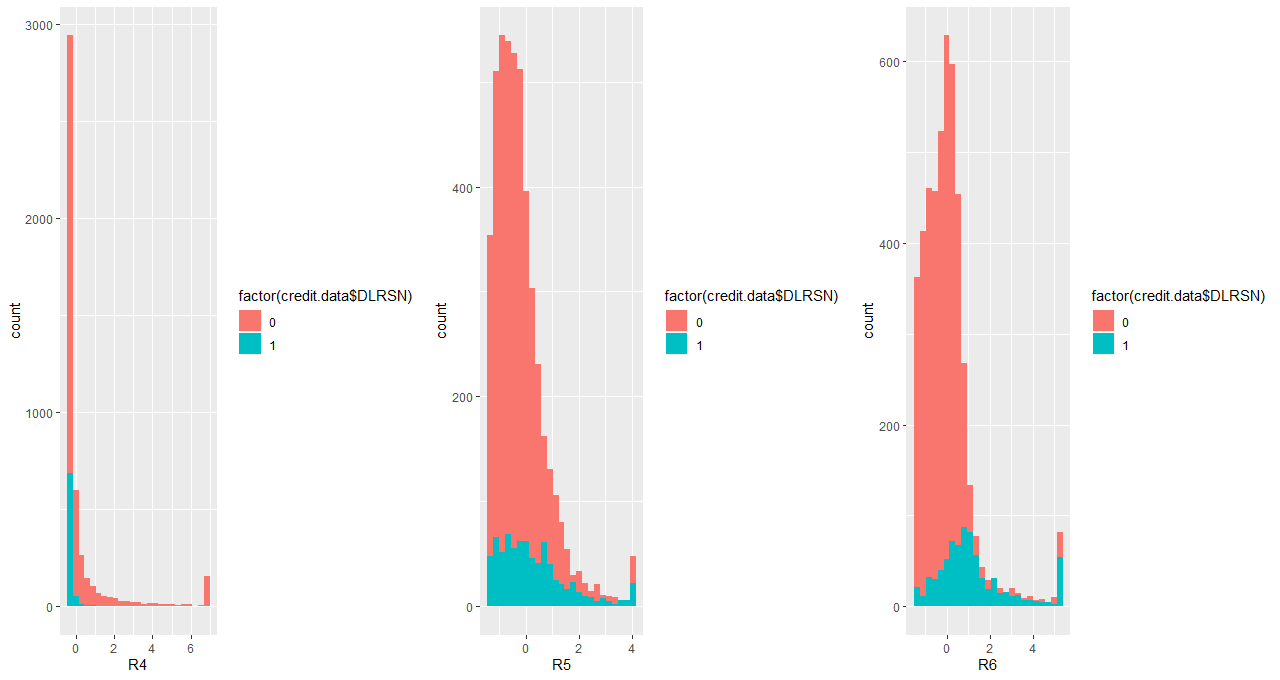
We can observe high skewness in variables R1, R4, R5, R6, R7 and R8

