# ANÁLISE DE RISCO DE CRÉDITO COM APLICAÇÃO DE REGRESSÃO LOGÍSTICA E REDES NEURAIS

Maria Aparecida Gouvêa[1]
Eric Bacconi Gonçalves[2]
Daielly Melina Nassif Mantovani[3]

▪ Artigo recebido em: 24/01/2013 ▪▪ Artigo aceito em: 08/07/2014 ▪▪▪ Segunda versão aceita em: 03/11/2014

## RESUMO

O objetivo deste estudo foi aplicar e comparar as técnicas regressão logística e redes neurais no desenvolvimento de modelos de predição de *credit scoring* com base em dados de uma grande instituição financeira brasileira. A questão-problema deste estudo é relevante, pois contribui para o aprimoramento das previsões e fornece apoio para as instituições financeiras tomarem decisões mais precisas sobre concessões de crédito.A base de dados correspondeu ao período de agosto de 2009 a fevereiro de 2010,período em que o Brasil vivenciou notável expansão da oferta de crédito no mercado.A partir de uma amostra de 20.000 dados, foram aplicadas as duas técnicas. A amostra foi dividida em três sub-amostras provenientes do mesmo universo de interesse: uma para construção do modelo (8.000 dados); a segunda para validação do modelo construído (6.000 dados) e a terceira também com 6.000 dados para testar o modelo obtido. Nas 3 sub-amostras houve uma distribuição equitativa de bons e maus clientes, classificados nestas categorias de acordo com padrões da instituição.Os dois modelos testados apresentaram estatísticas de desempenho satisfatórias e poderão ser empregados pela instituição bancária interessada na identificação de bons e maus pagadores de empréstimos.O processo de tomada de decisões de concessão de crédito bancário poderá ser agilizado com o apoio dos modelos analisados neste trabalho.

**Palavras-chave**: crédito, regressão logística, redes neurais

[1] Doutora em Administração pela Universidade de São Paulo (USP). Professora Livre Docente do Departamento de Administração da Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo (FEA/USP). Endereço: Endereço: Avenida Professor Luciano Gualberto 908, Sala G162, FEA/USP. CEP: 05508-010. São Paulo – SP. E-mail: magouvea@usp.br. Telefone: (11) 3091 6044.

[2] Mestre em Administração pela Universidade de São Paulo (FEA/USP). Endereço: Avenida Professor Luciano Gualberto 908, Sala G162, FEA/USP. CEP: 05508-010. São Paulo – SP. E-mail: eric.goncalves@telefonica.com.br. Telefone: (11) 3091 6044.

[3] Doutora em Administração pela Universidade de São Paulo (USP). Endereço: Avenida Professor Luciano Gualberto 908, Sala G162, FEA/USP. CEP: 05508-010. São Paulo – SP. E-mail: daimantovani@gmail.com. Telefone: (11) 97324 8820.

## CREDIT RISK ANALYSIS THROUGH LOGISTIC REGRESSION AND NEURAL NETWORKS

**ABSTRACT**

The objective of this study was to apply and compare the techniques logistic regression and neural networks in the development of models for predicting credit scoring based on data from a large Brazilian financial institution. The question-problem of this study is important, because it contributes to the improvement of forecasts and provides support for financial institutions in taking more precise decisions about credit concessions.The database corresponded to the period of August 2009 to February 2010, in which Brazil experienced a great expansion of credit offer to the market. From a sample of 20,000 data, the two techniques have been applied. The sample was divided into three smaller samples from the same universe of interest: one sample for model construction (8,000 data); the second one for model validation (6,000 data) and the third one also with 6,000 data to test the model obtained. In the three samples there was an equitable distribution of good and bad clients, classified into these categories according to the standards of the institution. The two modelstested showed satisfactory performance statistics and may be employed by banking institution interested in identifying good and bad payers of loans.The decision-making process in bank loan requests can be sped up with the support of the models analyzed in this work.

**Keywords**: credit, logistic regression, neural networks

# 1 INTRODUÇÃO

Os modelos de análise para concessão de crédito conhecidos como modelos de *credit scoring* baseiam-se em dados históricos da base de clientes existentes para avaliar se um futuro cliente terá mais chances de ser bom ou mau pagador. Os modelos de *credit scoring* são implantados nos sistemas das instituições, permitindo que a avaliação de crédito seja *on-line*.

Modelos que avaliam o crédito são de vital importância para o negócio de uma instituição financeira. Um cliente mal classificado pode causar prejuízos (no caso de classificar um cliente mau como bom) ou então privar a instituição de ganhos (no caso de classificar um cliente bom como mau).

Nenhum modelo consegue precisão absoluta, ou seja, acertar totalmente suas previsões. Qualquer avanço em termos de acurácia da previsão pode gerar ganhos financeiros para a instituição. Daí vem o interesse de se analisarem diferentes tipos de modelos e apontar quais apresentam maior precisão.

Dada sua aplicabilidade na previsão de variáveis, a regressão logística e as redes neurais têm sido empregadas em finanças e controle. Por esta razão, julgou-se oportuno apresentar as duas técnicas para utilização em um mesmo banco de dadospertencente a uma instituição financeira e comparar os seus resultados, no período de agosto de 2009 a fevereiro de 2010.Este período é relevante; pois o país passou por uma grande expansão da oferta de crédito.

No período de 1995 a 2003 houve expansão de 1,6% do volume de crédito; por sua vez, de 2004 a 2012, houve aumento de 209,9%, com ênfase para o crédito de pessoa física (30% do total), o que eleva o endividamento das famílias e o risco de crédito (SBICCA; FLORIANI; JUK, 2012). Um importante diferencial do presente estudo comparativamente aos disponíveis na literatura pesquisada é a possibilidade de extração de três amostras de grande magnitude, nas quais os dados foram processados em três momentos de análise: treinamento, validação e teste. Adicionalmente, os resultados obtidos por meio do uso das técnicas Regressão Logística e Redes Neurais corroboram outros estudos da literatura, o que eleva a segurança do uso desses modelos para análise de risco de crédito.

O objetivo deste estudo é a apresentação do uso de regressão logística e redes neurais para a classificação de bons e maus pagadores em financiamentos bancários, considerando-se o produto crédito pessoal. Complementarmente, pretende-se identificar a técnica com melhor aderência aos dados pesquisados provenientes de uma grande instituição financeira.

## 2 FUNDAMENTAÇÃO TEÓRICA

Nesse capítulo serão apresentados conceitos teóricos que darão sustentação ao desenvolvimento do tema deste trabalho.

### 2.1. Crédito ao Consumidor

A expressão crédito ao consumidor pode ser entendida como uma forma de comércio onde uma pessoa física obtém dinheiro, bens ou serviços e compromete-se a pagar por isso futuramente, acrescendo ao valor original um prêmio (juros) (SANTOS, 2000).

Atualmente, o crédito ao consumidor é uma grande indústria que opera no mundo. Empresas automobilísticas, bancos e outros segmentos utilizam as linhas de crédito ao consumidor como uma alternativa a mais para obter lucros.

Entretanto, tornar o crédito largamente disponível não significa distribuir crédito indistintamente para todos que o solicitam; existe um fator associado ao crédito ao consumidor que é fundamental na decisão de disponibilizar ou não o crédito: o risco.

