EXECUTIVE SUMMARY

GERMAN CREDIT SCORING DATA

PROBLEM & APPROACH:

This dataset contains 1000 records along with covariates governing whether they are risky for credit or not (0 or 1). The variable response in the dataset corresponds to the risk , 0 means bad and means good. Some basic EDA was performed, and categorical variables were encoded as factors to perform logistic regression .Our objective was to perform a logistic regression to predict the response variable, find best model, ROC curves ,AUC values and misclassification rates. Out of sample prediction was also done. Here, we have found out a suitable model after looking at models using variable selection techniques. We have sampled the data in a 70:30 ratio, ie 70% data for training and 30% data for testing.

RESULTS:

Parameters	Random Sample 1 (80:20)	RandomSample2(90:10)
Final Model	**response ~ status_chck_acc + duration + credit_history + purpose + saving_acctbonds + other_debtors + other_install_plans + installment_rate + credit_amount**	**response ~ status_chck_acc + duration + credit_history + purpose + saving_acctbonds + other_debtors + other_install_plans + installment_rate + credit_amount**
AIC	797.94	893.726
In-Sample AUC	0.8186	0.8193
Out-of-Sample AUC	0.7788	0.7326
Misclassification Rate(in-sample)	0.369	0.357
Misclassification Rate (Out-of-sample)	0.555	0.68
CV Score (AUC)	0.814	0.814
Mean Residual Deviance	741.937	837.726
CV Score (Asymmetric Misclassification rate)	0.547	0.5696
Regression Tree (Misclassification Rate)	0.375	0.332

Table 2: Summary of German Credit Scoring data model

We can find that both the models perform similarly, and the final model suggested by both the models were the same. Our final model is :

response ~ status_chck_acc + duration + credit_history + purpose + saving_acctbonds + other_debtors + other_install_plans + installment_rate + credit_amount

1.Exploratory Data Analysis:

- This dataset contains a set of 1000 observations under 21 attributes.
- The list of the variables is as follows:

```
> str(g_credit)
'data.frame':   1000 obs. of  21 variables:
 $ status_chck_acc      : Factor w/ 4 levels "A11","A12","A13",..: 1 2 4 1 1 4 4 2 4 2 ...
 $ duration             : int  6 48 12 42 24 36 24 36 12 30 ...
 $ credit_history       : Factor w/ 5 levels "A30","A31","A32",..: 5 3 5 3 4 3 3 3 3 5 ...
 $ purpose              : Factor w/ 10 levels "A40","A41","A410",..: 5 5 8 4 1 8 4 2 5 1 ...
 $ credit_amount        : int  1169 5951 2096 7882 4870 9055 2835 6948 3059 5234 ...
 $ saving_acctbonds     : Factor w/ 5 levels "A61","A62","A63",..: 5 1 1 1 1 5 3 1 4 1 ...
 $ present_employment   : Factor w/ 5 levels "A71","A72","A73",..: 5 3 4 4 3 3 5 3 4 1 ...
 $ installment_rate     : int  4 2 2 2 3 2 3 2 2 4 ...
 $ statussex            : Factor w/ 4 levels "A91","A92","A93",..: 3 2 3 3 3 3 3 3 1 4 ...
 $ other_debtors        : Factor w/ 3 levels "A101","A102",..: 1 1 1 3 1 1 1 1 1 1 ...
 $ present_residence    : int  4 2 3 4 4 4 4 2 4 2 ...
 $ property             : Factor w/ 4 levels "A121","A122",..: 1 1 1 2 4 4 2 3 1 3 ...
 $ age                  : int  67 22 49 45 53 35 53 35 61 28 ...
 $ other_install_plans  : Factor w/ 3 levels "A141","A142",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ housing              : Factor w/ 3 levels "A151","A152",..: 2 2 2 3 3 3 2 1 2 2 ...
 $ no_credits           : int  2 1 1 1 2 1 1 1 1 2 ...
 $ job                  : Factor w/ 4 levels "A171","A172",..: 3 3 2 3 3 2 3 4 2 4 ...
 $ no_people_maintenance: int  1 1 2 2 2 2 1 1 1 1 ...
 $ telephone            : Factor w/ 2 levels "A191","A192": 2 1 1 1 1 2 1 2 1 1 ...
 $ foreign_worker       : Factor w/ 2 levels "A201","A202": 1 1 1 1 1 1 1 1 1 1 ...
 $ response             : Factor w/ 2 levels "0","1": 1 2 1 1 2 1 1 1 1 2 ...
`
```

Table 9: Structure of Variables

There are 8 numerical variables including the response and 13 are categorical variables with various levels.

