German Credit Risk Analysis

Introduction

The German Credit dataset, developed by Prof. Dr. Hans Hofmann in 1994, is mainly used to predict the credit risk of any specific individual based on information of that individual. In this dataset, each entry represents a person who takes a credit by a bank. Each person is classified as good or bad credit risks according to the set of attributes. Our goal is to explore the dataset and then generate a model in order to accurately classify the credit risk.

Problem Approach

In the given problem, the team first performed basic exploratory data analysis in order to gain some better understanding of the data. The data was then split into 70:30 and subsequent steps of modeling were performed starting with variable selection, selecting final model and fitting logistic model. The in-sample and out-of-sample model performance was computed. Similar steps were then repeated for CART model. Finally, all the above steps were again repeated for another split ratio 80:20 in order to compare the new results with the previous one.

Major Results

- The logistic model provided **in-sample** misclassification rate as **35.28%** and **out-of-sample** misclassification rate as **37.33%.**
- The CART model provided **in-sample** misclassification rate as **30.42%** and **out-of-sample** misclassification rate as **46.67%**.
- On using different split ratio (80:20), we observe decrease in out-of-sample misclassification for both models.

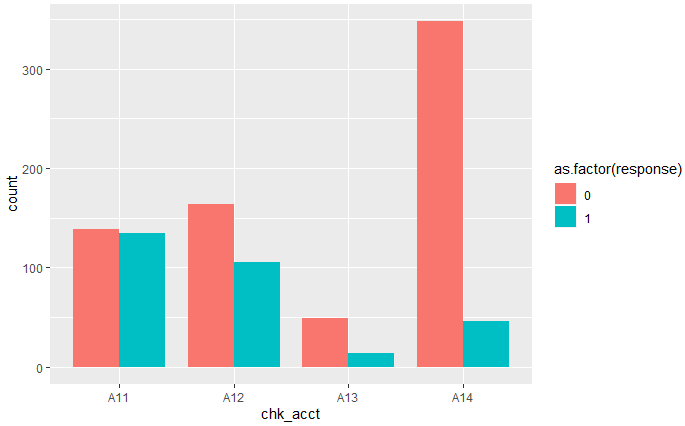
Data description

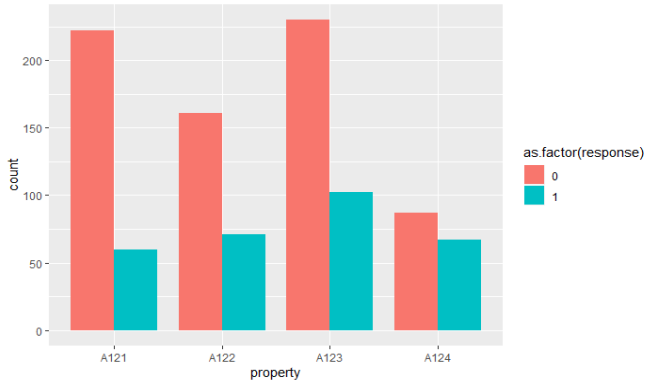
This dataset contains 1000 observations across 21 variables describing the information of an individual in the dataset. The response variable for this study is the "Response" column which takes two values to indicate the loan status – 1 is good and 2 is bad. The response column values were then deducted by 1 in order to convert it into binary variable. The overall dataset structure is as follows:

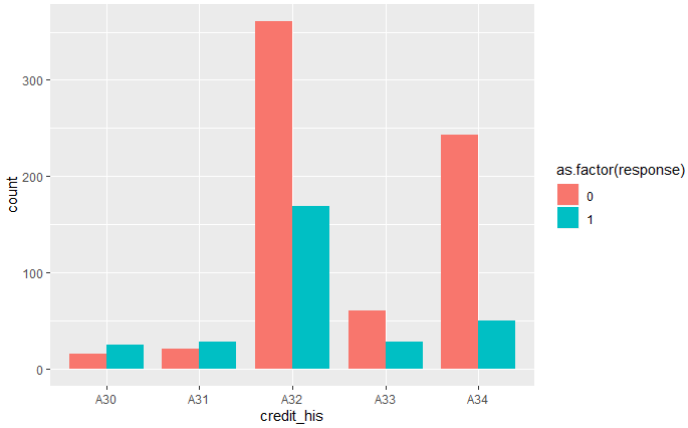
- 13 Categorical Variables
- 8 Numeric Variables

