TFG Codigo Modelos PDF

Diego Brito

Librerias

```
library(MASS)
library(glmnet)
library(rpart)
library(rpart.plot)
library(randomForest)
library(tidyverse)
library(readr)
library(psych)
library(ggplot2)
library(dplyr)
library(corrplot)
library(RColorBrewer)
library(gridExtra)
library(caret)
library(pROC)
library(car)
library(klaR)
```

```
Warning: package 'klaR' was built under R version 4.4.3
```

```
# library(MXM)
# library(parallel)
# library(doParallel)
setwd("C:\\Users\\diego\\OneDrive\\Escritorio\\UCM\\Cuarto\\Segundo Cuatri\\TFG")
```

Modelos

Ya tenemos la base de datos depurada, por tanto vamos a llamarla, y seguimos trabajando aqui

Base de datos depurada

```
load("DatosDepurados.Rda")
datos$WALLSMATERIAL_MODE <- as.factor(datos$WALLSMATERIAL_MODE)
datos$FONDKAPREMONT_MODE <- as.factor(datos$FONDKAPREMONT_MODE)
datos$TARGET <- as.factor(datos$TARGET)
datos$TARGET <- factor(datos$TARGET, levels = c(0,1), labels = c("PAYER", "NOT_PAYER"))
table(datos$TARGET)/nrow(datos)
```

```
     PAYER  NOT_PAYER
0.91334135 0.08665865
```

Train / test

```
set.seed(12345)
trainIndex <- createDataPartition(datos$TARGET, p=0.8, list=FALSE)
data_train <- datos[trainIndex,]
data_test <- datos[-trainIndex,]
data.frame(sort(colSums(is.na(data_train))))
```

```
                          sort.colSums.is.na.data_train...
EXT_SOURCE_3                                             0
EXT_SOURCE_2                                             0
DAYS_BIRTH                                               0
AMT_GOODS_PRICE                                          0
FLAG_OWN_CAR                                             0
EXT_SOURCE_1                                             0
CODE_GENDER                                              0
NAME_EDUCATION_TYPE                                      0
DAYS_EMPLOYED                                            0
REGION_RATING_CLIENT                                     0
AMT_CREDIT                                               0
NAME_INCOME_TYPE                                         0
```

NAME_CONTRACT_TYPE	0
AMT_CREDIT_RANGE	0
REGION_POPULATION_RELATIVE	0
NAME_HOUSING_TYPE	0
FLAG_WORK_PHONE	0
DEF_30_CNT_SOCIAL_CIRCLE	0
REG_CITY_NOT_LIVE_CITY	0
DAYS_REGISTRATION	0
REGION_RATING_CLIENT_W_CITY	0
FLAG_DOCUMENT_3	0
AGE_GROUP	0
EMPLOYMENT_YEAR	0
FLAG_PHONE	0
OWN_CAR_AGE	0
CNT_CHILDREN	0
DAYS_LAST_PHONE_CHANGE	0
FLAG_DOCUMENT_18	0
NAME_TYPE_SUITE	0
FLAG_DOCUMENT_16	0
WEEKDAY_APPR_PROCESS_START	0
REG_CITY_NOT_WORK_CITY	0
AMT_ANNUITY	0
WALLSMATERIAL_MODE	0
AMT_INCOME_TOTAL	0
HOUR_APPR_PROCESS_START	0
AMT_REQ_CREDIT_BUREAU_QRT	0
APARTMENTS_AVG	0
FLOORSMAX_AVG	0
FLAG_DOCUMENT_5	0
FLAG_DOCUMENT_2	0
FONDKAPREMONT_MODE	0
OBS_30_CNT_SOCIAL_CIRCLE	0
YEARS_EMPLOYED	0
TARGET	0

Forward

AIC

```
# null<-glm(TARGET~1,data=data_train,family=binomial)
# full<-glm(TARGET~.,data=data_train,family=binomial)
# AIC
```

```
#modeloForwAIC<-stats::step(null, scope=list(lower=null, upper=full), direction="forward",tra
# una vez sabemos el modelo podemos escribirlo de esta forma
modeloForwAIC <- glm(TARGET~EXT_SOURCE_2+EXT_SOURCE_3+EXT_SOURCE_1+                    NAI
    AMT_CREDIT_RANGE+NAME_HOUSING_TYPE, data = data_train, family = "binomial")
```

BIC

```
#modeloForwBIC<-stats::step(null, scope=list(lower=null, upper=full), direction="forward", k=
# de la misma forma hacemos lo mismo, esto se realiza porque cada vez que vamos a ejecutar ll
modeloForwBIC <- glm(TARGET~EXT_SOURCE_2+EXT_SOURCE_3+EXT_SOURCE_1+NAME_EDUCATION_TYPE+
                       REGION_RATING_CLIENT_W_CITY+
                       DAYS_LAST_PHONE_CHANGE+                      NAME_CONTRACT_TYPE+AMT_I
            FLAG_PHONE+FLOORSMAX_AVG, data = data_train,family = "binomial")
```

Discriminante

```
moddis <- lda(TARGET~., data=data_train)
```

```
Warning in lda.default(x, grouping, ...): variables are collinear
```

