DOI: http://dx.doi.org/10.5007/2175-8069.2013v10n20p139

Análise de risco de crédito com o uso de regressão logística

Credit risk analysis through logistic regression

Eric Bacconi Gonçalves
Mestre em Administração pela Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo – FEA/USP
Endereço: Av. Professor Luciano Gualberto, n° 908, Sala G-162
CEP: 05508-010 - São Paulo/SP – Brasil
E-mail: eric.goncalves@telefonica.com.br
Telefone: (11) 3091-6044

Maria Aparecida Gouvêa
Doutora em Administração pela Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo – FEA/USP
Professora Livre Docente do Departamento de Administração da Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo – FEA/USP
Endereço: Av. Professor Luciano Gualberto, n° 908, Sala G-162
CEP: 05508-010 - São Paulo/SP – Brasil
E-mail: magouvea@usp.br
Telefone: (11) 3091-6044

Daielly Melina Nassif Mantovani
Doutora em Administração pela Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo – FEA/USP
Professora do Programa de Pós-Graduação em Administração das Faculdades Metropolitanas Unidas – PPGA/FMU
Endereço: Rua Taguá, n° 150
CEP: 01508-010 - São Paulo/SP – Brasil
E-mail: daimantovani@gmail.com
Telefone: (11) 2985 8820

Artigo recebido em 06/06/2011. Revisado por pares em 01/10/2012. Reformulado em 12/06/2013. Recomendado para publicação em 25/07/2013 por Sandra Rolim Ensslin (Editora Científica). Publicado em 09/08/2013.

Resumo

O objetivo deste estudo foi aplicar a técnica de regressão logística no desenvolvimento de um modelo de predição de *credit scoring* com dados de uma instituição financeira. A partir de uma amostra de 20.000 dados, foram definidas três subamostras: uma para construção do modelo (8.000 dados) e duas para validação, cada uma com 6.000 dados. Nas 3 subamostras, houve distribuição equitativa de bons e maus clientes, classificados nessas categorias conforme padrões da instituição. O modelo de regressão logística apresentou adequados indicadores de ajuste aos dados, podendo ser utilizado no processo de tomada de decisões de concessão de crédito bancário.

Palavras-chave: Crédito. Regressão Logística. Risco.

Abstract

The objective of this study was to apply the logistic regression technique in the development of a model for predicting credit scoring using data from a financial institution. From a sample of 20,000 data, three sub-samples were defined: one sample for model construction (8,000 data) and two other ones for validation, each one with 6,000 data. In the three sub-samples, there was an equitable distribution of good and bad clients, classified into these categories according to institutional standards. The logistic regression model presented adequate indicators of data adjustment in the results, which can be used in the decision-making process of bank credit concessions.

Keywords: Credit. Logistic Regression. Risk.

1 Introdução

O crédito ao consumidor é uma grande indústria que opera no mundo, facilitando as transações de bens e serviços. Assim, grandes varejistas impulsionam suas vendas, fornecendo crédito. "O crédito ao consumidor é um negócio essencial. O maior desafio dessa indústria é tornar o crédito largamente disponível; assim, tantas pessoas quanto possíveis terão a oportunidade de utilizar essa poderosa ferramenta" (LEWIS, 1992, p. 2).

Associado ao crédito ao consumidor, existe o fator risco. Para administrar o risco de crédito, diversas metodologias de mensuração de exposição e mecanismos de gestão existem (LIMA *et al*., 2009). A análise discriminante, a regressão logística, as árvores de decisão e a programação linear e suas variações podem ser destacadas como técnicas para construção de modelos de risco de crédito (SAMEJIMA; DOYA; KAWATO, 2003). A gestão do risco de crédito é um fator estratégico nas instituições que oferecem produtos dessa natureza ao mercado. Os modelos matemáticos e estatísticos auxiliam nessa gestão, permitindo, com base em dados de contratos já existentes, criar equações que classifiquem bons e maus pagadores. A criação e implementação de um modelo estatístico, utilizado de maneira integrada com o desenvolvimento do banco de dados de clientes da instituição, leva a um modelo de *credit*

scoring dinâmico, bem ajustado e constantemente aprimorado, que implica um diferencial à organização.

Tendo em vista a relevância de modelos estatísticos no apoio às instituições bancárias no processo de tomada de decisões de concessão de crédito, o presente estudo propõe-se a responder à seguinte pergunta de pesquisa: qual a adequação da técnica de análise de regressão logística para a classificação de clientes segundo seu *status* de adimplência? Portanto, o objetivo deste artigo é a apresentação do uso de regressão logística para a classificação de bons e maus pagadores em financiamentos bancários, considerando-se o produto crédito pessoal.

2 Fundamentação Teórica

2.1 Crédito ao Consumidor e Avaliação do Risco de Crédito

Crédito, por definição, é "todo ato de vontade ou disposição de alguém de destacar ou ceder, temporariamente, parte do seu patrimônio a um terceiro, com a expectativa de que essa parcela volte a sua posse integralmente, após decorrido o tempo estipulado" (SCHRICKEL, 1995, p. 25). Patrimônio, por sua vez, pode ser entendido como dinheiro no caso de empréstimo monetário ou de bens e empréstimo para uso ou venda com pagamento parcelado, ou a prazo. Em um sentido mais amplo, o crédito pode ser definido como um instrumento de desenvolvimento da economia, à medida que se propõe a financiar os agentes econômicos, tais como o Estado, as empresas e as famílias de uma determinada nação ou de nações específicas, para que possam satisfazer suas necessidades de consumo e investimento (PALMUTI; PICCHIAI, 2012).

Devido ao fato de envolver a expectativa do retorno do patrimônio, deve-se entender que todo crédito está associado a um certo risco (SCHRICKEL, 1995, p. 24).

Como se trata de um ato de vontade, cabe ao cedente do patrimônio a decisão de cedê-lo ou não, tendo o direito de recusar se achar conveniente. Os bancos comerciais destacam-se nesse papel de cedentes, oferecendo capital por empréstimos e financiamentos para os agentes econômicos, estimulando o crescimento econômico (PALMUTI; PICCHIAI, 2012).

Apesar de existirem empréstimos a título gratuito, ou seja, não onerosos àquele que recebe o bem, normalmente associa-se a qualquer transação de empréstimo um preço remuneratório, a ser pago pelo tomador (SECURATO, 2002, p. 18). Esse preço, conhecido como taxa de juros, baseia-se na compensação dos riscos assumidos pelo cedente quanto à possível perda ou deterioração de seu patrimônio cedido.

Gitman (1997, p. 202) define risco como possibilidade de prejuízo financeiro. Ativos que possuem maiores possibilidades de prejuízo financeiro são mais arriscados que aqueles com menores possibilidades. Risco, dessa forma, pode ser entendido como incerteza ao se referir à "possibilidade de retornos associada a um dado ativo". Entretanto, Lima (2002, p. 20) aponta que "no risco, as probabilidades de ocorrência de um dado evento são conhecidas, enquanto na incerteza não há dados para calcularmos essas probabilidades".

Mais especificamente enfocado para uma instituição financeira,

> risco de crédito define-se como a medida numérica da incerteza com relação ao recebimento futuro de um valor contratado (ou compromissado), a ser pago por um tomador de um empréstimo, contraparte de um contrato ou emissor de um título

> carregado nos estoques da instituição, descontadas as expectativas de recuperação e realização de garantias. (DUARTE JR. *et al.*, 1999, p. 67).

Os principais subtipos desse risco são (FIGUEIREDO, 2001, p. 9):

- **Risco de inadimplência**: risco do não pagamento, por parte do tomador, de uma operação de crédito – empréstimo, financiamento, adiantamentos, operações de *leasing* – ou ainda a possibilidade de uma contraparte de um contrato ou emissor de um título não honrar seu crédito.
- **Risco de degradação de garantia**: risco de perdas em função das garantias oferecidas por um tomador deixarem de cobrir o valor de suas obrigações junto à instituição, em função de desvalorização do bem no mercado e dilapidação do patrimônio empenhado pelo tomador.
- **Risco de concentração de crédito**: possibilidade de perdas em função da concentração de empréstimos e financiamentos em poucos setores da economia, classes de ativos ou de empréstimos elevados para um único cliente ou grupo econômico.
- **Risco de degradação de crédito**: perda pela queda na qualidade creditícia do tomador de crédito, emissor de um título ou contraparte de uma transação, ocasionando uma diminuição no valor de suas obrigações. Esse risco pode acontecer em uma transação do tipo de aquisição de ações ou de títulos soberanos que podem perder valor.
- **Risco soberano**: risco de perdas envolvendo transações internacionais – aquisição de títulos, operações de câmbio – quando o tomador de um empréstimo ou emissor de um título não pode honrar seu compromisso por restrições do país sede.