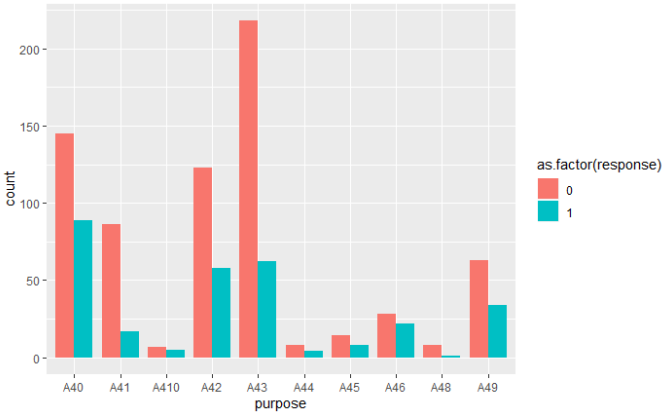
Data Exploration

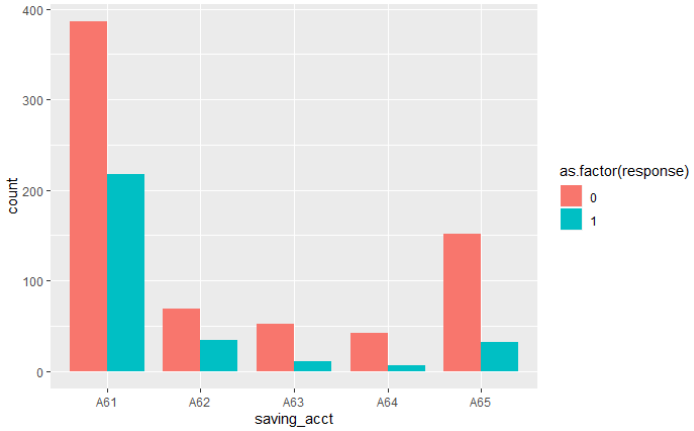
In order to better understand the data, barplots have been generated for each categorical variable (0: not-defaulters, 1: defaulters). For 3 continuous variables, scatter jitter plot has been plotted.

