School of Computer Science & Software Engineering

Bachelor of Computer Science

INFO411
Assignment 1

Name: Sivathorn Siralert
UOW ID: 6738564

Part 1

I first load in all the data and describe the columns

```
> describe(data)
data

 46  Variables      2500  Observations
-------------------------------------------------------------------------------
-----------------------------------------------------
functionary
       n  missing distinct     Info      Sum     Mean      Gmd
    2500        0        2    0.603      696   0.2784   0.4019

-------------------------------------------------------------------------------
-----------------------------------------------------
re.balanced..paid.back..a.recently.overdrawn.current.acount
       n  missing distinct     Info      Sum     Mean      Gmd
    2500        0        2    0.379     2129   0.8516   0.2529

-------------------------------------------------------------------------------
-----------------------------------------------------
FI3O.credit.score
       n  missing distinct     Info      Sum     Mean      Gmd
    2500        0        2    0.444     2049   0.8196   0.2958

-------------------------------------------------------------------------------
-----------------------------------------------------
gender
       n  missing distinct     Info      Sum     Mean      Gmd
    2500        0        2     0.75     1235    0.494   0.5001

-------------------------------------------------------------------------------
-----------------------------------------------------
X0..accounts.at.other.banks
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    3.048    1.606

Value          1     2     3     4     5
Frequency    490   467   499   521   523
Proportion 0.196 0.187 0.200 0.208 0.209

For the frequency table, variable is rounded to the nearest 0
-------------------------------------------------------------------------------
-----------------------------------------------------
credit.refused.in.past.
       n  missing distinct     Info      Sum     Mean      Gmd
    2500        0        2    0.347      334   0.1336   0.2316
```

```
--------------------------------------------------------------------------------
----------------------------------------------------
years.employed
       n  missing distinct     Info     Mean      Gmd 
    2500        0        5     0.96    3.011    1.588 

Value          1     2     3     4     5
Frequency    489   495   503   526   487
Proportion 0.196 0.198 0.201 0.210 0.195

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
savings.on.other.accounts
       n  missing distinct     Info     Mean      Gmd 
    2500        0        6    0.959    3.142    2.014 

Value          1     2     3     4     5     6
Frequency    585   559   490    36   443   387
Proportion 0.234 0.224 0.196 0.014 0.177 0.155

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
self.employed.
       n  missing distinct     Info      Sum     Mean      Gmd 
    2500        0        2    0.474      492   0.1968   0.3163 

--------------------------------------------------------------------------------
----------------------------------------------------
max..account.balance.12.months.ago
       n  missing distinct     Info     Mean      Gmd 
    2500        0        5     0.96    2.958    1.577 

Value          1     2     3     4     5
Frequency    490   544   517   479   470
Proportion 0.196 0.218 0.207 0.192 0.188

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
min..account.balance.12.months.ago
       n  missing distinct     Info     Mean      Gmd 
    2500        0        5     0.96    2.972    1.596 

Value          1     2     3     4     5
Frequency    518   481   539   476   486
Proportion 0.207 0.192 0.216 0.190 0.194
```

```
For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
-----------------------------------------------------
avrg..account.balance.12.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    2.985    1.594

Value          1     2     3     4     5
Frequency    500   510   506   495   489
Proportion 0.200 0.204 0.202 0.198 0.196

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
-----------------------------------------------------
max..account.balance.11.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96     2.99    1.607

Value          1     2     3     4     5
Frequency    514   498   481   514   493
Proportion 0.206 0.199 0.192 0.206 0.197

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
-----------------------------------------------------
min..account.balance.11.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    2.964    1.597

Value          1     2     3     4     5
Frequency    523   493   508   502   474
Proportion 0.209 0.197 0.203 0.201 0.190

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
-----------------------------------------------------
avrg..account.balance.11.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    2.989     1.58

Value          1     2     3     4     5
Frequency    503   481   518   537   461
Proportion 0.201 0.192 0.207 0.215 0.184

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
-----------------------------------------------------
max..account.balance.10.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96     2.95    1.614

Value          1     2     3     4     5
Frequency    529   532   466   482   491
Proportion 0.212 0.213 0.186 0.193 0.196

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
min..account.balance.10.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    3.003    1.595

Value          1     2     3     4     5
Frequency    487   521   487   507   498
Proportion 0.195 0.208 0.195 0.203 0.199

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
avrg..account.balance.10.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    3.027    1.583

Value          1     2     3     4     5
Frequency    464   517   512   502   505
Proportion 0.186 0.207 0.205 0.201 0.202

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
max..account.balance.9.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    3.014     1.63

Value          1     2     3     4     5
Frequency    516   487   483   474   540
Proportion 0.206 0.195 0.193 0.190 0.216

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
min..account.balance.9.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    2.986    1.604

Value          1     2     3     4     5
Frequency    495   538   479   483   505
Proportion 0.198 0.215 0.192 0.193 0.202
```

```
For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
avrg..account.balance.9.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    2.971    1.605

Value          1     2     3     4     5
Frequency    526   492   490   512   480
Proportion 0.210 0.197 0.196 0.205 0.192

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
max..account.balance.8.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    3.046    1.619

Value          1     2     3     4     5
Frequency    488   492   484   488   548
Proportion 0.195 0.197 0.194 0.195 0.219

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
min..account.balance.8.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    3.019    1.586

Value          1     2     3     4     5
Frequency    481   489   531   499   500
Proportion 0.192 0.196 0.212 0.200 0.200

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
avrg..account.balance.8.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    2.989    1.601

Value          1     2     3     4     5
Frequency    501   510   506   481   502
Proportion 0.200 0.204 0.202 0.192 0.201

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
max..account.balance.7.months.ago
```

```
       n  missing distinct     Info     Mean      Gmd 
    2500        0        5     0.96    3.004    1.606 

Value          1     2     3     4     5
Frequency    498   519   457   526   500
Proportion 0.199 0.208 0.183 0.210 0.200

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
min..account.balance.7.months.ago 
       n  missing distinct     Info     Mean      Gmd 
    2500        0        5     0.96    3.028    1.582 

Value          1     2     3     4     5
Frequency    476   491   508   536   489
Proportion 0.190 0.196 0.203 0.214 0.196

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
avrg..account.balance.7.months.ago 
       n  missing distinct     Info     Mean      Gmd 
    2500        0        5     0.96     3.03    1.613 

Value          1     2     3     4     5
Frequency    493   497   480   502   528
Proportion 0.197 0.199 0.192 0.201 0.211

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
max..account.balance.6.months.ago 
       n  missing distinct     Info     Mean      Gmd 
    2500        0        5     0.96    2.997    1.602 

Value          1     2     3     4     5
Frequency    507   494   493   512   494
Proportion 0.203 0.198 0.197 0.205 0.198

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
min..account.balance.6.months.ago 
       n  missing distinct     Info     Mean      Gmd 
    2500        0        5     0.96    3.028    1.606 

Value          1     2     3     4     5
Frequency    471   535   484   474   536
```

```
Proportion 0.188 0.214 0.194 0.190 0.214

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
avrg..account.balance.6.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    3.049    1.596

Value          1     2     3     4     5
Frequency    480   470   516   515   519
Proportion 0.192 0.188 0.206 0.206 0.208

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
max..account.balance.5.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    3.003     1.61

Value          1     2     3     4     5
Frequency    511   484   500   496   509
Proportion 0.204 0.194 0.200 0.198 0.204

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
min..account.balance.5.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96     2.99    1.597

Value          1     2     3     4     5
Frequency    499   505   515   483   498
Proportion 0.200 0.202 0.206 0.193 0.199

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
avrg..account.balance.5.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    2.979    1.612

Value          1     2     3     4     5
Frequency    513   524   468   493   502
Proportion 0.205 0.210 0.187 0.197 0.201

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
```

```
max..account.balance.4.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96     3.03    1.605

Value          1     2     3     4     5
Frequency    492   484   500   505   519
Proportion 0.197 0.194 0.200 0.202 0.208

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
------------------------------------------------------
min..account.balance.4.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    3.006    1.601

Value          1     2     3     4     5
Frequency    496   505   489   509   501
Proportion 0.198 0.202 0.196 0.204 0.200

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
------------------------------------------------------
avrg..account.balance.4.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    3.038     1.58

Value          1     2     3     4     5
Frequency    460   504   527   500   509
Proportion 0.184 0.202 0.211 0.200 0.204

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
------------------------------------------------------
max..account.balance.3.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    3.021    1.609

Value          1     2     3     4     5
Frequency    501   482   497   503   517
Proportion 0.200 0.193 0.199 0.201 0.207

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
------------------------------------------------------
min..account.balance.3.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    3.019    1.574

Value          1     2     3     4     5
```

```
Frequency    473   496   520   532   479
Proportion 0.189 0.198 0.208 0.213 0.192

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
-----------------------------------------------------
avrg..account.balance.3.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96     3.04    1.606

Value          1     2     3     4     5
Frequency    488   487   483   522   520
Proportion 0.195 0.195 0.193 0.209 0.208

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
-----------------------------------------------------
max..account.balance.2.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    2.996    1.578

Value          1     2     3     4     5
Frequency    482   507   531   498   482
Proportion 0.193 0.203 0.212 0.199 0.193

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
-----------------------------------------------------
min..account.balance.2.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96    3.011    1.623

Value          1     2     3     4     5
Frequency    515   481   497   475   532
Proportion 0.206 0.192 0.199 0.190 0.213

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
-----------------------------------------------------
avrg..account.balance.2.months.ago
       n  missing distinct     Info     Mean      Gmd
    2500        0        5     0.96     3.02    1.594

Value          1     2     3     4     5
Frequency    473   528   481   511   507
Proportion 0.189 0.211 0.192 0.204 0.203

For the frequency table, variable is rounded to the nearest 0
```

```
--------------------------------------------------------------------------------
----------------------------------------------------
max..account.balance.1.months.ago
       n  missing distinct     Info     Mean      Gmd 
    2500        0        5     0.96    2.973    1.591 

Value          1     2     3     4     5
Frequency    508   507   503   508   474
Proportion 0.203 0.203 0.201 0.203 0.190

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
min..account.balance.1.months.ago
       n  missing distinct     Info     Mean      Gmd 
    2500        0        5    0.959    2.952    1.589 

Value          1     2     3     4     5
Frequency    492   568   492   463   485
Proportion 0.197 0.227 0.197 0.185 0.194

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
avrg..account.balance.1.months.ago
       n  missing distinct     Info     Mean      Gmd 
    2500        0        5     0.96    3.011    1.576 

Value          1     2     3     4     5
Frequency    469   524   498   528   481
Proportion 0.188 0.210 0.199 0.211 0.192

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
credit.rating
       n  missing distinct     Info     Mean      Gmd 
    2500        0        4    0.916     1.58    1.146 

Value          0     1     2     3
Frequency    538   483   970   509
Proportion 0.215 0.193 0.388 0.204

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
----------------------------------------------------
```

