Analytics & Inteligência Artificial

Tema da aula

Regressão Logística

FIA Business School

NOSSOS DIFERENCIAIS | QUEM SOMOS

BUSINESS SCHOOL
Graduação, pós-graduação, MBA, Pós- MBA, Mestrado Profissional, Curso In *Company* e EAD

CONSULTING
Consultoria personalizada que oferece soluções baseada em seu problema de negócio

RESEARCH
Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais escolas de negócio do mundo, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 projetos de consultorias em organizações públicas e privadas.

NOTA MÁXIMA ENADE 2019 MELHOR ESCOLA DO BRASIL

Único curso de graduação em administração a receber as notas máximas

QS WORLD UNIVERSITY RANKINGS

A primeira escola brasileira a ser finalista da maior competição de MBA do mundo

JOHN MOLSON SCHOOL OF BUSINESS

Única *Business School* brasileira a figurar no *ranking* LATAM

UN GLOBAL COMPACT

Signatária do Pacto Global da ONU

ANAMBA ASSOCIAÇÃO NACIONAL DE MBA

Membro fundador da ANAMBA - Associação Nacional MBAs

ASSOCIATION AMBA ACCREDITED

Credenciada pela AMBA - *Association of MBAs*

Executive MBA Council

Credenciada ao *Executive MBA Council*

AACSB Business Education Alliance Member

Filiada a AACSB - *Association to Advance Collegiate Schools of Business*

EFMD

Filiada a EFMD - *European Foundation for Management Development*

GUIA DO MBA ESTADÃO GUIA DO MBA SELECIONADO

Referência em cursos de MBA nas principais mídias de circulação

O **Laboratório de Análise de Dados** – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de ***Big Data***, ***Analytics*** e **Inteligência Artificial**.


Profª Drª Alessandra Montini

O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil

Os diretores foram professores de grandes especialistas do mercado

+10 anos de atuação

+1000 alunos formados

Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria
- Larga experiência de mercado na resolução de *cases*
- Participação em Congressos Nacionais e Internacionais
- Professor assistente que acompanha o aluno durante todo o curso

Estrutura

- 100% das aulas realizadas em laboratórios
- Computadores para uso individual durante as aulas
- 5 laboratórios de alta qualidade (investimento +R$2MM)
- 2 Unidades próximas a estação de metrô (com estacionamento)

Conteúdo da Aula

1. Introdução
2. Regressão Logística Simples
3. Regressão Logística Múltipla
   i. Teste de hipótese sobre os parâmetros
   ii. Seleção de variáveis
   iii. Análise de desempenho
4. Exercícios

1. Introdução

Objetivo

O modelo de Regressão Logística tem como objetivo **predizer um evento binário**, segundo um conjunto de variáveis explicativas. Um evento binário é codificado como:

- 1, quando ocorre o evento de interesse;
- 0, caso contrário.

Dessa forma, na Regressão Logística, a **variável resposta** deve ser de natureza binária.

Case: Inadimplência em Banco

Exemplo
Identificar a probabilidade de uma pessoa que ainda não é cliente da instituição financeira se tornar inadimplente ao adquirir um crédito pessoal.

Aplicação
Segmento bancário

https://www.youtube.com/watch?v=dwIGfhhgKOc&feature=youtu.be

Case: Seguradora

Exemplo
Identificar a probabilidade de um indivíduo sofrer um sinistro, com base em seu estilo de vida.

Aplicação
Seguradoras

https://www.youtube.com/watch?v=qYSdjUPCwwY&feature=youtu.be

Case: Migração de Plano em Telecom

Exemplo
Identificar os clientes com maior propensão à migração de um plano controle para um plano pós-pago.

Aplicação
Telecomunicações

Case: Doenças Cardíacas

Exemplo
Identificar a probabilidade de um paciente ter problemas coronarianos, de acordo com seu hábito de vida: quantidade de horas de sono, quantidade de refeições diárias, frequência de consumo de frituras, frequência de consumo de doces, frequência de exercícios físicos, valor de colesterol total, valor de triglicérides etc.

Aplicação
Área Médica

Case: Fraude na transação de cartão de crédito

Exemplo
Identificar a probabilidade de uma transação de cartão de crédito ser uma fraude, com base em suas características.

Aplicação
Cartão de Crédito

2. Regressão Logística Simples

Exemplo

Considere o case: *Fraude na transação de cartão de crédito.*

A variável de interesse Y é uma variável aleatória, que pode assumir os valores Y = 0 ou Y = 1. Podemos considerar Y = 0 como “não fraude” e Y = 1 como “fraude”.

Y = 0

Y = 1

Definição da resposta

Considere o case: *Fraude na transação de cartão de crédito.*

A variável de interesse Y é uma variável aleatória, que pode assumir os valores Y = 0 ou Y = 1. Podemos considerar Y = 0 como “não fraude” e Y = 1 como “fraude”.

A Regressão Logística aloca uma nova observação em 1 dentre 2 possíveis grupos, por meio do cálculo de uma probabilidade ***p***, que corresponde à probabilidade de a variável resposta Y assumir o valor 1.
Pode ser denotada, também, como P(Y = 1).

Uma probabilidade é um valor que varia entre 0 e 1, ou seja:

$$0 \leq p \leq 1$$

Variável Explicativa

Considere o case: *Fraude na transação de cartão de crédito.*

A variável de interesse Y é uma variável aleatória, que pode assumir os valores Y = 0 ou Y = 1. Podemos considerar Y = 0 como “não fraude” e Y = 1 como “fraude”.

O cálculo desta probabilidade é feito por meio da variável explicativa (**X**). No exemplo, podemos determinar a probabilidade de fraude a depender do tempo de banco do funcionário.