### 2.2. Risco de Crédito

A concessão de crédito é atividade básica das instituições financeiras; entretanto, no desenvolver deste negócio, os bancos estão expostos a diversos tipos de riscos, entre eles o mais relevante é o risco de crédito (FERREIRA, *et al.*, 2012 p.42).Antes de qualquer sofisticação, resultante da engenharia financeira, o puro ato de emprestar uma quantia a alguém traz embutida em si a probabilidade de ela não ser recebida, a incerteza em relação ao retorno. Isto é, na essência, o risco de crédito, e que se pode definir como: o risco de uma contraparte, em um acordo de concessão de crédito, não honrar seu compromisso.

Segundo Caouette *et al.* (2000, p. 1), "se crédito pode ser definido como a expectativa de recebimento de uma soma em dinheiro em um prazo

determinado, então Risco de Crédito é a chance que esta expectativa não se concretize".

A atividade de concessão de crédito é função básica dos bancos; portanto, o risco de crédito toma papel relevante na composição dos riscos de uma instituição e pode ser encontrado tanto em operações onde existe liberação de dinheiro para os clientes como naquelas onde há apenas a possibilidade do uso, os limites pré-concedidos. Os principais tipos de operações de crédito de um banco são: empréstimos, financiamentos, descontos de títulos, adiantamento a depositantes, adiantamento de câmbio, operações de arrendamento mercantil *(leasing*), avais e fianças etc.Nessas operações, o risco pode se apresentar sob diversas formas; conhecê-las conceitualmente ajuda a direcionar o gerenciamento e a mitigação.Louzis *et al.* (2012) afirmam que a análise de risco de crédito é fundamental para a administração de uma instituição financeira.

No universo do crédito ao consumidor, a promessa de pagamento futuro envolve a ideia de risco. Como o futuro não pode ser corretamente predito, todo crédito ao consumidor envolve risco, pois nunca existe a certeza do pagamento (LEWIS, 1992). Cabe à análise de crédito estimar o risco envolvido para a concessão ou não do crédito.O risco máximo que a instituição pode aceitar depende da política adotada pela empresa.

## 2.3. Avaliação do 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 (SWIDERSKI, *et al.,* 2012), quer seja pelo aceite de clientes que irão gerar prejuízos ao negócio, quer seja pela recusa de clientes bons que gerariam lucros ao negócio.

Segundo Soares *et al.* (2013): avaliação de risco de crédito é um processo de quantificar a possibilidade de que os fluxos de caixa esperados com as operações de crédito não se confirmem. Essa avaliação é feita pelas expectativas de desempenho do proponente.

A avaliação do risco de um potencial cliente pode ser feita de duas maneiras: por meio de julgamento, uma forma mais subjetiva que envolve uma análise mais qualitativa; ou por meio da classificação do tomador via modelos de avaliação, envolvendo uma análise mais quantitativa.

Desde meados dos anos 2000, praticamente todas as grandes empresas que trabalham com concessão de crédito utilizam as duas formas combinadas. Na avaliação do risco de crédito por meio de classificação do tomador é que são utilizados os modelos chamados *credit scoring,* que permitem uma mensuração do risco do tomador de crédito, auxiliando na tomada de decisão de concessão ou não do crédito (CAMARGOS *et al.*, 2012).

## 2.4. Acordos da Basileia e Risco de Crédito

Em 1988, foi firmado um acordo chamado de Basileia I com o intuito de assegurar estabilidade financeira no sistema bancário dos países desenvolvidos (PEREIRA, 2006); esse acordo buscava o fortalecimento dos bancos participantes bem como de seus depositantes. Diante de novas pressões do mercado, o acordo inicial foi ampliado de forma a incluir mais países e trazer controles mais

rigorosos; vieram posteriormente, em 2004, o acordo Basileia II e, em 2010, o acordo Basileia III.

Yanaka e Holland (2010) apontam que de acordo com as normas vigentes os bancos devem constituir uma provisão específica de acordo com o nível de risco. Os autores ainda afirmam que no gerenciamento do risco de crédito o banco deve considerar a correlação entre eventos de inadimplência.

Segundo Bonfim (2009), o acordo da Basileia II promoveu um crescimento na utilização de modelos para avaliação de crédito, visto que este acordo propôs o uso deste tipo de modelo para avaliar as necessidades de capital do banco. O acordo permitiu aos bancos estimar o risco de crédito de duas formas: pela utilização de *ratings* externos de agências e pela construção dos próprios modelos de avaliação de crédito de acordo com informações internas (BASTOS, 2010).

### 2.5. Modelos de *credit scoring*

De acordo com Mileris (2012), a ideia de um modelo de *credit scoring* é condensar várias informações quantitativas e qualitativas dos proponentes em uma pontuação que reflita a capacidade de pagamento de cada indivíduo.

Nos anos cinquenta, os modelos de crédito foram difundidos na indústria bancária americana. Os primeiros modelos baseavam-se em pesos pré-estabelecidos para certas características determinadas, somando-se os pontos e obtendo-se um escore de classificação.

O crescimento do uso de modelos na década de 60 transformou os negócios no mercado americano (THOMAS, 2000).

Não somente empresas do segmento financeiro, mas também grandes varejistas começaram a fazer uso de modelos de *credit scoring* para efetuar vendas a crédito para seus consumidores.

No Brasil, a história é mais curta. As instituições financeiras passaram a utilizar maciçamente os modelos de *credit scoring* apenas em meados dos anos 90, pós estabilidade alcançada com a implantação do plano Real (CAMARGOS *et al.*, 2012).

Há sete passos a serem seguidos para se construir um modelo de *creditscoring*, a saber:

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 são fundamentais para o sucesso do modelo (TREVISANI *et al.*, 2004).

2 . 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 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, trabalha 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; ORESKY, *et al.,* 2012; FERREIRA *et al.,*2012; 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) 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, segundo critérios estatísticos, cada variável a ser utilizada no modelo.

5. Escolha e aplicação das técnicas a serem utilizadas para a construção do modelo: neste estudo serão utilizadas Regressão Logística e Redes Neurais; Hand e Henley (1997) destacam ainda Análise Discriminante, Regressão Linear, e Árvores de Decisão, como métodos utilizados na prática; alguns estudiosos também têm usado Análise de Sobrevivência (HARRISON; ANSELL, 2002; ANDREEVA, 2003); estudos anteriores evidenciam que não há um método sempre melhor que os demais, tudo dependendo da estrutura dos dados, variáveis utilizadas, como a técnica se ajusta aos dados (SADATRASOUL, 2013).

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.

## 2.6. Regressão logística

Regressão Logística é a técnica mais utilizada no mercado para o desenvolvimento de modelos de *credit scoring*(CROOK *et al.,* 2007). 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).

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.

Considere o caso em que as observações podem ser classificadas em apenas uma de duas categorias. Por exemplo, 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 e 0 se pertence à categoria dos maus (TSAI, 2012).

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:

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

$\beta' = (\beta_0, \beta_1, \beta_2, ..., \beta_n)$: 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):

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

A variável dependente binária neste estudo assumiu os valores 1 para bons clientes e 0 para maus clientes. Os resultados da técnica de regressão logística binária são os mesmos independentemente da categoria que foi codificada como 1. A expressão matemática do modelo obtida no processamento desta técnica permitirá prontamente o cálculo da probabilidade de o cliente ser bom pagador do empréstimo bancário; para a obtenção da probabilidade de se ter um mau cliente, bastará calcular a probabilidade complementar, ou seja, se a probabilidade de determinado cliente ser bom for 0,6, a probabilidade de o mesmo cliente ser mau pagador será 0,4.

