**Executive Summary**

The objective of the study on bankruptcy data was to classify the given data and predict bankruptcy. The Bankruptcy data was collected from COMPUSTAT for the years 1980 to 2000 and has 5436 observations with 13 variables. For the study, as there was no clear trend in bankruptcy, it was assumed that the data across the years can be pooled together and studied.

Of the 13 variables, one of them was “DLRSN”- a categorical variable indicating default, the dependent variable of the prediction. Overall, the bankruptcy is about 14% of the entire sample. After the initial EDA was performed, the data was separated randomly into test and train datasets with 20-80 split. The train dataset was used to create the logistic model. Variable selection was performed using stepwise and LASSO and AIC, BIC and AUC were the criteria used to measure model adequacy and the final model was based on BIC as it more aggressively prefers less complex models with minimal compromise on model selection metrics.

The final model on the train data was :

Logit(DLRSN)= **-2.58 + 0.25\*R1 + 0.59\*R2 - 0.38\*R3 – 0.43\*R4 + 0.31R6 - 0.45\*R7- 0.37\*R8 +0.33\*R9 -1.34\*R10**

Once the model was built, the accuracy was tested using ROC Curve, AUC and misclassification rate. There were two types of validation performed on the model, out of sample validation and k- folds cross validation. Also, initially an optimal p cut off was obtained using a symmetric cost function and then weight of 15 was added to penalize the False negative responses which brought down the pcut off down to 0.06.

Also, a general threshold that is applied to AUC is 0.7, above which, a model is considered to have a good explanatory power, and the model here had a good AUC.

The comparative statistics between the train-test with AUC and costs are shown below:

|  |  |  |
| --- | --- | --- |
| **Cut off probability =0.06** | **AUC** | **AMR** |
| In-sample (training data) | 0.88 | 0.35 |
| Out-of-sample (test data) | 0.86 | 0.35 |

As evidenced, the misclassification rates and AUC dropped slightly from the train to the test data as the actual model was built only on the train data.

5-fold clustering was performed followed by this and the mean residual deviance was obtained around 0.49 of the complete data set. A second model was created using a CART classification tree and the AUC value was calculated as a part of out of sample validation. The AUC value obtained from CART was lesser than that obtained from logistic regression thus we concluded that logistic using a logit link will better predict the bankruptcy companies in the given data set.

