Previsão de risco de crédito usando Regressão Logística

Fran mateus

25/02/2022

Bibliotecas e dataset:

```
library(caret)
```

```
## Carregando pacotes exigidos: ggplot2
```

```
## Carregando pacotes exigidos: lattice
```

```
library(ROCR)
library(e1071)
```

Dataset:

```
credito_dataset <- read.csv("credit_dataset_final.csv", header = TRUE, sep = ",")
head(credito_dataset)
```

```
##   credit.rating account.balance credit.duration.months
## 1             1               1                     18
## 2             1               1                      9
## 3             1               2                     12
## 4             1               1                     12
## 5             1               1                     12
## 6             1               1                     10
##   previous.credit.payment.status credit.purpose credit.amount savings
## 1                              3              2          1049       1
## 2                              3              4          2799       1
## 3                              2              4           841       2
## 4                              3              4          2122       1
## 5                              3              4          2171       1
## 6                              3              4          2241       1
##   employment.duration installment.rate marital.status guarantor
## 1                   1                4              1         1
## 2                   2                2              3         1
## 3                   3                2              1         1
## 4                   2                3              3         1
## 5                   2                4              3         1
## 6                   1                1              3         1
##   residence.duration current.assets age other.credits apartment.type
## 1                  4              2  21             2              1
## 2                  2              1  36             2              1
## 3                  4              1  23             2              1
## 4                  2              1  39             2              1
## 5                  4              2  38             1              2
## 6                  3              1  48             2              1
##   bank.credits occupation dependents telephone foreign.worker
## 1            1          3          1         1              1
## 2            2          3          2         1              1
## 3            1          2          1         1              1
## 4            2          2          2         1              2
## 5            2          2          1         1              2
## 6            2          2          2         1              2
```

```
summary(credito_dataset)
```

```
##  credit.rating account.balance credit.duration.months
##  Min.   :0.0   Min.   :1.000   Min.   : 4.0
##  1st Qu.:0.0   1st Qu.:1.000   1st Qu.:12.0
##  Median :1.0   Median :2.000   Median :18.0
##  Mean   :0.7   Mean   :2.183   Mean   :20.9
##  3rd Qu.:1.0   3rd Qu.:3.000   3rd Qu.:24.0
##  Max.   :1.0   Max.   :3.000   Max.   :72.0
##  previous.credit.payment.status credit.purpose  credit.amount      savings
##  Min.   :1.000                  Min.   :1.000   Min.   :  250   Min.   :1.000
##  1st Qu.:2.000                  1st Qu.:2.000   1st Qu.: 1366   1st Qu.:1.000
##  Median :2.000                  Median :3.000   Median : 2320   Median :1.000
##  Mean   :2.292                  Mean   :2.965   Mean   : 3271   Mean   :1.874
##  3rd Qu.:3.000                  3rd Qu.:4.000   3rd Qu.: 3972   3rd Qu.:3.000
##  Max.   :3.000                  Max.   :4.000   Max.   :18424   Max.   :4.000
##  employment.duration installment.rate marital.status    guarantor
##  Min.   :1.000       Min.   :1.000    Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000       1st Qu.:2.000    1st Qu.:1.000   1st Qu.:1.000
##  Median :2.000       Median :3.000    Median :3.000   Median :1.000
##  Mean   :2.446       Mean   :2.973    Mean   :2.372   Mean   :1.093
##  3rd Qu.:4.000       3rd Qu.:4.000    3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :4.000       Max.   :4.000    Max.   :4.000   Max.   :2.000
##  residence.duration current.assets       age        other.credits
##  Min.   :1.000      Min.   :1.000   Min.   :19.00   Min.   :1.000
##  1st Qu.:2.000      1st Qu.:1.000   1st Qu.:27.00   1st Qu.:2.000
##  Median :3.000      Median :2.000   Median :33.00   Median :2.000
##  Mean   :2.845      Mean   :2.358   Mean   :35.54   Mean   :1.814
##  3rd Qu.:4.000      3rd Qu.:3.000   3rd Qu.:42.00   3rd Qu.:2.000
##  Max.   :4.000      Max.   :4.000   Max.   :75.00   Max.   :2.000
##  apartment.type   bank.credits     occupation      dependents
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:1.000
##  Median :2.000   Median :1.000   Median :3.000   Median :1.000
##  Mean   :1.928   Mean   :1.367   Mean   :2.904   Mean   :1.155
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :3.000   Max.   :2.000   Max.   :4.000   Max.   :2.000
##    telephone     foreign.worker
##  Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.000
##  Median :1.000   Median :1.000
##  Mean   :1.404   Mean   :1.037
##  3rd Qu.:2.000   3rd Qu.:1.000
##  Max.   :2.000   Max.   :2.000
```

```
str(credito_dataset)
```

```
## 'data.frame':    1000 obs. of  21 variables:
##  $ credit.rating                 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ account.balance               : int  1 1 2 1 1 1 1 1 3 2 ...
##  $ credit.duration.months        : int  18 9 12 12 12 10 8 6 18 24 ...
##  $ previous.credit.payment.status: int  3 3 2 3 3 3 3 3 3 2 ...
##  $ credit.purpose                : int  2 4 4 4 4 4 4 4 3 3 ...
##  $ credit.amount                 : int  1049 2799 841 2122 2171 2241 3398 1361 1098 3758
...
##  $ savings                       : int  1 1 2 1 1 1 1 1 1 3 ...
##  $ employment.duration           : int  1 2 3 2 2 1 3 1 1 1 ...
##  $ installment.rate              : int  4 2 2 3 4 1 1 2 4 1 ...
##  $ marital.status                : int  1 3 1 3 3 3 3 3 1 1 ...
##  $ guarantor                     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ residence.duration            : int  4 2 4 2 4 3 4 4 4 4 ...
##  $ current.assets                : int  2 1 1 1 2 1 1 1 3 4 ...
##  $ age                           : int  21 36 23 39 38 48 39 40 65 23 ...
##  $ other.credits                 : int  2 2 2 2 1 2 2 2 2 2 ...
##  $ apartment.type                : int  1 1 1 1 2 1 2 2 2 1 ...
##  $ bank.credits                  : int  1 2 1 2 2 2 2 1 2 1 ...
##  $ occupation                    : int  3 3 2 2 2 2 2 2 1 1 ...
##  $ dependents                    : int  1 2 1 2 1 2 1 2 1 1 ...
##  $ telephone                     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ foreign.worker                : int  1 1 1 2 2 2 2 2 1 1 ...
```

