**BANA 7046**

**Homework 4**

Presented By:

Anjali Shalimar, Digvijay Kawale, Meenal Narsinghani

**Index:**

**Problem 1: Boston Housing Data**

1. Introduction
2. Exploratory Data Analysis
3. Linear Regression and Residual Analysis
4. Cross Validation
5. CART
6. Repeating the steps for other random data

**Problem 2: Predicting Bankruptcy**

1. Executive Summary
2. Exploratory Data Analysis
3. Generalized Regression Models
4. Classification Trees
5. Comparing Logistic Regression Model and Classification Tree Model
6. Understanding Model Performance by Varying Train/Test Data Size
7. Results and Conclusion

**1.1 Introduction**1

The Boston housing data set contains information collected by the U.S. Census Service concerning housing in the area of Boston. It has been used extensively throughout the literature to benchmark algorithms. The data set consists for 506 observations with 14 attributes. Out of all variables, one of them is binary, one of them is integer and rest 12 are numeric variables.

Using the random 70% of the Boston data set as our training data, we have done the Exploratory Data Analysis. Our final aim is to predict the Median value of homes using statistical methods such as Linear Regression model and regression trees (CART) with this training data. The results will be then compared for these different methods. For Linear Regression we have done the variable selection using methods like best subset, stepwise, and LASSO.

**1.2 Exploratory Data Analysis**

**Crim**: We observe that the mean of the ‘crim’ is greater than its median with a large difference. It is observed form Fig 2.2 that there are a lot of outliers in the quantitative variable crim. This is causing its distribution to be extremely positively skewed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **crim** | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 0.00632 | 0.08232 | 0.23912 | 3.27295 | 3.68567 | 67.92080 |

**Zn:** We observe that the mean of the ‘Zn’ is greater than its median with a large difference. The median itself has a zero value which tells us that 50% of the values are 0. It is observed in Fig 2.4 that there are lot of outliers in the variable ‘Zn’. Fig 2.3 also shows that it has positively skewed distribution.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Zn** | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 0 | 0 | 0 | 11.52 | 19 | 95 |

**Indus:** We observe that the mean of the ‘indus’ is greater than its median value. We find no outliers in the variable ‘indus’ which is observed in Fig 2.6.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **indus** | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 0.46 | 5.16 | 9.69 | 11.07 | 18.1 | 27.74 |

**Nox:** We observe that the mean of the ‘Nox’ is slightly greater than its media. We find no outliers in the variable ‘indus’ which is observed in Fig 2.8.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **nox** | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 0.385 | 0.448 | 0.538 | 0.5547 | 0.631 | 0.871 |

**Rm:** We observe that the mean of the ‘rm’ is almost equal to its median. We observe from figures 2.9 and 2.10 that the variable is normally distributed but there are lot of outliers both the sides.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **rm** | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 3.561 | 5.875 | 6.212 | 6.292 | 6.633 | 8.78 |

**Age:** We observe that the mean of the ‘Age’ is less than its median. Hence, it might have negatively skewed distribution. It is also evident from the Fig 2.11 that it has a negatively skewed distribution. From Fig 2.12 we do not see any outliers in this variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **age** | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 6 | 44.05 | 77.7 | 68.23 | 93.6 | 100 |

**Dis:** We observe that the mean of the ‘dis’ is greater than its median with a large difference. Hence, it might have positively skewed distribution. We observe in Fig 2.14 that there are outliers in the data that are causing its distribution to be positively skewed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dis** | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 1.130 | 2.088 | 3.112 | 3.826 | 5.213 | 12.127 |

**Rad:** We observe that the mean of the ‘rad’ is greater than its median with a large difference. Hence, it might have positively skewed distribution.From Fig 2.16there are no outliers observed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **rad** | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 1 | 4 | 5 | 9.515 | 24 | 24 |

**Tax:** We observe that the mean of the ‘tax’ is greater than its median with a large difference. It is observed from Fig 2.18 that there are no outliers in the variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **tax** | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 187 | 277 | 330 | 408.8 | 666 | 711 |

**Ptratio:** We observe that the mean of the ‘ptratio’ is less than its median. From Fig 2.20 we can observe that there are 2 outliers in the data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ptratio** | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 12.60 | 17.40 | 19.1 | 18.49 | 20.2 | 22 |

**Black:** We observe that the mean of the ‘Black’ is greater than its median with a large difference. From Fig 2.22 we observe that there are lot of outliers in the variable observations.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **black** | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 0.32 | 376.12 | 391.43 | 360.28 | 396.13 | 396.90 |

**Lsat:** We observe that the mean of the ‘Lsat’ is greater than its median. We can see from the Fig 2.23 that the distribution of variables is positively skewed and also from Fig 2.24 we can see that there are few outliers in the data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **lsat** | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 1.920 | 6.715 | 11.25 | 12.555 | 17.105 | 36.980 |

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**Medv**: We observe that the mean of the ‘medv’ is greater than its median. From Fig 2.25 it is observed that variable has a positively skewed distribution. Fig 2.26 tells us that there are lot of outliers in the variable.

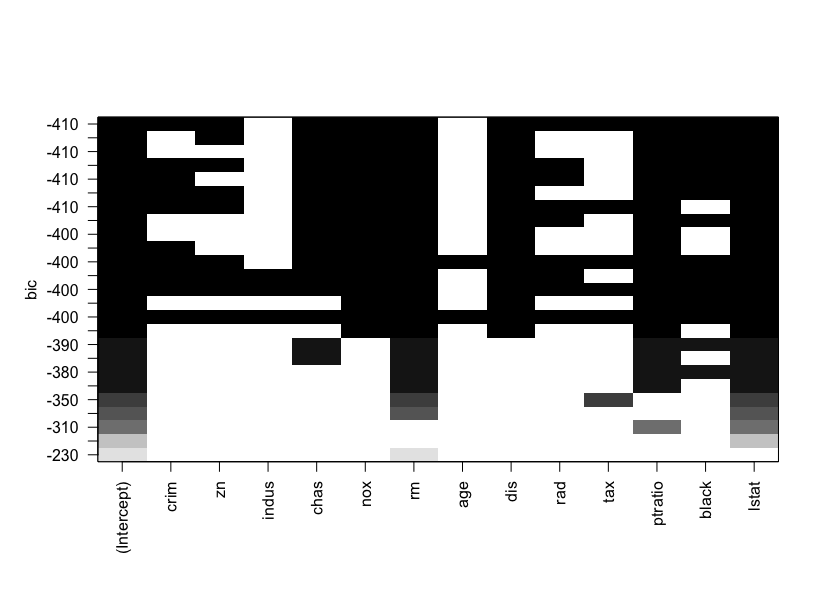
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **medv** | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 5 | 16.8 | 21.1 | 22.66 | 25.15 | 50 |

A screenshot of a cell phone

Description automatically generated

* 1. **Linear Regression and Residual Analysis**:

**Best Subset Method for Variable Selection:** After applying the best subset model for the given data set it gives the results that has number of variables ranging from 2 to 13. A total of 25 best subsets are given in the results by the best subset method. We will compare the values such as adjusted r square values and BIC values to select the best model from these 25 models. The minimum BIC value and maximum value of adjusted r square both are given by the model that has 11 variables which excludes indus and age.