The summary of the variables is as follows:

```
> summary(g_credit)
 status_chck_acc    duration    credit_history    purpose    credit_amount   saving_acctbonds present_employment
 A11:274         Min.   : 4.0   A30: 40        A43    :280   Min.   :  250   A61:603          A71: 62
 A12:269         1st Qu.:12.0   A31: 49        A40    :234   1st Qu.: 1366   A62:103          A72:172
 A13: 63         Median :18.0   A32:530        A42    :181   Median : 2320   A63: 63          A73:339
 A14:394         Mean   :20.9   A33: 88        A41    :103   Mean   : 3271   A64: 48          A74:174
                 3rd Qu.:24.0   A34:293        A49    : 97   3rd Qu.: 3972   A65:183          A75:253
                 Max.   :72.0                  A46    : 50   Max.   :18424
                                               (Other): 55
 installment_rate statussex other_debtors present_residence property        age        other_install_plans housing
 Min.   :1.000    A91: 50   A101:907      Min.   :1.000     A121:282   Min.   :19.00   A141:139            A151:179
 1st Qu.:2.000    A92:310   A102: 41      1st Qu.:2.000     A122:232   1st Qu.:27.00   A142: 47            A152:713
 Median :3.000    A93:548   A103: 52      Median :3.000     A123:332   Median :33.00   A143:814            A153:108
 Mean   :2.973    A94: 92                 Mean   :2.845     A124:154   Mean   :35.55
 3rd Qu.:4.000                            3rd Qu.:4.000                3rd Qu.:42.00
 Max.   :4.000                            Max.   :4.000                Max.   :75.00

  no_credits      job      no_people_maintenance telephone  foreign_worker response
 Min.   :1.000   A171: 22   Min.   :1.000         A191:596   A201:963       0:700
 1st Qu.:1.000   A172:200   1st Qu.:1.000         A192:404   A202: 37       1:300
 Median :1.000   A173:630   Median :1.000
 Mean   :1.407   A174:148   Mean   :1.155
 3rd Qu.:2.000              3rd Qu.:1.000
 Max.   :4.000              Max.   :2.000
```

Table 10: Summary of variables

Visualization of Variables:

We have used histograms to understand the distribution of numerical variables. The distribution of numerical as well as categorical variables (only some variables) with respect to response variable is also presented.