**Exploratory Data Analysis**

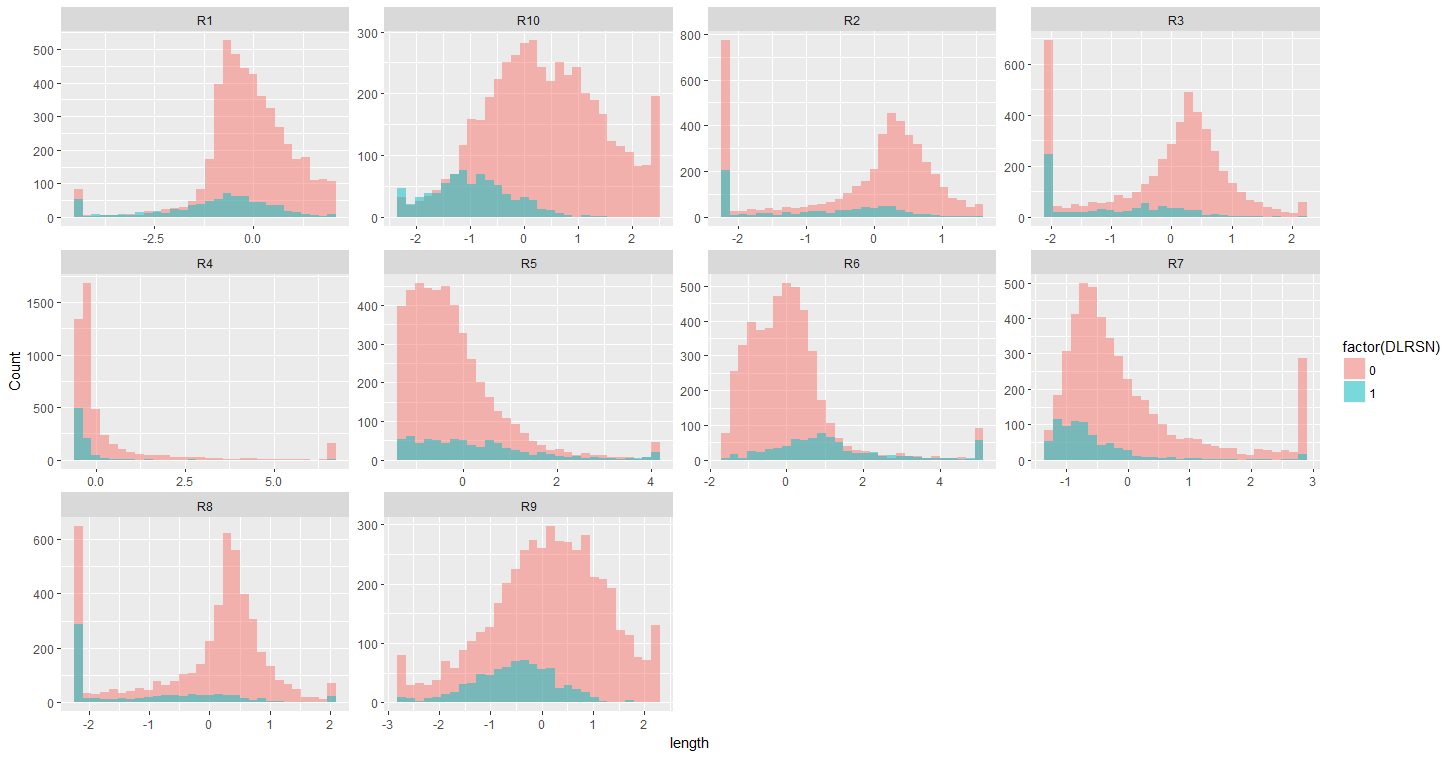
The data was initially read from a csv file and it contained 5436 rows with 13 variables. The dependent variable was a categorical variable, with 1 capturing bankruptcy. The dependent variables were:

* 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 **(Negative)**
* R7=Current Asset/Current Liability;
* R8=Net Income/Total Asset;
* R9=LOG(SALE);
* R10=LOG(Market Cap)

The variables left out were CUSIP, which is an ID, and Financial year, which was ignored for the study as it was seen that there is no clear trend of bankruptcies across years.

The other explanatory variables had the below summary statistics and distributions, with no missing values:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Stats** | **R1** | **R2** | **R3** | **R4** | **R5** | **R6** | **R7** | **R8** | **R9** | **R10** |
| Min | -4.38 | -2.24 | -2.06 | -0.42 | -1.36 | -1.5 | -1.23 | -2.2 | -2.76 | -2.2 |
| 1st Q | -0.75 | -1.08 | -1.07 | -0.38 | -0.87 | -0.65 | -0.77 | -1 | -0.66 | -0.6 |
| Median | -0.22 | 0.13 | 0.06 | -0.3 | -0.35 | 0 | -0.43 | 0.2 | 0.06 | 0.1 |
| Mean | -0.23 | -0.29 | -0.24 | 0.23 | -0.13 | 0.19 | -0.09 | -0.2 | 0.02 | 0.1 |
| 3rd Q | 0.48 | 0.51 | 0.5 | 0.02 | 0.28 | 0.6 | 0.17 | 0.5 | 0.81 | 0.9 |
| Max | 2.02 | 1.48 | 2.14 | 6.69 | 4.03 | 5.11 | 2.87 | 2 | 2.17 | 2.4 |



Of these, R2, R3, R4 and R8 have distributions deviating from normality.

