Research, Society and Development, v. 11, n. 7, e15211729761, 2022
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i7.29761

Modelagem com regressão logística para análise de concessão de crédito

Modeling with logistic regression for credit grant analysis

Modelación con regresión para análisis de otorgamento de crédito

Recebido: 29/04/2022 | Revisado: 06/05/2022 | Aceito: 13/05/2022 | Publicado: 19/05/2022

Rafaella Santos Beserra
ORCID: https://orcid.org/0000-0003-1291-6852
Universidade Federal Rural de Pernambuco, Brasil
E-mail: rafaellasantosbeserra@gmail.com
Nyedja Fialho Morais Barbosa
ORCID: https://orcid.org/0000-0003-1813-320X
Universidade Estadual da Paraíba, Brasil
E-mail: nyedja@servidor.uepb.edu.br
Ana Patricia Bastos Peixoto
ORCID: https://orcid.org/0000-0003-0690-1144
Universidade Estadual da Paraíba, Brasil
E-mail: anapatricia@servidor.uepb.edu.br
Erika Fialho Morais Xavier
ORCID: https://orcid.org/0000-0002-8217-7891
Fundação Oswaldo Cruz, Brasil
E-mail: erika.xavier@fiocruz.br
Jader Silva Jale
ORCID: https://orcid.org/0000-0001-7414-1154
Universidade Federal Rural de Pernambuco, Brasil
E-mail: jsjale1983@gmail.com
Silvio Fernando Alves Xavier Junior
ORCID: https://orcid.org/0000-0002-4832-0711
Universidade Estadual da Paraíba, Brasil
E-mail: profsilviofernando@gmail.com

Resumo
Com o avanço do *Big Data* e o crescente número de grandes massas de dados nas mais diversas áreas de estudo, técnicas de mineração de dados tornam-se cada vez mais necessárias para obtenção de informações estatísticas precisas e robustas. Este estudo teve como objetivo mostrar a eficiência da regressão logística como técnica de mineração de dados na obtenção de um modelo útil e estatisticamente eficaz na análise de clientes para a concessão do crédito bancário. Os dados utilizados são oriundos do repositório *Machine Learning Repository's* da Universidade da California-Irvin UCI, sendo divididos em dois grupos: treinamento e teste. O modelo ajustado foi selecionado com o método *stepwise* no programa R e atendeu as expectativas de qualidade do ajuste, com acurácia de aproximadamente 72% em discriminar clientes adimplentes de inadimplentes, sensibilidade de 87% dos 140 clientes adimplentes o modelo acertou 122 e especificidade de 38%. A curva ROC teve uma área de 0,847, sugerindo um ajuste eficaz.
Palavras-chave: Mineração de dados; Curva ROC; Probabilidade.

Abstract
With the advancement of Big Data and the growing number of large masses of data in the most diverse areas of study, data mining techniques become increasingly necessary to obtain accurate and robust statistical information. This study aimed to show the efficiency of logistic regression as a data mining technique in obtaining a useful and statistically effective model in the analysis of customers for granting bank credit. The data comes from the Machine Learning Repository's at the University of California-Irvin UCI. The database was divided into two groups: training and testing. The adjusted model was selected using the stepwise method in the R program. The model met the expectations of goodness of fit, with an accuracy of approximately 72% in discriminating non-defaulting from non-defaulting customers, sensitivity of 87% of the 140 non-defaulting customers, the model was correct 122 and specificity of 38%. The ROC curve had an area of 0.847, suggesting an effective fit.
Keywords: Data mining; ROC curve; Probability.

Resumen
Con el avance del Big Data y el creciente número de grandes masas de datos en las más diversas áreas de estudio, las técnicas de minería de datos se vuelven cada vez más necesarias para obtener información estadística precisa y robusta. Este estudio tuvo como objetivo mostrar la eficiencia de la regresión logística como técnica de minería de

datos en la obtención de un modelo útil y estadísticamente efectivo en el análisis de clientes para el otorgamiento de crédito bancario. Los datos provienen del repositorio de aprendizaje automático de la Universidad de California-Irvin UCI. La base de datos se dividió en dos grupos: entrenamiento y prueba. El modelo ajustado se seleccionó mediante el método stepwise en el programa R. El modelo cumplió con las expectativas de bondad de ajuste, con una precisión de aproximadamente 72% en discriminar clientes no morosos de no morosos, sensibilidad de 87% de los 140 no morosos. -clientes morosos, el modelo fue correcto 122 y especificidad del 38%. La curva ROC tenía un área de 0.847, sugiriendo un ajuste efectivo.
Palabras clave: Minería de datos; Curva ROC; Probabilidad.

1. Introdução

Com a expansão da economia no Brasil as instituições financeiras viram a necessidade de controle e gerenciamento eficaz do risco de crédito (Lima, 2011). Da Silva (2000), esclarece que o risco é uma expressão que serve para caracterizar os diversos fatores que poderão contribuir para que aquele que concedeu o crédito não receba do devedor na época acordada. A inadimplência é um fator relevante para as instituições financeiras e se não for bem administrada, pode desencadear um desequilíbrio na gestão dos negócios.

