Análise de Risco de Crédito com o Uso de Modelos de Regressão Logística e Redes Neurais

AUTORES
MARIA APARECIDA GOUVÊA
Universidade de São Paulo
magouvea@usp.br

ERIC BACCONI GONÇALVES
Universidade de São Paulo
ebacconi@uol.com.br

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. Há alguns anos ao fazer uma solicitação de crédito, o cliente preenchia uma proposta que seria avaliada por um ou mais analistas que apresentavam um parecer em relação ao pedido (SEMOLINI, 2002, p. 103). Apesar de eficaz, este processo era lento, por não permitir a análise de muitos pedidos. 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 de forma qualitativa. Este trabalho ilustra 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, podendo-se aumentar a eficiência do seu negócio. Neste estudo são apresentados, em um primeiro momento, conceitos de crédito e risco. Posteriormente, a partir de uma amostra de dados, fornecida por uma grande instituição financeira brasileira, estão desenvolvidos dois modelos, aplicando-se duas técnicas para classificação de clientes: Regressão Logística e Redes Neurais. Em uma etapa final, é apontado o modelo com melhores indicadores de qualidade e de ajuste aos dados.

Abstract
The decisions making of credit concession is based mainly on the evaluation of the risk of loans of potential credit products users. Some years ago, during a credit request, the client used to write a proposal that would be evaluated by one or more analysts that presented a decision related to the request (SEMOLINI, 2002, p. 103). Although efficient, this process was slow because it did not allow the evaluation of a lot of requests. With the technological advance, statistical models were developed in order to support the analysis of credit requests, which some decades ago used to be made through qualitative ways. This study shows the procedures to be adopted by a company in order to identify the best credit model to evaluate the risk of consumer loans. Use of the best fitted model will favor the definition of an adequate business strategy thereby increasing profits. The first phase of this study introduces concepts of credit and risk. Subsequently, with a sample set of applicants from a large Brazilian financial institution, two credit scoring models are built applying two techniques: Logistic Regression and Neural Networks. Finally, the model with the best quality and data adjust indicators is pointed out.

Palavras-chave: crédito, regressão logística, redes neurais

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 os vultuosos lucros que provinham da desvalorização da moeda (ROSA, 2000, p. 1). Após o fim do período inflacionário, percebeu-se a necessidade de se aumentarem 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.

Há alguns anos ao fazer uma solicitação de crédito, o cliente preenchia uma proposta que seria avaliada por um ou mais analistas que apresentavam um parecer em relação ao pedido (SEMOLINI, 2002, p. 103). Apesar de eficaz, este processo era lento, por não permitir a análise de muitos pedidos. Os modelos de análise para concessão de crédito começaram a ser adotados nas instituições financeiras com o objetivo de acelerar a avaliação das propostas.

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.

2. Objetivos do estudo

Pretende-se com este trabalho:

- Desenvolver e comparar dois modelos de *credit scoring*, mediante o uso de duas técnicas estatísticas: Regressão Logística e Redes Neurais;
- Identificar o melhor modelo para a classificação de clientes.

3. Fundamentação teórica

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

3.1. 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, p. 15).

Atualmente, o crédito ao consumidor é uma grande indústria que opera no mundo. Grandes varejistas impulsionam suas vendas, fornecendo crédito. 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 decisivo na decisão de disponibilizar ou não o crédito: o risco.

3.2. Risco de Crédito

O risco de crédito é a mais antiga forma de risco no mercado financeiro (FIGUEIREDO, 2001, p. 9). É conseqüência de uma transação financeira contratada entre um fornecedor de fundos (doador do crédito) e um usuário (tomador do crédito). Antes de qualquer sofisticação, produto 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.
No universo do crédito ao consumidor, a promessa de pagamento futuro envolve a idéia 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, p. 2). 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. O risco apresentado pelo solicitante é de extrema importância no processo de concessão de crédito, devendo ser considerados vários quesitos na sua avaliação.

3.3. 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, 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 avaliação do risco de um potencial cliente pode ser feita de duas maneiras:

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

Atualmente, 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 (concessão ou não do crédito).