Pré-processamento:

Conversão das variáveis numéricas que são categóricas em fatores:

```
to.factors <- function(df, variables){
  for (variable in variables){
    df[[variable]] <- as.factor(df[[variable]])
  }
  return(df)
}
```

Normalização das variáveis:

```
scale.features <- function(df, variables){
  for (variable in variables){
    df[[variable]] <- scale(df[[variable]], center = T, scale = T)
  }
  return(df)
}
```

Lista de variáveis numéricas

```
numeric.vars <- c("credit.duration.months", "age", "credit.amount")
```

```
credito_dataset_scaled <- scale.features(credito_dataset, numeric.vars)
```

Lista de variáveis categóricas:

```
categorical.vars <- c('credit.rating', 'account.balance', 'previous.credit.payment.status',
                      'credit.purpose', 'savings', 'employment.duration', 'installment.rate',
                      'marital.status', 'guarantor', 'residence.duration', 'current.assets',
                      'other.credits', 'apartment.type', 'bank.credits', 'occupation',
                      'dependents', 'telephone', 'foreign.worker')
```

Aplicando as conversões ao dataset

```
credito_dataset_final <- to.factors(df = credito_dataset_scaled, variables = categorical.var
s)
head(credito_dataset_final)
```

```
##   credit.rating account.balance credit.duration.months
## 1             1               1             -0.2407368
## 2             1               1             -0.9870788
## 3             1               2             -0.7382981
## 4             1               1             -0.7382981
## 5             1               1             -0.7382981
## 6             1               1             -0.9041519
##   previous.credit.payment.status credit.purpose credit.amount savings
## 1                              3              2    -0.7872630       1
## 2                              3              4    -0.1673006       1
## 3                              2              4    -0.8609500       2
## 4                              3              4    -0.4071375       1
## 5                              3              4    -0.3897785       1
## 6                              3              4    -0.3649800       1
##   employment.duration installment.rate marital.status guarantor
## 1                   1                4              1         1
## 2                   2                2              3         1
## 3                   3                2              1         1
## 4                   2                3              3         1
## 5                   2                4              3         1
## 6                   1                1              3         1
##   residence.duration current.assets         age other.credits apartment.type
## 1                  4              2 -1.28093214             2              1
## 2                  2              1  0.04034293             2              1
## 3                  4              1 -1.10476213             2              1
## 4                  2              1  0.30459795             2              1
## 5                  4              2  0.21651294             1              2
## 6                  3              1  1.09736299             2              1
##   bank.credits occupation dependents telephone foreign.worker
## 1            1          3          1         1              1
## 2            2          3          2         1              1
## 3            1          2          1         1              1
## 4            2          2          2         1              2
## 5            2          2          1         1              2
## 6            2          2          2         1              2
```

```
summary(credito_dataset_final)
```

```
##  credit.rating account.balance credit.duration.months.V1
##  0:300         1:274           Min.   :-1.401713
##  1:700         2:269           1st Qu.:-0.738298
##                3:457           Median :-0.240737
##                                Mean   : 0.000000
##                                3rd Qu.: 0.256825
##                                Max.   : 4.237315
##  previous.credit.payment.status credit.purpose  credit.amount.V1   savings
##  1: 89                           1:103          Min.   :-1.070320   1:603
##  2:530                           2:181          1st Qu.:-0.675138   2:103
##  3:381                           3:364          Median :-0.337170   3:111
##                                  4:352          Mean   : 0.000000   4:183
##                                                 3rd Qu.: 0.248340
##                                                 Max.   : 5.368078
##  employment.duration installment.rate marital.status guarantor
##  1:234               1:136            1:360          1:907
##  2:339               2:231            3:548          2: 93
##  3:174               3:157            4: 92
##  4:253               4:476
##
##
##  residence.duration current.assets       age.V1        other.credits
##  1:130              1:282          Min.   :-1.457102   1:186
##  2:308              2:232          1st Qu.:-0.752422   2:814

##  3:149              3:332          Median :-0.223912
##  4:413              4:154          Mean   : 0.000000
##                                    3rd Qu.: 0.568853
##                                    Max.   : 3.475658
##  apartment.type bank.credits occupation dependents telephone foreign.worker
##  1:179          1:633        1: 22      1:845      1:596     1:963
##  2:714          2:367        2:200      2:155      2:404     2: 37
##  3:107                       3:630
##                              4:148
##
##
```

Preparando os dados de treino e de teste

```
indexes <- sample(1:nrow(credito_dataset_final), size = 0.6 * nrow(credito_dataset_final))
train.data <- credito_dataset_final[indexes,]
test.data <- credito_dataset_final[-indexes,]
```

```
class(train.data)
```

```
## [1] "data.frame"
```

```
class(test.data)
```

```
## [1] "data.frame"
```

Separando os atributos e as classes

```
test.feature.vars <- test.data[,-1]
test.class.var <- test.data[,1]
```

```
class(test.feature.vars)
```

```
## [1] "data.frame"
```

```
class(test.class.var)
```

```
## [1] "factor"
```

Construindo o modelo de regressão logística (Fitting Generalized Linear Model)

```
formula.init <- "credit.rating ~ ."
formula.init <- as.formula(formula.init)
```

```
modelo_v1 <- glm(formula = formula.init, data = train.data, family = "binomial")
```