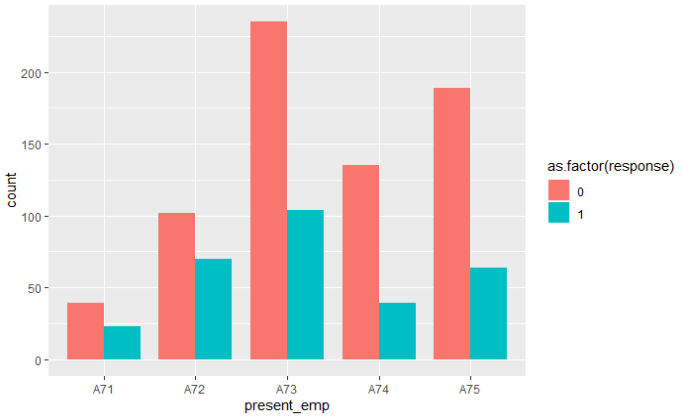


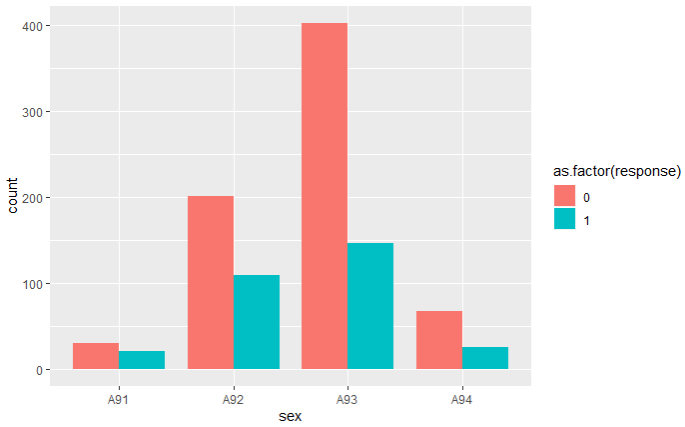


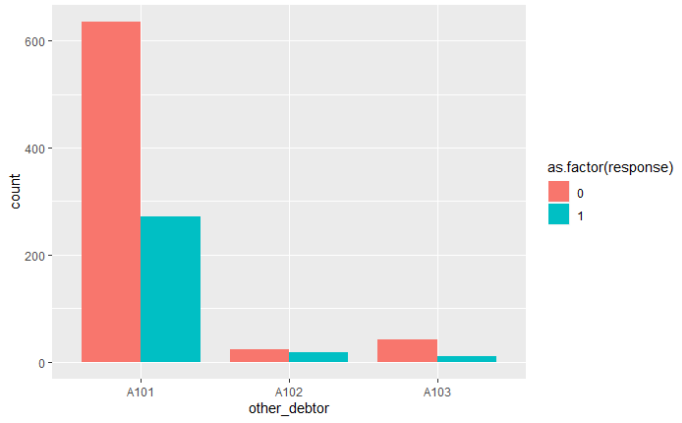


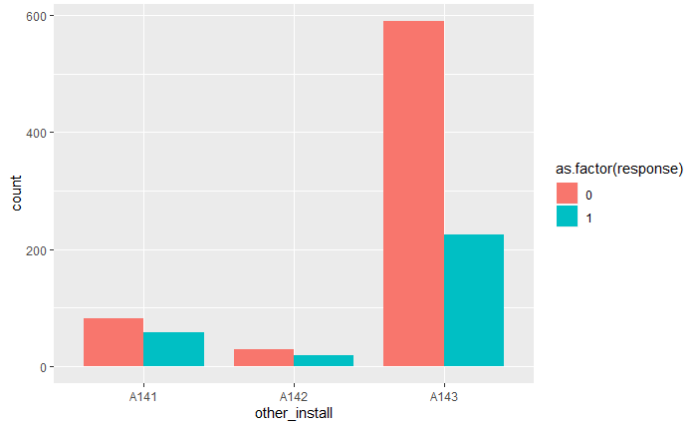


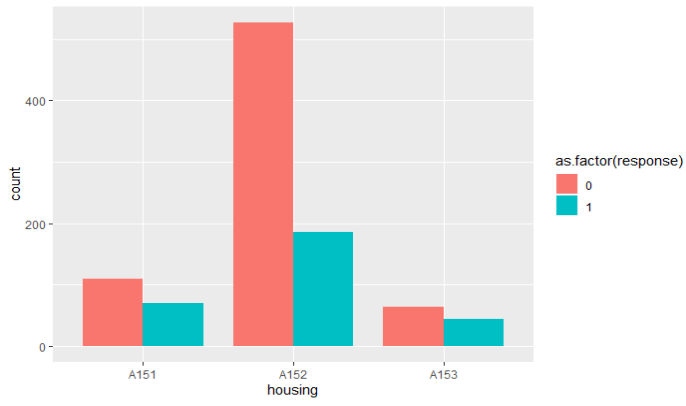

count
400
300
200
100
0
A61
A62
A63
A64
A65
saving_acct
as.factor(response)
0
1

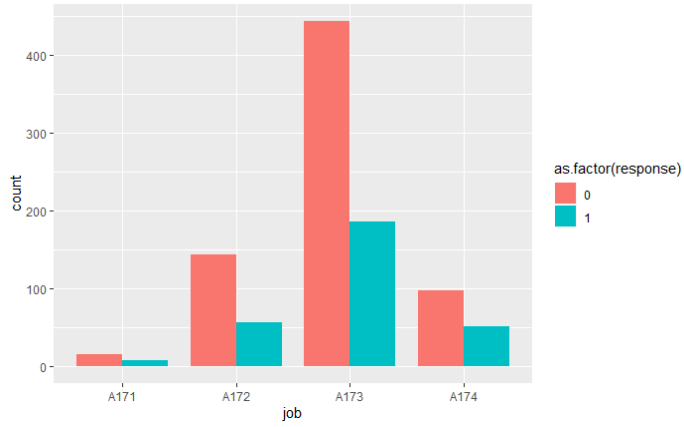

count
200
150
100
50
0
A71
A72
A73
A74
A75
present_emp
as.factor(response)
0
1

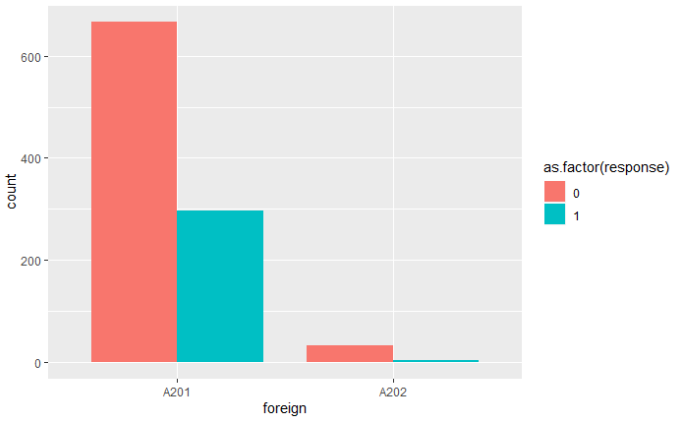

count
400
300
200
100
0
A91
A92
A93
A94
sex
as.factor(response)
0
1

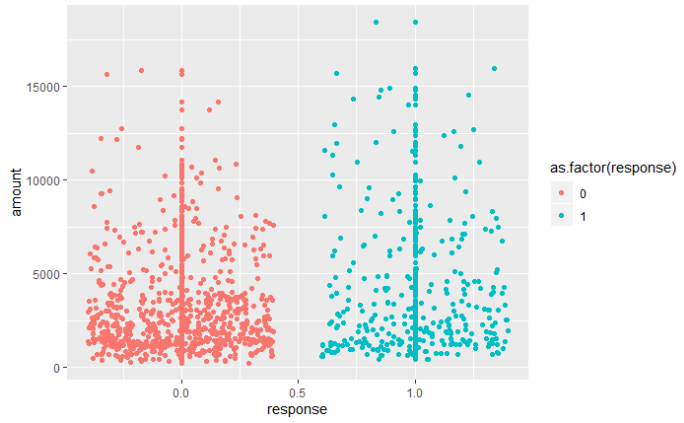

count
600
400
200
0
A101
A102
A103
other_debtor
as.factor(response)
0
1

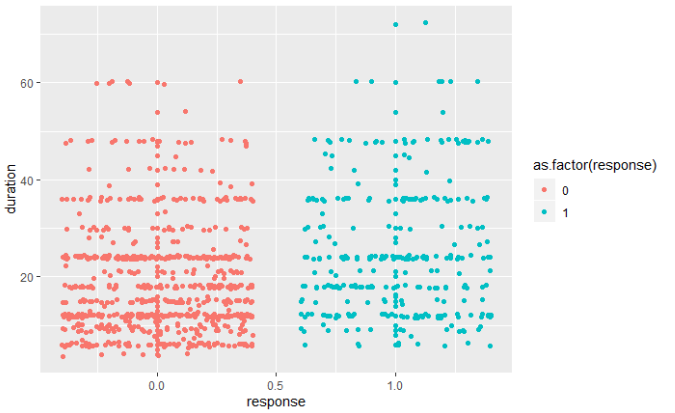

count
600
400
200
0
A141
A142
A143
other_install
as.factor(response)
