2018

# Credit Default Prediction

ANALYSIS OF GERMAN CREDIT DATA

AYUSHI CHUDGOR
RIN: 661853790

RENSSELAER POLYTECHNIC INSTITUTE

## Introduction:

The beginnings of 2007-08 credit crunch appeared early in 2007 was considered to be local problems among lower-quality U.S. mortgage lenders. An increase in subprime mortgage defaults in February 2007 had caused filing for Chapter 11 bankruptcy for April New Century Financial. One would ask what is subprime mortgage? Subprime mortgage is availability of loan to those people which would otherwise not be considered for loan. Lenders in financial markets did not wish to carry this risk and bundled this loans into different groups based on the assumed safety of loans. These groups are then called tranches and rated by different rating agencies such as Fitch Ratings, Standard & Poor's, and Moody's. Initially, all ratings agencies gave them "safe" AAA status usually given to sovereign loans. When the subprime loans started making losses, this forced them to reanalyze and downgrade them. These subprime securities are usually held by hedge funds backed by multinational banks such as BNP Paribas, Bear Sterns, UBS, Morgan Stanley etc., Trading floors of all multinational banks around the world traded or held these securities. After being downgraded, securitized products lost most of its value and banks had to face massive losses.

Multinational corporations were termed as "too big to fail" when in 2000s Gramm-Leach-Bliley Act was passed. This Congress approved legislation which allowed commercial and investment banks to merge and form institutions of unprecedented size and global reach. Common men held their life earnings with these banks and news of losses caused widespread panic. Financial systems are made one on factor often neglected – investor confidence. When this confidence started crumbling, it caused massive run on the banks. This panic snowballed into banks all around the world failing and booking huge losses on same day. We started to look at signs of systemic risk and entire countries like Italy, Greece on brink of bankruptcy. One has to ask what went wrong so terribly? It is often suggested if loan applicants had been better vetted, the crisis would not have as severe. It would have allowed systemic failing of financial systems around the world to be nipped in the bud.

In wake of Financial Crisis 2007-08, led to development of field of credit risk management. Credit risk is defined as probability of loss due to a borrower's failure to make required payments on given loan. A major challenge for financial institutions is mitigating losses by understanding the required loan loss reserves and capital adequacy. Global financial crisis of 2008 and following credit crunch that, put credit risk management in regulatory spotlight. Regulators started demanding more thorough testing of their credit models and demanded higher transparency from financial institutions. They wanted to know if a bank has thorough knowledge of customers and credit risk associated with them.

My interest in quantitative finance specifically lies in field of credit risk analytics. Today, banks are trying to leverage this information and make better decisions which support the

financials of the company. Through my term project, I aim to explore the very first step of evaluation of loan applicant's credit risk. I aim to predict if loan applicant's credit risk is - good or bad based on their available demographic and socioeconomic information.

## Data Description:

- Credit default prediction requires updated information of loan applicant. It is proprietary information not easily available, for this project I decided to use snapshot information available. This dataset is obtained from open- source UCI Machine Learning Repository.
- The dataset available on UCI is in two forms. The first form where it is records the categories while second one is in form of converted numeric values. For the purpose of this purpose, I had decided to use numeric dataset as it reduces one step of data preprocessing of converting categories into numeric values.
- The aim to classify whether credit is good or bad. Essentially, it is a classification problem. Data is required as numeric values to be able to convert them to factors.
- It has 20 attributes and 1000 instances.
- Attribute description for German Credit data:
  a. Attribute 1: (qualitative) Status of existing checking account
     A11 : … < 0 DM
     A12 : 0 <= … < 200 DM
     A13 : … >= 200 DM / salary assignments for at least 1 year
     A14 : no checking account
  b. Attribute 2: (numerical) Duration in month
  c. Attribute 3: (qualitative) Credit history
     A30 : no credits taken/all credits paid back duly
     A31 : all credits at this bank paid back duly
     A32 : existing credits paid back duly till now
     A33 : delay in paying off in the past
     A34 : critical account/other credits existing (not at this bank)
  d. Attribute 4: (qualitative) Purpose
     A40 : car (new)
     A41 : car (used)
     A42 : furniture/equipment
     A43 : radio/television
     A44 : domestic appliances
     A45 : repairs
     A46 : education
     A47 : (vacation - does not exist?)
     A48 : retraining
     A49 : business
     A410 : others
  e. Attribute 5: (numerical) Credit amount
  f. Attribute 6: (qualitative) Savings account/bonds
     A61 : … < 100 DM
     A62 : 100 <= … < 500 DM
     A63 : 500 <= … < 1000 DM
     A64 : .. >= 1000 DM