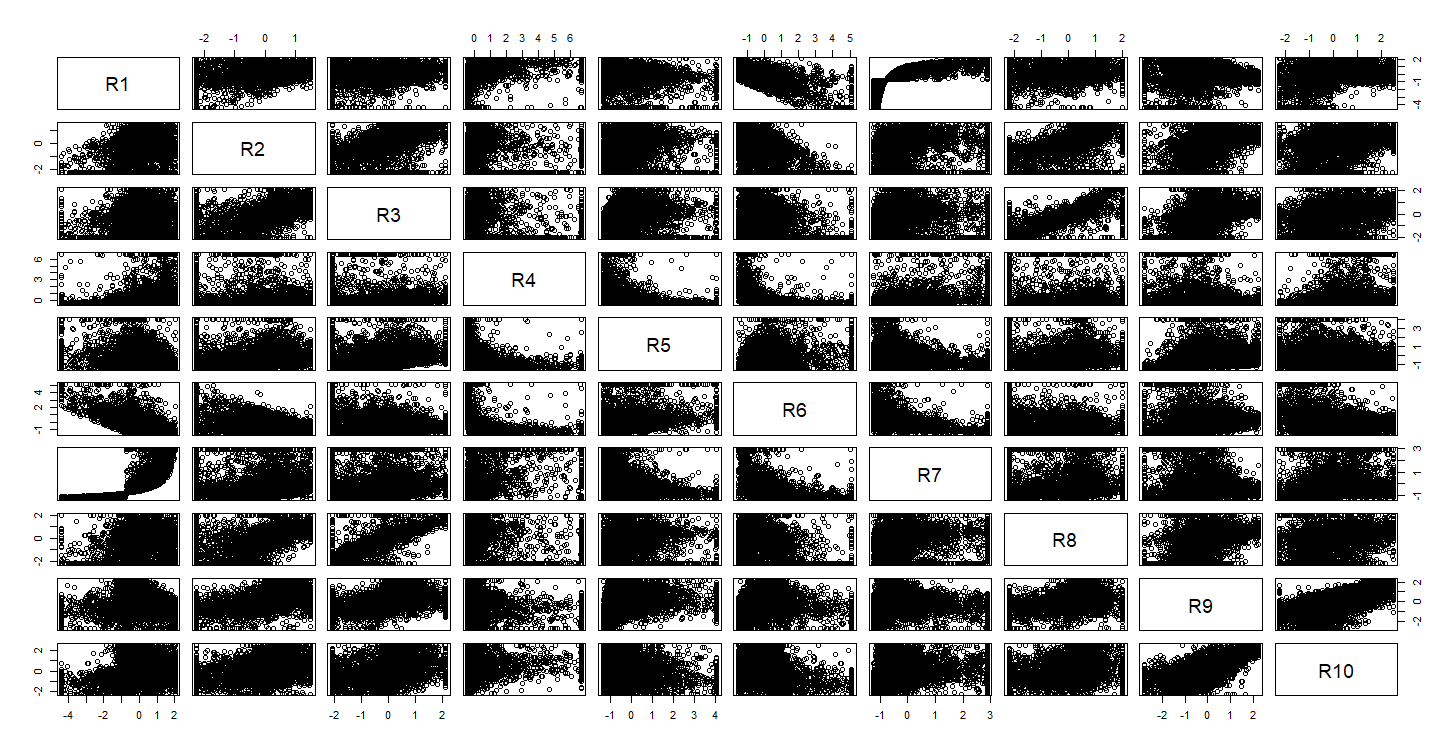
Then the correlations were explored between the covariates through correlation matrix and paired correlation plot as below.

The below highlighted variable pairs were observed to have high correlation:

R1-R7-**R6**(Negative), R2-R3- R8 and R9-R10.

