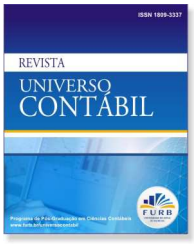

Revista Universo Contábil, ISSN 1809-3337
Blumenau, v. 8, n. 2, p. 84-102, abr./jun., 2012

doi:10.4270/ruc.2012214
Disponível em www.furb.br/universocontabil

APLICAÇÃO DE REGRESSÃO LOGÍSTICA E ALGORITMOS GENÉTICOS NA ANÁLISE DE RISCO DE CRÉDITO[1]

LOGISTIC REGRESSION AND GENETIC ALGORITHMS APPLIED TO CREDIT RISK ANALYSIS

APLICACIÓN DE LA REGRESIÓN LOGÍSTICA Y DE LOS ALGORITMOS GENÉTICOS EN EL ANÁLISIS DEL RIESGO DE CRÉDITO

Maria Aparecida Gouvêa
Doutora em Administração pela FEA/USP
Professora do Departamento de Administração da FEA/USP
Endereço: Rua Professor Luciano Gualberto, 908, FEA1,
CEP: 05508-900 – São Paulo/SP – Brasil
E-mail: magouvea@usp.br
Telefone: (11) 3091-6044

Eric Bacconi Gonçalves
Mestre em Administração pela FEA/USP
Endereço: Rua Professor Luciano Gualberto, 908, FEA1
CEP: 05508-900 – São Paulo/SP – Brasil
E-mail: eric.goncalves@telefonica.com.br
Telefone: (11) 3091-6044

Daielly Melina Nassif Mantovani
Doutoranda em Administração pela FEA/USP
Endereço: Rua Batista do Carmo 33
CEP: 01535-020 – São Paulo/SP – Brasil
E-mail: daimantovani@terra.com.br
Telefone: (11) 2985-8820

RESUMO

A tomada de decisões de concessão de crédito baseia-se fundamentalmente na avaliação do risco de inadimplência dos potenciais contratantes de produtos de crédito. Com o avanço tecnológico, modelos estatísticos foram desenvolvidos para dar sustentação à análise de solicitações de crédito, que há algumas décadas era realizada muitas vezes qualitativamente.

[1] Artigo recebido em 10.05.2011. Revisado por pares em 29.10.2011. Reformulado em 19.12.2011. Recomendado para publicação em 21.12.2012 por Ilse Maria Beuren (Editora). Publicado em 30.04.2012. Organização responsável pelo periódico: FURB.

O objetivo deste estudo é a apresentação do uso de regressão logística e algoritmos genéticos para a classificação de bons e maus pagadores em financiamentos bancários e a identificação do melhor modelo em termos de qualidade de ajuste. A partir de uma amostra de 14.000 dados, fornecida por uma grande instituição financeira brasileira, foram aplicadas as duas técnicas. A regressão logística apresentou melhor ajuste aos dados. Este estudo ilustrou os procedimentos a serem adotados por uma empresa para identificar o melhor modelo de concessão de crédito, a partir do qual é possível direcionar a estratégia da instituição no processo de avaliação de solicitações de empréstimos bancários.

Palavras-chave: Risco de crédito. Modelos de *credit scoring*. Regressão logística. Algoritmos genéticos.

ABSTRACT

The taking of decisions of credit concession is based basically on the evaluation of the insolvency risk of potential contractors of credit products. With the technological advance, statistical models have been developed to support the analysis of credit requests, which was many times carried through qualitatively some decades ago. The goal of this study is to present the use of logistic regression and genetic algorithms for sorting good and bad payers in bank financing and the identification of the best model in terms of goodness-of-fit. From a sample of 14,000 data, supplied by a great Brazilian financial institution, the two techniques were applied. Logistic regression presented the best goodness-of-fit. This work illustrated the procedures to be adopted by a company to identify the best model of credit concession, from which it is possible to direct the strategy of the institution in the evaluation process of bank loan requests.

***Keywords**: Credit risk. Credit scoring models. Logistic regresion. Genetic algorithms.*

RESUMEN

Las decisiones de concesión de crédito se basan principalmente en la evaluación del riesgo de impago de los contratistas potenciales de esos productos. Con los avances tecnológicos, modelos estadísticos fueron desarrollados para apoyar el análisis de las solicitudes de crédito, que hace unas pocas décadas se realizaban, con frecuencia, cualitativamente. El objetivo de este estudio es presentar el uso de la regresión logística y de los algoritmos genéticos para la clasificación de buenos y malos pagadores en la financiación bancaria y identificar el mejor modelo en términos de calidad de ajuste. A partir de una muestra de 14.000 datos, proporcionados por una gran institución financiera brasileña, las dos técnicas fueron aplicadas. La regresión logística mostró un mejor ajuste a los datos. Este estudio ha puesto de manifiesto los procedimientos que una empresa puede adoptar para identificar el mejor modelo de concesión de préstamos, que podrá dirigir la estrategia de la institución en la evaluación de las solicitudes de préstamos bancarios.

***Palabras clave**: Riesgo de crédito. Modelos de credit scoring. Regresión logística. Algoritmos genéticos.*

1 INTRODUÇÃO

Com a estabilidade da moeda, atingida no Plano Real em 1994, os empréstimos financeiros passaram a ser um bom negócio para os bancos, que já não obtinham vultuosos lucros provenientes da desvalorização da moeda (ROSA, 2000). Após o fim do período

inflacionário, houve necessidade de aumentar as alternativas de investimento para substituir a rentabilidade do período de inflação. Desde então as instituições têm se preocupado em aumentar suas carteiras de crédito. Entretanto, o empréstimo não poderia ser oferecido indiscriminadamente a todos aqueles clientes que o solicitassem, sendo necessárias formas de avaliar o candidato ao crédito.

Ao fazer uma solicitação de crédito, o cliente preenchia uma proposta que era avaliada por um ou mais analistas, que apresentavam um parecer em relação ao pedido (SEMOLINI, 2002). Este processo era lento, não permitindo a análise de muitos pedidos. Por esse motivo, as instituições financeiras começaram a adotar os modelos de *credit scoring* baseados em dados históricos de clientes existentes para avaliação da probabilidade de inadimplência de um futuro cliente. Os modelos de *credit scoring* são amplamente utilizados pelas instituições financeiras em geral e especificamente pelos bancos com algumas finalidades: reduzir os custos do processo de concessão de crédito; reduzir o risco de não recebimento do valor concedido; melhorar o processo de decisão de crédito e reduzir tempo e esforço empregados nessa atividade (ABDOU, 2009).

A disseminação do mercado de crédito nas últimas décadas impulsionou o desenvolvimento e aplicação de modelos de *credit scoring*, cada vez mais sofisticados, em instituições financeiras todo o mundo, tornando-se uma ferramenta crítica na operação dos bancos (ABDOU, 2009). Os modelos de *credit scoring* são específicos para aprovação em cada produto de crédito. Os produtos de crédito podem ser: crédito pessoal, cheque especial, financiamentos, entre outros. Nesse estudo, o produto em questão é o crédito pessoal.