Tenemos este problema

```
# Assume your data is in `df` # Step 1: Expand factor variables like LDA would
X_expanded <- model.matrix(~ . - 1, data = data_train)
# Step 2: Identify constant or near-zero variance columns
nzv <- nearZeroVar(X_expanded, saveMetrics = TRUE)
# View constant or near-constant columns
constant_or_nzv <- nzv[nzv$zeroVar | nzv$nzv, ]
print(constant_or_nzv)
```

```
                                    freqRatio percentUnique zeroVar  nzv
EXT_SOURCE_3                         47.74501  0.3979240471   FALSE TRUE
NAME_EDUCATION_TYPEIncomplete higher 24.94902  0.0009923293   FALSE TRUE
NAME_EDUCATION_TYPELower secondary   108.29826  0.0009923293   FALSE TRUE
NAME_INCOME_TYPEMaternity leave    40308.20000  0.0009923293   FALSE TRUE
NAME_INCOME_TYPEPensioner          25192.25000  0.0009923293   FALSE TRUE
NAME_INCOME_TYPEStudent            13435.40000  0.0009923293   FALSE TRUE
NAME_INCOME_TYPEUnemployed         11854.64706  0.0009923293   FALSE TRUE
```

```
AMT_CREDIT_RANGE900K-1M                    32.35750  0.0009923293   FALSE TRUE
NAME_HOUSING_TYPEMunicipal apartment       26.86479  0.0009923293   FALSE TRUE
NAME_HOUSING_TYPEOffice apartment         110.22848  0.0009923293   FALSE TRUE
NAME_HOUSING_TYPERented apartment          52.90372  0.0009923293   FALSE TRUE
EMPLOYMENT_YEAR20-30                       29.07700  0.0009923293   FALSE TRUE
EMPLOYMENT_YEAR30-40                      104.79843  0.0009923293   FALSE TRUE
EMPLOYMENT_YEAR40-50                     1428.40426  0.0009923293   FALSE TRUE
OWN_CAR_AGE                                23.99482  0.0292737142   FALSE TRUE
FLAG_DOCUMENT_181                         101.88208  0.0009923293   FALSE TRUE
NAME_TYPE_SUITEChildren                   111.65847  0.0009923293   FALSE TRUE
NAME_TYPE_SUITEGroup of people           1170.77907  0.0009923293   FALSE TRUE
NAME_TYPE_SUITEOther_A                    329.94581  0.0009923293   FALSE TRUE
NAME_TYPE_SUITEOther_B                    169.36855  0.0009923293   FALSE TRUE
NAME_TYPE_SUITESpouse, partner             25.16461  0.0009923293   FALSE TRUE
FLAG_DOCUMENT_161                          83.04754  0.0009923293   FALSE TRUE
WALLSMATERIAL_MODEBlock                    32.58540  0.0009923293   FALSE TRUE
WALLSMATERIAL_MODEMixed                   129.11362  0.0009923293   FALSE TRUE
WALLSMATERIAL_MODEMonolithic              173.34775  0.0009923293   FALSE TRUE
WALLSMATERIAL_MODEOthers                  183.73511  0.0009923293   FALSE TRUE
WALLSMATERIAL_MODEWooden                   56.20863  0.0009923293   FALSE TRUE
APARTMENTS_AVG                             23.29487  1.0880890715   FALSE TRUE
FLAG_DOCUMENT_51                           60.39080  0.0009923293   FALSE TRUE
FLAG_DOCUMENT_21                        16794.50000  0.0009923293   FALSE TRUE
FONDKAPREMONT_MODEnot specified            51.39043  0.0009923293   FALSE TRUE
FONDKAPREMONT_MODEorg spec account         53.53084  0.0009923293   FALSE TRUE
FONDKAPREMONT_MODEreg oper spec account    24.18695  0.0009923293   FALSE TRUE
```