Neste trabalho, inicialmente, todas as variáveis serão incluídas para construção do modelo; porém, no modelo logístico final, apenas algumas variáveis serão selecionadas. A escolha das variáveis será feita por meio do método *forward stepwise*, que é o mais largamente utilizado em modelos de regressão logística. A habilidade do método *forward stepwise* de acrescentar e eliminar variáveis já incluídas no modelo logístico faz deste o procedimento preferido por parte da maioria dos pesquisadores (HAIR JR. *et al.*, 2009, p. 180). As variáveis são selecionadas a cada passo, de acordo com critérios que otimizem o modelo, evitando-se problemas de multicolinearidade. Somente as variáveis realmente importantes para o modelo são selecionadas. A importância de cada variável é aferida em termos da sua contribuição para diferenciar os dois grupos obtidos na partição da variável dependente, que, no caso do presente estudo, apresenta a alocação do cliente tomador de empréstimo bancário nas categorias: bom e mau cliente.Para cada variável a ser incluída

no modelo é testada a sua contribuição incremental na diferenciação dos dois grupos diante das variáveis já presentes na equação logística.Para detalhes do método sugere-se a leitura de Neter *et al.* (1996, p. 348).

## 2.7. Redes neurais artificiais

Redes Neurais Artificiais são técnicas computacionais baseadas na estrutura neural de organismos inteligentes e que adquirem conhecimento por intermédio de experiências(AKKOÇ, 2012, p.170).

Redes neurais artificiais são desenvolvidas por meio de modelos matemáticos, onde as seguintes suposições são feitas (FAUSETT, 1994, p. 3):

1. O processamento das informações ocorre dentro dos chamados neurônios;

2. Os estímulos são transmitidos pelos neurônios por meio de conexões;

3. Cada conexão tem associada a si um peso, que, numa rede neural padrão, multiplica-se ao estímulo recebido;

4. Cada neurônio contribui para a função de ativação (geralmente não linear) para determinar o estímulo de saída (resposta da rede).

O modelo pioneiro de McCulloch e Pitts de 1943 (FAUSETT, 1994), para uma unidade de processamento (neurônio), pode ser resumido em:

- Sinais são apresentados à entrada;
- Cada sinal é multiplicado por um peso que indica sua influência na saída da unidade;
- É feita a soma ponderada dos sinais que produz um nível de atividade.

Se este nível excede um limite, a unidade produz uma saída.

No esquema, têm-se p sinais de entrada $X_1, X_2, ..., X_p$ e pesos correspondentes $W_1, W_2, ..., W_p$ e seja k o limite.

Neste modelo o nível de atividade é dado por:

$$a = \sum_{i=1}^{p} W_i X_i$$

A saída y é dada por:

$$y = 1, \text{ se } a \geq k$$

$$y = 0, \text{ se } a < k$$

Na definição de um modelo de redes neurais três características devem ser observadas: a forma que a rede tem, chamada arquitetura; o método para determinação dos pesos, chamado algoritmo de aprendizado; e a função de ativação.

Arquitetura refere-se ao formato da rede. Toda rede é dividida em camadas, usualmente classificadas em três grupos:

- Camada de Entrada: onde os padrões são apresentados à rede;
- Camadas Intermediárias ou Ocultas: onde é feita a maior parte do processamento, por meio das conexões ponderadas;
- Camada de Saída: onde o resultado final é concluído e apresentado.

Existem três tipos principais de arquitetura (HAYKIN, 1999, p. 46-48): redes *feedforward* com uma única camada, redes *feedforward* com múltiplas camadas, e redes recorrentes.

Redes *feedforward* com uma única camada: são o caso mais simples de rede, existindo apenas uma camada de entrada e uma camada de saída. As redes são alimentadas adiante, ou seja, apenas a camada de entrada fornece informações para a camada de saída. Algumas das redes que utilizam essa arquitetura são: Rede de Hebb, *perceptron*, ADALINE, entre outras.

Redes *feedforward* com múltiplas camadas: são aquelas que possuem uma ou mais camadas intermediárias. A saída de cada camada é utilizada como entrada para a próxima camada. Da mesma forma que a arquitetura anterior, este tipo de rede caracteriza-se apenas por alimentação adiante. As redes *multilayer perceptron* (MLP), MADALINE e de função de base radial são algumas das redes que utilizam esta arquitetura.

Redes Recorrentes: neste tipo de rede, a camada de saída possui ao menos uma ligação que realimenta a rede. As redes chamadas de BAM (*Bidirecional Associative Memory*) e ART1 e ART2 (*Adaptative Resonance Theory*) são redes recorrentes.

Uma das propriedades mais importantes de uma rede neural artificial é a capacidade de aprender por intermédio de exemplos e fazer inferências sobre o que aprendeu, melhorando gradativamente o seu desempenho(FERNEDA, 2006, p. 26). Esse aprendizado é realizado, ajustando-se os pesos por meio de um processo iterativo.

Algoritmo de aprendizado é um conjunto de regras para a solução de um problema de aprendizado. Há muitos tipos de algoritmos específicos para determinados modelos de redes neurais, os quais diferem entre si sobretudo pelo modo como os pesos são modificados.

Existem basicamente três tipos de aprendizado:

1. Aprendizado Supervisionado: a resposta esperada é indicada para a rede. Trata-se do exemplo deste trabalho onde *a priori* já se sabe se o cliente é bom ou mau;

2. Aprendizado Não Supervisionado: a rede deve basear-se apenas nos estímulos recebidos; a rede deve aprender a agrupar os estímulos;

3. Aprendizado por Reforço: o comportamento da rede é avaliado por um crítico externo.

Cada neurônio contribui para o estímulo de saída. A função de ativação desempenha o papel de restringir a amplitude de saída de um neurônio, em

geral [0,1] ou [-1,1] (HAYKIN, 1999, p. 37). Alguns exemplos de funções de ativação utilizadas são:

- Função Limiar: *f(x)*= 1 se x < k e 0, caso contrário
- Função Logística: *f(x)*= $\frac{1}{1+e^{(-\alpha x)}}$
- Função Tangente Hiperbólica: *f(x)*=tanh(x)

## 2.8. 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 (3000 clientes considerados bons e 3000 considerados maus para cada uma das duas). Os critérios que serão utilizados são apresentados a seguir.

- Taxa de acerto: 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.

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 calcular-se 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 (Ia). Esse indicador privilegiará os modelos que tenham altos índices de acerto para os dois tipos de clientes. Quanto maior for o indicador, melhor será o modelo.

-Teste de Kolmogorov-Smirnov (KS): segundo Crook *et al.* (2007) é uma importante medida de separação, muito utilizada na prática (PICININI *et al.*, 2003; OOGHE *et al.*, 2003; CHEN e WU, 2009). 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 consultadas de acordo com o nível de significância e tamanho da amostra (SIEGEL, 1975, p. 309-310).

Ambas as populações (bons e maus clientes) são divididas em intervalos iguais de pontos atribuídos em cada técnica (escores) e para cada um 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 (KS).

O exemplo apresentado na Tabela1, a seguir, foi adaptado de Lewis (1992, p. 144); o KS deste modelo hipotético é de 28%.