A questão que se configura no presente estudo é a carência de um procedimento de avaliação de solicitações de crédito com embasamento de uma modelagem estatística que ofereça bom ajuste aos dados de clientes bancários. O objetivo do estudo é a apresentação do uso de regressão logística e algoritmos genéticos para a classificação de bons e maus pagadores em financiamentos bancários. Busca-se também comparar os seus resultados, identificando-se a técnica com melhor aderência aos dados pesquisados provenientes de uma grande instituição financeira.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Risco de Crédito

O ponto principal para a concessão de crédito é a avaliação do risco. Se o risco for mal avaliado a empresa certamente irá perder dinheiro, quer seja pelo aceite de clientes que irão gerar prejuízos ao negócio, quer seja pela recusa de clientes bons que gerariam lucros ao negócio. Empresas que têm uma avaliação melhor que as concorrentes na concessão de crédito levam vantagem em relação às demais, por ficarem menos vulneráveis às conseqüências decorrentes de decisões equivocadas no fornecimento de crédito.

A literatura aponta diversas variáveis como preditoras do risco de um cliente, tais como estado civil, renda, idade, entre outras (ABDOU, 2009). O autor, no estudo realizado em bancos egípcios incluiu em seu modelo de análise as seguintes variáveis: montante do empréstimo, duração do empréstimo, tipo de empréstimo, motivo do empréstimo, idade, estado civil, gênero, dependentes, profissão, nível de escolaridade, *house status* (casa própria, alugada, financiada etc.), *utility bills* (contas básicas de casa, como água e energia), relatório CBE (*credit by examination*), reputação, garantias, visita de campo, estudo de viabilidade, *status* de cartão de crédito, relação com outros bancos, empréstimos em outros bancos, posse de veículo e documentos formais (ABDOU, 2009).

No estudo realizado em bancos britânicos, Bijak e Thomas (2012) incluíram no modelo as seguintes variáveis: estado civil, *status* residencial, número de filhos, tempo em que vive na moradia atual, tempo em que viveu na moradia anterior, telefone residencial, ocupação, tempo no emprego atual, tempo no emprego anterior, renda líquida, tempo em que

é correntista do banco, número de cartões de crédito, montante do empréstimo, prazo do empréstimo, motivo de empréstimo, posse de seguro (BIJAK; THOMAS, 2012).

Por sua vez, Yap et al. (2011) abordaram na construção de seu modelo as variáveis, gênero, idade, bairro de residência, ocupação, raça, estado civil, número de dependentes e número de carros. Mavri et al. (2008) consideraram as variáveis gênero, estado civil, nível de escolaridade, idade, tempo no emprego atual, renda mensal, posse de propriedades, cartões de crédito, atividade bancária e limite de crédito solicitado. Sustersic et al. (2009), na modelagem em um banco eslováco incluíram as variáveis gênero, número de empréstimos anteriores, existência de refinanciamentos, taxa de juros de empréstimos anteriores, fluxo de caixa do cliente, forma de pagamento do empréstimo, data de aprovação do último empréstimo.

Quaisquer que sejam as variáveis independentes incluídas no modelo como preditoras do *status* do cliente, a variável dependente dos modelos de *credit scoring* é uma variável binária que diferencia os bons e maus clientes (YAP et al., 2011). Os trabalhos mencionados possuem, cada um, variáveis específicas incluídas de acordo com as características e necessidades da instituição e do país estudado. Entretanto, nota-se que invariavelmente são consideradas na modelagem características demográficas do indivíduo e variáveis relacionadas (como idade, gênero e estado civil) ao seu *status* financeiro (como posses, garantias e comportamento em outros empréstimos).

A avaliação do risco de um potencial cliente pode ser feita por meio de:

a) julgamento - envolve uma análise mais qualitativa, pela avaliação simples de dados numéricos guiada pela própria perspectiva e experiência do gerente de crédito (ABDOU, 2009). A análise por julgamento, em geral, leva em conta as seguintes variáveis: caráter do cliente, capital, circunstância do empréstimo, capacitação e garantias adicionais (YAP et al., 2011);
b) classificação do tomador via modelos de avaliação - envolve uma análise mais quantitativa, com técnicas estatísticas convencionais e avançadas. As primeiras compreendem, por exemplo, a regressão linear, a análise discriminante, a análise *probit* e a regressão logística. As técnicas avançadas são representadas pelos algoritmos genéticos, modelos *fuzzy* e sistemas especialistas (ABDOU, 2009).

As empresas que trabalham com concessão de crédito podem utilizar as duas formas combinadas de avaliação do risco.

2.2 Modelos de *Credit Scoring*

A função de um modelo de *credit scoring* é mensurar o risco, sendo, portanto, uma ferramenta que auxilia na decisão de fornecer ou não o crédito para o solicitante. Informações cadastrais e comportamentos anteriores dos clientes são considerados na modelagem. O resultado do modelo é identificar perfis de clientes que sejam atrativos para a empresa conceder o crédito. Há sete passos para a construção de um modelo de *credit scoring*:

a) 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 são fundamentais para o sucesso do modelo (TREVISANI et al., 2004);
b) classificação dos clientes de acordo com o padrão de comportamento e definição da variável resposta (tipo de cliente) - as instituições têm sua própria política de crédito e os conceitos de bons e maus clientes podem mudar dependendo de cada uma. Na realidade, nessa classificação, além de clientes bons e maus, também existem os clientes excluídos, aqueles que possuem características peculiares e que não devem ser considerados (por exemplo, trabalha na instituição) e os clientes indeterminados, aqueles que estão na fronteira entre serem bons ou maus. Na

prática, as instituições consideram apenas os clientes bons e maus para fazer o modelo devido à maior facilidade de trabalhar com modelos de resposta binária. Esta tendência de trabalhar apenas com clientes bons e maus também é observada nos trabalhos acadêmicos (HAND; HENLEY, 1997; ROSA, 2000; SEMOLINI, 2002; OHTOSHI, 2003; LIMA et al., 2009).

c) seleção de amostra aleatória representativa da base histórica - com a base e a variável resposta definidas, selecionam-se amostras representativas de clientes bons e maus. É importante que as amostras de bons e maus clientes tenham o mesmo tamanho para se evitar qualquer viés devido à diferença de tamanhos. Não existe um número fixo para a amostra, mas Lewis (1992, p. 31) sugere pelo menos 1.500 clientes bons e 1.500 clientes maus para serem propiciados resultados robustos. Costuma-se trabalhar com duas amostras, uma para construção do modelo e outra para sua validação. Neste estudo foi possível o acesso aos dados de 14.000 clientes, sendo 7.000 bons e 7.000 maus;
d) análise descritiva e preparação dos dados - consiste em analisar, segundo critérios estatísticos, cada variável a ser utilizada no modelo;
e) escolha e aplicação das técnicas a serem utilizadas para a construção do modelo - neste estudo utilizaram-se Regressão Logística e Algoritmos Genéticos. Hand e Henley (1997) destacam ainda a Análise Discriminante, Regressão Linear e Árvores de Decisão, como métodos utilizados na prática. Alguns estudiosos também têm utilizado Análise de Sobrevivência (HARRISON; ANSELL, 2002; ANDREEVA, 2003). Não existe um método claramente melhor que os demais, tudo depende de como a técnica escolhida se ajusta aos dados;
f) definição dos critérios de comparação dos modelos - será definida a medida de comparação dos modelos, normalmente pelo indicador de acertos e a estatística de Kolmogorov-Smirnov (KS);
g) seleção e implantação do melhor modelo - por meio dos critérios previamente 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.

O Quadro 1 apresenta os resultados obtidos por estudos similares consultados.