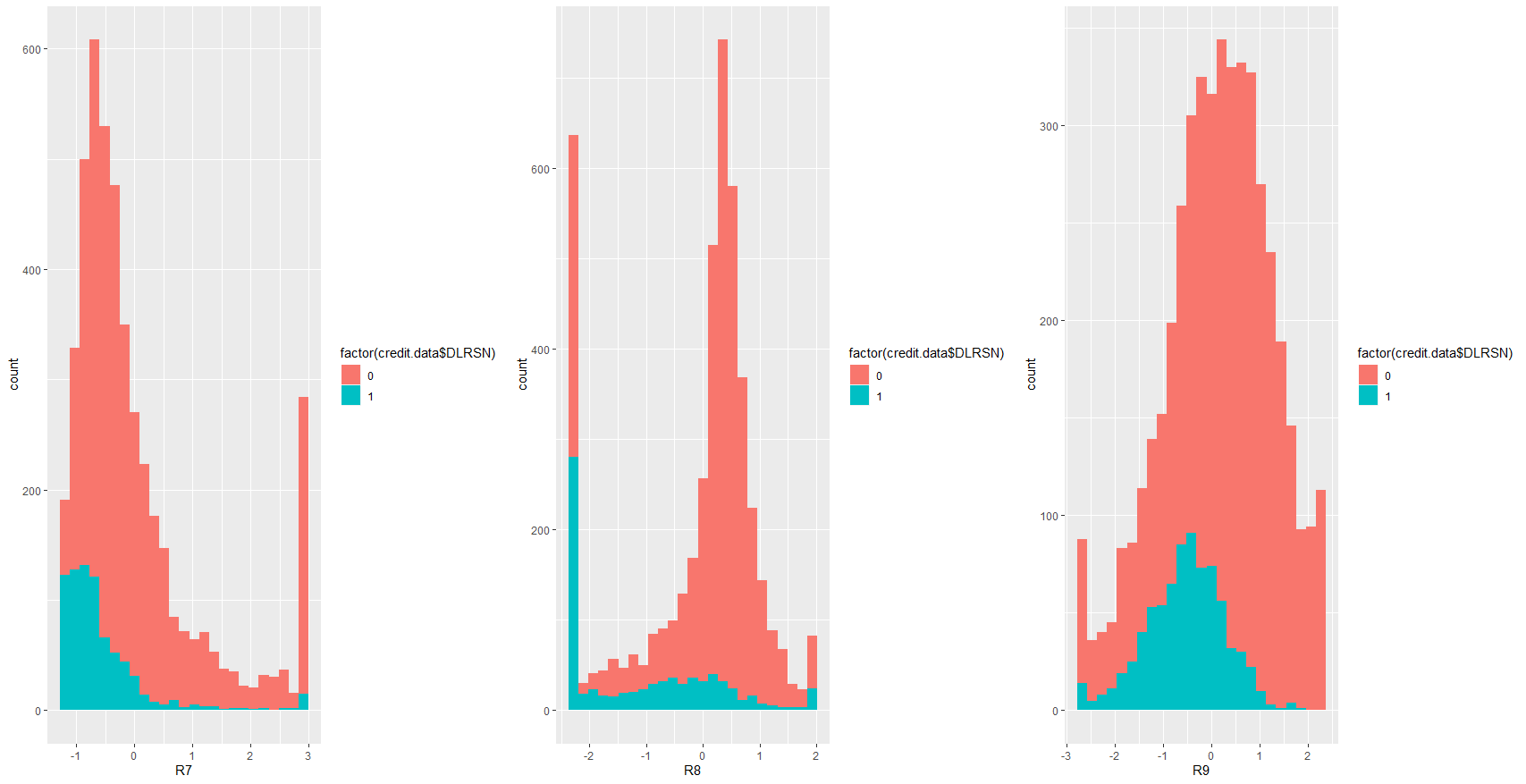
***Univariate Analysis***

Working Capital/Total Asset Retained Earning/Total Asset Earnings Before Interest & Tax/Total Asset