0
1

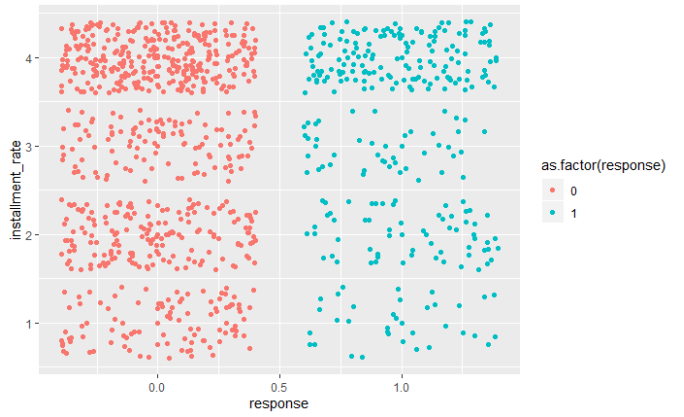

count
400
200
0
A151
A152
A153
housing
as.factor(response)
0
1

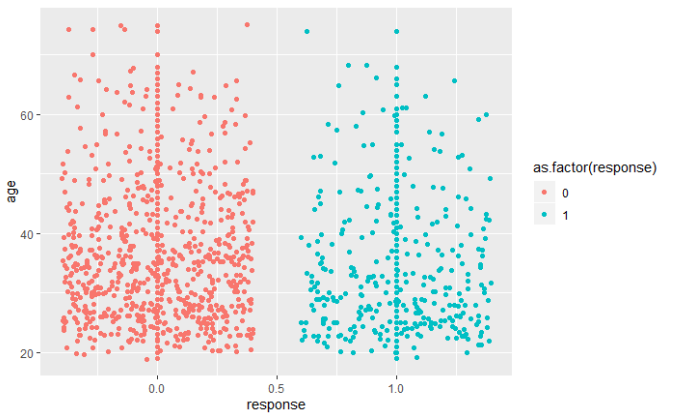

count
400
300
200
100
0
A171
A172
A173
A174
job
as.factor(response)
0
1


count
600
400
200
0
A201
A202
foreign
as.factor(response)
0
1


amount
15000
10000
5000
0
0.0
0.5
1.0
response
as.factor(response)
0
1


duration
60
40
20
0.0
0.5
1.0
response
as.factor(response)
0
1


installment_rate
4
3
2
1
0.0
0.5
1.0
response
as.factor(response)
0
1


age
60
40
20
0.0
0.5
1.0
response
as.factor(response)
0
1

Data Sampling

We have created train and test dataset by doing random sampling on the dataset with 70-30 ratio. Another sample with 80-20 ratio is also created to compare results.