Referência	**Técnica**	**Amostra**	**Porcentagem de acerto**
Yap et al. (2011)	Regressão logística	2765 casos de uma instituição local na Malásia	71,52%
Mavri et al. (2008)	Regressão logística	350 casos de um Banco Europeu – solicitação de cartão de crédito	71,87%
Šušteršič et al. (2009)	Algorítmos genéticos	581 casos de um Banco Eslováco – solicitação de empréstimos	76,5% no modelo 1 (seleção das amostras pelo modelo de Kohonen) 72,7% no modelo 2 (seleção aleatória das amostras de treinamento de validação)
Brown e Mues (2012)	Regressão Logística	Cinco conjuntos de dados: - Banco Benelux 1: 2974 casos - Banco Bebelux 2: 7190 casos - Banco Austrália: 547 casos - Banco Alemanha: 1000 casos - Banco Benelux 3: 1197 casos	- Benelux 1: 76,9% - Bebelux 2: 78,7% - Austrália: 90,6% - Alemanha: 76,7% - Benelux 3: 63,4%
Ulises e Carmon (2011)	Regressão Logística	- 200 casos – Fundo Rotativo de Ação e Cidadania Recife/Brasil	80%
Brito e Assaf Neto (2008)	Regressão Logística	- 60 empresas listadas na Bovespa classificadas solventes /insolventes	90%

Quadro 1 – Resultados obtidos por modelos de *credit scoring* na literatura pesquisada

2.3 Regressão Logística

A regressão logística é uma técnica de análise multivariada utilizada para aferição da probabilidade de ocorrência de um evento e para identificação das características dos elementos pertencentes a cada categoria estabelecida pela dicotomia da variável dependente. 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).

Esta técnica é preferida à análise discriminante por ser similar à regressão, com testes estatísticos diretos, habilidade de incorporar efeitos não-lineares e diversos tipos de diagnósticos. Na estimação dos coeficientes da função logística, esta técnica procura maximizar a verossimilhança de que um evento ocorra (HAIR JR et al., 2009).

No campo de análise de risco, é a técnica mais utilizada para a criação de modelos de *credit scoring* (BIJAK; THOMAS, 2012) juntamente com a análise discriminante (MAVRI et al., 2008; YAP et al., 2011).

2.4 Algoritmos Genéticos

Os algoritmos genéticos são uma família de modelos computacionais inspirados na evolução, que incorporam uma solução potencial para um problema específico numa estrutura semelhante à de um cromossomo e aplicam operadores de seleção, cruzamento (*cross-over)* e mutação a essas estruturas de forma a preservar informações críticas relativas à solução do problema. Normalmente, os AG's são vistos como otimizadores de funções, embora a quantidade de problemas para os quais os AG's se aplicam seja bastante abrangente.

A idéia dos algoritmos genéticos se assemelha à evolução das espécies proposta por Darwin: os algoritmos vão evoluindo com o passar das gerações e os candidatos à solução do problema que se quer resolver permanecem vivos e se reproduzem (BACK; LAITINEN; SERE, 1996). Surgiram na década de 1960 desenvolvidos pelo pesquisador John Holland, inspirado nos estudos acerca da evolução natural e tornaram-se a base da computação evolutiva (SILVA, 2011).

Essa técnica é eficiente para problemas de otimização e trabalha da seguinte forma: inicia-se com uma população de possíveis soluções ao problema e a partir delas são criadas novas gerações de soluções, partindo-se do princípio de que as novas soluções são uma evolução das anteriores, portanto, são melhores (SUSTERSIC et al., 2009).

O primeiro estágio trata da criação da população de soluções candidatas ao problema. Com esse conjunto inicial, passa-se a uma fase de avaliação, em que a *fitness function* verifica a adequação de cada uma delas ao problema Os valores obtidos pela função para cada solução é rankeado pelo algoritmo, sendo selecionadas as soluções com melhor desempenho para a fase seguinte, que é a reprodução. Nessa última fase, as soluções com melhor valor de *fitness* compõem um novo conjunto de soluções candidatas e o ciclo se repete com o refinamento das soluções (SUSTERSIC et al., 2009).

A estrutura das soluções candidatas tem uma representação cromossômica que sofre mutações, cruzamentos e seleção natural, criando-se soluções mais aptas após esse processo, tal qual as teorias da genética. As interações cessam quando as soluções encontradas tornam-se estáveis, isto é, não mais melhoram (SILVA, 2011). O fluxograma apresentado a seguir ilustra a dinâmica de um algoritmo genético.

Os algoritmos genéticos têm sido amplamente utilizados por sua capacidade de solucionar problemas complexos nas áreas de gestão, finanças, gestão bancária e logística (ABDOU, 2009).

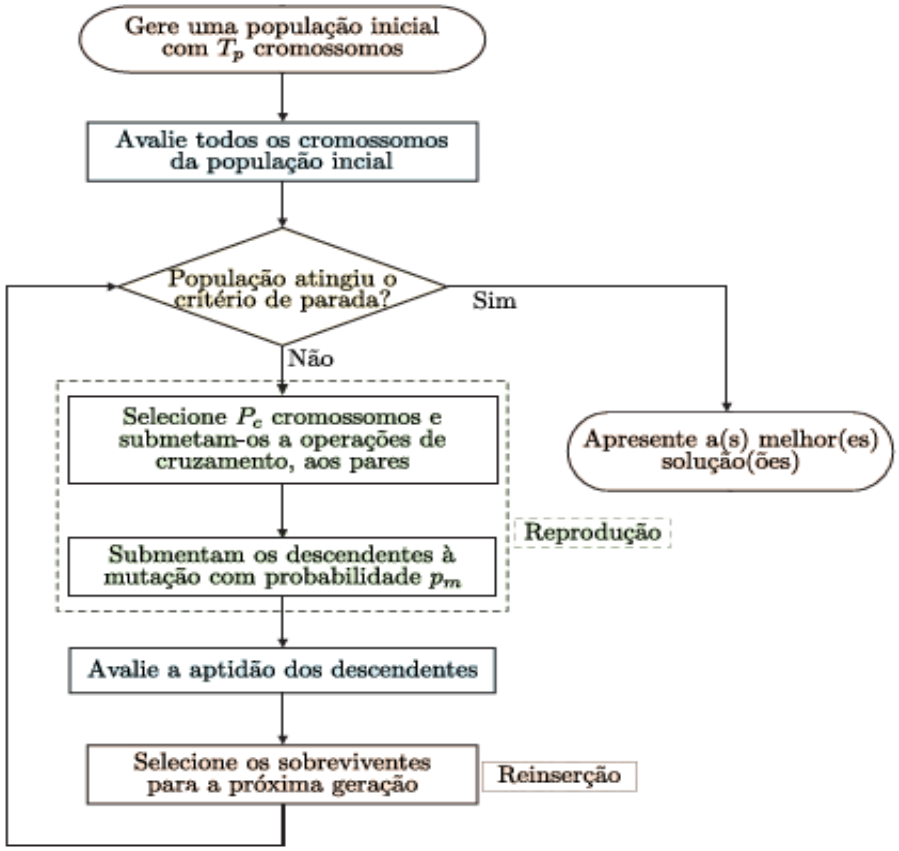

Figura 1 – Fases do algoritmo genético
Fonte: Silva (2011, p. 30).