otros

```
(modelo_rf <- randomForest(TARGET ~ ., data = data_train, ntree = 100))
```

```
Call:
 randomForest(formula = TARGET ~ ., data = data_train, ntree = 100)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 6

        OOB estimate of  error rate: 8.67%
Confusion matrix:
```

```
           PAYER NOT_PAYER  class.error
PAYER     183898       182 0.0009887006
NOT_PAYER  17290       176 0.9899232795
```

```
# Call:
#  randomForest(formula = TARGET ~ ., data = data_train, ntree = 100)
#                Type of random forest: classification
#                      Number of trees: 100
# No. of variables tried at each split: 6
#
#         OOB estimate of  error rate: 8.66%
#Confusion matrix: #            PAYER NOT_PAYER class.error
# PAYER     183884       186 0.001010485
# NOT_PAYER  17266       194 0.988888889
```

```
modelos<-list(modeloForwBIC, modeloForwAIC)
sapply(modelos,function(x) formula(x))
```

```
[[1]]
TARGET ~ EXT_SOURCE_2 + EXT_SOURCE_3 + EXT_SOURCE_1 + NAME_EDUCATION_TYPE +
    CODE_GENDER + FLAG_DOCUMENT_3 + FLAG_OWN_CAR + REG_CITY_NOT_LIVE_CITY +
    DEF_30_CNT_SOCIAL_CIRCLE + DAYS_EMPLOYED + REGION_RATING_CLIENT_W_CITY +
    DAYS_LAST_PHONE_CHANGE + NAME_CONTRACT_TYPE + AMT_REQ_CREDIT_BUREAU_QRT +
    FLAG_DOCUMENT_18 + FLAG_DOCUMENT_16 + OWN_CAR_AGE + AMT_ANNUITY +
    AMT_GOODS_PRICE + AMT_CREDIT + FLAG_WORK_PHONE + DAYS_REGISTRATION +
    NAME_INCOME_TYPE + FLAG_PHONE + FLOORSMAX_AVG

[[2]]
TARGET ~ EXT_SOURCE_2 + EXT_SOURCE_3 + EXT_SOURCE_1 + NAME_EDUCATION_TYPE +
    CODE_GENDER + FLAG_DOCUMENT_3 + FLAG_OWN_CAR + REG_CITY_NOT_LIVE_CITY +
    DEF_30_CNT_SOCIAL_CIRCLE + DAYS_EMPLOYED + REGION_RATING_CLIENT_W_CITY +
    DAYS_LAST_PHONE_CHANGE + NAME_CONTRACT_TYPE + WALLSMATERIAL_MODE +
    AMT_REQ_CREDIT_BUREAU_QRT + FLAG_DOCUMENT_18 + FLAG_DOCUMENT_16 +
    OWN_CAR_AGE + AMT_ANNUITY + AMT_GOODS_PRICE + AMT_CREDIT +
    FLAG_WORK_PHONE + DAYS_REGISTRATION + NAME_INCOME_TYPE +
    FLAG_PHONE + FLOORSMAX_AVG + FONDKAPREMONT_MODE + DAYS_BIRTH +
    NAME_TYPE_SUITE + FLAG_DOCUMENT_2 + FLAG_DOCUMENT_5 + YEARS_EMPLOYED +
    REGION_RATING_CLIENT + REGION_POPULATION_RELATIVE + HOUR_APPR_PROCESS_START +
    WEEKDAY_APPR_PROCESS_START + AMT_CREDIT_RANGE + NAME_HOUSING_TYPE
```

Analisis de modelos

Validacion cruzada

```
data_train$TARGET <- relevel(data_train$TARGET, ref = "NOT_PAYER")
# evaluamos los modelos glm
modelos<-list(moddis, modeloForwBIC, modeloForwAIC)
metodos<-list("lda", "glm", "glm")
titulos<-list("LDA", "Forward BIC", "Forward AIC")
vcrTodosModelos<-list()
predClassCounts <- list()
for (i in 1:length(modelos)){
  set.seed(12345)
  vcr<-train(formula(modelos[[i]]),
             data = data_train,
             method = metodos[[i]],
             family="binomial",
             trControl = trainControl(method="repeatedcv",
                                      number=5,
                                      repeats=5,
                                      summaryFunction=twoClassSummary,
                                      classProbs=TRUE,
                                      savePredictions = TRUE)
             )
  vcrTodosModelos[[ titulos[[i]] ]] <- vcr
  # Predict class probabilities on test data
  probs <- predict(vcr, newdata = data_test, type = "prob")
  # Apply 0.5 threshold to NOT_PAYER probability
  pred_classes <- ifelse(probs$NOT_PAYER >= 0.5, "NOT_PAYER", "PAYER")
  # Count predicted classes
  class_counts <- table(Predicted = pred_classes)
  predClassCounts[[ titulos[[i]] ]] <- class_counts
}
```

