Exploração de dados - Banco Czech

Bruno Santos Wance de Souza

Lucas de Jesus Matias
Luiz Cesar Costa Raymundo

21 de novembro de 2018

Contents

Pagamento de Empréstimo

Leitura dos dados

Os dados do csv gerado a partir da planilha foram carregados para a variável "pagamentoEprestimo".

```
pagamentoEmprestimo <-
    read.csv2("./dados/pagamento_emprestimo.csv", stringsAsFactors = FALSE)
```

Criação do modelo

A funcionalidade glm foi utilizada para geração do modelo de regressão e este vinculado à variável glmPagamento.

```
glm(data = pagamentoEmprestimo,
    formula = pagamento ~ estadocivil + idade + sexo, family = binomial) ->
  glmPagamento
```

Análise das variáveis

Os valores Ps das variáveis reijeitam a hipótese inicial de que são irrelevantes para o modelo, portanto foram consideradas úteis todas as variáveis para a predição.

```
summary(glmPagamento)
```

```
##
## Call:
## glm(formula = pagamento ~ estadocivil + idade + sexo, family = binomial,
##     data = pagamentoEmprestimo)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4892  -0.4015   0.4166   0.5905   2.1662
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.96591    1.12267  -1.751  0.07993 .
## estadocivil -2.95095    0.58293  -5.062 4.14e-07 ***
## idade        0.11614    0.04432   2.621  0.00877 **
## sexo         1.30123    0.43861   2.967  0.00301 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 212.70  on 179  degrees of freedom
## Residual deviance: 146.65  on 176  degrees of freedom
## AIC: 154.65
##
## Number of Fisher Scoring iterations: 5
```

Predição do modelo

Para testar o modelo, foi realizada a predição.

```
glmprobsPagamento <- predict(glmPagamento, type="response")
```

A predição acima de 0,5 foi considerada para o pagamento do empréstimo e menor ou igual a 0,5 como não pagamento. Foi testado pontos de corte menores e maiores, mas nenhum trouxe maior previsão que o ponto de corte 0,5.

```
nLinhasPagamento <- nrow(pagamentoEmprestimo)
glmpredPagamento <- rep(0, nLinhasPagamento)
glmpredPagamento[ glmprobsPagamento > 0.5 ] <- 1
```

Verificação da previsão

Aplicando a predição para os dados já possuídos, obtiveram-se 24 True Negatives, 125 True Positives, de um total de 180 registros. Os pagamentos forma previstos com aproximadamente 82,8% de sucesso.

```
table(glmpredPagamento, pagamentoEmprestimo$pagamento) -> tabelaPagamentoEmprestimo
tabelaPagamentoEmprestimo
```

```
##
## glmpredPagamento   0   1
##                0  24   5
##                1  26 125
```

```
(as.vector(tabelaPagamentoEmprestimo)[1] + as.vector(tabelaPagamentoEmprestimo)[4]) / nLinhasPagamento
```

```
## [1] 0.8277778
```

Conclusão

O modelo gerado obteve um sucesso de previsão de 82,8% de sucesso sobre os dados já possuídos.

Default de crédito

Leitura dos dados

Os dados do csv gerado a partir da planilha foram carregados para a variável "defaultCredito".

```
defaultCredito <-
    read.csv2("./dados/default_de_credito.csv", stringsAsFactors = FALSE)
```

Criação do modelo

A funcionalidade glm foi utilizada para a geração do modelo de regressão e este vinculado à variável glmDefaultCredito

```
glm(data = defaultCredito,
    formula = default ~ idade + educacao + t_emprego +
                        t_endereco + renda + divida + divida_cc +
                        outras_div, family = binomial) ->
  glmDefaultCredito
```

Análise das variáveis

Após análise inicial do modelo, verificamos que algumas variáveis não rejeitaram a hipótese original, por possuir o valor P muito elevado, não acrescentando relevância ao modelo.

```
summary(glmDefaultCredito)
```

```
##
## Call:
## glm(formula = default ~ idade + educacao + t_emprego + t_endereco +
##     renda + divida + divida_cc + outras_div, family = binomial,
##     data = defaultCredito)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2989  -0.6653  -0.3230   0.1586   2.8708
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.7619012  0.7422673  -2.374 0.017612 *
## idade        0.0305786  0.0204313   1.497 0.134483
## educacao     0.0830897  0.1440116   0.577 0.563963
## t_emprego   -0.2504407  0.0387710  -6.459 1.05e-10 ***
## t_endereco  -0.0967593  0.0270678  -3.575 0.000351 ***
## renda       -0.0003825  0.0111299  -0.034 0.972585
## divida       0.0737017  0.0380499   1.937 0.052748 .
## divida_cc    0.5574310  0.1286410   4.333 1.47e-05 ***
## outras_div   0.0491476  0.0966352   0.509 0.611040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 570.95  on 499  degrees of freedom
## Residual deviance: 401.37  on 491  degrees of freedom
## AIC: 419.37
```

```
##
## Number of Fisher Scoring iterations: 6
```

Modelo final

Removendo as variáveis não relevantes ao modelo, uma a uma, e reexecutando o modelo após a retirada de cada uma foi possível chegar a um modelo com variáveis relevantes.

```
glm(data = defaultCredito,
    formula = default ~ t_emprego + divida + divida_cc, family = binomial) ->
  glmDefaultCredito

summary(glmDefaultCredito)
```

```
##
## Call:
## glm(formula = default ~ t_emprego + divida + divida_cc, family = binomial,
##     data = defaultCredito)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2752  -0.6731  -0.3738   0.2857   2.5518
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.25358    0.27709  -4.524 6.07e-06 ***
## t_emprego   -0.22966    0.03090  -7.434 1.06e-13 ***
## divida       0.08066    0.02210   3.651 0.000262 ***
## divida_cc    0.50322    0.09776   5.148 2.64e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 570.95  on 499  degrees of freedom
## Residual deviance: 416.05  on 496  degrees of freedom
## AIC: 424.05
##
## Number of Fisher Scoring iterations: 5
```

Predição do modelo

Para testar o modelo, foi criada a predição.

```
glmprobsDefaultCredito <- predict(glmDefaultCredito, type="response")
```

A predição acima de 0,5 foi consierada como positiva para a resposta e menor ou igual a 0,5 como negativa.

```
nLinhasDefaultCredito <- nrow(defaultCredito)
glmpredDefaultCredito <- rep(0, nLinhasDefaultCredito)
glmpredDefaultCredito[ glmprobsDefaultCredito > 0.5 ] <- 1
```

Verificação da previsão

A predição foi comparada com os dados já possuídos, obtevem-se 350 True Negatives, 60 True Positives, de um total de 500. Foi possível prever os resultados com 82% de sucesso.

```r
table(glmpredDefaultCredito, defaultCredito$default) -> tabelaDefaultCredito
tabelaDefaultCredito
```

```
##
## glmpredDefaultCredito   0   1
##                     0 350  69
##                     1  21  60
```

```r
(as.vector(tabelaDefaultCredito)[1] + as.vector(tabelaDefaultCredito)[4]) / nLinhasDefaultCredito
```

```
## [1] 0.82
```

Conclusão

O modelo gerado obteve um sucesso de previsão de 82% de sucesso sobre os dados já possuídos.