A concessão de crédito ao solicitante é sempre uma decisão a ser tomada em condições de incerteza. Para Tavares (2009), a concessão de crédito, nos últimos anos, tem sido um dos principais componentes do crescimento do padrão de vida dos consumidores e do lucro das empresas. A concessão do crédito se dá, a partir do momento em que a instituição se sente segura a ponto de entregar seu capital ao solicitante, no intuito de que este voltará com acréscimo de remuneração. Se o credor puder mensurar o risco de crédito e as chances de o cliente incidir em perdas, terá maior convicção na decisão de crédito e favorecerá a redução dos índices de inadimplência (Marcelino, 2012).

Para avaliar o risco de inadimplência associada ao perfil do cliente são utilizados modelos estatísticos denominados *Credit Scoring* para estimar a probabilidade de não pagamento. Costa (2003) explica que o objetivo da análise de crédito é identificar os riscos a concessão de crédito, visando reduzir a probabilidade de insucesso na operação. O que propiciou os primeiros modelos de *credit scoring* (pontuação de crédito), que criam uma pontuação de crédito no intuito de ordenar ou classificar os clientes frente a probabilidade de pagar o empréstimo concedido, a probabilidade de risco de crédito (Moura, 2018).

Segundo Mays e Lynas (2004), no ano de 2002, o então presidente do Federal Reserve System (FED) Alan Greenspan fez um pronunciamento onde afirma que a tecnologia de *credit scoring* contribui para reduzir drasticamente o custo de avaliação do crédito além de melhorar a consistência, velocidade e acurácia na decisão de crédito. Os modelos utilizados são baseados nas informações do solicitante de crédito, das quais originam variáveis e que por meio de técnicas estatísticas passam a ter pontuações, que combinadas formam scores que mensuram a credibilidade do solicitante de crédito, um ponto de corte no qual procura prever quais serão os possíveis “bons” e “maus” pagadores (Lewis, 1992). A pontuação do *Credit Scoring* pode ser interpretada como a probabilidade de risco de crédito. A pontuação de crédito é um instrumento estatístico desenvolvido para que o analista avalie a probabilidade de que determinado cliente venha a tornar-se inadimplente no futuro (Marcelino, 2012).

Para estimar a probabilidade de inadimplência utilizam-se técnicas estatísticas de análise discriminante. Os escores são geralmente calculados atribuindo-se pesos a variáveis que caracterizam o solicitante e a operação de crédito. A seleção das variáveis e a determinação dos pesos são obtidas por meio de *softwares* estatísticos (Lima, 2011). Este estudo tem como objetivo desenvolver um modelo capaz de prever clientes como "maus pagadores" e "bons pagadores" por meio da Regressão Logística e, assim, analisar o desempenho do modelo na classificação dos clientes para a concessão do crédito bancário.

2. Metodologia

A análise de regressão logística foi desenvolvida em meados do século XIX, com o objetivo de descrever os problemas de crescimento demográfico. Paula (2004) aponta que a regressão logística tornou-se popularmente conhecida a partir dos anos 50, mas obteve destaque na literatura a partir dos trabalhos de Cox & Snell (1989), especialmente entre os estatísticos. Segundo Souza (2006) aspectos teóricos do modelo de regressão logísticas são amplamente discutidos na literatura, destacando-se (Kleinbaum & Klein, 2002), (Agresti, 2018), (Hosmer et al., 2013), (Cox & Hinkley, 1979), entre outros. A regressão Logística é umas das técnicas estatísticas que se caracteriza por descrever uma variável qualitativa binária, associada ao um conjunto de observações de variáveis independente, a fim de permitir formar um modelo de previsão. Esta técnica é amplamente utilizada em diversos tipos de problemas, como Paula (2004) explica: Mesmo quando a resposta não é originalmente binária, alguns pesquisadores têm dicotomizado a variável resposta de modo que a probabilidade de sucesso possa ser modelada por intermédio da regressão logística. Tudo isso se deve, principalmente, à facilidade de interpretação dos parâmetros de um modelo logístico, bem como pela possibilidade do uso desse tipo de metodologia em análise com objetivo de discriminação (Paula, 2004).

