# Credit Risk Scoring

Implementation Using Classification

Kanika Sharma 3/25/18 ESC Rennes School of Business

# CONTENT

# Acknowledgement

I sincerely thank Mr. Seongsoo Jang for his approval in carrying this project. His valuable feedback really helped me in writing the better and modified research paper.

**Abstract**

*In today's fast changing digital world, Credit risk scoring is considered as one of main competencies in the commercial banks. Primary objective of all commercial banks is collecting reserve funds of people and substances and allocating them as managing and account offices to mechanical, benefit and profitable companies. Also, it is said to be challenging and important data mining problems. Not refunding offices by clients, put banks in numerous inconveniences such as not being able to refund Central Bank advances, more offices compare to client's discount and not being competent of to give more offices. The significance of this matter and its part in financial development and increase work openings driven to different models for exploring customer's credits who want to get offices. Various Models has been discovered to validate credit risk applications requested by the public or companies or individualist research will deal with process of credit risk scoring using classification for better decision making*

*Keywords:* Credit Scoring; Classification Techniques;

# 1 RESEARCH GOAL

Credit Risk is one of most important topics among financial sectors. Every time a bank receives an application for the loan, the process is such that bank must find out whether the loan requester will be valid candidate, or will he pay the loan or not by looking at the metadata provided with the application. It can be judged by both judgemental methods r reedit risk scoring methods although the process has changed from last one decade. The objective of the process is to get the answers to following two questions: -

- If the applicant is a good credit risk, i.e. is likely to repay the loan, then not approving the loan to the person results in a loss of business to the bank
- If the applicant is a bad credit risk, i.e. is not likely to repay the loan, then approving the loan to the person results in a financial loss to the bank

The Objective of this research is to minimize the loss and maximize the profit for the bank by making a better decision-making system. In this process, loan requester's various backgrounds are checked like demographics, credit history, assets owned. The classification in the end will tell whether the applicant is good or bad customer. And only after his, the loan is approved by loan manager's research will be done by applying several Logistic Regression, Random Forest, Decision Tree on the data and finding the best suited classifier for the problem.

# 2 DATA UNDERSTANDING

The German Credit Dataset is available online for the analysis by various colleges. It contains 20 columns for 1000 loan requester or applicants. Below are the columns with them

- Creditability (kredit) where 0 and 1 define credit worthy values
- Account. Balance (laufkont) showing the balance of the current account
- Duration.of. Credit. Month (laufzeit) showing duration in months
- Payment.Status.of. Previous. Credit (moral) showing payment of previous credit
- Purpose (verw) showing purpose of credit
- Credit.Amount (Hoehe) showing the amount of credit in Deutsche Credit
- Value.Savings. Stocks (sparkont) showing the value of the stocks
- Length.of. current. Employment (beszeit) showing for how long employed by current employer
- Instalment.per.cent (rate) showing the instalment percentage of available income
- Sex...Marital.Status(famges) showing the marital status
- Guarantors(buerge) whether applicant have any guarantors
- Duration.in.Current. address(wohnzeit) showing the duration for which applicant is living in current address.
- Most. Valuable. available. asset (verm) showing the most available assets
- Age. Years(alter) showing the age of the applicant
- Concurrent.Credits (weitkred) showing further running credits
- Type. Of. apartment(wohn) showing type of apartment
- No.of.Credits.at.this. Bank. Bank(bishkred) showing Number of previous credits at this bank (including the running one)
- Occupation (beruf) showing occupation
- No.of. dependents(pers) showing number of persons entitled to maintainence.
- Telephone(telef) showing telephone
- Foreign.Worker (gastarb) showing whether the applicant is foreigner or not.

These variable needs to be classified and scored to reach that may potentially have any influence on Creditability.

- Account Balance: No account (1), None (No balance) (2), Some Balance (3)
- Payment Status: Some Problems (1), Paid Up (2), No Problems (in this bank) (3)
- Savings/Stock Value: None, Below 100 DM, [100, 1000] DM, Above 1000 DM
- Employment Length: Below 1 year (including unemployed), [1, 4), [4, 7), Above 7
- Sex/Marital Status: Male Divorced/Single, Male Married/Widowed, Female
- No of Credits at this bank: 1, More than 1
- Guarantor : None, Yes
- Concurrent Credits: Other Banks or Dept Stores, None
- Foreign Worker variable may be dropped from the study
- Purpose of Credit: New car, Used car, Home Related, Other

# 3 CLASSIFICATION

Classification is the way of describing target function variable that connects each attribute set x to pre-defined classes of dependent variable y and this target function is called classification model. This classification model is used for prediction and descriptive modelling. In our case study, it helps to distinguish between “good” and “bad” applicants. We will go through different classification algorithm.