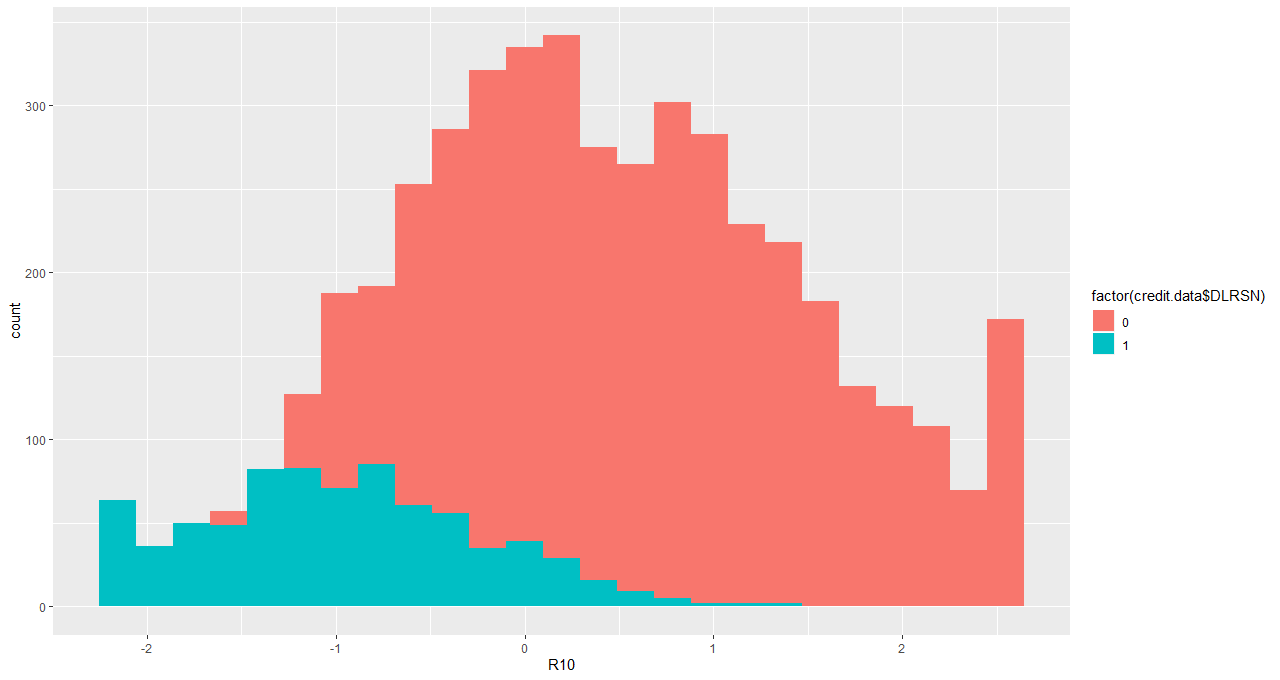


Market Capital/ Total Liability SALE/Total Asset Total Liability/Total Asset



Current Asset/Current Liability Net Income/Total Asset LOG(SALE) 

Log(Market Cap)



From the above analysis we see the bankruptcy indicator distribution to change significantly with the variables R10, R7, R8, R6 and R9.

**3.3 Variable Selection**

***Variable selection using forward step selection and AIC***

A forward step selection was used to find the model with the lowest AIC value. The lowest AIC value of **2147.12** was obtained while using the predictors R10, R7, R8, R2, R6, R9, R3, R4 in the logistic regression model

***Variable selection using forward step selection and BIC***

The approach of selecting the model with lowest BIC, gives a more conservative model with lower prediction variance. Hence a forward step selection was used to obtain the model with the lowest BIC (2154.55) and this model was chosen as the final prediction model. The most significant variables to determine bankruptcy were identified to be **R10, R7, R8, R2, R6, R9, and R3** *(Please see the section 3.1 to view the variable names)*

**3.4 In-sample Prediction**

***Identifying the optimal predicted probability cut off for classification***

A weight of **35:1** was used in the cost function to generate the optimal probability threshold for classifying a given observation. The threshold probability was obtained to be 0.03.