A avaliação do risco de um potencial cliente pode ser feita de duas maneiras:

1. Por meio de julgamento, como uma forma mais subjetiva que envolve uma análise mais qualitativa.
2. Por meio da classificação do tomador, via modelos de avaliação, envolvendo uma análise mais quantitativa.

Grande parte das empresas que trabalham com concessão de crédito utilizam as duas formas combinadas. Na avaliação do risco de crédito por meio de julgamento, o analista avalia a solicitação de empréstimo mediante ficha cadastral e/ou entrevista. Para esse tipo de avaliação, existem 4 "Cs" largamente mencionados na literatura pesquisada que devem ser considerados (SANTI FILHO, 1997; SCHRICKEL, 1995):

- **Caráter**: refere-se à intenção de pagar. O avaliador deve levar em consideração o cadastro do cliente, levantando informações sobre empréstimos anteriores, atuação na praça, existência de restrições.
- **Capacidade**: refere-se à habilidade de pagar. É considerado o aspecto mais subjetivo do risco, pois depende mais da percepção do analista do que da análise de dados cadastrais.

- **Capital**: refere-se ao potencial de "produzir" dinheiro. No caso de análise para pessoa física, o avaliador deve levar em consideração a renda do indivíduo e seu patrimônio para entender se ele possui meios de quitar o empréstimo.
- **Condições**: referem-se ao micro e macrocenário em que o tomador está inserido. Esse último aspecto foge do controle do tomador e requer a análise dos fatores externos que afetam a economia, como planos de ajuste da economia, bolsas de valores em queda (ou em alta), entre outros.

Alguns autores, como Securato (2002), consideram um quinto "C" (Colateral), que diz respeito às garantias que o devedor deve apresentar para viabilizar a operação de crédito.

Na avaliação do risco de crédito por meio de classificação do tomador é que são utilizados os modelos chamados de *credit scoring,* que permitem uma mensuração do risco do tomador de crédito, auxiliando na tomada de decisão (concessão ou não do crédito). Se o risco for mal avaliado, a empresa certamente irá perder dinheiro, seja pelo aceite de clientes que irão gerar prejuízos ao negócio, ou pela recusa de clientes bons que gerariam lucros ao negócio.

2.2. Modelo de Credit Scoring

Ao longo dos anos, muitos administradores de crédito buscaram uma forma de reduzir o processo de análise de crédito a uma fórmula numérica. Entretanto, até o desenvolvimento dos computadores, poucos avanços foram feitos na análise de grandes massas de dados.

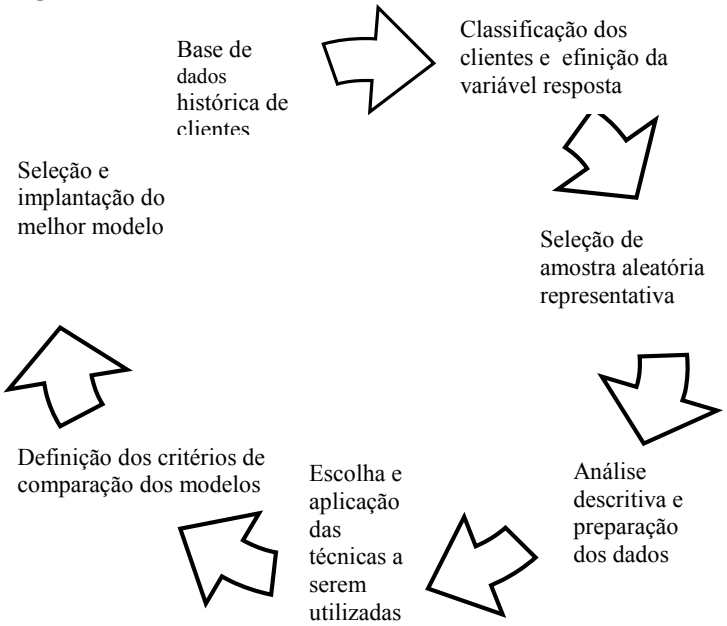
No Brasil, as instituições financeiras passaram a utilizar maciçamente os modelos de *credit scoring* apenas em meados dos anos 1990. Casa Nova (2013) destaca a importância de se utilizarem modelos estruturados e com bom ajuste para a análise de crédito, pois reduzem o risco de se conceder crédito a potenciais maus pagadores. Ademais, na visão da autora, há um custo mais elevado em se classificar erroneamente um mau pagador do que um bom pagador, isto é, há maior dano à organização se um mau cliente for classificado como bom pagador do que a situação contrária; por essa razão, além dos índices gerais de ajuste, é necessário observar com cuidado a taxa de acertos de classificação dos inadimplentes (CASA NOVA, 2013). Há sete passos para se construir um modelo de *credit scoring,* conforme mostra a Figura 1.

1. **Levantamento de uma base histórica de clientes**: a suposição básica para se construir um modelo de avaliação de crédito é que os clientes têm o mesmo padrão de comportamento ao longo do tempo; portanto, com base em informações passadas, são construídos os modelos; a disponibilidade e qualidade da base de dados, portanto, são fundamentais para o sucesso do modelo (TREVISANI *et al.*, 2004).
2. **Classificação dos clientes de acordo com o padrão de comportamento e a definição da variável resposta (tipo de cliente)**: as instituições têm sua própria política de crédito e os conceitos de bons e maus clientes podem variar; nessa classificação, além de clientes bons e maus, também existem os clientes excluídos, que possuem características peculiares e que não devem ser considerados (por exemplo, trabalham na instituição) e os clientes indeterminados, que estão na fronteira entre serem bons ou maus, não existindo, ainda, uma posição clara para eles; na prática e

nos trabalhos acadêmicos, consideram-se apenas os clientes bons e maus para o modelo, devido à maior facilidade de se trabalhar com resposta binária (HAND; HENLEY, 1997; ROSA, 2000; SEMOLINI, 2002; OHTOSHI, 2003; LIMA *et al.*, 2009).

3. **Seleção de amostra aleatória representativa da base histórica**: é importante que as amostras de bons e maus clientes tenham o mesmo tamanho para se evitar possível viés, devido à diferença de tamanhos; não existe um número fixo para a amostra; entretanto, Lewis (1992, p. 31) sugere uma amostra de pelo menos 1.500 clientes bons e 1.500 clientes maus para obter resultados robustos; costuma-se usar duas amostras para construção e para validação do modelo. Neste trabalho, foi possível o acesso aos dados de 20.000 clientes, sendo 10.000 bons e 10.000 maus.
4. **Análise descritiva e preparação dos dados**: consiste em analisar cada variável a ser utilizada no modelo segundo critérios estatísticos.
5. **Escolha e aplicação das técnicas a serem utilizadas para a construção do modelo**: neste estudo, será aplicada a análise de regressão logística; Hand e Henley (1997) destacam, ainda, Análise Discriminante, Regressão Linear e Árvores de Decisão; alguns estudiosos também têm usado Análise de Sobrevivência (HARRISON; ANSELL, 2002; ANDREEVA, 2003).
6. **Definição dos critérios de comparação dos modelos**: geralmente, usam-se o indicador de acertos e a estatística de Kolmogorov-Smirnov (KS).
7. **Seleção e Implantação do melhor modelo**: segundo critérios definidos, o melhor modelo é escolhido; para implantá-lo, a instituição deve adequar seus sistemas para receber o algoritmo final e programar a utilização do mesmo junto às demais áreas envolvidas.