**Tabela 1 - Exemplo de cálculo no teste de Kolmogorov-Smirnov**

| | Número de clientes | | Frequência Acumulada | | |
|---|---|---|---|---|---|
| Faixa de pontos | Bons | Maus | Bons | Maus | Diferença |
| 280 ou mais | 320 | 2 | 2% | 1% | 1% |
| 260-279 | 1291 | 4 | 10% | 2% | 8% |
| 250-259 | 1768 | 17 | 20% | 7% | 14% |
| 240-249 | 2295 | 26 | 34% | 15% | 20% |
| 230-239 | 2571 | 36 | 50% | 25% | 24% |
| 220-229 | 2714 | 42 | 66% | 38% | **28%** |
| 210-219 | 2787 | 81 | 83% | 62% | 21% |
| 200-209 | 2690 | 115 | 99% | 97% | 3% |
| Abaixo de 200 | 106 | 11 | 100% | 100% | 0% |

Fonte: Adaptado de Lewis(1992, p.144).

No caso deste trabalho, 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 clientes.

Analogamente ao resultado 28% destacado na Tabela 1, foram obtidos 6 valores da estatística KS, sendo 3 deles para a técnica de regressão logística (amostras de treinamento, validação e teste) e 3 para redes neurais (amostras de treinamento, validação e teste). Estes 6 resultados serão exibidos na Tabela7, na seção 4.4.

# 3 ASPECTOS METODOLÓGICOS

## 3.1. Descrição do estudo

Uma instituição financeira deseja conceder empréstimos a seus clientes e, para isso, necessita de uma ferramenta que avalie o grau de risco associado a cada empréstimo para auxiliar o processo de tomada de decisão. Para viabilizar este projeto, foram disponibilizadas informações do histórico de clientes que contrataram um crédito pessoal.

## 3.2. O produto de crédito em estudo

O produto em estudo é o 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.Os contratos de crédito pessoal podem ter juros pré ou pós-fixados. Os pré-fixados têm juros estabelecidos quando o cliente contrata o empréstimo e, no pós-fixado, a instituição financeira define um índice que vai ser o responsável pela correção das parcelas do empréstimo ao longo dos meses em que ele tem de ser pago, além dos juros. Nesse caso, o valor da parcela varia ao longo do pagamento de acordo com o indexador fixado no contrato.

Sobre o Crédito Pessoal é cobrado o IOF (Imposto sobre Operações Financeiras), conforme previsto na legislação, e a Taxa de Abertura ou

Renovação de Crédito. Para este estudo é abordada a modalidade com juros pré-fixados com prazos de empréstimos variando de 1 a 12 meses.

### 3.3. Os dados

Para a realização do estudo foram selecionados aleatoriamente, a partir do universo de clientes do banco em estudo, 10.000 contratos de crédito tidos como bons e 10.000 considerados maus, realizados no período de agosto de 2009 a fevereiro de 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. Trata-se de uma base de dados histórica com informações mensais de utilização do produto.

A instituição bancária focalizada neste trabalho concordou em fornecer parte dos dados de seu cadastro de clientes desde que fosse mantida confidencialidade tanto da própria instituição como dos clientes pesquisados. Por essa razão, os dados serão analisados em termos gerais sem nenhuma identificação individualizada; além disso, não será possível divulgar o nome da instituição bancária participante deste estudo. Foi concedidapela instituição bancária uma base de dados armazenados em uma planilha do *software* estatístico *Statistical Package for Social Sciences* (SPSS). Estes dados referem-se aos dois tipos de clientes (bons e maus) com 10000 casos em cada um, que foram selecionados do seu arquivo total de clientes com contratos de crédito pessoal no período definido para o estudo. Nesta seleção usou-se uma sequência de números aleatórios gerada pela funçãodo SPSS "amostra aleatória de casos" (*random sample of cases*) para cada situação: bons e maus clientes.

A definição de contratos "bons" e "maus" baseou-se na política interna da instituição bancária estudada. Será denominada como Variável Resposta a variável que separará os clientes nos dois grupos: bons (pagamento do empréstimo com até 20 dias de atraso) e maus (pagamento com 60 ou mais dias de atraso). Maiores detalhes dessa variável são apresentados na seção 3.4, a seguir.

No trabalho a amostra é dividida em três sub-amostras provenientes do mesmo universo de interesse: uma para construção do modelo, 8.000 dados (sendo 4.000 bons e 4.000 maus); a segunda para validação do modelo construído, 6.000 dados (sendo 3.000 bons e 3.000 maus) e a terceira também com 6.000 (com a mesma divisão equitativa) para testar o modelo obtido.

Cada sub-amostra tem a sua função específica (ARMINGER *et al.,* 1997, p. 294). A sub-amostra de construção do modelo é usada para estimação dos parâmetros do modelo, a sub-amostra de teste irá verificar o poder de predição dos modelos construídos, e a sub-amostra de validação, particularmente numa rede neural, tem a função de validar os parâmetros, evitando o "superajuste" (*overfitting*) do modelo. No modelo de regressão logística a amostra de validação terá o mesmo papel da amostra de teste: avaliar a predição do modelo.

### 3.4. As variáveis

As variáveis explanatórias disponíveis contêm características divididas em dois grupos: Variáveis Cadastrais e Variáveis de Utilização e Restrição. 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.

Para o desenvolvimento de um modelo de *credit 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. Para o produto em estudo, clientes com 60 ou mais dias de atraso foram considerados Maus (inadimplentes) e clientes com no máximo 20 dias de atraso como Bons. Os clientes que apresentam atrasos no intervalo entre bons e maus foram definidos como indeterminados e excluídos deste trabalho.

Convém reforçar que não fez parte do escopo do estudo comentar sobre o tratamento que o banco dá ao cliente que excedeu o limite de 60 dias e depois honrou sua dívida. O próprio banco informou a sua política de que tais clientes são ruins, independentemente de posteriormente se recuperarem.Mas a quitação da dívida, mesmo que tardia, favorece o cliente em um novo pedido de empréstimo a este banco, pois é feito um controle do histórico de cada cliente, para ser levado em conta em novos empréstimos posteriormente. No caso de ser um cliente com grande atraso para fazer a quitação, em empréstimo anterior, é também feita uma entrevista para avaliar qualitativamente a sua real situação e suas chances de pagar o novo empréstimo.

## 4 RESULTADOS

Nesta seção serão abordados os métodos de tratamento das variáveis, o perfil da amostra, a aplicação das duas técnicas estudadas e os resultados obtidos por intermédio de cada uma delas, comparando-se o desempenho destas. Para a análise descritiva, categorização dos dados e aplicação de regressão logística foi utilizado o *software* SPSS for *Windows* v.15.0; para a seleção das amostras e aplicação da rede neural foi utilizado o *software Enterprise Miner* v.6.1.

### 4.1. Tratamento das variáveis e perfil da amostra

As variáveis quantitativas foram categorizadas. Inicialmente foram identificados os decis destas 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. 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 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. 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 descrito 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.

A Tabela 2 apresenta o perfil da amostra pesquisada.

