UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

Andrigo Andrade Martins

Aplicação de Análise de Risco de Crédito com o uso das Técnicas de Regressão Logística e Árvores de Decisão

Belo Horizonte

2020

Andrigo Andrade Martins

Aplicação de Análise de Risco de Crédito com o uso das Técnicas de Regressão Logística e Árvores de Decisão

Monografia de especialização apresentada ao Instituto de Ciências Exatas Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Especialista em Estatística.

Orientador: Prof. Dr. Guilherme Lopes de Oliveira

Belo Horizonte

2020

:

Martins, Andrigo Andrade.

M386a Aplicação de análise de risco de crédito com o uso das técnicas de regressão logística e árvores de decisão [manuscrito] / Andrigo Andrade Martins. — 2020.
49.f. il.

Orientador: Guilherme Lopes de Oliveira.
Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências 48-49.

1. Estatística. 2. Árvore de decisão. 3. Análise de regressão. 4.. Curva característica de operação do receptor.5. Sistemas de avaliação de risco de crédito (Finanças). I. Oliveira, Guilherme Lopes de. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística .III. Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6ª Região nº 1510

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 218ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE ANDRIGO ANDRADE MARTINS.

Aos dezesseis dias do mês de dezembro de 2020, às 14:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso do aluno **Andrigo Andrade Martins**, intitulado: "*Aplicação de análise de risco de crédito com o uso das técnicas de regressão logística e árvores de decisão.*", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, o Presidente da Comissão, Professor Guilherme Lopes de Oliveira – Orientador, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado Aprovado condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 16 de dezembro de 2020.

Guilherme Lopes de Oliveira

Prof. Guilherme Lopes de Oliveira (Orientador)
DECOM/CEFET-MG

Jussiane Nader Gonçalves Assinado de forma digital por Jussiane Nader Gonçalves Dados: 2020.12.16 17:14:54 -03'00'

Profa. Jussiane Nader Gonçalves
Departamento de Estatística / UFMG

SERPRO
Assinado digitalmente por:
ROBERTO DA COSTA QUININO
CPF:/CNPJ: 80871291720
Assinado em: 16/12/2020
Sua autenticidade pode ser confirmada no endereço : <http://www.serpro.gov.br/assinador-digital>

Prof. Roberto da Costa Quinino
Departamento de Estatística / UFMG

Este trabalho é dedicado à minha família e à minha esposa que tanto me apoiaram para concretização deste objetivo.

AGRADECIMENTOS

Agradeço a Deus que depositou em mim a fé, a esperança e a certeza de que meus sonhos se realizariam e os meus esforços não seriam em vão.

Agradeço a meus pais por ter me presenteado com a vida e por me conduzirem por bons caminhos, pelo amor, carinho, dedicação e confiança que depositaram em mim contribuindo para a realização deste objetivo. E as minhas irmãs e minha esposa pelo apoio e compreensão.

Agradeço ao Banco do Brasil S/A, pela concessão da bolsa de estudos. Tenho o prazer de manifestar a minha gratidão a essa instituição que me ajudou financeiramente nessa caminhada.

Agradeço aos meus colegas de sala de aula, pelo apoio e pela amizade que foi cultivada durante esses meses e que permanecerá para sempre.

Aos professores do curso, que assim como eu, sacrificaram seus finais de semana em busca de novas experiências do conhecimento.

Deixo um agradecimento especial ao meu orientador pelo incentivo, dedicação e por todo apoio e paciência ao longo da elaboração desse trabalho.

Resumo

Com o aumento da concessão de crédito e a necessidade de uma resposta rápida e assertiva às requisições deste tipo de serviço, faz-se necessária a utilização de métodos para mensuração do risco de crédito com o intuito de tornar o processo de decisão mais confiável. Esse trabalho tem o objetivo de comparar e exemplificar a utilização de duas técnicas para este fim, sendo elas: regressão logística e árvore de decisão por meio do método CART. Para isso, selecionou-se uma base de dados relacionada à concessão de cartão de crédito por certa instituição bancária. A base de dados de interesse foi dividida em dois subconjuntos de treinamento (ajuste) e validação (verificação), ficando este último com vinte por cento do tamanho da base de dados inicial. As predições obtidas com o ajuste de ambos os métodos foram comparados nos dados de treinamento e de validação através de indicadores de desempenho, tais como sensibilidade, especificidade, área sob a curva ROC e taxa de acerto global. Os resultados mostram que o modelo de previsão produzido pela regressão logística foi o que se mostrou mais adequado.

Palavras-chave: Árvore de decisão. *Credit scoring*. Curva ROC. Regressão logística.

Abstract

With the increase in credit granting and the need for a quick and assertive response to requests for this type of service, it is necessary to use methods for measuring credit risk in order to make the decision process more reliable. This work aims to compare and exemplify the use of two techniques for this purpose, namely: logistic regression and decision tree using the CART method. For this, a database related to the granting of a credit card by a certain banking institution was selected. The database of interest was divided into two subsets of training (adjustment) and validation (verification), the latter being twenty percent of the size of the initial database. The predictions obtained by adjusting both methods were compared in the training and validation data through performance indicators, such as sensitivity, specificity and area under the ROC curve. The results were satisfactory and quite similar in both cases, however the logistic regression outperformed the CART decision tree in our application.

Keywords: Decision tree. Credit scoring. ROC curve. Logistic regression.

Sumário

1 INTRODUÇÃO 10
2 MODELO DE REGRESSÃO LOGÍSTICA 11
2.1 Estimativa dos Coeficientes: Método de Máxima Verossimilhança 13
2.2 Testando a Significância dos Coeficientes 13
2.3 Análise da Qualidade de Ajuste do Modelo Logístico 14
2.3.1 Teste de Hosmer – Lemeshow 14
2.3.2 Matriz de Classificação: Sensibilidade, Especificidade e a Curva ROC ... 15
2.4 Interpretação dos Coeficientes do Modelo de Regressão Logística 19
3 ÁRVORES DE DECISÃO 20
3.1 Processo de Indução de Árvores 20
3.2 Algoritmo de Indução de Árvores de Decisão 21
4 ESTUDO DE CASO 23
4.1 Análise Descritiva da Base de Dados 26
4.2 Ajuste do Modelo de Regressão Logística 32
4.2.1 Diagnóstico de Ajuste do Modelo Logístico 34
4.2.1.1 Função Resposta Estimada 34
4.2.1.2 Teste de Hosmer – Lemeshow 35
4.2.1.3 Curva ROC Regressão Logística 36
4.2.1.4 Matriz de Classificação Regressão Logística 37
4.2.2 Análise da Predição nos Dados de Validação Regressão Logística 39
4.3 Resultados da Análise da Árvore de Decisão CART 40
4.3.1 Diagnóstico de ajuste da Árvore de Decisão CART 42
4.3.1.1 Curva ROC da Árvore de Decisão CART 43
4.3.1.2 Matriz de Classificação da Árvore de Decisão CART 43
4.3.2 Análise da Predição nos Dados de Validação Árvore de Decisão CART . 44
5 CONCLUSÕES 46
REFERÊNCIAS 48

1 INTRODUÇÃO

Em toda concessão de crédito há um risco associado à inadimplência do tomador. Para mitigar esse risco, o agente cedente deve avaliar o potencial de retorno do tomador, identificando se o mesmo possui idoneidade e capacidade financeira suficiente para amortizar a dívida que pretende contrair.

A análise sobre o nível de risco de crédito requer a definição de critérios cuidadosos que possam indicar a possibilidade de inadimplência do tomador de um crédito ou do lançador de um título, remetendo para mecanismos de avaliação da sua saúde financeira. Neste momento, estará em evidência a necessidade de prever a probabilidade de o pagamento ocorrer (Caouette, Altman e Narayanan, 1999).

Essa questão assume maior relevância no caso de uma instituição financeira ao se considerar que os reflexos de uma correta mensuração dos níveis de risco sobre uma carteira de ativos podem representar um diferencial competitivo. Por outro lado, o insucesso de um modelo de avaliação de risco acarretará efeitos danoso à instituição. Os modelos de análise para concessão de crédito, conhecidos como modelos de *credit scoring*, se baseiam em dados históricos da base de clientes existentes para avaliar se, no futuro, um cliente terá mais chances de ser bom ou um mau pagador.

Os modelos de *credit scoring* são implantados nos sistemas das instituições, permitindo que a avaliação de crédito, tanto de pessoas físicas quanto de empresas, seja realizada de forma massificada. Para os clientes pessoas físicas se utilizam informações cadastrais e de comportamento. Já quando aplicados às empresas são utilizados índices financeiros. Os modelos de *credit scoring* são específicos para a aprovação em cada produto de crédito, dentre os quais podemos citar o crédito pessoal, cartão de crédito, cheque especial, entre outros. Neste estudo, o produto em questão diz respeito ao cartão de crédito concedido para pessoa física por certa instituição bancária.

As instituições financeiras que trabalham com concessão de crédito utilizam-se de modelos probabilísticos para avaliar o risco de inadimplência dos potenciais contratantes de produtos de crédito. Dentre esses modelos se destacam a regressão logística, árvores de decisão, análise discriminante, redes neurais e algoritmos genéticos (Sicsú, 2010). Estes métodos visam a identificação de um critério de avaliação do risco de crédito que tenha boa aderência aos dados disponíveis. A adoção de um bom critério permite o direcionamento da estratégia da instituição,

podendo aumentar a eficiência do seu negócio. Qualquer avanço nas técnicas, que resulte no aumento da precisão de um modelo de previsão, acarreta ganhos financeiros para a instituição.

Neste contexto, o objetivo deste estudo é a apresentar uma revisão detalhada a respeito do modelo de regressão logística e o modelo CART de árvore de decisão para a classificação de bons e maus pagadores, considerando-se uma aplicação ao produto cartão de crédito pessoa física concedido por uma instituição bancária. Ambas as técnicas utilizadas permitem a identificação de fatores importantes para a discriminação dos clientes da instituição.

2 MODELO DE REGRESSÃO LOGÍSTICA

Em diversos problemas que envolvem a análise estatística, o objetivo é a construção e verificação de modelos que relacionam um conjunto de p variáveis explicativas $X_1,\ X_2, \ldots\ X_P$, (variáveis independentes) à uma variável resposta Y (variável dependente). O conjunto de variáveis independentes pode conter apenas uma (regressão simples com $p = 1$) ou diversas (regressão múltipla $p > 1$) variáveis, as quais podem ser qualitativas (categóricas; nominais ou ordinais) ou quantitativas (numéricas; discretas ou contínuas) ou uma mistura entre variáveis de diferentes tipos no caso da regressão múltipla.

Num contexto geral, tal análise é feita por muito dos chamados modelos de regressão. Para cada tipo de variável resposta pode-se definir uma classe de modelos mais apropriada. Quando a variável resposta de interesse é dicotômica, atribuindo-se o valor 1 à ocorrência do evento de interesse (sucesso) e 0 à sua não-ocorrência (fracasso), tem-se como alternativa apropriada a classe do chamado modelo de regressão logística (Mingoti, 2005).

O modelo de regressão logística fornece uma maneira de se relacionar a probabilidade do sucesso do evento, ou seja, $P\ (Y = 1)$, com os valores das variáveis explicativas $X_1,\ X_2, \ldots\ X_P$. A equação matemática que define esta relação é dada por

$$\pi(x) = \frac{\exp\{\beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + \ldots + \beta_p\ x_p\}}{1 + \exp\{\beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + \ldots + \beta_p\ x_p\}}\ , \qquad (1)$$

onde $\pi\ (\boldsymbol{x}) = P\ (Y = 1 \mid \boldsymbol{x})$; $\boldsymbol{x} = (x_1,\ x_2, \ldots, x_p)$, são os valores observados para as variáveis aleatórias explicativas $X_1,\ X_2, \ldots, X_P$; e $\beta_0,\ \beta_1, \ldots, \beta_{P,}$ são os coeficientes (parâmetros) de regressão a serem estimados a partir dos dados observados. Pelas características da construção do modelo, tem-se que a variável resposta de interesse segue uma distribuição Bernoulli com probabilidade de sucesso dependente dos valores observados das variáveis aleatórias $x_1,\ x_2, \ldots, x_P$, isto é, $Y \sim Bernoulli\ (\pi\ (x))$. Dessa forma, o valor esperado (média) da variável

resposta, dados $x_1,\ x_2, \ldots, x_P$ é tal que $E\ (Y \mid x) = \pi(x)$. Usando manipulação algébrica apropriada sob a equação (1) é possível obter uma relação linear entre a probabilidade de sucesso e as variáveis explicativas, a qual é dada por

$$g(x) = \ln\left[\frac{\pi\ (x)}{1-\pi\ (x)}\right] = \beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + \ldots + \beta_p\ x_p \qquad (2)$$

A função $g(x)$ resultante dessa manipulação algébrica é comumente chamada de transformação logito. O logito, $g(x)$, é linear nos parâmetros, contínua e, dependendo da alcance de x, pode variar de $-\infty$ a $+\infty$. Portanto, neste modelo a relação direta entre as variáveis preditoras e a média da variável resposta não é linear. Em vez disso, a relação linear é assumida entre os coeficientes associados às variáveis $X_1,\ X_2, \ldots, X_P$ e a transformação logito correspondente como pode ser visto na equação (2). Como consequência da trasnformação, as probabilidades obtidas via regressão logística tem um formato curvilíneo apresentado na forma de S, conforme ilustrado na Figura 1 para o caso de uma única variável explicativa. O eixo vertical da figura corresponde às probabilidades $\pi\ (x)$ em função do valor correspondente na transformação logito $g(x)$ representada no eixo horizontal da Figura 1.

