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

JI-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: Credit Scoring, Machine Learning, Probabilidade de Default, Regressão Logístico.

Lista de ilustrações

Figura 1 – Machine Learning e suas aplicações 9
Figura 2 – Gráfico do modelo de regressão logística. 10

Lista de tabelas

Tabela 1 – Tabela 13

Sumário

1 INTRODUÇÃO 7
1.1 Objetivos 7

2 REFERENCIAL TEÓRICO 8
2.1 Credit Scoring 8
2.2 Breve Introdução sobre Machine Learning 8
2.3 Modelo de Regressão Logística 9
2.4 Estimação dos Parâmetros 10
2.5 Interpretação dos Parâmetros 11
2.6 Testes de Significância 11
2.6.1 Teste da Razão de Verossimilhança (TRV) 11
2.6.2 Teste de Wald 11
2.6.3 Medida da Qualidade do Ajuste do Modelo 12
2.6.4 Intervalo de Confiança 12
2.7 Seleção de Variáveis 12
2.7.1 Diagnóstico Resídual 13
2.7.2 Desempenho do Modelo 13

3 METODOLOGIA 14

4 RESULTADOS E DISCUSSÕES 15

5 CONSIDERAÇÕES FINAIS 16

REFERÊNCIAS 17

APÊNDICES 18

APÊNDICE A – DESCRIÇÃO DAS VARIÁVEIS 19

APÊNDICE B – SCRIPT EM R 20

1 Introdução

A economia em crescimento trás consigo a expansão das linhas de créditos oferecidas pelas instituições financeiras. E umas dessas linhas de créditos oferecidas, conhecemos pelo nome de cartão de crédito. Cartão de crédito em resumo, é uma forma de empréstimos com prazo de até 40 dias, disponível para fazer compras de bens e serviços. As taxas normalmente são padrão entre os bancos, mas o limite é definido com base na renda do solicitante do produto bancário.

Desta forma, torna-se mais importante para as instituições entender melhor os seus clientes para melhor oferecer as concessões de créditos e manter uma carteira de crédito bem gerida, evitar grandes prejuízos.

1.1 Objetivos

A presente pesquisa 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, estamos interessados em construir um modelo preditivo de uma amostra de trainamento e após verificar a eficência deste modelo em uma amostra teste. Este modelo propõe auxiliar na tomada de decisão sobre o risco de crédito (ou modelo de Credit Scoring) das concessões de cartões de créditos.

Em todo processo de tomada de decisão pareamos entre duas hipotéses possíveis, podendo-se cometer dois tipos de erros. Um é recusar a concessão de crédito de um solicitante que, apesar de ter um perfil de alto risco, honraria seu compromisso. Outro erro é oferecer o crédito ao solicitante que irá implicar me perdas no futuro. Portanto, é ideal que temos um modelo eficaz para minizar esses erros.

Neste contexto, vamos relacionar os seguintes objetivos específicos:

- Implementar um modelo de *Credit Scoring* por meio de Regressão Logística;
-
-

2 REFERENCIAL TEÓRICO

2.1 Credit Scoring

Segundo (SICSÚ, 2010), inúmeras tomadas de decisões precede a incertezas, com a concessão de crédito não se destingui disto, conceder crédito implica a possibilidade de perda. Razão está que o credor ao estimar a probabilidade de perda ajudará na sua tomada de decisão mais confiável. E este modelo de estimação chamado de credit scoring tem como objetivo de prever ou quantificar, na data da concessão de crédito, a probabilidade de perda em uma operação de crédito que denominanos **risco de crédito**.