Figura 1 – Ciclo de desenvolvimento de um modelo



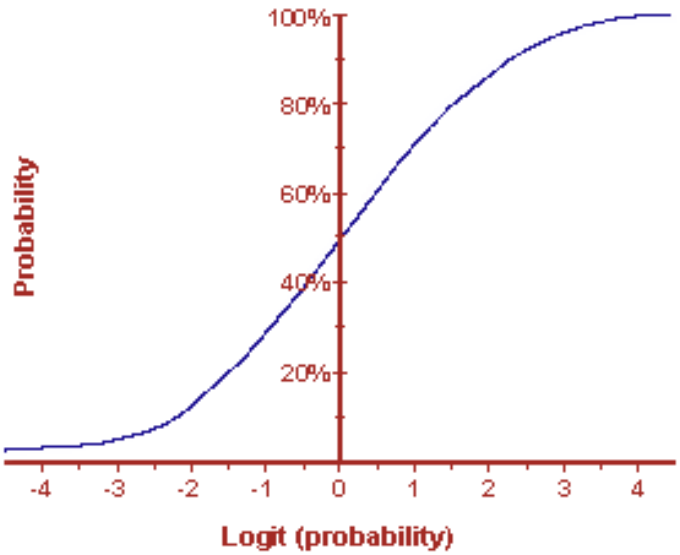
Fonte: Elabora pelos autores.

Neste trabalho, serão seguidos os quatro primeiros passos. Na quinta etapa, será empregada somente a técnica de regressão logística. Embora não se pretenda fazer uma comparação do desempenho de técnicas estatísticas, o sexto passo será adotado no sentido de uso dos indicadores para avaliação da qualidade de ajuste da técnica aplicada aos dados.

2.3 Regressão Logística

Regressão Logística é a técnica mais utilizada no mercado para o desenvolvimento de modelos de *credit scoring* (ROSA, 2000; OHTOSHI, 2003). Ao contrário da análise discriminante, não exige a suposição da normalidade das variáveis independentes e é mais robusta quando a mesma não é atendida (HAIR JR. *et al.*, 2009, p. 225). A regressão logística prediz a probabilidade de um evento ocorrer, a qual pode estar entre 0 e 1. A relação entre as variáveis independentes e a variável dependente se assemelha a uma curva em forma de S, conforme mostra a Figura 2.

Figura 2 – Curva logística



Fonte: Adaptado de Sharma (1996, p. 320).

A literatura da área traz exemplos de aplicação da regressão logística para a criação de modelos de c*redit scoring* com bom ajuste. Palmuti e Picchiai (2012) aplicaram a técnica a uma amostra de 1.635 cadastros de empreendedores (formais e informais) de uma instituição de crédito popular. O modelo dessa pesquisa considerou como variável dependente a "qualidade do crédito", que possui duas categorias: adimplente (1.305 casos) e inadimplente (330 casos), sendo inadimplente o cliente que se encontrava em atraso há um período superior a 30 dias na data da coleta dos dados. Como variáveis independentes, o estudo considerou: gênero, grau de formalidade do empreendedor, garantia oferecida, setor de atuação, escolaridade, valor do crédito, taxa de juros, idade, prazo de pagamento, renda declarada e valor da prestação. A regressão logística reteve no modelo as seguintes variáveis: valor do crédito, nível de formalização, renda, valor da prestação, prazo de pagamento e taxa de juros.

O modelo final teve taxa de acertos de 87,4%. Dessa forma, o modelo de Palmuti e Picchiai (2012) observou que as variáveis do perfil não possuem influência na qualidade do crédito, mas, sim, variáveis relacionadas diretamente ao risco (juros, valor da prestação, renda etc.).

Casa Nova (2013) realizou um estudo comparando diferentes técnicas para avaliação de insolvência/inadimplência de empresas de pequeno e médio porte dos setores industrial e de comércio. A autora obteve dados dos anos de 2001 a 2003, da Serasa Experian, totalizando 259 empresas do setor industrial e 360 empresas do setor de comércio. Foram aplicadas as técnicas de análise envoltória de dados, redes neurais e regressão logística, comparando-se os resultados de classificação (CASA NOVA, 2013).

As variáveis independentes do modelo de Casa Nova (2013) constituíram-se de indicadores contábeis, tais como endividamento, imobilização do patrimônio líquido, liquidez, ciclo financeiro, rentabilidade, giro de ativo, margem, saldo de tesouraria. A autora, além disso, agrupou os indicadores por meio da análise de componentes principais, obtendo para cada setor cinco fatores, que serviram de *input* para as análises de inadimplência. O modelo de regressão logística no caso das empresas industriais obteve taxa de acerto geral, variando de 62% a 68% nos anos de 2001 a 2003, para o modelo que utilizou escores fatoriais; porém, quando se considera a classificação das empresas em adimplentes e inadimplentes, observou-se que em torno de metade das empresas inadimplentes foram classificadas corretamente (43% em 2001, 53% em 2002 e 57% em 2003). No caso do setor comercial, o modelo que utilizou os indicadores originais obteve melhor taxa de acertos, variando de 62% a 68% nos anos de 2001 a 2003. Em relação às empresas más pagadoras, observou-se taxa de acerto variando de 46% a 58% nos anos de 2001 a 2003. A autora destaca que, no caso das empresas comerciais, "o modelo logístico não acrescentou informação relevante à decisão, pois se todas as empresas fossem classificadas como adimplentes, a taxa de acerto seria de 54%" (CASA NOVA, 2013, p. 91). As outras técnicas aplicadas pela autora obtiveram resultados equivalentes, o que destaca e reforça a importância de estudos adicionais nessa área para o caso brasileiro, envolvendo a criação e discussão de modelos de classificação e apoio à decisão para clientes finais, pessoa física e empresas.

Soares, Coutinho e Camargos (2012) realizaram um estudo similar ao de Casa Nova (2013), construindo um modelo de crédito por meio de um modelo logístico para classificar empresas com base em indicadores contábeis. Considerou uma base de dados de 72 empresas brasileiras não financeiras; usando como variável dependente a escala de *rating* de crédito da Standard & Poor's, agrupada em quatro categorias (considerando que quanto maior o *rating*, maior a probabilidade de a empresa saldar seus débitos), e como variáveis independentes, indicadores contábil-financeiros (rentabilidade, endividamento, cobertura de juros, imobilização, tamanho da empresa, o fato de estar ou não listada na BM&F Bovespa e o grau de governança corporativa). Apenas as variáveis governança corporativa, tamanho da empresa e índice de cobertura apresentaram discriminação significativa no modelo. A taxa de acertos geral foi de 59,7%, com acertos de 40% para a classe 4 de *rating*; 88% para a classe 3; 3,5% para a escala 2; e 0% para a escala 1 (SOARES; COUTINHO; CAMARGOS, 2012). Assim, observou-se baixa taxa de acerto para os níveis mais baixos de *rating*, fato que os autores colocam como decorrente do tamanho reduzido da amostra para essas classes (SOARES; COUTINHO; CAMARGOS, 2012).

Camargos, Araújo e Camargos (2012), de forma análoga, desenvolveram um modelo de regressão logística para a classificação de clientes de uma instituição bancária pública do estado de Minas Gerais. A amostra foi composta por 9.232 contratos de financiamento

realizados a micro e pequenas empresas no período de 1997 a 2005; desse total, 94,8% das empresas estavam adimplentes e 5,2% estavam inadimplentes no momento da coleta de dados. A variável dependente qualidade do crédito considerou como inadimplentes os clientes com atraso superior a 90 dias no pagamento de pelo menos uma das parcelas do financiamento; os adimplentes, por sua vez, foram considerados aqueles que não estavam atrasados em nenhuma parcela do financiamento. As variáveis independentes consideraram aspectos socioeconômicos dos sócios e avalistas e aspectos econômico-financeiros da empresa, como, por exemplo, valor do contrato, valor dos investimentos fixos do projeto, valor dos recursos próprios no projeto, valor do projeto, capital de giro, ano do faturamento, proporção do financiamento sobre o valor financiado, localização da empresa, setor de atividade, nível de informatização, mercado de atuação, escolaridade dos sócios, experiência dos sócios no negócio, proporção dos bens do avalista sobre o valor financiado, valor dos bens do avalista e renda do avalista (CAMARGOS; ARAÚJO; CAMARGOS, 2012).