A65 : unknown/ no savings account

g. Attribute 7: (qualitative) Present employment since
   A71 : unemployed
   A72 : … < 1 year
   A73 : 1 <= … < 4 years
   A74 : 4 <= … < 7 years
   A75 : .. >= 7 years
h. Attribute 8: (numerical) Installment rate in percentage of disposable income
i. Attribute 9: (qualitative) Personal status and sex
   A91 : male : divorced/separated
   A92 : female : divorced/separated/married
   A93 : male : single
   A94 : male : married/widowed
   A95 : female : single
j. Attribute 10: (qualitative) Other debtors / guarantors
   A101 : none
   A102 : co-applicant
   A103 : guarantor
k. Attribute 11: (numerical) Present residence since
l. Attribute 12: (qualitative) Property
   A121 : real estate
   A122 : if not A121 : building society savings agreement/life insurance
   A123 : if not A121/A122 : car or other, not in attribute 6
   A124 : unknown / no property
m. Attribute 13: (numerical) Age in years
n. Attribute 14: (qualitative) Other installment plans
   A141 : bank
   A142 : stores
   A143 : none
o. Attribute 15: (qualitative) Housing
   A151 : rent
   A152 : own
   A153 : for free
p. Attribute 16: (numerical) Number of existing credits at this bank
q. Attribute 17: (qualitative) Job
   A171 : unemployed/ unskilled - non-resident
   A172 : unskilled - resident
   A173 : skilled employee / official
   A174 : management/ self-employed/highly qualified employee/ officer
r. Attribute 18: (numerical) Number of people being liable to provide maintenance for
s. Attribute 19: (qualitative) Telephone
   A191 : none
   A192 : yes, registered under the customer's name
t. Attribute 21: (qualitative) foreign worker
   A201 : yes
   A202 : no

- It includes both quantitative and qualitative attributes in the dataset. The attributes included in the dataset are:

1. Quantitative Attributes:
   Duration (in months), Credit amount, installment rate percentage, present residence since, age(years), number of existing credits at bank, Number of people being liable to provide maintenance for,
2. Qualitative Attributes:
   Status of existing checking account, Credit history, Purpose, Savings accounts, Present employment since, Personal status and sex, Debtors/guarantors, property, installment plans, Housing, job, Telephone, Foreign worker.

- *Following is the snapshot of structure of my dataset used in model prediction:*

```
'data.frame':    1000 obs. of 21 variables:
 $ Account.Balance                    : int  1 1 2 1 1 1 1 1 4 2 ...
 $ Duration.of.Credit..month.         : int  18 9 12 12 12 10 8 6 18 24 ...
 $ Payment.Status.of.Previous.Credit : int  4 4 2 4 4 4 4 4 4 2 ...
 $ Purpose                            : int  2 0 9 0 0 0 0 0 3 3 ...
 $ Credit.Amount                      : int  1049 2799 841 2122 2171 2241
3398 1361 1098 3758 ...
 $ Value.Savings.Stocks               : int  1 1 2 1 1 1 1 1 1 3 ...
 $ Length.of.current.employment       : int  2 3 4 3 3 2 4 2 1 1 ...
 $ Instalment.per.cent                : int  4 2 2 3 4 1 1 2 4 1 ...
```

## Dataset Analysis:

- The dataset includes both quantitative and qualitative attributes in the dataset.
- The dataset is checked for missing values using is.na() and was found that it does not have any missing values.
- Using *summary(),* we check statistics of the dataset. It provides us with minimum value, maximum value, first and third quartile, mean and median of the attribute for entire dataset. We check it as it helps us in flooring/ capping operations.
- *Below is snapshot of summary for attribute 'credit.amount':*

| Credit.amount |
| --- |
| Min. : 250 |
| 1st Qu.: 1366 |
| Median: 2320 |
| Mean : 3271 |
| 3rd Qu.: 3972 |
| Max. :18424 |

- For quantitative attributes, using boxplots outliers were detected. Outliers were handled using flooring and capping.
- *Below is an example for flooring/capping for non-categorical attribute 'Duration.in.months':*