The weakest correlation with bankruptcy was seen with R4 and R5.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Correlation** | **DLRSN** | **R1** | **R2** | **R3** | **R4** | **R5** | **R6** | **R7** | **R8** | **R9** | **R10** |
|  |  |  |  |  |  |  |  |  |  |  |  |
| **DLRSN** | 1.00 | -0.25 | -0.20 | -0.27 | -0.14 | 0.17 | 0.32 | -0.18 | -0.28 | -0.21 | -0.42 |
| **R1** | -0.25 | 1.00 | 0.28 | 0.21 | 0.35 | -0.14 | -0.74 | 0.74 | 0.26 | 0.06 | 0.23 |
| **R2** | -0.20 | 0.28 | 1.00 | 0.73 | -0.16 | 0.05 | -0.38 | 0.08 | 0.72 | 0.64 | 0.48 |
| **R3** | -0.27 | 0.21 | 0.73 | 1.00 | -0.18 | 0.15 | -0.26 | -0.01 | 0.89 | 0.60 | 0.44 |
| **R4** | -0.14 | 0.35 | -0.16 | -0.18 | 1.00 | -0.24 | -0.34 | 0.56 | -0.14 | -0.29 | 0.19 |
| **R5** | 0.17 | -0.14 | 0.05 | 0.15 | -0.24 | 1.00 | 0.28 | -0.23 | 0.08 | 0.23 | -0.19 |
| **R6** | 0.32 | -0.74 | -0.38 | -0.26 | -0.34 | 0.28 | 1.00 | -0.54 | -0.33 | -0.06 | -0.30 |
| **R7** | -0.18 | 0.74 | 0.08 | -0.01 | 0.56 | -0.23 | -0.54 | 1.00 | 0.06 | -0.20 | 0.09 |
| **R8** | -0.28 | 0.26 | 0.72 | 0.89 | -0.14 | 0.08 | -0.33 | 0.06 | 1.00 | 0.53 | 0.41 |
| **R9** | -0.21 | 0.06 | 0.64 | 0.60 | -0.29 | 0.23 | -0.06 | -0.20 | 0.53 | 1.00 | 0.72 |
| **R10** | -0.42 | 0.23 | 0.48 | 0.44 | 0.19 | -0.19 | -0.30 | 0.09 | 0.41 | 0.72 | 1.00 |



**Model Building**

For the model building stage, the data was split for training and testing with 80-20 split. The initial model was run on all the covariates. This full model showed that few of the variables- R9 and R5 were insignificant at 95% significance.

A stepwise model selection procedure was used with both AIC and BIC as criteria. Also, LASSO was also used for variable selection. The model with least complexity and best AUC was obtained from stepwise model selection using BIC as a criterion.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameters** | **Full Model** |  | **Backward Selection-AIC** | **Backward Selection-BIC** |
| AIC | 2411.2 |  | 2394.6 | 2478.396 |
| BIC | 2481.4 |  | 2473.1 | 2468.1 |
| Variables-Excluded/Insignificant | R5,R9 |  | R5 | R5,R9 |

The final model was:

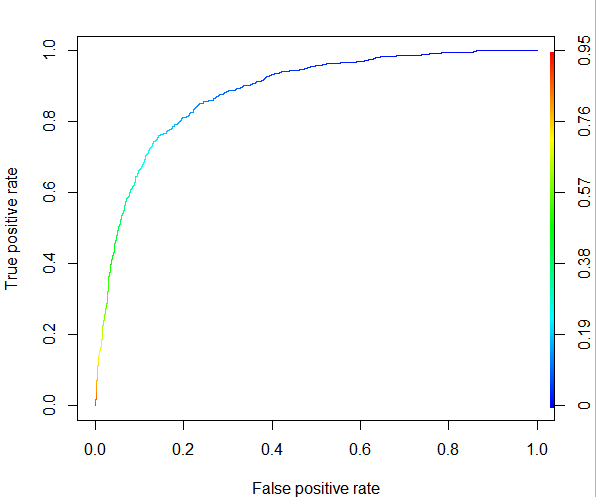
Logit(DLRSN)= **-2.58 + 0.25\*R1 + 0.59\*R2 - 0.38\*R3 – 0.43\*R4 + 0.31R6 - 0.45\*R7- 0.37\*R8 +0.33\*R9 -1.34\*R10**

**In-Sample Performance**

The ROC curve for the chosen model, having AUC of 0.88, is shown below. As per industry standard, the cut-off probability for bankruptcy used was 1/16, and responses were categorized for bankruptcy and studied against the actual responses in the dependent variable column. This yielded a misclassification rate of 0.33 and with a pcut of 1/16.