O modelo final de Camargos, Araújo e Camargos (2012) reteve as variáveis valor do financiamento, proporção dos bens do avalista sobre o valor financiado, valor dos investimentos fixos do projeto, tempo de atividade da empresa e proporção do faturamento sobre o valor financiado. Os coeficientes da análise indicam que quanto maior o valor financiado, maior a probabilidade de o cliente ser classificado como inadimplente; quanto maior a proporção dos bens do avalista sobre o valor financiado, menor a probabilidade de a empresa se tornar inadimplente; analogamente, quanto maior o valor dos investimentos fixos da empresa, menor a probabilidade de se tornar inadimplente; quanto maior o tempo de atividade da empresa, menor a probabilidade de inadimplência; e quanto maior a proporção do faturamento sobre o valor financiado, maior a probabilidade de inadimplência. A taxa de acerto geral obtida foi de 67,7%, enquanto que para os adimplentes foi de 68,7% e para os inadimplentes foi de 49,7%, considerando o ponto de corte de 0,06. Quando se altera o ponto de corte para 0,1, há aumento da proporção de acertos para o modelo geral (88,5%) e para os adimplentes (92,3%), porém uma redução vertiginosa dos acertos para os inadimplentes (19,8%) (CAMARGOS; ARAÚJO; CAMARGOS, 2012).

Ferreira, Celso e Barbosa Neto (2012) criaram um modelo de risco de crédito para clientes de um banco de varejo no estado de Minas Gerais, totalizando uma amostra de 74 clientes (83,8% adimplentes e 16,2% inadimplentes). As variáveis independentes foram a idade do cliente, renda, o tempo de relacionamento com o banco, o limite de cheque especial, o estado civil e a escolaridade. A taxa de acerto geral foi de 91,9% dos casos; para os adimplentes, houve acerto em 95,2% dos casos, e para os inadimplentes, 75% de acerto. O modelo *logit* binomial indicou que quanto maiores os valores da renda, tempo de relacionamento com o banco e o limite de cheque especial, maior a probabilidade de inadimplência. Em contraposição, quanto maior a idade, menor o risco de não pagamento (FERREIRA; CELSO; BARBOSA NETO, 2012).

Selau e Ribeiro (2009) realizaram um estudo com clientes de cartão de crédito de uma rede de farmácias no estado do Rio Grande do Sul, considerando uma amostra de 11.681 clientes. Foram considerados bons pagadores aqueles que tiveram atrasos de até 30 dias e maus pagadores aqueles que tiveram atrasos superiores a 60 dias. As variáveis independentes consideradas foram sexo, idade, estado civil, escolaridade, renda, tipo de renda, profissão, tipo de ocupação, CEP residencial e comercial, tempo de emprego, crédito com outras instituições, tipo de residência, número de filhos, pagamento de pensão. A taxa de acerto geral do modelo foi de 73% (SELAU; RIBEIRO, 2009).

2.3.1 Histórico

Segundo Lima (2002, p. 77), "a função logística surgiu em 1845, ligada a problemas de crescimento demográfico, problemas em que, até os dias de hoje, essa função é utilizada". Na década de 30, essa metodologia passou a ser aplicada no âmbito da biologia, e posteriormente nas áreas relacionadas a problemas econômicos e sociais. Paula (2002, p. 118) aponta que, apesar de o modelo de regressão logística ser conhecido desde os anos 1950, foi devido a trabalhos do estatístico David Cox, na década de 1970, que essa técnica se tornou bastante popular entre os usuários de Estatística.

A regressão logística tem sido uma das principais ferramentas na modelagem estatística de dados, sendo largamente utilizada em diversos tipos de problema. Paula (2002, p. 118) explica:

> Mesmo quando a resposta não é originalmente binária, alguns pesquisadores têm dicotomizado a variável resposta de modo que a probabilidade de sucesso possa ser modelada por intermédio da regressão logística. Tudo isso se deve, principalmente, à facilidade de interpretação dos parâmetros de um modelo logístico e também pela possibilidade do uso desse tipo de metodologia em análise com objetivo de discriminação (PAULA, 2002, p. 118).

2.3.2 Conceitos

Nos modelos de regressão logística, a variável dependente é, em geral, uma variável binária (nominal ou ordinal) e as variáveis independentes podem ser categóricas (desde que dicotomizadas após transformação) ou contínuas. Cumpre destacar que, na maioria dos casos apresentados na literatura estudada, a regressão logística é apresentada com variável resposta binária. Entretanto, há o caso em que a variável resposta é múltipla, ou seja, com mais de duas categorias (DESAI *et al.*, 1997); inclusive, alguns *softwares* apresentam a opção de utilização de variável resposta múltipla.

Considere o caso em que as observações podem ser classificadas em uma de duas categorias mutuamente exclusivas (1 ou 0). Como exemplo, as categorias poderiam representar um indivíduo que pode ser classificado como cliente bom ou mau.

A variável dependente binária Y pode assumir os valores: 1 - se o i-ésimo indivíduo pertence à categoria dos bons, ou 0 - se o i-ésimo indivíduo pertence à categoria dos maus.

Seja $X = (1, X_1, X_2, ..., X_n)$: vetor onde o primeiro elemento é igual a 1 (constante) e os demais representam as *n* variáveis independentes do modelo.

O modelo de regressão logística é um caso particular dos Modelos Lineares Generalizados (DOBSON, 1990; PAULA, 2002). A função que caracteriza esse modelo é dada por:

$$ln\left(\frac{p(X)}{1-p(X)}\right) = \beta' X = \mathrm{Z}\text{ , onde}$$

$\beta' = (\beta_0, \beta_1, \beta_2, ..., \beta_n)$: vetor de parâmetros associados às variáveis

p (X) = E (Y=1|X): probabilidade de o indivíduo ser classificado como bom, dado o vetor X.

Essa probabilidade é expressa por Neter *et al.* (1996, p. 580):

$$p(X) = E(Y) = \frac{e^{\beta' X}}{1 + e^{\beta' X}} = \frac{e^{Z}}{1 + e^{Z}}$$

2.3.3 Método de escolha das variáveis

Neste trabalho, inicialmente, todas as variáveis serão incluídas para construção do modelo; entretanto, no modelo logístico final, apenas algumas variáveis serão selecionadas. A escolha das variáveis será feita por intermédio do método *forward stepwise*, que é o mais largamente utilizado em modelos de regressão logística. No método *forward stepwise*, as variáveis são selecionadas a cada passo, de acordo com critérios que otimizem o modelo, reduzindo a variância e evitando problemas de multicolinearidade. Somente as variáveis realmente importantes para o modelo são selecionadas. Para detalhes da metodologia, sugere-se a leitura de Canton (1988, p. 28) e Neter *et al.* (1996, p. 348).

2.3.4 Pontos fortes e fracos da aplicação de regressão logística

Fensterstock (2005, p. 48) aponta as seguintes vantagens na utilização de técnicas estatísticas multivariadas na construção de modelos:

- Modelo gerado leva em consideração a correlação entre as variáveis, identificando relações que não seriam visíveis e eliminando variáveis redundantes.
- Consideram as variáveis individual e simultaneamente.
- Usuário pode verificar as fontes de erro e otimizar o modelo.

Particularmente no caso da regressão logística, conforme mencionado no início da seção 2.3, não se exige distribuição normal das variáveis independentes, diferentemente da análise discriminante, sendo a primeira mais flexível em suas condições de uso.

No mesmo texto, o autor também identifica as desvantagens desse tipo de técnica:

- Em muitos casos, a preparação das variáveis demanda muito tempo.
- No caso de muitas variáveis, o analista deve fazer uma pré-seleção das mais importantes, baseando-se em análises separadas.
- Alguns modelos resultantes são de difícil implementação.

2.4 Critérios de Avaliação de Performance

Para avaliar a *performance* do modelo, foram selecionadas duas amostras, uma de validação e outra de teste de mesmo tamanho (3.000 clientes considerados bons e 3.000 considerados maus para cada uma das duas). Os critérios que serão utilizados são apresentados a seguir.

2.4.1 Taxa de Acerto

Mede-se a taxa de acerto por meio da divisão do total de clientes classificados corretamente e pela quantidade de clientes que fizeram parte do modelo. De forma similar, pode-se quantificar a taxa de acertos dos bons e maus clientes.

