



**UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE BACHARELADO EM ESTATÍSTICA**

RAFAELLA SANTOS BESERRA

**MODELAGEM COM REGRESSÃO LOGÍSTICA PARA ANÁLISE
DE CONCESSÃO DE CRÉDITO**

**CAMPINA GRANDE - PB
2021**

RAFAELLA SANTOS BESERRA

**MODELAGEM COM REGRESSÃO LOGÍSTICA PARA ANÁLISE DE
CONCESSÃO DE CRÉDITO**

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. Silvio Fernando Alvez Xavier Júnior

Coorientador: Profa. Dra. Érika Fialho Morais Xavier

**CAMPINA GRANDE - PB
2021**

É expressamente proibido a comercialização deste documento, tanto na forma impressa como eletrônica. Sua reprodução total ou parcial é permitida exclusivamente para fins acadêmicos e científicos, desde que na reprodução figure a identificação do autor, título, instituição e ano do trabalho.

B554m Beserra, Rafaella Santos.
Modelagem com regressão logística para análise de concessão de crédito [manuscrito] / Rafaella Santos Beserra. - 2021.
34 p. : il. colorido.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2021.
"Orientação : Prof. Dr. Sílvio Fernando Alves Xavier Júnior, Departamento de Estatística - CCT."

1. Regressão logística. 2. Mineração de dados. 3. Curva ROC. 4. Probabilidade. I. Título

21. ed. CDD 519.5

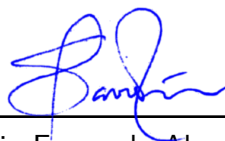
RAFAELLA SANTOS BESERRA

MODELAGEM COM REGRESSÃO LOGÍSTICA PARA ANÁLISE DE CONCESSÃO DE
CRÉDITO

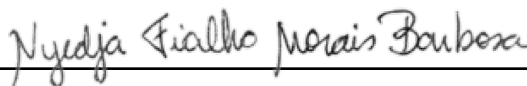
Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Trabalho aprovado em 07 de Outubro de 2021.

BANCA EXAMINADORA



Prof. Dr. Silvio Fernando Alvez Xavier Júnior
Universidade Estadual da Paraíba (UEPB)



Profa. Ma. Nyedja Fialho Morais Barbosa
Universidade Estadual da Paraíba (UEPB)



Profa. Dr. Ana Patricia Basto Peixoto de
Oliveira
Universidade Estadual da Paraíba (UEPB)

À minha família, em especial aos meus pais, Elizama e Robmilson

À minha irmã, Gabriella

À minha avó, Maria Dalva .

AGRADECIMENTOS

Primeiramente agradeço a Deus pela vida, por ser minha força, por nunca ter desistido de mim, todas as minhas realizações são pra honra e glória do Seu nome.

Aos meus pais que são a razão do meu viver, por todo apoio e sacrifícios que fizeram para proporcionar uma educação de qualidade que eles não tiveram oportunidades para mim e minha irmã, por me apoiar em todas as minhas decisões e acreditar que eu sou capaz, tudo o que eu sou e serei, devo a eles. A minha avó que é a minha maior apoiadora, que faz tudo o que for necessário por mim. E a minha irmã por sempre estar do meu lado. Aos meus amigos da vida, que sempre me incentivaram e nunca me deixaram desistir dos meus sonhos: Ana carolina, Nelson, Rayane, Silas, Matheus e Wendel.

Aos amigos que ganhei na universidade. Foram 5 longos anos de alegrias e choro, de vitórias e fracassos, e o mais importante a nossa união, a gente conseguiu juntos, muitos ficaram pelo caminho, mas a gente venceu: Álvaro, Eduardo, Maria Beatriz, Viviane Costa, Gilmar e Wylliam.

Aos meus orientadores professor Silvio Fernando e a professora Erika Fialho, por toda dedicação e apoio ao meu trabalho não me deixaram nem um momento sem apoio, pra mim são uns exemplos de professores, que admiro muito.

E a todos os professores do departamento de estatística por toda dedicação e ensinamento passados.

“Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito. Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes.”
(Martin Luther King)

RESUMO

Devido ao aumento do volume de dados ao longo dos anos, as técnicas de mineração de dados se tornam cada vez mais relevantes para obtenção de informações. Portanto, esse trabalho tem como objetivo mostrar a eficiência de uma técnica de mineração de dados que é a regressão logística, onde a característica principal é a abordagem das variáveis resposta categóricas nas suas formas binárias e múltiplas. Ao longo do trabalho são discutidos métodos para estimação do modelo de regressão e testes para avaliação desse modelo gerado. Os dados são oriundos do repositório MachineLearning Repository's da Universidade da Califórnia-Irvin UCI. O programa usado para modelar o modelo logístico foi *software R*. Os dados foram divididos em treinamento e teste, onde o modelo ajustado foi selecionado através do método stepwise com nível de significância de < 0.05 . O modelo atendeu as expectativas de qualidade do ajuste, as métricas do modelo foram uma acurácia de aproximadamente 72 % em discriminar clientes adimplentes de inadimplentes, sensibilidade de 87% dos 140 clientes adimplente o modelo acertou 122 e especificidade de 38%. Outra ferramenta utilizada para avaliar o modelo é a curva ROC, que teve uma área de 0,847 sugerindo que o modelo é bastante eficiente.

Palavras-chaves: Mineração de dados. Curva ROC. Probabilidade.

ABSTRACT

The volume of data increases over the years, as data mining techniques become increasingly relevant for obtaining information. Therefore, this work aims to show the efficiency of a data mining technique that is a logistic regression, where the main characteristic is the approach of categorical response variables in their binary and multiple forms. Throughout the work, methods for estimating the regression model and tests for evaluating this generated model are discussed. The data comes from the Machine Learning Repository's repository at the University of California-Irvine UCI. The program used to model the logistic model was *software R*. Data were divided into training and testing, where the model conducted was selected using the stepwise method with a significance level of <0.05 . The model met the Goodness-of-fit expectations, as model metrics were approximately 72 % accuracy in discriminate non-defaulting customers from non-defaulting customers, sensitivity of 87 % of 140 defaulting customers, the model was correct 122 and specificity of 38 %. Another tool to use to evaluate the model is a ROC curve, which had an area of 0,847 suggesting that the model is quite efficient.

Keywords: Data mining. ROC Curve. Probability.