Primeiramente, é estabelecida a população inicial de cromossomos (soluções candidatas) de forma aleatória, determinada pelo próprio algoritmo, de forma que haja pontos espalhados por todo o espaço de busca do algoritmo. A seguir, deve-se estabelecer uma medida de aptidão ao cromossomo, isto é, um escore que o diferencie, com base na qualidade que ele represente para a solução (SILVA, 2011). O valor de escore do cliente é dado por:

$$S_j = \sum_{i=1}^{72} w_i \left(p_{ij} \right)$$

Onde:

S_j = Escore obtido pelo cliente j

w_i = Peso relativo à categoria i

p_{ij} = Indicador binário igual a 1, se o cliente j possui a categoria i e 0, caso contrário.

Para se definir se o cliente é bom ou mau foi utilizada a seguinte regra:

Se $S_j \geq 0$, o cliente é considerado bom

Se $S_j < 0$, o cliente é considerado mau

O algoritmo deve encontrar o vetor $W=[w_1, w_2, ..., w_{72}]$ que resulte em um critério de classificação com uma boa taxa de acertos na predição do desempenho de pagamento do crédito. A seguir, parte-se para a seleção, que representa o papel da seleção natural no processo de evolução, selecionando os melhores cromossomos para as próximas fases. A seleção é essencial para se estabelecer a pressão seletiva adequada ao ambiente. A pressão determina quanto o ambiente é ou não favorável ao cromossomo, moldando o privilégio de

sobrevivência de um cromossomo sobre os demais. Quanto maior a pressão, maior a probabilidade de os cromossomos aptos se sobressaírem (SILVA, 2011).

A forma como os cromossomos são selecionados depende do operador de seleção escolhido. Alguns deles são (SILVA, 2011, p. 33-34):

a) seleção estocástica com reposição (roleta) - baseada na aptidão dos cromossomos, quanto mais apto, maior a probabilidade de ser selecionado;
b) seleção por torneio simples - um grupo é selecionado aleatoriamente e será escolhido o cromossomo com maior aptidão;
c) seleção por torneio estocástico - um grupo é selecionado aleatoriamente e será escolhido o cromossomo de forma estocástica, isto é, quanto mais apto, maior a chance de ser escolhido;
d) seleção por ordenação - os cromossomos são ordenados do menos ao mais apto e recebem cada um uma probabilidade associada;
e) seleção elitista - seleciona diretamente os melhores cromossomos da população.

A etapa seguinte consiste no cruzamento, isto é, na troca de material genético entre os cromossomos pais, para a geração dos cromossomos filhos. O operador de seleção escolhido pelo pesquisador seleciona pares de cromossomos na base de dados, que são cruzados (mesclados) gerando seus descendentes. Serão realizados tantos cruzamentos, quanto determinado pelo parâmetro taxa de cruzamento. Não se deve gerar cromossomos com genes repetidos. Há três tipos de processos de cruzamento (SILVA, 2011, p. 35):

a) cruzamento simples - um ponto aleatório dos cromossomos pais é escolhido e nesse ponto é feito um corte. A parte 1 do pai A é fundida à parte 2 do pai B, gerando o filho A. Analogamente, a parte 2 do pai A é fundida à parte 1 do pai B, gerando o filho B;
b) cruzamento múltiplo - dois ou mais pontos aleatórios de corte são estabelecidos;
c) cruzamento uniforme - cria-se uma máscara em que cada pai determina uma parte da sequência genética dos filhos.

A mutação altera características genéticas do cromossomo, criando novas características que não existiam na população em análise. Essa mutação ocorre de acordo com uma probabilidade definida pelo pesquisador e que deve ser pequena. Após a reprodução, ocorre a reinserção, que determina os cromossomos que serão levados para a próxima rodada do algoritmo, até que se atinja o critério de parada.

Silva (2011) estabelece alguns critérios para que se determine a parada do algoritmo: a) esgotamento do número máximo de gerações pré-estabelecido; b) esgotamento do tempo máximo de processamento estabelecido; c) encontro de um cromossomo com aptidão superior a um valor pré-determinado; d) estagnação da população.

3 ASPECTOS METODOLÓGICOS

Apresentam-se a seguir algumas informações sobre os dados coletados, as variáveis de interesse e os critérios de avaliação de *performance*. Para a aplicação de regressão logística foi utilizado o *software SPSS for Windows* v.11.0; para o algoritmo genético foi utilizado um programa desenvolvido pelos autores em *Visual Basic*.

3.1 Os Dados

Para a realização deste estudo foram disponibilizadas informações do histórico de clientes que contrataram um crédito pessoal junto a uma instituição financeira, a qual necessita de uma ferramenta que avalie o grau de risco associado a cada empréstimo para auxiliar o processo de tomada de decisão. A partir do universo de clientes deste banco, foram selecionados aleatoriamente 7.000 contratos de crédito bons e 7.000 considerados maus, no

período de agosto/2009 a fevereiro/2010, sendo que todos estes contratos já venceram, isto é, a amostra foi coletada após a data de vencimento da última parcela de todos os contratos.

No trabalho a amostra é dividida em duas sub-amostras provenientes do mesmo universo de interesse: uma para construção do modelo, 8.000 dados (sendo 4.000 bons e 4.000 maus) e outra para validação do modelo construído, 6.000 dados (sendo 3.000 bons e 3.000 maus). A sub-amostra de construção do modelo é usada para estimação dos parâmetros do modelo e a sub-amostra de validação tem como função verificar o poder de predição dos modelos construídos (ARMINGER; ENACHE; BONNE, 1997).

3.2 Definição da Variável Resposta

Para o desenvolvimento de um modelo de c*redit scoring* é preciso definir, num primeiro momento, o que a instituição financeira considera como um bom e mau pagador. Esta definição, da Variável Resposta, também denominada de Definição de *Performance*, está diretamente ligada à política de crédito da instituição. A instituição fornecedora dos dados amostrais adota a seguinte definição: clientes com 60 ou mais dias de atraso são considerados maus (inadimplentes) e clientes com no máximo 20 dias de atraso como bons.

A mensuração do atraso é calculada por meio da parcela paga com maior atraso pelo cliente. Por exemplo, um cliente que atrasou três parcelas por vinte dias consecutivos ainda assim é considerado um bom cliente, ao passo que um cliente que tenha atrasado uma parcela por sessenta dias é considerado mau. Os clientes que apresentam atrasos no intervalo entre bons e maus foram definidos como indeterminados e desconsiderados neste trabalho. Apesar de o universo de clientes da instituição bancária focalizada neste estudo ser constituído de três tipos de clientes (bons, maus e indeterminados), a amostra de 14.000 dados incluiu somente os clientes classificados como bons ou maus, sendo 7000 clientes em cada situação.

3.3 Variáveis Independentes

Foram consideradas as seguintes variáveis cadastrais (relacionadas ao perfil do cliente) e de utilização e restrição (relativas às restrições de crédito e apontamentos sobre outras operações de crédito do cliente): sexo, estado civil, fone residencial, fone comercial, tempo no emprego atual, salário do cliente, quantidade de parcelas a serem quitadas, primeira aquisição, tempo na residência atual, valor da parcela, valor total do empréstimo, tipo de crédito, idade, CEP residencial, CEP comercial, código da profissão, nome da profissão, salário do cônjuge.

3.4 Tratamento das Variáveis

Inicialmente, as variáveis quantitativas foram categorizadas. Para este tratamento, foram identificados os decis destas variáveis. Partindo-se dos decis, o passo seguinte foi analisá-los de acordo com a variável resposta (rótulo: TIPO). Foi calculada a distribuição de bons e maus clientes por decil e em seguida obtida a razão entre bons e maus, o chamado risco relativo (RR).