Fig. Before outlier treatment

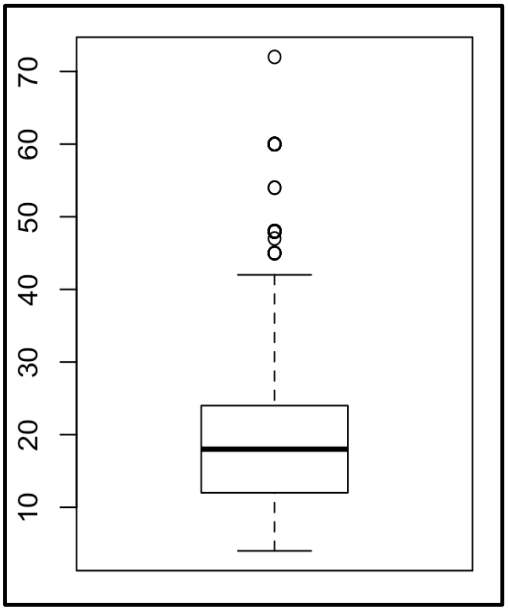
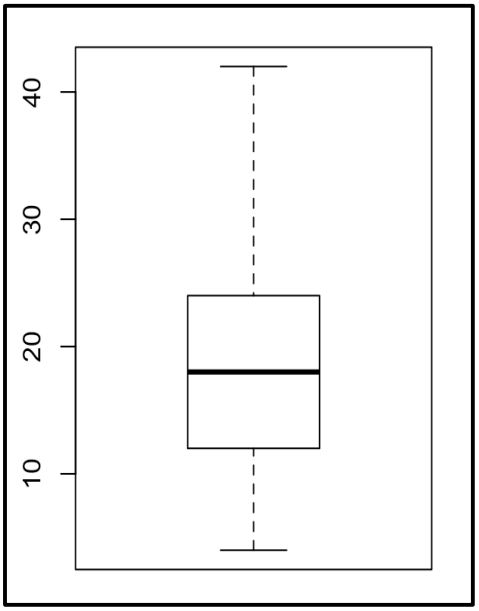
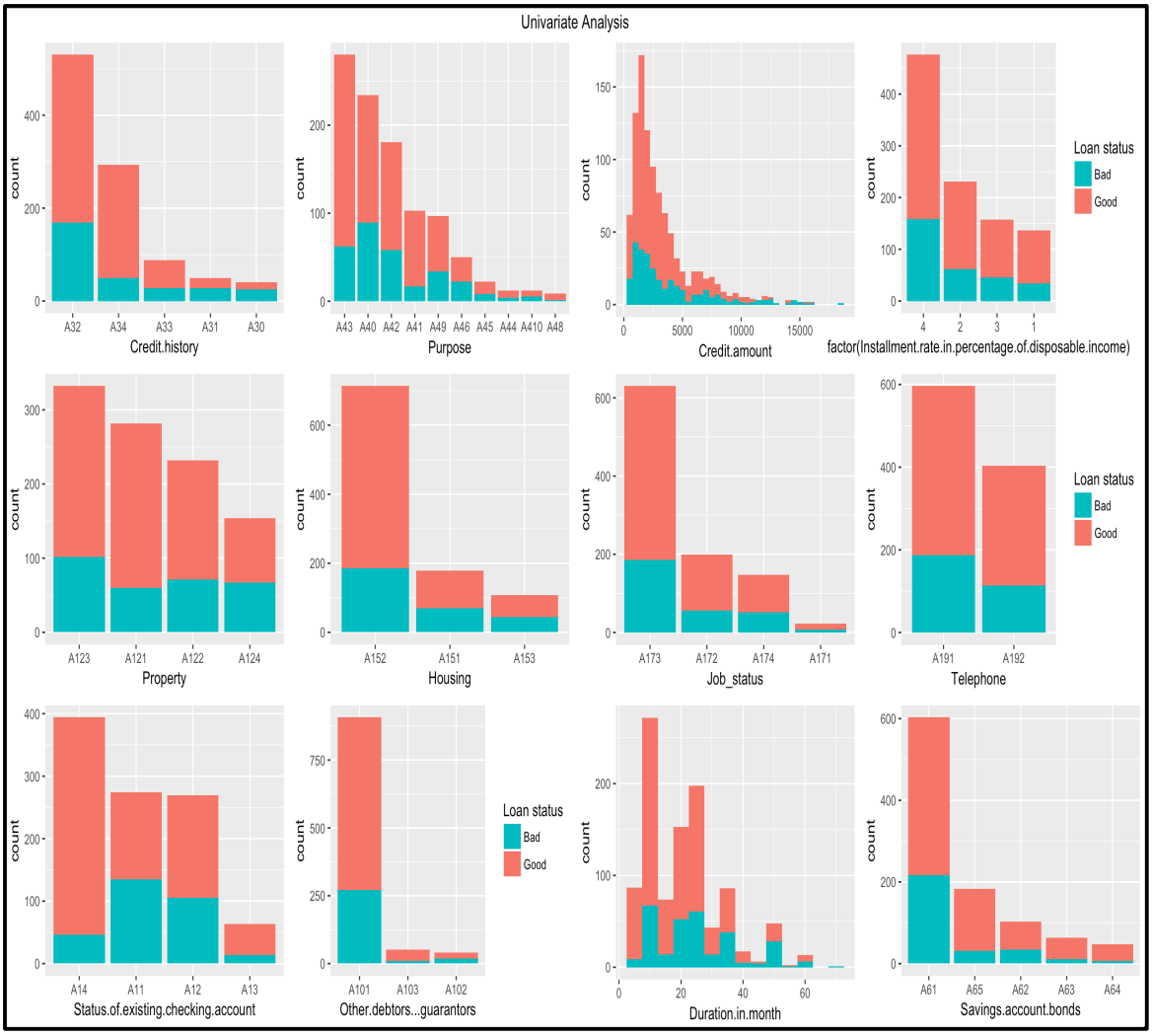