LISTA DE ILUSTRAÇÕES

Figura 1 – O grafico apresenta a curva do modelo da regressão logística.	15
Figura 2 – Gráfico com a classificação dos clientes por situação.	24
Figura 3 – Histograma de frequência das idades dos clientes.	24
Figura 4 – Boxplot do valor do crédito em relação a situação dos clientes.	25
Figura 5 – Gráfico de risco associada a outras variáveis explicativas: conta bancária, fiador, propriedades e habitação.	26
Figura 6 – Curva ROC do modelo logístico.	30
Figura 7 – Gráficos da análise de resíduos.	31

LISTA DE TABELAS

Tabela 1 – Descrição dos dados	22
Tabela 2 – Descrição dos dados	23
Tabela 3 – Estimativas e testes associados ao modelo de regressão logística selecionado.	27
Tabela 4 – Razão de chances (OR - odds ratio em inglês) e intervalo de confiança .	29
Tabela 5 – Matriz de confusão: Valores reais X valores preditos.	29
Tabela 6 – Testes de diagnóstico do modelo.	30

SUMÁRIO

1	INTRODUÇÃO	11
2	FUNDAMENTAÇÃO TEÓRICA	13
2.1	Marco Histórico	13
2.2	Modelos lineares generalizados	13
2.3	Regressão logística	14
2.4	Estimação dos parâmetros	16
2.5	Interpretação dos parâmetros	17
2.6	Seleção do modelo	18
2.7	Métodos de verificação da qualidade do ajuste	19
2.7.1	<i>Teste da razão de verossimilhança (TRV)</i>	19
2.7.2	<i>Teste de Wald</i>	20
2.8	Diagnóstico do Modelo.	20
2.8.1	<i>Curva ROC</i>	20
2.8.2	<i>Resíduo quantílico aleatorizado</i>	21
3	APLICAÇÃO	22
3.1	Descrição dos dados e análise exploratória.	22
3.2	Regressão logística	26
3.3	Avaliação do modelo.	29
4	CONCLUSÃO	32
	REFERÊNCIAS	33

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). 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 sente-se 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 a fim 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 de Credit Scoring são um processo baseado 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. O score é a mensuração da credibilidade 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). Dentre as técnicas estatísticas mais utilizadas na modelagem de credit scoring destacam-se: Regressão Linear, Análise Discriminante, Redes Bayesianas, Redes Neurais, Regressão Logística e Análise de Sobrevivência (HARRISON; ANSELL, 2002). Portanto, esse trabalho 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 FUNDAMENTAÇÃO TEÓRICA

Nesta seção são apresentadas os principais conceitos da análise do modelo de regressão logística.

2.1 Marco Histórico

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, 2010), (AGRESTI, 2018), (JR; LEMESHOW; STURDIVANT, 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. A regressão logística 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 e também pela possibilidade do uso desse tipo de metodologia em análise com objetivo de discriminação (PAULA, 2004).

A regressão logística é aplicada em várias áreas como a financeira, ambiental e epidemiológica e se destaca como uma forte ferramenta de análise de dados, capaz de estimar a probabilidade de ocorrência de um evento em estudo.

2.2 Modelos lineares generalizados

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

$$Y = Z\beta + \epsilon, \quad (2.1)$$

em que Z é uma matriz de dimensão $n \times p$ de especificação do modelo associada a um vector $\beta = (\beta_1, \dots, \beta_p)^T$ de parâmetros, e ϵ é um vetor de erros aleatórios com distribuição que se supõe $N_n = (0, \sigma^2 I)$.

Portanto 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 á família exponencial.
- ii **Componente sistemática:** É definida pelas variáveis explicativas $x_{i_1}, x_{i_2}, \dots, x_{i_p}$ que produzem um preditor linear N_i , dado por:

$$n_i = \sum_{j=1}^p \beta_j x_{ij} = x_i^T \beta, \quad (2.2)$$

onde β é o vetor de parâmetro desconhecido.

- iii **Função de ligação:** É a função que relaciona o componente aleatório ao componente sistemático do modelo linear, dada por:

$$g(u_i) = n_i. \quad (2.3)$$

A escolha da função de ligação depende do tipo de resposta e do estudo particular que se está a fazer. Em MLG há duas classes importantes, e uma delas é o modelo logit (caracterizada pela regressão logística), onde a variável resposta pode ser associada as variáveis aleatórias de Bernoulli.

2.3 Regressão logística

O modelo de Regressão Logística é bem semelhante ao modelo de Regressão Linear. Enquanto no modelo de regressão linear temos uma variável dependente contínua, na regressão logística a variável dependente é categórica e geralmente dicotômica, ou seja, só há duas respostas (sucesso ou fracasso). A regressão logística analisa a probabilidade de um evento ocorrer ou não, uma importante característica dos modelos de Regressão Logística é o fato das variáveis dependentes terem distribuição binomial (SOUZA, 2006). A variável dependente (Y) é escrita da seguinte forma:

$$Y_i = \begin{cases} 1, & \text{Sucesso} \\ 0, & \text{Fracasso} \end{cases} \quad (2.4)$$

Assim obtemos as probabilidades, como $\pi_i = P(Y = 1|X = x_i)$ para sucesso $1 - \pi_i = P(Y = 0|X = x_i)$ para fracasso. Para descrever a média condicional de Y dado X com a distribuição logística, é utilizada a notação π_i (JR; 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:

$$\pi_i = \pi(x_i) = P(Y = 1|X = x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_i + \dots + \beta_k x_k)} + \epsilon_i. \quad (2.5)$$

E a probabilidade de fracasso é definida como:

$$1 - \pi_i = 1 - \pi(x_i) = P(Y = 0|X = x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_i + \dots + \beta_k x_k)} + \epsilon_i. \quad (2.6)$$

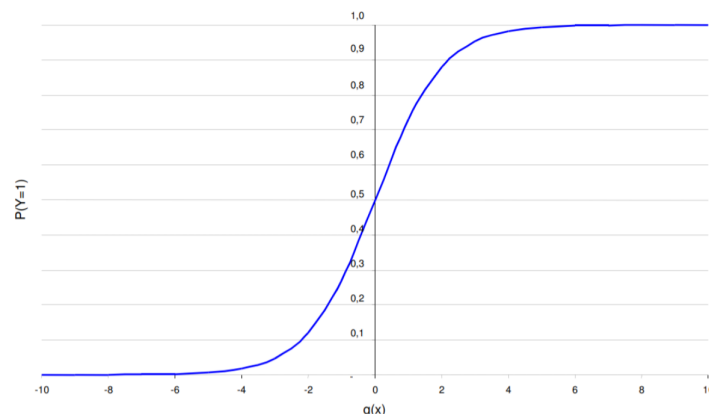
Tem-se os seguintes pressupostos para ϵ_i :

- i) $E(\epsilon_i|x_i) = 0$.
- ii) $Var(\epsilon_i|x_i) = \pi(x_i)[1 - \pi(x_i)]$.
- iii) $Cov(\epsilon_i, \epsilon_l) = 0$, se $i \neq l$

Os coeficientes $\beta_0, \beta_1, \dots, \beta_P$ são estimados a partir do conjunto de dados pelo método da máxima verossimilhança, em que encontra uma combinação de coeficientes que maximiza a probabilidade da amostra ter sido observada.