Logistic regression model without variable selection

When we developed a logistic model using all the variables and on applying different link functions, following points were observed –

- Many variables were found to be significant with p-values less than .05 for each of the link.
- Although most of the significant variables were same in each model, but there was still some minor difference in significance level and some variables.
- The summary statistics are given below:

Method	Link	Residual Deviance	Residual DF	Mean Residual Deviance	AIC
GLM	Logistic	628.26	651	0.965	726.26
	Probit	628.56	651	0.965	726.56
	Cloglog	626.05	651	0.961	724.05

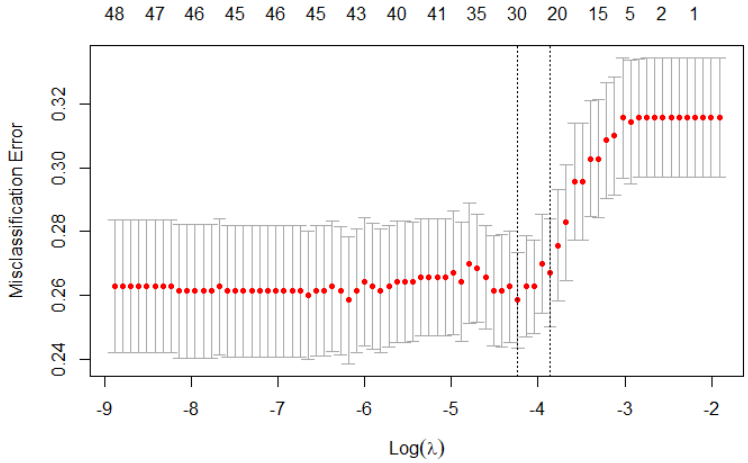
Logistic regression model with variable selection techniques

1. Stepwise Selection

When stepwise selection technique was used on the dataset with forward, backward and both ways selection, following summary table was obtained –

Selection Criteria	Variables Selected	AIC Value
AIC	Chk_account, duration, saving_acct, other_debtor, sex, installment_rate, amount, credit_his, purpose, other_install, telephone, foreign, housing	713.71
BIC	Chk_account, duration	767.32

2. LASSO selection



When variable selection was done using LASSO regression technique, we get the following variables: *chk_account, duration, credit_his, purpose, amount, saving_acct, present_emp, installment_rate, sex, other_debtor, age, other_install, housing and telephone.*

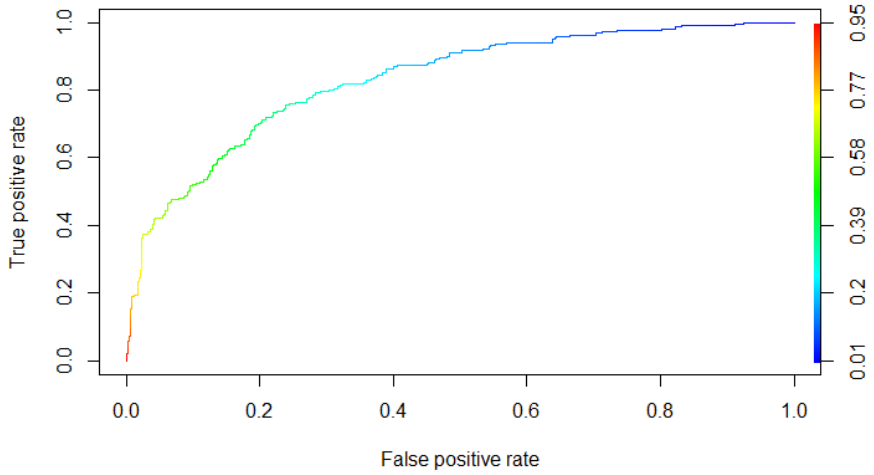
Only "present_emp" is the difference between two models. So, we are going with the stepwise selection criteria using AIC and using that as our final model.

Final Model: Response ~ Chk_account + duration + saving_acct + other_debtor + sex + installment_rate + amount + credit_his + purpose + other_install + telephone + foreign + housing

In-sample accuracy

- **ROC curve and AUC score**

Following ROC curve was obtained when model was applied on the training dataset with an **AUC score of 0.8298.**



- **Misclassification rate**

Using the probability cut-off as 1/6, confusion matrix was created as follows:

Misclassification Rate – 0.3528		**Predicted**	
		0	**1**
True	**0**	256	223
	1	24	197