Next i do some data cleaning and then examine the corelation

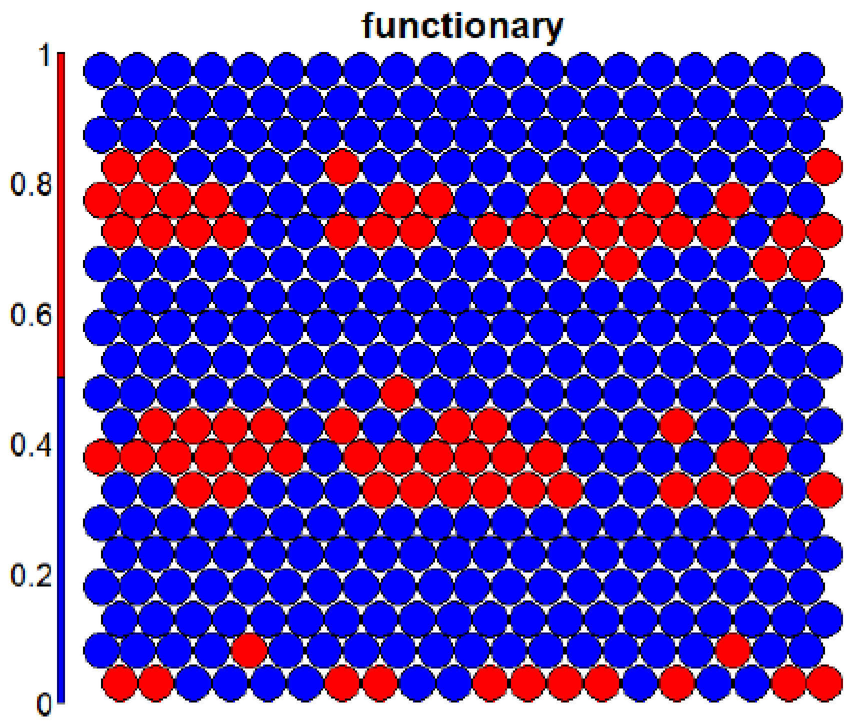
```
> head(corTable,6)
                                                             [,1]
credit.rating                                          1.00000000
functionary                                            0.31728279
FI3O.credit.score                                      0.27988701
re.balanced..paid.back..a.recently.overdrawn.current.acount 0.21822314
credit.refused.in.past.                                0.21783847
gender
```

