

O MODELO DE *Collection Score* COMO FERRAMENTA DE RECUPERAÇÃO DE CRÉDITO

Roberta Maria e Silva e Oliveira

Centro Universitário Jorge Amado

rmsoliveira@gmail.com

Carlos Alberto Lima da Silva

Universidade Estadual de Feira de Santana - Departamento de Saúde

carlosls.compos@gmail.com

Paulo Henrique Ferreira da Silva

Universidade Federal da Bahia - Departamento de Estatística

paulohenri@ufba.br

Francisco Louzada

Universidade de São Paulo - Instituto de Ciências Matemáticas e de Computação

louzada@icmc.usp.br

RESUMO

O controle do risco de crédito, que consiste no risco do tomador de um empréstimo não honrar seus compromissos conforme acordado, é prioridade atual das instituições financeiras. Neste artigo, como uma etapa na gestão do relacionamento com o cliente, propõe-se a aplicação de uma metodologia de classificação dos clientes de uma instituição financeira brasileira, no processo de recuperação de crédito inadimplido, quando implantados procedimentos de contato para cobrança por telefone, correspondência e/ou mensagem eletrônica. Essa avaliação do comportamento do cliente durante uma operação de crédito, com o intuito de recuperação do crédito inadimplido, é conhecida como *Collection Score*. Para alcançar esse objetivo, aplica-se um modelo estatístico de classificação conhecido como modelo de Regressão Logística Multinomial (RLM), cuja variável resposta é politômica, ou seja, com três ou mais níveis de classificação. A estimação dos parâmetros do modelo RLM é realizada pelo método da máxima verossimilhança, com seleção de variáveis/modelo pelo método *stepwise (backward e forward)* de acordo com o critério AIC, e teste de eficiência do modelo classificatório por meio de várias medidas de avaliação de desempenho, tais como, acurácia, sensibilidade, especificidade, eficiência, valores de predição positivo e negativo, medida F e coeficiente de correlação de Matthews, curva ROC, Brier score, medida de concordância com a estatística Kappa, teste de Wald, *odds ratio* e análise dos resíduos através da técnica de envelopamento representado com o gráfico quantil-quantil. Considera-se uma base de contratos de três operações de crédito com características semelhantes e que estavam inadimplentes em janeiro de 2017. O modelo final ajustado apresentou resultados satisfatórios, com probabilidade de acerto de classificação de 0,8308 e Brier score de 0,2519, que são indicativos da evidência de que o modelo produz previsões calibradas.

ABSTRACT

The control of credit risk, which is the risk that the borrower does not honor its commitments as agreed upon, is the current priority of financial institutions. In this paper, as a step in the Customer Lifecycle Management, it is proposed the application of a methodology to classify the customers of a Brazilian financial institution in the process of recovering delinquent credit when implementing contact procedures for collection by telephone, correspondence or e-mail. This evaluation of the customer's behavior during a credit operation to recover the defaulted credit is known as Collection Score. In order to achieve this goal, a statistical classification model known as the Multinomial Logistic Regression (RLM) model is applied, whose dependent variable is polytomic, i.e., with three or more classification levels. Thus, with a base on contracts of three credit operations with similar characteristics, which were in default in January 2017, we estimate the parameters of the proposed MLR model by the maximum likelihood method. Stepwise (backward and forward) model selection is made via Akaike's information criterion. The efficiency test of the classification model is performed via several measures of performance evaluation, such as accuracy, sensitivity, specificity, efficiency, positive and negative predictive values, F-measure and Matthews correlation coefficient, ROC curve, Brier score, measure of agreement with the Kappa statistic, Wald test, odds ratio and residual analysis through the enveloping technique represented with the quantile-quantile plot. The final fitted model showed satisfactory results, with the probability of correct classification (accuracy) of 0.8308 and Brier score of 0.2519, indicating that there is evidence that the model performs calibrated forecasts.

Palavras-chave: Acordos de Basileia, Gestão do Relacionamento com o Cliente, Inadimplência, Regressão logística multinomial.

1 INTRODUÇÃO

Atualmente, as instituições do mercado financeiro, tais como, operadoras de cartões de crédito, bancos, concessionárias de crédito, investem na retenção de clientes, visto que já se comprovou através de pesquisas de *marketing*, que é mais lucrativo conhecer e manter suas carteiras de clientes do que buscar clientes novos, cujo custo é muito elevado, conforme fundamentado por Pinho [1]. Para isso, é de suma importância o Gerenciamento do Relacionamento com o Cliente (GRC, ou ainda CRM, do inglês *Customer Relationship Management*), que se origina no *Marketing* de Relacionamento, que consiste em um conjunto de estratégias criadas para gerenciar a interação entre uma instituição e os seus clientes/consumidores.

O GRC envolve tecnologia de informação, processos de negócios e atitude empresarial, somando forças para gerar diferencial competitivo por meio do relacionamento com os clientes, com foco na forma de se relacionar e nas adaptações internas necessárias decorrentes disso. A implementação de *software* de gestão e de obtenção de conhecimento prevê maneiras mais eficientes de se efetivar o relacionamento cliente-empresa (Xavier & Dornelas [2]).

Neste contexto, o objetivo do GRC é fornecer elementos referenciais, funcionais e de projeção, que possibilitem à instituição financeira definir estratégias adequadas para explorar as potencialidades de seus clientes.

Em dezembro de 2010, o Comitê de Basileia divulgou dois documentos: "Basileia III: Uma estrutura reguladora global para bancos e sistemas bancários mais resilientes" (em inglês, *Basel III: A global regulatory framework for more resilient banks and banking systems*) e "Basileia III: Estrutura internacional para medição, padrões e monitoramento do

risco de liquidez” (em inglês, *Basel III: International framework for liquidity risk measurement, standards and monitoring*), conhecidos como Basileia III.

A Basileia III visa o aperfeiçoamento da capacidade das instituições financeiras absorverem choques provenientes do próprio sistema financeiro ou dos demais setores da economia, reduzindo o risco de transferência de crises financeiras para a economia real (Banco Central do Brasil, BCB).

Segundo Pinheiro *et al.* [3], essas medidas (ou novas regras) fazem com que as instituições financeiras tenham que planejar melhor suas ações de investimento e de concessão de crédito, priorizando a contratação de ativos que proporcionem uma melhor relação entre retorno e risco. Contudo, as ações não estão limitadas à revisão da política de investimento e crédito.

O Acordo de Basileia III entrou efetivamente em vigor no Brasil em 1º de outubro de 2013. A data marca o início de uma longa fase de transição para os novos padrões prudenciais fixados pelo Comitê de Basileia, fase esta que deve ser concluída integralmente somente em 2022, de acordo com a Associação Brasileira das Entidades dos Mercados Financeiro e de Capitais (ANBIMA [4]).

A crise econômica, que tem desestabilizado o Brasil desde o início de 2014, levou o país a uma das suas piores recessões econômicas desde os anos 1930, havendo recuo do Produto Interno Bruto (PIB) por mais de um ano consecutivo. Segundo o jornal *Financial Times*, em 2015 a economia brasileira recuou em 4,5%, sendo seu recorde naquele período. Em setembro de 2016, a taxa de desemprego alcançou 11,8%, atingindo 12 milhões de brasileiros. Esse cenário econômico e o aumento do desemprego podem impactar diretamente nos índices de inadimplência nas linhas de crédito das instituições financeiras no Brasil.

Segundo Oliveira & Louzada [5], para o lançamento de uma linha de crédito, as instituições financeiras primeiramente realizam uma pesquisa de mercado; em seguida, aplicam modelos de *Credit Score* para concessão de crédito a clientes novos; modelos de *Behavior Score*, que visam ampliar a fidelização e receitas com os clientes já captados; e por fim, modelos de *Collection Score*, que, através da Estatística, buscam aprimorar os processos de cobrança e recuperação de créditos inadimplentes.

Posto isso, as instituições financeiras buscam minorar os impactos de eventos fortuitos e malquistos na produção de resultados, gerenciando a concessão de crédito por meio, dentre outras, do GRC e aplicando modelos estatísticos de *Collection Score*.