Os Modelos lineares generalizados (MLG's) foram propostos por Nelder e Wedderburn (1972) como uma extensão do modelo linear clássico de regressão:

$$Y = Z\beta + \varepsilon, \tag{1}$$

em que Z é uma matriz de dimensão $n \times p$ de especificação do modelo associada a um vetor $\beta = (\beta_1, \ldots, \beta_P)^T$ de parâmetros, e ε é um vetor de erros aleatórios com distribuição supostamente $N_n = (0, \sigma^2 I)$. Dessa forma, os MLG's são modelos para análise de dados que apresentam uma estrutura não linear em um conjunto linear de parâmetros e a variável resposta segue uma distribuição com propriedades muito específicas: a família exponencial. Os MLG's são caracterizados pela seguinte estrutura:

i **Componente aleatória:** Onde a variável dependente e a explicativa são independentes com distribuição pertencente à família exponencial;

ii **Componente sistemática:** Definida através das variáveis explicativas $xi_1, xi_2, \ldots, xi_P$, que produzem um preditor linear N_i, dado por:

$$n_i = \sum_{i=1}^{n} \beta_J x_{ij = x_i^T \beta}, \tag{2}$$

onde β é o vetor de parâmetro desconhecido;

iii **Função de ligação:** Relaciona o componente aleatório ao componente sistemático do modelo linear, e é dada por

$$g(u_i) = n_i. \tag{3}$$

A escolha da função de ligação depende do tipo de resposta e do estudo particular que se pretende fazer. Nos modelos lineares generalizados uma classe bastante importante é o modelo *logit* (caracterizada pela regressão logística), onde a variável resposta pode ser associada as variáveis aleatórias de Bernoulli.

O modelo de Regressão Logística é bem semelhante ao modelo de Regressão Linear; entretanto, na regressão logística a variável dependente é categórica e geralmente dicotômica, e analisa-se a probabilidade de um evento ocorrer ou não ("sucesso" ou "fracasso"). Uma importante característica dos modelos de Regressão Logística é o fato das variáveis dependentes terem distribuição binomial (Souza, 2006). Assim, a variável dependente (Y) é escrita da seguinte forma:

$$Yi = \begin{cases} 1, Sucesso \\ 0, Fracasso \end{cases} \quad (4)$$

As probabilidades de sucesso e fracasso são, respectivamente, $\pi i = P(Y = 1 \mid X = xi)$ e $1 - \pi i = P(Y = 1 \mid X = xi)$. Para descrever a média condicional de Y dado X com a distribuição logística, utiliza-se a notação πi (Hosmer; Lemeshow & Sturdivant, 2013). Como $\pi(x)$ varia entre 0 e 1, uma representação linear para ela sobre todos os valores de x não é adequado, então considera-se a transformação logística de $\pi(x)$ sob a forma linear. A probabilidade do modelo logístico de sucesso é definida como:

$$\pi_i = \pi(x_i) = P(Y = 1 \mid X = x_i) = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_i + ... + \beta_k x_k)} + \varepsilon_i , \quad (5)$$

e a probabilidade de fracasso é definida como:

$$1\text{-}\pi_i = 1 - \pi(x_i) = P(Y = 1 \mid X = x_i) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)} + \varepsilon_i , \quad (6)$$

Pressupostos para ε_i :

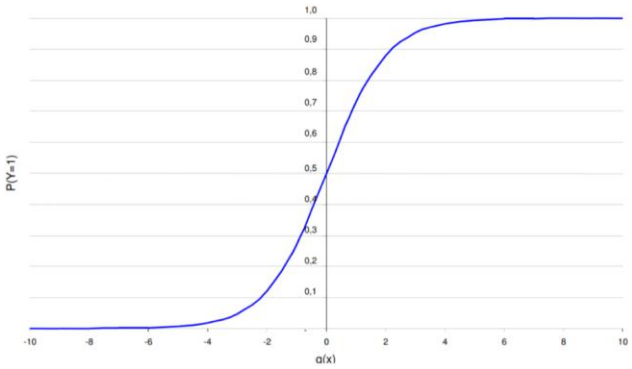
i $E(\varepsilon_i \mid xi) = 0;$

ii $Var(\varepsilon_i \mid xi) = \pi(xi)[1 - \pi(xi)];$

iii $Cov(\varepsilon_i, \varepsilon_l) = 0, se\, i \neq l.$

Os coeficientes $\beta_0, \beta_1, \ldots, \beta_P$ são estimados pelo método da máxima verossimilhança, que consiste em encontrar uma combinação de coeficientes que maximiza a probabilidade de a amostra ter sido observada. Fixando uma combinação de β e variando o valor de x, percebe-se que a curva logística possui um comportamento probabilístico em formato da letra 'S', uma característica da Regressão Logística de acordo com Hosmer et al. (2013), conforme ilustrado na Figura 1.

Figura 1 - O gráfico apresenta a curva do modelo da regressão logística.



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

Quando na Figura 1, $\beta_1 < 0, \pi(xi) \to 0$, quando $\beta_1 > 0, \pi(xi) \to 1$. Para o caso em que $\beta_1 = 0$ significa que a variável Y é independente da variável X.