**Tabela 2 - Perfil da amostra coletada no estudo**

| Variável | Distribuição percentual |
|---|---|
| Sexo (masculino, feminino.) | 55%, 45% |
| Estado civil (casado, solteiro, outros) | 52%, 37%, 11% |
| Posse de Fone Residencial (sim, não) | 67%, 33% |
| Posse de Fone Comercial (sim, não) | 71%, 29% |
| Tempo no emprego atual (até 24 meses, de 25 a 72, de 73 a 127, acima de 127 meses) | 15%, 39%, 27%, 19% |
| Salário do cliente (menos de R$ 6.499, de R$ 6.500 a R$ 9.499, de R$ 9.500 a R$ 15.749, de R$ 15.750 a R$ 20.149, de R$ 20.150 a R$ 29.999, R$ 30.000 ou mais) | 8%, 15%, 22%, 28%, 14%, 13% |
| Quantidade de parcelas a quitar (até 4, 5 a 6, 7 a 9, 10 a 12) | 17%, 42%, 25%, 16% |
| Primeira aquisição de empréstimo (sim, não) | 60%, 40% |
| Tempo na residência atual (até 12 meses, de 13 a 24, de 25 a 120, acima de 120 meses) | 16%, 35%, 21%, 28% |
| Valor da parcela(menos de R$ 1.249, de R$ 1.250 a R$ 1.599, de R$ 1.600 a R$ 2.599, R$ 2.600 ou mais) | 22%, 30%, 35%, 13% |
| Valor total do empréstimo (menos de R$ 2.999, de R$ 3.000 a R$ 3.999, de R$ 4.000 a R$ 4.999, de R$ 5.000 a R$ 7.999, de R$ 8.000 a R$ 17.999, R$ 18.000 ou mais) | 20%, 18%, 22%, 15%, 15%, 10% |
| Tipo de crédito (carnê, cheque) | 65%, 35% |
| Idade (até 25, de 26 a 40, de 41 a 58, acima de 58) | 10%, 25%, 35%, 30% |
| CEP residencial (zonas norte, sul, centro, leste, oeste) | 15%, 18%, 21%, 32%, 14% |
| CEP comercial (zonas norte, sul, centro, leste, oeste) | 22%, 19%, 15%, 25%, 19% |
| Profissão(autônomo, aposentado/pensionista, funcionário público, profissional liberal, profissional de comércio, profissional de indústria, profissional de serviços) | 11%, 17%, 12%, 16%, 13%, 21%, 10% |
| % valor do empréstimo / salário (até 27,9%, de 28% a 47,4%, de 47,5% a 64,9%, 65% ou mais) | 15%, 24%, 30%, 31% |
| % valor da parcela / salário (até 9,9%, de 10% a 13,4%, de 13,5% a 16,4%, de 16,5% a 22,4%, 22,5% ou mais) | 28%, 19%, 20%, 18%, 15% |
| Tipo de cliente (bom, mau) | 50%, 50% |

Fonte: Dados do estudo processados

## 4.2. Regressão logística

Para a estimação do modelo de regressão logística utilizou-se a amostra de 8000 casos divididos equitativamente nas categorias de bons e maus clientes.

Inicialmente, é interessante avaliar a relação logística entre cada variável independente e a variável dependente TIPO.

Foi utilizado o recurso de criação de variáveis *dummies*. Logo, considerando-se as variáveis da Tabela 2, usou-se a última categoria como referência e para cada categoria anterior gerou-se uma variável *dummy* que, por definição, assumiu os códigos 1 e 0. Por exemplo, para a variável tempo no emprego, tem-se primeira, segunda e terceira faixa de tempo, sendo a quarta faixa a categoria de referência; as três primeiras faixas são *dummies* e o cliente que, por exemplo, estiver há menos de 24 meses no emprego atual receberá código 1 na primeira *dummy* e 0 nas outras duas.Como a variável tempo no emprego tem quatro faixas, foram construídas três variáveis *dummies*. Se todas estas três variáveis *dummies* apresentassem diferença significante em relação à quarta faixa de tempo em termos de probabilidade de se ter um bom cliente de empréstimo bancário, as três teriam sido incluídas no modelo logístico com base no método *forward stepwise*. Entretanto, observa-se na Tabela 3 que somente as duas primeiras faixas de tempo no emprego foram mantidas no modelo logístico. Logo, não há diferença na probabilidade de o cliente ser um bom pagador do empréstimo se ele estiver trabalhando de 73 a 127 meses (terceira faixa de tempo no emprego) ou acima de 127 meses (quarta faixa). Por outro lado, se o cliente estiver em uma das duas primeiras faixas, a probabilidade de ser bom pagador do empréstimo será menor do que na situação da quarta faixa, uma vez que o sinal dos seus coeficientes logísticos é negativo.

Analogamente, para as demais variáveis, somente foram mantidas no modelo as *dummies* que apresentaram diferença significante na probabilidade de se ter um bom cliente, sempre comparativamente à última categoria de cada variável.

A Tabela 2 contém 18 variáveis independentes e uma variável dependente (Tipo de cliente). O total de categorias das 18 variáveis independentes é 71. Descontando-se a última categoria de cada uma, foram criadas 71 – 18 = 53 variáveis *dummies*. Mas, conforme explicação dada sobre o ocorrido com a variável tempo no emprego, algumas das 53 *dummies* não permaneceram no modelo logístico.

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*.

Neste estudo, Z é 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 3 apresenta as variáveis selecionadas e as estatísticas geradas pelo modelo logístico.

**Tabela 3 – Modelo de regressão logística processado a partir da amostra coletada**

| 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 | -0,4848 | 41,6169 | 0,0000 | -0,0598 | 0,6158 |
| Segunda faixa de tempo de | -0,2166 | 12,6825 | 0,0004 | -0,031 | 0,8053 |
| Primeira faixa de número de | 1,6733 | 276,6224 | 0,0000 | 0,1574 | 5,3296 |
| Segunda faixa de número de | 0,9658 | 169,084 | 0,0000 | 0,1227 | 2,627 |
| Penúltima faixa de número de | 0,3051 | 20,2011 | 0,0000 | 0,0405 | 1,3568 |
| Segunda faixa de tempo de | -0,3363 | 11,2356 | 0,0008 | -0,0289 | 0,7144 |
| Penúltima faixa de tempo de | -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 | 0,9633 | 62,1252 | 0,0000 | 0,0736 | 2,6203 |
| Segunda faixa de valor do | 0,5915 | 24,7781 | 0,0000 | 0,0453 | 1,8067 |
| Terceira faixa de valor do | 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 | -0,3549 | 9,3714 | 0,0022 | -0,0258 | 0,7012 |
| Primeira categoria de CEP | -0,29 | 8,1718 | 0,0043 | -0,0236 | 0,7483 |
| Segunda categoria de CEP | -0,2888 | 20,231 | 0,0000 | -0,0405 | 0,7492 |
| Terceira categoria de CEP | -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 empréstimo/ | 0,2481 | 9,0609 | 0,0026 | 0,0252 | 1,2816 |
| Terceira faixa de empréstimo / | 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: Dados do estudo processados

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.

