Support Vector Machine

Michael In der Au

22.06.2018

Contents

1 Vorbereitung

1.1 Benötigte Pakete

```r
#required packages
library(e1071)
library(mice)
library(caret)
library(ggplot2)
library(dplyr)
library(Hmisc)
```

2 Einlesen der Ausgangsdaten

```r
#read train data
train <- read.csv("data/cs-training.csv", header = T, sep = ",", dec = ".")
#exclude ids
train <- train[-c(1)]
#read test data
test <- read.csv("data/cs-test.csv", header = T, sep = ",", dec = ".")
#exclude ids
test <- test[-c(1)]

#view data
summary(train)
```

```
##  SeriousDlqin2yrs  RevolvingUtilizationOfUnsecuredLines      age
##  Min.   :0.00000   Min.   :    0.00                     Min.   :  0.0
##  1st Qu.:0.00000   1st Qu.:    0.03                     1st Qu.: 41.0
##  Median :0.00000   Median :    0.15                     Median : 52.0
##  Mean   :0.06684   Mean   :    6.05                     Mean   : 52.3
##  3rd Qu.:0.00000   3rd Qu.:    0.56                     3rd Qu.: 63.0
##  Max.   :1.00000   Max.   :50708.00                     Max.   :109.0
##
##  NumberOfTime30_59DaysPastDueNotWorse   DebtRatio        MonthlyIncome
##  Min.   : 0.000                       Min.   :      0.0   Min.   :      0
##  1st Qu.: 0.000                       1st Qu.:      0.2   1st Qu.:   3400
##  Median : 0.000                       Median :      0.4   Median :   5400
##  Mean   : 0.421                       Mean   :    353.0   Mean   :   6670
##  3rd Qu.: 0.000                       3rd Qu.:      0.9   3rd Qu.:   8249
##  Max.   :98.000                       Max.   :329664.0   Max.   :3008750
##                                                          NA's   :29731
##  NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate
##  Min.   : 0.000                  Min.   : 0.000
##  1st Qu.: 5.000                  1st Qu.: 0.000
##  Median : 8.000                  Median : 0.000
##  Mean   : 8.453                  Mean   : 0.266
##  3rd Qu.:11.000                  3rd Qu.: 0.000
##  Max.   :58.000                  Max.   :98.000
##
##  NumberRealEstateLoansOrLines NumberOfTime60_89DaysPastDueNotWorse
##  Min.   : 0.000               Min.   : 0.0000
##  1st Qu.: 0.000               1st Qu.: 0.0000
##  Median : 1.000               Median : 0.0000
##  Mean   : 1.018               Mean   : 0.2404
##  3rd Qu.: 2.000               3rd Qu.: 0.0000
##  Max.   :54.000               Max.   :98.0000
##
##  NumberOfDependents
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 0.000
##  Mean   : 0.757
##  3rd Qu.: 1.000
##  Max.   :20.000
##  NA's   :3924
```

```
summary(test)
```

```
##  SeriousDlqin2yrs RevolvingUtilizationOfUnsecuredLines      age
##  Mode:logical     Min.   :    0.000                    Min.   : 21.00
##  NA's:101503      1st Qu.:    0.030                    1st Qu.: 41.00
##                   Median :    0.153                    Median : 52.00
##                   Mean   :    5.310                    Mean   : 52.41
##                   3rd Qu.:    0.564                    3rd Qu.: 63.00
##                   Max.   :21821.000                    Max.   :104.00
##
##  NumberOfTime30_59DaysPastDueNotWorse   DebtRatio
##  Min.   : 0.0000                      Min.   :     0.00
##  1st Qu.: 0.0000                      1st Qu.:     0.17
##  Median : 0.0000                      Median :     0.36
##  Mean   : 0.4538                      Mean   :   344.48
##  3rd Qu.: 0.0000                      3rd Qu.:     0.85
##  Max.   :98.0000                      Max.   :268326.00
##
##  MonthlyIncome     NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate
##  Min.   :      0   Min.   : 0.000                  Min.   : 0.0000
##  1st Qu.:   3408   1st Qu.: 5.000                  1st Qu.: 0.0000
##  Median :   5400   Median : 8.000                  Median : 0.0000
##  Mean   :   6855   Mean   : 8.454                  Mean   : 0.2967
##  3rd Qu.:   8200   3rd Qu.:11.000                  3rd Qu.: 0.0000
##  Max.   :7727000   Max.   :85.000                  Max.   :98.0000
##  NA's   :20103
##  NumberRealEstateLoansOrLines NumberOfTime60_89DaysPastDueNotWorse
##  Min.   : 0.000               Min.   : 0.0000
##  1st Qu.: 0.000               1st Qu.: 0.0000
##  Median : 1.000               Median : 0.0000
##  Mean   : 1.013               Mean   : 0.2703
##  3rd Qu.: 2.000               3rd Qu.: 0.0000
##  Max.   :37.000               Max.   :98.0000
##
##  NumberOfDependents
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 0.000
##  Mean   : 0.769
##  3rd Qu.: 1.000
##  Max.   :43.000
##  NA's   :2626
```

