Regressão Logística

# Carregar Dados

TEBATRANSF <- read\_excel("TEBATRANSF.xlsx")  
tt=TEBATRANSF[,-1]

# Separando X e Y

tt$cancelsim=ifelse(tt$cancel=="sim",1,0) # Variavel a ser prevista  
tt=tt[,-10] #Variaveveis previsoras

# Separar texte e treino

set.seed(123)  
index=sample(1:2000, 1200 )  
lrn= tt[index,] #arquivo para desenvolvimento   
tst=tt[-index,] #arquivo para teste

# Criar modelo

mod1=glm(data = lrn, cancelsim~.,family = binomial() )  
summary(mod1)

##   
## Call:  
## glm(formula = cancelsim ~ ., family = binomial(), data = lrn)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.37301 -0.58238 -0.30429 -0.08389 2.70625   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 7.31615 3.21496 2.276 0.02287 \*   
## idade -0.04307 0.01338 -3.220 0.00128 \*\*   
## klinhas 0.10801 0.12457 0.867 0.38589   
## Ltempcli -3.28898 0.44001 -7.475 7.73e-14 \*\*\*  
## Lrenda -0.41845 1.07523 -0.389 0.69715   
## Sfatura 0.11789 0.01112 10.597 < 2e-16 \*\*\*  
## temp\_rsd 0.02289 0.04362 0.525 0.59978   
## localB 2.09913 0.24530 8.557 < 2e-16 \*\*\*  
## localC -0.48927 0.27177 -1.800 0.07181 .   
## localD 1.32211 0.22050 5.996 2.02e-09 \*\*\*  
## tvcabosim 0.27492 0.18989 1.448 0.14769   
## debautsim -0.28410 0.17701 -1.605 0.10850   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1303.85 on 1199 degrees of freedom  
## Residual deviance: 870.38 on 1188 degrees of freedom  
## AIC: 894.38  
##   
## Number of Fisher Scoring iterations: 6

# Selecao de variaveis com base no criterio AIC

mod2=step(mod1)

## Start: AIC=894.38  
## cancelsim ~ idade + klinhas + Ltempcli + Lrenda + Sfatura + temp\_rsd +   
## local + tvcabo + debaut  
##   
## Df Deviance AIC  
## - Lrenda 1 870.53 892.53  
## - temp\_rsd 1 870.66 892.66  
## - klinhas 1 871.13 893.13  
## <none> 870.38 894.38  
## - tvcabo 1 872.51 894.51  
## - debaut 1 872.99 894.99  
## - idade 1 881.11 903.11  
## - Ltempcli 1 931.92 953.92  
## - local 3 1002.64 1020.64  
## - Sfatura 1 1009.99 1031.99  
##   
## Step: AIC=892.53  
## cancelsim ~ idade + klinhas + Ltempcli + Sfatura + temp\_rsd +   
## local + tvcabo + debaut  
##   
## Df Deviance AIC  
## - temp\_rsd 1 870.80 890.80  
## - klinhas 1 871.13 891.13  
## <none> 870.53 892.53  
## - tvcabo 1 872.65 892.65  
## - debaut 1 873.14 893.14  
## - idade 1 883.08 903.08  
## - Ltempcli 1 971.16 991.16  
## - local 3 1003.85 1019.85  
## - Sfatura 1 1023.34 1043.34  
##   
## Step: AIC=890.8  
## cancelsim ~ idade + klinhas + Ltempcli + Sfatura + local + tvcabo +   
## debaut  
##   
## Df Deviance AIC  
## - klinhas 1 871.38 889.38  
## <none> 870.80 890.80  
## - tvcabo 1 872.85 890.85  
## - debaut 1 873.49 891.49  
## - idade 1 883.22 901.22  
## - Ltempcli 1 971.73 989.73  
## - local 3 1004.58 1018.58  
## - Sfatura 1 1024.25 1042.25  
##   
## Step: AIC=889.38  
## cancelsim ~ idade + Ltempcli + Sfatura + local + tvcabo + debaut  
##   
## Df Deviance AIC  
## - tvcabo 1 873.34 889.34  
## <none> 871.38 889.38  
## - debaut 1 874.06 890.06  
## - idade 1 883.35 899.35  
## - Ltempcli 1 985.34 1001.34  
## - local 3 1005.06 1017.06  
## - Sfatura 1 1032.29 1048.29  
##   
## Step: AIC=889.34  
## cancelsim ~ idade + Ltempcli + Sfatura + local + debaut  
##   
## Df Deviance AIC  
## <none> 873.34 889.34  
## - debaut 1 876.02 890.02  
## - idade 1 885.64 899.64  
## - Ltempcli 1 986.74 1000.74  
## - local 3 1009.26 1019.26  
## - Sfatura 1 1034.39 1048.39

summary(mod2)