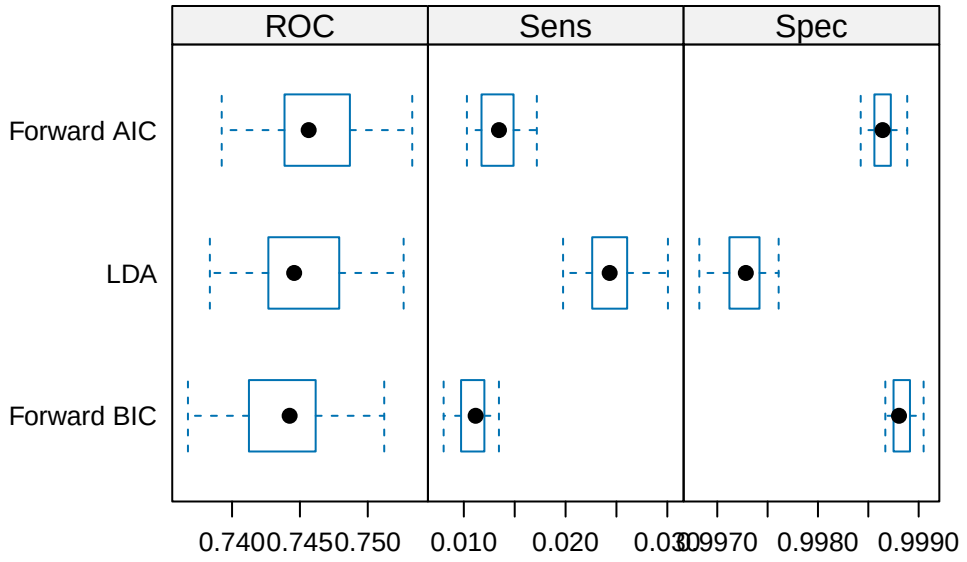
```
Warning in train.default(x, y, weights = w, ...): The metric "Accuracy" was not
in the result set. ROC will be used instead.
```

```
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
```

```
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
Warning in lda.default(x, grouping, ...): variables are collinear
```

```
Warning in train.default(x, y, weights = w, ...): The metric "Accuracy" was not
in the result set. ROC will be used instead.
Warning in train.default(x, y, weights = w, ...): The metric "Accuracy" was not
in the result set. ROC will be used instead.
```

```
bwplot(resamples(vcrTodosModelos), metric=c("ROC", "Sens", "Spec"), scales = list(x = list(re
```




```
summary(resamples(vcrTodosModelos), metric=c("ROC", "Sens", "Spec"))
```

```
Call:
summary.resamples(object = resamples(vcrTodosModelos), metric =
 c("ROC", "Sens", "Spec"))

Models: LDA, Forward BIC, Forward AIC
Number of resamples: 25

ROC
                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
LDA         0.7383652 0.7426743 0.7445765 0.7451961 0.7478972 0.7526424    0
Forward BIC 0.7367485 0.7412434 0.7442415 0.7441938 0.7461598 0.7512106    0
Forward AIC 0.7392366 0.7438641 0.7456450 0.7461731 0.7486817 0.7532757    0

Sens
                  Min.     1st Qu.     Median       Mean    3rd Qu.       Max.
LDA         0.019753793 0.022616662 0.02433438 0.02447037 0.02605210 0.03006012
Forward BIC 0.008016032 0.009730967 0.01116199 0.01090116 0.01202061 0.01345163
Forward AIC 0.010303377 0.011737761 0.01345548 0.01338600 0.01488692 0.01717721
            NA's
```

```
LDA            0
Forward BIC    0
Forward AIC    0
```

```
Spec
                 Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
LDA         0.9968220 0.9971208 0.9972838 0.9972784 0.9974196 0.9976097    0
Forward BIC 0.9986691 0.9987505 0.9988049 0.9988266 0.9989135 0.9990493    0
Forward AIC 0.9984246 0.9985604 0.9986419 0.9986430 0.9987234 0.9988864    0
```

Matrices de confusion

```
# Get predicted probabilities
probs <- predict(vcrTodosModelos[["Forward BIC"]], data_train, type = "prob")
# Use ROC analysis to find best threshold
# roc_obj <- roc(data_train$TARGET, probs)
# best_thresh <- coords(roc_obj, "best", ret = "threshold", best.method = "youden")
#
# print(best_thresh)
```

Train

```
probs <- predict(vcrTodosModelos[["Forward BIC"]], data_train, type = "prob")

cm2<-confusionMatrix(data=as.factor(ifelse(probs$NOT_PAYER>=0.08,"NOT_PAYER","PAYER")),
                     reference=data_train$TARGET, positive="NOT_PAYER")

cm2$table
```

```
          Reference
Prediction  NOT_PAYER  PAYER
  NOT_PAYER     12437  64953
  PAYER          5029 119127
```

```
cm2$overall[1:2]
```

```
 Accuracy     Kappa
0.6527741 0.1407215
```

```
cm2$byClass[1:2]
```

```
Sensitivity Specificity
  0.7120692   0.6471480
```

Test

```
probs <- predict(vcrTodosModelos[["Forward BIC"]], data_test, type = "prob")

cm_test<-confusionMatrix(data=as.factor(ifelse(probs$NOT_PAYER>=0.08,"NOT_PAYER","PAYER")),
                         reference=data_test$TARGET, positive="NOT_PAYER")
```

```
Warning in confusionMatrix.default(data = as.factor(ifelse(probs$NOT_PAYER >= :
Levels are not in the same order for reference and data. Refactoring data to
match.
```

```
cm_test$table
```

```
            Reference
Prediction  PAYER NOT_PAYER
  PAYER     29806      1233
  NOT_PAYER 16213      3133
```

```
cm_test$overall[1:2]
```

```
 Accuracy     Kappa
0.6537462 0.1430914
```

```
cm_test$byClass[1:2]
```

```
Sensitivity Specificity
  0.7175905   0.6476890
```

Curva ROC

Train

```
probs <- predict(modeloForwBIC, data_train, type = "response")
curvaROC<-roc(data_train$TARGET, probs)
```