Interpretação dos parâmetros

Uma medida de fácil interpretação para os coeficientes do modelo de regressão logística é a razão das chances, através da função *odds ratio* – OR. A razão de chance compara a probabilidade de dois eventos ocorrerem, ou seja, a probabilidade de sucesso de um evento ocorrer sobre a probabilidade de fracasso. "Em regressão logística, uma razão de chances é a razão da chance de sucesso sob a condição 2 sobre a chance de sucesso sob a condição 1 nos regressores [...]" (Walpole, 2009). A razão de chance denominada por Ψ é definida por:

$$\psi = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} \tag{7}$$

em que,

$\pi(1) = Variável\ independente\ x = 1.$

$\pi(0) = Variável\ independente\ x = 0.$

O logaritmo da razão de chance (*"log-odds")* é:

$$\ln(\psi) = \ln\left[\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}\right] = g(1) - g(0). \tag{8}$$

Substituindo pela expressão do modelo de regressão logística, a razão de chance é definida como:

$$\psi = \frac{\left(\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}\right)\left(\frac{1}{1+\exp(\beta_0+\beta_1)}\right)}{\left[\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right]\left[\frac{1}{1+\exp(\beta_0)}\right]} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1), \tag{9}$$

e o logaritmo da razão de chances é definida como:

$$\ln(\psi) = \ln[\exp(\beta_1)] = \beta_1. \tag{10}$$

Os valores das razões de chance são constantes, o valor 1 indica que a ocorrência do evento é igualmente provável entre os grupos em estudo. Uma razão de chance maior que 1 indica que o evento tem maior probabilidade de ocorrer na primeira classe, caso seja menor do que 1, indica a probabilidade de menor ocorrência na primeira classe em elação a segunda.

Curva ROC

Em um teste de diagnóstico existem dois tipos de erro que podem ocorrer na decisão: a escolha de uma falha (no sentido de declarar um inadimplente como adimplente) ou a escolha de um falso alarme (declarar uma pessoa adimplente como inadimplente). Para contornar estes tipos de situações, foi necessário desenvolver medidas alternativas de diagnóstico com propriedades mais robustas do que a sensibilidade e a especificidade. A análise ROC (*Receiver Operating Characteristic*) foi a técnica desenvolvida para tornear este tipo de problema (Braga, 2001).

A curva ROC ilustra a relação entre sensibilidade e especificidade, e pode ser utilizada para decidir um bom ponto de corte. Segundo Costa (2013) uma curva ROC é um gráfico de linha que plota a sensibilidade do teste em função da probabilidade de um resultado falso-positivo (1- especificidade) para uma série de diferentes pontos de corte. Quanto mais perto a linha está no canto superior esquerdo do gráfico, mais preciso é o teste. Além disso, o ponto que se encontra mais

próximo desse canto é normalmente escolhido como o corte que maximiza simultaneamente tanto a sensibilidade como a especificidade (Pagano & Gauvreau, 2011).

Resíduo quantílico aleatorizado

A análise dos resíduos compreende a um conjunto de técnicas baseadas na leitura e interpretação dos resíduos, e são utilizadas para auxiliar na verificação da validade das suposições de um modelo de regressão e, consequentemente, analisar a aderência e a adequação da distribuição considerada na formulação do modelo (Pereira, 2019).

Em situações de regressão não normais, como regressão logística, os resíduos, como geralmente definidos, podem estar tão longe de normalidade e de ter variâncias iguais por não ter uso prático (Dunn & Smyth, 1996). Os resíduos quantílicos aleatorizados apresentam distribuição Normal, independente da distribuição da variável resposta e de sua dispersão. O resíduo quantílico aleatorizado, ajustado um MLG, é definido por:

$$r_i^q = \Phi^{-1}\{F(y_i; \mu_i, \phi)\}, \quad (11)$$

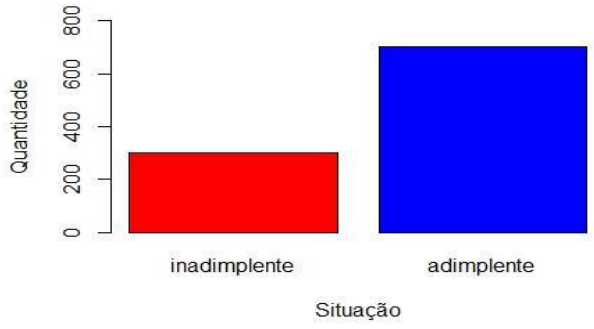
sendo Φ a função de distribuição acumulada da Normal padrão. Se os parâmetros do modelo são consistentemente estimados, então r_i^q converge para uma distribuição Normal padrão. Se Y é discreta, então um recurso de aleatorização é aplicado de tal forma que, também nesse caso, se os parâmetros do modelo são consistentemente estimados, então r_i^q converge para uma distribuição Normal padrão.

3. Resultados e Discussão

Nesta seção são apresentados os resultados obtidos através do software R (Team, 2021), onde foi aplicado o modelo de regressão logística na base de dados *German Credit Data*, disponibilizada pela Universidade da California-Irvin UCI em seu repositório Machine Learning Repository's. O banco de dados possui informações financeiras e pessoais de 1000 clientes de um cartão de crédito da Alemanha, sendo composto de vinte variáveis explicativas e uma variável dependente categórica binária que está denominada como risco. As pontuações de crédito estão convertidas para DM.