Ao fixarmos uma combinação de β e variarmos 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 Jr, Lemeshow e Sturdivant (2013), conforme ilustrado na Figura 1

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



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

2.4 Estimação dos parâmetros

A estimação dos parâmetros da regressão logística pode ser feita com o método de máxima verossimilhança, maximizando a função de probabilidade. O método de máxima verossimilhança consiste em encontrar o valor do parâmetro que maximiza a função de verossimilhança, apresentando um estimador razoável para o parâmetro avaliado (SOUZA, 2006). A função de verossimilhança é dada por

$$L(\beta) = \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{(1-y_i)}. \quad (2.7)$$

E o logaritmo da função de verossimilhança é

$$l(\beta) = \sum_{i=1}^n [y_i \ln \pi(x_i) + (1 - y_i) \ln (1 - \pi(x_i))]. \quad (2.8)$$

Então para encontrar o valor de β que maximiza o logaritmo da função de verossimilhança, deriva-se $l(\beta)$ em relação aos parâmetros de regressão e utiliza-se o algoritmo de otimização de Newton- Raphson, ou seja,

$$\frac{\partial l(\beta)}{\partial \beta_0} = \sum_{i=1}^n y_i - \sum_{i=1}^n \pi(x_i) = 0 \quad \frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^n x_{ij} y_i - \sum_{i=1}^n x_{ij} \pi(x_i) = 0, \quad (2.9)$$

para $j \in 1, \dots, p$

Com base nas propriedades de somatório, segue-se que o estimador pelo método da máxima verossimilhança $\hat{\beta}$, é a solução das equações de verossimilhança.

$$\sum_{i=1}^n (y_i - \pi(x_i)) = 0 \quad \sum_{i=1}^n x_{ij} (y_i - \pi(x_i)) = 0, \text{ para } j \in 1, \dots, p. \quad (2.10)$$

Dessa forma, o vetor escore $U(\beta)$, pode ser escrito como

$$U(\beta) = X^T y - X^T \pi = X^T (y - \pi). \quad (2.11)$$

A matriz de informação de Fisher é dada por:

$$I(\beta) = E \left[\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} \right] = X^T Q X, \quad (2.12)$$

onde $Q = \text{diag}[(x)(1(x))]$, X é a matriz de dados e sua inversa $[I(\beta)]^{-1}$ é a matriz de variâncias e covariâncias das estimativas de máxima verossimilhança dos parâmetros (SILVA, 1992).

De acordo com Souza (2006), a solução para as equações de verossimilhança é obtida usando o método iterativo de Newton-Raphson. O conjunto de equações iterativas é dado por:

$$\beta^{(t+1)} = \beta^{(t)} + [I(\beta^{(t)})]^{-1} U(\beta^{(t)}); t = 0, 1, 2, \dots = \beta^{(t)} + [X^T Q^{(t)} X]^{-1} X^T (y - \pi^{(t)}), \quad (2.13)$$

sendo que $\beta^{(t)}$ e $\beta^{(t+1)}$ são vetores de parâmetros estimados nos passos t e $t + 1$, respectivamente.

2.5 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:

$$\Psi = \frac{\pi(1) / [1 - \pi(1)]}{\pi(0) / [1 - \pi(0)]}, \quad (2.14)$$

onde

$$\pi(1) = \text{Variável independente } x = 1.$$

$$\pi(0) = \text{Variável independente } x = 0.$$

O logaritmo da razão de chance (“log-odds”) é:

$$\ln(\Psi) = \ln \left[\frac{\pi(1) / [1 - \pi(1)]}{\pi(0) / [1 - \pi(0)]} \right] = g(1) - g(0). \quad (2.15)$$

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

$$\Psi = \frac{\left(\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} \right) / \left(\frac{1}{1 + \exp(\beta_0 + \beta_1)} \right)}{\left[\frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \right] / \left[\frac{1}{1 + \exp(\beta_0)} \right]} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1). \quad (2.16)$$

E o logaritmo da razão de chances é definido como:

$$\ln(\Psi) = \ln[\exp(\beta_1)] = \beta_1. \quad (2.17)$$

Os valores das razões de chance são constantes, onde 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, já se a razão de chance for menor que 1 indica a probabilidade de menor ocorrência na primeira classe em relação a segunda.

2.6 Seleção do modelo

Um dos problemas importantes na análise de regressão é selecionar as variáveis para o modelo. Portanto existem vários procedimentos para a seleção de modelos de regressão. Alguns serão descritos a seguir.

i Método Forward

Esse procedimento parte da suposição de que não há variável no modelo, apenas o intercepto. A ideia do método é adicionar uma variável de cada vez. A primeira variável selecionada é aquela com maior correlação com a resposta. O processo é repetido, ou seja, variável com maior correlação parcial com a variável resposta é adicionada no modelo se sua estatística F parcial for maior que o ponto crítico, até que não seja incluída mais nenhuma variável explicativa no modelo (FERREIRA, 2012).

ii. Método Backward

O método Backward faz o caminho oposto do método forward, ele incorpora inicialmente todas as variáveis e depois, por etapas, cada uma pode ser ou não eliminada. A decisão de retirada da variável é tomada baseando-se em testes F parciais, que são calculados para cada variável como se ela fosse a última a entrar no modelo (FERREIRA, 2012).

iii. Método Stepwise

Esse método é uma mistura dos métodos forward e backward, inicia com o forward porém a cada variável adicionada as variáveis anteriores são revisadas e verifica-se se seu poder de explicação do modelo permanece significativo.

- Iniciamos com uma variável: aquela que tiver maior correlação com a variável resposta.
- A cada passo do forward, depois de incluir uma variável, aplica-se o backward para ver se será descartada alguma variável.
- Continuamos o processo até não incluir ou excluir nenhuma variável (FERREIRA, 2012).

iv. Critério de informação de Akaike (AIC)

O Critério de Informação de Akaike (AIC) foi proposto por Akaike (1974). A ideia básica deste método é selecionar um modelo que seja parcimonioso. O AIC é definido por:

$$AIC = -2\ln(L_p) + 2[(p + 1) + 1].$$

onde L_p é a função de máxima verossimilhança do modelo e p é o número de variáveis explicativas no modelo. Quanto menor for o valor encontrado, melhor será o ajuste do modelo.

v. Critério de informação Bayesiano

O Critério de Informação Bayesiano (BIC) foi proposto por Schwarz (1978). Ele é bem parecido com o AIC, também é baseado na função de verossimilhança. Porém, ao analisar a estrutura de penalidade, ele é mais rigoroso que o AIC, pois em alguns modelos o BIC é sensível ao aumento da verossimilhança. O BIC é definido por:

$$BIC = -2\ln(L_p) + [(p + 1) + 1]\ln(n),$$

onde L_p é a função de verossimilhança e p é o número de variáveis explicativa no modelo.