3.4. Modelos de *credit scoring*

O pioneiro dos modelos de crédito foi Henry Wells, executivo da Spiegel Inc. que desenvolveu um modelo de escore para crédito durante a Segunda Guerra Mundial (LEWIS, 1992, p. 19). Wells necessitava de ferramentas que permitissem aos analistas inexperientes fazer avaliação de crédito, pois muitos de seus funcionários experientes foram recrutados para a Guerra.
Nos anos cinqüenta, os modelos de escore 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, p. 154).

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. Varejistas como a Wards, Blomingdale's e J.C. Penney aparecem entre as pioneiras neste segmento.
Atualmente, aproximadamente 90% das empresas americanas que oferecem algum tipo de crédito ao consumidor utilizam modelos de *credit scoring.*
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.
Há sete passos a serem seguidos para se construir um modelo de *credit scoring*, 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

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, que são aqueles que estão na fronteira entre serem bons ou maus, não existindo, ainda, uma posição clara para eles. 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 (ROSA, 2000; OHTOSHI, 2003; SEMOLINI, 2002; HAND; HENLEY, 1997; entre outros).

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 qualquer possível viés devido à diferença de tamanhos. Não existe um número fixo para a amostra; entretanto, Lewis (1992, p. 31) sugere uma amostra de 1.500 clientes bons e 1.500 clientes maus para serem propiciados resultados robustos. Costuma-se trabalhar com três amostras, uma para construção do modelo, uma para validação e outra para teste do modelo.

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 trabalho serão utilizadas Regressão Logística e Redes Neurais. Hand e Henley (1997) destacam ainda Análise de Discriminante, Regressão Linear, e Árvores de Decisão, como métodos utilizados na prática. Recentemente alguns estudiosos também têm utilizado Análise de Sobrevivência (HARRISON; ANSELL, 2002; ANDREEVA, 2003). Não há um método claramente melhor que os demais, pois depende de como a técnica se ajusta aos dados.

6. Definição dos critérios de comparação dos modelos

Aqui será definida a medida de comparação dos modelos, normalmente pelo índice de acertos e a estatística de Kolmogorov-Smirnov (KS).

7. Seleção e Implantação do melhor modelo

Por meio dos critérios previamente definidos, o melhor modelo é escolhido. Com isso deve-se programar a implantação do modelo. A instituição deve adequar seus sistemas para receber o algoritmo final e programar a utilização do mesmo junto às demais áreas envolvidas.

3.5. Regressão logística

Regressão Logística é a técnica mais utilizada no mercado para o desenvolvimento de modelos de *credit scoring* (ROSA, 2000; OHTOSHI, 2003). Apresenta vantagem em relação à Análise Discriminante, por não pressupor dados de entrada com distribuição normal, embora seja desejável que as variáveis tenham essa distribuição (HAIR *et al*, 1998, p. 231).

3.5.1. Conceitos

Nos modelos de regressão logística, a variável dependente é, em geral, uma variável binária (nominal ou ordinal) e as variáveis independentes podem ser categóricas (desde que dicotomizadas após transformação) ou contínuas.

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

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

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

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

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

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

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

3.5.2. Método de escolha das variáveis

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

3.6. Redes neurais artificiais

Redes Neurais Artificiais são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento por intermédio de experiências.

3.6.1. Conceitos

Um modelo de rede neural artificial processa certas características e produz respostas similarmente ao cérebro humano. 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, 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:

$y = 1$, *se* $a \geq k$

$y = 0$, *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.

3.6.1.1 Arquitetura

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; podem ser consideradas como extratoras de características;
- Camada de Saída: onde o resultado final é concluído e apresentado.

Existem basicamente 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.