Com a análise exploratória dos dados observa-se (Figura 2) que a porcentagem de inadimplentes representa 30% da amostra, correspondendo a 300 clientes, enquanto a porcentagem de clientes que cumpriram o contrato de crédito representa 70%, ou seja, 700 indivíduos.

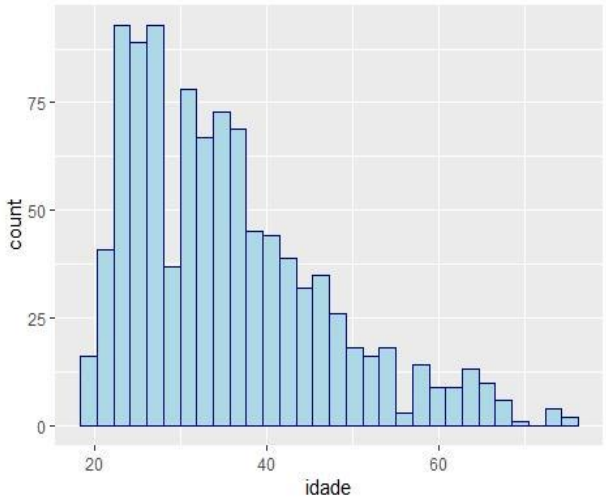
Figura 2 - Gráfico com a classificação dos clientes por situação.



Fonte: Elaboração dos autores (2022).

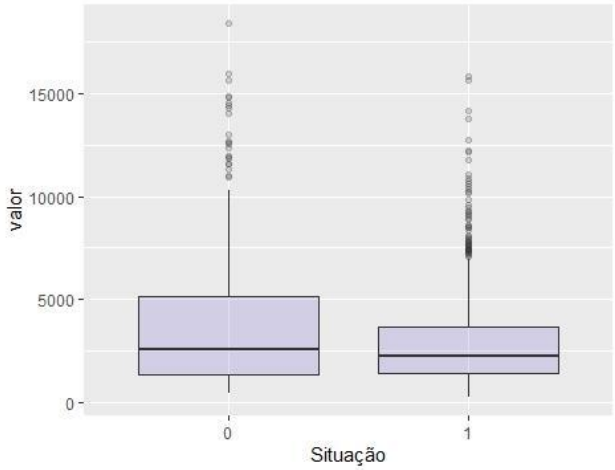
A idade dos clientes que compunham a amostra variou entre 19 e 75 anos, com média de 35 anos, onde 95% dos indivíduos analisados (Figura 3) apresentaram idade entre 19 e 42 anos. Com relação aos valores de crédito concedidos (Figura 4), observa-se que o valor mínimo de crédito concedido foi de 250 e o valor o máximo 18428, em que o 0 representa os clientes inadimplentes e o 1 os clientes adimplentes. Verificou-se também que há uma variância menor no valor do crédito concedido aos clientes adimplentes em relação aos clientes inadimplentes e há alguns valores discrepantes em relação aos valores de crédito.

Figura 3 - Histograma de frequência de idades dos clientes.



Fonte: Elaboração dos autores (2022).

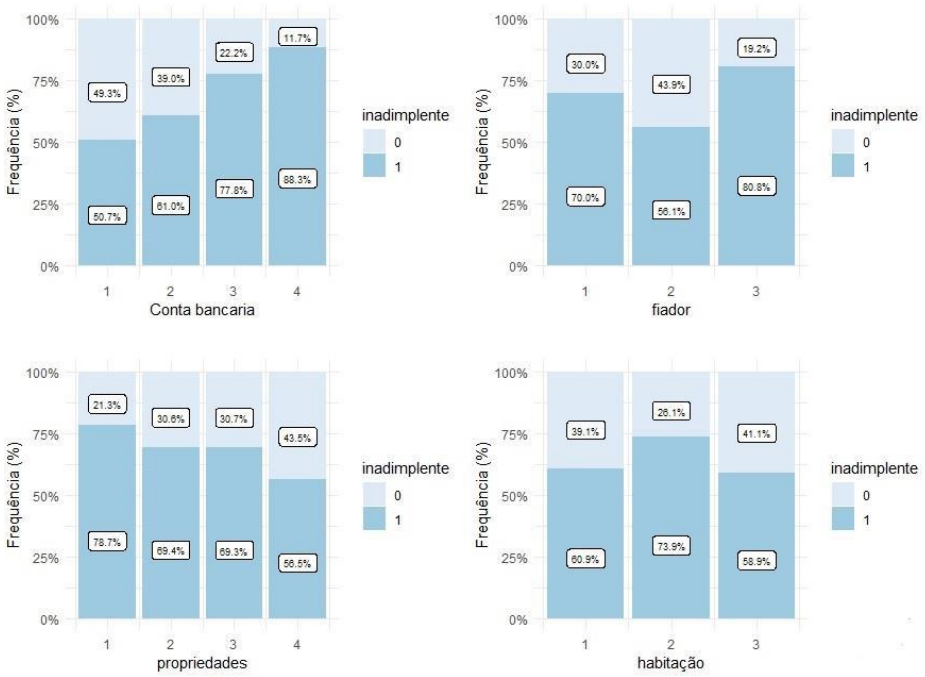
Figura 4 - Boxplot do valor do crédito em relação a situação dos clientes.