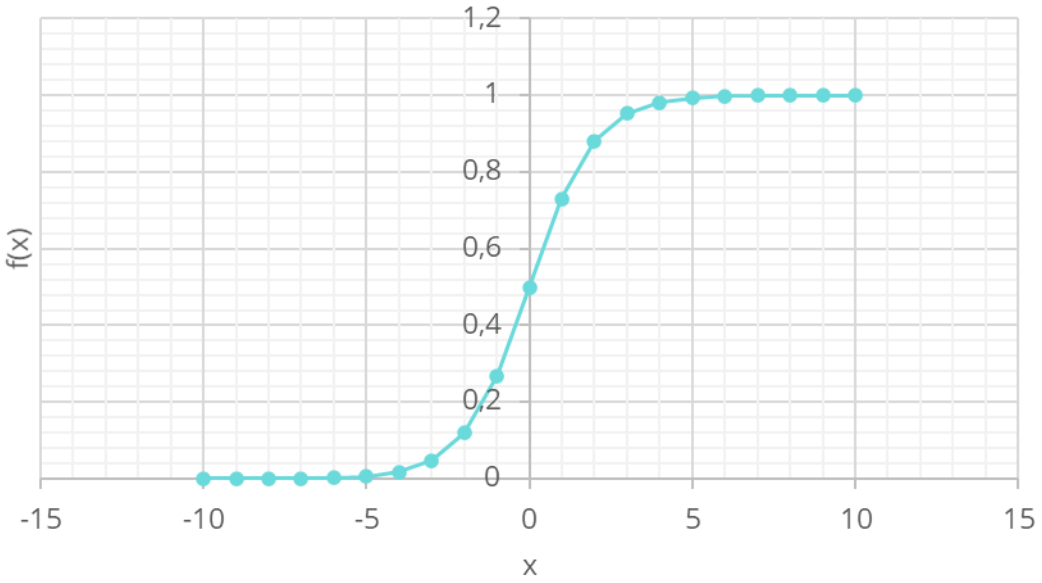
A variável X pode assumir qualquer valor. Ou seja, tecnicamente, X pode variar de menos infinito a mais infinito.

$$-\infty \leq X \leq +\infty$$

Função Logística

A **Função Logística** - f(X) - é uma função que assume valores entre 0 e 1, sendo que a variável X pode assumir qualquer valor.

O gráfico representa a **Função Logística**. Os valores de X estão variando entre -10 e 10 e os valores de f(X) variando entre 0 e 1.



A **Função Logística** é dada por:

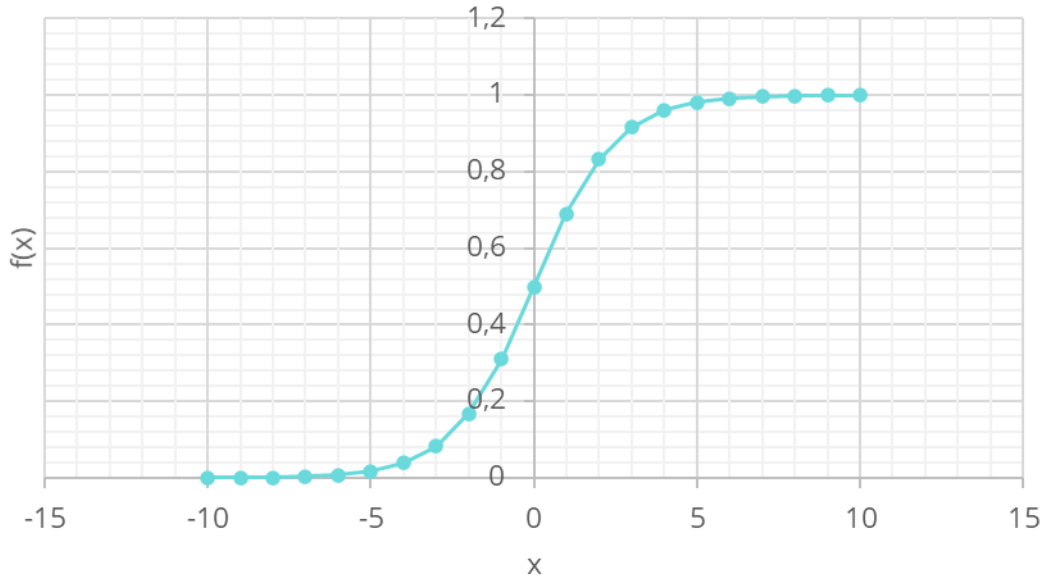
$$f(X) = \frac{e^{\beta X}}{1 + e^{\beta X}}$$

Sendo que ***β*** é um valor fixado; e o número ***e*** representa a base do logaritmo neperiano, ou seja, aproximadamente 2,7182.

Função Logística

A **Função Logística** - f(X) - é uma função que assume valores entre 0 e 1, sendo que a variável X pode assumir qualquer valor.

Quando o coeficiente β da função logística é **positivo**, o valor da função *f* **cresce** à medida que se aumenta o valor de X. A figura abaixo apresenta a função logística com $\beta = 0{,}8$.

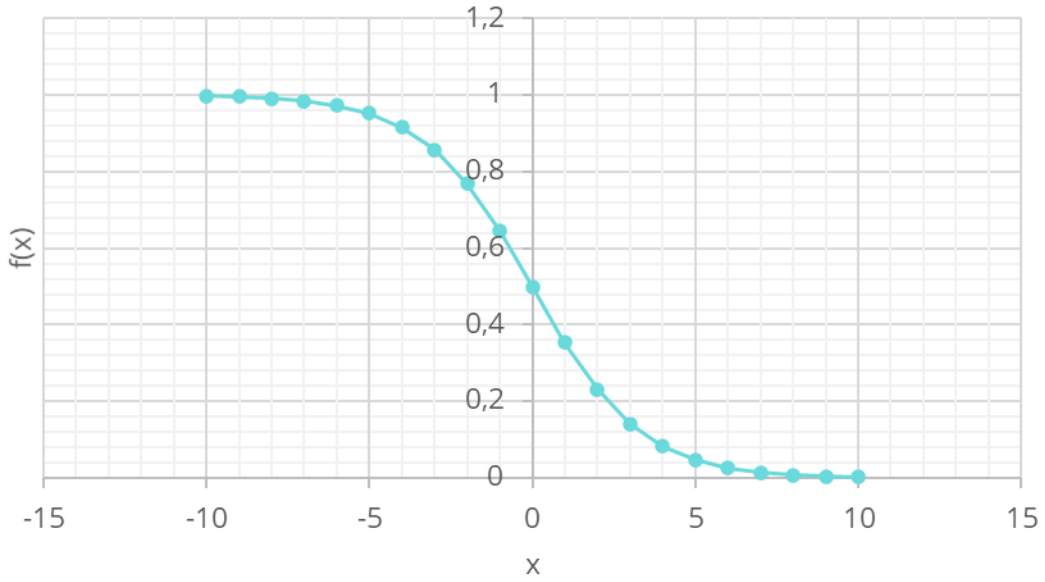


$$f(X) = \frac{e^{0,8*X}}{1 + e^{0,8*X}}$$

Função Logística

A **Função Logística** - f(X) - é uma função que assume valores entre 0 e 1, sendo que a variável X pode assumir qualquer valor.