Logistic regression provided us the possibility to classify outcome in binary 0,1 while segregating the issues with linear regression and discriminant analysis. Examples of binary outcomes include the presence and absence of defects, attendance and absenteeism, and student retention versus dropout. Fisher used discriminant function analysis as his approach to predict a binary outcome (which you can think of as group membership). This approach essentially involves finding the linear combination of predictors that maximizes the difference between groups, and thus serves to predict group membership. Discriminant analysis has been a very useful tool but suffers from being technically applicable only with continuous predictors (though many of my textbooks show the use of dichotomous predictors anyway). Logistic regression provides a superior alternative.

This is a function that takes as input the client characteristics and outputs the probability of default.

$$p = expr\ (\beta 0 + \beta 1 \cdot x1 + \cdots + \beta n \cdot xn)$$
$$1 + expr\ (\beta 0 + \beta 1 \cdot x1 + \cdots + \beta n \cdot xn)$$

where in the above □ p is the probability of default □ xi is the explanatory factor i □ βi is the regression coefficient of the explanatory factor i □ n is the number of explanatory variables for each of the existing data points it is known whether the client has gone into default or not (i.e. p=1 or p=0). The aim in the here is to find the coefficients β0, βn such that the model’s probability of default equals to the observed probability of default. Typically, this is done through maximum likelihood

Decision trees are very popular tools for classification and prediction problems. A decision tree is a classifier which recursively partitions the instance space or the variable set. Decision trees are represented as a tree structure where each node can be classified as either a leaf node or a decision node. A leaf node holds the value of the target attribute, while a decision node specifies the rule to be implemented on a single attribute-value. Each decision node gets splits up into two or more nodes according to a classification function of the input attributes-values. The Gini index or the impurity value and the information gain is calculated at every node.

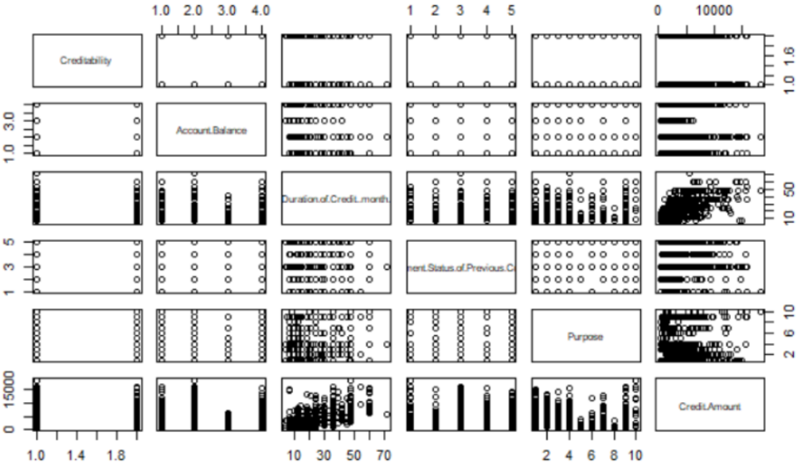
Random forests are collections of decision trees that provide predictions into the structure of data. They are a tool that pulls the power of multiple decision trees in judicious randomization, and ensemble learning to produce predictive models. They provide variable rankings, missing values, segmentations, and reporting for each record to ensure deep data understanding. After each tree is built, all the data is run down the tree. Random forests run very efficiently on large databases, producing accurate results. They handle multiple variables without deletion, giving estimates of the importance of the variables to solve the classification problems.

# 4 Data Cleaning and exploration

Since this file is available online, it is mandatory to do perform some data pre-processing steps and test the quality of the data otherwise it is possible that dataset might not be in suitable format for the classifiers. This procedure will not only cover the insights for the data too but also on the trustworthiness. The data set has all columns as numerical.
We need to convert integer as well as categorical columns into Factor variable for this analysis.

## 4.1 Scatter Plot

Let's do some exploratory data analysis by using PAIRS. When it comes to seeing the correlation among various variables, Scatter plot is the best option.
Rather than using all the variables, only the relevant ones can be used here.



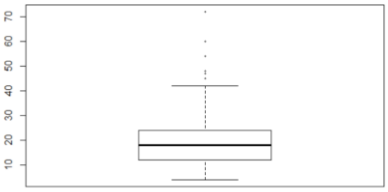
## 4.2 Descriptive Statistics

Descriptive statistics can be performed by aggregating the summary for relevant columns.

```
                    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  Account Balance   1.000   1.000   2.000   2.577   4.000   4.000
Credit Amount       250     1366    2320    3271    3972   18424
Duration of Cr M     4.0     12.0    18.0    20.9    24.0    72.0
```