3.6.1.2. Processo de Aprendizado

A propriedade mais importante das redes neurais é a habilidade de "aprender" de acordo com o ambiente e com isso melhorar seu desempenho (CASTRO JR., 2003, p. 92). Esse aprendizado é realizado, ajustando-se os pesos por meio de um processo iterativo. O objetivo do processo é a obtenção de um algoritmo de aprendizado que permita uma solução generalizada para certa classe de problema.
Denomina-se algoritmo de aprendizado um conjunto de regras bem definidas para a solução de um problema de aprendizado. Existem muitos tipos de algoritmos específicos para determinados modelos de redes neurais. Estes algoritmos diferem entre si principalmente pelo modo como os pesos são modificados.
Existem basicamente três tipos de aprendizado:

1. Aprendizado Supervisionado: neste tipo de aprendizado, é indicada para a rede qual a resposta esperada. Trata-se do exemplo deste trabalho onde a priori já se sabe se o cliente é bom ou mau;
2. Aprendizado Não Supervisionado: neste tipo de aprendizado, a rede deve basear-se apenas nos estímulos recebidos; a rede deve aprender a agrupar os estímulos;
3. Aprendizado por Reforço: neste tipo de aprendizado, o comportamento da rede é avaliado por um crítico externo.

Cada tipo de aprendizado possui vários algoritmos possíveis de serem utilizados.

3.6.1.3. Funções de Ativação

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)

3.7. 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.

3.7.1. 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.

3.7.2. Teste de Kolmogorov-Smirnov

O outro critério bastante utilizado na prática (PICININI *et al*, 2003; OOGHE *et al*, 2001; Pereira, 2004) a ser abordado neste trabalho é o teste de Kolmogorov-Smirnov (KS).
O teste de KS 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). Este teste se baseia na distribuição acumulada dos escores dos clientes considerados como bons e maus.
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 (ver SIEGEL, 1975, p. 309-310). Neste trabalho, como as amostras são grandes, a tendência é que todos os modelos rejeitem a hipótese de igualdade nas distribuições. Será considerado melhor modelo o de maior valor no teste, pois este resultado indica uma separação maior entre bons e maus.

4. Aspectos metodológicos

4.1. 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.

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

4.3. 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 2002 a fevereiro de 2003, 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. Com esta estrutura pode-se acompanhar o andamento do contrato e o momento o cliente deixou de pagar uma ou mais parcelas.
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 eqüitativa) 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.

4.4. 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.
Tanto as Variáveis Cadastrais como as de Utilização e Restrição são coletadas no momento em que o cliente contrata o produto.
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. 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 denominados Indeterminados representam um grupo cujo comportamento de crédito não é suficientemente claro para indicá-los como bons ou maus pagadores. Na prática, estes clientes que não estão claramente definidos como bons ou maus são analisados qualitativamente pelo analista de crédito.

5. Aplicação

Nesta seção serão abordados os métodos de tratamento das variáveis, 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.11.0; para a seleção das amostras e aplicação da rede neural foi utilizado o *software Enterprise Miner* v.4.1.

5.1. Tratamento das variáveis

Inicialmente, as variáveis quantitativas foram categorizadas.
Para a categorização das variáveis contínuas, 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. Conforme Pereira (2004, p. 49), existem duas razões para se fazer uma "nova categorização" das variáveis qualitativas. A primeira é evitar categorias com um número muito pequeno de observações, o que pode levar a estimativas pouco robustas dos parâmetros associados a elas. A segunda é eliminar parâmetros do modelo; se duas categorias têm risco próximo, é razoável agrupá-las 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 adotado por Rosa (2000, p. 17), que é explicado por Hand e Henley (1997, p. 527);
- A variável salário do cônjuge foi removida da análise por ter muitos dados faltantes;
- Foram criadas duas novas variáveis, percentual do valor do empréstimo sobre o salário e percentual do valor da parcela sobre o salário, categorizadas em faixas.

A Tabela 1 apresenta as variáveis utilizadas.