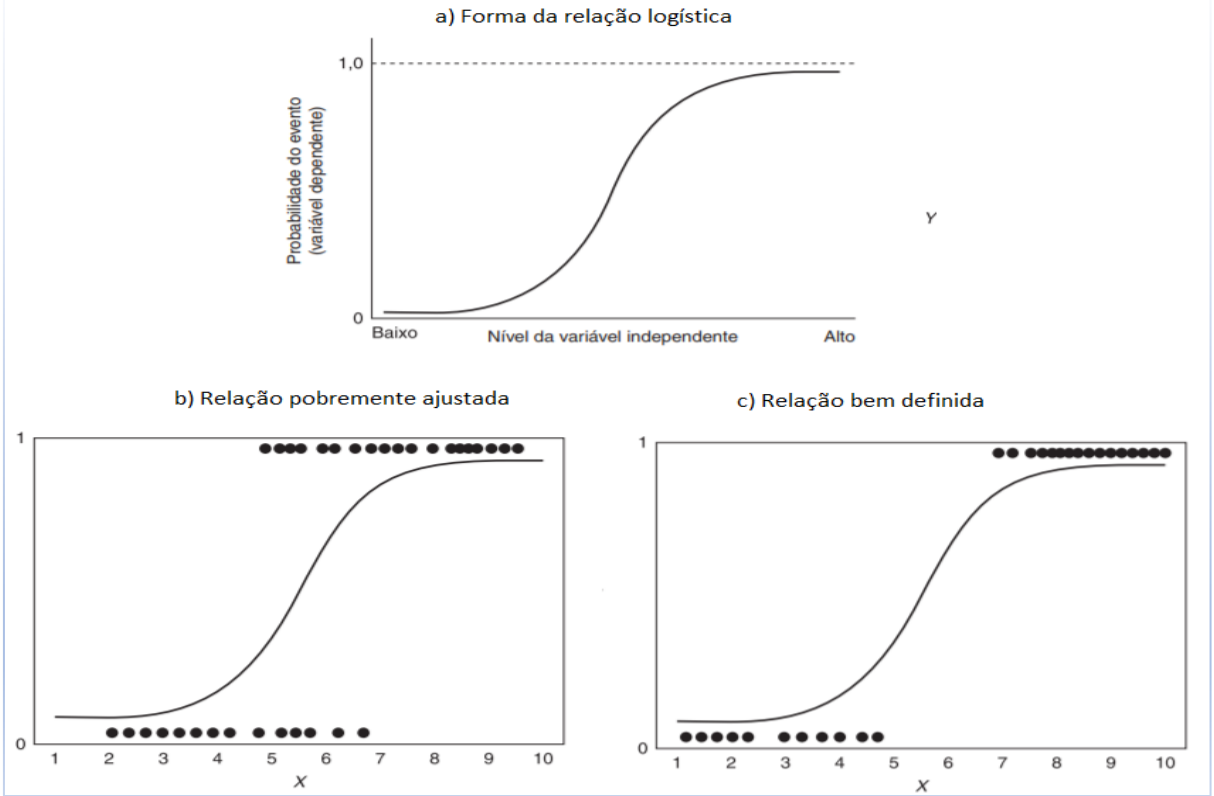

Figura 1- Exemplos de ajustes do modelo logístico com uma variável preditora Fonte: Adaptado de Hair *et al*. (2009)

2.1 Estimativa dos Coeficientes: Método de Máxima Verossimilhança

O ajuste de um modelo de regressão diz respeito a estimação de seus parâmetros desconhecidos, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_P)$, com base nos dados observados. O método da máxima verossimilhança fornece estimadores para $\boldsymbol{\beta}$ que maximizam a probabilidade de se obter os dados observados da amostra. A maximização é feita por meio da chamada função de verossimilhança, a qual pode ser interpretada como sendo a probabilidade dos dados observados como uma função dos parâmetros do modelo, condicional nos valores observados para as covariáveis $x_1, x_2, \ldots, x_P$.

No caso do modelo de regressão logística, temos que Y assume os valores 0 ou 1, sendo $\pi(\boldsymbol{x}) = P(Y = 1 \mid x)$ *e* $1 - \pi(\boldsymbol{x}) = P(Y = 0 \mid \boldsymbol{x})$. Com o pressuposto de que as unidades amostrais $i = 1, \ldots, n$, onde n representa o tamanho da amostra, são independentes entre si e, além disso, como cada uma delas está associada ao respectivo modelo $Bernoulli\left(\pi(\boldsymbol{x_i})\right)$, a função de verossimilhança fica dada por

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1 - y_i} \quad (3)$$

O princípio da máxima verossimilhança assegura que se use como estimador de β o valor que maximiza a equação (3). Os detalhes matemáticos envolvendo o processo de definição das equações de maximização da função de verossimilhança $l(\beta)$ não serão descritos neste trabalho (para detalhes veja, *e.g.*, Hosmer, Lemeshow e Sturdivant, (2013)). Vale ressaltar que as equações resultantes para estimação dos parâmetros via maximização da função $l(\beta)$ são não lineares nos parâmetros do modelo. Por isso, a obtenção da solução é feita usando métodos especiais iterativos de estimação. Neste trabalho, o processo de estimação é feito através da função *glm()* do *software* R usando o seu método *default* de estimação, denominado *iteratively reweighted least squares method* (McCullagh e Nelder, 1989).

2.2 Testando a Significância dos Coeficientes

Depois de estimados os coeficientes β, devemos analisar a relevância das variáveis associadas a cada um destes coeficientes. Essa análise normalmente envolve testes de hipóteses para verificar se cada um dos parâmetros são significativos ou não. Para este fim, utilizaremos o resultado associado ao teste de Wald, que é o teste *default* disponível através da função de ajuste *glm()* do *software* R.

Para $j = 1, \dots, p$, as hipóteses do teste são $H_0 : \beta_j = 0$ *versus* $H_1 : \beta_j \neq 0$. Sendo $\hat{\beta}_j$ a estimativa de máxima verossimilhança e $SE\ (\hat{\beta}_j)$ o erro padrão do estimador para o parâmetro β_j, a estatística do teste de Wald é dada por

$$w_j = \frac{\widehat{\beta}_j}{SE\ (\widehat{\beta}_j)} \tag{4}$$

O teste se baseia na distribuição assintótica dos estimadores de máxima verossimilhança dos parâmetros do modelo. Mais especificamente, sob a hipótese $H_0 : \beta_j = 0$, a estatística w dada na equação (4), tem distribuição normal – padrão.

2.3 Análise da Qualidade de Ajuste do Modelo Logístico

2.3.1 Teste de Hosmer – Lemeshow

Existem várias medidas que podem ser usadas para quantificar a qualidade global de ajuste do modelo logístico dentre elas se destaca o teste baseado na estatística de Hosmer – Lemeshow (Hosmer e Lemeshow, 1980), o qual será utilizado neste trabalho.

O teste de Hosmer – Lemeshow é muito utilizado em regressão logística com a finalidade de testar a bondade do ajuste, ou seja, o teste comprova se o modelo proposto pode explicar bem o que se observa. O teste avalia o modelo ajustado através das distâncias entre as probabilidades ajustadas e as probabilidades observadas. As hipóteses do teste são:

H_0: o modelo se ajusta bem aos dados

H_1: o modelo não está bem ajustado aos dados

A bondade do teste é baseada na divisão da amostra segundo suas probabilidades ajustadas com base nos valores dos parâmetros estimados pela regressão logística. Para tal, os valores ajustados são dispostos em ordem crescente e, em seguida, separados em G grupos de tamanho aproximadamente igual de modo que cada conjunto tenha um número de observações $n_1, n_2, n_3, \dots, n_G$. Em casos gerais, Hosmer e Lemeshow (1980) propõem que seja utilizado $G = 10$. Sejam $e_1, e_2, e_3, \dots, e_G$ o número esperado de elementos amostrais em cada conjunto calculados com base no modelo logístico ajustado. A estatística do teste Hosmer – Lemeshow é calculada com base na comparação dos valores de n_g com e_g, $g = 1, 2, 3, \dots, G$, dispostos em uma tabela de contingência.

Valores elevados, da estatística de Hosmer – Lemeshow indicam um modelo não adequado. Segundo Hosmer, Lemeshow e Sturdivant (2013), se H_0 for verdadeira, ou seja, se o modelo está bem ajustado, a distribuição da estatística do teste se aproxima de uma distribuição qui-quadrado com $G - 2$ graus de liberdade. Isto fornece uma maneira de se tomar uma decisão sobre a rejeição ou não da hipótese nula usando métodos da região crítica ou p-valor apropriados, isto é, obtidos com base na distribuição qui-quadrado com $G - 2$ graus de liberdade. Um p-valor alto, que leva à não rejeição de H_0, indica que não existem diferenças significativas entre as probabilidades previstas pelo modelo e os valores observados no banco de dados em análise, caracterizando um bom ajuste do modelo aos dados.

2.3.2 Matriz de Classificação: Sensibilidade, Especificidade e a Curva ROC

Além do teste de significância descrito na Seção 2.3.1, uma maneira de avaliar os resultados de um modelo de regressão logística é através dos resultados de uma matriz de classificação 2 x 2 que distribui os elementos da amostra entre sucesso e fracasso com base nos valores das probabilidades de sucesso estimadas e contrasta as classificações estimadas com as classificações reais (observadas com base nos valores da variável resposta na amostra). De forma geral, a matriz de classificação tem a seguinte forma:

		Valores reais (esperados)	
		Positivo	Negativo
Valores previstos (modelo)	Positivo	VP	FP
	Negativo	FN	VN

Figura 2 - Matriz de classificação Fonte: Elaboração própria

Na Figura 2, os termos VP (verdadeiro positivo) e VN (verdadeiro negativo) indicam que o valor previsto coincide com o valor observado correspondente a um sucesso ou a um fracasso, respectivamente; FP (falso positivo) indica que cometemos o que chamamos de Erro do Tipo I – quando um fracasso observado é classificado como sucesso; e FN (falso negativo) indica a ocorrência do que chamamos de Erro do Tipo II – quando um sucesso observado é classificado como sendo um fracasso. No contexto de análise de risco de crédito, assumindo que $Y = 1$ caracteriza um cliente não credível, estes erros podem ser interpretados como descrito abaixo:

- **Erro tipo I (FP)** - ocorre quando se recusa uma operação que, caso fosse realizada, seria um bom negócio para o credor (um cliente credível $Y = 0$ ser classificado como não credível $Y = 1$). Ou seja, cometer o erro tipo I implica perda de um bom cliente. O custo dessa perda é de difícil avaliação.
- **Erro tipo II (FN)** - ocorre quando se aprova uma operação que se tornará problemática para o credor (um cliente não credível $Y = 1$ ser classificado como credível $Y = 0$). Ou seja, cometer o erro tipo II implica perda financeira para o credor.

Uma vez que os valores preditos pelo modelo estão na escala das probabilidades de sucesso, para realizar a classificação das unidades amostrais com base no modelo ajustado é necessária a definição uma probabilidade de referência chamada de ponto de corte. A probabilidade estimada para cada indivíduo é comparada com o ponto de corte pré-estabelecido. Se a probabilidade estimada exceder o ponto de corte, então assume-se que o resultado predito para a variável resposta deve ser igual a 1 (sucesso); caso contrário, deve ser igual a 0 (fracasso).

A porcentagem de sucessos corretamente previstos pelo modelo (VP) dentre o total de sucessos observados (VP+FN) é chamada de sensibilidade e a porcentagem de fracassos corretamente previstos pelo modelo (VN) dentre o total de fracassos observados no bando de dados (VN+FP) é chamada de especificidade. Sendo assim, no contexto de análise de risco de crédito considerado para aplicação nesse trabalho, a sensibilidade pode ser entendida como a capacidade de identificar os clientes não credíveis, ou seja, corresponde à probabilidade de o teste classificar corretamente um mau cliente (um cliente não credível $Y = 1$ ser classificado como não credível $\hat{Y} = 1$). A especificidade pode ser entendida como a capacidade de identificar os clientes credíveis, ou seja, corresponde à probabilidade de o teste classificar corretamente um bom cliente (um cliente credível $Y = 0$ ser classificado como credível $\hat{Y} = 0$). Em geral, quanto maior (mais próximo de 1) é o ponto de corte, maior é a especificidade do

modelo, mas menor é a sua sensibilidade. Assim, na escolha do ponto de corte, levamos em consideração a intenção do modelo como critério de classificação.

Neste trabalho, a análise da capacidade preditiva do modelo será feita considerando-se diferentes pontos de cortes. Em geral, o ponto de corte 0,5 é utilizado por se tratar do valor que divide a escala das probabilidades exatamente ao meio, mas outras escolhas podem ser feitas em cada contexto. Como salientado por Paiva (2015), se o objetivo for escolher um ponto de corte ideal para efeitos de classificação, pode-se selecionar um ponto de corte que maximiza a sensibilidade e especificidade ao mesmo tempo. A definição do valor que corresponde a essa maximização conjunta pode ser feita facilmente através de um gráfico similar ao que aparece na Figura 3, onde temos uma comparação das curvas de sensibilidade e especificidade com relação a cada ponto de corte possível. Para o exemplo mostrado na Figura 3, uma escolha "ótima" para um ponto de corte pode ser aproximadamente 0,20, que é aproximadamente onde as curvas de sensibilidade e especificidade se cruzam neste caso. Para o estudo de caso abordado neste trabalho, os pontos de cortes utilizados serão definidos oportunamente ao analisarmos os resultados do ajuste obtidos.