##   
## Call:  
## glm(formula = cancelsim ~ idade + Ltempcli + Sfatura + local +   
## debaut, family = binomial(), data = lrn)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.48985 -0.57591 -0.30731 -0.08443 2.73478   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 6.65609 0.86555 7.690 1.47e-14 \*\*\*  
## idade -0.04364 0.01267 -3.444 0.000574 \*\*\*  
## Ltempcli -3.43924 0.35218 -9.766 < 2e-16 \*\*\*  
## Sfatura 0.11794 0.01049 11.239 < 2e-16 \*\*\*  
## localB 2.12125 0.24439 8.680 < 2e-16 \*\*\*  
## localC -0.48770 0.27098 -1.800 0.071894 .   
## localD 1.32915 0.21920 6.064 1.33e-09 \*\*\*  
## debautsim -0.28769 0.17656 -1.629 0.103214   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1303.85 on 1199 degrees of freedom  
## Residual deviance: 873.34 on 1192 degrees of freedom  
## AIC: 889.34  
##   
## Number of Fisher Scoring iterations: 6

# Criar previsoes

tst$ps=predict(mod2, newdata = tst, type = "response")  
print(head(tst), digits=3)

## # A tibble: 6 x 11  
## idade klinhas Ltempcli Lrenda Sfatura temp\_rsd local tvcabo debaut  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <chr> <chr>   
## 1 35 1 2.71 3.69 24.4 4.8 A nao nao   
## 2 27 1 2.89 3.75 35.7 4.8 D nao sim   
## 3 30 1 3.14 3.76 29.3 8.1 B nao nao   
## 4 39 2 3.04 3.84 24.4 2.8 A nao nao   
## 5 45 3 2.89 3.85 16.9 8 A sim sim   
## 6 33 1 3.18 3.81 14.4 4.3 C sim sim   
## # … with 2 more variables: cancelsim <dbl>, ps <dbl>

# Fazer testes

## Hosmer Lemeshow

library(ResourceSelection)

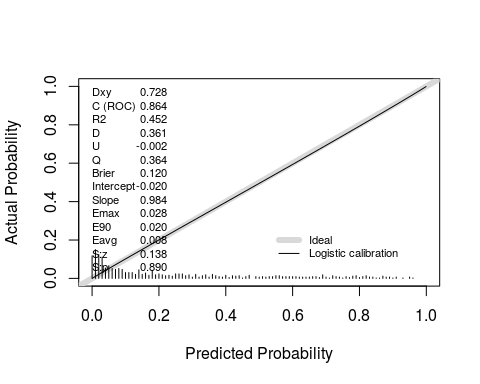
## ResourceSelection 0.3-5 2019-07-22

hoslem.test(mod2$y, fitted(mod2), g=10)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: mod2$y, fitted(mod2)  
## X-squared = 3.1873, df = 8, p-value = 0.9221

## Spiegelhalter

library (rms)  
val.prob (tst$ps, tst$cancelsim, smooth = F)



## Dxy C (ROC) R2 D D:Chi-sq   
## 0.728481114 0.864240557 0.451974209 0.361106357 289.885085879   
## D:p U U:Chi-sq U:p Q   
## NA -0.002436801 0.050559016 0.975037343 0.363543159   
## Brier Intercept Slope Emax E90   
## 0.119721123 -0.020023035 0.984019371 0.027986824 0.019574397   
## Eavg S:z S:p   
## 0.008116074 0.138458673 0.889877929

## Zoyowsky

library(arules)

## Loading required package: Matrix

##   
## Attaching package: 'arules'

## The following object is masked from 'package:car':  
##   
## recode

## The following objects are masked from 'package:base':  
##   
## abbreviate, write

kp=discretize(tst$ps, method = 'frequency', breaks = 5)  
m=table(kp, tst$cancelsim)  
mp=prop.table(m,1)  
print(mp,digits=2)

##   
## kp 0 1  
## [0.000386,0.0292) 0.9938 0.0063  
## [0.0292,0.0852) 0.9187 0.0813  
## [0.0852,0.212) 0.8688 0.1313  
## [0.212,0.477) 0.6813 0.3187  
## [0.477,0.965] 0.3063 0.6937

# Matriz de classificação

PC=0.50 # Ponto de corte para classificacao  
klas=ifelse(tst$ps>=PC,"prev\_sim","prev\_nao")  
table(tst$cancelsim,klas)

## klas  
## prev\_nao prev\_sim  
## 0 557 46  
## 1 88 109

# Criar previsao

novo.indiv=data.frame(idade=51,Ltempcli=3,Sfatura=25,local="A", debaut = "nao")  
novo.p=predict(mod2, novo.indiv, type="response")  
novo.p

## 1   
## 0.05026981

# Criar previsoes com base em faixas de frequência

library(arules)  
xx=c("0.00 a 0.50","0.50 a 0.75", "0.75 a 1.00")  
kps=discretize(tst$ps, method = "fixed", breaks = c(0,.50,.75,1),labels = xx)  
class=table(kps,tst$cancelsim)  
class