Mean Residual Deviance: 0.5534455

|  |  |  |
| --- | --- | --- |
| **Actual Values** | **Predicted** | **Values** |
|  | **0** | **1** |
| **0** | 2301 | 1421 |
| **1** | 50 | 576 |



**Out of Sample Performance**

The performance of the chosen model was now tested on the 20% test data, and results compared as below:

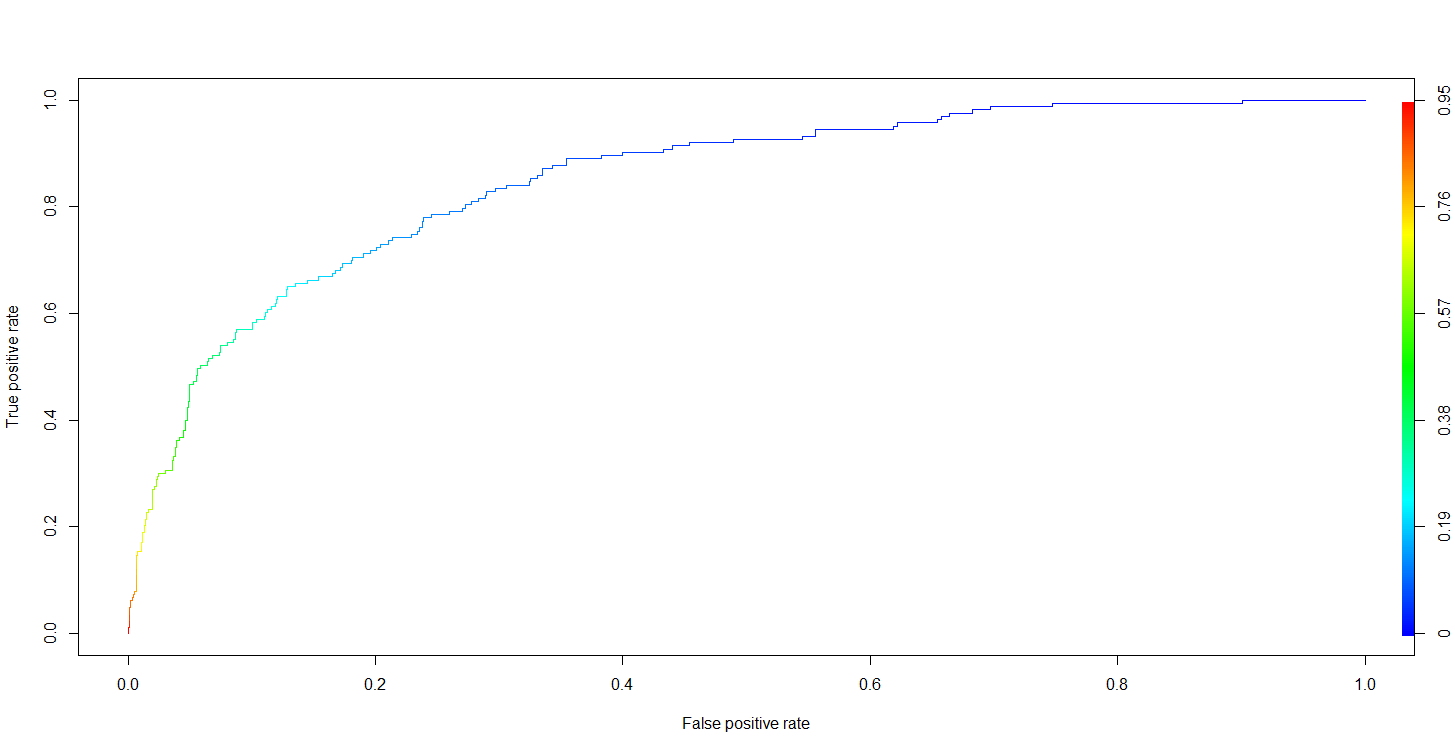
|  |  |  |
| --- | --- | --- |
| **Cut off probability =1/16** | **AUC** | **AMR** |
| In-sample (training data) | 0.88 | 0.35 |
| Out-of-sample (test data) | 0.86 | 0.35 |