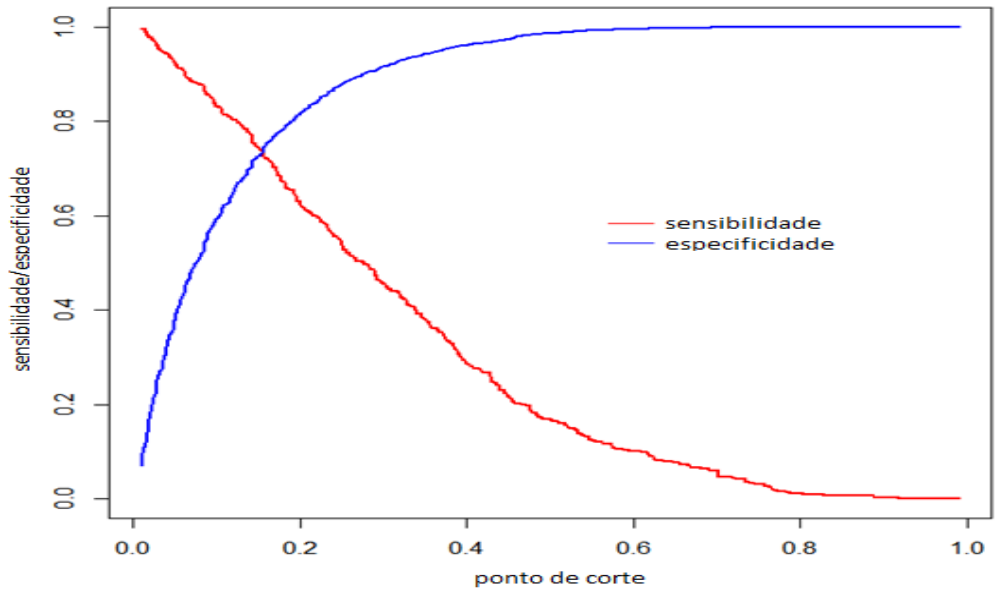

Figura 3 – Exemplo de gráfico de sensibilidade e especificidade Fonte: Paiva (2015)

Ainda no contexto de avaliar o poder de discriminação do modelo, a curva ROC (*Receiver Operating Characteristc*) é uma ferramenta gráfica que permite avaliar o desempenho de um modelo de regressão binária tendo como base uma comparação entre a sensibilidade e a especificidade. Mais especificamente, faz-se um gráfico contendo a

sensibilidade no eixo vertical e (1- especificidade) no eixo horizontal obtidas a partir da utilização de vários pontos de corte.

A área sob a curva ROC varia de 0,5 a 1,0 e é utilizada como uma medida para o poder discriminatório do modelo, ou seja, a capacidade do modelo em discriminar entre aqueles indivíduos nos quais o evento de interesse ocorreu ($Y = 1$) *versus* aqueles em o evento não ocorreu ($Y = 0$). Como exemplo, a Figura 4 traz três curvas ROC com níveis de discriminação diferentes: baixo, médio e elevado. Quanto mais próxima do canto superior esquerdo, maior será o seu poder discriminante, ou seja, maior será a capacidade do teste em distinguir sucessos e fracassos nos dados observados.

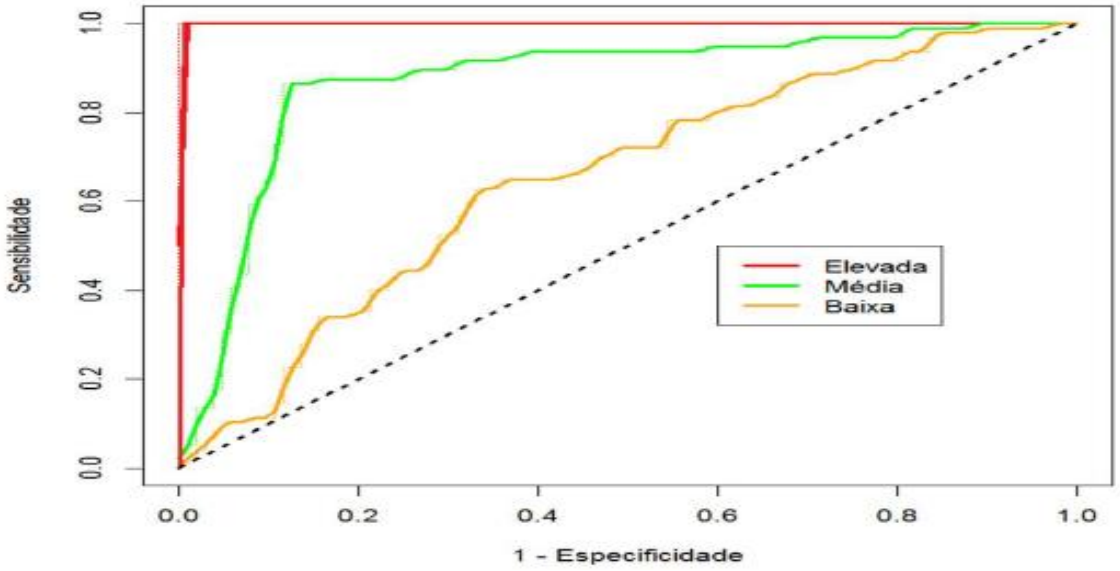

Figura 4 - Níveis de discriminação Fonte: Cristiano (2017)

Não existe um concenso na literatura a respeito de um valor de referência para a área sob a curva ROC que indicaria uma boa discriminação em Hosmer, Lemeshow e Sturdivant (2013), fornecem um critério de avaliação, o qual é apresentado abaixo e será considerado neste trabalho:

Quadro 1 - Critérios de classificação da curva ROC

ROC = 0,5	não possui poder de discriminação
0,5 < ROC < 0,7	baixo poder de discriminação
0,7 <= ROC < 0,8	aceitável poder de discriminação
0,8 <= ROC < 0,9	excelente poder de discriminação
ROC >= 0,9	poder de discriminação acima do normal

2.4 Interpretação dos Coeficientes do Modelo de Regressão Logística

Após a definição de um modelo que seja significativo e tenha capacidade discriminatória aceitável para o contexto prático de interesse, uma análise interessante diz respeito à interpretação dos parâmetros do modelo $\beta_1, \beta_2, \ldots, \beta_P$. No modelo de regressão logística esta interpretação é obtida comparando a probabilidade de sucesso $\pi(x)$ com a probabilidade de fracasso $1 - \pi(x)$, usando a função *Odds Ratio* - OR (razão de chances). Essa função é obtida a partir da função *odds.*

Considere o caso de um modelo com apenas uma variável explicativa X. Para um indivíduo particular, a chance de sucesso do evento, dado que o valor da variável preditora para esse indivíduo é $X = x_0$, é dada por

$$o(x_0) = \frac{\pi(x_0)}{1-\pi(x_0)} = \frac{\frac{e^{\beta_0 + \beta_1 x_0}}{1+e^{\beta_0+\beta_1 x_0}}}{1-\frac{e^{\beta_0+\beta_1 x_0}}{1+e^{\beta_0+\beta_1 x_0}}} = \frac{\frac{e^{\beta_0+\beta_1 x_0}}{1+e^{\beta_0+\beta_1 x_0}}}{\frac{1}{1+e^{\beta_0+\beta_1 x_0}}} = e^{\beta_0 + \beta_1 x_0}. \quad (5)$$

Similarmente, para um outro indivíduo tal que $X = x_1$, tem – se

$$\frac{\pi(x_1)}{1-\pi(x_1)} = e^{\beta_0 + \beta_1 x_1}. \quad (6)$$

A comparação destes dois indivíduos através da razão entre as chances dadas nas equações (5) e (6) é dada por

$$OR(x_1, x_0) = \frac{e^{\beta_0 + \beta_1 x_1}}{e^{\beta_0 + \beta_1 x_0}} = e^{\beta_1 (x_1 - x_0)}.$$

Desta forma, se consideramos, sem perda de generalidade, que $x_1 = x_0 + 1$, então

$$OR(x_0 + 1, x_0) = e^{\beta_1 (x_0 + 1 - x_0) = e^{\beta_1}}. \quad (7)$$

Portanto, a razão de chances $OR(x_0 + 1, x_0)$ pode ser interpretada como sendo a magnitude com qual a chance de $Y = 1$ se modifica pelo acréscimo de 1 unidade na covariável X. A chance de ocorrência do evento será aumentada se $\beta_1 > 0$, diminuirá se $\beta_1 < 0$ e será considerada igual se $\beta_1 = 0$.

No caso de um modelo com múltiplas covariáveis, a interpretação é feita separadamente para cada covariável na forma descrita acima, considerando que as demais covariáveis sejam mantidas fixadas.

3 ÁRVORES DE DECISÃO

As árvores de decisão são métodos que utilizam uma representação gráfica baseada em ramificações, cujo objetivo é identificar grupos de indivíduos com características de interesse em comum. Para tal, é utilizado um método recursivo que divide a amostra inicial em subamostras, baseando-se em resultados observados das variáveis explicativas e em suas interações. Formam-se, assim, grupos para os quais a variável resposta apresenta comportamento homogêneo dentro dos grupos e heterogêneo entre eles (Breiman *et al.*, 1984).

Uma árvore de decisão é chamada de árvore de classificação se a variável resposta for categórica, ou árvore de regressão, se ela for numérica (Taconeli, 2008). Neste trabalho, serão induzidas árvores de classificação, pois a variável resposta é dicotômica (sim ou não para o pagamento do cartão de crédito do próximo mês).

3.1 Processo de Indução de Árvores

A indução de árvore de decisão é uma técnica de modelagem não-paramétrica, que faz divisões recursivas num espaço finito multidimensional definido por variáveis independentes, em zonas que são tão homogêneas quanto possível em termos da resposta do atributo alvo. O resultado da análise é uma estrutura hierárquica chamada árvore de decisão com ramos e folhas, que contém as regras para predizer novos casos (Tan, Steinbach e Kumar, 2009). A árvore de decisão é a representação gráfica do modelo criado, semelhante a uma árvore em sentido invertido.

O processo de indução de árvores é iniciado por meio de uma amostra, denominada nó raiz, que é dividida em subamostras, denominadas nós filhos ou nós intermediários, os quais são chamados também de nós de decisão. Cada nó passa por um teste sobre uma ou mais variáveis independentes (atributos) e os resultados desses testes formam os ramos da árvore. Esses testes, na maioria dos casos, consistem na comparação dos valores do atributo com um valor de referência (constante). Se o atributo em teste num determinado nó é do tipo nominal, o número de ramos a partir do nó de decisão poderá ser igual ou menor ao número de categorias que o atributo possa assumir. Já para atributos em testes do tipo contínuo, o nó de decisão se ramifica em duas subamostras, fazendo a comparação do tipo maior ou menor que um certo valor de referência para o atributo. Quando uma subamostra não puder mais ser subdivida segundo algum critério de parada, ela é então denominada de nó final ou nó folha. Os nós folha

são dispostos na extremidade inferior da árvore, onde cada folha representa um valor de predição para a variável dependente (atributo alvo) ou uma distribuição de probabilidade dos seus possíveis valores. Esse processo de definição da árvore é dito recursivo devido a cada subamostra gerar novas subamostras. A estrutura de uma árvore de decisão está ilustrada na Figura 5.

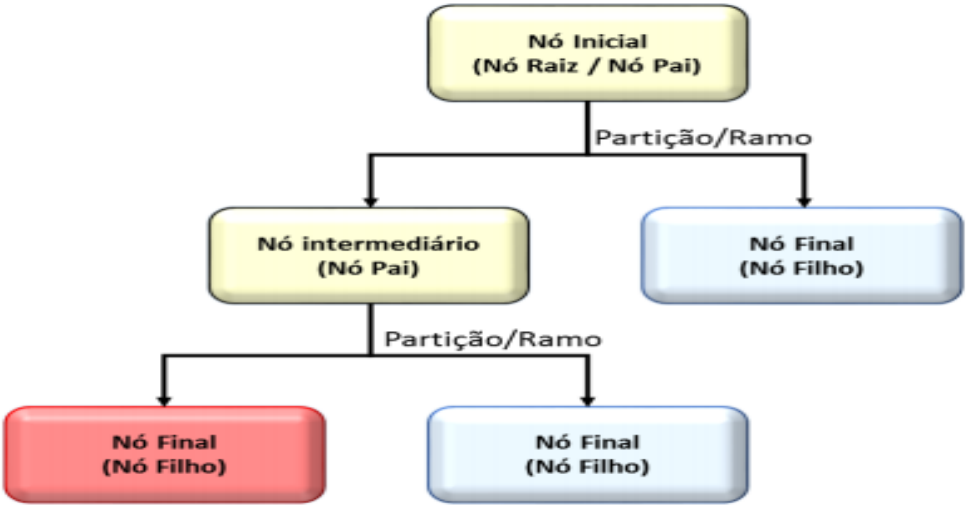

Figura 5 - Estrutura de uma árvore de decisão Fonte: Dantas e Donadia (2013)

3.2 Algoritmo de Indução de Árvores de Decisão

Existem diferentes propostas de algoritmos para indução de uma árvore de decisão, por exemplo: CHAID (*Chi-square Automatic Interaction Detection*) (Kass, 1980), CART (*Classification and Regression Trees*) (Breiman *et al*., 1984), ID3 (*Iteractive Dichotomizer* 3) (Quinlan, 1986), C4.5 (Quinlan, 1993). Neste trabalho, o algoritmo escolhido para a indução da árvore de decisão é chamado de CART (*Classification and Regression Trees*). Ele foi escolhido por ser de fácil implementação e estar disponível no pacote *rpart* do *software* R. As árvores geradas pelo algoritmo CART e baseada na técnica recursiva de divisão binária, as quais podem ser percorridas da sua raiz até as folhas respondendo as questões simples do tipo “sim” ou “não”.

Em cada passo do algoritmo, para selecionar qual variável (atributo) e de que forma essa irá dividir o nó, tem-se que identificar a partição que produz maior redução de heterogeneidade dos dados (impureza) no que diz respeito à distribuição das classes da variável resposta dentro dos nós criados. Para tal, faz-se necessária a utilização de alguma medida de impureza. Segundo Fonseca (1994), entre os critérios disponíveis para tal estão o método da entropia, o critério de

Gini, o método da paridade, a técnica de Laplace e também a escolha randômica. Dentre estes optou-se pela utilização do critério de Gini, visto que é bastante utilizado, mais favorável à otimização numérica, além de ser a medida padrão do algoritmo CART implementado no *software* R. (Hastie, Tibshirani e Friedman, 2001). O índice de Gini, avaliado em um nó t, é dado por

$$\phi(t) = 1 - \sum_{k=1}^{K} p_{t,k}^2 ,$$

em que $p_{t,k}$ representa a proporção de indivíduos de cada classe do nó t e K é o número de classes da variável resposta. No caso da regressão logística $K = 2$.