Neste artigo, identifica-se quais são os elementos que influenciam com eficiência na utilização do GRC para o controle e redução da inadimplência ou para o aumento e acompanhamento da adimplência em uma determinada instituição financeira no Brasil, visando contribuir nos processos de recuperação de clientes, a fim de minimizar os impactos de eventos inesperados e indesejados na geração de seus resultados, conforme preconizam o Acordo de Basileia (regulamento mundial) e as Resoluções CMN (Conselho Monetário Nacional) nº 2.682 de 22/12/1999 e CMN nº 3.721 de 30/04/2009 divulgadas pelo BCB.

Para recuperar clientes e os créditos inadimplidos, demonstra-se como classificar o cliente que paga uma prestação de um contrato de crédito em atraso a partir do momento da intervenção realizada pela equipe de recuperação de crédito da instituição financeira em questão, a qual forneceu os dados para o presente estudo de forma sigilosa. Devido a essa cláusula de sigilo, o nome da instituição, bem como sua localização não são divulgados. Desta forma, pode-se identificar as premissas que auxiliam na eficácia do GRC, na meta de controle/redução da inadimplência por meio da modelagem de regressão logística multinomial (RLM).

Este artigo está organizado da seguinte forma. A Seção 2 descreve a metodologia utilizada no desenvolvimento deste trabalho. A Seção 3 apresenta e discute os principais resultados alcançados. Considerações finais e conclusão compõem a última seção do artigo.

2 MATERIAL E MÉTODOS

2.1 POPULAÇÃO E DESENHO DO ESTUDO

As instituições financeiras costumam realizar campanhas para redução da inadimplência, como também utilizam de diversas formas de contato com seus clientes a fim de mantê-los informados e atentos quanto ao cumprimento dos prazos para pagamento das prestações de seus produtos e/ou serviços contratados. Neste contexto, é necessário que essas instituições, com o auxílio da tecnologia da informação, disponham de sistema que armazene os dados dos clientes de forma detalhada, de modo que possibilite contatá-los da maneira que eles melhor desejarem. Por exemplo, por constar que o cliente prefere se comunicar por *e-mail* ou pelo celular em determinados dias e horários.

Questiona-se neste artigo como o GRC pode auxiliar no controle/redução da inadimplência em uma instituição financeira no Brasil, buscando identificar quais princípios que corroboram a sua efetividade.

Para alcançar tal objetivo, aplica-se o método estatístico conhecido como *Collection Score* (ou ainda, modelo de cobrança), que é uma prática dentro das instituições financeiras para recuperação de dívidas existentes. Segundo Souza [6], o modelo de *Collection Score* tem por finalidade identificar a probabilidade de pagamento dos clientes que já se tornaram inadimplentes. Isto significa que a população-alvo do modelo de cobrança é aquela formada por clientes que não cumpriram com as suas obrigações de pagamentos nos prazos combinados com as instituições credoras.

Portanto, esse modelo indica a probabilidade de cada cliente pagar ou não as suas dívidas, definindo como *bom pagador* o cliente que sai da situação de inadimplência e como *mau pagador* aquele que não consegue restabelecer sua situação original (isto é, de cliente adimplente). Tal modelo possibilita também priorizar quais os clientes que devem ser cobrados, reduzindo, assim, o número de contatos desnecessários com o cliente, tais como telefonemas ou emails, evitando assim o atrito com clientes bons pagadores (com menos de 30 dias de atraso), e aumentando a relação entre custo e benefício durante as atividades das equipes de cobrança.

Em concordância com Machado [7], os modelos de *Collection Score* também podem ser utilizados para monitorar e administrar o portfólio de clientes inadimplentes, uma vez que, além de identificarem clientes em risco, são capazes de reconhecer clientes “curados”, gerando aumento de receita e de relacionamento, reduzindo a rotatividade e o cancelamento (*churn*). Isso é muito importante para as instituições, uma vez que manter clientes é mais barato do que obter clientes novos. Neste sentido, o relacionamento de longo prazo ganha dimensão importante devido ao benefício mútuo.

O presente trabalho aplica o modelo de *Collection Score* numa linha de crédito de uma instituição financeira no Brasil, a qual disponibilizou uma base de dados de clientes que apresentavam até 270 dias de inadimplência no final da primeira quinzena de janeiro de 2017 em três tipos de operações de crédito, com safras de 2014 a 2016, permitindo que fossem observadas as intervenções realizadas pelas equipes de cobrança nos meses de janeiro e fevereiro de 2017, conforme o passo-a-passo a seguir:

- **Passo 1.** Identificar o momento de inadimplência;
- **Passo 2.** Identificar se houve pagamento espontâneo (isto é, sem intervenção das equipes de cobrança);
- **Passo 3.** Identificar o primeiro momento de cobrança em até 30 dias após o Passo 1;
- **Passo 4.** Identificar se houve pagamento após a ocorrência do Passo 3.

2.2 MODELAGEM

Segundo Hosmer *et al.* [8], considera-se, para cada indivíduo i , um conjunto de p variáveis independentes designadas pelo vetor $\mathbf{x}_i = (x_{1,i}, x_{2,i}, \dots, x_{p,i})^\top$, com $i = 1, 2, \dots, N$. Para o momento, assume-se que cada uma dessas variáveis é pelo menos escalada em intervalos. Definindo que a probabilidade condicional de inadimplência esteja presente e seja denotada por $\pi(\mathbf{x}_i) = P(Y_i = 1 | \mathbf{x}_i)$, o *logit* do modelo de regressão logística múltipla é dado pela equação:

$$g(\mathbf{x}_i) = \log \left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_p x_{p,i}.$$

Sendo assim, segue que:

$$\pi(\mathbf{x}_i) = \frac{\exp\{g(\mathbf{x}_i)\}}{1 + \exp\{g(\mathbf{x}_i)\}}.$$

Se algumas das variáveis independentes são discretas, ou variáveis de escala nominal, como sexo, estado civil, escolaridade, raça, e assim por diante, não é apropriado incluí-las no modelo como se fossem variáveis contínuas. Os números usados para representar os vários níveis dessas variáveis de escala nominal são meramente identificadores e não têm significado numérico. Nesta situação, o método de escolha é usar um conjunto de variáveis categóricas, conhecidas como variáveis *dummy* (Hosmer *et al.* [8]).

Para o modelo de *Collection Score*, a variável resposta (ou dependente) é se os clientes efetuaram o pagamento da prestação em atraso com ou sem a intervenção das equipes de cobrança. Para isso, utiliza-se o modelo RLM, usual para análise de eventos politômicos, isto é, com mais de duas respostas categóricas. Com o auxílio da função *logit*, que descreve a probabilidade de ocorrência de um evento, o modelo informa qual a probabilidade de um cliente ser classificado como *Ótimo*, *Bom*, *Ruim* ou *Péssimo*, em que:

- **Categoria 0. Péssimo:** O cliente inadimplente não efetuou o pagamento após a intervenção das equipes internas de cobrança;
- **Categoria 1. Ruim:** O cliente inadimplente não efetuou o pagamento sem a intervenção das equipes internas de cobrança;
- **Categoria 2. Bom:** O cliente inadimplente efetuou o pagamento após a intervenção das equipes internas de cobrança;
- **Categoria 3. Ótimo:** O cliente inadimplente efetuou o pagamento sem a intervenção das equipes internas de cobrança.