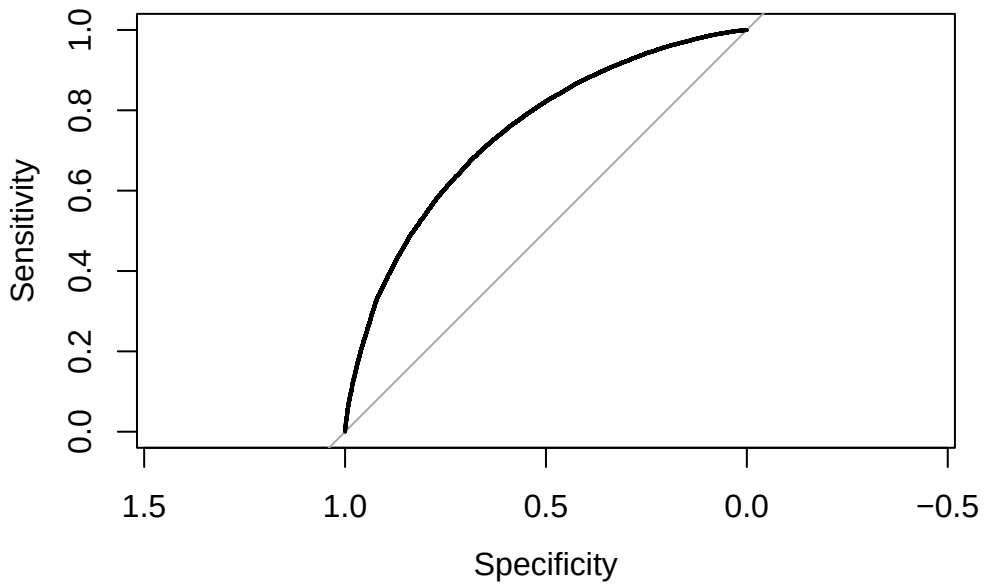
```
Setting levels: control = NOT_PAYER, case = PAYER
```

```
Setting direction: controls > cases
```

```
curvaROC$auc
```

```
Area under the curve: 0.7448
```

```
plot(curvaROC)
```



Test

```
probs_test <- predict(modeloForwBIC, data_test, type = "response")
curvaROC_test<-roc(data_test$TARGET, probs_test)
```

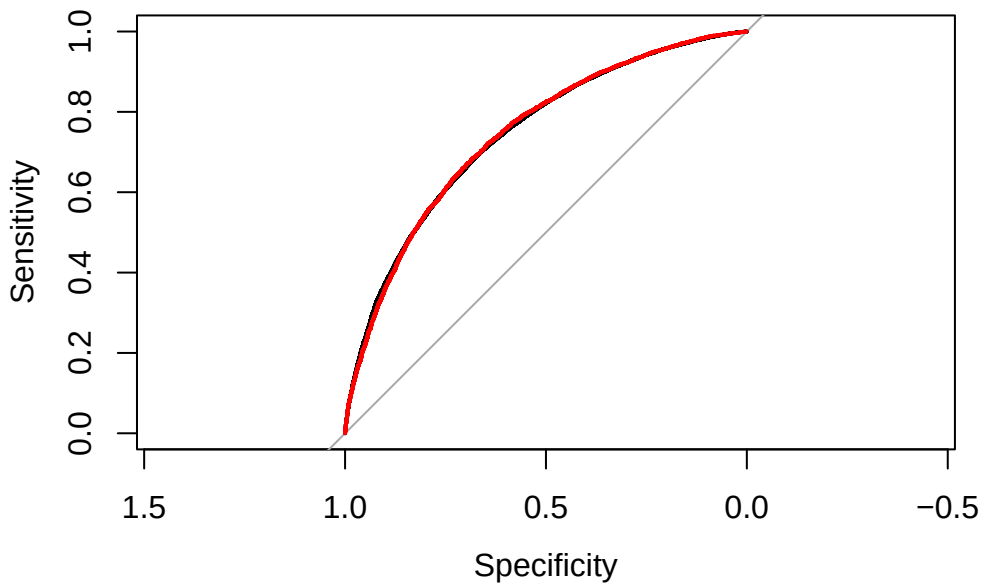
```
Setting levels: control = PAYER, case = NOT_PAYER
```

```
Setting direction: controls < cases
```

```
curvaROC_test$auc
```

```
Area under the curve: 0.7448
```

```
plot(curvaROC)
plot(curvaROC_test, add=T, col="red")
```




```
anova(modeloForwBIC)
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: TARGET

Terms added sequentially (first to last)

                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                      201545     118808
```

```
EXT_SOURCE_2                 1   5398.8    201544     113409 < 2.2e-16 ***
EXT_SOURCE_3                 1   3685.8    201543     109723 < 2.2e-16 ***
EXT_SOURCE_1                 1    789.7    201542     108934 < 2.2e-16 ***
NAME_EDUCATION_TYPE          4    608.3    201538     108325 < 2.2e-16 ***
CODE_GENDER                  1    222.9    201537     108102 < 2.2e-16 ***
FLAG_DOCUMENT_3              1    289.8    201536     107813 < 2.2e-16 ***
FLAG_OWN_CAR                 1    231.1    201535     107582 < 2.2e-16 ***
REG_CITY_NOT_LIVE_CITY       1     74.8    201534     107507 < 2.2e-16 ***
DEF_30_CNT_SOCIAL_CIRCLE     1    104.6    201533     107402 < 2.2e-16 ***
DAYS_EMPLOYED                1    150.8    201532     107251 < 2.2e-16 ***
REGION_RATING_CLIENT_W_CITY  2    100.4    201530     107151 < 2.2e-16 ***
DAYS_LAST_PHONE_CHANGE       1     61.0    201529     107090 5.852e-15 ***
NAME_CONTRACT_TYPE           1     71.1    201528     107019 < 2.2e-16 ***
AMT_REQ_CREDIT_BUREAU_QRT    1     35.8    201527     106983 2.231e-09 ***
FLAG_DOCUMENT_18             1     43.5    201526     106940 4.329e-11 ***
FLAG_DOCUMENT_16             1     37.7    201525     106902 8.286e-10 ***
OWN_CAR_AGE                  1     31.9    201524     106870 1.594e-08 ***
AMT_ANNUITY                  1     23.5    201523     106847 1.224e-06 ***
AMT_GOODS_PRICE              1    131.4    201522     106715 < 2.2e-16 ***
AMT_CREDIT                   1    302.0    201521     106413 < 2.2e-16 ***
FLAG_WORK_PHONE              1     54.8    201520     106358 1.364e-13 ***
DAYS_REGISTRATION            1     17.6    201519     106341 2.775e-05 ***
NAME_INCOME_TYPE             7    328.5    201512     106012 < 2.2e-16 ***
FLAG_PHONE                   1     12.8    201511     106000 0.0003396 ***
FLOORSMAX_AVG                1     50.4    201510     105949 1.268e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Otros modelos AUC, Matrices

Train