O índice de Gini assume seu valor máximo quando todas as classes da variável resposta possuem igual distribuição dentro do nó e assume o seu valor mínimo quando este é composto por apenas uma das suas classes. Depois de dividir em nó t em dois subconjuntos t_1 e t_2 com tamanhos n_1 e n_2, o índice de Gini dos dados divididos é definido como

$$\phi(t_1, t_2) = \frac{n_1}{n} \phi(t_1) + \frac{n_2}{n} \phi(t_2) \quad (8)$$

A regra selecionada para a primeira partição da amostra é aquela que maximizar, entre todas as candidatas, a redução do índice de Gini do nó pai com relação aos nós filhos. A partir desta regra, a amostra inicial será dividida em duas subamostras, nas quais o mesmo procedimento é aplicado sucessivamente. O algoritmo de indução da árvore de decisão é executado até que as divisões trouxerem ganhos em relação à medida do índice de Gini, comparando o valor para esta medida com um valor pré-estabelecido pelo usuário, denominado critério de parada.

O algoritmo CART, inicialmente, expande a árvore até que a quantidade máxima de nós finais seja alcançada (exaustão, saturação). O usuário pode estabelecer critérios para limitar a expansão dos nós da árvore como, por exemplo, determinar um número mínimo de elementos para que um nó possa ser particionado. A expansão exaustiva da árvore pode ocasionar partições que pouco contribuem para a explicação da variável resposta e, por vezes, refletem apenas ruídos ou erros. Quanto maior for a árvore, menor será o erro de classificação que ela apresentará, dado que o algoritmo procura separar a amostra em nós cada vez mais homogêneos, ficando assim muito específico para o conjunto de dados usados para indução da árvore (*overfitting*). Logo, essa árvore estendida excessivamente pode não apresentar um bom ajuste com relação à predição da variável resposta em novas amostras. A condição de *overfitting* pode ser evitada com o uso de podas da árvore, a fim de torná-la mais generalista e menos complexa.

Denomina-se poda o processo que consiste em desfazer as partições que menos contribuem para a explicação da variável resposta, resultando em uma sequência de subárvores de diferentes tamanhos (determinado pelo número de nós finais) que variam do tamanho máximo atingido pela árvore original até a menor árvore possível, isto é, aquela formada apenas pelo nó raiz. Para cada uma dessas subárvores é calculada um valor para uma função do tipo custo-complexidade (Breiman *et al*., 1984), a qual é calculada durante o processo de poda usando a técnica de validação cruzada. Ao final deste processo, será escolhida como árvore final aquela que permita minimizar o valor do erro de predição calculado através da função custo-complexidade utilizada. Detalhes sobre o processo de poda podem ser encontrados em, *e.g.*, Lauretto (2010) e Raimundo *et al.* (2008).

Enfim, em termos gerais, as características do algoritmo CART podem ser resumidas como:

- **Vantagens**: Não precisa realizar qualquer tipo de categorização, pois o algoritmo não faz restrições quanto às escalas das variáveis explicativas, podendo estas serem numéricas (discretas ou contínuas), ordinais e nominais. Por utilizar partições binárias, as variáveis podem aparecer em diferentes níveis do modelo, permitindo reconhecer diversas interações com outras variáveis.
- **Desvantagens**: Por utilizar partições binárias, pode aumentar a complexidade da árvore (muitos níveis de profundidade), o que pode dificultar a apresentação e interpretação dos resultados. Após a construção da árvore de decisão, esta pode se tornar demasiadamente específica aos dados utilizados e, via de regra, com alta complexidade. Esta condição é chamada de *overfitting*, ou super-ajuste, e pode ser evitada com o uso de podas da árvore, que a tornam mais generalista e menos complexa.

4 ESTUDO DE CASO

O estudo de caso apresentado neste trabalho tem o objetivo de ajustar um modelo de regressão logística e um modelo de árvore de decisão CART para a predição da probabilidade de inadimplência dos clientes de cartão de crédito de uma instituição financeira. Os dados foram coletados de um repositório público disponível na página da UCI (*Machine Learning Repository*), que é um conjunto de bancos de dados, teorias de domínio e geradores de dados usados pela comunidade de aprendizado de máquina para a análise empírica de algoritmos de aprendizado de máquina. O arquivo foi criado como um arquivo ftp em 1987 por *David Aha* e

outros estudantes da *UC Irvine*. Desde então, tem sido amplamente utilizado por estudantes, educadores e pesquisadores em todo o mundo como fonte primária de conjuntos de dados de aprendizado de máquina. O conjunto de dados desse trabalho se trata do histórico de pagamentos padrão mensais dos clientes usuários de cartão de crédito de uma instituição financeira em Taiwan, de abril de 2005 a setembro de 2005.

O objetivo geral da classificação é prever, a partir de um conjunto de atributos reunidos pela instituição, se o cliente está propenso a atrasar o pagamento padrão no próximo mês (cliente entrar em *default*) ou não (cliente não entrar em *default*). Portanto, como estamos interessados em analisar o risco de um cliente entrar em *default*, a análise é proposta para a classificação dos clientes em clientes credíveis ($Y = 0$) ou não credíveis ($Y = 1$).

O conjunto de dados de interesse contém informações para um total de 30.000 clientes de cartão de crédito da instituição a respeito de seu *status* quando ao *default*, fatores demográficos, dados de situação do crédito, saldo da conta bancária e histórico de pagamentos no período de abril de 2005 a setembro de 2005. Uma descrição completa de cada variável considerada neste estudo é apresentada na Tabela 1 com as respectivas categorias/escalas de medição consideradas. Do total de clientes na base de dados, 23.364 (77,88%) não entraram em *default*, sendo considerados clientes credíveis. Por outro lado, 6.636 (22,12%) clientes se caracterizam como clientes em *default*, sendo considerados como clientes não credíveis.

A base de dados original passou por um processo de limpeza com remoção dos indivíduos com valores ausentes para alguma das variáveis de interesse. Além disso, no processo de verificação de consistência do banco de dados, foi removido apenas um indivíduo que tinha valor muito discrepante na variável VLR_PAG_1, além de ter filtrado indivíduos entre -15 e 700 (4 observações) com valores muito discrepantes (*outliers*), da variável SALDO que foram removidos da análise, a fim de evitar efeitos meramente influenciados por tais observações. Vale ressaltar que nem sempre valores atípicos devem ser removidos da análise estatística, pois podem se tratar de *outliers* genuínos. No entanto, para o conjunto de dados em questão, cujo tamanho amostral é considerável, julgamos que tal procedimento não gera grande impacto e exclui apenas uma parcela insignificante dos clientes típicos da instituição. Ao final, restaram um total de 29.476 clientes, sendo 22.878 (77,61%) clientes credíveis ($Y = 0$) e 6.598 (22,39%) clientes não credíveis ($Y = 1$).

Tabela 1 - Descrição das variáveis e suas características

Variáveis	Descrição	Tipo	Categorias/Escalas
LIMITE_CAR	valor do limite de crédito concedido	quantitativa contínua	em milhares de dólares
SEXO	sexo	qualitativa categórica	1 - masculino
			2 - feminino
EDU	educação/grau de instrução	categórica	1 - pós - graduação
			2 - universidade
			3 - ensino médio
EST_CIVIL	estado civil	categórica	1 - casado
			2 - solteiro
			3 - outros
IDADE	idade do cliente	quantitativa-discreta	anos completos
STATUS_PAG_1	situação do pagamento dos clientes no mês anterior	categórica	1 - quitado
			2 - crédito rotativo
			3 - atraso ≤ 90dias
			4 - atraso > 90 dias
SALDO	saldo da conta do cliente	contínua	em milhares de dólares
VLR_PAG_1	valor da fatura paga em setembro de 2005	contínua	em milhares de dólares
VLR_PAG_2	valor da fatura paga em agosto de 2005	contínua	em milhares de dólares
VLR_PAG_3	valor da fatura paga em julho de 2005	contínua	em milhares de dólares
VLR_PAG_4	valor da fatura paga em junho de 2005	contínua	em milhares de dólares
VLR_PAG_5	valor da fatura paga em maio de 2005	contínua	em milhares de dólares
VLR_PAG_6	valor da fatura paga em abril de 2005	contínua	em milhares de dólares
Y	*default* próximo mês	categórica	0 - não
			1 - sim

4.1 Análise Descritiva da Base de Dados

No processo de análise estatística, mesmo que o objetivo final seja, por exemplo, a construção de um modelo de regressão logística, é importante analisar descritivamente as variáveis que compõem o banco de dados. Isto permite conhecer o comportamento individual de cada variável, bem como a sua associação com as demais variáveis que potencialmente irão compor o modelo. A análise descritiva possibilita uma noção prévia do perfil dos clientes que fazem parte da amostra. Nesta seção, é apresentada uma breve descrição das variáveis descritas na Tabela 1.

A análise do gráfico *box-plot*, Figura 6, evidencia que os clientes que entraram em *default* são um pouco mais homogêneos, visto que eles apresentam uma distância um pouco menor entre o 1° e 3° quartis. Nos clientes que não entraram em *default*, observa-se que o 3° quartil mais distante da mediana, indicando que no grupo de clientes não credíveis existe uma maior concentração em valores menores do saldo em conta. Os dois grupos apresentam diversos valores discrepantes, *outliers*, na cauda superior da distribuição dos dados, mas em nenhum dos grupos se observou clientes com valores atípicos na direção de valores negativos para o saldo.

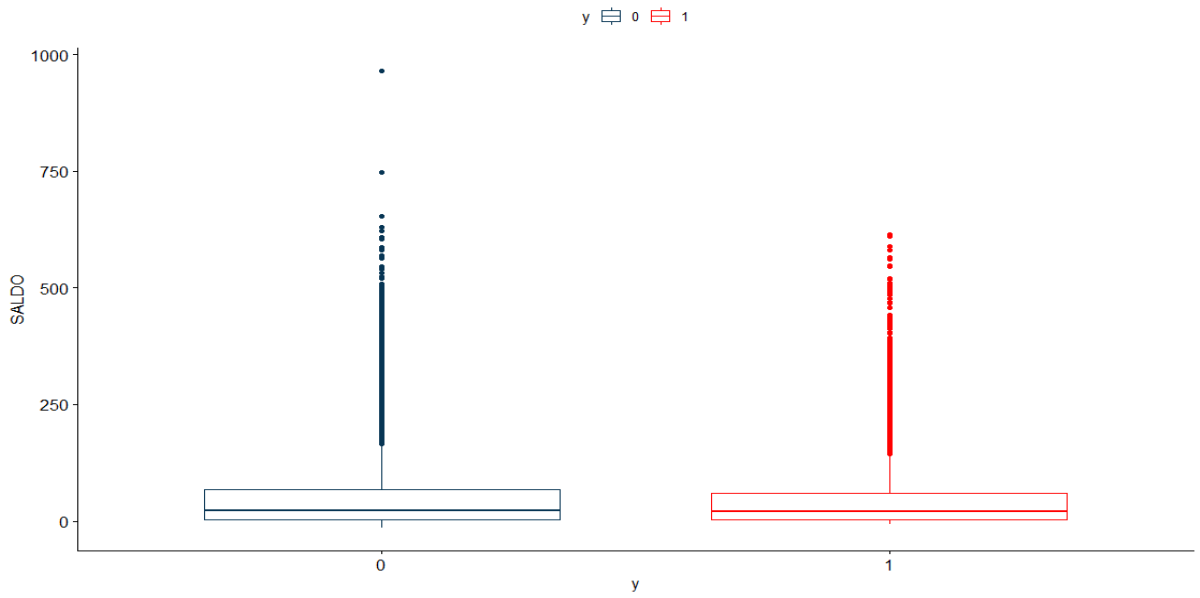

Figura 6 - Gráfico *box-plot* variável SALDO por tipo de cliente (credível 0 e não credível 1)

A análise da Tabela 2 também destaca a assimetria na distribuição da variável SALDO, dado que as médias e medianas são bastante discrepantes nos dois grupos. Como esperado, a

grande concentração de valores atípicos na cauda superior da distribuição dos dados provoca uma elevação da média amostral da variável em ambos os grupos.

Tabela 2 - Tabela de medidas descritivas variável SALDO

Y	minimo	1° quartil	mediana	média	3° quartil	máximo	desvio-padrão
0	-15,31	3,68	23,05	51,72	68,52	964,51	73,22
1	-6,67	2,97	20,04	48,30	59,09	613,86	73,63

Através dos gráficos *box-plot* que contam na Figura 7 observa-se que a variável LIMITE_CAR tem uma grande concentração em valores no intervalo (0,500), tanto para clientes credíveis quanto não credíveis. Os clientes que entraram em *default* têm o 3° quartil mais distante da mediana. Os clientes que não entraram em *default*, apresentam uma distância um pouco maior entre o 1° e 3° quartis, indicando maior variabilidade. Os dois grupos apresentam diversos valores discrepantes, na cauda superior da distribuição dos dados. Como evidenciado na Tabela 3, a variabilidade da variável LIMITE_CAR é maior no grupo dos clientes credíveis, os quais apresentam média e mediana maior que os clientes não credíveis.