Quando o coeficiente β da função logística é **negativo**, o valor da função *f* **decresce** à medida que se aumenta o valor de X. A figura abaixo apresenta a função logística com β = -0,6.



$$f(X) = \frac{e^{-0,6*X}}{1 + e^{-0,6*X}}$$

Modelo de Regressão Logística

A Função Logística pode ser utilizada para se obter a **probabilidade** (*p*) associada ao evento binário, no modelo de Regressão Logística.

No exemplo de fraude, a probabilidade de fraude é dependente da variável X_1 = *tempo de relacionamento*. Dessa forma:

$$p = P(Y = 1) = f(X_1) = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}$$

Modelo de Regressão Logística

A Função Logística pode ser utilizada para se obter a **probabilidade** (*p*) associada ao evento binário, no modelo de Regressão Logística.

A probabilidade de uma transação ser fraudulenta também pode ser escrita como:

$$p = P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1)}}$$

3. Regressão Logística Múltipla

Ainda no exemplo anterior, a probabilidade de fraude (***p***) pode ser calculada considerando diversas variáveis explicativas, tais como:

X_1 = *tempo de relacionamento,*

X_2 = *valor,*

X_3 = *idade,*

X_4 = *renda.*

Neste caso, a probabilidade de fraude pode ser calculada por meio da função logística, da seguinte forma:

$$p = P(Y = 1) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4}}$$

Ainda no exemplo anterior, a probabilidade de fraude (***p***) pode ser calculada considerando diversas variáveis explicativas, tais como:

X_1 = *tempo de relacionamento,*

X_2 = *valor,*

X_3 = *idade,*

X_4 = *renda.*

A probabilidade de fraude também pode ser escrita como:

$$p = P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)}}$$

Case: Aumento do time de vendas

Com o objetivo de aumentar seu time de vendas, uma indústria de vendas diretas de cosméticos deseja identificar o perfil das pessoas que poderiam tornar-se consultores de seus produtos.

Seja Y o evento “tornar-se consultor” (1 = Sim e 0 = Não) e as variáveis explicativas:

Sexo: M (masculino) e F (feminino)

Idade: idade do cliente

Cidade: RJ ou SP

As primeiras linhas da base de dados são apresentadas na tabela. A coluna *Resposta* assume **valor = 1** para indivíduos que **tornaram-se consultores** e **valor = 0** para indivíduos que **não se tornaram consultores.**

Sexo	Idade	Cidade	Resposta
M	49	SP	0
M	32	SP	0
F	48	SP	0
F	32	RJ	0
F	64	SP	0
F	56	SP	0
F	69	SP	0

Arquivo: Cosmeticos.xlsx

Case: Aumento do time de vendas

Com o objetivo de aumentar seu time de vendas, uma indústria de vendas diretas de cosméticos deseja identificar o perfil das pessoas que poderiam tornar-se consultores de seus produtos.

Seja Y o evento “tornar-se consultor” (1 = Sim e 0 = Não) e as variáveis explicativas:

Sexo: M (masculino) e F (feminino)

Idade: idade do cliente

Cidade: RJ ou SP

A probabilidade de o indivíduo se tornar consultor de cosméticos, denotada como p = P(Y=1), pode ser escrita como:

$$p = \frac{e^{\beta_0+\beta_1 sexo+\beta_2 idade+\beta_3 cidade}}{1+e^{\beta_0+\beta_1 sexo+\beta_2 idade+\beta_3 cidade}}$$

Arquivo: Cosmeticos.xlsx

Teste de hipótese sobre os parâmetros do modelo

Seja a hipótese nula (H0) correspondente a assumir que os valores β_i são iguais a zero ($\beta_i = 0$), para i = 1, 2 e 3. Podemos testar as seguintes hipóteses alternativas:

H_1: caso a variável sexo seja importante, o parâmetro β_1 será diferente de zero ($\beta_1 \neq 0$)

H_1: caso a variável idade seja importante, o parâmetro β_2 será diferente de zero ($\beta_2 \neq 0$)

H_1: caso a variável cidade seja importante, o parâmetro β_3 será diferente de zero ($\beta_3 \neq 0$)

A probabilidade de o indivíduo se tornar consultor de cosméticos, denotada como p = P(Y=1), pode ser escrita como:

$$p = \frac{e^{\beta_0+\beta_1 sexo+\beta_2 idade+\beta_3 cidade}}{1+e^{\beta_0+\beta_1 sexo+\beta_2 idade+\beta_3 cidade}}$$

Arquivo: Cosmeticos.xlsx

Teste de hipótese sobre os parâmetros do modelo

Para verificar se a variável **sexo** deve fazer parte do modelo, deve-se testar a hipótese:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

```
Call:
glm(formula = Resposta ~ Sexo + Idade + Cidade, family = binomial(link = "logit"),
    data = cosmeticos)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.96839  -0.25510   0.02773   0.31641   2.39982

Coefficients:
            Estimate Std. Error z value      Pr(>|z|)
(Intercept)   8.0459     1.5376   5.233 0.00000016695 ***
SexoM         1.5793     0.5497   2.873       0.00407 **
Idade        -0.1958     0.0395  -4.956 0.00000072044 ***
CidadeSP     -3.2332     0.5477  -5.903 0.00000000357 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Arquivo: Cosmeticos.xlsx

Como o nível descritivo (**0,0047**) < **0,10**, rejeita-se a hipótese H_0, evidenciando que $\beta_1 \neq 0$. Ou seja, a variável *sexo* deve fazer parte do modelo.

Teste de hipótese sobre os parâmetros do modelo

Para verificar se a variável **idade** deve fazer parte do modelo, deve-se testar a hipótese:

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 \neq 0$$

```
Call:
glm(formula = Resposta ~ Sexo + Idade + Cidade, family = binomial(link = "logit"),
    data = cosmeticos)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.96839  -0.25510   0.02773   0.31641   2.39982

Coefficients:
            Estimate Std. Error z value      Pr(>|z|)
(Intercept)   8.0459     1.5376   5.233 0.00000016695 ***
SexoM         1.5793     0.5497   2.873       0.00407 **
Idade        -0.1958     0.0395  -4.956 0.00000072044 ***
CidadeSP     -3.2332     0.5477  -5.903 0.00000000357 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como o nível descritivo (**0,00000072**) < **0,10**, rejeita-se a hipótese H_0, evidenciando que $\beta_2 \neq 0$. Ou seja, a variável *idade* deve fazer parte do modelo.