Variáveis com coeficiente logístico estimado negativo (positivo) indicam que a categoria focalizada, em relação à referência, está associada com diminuição (aumento) na desigualdade e, por conseguinte, diminuição (aumento) na probabilidade de se ter um bom cliente.Pela Tabela 3, tem-se:

- os homens têm menor probabilidade de serem bons clientes em relação às mulheres;

- idem para solteiros em relação à categoria outros; os casados não entraram no modelo, indicando que este estado civil tem igual probabilidade de ser bom cliente em relação a outros;

- quanto menor o tempo de emprego, menor a chance de ser bom cliente;

- quanto menor o número de parcelas, maior a chance de ser bom cliente;

- moradores no endereço atual na segunda e terceira faixa de tempo têm menor chance de serem bons clientes em relação aos que moram há mais tempo;

- aqueles cujo valor da parcela é o menor em relação aos de maior valor têm menor chance de serem bons clientes;

- quanto menor o valor do empréstimo, maior a chance de ser bom cliente;

- aqueles com tipo de crédito carnê têm menor chance de serem bons clientes;

- quanto mais jovem, menor a chance de ser bom cliente;

- clientes moradores na zona norte têm menor chance de serem bons pagadores da dívida em relação aos da zona oeste, usada como referência;

- clientes nos endereços comerciais das zonas norte, sul e centro têm menor chance de serem bons pagadores da dívida em relação aos da zona oeste, usada como referência;

- a primeira, terceira, quinta e sexta categorias de profissão têm maior chance de serem bons clientes do que a sétima categoria, usada como referência;

- a primeira e terceira faixas da relação empréstimo / salário têm clientes com maior chance de pagar o empréstimo do que os da quarta faixa, usada como referência;

- aqueles que adquiriram empréstimo pela primeira vez têm menor chance de serem bons clientes.

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.Os coeficientes de todas as variáveis incluídas no modelo são estatisticamente diferentes de zero.

Há dois testes de significância do modelo final: teste Qui-quadrado da mudança no valor de –2LL e o teste de Hosmer e Lemeshow.A Tabela 4 apresenta o valor inicial de –2LL, com apenas a constante no modelo, seu valor final, a diferença "*improvement*" e o nível descritivo.

**Tabela 4 – Resultados do 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: Dados do estudo processados

No modelo de 28 variáveis, a redução na medida –2LL foi estatisticamente significante.

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. Este resultado conduz à não rejeição da hipótese nula do teste, endossando a aderência do modelo aos dados.

## 4.3. Rede neural

Neste trabalho, será utilizada uma rede com aprendizado supervisionado, pois já se conhece previamente se o cliente em questão é bom ou mau. Segundo Potts (1998, p. 44), a estrutura de rede neural mais utilizada para este tipo de problema é *multilayer perceptron* (MLP), que se trata de uma rede com arquitetura *feedforward* com múltiplas camadas. A literatura consultada (ARMINGER *et al.,* 1997; ARRAES *et al.,* 1999; OLSON *et al.,* 2012; HOFMANN *et al.,* 2007) comprova esta afirmação. Neste estudo será adotada uma rede MLP. As redes MLP podem ser treinadas utilizando-se os seguintes algoritmos: Gradiente Descendente Conjugado, Levenberg-Marquardt, *Back propagation, Quick propagation* ou Delta-bar-Delta. O mais utilizado (LEMOS, *et al*., 2005, p. 230) é o algoritmo *Back propagation*, que será detalhado posteriormente. Para compreensão dos demais, sugere-se a leitura de Fausett (1994) e Haykin (1999).

O modelo implementado tem uma camada de neurônios de entrada; um único neurônio nacamada de saída, que corresponde ao resultado se o cliente é bom ou mau na classificação da rede e uma camada intermediária com três neurônios, pois foi a rede que apresentou melhores resultados, tanto no quesito de maior percentual de acertos, quanto no quesito de redução do erro médio. Redes que possuíam um, dois ou quatro neurônios, também foram testadas.

Cada neurônio da camada escondida é um elemento de processamento que recebe n entradas ponderadas por pesos $W_i$. A soma ponderada das entradas é transformada por meio de uma função de ativação não linear f(.).

A função de ativação utilizada neste estudo será a função logística, $\frac{1}{1+e^{(-g)}}$, onde $g = \sum_{i=1}^{p} W_i X_i$ é a soma ponderada das entradas do neurônio.

O treinamento da rede consiste em encontrar o conjunto de pesos $W_i$ que minimiza uma função de erro. Neste trabalho, será utilizado para o treinamento o algoritmo *Back propagation*. Neste algoritmo a rede opera em uma sequência de dois passos. Primeiro, um padrão é apresentado à camada de entrada da rede. A atividade resultante flui através da rede, camada por

camada, até que a resposta seja produzida pela camada de saída. No segundo passo, a saída obtida é comparada à saída desejada para esse padrão particular. Se esta não estiver correta, o erro é calculado. O erro é propagado a partir da camada de saída até a camada de entrada, e os pesos das conexões das unidades das camadas internas vão sendo modificados, conforme o erro é retropropagado. Esse processo é repetido nas sucessivas iterações até o critério de parada ser atingido.

O erro médio do conjunto de dados de validação foi o critério de parada adotado neste modelo. Esse erro é calculado por intermédio do módulo da diferença entre o valor que a rede localizou e o esperado; calcula-se a sua média para os 8000 casos (amostra de treinamento) ou 6000 casos (amostra de validação).

O processamento detectou que a estabilidade do modelo ocorreu após a nonagésima quarta iteração. Na amostra de validação o erro foi um pouco maior (0,62 x 0,58), o que é comum visto que o modelo é ajustado com base na primeira amostra.

No início, a má classificação é de 50%, porseralocação casual; com mais iterações, é obtidaataxa de 30,6% de erro para a amostra de treino e 32,3% para a de validação (Tabela 5).

**Tabela 5–Resultados das classificações obtidas pela técnica de Redes Neurais**

| Estatísticas obtidas | Treino | Validação |
|---|---|---|
| Classificação incorreta de casos | 0,306 | 0,323 |
| Erro médio | 0,576 | 0,619 |
| Erro quadrático médio | 0,197 | 0,211 |
| Graus de liberdade do modelo | 220 | |
| Graus de liberdade do erro | 7780 | |
| Graus de liberdade total | 8000 | |

Fonte: Dados do estudo processados

## 4.4. Avaliação da performance dos modelos

Após obtidos os modelos, foram escoradas as três amostras e calculados o Ia e o KS para cada um dos modelos. Os resultados são apresentados nas tabelas 6 e 7.

**Tabela 6 - Resultados comparativos de classificação entre os modelos Logístico e Redes Neurais**

| **Regressão logística** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **Treinamento** | | | **Validação** | | | **Teste** | | |
| | | 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 |
| **Rede neural** | | | | | | | | | | |
| | | **Treinamento** | | | **Validação** | | | **Teste** | | |
| | | Predito → | | | Predito → | | | Predito → | | |
| | | Mau | Bom | % Acerto | Mau | Bom | % Acerto | Mau | Bom | % Acerto |
| Observado ↓ | Mau | 3000 | 1000 | 75,0 | 2236 | 764 | 74,5 | 2255 | 745 | 75,2 |
| | Bom | 1280 | 2720 | 68 | 1080 | 1920 | 64,0 | 1046 | 1954 | 65,1 |
| | Total | 4280 | 3720 | 71,5 | 3316 | 2684 | 69,3 | 3301 | 2699 | 70,2 |

Fonte: Dados do estudo processados

Os dois modelos apresentaram bons resultados de classificação, pois, segundo Picinini *et al.* (2003, p. 465): "Modelos de *credit scoring* com taxas de acerto acima de 65% são considerados bons por especialistas". Os percentuais de acerto foram similares.

Outro resultado interessante é que os modelos apresentaram maior taxa de acerto nos clientes maus, sendo superior a 70% a taxa de acerto para clientes maus nas três amostras dos dois modelos.Os resultados são compatíveis com os achados do estudo de Feldman e Gross (2005), em que se aplicou o modelo de *classification trees.* Assim como no presente estudo, a taxa de acertos foi satisfatória. A adequação do modelo é fundamental, tendo em vista que a aceitação dos maus clientes apresenta maior custo à instituição do que a rejeição de um bom cliente (FELDMAN; GROSS, 2005). A Tabela 7, a seguir, apresenta os resultados dos critérios Ia e KS.