This allows me to conclude that i need to use functionary, FI30 credit score, rebalanced (paid back) a recently overdrawn current account, credit refused in past and gender as they have the highest correlation compared to the other attributes.

Part 2

Using those 5 attributes, i plot a SOM for each attribute.

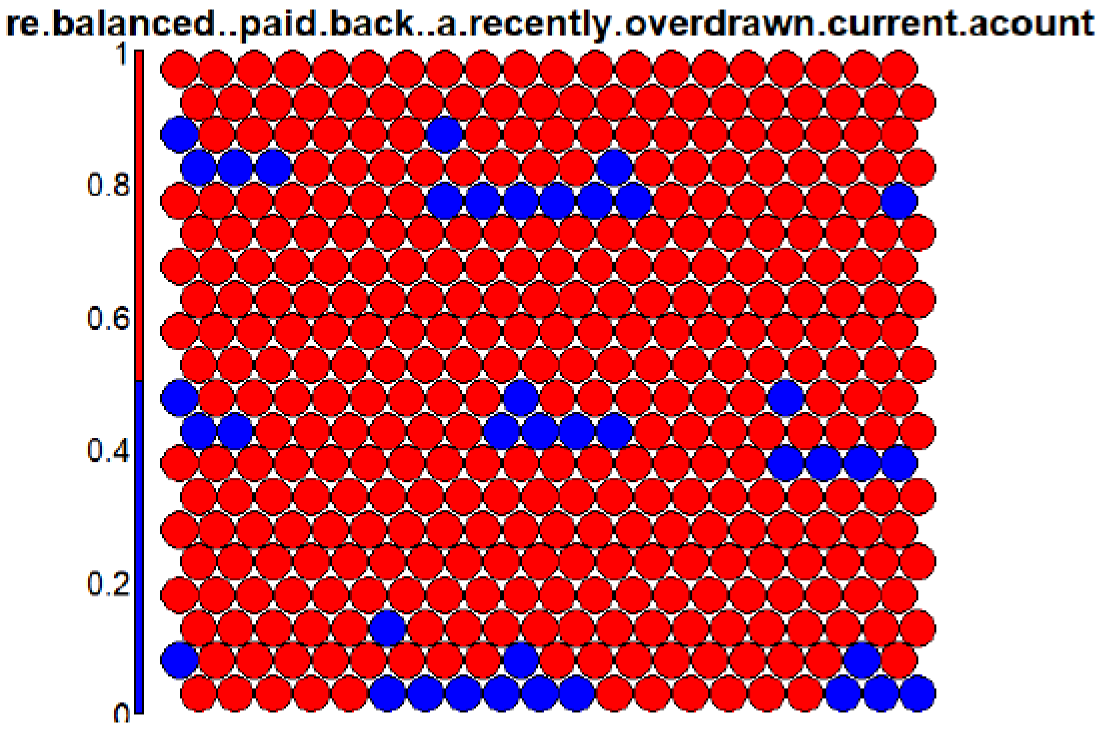
Functionary



Legend:
red-0/non functional
blue-1/fucntional

Most circles are blue, and are more clustered tgt compared to red, showing that most account are functional
functioning=likely to have higher credit rating

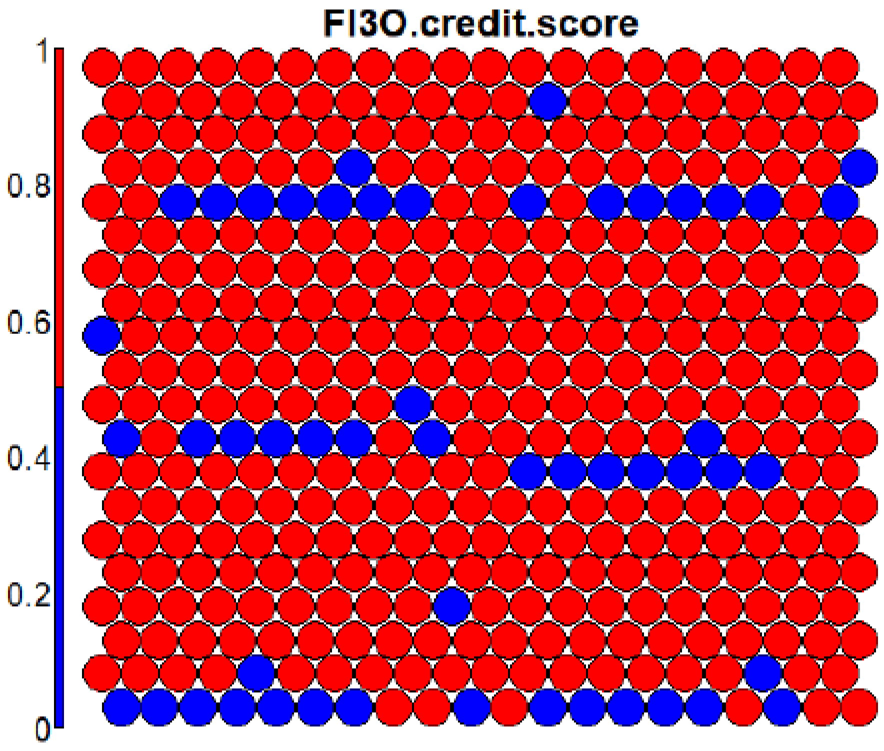
Paid back overdrawn current account



Legend:
red-1/paid back
Blue-0/not p[aid back

Most circles are red, and are more clustered compared to blue, showing that most users paid back overdrawn current account
Paid back=likely to have higher credit rating