Arquivo: Cosmeticos.xlsx

Teste de hipótese sobre os parâmetros do modelo

Para verificar se a variável **cidade** deve fazer parte do modelo, deve-se testar a hipótese:

$$H_0: \beta_3 = 0$$
$$H_1: \beta_3 \neq 0$$

```
Call:
glm(formula = Resposta ~ Sexo + Idade + Cidade, family = binomial(link = "logit"),
    data = cosmeticos)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.96839  -0.25510   0.02773   0.31641   2.39982

Coefficients:
            Estimate Std. Error z value      Pr(>|z|)
(Intercept)   8.0459     1.5376   5.233 0.00000016695 ***
SexoM         1.5793     0.5497   2.873       0.00407 **
Idade        -0.1958     0.0395  -4.956 0.00000072044 ***
CidadeSP     -3.2332     0.5477  -5.903 0.00000000357 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Arquivo: Cosmeticos.xlsx

Como o nível descritivo (**0,00000000357**) < **0,10**, rejeita-se a hipótese H_0, evidenciando que $\beta_3 \neq 0$. Ou seja, a variável *cidade* deve fazer parte do modelo.

Modelo com várias covariáveis

Output do modelo no R

```
Call:
glm(formula = Resposta ~ Sexo + Idade + Cidade, family = binomial(link = "logit"),
    data = cosmeticos)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.96839  -0.25510   0.02773   0.31641   2.39982

Coefficients:
            Estimate Std. Error z value      Pr(>|z|)
(Intercept)   8.0459     1.5376   5.233 0.00000016695 ***
SexoM         1.5793     0.5497   2.873       0.00407 **
Idade        -0.1958     0.0395  -4.956 0.00000072044 ***
CidadeSP     -3.2332     0.5477  -5.903 0.00000000357 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Coeficientes estimados

Arquivo: Cosmeticos.xlsx

Sexo = “M” – O peso positivo indica que se o profissional for do sexo masculino, a probabilidade de se tornar um consultor aumenta.

Idade – O peso negativo indica que, à medida que a idade aumenta, a probabilidade de se tornar um consultor diminui.

Cidade = “SP” – O peso negativo indica que se o profissional for da cidade de “São Paulo”, a probabilidade de se tornar um consultor diminui.

Modelo final

Tendo em vista que todas as variáveis são importantes, pode-se calcular a **probabilidade final** de o indivíduo se tornar um consultor, como:

$$p = \frac{e^{8{,}0459+1{,}5793*SexoM-0{,}1958*Idade-3{,}2332*CidadeSP}}{1+e^{8{,}0459+1{,}5793*SexoM-0{,}1958*Idade-3{,}2332*CidadeSP}}$$

Arquivo: Cosmeticos.xlsx

Cálculo da probabilidade

Tendo em vista que todas as variáveis são importantes, pode-se calcular a **probabilidade final** de o indivíduo se tornar um consultor, como:

$$p = \frac{e^{8{,}0459+1{,}5793*SexoM-0{,}1958*Idade-3{,}2332*CidadeSP}}{1+e^{8{,}0459+1{,}5793*SexoM-0{,}1958*Idade-3{,}2332*CidadeSP}}$$

Por exemplo, para um indivíduo do gênero **masculino**, com **49 anos de idade** e que mora na **cidade de São Paulo**, teríamos:

$$p = \frac{e^{8{,}0459+1{,}5793*1-0{,}1958*49-3{,}2332*1}}{1+e^{8{,}0459+1{,}5793*1-0{,}1958*49-3{,}2332*1}} = 0{,}0391$$

Arquivo: Cosmeticos.xlsx

Case: Consórcio

Suponha que uma instituição financeira deseja estimar a probabilidade de um cliente contratar um consórcio de automóvel, p = P(Y=1). A variável resposta é dada por: 1 se contratou o consórcio; 0 se não contratou o consórcio.

Contratou	DI	Financiamento	Poupança	Salário	CC
1	R$ 29.537,63	R$ 4.923.797,68	R$ 648,68	R$ 90.943,36	R$ 5.390,57
1	R$ 196.755,37	R$ 4.006.518,12	R$ 2.208,24	R$ 81.735,48	R$ 4.632,16
1	R$ 120.872,58	R$ 3.583.136,88	R$ 7.747,86	R$ 85.892,78	R$ 5.808,67
1	R$ 215.312,00	R$ 3.516.259,70	R$ 5.584,00	R$ 3.770,16	R$ 3.622,61
1	R$ 738.038,14	R$ 3.248.257,80	R$ 67.986,18	R$ 1.317,92	R$ 9.560,52
1	R$ 626.747,30	R$ 3.146.135,04	R$ 99.229,76	R$ 22.481,71	R$ 7.166,84
0	R$ 172.040,20	R$ 3.131.560,05	R$ 1.967,62	R$ 30.223,13	R$ 2.398,16
...	...	...	...	...	...