Tabela 1 - Variáveis categorizadas

Variável	**Categoria**	**Nome da variável**
Sexo	Masculino Feminino	V_SEXO_M V_SEXO_F
Estado civil	Casado Solteiro Outros	V_EST_C V_EST_S V_EST_O
Fone residencial	Sim Não	V_FN_R_S V_FN_R_N
Fone comercial	Sim Não	V_FN_C_S V_FN_C_N
Tempo no emprego atual	Até 24 meses, 25 a 72, 73 a 127, Acima de 127	V_TP_E1 a V_TP_E4
Salário	Até R 650, + 650 a 950, +950 a 1575, +1575 a 2015, +2015 a 3000, Acima de R 3000	V_SAL_F1 a V_SAL_F6
Quantidade de parcelas	Até 4, 5 a 6, 7 a 9, 10 a 12	V_Q_PC_1 a V_Q_PC_4
Primeira aquisição	Sim Não	V_PR_AQ_S V_PR_AQ_N
Tempo na residência atual	Até 12 meses, 13 a 24, 25 a 120, Acima de 120	V_TP_R1 a V_TP_R4
Valor da parcela	Até R 125, +125 a 160, +160 a 260, Acima de R 260	V_VL_PR1 a V_VL_PR4
Valor total do empréstimo	Até R 300, +300 a 400, +400 a 500, +500 a 800, +800 a 1800, Acima de R 1800	V_VL_EM1 a V_VL_EM6
Tipo de crédito	Carnê Cheque	V_CRE_CN V_CRE_CH
Idade	Até 25 anos, 26 a 40, 41 a 58, Acima de 58 anos	V_IIDADE1 a V_IDADE4
Faixa de CEP residencial	1 2 3 4 5	V_CEP_F1 a V_CEP_F5
Faixa de CEP comercial	1 2 3 4 5	V_CEC_F1 a V_CEC_F5
Código de profissão	1 2 3 4 5 6 7	V_COD_P1 a V_COD_P7
% Valor da parcela/Salário	Até 10%, 10.1% a 13.5%, 13.6% a 16.5%, 16.6% a 22.5%, Acima de 22.5%	V_FX_P1 a V_FX_P5
% Valor do Emprést/Salário	Até 28%, 28.1% a 47.5%, 47.6% a 65%, Acima de 65%	V_FX_E1 a V_FX_E4
Tipo de cliente	1 = Bom 0 = Mau	TIPO

5.2. Regressão logística

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

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

Como um dos objetivos desta análise é identificar quais variáveis são mais eficientes na caracterização dos dois tipos de clientes bancários, um procedimento *stepwise* foi empregado. O método de seleção escolhido foi o já mencionado *forward stepwise*.
Das 53 variáveis independentes disponíveis, considerando-se k-1 *dummies* para cada variável de k níveis, foram incluídas 28 variáveis no modelo.
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 2 apresenta, por variável, as estimativas dos coeficientes logísticos, os desvios-padrão das estimativas, as estatísticas de Wald, os graus de liberdade e os níveis descritivos dos testes de significância das variáveis independentes.