2.7 Métodos de verificação da qualidade do ajuste

Após estimar os coeficientes é necessário assegurar a significância das variáveis no modelo. Isto envolve a formulação de teste de hipóteses para determinar se as variáveis explicativas são significativamente relacionadas com a variável dependente. Veremos a seguir alguns testes:

2.7.1 Teste da razão de verossimilhança (TRV)

O TRV, testa simultaneamente se os coeficientes, com exceção do β_0 , são todos nulos. A comparação entre valores observados e preditos usando a função de verossimilhança é baseada na seguinte expressão:

$$D = -2\ln \left[\frac{\text{verossimilhança do modelo ajustado}}{\text{verossimilhança do modelo saturado}} \right].$$

O motivo de se utilizar “- 2ln” é matemático e é necessário para se obter uma distribuição que é conhecida e, portanto, pode ser utilizada para fins de teste hipótese. Este teste é chamado teste da razão de verossimilhança (PAIVA, 2015).

$$D = -2 \sum_{i=1}^n [y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1 - y_i) \ln\left(\frac{1 - \hat{\pi}_i}{1 - y_i}\right)], \quad (2.18)$$

sendo $\hat{\pi}_i$ a estimativa de verossimilhança de π_i .

A função D, também conhecida como “deviance”, sempre será positiva e quanto menor, melhor será o ajuste do modelo. Um modelo é dito saturado quando se contém todas as variáveis, um modelo ajustado corresponde apenas as variáveis ajustada ao modelo. As hipóteses testadas são:

$$H_0 : \beta_1 = \dots = \beta_t = 0$$

$$H1 : E_{j=1,\dots,p} \beta_0 \neq 0$$

Para estimar a significância de uma variável independente, deve-se comparar o valor do modelo com ou sem a variável. A mudança do valor de com a inclusão da variável é obtida por:

$$G = D \left[\frac{\text{modelo sem a variável}}{\text{modelo com a variável}} \right].$$

Ao rejeitar-se a hipótese nula H_0 , tem-se que a variável independente testada é significativa ao modelo.

2.7.2 Teste de Wald

O teste Wald também é utilizado na regressão logística para verificar a significância estatística dos coeficientes no modelo. Desta forma, este teste verifica se cada uma das variáveis independentes apresenta uma relação estatisticamente significativa com a variável dependente. As Hipótese do teste são:

$$H_0 : \beta_k = 0$$

vs

$$H_1 : \beta_k \neq 0$$

O teste de Wald é obtido comparando-se a estimativa de máxima verossimilhança de um coeficiente com a estimativa do seu erro padrão:

$$W = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \quad (2.19)$$

Jr e Donner (1977) examinaram o teste de Wald e observaram que ele tem um comportamento inconsistente, falhando na rejeição da hipótese nula quando o coeficiente β_1 é significativo. Jennings (1986), também encontrou esse mesmo comportamento inconsistente. Portanto, é indicado o uso do teste de verossimilhança.

2.8 Diagnóstico do Modelo.

As técnicas de análise de diagnóstico têm como objetivo avaliar a qualidade de um modelo.

2.8.1 Curva ROC

Num 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

este tipo 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 tornar 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).

2.8.2 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, conseqüentemente, 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, o resíduo quantílico fica definido por:

$$r_i^q = \Phi^{-1}\{F(y_i; \mu_i, \phi)\}, \quad (2.20)$$

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 APLICAÇÃO

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 Califórnia-Irvin UCI em seu repositório Machine Learning Repository's.

3.1 Descrição dos dados e análise exploratória.

Este banco de dados possui informações financeiras e pessoais de 1000 clientes de um cartão de crédito da Alemanha, as variáveis contidas no banco de dados estão apresentadas nas Tabela 1 e Tabela 2 onde duração, idade e valor são quantitativas e o restante são qualitativas, ao todo somam vinte variáveis explicativas e uma variável dependente categórica binária que esta denominada como risco. As pontuação de crédito estão convertidos para DM.

Tabela 1 – Descrição dos dados

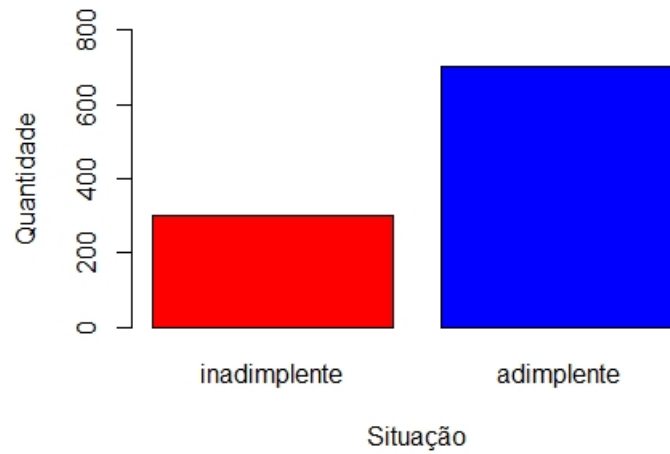
Variável	Descrição	Valor assumido
Status	Status da conta corrente	1: Sem conta corrente, 2: < 0 DM, 3: $0 \leq \dots < 200DM$, 4 : $\geq 200DM$
Duração	Duração do crédito em mês	Meses
Credit_history	Historico de cumprimento de contratos de créditos	0: atraso, 1: conta crítica, 2:nenhum crédito recebido anteriores 3: Crédito existente reembolso devidamente até agora, 4: Todos os creditos nesse banco reembolso devido.
Objetivo	Finalidade para que o crédito é necessario	0: Outros 1: Carro(novo) 2: Carro(usado), 3: Movéis, 4: Radio/ TV, 5: Eletrodomésticos, 6: Reparos, 7: Educação, 8: Ferías, 9: Retreinamentos, 10: Negócios.
Valor	Valor de Crédito em DM	Valor
Poupança	Poupança do devedor	1: Sem conta, 2: < 100 DM, 3: $100 \leq \dots < 500DM$, 4 : $500 \leq \dots < 1000DM$, 5 : $> 1000DM$.
Emprego_duracao	Duração do emprego devedor no emprego atual.	1: desempregado, 2: < 1 ano, 3: $1 \leq \dots$ 4 anos, 4: $4 \leq \dots < 7$ anos, 5: ≥ 7 anos.
Taxa	Créd das prestações como uma porcentagem da renda do devedor	1: ≥ 35 , 2 : $25 \leq \dots < 35$, 3 : $20 \leq \dots < 25$, 4 : < 20 .

