# German Credit Case

Tuesday, June 7 2022

Niccolo Cherubini & Ivan Kostine

# Introduction

Case study of the german credit data set:

- 1000 past credit applicants information
- 30 variables + 1 Response

## Main Goal

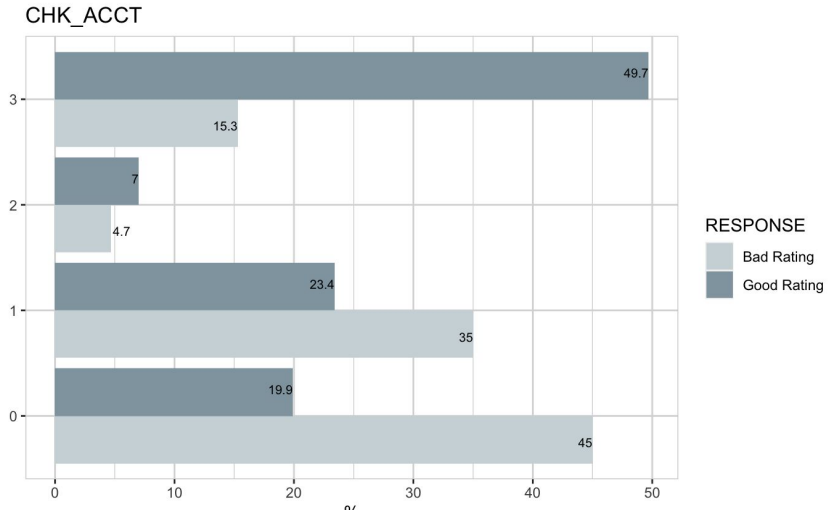
- Explore and understand Data
- Apply different models to predict new applicants credit risk.

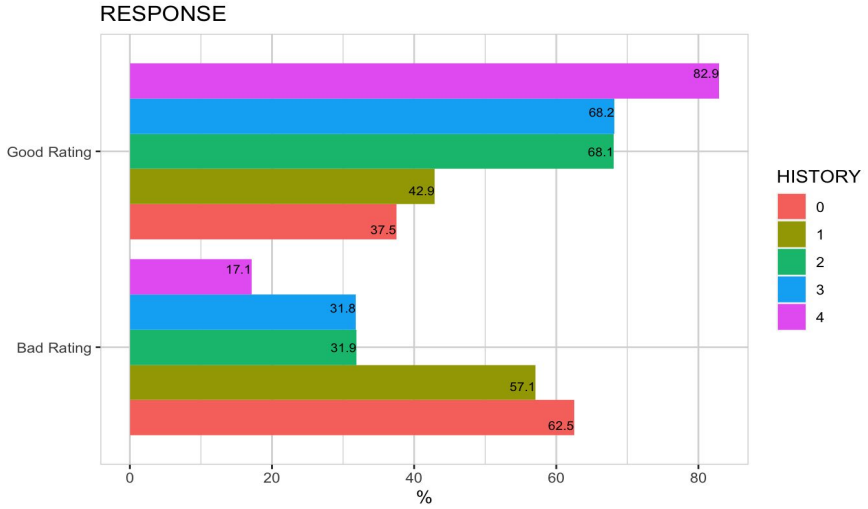
# Data Exploration

## Categorical Features

28 features:

- 18 binary (2 levels)
- 10 with more than 2 levels



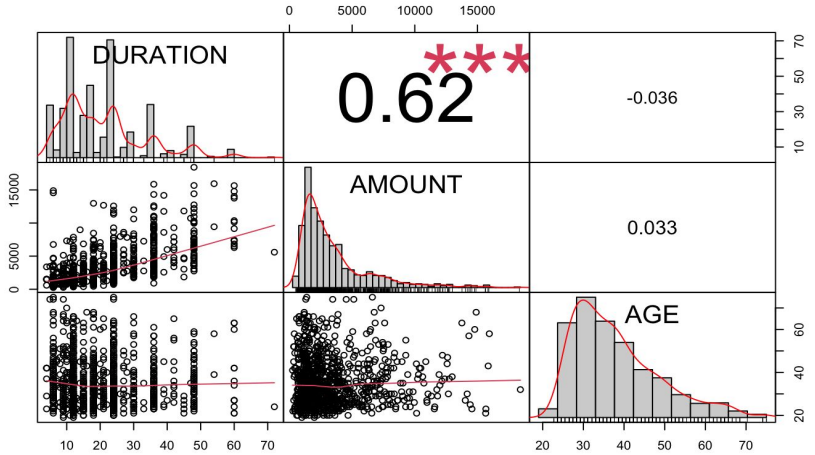



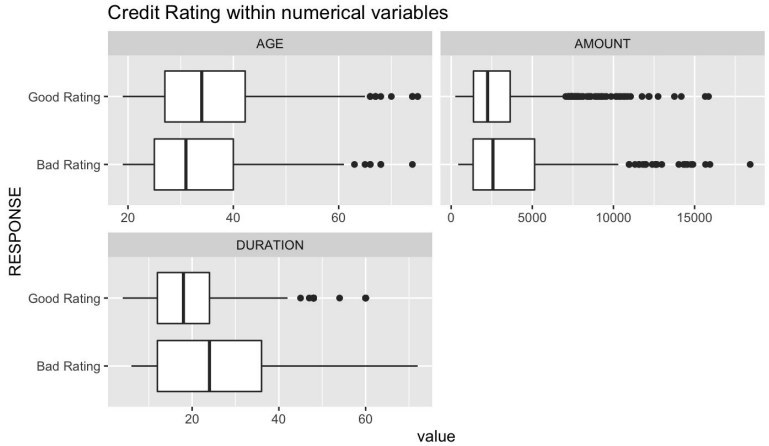
# Data Exploration

## Numerical Features

3 features:

- Credit duration
- Credit amount
- Age of the applicant



# Data Exploration

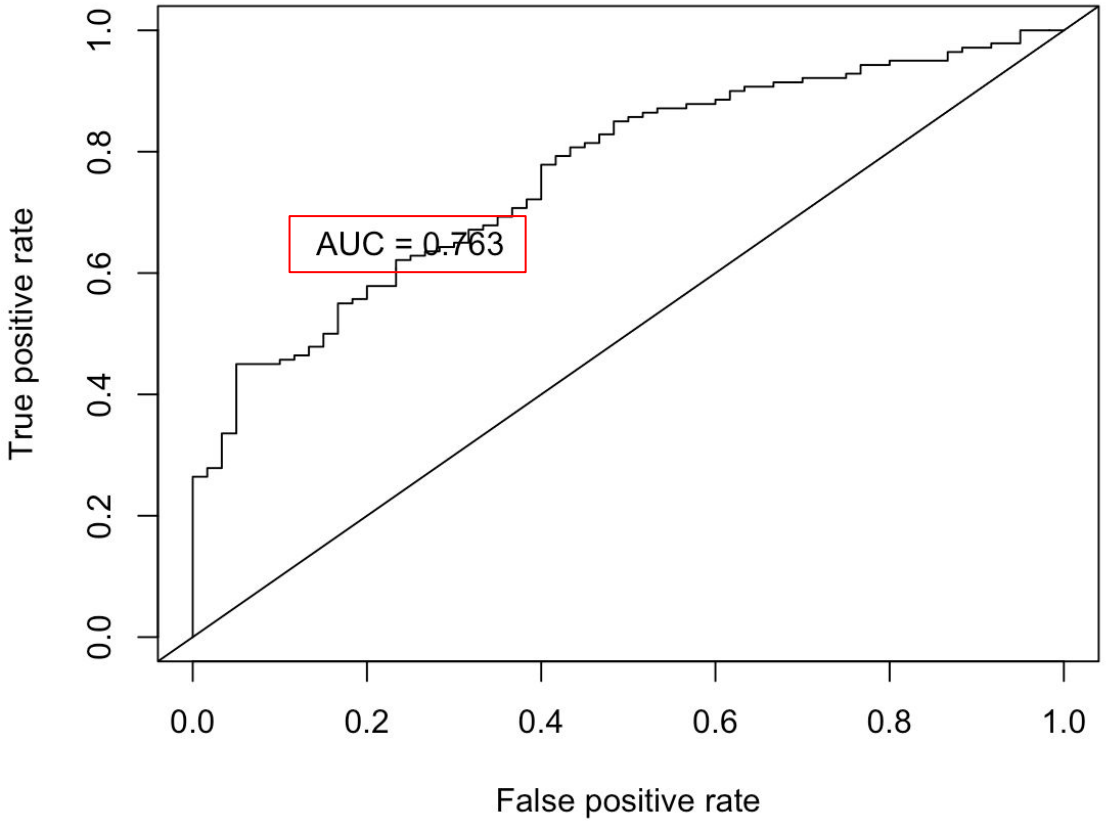
## Response Variable

- 1 if applicant has a Good Rating
- 0 if applicant has a Bad Rating

# Models

```
            Reference
Prediction     Bad Rating Good Rating
  Bad Rating           32          13
  Good Rating          28         127

               Accuracy : 0.795
                 95% CI : (0.732, 0.849)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 0.00161

                  Kappa : 0.474

 Mcnemar's Test P-Value : 0.02878

            Sensitivity : 0.907
            Specificity : 0.533
         Pos Pred Value : 0.819
         Neg Pred Value : 0.711
             Prevalence : 0.700
         Detection Rate : 0.635
   Detection Prevalence : 0.775
      Balanced Accuracy : 0.720

       'Positive' Class : Good Rating
```

## Linear Discriminant Analysis



# Logistic Regression

```
glm(formula = RESPONSE ~ ., family = binomial(), data = train.df)
```

Confusion Matrix and Statistics

```
          Reference
Prediction FALSE TRUE
     FALSE    34   24
     TRUE     26  116
```

```
               Accuracy : 0.75
                 95% CI : (0.684, 0.8084)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 0.06955

                  Kappa : 0.399

 Mcnemar's Test P-Value : 0.88754

            Sensitivity : 0.8286
            Specificity : 0.5667
         Pos Pred Value : 0.8169
         Neg Pred Value : 0.5862
             Prevalence : 0.7000
         Detection Rate : 0.5800
   Detection Prevalence : 0.7100
      Balanced Accuracy : 0.6976

       'Positive' Class : TRUE
```

Before AIC Backward Selection:

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         7.985e-01  1.131e+00   0.706 0.480089
CHK_ACCT1           6.469e-01  2.467e-01   2.622 0.008740 **
CHK_ACCT2           1.407e+00  4.398e-01   3.198 0.001384 **
CHK_ACCT3           1.755e+00  2.617e-01   6.706    2e-11 ***
DURATION           -2.364e-02  1.070e-02  -2.209 0.027166 *
HISTORY1            8.407e-02  6.118e-01   0.137 0.890700
HISTORY2            4.447e-01  4.827e-01   0.921 0.356969
HISTORY3            8.433e-01  5.157e-01   1.635 0.101972
HISTORY4            1.709e+00  4.884e-01   3.499 0.000466 ***
NEW_CAR1           -5.038e-01  4.268e-01  -1.180 0.237875
USED_CAR1           1.102e+00  5.376e-01   2.050 0.040341 *
FURNITURE1          2.716e-01  4.404e-01   0.617 0.537454
RADIO.TV1           3.755e-01  4.309e-01   0.871 0.383498
EDUCATION1         -6.979e-01  5.518e-01  -1.265 0.205931
RETRAINING1         3.081e-02  4.913e-01   0.063 0.949993
AMOUNT             -1.483e-04  5.026e-05  -2.950 0.003179 **
SAV_ACCT1           3.926e-01  3.436e-01   1.143 0.253146
SAV_ACCT2           4.506e-01  4.465e-01   1.009 0.312797
SAV_ACCT3           9.325e-01  5.727e-01   1.628 0.103484
SAV_ACCT4           8.436e-01  2.920e-01   2.890 0.003858 **
EMPLOYMENT1        -2.868e-01  4.811e-01  -0.596 0.551003
EMPLOYMENT2         4.290e-01  4.660e-01   0.921 0.357287
EMPLOYMENT3         7.842e-01  5.021e-01   1.562 0.118336
EMPLOYMENT4         7.667e-04  4.674e-01   0.002 0.998691
INSTALL_RATE2      -4.323e-01  3.500e-01  -1.235 0.216792
INSTALL_RATE3      -6.862e-01  3.948e-01  -1.738 0.082169 .
INSTALL_RATE4      -9.869e-01  3.469e-01  -2.845 0.004446 **
MALE_DIV1          -4.403e-02  4.401e-01  -0.100 0.920323
MALE_SINGLE1        6.436e-01  2.425e-01   2.654 0.007962 **
MALE_MAR_or_WID1    2.438e-01  3.591e-01   0.679 0.497290
CO.APPLICANT1      -4.646e-01  4.659e-01  -0.997 0.318615
GUARANTOR1          1.080e+00  4.966e-01   2.175 0.029659 *
PRESENT_RESIDENT2  -7.747e-01  3.331e-01  -2.326 0.020042 *
PRESENT_RESIDENT3  -5.702e-01  3.724e-01  -1.531 0.125696
PRESENT_RESIDENT4  -3.545e-01  3.378e-01  -1.049 0.294027
REAL_ESTATE1        1.405e-01  2.429e-01   0.578 0.562959
PROP_UNKN_NONE1    -4.205e-01  4.483e-01  -0.938 0.348227
AGE                 5.164e-03  1.050e-02   0.492 0.622715
OTHER_INSTALL1     -5.739e-01  2.382e-01  -2.410 0.015961 *
RENT1              -3.706e-01  5.564e-01  -0.666 0.505357
OWN_RES1           -1.369e-01  5.232e-01  -0.262 0.793631
NUM_CREDITS2       -4.280e-01  2.759e-01  -1.551 0.120848
NUM_CREDITS3       -1.728e-01  7.915e-01  -0.218 0.827143
NUM_CREDITS4       -6.465e-01  1.167e+00  -0.554 0.579495
JOB1                1.160e-02  7.420e-01   0.016 0.987528
JOB2               -1.533e-01  7.081e-01  -0.216 0.828607
JOB3                1.123e-01  7.161e-01   0.157 0.875424
NUM_DEPENDENTS2    -2.735e-01  2.808e-01  -0.974 0.330176
TELEPHONE1          2.701e-01  2.260e-01   1.195 0.232041
FOREIGN1            2.018e+00  9.128e-01   2.211 0.027028 *
```

After AIC Backward Selection:

```
Step:  AIC=786.05
RESPONSE ~ CHK_ACCT + DURATION + HISTORY + NEW_CAR + USED_CAR +
    EDUCATION + AMOUNT + SAV_ACCT + EMPLOYMENT + INSTALL_RATE +
    MALE_SINGLE + GUARANTOR + PRESENT_RESIDENT + OTHER_INSTALL +
    TELEPHONE + FOREIGN

                   Df Deviance    AIC
<none>                  722.05 786.05
- TELEPHONE         1   724.43 786.43
- PRESENT_RESIDENT  3   728.73 786.73
- INSTALL_RATE      3   731.27 789.27
- SAV_ACCT          4   733.52 789.52
- USED_CAR          1   727.74 789.74
- DURATION          1   727.82 789.82
- OTHER_INSTALL     1   728.34 790.34
- EDUCATION         1   728.72 790.72
- GUARANTOR         1   728.98 790.98
- MALE_SINGLE       1   729.51 791.51
- EMPLOYMENT        4   735.54 791.54
- FOREIGN           1   729.60 791.60
- NEW_CAR           1   732.57 794.57
- AMOUNT            1   733.61 795.61
- HISTORY           4   748.40 804.40
- CHK_ACCT          3   780.85 838.85
```

# Support Vector Machines

```
Support Vector Machines with Radial Basis Function Kernel

800 samples
 30 predictor
  2 classes: 'Bad Rating', 'Good Rating'

No pre-processing
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 533, 534, 533
Resampling results across tuning parameters:

  C      sigma  Accuracy   Kappa
    0.1  1      0.6999991  0
    0.1  2      0.6999991  0
    0.1  3      0.6999991  0
    0.1  4      0.6999991  0
    1.0  1      0.6999991  0
    1.0  2      0.6999991  0
    1.0  3      0.6999991  0
    1.0  4      0.6999991  0
   10.0  1      0.6999991  0
   10.0  2      0.6999991  0
   10.0  3      0.6999991  0
   10.0  4      0.6999991  0
  100.0  1      0.6999991  0
  100.0  2      0.6999991  0
  100.0  3      0.6999991  0
  100.0  4      0.6999991  0