**Fig 3.1 BIC values of Models using Best Subset Method**

**Stepwise Regression Method for Variable Selection:** The stepwise regression consists of iteratively adding and removing predictors, in the predictive model, in order to find the subset of variables in the data set resulting in the best performing model, that is a model that lowers prediction error. Using the stepwise regression method for model selection, the results suggest that the minimum AIC value is observed in the model which has 11 variables that exclude indus and age. The results are aligned with our Best Subset method of variable selection.

**LASSO Method for Variable Selection:** In [statistics](https://en.wikipedia.org/wiki/Statistics) LASSO is a [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) method that performs both [variable selection](https://en.wikipedia.org/wiki/Variable_selection) and [regularization](https://en.wikipedia.org/wiki/Regularization_(mathematics)) in order to enhance the prediction accuracy and interpretability of the [statistical model](https://en.wikipedia.org/wiki/Statistical_model) it produces. We will be using LASSO to select the variables to be used for prediction of housing prices. By taking the minimum value of lambda as its optimal value the LASSO method of variable selection suggests taking 6 variables – crim, chas, rm, ptratio, black and lstat. Fig 3.2 shows the plot log lambda vs the coefficients. Fig 3.3 shows the plot of log lambda vs Mean Squared Error. The dotted lines represent the optimal value of the lambda.

A screenshot of a social media post

Description automatically generated

**Fig 3.2 Plot of Log Lambda vs Coefficients**

A screenshot of a social media post

Description automatically generated

**Fig 3.3 Plot of Log Lambda vs the Mean-squared Error of Models**

We go forward with choosing the model suggested by the LASSO method of variable selection. Below is the table of variable and coefficients.

|  |  |
| --- | --- |
| **Variables** | **Coefficients** |
| Intercept | 9.77 |
| crim | -0.04866 |
| chas | 3.705 |
| rm | 4.97 |
| Ptratio | -0.8864 |
| Black | 0.01 |
| lstat | -0.454 |

**Table 3.1 Variables and Coefficients for the selected model**

**Residual Diagnosis:** The mean squared error of the model 26.4995. The values of R square and adjusted R square are 0.701 and 0.6958 respectively. Fig 3.4 shows that the residuals are randomly distributed. Fig 3.5 shows that the residuals are normally distributed.

**A close up of a mans face

Description automatically generated**

**Fig 3.4 Distribution of Residuals**

**A screenshot of a cell phone

Description automatically generated**

**Fig 3.5 QQ-plot of Residuals**

After testing the out of sample data (remaining 30% testing data) using the selected final model, the mean square error is found to be 24.97. This mean square is error is very close to the in sample mean square error which tells us that the model prediction is good for out of sample data.

**1.4 Cross Validation:**

Cross validation1 is an alternative approach to training/testing split. For k-fold cross validation, the dataset is divided into k parts. Each part serves as the test set in each iteration and the rest serve as training set. The out-of-sample performance measures from the k iterations are averaged.

We have used 3-fold cross validation to check the performance of the model. The function used to do this was cv.glm() in R. The function has mean square error as its default cost function. After applying it on the Boston data set, it gives a mean square error of 28.614. The residual mean square error when calculated directly for the out of sample Boston testing data was 24.97. The mean square errors using both the methods are very close. The mean square error of in sample data was 26.4995. This is closer to the value yielded by cross validation than the out of sample data value.

**1.5 CART:**

CART2 is classification and regression trees. We will be using regression tree to predict the housing prices using the Boston data set. CART was performed in R using the rpart function in the rpart library which directly gives us the regression tree. Fig 5.1 shows the results and the regression tree using the Boston training data.

Using the below regression tree, we will predict the housing prices for the in-sample data and the out of sample data. It is observed that using the regression tree the in sample mean square error is 18.75 and the out of sample mean square error is 13.99. It tells us that the predictions using the CART method are better than the predictions done linear regression model.

A close up of a map

Description automatically generated

**Fig 5.1 Regression Tree using the CART method to predict housing prices**

* 1. **Repeating the steps for other random data**2**:**

Initially the same regression model was fit (using the 6 variables) for the new random data set. The in-sample MSE for the model is 26.44 compared to the in-sample MSE of 26.49 yielded by the earlier data. The out of sample MSE for the model is 27.48 compared to the out of sample MSE of 24.97 yielded by the earlier data. The 3-fold cross validation is going to give the same results always as it used the whole data set for the model.

The regression tree approach gives a completely different tree after changing the data. The in-sample MSE for the model is 16.99 compared to the in-sample MSE of 18.75 yielded by the earlier data. The out of sample MSE for the model is 20.75 compared to the out of sample MSE of 13.99 yielded by the earlier data.

The results are almost consistent after changing the training and testing data in both the methods – Linear regression and CART. The predictions are better using the CART method when compared with the predictions using the Linear Regression Approach.