Tabela 2 - Modelo de Regressão Logística

Variável	Coeficiente logístico estimado	Desvio-padrão	Wald	Graus de liberdade	Nível descritivo	R - Correlação parcial	Exp(B)
V_SEXO_M	-0,314	0,053	35,0381	1	0,0000	-0,0546	0,7305
V_EST_S	-0,1707	0,0556	9,4374	1	0,0021	-0,0259	0,8431
V_TP_E1	-0,4848	0,0751	41,6169	1	0,0000	-0,0598	0,6158
V_TP_E2	-0,2166	0,0608	12,6825	1	0,0004	-0,031	0,8053
V_Q_PC_1	1,6733	0,1006	276,6224	1	0,0000	0,1574	5,3296
V_Q_PC_2	0,9658	0,0743	169,084	1	0,0000	0,1227	2,627
V_Q_PC_3	0,3051	0,0679	20,2011	1	0,0000	0,0405	1,3568
V_TP_R2	-0,3363	0,1003	11,2356	1	0,0008	-0,0289	0,7144
V_TP_R3	-0,1451	0,0545	7,0946	1	0,0077	-0,0214	0,865
V_VL_PR1	-0,2035	0,0878	5,3672	1	0,0205	-0,0174	0,8159
V_VL_EM1	0,9633	0,1222	62,1252	1	0,0000	0,0736	2,6203
V_VL_EM2	0,5915	0,1188	24,7781	1	0,0000	0,0453	1,8067
V_VL_EM3	0,4683	0,0889	27,7693	1	0,0000	0,0482	1,5972
V_CRE_CN	-1,34	0,0853	246,7614	1	0,0000	-0,1486	0,2618
V_IDADE1	-0,7429	0,1371	29,3706	1	0,0000	-0,0497	0,4757
V_IDADE2	-0,6435	0,0902	50,924	1	0,0000	-0,0664	0,5254
V_IDADE3	-0,2848	0,0808	12,4401	1	0,0004	-0,0307	0,7522
V_CEP_F1	-0,3549	0,1159	9,3714	1	0,0022	-0,0258	0,7012
V_CEC_F1	-0,29	0,1014	8,1718	1	0,0043	-0,0236	0,7483
V_CEC_F2	-0,2888	0,0642	20,231	1	0,0000	-0,0405	0,7492
V_CEC_F3	-0,2662	0,074	12,9248	1	0,0003	-0,0314	0,7663
V_COD_P1	0,3033	0,0945	10,3013	1	0,0013	0,0274	1,3543
V_COD_P3	0,5048	0,0889	32,2381	1	0,0000	0,0522	1,6566
V_COD_P7	0,4752	0,1048	20,5579	1	0,0000	0,0409	1,6084
V_COD_P8	0,1899	0,0692	7,534	1	0,0061	0,0223	1,2091
V_FX__E1	0,2481	0,0824	9,0609	1	0,0026	0,0252	1,2816
V_FX__E3	0,164	0,0664	6,0906	1	0,0136	0,0192	1,1782
V_PR_AQ_N	-0,6513	0,0526	153,5677	1	0,0000	-0,1169	0,5213
Constante	0,5868	0,0903	42,2047	1	0,0000		

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 indicam que a categoria focalizada, em relação à referência, está associada com diminuição na desigualdade e, por conseguinte, diminuição na probabilidade de se ter um bom cliente. As variáveis que mais afetam

positivamente a probabilidade de se ter um bom cliente são V_Q_PC_1, V_Q_PC_2 E V_VL_EM1. No extremo oposto, as variáveis com maior impacto negativo sobre esta probabilidade são V_CRE_CN, V_PR_AQ E V_IDADE2. Pela tabela 2, os coeficientes de todas as variáveis incluídas no modelo logístico 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 3 apresenta o valor inicial de –2LL, com apenas a constante no modelo, seu valor final, a diferença "*improvement*" e o nível descritivo.

Tabela 3 - 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

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.

5.3. 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 utilizado 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; ZERBINI, 2000; CASTRO JR., 2003; OHTOSHI, 2003) 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 comum (CASTRO JR., 2003, p. 142) 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 camada 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 seqüência 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%, por ser alocação casual; com mais iterações, é obtida a taxa de 30,6% de erro para a amostra de treino e 32,3% para a de validação (Tabela 4).

Tabela 4 - Estatísticas da Rede Neural adotada

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	

5.4. 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 5 e 6.

Tabela 5 - Resultados de classificação

Regressão logística										
		Treinamento			**Validação**			**Teste**		
		Predito →			Predito →			Predito →		
		Mau	Bom	% Acerto	Mau	Bom	% Acerto	Mau	Bom	% Acerto
Observado ↓	Mau	2833	1167	70.8	2111	889	70.4	2159	841	72.0
	Bom	1294	2706	67.7	1078	1922	64.1	1059	1941	64.7
	Total	4127	3873	69.2	3189	2811	67.2	3218	2782	68.3
Rede neural										
		Treinamento			**Validação**			**Teste**		
		Predito →			Predito →			Predito →		
		Mau	Bom	% Acerto	Mau	Bom	% Acerto	Mau	Bom	% Acerto
Observado ↓	Mau	2979	1021	74.5	2236	764	74.5	2255	745	75.2
	Bom	1430	2570	64.3	1177	1823	60.8	1193	1807	60.2
	Total	4409	3591	69.4	3413	2587	67.7	3448	2552	67.7

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 muito 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. A Tabela 6, a seguir, apresenta os resultados dos critérios Ia e KS.