Visualizando os detalhes do modelo

```
summary(modelo_v1)
```

```
##
## Call:
## glm(formula = formula.init, family = "binomial", data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7148  -0.7341   0.4057   0.7100   1.9961
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     0.15208    0.93337   0.163 0.870565
## account.balance2                0.54619    0.26990   2.024 0.043007 *
## account.balance3                1.87908    0.28121   6.682 2.35e-11 ***
## credit.duration.months         -0.32406    0.14712  -2.203 0.027610 *
## previous.credit.payment.status2 0.73008    0.39225   1.861 0.062705 .
## previous.credit.payment.status3 1.47417    0.40971   3.598 0.000321 ***
## credit.purpose2                -0.74631    0.48368  -1.543 0.122835
## credit.purpose3                -1.12913    0.46079  -2.450 0.014269 *
## credit.purpose4                -1.27091    0.44744  -2.840 0.004506 **
## credit.amount                  -0.39751    0.16699  -2.380 0.017293 *
## savings2                        0.28249    0.36914   0.765 0.444119
## savings3                        0.47424    0.39437   1.203 0.229155
## savings4                        1.01348    0.34558   2.933 0.003360 **
## employment.duration2            0.37583    0.30584   1.229 0.219130
## employment.duration3            0.51739    0.35200   1.470 0.141608
## employment.duration4            0.38679    0.35585   1.087 0.277054
## installment.rate2              -0.76023    0.40437  -1.880 0.060105 .
## installment.rate3              -1.04824    0.43339  -2.419 0.015576 *
## installment.rate4              -1.17954    0.39105  -3.016 0.002559 **
## marital.status3                 0.74425    0.26004   2.862 0.004210 **
## marital.status4                 0.61745    0.39296   1.571 0.116121
## guarantor2                      0.47307    0.38175   1.239 0.215260
## residence.duration2            -0.45876    0.36654  -1.252 0.210717
## residence.duration3            -0.20346    0.40224  -0.506 0.612983
## residence.duration4             0.04985    0.36430   0.137 0.891164
## current.assets2                -0.12528    0.31963  -0.392 0.695099
## current.assets3                 0.17535    0.29444   0.596 0.551500
## current.assets4                -0.73506    0.51219  -1.435 0.151253
## age                             0.03185    0.12775   0.249 0.803110
## other.credits2                  0.14069    0.28565   0.493 0.622342
## apartment.type2                 0.34116    0.29966   1.138 0.254919
## apartment.type3                 0.58714    0.57694   1.018 0.308836
## bank.credits2                  -0.25348    0.30100  -0.842 0.399727
## occupation2                    -0.50554    0.67772  -0.746 0.455700
## occupation3                    -0.36000    0.64504  -0.558 0.576774
## occupation4                    -0.64015    0.70374  -0.910 0.363015
## dependents2                    -0.01848    0.32113  -0.058 0.954112
## telephone2                      0.31903    0.26502   1.204 0.228673
## foreign.worker2                 1.80139    0.97851   1.841 0.065628 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 741.31  on 599  degrees of freedom
```

```
## Residual deviance: 553.80  on 561  degrees of freedom
## AIC: 631.8
##
## Number of Fisher Scoring iterations: 5
```

Fazendo previsões e analisando o resultado

```
previsoes <- predict(modelo_v1, test.data, type = "response")
previsoes <- round(previsoes)
```

```
View(previsoes)
```