Fig. After outlier treatment



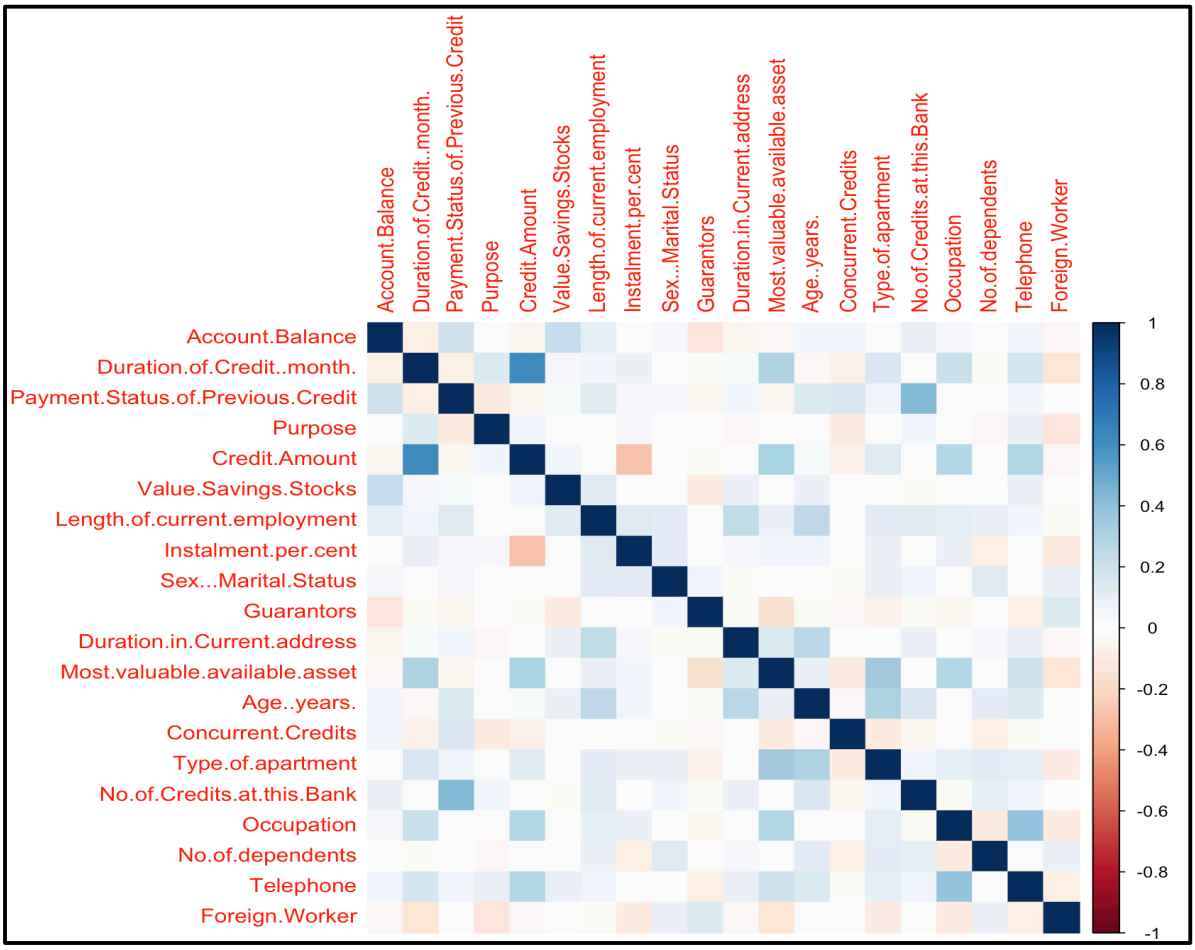

- Using *ggplot2()* and *gridExtra()* library, each attribute was plotted with division of values to factor of credibility if it is bad or good.
- *Below are those categorical attributes plotted using bar plot:*




- Some of the interpretation regarding high default from above plot is as follows:

| **Attribute** | **Value that point to reason for high default** |
|---|---|
| Job | A173 – Skilled employee / Official |
| Housing | A152 – Own house |
| Checking Account | A11, A12 (DM less than 0 and 200) |
| Credit History | A32 (Existing credits paid back till now); A34 (Other credits) |
| Property (loan) | A123 (car or other) |
| Purpose | A40 (New car) ； A42 (Furniture or equipment) |
| Duration of loan | 10 months |
| Instalment Rate | Instalment rate of 4% of disposable income |
| Credit Amount | For credits above 7500 & for credits between 1000 & 2000 |

- Predicting whether credit is good or bad is a classification problem. Before we need to split the data for training and testing purpose. For this step, I take randomly 70% data for training and rest 30% for testing. To ensure we get same results every time, we also set seed for our program.
- The dataset has 20 attributes & need to check the pairwise correlation between them before we proceed.
- Correlation matrix is plotted using the *corrplot()* library. Positive correlations are displayed in blue while negative correlations in red. Color intensity is proportional to the strength of correlation coefficients.
- *The figure represents pairwise correlation matrix between dataset attributes:*



## Model Development:

1. **Logistic Regression:**

- Logistic regression is a method used for fitting a regression curve when y is a categorical variable.
- The predictors can be continuous, categorical or a mix of both.
- We can simplest case scenario *y* is binary meaning that it can assume either the value 1 or 0. The function to be called is *glm()* and the fitting process is not different from that used in linear regression.
- The AIC of 718.7 obtained is best after applying StepAIC to remove the insignificant variables.