2.2 Breve Introdução sobre Machine Learning

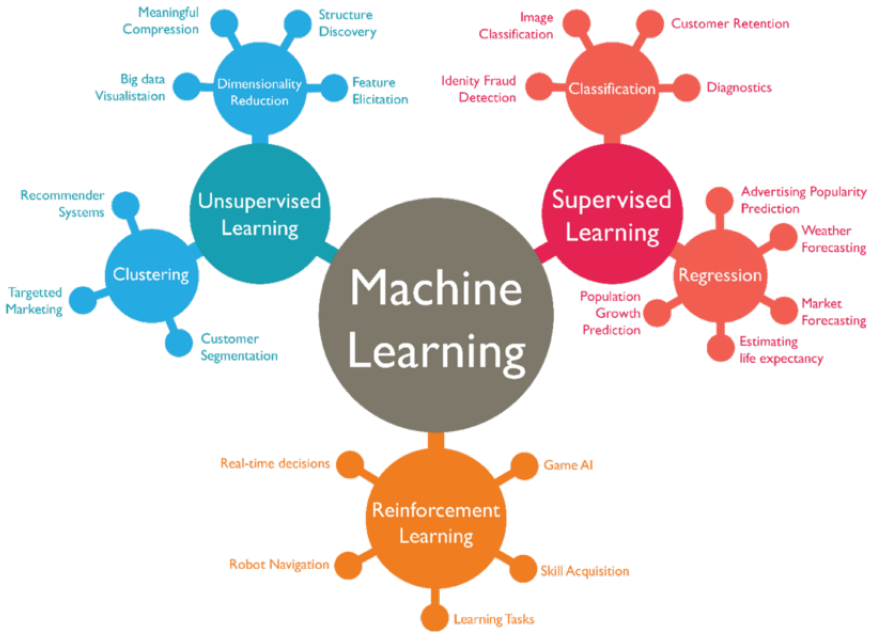
Uma definição básica sobre Machine Learning (Aprendizado de Máquina) é englobar um conjunto de regras com algoritmos e procedimentos que tem como objetivo de extrair informações apartir dos dados e dessas informações tomar uma decisão.

Segundo (GOODFELLOW; BENGIO; COURVILLE, 2016), os algoritmos de Machine Learning podem ser amplamente categorizados pelos tipos de aprendizagem, sitentizando 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 dessas 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 à apresentação da variavél resposta ao algoritmo;
- 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.

A Figura 1 representa a composição da área de *machine learning* com as devidas técnicas utilizadas para aprendizagem.

Figura 1 – Machine Learning e suas aplicações



Fonte: Learning (2022)

2.3 Modelo de Regressão Logística

A regressão logística tem como principal uso, modelar uma variável binária $(0, 1)$, com base em mais variáveis, estas chamadas de variáveis explicativas ou preditoras. E comumentemente 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.

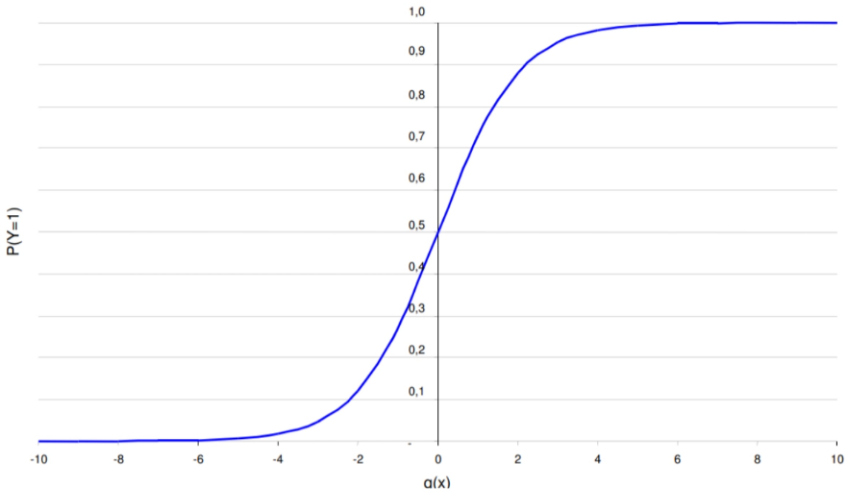
A resposta do modelo dito está conveniente a uma distribuição subjacente, ou seja, segue uma distribuição de Bernoulli. Concordantemente com (BOLFARINE; SANDOVAL, 2010), esta distribuição é um distribuição particular da distribuição Binomial que a função de probabilidade pode ser expressa:

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

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

Mesmo usando a regressão linear para utilizar para obter uma estimativa de probabilidade do resultado, quebramos um pressuposto, pois algumas estimativas podem estar

Figura 2 – Gráfico do modelo de regressão logística.



Fonte: https://aprenderdatascience.com/regressao-logistica/

fora do intervalo $[0, 1]$. Pois está expressão acarreta que é possível para $E(Y|x)$ assumir qualquer valor.

Presuma que o modelo linear tradicional tenha a forma:

$$y_i = \mathbf{x'}_i\beta + \varepsilon_i \tag{2.2}$$

em que $\mathbf{x'}_i = [1, x_{i1}, x_{i2}, \cdots, x_{ik}]$, $\beta' = [\beta_0, \beta_1, \beta_2, \cdots, \beta_k]$ e a variável resposta tem valores entre o intervalo $[0, 1]$.