Tabela 2 – Descrição dos dados

Variável	Descrição	Valor assumido
Status_sex	Informação combinadas entre sexo e estado civil	1: Masculino/ divorciado, 2: feminino não solteiro ou masculino solteiro, 3: Masculino Casado, 4: Feminino solteiro.
Bueber	Existe outro devedor ou fiador para o crédito ?	1: Nenhum, 2: co-requerente, 3: Fiador.
Present_residence	Período em anos que o devedor vive na residência atual	1: < 1 ano, 2: $1 \leq \dots \leq 4$ anos, 3: $4 \leq \dots < 7$ anos, 4: ≥ 7 anos.
Propriedade	As propriedades mais valiosas do devedor	1: sem propriedade, 2: carro ou outro, 3: edifício soc 4: Imóveis.
Idade	Idade em anos	Anos
Habitação		1: de graça, 2: aluguel, 3: próprio
Numero_cred	Quantidade de crédito incluindo atual	1: 1, 2: 2-3, 3: 4-5, 4: ≥ 6
Emprego	Emprego atual	1: desempregado, 2: não qualificado, 3: funcionario, 4: gerente.
Dependente	Número de pessoas que dependem do devedor	1: 3 ou mais, 2: 0 a 2
Telefone	Existe telefone fixo cadastrado no nome do devedor	1: Não, 2: Sim
Gastarb	O devedor é um trabalhador estrangeiro ?	1: Sim, 2: Não
Risco	O contrato de crédito foi cumprido?	0: Não, 1: Sim.

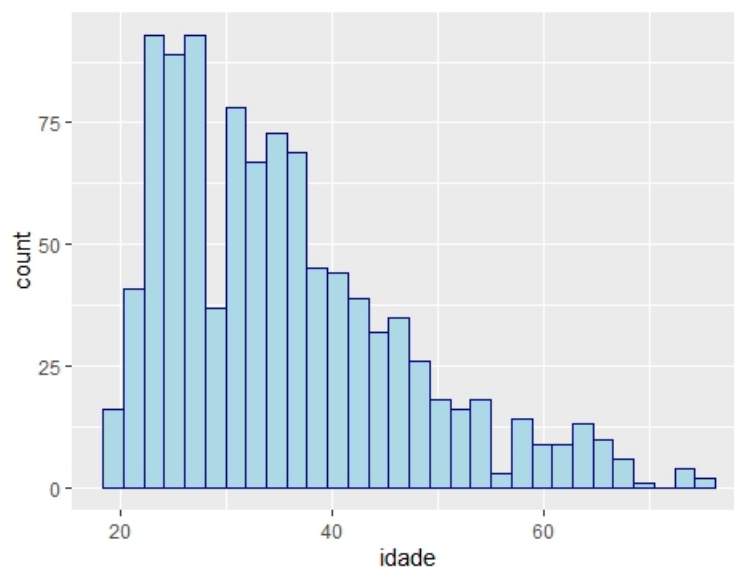
Seguindo com a análise exploratória dos dados observa-se através do gráfico de barras na Figura 2 como a variável dependente esta dividida, o número de inadimplente representa 30 % clientes que não cumpriram o contrato de crédito, ou seja, 300 clientes e adimplente, clientes que cumpriram o contrato de crédito foram 70 %, 700 clientes.

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



Realizando uma análise descritiva dos dados numéricos observa-se através do histograma de idade na Figura 3 a amplitude de idades dos clientes está entre 19 a 75 anos, com a média de 35 anos, porém 95% dos clientes possuem a idade de até 42 anos. Ou seja a maioria dos clientes são pessoas com faixa etária baixas.

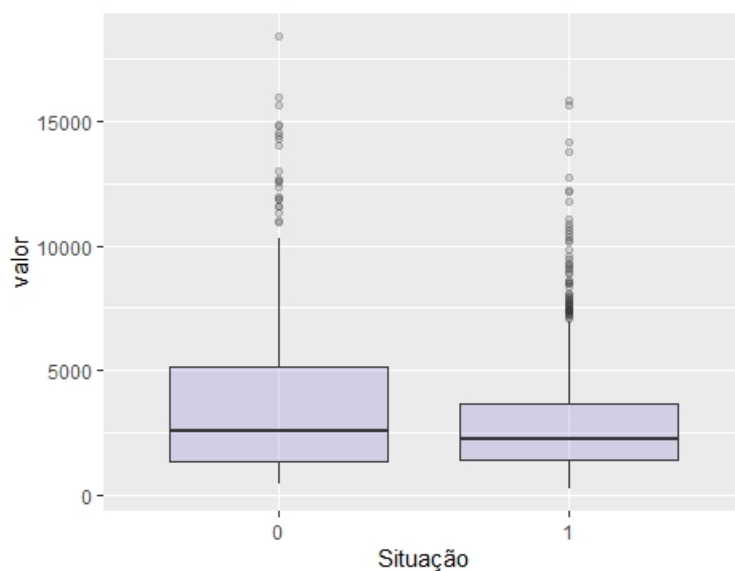
Figura 3 – Histograma de frequência das idades dos clientes.



Na Figura 4 pode-se observar através dos boxplots 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 ha uma variança menor no