Tabela 6 - Índices de comparação

Ia		**Amostra**	
	Treinamento	**Validação**	**Teste**
Regressão logística	47.9	45.1	46.6
Rede neural	47.9	45.3	45.3
KS		**Amostra**	
	Treinamento	**Validação**	**Teste**
Regressão logística	38	35	37
Rede neural	39	35	35

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 àquele 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, foi eleito o modelo construído por **regressão logística**, pois, apesar de ter resultados similares aos obtidos por redes neurais, apresentou melhores resultados na amostra de teste, sugerindo ser o mais adequado para a aplicação em outras bases de dados.

6. Conclusões e recomendações

O objetivo deste estudo foi desenvolver modelos de predição de *credit scoring* com base em dados de uma grande instituição financeira usando Regressão Logística e Redes Neurais.

Os dois modelos 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 obteve resultados levemente superiores. O modelo proposto por este estudo para que a instituição pontue seus clientes é o modelo logístico com 28 variáveis exibidas na Tabela 2.

Não foi objeto deste estudo uma abordagem mais profunda das técnicas. As redes neurais apresentam uma grande gama de estruturas e variações que podem ser melhor exploradas.

Técnicas novas em problema de risco de crédito, como análise de sobrevivência, também merecem atenção em estudos futuros.

Referências bibliográficas

ANDREEVA, G. (2003) *European generic scoring models using logistic regression and survival analysis*. Bath: Young OR Conference.

ARMINGER, G., ENACHE, D. & BONNE, T. (1997) Analyzing Credit Risk Data: A Comparison of Logistic Discrimination, Classification Trees and Feedforward Networks. *Computational Statistics*, 12(2): 293-310. Berlim: Springer-Verlag.

ARRAES, D., SEMOLINI, R. & PICININI, R. (1999) *Arquiteturas de Redes Neurais Aplicadas a Data Mining no Mercado Financeiro*. Uma Aplicação para a Geração de Credit Ratings. São José dos Campos: IV Congresso Brasileiro de Redes Neurais.

BERRY, M. & LINOFF, G. (1997) *Data Mining Techniques*. New York: Wiley.

CANTON, A. W. P. (1988) *Aplicação de modelos estatísticos na avaliação de produtos*. Tese (Livre-Docência). Departamento de Administração. Universidade de São Paulo. FEA/USP.

CAOUETTE, J., ALTMANO, E. & NARAYANAN, P. (2000) *Gestão do Risco de Crédito*. Rio de Janeiro: Qualitymark.

CASTRO JR., F. H. F. (2003) *Previsão de Insolvência de Empresas Brasileiras Usando Análise Discriminante, Regressão Logística e Redes Neurais*. Dissertação de Mestrado. Departamento de Administração Universidade de São Paulo. FEA/USP.