Em algumas situações, é muito mais importante identificar um cliente bom do que um cliente mau (ou vice-versa); nesses casos, é comum dar-se um peso para a taxa de acertos mais adequada e se calcular uma média ponderada da taxa de acertos.

Neste trabalho, como não se têm informações *a priori* sobre o que seria mais atrativo para a instituição financeira (identificação de bons ou maus clientes), utilizar-se-á o produto entre as taxas de acerto de bons e maus clientes como um indicador de acerto para se avaliar a qualidade do modelo logístico (Ia).

2.4.2 Teste de Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov (KS) é muito utilizado na prática (PICININI; OLIVEIRA, MONTEIRO, 2003; OOGHE; CAMERLYNCK; BALCAEN, 2003; PEREIRA, 2004). É uma técnica não paramétrica para determinar se duas amostras foram extraídas da mesma população (ou de populações com distribuições similares) (SIEGEL, 1975, p. 144). Baseia-se na distribuição acumulada dos escores dos clientes considerados como bons e maus. Para se verificar se as amostras possuem a mesma distribuição, há tabelas estatísticas com o nível de significância e tamanho da amostra (SIEGEL, 1975, p. 309-310).

3 Aspectos Metodológicos

3.1 Descrição e Classificação do Estudo

Uma instituição a ser estudada necessita de uma ferramenta que avalie o grau de risco associado a cada empréstimo para auxiliar no processo de tomada de decisão, pois isso facilitaria que todos os clientes fossem classificados como bons ou maus pagadores, a fim de poder estimar a distribuição de perdas de sua carteira de crédito, obter um *credit rating* e direcionar o gerenciamento das operações de acordo com o risco de inadimplência dos contratantes. Para viabilizar este estudo, foram disponibilizadas, então, informações do histórico de clientes que contrataram um crédito pessoal.

O crédito pessoal é uma operação rápida e prática de crédito ao consumidor. Não é preciso declarar a finalidade que será dada ao empréstimo, o qual é concedido de acordo com a capacidade de crédito do solicitante. Outra característica do produto em questão é a não exigência de bens como garantia de pagamento. Para este estudo, é abordada a modalidade com juros pré-fixados, com prazos de empréstimos variando de 1 a 12 meses.

Considerando essa problemática, a presente pesquisa, quanto aos seus objetivos, pode ser classificada como quantitativa e descritiva. Uma pesquisa quantitativa é aquela que emprega métodos matemáticos e estatísticos na coleta e no tratamento dos dados, assim como na interpretação dos resultados obtidos (RICHARDSON, 1999), o que é coerente com o objetivo proposto de se utilizar a técnica de regressão logística para construir um modelo de análise de crédito. Adicionalmente, o estudo pode ser considerado descritivo, pois busca

definir e classificar a relação entre variáveis (GIL, 2002) pertinentes ao assunto sobre risco de crédito.

3.2 Os Dados

Para a realização do estudo, foram selecionados aleatoriamente, a partir do universo de clientes do banco em estudo, 20.000 contratos de crédito, metade deles considerados bons e a outra metade como maus, realizados no período de agosto de 2009 a fevereiro de 2010, sendo que todos esses contratos já venceram, isto é, a amostra foi coletada após a data de vencimento da última parcela de todos os contratos.

A amostra foi dividida em três subamostras provenientes do mesmo universo de interesse: uma para a construção do modelo, com 8.000 dados (sendo 4.000 bons e 4.000 maus), e duas amostras para validação do modelo construído, cada uma com 6.000 dados (sendo 3.000 bons e 3.000 maus). Cada subamostra tem a sua função específica (ARMINGER; ENACHE, BONNE, 1997, p. 294). A subamostra de construção do modelo é usada para estimação dos parâmetros do modelo; já as subamostras de validação, têm a função de validar os parâmetros e verificar o poder de predição do modelo construído.

3.3 As Variáveis

As variáveis explanatórias ou independentes disponíveis contêm características divididas em dois grupos: Variáveis Cadastrais e Variáveis de Utilização e Restrição. As Variáveis Cadastrais estão relacionadas ao cliente, e as Variáveis de Utilização e Restrição são relativas às restrições de crédito e apontamentos sobre outras operações de crédito do cliente existentes no mercado. Os dois grupos de variáveis são coletados no momento em que o cliente contrata o produto. As variáveis cadastrais utilizadas como *input* do modelo foram: gênero, estado civil, tempo no emprego, tempo de residência fixa, idade, CEP comercial, CEP residencial, telefone comercial e residencial, salário do cônjuge e profissão. As variáveis relacionadas ao contrato de empréstimo foram: número de parcelas, valor da parcela sobre o salário, valor do empréstimo, tipo de crédito, relação entre empréstimo e salário e aquisição do empréstimo.

As variáveis quantitativas foram categorizadas, isto é, transformadas para variáveis categóricas. Inicialmente, foram identificados os decis dessas variáveis. Partindo-se dos decis, o passo seguinte foi analisá-los de acordo com a variável resposta. Foi calculada a distribuição de bons e maus clientes por decil e, em seguida, calculada a razão entre bons e maus, o chamado risco relativo (RR). Grupos que apresentaram risco relativo (RR) semelhante foram reagrupados a fim de se diminuir o número de categorias por variável.

Também para as variáveis qualitativas, foi calculado o risco relativo para se diminuir o número de categorias, quando possível. Conforme Pereira (2004, p. 49), existem duas razões para se fazer uma "nova categorização" das variáveis qualitativas. A primeira é evitar categorias com um número muito pequeno de observações, o que pode levar a estimativas pouco robustas dos parâmetros associados a elas; a segunda é eliminar parâmetros do modelo; se duas categorias têm risco próximo, é razoável agrupá-las em uma única classe.

O RR, além de auxiliar no agrupamento das categorias, ajuda a entender se a categoria em questão está mais ligada a clientes bons ou ruins. Esse método de agrupamento de categorias é explicado por Hand e Henley (1997, p. 527).

Ao trabalhar-se com as variáveis disponibilizadas, os seguintes cuidados foram tomados:

- As variáveis sexo, primeira aquisição e tipo de crédito não foram recodificadas por já se tratarem de variáveis binárias.
- A variável profissão foi agrupada conforme a similaridade da natureza das ocupações.
- As variáveis telefone comercial e telefone residencial foram recodificadas na forma binária como posse ou não.
- As variáveis CEP comercial e CEP residencial foram agrupadas inicialmente de acordo com os três primeiros dígitos; em seguida, foi calculado o risco relativo de cada faixa e posteriormente houve o reagrupamento de acordo com risco relativo semelhante, procedimento idêntico ao adotado por Rosa (2000, p. 17), que é explicado por Hand e Henley (1997, p. 527).
- A variável salário do cônjuge foi removida da análise por ter muitos dados faltantes.
- Foram criadas duas novas variáveis: percentual do valor do empréstimo sobre o salário e percentual do valor da parcela sobre o salário, categorizadas em faixas.

Para o desenvolvimento de um modelo de c*redit scoring*, é preciso definir, também, o que a instituição financeira considera como um bom e um mau pagador. Essa definição, da variável resposta ou variável dependente, também denominada de Definição de *Performance*, está diretamente ligada à política de crédito da instituição. Para o produto em estudo, clientes com 60 ou mais dias de atraso foram considerados maus pagadores (inadimplentes) e clientes com no máximo 20 dias de atraso como bons pagadores, segundo critérios da instituição bancária em estudo. Os clientes que apresentam atrasos no intervalo entre bons e maus foram definidos como indeterminados e excluídos deste trabalho. Na revisão da literatura, observou-se que os diferentes trabalhos apresentaram formas distintas de definição das categorias da variável dependente. Palmuti e Picchiai (2012), por exemplo, consideram inadimplentes clientes com atrasos superiores a 30 dias; já Camargos, Araújo e Camargos (2012) consideraram inadimplentes os clientes com atrasos superiores a 90 dias, de acordo com a natureza dos clientes estudados (pessoas físicas, empreendedores, micro e pequenas empresas e empresas de grande porte). Dessa maneira, no presente estudo, optou-se pela classificação corrente aplicada pelo banco fornecedor dos dados de análises, pois o objetivo final é que o modelo logístico desenvolvido pudesse ser incorporado ao processo de concessão de crédito da instituição.