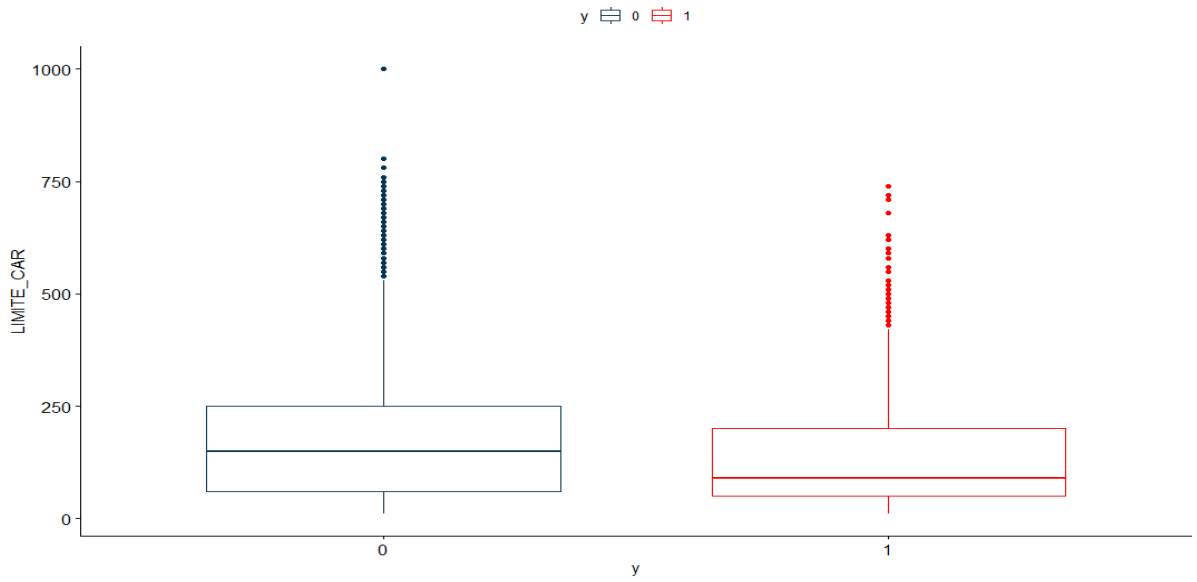

Figura 7 - Gráfico de dispersão variável LIMITE_CAR por tipo de cliente (credível 0 e não credível 1)

Tabela 3 - Tabela de medidas descritivas variável LIMITE_CAR

Y	minimo	1° quartil	mediana	média	3° quartil	máximo	desvio-padrão
0	10	60	150	178,05	250	1000	131,93
1	10	50	90	130,11	200	740	115,46

Com base na Figura 8, é possível perceber que a maior parte dos clientes se concentram entre as idades de 20 a 50 anos, nos dois grupos. Observa-se que a variável IDADE apresenta um pico próximo dos 30 anos, depois um decaimento. Isso mostra que a distribuição não é simétrica. Os dois grupos são bem homogêneos conforme demonstrado na análise descritiva que consta na Tabela 4.

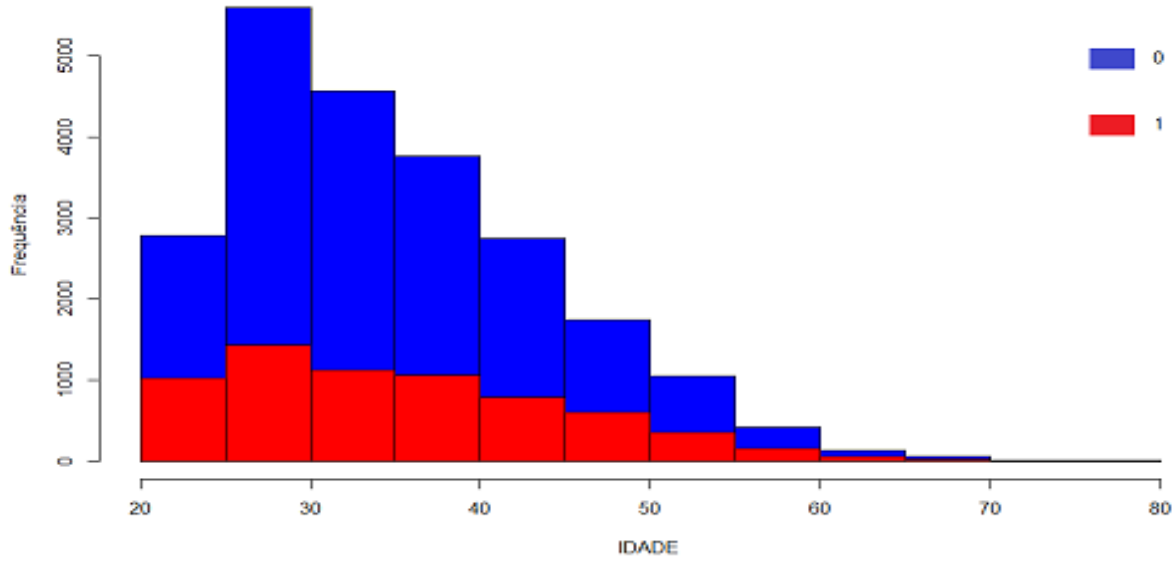

Figura 8 - Gráfico histograma variável IDADE por tipo de cliente (credível 0 não credível 1)

Tabela 4 - Tabela de medidas descritivas variável IDADE

Y	minimo	1° quartil	mediana	média	3° quartil	máximo	desvio-padrão
0	21	28	34	35,40	41	79	9,07
1	21	28	34	35,71	42	75	9,69

A análise dos gráficos *box-plot* mostrados na Figura 9 evidencia que, para ambos os tipos de clientes, todas as variáveis relacionadas aos valores pagos em faturas mensais apresentam alta concentração em valores baixos, com grande presença de valores atípicos na cauda superior da distribuição dos dados. A Tabela 5 traz algumas medidas descritivas das variáveis VLR_PAG_1 a VLR_PAG_6.

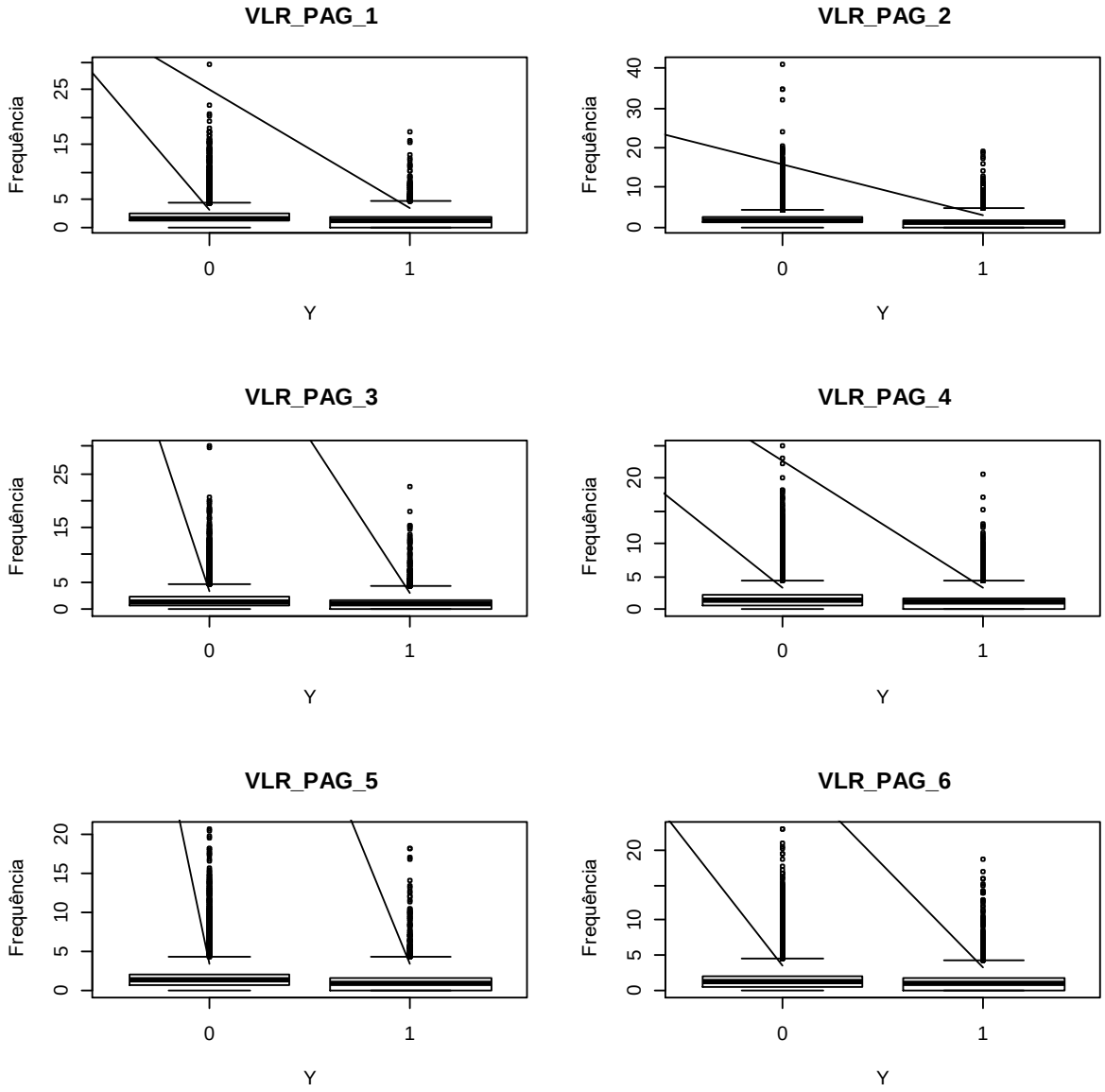

Figura 9 - Gráficos *box-plot* variável VLR_PAG_1 a VLR_PAG_6. Para fins de visualização, nesta figura foi considerada a raiz quadrada dos valores observados para as variáveis.

Tabela 5 - Tabela de medidas descritivas das variáveis VLR_PAG_1 a VLR_PAG_6

variáveis	minimo	1° quartil	mediana	média	3° quartil	máximo	desvio-padrão
VLR_PAG_1	0,00	1,00	2,10	5,62	5,00	873,55	16,28
VLR_PAG_2	0,00	0,82	2,00	5,88	5,00	1684,25	23,08
VLR_PAG_3	0,00	0,39	1,80	5,17	4,50	896,04	17,52
VLR_PAG_4	0,00	0,30	1,50	4,82	4,01	621,00	15,73
VLR_PAG_5	0,00	0,26	1,50	4,79	4,03	426,52	15,24
VLR_PAG_6	0,00	0,13	1,50	5,18	4,00	528,66	17,68

A variável denominada SEXO, conforme Tabela 6, é possível observar que 24,42% dos clientes que entraram em *default* é da categoria masculino, e 21,05% pertence a categoria feminino. Quanto aos clientes que não entraram em default 75,58% é da categoria masculino, e 78,95% pertence a categoria feminino. Pode ser notado que a categoria feminino representa 60,30% do total de clientes, contra 39,70 da categoria masculino.

Tabela 6 - Distribuição de clientes em função de Y conforme variável SEXO

SEXO		*default* próximo mês		Total
		Não	Sim	
masculino_1	quantidade	8.846	2.857	11.703
	% na categoria	75,58	24,42	100,00
feminino_2	quantidade	14.032	3.741	17.773
	% na categoria	78,95	21,05	100,00
Total	quantidade	22.878	6.598	29.476
	% na categoria	77,61	22,39	100,00

A segunda variável foi denominada de EST_CIVIL, está variável apresenta 3 categorias conforme Tabela 7 é possível observar que 23,75% dos clientes que entraram em *default* são da categoria casados, e 21,12% da categoria solteiros. Além disso, a categoria outros representa apenas 1,06% do total de clientes da variável. Os clientes dessa categoria são considerados divorciados e viúvos.

Tabela 7 - Distribuição de clientes em função de Y conforme variável EST_CIVIL

EST_CIVIL		*default* próximo mês		Total
		Não	Sim	
casados_1	quantidade	10.234	3.189	13.423
	% na categoria	76,25	23,75	100,00
solteiros_2	quantidade	12.413	3.325	15.738
	% na categoria	78,88	21,12	100,00
outros_3	quantidade	231	84	315
	% na categoria	73,33	26,67	100,00
Total	quantidade	22.878	6.598	29.476
	% na categoria	77,61	22,39	100,00

Nessa variável denominada EDU, apresenta 3 categorias. Conforme a Tabela 8, mostra que 19,24% dos clientes que entraram em *default* são da categoria pós-graduação, esse percentual aumenta na categoria universidade para 23,74%, e também na categoria ensino

médio 25,30%, ou seja, a proporção de clientes por categoria que entraram em *default* mostra que a categoria ensino médio é maior que todas as outras.

Tabela 8 - Distribuição de clientes em função de Y conforme variável EDU

EDU		*default* próximo mês		Total
		Não	Sim	
pós-graduação_1	quantidade	8.544	2.036	10.580
	% na categoria	80,76	19,24	100,00
universidade _2	quantidade	10.694	3.329	14.023
	% na categoria	76,26	23,74	100,00
ensino médio_3	quantidade	3.640	1.233	4.873
	% na categoria	74,70	25,30	100,00
Total	quantidade	22.878	6.598	29.476
	% na categoria	77,61	22,39	100,00

Na variável STATUS_PAG_1, apresenta 4 categorias. As informações da Tabela 9 mostram que os clientes da categoria atraso de 1 até 90 dias, apresentam uma proporção elevada de clientes que entraram em *default* 50,43%. Além disso, a categoria atraso maior que 90 dias tem uma proporção muito elevada de clientes que entraram em *default* 64,03%, porém representa apenas 0,48% do total de clientes da variável.