Como descrito em Bittencourt [9], com adaptações para este estudo, é definido o modelo RLM cuja variável resposta Y_i assume quatro níveis, sendo eles: 0, 1, 2 e 3, para $i = 1, 2, \dots, N$. Agora, o modelo logístico tem três funções *logit*: a razão entre $Y_i = 1$ e $Y_i = 0$, a razão entre $Y_i = 2$ e $Y_i = 0$, e a razão entre $Y_i = 3$ e $Y_i = 0$. Neste caso, o nível $Y_i = 0$ é assumido como base (ou referência). Os parâmetros β 's são estimados pelo método da máxima verossimilhança. Assim, seguem as funções *logit* dadas por:

$$\begin{aligned} g_1(\mathbf{x}_i) &= \log \left(\frac{P(Y_i = 1 | \mathbf{x}_i)}{P(Y_i = 0 | \mathbf{x}_i)} \right) = \beta_{10} + \beta_{11} x_{1,i} + \beta_{12} x_{2,i} + \dots + \beta_{1p} x_{p,i}, \\ g_2(\mathbf{x}_i) &= \log \left(\frac{P(Y_i = 2 | \mathbf{x}_i)}{P(Y_i = 0 | \mathbf{x}_i)} \right) = \beta_{20} + \beta_{21} x_{1,i} + \beta_{22} x_{2,i} + \dots + \beta_{2p} x_{p,i}, \\ g_3(\mathbf{x}_i) &= \log \left(\frac{P(Y_i = 3 | \mathbf{x}_i)}{P(Y_i = 0 | \mathbf{x}_i)} \right) = \beta_{30} + \beta_{31} x_{1,i} + \beta_{32} x_{2,i} + \dots + \beta_{3p} x_{p,i}. \end{aligned}$$

2.3 BANCO DE DADOS

Os dados do sistema de recuperação de crédito da instituição financeira em questão foram disponibilizados através de um arquivo com extensão CSV (*comma-separated values*). Visando manter o sigilo dos dados, as variáveis foram codificadas e, em seguida, ordenadas, categorizadas e/ou relativizadas (ver Tabela 1). Numa análise preliminar da base de dados, identificou-se que havia mais de um contrato inadimplente para um mesmo cliente. Baseado na dissertação de mestrado de Machado [7], após adaptações, foram utilizadas as seguintes premissas como critérios de desempate:

- **Premissa 1:** Deixar o contrato com a data de inadimplência mais antiga;
- **Premissa 2:** Caso o cliente tenha mais de um contrato com a mesma data de inadimplência, considerar aquele com maior tempo de contratação;
- **Premissa 3:** Caso persista o empate, considerar aquele com maior exposição no momento de inadimplência.

Após a aplicação das premissas acima, a base de dados foi finalizada com $N = 214$ clientes com um contrato para cada e três operações de crédito efetivadas entre os anos de 2014 e 2016.

TABELA 1: Variáveis disponíveis no banco de dados.

Variável (Código)	Classificação	Escala	Função no Modelo RLM
VAR002	Qualitativa	Nominal	Variável Independente
VAR004	Qualitativa	Nominal	Variável Independente
VAR005	Qualitativa	Nominal	Variável Independente
VAR009	Qualitativa	Ordinal	Variável Independente
VAR010	Qualitativa	Ordinal	Variável Independente
VAR011	Quantitativa	Contínua	Variável Independente
VAR012	Quantitativa	Discreta	Variável Independente
VAR013	Quantitativa	Discreta	Variável Independente
VAR014	Quantitativa	Contínua	Variável Independente
VAR015	Quantitativa	Contínua	Variável Independente
VAR016	Quantitativa	Contínua	Variável Independente
VAR017	Qualitativa	Nominal	Variável Independente
VAR018	Qualitativa	Nominal	Variável Independente
VAR027	Qualitativa	Ordinal	Variável Independente
VAR028	Quantitativa	Contínua	Variável Independente
VAR030	Quantitativa	Contínua	Variável Independente
VAR024	Qualitativa	Nominal	Variável Dependente

2.4 ANÁLISE DE DADOS

A análise dos dados e o ajuste do modelo de classificação proposto foram realizados com o auxílio do *software* estatístico R (R Core Team [10]). Para isso, a base de dados foi dividida, através de sorteio, pelo método de amostragem aleatória estratificada proporcional, em que 70% dos clientes foram para o banco de dados de treinamento do modelo e 30%, para o banco de dados de teste do modelo.

No processo de desenvolvimento do modelo, definiu-se como classe de referência a categoria 0 (*Péssimo*). Para estimar o modelo inicial, foi empregada a função *multinom(.)* do pacote “nnet” (Ripley & Venables [11]) do *software* R. Para a obtenção do modelo final, considerou-se o método de seleção passo-a-passo *stepwise* (*backward* e *forward*) segundo o critério de informação de Akaike (AIC, do inglês *Akaike information criterion*), mediante o uso da função *stepAIC(.)* do pacote “MASS” (Ripley [12]) do R. Com a definição do modelo final, foi gerada a chamada “matriz de confusão” a partir do banco de dados de teste, em

que se estabeleceu como predições negativas as classificações *Péssimo* (0) e *Ruim* (1), e como predições positivas as classificações *Bom* (2) e *Ótimo* (3). Em seguida, conforme a literatura, avaliou-se a capacidade preditiva do modelo por meio das seguintes medidas de desempenho: acurácia (ACC), sensibilidade (SEN), especificidade (SPE), eficiência (EFF), valor preditivo positivo (VPP), valor preditivo negativo (VPN), medida F e coeficiente de correlação de Matthews (MCC), além da curva ROC (*receiver operating characteristic curve*), a AUC (*area under the curve*), o Brier score e a estatística Kappa.

A matriz de confusão 2×2 é apresentada na Tabela 2, em que: VN = Verdadeiro Negativo, FP = Falso Positivo, FN = Falso Negativo, e VP = Verdadeiro Positivo.

TABELA 2: Matriz de confusão para classificação binária.

		Predição	
		Negativa	Positiva
Observação	Negativa	VN	FP
	Positiva	FN	VP

Segundo Silva *et al.* [13], a ACC é a probabilidade de ocorrer uma classificação correta; a SEN é a proporção de verdadeiros dentre todos os exemplos cuja classe esperada é a classe positiva; a SPE é a proporção de rejeições corretas; a EFF é a média aritmética entre a SEN e a SPE; o VPP é a proporção de acertos ou de VP's dentre todos os classificados como positivos; o VPN é a proporção de rejeições ou de VN's dentre os exemplos classificados como negativos; a medida F é a média harmônica entre a SEN e o VPP; e, por fim, o MCC é a taxa de classificação geral do modelo, que retorna um valor no intervalo $[-1, +1]$, em que um MCC de +1 corresponde a uma predição perfeita, 0 representa uma predição aleatória média, e -1 uma predição inversa. A seguir, são exibidas as expressões utilizadas para o cálculo dessas medidas de performance:

$$ACC = \frac{VN + VP}{VN + FP + FN + VP}, \quad SEN = \frac{VP}{VP + FN}, \quad SPE = \frac{VN}{VN + FP}, \quad EFF = \frac{SEN + SPE}{2},$$

$$VPP = \frac{VP}{VP + FP}, \quad VPN = \frac{VN}{VN + FN}, \quad F = \frac{2 \times SEN \times VPP}{SEN + VPP} \quad e$$

$$MCC = \frac{(VP \times VN) - (FP \times FN)}{\sqrt{(VP + FP) \times (VP + FN) \times (VN + FP) \times (VN + FN)}}.$$

Como descrito em Martinez *et al.* [14], a AUC (ou ainda, área sob a curva ROC) é uma medida da capacidade do modelo utilizada para discriminar os sujeitos com a característica de interesse *versus* os sujeitos sem a característica de interesse. Tal medida assume valores entre 0 e 1. Para classificar o poder discriminante do modelo RLM, utilizou-se os valores indicativos da AUC apresentados por Hosmer *et al.* [8], conforme é mostrado na Tabela 3. Os gráficos da curva ROC foram obtidos através da função *plot.roc(.)* do pacote “pROC” (Robin *et al.* [15]) do *software* R. O intervalo de confiança de 95% (IC 95%) foi calculado com 2.000 repetições estratificadas, e a curva ROC suavizada foi adicionada aplicando-se o parâmetro *smooth* na função *plot.roc(.)*.

TABELA 3: Classificação do poder discriminante de um modelo de classificação, de acordo com a AUC.

AUC	Poder Discriminante do Modelo
0,5	Sem Poder Discriminativo
]0,5; 0,7[Discriminação Fraca
]0,7; 0,8[Discriminação Aceitável
]0,8; 0,9[Discriminação Boa
$\geq 0,9$	Discriminação Excelente

Por fim, concluindo a avaliação da capacidade preditiva do modelo, foi utilizada a medida conhecida como *Brier score* (Brier [16]), que se trata de uma função de pontuação que mede a precisão das previsões probabilísticas de um modelo. Assim, quanto menor for a pontuação Brier para um conjunto de previsões, melhores as previsões estão ajustadas. O conjunto de possíveis resultados pode ser de natureza binária ou categórica, e as probabilidades atribuídas a esse conjunto de resultados devem somar um (sendo que cada probabilidade individual está no intervalo $[0, 1]$). Neste trabalho, o valor do *Brier score* foi obtido através da função *brierscore(.)* do pacote “scoring” (Merkle [17]) do R.