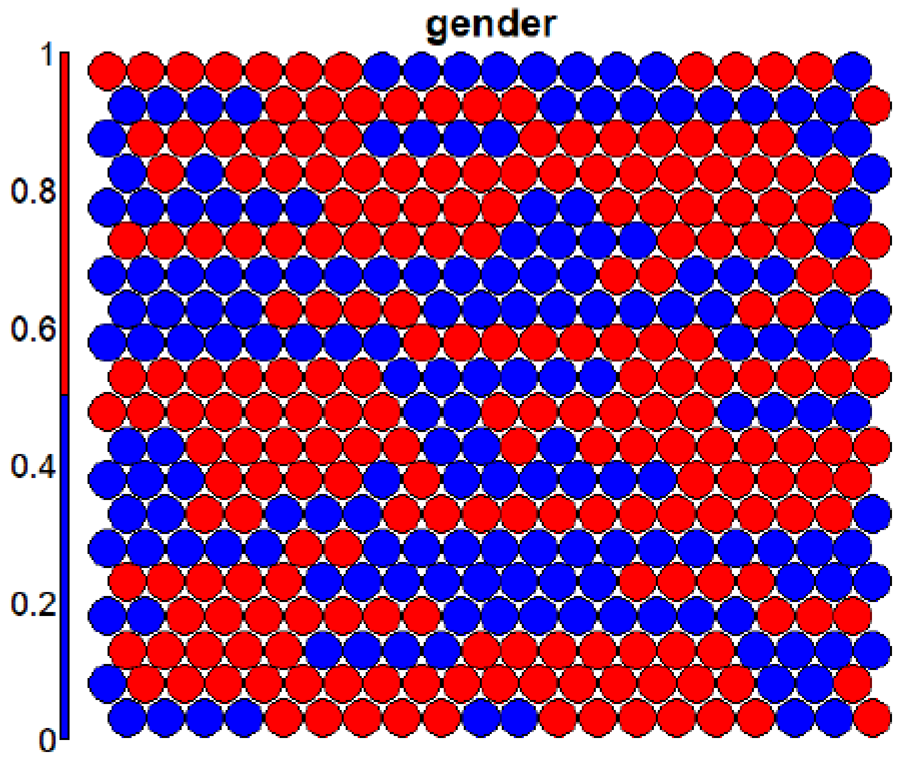
FI3O creit score



Legend:
red-1/credit
Blue-0/no credit

Most circles are red, and are more clustered than blue, showing that most users have high credit
High credit score=likely to have higher credit rating

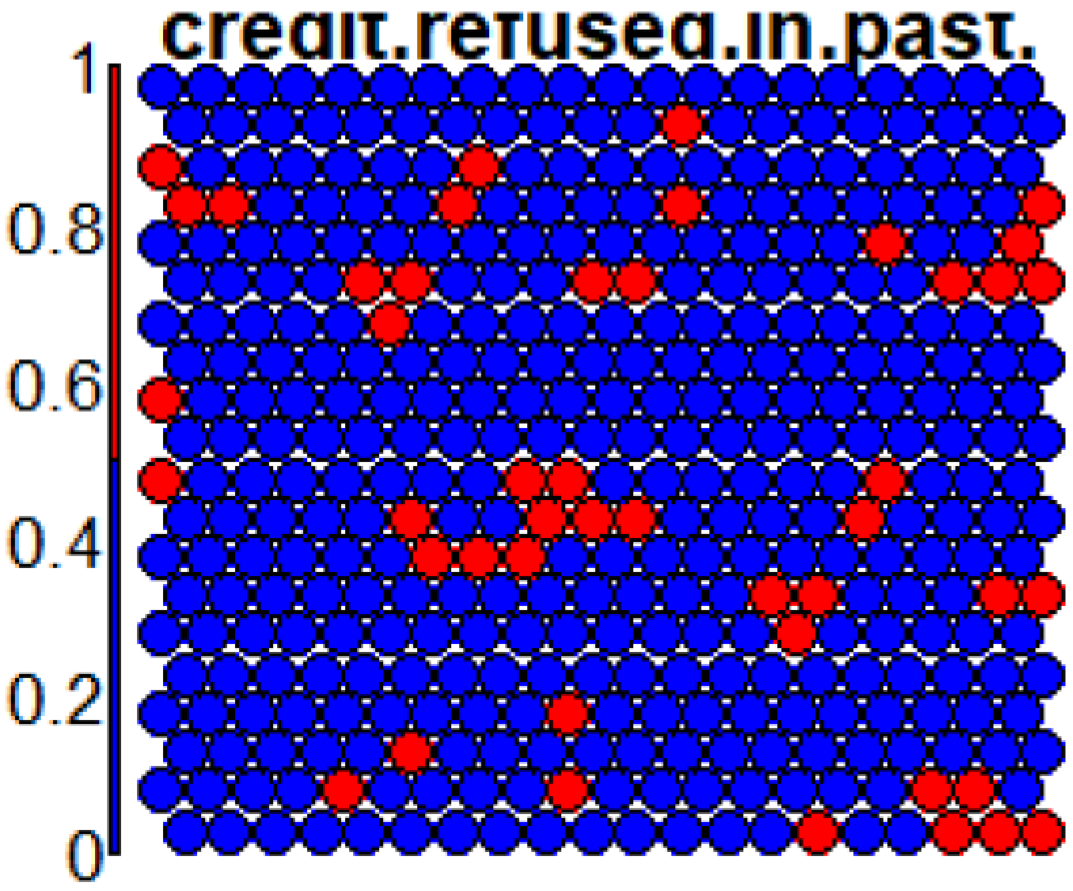
Gender



Legend:
red-1/emale
Blue-0/male

There is even spread of blue and red(looks like even)
Shows that theres equal chance for male or female to have high credit rating

