

DOUGLAS VINÍCIUS GONÇALVES ARAÚJO

**Modelo de Regressão Logística Aplicada a
Previsão de Inadimplência sobre Cartão de
Crédito de uma Instituição Financeira**

JI-PARANÁ

2022

DOUGLAS VINÍCIUS GONÇALVES ARAÚJO

Modelo de Regressão Logística Aplicada a Previsão de Inadimplência sobre Cartão de Crédito de uma Instituição Financeira

Relatório de Estágio Supervisionado apresentado como Trabalho de Pesquisa à Coordenação do Curso de Bacharelado em Estatística da Universidade Federal de Rondônia.

UNIVERSIDADE FEDERAL DE RONDÔNIA – UNIR
DEPARTAMENTO DE MATEMÁTICA E ESTATÍSTICA
RELATÓRIO DE PESQUISA

Jl-PARANÁ

2022

"Os livros servem para nos lembrar quanto somos estúpidos e tolos. São o guarda pretoriano de César, cochichando enquanto o desfile ruge pela avenida: Lembre-se, César, tu és mortal. A maioria de nós não pode sair correndo por aí, falar com todo mundo, conhecer todas as cidades do mundo, não temos tempo, dinheiro ou tantos amigos assim. As coisas que você está procurando, Montag, estão no mundo, mas a única possibilidade que o sujeito comum terá de ver noventa e nove por cento delas está num livro".

- Fahrenheit 451 de Ray Douglas Bradbury

Resumo

O objetivo deste trabalho tem como aplicar uma análise de regressão logística a dados de cartões de crédito de uma instituição financeira do estado de Rondônia, de forma gerar um modelo logístico capaz de prever a probabilidade de inadimplência ou risco de o tomador não honrar com o crédito.

Palavras-chaves: Risco de Crédito, Probabilidade de Default, Modelo de Regressão Logístico.

Lista de ilustrações

Figura 1 – Machine Learning e suas aplicações	10
---	----

Lista de tabelas

Tabela 1 – GroupRows	20
--------------------------------	----

Lista de abreviaturas e siglas

ABNT	Associação Brasileira de Normas Técnicas
abnTeX	ABsurdas Normas para TeX

Lista de símbolos

Γ	Letra grega Gama
Λ	Lambda
ζ	Letra grega minúscula zeta
$f(x; \theta)$	Função de Densidade de Probabilidade
Π	Produtório

Sumário

1	INTRODUÇÃO	9
1.1	Objetivos	9
2	REFERENCIAL TEÓRICO	10
2.1	Modelo de Credit Scoring	10
2.2	Breve Introdução sobre Machine Learning	10
2.3	Modelo de Regressão Logística	11
2.3.1	Interpretação dos Parâmetros	12
2.3.2	Estimação dos Parâmetros	12
2.3.3	Testes de Significância	12
2.3.4	Seleção de Variáveis	12
2.3.5	Desempenho dos Modelos	12
3	METODOLOGIA	13
4	RESULTADOS E DISCUSSÕES	14
5	CONSIDERAÇÕES FINAIS	15
	REFERÊNCIAS	16
	APÊNDICES	17
	APÊNDICE A – DESCRIÇÃO DAS VARIÁVEIS	18
	APÊNDICE B – SCRIPT EM R	19
	APÊNDICE C – SCRIPT EM PYTHON	21

1 Introdução

1.1 Objetivos

O objetivo deste trabalho é desenvolver um modelo de previsão de risco de inadimplência dos tomadores de cartões de créditos de uma instituição Financeira do Estado de Rondônia. Resumidamente, em poucas palavras, estamos interessados em construir um modelo preditivo que propõe efetivamente a decisão sobre o risco de crédito (ou modelo de Credit Scoring).