**Tabela 7 – Resultados obtidos para os índices de comparação Ia e KS em ambos os modelos**

| Ia | Amostra | | |
|---|---|---|---|
| | **Treinamento** | **Validação** | **Teste** |
| Regressão logística | 51,3 | 48,2 | 49,2 |
| Rede neural | 51,0 | 47,7 | 49,0 |
| **KS** | **Amostra** | | |
| | **Treinamento** | **Validação** | **Teste** |
| Regressão logística | 38 | 35 | 37 |
| Rede neural | 39 | 35 | 35 |

Fonte: Dados do estudo processados

Os valores KS podem ser considerados bons. Picinini *et al.* (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".

Na escolha do modelo mais adequado para estes dados, analisando sob o prisma dos indicadores Ia e KS, seria interessante aplicar um teste estatístico que pudesse comparar os resultados obtidos pelas duas técnicas estudadas neste trabalho.

Cogitou-se em aplicar o tradicional teste de igualdade de proporções para comparar as técnicas com base nos indicadores Ia e KS. Mas a natureza destes indicadores é diferente do conceito de proporção. Com relação ao índice Ia, conforme explicado na seção 2.8, foi usado o produto das taxas de acerto de bons e maus clientes pelo fato de que a instituição financeira focalizada no estudo não declarou qual seria seu principal interesse. Quanto à estatística KS, trata-se da maior diferença em uma distribuição da frequência acumulada entre dois grupos comparados em diversos intervalos de valores, conforme exemplo na Tabela 1.

Por outro lado, a Tabela 6 apresenta taxas de acertos para bons e maus clientes geradas em ambas as técnicas para três amostras: treinamento, validação e teste. Neste caso, é pertinente a aplicação do teste de igualdade de proporções.

A Tabela 8 apresenta o resultado de seis testes de igualdade de proporções com base nos dados da Tabela 6.

**Tabela 8 – Teste de igualdade de proporções em ambos os modelos**

| Teste | Amostra | Regressão Logística | Redes Neurais | Z observado | Decisão |
|---|---|---|---|---|---|
| - % Acertos de maus clientes | Treinamento | 72,8 | 75,0 | -2,215 | $H_0$ falsa |
| - % Acertos de bons clientes | Treinamento | 70,4 | 68,0 | 2,325 | $H_0$ falsa |
| - % Acertos de maus clientes | Validação | 72,3 | 74,5 | -1,958 | $H_0$ verdadeira |
| - % Acertos de bons clientes | Validação | 66,7 | 64,0 | 2,198 | $H_0$ falsa |
| - % Acertos de maus clientes | Teste | 72,5 | 75,2 | -2,350 | $H_0$ falsa |
| - % Acertos de bons clientes | Teste | 67,8 | 65,1 | 2,215 | $H_0$ falsa |

Fonte: Dados do estudo processados

Em cada teste a hipótese $H_0$ refere-se à igualdade das proporções em ambas as técnicas.

Observa-se nos testes 1 e 5 que a técnica de redes neurais produziu taxas de acertos de maus clientes estatisticamente maiores dos que as taxas geradas pela regressão logística.

Os testes 2, 4 e 6 revelam que a regressão logística apresentou taxas estatisticamente maiores de acertos de bons clientes do que a técnica de redes neurais.

As decisões a respeito de cada hipótese basearam-se em um nível de significância de 0,05, cujo valor crítico corresponde a 1,96. Observe-se que no teste 3 o valor observado da estatística Z ficou muito próximo, em módulo, do valor crítico. Usando-se o nível de significância mais restrito igual a 0,03, obtém-se o valor crítico de 2,17, para o qual os testes 1, 2, 4, 5 e 6 mantêm a decisão de rejeição de $H_0$.

Diante dos resultados encontrados, não há evidências da superioridade de nenhum dos dois modelos sobre o outro, pois regressão logística foi mais eficiente na previsão dos bons clientes, e redes neurais na previsão dos maus clientes.

As instituições financeiras têm uma determinada perda para cada tipo de erro:

- % acerto de bons clientes:a técnica de regressão logística reduziu a probabilidade de que bons clientes fossem classificados como maus; o erro de classificar bons clientes como maus pode ser contabilizado como "vendas não realizadas" ou "negócios perdidos"; neste caso, o banco poderia tercaptado juros com os empréstimos e não o fez.

- % acerto de maus pagadores: a técnica de redes neuraisreduziu a probabilidade de que maus clientes fossem classificados como bons; tal erro pode ser contabilizado como "calote"; neste caso, o banco poderia sofrer um grande prejuízo, com o risco de perda inclusive do valor emprestado.

As duas técnicas focalizadas neste trabalho apresentaram taxas maiores de acertos de maus clientes comparativamente aos bons pagadores de empréstimos bancários. Ainda que pareça ser mais importante detectar os maus pagadores de empréstimos bancários, o ideal seria ter um modelo com bom desempenho nas duas taxas de acertos, o que foi investigado por meio do índice Ia referente ao produto das mesmas. Conforme a Tabela 7, este índice apresentou resultados semelhantes nas duas técnicas, com ligeira vantagem para a regressão logística, a qual se deve aos acertos maiores do que a técnica de redes neurais na previsão dos bons clientes.

Como a instituição objeto de investigação neste estudo não se pronunciou sobre a sua maior preocupação em termos de bons e maus clientes, conclui-se que ambos os modelos atendem as suas necessidades de previsão.

## 5 CONCLUSÕES

O objetivo deste estudo foi aplicar e comparar as técnicas regressão logística e redes neurais no desenvolvimento de modelos de predição de *credit scoring* com base em dados de uma grande instituição financeira.

A expressiva elevação do volume de crédito no mercado, em especial entre 2004 e 2012, atraiu o interesse por esse tipo de modelo, tendo em vista que 30% do crédito se concentrou em clientes pessoa física, aumentando o endividamento das famílias (de 18,39% em 2005 para 43,27% em 2012), segundo o BACEN (2012) e o risco de inadimplência (SBICCA; FLORIANI; JUK, 2012).

Os dois modelos apresentaram resultados satisfatórios para a base de dados em questão (taxa de acerto acima de 65%), que foi fornecida por um grande banco de varejo que atua no Brasil. Apesar de o modelo de regressão logística gerar resultados levemente superioresem termos da estatística Ia, conforme a Tabela 7, tal aparente superioridade não se sustenta, quando são feitos os testes de igualdade de proporções separadamente para bons e maus clientes, conforme a Tabela 8. Por essa razão, conclui-se que as duas técnicas aplicadas à base de dados da instituição financeira deste estudo mostraram-se adequadas aos seus objetivos de previsão.

A Tabela 9 compara resultados de outros estudos da literatura pesquisada. Note-se que as quatro técnicas apresentadas na Tabela 9 se mostraram superior em pelo menos um trabalho, o que reforça o ponto mencionado no item 2.5: não há uma técnica que seja sempre superior às demais.