```
probs <- predict(vcrTodosModelos[["LDA"]], data_train, type = "prob")

cm2<-confusionMatrix(data=as.factor(ifelse(probs$NOT_PAYER>=0.08,"NOT_PAYER","PAYER")),
                     reference=data_train$TARGET, positive="NOT_PAYER")

cm2$table
```

```
          Reference
Prediction  NOT_PAYER  PAYER
```

```
  NOT_PAYER     12301  62154
  PAYER          5165 121926
```

```
cm2$overall[1:2]
```

```
 Accuracy     Kappa
0.6659869 0.1480381
```

```
cm2$byClass[1:2]
```

```
Sensitivity Specificity
  0.7042826   0.6623533
```

Test

```
probs <- predict(vcrTodosModelos[["LDA"]], data_test, type = "prob")

cm_test<-confusionMatrix(data=as.factor(ifelse(probs$NOT_PAYER>=0.08,"NOT_PAYER","PAYER")),
                      reference=data_test$TARGET, positive="NOT_PAYER")
```

```
Warning in confusionMatrix.default(data = as.factor(ifelse(probs$NOT_PAYER >= :
Levels are not in the same order for reference and data. Refactoring data to
match.
```

```
cm_test$table
```

```
           Reference
Prediction  PAYER NOT_PAYER
  PAYER     30529      1318
  NOT_PAYER 15490      3048
```

```
cm_test$overall[1:2]
```

```
 Accuracy     Kappa
0.6664087 0.1464234
```

```
cm_test$byClass[1:2]
```

```
Sensitivity Specificity
  0.6981219   0.6633999
```

Curva ROC

Train

```
probs <- predict(vcrTodosModelos[["LDA"]], data_train, type = "prob")
str(probs)
```

```
'data.frame':   201546 obs. of  2 variables:
 $ NOT_PAYER: num  0.6281 0.031 0.0268 0.0427 0.1045 ...
 $ PAYER    : num  0.372 0.969 0.973 0.957 0.895 ...
```

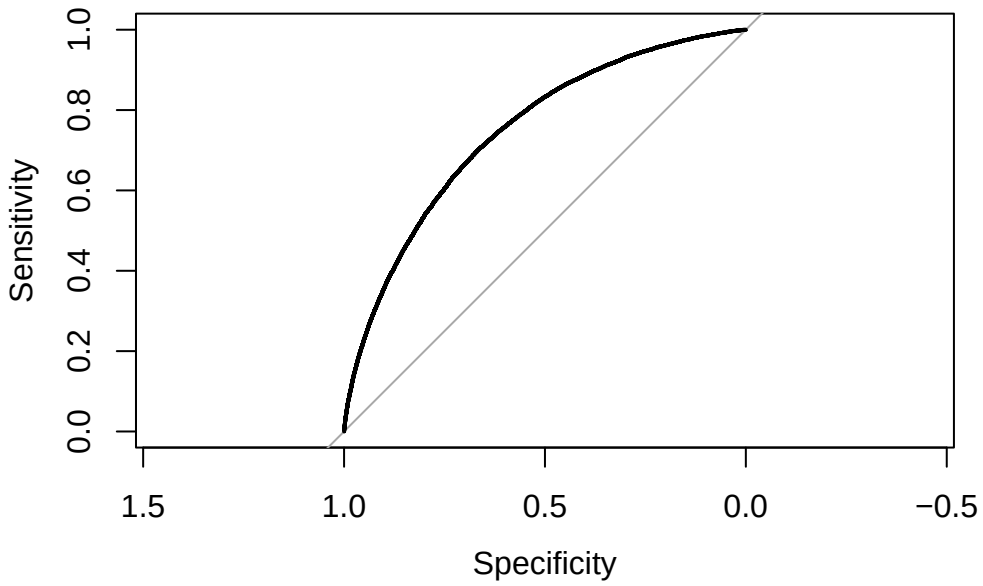
```
probs_pos <- probs[, "NOT_PAYER"]
curvaROC <- roc(data_train$TARGET, probs_pos, levels = c("PAYER", "NOT_PAYER"))
```