Grupos com RR semelhante foram reagrupados para se diminuir o número de categorias por variável. Para as variáveis qualitativas foi calculado o risco relativo para se diminuir o número de categorias, quando possível. Conforme Pereira (2004, p. 49), há 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 segunda é a redução de parâmetros do modelo; se duas categorias apresentam risco próximo, é razoável agrupá-las numa 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. Quando o resultado é muito acima de 1,

indica que essa característica está mais ligada ao perfil de bom cliente; para o resultado menor que 1 interpreta-se que a característica está relacionada aos maus clientes. No caso de a razão ser igual a 1, infere-se que essa característica não discrimina bons e maus clientes (HAND; HENLEY, 1997). Foram obtidas 72 categorias das variáveis originais para serem utilizadas na construção dos modelos.

3.5 Implementação dos Modelos

3.5.1 Regressão logística

As variáveis independentes focalizadas neste estudo foram codificadas na forma de variáveis *dummies*. Para cada variável o número de categorias (k) determinou o número de variáveis *dummies* (k-1) incluídas no processamento da regressão logística. O procedimento empregado para a inclusão das variáveis independentes no modelo logístico foi o *forward stepwise*. Em modelos *forward stepwise* inicia-se apenas com o termo da constante, exceto quando se omite este parâmetro na especificação da modelagem, e em cada passo é introduzida a variável com o menor nível de significância para o escore estatístico, desde que este seja menor do que um valor de remoção (*cutoff*), definido como 0,05 neste trabalho.

A estatística L (em inglês, *likelihood*) é a função de verossimilhança definida como a probabilidade de se obterem os resultados da amostra, segundo as estimativas dos parâmetros do modelo. Convencionou-se usar a expressão –2LL (-2 multiplicado pelo logaritmo decimal da probabilidade). Quanto menor o resultado –2LL, maior a qualidade do ajuste.

A probabilidade de o cliente ser bom pagador é dada, segundo o modelo logístico, por:

$$p = \frac{e^z}{1+e^z}$$

Considerando-se k-1 *dummies* para cada variável de k níveis, foram geradas 53 variáveis independentes, das quais foram incluídas 28 no modelo. Neste estudo, Z é a combinação linear das 28 variáveis independentes ponderadas pelos coeficientes logísticos.

3.5.2 Algoritmos genéticos

Cada uma das 72 categorias de variável recebeu um peso aleatório inicial. Foi incluída uma constante aditiva incorporada à equação linear.

Seguindo as fases de um algoritmo genético, têm-se:

a) Início - gerou-se uma população de 200 casos, cada cromossomo contendo 73 genes. O peso inicial w_i de cada um dos genes foi gerado aleatoriamente no intervalo [-1,1] (PICININI; OLIVEIRA; MONTEIRO, 2003);
b) Função de Aptidão (Fitness) - cada cliente foi associado ao cálculo de um escore e classificado como bom ou mau. Comparando-se com a informação já conhecida *a priori* sobre a natureza do cliente, pode-se calcular a precisão de cada cromossomo. O indicador de acertos (Ia) será a função de aptidão, ou seja, quanto maior o indicador, melhor será o cromossomo;
c) Seleção - foi usado um elitismo de 10%, ou seja, para cada nova geração, os vinte melhores cromossomos são mantidos e os outros cento e oitenta são formados por cruzamento e mutação. Essa medida foi utilizada por garantir que os melhores cromossomos da população corrente sejam incluídos na geração seguinte;
d) Cruzamento (*Cross-Over*) - para a escolha dos pais para o cruzamento, foi utilizado o método conhecido como roleta (*roulette wheel*) para seleção dentre os vinte cromossomos que foram mantidos (CHEN; HUANG, 2003). Neste método, cada indivíduo recebe uma probabilidade de ser sorteado de acordo com seu valor de função de aptidão. Para o processo de troca de material genético, foi usado um

método conhecido como cruzamento uniforme (PAPPA, 2002). Cada gene do cromossomo filho é escolhido aleatoriamente entre os genes de um dos pais, enquanto o segundo filho recebe os genes complementares do segundo pai;

e) Mutação - no processo de mutação, cada gene do cromossomo é avaliado independentemente e tem probabilidade de 0,5% de sofrer mutação. Sempre que um gene for escolhido para a mutação, a alteração genética é realizada, adicionando-se um pequeno valor escalar k neste gene;
f) Verificação do critério de parada - como critério de parada, foi definido um número máximo de gerações igual a 600. Após as seiscentas iterações, o cromossomo com maior aptidão será a solução (método elitista).

3.6 Critérios de Avaliação de *Performance*

Os critérios de avaliação de *performance* indicam quão adequado um modelo é. Para avaliar a *performance* do modelo foi selecionada uma amostra de validação. Além disso, existem outros critérios que serão utilizados, apresentados nos tópicos seguintes.

3.6.1 Indicador de Acertos

Mede-se a taxa de acerto por meio da divisão do total de clientes classificados corretamente, pela quantidade de clientes que fizeram parte do modelo.

De forma similar, pode-se quantificar a taxa de acertos dos bons e maus clientes.

$$Tab = \frac{Ab}{Nb}$$

Onde:
Tab =Taxa de acertos de clientes bons
Ab = Indivíduos bons corretamente classificados
Nb = Número total de clientes bons

$$Tam = \frac{Am}{Nm}$$

Onde:
Tam = Taxa de acertos de clientes maus
Am = Indivíduos maus corretamente classificados
Nm = Número total de clientes maus

Neste estudo, como não se têm informações *a priori* sobre o que seria mais atrativo para a instituição financeira (identificação de bons ou maus clientes), utilizou-se o produto entre as taxas de acerto de bons e maus clientes como um indicador de acerto para se avaliar a qualidade do modelo. Esse indicador privilegia os modelos que tenham altos índices de acerto para os dois tipos de clientes. Quanto maior for o indicador, melhor será o modelo.

$$Ia = Tab * Tam$$

Onde:
Ia = Indicador de acertos
Tab = Taxa de acertos de clientes bons
Tam =Taxa de acertos de clientes maus

3.6.2 Teste de Kolmogorov-Smirnov

Outro critério bastante utilizado na prática (OOGHE; CAMERLYNCH; BALCAEN, 2003; PICININI; OLIVEIRA; MONTEIRO, 2003; PEREIRA, 2004) é o teste de Kolmogorov-Smirnov (KS). Trata-se de uma técnica não paramétrica para determinar se duas amostras foram extraídas da mesma população (SIEGEL, 1975).

Este teste se baseia na distribuição acumulada dos escores dos clientes considerados como bons e maus. Ambas as amostras são divididas em intervalos iguais dos escores e para cada uma é determinada a frequência acumulada. Em cada intervalo calcula-se a diferença entre as frequências acumuladas e o teste se dá focando a maior diferença entre elas. Para se verificar se as amostras possuem a mesma distribuição, existem tabelas que são consultadas de acordo com o nível de significância e tamanho da amostra (SIEGEL, 1975). Neste estudo, como as amostras são grandes, a tendência é que todos os modelos rejeitem a hipótese de igualdade nas distribuições. Será considerado melhor modelo aquele que possuir o maior valor no teste, pois este resultado indica uma separação maior entre bons e maus.