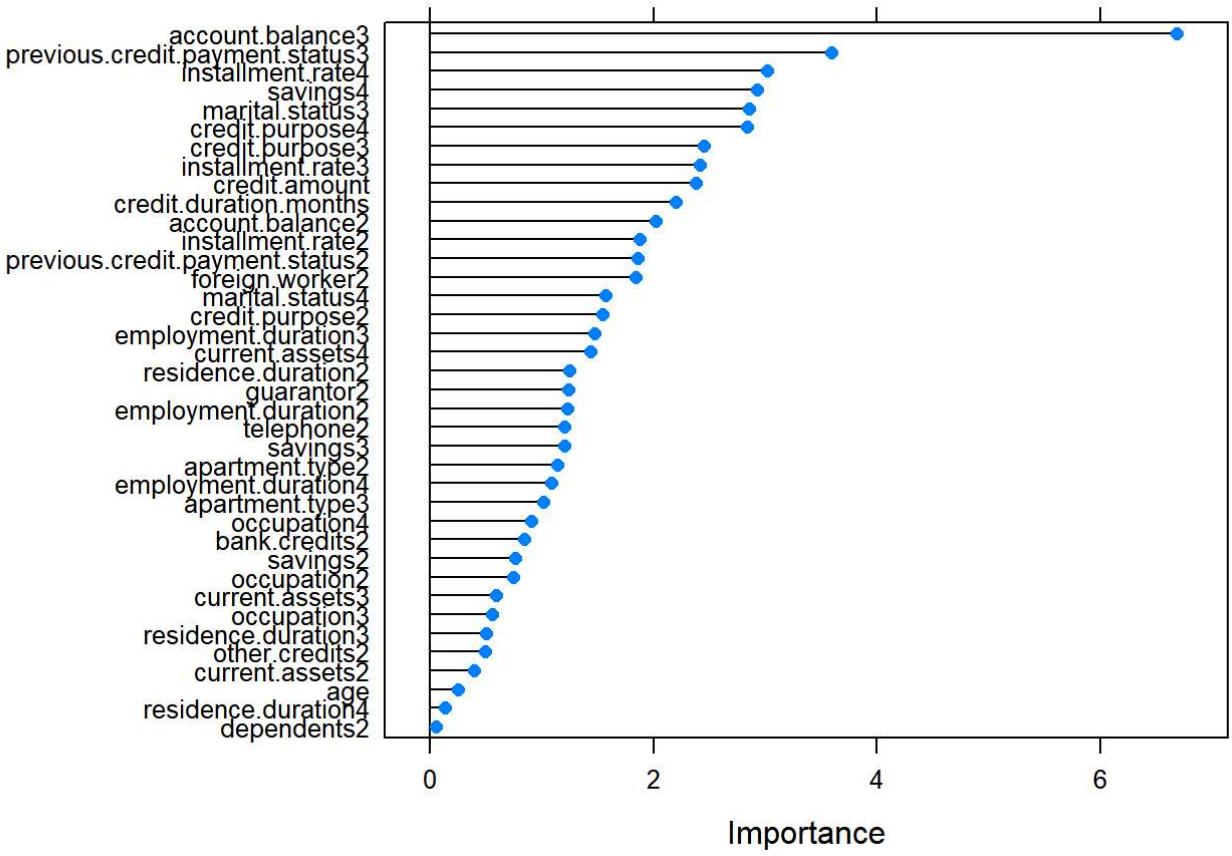
Confusion Matrix para comparar o valor observado com as previsões do modelo

```
confusionMatrix(table(data = previsoes, reference = test.class.var), positive = '1')
```

```
## Confusion Matrix and Statistics
##
##     reference
## data   0   1
##    0  63  39
##    1  52 246
##
##                Accuracy : 0.7725
##                  95% CI : (0.7282, 0.8127)
##     No Information Rate : 0.7125
##     P-Value [Acc > NIR] : 0.004056
##
##                   Kappa : 0.4253
##
##  Mcnemar's Test P-Value : 0.208413
##
##             Sensitivity : 0.8632
##             Specificity : 0.5478
##          Pos Pred Value : 0.8255
##          Neg Pred Value : 0.6176
##              Prevalence : 0.7125
##          Detection Rate : 0.6150
##    Detection Prevalence : 0.7450
##       Balanced Accuracy : 0.7055
##
##        'Positive' Class : 1
##
```

Feature Selection para descobrir quais são as variáveis mais relevantes para esse modelo

```
formula <- "credit.rating ~ ."
formula <- as.formula(formula)
control <- trainControl(method = "repeatedcv", number = 10, repeats = 2)
model <- train(formula, data = train.data, method = "glm", trControl = control)
importance <- varImp(model, scale = FALSE)
plot(importance)
```