Tabela 9 - Distribuição de clientes em função de Y conforme variável STATUS_PAG_1

STATUS_PAG_1		*default* próximo mês		Total
		Não	Sim	
quitado_1	quantidade	6.977	1.312	8.289
	% na categoria	84,17	15,83	100,00
crédito_rotativo _2	quantidade	12.578	1.866	14.444
	% na categoria	87,08	12,92	100,00
atraso ≤ 90dias _3	quantidade	3.273	3.331	6.604
	% na categoria	49,57	50,43	100,00
atraso > 90 dias _4	quantidade	50	89	139
	% na categoria	35,97	64,03	100,00
Total	quantidade	22.878	6.598	29.476
	% na categoria	77,61	22,39	100,00

4.2 Ajuste do Modelo de Regressão Logística

Selecionada a amostra de interesse, é usual utilizar uma parte dela para ajuste (treinamento) e outra parte para validação (teste) do modelo. Neste trabalho, a amostra foi aleatoriamente dividida em dois grupos correspondentes a 80% e 20% do total para comporem os grupos de treinamento e de validação, respectivamente. Assim, os dados de 23.580 clientes compuseram a amostra de treinamento e as informações de 5.896 clientes foram reservadas para validação do modelo. Tanto a seleção aleatória dos grupos de treinamento e validação quanto o ajuste do modelo foram feitos utilizando o *software* R (*Core Team,* 2019).

Ao ajustar o modelo, inicialmente, todas as variáveis apresentadas na Tabela 1 foram incluídas. O procedimento de retirada das variáveis que se mostraram não significativas, ao nível de significância de 5%, foi feito passo a passo a partir do modelo completo, obedecendo a ordem de magnitude dos valores-p observados segundo o teste de Wald descrito na Seção 2.2. A primeira variável a ser retirada do modelo foi EDU por corresponder o maior p-valor. Baseado no mesmo critério, a próxima variável a ser excluída do estudo será a variável IDADE, indicando que ela não tem diferenciação entre as categorias *default* e não *default*. Essa falta de significância é observada na análise descritiva. Prosseguindo com o processo de escolha das variáveis, obteve-se que a variável VLR_PAG_3, VLR_PAG_4, VLR_PAG_5 e VLR_PAG_6, não foram significativas. Optou-se por considerar apenas a informação das variáveis VLR_PAG_1 e VLR_PAG_2.

O resultado do ajuste para o modelo resultante é apresentado na Tabela 10. De acordo com os resultados do teste de significância, para o nível de significância de 5%, os coeficientes de todos os termos do modelo foram considerados estatisticamente significativos, ou seja, a hipótese nula de que o coeficiente é igual a zero é rejeitada. A exceção diz respeito ao coeficiente para a categoria EST_CIVIL_OUTROS, indicando que não existe diferença significativa entre aos clientes casados (categoria basal da variável) e clientes divorciados ou viúvos (EST_CIVIL_OUTROS).

Tabela 10 - Tabela do modelo final com as variáveis selecionadas

Variáveis	Estimativa do Coeficiente	Erro Padrão	Teste Wald	p-valor	Razão de Chance
LIMITE_CAR	-0,003	0,000	-18,824	0,000	0,996
SEXO_FEMININO	-0,150	0,034	-4,326	0,000	0,860
EST_CIVIL_SOLTEIROS	-0,192	0,034	-5,549	0,000	0,825
EST_CIVIL_OUTROS	-0,071	0,159	-0,446	0,656	0,931
STATUS_PAG_CRÉD. ROTATIVO	-0,712	0,051	-13,949	0,000	0,490
STATUS_PAG_ATRASO ≤ 90 DIAS	1,255	0,047	26,278	0,000	3,507
STATUS_PAG_ATRASO > 90 DIAS	1,338	0,206	6,467	0,000	3,812
SALDO	0,003	0,000	12,081	0,000	1,003
VLR_PAG_1	-0,011	0,002	-5,075	0,000	0,988
VLR_PAG_2	-0,012	0,002	-5,514	0,000	0,987
CONSTANTE	-0,708	0,055	-12,832	0,000	-

Após análise a de significância dos coeficientes do modelo, proceder-se-á à interpretação dos coeficientes estimados para cada variável em termos de razão de chances (*odds ratio*) descrita na Seção 2.4. Os valores observados estão apresentados na sexta coluna na Tabela 10.

Para a primeira variável LIMITE_CAR. A *odds ratio* indica que, mantendo as demais variáveis constantes, um cliente com uma unidade (1.000 dólares) a mais no limite do cartão de crédito tem uma chance 0,99 vezes menor de entrar em *default* se comparado a um cliente sem esta unidade adicional no limite de crédito.

Na variável SEXO. A *odds ratio* indica que a probabilidade dos clientes do sexo feminino entrarem em *default* é 0,86 vezes a chance dos clientes do sexo masculino, mantidas todas as outras variáveis constantes, ou seja, 14% inferior.

Na interpretação da variável EST_CIVIL, (i) *odds ratio* quando comparamos a probabilidade da categoria solteiro entrarem em default é 0,82 vezes que os clientes da categoria casado, mantida todas as outras variáveis constantes, (ii) a *odds ratio* quando comparamos a probabilidade da categoria outros entrarem em *default* é 0,93 vezes inferior que os clientes da categoria casado, no entanto este efeito não é significativo (p-valor = 0,656).

A variável STATUS_PAG_1, (i) a *odds ratio* quando comparamos a probabilidade categoria crédito rotativo entrarem em *default* é 0,49 vezes que os clientes da categoria quitado, mantida todas as outras variáveis constantes e (ii) a *odds ratio* quando comparamos a probabilidade da categoria atraso até 90 dias entrarem em default é 3,50 vezes superior que os

clientes da categoria quitados e (iii) a *odds ratio* quando comparamos a probabilidade da categoria atraso maior que 90 dias entrarem em default é 3,81 vezes superior que os clientes da categoria quitados.

Com a variável SALDO, a *odds ratio* indica que quando se aumenta o saldo da conta do cliente em 1 unidade (1.000 dólares) a chance do cliente entra em *default* é 0,003 vezes maior. Apesar de ter uma magnitude pequena, temos que o efeito da variável SALDO é significativo. O sinal positivo do coeficiente estimado para esta variável indica que clientes com mais saldo na conta tem uma chance maior de entrarem em *default* (não pagar o débito no mês seguinte) do que clientes com saldos mais baixos. Este é um resultado inconsistente com o que se espera na prática e pode ter sido gerado como efeito do comportamento numérico da variável entre clientes credíveis e não credíveis no banco de dados utilizado ou confundimento gerado por uma possível multicolinearidade entre as covariáveis consideradas no ajuste. Esta última hipótese parece ser menos factível pois, mesmo para um modelo que inclui apenas a variável saldo como covariável (resultado omitido do texto), tem-se que o coeficiente associado é negativo e significativo.

Na variável VLR_PAG_1, a *odds ratio* mostra que mantendo as demais variáveis constantes, quando aumenta o valor do pagamento do mês passado, em 1 unidade (1.000 dólares no valor da fatura) a chance do cliente entra em *default* é 0,98 vezes, ou seja, a cada unidade paga a mais diminui em quase 2% a chance do cliente entra em *default*.

A variável VLR_PAG_2, a *odds ratio* mostrou-se similar a VLR_PAG_1, ou seja, em 1 unidade (1.000 dólares no valor da fatura) a chance do cliente entra em *default* é 0,98 vezes, ou seja, a cada unidade paga a mais diminui em quase 2% a chance do cliente entra em *default*.

4.2.1 Diagnóstico de Ajuste do Modelo Logístico

Uma vez ajustado o modelo, é necessário avaliar a sua adequabilidade e a sua qualidade preditiva. Nesta seção, apresentamos algumas das medidas e métodos usualmente utilizados nesta avaliação.

4.2.1.1 Função Resposta Estimada

Na Figura 10 é apresentada a relação entre as probabilidades estimadas e os escores associados à curva logística estimada. Os escores são definidos como as transformações

logística da função apresentada na equação (2). Temos que a função resposta é monotônica, porém com uma forma de S (sigmoidal) não muito marcada, o que é desejado para indicar uma boa qualidade do ajuste (veja discussão feita com base na Figura 1). A curva obtida no ajuste, Figura 10, tem uma cauda mais extensa à esquerda, levando à uma concentração maior de probabilidades estimadas em valores pequenos, abaixo de 0,50. Este resultado pode indicar que o ajuste do modelo não é adequado aos dados observados que foram considerados. Contudo, cabe ressaltar que esta é uma análise descritiva e que pode ser utilizada conjuntamente com outras técnicas de análise da qualidade do ajuste, como as que serão discutidas nas seções seguintes.

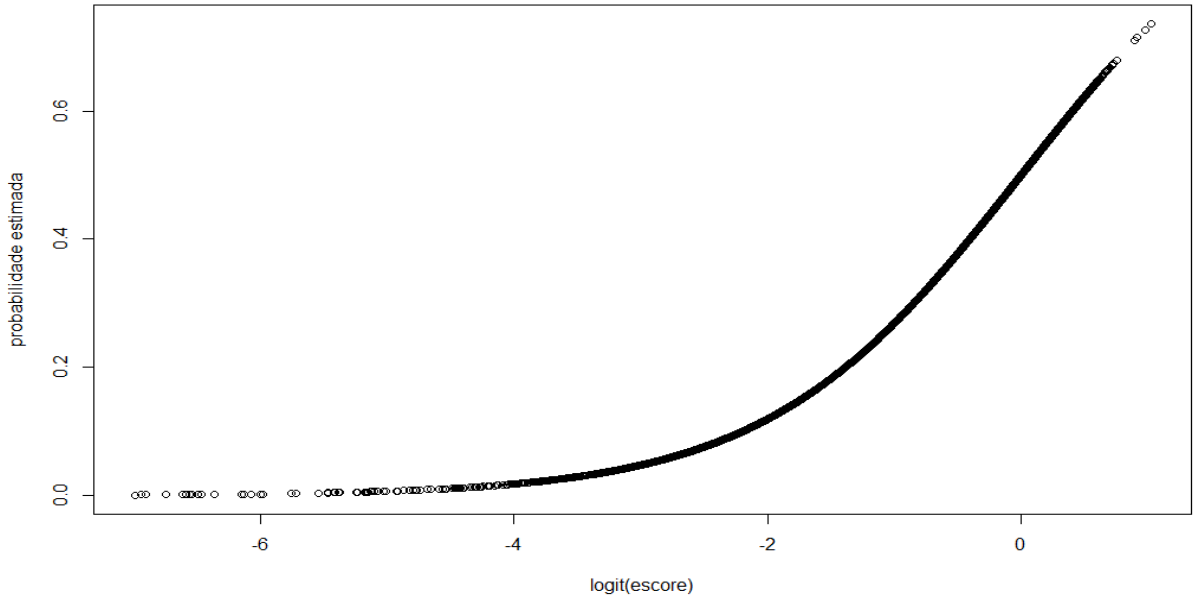

Figura 10 – Gráfico da curva logística modelo ajustado

4.2.1.2 Teste de Hosmer – Lemeshow

Um teste paramétrico para verificar a adequabilidade do ajuste de um modelo de regressão logistica é o Teste de Hosmer – Lemeshow descrito na Seção 2.3.1. O detalhamento do cálculo do teste Hosmer – Lemeshow referente ao modelo constante na Tabela 10 é apresentado na Tabela 11. Os 23.580 clientes do conjunto de treinamento foram divididos em 10 grupos e a estatística de teste obtida foi de 10,609. Como, o p-valor foi de 0,224, não se pode rejeitar a hipótese nula de que o modelo se ajusta bem aos dados, ao nível de significância de 5%.

Tabela 11 – Cálculo da Estatística do Teste de Hosmer – Lemeshow

Grupo	O-Pagamento padrão: Sim	O-Pagamento padrão: Não	Total	E-Pagamento padrão: Sim	E-Pagamento padrão: Não	Total
1	2.223	135	2.358	2.230,023	127,976	0,408
2	2.147	211	2.358	2.131,040	226,959	1,242
3	2.100	258	2.358	2.073,504	284,495	2,806
4	2.028	330	2.358	2.036,406	321,593	0,254
5	2.015	343	2.358	2.005,756	352,243	0,285
6	1.941	417	2.358	1.976,124	381,875	3,855
7	1.910	448	2.358	1.928,181	429,818	0,941
8	1.760	598	2.358	1.741,732	616,267	0,733
9	1.208	1.150	2.358	1.203,674	1.154,325	0,032
10	970	1.388	2.358	975,556	1.382,443	0,054
Total Geral	18.302	5.278	23.580	Teste Hosmer – Lemeshow		10,609
				Graus de liberdade		8
				p-valor		0,224

4.2.1.3 Curva ROC Regressão Logística

A curva ROC, detalhada no item 2.3.2, foi calculada para o modelo ajustado com os dados de treinamento, 80% dos dados totais. O resultado é apresentado na Figura 11 e traz a curva da sensibilidade e a curva equivalente a (1 – especificidade), que representa os falsos alarmes (clientes não credíveis classificados como credíveis, temos que a área abaixo da curva ROC é de 74,61%. Conclui-se então que a discriminação é aceitável dentro do critério apresentado na Seção 2.3.2.