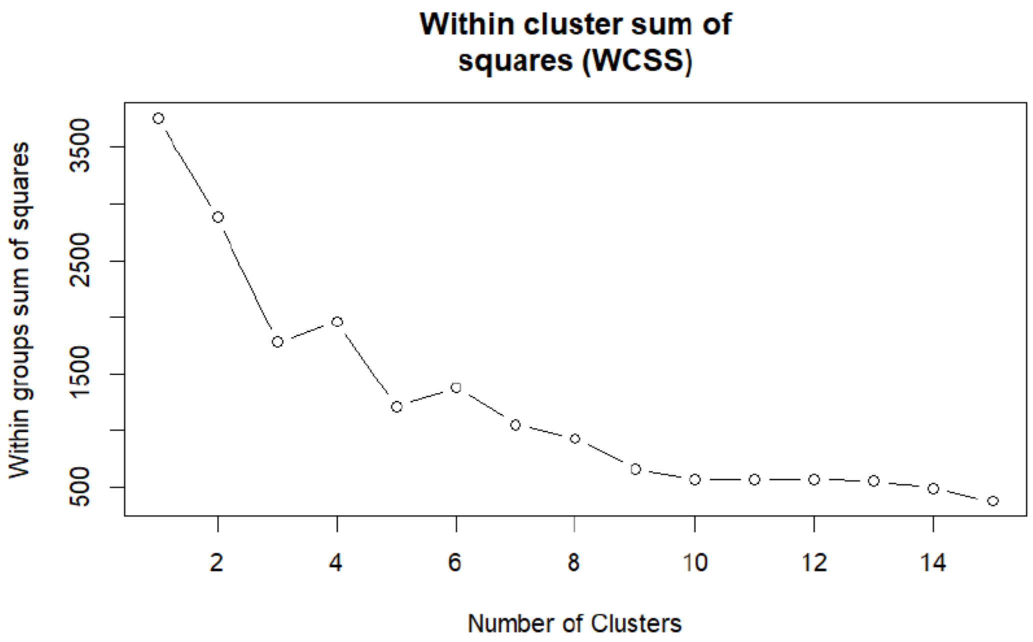
Credit refused in the apst



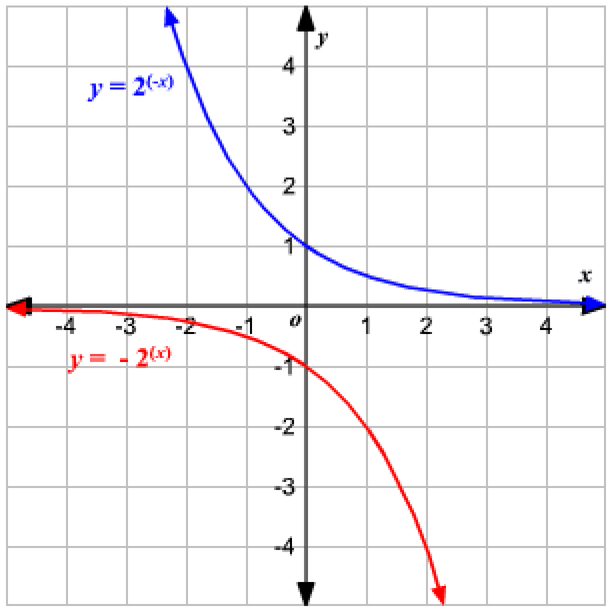
Legend:
red-1/refused
Blue-0/not refused

Most fo the map is blue, and are more clustered than red, showing that most ppl have no credit refused
No credit refused= more likely to have higher credit rating

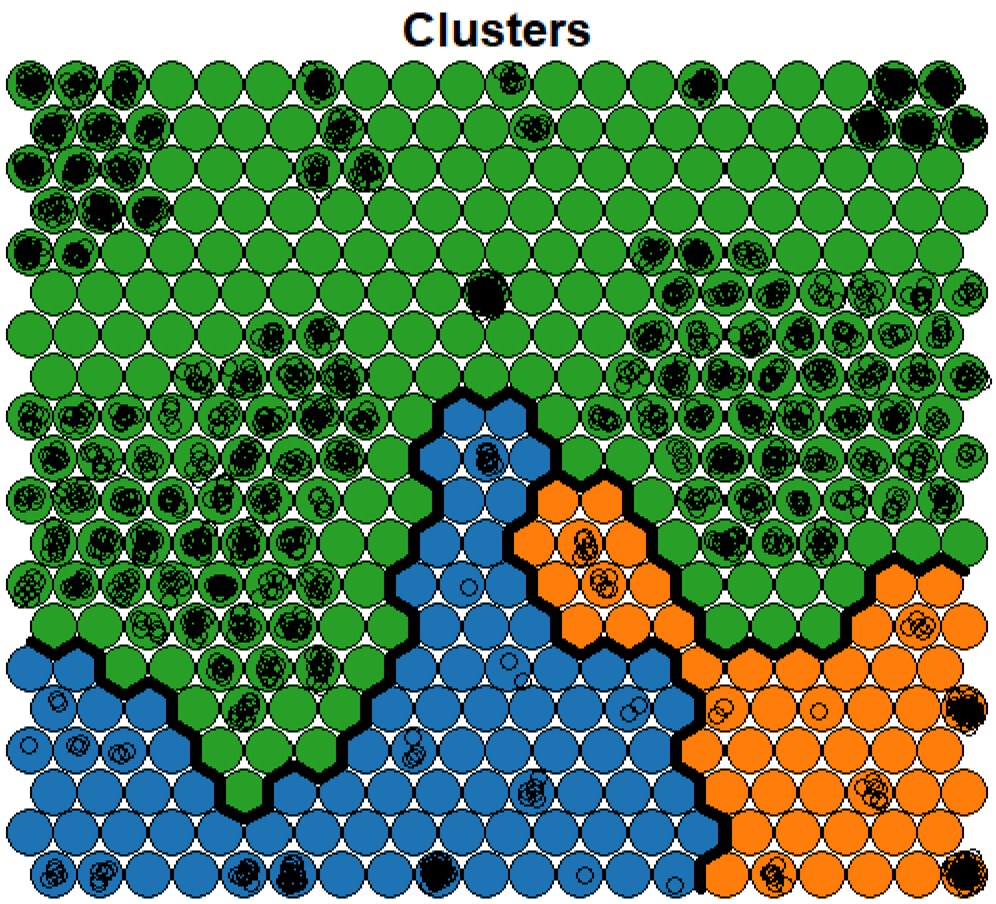
Part 3
I will now visualise the dataset using k mean clustering to demo the relationship b/w the WCSS(y axis) and the number of clusters(x axis)