**References:** 1. <https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html>

2. <https://xiaoruizhu.github.io/Data-Mining-R/lecture/5_Tree.html>

3. <https://xiaoruizhu.github.io/Data-Mining-R/lecture/3_LinearReg.html>

1. **Predicting Bankruptcy**

**2.1 Executive Summary**4

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, BIC and LASSO 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). Different link functions are considered while building the model. Classification tree is also used to test the prediction accuracy for 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)*

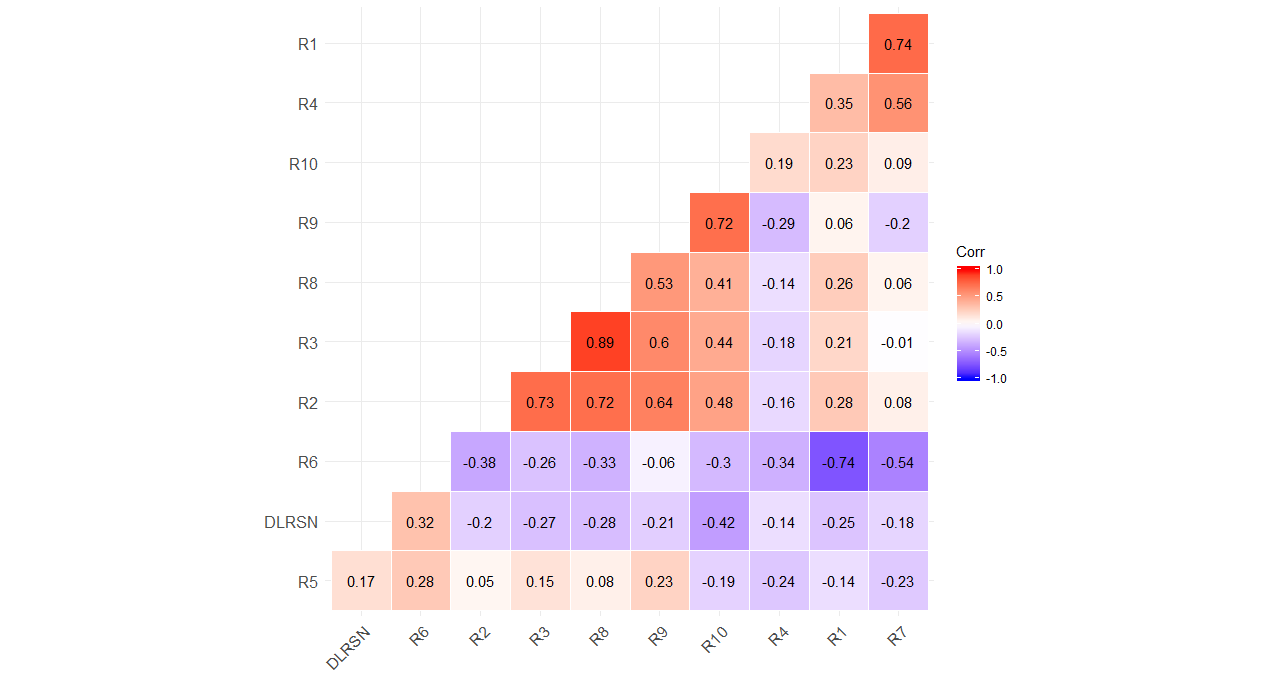
***R10****=LOG(Market Cap)*

The results obtained from classification tree and logistic regression model are compared to identify the best approach for the given split of the data.