Neste contexto, vamos relacionar os seguintes objetivos específicos:

2 REFERENCIAL TEÓRICO

2.1 Modelo de Credit Scoring

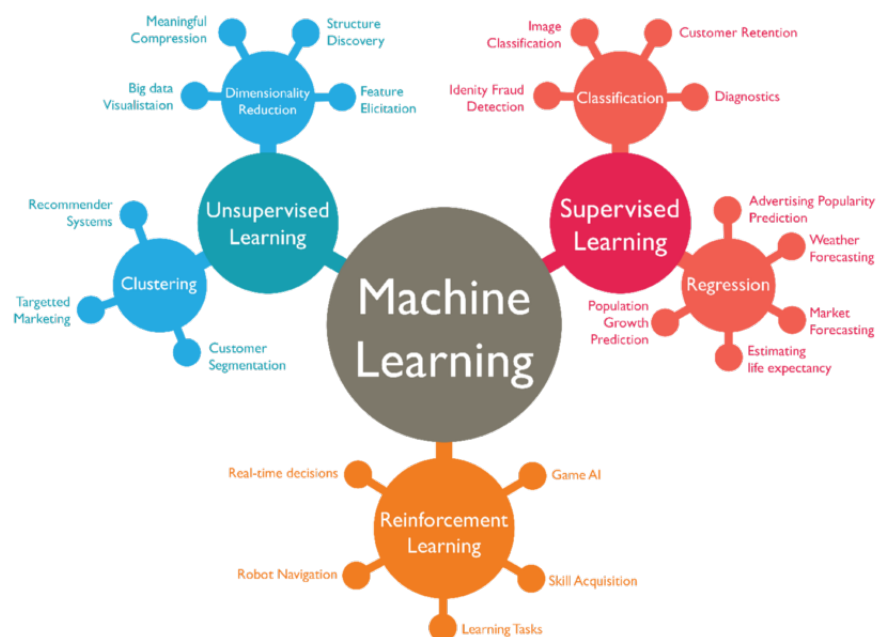
2.2 Breve Introdução sobre Machine Learning

Uma definição básica sobre Machine Learning (Aprendizado de Máquina) é englobar um conjunto de regras com algoritmos e procedimentos que tem como objetivo de extrair informações a partir dos dados e dessas informações tomar uma decisão.

Segundo (GOODFELLOW; BENGIO; COURVILLE, 2016), os algoritmos de Machine Learning podem ser amplamente categorizados pelos tipos de aprendizagem, sintetizando essas diferenças no tipo de experiência durante o aprendizado do algoritmo.

- Supervisionado: O algoritmo procura relação entre as variáveis preditoras e a variável resposta de um *dataset*. Através dessa associação é possível realizar previsões quando o algoritmo é apresentado novos dados;
- Não-Supervisionado: aqui o algoritmo tem como objetivo agrupar os dados com base em características similares, descartando a apresentação da variável resposta ao algoritmo;

Figura 1 – Machine Learning e suas aplicações



- Aprendizagem por reforço: o algoritmo aprende com base nas interações com o ambiente. Não são apresentadas as ações que devem ser tomadas, apenas as consequências das ações.

2.3 Modelo de Regressão Logística

A regressão logística tem como principal uso modelar de uma variável binária (0,1), com base em mais variáveis, estas chamadas de variáveis explicativas ou preditoras. E comumente a variável resposta ou dependente, assim chama-se a variável binária do modelo. Conforme (HILBE, 2016), o melhor modelo ajustado aos dados é assumido que:

- Não há correlação entre as variáveis preditoras;
- Estejam significativamente relacionados com a resposta;
- Que as observações dos dados não interferem entre si.

A resposta do modelo dito está conveniente a uma distribuição subjacente, ou seja, segue uma distribuição de Bernoulli. Concordantemente com (BOLFARINE; SANDOVAL, 2001), esta distribuição é um distribuição particular da distribuição Binomial que a função de probabilidade pode ser expressa:

$$f(x; \theta) = \theta_i^{x_i} (1 - \theta_i)^{1-x_i}, \quad x_i = 0, 1, \quad (2.1)$$

em que $i = 1, \dots, n$. Estes modelos são comumente empregados em situações que a resposta é dicotômica.

Porque não utilizar o modelo de regressão linear? Suponhamos uma situação, estamos tentando prever a condição médica de um paciente com três diagnósticos possíveis: acidente vascular cerebral (AVC), overdose de drogas e convulsões epiléticas. Podemos dar a essas condições valores como uma variável de resposta quantitativa:

$$Y = \begin{cases} 1, & \text{se AVC} \\ 2, & \text{se overdose} \\ 3, & \text{se convulsões} \end{cases}$$

Com essa conversão implica um ordenação dos resultados possíveis de Y , mas não há ordenação, pois se houvesse um ordenamento natural de leve, moderado e grave, da consideração a diferença de leve a moderado e entre moderado e grave seriam semelhantes os intervalos. Infelizmente, em geral, não há uma maneira de converter uma variável resposta qualitativa com mais de dois níveis em uma resposta quantitativa pronta para regressão linear.