## 4.3 Skewness for the Duration

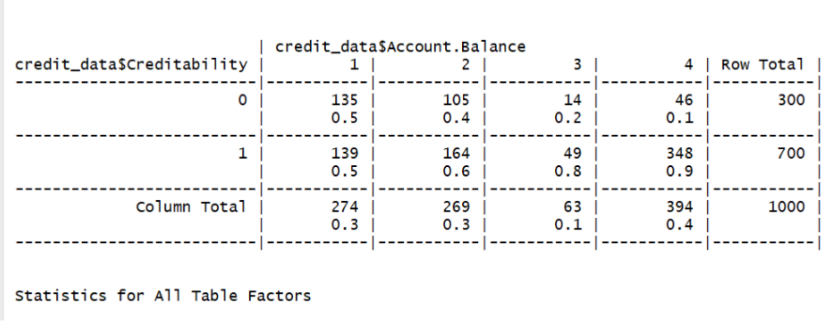
The skewness of the duration of credit can be seen to be skewed in range from 10 to 25. Let's see more of the plots to explore more about the data .



Skewness of Duration.of.Credit..mont

## 4.4 Cross Tables

Cross tables are one of the most famous tables to do exploratory data analysis on such data. It gives us joint and marginal probabilities of the variables depicts the cross-tabulation output from SPSS. For our case study, it can be utilised to know how much of effect does our independent variables have on column “creditability”.

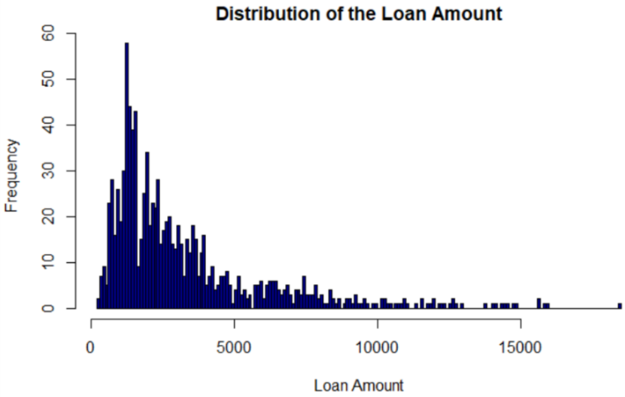


```
                          | credit_data$Account.Balance 
credit_data$Creditability |         1 |         2 |         3 |         4 | Row Total | 
--------------------------|-----------|-----------|-----------|-----------|-----------|
                        0 |       135 |       105 |        14 |        46 |       300 | 
                          |       0.5 |       0.4 |       0.2 |       0.1 |           | 
--------------------------|-----------|-----------|-----------|-----------|-----------|
                        1 |       139 |       164 |        49 |       348 |       700 | 
                          |       0.5 |       0.6 |       0.8 |       0.9 |           | 
--------------------------|-----------|-----------|-----------|-----------|-----------|
             Column Total |       274 |       269 |        63 |       394 |      1000 | 
                          |       0.3 |       0.3 |       0.1 |       0.4 |           | 
--------------------------|-----------|-----------|-----------|-----------|-----------|


Statistics for All Table Factors
```

For example, 30% of 1000 applicants have no account and another 30% have no balance while 40% have some balance in their account. Out of applicants with no balance, creditable are only 135 and rest 139 are Non-creditable.

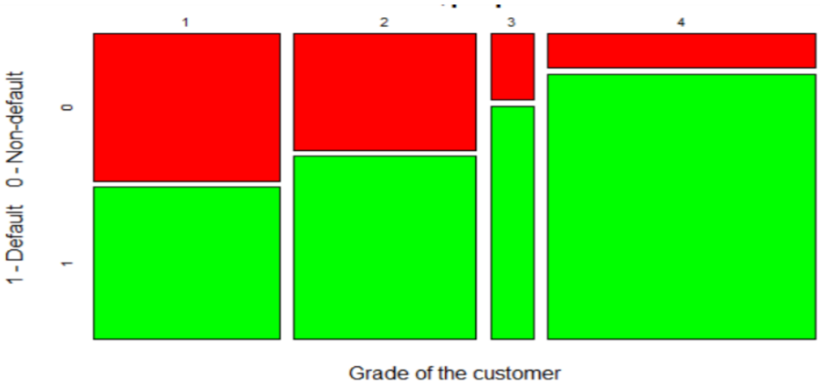
## 4.5 Histograms

This histogram of the loan Amount shows the distribution of the data.



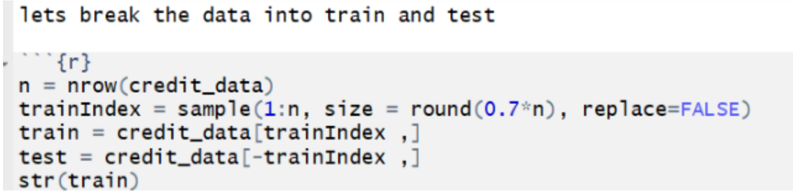
It is clearly evident that the Loan Amount is showing longer trail on left hand side , which means skewed towards LHS. Majority of Loan amount requested lies in the range of 500-5000 .