Assumiremos qeu 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:

$$E(y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i \tag{2.3}$$

o que implica em $E(y_i) = \mathbf{x'_i}\beta = \pi_i$, ou seja, o valor esperado da função resposta é apenas a probabilidade de que a variável resposta assuma o valor 1. Como o nosso valor esperado é π_i e o resultado dicotômico da variável deve ser maior ou igual a zero e menor do que um $(0 \leq E(y_i) = \pi_i \leq 1)$. Pode ser visto na Figura 2 que a curva é em forma de "S"e se assemelhace a um gráfico

2.4 Estimação dos Parâmetros

Para que se tenha um modelo ajustado, é imprescedível que seja feito a estimação dos parâmetros da regressão. Com isso utiliza-se o método de estimação de máxima de

verossimilhança. Este método, a partir de um conjunto e um modelo estatístico, estima os valores dos parâmetros do modelo que mais máxima a probabilidade dos dados observados, ou seja, busca parâmetros que maximizem a função de verossimilhança. Condizente (BOLFARINE; SANDOVAL, 2010), a definição da função de verossimilhança é:

Definição 2.4.1. Definição Sejam $X_1, \ldots, X_n$ uma amostra aleatória de tamanho n da variável aleatória X com função densidade $f(x|\theta)$, com $\theta \in \Theta$ é o espaço paramétrico. A função de verossimilhança de θ compatível com à amostra aleatória observada é dada por

$$L(\theta; x) = \prod_{i=1}^{n} f(x_i|\theta) \tag{2.4}$$

O estimador de máxima verossimilhança de θ é o valor $\theta \in \Theta$ que maximiza a função de verossimilhança $L(\theta; x)$.

Aplicando o logaritmo natural a função de verossimilhança

$$l(\theta; x) = \log L(\theta; x), \tag{2.5}$$

verificamos que o valor de θ

2.5 Interpretação dos Parâmetros

2.6 Testes de Significância

Depois de estimar os coeficientes, é necessário um avaliação da significância das variáveis no modelo, geralmente envolve a formulação e teste de hipótese estatística para determinar se as variáveis independentes estão "significativamente"relacionadas com a variável resposta (JR; LEMESHOW; STURDIVANT, 2013).

2.6.1 Teste da Razão de Verossimilhança (TRV)

2.6.2 Teste de Wald

O teste de Wald é usado para verificar a significância dos coeficientes no modelo. Neste caso, o teste verifica se cada uma das variáveis explicativas apresenta uma relação estatisticamente significativa com a variável resposta, ou seja, o teste compara entre a estimativa de máxima verossimilhança do parâmetro $\hat{\beta}_j$ e a estimativa de seu erro padrão. As hipóteses formuladas do teste são:

$$\begin{aligned} H_0 &: \beta_j = 0 \\ &vs \\ H_1 &: \beta_j \neq 0 \end{aligned} \tag{2.6}$$

A estatística de teste Wald para a regressão logística é

$$W_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

e tem distribuição normal padrão. Se não rejeitarmos H_0, temos que a variável x_j não explica a variável resposta.

Conforme (JR; DONNER, 1977), verificaram o desempenho do teste de Wald e descobriram que se comporta de maneira aberrante, em outras palavras, muitas vezes falha em rejeitar a hipótese nula quando o coeficiente era significativo. Por isso, eles recomendaram qeu o teste da razão de verossimilhança seja o ideal.

2.6.3 Medida da Qualidade do Ajuste do Modelo

2.6.4 Intervalo de Confiança

2.7 Seleção de Variáveis

Um dos problemas primordiais na análise de regressão é selecionar as variáveis para o modelo. Está questão sobre se deve incluir todas variáveis regressoras disponíveis ou apenas um subconjunto destas variáveis ao modelo. Após a decisão, as próximas etapas é a significância e adequação do modelo ajustado devem ser verificadas e a análise de resíduos deve ser conduzida.

i) **Método Forward** (*"passo a frente"*): esse procedimento caracteriza-se por considerar que não há variável no modelo, apenas o intercepto. Será adicionado uma variável de cada vez. A primeira variável selecionada é aquela com maior correlação com a resposta. Etapas se sucedem, quando uma variável por vez pode vir a ser incorporada no modelo se sua estatística F parcial for maior que o ponto crítico, o processo é interrompido quando não houver inclusão (CHARNET et al., 2008).

ii) **Método Backward** (*"passo atrás"*): esse método faz o caminho inverso do método forward, ele incorpora, inicialmente, todas as variáveis e depois, sequencialmente, cada uma pode ser ou não eliminada. A decisão de eliminação da variável é tomada baseando em testes F parciais, que são calculadas por cada variável como se ela fosse última a entrar no modelo (CHARNET et al., 2008).