4 ANÁLISE DOS RESULTADOS

4.1 Regressão logística

A técnica de regressão logística foi empregada para o alcance do objetivo de determinar se diferenças nas características sócio-demográficas dos clientes do banco em questão podem distinguir entre os bons e os maus pagadores de empréstimos bancários. Para a estimação do modelo de regressão logística utilizou-se a amostra de 8000 casos divididos eqüitativamente nas categorias de bons e maus clientes. A Tabela 1 apresenta as variáveis selecionadas e as estatísticas geradas pelo modelo logístico.

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		

Variáveis com coeficiente logístico estimado negativo indicam que a categoria focalizada, em relação à referência, está associada com 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 outro já experiente, tem menor probabilidade de se comportar como bom solicitante de apoio financeiro. Pelo coeficiente de correlação parcial (R), as variáveis que mais afetam positivamente a probabilidade de se ter um bom cliente são primeira faixa de número de parcelas, segunda faixa de número de parcelas e primeira faixa de valor do empréstimo. Por outro lado, as variáveis que mais afetam negativamente a probabilidade de se ter um bom cliente são o tipo de crédito ser carnê, ser a primeira aquisição de empréstimo e a segunda faixa de idade.

Há dois testes estatísticos para se avaliar a significância do modelo final: teste Qui-quadrado da mudança no valor de –2LL e o teste de Hosmer e Lemeshow. O primeiro testa a hipótese estatística de que os coeficientes para todos os termos no modelo final, exceto a constante, são iguais a zero. Os valores inicial e final da estatística -2LL foram, respectivamente, 11090,355 e 5264,686. A diferença denominada *improvement* de 5825,669, conforme o teste Qui-quadrado, apresentou nível descritivo de 0,0001. Portanto, no modelo de 28 variáveis, constatou-se que a redução na medida –2LL foi estatisticamente significante.

Um teste alternativo para avaliar a significância do modelo logístico final é o pseudo R^2 definido como um quociente, no qual o numerador é a diferença entre os valores inicial e final da estatística -2LL e o denominador é o seu valor inicial. Neste estudo o valor do pesudo R^2 foi 0,525, considerado significante ao nível de 0,05.

Outro coeficiente indicativo da precisão de ajuste do modelo aos dados é o Nagelkerke, medida comparável ao R^2 da regressão linear múltipla. Neste trabalho obteve-se o valor 0,711, também estatisticamente significante ao nível de 0,05.

O teste de Hosmer e Lemeshow considera a hipótese estatística de que as classificações em grupo previstas são iguais às observadas. Portanto, 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. Este resultado conduz à não rejeição da hipótese nula do teste, endossando a aderência do modelo aos dados.

Assim, o modelo de regressão logística apresentou resultados estatisticamente significantes, os quais reforçam a sua adequação.

4.2 Algoritmos genéticos

Na literatura consultada, há duas maneiras de avaliar o cliente por meio de algoritmos genéticos. A primeira, adotada por Chen, Huang e Chen (2002) e Fidelis, Lopes e Freitas (2000), soluciona o problema por meio de uma seqüência de regras tal qual uma árvore de decisão, ou seja, uma série de regras encadeadas que determinam se o cliente é bom ou mau, dependendo do caminho (ou galho da árvore) percorrido.

Na segunda forma, que será adotada neste trabalho, o algoritmo genético foi utilizado para encontrar uma equação discriminante que permita pontuar os clientes e, depois, separar os bons e maus clientes de acordo com o escore obtido. A equação pontua os clientes e os de maior pontuação são considerados bons. Esse caminho foi adotado por Kishore et al. (2000) e Picinini, Oliveira e Monteiro (2003).

O algoritmo foi executado três vezes conforme a configuração apontada na seção anterior, de forma que as soluções obtidas alcançassem estabilidade, ou seja, não houvesse melhora significativa na taxa de acertos. Aqui serão apresentados os resultados do algoritmo que obteve o maior Indicador de acertos (Ia).

As variáveis com peso muito pequeno foram descartadas. No trabalho de Picinini, Oliveira e Monteiro (2003) os autores consideraram que as variáveis com peso inferior a 0,15 ou superior a -0,15 seriam descartadas por possuírem um peso não significativo para o

modelo. Neste trabalho, depois de feita uma análise de sensibilidade, decidiu-se considerar como significativas para o modelo as variáveis com peso superior a 0,10 ou inferior a -0,10. Essa regra não foi aplicada para a constante, que se mostrou importante para o modelo mesmo com o valor abaixo do ponto de corte.

O peso das variáveis está na Tabela 2. Os pesos negativo e positivo indicam maior relação da variável com os clientes considerados, respectivamente, maus e bons. A variável com maior peso negativo foi tipo de crédito: carnê e com maior peso positivo foi primeira faixa de número de parcelas.

Tabela 2 – Pesos finais das variáveis

Pesos negativos		**Pesos positivos**	
Variáveis	**Peso**	**Variáveis**	**Peso**
Tipo de crédito: carnê	-0,99	Primeira faixa de número de parcelas	1,42
Sem telefone comercial	-0,98	Sexo feminino	0,97
Segunda faixa de idade	-0,98	Sétima categoria de profissão	0,95
Primeira faixa de salário	-0,95	Terceira faixa de relação entre empréstimo e salário	0,95
Segunda categoria de profissão	-0,91	Estado civil casado	0,93
Quarta faixa de número de parcelas	-0,88	Quarta faixa de idade	0,89
Quarta faixa de salário	-0,87	Segunda faixa de número de parcelas	0,88
Terceira faixa de relação entre parcela e salário	-0,80	Quinta faixa de relação entre parcela e salário	0,88
Segunda categoria de CEP residencial	-0,79	Primeira faixa de valor do empréstimo	0,83
Quinta faixa de valor do empréstimo	-0,76	Tipo de crédito: cheque	0,81
Terceira faixa de número de parcelas	-0,65	Quarta faixa de tempo de residência fixa	0,75
Terceira faixa de salário	-0,61	Segunda faixa de valor do empréstimo	0,59
Quarta faixa de valor do empréstimo	-0,59	Estado civil outros	0,58
Segunda categoria de CEP comercial	-0,59	Quinta categoria de CEP residencial	0,57
Quarta categoria de profissão	-0,56	Quarta faixa de tempo de emprego	0,56
Primeira faixa de tempo de emprego	-0,55	Primeira faixa de relação entre parcela e salário	0,55
Com telefone residencial	-0,54	Sexta faixa de salário	0,47
Primeira faixa de idade	-0,54	Terceira categoria de profissão	0,45
Terceira categoria de CEP comercial	-0,50	Quarta faixa de valor da parcela	0,41
Segunda faixa de tempo de emprego	-0,45	Terceira faixa de tempo de emprego	0,39
Segunda faixa de relação entre parcela e salário	-0,45	Terceira faixa de tempo de residência fixa	0,39
Quarta categoria de CEP residencial	-0,44	Segunda faixa de valor da parcela	0,34
Primeira faixa de relação entre empréstimo	-0,42	Nona categoria de profissão	0,33
Quarta faixa de relação entre empréstimo e salário	-0,39	Sexo masculino	0,29
Sexta faixa de valor do empréstimo	-0,28	Terceira faixa de valor do empréstimo	0,25
Terceira categoria de CEP residencial	-0,28	Não é primeiro empréstimo	0,24
Primeira aquisição de empréstimo	-0,28	Primeira faixa de tempo de residência fixa	0,19
Primeira categoria de CEP residencial	-0,23	Estado civil solteiro	0,14
Primeira categoria de CEP comercial	-0,22	Quarta categoria de CEP comercial	0,13
Quinta categoria de CEP comercial	-0,21	Primeira categoria de profissão	0,13
Segunda faixa de tempo de residência fixa	-0,14		
Segunda faixa de salário	-0,12		
Oitava categoria de profissão	-0,12		