Se tivermos uma resposta qualitativa binária (dois níveis), por exemplo, duas condições médicas do paciente e utilizando a variável *dummy* para codificar a resposta:

$$Y = \begin{cases} 1, & \text{se AVC} \\ 2, & \text{se overdose} \end{cases}$$

Mesmo usando a regressão linear para utilizar para obter uma estimativa de probabilidade do resultado, quebramos um pressuposto, pois algumas estimativas podem estar fora do intervalo $[0, 1]$.

E uma forma capaz de ter uma linha em forma de "S" para prever as probabilidades e descrever essa linha curva com os coeficientes da regressão linear.

Presuma que o modelo linear tradicional tenha a forma:

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i \quad (2.2)$$

em que $\mathbf{x}'_i = [1, x_{i1}, x_{i2}, \dots, x_{ik}]$, $\boldsymbol{\beta}' = [\beta_0, \beta_1, \beta_2, \dots, \beta_k]$ e a variável resposta tem valores entre o intervalo $[0, 1]$. Assumiremos que a variável resposta é uma variável aleatória com distribuição de Bernoulli com função de probabilidade dita anteriormente pela equação 2.1.

Uma vez que a $E(\varepsilon_i) = 0$, o valor esperado da variável resposta é:

$$E(y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i \quad (2.3)$$

o que implica em

$$E(y_i) = \mathbf{x}'_i \boldsymbol{\beta} = \pi_i$$

2.3.1 Interpretação dos Paramêtros

2.3.2 Estimação dos Paramêtros

2.3.3 Testes de Significância

2.3.4 Seleção de Variáveis

2.3.5 Desempenho dos Modelos

3 Metodologia

4 Resultados e Discussões

5 Considerações Finais

Referências

BOLFARINE, H.; SANDOVAL, M. C. *Introdução à inferência estatística*. [S.l.]: SBM, 2001. v. 2. Citado na página [11](#).

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. [S.l.]: MIT press, 2016. Citado na página [10](#).

HILBE, J. M. *Practical guide to logistic regression*. [S.l.]: crc Press, 2016. Citado na página [11](#).

Apêndices

APÊNDICE A – DESCRIÇÃO DAS VARIÁVEIS

Variável	Descrição da Variável	Tipo de Variável	Nº de Categorias	Categorias
Sexo				
Idade				
Escolaridade				
Renda				
SM30				
SM60				
SM90				
SM180				
SM360				
Limite				
Status				

APÊNDICE B – Script em R

```
##%#####%##
#####
####          REGRESSÃO LOGISTICA          ####
#####
##%#####%##

library(kableExtra)
dt <- mtcars[1:5, 1:6]

text_tbl<-data.frame(
  Items=c("Item1","Item2","Item3"),
  Features=c("bla bla", "...", "bla ...")
)

kbl(text_tbl,booktabs=T) %>%
  kable_styling(full_width=F) %>%
  column_spec(1,bold=T,color="red") %>%
  column_spec(2,width="30em")

kbl(mtcars[1:10,1:6], caption = "GroupRows", booktabs = T) %>%
  kable_styling() %>%
  pack_rows("Group1",4,7) %>%
  pack_rows("Group2",8,10)
```

Items	Features
Item1	bla bla
Item2	...
Item3	bla ...

Tabela 1 – GroupRows

	mpg	cyl	disp	hp	drat	wt
Mazda RX4	21.0	6	160.0	110	3.90	2.620
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875
Datsun 710	22.8	4	108.0	93	3.85	2.320
Group1						
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440
Valiant	18.1	6	225.0	105	2.76	3.460
Duster 360	14.3	8	360.0	245	3.21	3.570
Group2						
Merc 240D	24.4	4	146.7	62	3.69	3.190
Merc 230	22.8	4	140.8	95	3.92	3.150
Merc 280	19.2	6	167.6	123	3.92	3.440

APÊNDICE C – Script em Python