Since the target is to flag bankruptcy, when an asymmetric cost function was introduced to penalize the FN at twice the rate of FP,

For the out of sample data testing now, the model was fit, responses again categorized using the same cut off probability of 1/16 and the above metrics captured.

**# Misclassification Rate: 0.338**

**# FNR :0.079**



The AUC was 0.86 for the above curve.

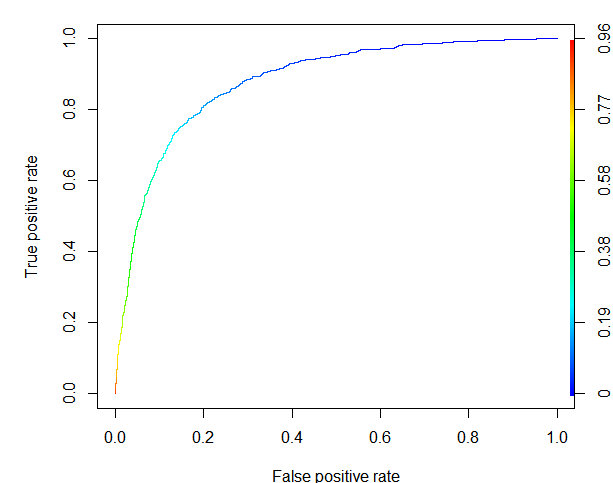
Grid search was also performed using a symmetric cost function and an optimal pcut off of 0.41 was obtained on which a misclassification rate of 0.11 was obtained.

**5-fold cross validation**

Further 5-fold cross validation was performed on the complete data set and a delta 0.4829318 was obtained.

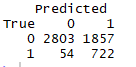
Also, asymmetric cost function with pcut-off of 0.06(obtained by grid method through unequal weights) was used in cross validation and mean residual deviance of 0.4829318 was reported.

ROC curve on complete dataset



AUC: 0.8793651

Confusion Matrix with Asymmetric cost function(15:1) – Complete dataset



Mean Classification Rate: 0.3515453 (Asymmetric)

**CART**

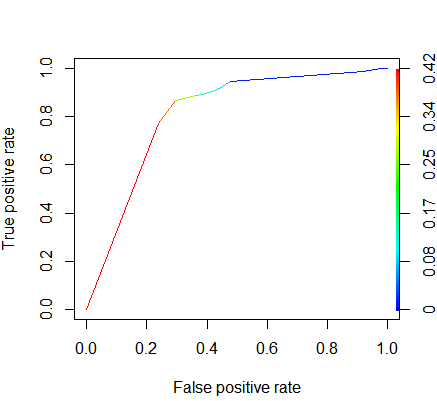
Further a classification tree was applied on the bankruptcy data set to evaluate and compare the performance of tree classification on this data set.

In sample Validation



Misclassification rate: 0.3860294

AUC: 0.8402809



Out of Sample Validation



Misclassification rate: 0.3341766

AUC: 0.8101457

Also, we finally repeated the entire process with another random sample of 80:20, but the results were similar if not the same.

**Conlcusion**

A logistic classification model performed better than a model created by using Cart algorithm and thus the model through the initial exploration was used.