Há uma concordância parcial de resultados das duas técnicas, comparando-se os coeficientes das Tabelas 1 e 2. As variáveis que ocupam o primeiro posto do *ranking* são as mesmas nos dois modelos - maior coeficiente positivo: primeira faixa de número de parcelas; maior coeficiente negativo: tipo de crédito carnê. Para os demais postos algumas diferenças são percebidas, até porque nem todas as variáveis foram mantidas na regressão logística pelo método *forward stepwise*. Exemplo de algumas semelhanças e diferenças:

a) segunda faixa de número de parcelas – semelhança (impacto positivo nos dois modelos); diferença (segundo lugar na logística e sétimo lugar nos algoritmos

genéticos);

b) primeira faixa de valor do empréstimo – semelhança (impacto positivo nos dois modelos); diferença (terceiro lugar na logística e nono lugar nos algoritmos genéticos);
c) primeira aquisição de empréstimo – semelhança (impacto negativo nos dois modelos); diferença (segundo lugar na logística e vigésimo sétimo lugar nos algoritmos genéticos);
d) segunda faixa de idade – semelhança (impacto negativo nos dois modelos); diferença (terceiro lugar nos dois modelos).

4.3 Avaliação da *Performance* dos Modelos

Após obtidos os modelos, foram escoradas as duas amostras e calculados o Ia e o KS para cada um dos modelos. Os resultados são apresentados na Tabela 3.

Tabela 3 - Resultados de classificação

Regressão logística							
Construção					**Validação**		
		Predito →			Predito →		
		Mau	Bom	% Acerto	Mau	Bom	% Acerto
Observado ↓	Mau	2913	1087	72,8	2169	831	72,3
	Bom	1184	2816	70,4	999	2001	66,7
	Total	4097	3903	71,6	3168	2832	69,5
Algoritmo genético							
Construção					**Validação**		
		Predito →			Predito →		
		Mau	Bom	% Acerto	Mau	Bom	% Acerto
Observado ↓	Mau	2775	1225	69,4	2073	927	69,1
	Bom	1264	2736	68,4	1063	1937	64,6
	Total	4039	3961	68,9	3136	2864	66,8

A Tabela 3 mostra os resultados de classificação obtidos pelos dois modelos. Ambos apresentaram bons resultados de classificação. Segundo Picinini, Oliveira e Monteiro (2003, p. 465), "modelos de *credit scoring* com taxas de acerto próximas ou acima de 65% são considerados bons por especialistas".

Os percentuais de acerto foram um pouco inferiores para o modelo de algoritmos genéticos. Outro resultado interessante é que os dois modelos apresentaram uma taxa de acertos maior para os clientes tidos como maus. Não há um consenso sobre o que representa melhor resultado para uma instituição bancária: maior taxa de acertos do modelo estatístico na classificação do mau cliente ou do bom cliente. A detecção correta do mau cliente protege a instituição de transtornos decorrentes do não cumprimento pelo cliente do pré-estabelecido no contrato de empréstimo.

Os resultados obtidos pelos modelos logístico e algoritmo genético representam menor risco de a instituição enfrentar esse tipo de problema. Por outro lado, a menor sensibilidade de ambos os modelos na identificação de um bom cliente pode incorrer na perda definitiva deste cliente que, ao ser rejeitado no seu pedido de empréstimo, provavelmente procurará uma instituição concorrente que o acolha. Se os modelos aplicados neste trabalho reproduzissem o contrário, com maior acerto de clientes bons, haveria o inconveniente de se errar mais na detecção do mau cliente, concedendo-se indevidamente o crédito a um cliente que poderia criar problemas futuros por ocasião da quitação da dívida.

A Tabela 4 apresenta os resultados dos critérios Ia e KS que foram os escolhidos para

comparar os modelos.

Tabela 4 - Índices de Comparação

Ia	**Amostra**	
	Construção	**Validação**
Regressão logística	51,3	48,2
Algoritmo genético	47,5	44,6
KS	**Amostra**	
	Construção	**Validação**
Regressão logística	38	37
Algoritmo genético	34	32

Os valores do critério Ia na Tabela 4 foram obtidos pelo produto das taxas de acertos na classificação dos clientes bons e dos clientes maus, apresentadas na Tabela 3. O modelo de regressão logística apresentou melhor desempenho para estes indicadores nas duas amostras consideradas, revelando-se a melhor modelagem segundo este critério.

Os valores KS dos dois modelos podem ser considerados bons. Novamente, Picinini, Oliveira e Monteiro (2003, p. 465) explicam que "o teste de Kolmogorov-Smirnov (KS) é utilizado no mercado financeiro na comparação de modelos de *credit scoring*, sendo que o mercado considera um bom modelo aquele que apresente um valor de KS superior a 30". Aqui novamente o modelo de regressão logística apresentou um resultado superior ao obtido pelo algoritmo genético.

Na escolha do modelo mais adequado para estes dados, analisando sob o prisma dos indicadores Ia e KS, foi eleito o modelo construído por regressão logística, pois este modelo apresentou melhores resultados na amostra de validação, sugerindo ser o mais adequado para a aplicação em outras bases de dados. Contudo, deve ser ressaltado, mais uma vez, que a adoção de qualquer um dos modelos traria bons resultados à instituição financeira.

5 CONCLUSÕES

Neste artigo buscou-se apresentar a aplicação das técnicas regressão logística e algoritmos genéticos no contexto de *credit scoring* e compará-las em termos de qualidade de ajuste aos dados. Como todo mecanismo matemático, ambos os modelos apresentaram taxas de acertos e de erros na classificação dos clientes como bons e maus pagadores de empréstimo bancário. Na avaliação das taxas de erros, a instituição deve considerar os casos em que erroneamente bons pagadores foram classificados como maus pagadores e vice-versa.

As políticas de crédito devem ser analisadas à luz da orientação estratégica da empresa. Se a prioridade for aumentar a participação no mercado, a instituição pode decidir pela concessão de empréstimo mesmo a clientes classificados como maus pagadores pelo modelo. Por outro lado, se a participação de mercado for considerada conveniente, a empresa pode optar por minimizar perdas com inadimplência e decidir pela não concessão de crédito até mesmo para os clientes com previsão de serem bons pagadores, mas que estejam com pequena probabilidade estimada para essa situação favorável a eles.

Apesar dos erros detectados, os modelos de *credit scoring* auxiliam as empresas na decisão de concessão de crédito e agilizam todo o processo. No desenvolvimento de modelos de avaliação de crédito alguns cuidados devem ser tomados a fim de garantir a qualidade do modelo, e a aplicabilidade posterior. Precauções na amostragem, definição clara nos critérios na classificação de clientes bons e maus e tratamento das variáveis da base de dados antes da aplicação das técnicas foram cuidados tomados neste estudo, visando otimizar resultados.

Os dois modelos gerados pelas técnicas de regressão logística e algoritmos genéticos apresentaram resultados satisfatórios para a base de dados em questão, que foi fornecida por um grande banco de varejo que atua no Brasil. O modelo de regressão logística apresentou

resultados superiores ao modelo baseado em algoritmos genéticos. O percentual de acerto total para a amostra de validação foi para a regressão logística e algoritmos genéticos, respectivamente, igual a 69,5% e 66,8%. O modelo proposto por este estudo para que a instituição pontue seus clientes é:

$$p = \frac{e^z}{1+e^z} \quad \text{, onde}$$

p = probabilidade de o cliente ser considerado bom
$z = B_0 + B_1.X_1 + B_2.X_2 + ...+ B_{28}.X_{28}$.