Tabela 1 – Tabela

	Verdadeiro	
Predito	Positivo	Negativo
Positivo	VP	FP
Negativo	FN	VN

iii) **Método Stepwise** (*"passo a passo"*): este procedimento é uma generalização do procedimento "passo a frente". As etapas de inclusão e retirada da variável do modelo são efetudadas conforme descrito nos procedimento anteriores.A etapa chega ao final quando nenhuma variável é incluída ou descartada (CHARNET et al., 2008).

E

2.7.1 Diagnóstico Resídual

2.7.2 Desempenho do Modelo

A matriz de confusão resume o desempenho de classificação de um modelo em relação a alguns dados de testes. É uma matriz bidimensional, organizada em uma dimensão pela verdadeira classe de um objeto e na outra pela classe qeu o classificador atribui. A Tabela **??** apresenta como uma matriz de confusão

3 Metodologia

Antemão ao processo de modelagem, foi realizado uma análise exploratória dos dados para entender melhor o conjunto de dados e selecionar possíveis variáveis explicativas. Além disso, existe a necessidade de investigar se há dependências entre as variáveis com efeitos nocivos de multicolinearidade ao modelo estimado.

Após, faz se necessário uma análise minuciosa dos outliers.

4 Resultados e Discussões

5 Considerações Finais

Referências

BOLFARINE, H.; SANDOVAL, M. C. *Introdução à inferência estatística*. [S.l.]: SBM, 2010. v. 2ł ed. Citado 2 vezes nas páginas 9 e 11.

CHARNET, R. et al. Análise de modelos de regressão linear com aplicações. *Editora da Unicamp*, 2008. Citado 2 vezes nas páginas 12 e 13.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. [S.l.]: MIT press, 2016. Citado na página 8.

HILBE, J. M. *Practical guide to logistic regression*. [S.l.]: crc Press, 2016. Citado na página 9.

JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. *Applied logistic regression*. [S.l.]: John Wiley & Sons, 2013. v. 398. Citado na página 11.

JR, W. W. H.; DONNER, A. Wald's test as applied to hypotheses in logit analysis. *Journal of the american statistical association*, Taylor & Francis, v. 72, n. 360a, p. 851–853, 1977. Citado na página 12.

LEARNING, A. I. T. M. 2022. Acesso em 26 de novembro de 2022. Disponível em: <https://becominghuman.ai/an-introduction-to-machine-learning-7db04da817c4>. Citado na página 9.

SICSÚ, A. L. *Credit Scoring: desenvolvimento, implantação, acompanhamento*. [S.l.]: Blucher, 2010. Citado na página 8.

Apêndices

APÊNDICE A – DESCRIÇÃO DAS VARIÁVEIS

Variável	Descrição da Variável	Tipo de Variável	Nº de Categorias	Categorias
Sexo	Sexo			
Estado Civl				
Escolaridade				
Idade				
Renda				
Patrimônio				
SM30				
SM60				
SM90				
SM180				
SM360				
Empréstimos				
Capital				
Aplicação				
Limite				
STATUS				

APÊNDICE B – Script em R

```
#############################################################
####                REGRESSÃO LOGISTICA                  ####
#############################################################
library(tidyverse)  #
library(rJava)      #
library(xlsx)       #
library(readxl)     #
library(dlookr)     #


Dados <- read_xlsx(".../Dataset.xlsx", sheet = 2)


cr <- select(Dados, -ID)

### Pré-processamento dos dados ###

# Verificando as variáveis
str(cr)
glimpse(cr)
  # Variáveis
# Sexo:  Masculino,  Feminino;
# Estado Civil: Solteiro, Casado, Divorciado, Viuvo e União Estável;
# Instrução: Não Informado, Sem instrução, 1ž Grau, 2ž Grau, Superior Incompleto, Superior completo e
# Pós-Graduação;
# Inadimplência: 0 = Não, 1 = Sim;


# Converter "SEXO","ESTADO_CIVIL","ESCOLARIDADE" e "STATUS" para fatores.
cr <- mutate_at(cr, vars(SEXO, ESTADO_CIVIL, ESCOLARIDADE, STATUS), as.factor)
```