4 Resultados

4.1 Modelo de Regressão Logística

Para a estimação do modelo de regressão logística, utilizou-se a amostra de 8.000 casos divididos equitativamente nas categorias de bons e maus clientes. Das 53 variáveis independentes, considerando-se k-1 *dummies* para cada variável de *k* níveis, foram incluídas 28 variáveis no modelo, de acordo com o método *forward stepwise*. Seja Z a combinação linear das 28 variáveis independentes ponderadas pelos coeficientes logísticos:

$$Z = B_0 + B_1.X_1 + B_2.X_2 + ... + B_{28}.X_{28}$$

A Tabela 1 apresenta as variáveis selecionadas e as estatísticas geradas pelo modelo logístico, considerando significância de 5%.

Tabela 1 – Modelo de regressão logística

Variável	Coeficiente logístico estimado (B)	*Wald*	Significância	R – correlação parcial	Exp (B)
Sexo masculino	-0,314	35,0381	0,0000	-0,0546	0,7305
Estado civil solteiro	-0,1707	9,4374	0,0021	-0,0259	0,8431
Primeira faixa de tempo de emprego	-0,4848	41,6169	0,0000	-0,0598	0,6158
Segunda faixa de tempo de emprego	-0,2166	12,6825	0,0004	-0,031	0,8053
Primeira faixa de número de parcelas	1,6733	276,6224	0,0000	0,1574	5,3296
Segunda faixa de número de parcelas	0,9658	169,084	0,0000	0,1227	2,627
Penúltima faixa de número de parcelas	0,3051	20,2011	0,0000	0,0405	1,3568
Segunda faixa de tempo de residência fixa	-0,3363	11,2356	0,0008	-0,0289	0,7144
Penúltima faixa de tempo de residência fixa	-0,1451	7,0946	0,0077	-0,0214	0,865
Primeira faixa de valor da parcela	-0,2035	5,3672	0,0205	-0,0174	0,8159
Primeira faixa de valor do empréstimo	0,9633	62,1252	0,0000	0,0736	2,6203
Segunda faixa de valor do empréstimo	0,5915	24,7781	0,0000	0,0453	1,8067
Terceira faixa de valor do empréstimo	0,4683	27,7693	0,0000	0,0482	1,5972
Tipo de crédito: carnê	-1,34	246,7614	0,0000	-0,1486	0,2618
Primeira faixa de idade	-0,7429	29,3706	0,0000	-0,0497	0,4757
Segunda faixa de idade	-0,6435	50,924	0,0000	-0,0664	0,5254
Terceira faixa de idade	-0,2848	12,4401	0,0004	-0,0307	0,7522
Primeira categoria de CEP residencial	-0,3549	9,3714	0,0022	-0,0258	0,7012
Primeira categoria de CEP comercial	-0,29	8,1718	0,0043	-0,0236	0,7483
Segunda categoria de CEP comercial	-0,2888	20,231	0,0000	-0,0405	0,7492
Terceira categoria de CEP comercial	-0,2662	12,9248	0,0003	-0,0314	0,7663
Primeira categoria de profissão	0,3033	10,3013	0,0013	0,0274	1,3543
Terceira categoria de profissão	0,5048	32,2381	0,0000	0,0522	1,6566
Quinta categoria de profissão	0,4752	20,5579	0,0000	0,0409	1,6084
Sexta categoria de profissão	0,1899	7,534	0,0061	0,0223	1,2091
Primeira faixa de relação entre empréstimo e salário	0,2481	9,0609	0,0026	0,0252	1,2816
Terceira faixa de relação entre empréstimo e salário	0,164	6,0906	0,0136	0,0192	1,1782
Primeira aquisição de empréstimo	-0,6513	153,5677	0,0000	-0,1169	0,5213
Constante	0,5868	42,2047	0,0000		

Fonte: Elaborado pelos autores.

4.2 Coeficientes Logísticos das Variáveis Independentes

Com variáveis categóricas, a avaliação do efeito de uma particular categoria deve ser feita em comparação com uma categoria de referência. O coeficiente para a categoria de referência é 0. Para exemplificação, será interpretado o coeficiente da variável Primeira faixa

de número de parcelas, sendo análogas as considerações para as demais. A variável quantidade de parcelas tem 4 níveis. Portanto, devem ser consideradas 3 variáveis *dummies*. Todas as 3 foram incluídas no modelo *stepwise*. A variável Primeira faixa de número de parcelas representa a primeira faixa da escala ordinal para quantidade de parcelas, com os códigos 1 para o nível mais baixo e 0, caso contrário. Analogamente, a variável Segunda faixa de número de parcelas corresponde à segunda faixa, com os códigos 1 para o segundo nível e 0, caso contrário. A categoria referência é o nível mais alto, no caso, a quarta faixa. O coeficiente logístico para Primeira faixa de número de parcelas é positivo, indicando que clientes com empréstimo bancário com menos parcelas (primeira faixa) têm maior probabilidade de serem bons clientes comparativamente àqueles com empréstimo a ser pago com número superior de parcelas (quarta faixa). Observam-se coeficientes positivos e decrescentes para as variáveis Primeira faixa de número de parcelas, Segunda faixa de número de parcelas e Penúltima faixa de número de parcelas. Portanto, quanto maior o número de parcelas, menor a probabilidade de se ter um bom cliente.

Analogamente, foram observados coeficientes positivos e decrescentes para as variáveis: valor do empréstimo e relação entre empréstimo e salário. Em ambas, a categoria mais alta de faixa recebeu o código 0. Portanto, quanto menor o valor do empréstimo e quanto menor a relação percentual entre o valor do empréstimo e o salário do cliente, maior a probabilidade de que este seja adimplente.

Outra variável com coeficientes positivos é a profissão do cliente bancário. Foram criados, inicialmente, 7 códigos para essa variável, conforme a similaridade das ocupações. A sétima categoria recebeu o código 0 na etapa de recodificação dessa variável qualitativa. No modelo, considerando-se o método *forward stepwise*, somente foram incluídas as categorias de profissão primeira, terceira, quinta e sexta, todas com coeficiente positivo. Portanto, o cliente cuja profissão estiver em uma dessas quatro categorias terá maior probabilidade de pagar o empréstimo bancário em comparação àquele com a profissão enquadrada na sétima categoria. Por outro lado, as profissões referentes às categorias segunda e quarta, não incluídas pelo método *forward stepwise*, não apresentam diferença estatística em relação à sétima categoria em termos de probabilidade de serem adimplentes.

Variáveis com coeficiente logístico estimado negativo indicam que a categoria focalizada, em relação à referência, está associada com a diminuição na probabilidade de se ter um bom cliente. Por exemplo, para a variável Primeira aquisição de empréstimo, um cliente na situação de ter o primeiro empréstimo concedido, em comparação a um cliente experiente na obtenção de empréstimos, tem menor probabilidade de se comportar como bom solicitante de apoio financeiro.

A seguir, são listadas outras variáveis com coeficientes negativos e crescentes e suas respectivas interpretações:

- Sexo: o cliente do sexo masculino tem menor probabilidade de adimplência.
- Estado civil: os solteiros têm menor probabilidade de serem bons clientes em relação ao estado civil "outros"; porém, os casados têm a mesma probabilidade que os de outro estado civil de serem adimplentes.
- Quanto menos tempo no atual emprego, quanto menos tempo no endereço atual e quanto menor a idade, menor a probabilidade de adimplência.
- Em termos de valor de cada parcela do empréstimo, valores muito baixos (primeira faixa) reduzem a chance de se ter um cliente adimplente; entretanto, valores a partir da

segunda faixa não se diferenciam do valor mais alto (última faixa) na probabilidade de se saldar a dívida.

- O tipo de crédito com a categoria carnê, em relação à categoria cheque, apresenta menor chance de adimplência.
- também foram incluídas variáveis referentes ao CEP residencial e ao CEP comercial, e as categorias presentes no modelo logístico receberam coeficientes negativos, indicando menor chance de adimplência em relação às respectivas categorias usadas como referência.