CHEN, M.-C., HUANG, S.-H. & CHEN, C.-M. (2002) *Credit Classification Analysis through the Genetic* Programming Approach. Taipei: Proceedings of the 2002 International Conference in Information Management. Tamkang University.
DOBSON, A. (1990) *An Introduction to Generalized Linear Models*. London: Chapman & Hall.
FAUSETT, L. (1994) *Fundamentals of Neural Networks*. Englewood-Cliffs: Prentice-Hall.
FENSTERSTOCK, F. (2005) Credit Scoring and the Next Step. *Business Credit*, 107(3): 46-49. New York: National Association of Credit Management.
FIGUEIREDO, R. P. (2001) *Gestão de Riscos Operacionais em Instituições Financeiras* – Uma Abordagem Qualitativa. Dissertação de Mestrado. Belém: Universidade da amazônia UNAMA.
FRITZ, S. & HOSEMANN, D. (2000) Restructuring the Credit Process: Behaviour Scoring for German Corporates. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 9(1): 9-21. Nottingham: John Wiley & Sons.
HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. (1998) *Multivariate Data Analysis*, New Jersey: Prentice-Hall.
HAND, D. J. & HENLEY, W. E. (1997) Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of Royal Statistical Society*: Series A (160): 523-541. London: Royal Statistical Society.
HARRISON, T. & ANSELL, J. (2002) Customer retention in the insurance industry: using survival analysis to predict cross-selling opportunities. *Journal of Financial Services Marketing*, 6(3): 229-239. London: Henry Stewart Publications.
HAYKIN, S. (1999) *Redes Neurais Princípios e Prática*. Porto Alegre: Bookman.
HUANG, Z., CHEN, H., HSU, C-J., CHEN, W. & WU, S. (2004) Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, 37(4): 543-558. St. Louis: Elsevier Science.
LEWIS, E. M. (1992) *An Introduction to Credit Scoring*. San Rafael: Fair Isaac and Co., Inc.
NANDA, S. & PENDHARKAR, P. (2001) Linear models for minimizing misclassification costs in bankruptcy prediction. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 10(3): 155-168. Nottingham: John Wiley & Sons.
NETER, J., KUTNER, M. H., NACHTSHEIN, C. J. & WASSERMAN, W. (1996) *Applied Linear Statistical Models*. Chicago: Irwin.
OHTOSHI, C. (2003) *Uma Comparação de Regressão Logística, Árvores de Classificação e Redes Neurais*: Analisando Dados de Crédito. Dissertação de Mestrado. Departamento de Estatística. Universidade de São Paulo. IME/USP.
OOGHE, H., CAMERLYNCK, J. & BALCAEN, S. (2001) *The Ooghe-Joos-De Vos Failure Prediction Models*: A Cross-Industry Validation. Working paper. Department of Corporate Finance. University of Ghent.
PAULA, G. A. (2002) *Modelos de Regressão com Apoio Computacional*. Material disponível em http://www.ime.usp.br/~giapaula/livro.pdf acesso em 05/12/2004.
PEREIRA, G. H. A. (2004) *Modelos de risco de crédito de clientes*: Uma aplicação a dados reais. Dissertação de Mestrado. Departamento de Estatística. Universidade de São Paulo. IME/USP.
PICININI, R., OLIVEIRA, G. M. B. & MONTEIRO, L. H. A. (2003) *Mineração de Critério de Credit Scoring Utilizando Algoritmos Genéticos*. Bauru: VI Simpósio Brasileiro de Automação Inteligente: 463-466.
POTTS, W. J. E. (1998) *Data Mining Primer Overview of Applications and Methods*. Carrie: SAS Institute Inc.
ROSA, P. T. M. (2000) *Modelos de Credit Scoring: Regressão Logística, CHAID e REAL*. Dissertação de Mestrado. Departamento de Estatística. Universidade de São Paulo. IME/USP.

SANTOS, J. O. (2000) *Análise de Crédito*: Empresas e Pessoas Físicas. São Paulo: Atlas.
SEMOLINI, R. (2002) *Support Vector Machines, Inferência Transdutiva e o Problema de Classificação*. Dissertação de Mestrado. Departamento de Engenharia Elétrica. Universidade Estadual de Campinas. FEEC/UNICAMP.
SIEGEL, S. (1975) *Estatística Não-Paramétrica para as Ciências do Comportamento*. São Paulo: McGraw-Hill.
THOMAS, L. (2000) A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers. *International Journal of Forecasting*, 16(2): 149-172. London: Elsevier.
TREVISANI, A. T., GONÇALVES, E. B., D'EMÍDIO, M. & HUMES, L. L. (2004) *Qualidade de Dados* – Desafio Crítico para o Sucesso do Business Intelligence. Itajaí: XVIII Congresso Latino Americano de Estratégia.
ZERBINI, M. B. A. A. (2000) *Três Ensaios sobre Crédito*. Tese de Doutorado. Departamento de Economia. Universidade de São Paulo. FEA/USP.