Lab04 - Previsão de Churn com Regressão Logística

Machine Learning usando o R

Thalles Quinaglia Liduares

2022-08-10

Exercicio proposto

Na Seção 02 do nosso Curso de Machine Learning usando o `R` exploramos o data set `Telco-Customer-Churn.csv` que fazia referência a dados de uma operadora de telecomunicações. Dentro do data set, havia uma variável qualitativa que media `churn` . Com base nesse data set, pede-se que você construa um modelo de regressão logística que explique a probabilidade de churn dentro desse conjunto de dados.

Upload dados

```
setwd("C:\\Program Files\\R\\Dados")

data<-read.csv(file="Telco-Customer-Churn.csv", stringsAsFactors = TRUE)

attach(data)
```

Upload pacotes

```
library(DescTools)
library(gtsummary)
library(caTools)
library(caret)
library(car)
library(plyr)
```

Analise exploratória dos dados

Como demonstrado pelos valores descritos abaixo, a taxa de cancelamento dos serviços de telefonia é de aproximadamente 27% para mulheres e 26.1% para homens.

```
table(Churn, gender)
```

```
##      gender
## Churn Female Male
##   No    2549 2625
##   Yes    939  930
```

Análise de existência de multicolinearidade

```
data1<-data[,-1] # Exclusão do ID dos clientes

reg1<-lm(as.numeric(Churn)~., data1)

#car::vif(reg1)
```

A mensagem de erro acima é um indicativo de existência de multicolinearidade. Portanto, existe correlação perfeita entre algumas variaveis da base de dados.

Serão inclusas no modelo as seguintes variáveis, em seguida será analisada o VIF para o teste de multicolinearidade.

```
reg2<-lm(as.numeric(Churn)~MonthlyCharges+tenure+gender+PhoneService+Dependents
         +Contract)

car::vif(reg2)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## MonthlyCharges 1.309309  1        1.144251
## tenure         2.224921  1        1.491617
## gender         1.000534  1        1.000267
## PhoneService   1.077744  1        1.038144
## Dependents     1.076335  1        1.037465
## Contract       2.134200  2        1.208673
```

Os respectivos valores de VIF variam de 1.00 a 2.22, logo, há evidências de ausência de multicolinearidade.

Partição da amostra entre 80% treino e 20% teste

```
set.seed(111)

part_data<-floor(0.80*nrow(data))

treino_data <-sample(seq_len(nrow(data)), size = part_data)

treino<-data[treino_data, ]

teste<-data[-treino_data,]
```

Regressão logistica

```
reg_log<-glm(Churn~gender+Dependents+MonthlyCharges+tenure+Contract+PhoneService,
           family = binomial(link="logit"),
           data=treino)

summary(reg_log)
```

```
##
## Call:
## glm(formula = Churn ~ gender + Dependents + MonthlyCharges +
##     tenure + Contract + PhoneService, family = binomial(link = "logit"),
##     data = treino)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8794  -0.6977  -0.3047   0.7899   3.2992
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.954215   0.132089  -7.224 5.05e-13 ***
## genderMale       -0.005772   0.070752  -0.082   0.9350
## DependentsYes    -0.257498   0.087647  -2.938   0.0033 **
## MonthlyCharges    0.033517   0.001706  19.652  < 2e-16 ***
## tenure           -0.039369   0.002325 -16.934  < 2e-16 ***
## ContractOne year -0.954616   0.112866  -8.458  < 2e-16 ***
## ContractTwo year -1.834464   0.183766  -9.983  < 2e-16 ***
## PhoneServiceYes  -0.961492   0.130236  -7.383 1.55e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6529.6  on 5633  degrees of freedom
## Residual deviance: 4846.3  on 5626  degrees of freedom
## AIC: 4862.3
##
## Number of Fisher Scoring iterations: 6
```

O gênero dos individuos não possui significância estatística para explicar a ocorrência de *Churn*, ao passo que as demais variáveis explicativas consideradas na regressão apresentam significância estatística.