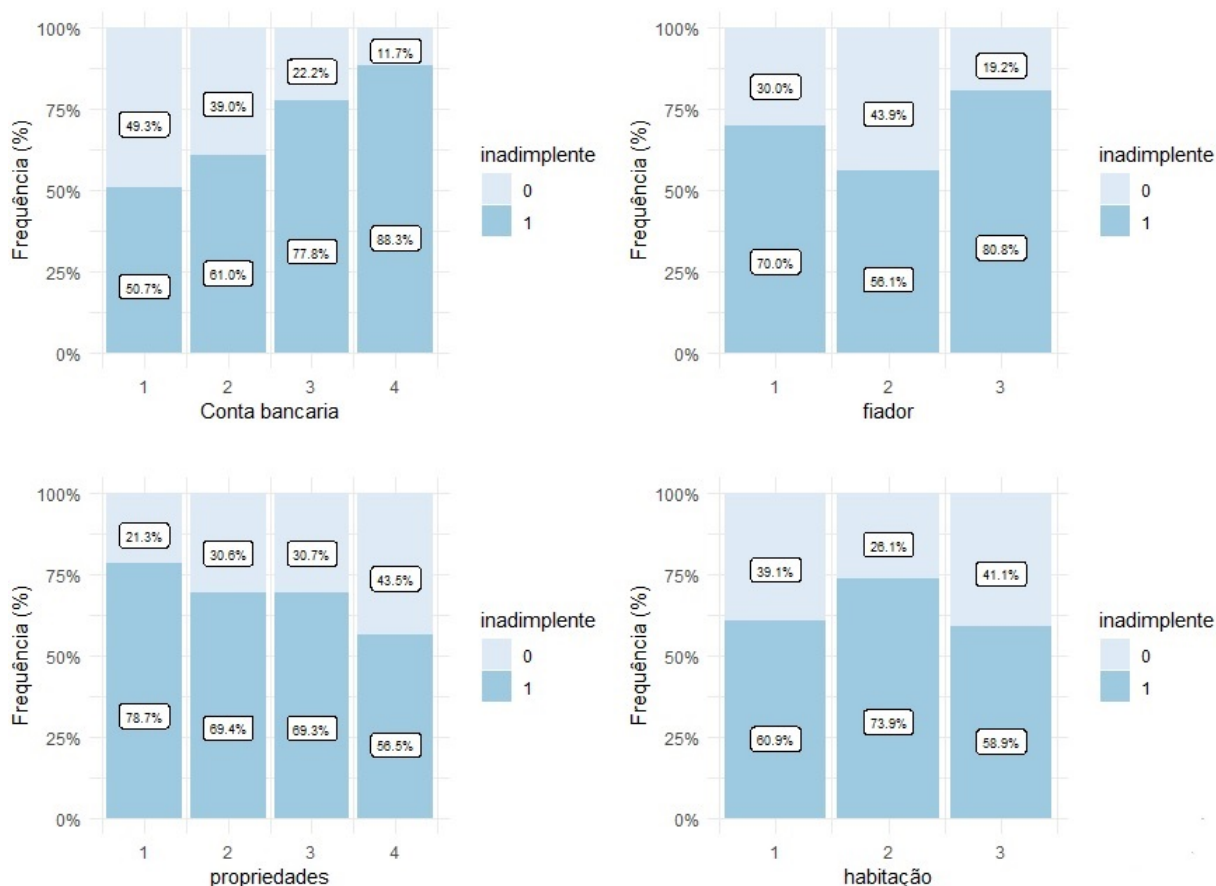
valor do crédito concedidos aos clientes adimplentes em relação aos clientes inadimplentes e há alguns valores discrepantes em relação aos valores de crédito.

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



Na Figura 5 observa-se através dos gráficos de barras a relação da variável resposta risco em relação a 4 variáveis explicativas: conta bancária, fiador, propriedades e habitação. Observa-se na variável conta bancária que dos clientes que possui conta > 200 DM, 88 % são os clientes adimplente, 80% dos clientes que possuem um fiador são adimplente e em relação a propriedade, a maioria dos clientes adimplente não possuem propriedades, e a maioria moram de aluguel.

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



3.2 Regressão logística

Inicialmente para a construção do modelo de regressão logística, dividiu-se os dados de forma aleatória em 80% para treinamento (criando um modelo preditivo) e 20% para teste (avaliar o modelo), onde a amostra de treinamento possui 800 observações divididas em 560 bons pagadores (adimplentes) e 240 maus pagadores (inadimplentes). Foi construído um modelo com a função R `glm()`, para o modelo linear generalizado com todas as variáveis explicativas, observando os seus níveis de significância. Para a seleção do melhor modelo logístico, foi utilizado o método de seleção de covariáveis *stepwise (both directions)*, este método, utiliza o Critério de Informação de Akaike (AIC) na combinação das variáveis dos diversos modelos simulados para selecionar o modelo mais ajustado. Quanto menor o AIC, melhor o ajuste do modelo. Portanto o conjunto de dados possuía 20 variáveis explicativas após a aplicação do método *stepwise* somente 15 variáveis foram significativas ao modelo, assim as seguintes variáveis foram selecionadas para o modelo: `status`, `duração`, `cred_hist`, `objetivos`, `valor`, `poupança`, `taxa`, `status_sex`, `buerge`, `residence`, `propriedade`, `idade`,

plano, habilitação e gastarb.

Na Tabela 3 está apresenta as variáveis selecionadas pelo método stepwise e as estimativas do coeficiente beta associadas a cada variável preditora, considerando nível de 5% de significância, sendo elas: Status ($0 < \dots < 200$ DM e ≥ 200 DM), Credit_hist (Todos os creditos nesse banco reembolso devido), objetivo (carro novo, carro usado, movéis, ferias e retreinamento), valor, poupança ($500 \dots < 1000$ DM, ≥ 1000 DM), taxa (<20), status_sex (Masculino casado) residence (1 ... 4 anos e ≥ 7 anos), idade, propriedade (Imovéis), habitação (aluguel) e gastarb (não).

Tabela 3 – Estimativas e testes associados ao modelo de regressão logística selecionado.

Variável	Coefficiente estimado	Erro padrão	Estatística Z	Pr(> z)
(Intercept)	-7,070e-01	1,075e+00	-0,658	0,510833
status ₂	8,070e-01	2,486e-01	3,246	0,001170 **
status ₃	1,361e+00	4,372e-01	3,112	0,001858 **
status ₄	1,970e+00	2,669e-01	7,381	1,57e-13 ***
duracao	-2,818e-02	1,058e-02	-2,663	0,007733 **
cred_hist ₄	1,317e+00	4,807e-01	2,739	0,006157 **
objetivos ₁	1,728e+00	4,150e-01	4,163	3,14e-05 ***
objetivos ₂	8,117e-01	3,004e-01	2,702	0,006888 **
objetivos ₃	9,581e-01	2,891e-01	3,314	0,000918 ***
valor	-1,091e-04	4,865e-05	-2,243	0,024926 *
poupanca ₄	1,373e+00	6,053e-01	2,268	0,023311 *
poupanca ₅	9,505e-01	2,981e-01	3,188	0,001432 **
taxa ₄	-9,075e-01	3,439e-01	-2,639	0,008321 **
status_sex ₂	8,817e-01	4,343e-01	2,030	0,042363 *
status_sex ₃	1,498e+00	4,272e-01	3,506	0,000454 ***
buerge ₂	-9,777e-01	4,768e-01	-2,051	0,040306 *
residence ₂	-7,060e-01	3,239e-01	-2,180	0,029268 *
propriedade ₄	-1,203e+00	4,967e-01	-2,422	0,015447 *
idade	2,459e-02	9,945e-03	2,473	0,013415 *
habitacao ₂	5,754e-01	2,638e-01	2,181	0,029209 *
gastarb ₂	-1,718e+00	7,625e-01	-2,253	0,024238 *

Pode-se observar que os coeficiente das estimativas das variáveis que foram positivos significa que a variável causa um aumento na probabilidade do cliente não ser inadimplente, que no caso foram as seguintes variáveis:

- i. Conta corrente $0 \leq \dots < 200$ DM e conta corrente ≥ 200 DM
- ii. Histórico de cumprimento de contrato (todos os creditos nesse banco reembolso devidamente).
- iii. Objetivos (Carro novo, carro usado e moveís)

- iv. Poupança de $200 \leq \dots < 500$ 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.