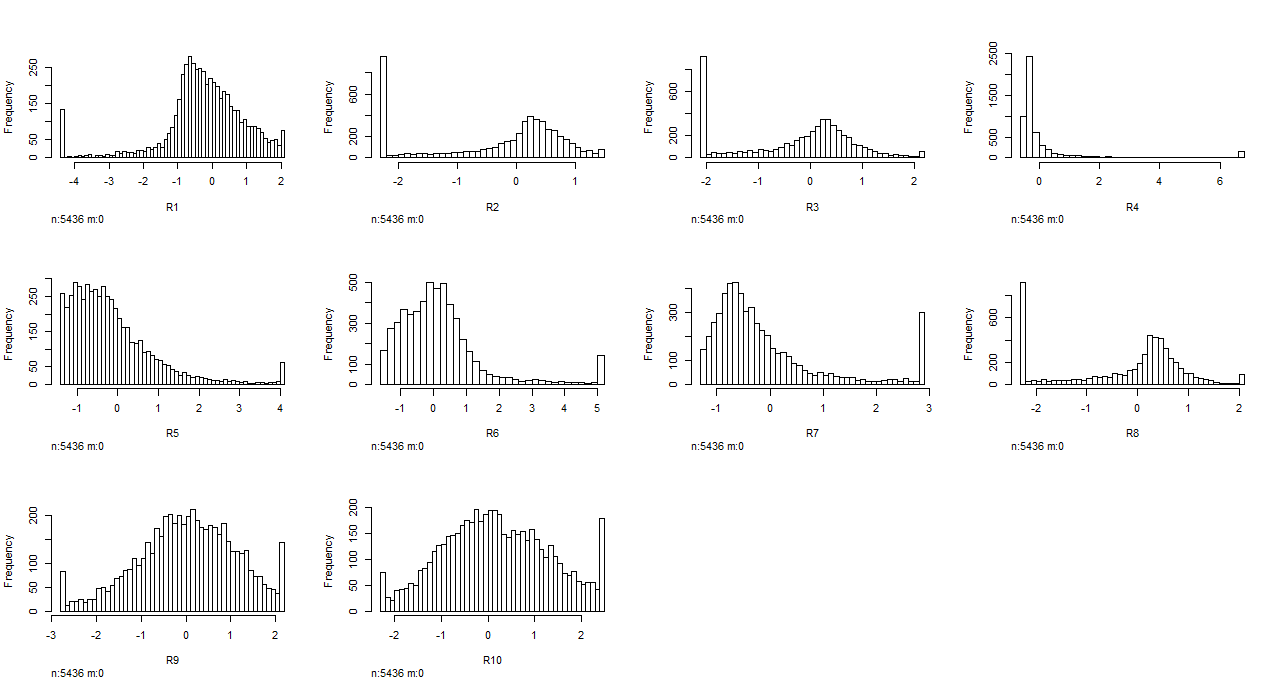


**2.2 Exploratory Data Analysis**

***Correlation Matrix***



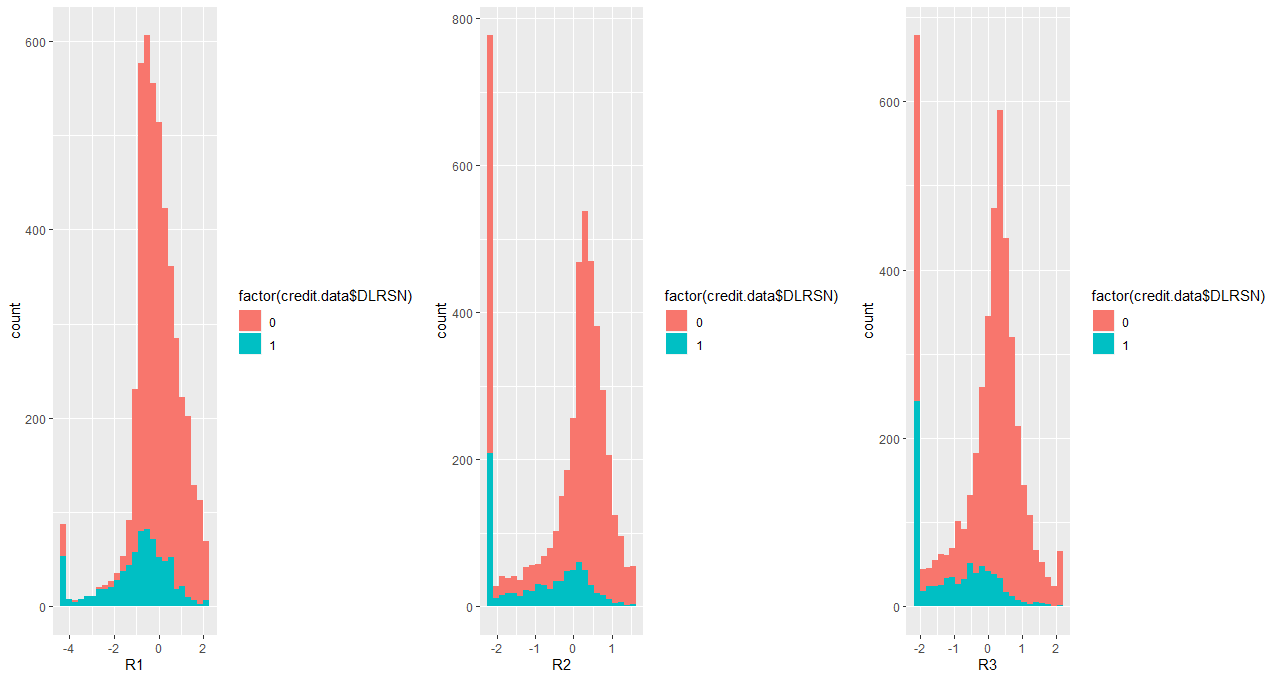
***Understanding the distribution of Numeric Variables***



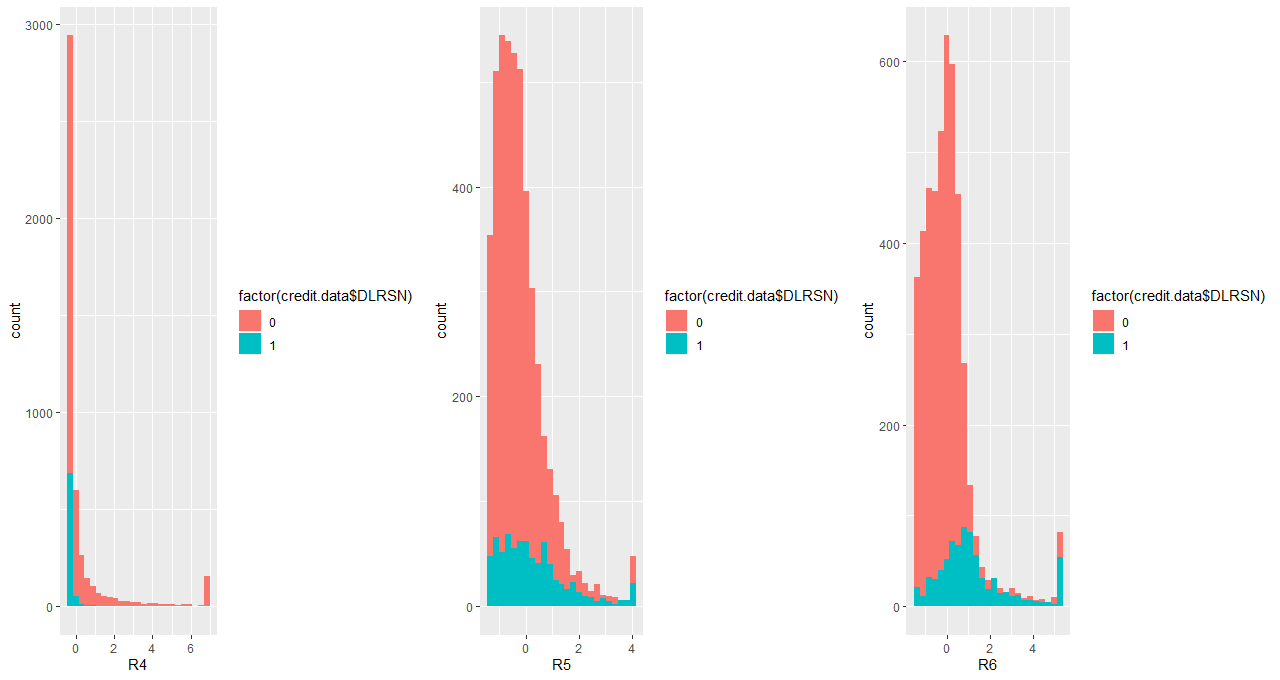
We can observe high skewness in variables R1, R4, R5, R6, R7 and R8