*Start: AIC=816.31*
*Default_status ~ Duration.in.month + Credit.amount +*
*Installment.rate.in.percentage.of.disposable.income +*
*Present.residence.since + Age.in.Years + Number.of.existing.credits.at.this.bank +*
*Number.of.people.being.liable.to.provide.maintenance.for*

*Final Step: AIC=812.7*
*Default_status ~ Duration.in.month + Credit.amount +*
*Installment.rate.in.percentage.of.disposable.income +*
*Age.in.Years*

- To further reduce multicollinearity between the variable, we use variance inflation factor *vif()*.
- Variables with high VIF and are insignificant were removed. If the VIF values are all less than the required threshold than their P-values were checked. Variables with high P-values and VIF are selected for removal. Every time an iteration is done AIC values are checked to see if there is any abnormal increase.
- We stop removing variables when AIC starts to increase. The intuition behind it is from that particular step all variables are significant.

*Call:*

*glm(formula = Creditability ~ Value.Savings.Stocks + Length.of.current.employment +*
*Duration.of.Credit..month. + Credit.Amount + Age..years.+ No.of.Credits.at.this.Bank ,*
*family = binomial(link = "logit"), data = xtrain)*

*Deviance Residuals:*

*Min 1Q Median 3Q Max*

*-2.4786 -0.7860 0.4505 0.7467 1.7556*

*Coefficients:*

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| *(Intercept)* | *-4.181e+00* | *1.216e+00* | *-3.440* | *0.000582* | |
| *Duration.of.Credit..month.* | *-2.437e-02* | *1.113e-02* | *-2.190* | *0.028530* | |
| *Credit.Amount* | *-5.948e-05* | *5.174e-05* | *-1.150* | *0.250287* | |
| *Value.Savings.Stocks* | *1.946e-01* | *6.743e-02* | *2.886* | *0.003906* | ** |
| *Length.of.current.employment* | *1.709e-01* | *8.780e-02* | *1.946* | *0.051648* | . |
| *Age..years.* | *-2.329e-03* | *1.046e-02* | *-0.223* | *0.823768* | |
| *No.of.Credits.at.this.Bank* | *-2.277e-01* | *1.917e-01* | *-1.188* | *0.235022* | |

*(Dispersion parameter for binomial family taken to be 1)*
*Null deviance: 856.9 on 699 degrees of freedom*
*Residual deviance: 676.7 on 679 degrees of freedom*
*AIC: 718.7*

*Number of Fisher Scoring iterations: 5*

Applying this model on test data, the mean is calculated by finding the difference between quality variable of test dataset and squared value of predicted quality. Mean squared error of 0.53 which is high but still acceptable.

- Accuracy obtained is:

| Threshold | Accuracy |
|---|---|
| 50% | 61% |
| 75% | 45% |

- From this we can interpret that as we increase our threshold from 50% to 75%, the accuracy dramatically drops from 61% to 45% which is not desirable.

2. **Discriminant Analysis:**

- Linear discriminant analysis is a method used to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination is used as a linear classifier for dimensionality reduction and then classification.
- It is carried out in R using MASS package.
- When we fit our dataset to training algorithm, we get the following output:
  *Call:*

*lda(Creditability ~ Value.Savings.Stocks + Length.of.current.employment +*
*Duration.of.Credit..month. + Credit.Amount + Age..years., data = xtrain)*

*Prior probabilities of groups:*

| 0 | 1 |
|---|---|
| 0.3014286 | 0.6985714 |

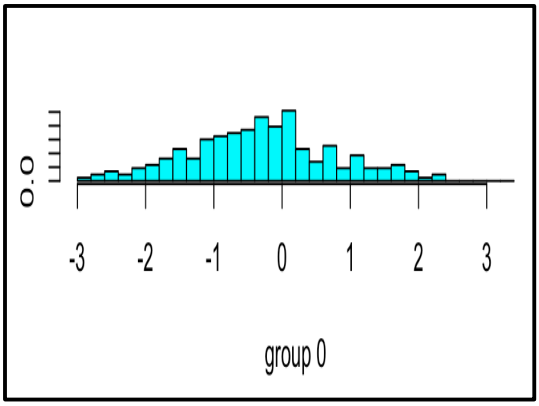
*Group means:*

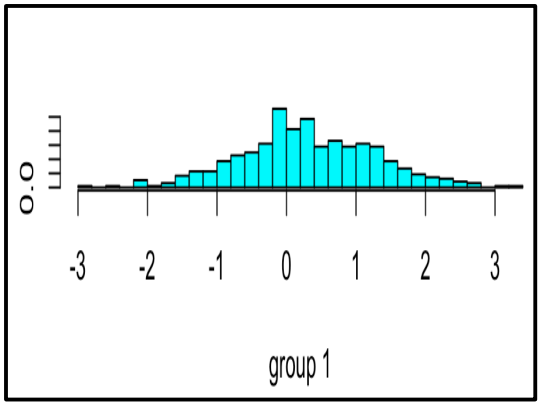
| | Value.Savings.Stocks | Length.of.current.employment | Duration.of.Credit..month. | Credit.Amount | Age..years. |
|---|---|---|---|---|---|
| 0 | 1.758294 | 3.161137 | 25.02844 | 3782.009 | 33.62085 |
| 1 | 2.261759 | 3.443763 | 19.65031 | 3082.577 | 35.40900 |

*Coefficients of linear discriminants:*

| | LD1 |
|---|---|
| Value.Savings.Stocks | 3.393937e-01 |
| Length.of.current.employment | 3.088581e-01 |
| Duration.of.Credit..month. | -7.191443e-02 |
| Credit.Amount | 4.657723e-05 |
| Age..years. | 8.708964e-03 |

- Following are the plots obtained of creditability variable after they are trained using *lda()* from MASS library:




- Confusion Matrix:

| Pred | 0 | 1 |
|---|---|---|
| 0 | 31 | 42 |
| 1 | 59 | 174 |

- Accuracy is calculated from above confusion matrix as follows:
- Accuracy = sum of diagonal elements divided by total number of elements:
- Accuracy = 205/300 = 68.33%
- The accuracy is improved to 68.33% in linear discriminant analysis compared to 61% in logistic regression.