Analise preditiva do modelo

```
pred_reg_log<-predict(reg_log, newdata=teste, type = 'response')
```

Ajuste do modelo

```
PseudoR2<-DescTools::PseudoR2(reg_log, which="Nagelkerke")

round(PseudoR2,3)*100
```

```
## Nagelkerke
##       37.6
```

O Pseudo R^2 é igual a 37.6%.

Teste Anova

```
car::Anova(reg_log,type="II", test="Wald")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Churn
##                Df    Chisq Pr(>Chisq)
## gender          1   0.0067   0.934984
## Dependents      1   8.6313   0.003304 **
## MonthlyCharges  1 386.2176  < 2.2e-16 ***
## tenure          1 286.7583  < 2.2e-16 ***
## Contract        2 138.7858  < 2.2e-16 ***
## PhoneService    1  54.5039  1.551e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Os coeficientes no modelo de regressão logistica não são diretamente interpretáveis. Logo, obtém-se as razões de chances com IC 95%.

```
exp(coef(reg_log))
```

```
##      (Intercept)       genderMale    DependentsYes   MonthlyCharges
##        0.3851142        0.9942450        0.7729830        1.0340855
##           tenure ContractOne year ContractTwo year  PhoneServiceYes
##        0.9613963        0.3849601        0.1596991        0.3823222
```

```
exp(confint(reg_log))
```

```
## Waiting for profiling to be done...
```

```
##                       2.5 %    97.5 %
## (Intercept)      0.2967707 0.4982028
## genderMale       0.8655052 1.1421864
## DependentsYes    0.6504219 0.9171872
## MonthlyCharges   1.0306728 1.0375883
## tenure           0.9569923 0.9657553
## ContractOne year 0.3077586 0.4791507
## ContractTwo year 0.1100075 0.2264630
## PhoneServiceYes  0.2964027 0.4939956
```

Plot da tabela de resultados