The graph of Number of Credits vs the creditability below show clear picture of the segmentation.



## 4.5 Creating Train and Test Data

Also, before proceeding further for algorithms, it is considered best practice to divide the dataset in training and test set. Once we train the model on training dataset, we can check the accuracy on test dataset. It is general tendency to divide the data into the ratio of 70:30.

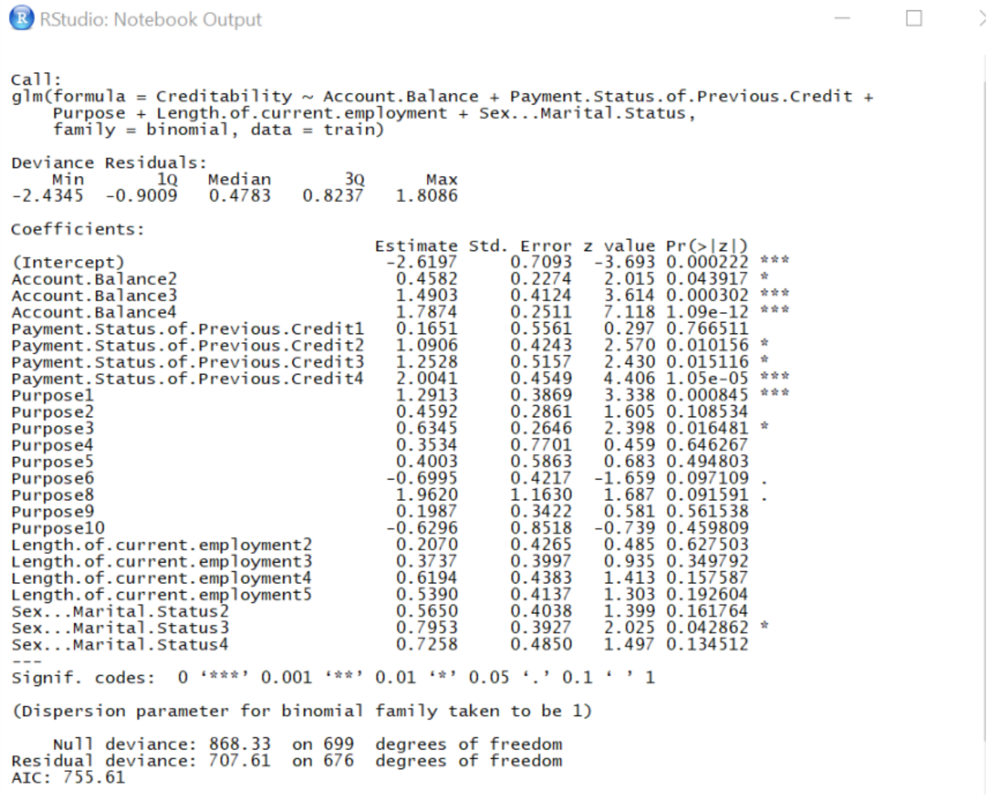


~~~
lets break the data into train and test

```{r}
n = nrow(credit_data)
trainIndex = sample(1:n, size = round(0.7*n), replace=FALSE)
train = credit_data[trainIndex ,]
test = credit_data[-trainIndex ,]
str(train)
~~~

# 5 Implementing Classification

Let’s apply above mentioned algorithms one by one on the data. First algorithm to be applied is logistic Regression. We shall consider only the relevant variables. Now, we fit out model to test and try to do the prediction and we will see how many false and true positive values we have in this method.

## 5.1 Logistic Regression with relevant Variables

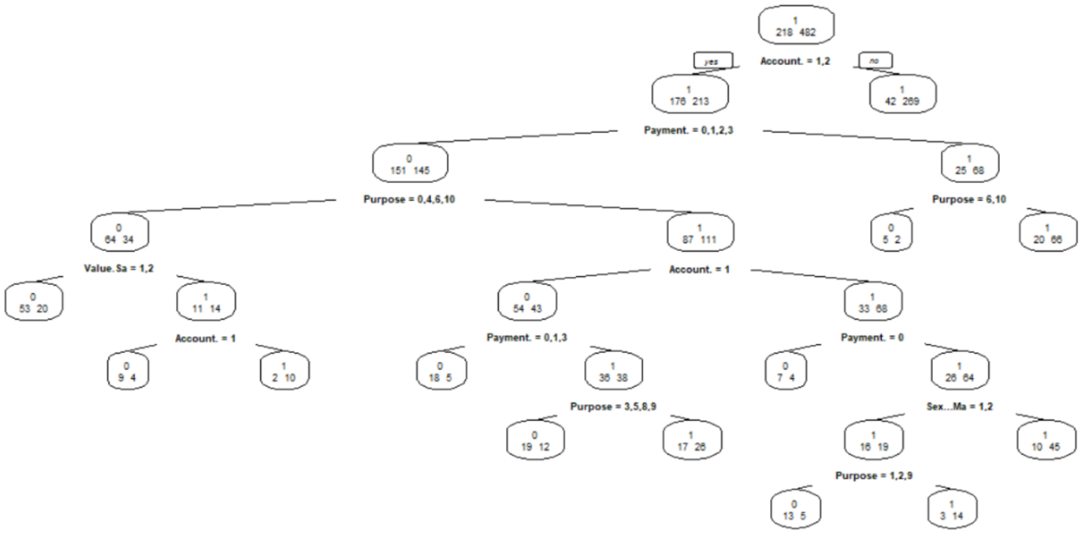