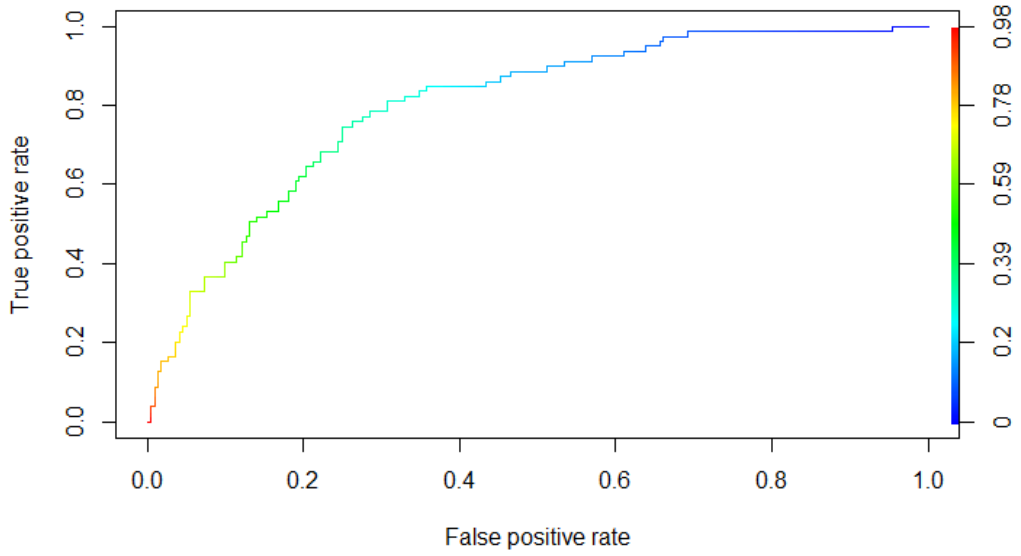
- **Mean Residual Deviance: 0.968**

  AIC: 713.71

Out-of-sample accuracy

- **ROC curve and AUC score**

Following ROC curve was obtained when model was applied on the training dataset with an **AUC score of 0.80.**



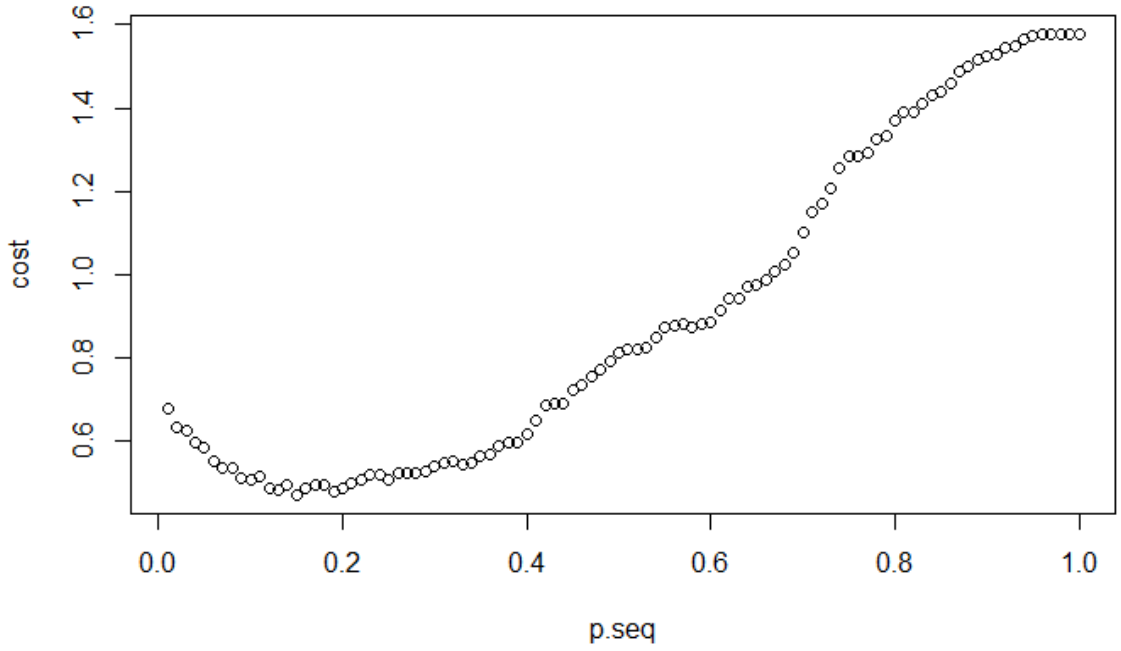
- **Misclassification rate**

Misclassification rate of 0.3733 was obtained from the model using the same probability cut-off as used in the in-sample accuracy calculation.

Misclassification Rate – 0.3733		**Predicted**	
		0	**1**
True	**0**	118	103
	1	9	70

Optimal Probability Cut-off

Using the asymmetric cost as 1:5, we apply the grid search method to find the optimal probability cut-off. The optimal probability cut-off comes as **0.15.**



Cross Validation

In order to predict average model error accuracy, we apply 3-fold cross validation using different types of cost functions. The results are summarized below:

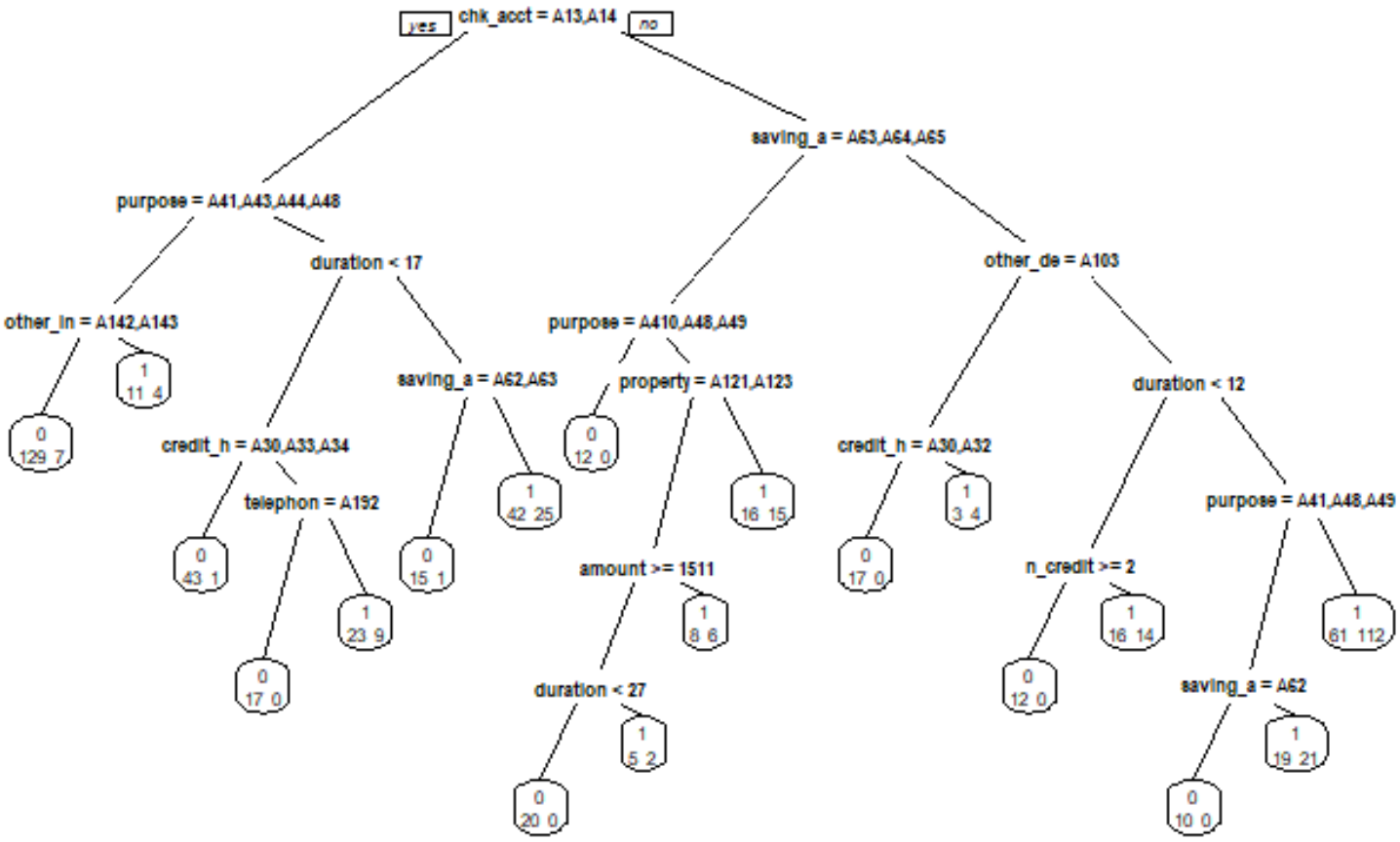
Cost Function	Model Error
Asymmetric Cost(1:5)	0.528
AUC	0.791
Default (Mean Squared Error)	0.164

Both AUC and Asymmetric cost performed worse as compared to the original model in terms of model accuracy.

Classification Tree

This time classification tree is used to fit our training data and the similar steps like above are repeated to check model in-sample and out-of-sample accuracy.

Tree Diagram:



In-Sample Prediction:

Using the 5:1 asymmetric loss, the following confusion matrix is created with Misclassification rate coming as 0.3042.

Misclassification Rate – 0.3042		**Predicted**	
		0	**1**
True	**0**	275	204
	1	9	212

Out-of-Sample Prediction:

The Out-of-Sample misclassification rate comes as 0.4667.

Misclassification Rate – 0.4667		Predicted	
		0	**1**
True	**0**	101	120
	1	20	59

CART in general will give better in-sample accuracy which is visible from our results. However, by doing so it runs into the risk of overfitting the data which may sometimes result in high out-of-sample misclassification rate as can be seen above.

Repeating above steps using different split ratio(80:20)

Logistic Regression using all variables:

Method	Link	Mean Residual Deviance	AIC
GLM	Logistic	0.958	817.45
	Probit	0.956	816.49
	Cloglog	0.967	816.92

The mean residual deviance has not changed. However, AIC value has now gone up for each link type.

Variable Selection:

- Using AIC, we get the same variables except "Housing" which is now not selected.
- Using BIC, we also get the same variables.