As 28 variáveis são dadas na Tabela 1.

Não foi objeto deste estudo uma abordagem mais profunda das técnicas focalizadas, mas sim ilustrar e difundir suas aplicações e utilidade na área de risco de crédito. A técnica de algoritmos genéticos é bastante flexível e ainda não tanto pesquisada em problemas de concessão de crédito. Conforme exposto no Quadro 1, boa parte dos estudos na área utiliza a técnica da Regressão Logística, tida como menos sofisticada que os algoritmos genéticos. O trabalho demonstrou que seus resultados, mesmo sendo uma técnica mais avançada, foram ligeiramente inferiores àqueles obtidos pela técnica mais popular (regressão logística). Isto indica que modelos menos complexos podem atender satisfatoriamente a necessidade das instituições financeiras para classificação dos clientes.

Todavia, deve-se considerar que os algoritmos genéticos podem ser aplicados de formas diversas a fim de otimizar o resultado obtido, alterando-se as medidas de seleção e os parâmetros da técnica, assim como ocorre em quaisquer técnicas estatísticas (a alteração dos parâmetros altera os resultados obtidos). Assim, sugere-se para estudos futuros a estimação de modelos distintos baseados em diferentes parâmetros dos algorítmos genéticos, para se avaliar a qualidade dos resultados de cada um deles. Técnicas novas neste tipo de problema, como análise de sobrevivência e o *data mining*, também merecem atenção em estudos futuros.

REFERÊNCIAS

ABDOU, H. A. Genetic programming for credit scoring: the case of Egyptian public sector banks. **Expert Systems with Applications**, v. 36, n. 9, p. 11402-11417, nov. 2009. http://dx.doi.org/10.1016/j.eswa.2009.01.076

ANDREEVA, G. European generic scoring models using logistic regression and survival analysis. In: YOUNG OR CONFERENCE, 2003, Bath. **Anais…** Bath: Young OR, 2003.

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.

BACK, B.; LAITINEN, T.; SERE, K. Neural networks and genetic algorithms for bankruptcy predictions. In: WORLD CONFERENCE ON EXPERT SYSTEMS, 3., Seul. **Anais…** Seul: Science Direct, 1996. http://dx.doi.org/10.1016/S0957-4174(96)00055-3

BIJAK, K.; THOMAS, L. C. Does segmentation always improve model performance in credit scoring? **Expert Systems with Applications**, v. 39, n. 3, p. 2433-2442, fev. 2012. http://dx.doi.org/10.1016/j.eswa.2011.08.093

BRITO, G. A. S.; ASSAF NETO, A. Modelo de classificação de risco de crédito de empresas. **Revista Contabilidade & Finanças**, v. 19, n. 46, p. 18-29, abr. 2008. http://dx.doi.org/10.1590/S1519-70772008000100003

BROWN, I.; MUES, C. An experimental comparison of classification algorithms for

imbalanced credit scoring data sets. **Expert Systems with Applications**, v. 39, n. 3, p. 3446-3453, fev. 2012. http://dx.doi.org/10.1016/j.eswa.2011.09.033

CHEN, M. C.; HUANG, S. H; CHEN, C. M. Credit classification analysis through the genetic programming approach. In: INTERNATIONAL CONFERENCE IN INFORMATION MANAGEMENT, 2002, Taipei. **Anais…** Taipei: Tamkang University, 2002.

CHEN, M. C.; HUANG, S. H. Credit scoring and rejected instances reassigning through evolutionary computation techniques. **Expert Systems with Applications**, St. Louis, v. 24, n. 4, p. 433-441, 2003. http://dx.doi.org/10.1016/S0957-4174(02)00191-4

FIDELIS, M. V.; LOPES, H. S.; FREITAS, A. A. Discovering comprehensible classification rules with a genetic algorithm. In: CONGRESS ON EVOLUTIONARY COMPUTATION, 2000, La Jolla. **Anais...** La Jolla: IEEE, 2000. http://dx.doi.org/10.1109/CEC.2000.870381

HAIR JR., J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **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. http://dx.doi.org/10.1111/j.1467-985X.1997.00078.x

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. http://dx.doi.org/10.1057/palgrave.fsm.4770054

KISHORE, J. K.; PATNAIK, L. M.; MANI, V.; AGRAWAL, V. K. Application of genetic programming for multicategory pattern classification. **IEEE Transactions on Evolutionary Computation**, Birmingham, v. 4, n. 3, p. 242-257, 2000. http://dx.doi.org/10.1109/4235.873235

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

LIMA, F. G.; PERERA, L. C. J.; KIMURA, H.; SILVA FILHO, A. C. 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.

MAVRI, M.; ANGELIS, V.; IOANNOU, G.; GAKI, E.; KOUFODONTIS, I. A two-stage dynamic credit scoring model, based on customers' profile and time horizon. **Journal of Financial Services Marketing,** v. 13, n. 1, p. 17-27, maio, 2008. http://dx.doi.org/10.1057/fsm.2008.2

OHTOSHI, C. **Uma comparação de regressão logística, árvores de classificação e redes neurais:** analisando dados de crédito. 2003. Dissertação (Mestrado em Estatística) – Programa 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**, v. 46, p. 39-70, 2003.

PAPPA, G. L. **Seleção de atributos utilizando algoritmos genéticos multiobjetivos**. 2002. Dissertação (Mestrado em Informática) – Programa de Pós-graduação em Informática, Pontifícia Universidade do Paraná, Curitiba, 2002.

PEREIRA, G. H. A. **Modelos de risco de crédito de clientes:** uma aplicação a dados reais. 2004. Dissertação (Mestrado em Estatística) – Programa 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...** Bauru: Universidade de Brasília, 2003.

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

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

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

SILVA, S. F. **Seleção de características por meio de algoritmos genéticos para aprimoramento de rankings e de modelos de classificação**. 2011. 115f. Tese (Doutorado em Estatística) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2011.

ŠUŠTERŠIČ, M.; MRAMOR, D.; ZUPAN, J. Consumer credit scoring models with limited data. **Expert Systems with Applications**, v. 36, n. 3, p. 4736-4744, abr. 2009. http://dx.doi.org/10.1016/j.eswa.2008.06.016

TREVISANI, A. T., GONÇALVES, E. B., D'EMÍDIO, M.; HUMES, L. L. Qualidade de dados: desafio crítico para o sucesso do business intelligence. In: CONGRESSO LATINO AMERICANO DE ESTRATÉGIA, 18., 2004, Itajaí. **Anais...** Itajaí: Sociedade Latinoamericana de Estratégia, 2004.

ULISES, C.; CARMONA, D. M. Application of credit scoring models in the analysis of insolvency of a Brazilian microcredit institution. **Journal of Modern Accounting and Auditing**, v. 7, n. 8, p. 799-812, 2011.

YAP, B. W.; ONG, S. H.; HUSAIN, N. H. M. Using data mining to improve assessment of credit worthiness via credit scoring models. **Expert Systems with Applications**, v. 38, n. 10, p. 13274-13283, set 2011. http://dx.doi.org/10.1016/j.eswa.2011.04.147