O *Brier score* é definido como:

$$BS = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^R (f_{ik} - o_{ik})^2,$$

em que R é o número de classes, N é o número de clientes, f_{ik} é a probabilidade prevista do i -ésimo cliente pertencer à k -ésima classe (previsões probabilísticas), $o_{ik} = 1$ se a classe real $y_i = k$, e $o_{ik} = 0$ se a classe real $y_i \neq k$, com $k = 0, 1, 2, 3$ e $i = 1, 2, \dots, 65$, para o banco de teste.

Também se verificou a medida de concordância entre os dados observados e os dados preditos pelo modelo final através da estatística Kappa, a qual informa a consistência ou concordância dos resultados quando a mensuração se repete nas mesmas condições. Para o cálculo da estatística Kappa e do seu IC 95% no *software* R, utilizou-se a função *Kappa(.)* do pacote “vcd” (Meyer *et al.* [18]). Baseado na matriz de confusão, segundo Matos *et al.* [19], após adaptações para este estudo, a estatística Kappa é dada por:

$$\kappa = \frac{P(\mathbf{A}) - P(\mathbf{E})}{1 - P(\mathbf{E})},$$

em que:

$$P(\mathbf{A}) = \frac{VN + VP}{VN + FP + FN + VP} \quad \text{e} \quad P(\mathbf{E}) = \frac{P(\text{“Negativas”}) + P(\text{“Positivas”})}{VN + FP + FN + VP},$$

com

$$P(\text{“Negativas”}) = \frac{(VN + FN) \times (VN + FP)}{VN + FP + FN + VP} \quad \text{e} \quad P(\text{“Positivas”}) = \frac{(FN + VP) \times (FP + VP)}{VN + FP + FN + VP}.$$

Obteve-se também o teste z bicaudal (ou bilateral) para o cálculo do p -valor dos coeficientes, o teste de Wald e a *odds ratio* (ou razão de chances) dos coeficientes do modelo com seus respectivos IC 95%. Neste caso, o cálculo do p -valor dos coeficientes é dado por (Marôco [20]): $P(|Z| > z \mid H_0)$, em que $Z \sim N(0, 1)$ e $H_0 : \beta_{lj} = 0$ ($l = 1, 2, 3$ e $j = 0, 1, \dots, p$) representa a hipótese nula; o teste de Wald é dado por (Bittencourt [9]): $W = \left(\frac{\hat{\beta}_{lj}}{\sigma_{\hat{\beta}_{lj}}} \right)^2 \stackrel{a}{\sim} \chi_1^2$, em que χ_1^2 denota a distribuição qui-quadrado com 1 grau de liberdade e $\sigma_{\hat{\beta}_{lj}}$ representa o erro-padrão (EP) do estimador de máxima verossimilhança para o coeficiente β_{lj} (isto é, $\hat{\beta}_{lj}$); e, por fim, a *odds ratio* (OR) é dada por (Marôco [20]):

$$OR(l, 0 \mid x_m) = \exp\{\beta_{lm}\} = \frac{P(Y = c \mid X_l = x_l + 1)P(Y = 0 \mid X_l = x_l + 1)}{P(Y = c \mid X_l = x_l)P(Y = 0 \mid X_l = x_l)},$$

em que as razões de chances são calculadas para cada uma das 3 classes (isto é, $l = 1, 2, 3$) relativamente à classe de referência 0 e à variável independente x_m ($m = 1, 2, \dots, p$).

Para a análise dos resíduos do modelo RLM ajustado, utilizou-se o gráfico quantil-quantil (*QQ plot*), que checa a adequação da distribuição de frequência dos dados a uma certa distribuição de probabilidades. Neste trabalho, verificou-se se os resíduos do modelo ajustado apresentam distribuição normal, por classificação, por meio da técnica de envelope simulado (Zeviani [21]).

3 RESULTADOS E DISCUSSÃO

Devido ao sigilo das informações financeiras dos clientes, apenas algumas de suas características estão resumidas na Tabela 4, onde consta a frequência (absoluta, percentual e percentual acumulada) de clientes por sexo, safra, operação de crédito, segmento de cliente, *rating* e resumo do risco de crédito dos contratos. Observa-se que, do total de $N = 214$ clientes, 51,4% são do sexo feminino, 64,0% têm contratos da operação tipo dois (OP02), 63,1% dos contratos iniciaram no ano de 2016 e a maioria dos clientes (94,4%) estão distribuídos em apenas dois segmentos (SG02 e SG04). Quanto ao *rating* e risco de crédito, 68,7% dos clientes estão classificados entre os *ratings* A e G, ou seja, contratos com 01 a 180 dias de atraso; porém, dentre esses, 31,3% são considerados de baixo risco de crédito, pois há possibilidade maior de recuperação do crédito inadimplido.

TABELA 4: Frequência de clientes por sexo, safra, operação de crédito, segmento de cliente, *rating* e resumo do risco de crédito dos contratos de uma instituição financeira ($N = 214$). Brasil, 2017.

Variável		Frequência	%	% Acumulada
Sexo				
Feminino		110	51,4	51,4
Masculino		104	48,6	100,0
Safra				
2014		17	7,9	7,9
2015		62	29,0	36,9
2016		135	63,1	100,0
Tipo de Crédito				
OP01		34	15,9	15,9
OP02		137	64,0	79,9
OP03		43	20,1	100,0
Perfil do Cliente				
SG01		4	1,9	1,9
SG02		93	43,5	45,4
SG03		8	3,7	49,1
SG04		109	50,9	100,0
<i>Rating</i> Dias de Atraso				
A	01 a 14	65	30,4	30,4
B	15 a 30	2	0,9	31,3
C	31 a 60	7	3,3	34,6
D	61 a 90	9	4,2	38,8
E	91 a 120	21	9,8	48,6
F	121 a 150	19	8,9	57,5
G	151 a 180	24	11,2	68,7
H	≥ 181	67	31,3	100,0
Risco de Crédito				
Alto: <i>Rating</i> D ao H		140	65,4	65,4
Médio: <i>Rating</i> C		7	3,3	68,7
Baixo: <i>Rating</i> A e B		67	31,3	100,0

Fonte: Sistema de Recuperação de Crédito - Instituição Financeira.

Na Tabela 5 é mostrada a frequência (absoluta, percentual e percentual acumulada) de clientes por percentual da dívida total *versus* valor do contrato, por percentual do valor vencido *versus* valor do contrato, e por percentual liquidado do contrato. Constata-se que 35,5% dos clientes têm o percentual da dívida entre 83% e 110% do valor do contrato, pois nesse intervalo percentual constam os juros e os encargos devido ao atraso. Em janeiro de 2017, 28,0% dos clientes tinham dívidas que representavam de 2% a 8% do valor vencido *versus* o valor do contrato; e 26,2% dos clientes tinham dívidas que representavam de 95% a 127% do valor vencido *versus* o valor do contrato. Além disso, 35,5% dos clientes

havam liquidado de -10% a 17% do valor dos empréstimos contratados, ou seja, estavam inadimplentes da maior parte do valor contratado. Outro item em destaque: até o momento da coleta do banco de dados, 24,8% dos clientes tinham liquidado entre 46% e 96% de seus contratos.

TABELA 5: Frequência de clientes por % da dívida total *versus* valor do contrato, por % do valor vencido *versus* valor do contrato, e por % liquidado do contrato com uma instituição financeira ($N = 214$). Brasil, 2017.

Intervalo	Frequência	%	% Acumulada
% da Dívida Total <i>versus</i> Valor do Contrato			
4 - 55	56	26,2	26,2
55 - 83	34	15,9	42,1
83 - 110	76	35,5	77,6
110 - 230	48	22,4	100,0
% do Valor Vencido <i>versus</i> Valor do Contrato			
2 - 8	60	28,0	28,0
8 - 54	44	20,6	48,6
54 - 95	54	25,2	73,8
95 - 127	56	26,2	100,0
% Liquidado do Contrato			
-130 - -10	48	22,4	22,4
-10 - 17	76	35,5	57,9
17 - 46	37	17,3	75,2
46 - 96	53	24,8	100,0

Fonte: Sistema de Recuperação de Crédito - Instituição Financeira.