## 3. Decision Trees(Classification):

- Decision trees are used in that situation where data is divided into groups than investigating a numerical response and the relationship predicted response has to a set of descriptor variables.
- The algorithm underlying this method is to first find the variable that does the best job of splitting the data into two groups. This step is repeated with the other variables and produces a tree graph, where each split represents a decision. The tree where the maximum no. of observations is correctly classified is selected.
- For decision trees, we use 2 levels to predict quality which is 'good and bad'. I made a new variable that stores response of these 2 levels. Here, creditability is made a factor for the model. I plot the decision tree using *rpart()* library available.
- The results of training decision trees is displayed using *printcp()*.

*Classification tree:*
*rpart(formula = Creditability ~ ., data = xtrain, method = "class")*

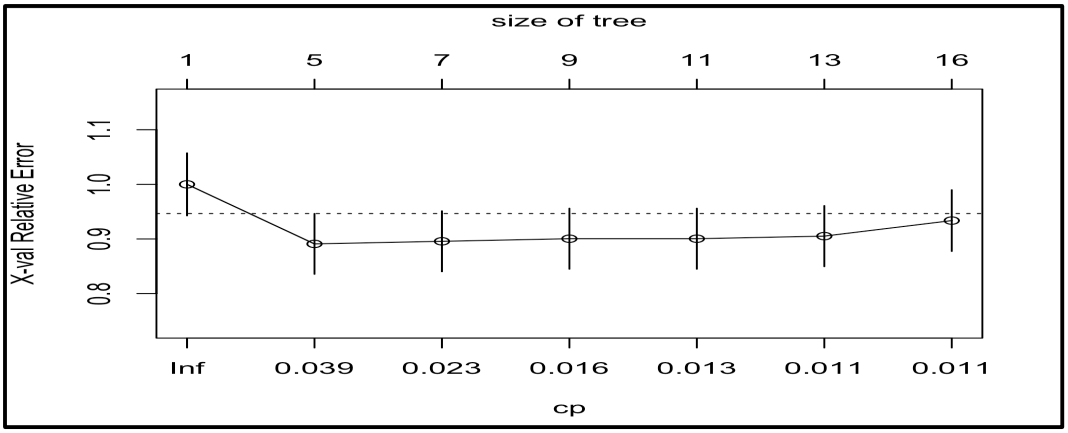
*Variables actually used in tree construction:*
*[1] Account.Balance Age..years. Concurrent.Credits*
*[4] Credit.Amount Duration.in.Current.address Duration.of.Credit..month.*
*[7] Instalment.per.cent Length.of.current.employment*
*Most.valuable.available.asset*
*[10] Payment.Status.of.Previous.Credit*

