UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE
DEPARTAMENTO DE ADMINISTRAÇÃO

ANÁLISE DE RISCO DE CRÉDITO COM O USO DE MODELOS DE REGRESSÃO LOGÍSTICA, REDES NEURAIS E ALGORITMOS GENÉTICOS.

Eric Bacconi Gonçalves

Orientadora: Profª Drª Maria Aparecida Gouvêa

SÃO PAULO
2005

Prof. Dr. Adolpho José Melfi
Reitor da Universidade de São Paulo

Profa. Dra. Maria Tereza Leme Fleury
Diretora da Faculdade de Economia, Administração e Contabilidade

Prof. Dr. Eduardo Pinheiro Gondim de Vasconcellos
Chefe do Departamento de Administração

Prof. Dr. Isak Kruglianskas
Coordenador do Programa de Pós-Graduação de Administração

ERIC BACCONI GONÇALVES

ANÁLISE DE RISCO DE CRÉDITO COM O USO DE MODELOS DE REGRESSÃO LOGÍSTICA, REDES NEURAIS E ALGORITMOS GENÉTICOS.

Dissertação apresentada ao Departamento de Administração da Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo como requisito para a obtenção do título de Mestre em Administração.

Orientadora: Profª Drª Maria Aparecida Gouvêa

SÃO PAULO

2005

Dissertação defendida e aprovada no Departamento de Administração da Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo – Programa de Pós-Graduação em Administração, pela seguinte banca examinadora:

FICHA CATALOGRÁFICA

Elaborada pela Seção de Processamento Técnico do SBD/FEA/USP

Gonçalves, Eric Bacconi
Análise de risco de crédito com o uso de modelos de regressão logística, redes neurais e algoritmos genéticos / Eric Bacconi Gonçalves. -- São Paulo, 2005.
96 p.

Dissertação (Mestrado) – Universidade de São Paulo, 2005.
Bibliografia.

1. Administração financeira 2. Estatística aplicada 3. Redes neurais 4. Algoritmos genéticos I. Universidade de São Paulo. Faculdade de Economia, Administração e Contabilidade. II. Título.

CDD – 658.15

Aos meus pais Simião e Cleide e à minha esposa Mariana que me incentivam e apóiam em todos os momentos.

AGRADECIMENTOS

A Deus por ter permitido que eu chegasse até aqui.

À professora Cida, por confiar no meu potencial, acreditar no tema deste trabalho e compartilhar seu conhecimento.

À minha esposa Mariana que faz que eu entenda o significado da palavra companheira.

Aos meus pais Simião e Cleide e à minha irmã Paula por sempre me apoiarem.

Aos meus familiares, que sempre torcem por mim.

Ao professor Ronaldo Zwicker, membro da banca, que desde as primeiras disciplinas da pós incentiva os alunos a “pensar”.

Ao professor e brilhante estatístico Adolpho Canton, que fez parte da minha qualificação e mostrou a importância da Estatística no contexto da Administração.

Ao professor e amigo Ricardo Furtado, membro da banca, por dividir seu conhecimento e colaborar com dicas preciosas na preparação deste trabalho.

Aos amigos Robinson Semolini e Wander Mion que cooperaram transmitindo suas experiências práticas.

Aos mestres Alexandre Trevisani, Leila Humes e Marcelo D´Emídio, companheiros nas disciplinas cursadas.

Aos meus colegas de trabalho, que me ajudaram sempre que possível.

Aos funcionários, professores e alunos da FEA que tornaram agradáveis estes dois anos e meio que passei cursando o mestrado.

“***O saber não ocupa lugar***”

Anônimo

RESUMO

Praticamente todas as grandes instituições brasileiras que trabalham com concessão de crédito utilizam-se de modelos para avaliar o risco de inadimplência dos potenciais contratantes de produtos de crédito. Qualquer avanço nas técnicas, que resulte no aumento da precisão de um modelo de previsão, acarreta ganhos financeiros para a instituição.