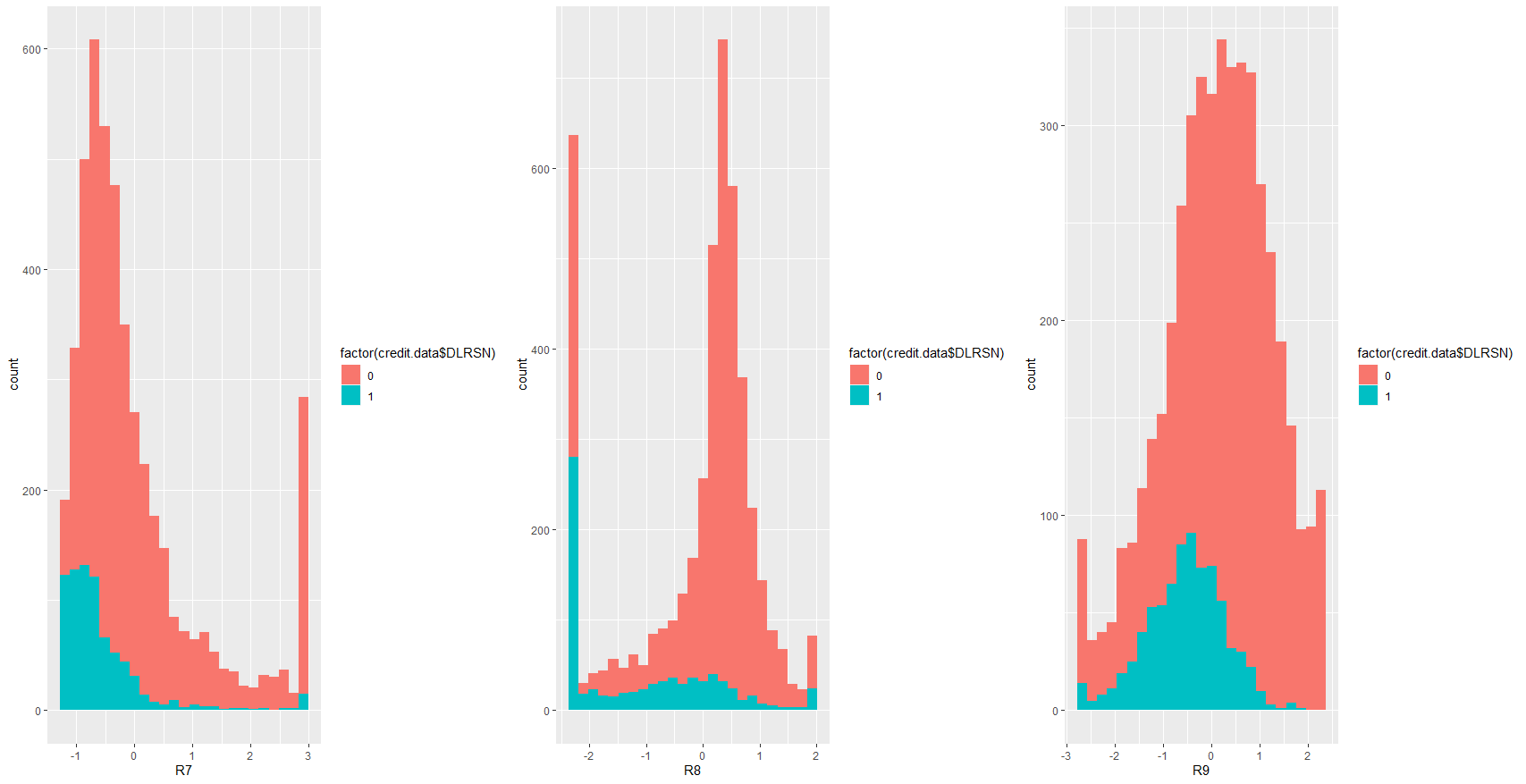
***Univariate Analysis***

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

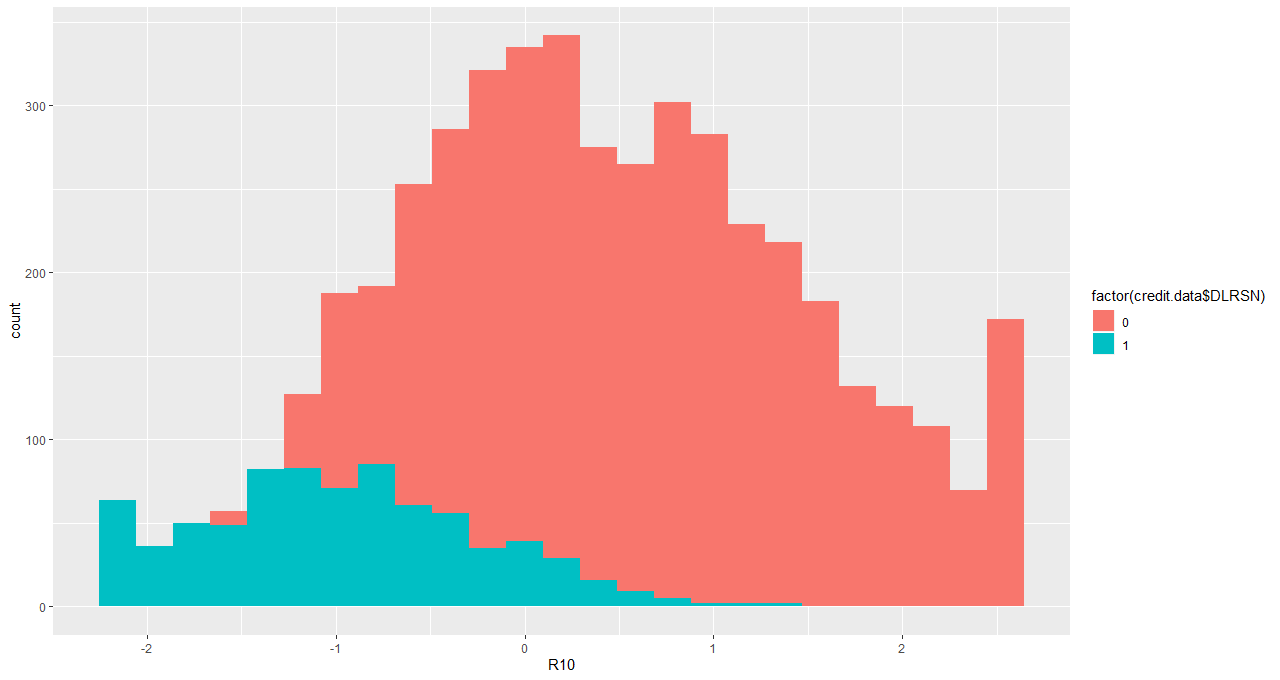


Market Capital/ Total Liability SALE/Total Asset Total Liability/Total Asset



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.

**2.3 Generalized Regression Models**6

Generalized regression models (based on the link functions such as Probit, Logit, complementary log-log) were built on the train data set. The model which corresponds to the best prediction performance was identified using BIC value, AUC, misclassification rate and mean residual deviance on the training data.

***Comparing model performance indicators for various link functions (on training data)***

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **AUC** | **Misclassification Rate** | **Residual Deviance** |
| Logit | 0.651 | 0.50 | 2155.300 |
| Probit | 0.701 | 0.490 | 2160.600 |
| Comp log-log | 0.668 | 0.679 | 2183.800 |

It was observed that the probit link function produces the highest in sample AUC and lowest misclassification rate in the training data set.