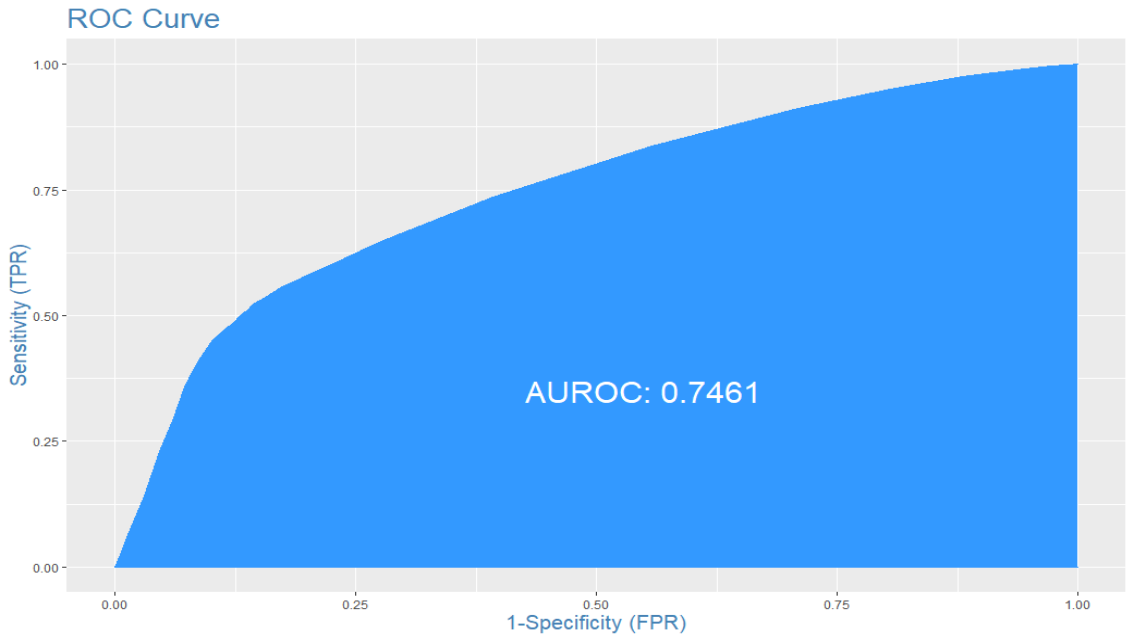

Figura 11 – Gráfico da curva ROC do modelo de regressão logística ajustado

4.2.1.4 Matriz de Classificação Regressão Logística

Outra informação utilizada para avaliar a qualidade do modelo é a matriz de classificação, a qual apresenta o total de erros e acertos do modelo ajustado em ambos os grupos de clientes credíveis e não credíveis. A partir do valor de probabilidade preditos pelo modelo de regressão logística, os clientes são classificados em cada um dos grupos de acordo com um limiar para esta probabilidade chamado de ponto de corte. Se a probabilidade predita pelo modelo é superior ao limiar especificado, então o cliente é classificado como não credível (*default*, $Y = 1$) e, caso contrário, é classificado como credível (não *default*, $Y = 0$). Foram elaboradas várias tabelas com pontos de corte distintos conforme discutido na sequência.

Em geral, o ponto de corte 0,50 é utilizado por se tratar do valor que divide a escala das probabilidades exatamente ao meio, mas outras escolhas podem ser feitas em cada contexto. Neste trabalho, além do ponto 0,50, nós consideramos o ponto de corte que maximiza simultaneamente a sensibilidade e especificidade do modelo ajustado. Para este fim, após uma análise similar àquela ilustrada na Figura 3, definiu-se o ponto de corte 0,19 por ser o valor onde, aproximadamente, as curvas de sensibilidade e especificidade se cruzam. Para fins de comparação, considerou-se também os pontos de corte 0,35 e 0,15. Os resultados são apresentados nas Tabelas 12 a 15.

Com a Tabela 12 vemos que o ponto de corte 0,50 retorna que, no geral, 80,07% dos clientes estão bem classificados, sendo a sensibilidade e a especificidade iguais a 39,64% e 91,73%, respetivamente.

Tabela 12 - Matriz de classificação ponto de corte de 0,50

Observado			Predito			
			default próximo mês			
			Não	Sim	% de acerto	% de erro
default próximo mês	Não	18.302	16.789	3.186	91,73	8,27
	Sim	5.278	1.513	2.092	39,64	60,36
Total			**18.302**	**5.278**	**80,07**	**19,93**

Com relação à Tabela 13, vê-se que a diminuição do ponto de corte para 0,35 têm-se uma taxa de acerto global de 78,83%, sendo a sensibilidade e a especificidade iguais a 49,22% e 87,36%, respetivamente.

Tabela 13 - Matriz de classificação ponto de corte de 0,35

Observado			Predito			
			default próximo mês			
			Não	Sim	% de acerto	% de erro
default próximo mês	Não	18.302	15.989	2.680	87,36	12,64
	Sim	5.278	2.313	2.598	49,22	50,78
Total			**18.302**	**5.278**	**78,83**	**21,17**

Para o ponto de corte 0,19, Tabela 14, o modelo mostra uma taxa de acerto global de 73,04%, sendo a sensibilidade e a especificidade iguais a 61,25% e 76,43%, respetivamente.

Tabela 14 - Matriz de classificação ponto de corte de 0,19

Observado			Predito			
			default próximo mês			
			Não	Sim	% de acerto	% de erro
default próximo mês	Não	18.302	13.989	2.045	76,43	23,57
	Sim	5.278	4.313	3.233	61,25	38,75
Total			**18.302**	**5.278**	**73,04**	**26,96**

Enfim, temos que com o ponto de corte 0,15 um total de 58,30% dos clientes são bem classificados, sendo a sensibilidade e a especificidade iguais a 79,14% e 52,29%, respetivamente.

Tabela 15 - Matriz de classificação ponto de corte de 0,15

Observado			Predito			
			default proximo mês			
			Não	Sim	% de acerto	% de erro
default próximo mês	Não	18.302	9.570	1.101	52,29	47,71
	Sim	5.278	8.732	4.177	79,14	20,86
Total			**18.302**	**5.278**	**58,30**	**41,70**

De um modo geral, analisando os resultados apresentados nas Tabelas 12 a 15, nota-se que a medida em que o ponto de corte aumenta, o percentual de acerto dos clientes que não entraram em *default* apresenta aumento, mas, por outro lado, o percentual de acerto dos clientes que entraram em *default* diminui. Seria interessante escolher um ponto de corte que forneça um maior equilíbrio na classificação em ambos os sentidos. Isto porque, para um modelo com uma sensibilidade muito alta e uma especificidade baixa, teria - se uma boa classificação dos clientes não credíveis, mas, ao mesmo tempo, uma má classificação dos clientes credíveis. Neste cenário, a instituição teria vantagem em termos de evitar possíveis perdas com clientes potencialmente não credíveis, mas deixaria de ter possíveis ganhos com clientes credíveis para os quais ela deveria conceder o crédito. Numa situação contrária, para um modelo com uma sensibilidade baixa e uma especificidade alta, a empresa tenderia a errar muito ao conceder crédito para clientes potencialmente não credíveis, mas, ao mesmo tempo, ela não teria perdas no sentido de deixar de conceder crédito a clientes bons pagadores. Buscando um equilíbrio entre as duas situações, o ponto de corte 0,19 parece ser o mais adequado para o modelo de regressão logística ajustado.

4.2.2 Análise da Predição nos Dados de Validação Regressão Logística

Considerando o ponto de corte de 0,19 foi feita a predição para os dados de validação, 20% dos dados totais, usando o modelo ajustado com os dados de treinamento, 80% dos dados totais. O desempenho nos dados de teste é semelhante ao desempenho nos dados de validação,

comparando-se os percentuais de acerto mostrados nas Tabelas 14 e 16. Isto indica que o modelo fornece predições consistentes.

Tabela 16 - Matriz de classificação para 20% da amostra ponto de corte de 0,19

Observado		Predito			
		default proximo mês			
		Não	Sim	% de acerto	% de erro
default próximo mês	Não 4.576	3.451	535	75,42	24,58
	Sim 1.320	1.125	785	59,47	40,53
Total		**4.576**	**1.320**	**71,85**	**28,15**

4.3 Resultados da Análise da Árvore de Decisão CART

Nesta seção, apresentamos o resultado para o ajuste do modelo utilizando o método CART para definição de uma árvore de classificação. Foi utilizada sua implementação disponível no pacote *rpart* do *software* R. Inicialmente, todas as variáveis constantes na Tabela 1 foram incluídas no ajuste. O processo de indução da árvore final utilizou o índice de Gini como medida de impureza, seguindo o procedimento descrito na Seção 3.

Através da aplicação do método é possível obter uma medida geral de importância de cada variável. Segundo a documentação do pacote *rpart* (Therneau e Atkinson, 2019), o valor de importância de uma variável é a soma das medidas de bondade da partição (árvore final) para cada partição na qual a variável foi uma variável primária, somado com as medidas de bondade em todas as divisões nas quais ela foi uma variável secundária. As três variáveis com representação superior a 1% com relação a soma de todos os valores de importância são mostradas na Tabela 17. A variável com maior relevância para o modelo é a variável STATUS_PAG_1, seguida por SALDO e VLR_PAG_1. Ao longo do processo de indução da árvore, apenas as duas primeiras dentre estas variáveis se mostraram relevantes e permaneceram na formação da árvore final, a qual é apresentada na Figura 12.

Tabela 17 - Tabela das variáveis importantes com valores e porcentagem

STATUS_PAG_1	SALDO	VLR_PAG_1
1.130,85 (69,8%)	322,30 (19,9%)	155,68 (9,6%)

A árvore final Figura 12 apresenta três folhas (nós finais) que podem ser utilizados para classificação dos clientes em credíveis ou não credíveis. Cada nó da árvore (representado por um retângulo) traz a informação sobre: linha superior, a classificação binária dos clientes alocados naquele nó, onde 0 representa cliente credível e 1 representa cliente não credível; na linha central, o percentual dos indivíduos alocados naquele nó que são de fato clientes credíveis (0) e não credíveis (1), respectivamente; e na linha inferior, o percentual do total de indivíduos que foram alocadas para aquele nó.

Vemos que a variável com maior valor de importância, STATUS_PAG_1, gera a primeira divisão nos dados: clientes pertencentes às categorias 1 e 2 desta variável são classificados como 0, ou seja, clientes credíveis. Esta partição determina a Folha 2, localizada à esquerda e que englobou 77% dos clientes (18.180). Destes, 86% foram classificados corretamente de acordo com sua situação real observada. Na sequência, a variável SALDO foi utilizada para definir as outras duas folhas da árvore induzida: clientes com saldo inferior a 1.000 dólares foram classificados como credíveis (Folha 4) e os demais foram classificados como não credíveis (Folha 5). A Folha 4 englobou 6% do total de clientes (1.451), sendo que para 28% deles a classificação do modelo foi feita de forma errada. Os demais 17% dos clientes (3.949) foram alocados à Folha 5 e classificados como não credíveis, tendo um percentual de acerto de 59% dentre eles.

Assim como no resultado obtido via regressão logística, nota-se uma situação inconsistente quanto ao efeito da variável SALDO na classificação: clientes com saldo mais elevado apresentam uma chance maior de ser um cliente não credível.

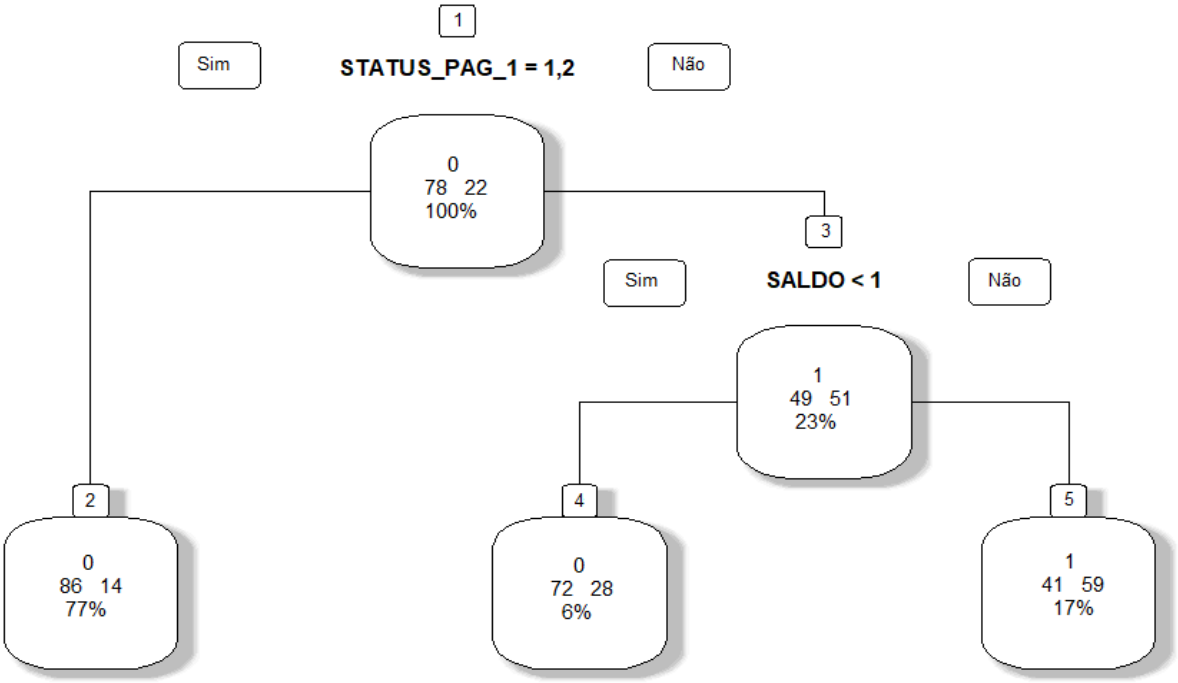

Figura 12 - Gráfico árvore de decisão do modelo ajustado com classificação 0 e 1, total de acertos e erros e porcentagem de dados

Vale ressaltar que a árvore final mostrada na Figura 12 foi analisada quanto à possibilidade de poda, mas o ajuste não se modifica neste caso. Para diferentes configurações do algoritmo, sempre temos apenas as variáveis STATUS_PAG_1 e SALDO fazendo parte da classificação da forma mostrada na Figura 12. Este, portanto, é considerado como sendo o melhor ajuste de uma árvore de decisão para os dados considerados. Na sequência, apresentamos uma avaliação sobre a qualidade do ajuste obtido.

4.3.1 Diagnóstico de Ajuste da Árvore de Decisão CART

Uma vez ajustado o modelo, é necessário avaliar a sua adequabilidade e a sua qualidade preditiva assim como foi feito para o modelo de regressão logística. Nesta seção apresentamos algumas das medidas e métodos usualmente utilizados nesta avaliação.