3 Preprocessing

3.1 Imputation der NAs

```
#imputation using MICE package
imp <- mice(train, m=5, maxit=2, method='pmm', seed = 123)
train_imputed <- complete(x = imp,action =  1)
#summary(train_imputed)
imp <- mice(test, m=5, maxit=2, method='pmm', seed = 123)
test_imputed <- complete(x = imp,action =  1)

#export imputed data as .csv for future usage
write.csv(test_imputed,"data/test_imputed.csv")
write.csv(train_imputed,"data/train_imputed.csv")
```

3.2 Kontrolle der imputierten Daten

```
#read imputed data
train <- read.csv("data/train_imputed.csv")
test <- read.csv("data/test_imputed.csv")

#imputed data
summary(train)
```

```
##        X           SeriousDlqin2yrs  RevolvingUtilizationOfUnsecuredLines
##  Min.   :     1   Min.   :0.00000   Min.   :    0.00
##  1st Qu.: 37501   1st Qu.:0.00000   1st Qu.:    0.03
##  Median : 75001   Median :0.00000   Median :    0.15
##  Mean   : 75001   Mean   :0.06684   Mean   :    6.05
##  3rd Qu.:112500   3rd Qu.:0.00000   3rd Qu.:    0.56
##  Max.   :150000   Max.   :1.00000   Max.   :50708.00
##       age        NumberOfTime30_59DaysPastDueNotWorse   DebtRatio
##  Min.   :  0.0   Min.   : 0.000                       Min.   :     0.0
##  1st Qu.: 41.0   1st Qu.: 0.000                       1st Qu.:     0.2
##  Median : 52.0   Median : 0.000                       Median :     0.4
##  Mean   : 52.3   Mean   : 0.421                       Mean   :   353.0
##  3rd Qu.: 63.0   3rd Qu.: 0.000                       3rd Qu.:     0.9
##  Max.   :109.0   Max.   :98.000                       Max.   :329664.0
##  MonthlyIncome     NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate
##  Min.   :      0   Min.   : 0.000                  Min.   : 0.000
##  1st Qu.:   3029   1st Qu.: 5.000                  1st Qu.: 0.000
##  Median :   5000   Median : 8.000                  Median : 0.000
##  Mean   :   6214   Mean   : 8.453                  Mean   : 0.266
##  3rd Qu.:   7792   3rd Qu.:11.000                  3rd Qu.: 0.000
##  Max.   :3008750   Max.   :58.000                  Max.   :98.000
##  NumberRealEstateLoansOrLines NumberOfTime60_89DaysPastDueNotWorse
##  Min.   : 0.000               Min.   : 0.0000
##  1st Qu.: 0.000               1st Qu.: 0.0000
##  Median : 1.000               Median : 0.0000
##  Mean   : 1.018               Mean   : 0.2404
##  3rd Qu.: 2.000               3rd Qu.: 0.0000
##  Max.   :54.000               Max.   :98.0000
##  NumberOfDependents
##  Min.   : 0.0000
##  1st Qu.: 0.0000
##  Median : 0.0000
##  Mean   : 0.7499
##  3rd Qu.: 1.0000
##  Max.   :20.0000
```

```r
summary(test)
```

```
##        X          SeriousDlqin2yrs RevolvingUtilizationOfUnsecuredLines
##  Min.   :     1   Mode:logical     Min.   :    0.000
##  1st Qu.: 25377   NA's:101503      1st Qu.:    0.030
##  Median : 50752                    Median :    0.153
##  Mean   : 50752                    Mean   :    5.310
##  3rd Qu.: 76128                    3rd Qu.:    0.564
##  Max.   :101503                    Max.   :21821.000
##       age          NumberOfTime30_59DaysPastDueNotWorse   DebtRatio
##  Min.   : 21.00   Min.   : 0.0000                      Min.   :     0.00
##  1st Qu.: 41.00   1st Qu.: 0.0000                      1st Qu.:     0.17
##  Median : 52.00   Median : 0.0000                      Median :     0.36
##  Mean   : 52.41   Mean   : 0.4538                      Mean   :   344.48
##  3rd Qu.: 63.00   3rd Qu.: 0.0000                      3rd Qu.:     0.85
##  Max.   :104.00   Max.   :98.0000                      Max.   :268326.00
##  MonthlyIncome     NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate
##  Min.   :      0   Min.   : 0.000                  Min.   : 0.0000
##  1st Qu.:   3420   1st Qu.: 5.000                  1st Qu.: 0.0000
##  Median :   5416   Median : 8.000                  Median : 0.0000
##  Mean   :   6877   Mean   : 8.454                  Mean   : 0.2967
##  3rd Qu.:   8200   3rd Qu.:11.000                  3rd Qu.: 0.0000
##  Max.   :7727000   Max.   :85.000                  Max.   :98.0000
##  NumberRealEstateLoansOrLines NumberOfTime60_89DaysPastDueNotWorse
##  Min.   : 0.000               Min.   : 0.0000
##  1st Qu.: 0.000               1st Qu.: 0.0000
##  Median : 1.000               Median : 0.0000
##  Mean   : 1.013               Mean   : 0.2703
##  3rd Qu.: 2.000               3rd Qu.: 0.0000
##  Max.   :37.000               Max.   :98.0000
##  NumberOfDependents
##  Min.   : 0.0000
##  1st Qu.: 0.0000
##  Median : 0.0000
##  Mean   : 0.7618
##  3rd Qu.: 1.0000
##  Max.   :43.0000
```