Fonte: Elaboração dos autores (2022).

Na Figura 5 observa-se a relação da variável resposta risco em relação a 4 variáveis explicativas: conta bancária, fiador, propriedades e habitação. Do perfil dos clientes adimplentes analisados, em relação às variáveis mencionadas, observa-se que 88% dos indivíduos possui conta > 200 DM, 80% possui fiador, a maioria não possui propriedades e moram de aluguel.

Figura 5 - Gráfico de risco associada a outras variáveis explicativas: conta bancária, fiador, propriedades e habitação.



Fonte: Elaboração dos autores (2022).

Regressão logística

Inicialmente, para a construção do modelo de regressão logística dividiu-se os dados de forma aleatória, correspondendo a 80% da amostra para treinamento (criando um modelo preditivo) e 20% para teste (com o objetivo de avaliar o modelo), onde a amostra de treinamento possui 800 observações divididas em 560 bons pagadores (adimplentes) e 240 maus pagadores (inadimplentes). Para a seleção do melhor modelo logístico, utilizou-se o método de seleção de covariáveis *stepwise (both directions*), o qual utiliza o Critério de Informação de Akaike (AIC) na combinação das variáveis dos diversos modelos simulados para selecionar o modelo mais ajustado de modo que quanto menor o AIC, melhor o ajuste do modelo. Após a aplicação do método *stepwise* somente 15, das 20 variáveis, foram significativas ao modelo. A Tabela 1 apresenta as variáveis selecionadas e suas respectivas estimativas do coeficiente beta, considerando nível de 5% de significância.

Tabela 1 – Variáveis significativas do modelo pelo método *Stepwise*.

Variável	Coeficiente estimado	Erro padrão	Estatística Z	Pr(>|z|)
(Intercept)	-7,070e-01	1,075e+00	-0,658	0,510833
status2	8,070e-01	2,486e-01	3,246	0,001170**
status3	1,361e+00	4,372e-01	3,112	0,001858**
status4	1,970e+00	2,669e-01	7,381	1,57e-13***
Duração	-2,818e-02	1,058e-02	-2,663	0,007733**
cred_hist4	1,317e+00	4,807e-01	2,739	0,006157**
objetivos1	1,728e+00	4,150e-01	4,163	3,14e-05***
objetivos2	8,117e-01	3,004e-01	2,702	0,006888**
objetivos3	9,581e-01	2,891e-01	3,314	0,000918***
Valor	-1,091e-04	4,865e-05	-2,243	0,024926*
poupanca4	1,373e+00	6,053e-01	2,268	0,023311*
poupanca5	9,505e-01	2,981e-01	3,188	0,001432**
taxa4	-9,075e-01	3,439e-01	-2,639	0,008321**
status_sex2	8,817e-01	4,343e-01	2,030	0,042363*
status_sex3	1,498e+00	4,272e-01	3,506	0,000454***
buerge2	-9,777e-01	4,768e-01	-2,051	0,040306*
residence2	-7,060e-01	3,239e-01	-2,180	0,029268*
propriedade4	-1,203e+00	4,967e-01	-2,422	0,015447*
Idade	2,459e-02	9,945e-03	2,473	0,013415*
habitacao2	5,754e-01	2,638e-01	2,181	0,029209*
gastarb2	-1,718e+00	7,625e-01	-2,253	0,024238*

Fonte: Elaboração dos autores (2022).

Os coeficientes das estimativas das variáveis (Tabela 1) que apresentaram valor positivo revelam que a respectiva variável causa um aumento na probabilidade de o cliente não ser inadimplente. Neste caso, é possível verificar que, de acordo com as informações obtidas, as variáveis que contribuem para o aumento da probabilidade de encontrar um bom pagador foram:

i Conta corrente $0 \leq ... < 200$ DM e conta corrente ≥ 200 DM
ii Histórico de cumprimento de contrato (todos os créditos nesse banco reembolsados devidamente).
iii Objetivos (Carro novo, carro usado e imóveis).
iv Poupança de $200 \leq ... < 500$ e poupança de > 1000 DM.
v Status (masculino casado).
vi Idade
vii Habitação (aluguel).

Em contrapartida, as variáveis com o coeficiente negativo indicam a redução na probabilidade do cliente se tornar um bom pagador. Estes indicam as características dos clientes que aumentam o risco de inadimplência, sendo estes:

i Duração em meses do crédito.
ii Valor do crédito.
iii Taxa de juros < 20
iv Co-requerente
v Período que o devedor reside na residência atual (1 a 4 anos)
vi Propriedade mais cara do devedor imóveis
vii Não é estrangeiro.