```
Setting direction: controls < cases
```

```
curvaROC$auc
```

```
Area under the curve: 0.7466
```

```
plot(curvaROC)
```



Test

```
probs <- predict(vcrTodosModelos[["LDA"]], data_test, type = "prob")
str(probs)
```

```
'data.frame':   50385 obs. of  2 variables:
 $ NOT_PAYER: num  0.1033 0.0292 0.1087 0.0326 0.0216 ...
 $ PAYER    : num  0.897 0.971 0.891 0.967 0.978 ...
```

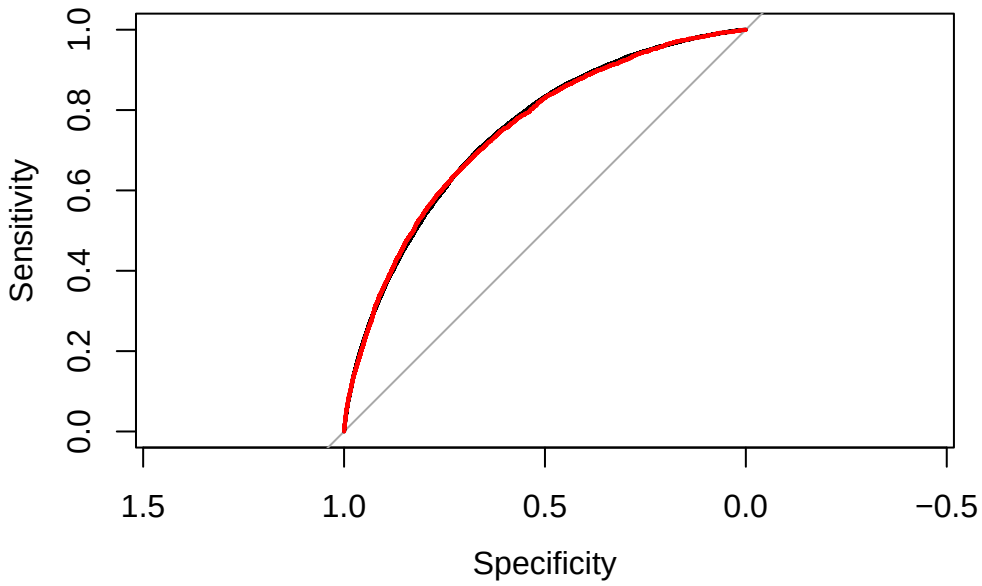
```
probs_pos <- probs[, "NOT_PAYER"]
curvaROC_test <- roc(data_test$TARGET, probs_pos, levels = c("PAYER", "NOT_PAYER"))
```

```
Setting direction: controls < cases
```

```
curvaROC$auc
```

```
Area under the curve: 0.7466
```

```
plot(curvaROC)
plot(curvaROC_test, add=T, col="red")
```




```
coeffs <- moddis$scaling
importance <- abs(coeffs[,1])
sorted_importance <- sort(importance, decreasing = TRUE)
print(sorted_importance)
```

```
                NAME_INCOME_TYPEUnemployed
                              3.020513e+01
           NAME_INCOME_TYPEMaternity leave
                              5.447024e+00
                        FLAG_DOCUMENT_21
                              4.345755e+00
                              EXT_SOURCE_3
                              2.912404e+00
                              EXT_SOURCE_2
                              2.688135e+00
           REGION_POPULATION_RELATIVE
                              2.085941e+00
                            AGE_GROUP30-40
                              1.475423e+00
                            AGE_GROUP40-50
                              1.395338e+00
                            AGE_GROUP20-30
```

```
                                1.386304e+00
            NAME_EDUCATION_TYPELower secondary
                                1.336682e+00
                          AGE_GROUP50 above
                                1.264941e+00
                   NAME_INCOME_TYPEStudent
                                1.182644e+00
                                EXT_SOURCE_1
                                1.149213e+00
NAME_EDUCATION_TYPESecondary / secondary special
                                1.004006e+00
                   NAME_INCOME_TYPEPensioner
                                7.826655e-01
          NAME_EDUCATION_TYPEIncomplete higher
                                7.267212e-01
           NAME_EDUCATION_TYPEHigher education
                                7.067526e-01
                 REGION_RATING_CLIENT_W_CITY3
                                6.857041e-01
                    AMT_CREDIT_RANGE400K-500K
                                5.614867e-01
                    AMT_CREDIT_RANGE500K-600K
                                5.559021e-01
                           FLAG_DOCUMENT_181
                                5.218544e-01
               NAME_TYPE_SUITEGroup of people
                                5.021683e-01
                    AMT_CREDIT_RANGE300K-400K
                                4.808357e-01
                           FLAG_DOCUMENT_161
                                4.489856e-01
                    AMT_CREDIT_RANGE600K-700K
                                4.410201e-01
                       REGION_RATING_CLIENT3
                                4.395788e-01
                        EMPLOYMENT_YEAR30-40
                                4.043022e-01
                 REGION_RATING_CLIENT_W_CITY2
                                3.954369e-01
                    AMT_CREDIT_RANGE200K-300K
                                3.828897e-01
                                CODE_GENDERM
                                3.817189e-01
```