*Root node error: 211/700 = 0.30143*

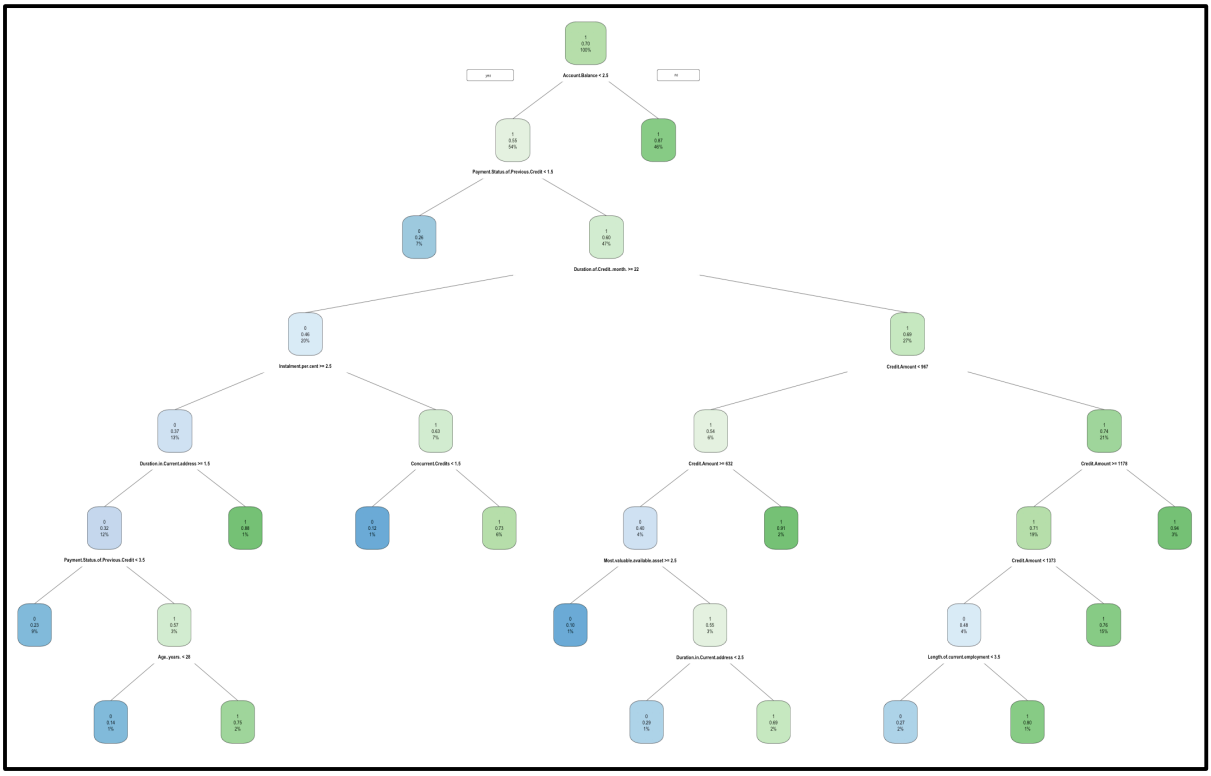
*n= 700*

| | CP | nsplit | rel error | xerror | xstd |
|---|---|---|---|---|---|
| *1* | *0.054502* | *0* | *1.00000* | *1.00000* | *0.057539* |
| *2* | *0.028436* | *4* | *0.77725* | *0.90995* | *0.055944* |
| *3* | *0.018957* | *6* | *0.72038* | *0.90521* | *0.055853* |
| *4* | *0.014218* | *8* | *0.68246* | *0.87678* | *0.055291* |
| *5* | *0.011848* | *10* | *0.65403* | *0.86256* | *0.055001* |
| *6* | *0.011058* | *12* | *0.63033* | *0.86256* | *0.055001* |
| *7* | *0.010000* | *15* | *0.59716* | *0.89573* | *0.055669* |

- *Plotcp()* is used to visualize cross validation results. It is interpreted as the line drawn by the *plotcp*() represents the highest cross-validated error less than the minimum cross-validated error plus the standard deviation of the error at that tree.




- Decision tree generated is plotted using *rpart.plot()* library.



- <u>Confusion matrix</u>:

| **Pred** | **0** | **1** |
|---|---|---|
| **0** | 33 | 29 |
| **1** | 56 | 182 |

- Accuracy is calculated from above confusion matrix as follows:
- Accuracy = sum of diagonal elements divided by total number of elements:
- Accuracy = 215/300 = 71.67%

4. **Random Forests:**

- Decision trees are simplistic and most of times outperformed by other algorithms. Random Forests are one way to improve its performance.
- The algorithm starts by building out trees similar to the way a normal decision tree algorithm works.
- However, every time a split has to made, it uses only a small random subset of features to make the split instead of the full set of features. It builds multiple trees using the same process, and then takes the average of all the trees to arrive at the final model.
- This works by reducing the amount of correlation between trees, and thus helping reduce the variance of the final tree.
- First, we provide 70% of data as test data for training for 300 training trees. The results of training random forests obtained are displayed:

*Call:*

*randomForest(formula = Creditability ~ ., data = xtrain, ntree = 300, importance = T, proximity = T, mtry = 5)*

*Type of random forest: classification*

*Number of trees: 300*

*No. of variables tried at each split: 4*

*OOB estimate of error rate: 24.14%*

*Confusion matrix:*

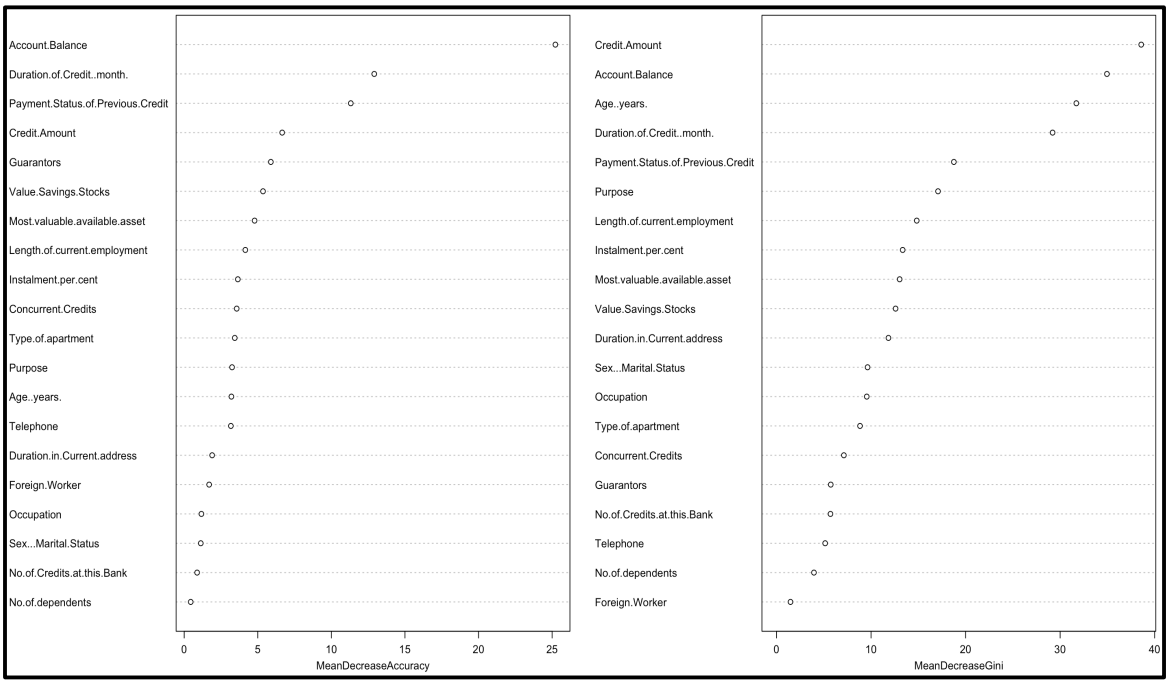
| | *0* | *1* | *class.error* |
|---|---|---|---|
| *0* | *86* | *125* | *0.59241706* |
| *1* | *44* | *445* | *0.08997955* |