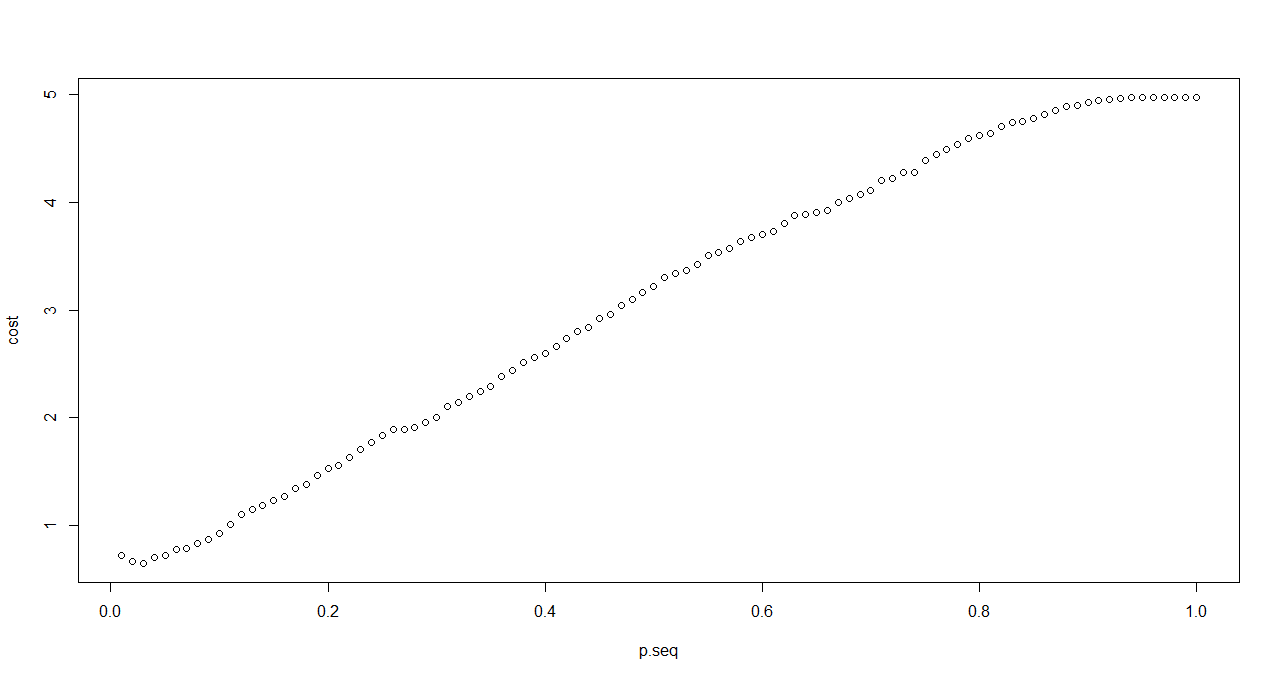


Fig 3.1 Identifying the probability with lowest cost

***ROC Curve and In-sample AUC***

The in-sample AUC was obtained to be 0.7

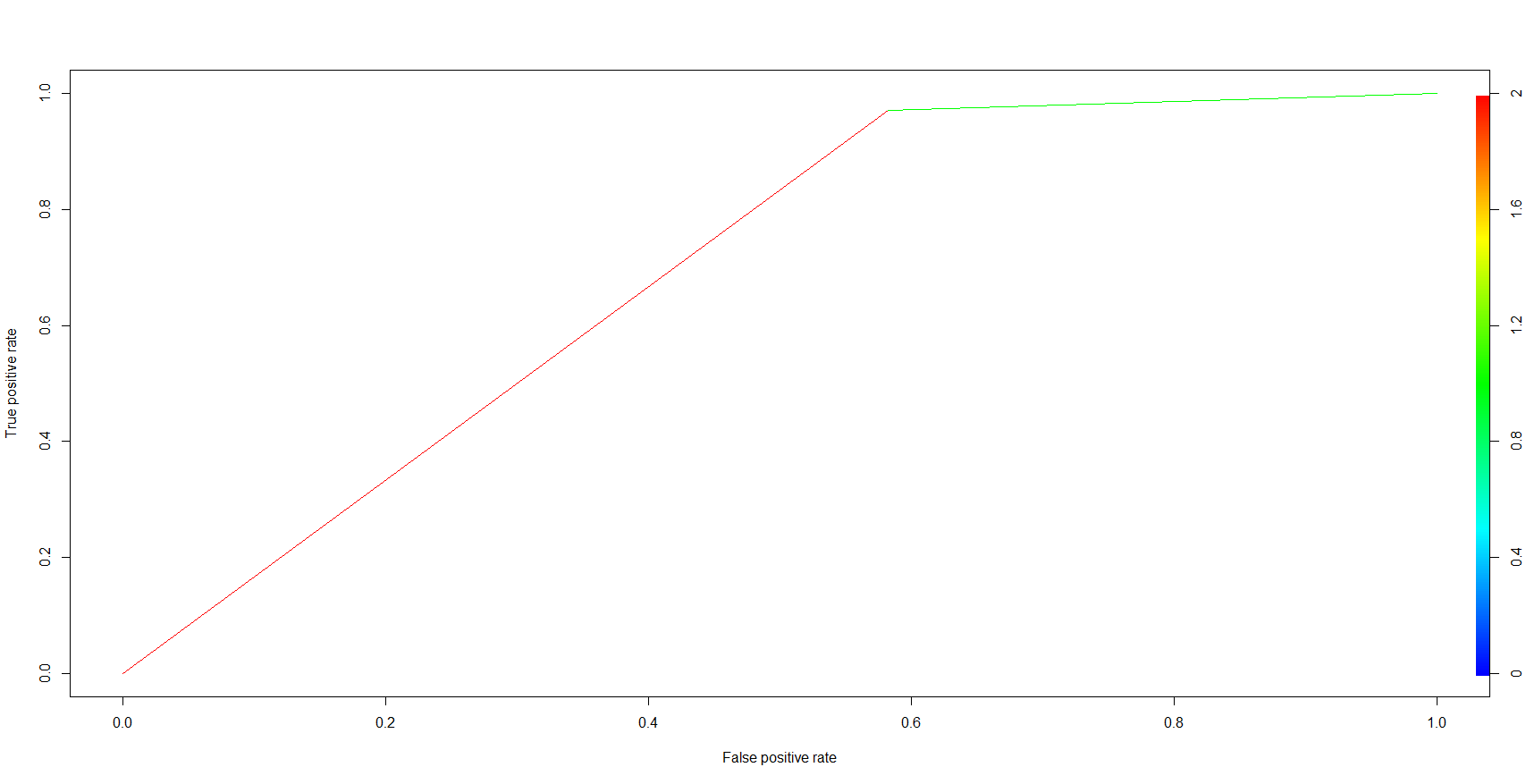


Fig 3.2 ROC curve for training data

***Contingency table and Misclassification Rate***

|  |  |  |
| --- | --- | --- |
|  | **Predicted** | |
| **TRUE** | **0** | **1** |
| **0** | 1364 | 1900 |
| **1** | 16 | 525 |