Neste estudo, a técnica de RLM foi utilizada para estimar a probabilidade de cada cliente ser classificado como *Ótimo*, *Bom*, *Ruim* ou *Péssimo*, a partir do momento que recebeu (ou não) contato por parte da equipe de cobrança e efetuou (ou não) o pagamento de uma prestação em atraso. Em concordância com Bittencourt [9], o modelo ajustado é estatisticamente significativo, com graus de liberdade = 15, AIC = 281,0795, Deviance = 251,0795 e Log-verossimilhança = 125,5398, uma vez que os resultados obtidos indicam que o modelo estimado pode ser utilizado na classificação dos clientes inadimplentes quanto à probabilidade de se tornarem bons pagadores após (ou não) intervenção realizada por equipes internas de cobrança.

As estimativas dos coeficientes do modelo RLM para as variáveis: idade do cliente, em anos completos (VAR030); tempo, em meses, desde o início do contrato (VAR009); percentual do valor vencido *versus* o valor do contrato (VAR015) e renda do cliente, em reais (VAR028); assim como para as classificações: *Ruim*, *Bom* e *Ótimo*, relativamente à categoria de referência *Péssimo*, estão disponíveis na Tabela 6. Os resultados apresentados para a estatística de Wald (W) indicam que há evidências suficientes para rejeitar a hipótese nula de que o coeficiente β é igual a zero, visto que o p -valor obtido é menor que 0,05 para quase todas as variáveis em cada classificação. Apenas dois e três parâmetros β não resultaram significativos ao nível de 5% nas equações (1) e (2), respectivamente, enquanto que, para a equação (3), todos os coeficientes β foram significativos. Pelo modelo ajustado, é possível verificar que a passagem da classificação de referência 0 (*Péssimo*) para a classificação 2 (*Bom*) não é afetada pelas variáveis VAR030 ($\hat{\beta}_{VAR030} = 0,0349$; p -valor = 0,3014) e VAR009 ($\hat{\beta}_{VAR009} = 0,0872$; p -valor = 0,0575). Porém, a probabilidade de passar da classificação de referência 0 para a classificação 3 (*Ótimo*) é afetada significativamente por todas as variáveis independentes do modelo (p -valor < 0,05 para todas as variáveis). Para a variável VAR030, a razão de chances de passar da classificação de referência *Péssimo* para a classificação *Ótimo* é 1,1055, isto é, para cada incremento (ano) na variável VAR030, as chances do cliente passar a ser classificado como *Ótimo* aumentam em 10,55%. Relativamente à variável VAR028, a razão de chances é 1,3582, isto é, para cada incremento (na renda,

em reais) na variável VAR028, as chances do cliente passar a ser classificado como *Ótimo* aumentam em 35,82%. Para a variável VAR009, a razão de chances é 1,1162, ou seja, a cada incremento (mês) na variável VAR009, as chances do cliente passar a ser classificado como *Ótimo* aumentam em 11,62%. Por fim, a variável VAR015 apresenta razão de chances de 0,9348, isto é, para cada incremento (percentual) na variável VAR015, as chances do cliente passar da classificação de referência *Péssimo* para a classificação *Ótimo* diminuem em 6,52%.

De acordo com Bittencourt [9], a partir das funções lineares $g_l(\mathbf{x}_i)$, $l = 1, 2, 3$ e $i = 1, 2, \dots, 65$ (para o banco de teste), cujos parâmetros β são estimados por máxima verossimilhança (a partir do banco de treinamento), é possível calcular as probabilidades condicionais de ocorrência de cada categoria da variável resposta Y_i dado um vetor de observações \mathbf{x}_i , como segue.

Sejam as funções lineares $g_l(\mathbf{x}_i)$, $l = 1, 2, 3$ e $i = 1, 2, \dots, 65$, definidas por:

$$g_1(\mathbf{x}_i) = -2,8071 + 0,0220 \text{ VAR030}_i + 0,1152 \text{ VAR009}_i + 0,0261 \text{ VAR015}_i - 0,0745 \text{ VAR028}_i, \quad (1)$$

$$g_2(\mathbf{x}_i) = -2,5593 + 0,0349 \text{ VAR030}_i + 0,0872 \text{ VAR009}_i - 0,0387 \text{ VAR015}_i + 0,3137 \text{ VAR028}_i, \quad (2)$$

$$g_3(\mathbf{x}_i) = -5,7623 + 0,1003 \text{ VAR030}_i + 0,1100 \text{ VAR009}_i - 0,0674 \text{ VAR015}_i + 0,3062 \text{ VAR028}_i. \quad (3)$$

Então, a probabilidade de observar cada uma das classificações possíveis em função das variáveis independentes no modelo final, é dada por:

$$P(Y_i = 0 | \mathbf{x}_i) = (1 + e^{-2,8071+0,0220 \text{ VAR030}_i+0,1152 \text{ VAR009}_i+0,0261 \text{ VAR015}_i-0,0745 \text{ VAR028}_i} + e^{-2,5593+0,0349 \text{ VAR030}_i+0,0872 \text{ VAR009}_i-0,0387 \text{ VAR015}_i+0,3137 \text{ VAR028}_i} + e^{-5,7623+0,1003 \text{ VAR030}_i+0,1100 \text{ VAR009}_i-0,0674 \text{ VAR015}_i+0,3062 \text{ VAR028}_i})^{-1}, \quad (4)$$

$$P(Y_i = 1 | \mathbf{x}_i) = e^{-2,8071+0,0220 \text{ VAR030}_i+0,1152 \text{ VAR009}_i+0,0261 \text{ VAR015}_i-0,0745 \text{ VAR028}_i} \times (1 + e^{-2,8071+0,0220 \text{ VAR030}_i+0,1152 \text{ VAR009}_i+0,0261 \text{ VAR015}_i-0,0745 \text{ VAR028}_i} + e^{-2,5593+0,0349 \text{ VAR030}_i+0,0872 \text{ VAR009}_i-0,0387 \text{ VAR015}_i+0,3137 \text{ VAR028}_i} + e^{-5,7623+0,1003 \text{ VAR030}_i+0,1100 \text{ VAR009}_i-0,0674 \text{ VAR015}_i+0,3062 \text{ VAR028}_i})^{-1}, \quad (5)$$

$$P(Y_i = 2 | \mathbf{x}_i) = e^{-2,5593+0,0349 \text{ VAR030}_i+0,0872 \text{ VAR009}_i-0,0387 \text{ VAR015}_i+0,3137 \text{ VAR028}_i} \times (1 + e^{-2,8071+0,0220 \text{ VAR030}_i+0,1152 \text{ VAR009}_i+0,0261 \text{ VAR015}_i-0,0745 \text{ VAR028}_i} + e^{-2,5593+0,0349 \text{ VAR030}_i+0,0872 \text{ VAR009}_i-0,0387 \text{ VAR015}_i+0,3137 \text{ VAR028}_i} + e^{-5,7623+0,1003 \text{ VAR030}_i+0,1100 \text{ VAR009}_i-0,0674 \text{ VAR015}_i+0,3062 \text{ VAR028}_i})^{-1}, \quad (6)$$

$$P(Y_i = 3 | \mathbf{x}_i) = e^{-5,7623+0,1003 \text{ VAR030}_i+0,1100 \text{ VAR009}_i-0,0674 \text{ VAR015}_i+0,3062 \text{ VAR028}_i} \times (1 + e^{-2,8071+0,0220 \text{ VAR030}_i+0,1152 \text{ VAR009}_i+0,0261 \text{ VAR015}_i-0,0745 \text{ VAR028}_i} + e^{-2,5593+0,0349 \text{ VAR030}_i+0,0872 \text{ VAR009}_i-0,0387 \text{ VAR015}_i+0,3137 \text{ VAR028}_i} + e^{-5,7623+0,1003 \text{ VAR030}_i+0,1100 \text{ VAR009}_i-0,0674 \text{ VAR015}_i+0,3062 \text{ VAR028}_i})^{-1}. \quad (7)$$