Prediction Accuracy:

	Metrics	70:30 Split	80:20 Split
In-Sample	AUC	0.8298	0.8268
	Misclassification	35.28%	36.25%
Out-of-Sample	AUC	0.80	0.805
	Misclassification	37.33%	36.50%

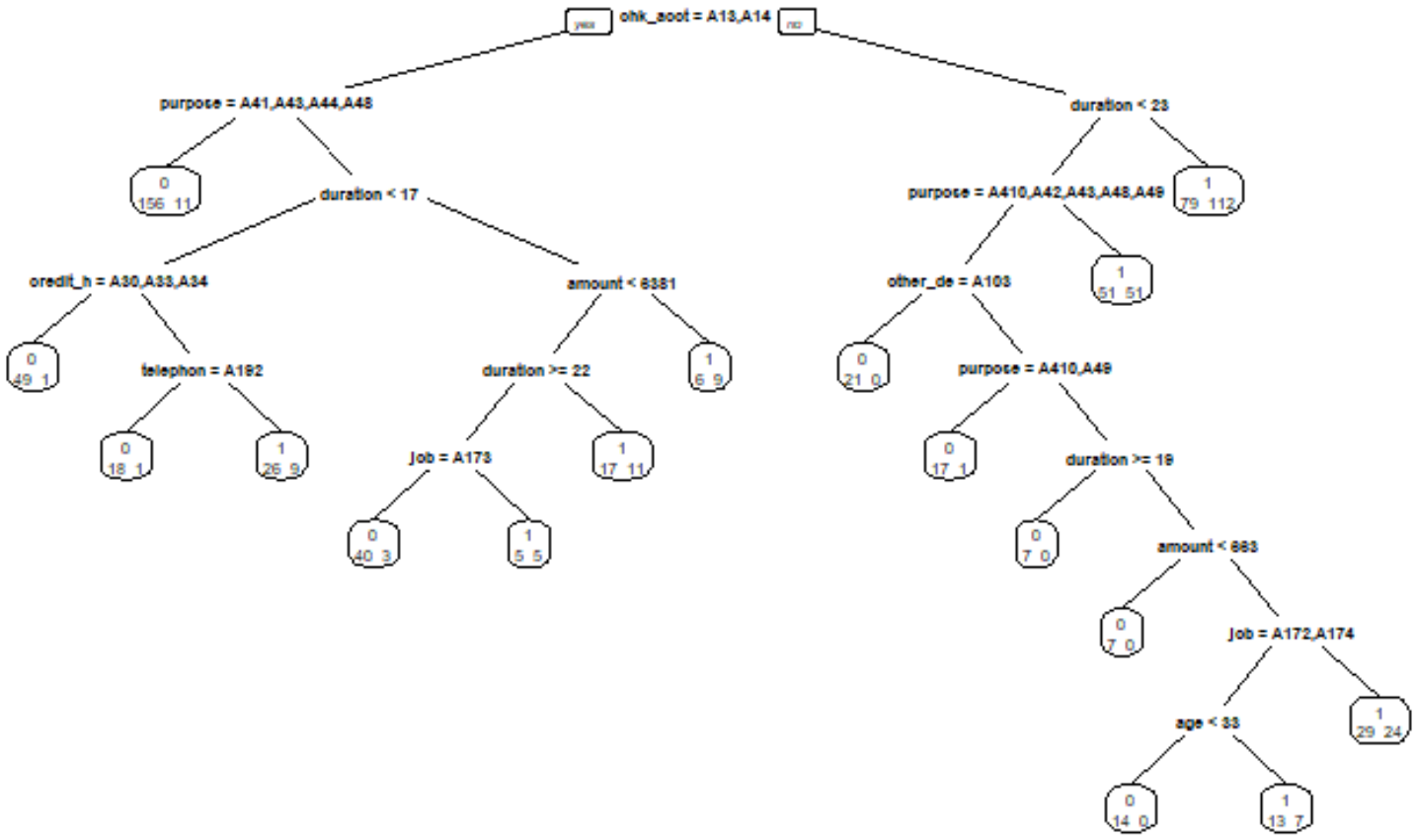
AUC is more or less the same. The Misclassification rate for out-of-sample however has gone down and the **optimal probability cut-off** has also decreased from earlier value of **0.15 to 0.12.**

Cross Validation:

Cost Function	Model Error(70:30)	Model Error(80:20)
Asymmetric Cost(1:5)	0.528	0.538
AUC	0.791	0.788
Default (Mean Squared Error)	0.164	0.163

No noticeable difference in the above values.

Classification Tree:



An interesting observation - The tree diagram has changed after changing the split ratio since there will be now different observations in our training data as compared to the previous one.

Classification Tree Prediction Accuracy:

	For 70:30 Split	**For 80:20 Split**
In-sample Misclassification Error	0.3042	0.3037
Out-of-sample Misclassification Error	0.4667	0.4350

The CART seems to perform better with higher split ratio in its favor.