Um conceito importante para entender e interpretar os coeficientes beta logísticos, é a razão de chances. Ele representa a razão das chances de um evento ocorrer (evento = ser adimplente) dada a presença do preditor x . O resultado na Tabela 4 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 tem 5 vezes mais chance de não se tornar inadimplente do que os outros clientes. Na Tabela 4 encontra-se também os intervalos de 95% de confiança para os parâmetros do modelo, com base na estatística de Wald, se o intervalo de confiança incluir o valor de 1, isso implica que não existe diferença entre os grupos estudados.

Tabela 4 – Razão de chances (OR - odds ratio em inglês) e intervalo de confiança .

Variável	OR	2.5 %	97.5 %
(Intercept)	0,493	-2,814	1,400
Status ₂	2,241	0,320	1,294
Status ₃	3,898	0,504	2,217
Status ₄	7,170	1,447	2,493
Duração	0,972	-0,049	-0,007
Cred_hist ₄	3,732	0,375	2,259
Objetivos ₁	5,627	0,914	2,541
Objetivos ₂	2,252	0,223	1,401
Objetivos ₃	2,607	0,392	1,525
Valor	1,000	0,000	0,000
Poupança ₄	3,947	0,187	2,559
Poupança ₅	2,587	0,366	1,535
Taxa ₄	0,404	-1,582	-0,233
Status_sex ₂	2,415	0,030	1,733
Status_sex ₃	4,472	0,661	2,335
buerge ₂	0,376	-1,912	-0,043
residence ₂	0,494	-1,341	-0,071
Propriedade ₄	0,300	-2,176	-0,229
Idade	1,025	0,005	0,044
Habitação ₂	1,778	0,058	1,092
Gastarb ₂	0,179	-3,213	-0,224

3.3 Avaliação do modelo.

A partir do modelo logístico podemos realizar previsões das probabilidades utilizando os dados de teste para avaliar o desempenho do nosso modelo de regressão logística, assim criando uma matriz de confusão para analisar a performance do modelo criado, como observa-se na Tabela 5.

Tabela 5 – Matriz de confusão: Valores reais X valores preditos.

		Valores Reais	
		Inadimplente	Adimplente
Predito	Inadimplente	23	18
	Adimplente	37	122

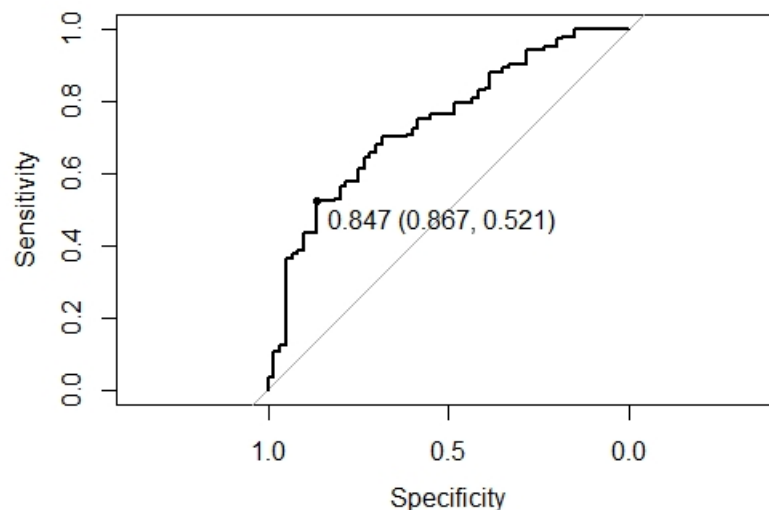
Os resultados dos testes de diagnóstico do modelo encontram-se na Tabela 6. Observa-se que o modelo teve 72% de acurácia o que é muito bom, pode-se considerar que há alta precisão dos resultados e baixo risco de erro, sensibilidade de 87% , dos 140 clientes adimplente o modelo conseguiu acertar 122, apresentou também uma especificidade de 38% dos 60 inadimplentes o modelo acertou 23 e a probabilidade do cliente ser adimplente, dado que o modelo a classificou como adimplente é de 76%.

Tabela 6 – Testes de diagnóstico do modelo.

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

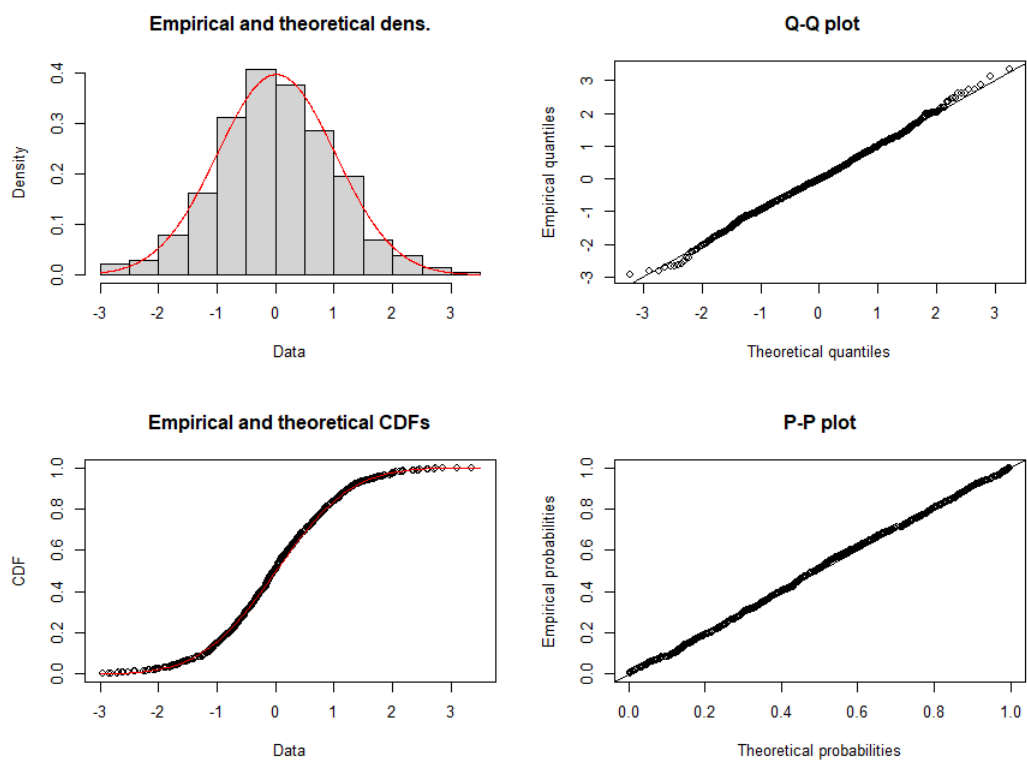
Outra ferramenta utilizada para avaliar a qualidade de ajuste do modelo é a curva ROC. A curva ROC para o modelo em questão encontra-se na Figura 6, na qual mostra o ganho em sensibilidade a medida que a taxa de falso-positivo (1-especificidade) aumenta e tem uma área sob a curva de 0,847, sugerindo que o modelo é bastante eficiente em discriminar clientes que são adimplente dos inadimplentes.