Arquivo: Consorcio.xlsx

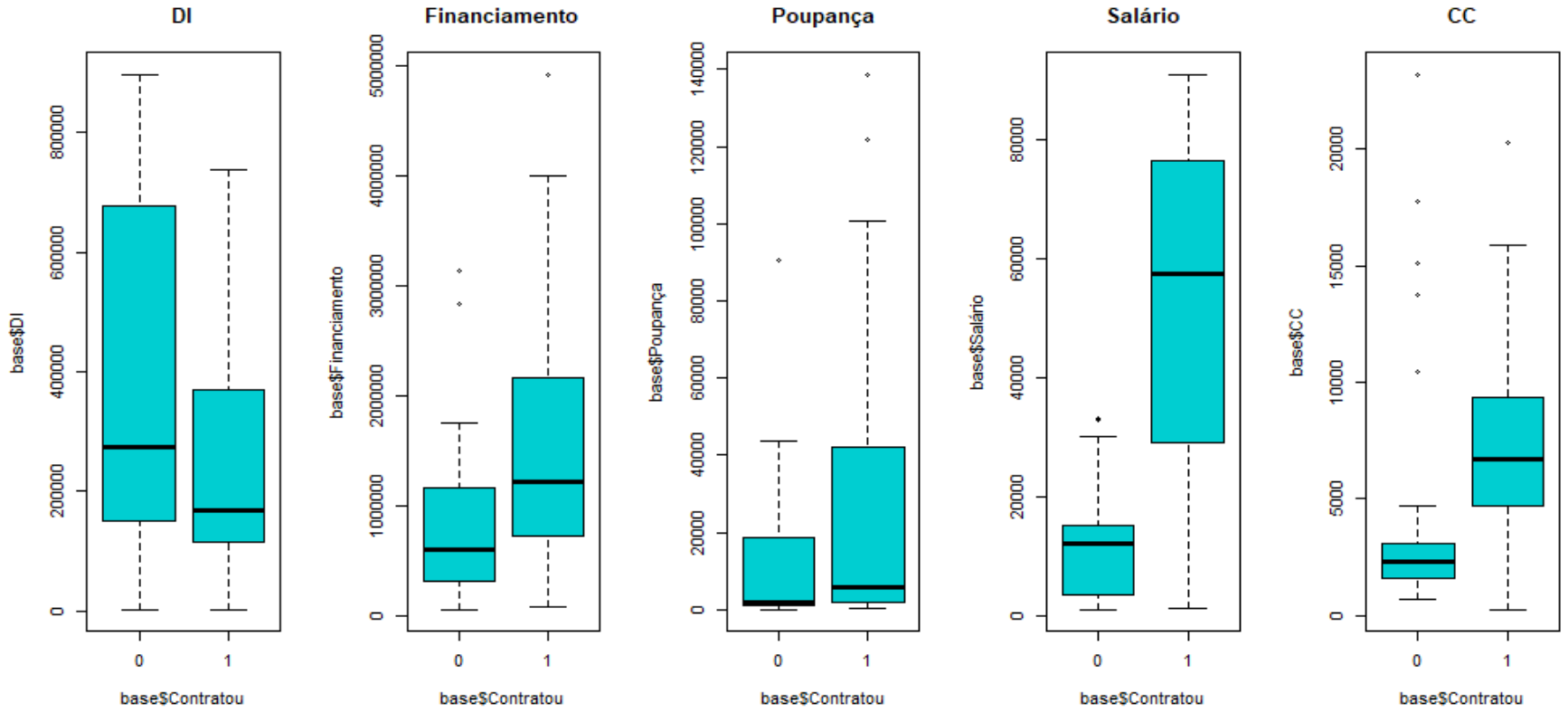
Case: Consórcio

Suponha que uma instituição financeira deseja estimar a probabilidade de um cliente contratar um consórcio de automóvel, p = P(Y=1). A variável resposta é dada por: 1 se contratou o consórcio; 0 se não contratou o consórcio.

(a) Faça a análise exploratória univariada e interprete todas as variáveis do banco de dados.
(b) Faça a análise bivariada das variáveis explicativas (covariáveis) *versus* variável resposta. Você acredita que:
    i. O valor investido no fundo DI (variável *DI*) é menor, em geral, para quem contratou consórcio?
    ii. O valor do financiamento imobiliário (variável *Financiamento*) é maior, em geral, para quem contratou consórcio?
    iii. O valor aplicado na poupança (variável *Poupança*) é maior, em geral, para quem contratou consórcio?
    iv. O salário do cliente (variável *Salário*) é maior, em geral, para quem contratou consórcio?
    v. O gasto no cartão de Crédito (variável *CC*) é menor, em geral, para quem contratou consórcio?
(c) Obtenha o modelo de regressão logística, utilizando 90% de confiança.
(d) Obtenha a probabilidade estimada.
(e) Obtenha a tabela de classificação.

Arquivo: Consorcio.xlsx

Análise exploratória bivariada



Arquivo: Consorcio.xlsx

Case: Consórcio

O modelo de regressão logística obtido é apresentado abaixo. As covariáveis devem fazer parte do modelo? Considere 90% de confiança.

```
Call:
glm(formula = Contratou ~ DI + Financiamento + Poupança + Salário +
    CC, family = binomial(link = "logit"), data = base)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.62961  -0.37557   0.04331   0.14503   3.03149

Coefficients:
                  Estimate    Std. Error z value Pr(>|z|)
(Intercept)   -3.4498742083  1.3517507781  -2.552 0.010706 *
DI            -0.0000042691  0.0000024702  -1.728 0.083943 .
Financiamento  0.0000005792  0.0000004420   1.310 0.190039
Poupança       0.0000647821  0.0000242716   2.669 0.007607 **
Salário        0.0001025703  0.0000263882   3.887 0.000102 ***
CC             0.0000679123  0.0001029244   0.660 0.509365
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

O modelo de Regressão Logística pode ser ajustado utilizando a função **glm** (“generalized linear models”), com a família *binomial (link = logit)*

Arquivo: Consorcio.xlsx

Case: Consórcio

O modelo de regressão logística obtido é apresentado abaixo. As covariáveis devem fazer parte do modelo? Considere 90% de confiança.

```
Call:
glm(formula = Contratou ~ DI + Financiamento + Poupança + Salário +
    CC, family = binomial(link = "logit"), data = base)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.62961  -0.37557   0.04331   0.14503   3.03149

Coefficients:
                 Estimate    Std. Error z value Pr(>|z|)
(Intercept)   -3.4498742083  1.3517507781  -2.552 0.010706 *
DI            -0.0000042691  0.0000024702  -1.728 0.083943 .
Financiamento  0.0000005792  0.0000004420   1.310 0.190039
Poupança       0.0000647821  0.0000242716   2.669 0.007607 **
Salário        0.0001025703  0.0000263882   3.887 0.000102 ***
CC             0.0000679123  0.0001029244   0.660 0.509365
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Arquivo: Consorcio.xlsx

1º passo

Retirar a variável com maior p-valor, caso haja alguma com p-valor > 0,10 (90% de confiança).

Neste caso, retiramos a variável *CC* e ajustamos o modelo novamente.