As probabilidades descritas em (4)-(7) são utilizadas para estabelecer uma regra para discriminação das classes. A regra de classificação para alocar uma observação \mathbf{x}_i numa das classes é dada por:

$$\begin{aligned} \mathbf{x}_i \in 0 & \text{ se } P(Y_i = 0 | \mathbf{x}_i) > [P(Y_i = 1 | \mathbf{x}_i), P(Y_i = 2 | \mathbf{x}_i) \text{ e } P(Y_i = 3 | \mathbf{x}_i)], \\ \mathbf{x}_i \in 1 & \text{ se } P(Y_i = 1 | \mathbf{x}_i) > [P(Y_i = 0 | \mathbf{x}_i), P(Y_i = 2 | \mathbf{x}_i) \text{ e } P(Y_i = 3 | \mathbf{x}_i)], \\ \mathbf{x}_i \in 2 & \text{ se } P(Y_i = 2 | \mathbf{x}_i) > [P(Y_i = 0 | \mathbf{x}_i), P(Y_i = 1 | \mathbf{x}_i) \text{ e } P(Y_i = 3 | \mathbf{x}_i)], \\ \mathbf{x}_i \in 3 & \text{ se } P(Y_i = 3 | \mathbf{x}_i) > [P(Y_i = 0 | \mathbf{x}_i), P(Y_i = 1 | \mathbf{x}_i) \text{ e } P(Y_i = 2 | \mathbf{x}_i)]. \end{aligned}$$

TABELA 6: Coeficientes do modelo final de RLM ajustado, que relaciona a classificação dos clientes quanto à probabilidade de pagamento após intervenção da equipe de cobrança versus as variáveis VAR030, VAR009, VAR015 e VAR028. A classificação de referência é a categoria 0 (Pessimismo).

Classificação	Variável	β	IC 95% para β	EP	W	p-valor	OR (e^β)	IC 95% para e^β
Ruim	Intercepto	-2,8071	[-5,2316; -0,3826]	1,2370	5,1494	0,0233	0,0604	[0,0053; 0,6821]
	VAR030	0,0220	[-0,0206; 0,0647]	0,0217	1,0282	0,3106	1,0223	[0,9796; 1,0668]
	VAR009	0,1152	[0,0291; 0,2013]	0,0439	6,8713	0,0088	1,1220	[1,0295; 1,2229]
	VAR015	0,0261	[0,0123; 0,0399]	0,0070	13,7233	0,0002	1,0264	[1,0123; 1,0407]
	VAR028	-0,0745	[-0,2836; 0,1347]	0,1067	0,4871	0,4852	0,9282	[0,7530; 1,1442]
Bom	Intercepto	-2,5593	[-5,7028; 0,5842]	1,6038	2,5464	0,1105	0,0774	[0,0033; 1,7935]
	VAR030	0,0349	[-0,0313; 0,1012]	0,0338	1,0681	0,3014	1,0355	[0,9692; 1,1065]
	VAR009	0,0872	[-0,0028; 0,1772]	0,0459	3,0673	0,0575	1,0911	[0,9972; 1,1938]
	VAR015	-0,0387	[-0,0655; -0,0119]	0,0137	8,0243	0,0046	0,9620	[0,9366; 0,9881]
	VAR028	0,3137	[0,0296; 0,5978]	0,1450	4,6824	0,0305	1,3684	[1,0300; 1,8180]
Ótimo	Intercepto	-5,7623	[-9,7490; -1,7756]	2,0341	8,0252	0,0046	0,0031	[0,0001; 0,1694]
	VAR030	0,1003	[0,0205; 0,1800]	0,0407	6,0755	0,0137	1,1055	[1,0208; 1,1973]
	VAR009	0,1100	[0,0143; 0,2057]	0,0488	5,0728	0,0243	1,1162	[1,0144; 1,2283]
	VAR015	-0,0674	[-0,1242; -0,0106]	0,0290	5,4083	0,0200	0,9348	[0,8832; 0,9895]
	VAR028	0,3062	[0,0023; 0,6100]	0,1550	3,8994	0,0483	1,3582	[1,0023; 1,8405]

Fonte: Sistema de Recuperação de Crédito - Instituição Financeira.

A partir do banco de dados de teste, foi gerada a matriz de confusão conforme é mostrado na Tabela 7, em que definiu-se como predição negativa as classificações 0 e 1 (*Péssimo* e *Ruim*, respectivamente), e como predição positiva as classificações 2 e 3 (*Bom* e *Ótimo*, respectivamente). Assim, a ACC do modelo ajustado é de 0,8308, demonstrando a sua utilidade para classificar novas observações. Tal valor é consideravelmente superior à taxa de classificação geral do modelo: $MCC = 0,5893$, que indica predição aleatória média a perfeita. O modelo estimado apresenta elevada SPE de 0,9111 e $SEN = 0,6500$, com $EFF = 0,7806$. Para as medidas VPP e VPN, o modelo RLM ajustado apresenta precisão de 0,7647 e 0,8542, respectivamente. Por fim, a medida F do modelo final é de 0,7027.

TABELA 7: Matriz de confusão (banco de dados de teste).

		Predição	
		Negativa	Positiva
Observação	Negativa	41	4
	Positiva	7	13

Ainda de acordo com os resultados exibidos na Tabela 7, o modelo RLM ajustado apresenta capacidade discriminante fraca para clientes classificados como *Péssimo* ($AUC_0 = 0,5885$), enquanto que, para clientes classificados como *Ruim*, *Bom* ou *Ótimo*, o modelo apresenta capacidade discriminante aceitável, com $AUC_1 = 0,7827$, $AUC_2 = 0,7263$ e $AUC_3 = 0,7960$.

Nas Figuras 1 e 2 são mostradas as curvas ROC, por classificação do cliente *versus* observações negativas e positivas, para o banco de dados de teste, com as respectivas AUC's. Destacam-se os gráficos inferiores da Figura 1, com AUC de 0,8030 (painel esquerdo) e 0,7080 (painel direito).

A estatística Kappa calculada ($\kappa = 0,5855$; IC 95% = [0,3677; 0,8034]) traz evidências de uma concordância moderada entre as classificações observadas e as predições. O Brier score obtido é de, aproximadamente, 0,2519, o que é indicativo de que, para o banco de dados de teste, as previsões estão calibradas.

Analisando a Figura 3, que contém os QQ plots com envelopamento dos resíduos do modelo RLM ajustado, por classificação do cliente, percebe-se que há evidências favoráveis quanto ao bom ajuste do modelo devido ao fato dos resíduos estarem distribuídos/localizados dentro dos envelopes simulados, visto que o eixo sinalizado como *Componente do Desvio* representa os valores dos resíduos do modelo ajustado, por classe.

4 CONSIDERAÇÕES FINAIS E CONCLUSÃO

O GRC é sustentado por um conjunto de informações qualitativas e quantitativas dos clientes, permitindo que se possam traçar diretrizes de cobrança eficazes e ao menor custo possível.

Neste estudo, propôs-se a elaboração de um modelo de recuperação de crédito para classificar clientes quanto à probabilidade de se tornarem adimplentes tendo havido, ou não, intervenção realizada por equipe de cobrança. Como descrito na Seção 2, foi aplicado o modelo RLM para classificar os clientes de acordo com sua probabilidade de pagamento de uma prestação em atraso (*Collection Score*). Para tanto, a metodologia utilizada é bastante criteriosa devido à constante mudança de perfil dos clientes inadimplentes e das políticas econômicas governamentais regulamentadas pelo CMN. As variáveis independentes no modelo final indicam as premissas que auxiliam no controle e redução da inadimplência, corroborando a eficácia do GRC. Ademais, há evidências de que a inadimplência diminui após o contato realizado pela equipe de recuperação de crédito.