Figura 6 – Curva ROC do modelo logístico.



Também é importante a análise de resíduos no diagnóstico do modelo. Como a falta de normalidade dos resíduos é particularmente notável na modelagem de dados discretos, aplicou-se os resíduos quantílico aleatorizados para ajustar o modelo pra convergir pra uma distribuição Normal padrão, como mostra a Figura 7, onde observou-se que os residuos convergiram para uma distribuição normal.

Figura 7 – Gráficos da análise de resíduos.



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, por meio do método stepwise foram selecionadas as variáveis significativa ao modelo, criando um modelo com suas estimativas, que através dela foi realizada a razão de chance, onde 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 foi realizada as previsões com os dados teste para validar o desempenho do modelo, o que apresentou um bom desempenho, uma taxa de acerto de 72% e sensibilidade de 87%, além da curva ROC surgir que o modelo é bastante eficiente em discriminar 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.

REFERÊNCIAS

- AGRESTI, A. *An introduction to categorical data analysis*. [S.l.]: John Wiley & Sons, 2018. Citado na página 13.
- BRAGA, A. Curvas roc: aspectos funcionais e aplicações. 2001. Citado na página 21.
- COSTA, R. R. *Análise empresarial avançada para crédito*. [S.l.]: Qualitymark Editora Ltda, 2003. Citado na página 11.
- COSTA, R. S. d. Teste de diagnóstico baseado em análise de regressão logística. 2013. Citado na página 21.
- COX, D. R.; HINKLEY, D. V. *Theoretical statistics*. [S.l.]: CRC Press, 1979. Citado na página 13.
- DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996. Citado na página 21.
- FERREIRA, A. Disciplina de modelos lineares 2012-2. *Universidade do Estado*, 2012. Citado na página 18.
- HARRISON, T.; ANSELL, J. Customer retention in the insurance industry: using survival analysis to predict cross-selling opportunities. *Journal of Financial Services Marketing*, Springer, v. 6, n. 3, p. 229–239, 2002. Citado na página 12.
- JENNINGS, D. E. Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, Taylor & Francis, v. 81, n. 394, p. 471–476, 1986. Citado na página 20.
- JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. *Applied logistic regression*. [S.l.]: John Wiley & Sons, 2013. v. 398. Citado 3 vezes nas páginas 13, 14 e 15.
- 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 20.
- KLEINBAUM, D. G.; KLEIN, M. *Logistic regression: a self-learning text*. [S.l.]: Springer Science & Business Media, 2010. Citado na página 13.
- LEWIS, E. M. *An introduction to credit scoring*. [S.l.]: Fair, Isaac and Company, 1992. Citado na página 11.
- LIMA, F. A. P. d. *Práticas em gestão de sistemas de credit scoring e portfólio de crédito em instituições financeiras brasileiras*. Tese (Doutorado), 2011. Citado 2 vezes nas páginas 11 e 12.
- MARCELINO, J. A. Credit scoring: uma ferramenta para análise de crédito em uma instituição de microcrédito produtivo e orientado. 2012. Citado na página 11.
- MAYS, F. E.; LYNAS, N. *Credit scoring for risk managers: The handbook for lenders*. [S.l.]: Thomson/South-Western OH, USA, 2004. Citado na página 11.

- MOURA, G. M. Regressão logística aplicada a análise de risco de crédito. 2018. Citado na página 11.
- NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972. Citado na página 13.
- PAGANO, M.; GAUVREAU, K. Princípios de bioestatística. In: *Princípios de bioestatística*. [S.l.: s.n.], 2011. p. xv–506. Citado na página 21.
- PAIVA, C. C. V. Previsão da inadimplência através da regressão logística. Universidade Federal de Minas Gerais, 2015. Citado na página 19.
- PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2004. Citado na página 13.
- PEREIRA, M. A. A. Modelos não lineares assimétricos com efeitos mistos. Universidade Federal de São Carlos, 2019. Citado na página 21.
- SCHWARZ, G. Estimating the dimension of a model. *The annals of statistics*, JSTOR, p. 461–464, 1978. Citado na página 19.
- SILVA, G. *Modelos logísticos para dados binários*. Tese (Doutorado) — Dissertação de Mestrado, IME-USP, 1992. Citado na página 16.
- SILVA, J. P. da. *Gestão e análise de risco de crédito*. [S.l.]: Editora Atlas SA, 2000. Citado na página 11.
- SOUZA, É. C. d. *Análise de influência local no modelo de regressão logística. 2006, 101 f.* Tese (Doutorado) — Dissertação (Mestrado em Agronomia)-Escola Superior de Agricultura Luiz de . . . , 2006. Citado 4 vezes nas páginas 13, 14, 16 e 17.
- TAVARES, M. d. C. A crise financeira atual. *Paper Itamaraty*, v. 30, n. 04, 2009. Citado na página 11.
- TEAM, R. C. *R: A language and environment for statistical computing (R Version 4.0.3, R Foundation for Statistical Computing, Vienna, Austria, 2020)*. 2021. Disponível em: <<https://www.r-project.org/>>. Acesso em: 12 de Agosto de 2021. Citado na página 22.
- WALPOLE, R. E. *Probabilidade & Estatística para engenharia e ciências*. [S.l.]: Pearson Prentice Hall, 2009. Citado na página 17.