Na Tabela 2 encontra-se a razão de chances, que representa a chance de um evento ocorrer (ser adimplente) dada a presença do preditor X, bem como os intervalos de confiança ao nível de 95% para os parâmetros do modelo, onde com base na estatística de Wald, se o intervalo de confiança incluir o valor de 1, implica que não existe diferença entre os grupos estudados.

Tabela 2 – Variáveis estudadas, Razão de Chances (OR) e intervalos de confiança.

Variável	OR	2.5%	97.5%
(Intercepto)	0,493	-2,814	1,400
Status2	2,241	0,320	1,294
Status3	3,898	0,504	2,217
Status4	7,170	1,447	2,493
Duração	0,972	-0,049	-0,007
Cred_hist4	3,732	0,375	2,259
Objetivos1	5,627	0,914	2,541
Objetivos2	2,252	0,223	1,401
Objetivos3	2,607	0,392	1,525
Valor	1,000	0,000	0,000
Poupança4	3,947	0,187	2,559
Poupança5	2,587	0,366	1,535
Taxa4	0,404	-1,582	-0,233
Status_sex2	2,415	0,030	1,733
Status_sex3	4,472	0,661	2,335
buerge2	0,376	-1,912	-0,043
residence2	0,494	-1,341	-0,071
Propriedade4	0,300	-2,176	-0,229
Idade	1,025	0,005	0,044
Habitação2	1,778	0,058	1,092
Gastarb2	0,179	-3,213	-0,224

Fonte: Elaboração dos autores (2022).

O resultado obtido (Tabela 2) evidencia que se um cliente tiver uma conta corrente ≥ 200 DM tem 7 vezes mais chance de não se tornar inadimplente do que os clientes que não possuem conta corrente, observa-se também que os clientes que utilizaram o crédito para a compra de um carro novo têm 5 vezes mais chance de não se tornar inadimplente do que os outros clientes.

Avaliação do modelo

A partir do modelo logístico foi possível realizar predições das probabilidades utilizando os dados de teste para avaliar o desempenho do modelo de regressão logística, por meio da criação de uma matriz de confusão (Tabela 3).

Tabela 3 – Valores preditos para os clientes adimplentes e inadimplentes.

Valores Reais		Inadimplente	Adimplente
Predito	Inadimplente	23	18
	Adimplente	37	122

Fonte: Elaboração dos autores (2022).

Dos resultados dos testes de diagnóstico do modelo (Tabela 4), observa-se 72% de acurácia, o que indica alta precisão dos resultados e baixo risco de erro; sensibilidade de 87%, onde dos 140 clientes adimplentes o modelo foi capaz de detectar

122; especificidade de 38% de modo que, dos 60 inadimplentes o modelo acertou 23; e a probabilidade de o cliente ser adimplente, dado que o modelo a classificou como adimplente é de 76%.

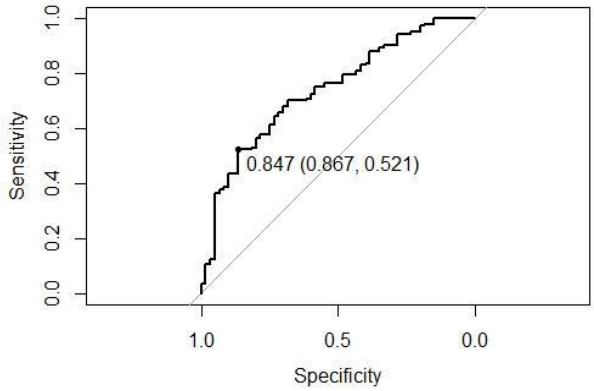
Tabela 4 – Resultados percentuais dos testes de diagnóstico do modelo.

Indicadores	**Valores**
Acurácia	72%
Sensibilidade	87%
Especificidade	38%
Valor predito positivo	76 %
Valor predito negativo	56%

Fonte: Elaboração dos autores (2022).

A curva ROC para o modelo (Figura 6) apresentou o ganho em sensibilidade à medida que a taxa de falso-positivo (1-especificidade) aumenta, revelando área sob a curva de 0,847, o que sugere que o modelo é eficiente em discriminar clientes que são adimplentes dos inadimplentes.