Destarte que o modelo de *Collection Score* deve ser sempre revisado, reestruturado e reavaliado, pois sua eficácia possibilita a mudança de perfil do público-alvo, tornando os indivíduos da amostra cada vez mais com escores piores, sendo, assim, necessária a

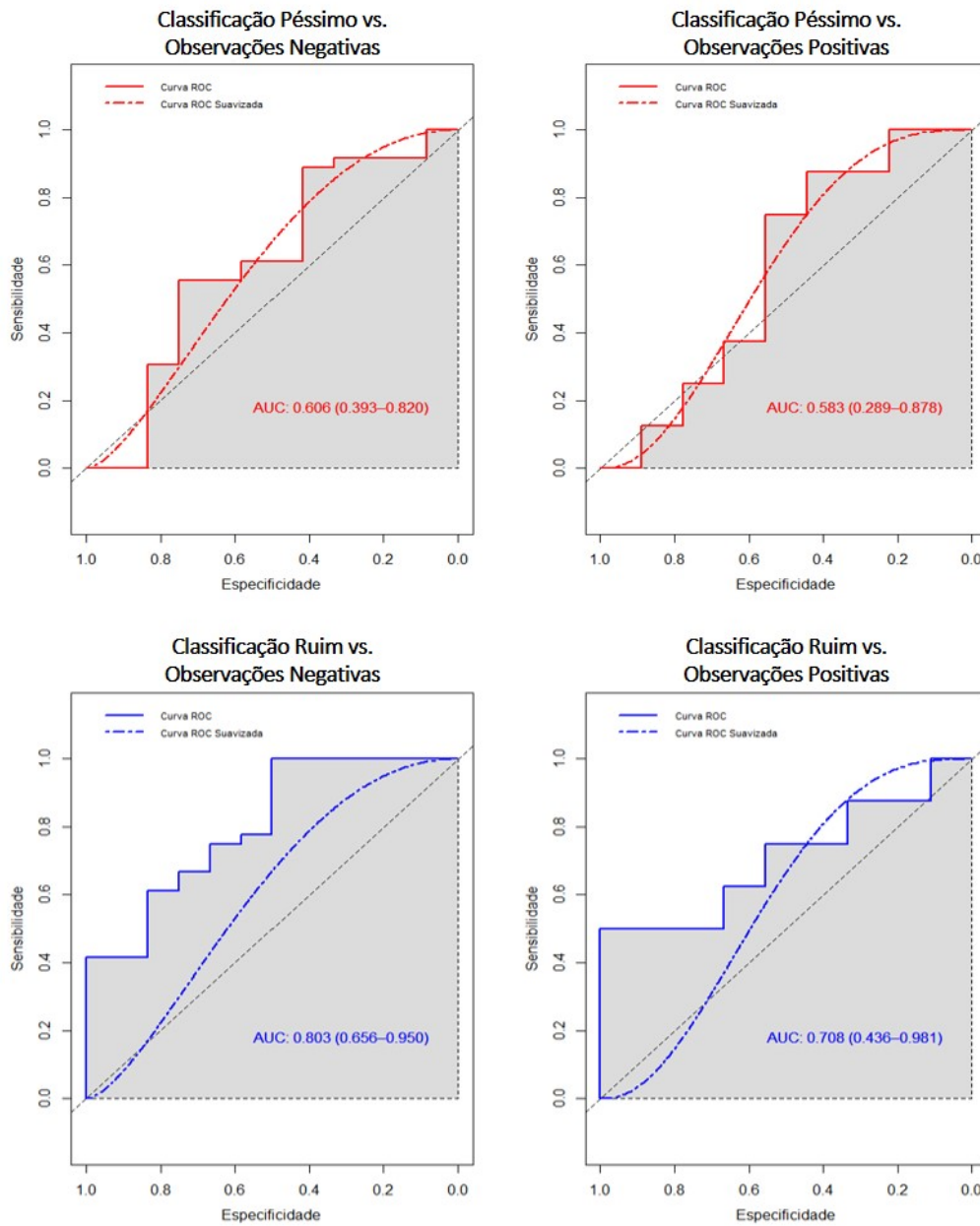


FIGURA 1: Curva ROC por classificação do cliente (*Péssimo* e *Ruim*) versus observações negativas e positivas, com a respectiva AUC e IC 95% (banco de dados de teste).

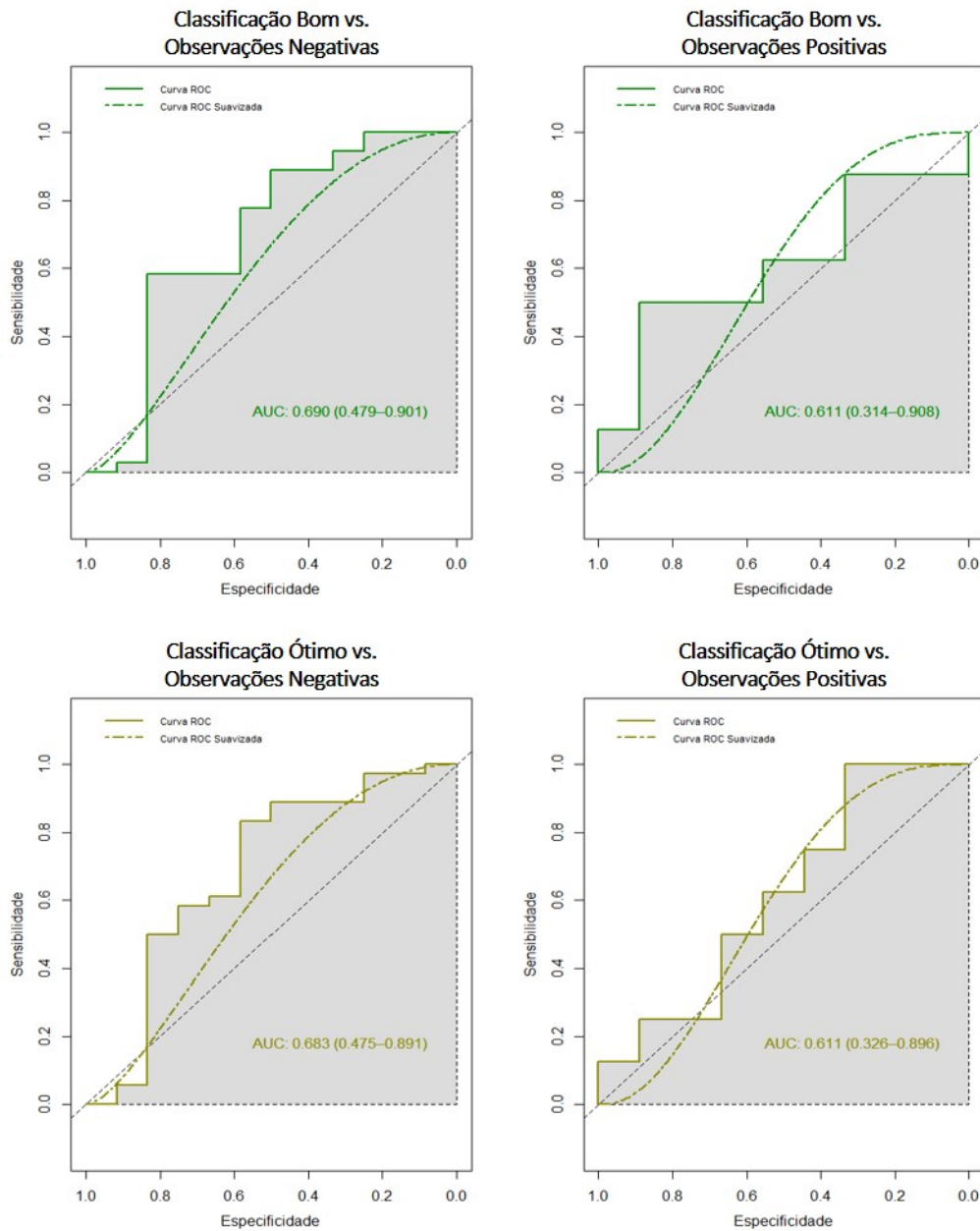


FIGURA 2: Curva ROC por classificação do cliente (*Bom* e *Ótimo*) versus observações negativas e positivas, com a respectiva AUC e IC 95% (banco de dados de teste).