Case: Consórcio

O modelo de regressão logística obtido é apresentado abaixo. As covariáveis devem fazer parte do modelo? Considere 90% de confiança.

```
Call:
glm(formula = Contratou ~ DI + Financiamento + Poupança + Salário,
    family = binomial(link = "logit"), data = base)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.66397  -0.35712   0.04811   0.18997   2.99883

Coefficients:
                  Estimate    Std. Error z value  Pr(>|z|)
(Intercept)   -2.9476599182  1.0752024912  -2.741   0.00612 **
DI            -0.0000048346  0.0000023773  -2.034   0.04199 *
Financiamento  0.0000005451  0.0000004395   1.240   0.21489
Poupança       0.0000702796  0.0000234092   3.002   0.00268 **
Salário        0.0001002615  0.0000252699   3.968 0.0000726 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Arquivo: Consorcio.xlsx

2º passo

Se ainda restar alguma variável com p-valor > 0,10, retirar novamente a de maior p-valor.

Neste caso, retiramos a variável *Financiamento* e ajustamos o modelo novamente.

Case: Consórcio

O modelo de regressão logística obtido é apresentado abaixo. As covariáveis devem fazer parte do modelo? Considere 90% de confiança.

```
Call:
glm(formula = Contratou ~ DI + Poupança + Salário, family = binomial(link = "logit"),
    data = base)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8009  -0.4484   0.0593   0.2024   2.7954

Coefficients:
               Estimate   Std. Error z value  Pr(>|z|)
(Intercept) -2.370693277  0.877561334  -2.701   0.00690 **
DI          -0.000004758  0.000002221  -2.142   0.03216 *
Poupança     0.000071786  0.000022639   3.171   0.00152 **
Salário      0.000101665  0.000025074   4.055 0.0000502 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Todos os p-valores associados aos parâmetros das variáveis explicativas são < 0,10.

Arquivo: Consorcio.xlsx

Case: Consórcio

A probabilidade de um cliente contratar o consórcio é dada por:

$$p = \frac{e^{-2,3707-0,0000047*DI+0,0000717*Poupança+0,0001016*Salario}}{1+e^{-2,3707-0,0000047*DI+0,0000717*Poupança+0,0001016*Salario}}$$

Arquivo: Consorcio.xlsx

Case: Consórcio

Cálculo da probabilidade estimada no **Excel**:

Contratou	DI	Financiamento	Poupança	Salário	CC	Prob. Estimada
1	R$ 29.537,63	R$ 4.923.797,68	R$ 648,68	R$ 90.943,36	R$ 5.390,57	0,99887
1	R$ 196.755,37	R$ 4.006.518,12	R$ 2.208,24	R$ 81.735,48	R$ 4.632,16	0,99430
1	R$ 120.872,58	R$ 3.583.136,88	R$ 7.747,86	R$ 85.892,78	R$ 5.808,67	0,99824
1	R$ 215.312,00	R$ 3.516.259,70	R$ 5.584,00	R$ 3.770,16	R$ 3.622,61	0,06843
1	R$ 738.038,14	R$ 3.248.257,80	R$ 67.986,18	R$ 1.317,92	R$ 9.560,52	0,29570
1	R$ 626.747,30	R$ 3.146.135,04	R$ 99.229,76	R$ 22.481,71	R$ 7.166,84	0,98298

Variável	Estimativa
(Intercept)	-2,370693277
DI	-0,000004758
Poupança	0,000071786
Salário	0,000101665

Arquivo: Consorcio.xlsx; aba “Cálculo da Probabilidade”

Case: Consórcio

Cálculo da probabilidade estimada no **R**, utilizando *predict(modelo, base, type = "response")*

Contratou	DI	Financiamento	Poupança	Salário	CC	probabilidade	predito
1	29537.631	4923797.7	648.681	90943.36	5390.5723	0.99886626	1
1	196755.372	4006518.1	2208.237	81735.48	4632.1640	0.99429809	1
1	120872.580	3583136.9	7747.857	85892.78	5808.6687	0.99824344	1
1	215312.000	3516259.7	5583.999	3770.16	3622.6124	0.06842748	0
1	738038.139	3248257.8	67986.180	1317.92	9560.5224	0.29563623	0
1	626747.304	3146135.0	99229.760	22481.71	7166.8360	0.98297481	1

Deve-se definir um critério de alocação dos indivíduos nos grupos *"contratou consórcio"* (Y=1) e *"não contratou consórcio"* (Y=0). Será considerando como ponto de corte a probabilidade estimada **p > 0,5**.

Arquivo: Consorcio.xlsx

Case: Consórcio

Após a obtenção da probabilidade de contratação, é importante obter a **tabela de classificação**.

Output do R:

		base$predito	
		0	1
base$Contratou	0	37	3
	1	5	44
	#Total cases	42	47

Arquivo: Consorcio.xlsx

Case: Consórcio

Após a obtenção da probabilidade de contratação, é importante obter a **tabela de classificação**.

Output do R:

		base$predito	
		0	1
base$Contratou	0	37	3
	1	5	44
	#Total cases	42	47

- 44 indivíduos que contrataram foram classificados **corretamente**.
- 5 indivíduos que contrataram foram classificados **incorretamente**.

Arquivo: Consorcio.xlsx

Case: Consórcio

Após a obtenção da probabilidade de contratação, é importante obter a **tabela de classificação**.

Output do R:

		base$predito	
		0	1
base$Contratou	0	37	3
	1	5	44
	#Total cases	42	47

- 37 indivíduos que não contrataram foram classificados **corretamente**.
- 3 indivíduos que não contrataram foram classificados **incorretamente**.

Arquivo: Consorcio.xlsx

Case: Consórcio

Após a obtenção da probabilidade de contratação, é importante obter a **tabela de classificação**.

Output do R:

		base$predito	
		0	1
base$Contratou	0	37	3
	1	5	44
	#Total cases	42	47

O percentual global de classificação correta (ou **acurácia**) é dado por: $\frac{37+44}{37+3+5+44} = 91,01\%$

Arquivo: Consorcio.xlsx

Análise de Desempenho

A **tabela de classificação** apresenta o cruzamento da variável resposta observada com a variável resposta predita pelo modelo. Ela também é conhecida como **matriz de confusão**.

Um bom ajuste de modelo apresenta grande concentração de casos na diagonal principal.