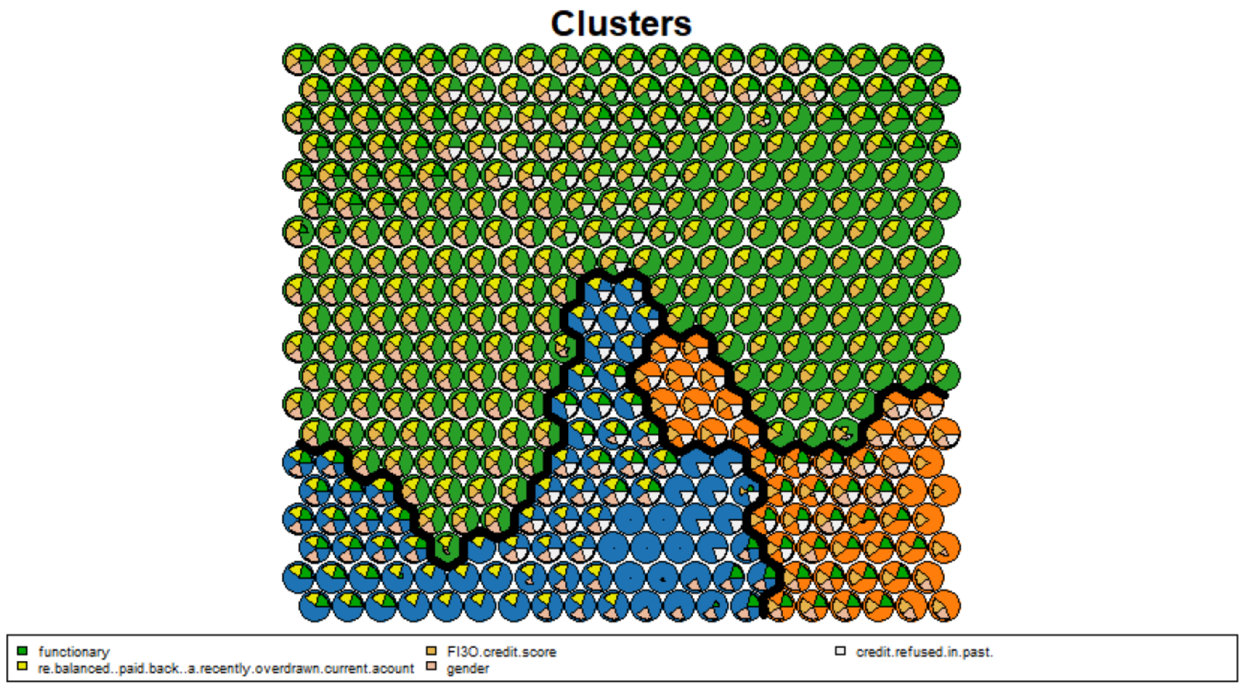


The relationship appears to be expoinantial looking like this



(just an example of exponential graph)
When number of clusters increase, there is an exponential drop in WCSS.
From this we determine that 3 is the best number of clusters to use.


Clusters

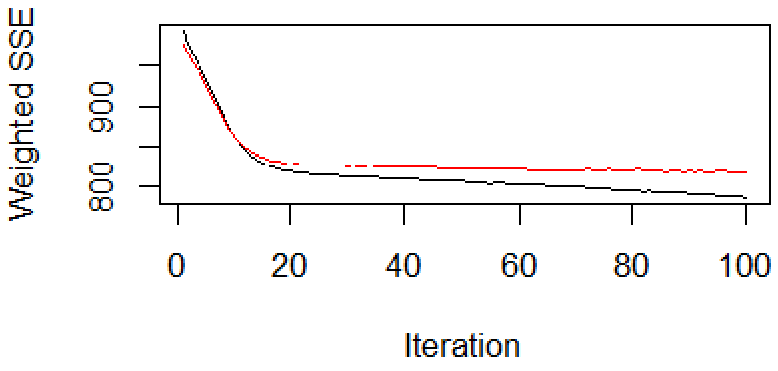

Clusters
functionary
re.balanced..paid.back..a.recently.overdrawn.current.acount
FI3O.credit.score
gender
credit.refused.in.past.

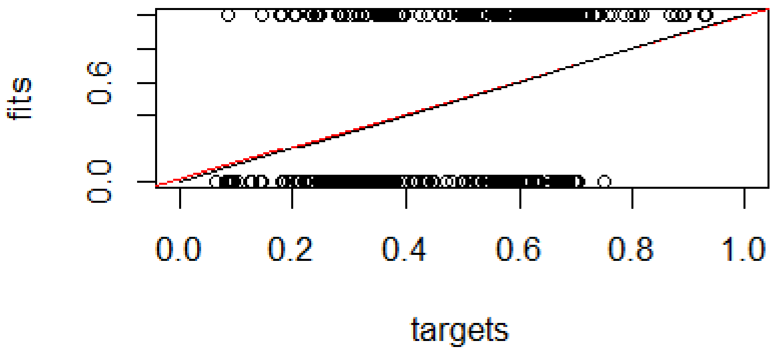
From all this information, we can deduce that kohonen is not 100% accurate. I hypothesise that thai is because the input quality affects the SOM topology, affectignt heir ability to map vectors, as well as how many times we iterate and many more