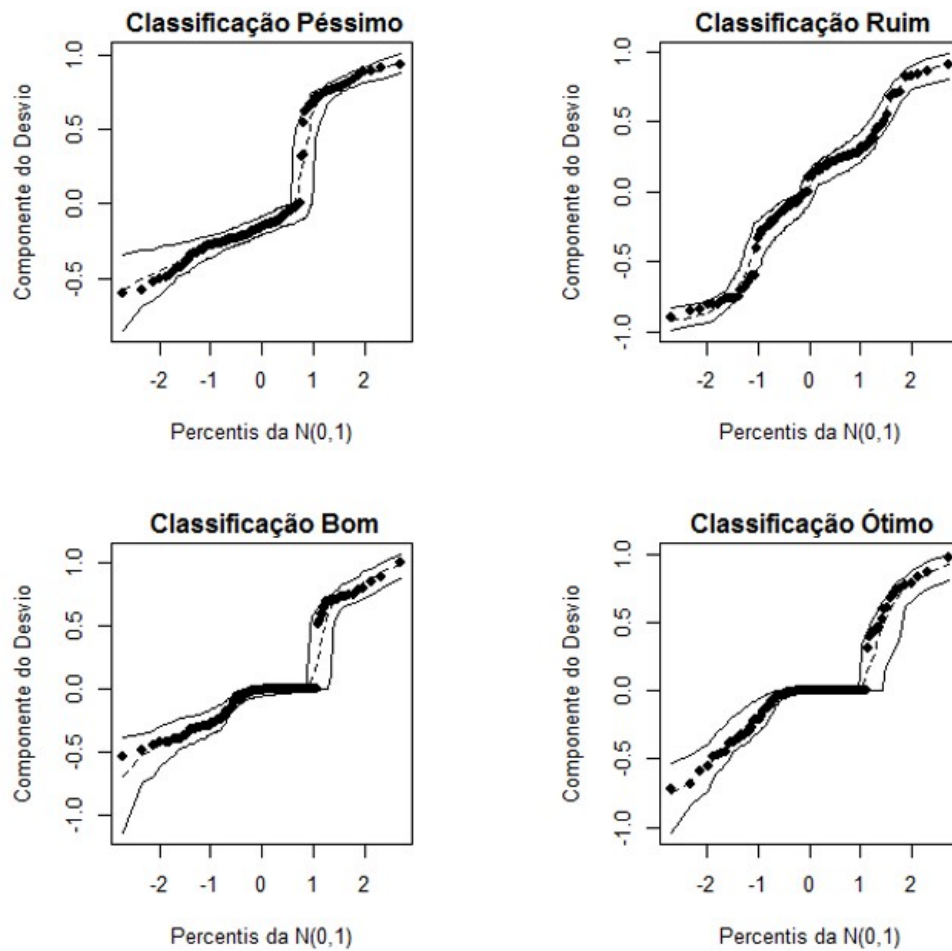


FIGURA 3: QQ plot da distribuição normal com envelope simulado para os resíduos do modelo RLM ajustado, por classificação do cliente (*Pessimista*, *Ruim*, *Bom* e *Ótimo*).

criação de um novo modelo. Portanto, sugere-se também aplicar a modelagem de regressão logística ordinal e comparar com os resultados da modelagem realizada neste estudo, a fim de se verificar o quanto a sequência de contatos realizados pela equipe de cobrança pode impactar na classificação dos clientes e no índice de recuperação de créditos inadimplidos. Assim como recomenda-se empregar outras técnicas de classificação bastante flexíveis, como as máquinas de vetores de suporte (do inglês *support vector machines*) e as florestas aleatórias (do inglês *random forests*), ambas descritas em detalhes em James *et al.* [22].

REFERÊNCIAS

- [1] A. Pinho, *Estratégias de retenção de clientes no marketing de relacionamento. Congresso Knowledge Management Brasil. São Paulo-SP, 2007.* [Online]. Disponível em: <http://pt.slideshare.net/AndersonGP/estrategias-de-reteno-de-clientes-no-marketing-de-relacionamento>
- [2] R. O. Xavier e J. S. Dornelas, “O papel do gerente num contexto de mudança baseada no uso da tecnologia CRM,” *Revista de Administração Contemporânea*, vol. 10, pp. 9 – 30, 2006. [Online]. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-65552006000100002&nrm=iso
- [3] F. A. P. Pinheiro, J. R. F. Savóia, e J. R. Securato, “Basileia III: Impacto para os Bancos no Brasil,” *Revista Contabilidade & Finanças*, vol. 26, pp. 345 – 361, 2015. [Online]. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1519-70772015000300345&nrm=iso
- [4] ANBIMA, *Basileia III: Relatório Basileia - Adequação no Brasil nos Acordos de Basileia*, 2010. [Online]. Disponível em: <http://portal.anbima.com.br/informacoes-tecnicas/estudos/perspectivas/Documents/Perspectivas%20ANBIMA%20Basileia%20III.pdf>
- [5] M. Oliveira e F. Louzada, “Risco de recuperação: Uma aplicação de modelo de riscos competitivos latentes em créditos inadimplentes,” *Revista Tecnologia de Crédito*, no. 88, pp. 45–54, 2014.
- [6] R. B. Souza, “O modelo de collection scoring como ferramenta para a gestão estratégica do risco de crédito,” Tese de Doutorado, Escola de Administração de Empresas de São Paulo, Fundação Getúlio Vargas, 2000.
- [7] A. Machado, *Collection Scoring via Regressão Logística e Modelo de Riscos Proporcionalis de Cox. Dissertação (Mestrado em Estatística) - Departamento de Estatística do Instituto de Ciências Exatas, UnB, 2015.* [Online]. Disponível em: http://repositorio.unb.br/bitstream/10482/19575/1/2015_AlineRodriguesMachado.pdf
- [8] D. Hosmer, S. Lemeshow, e R. Sturdivant, *Applied Logistic Regression*, 3rd ed. Wiley & Sons, 2013.
- [9] H. Bittencourt, “Regressão logística politômica: revisão teórica e aplicações,” *ACTA SCIENTIAE - Revista de Ensino de Ciências e Matemática*, vol. 5, no. 1, pp. 77–78, 2003.
- [10] R. C. Team, “R: A language and environment for statistical computing. (Version 3.3.1 (2016-06-21)). Vienna, Austria: R Foundation for Statistical Computing,” 2016.
- [11] B. Ripley e W. Venables, “nnet: Feed-forward neural networks and multinomial log-linear models,” *R package version*, vol. 7, no. 5, 2011.
- [12] B. Ripley, “Support functions and datasets for Venables and Ripley’s MASS. package MASS,” 2017.

- [13] L. A. Silva, S. M. Peres, e C. Boscarioli, *Introdução à mineração de dados: com aplicações em R*. Elsevier Brasil, 2017.
- [14] E. Martinez, F. Louzada, e B. Pereira, “A curva ROC para testes diagnósticos,” *CADERNOS Saúde Coletiva (UFRJ)*, vol. 11, no. 1, pp. 7–31, 2003.
- [15] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, M. Müller, e S. Siegert, “pROC: display and analyze ROC curves,” *R Package Version*, vol. 1, 2018.
- [16] G. Brier, *Verification of Forecasts Expressed in Terms of Probability*, 1950. [Online]. Disponível em: <https://docs.lib.noaa.gov/rescue/mwr/078/mwr-078-01-0001.pdf>
- [17] E. Merkle, *Proper scoring rules. R package version 0.5-1*, 2015. [Online]. Disponível em: <https://cran.r-project.org/package=scoring>
- [18] D. Meyer, A. Zeileis, K. Hornik, F. Gerber, e M. Friendly, *Visualizing Categorical Data. R package version 1.4-3*, 2016. [Online]. Disponível em: <https://cran.r-project.org/package=vcd>
- [19] P. Matos, L. Lombardi, R. Ciferri, T. Pardo, C. Ciferri, e M. Vieira, *Métricas de Avaliação. Relatório Técnico*, 2009. [Online]. Disponível em: <http://conteudo.icmc.usp.br/pessoas/taspardo/TechReportUFSCar2009a-MatosEtAl.pdf>
- [20] J. Marôco, *Análise Estatística com o SPSS Statistics*, 6ª ed. Report Number Lda, 2014.
- [21] W. Zeviani, *Como fazer e interpretar o gráfico quantil-quantil*, 2012. [Online]. Disponível em: <https://www.r-bloggers.com/lang/portuguese/1179>
- [22] G. James, D. Witten, T. Hastie, e R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.