3.3 Datenbereinigung

3.3.1 RevolvingUtilizationOfUnsecuredLines

```
#-RevolvingUtilizationOfUnsecuredLines
#-- (total balance) / (total credit limit)

# the closer this value is to 100% the more the consumer is using the credit limit
summary(train$RevolvingUtilizationOfUnsecuredLines)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     0.00     0.03     0.15     6.05     0.56 50708.00
```

```
mis <-train %>%
  filter(train$RevolvingUtilizationOfUnsecuredLines > 1)
summary(mis)
```

```
##        X          SeriousDlqin2yrs RevolvingUtilizationOfUnsecuredLines
##  Min.   :   163   Min.   :0.0000   Min.   :    1.00
##  1st Qu.: 38500   1st Qu.:0.0000   1st Qu.:    1.02
##  Median : 76727   Median :0.0000   Median :    1.07
##  Mean   : 75812   Mean   :0.3725   Mean   :  259.77
##  3rd Qu.:112448   3rd Qu.:1.0000   3rd Qu.:    1.30
##  Max.   :149974   Max.   :1.0000   Max.   :50708.00
##       age        NumberOfTime30_59DaysPastDueNotWorse   DebtRatio
##  Min.   :21.00   Min.   : 0.000                       Min.   :    0.001
##  1st Qu.:34.00   1st Qu.: 0.000                       1st Qu.:    0.181
##  Median :43.00   Median : 1.000                       Median :    0.374
##  Mean   :44.06   Mean   : 1.016                       Mean   :  245.169
##  3rd Qu.:52.00   3rd Qu.: 2.000                       3rd Qu.:    0.806
##  Max.   :88.00   Max.   :10.000                       Max.   :21395.000
##  MonthlyIncome    NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate
##  Min.   :     0   Min.   : 0.000                  Min.   : 0.0000
##  1st Qu.:  2500   1st Qu.: 3.000                  1st Qu.: 0.0000
##  Median :  3960   Median : 6.000                  Median : 0.0000
##  Mean   :  4982   Mean   : 6.374                  Mean   : 0.6378
##  3rd Qu.:  6020   3rd Qu.: 8.000                  3rd Qu.: 1.0000
##  Max.   :141500   Max.   :40.000                  Max.   :15.0000
##  NumberRealEstateLoansOrLines NumberOfTime60_89DaysPastDueNotWorse
##  Min.   : 0.000               Min.   :0.0000
##  1st Qu.: 0.000               1st Qu.:0.0000
##  Median : 0.000               Median :0.0000
##  Mean   : 0.682               Mean   :0.4324
##  3rd Qu.: 1.000               3rd Qu.:1.0000
##  Max.   :10.000               Max.   :7.0000
##  NumberOfDependents
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.9124
##  3rd Qu.:2.0000
##  Max.   :8.0000
```

```
#percentage of regressor > 1 in train data
nrow(mis)/nrow(train)*100
```

```
## [1] 2.214
```

```
#apply coded value -1 to outliers
train$RevolvingUtilizationOfUnsecuredLines[train$RevolvingUtilizationOfUnsecuredLines > 1] <- -1

summary(train$RevolvingUtilizationOfUnsecuredLines)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## -1.00000  0.02485  0.13540  0.27492  0.50693  1.00000 
```