***Variable selection for logistic regression model using forward step selection and AIC***

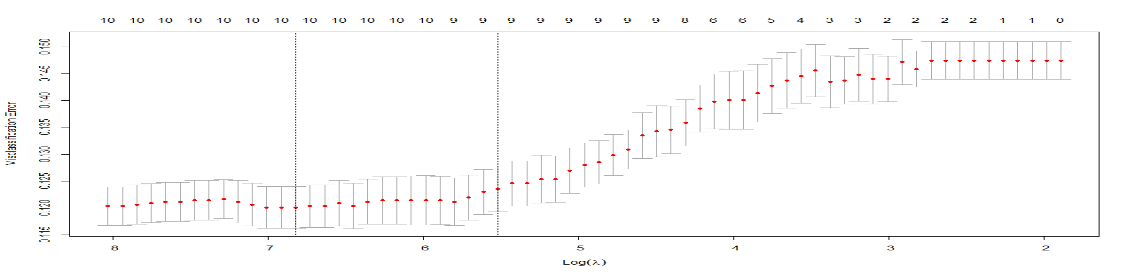
A forward step selection was used to find the model with the lowest AIC value. The lowest AIC value of **2147.12** was obtained while using the predictors R10, R7, R8, R2, R6, R9, R3, R4 in the logistic regression model

***Variable selection for logistic regression model 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)*

***Logistic regression model performance using LASSO***

Lasso model performance was identified for the value of lambda corresponding to the lowest misclassification rate in the training data. Corresponding to this value of lambda the model performance indicators were noted on the training data. In sample AUC was obtained to be 0.5851472 and in sample misclassification rate was 0.704825



*Figure 2.1*

***Selecting the Best Logistic Regression Model***

The model having the lowest BIC value was selected as the final prediction model as this gives this most conservative model.