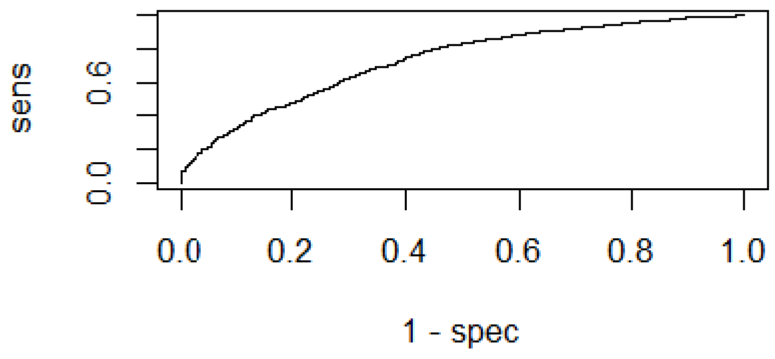
Part 4

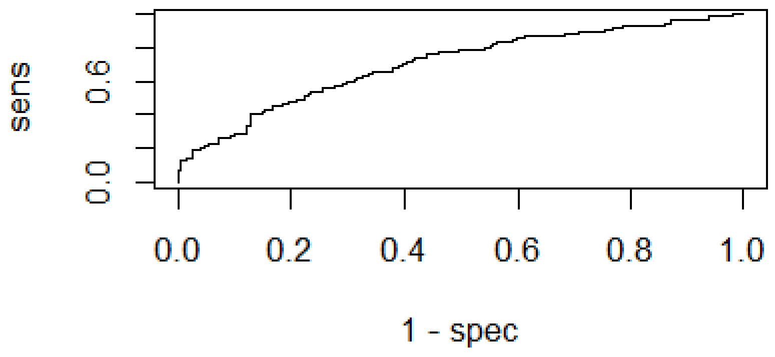
We will conduct the MLP with 100 iterations, learning parameter to 0.01, number of nodes to 20 after splitting the dataset to a ratio of 0.2

```
# Train the MLP model
model = mlp(trainSet$inputsTrain,
            trainSet$targetsTrain,
            size=c(20),
            learnFuncParams=c(0.001),
            maxit=100,
            inputsTest = trainSet$inputsTest,
            targetsTest = trainSet$targetsTest)
#
# Predict the test
# The predict() function in R is used to predict the values
# based on the input data.
predictTestSet = predict(model, trainSet$inputsTest)
```









After fine tuning, we perform the following

```
>confusionMatrix(trainSet$targetsTrain, fitted.values(model))
      Predictions
Targets 1   2     3
          1 214   164   22
          2 120   612   51
          3 44    175   178
```

```
>confusionMatrix(trainSet$targetsTest, predictTestSet)
      Predictions
Targets 1   2     3
          1 52    19    2
          2 27    150   21
          3 14    50    40
```