```
Call:
glm(formula = Creditability ~ Account.Balance + Payment.Status.of.Previous.Credit +
    Purpose + Length.of.current.employment + Sex...Marital.Status,
    family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4345  -0.9009   0.4783   0.8237   1.8086

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -2.6197     0.7093  -3.693 0.000222 ***
Account.Balance2                     0.4582     0.2274   2.015 0.043917 *
Account.Balance3                     1.4903     0.4124   3.614 0.000302 ***
Account.Balance4                     1.7874     0.2511   7.118 1.09e-12 ***
Payment.Status.of.Previous.Credit1   0.1651     0.5561   0.297 0.766511
Payment.Status.of.Previous.Credit2   1.0906     0.4243   2.570 0.010156 *
Payment.Status.of.Previous.Credit3   1.2528     0.5157   2.430 0.015116 *
Payment.Status.of.Previous.Credit4   2.0041     0.4549   4.406 1.05e-05 ***
Purpose1                             1.2913     0.3869   3.338 0.000845 ***
Purpose2                             0.4592     0.2861   1.605 0.108534
Purpose3                             0.6345     0.2646   2.398 0.016481 *
Purpose4                             0.3534     0.7701   0.459 0.646267
Purpose5                             0.4003     0.5863   0.683 0.494803
Purpose6                            -0.6995     0.4217  -1.659 0.097109 .
Purpose8                             1.9620     1.1630   1.687 0.091591 .
Purpose9                             0.1987     0.3422   0.581 0.561538
Purpose10                           -0.6296     0.8518  -0.739 0.459809
Length.of.current.employment2        0.2070     0.4265   0.485 0.627503