Tabela de classificação, avaliada no ponto de corte:

		Variável resposta predita		
		0	**1**	**Total**
Variável resposta observada	**0**	VN	FP	VN + FP
	1	FN	VP	FN + VP
	Total	VN + FN	FP + VP	VN + FN + FP + VP

VP = verdadeiro positivo; VN = verdadeiro negativo; FP = falso positivo; FN = falso negativo

Análise de Desempenho

Tabela de classificação avaliada no ponto de corte:

		Variável resposta predita		
		0	**1**	**Total**
Variável resposta observada	**0**	VN	FP	VN + FP
	1	FN	VP	FN + VP
	Total	VN + FN	FP + VP	VN + FN + FP + VP

VP = verdadeiro positivo; VN = verdadeiro negativo; FP = falso positivo; FN = falso negativo

Os índices de **acurácia**, **sensibilidade** e **especificidade** variam de 0 a 1 (ou de 0% a 100%).

Considerando que um evento tenha probabilidade de ocorrência de 50%. Valores acima de 50% indicam acerto superior ao aleatório (ausência de modelo). Valores acima de 60% são considerados índices satisfatórios. Já valores acima de 70%-75% indicam ótimo desempenho.

Acurácia

$$Acur = \frac{VP + VN}{VP + VN + FP + FN}$$

Sensibilidade

$$Sensib = \frac{VP}{VP + FN}$$

Especificidade

$$Espec = \frac{VN}{FP + VN}$$

Tabela de classificação – Case de ***Consórcio***:

		Variável resposta predita		
		0	**1**	**Total**
Variável resposta observada	**0**	37	3	40
	1	5	44	49
	Total	42	47	89

Acurácia

$$Acur = \frac{37 + 44}{89} = 91{,}0\%$$

Sensibilidade

$$Sensib = \frac{44}{49} = 89{,}8\%$$

Especificidade

$$Espec = \frac{37}{40} = 92{,}5\%$$

A acurácia de 91,0% mostra que, em geral a cada 100 clientes, o modelo **acerta** se haverá contratação de consórcio ou não para 91 deles.

Os percentuais de sensibilidade e especificidade são próximos, o que indica que o modelo apresenta capacidades semelhantes de acertar tanto os clientes que **contratam** quanto os que **não contratam** consórcio.

Arquivo: Consorcio.xlsx

4. Exercícios

Case: Cancelamento Telecom

Um diretor encarregado da tarefa de retenção de clientes de uma Telecom deseja criar um modelo para calcular a probabilidade de o cliente cancelar sua conta, a fim de realizar ações de retenção ativa.

ID	Score_Serasa	Sexo	Idade	Tempo_relacionamento	Possui_internet	Salario_anual	Cancelou
15634602	619	Female	42	2	1	101348,88	1
15647311	608	Female	41	1	0	112542,58	0
15619304	502	Female	42	8	1	113931,57	1
15701354	699	Female	39	1	0	93826,63	0
15737888	850	Female	43	2	1	79084,1	0
15574012	645	Male	44	8	1	149756,71	1
15592531	822	Male	50	7	1	10062,8	0
15656148	376	Female	29	4	1	119346,88	1
15792365	501	Male	44	4	0	74940,5	0
15592389	684	Male	27	2	1	71725,73	0
...	...	...	...	...	...	...	...

Arquivo: Cancelamento_Telecom.xlsx

Case: Cancelamento Telecom

Um diretor encarregado da tarefa de retenção de clientes de uma Telecom deseja criar um modelo para calcular a probabilidade de o cliente cancelar sua conta, a fim de realizar ações de retenção ativa.

(a) Faça a análise exploratória univariada e interprete todas as variáveis do banco de dados.
(b) Faça a análise bivariada das variáveis explicativas (covariáveis) vs. a variável resposta. Interprete os resultados.
(c) Obtenha o modelo de regressão logística utilizando 90% de confiança.
(d) Obtenha a probabilidade estimada.
(e) Obtenha a tabela de classificação utilizando o ponto de corte de 0,2. Qual o percentual de classificação correta? E a sensibilidade e especificidade?

Arquivo: Cancelamento_Telecom.xlsx

Case: Avaliação de risco de um empréstimo bancário

Considere dados históricos provenientes de um banco de varejo, referentes a 5.000 propostas de crédito geradas durante solicitações de empréstimo. A base traz informações como idade, nível de instrução, tempo de experiência, tempo no endereço e renda; além da variável resposta "*classif*" (0=bom pagador, 1=mau pagador). Deseja-se avaliar o potencial dessas variáveis para predizer se um cliente será um bom ou mau pagador.

idade	experiencia	tempo_endereco	renda	debito_renda	cred_deb	outros_debitos	classif
41	17	12	35,9	11,9	0,5041078	3,7679922	0
30	13	8	46,7	17,88	1,35269352	6,99726648	0
40	15	14	61,8	10,64	3,43899696	3,13652304	0
41	15	14	72	29,67	4,165668	17,196732	0
57	7	37	25,6	15,86	1,49819904	2,56196096	0
45	0	13	28,1	4,28	0,92486092	0,27781908	0
36	1	3	19,6	12,82	1,21113104	1,30158896	1
39	20	9	80,5	12,32	1,8545912	8,0630088	0
43	12	11	68,7	6,82	1,4290287	3,2563113	0
....	...	...	...	...	...	...	...

Arquivo: Emprestimo_Bancario.xlsx

Case: Avaliação de risco de um empréstimo bancário

Considere dados históricos provenientes de um banco de varejo, referentes a 5.000 propostas de crédito geradas durante solicitações de empréstimo. A base traz informações como idade, nível de instrução, tempo de experiência, tempo no endereço e renda; além da variável resposta "*classif*" (0=bom pagador, 1=mau pagador). Deseja-se avaliar o potencial dessas variáveis para predizer se um cliente será um bom ou mau pagador.

(a) Faça a análise exploratória univariada e interprete todas as variáveis do banco de dados.
(b) Há algum outlier que mereça sua atenção? E com relação a dados faltantes?
(c) Qual o % de empréstimos não pagos (ou "% de maus", ou "% de default")?
(d) Qual a correlação entre as variáveis presentes na base?
(e) Quais são as variáveis que melhor explicam se o empréstimo foi pago ou não?
(f) Ajuste um modelo logístico e avalie quais variáveis apareceram como significantes. Adote 10% de significância.
(g) Calcule o score de propensão a default como uma coluna nova na base.
(h) Qual a sugestão inicial de ponto de corte para predizer se o empréstimo será pago ou não?
(i) Considerando essa sugestão, construa a matriz de confusão. Qual a taxa de classificação correta?