***In-Sample Performance of the Selected Model***

***Grid Search Algorithm: 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.

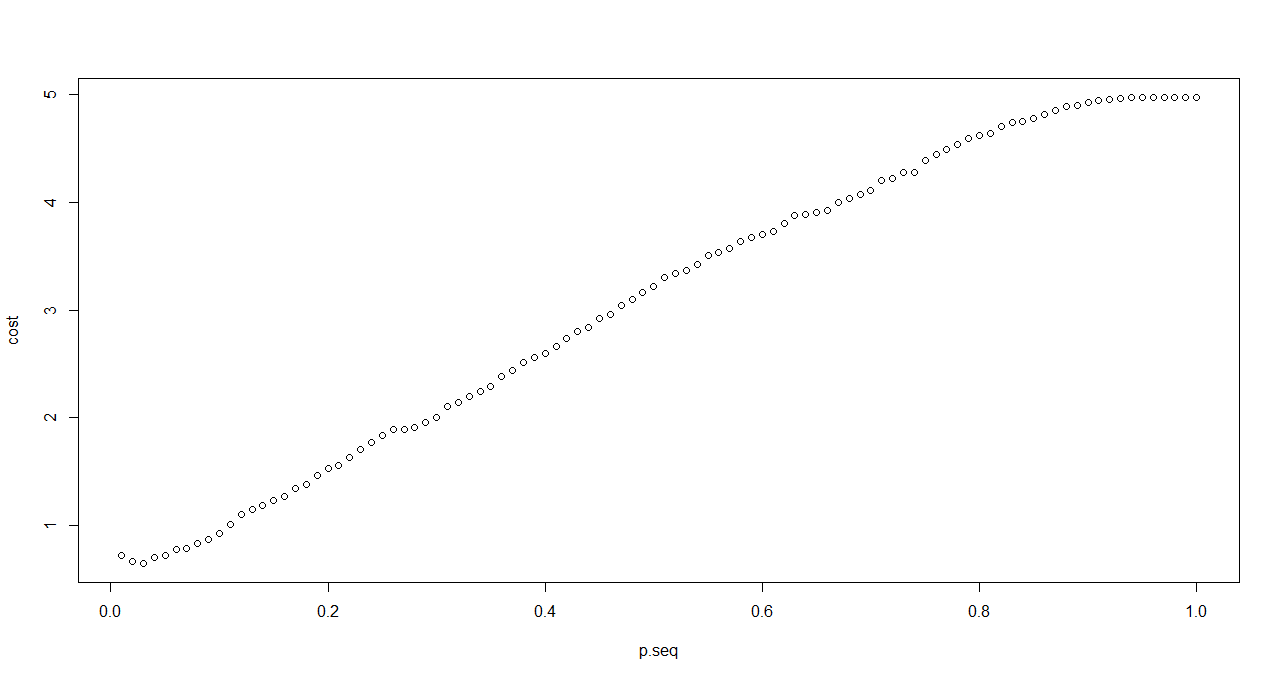


Fig 2.2 Identifying the probability with lowest cost

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

The in-sample AUC was obtained to be 0.7

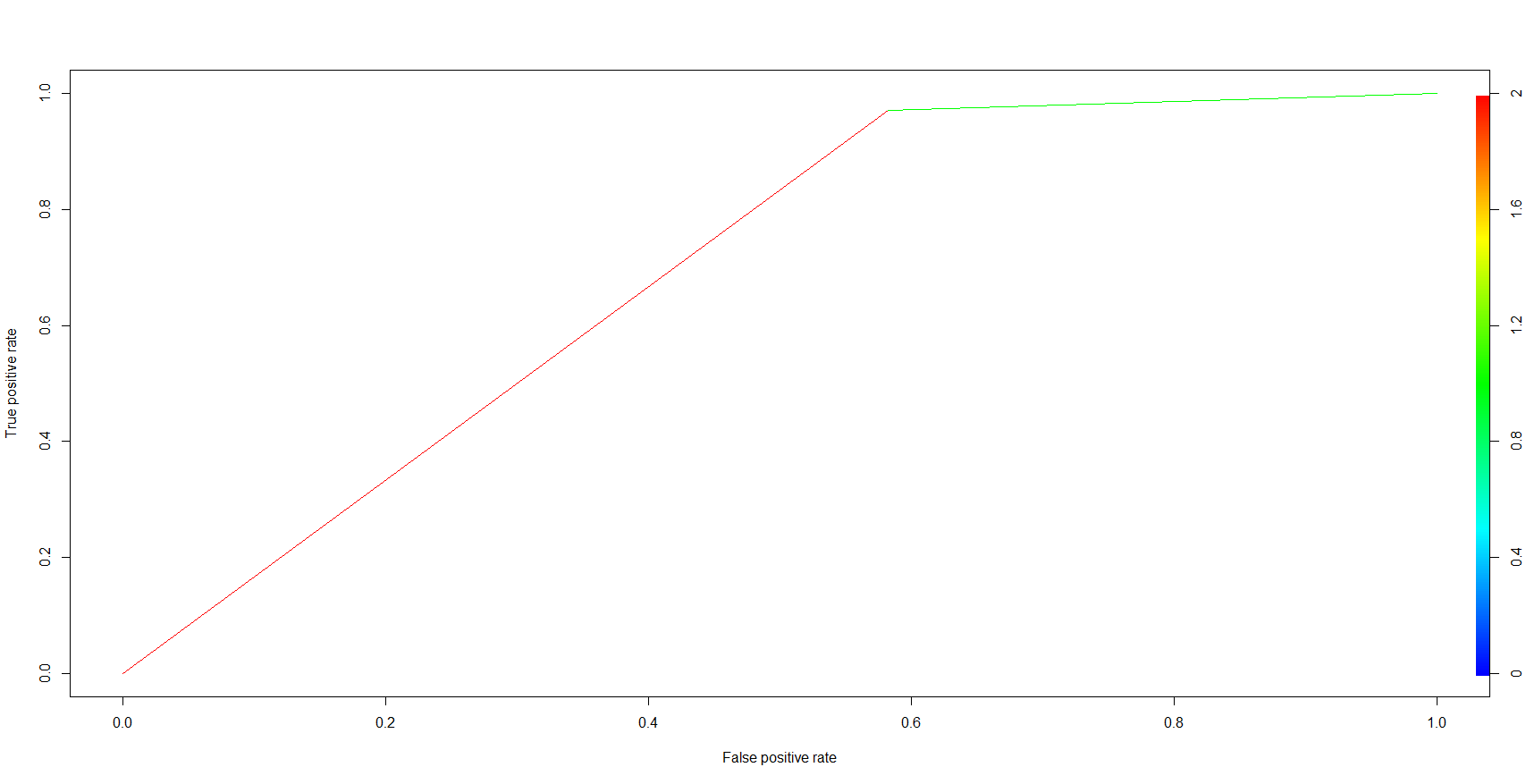


Fig 2.3 ROC curve for training data

***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**

***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

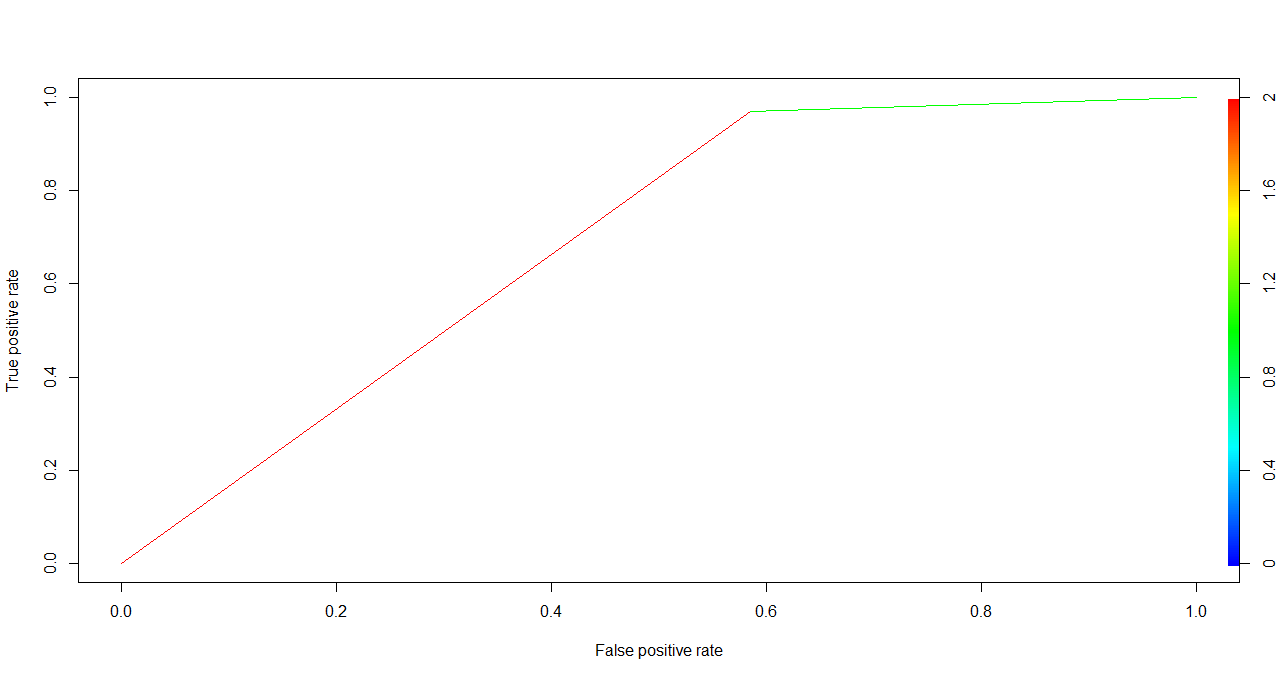


Fig 2.4 ROC Curve for testing data

***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**

***Cross Validation on 100% data***

|  |  |  |
| --- | --- | --- |
|  | **Asymmetric Misclassification Rate** | **AUC** |
| **Adj.CV Score** | 0.501 | 0.69 |

**2.4 Classification Trees** 5

A classification tree was built on the training dataset (70% data). The diagonal elements of the loss matrix were specified as (35,1) like the cost function weights used in the generalized linear regression model.