3.3.2 Age

```
#-age

summary(train$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0    41.0    52.0    52.3    63.0   109.0 
```

```
mis <- train %>%
  filter(train$age == 0)
nrow(mis)
```

```
## [1] 1
```

```
#omit line with age = 0
train <- subset(train, age > 0)

summary(train$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    21.0    41.0    52.0    52.3    63.0   109.0 
```

3.3.3 NumberOfTime30__59DaysPastDueNotWorse

```
#-NumberOfTime30_59DaysPastDueNotWorse

summary(train$NumberOfTime30_59DaysPastDueNotWorse)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.000   0.000   0.421   0.000  98.000 
```

```
summary(train$NumberOfTime60_89DaysPastDueNotWorse)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.0000  0.0000  0.2404  0.0000 98.0000 
```

```
summary(train$NumberOfTimes90DaysLate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.000   0.000   0.266   0.000  98.000 
```

```
#it can be assumed that 96 and 98 are coded values of some kind,
#because both values have their own meaning they cant be ommited
# and have to be encoded
nrow(subset(train, train$NumberOfTime30_59DaysPastDueNotWorse >=15))
```

```
## [1] 269
```

```
n_96 <- nrow(subset(train, train$NumberOfTime30_59DaysPastDueNotWorse ==96))
n_98 <- nrow(subset(train, train$NumberOfTime30_59DaysPastDueNotWorse ==98))

(n_96+n_98)/nrow(train)*100
```

```
## [1] 0.1793345
```

```
train$NumberOfTime30_59DaysPastDueNotWorse[train$NumberOfTime30_59DaysPastDueNotWorse >= 15]<- -1

summary(train$NumberOfTime30_59DaysPastDueNotWorse)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.0000  0.0000  0.0000  0.2436  0.0000 13.0000
```

3.3.4 NumberOfTime60_89DaysPastDueNotWorse, NumberOfTimes90DaysLate

```
#the same approach applies to NumberOfTime60_89DaysPastDueNotWorse and
#NumberOfTimes90DaysLate
nrow(subset(train, train$NumberOfTime60_89DaysPastDueNotWorse >=15))
```

```
## [1] 269
```

```
n_96 <- nrow(subset(train, train$NumberOfTime60_89DaysPastDueNotWorse ==96))
n_98 <- nrow(subset(train, train$NumberOfTime60_89DaysPastDueNotWorse ==98))

(n_96+n_98)/nrow(train)*100
```

```
## [1] 0.1793345
```

```
train$NumberOfTime60_89DaysPastDueNotWorse[train$NumberOfTime60_89DaysPastDueNotWorse >= 15] <- -1

summary(train$NumberOfTime60_89DaysPastDueNotWorse)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.00000  0.00000  0.00000  0.06291  0.00000 11.00000
```

```
summary(train$NumberOfTimes90DaysLate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.266   0.000  98.000
```

```
nrow(subset(train, train$NumberOfTimes90DaysLate >=19))
```

```
## [1] 269
```

```
n_96 <- nrow(subset(train, train$NumberOfTimes90DaysLate ==96))
n_98 <- nrow(subset(train, train$NumberOfTimes90DaysLate ==98))

(n_96+n_98)/nrow(train)*100
```

```
## [1] 0.1793345
```

```
train$NumberOfTimes90DaysLate[train$NumberOfTimes90DaysLate >= 19] <- -1

summary(train$NumberOfTimes90DaysLate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.0000  0.0000  0.0000  0.0885  0.0000 17.0000
```