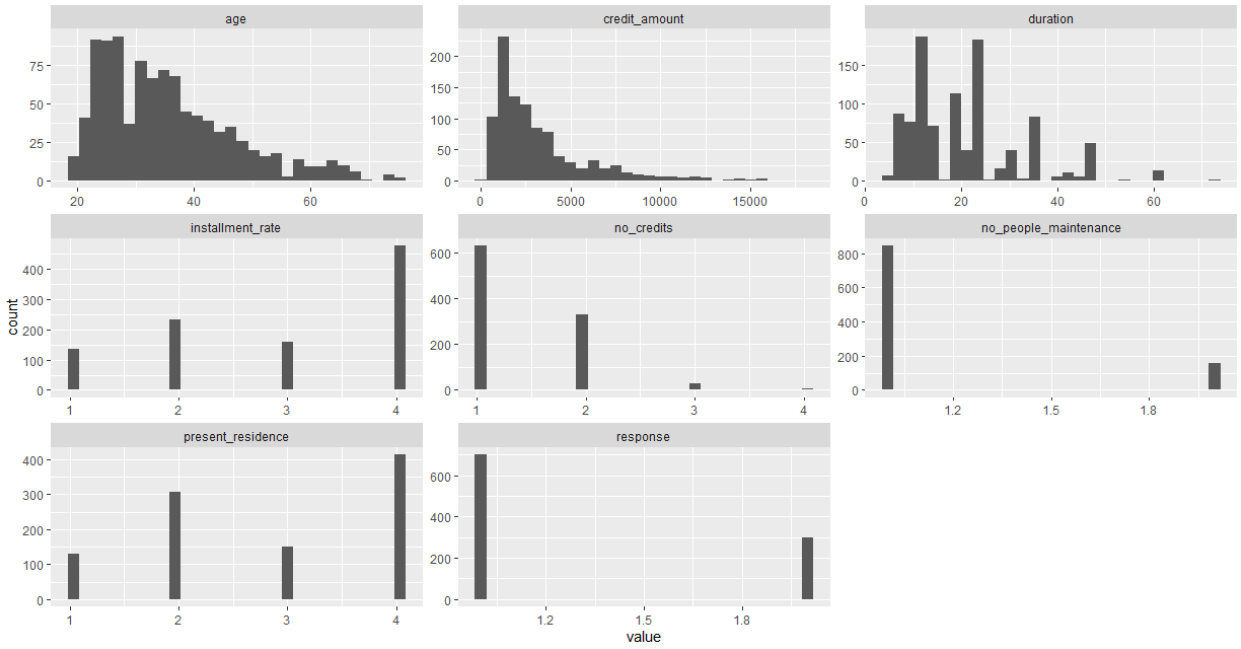


Fig 6: Histograms of numerical variables

We can find that some of the numerical variables need to be converted into factors as they are different categories. Also, Histograms of age, duration and credit amount do not seem to be normally distributed. We could check for outliers for the variables using boxplots.

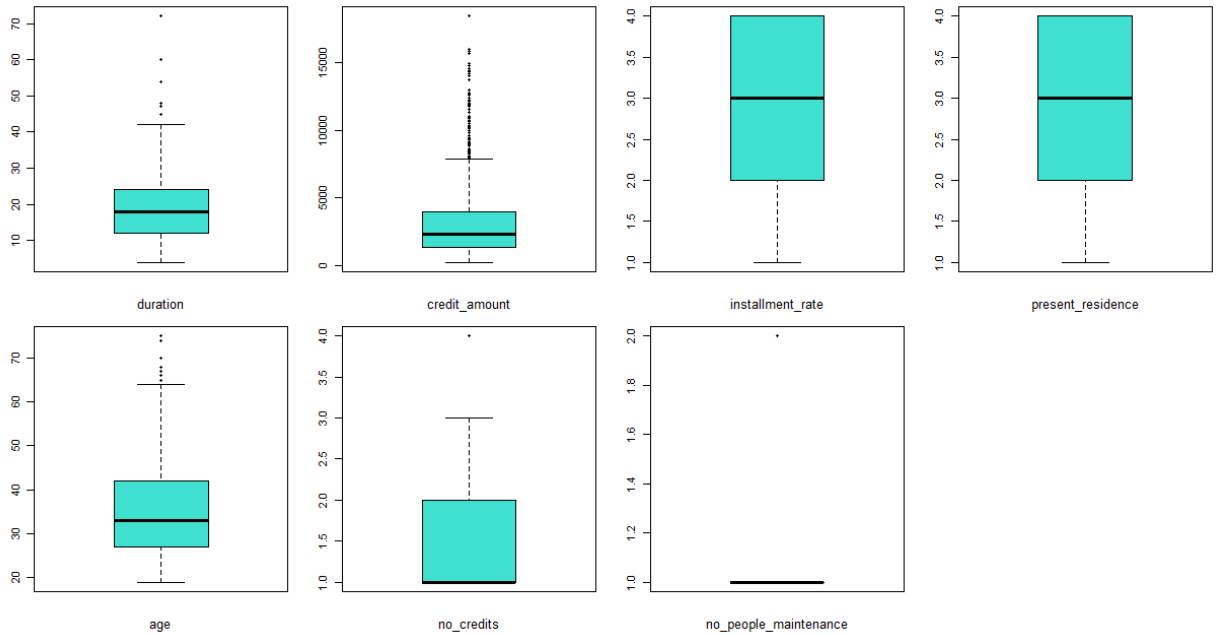


Fig 7: Boxplots of variables

We can observe that some of the variables are having outliers. But, for the purpose of this analysis, we are not removing any outliers.