```
gtsummary::tbl_regression(reg_log, exponentiate=TRUE)
```

Characteristic	**OR**[1]	**95% CI**[1]	**p-value**
gender			
Female	—	—	

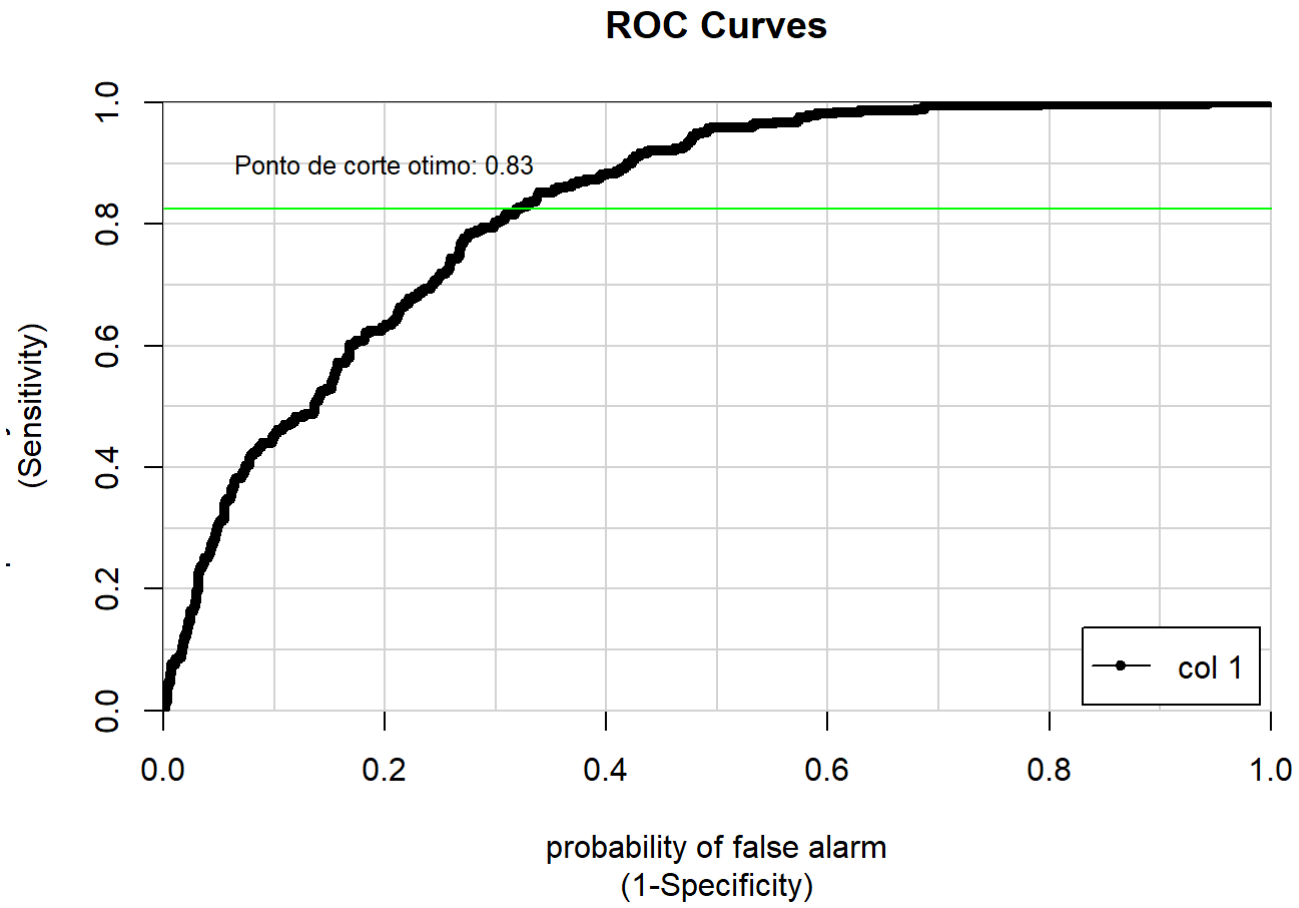
[1] OR = Odds Ratio, CI = Confidence Interval

Characteristic	OR[1]	95% CI[1]	p-value
Male	0.99	0.87, 1.14	>0.9
Dependents			
No	—	—	
Yes	0.77	0.65, 0.92	0.003
MonthlyCharges	1.03	1.03, 1.04	<0.001
tenure	0.96	0.96, 0.97	<0.001
Contract			
Month-to-month	—	—	
One year	0.38	0.31, 0.48	<0.001
Two year	0.16	0.11, 0.23	<0.001
PhoneService			
No	—	—	
Yes	0.38	0.30, 0.49	<0.001

[1] OR = Odds Ratio, CI = Confidence Interval

Curva ROC

```
reg_AUC<-colAUC(pred_reg_log, teste$Churn, plotROC = TRUE)
abline(h=reg_AUC, col='green')
text(.2,.9, cex=.8, labels=paste("Ponto de corte otimo:", round(reg_AUC,2)))
```



Com base no ponto de corte ótimo estimado irei plotar a *Confusion Matrix* para analise de acurácia do modelo.

```
Churn_prob<-ifelse(pred_reg_log>0.82,1,0)

Churn_class<-factor(Churn_prob)


teste$Churn<-as.factor(mapvalues(teste$Churn, from=c("No","Yes"), to=c(0,1)))

confusionMatrix(Churn_class, teste$Churn)
```

```
## Warning in confusionMatrix.default(Churn_class, teste$Churn): Levels are not in
## the same order for reference and data. Refactoring data to match.
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1040  369
##          1    0    0
##
##                Accuracy : 0.7381
##                  95% CI : (0.7143, 0.7609)
##     No Information Rate : 0.7381
##     P-Value [Acc > NIR] : 0.514
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.7381
##          Neg Pred Value :    NaN
##              Prevalence : 0.7381
##          Detection Rate : 0.7381
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##
```

O modelo de regressão logistica proposto nesta análise aponta para ocorrência de *Churn* em 369 casos dentre um total de 1409 com nível de acurácia igual a 73.8%.