As variáveis que mais afetam positivamente a probabilidade de se ter um bom cliente, pode-se perceber, são as da primeira faixa de número de parcelas, da segunda faixa de número de parcelas e da primeira faixa de valor do empréstimo. Os coeficientes de todas as variáveis incluídas no modelo são estatisticamente diferentes de zero ao nível de significância de 5%.

4.3 Testes de Significância do Modelo Logístico

Há dois testes de significância do modelo final: teste Qui-quadrado da mudança no valor de –2LL (Teste Omnibus) e o teste de Hosmer e Lemeshow. Esses testes possibilitam analisar, após a inclusão das variáveis independentes, se o modelo pode ser considerado capaz de realizar as predições com a acurácia desejada.

A estatística de referência L é a função de verossimilhança definida como a probabilidade de obter os resultados da amostra, dadas as estimativas dos parâmetros do modelo logístico. Como essa probabilidade é um valor menor do que 1, convencionou-se usar a expressão –2LL (-2 multiplicado pelo logaritmo decimal da probabilidade – em inglês, *likelihood*). Assim, o resultado –2LL é uma medida da qualidade de ajuste do modelo estimado aos dados. Quanto menor o valor de -2LL, maior a qualidade do ajuste.

A Tabela 2 apresenta os resultados do teste Omnibus, com o valor inicial de –2LL, com apenas a constante no modelo, o seu valor final, a diferença "*improvement*" e o nível descritivo.

Tabela 2 – Teste Omnibus: Teste Qui-quadrado da mudança em –2LL

-2LL	Qui-quadrado *(improvement)*	Graus de liberdade	Nível descritivo
11090,355			
9264,686	1825,669	28	0,0000

Fonte: Elaborado pelos autores.

No modelo de 28 variáveis, a redução na medida –2LL foi estatisticamente significante. O teste Omnibus objetiva demonstrar a capacidade preditiva do modelo, sendo que o resultado, descrito na Tabela 2, foi 1825,669, significativo estatisticamente, confirmando que as variáveis independentes contribuem para melhorar a qualidade das predições no contexto do empréstimo bancário, objeto de atenção neste estudo.

O teste de Hosmer e Lemeshow considera a hipótese estatística de que as classificações em grupo previstas são iguais às observadas. Trata-se de um teste do ajuste do modelo aos dados. A estatística Qui-quadrado apresentou o resultado 3,4307, com 8 graus de liberdade e nível descritivo igual a 0,9045. Esse resultado conduz à não rejeição da hipótese nula do teste, endossando a aderência do modelo aos dados.

4.4 Avaliação da Performance do Modelo Logístico

Após obtido o modelo logístico, foram escoradas as três amostras e calculados o Ia e o KS. Os resultados são apresentados nas Tabelas 3 e 4.

Tabela 3 – Resultados de classificação

Regressão logística										
		Construção			**Validação**			**Validação**		
		Predito →			Predito →			Predito →		
		Mau	Bom	% Acerto	Mau	Bom	% Acerto	Mau	Bom	% Acerto
Observado ↓	Mau	2913	1087	72,8	2169	831	72,3	2175	825	72,5
	Bom	1184	2816	70,4	999	2001	66,7	965	2035	67,8
	Total	4097	3903	71,6	3168	2832	69,5	3140	2860	70,2

Fonte: Elaborado pelos autores.

O modelo logístico apresentou bons resultados de classificação, pois, segundo Picinini, Oliveira e Monteiro (2003) e Selau e Ribeiro (2009), modelos de *credit scoring* com taxas de acerto acima de 65% são considerados bons por especialistas. Cumpre observar que o modelo logístico apresentou maior taxa de acerto nos clientes maus, sendo superior a 70% a taxa de acerto para clientes maus nas três amostras. A taxa de acertos de 70% para os maus pagadores evidencia a qualidade do modelo logístico, pois, como mencionado por Casa Nova (2013), há maior custo da má classificação dos inadimplentes do que dos adimplentes.

O índice de acertos da amostra total obtido foi próximo àqueles obtidos por Casa Nova (2013), Camargos, Araújo e Camargos (2012) e Selau e Ribeiro (2009) em seus estudos. Palmuti e Picchiai (2012) e Ferreira, Celso e Barbosa Neto (2012), cujos estudos são mais próximos da presente pesquisa, tendo em vista que os primeiros estudaram empreendedores e os segundos pessoas físicas, obtiveram maiores taxas de classificações corretas do que as obtidas neste estudo – 87,4% e 91,9%. Todavia, vale ressaltar que as amostras utilizadas por ambos são menores do que as utilizadas neste estudo e que ambos não apresentaram resultados para os modelos de validação.

A Tabela 4 apresenta os resultados dos critérios Ia e KS.

Tabela 4 – Critérios de avaliação

Ia	**Amostra**		
	Construção	**Validação**	**Validação**
Regressão logística	51,3	48,2	49,2
KS	**Amostra**		
	Construção	**Validação**	**Validação**
Regressão logística	38	35	37

Fonte: Elaborado pelos autores.

Os valores KS podem ser considerados bons. Picinini, Oliveira e Monteiro (2003, p. 465) explicam: "o teste de Kolmogorov-Smirnov (KS) é utilizado no mercado financeiro como um dos indicadores de eficiência de modelos de *credit scoring*, sendo que o mercado considera um bom modelo aquele que apresente um valor de KS igual ou superior a 30".

Selau e Ribeiro (2009) explicam que o teste KS busca encontrar a diferença máxima entre duas distribuições acumuladas; assim, na prática se houver uma diferença de pelo menos 30 entre a distribuição dos adimplentes e dos inadimplentes, o modelo consegue discriminar satisfatoriamente os dois grupos, pois a diferença é grande o suficiente para diferenciar os grupos. Portanto, foram aceitáveis todos os resultados gerados na aplicação da regressão logística aos dados fornecidos pela instituição financeira, o que implica na viabilidade de implantação do modelo proposto para análise de crédito da instituição, podendo ter impactos positivos na estratégia de concessão de crédito pessoal desse banco em específico.

5 Considerações Finais

Considerando-se a relevância de modelos estatísticos no apoio à decisão de concessão de crédito, o objetivo deste estudo foi aplicar a regressão logística no desenvolvimento de um modelo de predição de *credit scoring* com base em dados de uma grande instituição financeira.

Os resultados obtidos por meio da regressão logística apresentaram boa qualidade de ajuste aos dados nas três amostras pesquisadas: uma de construção e duas de validação do modelo. O modelo proposto por este estudo para que a instituição pontuasse seus clientes é o modelo logístico com 28 variáveis, exibidas na Tabela 1. Houve maior índice de acertos na predição dos maus clientes, o que indica bom ajuste do modelo. Apesar dos bons índices de acertos de classificação dos clientes, devem ser analisadas as taxas de erros nos dois sentidos: os casos em que erroneamente bons pagadores foram classificados como maus pagadores e vice-versa.

A decisão final sobre a concessão de crédito deve ser tomada com base no modelo logístico e na orientação estratégica da empresa. Se a prioridade for aumentar a participação no mercado, a instituição pode modificar sua classificação de clientes bons e maus pagadores, por exemplo, considerando como bons pagadores clientes com atrasos maiores do que 20 dias. Por outro lado, se a participação de mercado for considerada conveniente, a empresa pode optar por minimizar perdas com inadimplência e considerar como maus pagadores clientes com atrasos inferiores a 90 dias, restringindo, assim, o crédito.

Em geral os resultados encontrados foram semelhantes aos obtidos por outros pesquisadores em estudos de *credit scoring*, em situações semelhantes e distintas às do presente artigo. Apenas os estudos de Palmuti e Picchiai (2012) e Ferreira, Celso e Barbosa Neto (2012) tiveram ajuste superior ao da pesquisa; porém, as amostras utilizadas nesses estudos são menores e os trabalhos não deixaram claros os resultados para amostras de validação. Diante do exposto, o modelo logístico mostra-se como um instrumento promissor para análise de crédito, podendo ser aprimorado com a inclusão de outras variáveis e podendo levar ao aperfeiçoamento dos cadastros de cliente, de modo a incluir tais variáveis relevantes nos registros.