Length.of.current.employment3        0.3737     0.3997   0.935 0.349792
Length.of.current.employment4        0.6194     0.4383   1.413 0.157587
Length.of.current.employment5        0.5390     0.4137   1.303 0.192604
Sex...Marital.Status2                0.5650     0.4038   1.399 0.161764
Sex...Marital.Status3                0.7953     0.3927   2.025 0.042862 *
Sex...Marital.Status4                0.7258     0.4850   1.497 0.134512
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 868.33  on 699  degrees of freedom
Residual deviance: 707.61  on 676  degrees of freedom
AIC: 755.61
```

The Summary shows the various variables with their coefficients. the positive the value of these coefficients, the higher the risk is. The p value and coefficients of various variables can be analysed from this result.

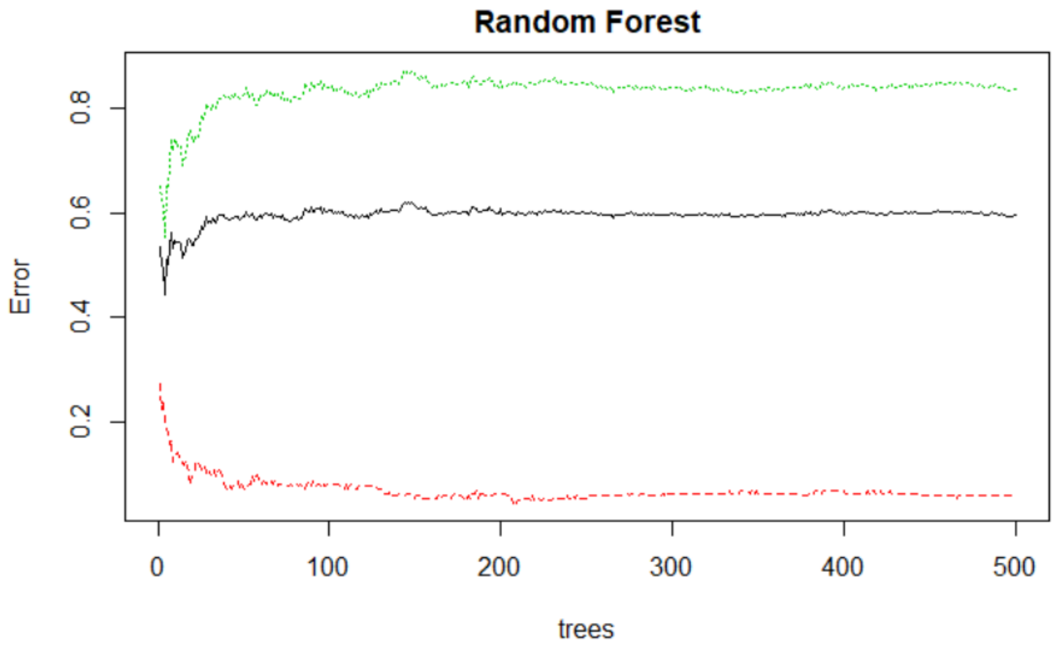
## 5.2 Decision Tree

Tree is generated with value of n as 700. Looking at the tree, the root node is 218 candidates will be credit worthy and 48é will not be. Travelling down the branches, if the candidate has account balance of 1, then we move left and find the spilt on the Payment and mater on Purpose and Value.

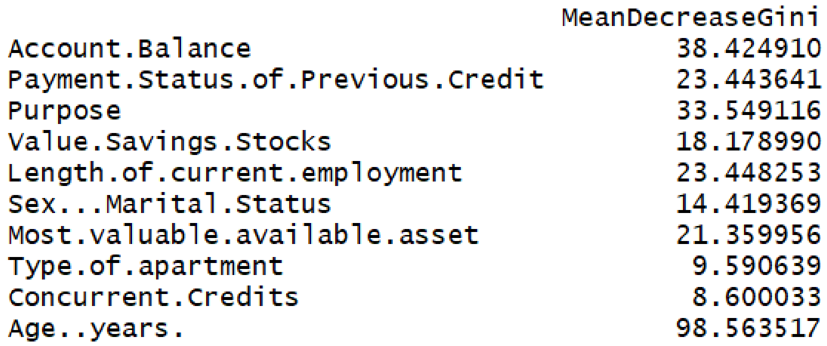


## 5.3 Random Forest

First Dry Run of random forest is run with same variables as for Logistic Model.

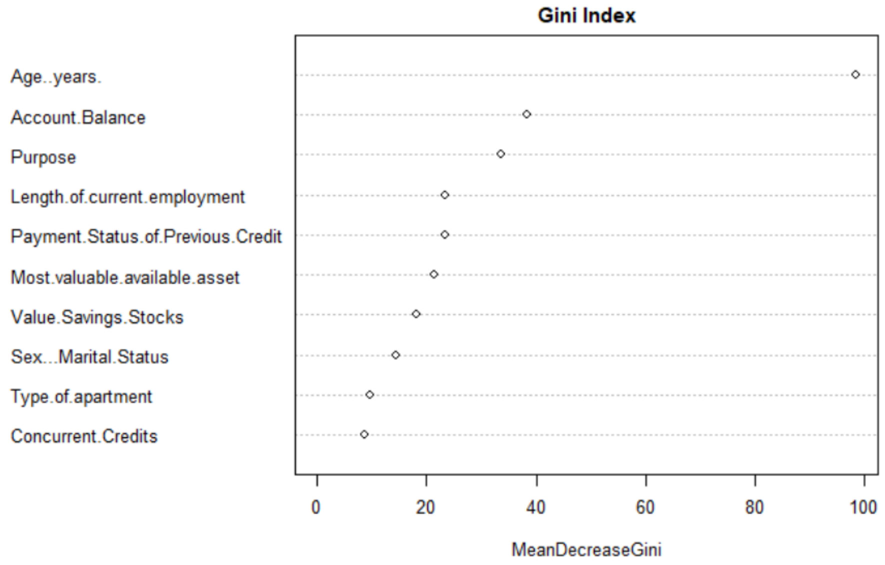


Here we can see Error plot of Random forest. The error of various classes is depicted in the graph. Where the black indicates the out of bags samples. It is evident that the error is lowest after 200 trees. Let's look at the MiniDecreaseGini also.