accuracy: (52+150+40)/(52+19+2+27+150+21+14+50+40)=240/375=64%

Code

```
# Preprocessing
library(kohonen)
library(dummies)
library(ggplot2)
library(sp)
library(maptools)
library(reshape2)
library(rgeos)
library(sf)
library(terra)
library(MASS)
library(Hmisc)
library(RSNNS)

# Colour palette definition
pretty_palette <- c("#1f77b4", '#ff7f0e', '#2ca02c', '#d62728', '#9467bd',
'#8c564b', '#e377c2')

### DATA PREPARATION
data <- read.csv("./creditworthiness.csv")
describe(data)
classifiedData = subset(data, data[,46] > 0)
unknownData = subset(data, data[,46] == 0)
corTable = abs(cor(classifiedData, y=classifiedData$credit.rating))
corTable = corTable[order(corTable, decreasing = TRUE),,drop = FALSE]
head(corTable,6)

# ------------------- SOM TRAINING ---------------------------

#choose the variables with which to train the SOM
#the following selects column 1,2,3,4,6
interestedFeatures <- data[, c(1,2,3,4,6)]

#data_train <- data[, c(1:45)]
data_train <- classifiedData[, c(1:45)]

# now train the SOM using the Kohonen method
data_train_matrix <- as.matrix(scale(data_train))
names(data_train_matrix) <- names(data_train)
require(kohonen)
x_dim=20
y_dim=20
small_areas <-FALSE
if (small_areas){
  # larger grid for the small areas example (more samples)
  som_grid <- somgrid(xdim = x_dim, ydim=y_dim, topo="hexagonal")
} else {
  som_grid <- somgrid(xdim = x_dim/2, ydim=y_dim/2, topo="hexagonal")
}
# Train the SOM model!
if (packageVersion("kohonen") < 3){
  system.time(som_model <- som(data_train_matrix,
                               grid=som_grid,
                               rlen=1000,
                               alpha=c(0.8,0.01),
                               n.hood = "circular",
                               keep.data = TRUE ))
}else{
  system.time(som_model <- som(data_train_matrix,
                               grid=som_grid,
                               rlen=1000,
                               alpha=c(0.8,0.01),
                               mode="online",
                               normalizeDataLayers=false,
                               keep.data = TRUE ))
}
summary(som_model)
rm(som_grid, data_train_matrix)

# -------------------- SOM VISUALISATION -----------------

source('./coolBlueHotRed.R')
# Plot the heatmap for a variable at scaled / normalised values
var <- 1 # Functionary
var_unscaled <- aggregate(as.numeric(data_train[,var]),
by=list(som_model$unit.classif), FUN=mean, simplify=TRUE)[,2]
plot(som_model, type = "property", property=var_unscaled,
main=names(data_train)[var], palette.name=coolBlueHotRed)
rm(var_unscaled, var)

var <- 2 # FI3O.credit.score
var_unscaled <- aggregate(as.numeric(data_train[,var]),
by=list(som_model$unit.classif), FUN=mean, simplify=TRUE)[,2]
plot(som_model, type = "property", property=var_unscaled,
main=names(data_train)[var], palette.name=coolBlueHotRed)
rm(var_unscaled, var)

var <- 3 # Rebalance.payback
var_unscaled <- aggregate(as.numeric(data_train[,var]),
by=list(som_model$unit.classif), FUN=mean, simplify=TRUE)[,2]
plot(som_model, type = "property", property=var_unscaled,
main=names(data_train)[var], palette.name=coolBlueHotRed)
rm(var_unscaled, var)

var <- 4 # credit.refused.in.past.
```

```
var_unscaled <- aggregate(as.numeric(data_train[,var]),
by=list(som_model$unit.classif), FUN=mean, simplify=TRUE)[,2]
plot(som_model, type = "property", property=var_unscaled,
main=names(data_train)[var], palette.name=coolBlueHotRed)
rm(var_unscaled, var)

var <- 6 #Gender
var_unscaled <- aggregate(as.numeric(data_train[,var]),
by=list(som_model$unit.classif), FUN=mean, simplify=TRUE)[,2]
plot(som_model, type = "property", property=var_unscaled,
main=names(data_train)[var], palette.name=coolBlueHotRed)
rm(var_unscaled, var)

source('./plotHeatMap.R')
plotHeatMap(som_model, classifiedData, variable=0)

genderT = with(classifiedData, table(credit.rating, gender))
barplot(genderT, beside = TRUE,
        legend = c("Credit Rating A", "Credit Rating B", "Credit Rating
C"),
        col = c("darkgreen","yellow", "red"),
        main = "Gender vs Credit Rating",
        sub="0 = Male, 1 = Female")
selfEmployed = with(classifiedData, table(credit.rating, self.employed.))
barplot(selfEmployed, beside = TRUE,
        legend = c("Credit Rating A", "Credit Rating B", "Credit Rating
C"),
        col = c("darkgreen","yellow", "red"),
        main = "Self Employed vs Credit Rating",
        sub="0 = No, 1 = Yes")
functional = with(classifiedData, table(credit.rating, functionary))
barplot(functional, beside = TRUE,
        legend = c("Credit Rating A", "Credit Rating B", "Credit Rating
C"),
        col = c("darkgreen","yellow", "red"),
        main = "Functionary vs Credit Rating",
        sub="0 = No, 1 = Yes")
genderT = with(classifiedData, table(credit.rating, gender))
genderT
barplot(genderT, beside = TRUE,
        legend = c("Credit Rating A", "Credit Rating B", "Credit Rating
C"),
        col = c("darkgreen","yellow", "red"),
        main = "Gender vs Credit Rating",
        sub="0 = Male, 1 = Female")
selfEmployed = with(classifiedData, table(credit.rating, self.employed.))
barplot(selfEmployed, beside = TRUE,
        legend = c("Credit Rating A", "Credit Rating B", "Credit Rating
C"),
        col = c("darkgreen","yellow", "red"),
        main = "Self Employed vs Credit Rating",
        sub="0 = No, 1 = Yes")
# generate a contingency table and plot a barplot
FI30T = with(classifiedData, table(credit.rating, FI3O.credit.score))
FI30T
barplot(genderT, beside = TRUE,
        legend = c("Credit Rating A", "Credit Rating B", "Credit Rating
C"),
        col = c("darkgreen","yellow", "red"),
        main = "FI30 vs Credit Rating",
        sub="0 = Not OK, 1 = Ok")

# show the WCSS metric for kmeans for different clustering sizes.
# Can be used as a "rough" indicator of the ideal number of clusters
mydata <- matrix(unlist(som_model$codes), ncol = length(data_train),
                 byrow = FALSE)
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(mydata,
                                     centers=i)$withinss)
par(mar=c(5.1,4.1,4.1,2.1))
plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares", main="Within cluster sum of
squares (WCSS)")

# Form clusters on grid
## use hierarchical clustering to cluster the codebook vectors
som_cluster <- cutree(hclust(dist(mydata)), 3)
# Show the map with different colours for every cluster

plot(som_model, type="mapping", bgcol = pretty_palette[som_cluster], main
     = "Clusters")
add.cluster.boundaries(som_model, som_cluster)
#show the same plot with the codes instead of just colours
plot(som_model, type="codes", bgcol = pretty_palette[som_cluster], main =
       "Clusters")
add.cluster.boundaries(som_model, som_cluster)


# ------------------ Clustering SOM results -------------------

# show the WCSS metric for kmeans for different clustering sizes.
# Can be used as a "rough" indicator of the ideal number of clusters
mydata <- matrix(unlist(som_model$codes), ncol = length(data_train),
                 byrow = FALSE)
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(mydata,
                                     centers=i)$withinss)
par(mar=c(5.1,4.1,4.1,2.1))
```

```
plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares", main="Within cluster sum of
squares (WCSS)")

# Form clusters on grid
## use hierarchical clustering to cluster the codebook vectors
som_cluster <- cutree(hclust(dist(mydata)), 3)
# Show the map with different colours for every cluster

plot(som_model, type="mapping", bgcol = pretty_palette[som_cluster], main
     = "Clusters")
add.cluster.boundaries(som_model, som_cluster)
#show the same plot with the codes instead of just colours
plot(som_model, type="codes", bgcol = pretty_palette[som_cluster], main =
       "Clusters")
add.cluster.boundaries(som_model, som_cluster)




# To train the MLP model to classified based on the following
# interested columns.
interestedColumns = c(1, 2, 3, 4, 6, 9)
# Seperate value from targets
trainValues = classifiedData[,interestedColumns]
unknownValues = unknownData[,interestedColumns]
# Use decodeClassLabels() to decode class labels from a
# numerical or levels vector to a binary matrix.
trainTargets = decodeClassLabels(classifiedData[,46])
# Split the data into training and testing data set
trainSet = splitForTrainingAndTest(trainValues, trainTargets, ratio =
                                    0.2)
# Normalized the training data set
trainSet = normTrainingAndTestSet(trainSet)
# Train the MLP model
model = mlp(trainSet$inputsTrain,
            trainSet$targetsTrain,
            size=c(20),
            learnFuncParams=c(0.001),
            maxit=100,
            inputsTest = trainSet$inputsTest,
            targetsTest = trainSet$targetsTest)
#
# Predict the test
# The predict() function in R is used to predict the values
# based on the input data.
predictTestSet = predict(model, trainSet$inputsTest)
# Predict the unknown set
predictUnknownSet = predict(model, unknownValues)
```

```
# Compute the confusion matrix
confusionMatrix(trainSet$targetsTrain, fitted.values(model))
confusionMatrix(trainSet$targetsTest, predictTestSet)
# interpreting the unknown data set (prediction)
head(trainTargets)
head(classifiedData[,46])
head(predictUnknownSet)
# Plot
par(mar=c(5.1,4.1,4.1,2.1))
par(mfrow=c(2,2))
plotIterativeError(model)
plotRegressionError(predictTestSet[,2], trainSet$targetsTest[,2])
plotROC(fitted.values(model)[,2], trainSet$targetsTrain[,2])
plotROC(predictTestSet[,2],trainSet$targetsTest[,2])
summary(model)
model
weightMatrix(model)
extractNetInfo(model)
```