Misclassification Rate = **0.50**

True Positive Rate = **0.97**

False Positive Rate = **0.58**

**3.4 Out of Sample Prediction**

30% of the bankruptcy data was used as the test data set. The out of sample prediction accuracy gives a clearer picture into the prediction power of the model. A predicted probability threshold of 0.03 was used for classification (according to the optimal value obtained using the training data).

***ROC Curve and out of sample AUC***

The out of sample AUC was obtained to be 0.69

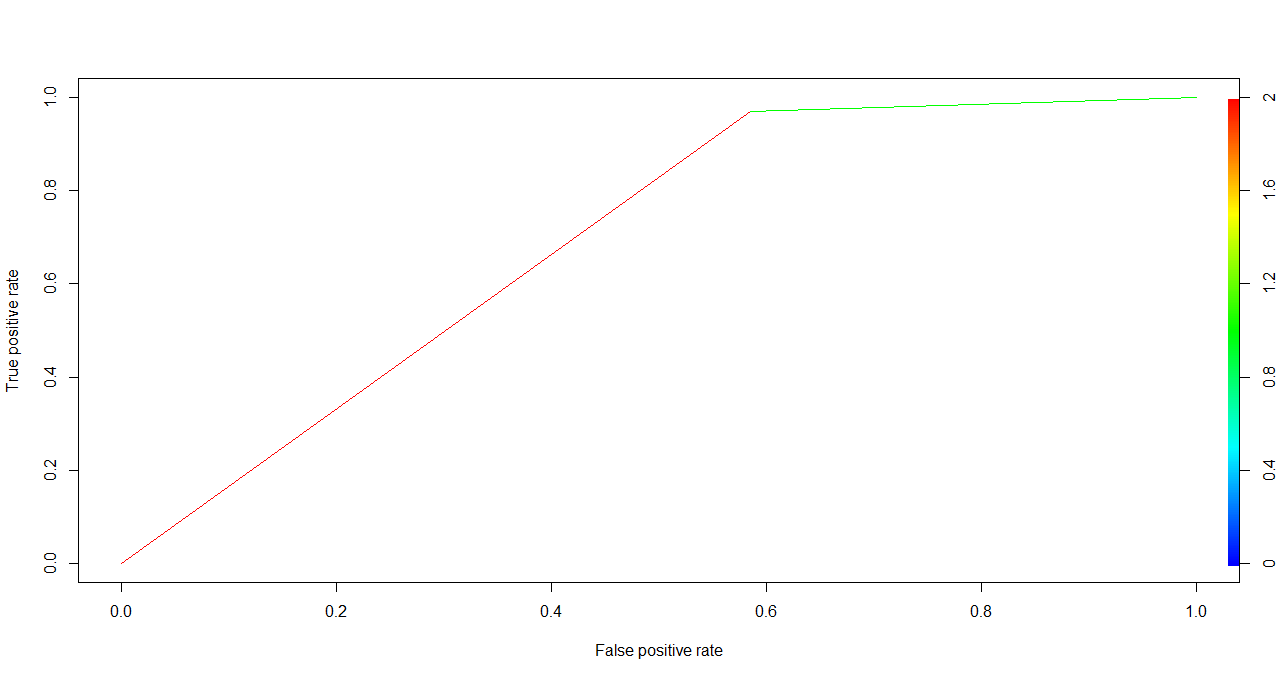


Fig 3.3 ROC Curve for testing data

***Contingency table and Misclassification Rate***

|  |  |  |
| --- | --- | --- |
|  | **Predicted** | |
| **TRUE** | **0** | **1** |
| **0** | 580 | 816 |
| **1** | 7 | 228 |

Misclassification Rate = **0.504**

True Positive Rate = **0.9702**

False Positive Rate = **0.584**

**3.5 Results and Conclusion**

The model based on step BIC was used to predict the bankruptcy. The model gives a very high True Positive Rate based on the assigned weights of the cost function.