```
                                   MeanDecreaseGini
Account.Balance                           38.424910
Payment.Status.of.Previous.Credit         23.443641
Purpose                                   33.549116
Value.Savings.Stocks                      18.178990
Length.of.current.employment              23.448253
Sex...Marital.Status                      14.419369
Most.valuable.available.asset             21.359956
Type.of.apartment                          9.590639
Concurrent.Credits                         8.600033
Age..years.                               98.563517
```

The MeanDecreaseGini measures the Gini importance = how important the features are *over all splits* done in the tree/forest - whereas for everyone split the Gini importance indicates how much the Gini criterion = "unequality/heterogeneity" was reduced using this split. Now it is clear that age and account balance and Purpose contributed the most to the obtaining such splits, hence they are most important for this model.



ntree: number of trees to grow; 700; mtry: number of variables selected at each split, By default, mtry = floor of (sqrt of number of independent variables) for classification model.

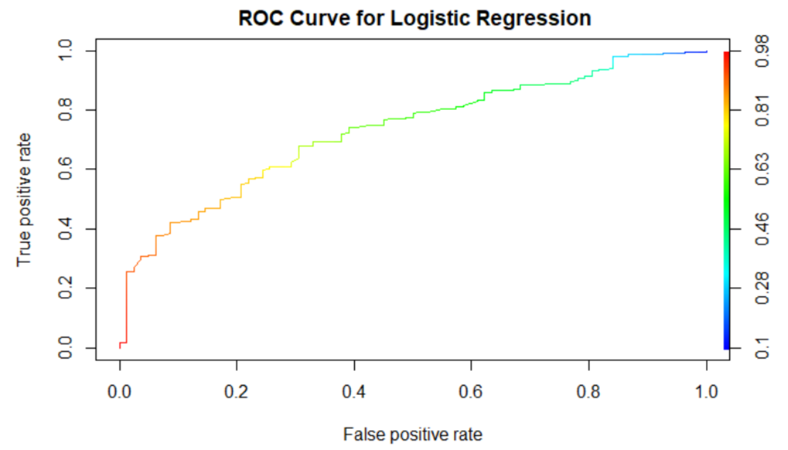
# 6 MODEL ASSESSMENT VIA ROC

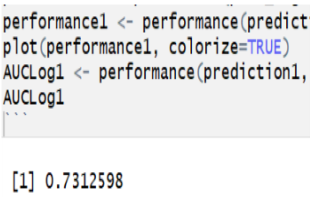
Model evaluation is performed to ensure that a fitted model can accurately predict responses for future or unknown subjects. Without model evaluation, we might train models that over-fit in the training data. To prevent overfitting, we can employ packages, such as caret, miner, and roc to evaluate the performance of the fitted model. Furthermore, model evaluation can help select the optimum model, which is more robust and can accurately predict responses for future subjects.

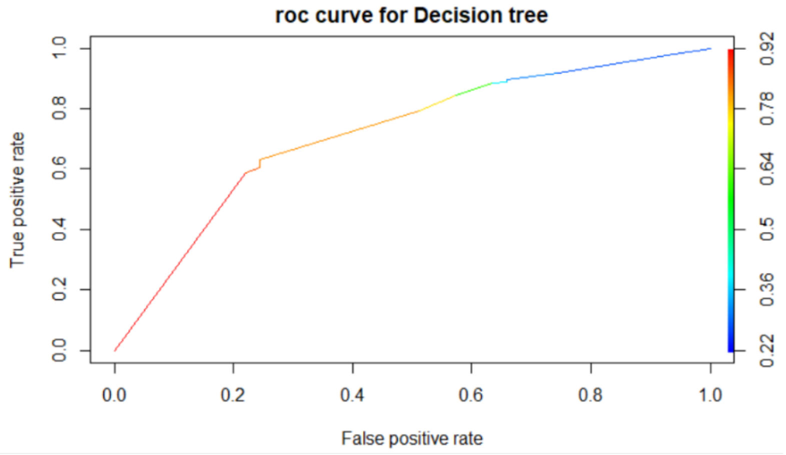
Accuracy is checked via ROC curve. ROC curve includes in one graph the performance of the model for all possible cut-off values. ROC stands for “Receiver Operating Characteristic”. The ROC curve as shown below considers systematically all cut-off values for the PD from 0 to 100%. For each cut-off value it should be able to measure the number of Goods and Bads. We shall also see AUC that is area under curve for the accuracy's greater the value of AUC, the better the model is.

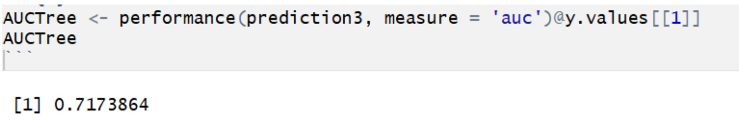
Let's see the AUC for all the algorithms too.

## 6.1 Logistic Regression with relevant Variables





```
performance1 <- performance(predict
plot(performance1, colorize=TRUE)
AUCLog1 <- performance(prediction1,
AUCLog1
```
```

```
[1] 0.7312598
```

The AUC value is 0.7312.

Now looking at this above ROC Curve, we see that the curve is more inclined towards right hand side and it is not so much in the top side of the area. Neither does it follow the 45 degrees of the angle of ROC space.

Let's move onto CART (Classification and Regression Trees) to see if we can make better predictions on this highly unbalanced data with Decision Trees.