Arquivo: Emprestimo_Bancario.xlsx

Case: Fatores de influência no valor da remuneração mensal

Considere os dados de 534 profissionais, seus salários e algumas informações sociodemográficas. A ideia é tentar entender os fatores que mais influenciam na possibilidade do profissional ganhar um salário superior a 10 s.m. (salários mínimos). A partir da análise dessa base de dados, vamos tentar entender alguns elementos em torno desse tema, com a seguinte base de dados:

EDUCACAO	SUL	SEXO	ANOS_EXPERIENCIA	SALARIO	IDADE	FLAG_CASADO
8	0	1	21	5,1	35	1
9	0	1	42	4,95	57	1
12	0	0	1	6,67	19	0
12	0	0	4	4	22	0
12	0	0	17	7,5	35	1
13	0	0	9	13,07	28	0
10	1	0	27	4,45	43	0
12	0	0	9	19,47	27	0
16	0	0	11	13,28	33	1
12	0	0	9	8,75	27	0
12	0	0	17	11,35	35	1
...	...	...	...	...	...	...

Arquivo: Fatores_Impacto_Salario.xlsx

Case: Fatores de influência no valor da remuneração mensal

Considere os dados de 534 profissionais, seus salários e algumas informações sociodemográficas. A ideia é tentar entender os fatores que mais influenciam na possibilidade do profissional ganhar um salário superior a 10 s.m. (salários mínimos). A partir da análise dessa base de dados, vamos tentar entender alguns elementos em torno desse tema, com a seguinte base de dados:

(a) Faça a análise exploratória univariada e interprete todas as variáveis do banco de dados.
(b) Há algum outlier que mereça atenção? E com relação a dados faltantes?
(c) Qual o % de pessoas na situação de interesse?
(d) Qual a correlação entre as variáveis presentes na base?
(e) Quais são as variáveis que melhor explicam se o profissional ganha ou não mais de 10 s.m.?
(f) Ajuste um modelo logístico e avalie quais variáveis apareceram como significantes. Adote 10% de significância.
(g) Calcule a probabilidade de um indivíduo ganhar mais que 10 s.m. como uma coluna nova na base.
(h) Qual a sugestão inicial de ponto de corte para predizer se a pessoa ganhará ou não mais de 10 s.m.?
(i) Considerando essa sugestão, construa a matriz de confusão. Qual a taxa de classificação correta?

Arquivo: Fatores_Impacto_Salario.xlsx

Case: Cancelamento de seguro auto

Considere os dados de 2.143 indivíduos com seguro de automóvel, em uma certa seguradora. Deseja-se entender quais características influenciam o cancelamento do seguro, construindo um modelo que explique a probabilidade de cancelamento, a fim de estabelecer estratégias de retenção mais efetivas.

Churn	Renda	Reclamacoes	Educacao	Tempo_cliente	Classe_idade	Idade_carro	Debito_autom
1	03_>R$ 5000	04_>5	04_Pós-graduação	01_Até 1 ano	03_36 a 55	02_1 a 3 anos	01_Não
1	03_>R$ 5000	04_>5	04_Pós-graduação	01_Até 1 ano	03_36 a 55	02_1 a 3 anos	01_Não
1	03_>R$ 5000	03_3 a 5	04_Pós-graduação	01_Até 1 ano	02_26 a 35	01_Zero	01_Não
1	03_>R$ 5000	03_3 a 5	04_Pós-graduação	01_Até 1 ano	04_>55	02_1 a 3 anos	01_Não
1	02_Entre R$ 2001 e R$ 5000	04_>5	03_Curso superior	02_1 a 3 anos	03_36 a 55	04_7 a 9 anos	01_Não
1	02_Entre R$ 2001 e R$ 5000	03_3 a 5	03_Curso superior	01_Até 1 ano	03_36 a 55	04_7 a 9 anos	02_Sim
1	02_Entre R$ 2001 e R$ 5000	03_3 a 5	03_Curso superior	01_Até 1 ano	03_36 a 55	05_>=10 anos	01_Não
1	01_Até R$ 2000	04_>5	01_Ensino Fundamental	02_1 a 3 anos	03_36 a 55	01_Zero	02_Sim
1	03_>R$ 5000	03_3 a 5	04_Pós-graduação	01_Até 1 ano	01_Ate 25	02_1 a 3 anos	01_Não
...	...	...	...	...	...	...	...

Arquivo: Cancelamento_Seguro_Auto.xlsx

Case: Cancelamento de seguro auto

Considere os dados de 2.143 indivíduos com seguro de automóvel, em uma certa seguradora. Deseja-se entender quais características influenciam o cancelamento do seguro, construindo um modelo que explique a probabilidade de cancelamento, a fim de estabelecer estratégias de retenção mais efetivas.

(a) Faça a análise exploratória univariada e interprete todas as variáveis do banco de dados. Qual o % churn (ou % de cancelamento)?
(b) Quais são as variáveis que melhor explicam o cancelamento?
(c) Ajuste um modelo de regressão logística e avalie quais variáveis apareceram como significantes. Adote 10% de significância.
(d) A partir do modelo, quais são as variáveis que melhor explicam se o segurado cancelou ou não?
(e) Calcule a probabilidade de propensão ao churn como uma coluna nova na base.
(f) Qual a sugestão inicial de ponto de corte para predizer se o cliente cancelará o seguro ou não?
(g) Considerando essa sugestão, construa a matriz de confusão. Qual a taxa de classificação correta?

Arquivo: Cancelamento_Seguro_Auto.xlsx

Referências

- Agresti, A. (2002). *Categorical data analysis* (Vol. 359). Wiley-interscience.
- Conover, W. J. (1999). *Practical nonparametric statistics*. New York: Wiley.
- Cramér, H. (1945). *Mathematical methods of statistics* (Vol. 9). Princeton University Press.
- Hosmer, D. W. e Lemeshow, S. (2000). *Applied Logistic Regression*, 2ª ed. New York: Wiley.