**Tabela 9 - Precisão da classificação dos modelos construídos na literatura pesquisada**

| | Regressão Logística | Redes Neurais | Algoritmos Genéticos | Árvores de Classificação |
|---|---|---|---|---|
| Arminger *et al.* (1997) | 67,6 | 65,2 | | 66,4 |
| Maher e Sen (1998), | 61,7 | 66,7 | | |
| Arraes *et al.* (1999) | 84,8 | 85,4 | | |
| Chen *et al.* (2002) | | 91,9 | 92,9 | |
| Picinini *et al.* (2003) | 63,5 | 64,4 | 67,5 | |
| Lemos, *et al*. (2005) | | 90,0 | | 71,9 |
| Akkoç (2012) | 57,8 | 58,6 | | |
| Olson *et al.*(2012) | 81,3 | 79,8 | | 94,8 |
| Swiderski, *et al.*(2012) | 82,0 | | | 81,7 |

Fonte: Dos autores.

As duas técnicas forneceram melhores ajustes na predição dos maus clientes.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 decisões de crédito devem ser tomadas 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.

Ainda que os modelos estatísticos apresentem limitações por se tratarem de uma abstração da realidade, o processo de tomada de decisões de concessão de crédito poderá ser agilizado com o apoio do modelo logístico e/ou do modelo de redes neurais recomendados neste trabalho.Ressalta-se que os dados contemplados no estudo compreendem o período de 2009 a 2010 e que a realidade econômico-social atual se alterou, com a elevação da taxa básica de juros e da inflação, o que reflete no mercado de crédito. Portanto, os resultados apresentados devem ser interpretados de acordo com a realidade do período. Por outro lado, os modelos estatísticos empregados nesse estudo apresentam-se como ferramenta valiosa para a tomada de decisões ainda no atual ambiente, mais volátil e menos estável; pois agregam informação ao processo decisório.

## REFERÊNCIAS

AKKOÇ, S. An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS)

model for credit scoring analysis: The case of Turkish credit card data. **European Journal of Operational Research,**v. 222, n.1, p.168–178, 2012.

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.

ARRAES, D., SEMOLINI, R.; PICININI, R. Arquiteturas de redes neurais aplicadas a data mining no mercado financeiro: uma aplicação para a geração de credit ratings. In: CONGRESSO BRASILEIRO DE REDES NEURAIS, 4, 1999, São José dos Campos. **Anais...** São José dos Campos: Sociedade Brasileira de Redes Neurais, 1999.

BASTOS. J. Forecasting bank loans loss-given-default. **Journal of Banking and Finance**, v.34, p.2510-2517, 2010

BONFIM, D. Credit risk drivers : Evaluating the contribution of firm level information and of macroeconomic dynamics. **Journal of Banking and Finance**, v.33, p.281-299, 2009

CAMARGOS, M. A.; CAMARGOS, M. C. S.; ARAÚJO, E. A. T. 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.

CAOUETTE, J., ALTMANO, E.; NARAYANAN, P. **Gestão do risco de crédito**. Rio de Janeiro: Qualitymark, 2000.

CHEN, C.; WU, C. Small trades and volatility increases after stock splits**. International Review of Economics & Finance**, v.18, n.4, p.592–610., 2009

CHEN, M.-C.; HUANG, S.-H; CHEN, C.-M. Credit Classification Analysis through the Genetic Programming Approach, Taipei: **Proceedings of the 2002 International Conference in Information Management**, Tamkang University, 2002

CROOK, J. N., EDELMAN, D. B.; THOMAS, L**.** C. (2007). Recent developments in consumer credit risk assessment.**European Journal of Operational Research**, v.183, n.3, p.1447–1465, 2007

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

FAUSETT, L. **Fundamentals of neural networks**. Englewood-Cliffs: Prentice-Hall, 1994.

FELDMAN; D.; GROSS, S. Mortgage default: classification trees analysis. The Journal of Real State Finance. v.30, p.369-396, 2005.

FERNEDA, E. Redes neurais e sua aplicação em sistemas de recuperação de informação. Ciência da Informação, v. 35, n. 1, p. 25–30. 2006.

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.

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, s. A, 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.

HAYKIN, S. **Redes neurais princípios e prática**. Porto Alegre: Bookman, 1999.

HOFFMANN, F., BAESENS, B., MUES, C., VAN GESTEL, T., & VANTHIENEN, J. Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms**. European Journal of Operational Research**, v.177,n.1, p.540–555, 2007.

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

LEMOS, E. P.; STEINER, M. T. A.; NIEVOLA, J. C. Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining . **Revista de Administração da USP,** São Paulo, v. 40, n. 3, p. 225-234, 2005.

LOUZIS, D. P., VOULDIS, A. T., & METAXAS, V. L.. Macroeconomic and bank-specific determinants of non-performing loans in Greece : A comparative study of mortgage , business and consumer loan portfolios. **Journal of Banking and Finance**, v.36, n.4, 2012

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.

MAHER, J.J.; SEN, T.K. Predicting bond ratings using neural networks: a comparison with logistic regression. **Intelligent Systems in Accounting, Finance and Management**. v. 6, n.1, p.59-72, 1998.

MILERIS, R.. Macroeconomic Determinants of Loan Portfolio Credit Risk in Banks**, Inzinerine Ekonomika-Engineering Economics,**v.23,n.5, p.496–504, 2012.

NETER, J., KUTNER, M. H., NACHTSHEIN, C. J.; WASSERMAN, W. **Applied linear statistical models.** Chicago: Irwin, 1996.

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.

OLSON, D. L., DELEN, D., MENG, Y. Comparative analysis of data mining methods for bankruptcy prediction. **Decision Support Systems**, v.52, n.2, p.464–473, 2012.

ORESKI, S., ORESKI, D., ORESKI, G. Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. **Expert Systems with Applications**, v.39,n.16, p.12605–12617, 2012.

PAULA, G. A. **Modelos de Regressão com Apoio Computacional**. São Paulo: IME/USP, 2002. Disponível em: <http://www.ime.usp.br/~giapaula/livro.pdf>. Acesso em: abr. 2011.

PEREIRA, J. M. Gestão do risco operacional: uma avaliação do novo acordo de capitais - Basiléia II, **Revista Contemporânea de Contabilidade,** v. 3, n. 6 103-124, 2006

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.

SANTOS, J. O. **Análise de crédito**: empresas e pessoas físicas. São Paulo: Atlas, 2000.

SADATRASOUL, S. M., GHOLAMIAN, M. R., SIAMI, M.; HAJIMOHAMMADI, Z. Credit scoring in banks and financial institutions via data mining techniques : A literature review, **Journal of AI and Data Mining** v.1 n.2, 119–129, 2013.

SBICCA, A.; FLORIANI, V.; JUK, Y. Expansão do crédito no Brasil e a vulnerabilidade do consumidor. **Revista Economia & Tecnologia**, v.08, n.04, p.5-16, 2012.

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.

SWIDERSKI, B., KUREK, J., & OSOWSKI, S. Multistage classification by using logistic regression and neural networks for assessment of financial condition of company. **Decision Support Systems**, v.52,n.2, p.539–547, 2012.

THOMAS, L. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. **International Journal of Forecasting**, London, v.16, n.2, p.149-172, 2000.

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.

TSAI, B. Comparison of Binary Logit Model and Multinomial Logit Model in Predicting Corporate Failure. **Review of Economics & Finance**, v.1994, p.99–111, 2010

YANAKA, G.; HOLLAND, M. Basiléia II e a Exigência de Capital para Risco de Credito de Bancos no Brasil. **Revista Brasileira de Finanças**, v. 8, p. 5-21, 2010.