Generalized Linear Regression:

Our goal is to create a generalized linear regression model with different link functions to analyse the effects of covariates on the response variable. We have encoded all categorical variables as factors for regression. Then, we found out a suitable model after looking at models using different variable selection criteria like AIC,BIC and LASSO regression . We have sampled the data in an 80:20 ratio, i.e., 80% data for training and 20% data for testing.

We tried different link functions - logistic, probit, complementary log-log link and compared the results from all the three models. The model performances were as below:

Parameters	Logit Model	Probit Model	Clog log Model
AIC	796.84	797.37	795.74
Residual Deviance	698.84	699.37	697.74

Table 11: Comparison of different Link functions

We can see that all the link functions provide similar results.

Variable Selection:

We used various variable selection criteria to find out the best model. We got the following models with stepwise AIC ,BIC and LASSO:

Model Type	Model	AIC value
AIC	**response ~ status_chck_acc + duration + credit_history + purpose + saving_acctbonds + other_debtors + present_employment + foreign_worker + other_install_plans + installment_rate + statussex + no_people_maintenance + credit_amount + age**	**787.79**
BIC	**response ~ status_chck_acc + duration**	**847.51**
LASSO	**response ~ status_chck_acc + duration + credit_history + purpose + saving_acctbonds + other_debtors + present_employment + foreign_worker + other_install_plans + installment_rate + statussex + no_people_maintenance + credit_amount + age**	**787.79**

Table 12: Variable Selection model parameters (Random Sample 1)

The stepwise model using BIC included only a very small number of variables as it penalizes the model for including more variables. Variables were selected after considering AIC performance criteria. Even though many predictor variables appeared in the model suggested by AIC , some of them were not significant when we observed the model characteristic p values.

We removed those insignificant variables and created our final model which is:

response ~ status_chck_acc + duration + credit_history + purpose + saving_acctbonds + other_debtors + other_install_plans + installment_rate + credit_amount

AIC of the final model was calculated to be **797.94.**

IN SAMPLE RESULTS:

All results provided below are for in sample. The cutoff probability was kept as 0.1667.
ROC curve is presented below, and the AUC value is 0.8186.

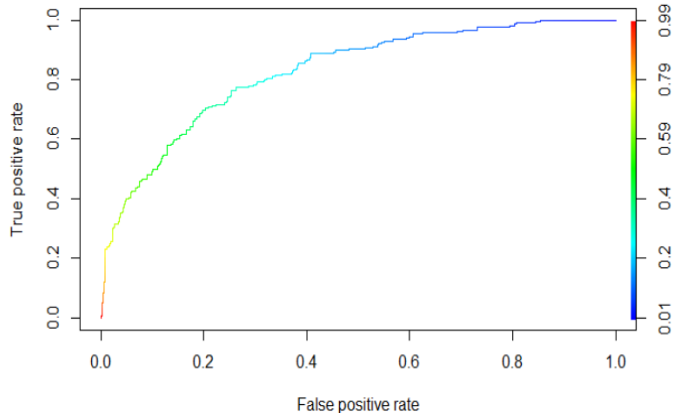


Fig 8: ROC curve (in-sample)

Confusion Matrix obtained is presented below and the Misclassification Rate came out to be 0.369 .

	Predicted	
Truth	0	1
0	295	268
1	27	210

The mean residual deviance was calculated to be **741.937.**

ii) OUT OF SAMPLE RESULTS:
Using final logistic linear model built from on the 80% training data and 20% testing data, the out of sample results are presented below. The cutoff probability was kept as 0.1667.
ROC curve is presented below, and the AUC value is 0.7788.