4.3.1.1 Curva ROC da Árvore de Decisão CART

A curva ROC foi calculada para o modelo ajustado com os dados de treinamento, 80% dos dados totais. O resultado é apresentado na Figura 13, temos que a área abaixo da curva ROC é de 69,63%. Conclui-se então que a discriminação é aceitável de acordo com o critério especificado na Seção 2.3.2.

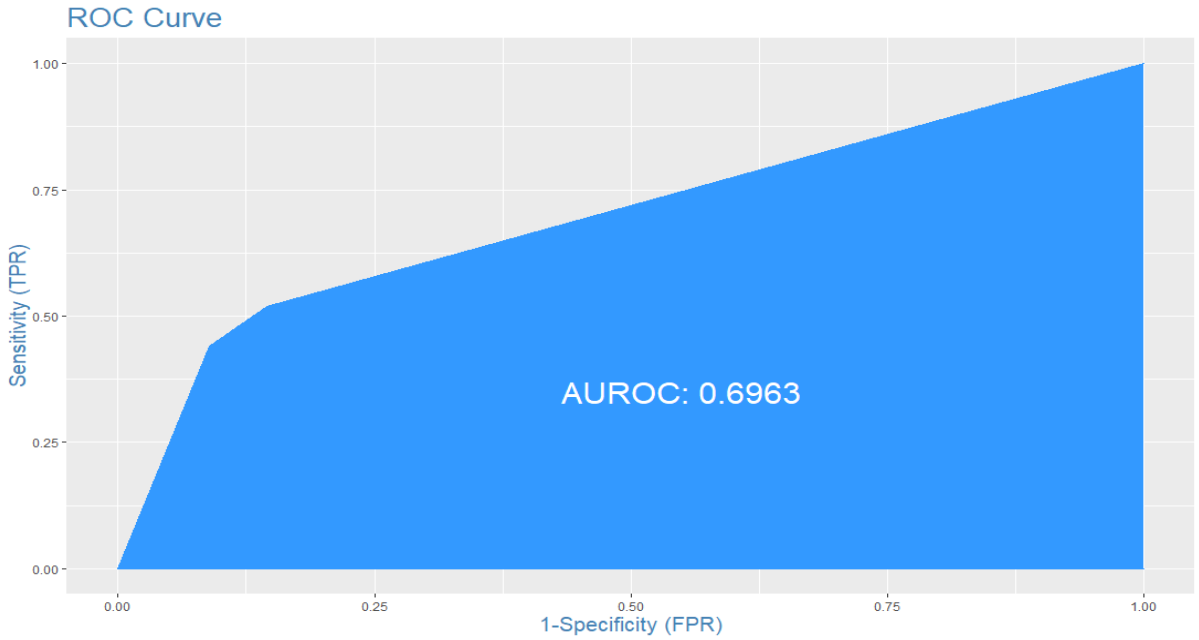

Figura 13 – Gráfico da curva ROC do modelo de árvore de decisão ajustado

4.3.1.2 Matriz de Classificação da Árvore de Decisão CART

Outra informação utilizada para avaliar a qualidade do ajuste é a matriz de classificação, a qual apresenta o total de erros e acertos do modelo ajustado em ambos os grupos de clientes credíveis e não credíveis para um determinado ponto de corte pre-determinado. Para avaliação da adequação da árvore de decisão ajustada foram considerados apenas os pontos de corte 0,50 e 0,19. Este último é aproximadamente o ponto em que as curvas de sensibilidade e especificidade se cruzam no caso da árvore de decisão, seguindo o que foi discutido na figura 3. Os resultados são apresentados, respectivamente, nas Tabelas 18 e 19. Verifica-se que, para o ponto de corte de 0,50, temos 80,63% dos clientes que estão bem classificados, sendo a sensibilidade e a especificidade iguais a 44,14% e 91,15%, respectivamente. E considerando o

ponto de corte de 0,19 os resultados são 77,98% dos clientes que estão bem classificados, sendo a sensibilidade e a especificidade 51,97% e 85,48% respetivamente. Os resultados com o ponto de corte 0,19 parecem ser mais adequado por deixarem equilibradas a sensibilidade e especificidade.

Tabela 18 - Matriz de classificação ponto de corte de 0,50

Observado			Predito			
			default próximo mês			
			Não	Sim	% de acerto	% de erro
default próximo mês	Não	18.302	16.683	2.948	91,15	8,85
	Sim	5.278	1.619	2.330	44,14	55,86
Total			**18.302**	**5.278**	**80,63**	**19,37**

Tabela 19 - Matriz de classificação ponto de corte de 0,19

Observado			Predito			
			default proximo mês			
			Não	Sim	% de acerto	% de erro
default próximo mês	Não	18.302	15.645	2.535	85,48	14,52
	Sim	5.278	2.657	2.743	51,97	48,03
Total			**18.302**	**5.278**	**77,98**	**22,02**

4.3.2 Análise da Predição nos Dados de Validação da Árvore de Decisão CART

Considerando o ponto de corte de 0,19 foi feita a predição para os dados de validação, 20% dos dados totais, usando o modelo ajustado com os dados de treinamento, 80% dos dados totais. O desempenho nos dados de teste é semelhante ao desempenho nos dados de validação, comparando-se os respectivos percentuais de acerto, conforme resultados das Tabelas 19 e 20, mostrando consistência na capacidade preditiva da árvore ajustada.

Tabela 20 - Matriz de classificação para 20% da amostra ponto de corte de 0,19

Observado		Predito			
		default proximo mês			
		Não	Sim	% de acerto	% de erro
default próximo mês	Não 4.576	3.910	643	85,44	14,56
	Sim 1.320	666	677	51,28	48,72
Total		**4.576**	**1.320**	**77,80**	**22,20**

5 CONCLUSÕES

Este trabalho teve como objetivo aplicar os métodos estatístico de regressão logística e árvore de decisão na classificação de clientes usuários de cartão de crédito a fim de comparar a precisão preditiva da probabilidade de *default* no pagamento do próximo mês.

Os modelos de árvore de decisão são mais simples de se ajustar no sentido de não ser necessária uma pré-seleção de variáveis nem suposição de uma distribuição de probabilidades para a modelagem da variável resposta de interesse. Neste trabalho, o algoritmo CART foi utilizado na indução da árvore de decisão. O algoritmo busca as variáveis mais importantes de acordo com a medida de importância calculada para cada uma. O método CART limita o tamanho da árvore usando técnicas de validação cruzada para evitar sobreajuste dos dados**.** Outra característica interessante deste modelo é sua fácil apresentação e interpretação dos resultados.

Para o ajuste no modelo de regressão logística, assume-se que a variável resposta segue um modelo de probabilidade Bernoulli, cuja probabilidade de sucesso corresponde à probabilidade de ocorrência do evento de interesse. No processo de construção do modelo de regressão é recomendado realizar uma seleção prévia das variáveis. As variáveis são incluídas no modelo para serem avaliadas quanto a sua significância, verificada através de testes estatísticos apropriados. No caso da regressão, procedimentos inferenciais, como o teste de Hosmer – Lemeshow, podem ser utilizados para verificar a adequabilidade do modelo ajustado aos dados observados.

Com relação ao ajuste feito para o conjunto de dados reais selecionados, conclui-se que os dois métodos são aplicáveis para prever a probabilidade de inadimplência (*default*). A avaliação da capacidade preditiva dos modelos foi feita utilizando a medida da área sobre a curva ROC e a matriz de classificação, da qual se obtém medidas de sensibilidade e especificidade. Os resultados se mostraram satisfatórios e bastante similares nos dois casos, porém a regressão logística teve desempenho superior à árvore de decisão CART na nossa aplicação.

Na prática, quando esses modelos são aplicados de forma adequada, impactam diretamente na operação e lucratividade da empresa. No contexto da aplicação considerada no estudo de caso (Seção 4), vale discutir a respeito da utilidade dos modelos pelo fato deles envolverem variáveis associadas ao uso prévio de crédito pelo cliente. Na instituição financeira, ele serviria para avaliar e prever melhor a probabilidade de inadimplência dos clientes atuais

que já utilizam o cartão de crédito, bem como identificar os principais fatores que determinam essa probabilidade. Isso informaria as decisões do emissor sobre a quem conceder um cartão de crédito e qual limite de crédito fornecer baseado na situação anterior. Também ajudaria o emissor a ter um melhor entendimento de seus clientes atuais e potenciais, o que informaria sua estratégia futura, incluindo seu planejamento de oferta de produtos de crédito direcionados a seus clientes. No âmbito do problema relacionado a novos clientes do produto cartão de crédito, existem algumas limitações importantes a serem consideradas, uma vez que o modelo exige mensurações de características com uso prévio do cartão de crédito. Dessa forma, para novos clientes, o uso dos modelos da forma em que foram ajustados careceria da informação referente as relações de crédito dos clientes com outras instituições financeiras, por exemplo, quando este tipo de informação estiver disponível.

Enfim, tanto o método da regressão logística, quando a árvore de decisão tem grande potencial de utilização no contexto prático da análise de risco de crédito. Alguns detalhes técnicos necessários à aplicação dos métodos foram apresentados neste trabalho com aplicação a um único conjunto de dados. Diversos outros exemplos podem ser encontrados na literatura, evidenciando a versatilidade e generalidade da aplicação de ambos os métodos.

REFERÊNCIAS

Breiman *et al*. **Classification and Regression Trees**. California: Wadsworth International, 1984, 358 p.

Caouette, J. B., Altman, E. I. e Narayanan, P. **Gestão do risco de crédito: o próximo grande desafio financeiro**. Rio de Janeiro: Qualitymark, 1999.

Cristiano, M. V. M. B. **Sensibilidade e Especificidade na Curva ROC Um Caso de Estudo**. 2017. Dissertação (Mestrado em Gestão de Sistemas de Informação Médica) - Escola Superior de Tecnologia e Gestão, Instituto Politécnico, Leiria, 2017.

Dantas D. e Donadia E. A. **Comparação entre as técnicas de Regressão Logística, Árvore de Decisão, Bagging e Random Forest Aplicadas a um Estudo de Concessão de Crédito**. 2013. Trabalho de Conclusão de Curso (Bacharel em Estatística) - Departamento de Estatística, Universidade Federal do Paraná, Curitiba, 2013.

Fonseca, J. M. M. R. **Indução de Árvores de Decisão**. 1994. Dissertação (Mestrado) - Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Lisboa, 1994.

Hair *et al*. Análise multivariada de dados. 6. ed. São Paulo: Bookman, 2009, 284 – 285 p.

Hastie, T. e Tibshirani, R.; Friedman, J. **The Elements of Statistical Learning**. New York: Springer, 2001, 731 p.

Hosmer, D. W. e Lemeshow, S. Goodness of fit tests for the multiple logistic regression model, **Communications in Statistics - Theory and Methods**, v. 9, n. 10, p. 1043-1069, 1980. DOI: 10.1080 / 03610928008827941. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/03610928008827941. Acesso em: 02 jul. 2020.

Hosmer, D. W., Lemeshow, S. e Sturdivant, R. X. **Applied Logistic Regression**. 3. ed. Hoboken: John Wiley & Sons, 2013, 500 p.

Kass, G. An exploratory technique for investigating large quantities of categorical data. **Applied Statistics**. v. 29, n. 2, p. 119-127, 1980. DOI: https://doi.org/10.2307/2986296. Disponível em: https://www.jstor.org/stable/2986296. Acesso em: 25 nov. 2020.

Lauretto, M. S. **Árvores de Decisão**. Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, São Paulo, 2010. Disponível em: https://edisciplinas.usp.br/pluginfile.php/4469825/mod_resource/content/1/ArvoresDecisao_normalsize.pdf. Acesso em 10 de nov. de 2020.

McCullagh P. e Nelder, J. A. **Generalized Linear Models**. London: Chapman and Hall, 1989.

Mingoti, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. 1. ed. Belo Horizonte: UFMG, 2005, 297 p.

Paiva, C. C. V. **Previsão da Inadimplência através da Regressão Logística**. 2015. Monografia (Especialização em Estatística) - Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, 2015.

Quinlan, J. R. Introduction of decision trees. **Machine Learning**, vol. 1, n. 1, p. 81-106, 1986. DOI: https://doi.org/10.1007/BF00116251. Disponível em: https://link.springer.com/article/10.1007/BF00116251#citeas. Acesso em: 25 nov. 2020.

Quinlan, J. R. **C4.5**: **Programs for machine learning**. San Mateo: Morgan Kaufmann Publishers, 1993, 302 p.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2019. Disponível em: https://www.R-project.org/. Acesso em: 20 jun. 2020.

Raimundo *et. al*. **O Algoritmo de Classificação CART em uma Ferramenta de Data Mining.** Unidade Acadêmica de Ciências, Engenharias e Tecnologias - Universidade do Extremo Sul Catarinense, Criciúma; Curso de Engenharia da Computação - Universidade Federal do Pampa, Pelotas, 2008. Disponível em: https://core.ac.uk/download/pdf/236392249.pdf. Acesso em: 10 nov. 2020.

Sicsú, A. L. **Credit Scoring: desenvolvimento, implantação, acompanhamento**. 4. ed. São Paulo: Blucher, 2010, 84 p.

Taconeli, C. A. **Árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridade e entropia**. Tese (Doutorado em Agronomia) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2008.

Tan, P., Steinbach, M. e Kumar, V. **Introdução ao Data Mining**. Rio de Janeiro: Ciência Moderna, 2009.

Therneau T. M., Atkinson E. J. e Foundation M. rpart: **An Introduction to Recursive Partitioning Using the RPART Routines**. R package version 3.6.0, 2019. Disponível em: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf. Acesso em: 25 nov. 2020.

Yeh, I. C. e Lien, C. H. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. **Expert Systems with Applications**, v. 36 n. 2, p. 2473-2480, 2009. Disponível em: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients. Acesso em: 05 jun. 2020.