3.3.5 Debt ratio

```
#- DebtRatio
#-- (total debts) / (monthly income)
#-- thus, values > 1 indicate more debts than income

summary(train$DebtRatio)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      0.0      0.2      0.4    353.0      0.9 329664.0
```

```
summary(train$MonthlyIncome)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    3028    5000    6214    7792 3008750
```

```
#monthly income is the denominator of debt ratio thus it cannot be 0
#percentage of regressor > 1 in train data
n_inc0 <- nrow(subset(train, train$MonthlyIncome ==0))
n_inc0/nrow(train)*100
```

```
## [1] 2.582017
```

```
#if the monthly salary is equal to zero it is replaced by -1

index <- train$MonthlyIncome == 0
train$DebtRatio[index] <- -1
summary(train$MonthlyIncome)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    3028    5000    6214    7792 3008750
```

```
#if the monthly income is missing, it is replaced by 1

train$MonthlyIncome[is.na(train$MonthlyIncome)] <- 1

summary(train$DebtRatio)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     -1.00      0.16      0.35    275.35      0.73 307001.00
```

3.3.6 Monthly income

```
#-Monthly income

summary(train$MonthlyIncome)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    3028    5000    6214    7792 3008750
```

```
n_inc50k <- nrow(subset(train, train$MonthlyIncome >50000))
n_inc50k/nrow(train)*100
```

```
## [1] 0.2226682
```

```
#omit outliers
train <- subset(train, MonthlyIncome < 50000)
```

3.3.7 NumberOfOpenCreditLinesAndLoans

```
#-NumberOfOpenCreditLinesAndLoans

summary(train$NumberOfOpenCreditLinesAndLoans)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   5.000   8.000   8.447  11.000  58.000
```

```
#omit outliers in the 99th percentile

nrow(train[train$NumberOfOpenCreditLinesAndLoans < quantile(train$NumberOfOpenCreditLinesAndLoans, 0.99
```

```
## [1] 147760
```

```
train <- train[train$NumberOfOpenCreditLinesAndLoans < quantile(train$NumberOfOpenCreditLinesAndLoans, 
summary(train$NumberOfOpenCreditLinesAndLoans)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     5.0     8.0     8.2    11.0    23.0
```

3.3.8 NumberOfDependents

```
#-NumberOfDependents

summary(train$NumberOfDependents)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.7493  1.0000 20.0000
```

```
#omit outliers in the 99th percentile

nrow(train[train$NumberOfDependents < quantile(train$NumberOfDependents, 0.99),])
```

```
## [1] 143923
```

```
train <- train[train$NumberOfDependents < quantile(train$NumberOfDependents, 0.99),]
summary(train$NumberOfDependents)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.653   1.000   3.000
```