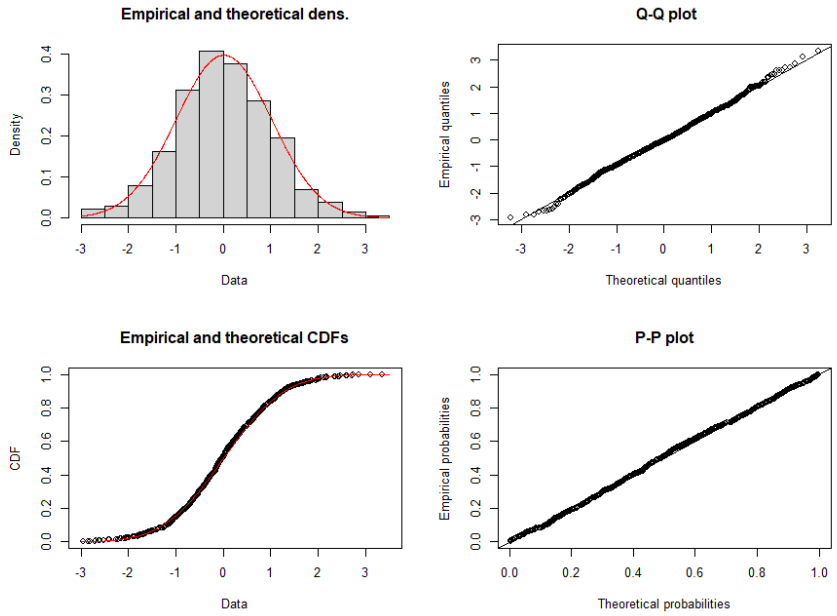
Figura 6 - Curva ROC para o modelo logístico.



Fonte: Elaboração dos autores (2022).

Como a falta de normalidade dos resíduos é particularmente notável na modelagem de dados discretos, aplicou-se os resíduos quantílicos aleatorizados para ajustar o modelo à uma distribuição Normal padrão (Figura 7), o que pode ser comprovado por meio da análise dos gráficos de resíduos.

Figura 7 - Gráficos da análise de resíduos.



Fonte: Elaboração dos autores (2022).

4. Conclusão

Este trabalho apresentou uma aplicação do modelo de regressão logística em um conjunto de dados com características de alguns clientes de um cartão de crédito, com o objetivo de analisar qual perfil de cliente aumenta a probabilidade de não se tornar inadimplente. Os dados foram divididos em treinamentos e teste onde, por meio do método stepwise foram selecionadas as variáveis significativas ao modelo, criando um modelo com suas estimativas. De acordo com o cálculo da razão de chance, observou-se que um cliente que possui uma conta corrente > 200 DM tem 7 vezes chance de não se tornar inadimplente dos que não possui. Em seguida estimou-se previsões com os dados teste para validar o desempenho do modelo, o que apresentou um bom desempenho, resultando numa taxa de acerto de 72% e sensibilidade de 87%, além da curva ROC sugerir que o modelo é bastante eficiente em descriminar clientes que são adimplentes dos inadimplentes. Como trabalhos futuros pretende-se aplicar o estudo com outras técnicas de *machine learning* visando combater prejuízos em empresas, bancos e instituições.

Agradecimentos

Os autores agradecem as contribuições dos revisores para a melhoria na qualidade do artigo.

Referências

Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.

Braga, A. C. (2001). Curvas ROC: aspectos funcionais e aplicações.

Costa, R. R. (2003). *Análise empresarial avançada para crédito*. Qualitymark Editora Ltda.

Costa, R. S. D. (2013). Teste de diagnóstico baseado em análise de regressão logística.

Cox, D. R., & Hinkley, D. V. (1979). *Theoretical statistics*. CRC Press.

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236-244.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Kleinbaum, D. G., & Klein, M. (2002). Analysis of matched data using logistic regression. *Logistic regression: A self-learning text*, 227-265.

Lewis, E. M. (1992). *An introduction to credit scoring*. Fair, Isaac and Company.

Lima, F. A. P. D. (2011). *Práticas em gestão de sistemas de credit scoring e portfólio de crédito em instituições financeiras brasileiras* (Tese de Doutorado).

Marcelino, J. A. (2012). Credit scoring: uma ferramenta para análise de crédito em uma instituição de microcrédito produtivo e orientado.

Mays, F. E., & Lynas, N. (2004). *Credit scoring for risk managers: The handbook for lenders*. Thomson/South-Western.

Moura, G. M. (2018). Regressão Logística aplicada a análise de risco de crédito. (Monografia, Universidade Federal do Rio Grande).

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370-384.

Pagano, M., & Gauvreau, K. (2011). Princípios de bioestatística. In *Princípios de bioestatística* (pp. xv-506).

Paula, G. A. (2004). *Modelos de regressão: com apoio computacional* (pp. 28-55). IME-USP.

Pereira, M. A. A. (2019). Modelos não lineares assimétricos com efeitos mistos.

Da Silva, J. P. (2000). *Gestão e análise de risco de crédito*. Editora Atlas SA.

Souza, É. C. D. (2006). *Análise de influência local no modelo de regressão logística* (Tese de Doutorado, Universidade de São Paulo).

Tavares, M.D.C. (2009). A crise financeira atual. *Paper Itamaraty*, *30*(04).

Team, R. C. (2021). R: A language and environment for statistical computing (R Version 4.0. 3, R Foundation for Statistical Computing, Vienna, Austria, 2020).: https://www.r-project.org/.

Walpole, R. E. (2009). *Probabilidade & Estatística para engenharias e ciências.* Pearson Prentice Hall.