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were sigma = 4 and C = 0.1.
```

C=0.1. sigma = 4

| Prediction | Reference: Bad Rating | Reference: Good Rating |
|---|---|---|
| Bad Rating | 0 | 0 |
| Good Rating | 60 | 140 |

C = 10, sigma = default

| Prediction | Reference: Bad Rating | Reference: Good Rating |
|---|---|---|
| Bad Rating | 60 | 0 |
| Good Rating | 0 | 140 |

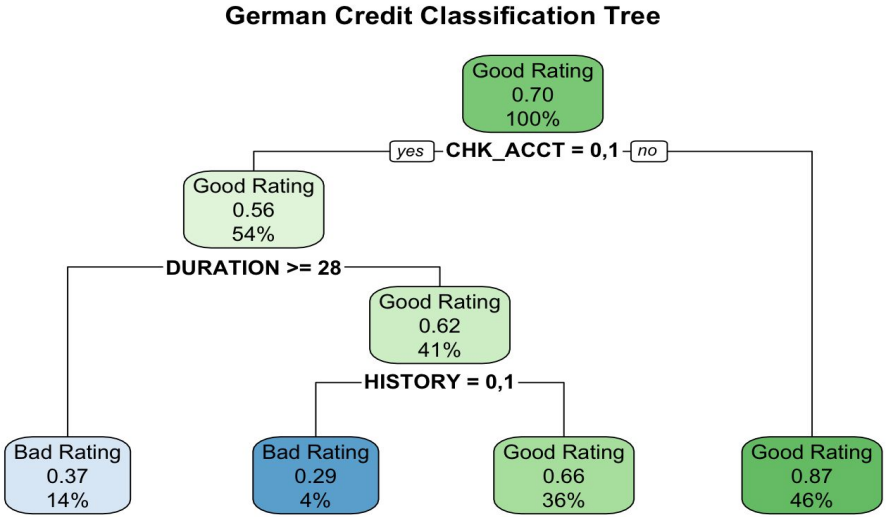
```
Support Vector Machines with Linear Kernel

800 samples
 30 predictor
  2 classes: 'Bad Rating', 'Good Rating'

No pre-processing
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 533, 534, 533
Resampling results across tuning parameters:

  C      Accuracy   Kappa
    0.1  0.7037538  0.1789957
    1.0  0.7187491  0.2128241
   10.0  0.7149850  0.1998996
  100.0  0.7187444  0.2310682

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was C = 1.
```

| Prediction | Reference: Bad Rating | Reference: Good Rating |
|---|---|---|
| Bad Rating | 36 | 12 |
| Good Rating | 24 | 128 |

```
               Accuracy : 0.82
                 95% CI : (0.76, 0.871)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 7.54e-05

                  Kappa : 0.545

 Mcnemar's Test P-Value : 0.0668

            Sensitivity : 0.914
            Specificity : 0.600
         Pos Pred Value : 0.842
         Neg Pred Value : 0.750
             Prevalence : 0.700
         Detection Rate : 0.640
   Detection Prevalence : 0.760
      Balanced Accuracy : 0.757

       'Positive' Class : Good Rating
```

We can observe that the SVM using Radial Kernel makes only good predictions (no error), it probably means that the model is overfitting and therefore it is not optimal prediction. The SVM linear has a quite good balanced accuracy

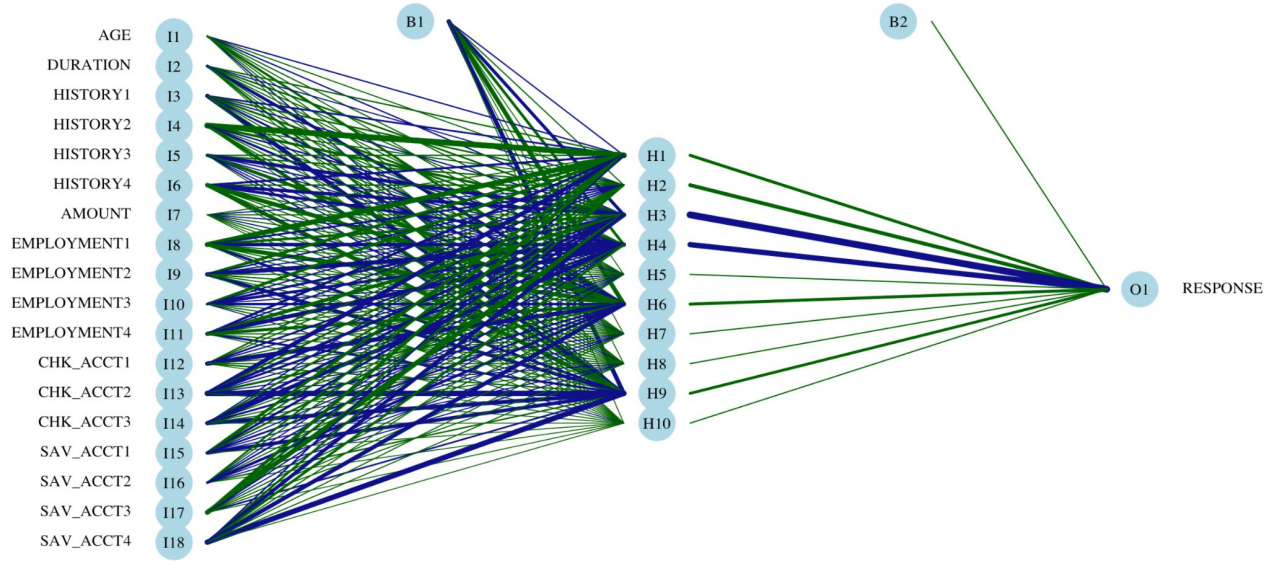
# CART