3.4 Abschluss

```
summary(train)
```

```
##        X          SeriousDlqin2yrs RevolvingUtilizationOfUnsecuredLines
##  Min.   :     1   Min.   :0.0000   Min.   :-1.00000
##  1st Qu.: 37477   1st Qu.:0.0000   1st Qu.: 0.02424
##  Median : 75033   Median :0.0000   Median : 0.13244
##  Mean   : 74995   Mean   :0.0659   Mean   : 0.27390
##  3rd Qu.:112465   3rd Qu.:0.0000   3rd Qu.: 0.50534
##  Max.   :150000   Max.   :1.0000   Max.   : 1.00000
##       age         NumberOfTime30_59DaysPastDueNotWorse   DebtRatio
##  Min.   : 21.00   Min.   :-1.000                       Min.   :    -1.00
##  1st Qu.: 41.00   1st Qu.: 0.000                       1st Qu.:      0.15
##  Median : 52.00   Median : 0.000                       Median :      0.35
##  Mean   : 52.39   Mean   : 0.239                       Mean   :   276.09
##  3rd Qu.: 63.00   3rd Qu.: 0.000                       3rd Qu.:      0.74
##  Max.   :109.00   Max.   :13.000                       Max.   :307001.00
##  MonthlyIncome   NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate
##  Min.   :    0   Min.   : 0.000                  Min.   :-1.00000
##  1st Qu.: 3000   1st Qu.: 5.000                  1st Qu.: 0.00000
##  Median : 5000   Median : 8.000                  Median : 0.00000
##  Mean   : 5841   Mean   : 8.189                  Mean   : 0.08791
##  3rd Qu.: 7600   3rd Qu.:11.000                  3rd Qu.: 0.00000
##  Max.   :49750   Max.   :23.000                  Max.   :17.00000
##  NumberRealEstateLoansOrLines NumberOfTime60_89DaysPastDueNotWorse
##  Min.   : 0.0000              Min.   :-1.00000
##  1st Qu.: 0.0000              1st Qu.: 0.00000
##  Median : 1.0000              Median : 0.00000
##  Mean   : 0.9913              Mean   : 0.06203
##  3rd Qu.: 2.0000              3rd Qu.: 0.00000
##  Max.   :15.0000              Max.   :11.00000
##  NumberOfDependents
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.653
##  3rd Qu.:1.000
##  Max.   :3.000
```

```
#remove ids
train <- train[-1]
test <- test[-1]
```

4 Modellierung

```
#seed for reproducibility
set.seed(123)
```

```
#model subset
train <- head(train,5000)
test <- head(test,5000)
```

4.1 Paket e1071

```
#seed for reproducibility
set.seed(123)

#svm classifier using e1071
```

```r
library(e1071)
library(caret)
```

```r
classifier_rbf <- svm(formula = SeriousDlqin2yrs ~ .,
                  data = train,
                  type = "C-classification",
                  kernel = "radial")

#train set prediction
pred_train <- predict(classifier_rbf,
                      newdata = train[-1])
```

```r
(cm = table(train[,1], pred_train))
```

```r
#test set prediction
pred_test <- predict(classifier_rbf,
                     newdata = test[-1])

summary(pred_test)
```

```
##      0      1
## 100921    582
```

4.2 Paket kernlab

```r
library(kernlab)
```

```r
mod <- ksvm(as.factor(train$SeriousDlqin2yrs)~.,
            data = train,
            kernel = "rbfdot",
            prob.model = TRUE)
```

```r
#model overview
mod
```

```
## Support Vector Machine object of class "ksvm"
##
## SV type: C-svc  (classification)
##  parameter : cost C = 1
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  0.130940534642902
##
## Number of Support Vectors : 951
##
## Objective Function Value : -530.298
## Training error : 0.0438
## Probability model included.
```

```r
#number of support vectors
mod@nSV
```

```
## [1] 951
```

```r
#line number of support vectors in the trainset
# mod@alphaindex
```

```r
#alpha values
# mod@alpha
#hyperplane coefficiants
# mod@coef
#negative intercept
mod@b
```

```
## [1] 0.5257232
```

```r
#error of the seperating hyperplane on the trainset
mod@error
```

```
## [1] 0.0438
```

```r
# prediction
u <- predict(mod, train[-1])

pred <- predict(mod, newdata = train[-1])
head(pred)
```

```
## [1] 0 0 0 0 0 0
## Levels: 0 1
```

```r
#confusion matrix

(cm = table(train[,1], pred))
```

```
##    pred
##           0      1
##   0 133908    530
##   1   8846    639
```

```r
#z scores
# mod@xmatrix
#scaled values
mod@scaling$scaled
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```r
#coefficiants
mod@coef[[1]][1]
```

```
## [1] 1
```

```r
mod@coef[[1]][2]
```

```
## [1] -0.5869333
```

```r
#prediction on the test set
pred.test <- predict(mod, test, type = "response")
summary(pred.test)
```

```
##      0      1
## 100893    610
```