## 6.2 Decision Tree


```

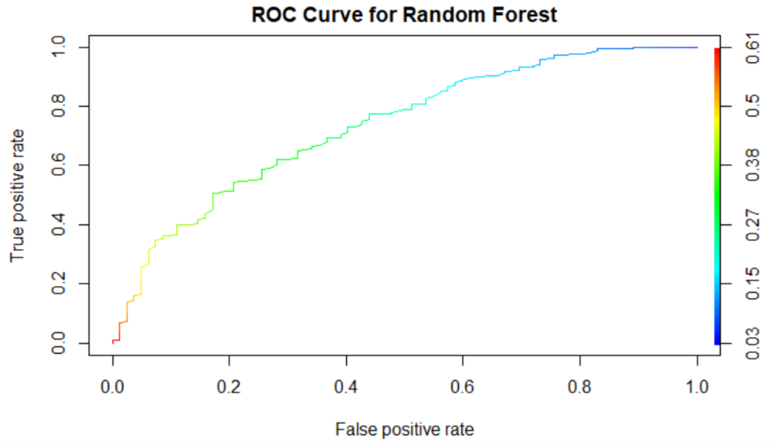


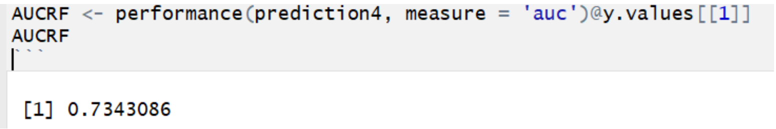
```
AUCTree <- performance(prediction3, measure = 'auc')@y.values[[1]]
AUCTree
```

```
[1] 0.7173864
```

Well the AUC value is bit lower than that of logistic, hence we will not use this model for your case study. Well the ROC curve is quite like that of Logistic .it' cleaner although the area coverage and angle is quite similar.

## 6.3 Random Forest



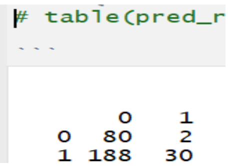


```
AUCRF <- performance(prediction4, measure = 'auc')@y.values[[1]]
AUCRF
```

```
[1] 0.7343086
```

The AUC for this model comes out to be 0.73430 which is quite close to AUC value for Logistic Model. Well in this graph, we can see bit better angle than the other graphs.

We can also check sensitivity a specificity by adjusting the threshold value of the AUC model. Sensitivity and specificity helps us to statistically measure the performance of the classifier. Sensitivity: (also called the true positive rate) measures the proportion of positives that are correctly identified. Specificity: (also called the true negative rate) measures the proportion of negatives that are correctly identified. with the threshold value of 0.5, we get this,



```
# table(pred_r
```

```
      0   1
  0  80   2
  1 188  30
```

This means Sensitivity is 30/ 218 =0.137

Specificity is 80/82=0.97.

# 7 CONCLUSION

The Random forest model, at a cut off 0.50, give us the Specificity with drastic results of 97%. And the Overall Accuracy of the model is also not compromised much as it is still considerably better at 73%.

# 8 COST PROFIT CONSIDERATION IN FINANCIAL SECTOR

These may be the best scores we can come up with using these models, but are the results acceptable for determining the credit-worthiness of a loan applicant? That depends on the credit standards being used by the lending institution. these statistical decisions must be translated into profit consideration for the bank

At best, it looks like our models give us an 73% chance of lending to good credit risks. Let us assume that a correct decision of the bank would result in 20% profit at the end of 5 years. An If the bank can predict an applicant as credit worthy or not credit worthy, then this decision considered to be good or correct decision. Out of 1000 applicants, 70% are creditworthy. A loan manager without any model would incur [0.7*0.20 + 0.3 (-1)] = - 0.16 or 0.16-unit loss

For every $1 million in loans, at best we might expect to be repaid $730,000. On average, we would expect to recover around $700,000 in principal. In other words, according to our analysis, there is between a 65% and 75% chance we will recapture our $1 million loan, depending on the modelling method we use.

As we add loan applicants to our data bases, we would want them to cluster in the darkest area of the high-density plot if we are going to consider them good credit risks.

Unless we charge a lot of interest to cover our losses, we might need better models.

Banks and Financial Institutions can use this model to create a Loan Acceptance Strategy for every Loan Applications and minimise the Bad Loan Error Rate from their portfolio.

# 9 REFERENCES

German Credit Data Explained. Retrieved from https://onlinecourses.science.psu.edu/stat857/node/216

V. Cherkassky and F. Mulier. *Learning from Data: Concepts, theory and Methods.*Wiley Interscience, 1998.

P.Dominogos. (1999). *Data Minining and Knowledge Discovery* , The Role of Occam's Razor in Knowledge Discovery.