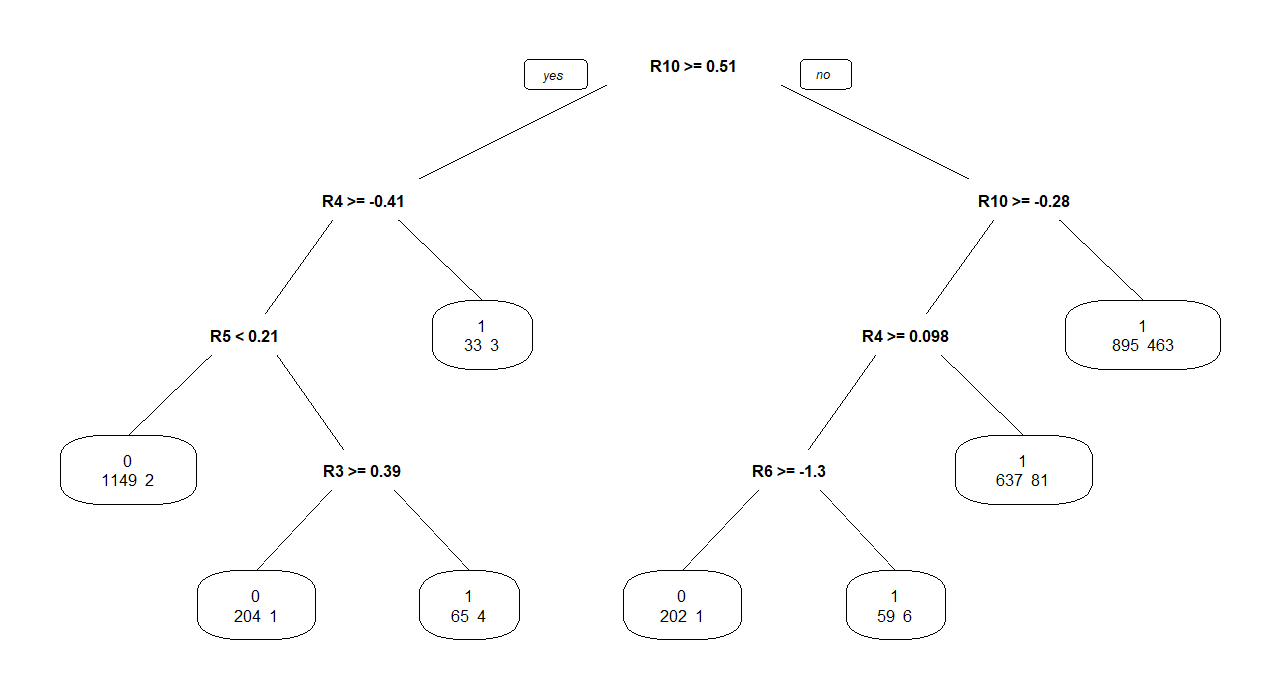
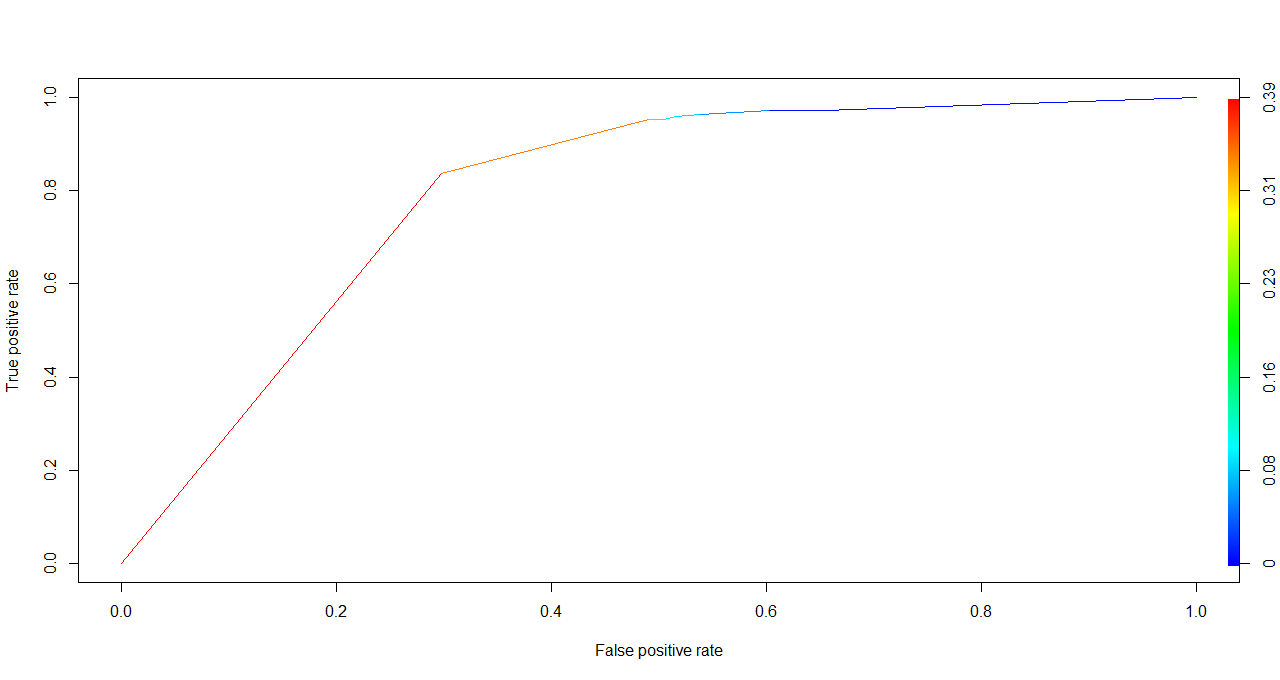


Fig 2.5 Classification Tree

***Out of Sample Prediction Performance***

***AUC***

Out of sample AUC was obtained to be 0.7965478



***Contingency table and Misclassification Rate***

|  |  |  |
| --- | --- | --- |
|  | **Predicted** | |
| **TRUE** | **0** | **1** |
| **0** | 653 | 763 |
| **1** | 8 | 207 |

Misclassification Rate = **0.4727161**

**2.5 Comparing Logistic Regression Model and Classification Tree Model**



For the given split of training and test data, the classification tree performs better in predicting bankruptcy.

**2.6 Understanding Model Performance by Varying Train/Test Data Size**

The given data was split into training set (80%) and testing set (20%) using the split ratio of 80 : 20.The above procedure was repeated to understand the performance of the logistic regression models and classification trees using the new split. The best AIC model now has additional variables, though the best BIC model remains the same. The optimal cut off value for the BIC model remains the same (0.02). The out of sample AUC value slightly decreases for the logistic regression model (0.651) with a slight increase in the misclassification rate. A similar performance in observed in the classification tree where the AUC slightly decreases with a corresponding increase in out of sample misclassification rate. We also observe that the classification tree slightly overfits the data.

**2.7 Results and Conclusion**

The logistic regression 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. The performance was compared with that of a classification tree model. It was observed that the classification tree had a better prediction performance compared to the logistic regression model for the given split of the test data and the given weights on the cost function.

**References:** 4. UC Course Files – Data Mining (BANA 7046 -001 )

5. <https://xiaoruizhu.github.io/Data-Mining-R/lecture/5_Tree.html>

6. <https://xiaoruizhu.github.io/Data-Mining-R/lecture/4_LogisticReg.html>