- To improve the model, different values were tried for ranging from 3 to 7 for 'mtry'. *Mtry* refers to the number of variables randomly sampled as candidates at each split. The best output received at *mtry* value of 5.
- After training step, we need to apply to the trained model to test data using *predict()*. Output received is as follows:
- Confusion matrix:

| Pred | 0 | 1 |
|---|---|---|
| 0 | 39 | 19 |
| 1 | 50 | 192 |

- Accuracy is calculated from above confusion matrix as follows:
- Accuracy = sum of diagonal elements divided by total number of elements
- Accuracy = 231/300 = 77%

- Below is variable importance plot for train data model. From this we interpret, the account balance, duration of credit month and payment status of previous credit are the most important predictor variables used.




- The main aim is to reduce the loss to be bared by the bank. Every time a credit decision is incorrectly predicted, it results in loss of profit but every correct decision yields full profit. This is how cost matrix is used coupled with confusion matrix helps us predict required amount of profit/loss.
- Representation of cost matrix is given as:

| Good = 1, Bad = 2 | Actuals | 1 | 2 |
|---|---|---|---|
| Predicted | 1 | 0 | 1 |
| | 2 | 3 | 0 |

- In this situation, it is worse to classify a customer as good when they are bad (3) than classifying a customer as bad when they are good (1).

**Conclusion:**

- The following reasons is why I choose random forest model whether credit approval decision is good or bad:
  a. Logistic regression provided unsuitable results even after selecting best features for modeling using StepAIC and variance inflation factor.
  b. Decision trees have high variance when they utilize different training and test sets of the same data, as it leads to over fitting of training data. Leading to poor performance on unseen data and limiting its use in predictive modeling.

c. Random forest aims to reduce the correlation captured in decision trees by choosing only a subsample of the feature space at each split. It uses a stopping criteria for node splits while pruning the trees and thereby making them uncorrelated.
d. From above model predictions, summary of accuracy provided by these models developed are as follows:

| Methods | Accuracy |
|---|---|
| 1. Logistic Regression | 50% Threshold: 61%<br>75% Threshold: 45% |
| 2. Discriminant Analysis | 68.33% |
| 3. Decision Trees | 71.67% |
| 4. Random Forests | 77% |

- Initially, the model prediction was carried out using outliers present in data. It affected the accuracy of model prediction severely. Upon outlier treatment, my model started performing better than its prior performance. From this we learn that, outliers can be problematic even if it is not bad values.
- Upon completion of this assignment, I learnt data preprocessing specifically to type of attributes needs to be carefully handled depending on problem trying to be addressed.
- Starting the implements from classification methods like regression to random forests helps better understand working of the data. This in turn helps in correct selection of assumptions required in prediction of model.
- Upon selecting random forests model, it will be able to correctly predict with 77% accuracy whether credit approval decision is good or bad.
- For future work, for example in regression model can be improved by working on sensitivity and specificity rather than just focus on accuracy. This selection depends on business problem trying to be solved and threshold level required.
- The model can be further improved used advanced analytics techniques such as ensemble methods (bagging/boosting), neural networks.
- For selecting appropriate attributes for logistic regression to further improve its accuracy, more advanced feature selection techniques such as PCA could be used.

***References:***


1. *An Introduction to Statistical Learning with Applications in R,* *James, G., Witten, D., Hastie, T., Tibshirani, R.*
2. *https://www.r-bloggers.com/how-to-implement-random-forests-in-r/*
3. *https://www.r-bloggers.com/using-decision-trees-to-predict-infant-birth-weights/*
4. *https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/*
5. *Data analytics lecture notes*
6. *Packages Documentation: rpart, rpart.plot, MASS, gridExtra, ggplot2,caret,cars*
7. *https://files.stlouisfed.org/files/htdocs/publications/review/08/09/Mizen.pdf*
8. *Dataset: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)*