##   
## kps 0 1  
## 0.00 a 0.50 557 88  
## 0.50 a 0.75 39 58  
## 0.75 a 1.00 7 51

print(prop.table(class,1), digits=3)

##   
## kps 0 1  
## 0.00 a 0.50 0.864 0.136  
## 0.50 a 0.75 0.402 0.598  
## 0.75 a 1.00 0.121 0.879

# Calcular indicadores

library(hmeasure)  
HMeasure(tst$cancelsim, tst$ps)$metric

## H Gini AUC AUCH KS MER MWL  
## scores 0.4470573 0.7284811 0.8642406 0.8728186 0.5683006 0.16 0.1602563  
## Spec.Sens95 Sens.Spec95 ER Sens Spec Precision  
## scores 0.4626866 0.5025381 0.1675 0.5532995 0.9237148 0.7032258  
## Recall TPR FPR F Youden TP FP TN FN  
## scores 0.5532995 0.5532995 0.07628524 0.6193182 0.4770143 109 46 557 88

# Cross validation

library(boot)

##   
## Attaching package: 'boot'

## The following object is masked from 'package:survival':  
##   
## aml

## The following object is masked from 'package:lattice':  
##   
## melanoma

## The following object is masked from 'package:car':  
##   
## logit

mod3=glm(data = tt, cancelsim~.,family = binomial())  
mod4=step(mod3)

## Start: AIC=1491.63  
## cancelsim ~ idade + klinhas + Ltempcli + Lrenda + Sfatura + temp\_rsd +   
## local + tvcabo + debaut  
##   
## Df Deviance AIC  
## - Lrenda 1 1467.8 1489.8  
## - debaut 1 1467.8 1489.8  
## - klinhas 1 1468.6 1490.6  
## - tvcabo 1 1468.8 1490.8  
## - temp\_rsd 1 1469.2 1491.2  
## <none> 1467.6 1491.6  
## - idade 1 1485.8 1507.8  
## - Ltempcli 1 1579.2 1601.2  
## - local 3 1674.6 1692.6  
## - Sfatura 1 1726.1 1748.1  
##   
## Step: AIC=1489.77  
## cancelsim ~ idade + klinhas + Ltempcli + Sfatura + temp\_rsd +   
## local + tvcabo + debaut  
##   
## Df Deviance AIC  
## - debaut 1 1468.0 1488.0  
## - klinhas 1 1468.7 1488.7  
## - tvcabo 1 1469.0 1489.0  
## - temp\_rsd 1 1469.4 1489.4  
## <none> 1467.8 1489.8  
## - idade 1 1488.9 1508.9  
## - Ltempcli 1 1640.7 1660.7  
## - local 3 1675.8 1691.8  
## - Sfatura 1 1753.1 1773.1  
##   
## Step: AIC=1487.97  
## cancelsim ~ idade + klinhas + Ltempcli + Sfatura + temp\_rsd +   
## local + tvcabo  
##   
## Df Deviance AIC  
## - klinhas 1 1468.9 1486.9  
## - tvcabo 1 1469.2 1487.2  
## - temp\_rsd 1 1469.6 1487.6  
## <none> 1468.0 1488.0  
## - idade 1 1489.2 1507.2  
## - Ltempcli 1 1640.8 1658.8  
## - local 3 1675.8 1689.8  
## - Sfatura 1 1753.2 1771.2  
##   
## Step: AIC=1486.86  
## cancelsim ~ idade + Ltempcli + Sfatura + temp\_rsd + local + tvcabo  
##   
## Df Deviance AIC  
## - tvcabo 1 1470.0 1486.0  
## - temp\_rsd 1 1470.5 1486.5  
## <none> 1468.9 1486.9  
## - idade 1 1489.3 1505.3  
## - Ltempcli 1 1661.8 1677.8  
## - local 3 1676.8 1688.8  
## - Sfatura 1 1768.5 1784.5  
##   
## Step: AIC=1486.02  
## cancelsim ~ idade + Ltempcli + Sfatura + temp\_rsd + local  
##   
## Df Deviance AIC  
## - temp\_rsd 1 1471.7 1485.7  
## <none> 1470.0 1486.0  
## - idade 1 1490.9 1504.9  
## - Ltempcli 1 1662.3 1676.3  
## - local 3 1678.6 1688.6  
## - Sfatura 1 1769.7 1783.7  
##   
## Step: AIC=1485.65  
## cancelsim ~ idade + Ltempcli + Sfatura + local  
##   
## Df Deviance AIC  
## <none> 1471.7 1485.7  
## - idade 1 1492.2 1504.2  
## - Ltempcli 1 1663.7 1675.7  
## - local 3 1680.0 1688.0  
## - Sfatura 1 1770.7 1782.7

#mod4  
set.seed(11)  
cvglm=cv.glm(data = tt, glmfit = mod4, K = 10)  
cvglm$delta[1]

## [1] 0.1178127

cv.glm(data = tt, glmfit = mod4)$delta[1]

## [1] 0.1175945