```
Confusion Matrix and Statistics

             Reference
Prediction    Bad Rating Good Rating
  Bad Rating          33          20
  Good Rating         57         190

               Accuracy : 0.743
                 95% CI : (0.69, 0.792)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 0.0561

                  Kappa : 0.308

 Mcnemar's Test P-Value : 4.09e-05

            Sensitivity : 0.905
            Specificity : 0.367
         Pos Pred Value : 0.769
         Neg Pred Value : 0.623
             Prevalence : 0.700
         Detection Rate : 0.633
   Detection Prevalence : 0.823
      Balanced Accuracy : 0.636

       'Positive' Class : Good Rating
```

- Pruned Tree: 3 nodes
- Really high sensitivity
- Low balanced accuracy

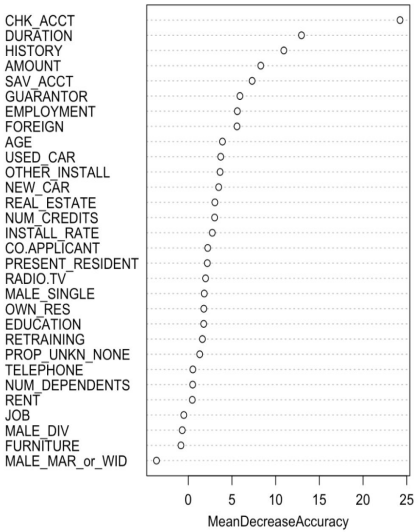
# Neural Network

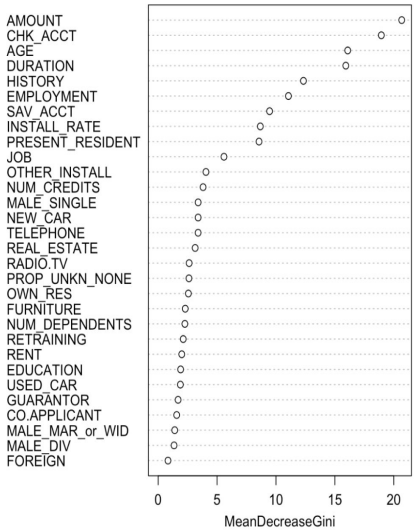



```
Confusion Matrix and Statistics

                Reference
Prediction      Bad Rating Good Rating
  Bad Rating            42          32
  Good Rating           48         178

               Accuracy : 0.733
                 95% CI : (0.679, 0.783)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 0.1149

                  Kappa : 0.331

 Mcnemar's Test P-Value : 0.0935

            Sensitivity : 0.848
            Specificity : 0.467
         Pos Pred Value : 0.788
         Neg Pred Value : 0.568
             Prevalence : 0.700
         Detection Rate : 0.593
   Detection Prevalence : 0.753
      Balanced Accuracy : 0.657

       'Positive' Class : Good Rating
```

- high sensitivity, low specificity -> balanced accuracy not so good. Model predicts too many Good Ratings

# Random Forest




```
Confusion Matrix and Statistics

          Reference
Prediction    Bad Rating Good Rating
  Bad Rating          65          63
  Good Rating         25         147

               Accuracy : 0.707
                 95% CI : (0.652, 0.758)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 0.428

                  Kappa : 0.377

 Mcnemar's Test P-Value : 8.01e-05

            Sensitivity : 0.700
            Specificity : 0.722
         Pos Pred Value : 0.855
         Neg Pred Value : 0.508
             Prevalence : 0.700
         Detection Rate : 0.490
   Detection Prevalence : 0.573
      Balanced Accuracy : 0.711

       'Positive' Class : Good Rating
```

- Variables importance:
  - checking account
  - duration (month)
  - credit history
  - age
  - employment
  - savings account

- Model fit using Cross-validation:
  - good specificity and sensitivity -> good balanced accuracy

# Results

- High amount of good ratings in the initial data
- All the models tend to predict good ratings
- Best models:
  - Random Forest
  - SVM Linear

## Business advice:

- Review the credit rating system, what makes a good or a bad rating.
- Use our Random Forest or SVM linear models to make future applicants credit rating predictions.