Construindo um novo modelo com as variáveis selecionadas

```
formula.new <- "credit.rating ~ account.balance + credit.purpose + previous.credit.payment.st
atus + savings + credit.duration.months"
formula.new <- as.formula(formula.new)
modelo_v2 <- glm(formula = formula.new, data = train.data, family = "binomial")
```

Visualizando o novo modelo

```
summary(modelo_v2)
```

```
##
## Call:
## glm(formula = formula.new, family = "binomial", data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5510  -0.8803   0.4644   0.7854   1.9411
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.3281     0.4737  -0.693  0.48853
## account.balance2                 0.5027     0.2466   2.038  0.04154 *
## account.balance3                 1.7978     0.2593   6.934 4.08e-12 ***
## credit.purpose2                 -0.5903     0.4366  -1.352  0.17632
## credit.purpose3                 -0.8864     0.4067  -2.179  0.02931 *
## credit.purpose4                 -0.9557     0.4023  -2.376  0.01751 *
## previous.credit.payment.status2  0.7886     0.3394   2.323  0.02016 *
## previous.credit.payment.status3  1.4332     0.3537   4.052 5.09e-05 ***
## savings2                         0.1761     0.3364   0.523  0.60070
## savings3                         0.4423     0.3640   1.215  0.22424
## savings4                         0.8319     0.3133   2.655  0.00792 **
## credit.duration.months          -0.4958     0.1037  -4.781 1.75e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 741.31  on 599  degrees of freedom
## Residual deviance: 597.48  on 588  degrees of freedom
## AIC: 621.48
##
## Number of Fisher Scoring iterations: 5
```

Prevendo e Avaliando o modelo

```
previsoes_new <- predict(modelo_v2, test.data, type = "response")
previsoes_new <- round(previsoes_new)
```

Confusion Matrix

```
confusionMatrix(table(data = previsoes_new, reference = test.class.var), positive = '1')
```

```
## Confusion Matrix and Statistics
##
##     reference
## data   0   1
##    0  55  40
##    1  60 245
##
##                Accuracy : 0.75
##                  95% CI : (0.7046, 0.7917)
##     No Information Rate : 0.7125
##     P-Value [Acc > NIR] : 0.05313
##
##                   Kappa : 0.3564
##
##  Mcnemar's Test P-Value : 0.05743
##
##             Sensitivity : 0.8596
##             Specificity : 0.4783
##          Pos Pred Value : 0.8033
##          Neg Pred Value : 0.5789
##              Prevalence : 0.7125
##          Detection Rate : 0.6125
##    Detection Prevalence : 0.7625
##       Balanced Accuracy : 0.6690
##
##        'Positive' Class : 1
##
```

Avaliando a performance do modelo

Plot do modelo com a melhor acurácia

```
modelo_final <- modelo_v2
previsoes <- predict(modelo_final, test.feature.vars, type = "response")
previsoes_finais <- prediction(previsoes, test.class.var)
```