```
            REGION_RATING_CLIENT2
                         3.375708e-01
                        FLAG_OWN_CARY
                         3.127267e-01
                EMPLOYMENT_YEAR40-50
                         3.001720e-01
                       FLOORSMAX_AVG
                         2.771353e-01
NAME_HOUSING_TYPEOffice apartment
                         2.618962e-01
                          AMT_CREDIT
                         2.600076e-01
               EMPLOYMENT_YEAR20-30
                         2.583529e-01
             NAME_TYPE_SUITEOther_B
                         2.474244e-01
          AMT_CREDIT_RANGE100K-200K
                         2.382407e-01
          AMT_CREDIT_RANGE700K-800K
                         2.252928e-01
                  FLAG_DOCUMENT_31
                         2.231567e-01
                  FLAG_DOCUMENT_51
                         2.210624e-01
           DEF_30_CNT_SOCIAL_CIRCLE
                         2.194826e-01
            NAME_INCOME_TYPEWorking
                         2.097342e-01
            REG_CITY_NOT_LIVE_CITY1
                         2.030575e-01
NAME_HOUSING_TYPERented apartment
                         1.859280e-01
          AMT_CREDIT_RANGE800K-900K
                         1.796256e-01
                  FLAG_WORK_PHONE1
                         1.743235e-01
       NAME_TYPE_SUITEUnaccompanied
                         1.544677e-01
      NAME_HOUSING_TYPEWith parents
                         1.490864e-01
       WALLSMATERIAL_MODEMonolithic
                         1.453681e-01
           AMT_CREDIT_RANGE1M Above
```

```
                          1.411061e-01
              WALLSMATERIAL_MODEOthers
                          1.337589e-01
   NAME_HOUSING_TYPEMunicipal apartment
                          1.179861e-01
               AMT_CREDIT_RANGE900K-1M
                          1.151345e-01
FONDKAPREMONT_MODEreg oper spec account
                          1.122304e-01
   NAME_INCOME_TYPECommercial associate
                          1.036708e-01
                NAME_TYPE_SUITEFamily
                          1.021477e-01
              WALLSMATERIAL_MODEPanel
                          9.489312e-02
     FONDKAPREMONT_MODEorg spec account
                          8.719861e-02
              WALLSMATERIAL_MODEBlock
                          7.668154e-02
       WEEKDAY_APPR_PROCESS_STARTMONDAY
                          7.262149e-02
                 EMPLOYMENT_YEAR5-10
                          7.132580e-02
                       YEARS_EMPLOYED
                          6.839381e-02
              WALLSMATERIAL_MODEWooden
                          6.622466e-02
       WEEKDAY_APPR_PROCESS_STARTSUNDAY
                          6.540212e-02
              NAME_TYPE_SUITEChildren
                          6.478319e-02
                           FLAG_PHONE1
                          6.299659e-02
     NAME_HOUSING_TYPEHouse / apartment
                          5.600838e-02
        FONDKAPREMONT_MODEnot specified
                          5.117367e-02
      WEEKDAY_APPR_PROCESS_STARTTUESDAY
                          3.937959e-02
         NAME_TYPE_SUITESpouse, partner
                          3.732481e-02
     WEEKDAY_APPR_PROCESS_STARTSATURDAY
                          3.692898e-02
```

```
NAME_CONTRACT_TYPERevolving loans
                     3.249683e-02
        AMT_REQ_CREDIT_BUREAU_QRT
                     3.243485e-02
          REG_CITY_NOT_WORK_CITY1
                     3.100387e-02
                    APARTMENTS_AVG
                     3.010494e-02
                     CNT_CHILDREN
                     2.998628e-02
   WALLSMATERIAL_MODEStone, brick
                     2.653276e-02
             EMPLOYMENT_YEAR10-20
                     2.310469e-02
FONDKAPREMONT_MODEreg oper account
                     2.120956e-02
WEEKDAY_APPR_PROCESS_STARTWEDNESDAY
                     1.706198e-02
 WEEKDAY_APPR_PROCESS_STARTTHURSDAY
                     1.406181e-02
                  AMT_INCOME_TOTAL
                     1.286044e-02
           NAME_TYPE_SUITEOther_A
                     1.168900e-02
          WALLSMATERIAL_MODEMixed
                     1.080305e-02
     NAME_INCOME_TYPEState servant
                     7.643848e-03
                      OWN_CAR_AGE
                     4.730482e-03
         OBS_30_CNT_SOCIAL_CIRCLE
                     4.414556e-03
          HOUR_APPR_PROCESS_START
                     3.657671e-03
                    DAYS_EMPLOYED
                     2.611506e-04
                EMPLOYMENT_YEAR0-5
                     1.358216e-04
           DAYS_LAST_PHONE_CHANGE
                     7.116356e-05
                       DAYS_BIRTH
                     1.639156e-05
                      AMT_ANNUITY
```

```
     9.853604e-06
DAYS_REGISTRATION
     9.701877e-06
  AMT_GOODS_PRICE
     2.866606e-06
```