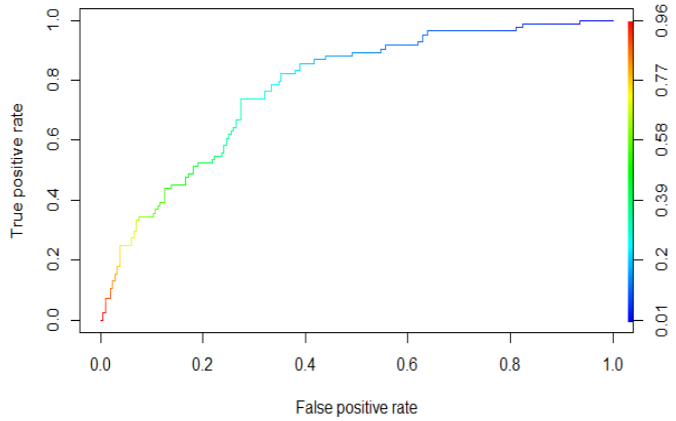


Fig 9: ROC curve (out-of-sample)

Using a 5:1 penalty for misclassification . Asymmetric Misclassification rate (to penalize false positives) seemed to be 0.555.

CROSS VALIDATION:

We performed 5-fold Cross validation on the original data and found out the performance characteristics of the model. We tried both AUC and asymmetric misclassification rate as the cost functions. We got the following CV scores :

AUC : **0.814**

Asymmetric misclassification rate : **0.541**

We can observe that we are getting almost similar results from the logistic regression model. We obtain a slightly higher value of AUC which is an indicator of a better model. As cross validation reduces the chances of overfitting, it can give a better performance model.

Classification Tree:

Here, we have an asymmetric cost function. It means that false negatives will cost more than false positives (predicting 1 when truth is 0). Here we assume that false negative cost 5 times of false positive.

The Classification tree is as shown below:

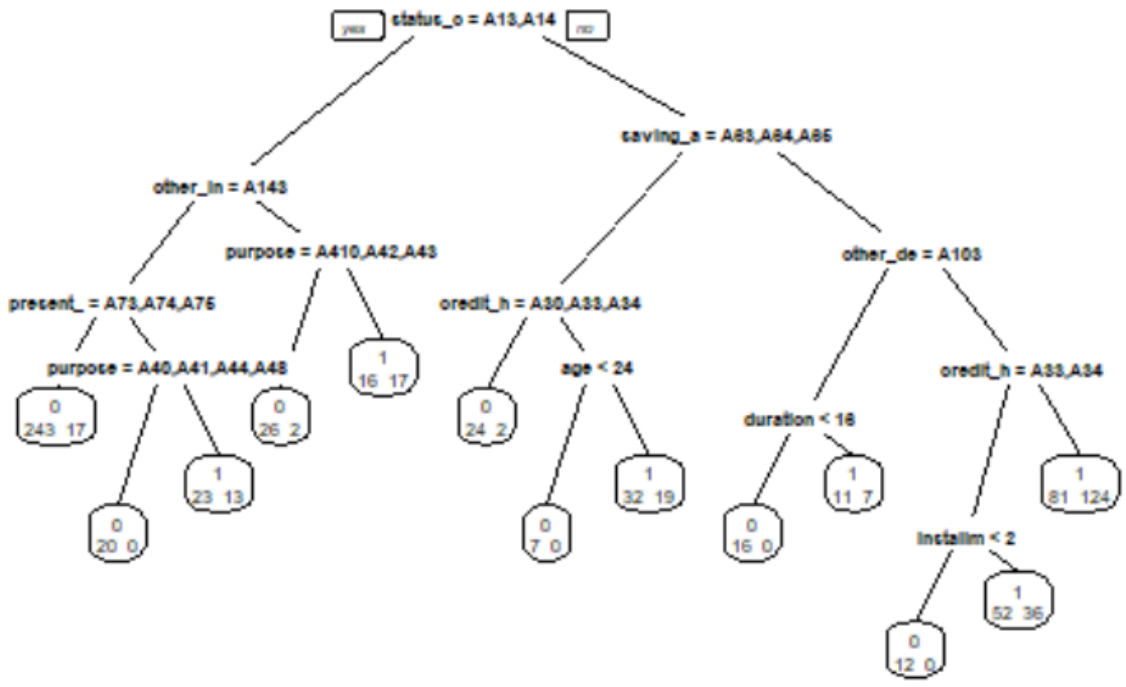


Fig 10: Classification Tree

The above classification tree can be used to predict the response variable.