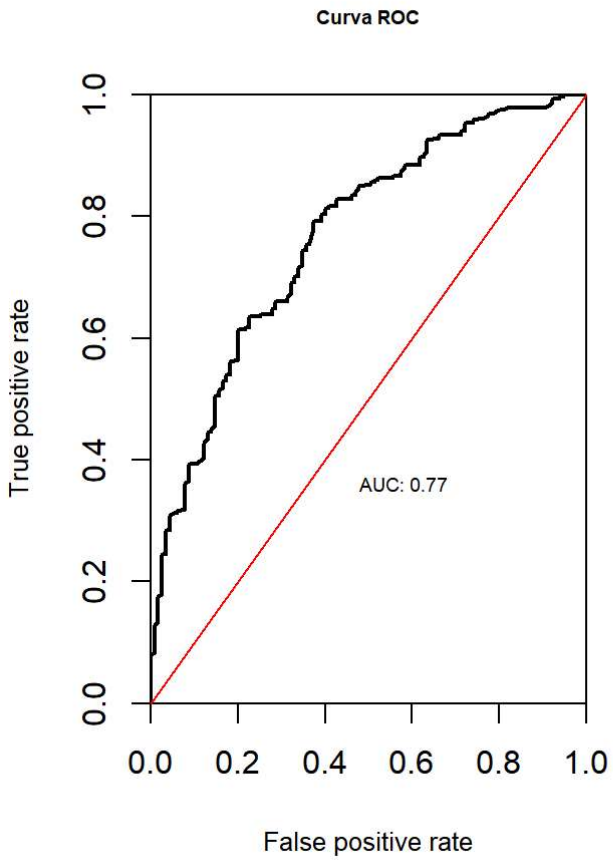
Função para Plot ROC

```
plot.roc.curve <- function(predictions, title.text){
  perf <- performance(predictions, "tpr", "fpr")
  plot(perf,col = "black",lty = 1, lwd = 2,
       main = title.text, cex.main = 0.6, cex.lab = 0.8,xaxs = "i", yaxs = "i")
  abline(0,1, col = "red")
  auc <- performance(predictions,"auc")
  auc <- unlist(slot(auc, "y.values"))
  auc <- round(auc,2)
  legend(0.4,0.4,legend = c(paste0("AUC: ",auc)), cex = 0.6, bty = "n", box.col = "white")

}
```

Plot

```
par(mfrow = c(1, 2))
plot.roc.curve(previsoes_finais, title.text = "Curva ROC")
```



Fazendo previsões em novos dados

Novos dados

```
account.balance <- c(1, 3, 3, 2)
credit.purpose <- c(4, 2, 3, 2)
previous.credit.payment.status <- c(3, 3, 2, 2)
savings <- c(2, 3, 2, 3)
credit.duration.months <- c(15, 12, 8, 6)
```

Cria um dataframe

```
novo_dataset <- data.frame(account.balance,
                           credit.purpose,
                           previous.credit.payment.status,
                           savings,
                          credit.duration.months)
```

```
View(novo_dataset)
```

Separa variáveis explanatórias numéricas e categóricas

```
new.numeric.vars <- c("credit.duration.months")
new.categorical.vars <- c('account.balance', 'previous.credit.payment.status',
                          'credit.purpose', 'savings')
```

Aplica as transformações

```
novo_dataset_final <- to.factors(df = novo_dataset, variables = new.categorical.vars)
str(novo_dataset_final)
```

```
## 'data.frame':    4 obs. of  5 variables:
##  $ account.balance               : Factor w/ 3 levels "1","2","3": 1 3 3 2
##  $ credit.purpose                : Factor w/ 3 levels "2","3","4": 3 1 2 1
##  $ previous.credit.payment.status: Factor w/ 2 levels "2","3": 2 2 1 1
##  $ savings                       : Factor w/ 2 levels "2","3": 1 2 1 2
##  $ credit.duration.months        : num  15 12 8 6
```

```
novo_dataset_final <- scale.features(novo_dataset_final, new.numeric.vars)
str(novo_dataset_final)
```

```
## 'data.frame':    4 obs. of  5 variables:
##  $ account.balance               : Factor w/ 3 levels "1","2","3": 1 3 3 2
##  $ credit.purpose                : Factor w/ 3 levels "2","3","4": 3 1 2 1
##  $ previous.credit.payment.status: Factor w/ 2 levels "2","3": 2 2 1 1
##  $ savings                       : Factor w/ 2 levels "2","3": 1 2 1 2
##  $ credit.duration.months        : num [1:4, 1] 1.178 0.434 -0.558 -1.054
##   ..- attr(*, "scaled:center")= num 10.2
##   ..- attr(*, "scaled:scale")= num 4.03
```

```
View(novo_dataset_final)
```

Previsões

```
previsao_novo_cliente <- predict(modelo_final, newdata = novo_dataset_final, type = "respons
e")
```

```
round(previsao_novo_cliente)
```

```
## 1 2 3 4
## 0 1 1 1
```