Embora haja limitações inerentes às modelagens estatísticas, no sentido de serem uma abstração da realidade e de incorporarem elementos de incerteza, os resultados proporcionados pela regressão logística foram satisfatórios e podem subsidiar a instituição na atividade de avaliação de pedidos de concessão de crédito bancário. Sugere-se para estudos futuros a replicação do modelo de regressão logística a outras bases de dados de diferentes instituições bancárias, bem como a clientes de outros tipos de produtos financeiros, que não

sejam apenas o crédito pessoal; adicionalmente, outras técnicas para *credit scoring*, tais como análise discriminante, redes neurais, análise envoltória de dados, algoritmos genéticos etc., que podem ser aplicadas com o objetivo de se compararem os ajustes para cada tipo de produto financeiro, podendo-se, então, selecionar o melhor modelo para a prática da organização.

Referências

ANDREEVA, G. European generic scoring models using logistic regression and survival analysis. **Journal of Operational Research Society**, v.57, n.10, p. 1180-1187, 2006.

ARMINGER, G.; ENACHE, D.; BONNE, T. Analyzing credit risk data: a comparison of logistic discrimination, classification trees and feedforward networks. **Computational Statistics**, Berlim, v. 12, n. 2, p. 293-310. 1997.

CAMARGOS, M. A.; ARAÚJO, E. A. T.; CAMARGOS, M. C. S. A inadimplência em um programa de crédito de uma instituição financeira pública de Minas Gerais: uma análise utilizando regressão logística. **REGE**, v. 19, n. 3, p. 473–492. 2012.

CANTON, A. W. P. **Aplicação de modelos estatísticos na avaliação de produtos**. 1988. 253 f. Tese (Livre-Docência) – Concurso para Livre Docente do Departamento de Administração, Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo, São Paulo, 1988.

CASA NOVA, S. P. C. Quanto pior, melhor : Estudo da utilização da análise por envoltória de dados em modelos de análise de inadimplência/insolvência de empresas. **Revista Contemporanea de Contabilidade**, v. 10, n. 19, p. 71–96. 2013.

DESAI, V. S. *et al*. Credit scoring models in the credit union environment using neural networks and genetic algorithms. **IMA J. Mathematics Applied in Business and Industry**, Oxford, v. 8, p. 323-346. 1997.

DOBSON, A. **An introduction to generalized linear models**. London: Chapman & Hall, 1990.

DUARTE JR. *et al*.. Gerenciamento de riscos corporativos: classificação, definições e exemplos. **Resenha BM&F**, São Paulo, n. 134. 1999.

FENSTERSTOCK, F. Credit scoring and the next step. **Business Credit**, New York, v. 107, n. 3, p. 46-49. 2005.

FERREIRA, M. A. M.; CELSO, A. S. S.; BARBOSA NETO, J. E. Aplicação do modelo logil binomial na análise do risco de crédito em uma instituição bancária. **Revista de Negócios**, v. 17, n. 1, p. 41–59. 2012.

FIGUEIREDO, R. P. **Gestão de riscos operacionais em instituições financeiras**: uma abordagem qualitativa. 2001. 187 f. Dissertação (Mestrado em Administração). Curso de Pós-graduação em Administração, Universidade da Amazônia, Belém, 2001.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GITMAN, L. J. **Princípios de Administração Financeira**. São Paulo: Harbra, 1997.

HAIR JR., J. F. *et al*. **Análise multivariada de dados**. São Paulo: Bookman, 2009.

HAND, D. J.; HENLEY, W. E. Statistical classification methods in consumer credit scoring: a review. **Journal of Royal Statistical Society**, London, v. 160, p. 523-541. 1997.

HARRISON, T.; ANSELL, J. Customer retention in the insurance industry: using survival analysis to predict cross-selling opportunities. **Journal of Financial Services Marketing**, London, v. 6, n. 3, p. 229-239. 2002.

LEWIS, E. M. **An introduction to credit scoring**. San Rafael: Fair Isaac and Co. Inc, 1992.

LIMA, J. **A análise econômico-financeira de empresas sob a ótica da estatística multivariada.** 2002. 192 f. Dissertação (Mestrado em Métodos Numéricos em Engenharia) – Curso de Pós-graduação em Tecnologia e Ciências Exatas, Universidade Federal do Paraná, Curitiba, 2002.

LIMA, F. G. *et al*. Aplicação de redes neurais na análise e na concessão de crédito ao consumidor. **Revista de Administração da USP**, São Paulo, v. 44, n. 1, p. 34-45. 2009.

NETER, J. *et al*. **Applied linear statistical models**. Chicago: Irwin, 1996.

OHTOSHI, C. **Uma comparação de regressão logística, árvores de classificação e redes neurais:** analisando dados de crédito. 2003. 147 f. Dissertação (Mestrado em Estatística) – Curso de Pós-graduação em Estatística, Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo, 2003.

OOGHE, H.; CAMERLYNCK, J.; BALCAEN, S. The Ooghe-Joos-De Vos failure prediction models: a cross-industry validation. **Brussels Economic Review**, Brussels, v. 46, p. 39-70. 2003.

PALMUTI, C. S.; PICCHIAI, D. Mensuração do risco de crédito por meio de análise estatística multivariada. **Revista Economia Ensaios**, v. 26, n. 2, p. 7–22. 2012.

PAULA, G. A. **Modelos de Regressão com Apoio Computacional**. Instituto de Matemática e Estatística, 2002. Disponível em: <http://www.ime.usp.br/~giapaula/livro.pdf>. Acesso em: 14 abr. 2011.

PEREIRA, G. H. A. **Modelos de risco de crédito de clientes:** uma aplicação a dados reais. 2004. 104 f. Dissertação (Mestrado em Estatística) – Curso de Pós-graduação em Estatística, Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo, 2004.

PICININI, R., OLIVEIRA, G. M. B.; MONTEIRO, L. H. A. Mineração de critério de credit scoring utilizando algoritmos genéticos. In: SIMPÓSIO BRASILEIRO DE AUTOMAÇÃO INTELIGENTE, 6., 2003, Bauru. **Anais...** SBAI, 2003. CD-ROM.

RICHARDSON, R. J. **Pesquisa social**: métodos e técnicas. 3. ed. São Paulo: Atlas, 1999.

ROSA, P. T. M. **Modelos de Credit Scoring:** Regressão Logística, CHAID e REAL. 2000. 125 f. Dissertação (Mestrado em Estatística) – Curso de Pós-graduação em Estatística, Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo, 2000.

SAMEJIMA, K.; DOYA, K.; KAWATO, M. Inter-module credit assignment in modular reinforcement learning. **Neural Networks**, v. 16, n. 7, p. 985-994. 2003.

SANTI FILHO, A. **Avaliação de Riscos de Crédito**. São Paulo: Atlas, 1997.

SCHRICKEL, W. K. **Análise de crédito**: concessão e gerência de empréstimos. São Paulo: Atlas, 1995.

SECURATO, J. R. **Crédito**: análise e avaliação do risco. São Paulo: Saint Paul, 2002.
SELAU, L. P. R.; RIBEIRO, J. L. D. Uma sistemática para construção e escolha de modelos de previsão de risco de crédito. **Revista Gestão e Produção**, v. 16, n. 3, p. 398–413. 2009.

SEMOLINI, R. **Support vector machines, inferência transdutiva e o problema de classificação**. 2002. 142 f. Dissertação (Mestrado em Engenharia Elétrica) – Curso de Pós-graduação em Engenharia Elétrica, Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas, Campinas, 2002.

SHARMA, S. **Applied multivariate techniques**. New York: John Wiley and Sons, 1996.

SIEGEL, S. E**statística não paramétrica para as ciências do comportamento**. São Paulo: McGraw-Hill, 1975.

SOARES, G. O. G.; COUTINHO, E. S.; CAMARGOS, M. A. Determinantes do Rating de Crédito de Companhias Brasileiras. **Revista Contabilidade Vista & Revista**, v. 23, n. 3, p. 109–143. 2012.

TREVISANI, A. T. *et al*. Qualidade de dados: desafio crítico para o sucesso do business intelligence. In: CONGRESSO LATINO AMERICANO DE ESTRATÉGIA, 18., 2004, Itajaí. **Anais...** SLADE, 2004. CD-ROM.