The asymmetric misclassification table for out of sample performance is as below:

Truth	Predicted 0	1
0	78	59
1	16	47

The misclassification rate is calculated to be 0.375.

Then, we used plotcp() function to prune the tree and the resulting plot we observed is as below:

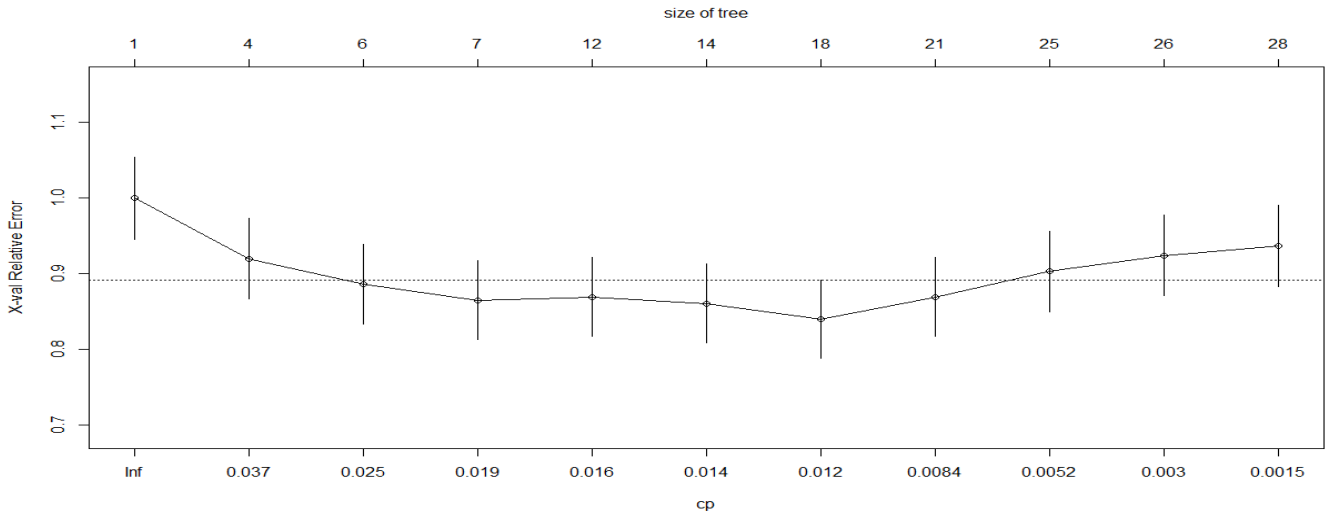


Fig 11: Relationship between cross-validation error in the training set and size of tree

Comparing CART to the logistic regression model:

When we compared the CART model to the logistic regression model we obtained, the misclassification table was calculated to be 0.375.

Truth	Predicted 0	1
0	71	66
1	9	54

<u>Random Sample 2 (90:10 split)</u>

We then repeated all the modelling techniques by using another random sample which was split in a 90:10 ratio into training and testing data respectively.

We tried different link functions - logistic, probit, complementary log-log link and compared the results from all the three models. The model performances were as below:

Parameters	Logit Model	Probit Model	Clog log Model
AIC	892.42	891.72	890.01
Residual Deviance	794.42	793.72	792.01

Table 13: Comparison of different link functions (Random Sample 2)

We performed all the variable selection methods as before and the results we obtained is as tabulated below:

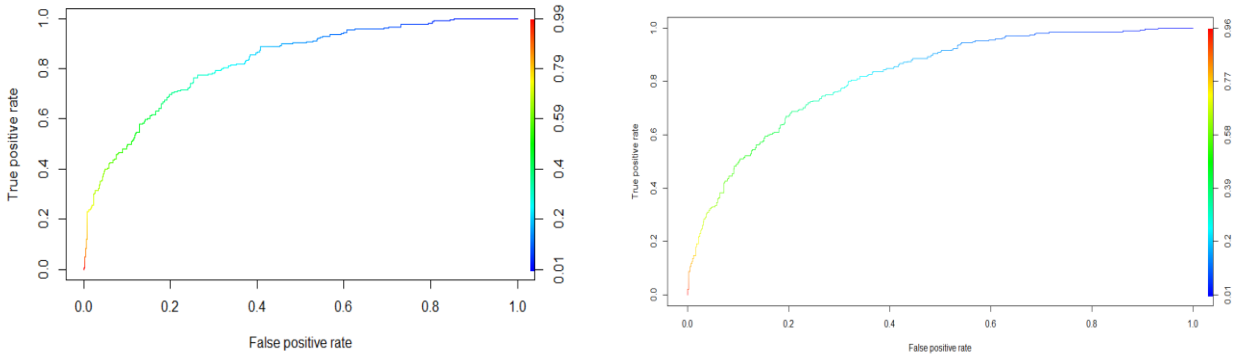
Parameters	Random Sample 1 (80:20)	Random Sample2(90:10)
Final Model	**response ~ status_chck_acc + duration + credit_history + purpose + saving_acctbonds + other_debtors + other_install_plans + installment_rate + credit_amount**	**response ~ status_chck_acc + duration + credit_history + purpose + saving_acctbonds + other_debtors + other_install_plans + installment_rate + credit_amount**
AIC	797.94	893.726
In-Sample AUC	0.8186	0.8193
Out-of-Sample AUC	0.7788	0.7326
Misclassification Rate(in-sample)	0.369	0.357
Misclassification Rate (Out-of-sample)	0.555	0.68
CV Score (AUC)	0.814	0.814
Mean Residual Deviance	741.937	837.726

CV Score (Asymmetric Misclassification rate)	0.547	0.5696
Classification Tree (Misclassification Rate)	0.375	0.332

Table 14: Comparison of Model with different random samples

ROC CURVES:

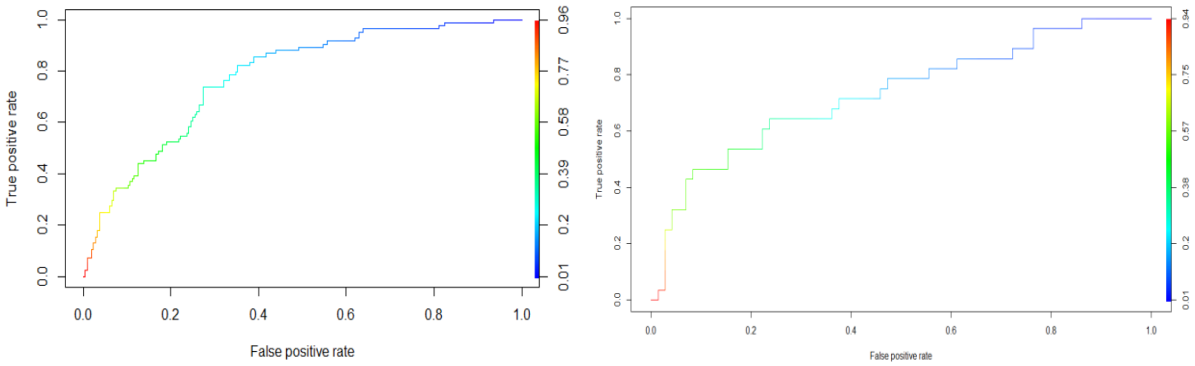
IN SAMPLE:



Random Sample 1 Random Sample 2

Fig 12: ROC Curves for different random samples (in-sample)

OUT OF SAMPLE:



Random Sample 1 Random Sample 2

Fig 13: ROC Curves for different random samples (out-of-sample)

We can find that both the models perform similarly, and the final model suggested by both the models were the same. Our final model is :

response ~ status_chck_acc + duration + credit_history + purpose + saving_acctbonds + other_debtors + other_install_plans + installment_rate + credit_amount