Neste trabalho 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 três modelos, aplicando-se três técnicas para a classificação de clientes: Regressão Logística, Redes Neurais e Algoritmos Genéticos. Em uma etapa final, são avaliadas e comparadas a qualidade e performance dos modelos desenvolvidos, onde é apontado qual o modelo que melhor se ajusta aos dados.

Os resultados obtidos pelos modelos de regressão logística e rede neural são satisfatórios e bastante próximos, sendo o primeiro ligeiramente superior. O modelo embasado por algoritmos genéticos apresenta também bons resultados embora num patamar inferior aos dois já citados.

Este trabalho ilustra os procedimentos a serem adotados por uma empresa para identificar o melhor modelo de concessão de crédito que tenha boa aderência aos seus dados. A adoção do melhor modelo detectado permite o direcionamento da estratégia da instituição, podendo aumentar a eficiência do seu negócio.

Palavras-chave: risco de crédito, modelos de *credit scoring*, regressão logística, redes neurais, algoritmos genéticos.

ABSTRACT

Most of the large Brazilian institutions which work with credit concession use credit models to evaluate the risk of consumer loans. Any improvement in techniques that results in the precision increase of a prediction model, will provide financial gains to the institution.

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, three credit scoring models are built applying three different techniques: Logistic Regression, Neural Networks and Genetic Algorithms. Finally, the quality and the performance of these models are evaluated and compared, and the best one is identified.

The results obtained by the logistic regression model and neural network model are good and very similar, but the first one is slightly better. The results obtained with the genetic algorithm model are also good, but a little bit inferior.

This study shows proceedings to be adopted by a financial institution in order to identify the best credit model to evaluate the risk of consumer loans. The use of the proper model will help the definition of an adequate business strategy and increase profits.

Keywords: credit risk, credit scoring models, logistic regression, neural networks, genetic algorithms.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

C A P Í T U L O 1 - I N T R O D U Ç Ã O

1.1 CENÁRIO

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. Com isso, 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.*

Os modelos de *credit scoring* são específicos para a aprovação em cada produto de crédito, sendo que os produtos de crédito podem ser: crédito pessoal, cheque especial, empréstimos para financiamentos, entre outros. Nesse estudo, o produto em questão será o crédito pessoal.

1.2 OBJETIVOS DO ESTUDO

1.2.1 Objetivos Gerais

Com base nos dados provenientes de uma amostra, pretende-se:

- Desenvolver três modelos de *credit scoring*, mediante o uso de três técnicas estatísticas/computacionais:

1. Regressão Logística;
2. Redes Neurais;
3. Algoritmos Genéticos;

- Comparar os modelos desenvolvidos em termos de indicadores de qualidade de ajuste e previsão;
- Propor um modelo para a classificação de clientes.

1.2.2 Objetivos Específicos

Para o alcance dos objetivos gerais, são definidos especificamente os seguintes objetivos:

- Selecionar as variáveis a serem utilizadas em cada um dos três modelos;
- Definir critérios para aferir o poder de discriminação das variáveis;
- Identificar as variáveis com maior poder de discriminação dos clientes catalogados nos grupos de bons e maus pagadores;
- Definir critérios para comparação da eficiência dos modelos;
- Comparar os resultados obtidos pelos três modelos;

- Identificar qual modelo apresentou-se como o mais indicado para a discriminação dos clientes.

1.3 JUSTIFICATIVA DO TRABALHO

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. Sabendo disto, qualquer avanço em termos de acuracidade da previsão gera ganhos financeiros para a instituição. Daí vem o interesse de analisar diferentes tipos de modelo e apontar quais apresentam uma maior precisão.

Na literatura pesquisada, principalmente no Brasil, encontram-se poucos estudos que abordam os algoritmos genéticos como ferramenta para construção de modelos de *credit scoring*. Em contrapartida, redes neurais e regressão logística são largamente empregadas neste tipo de problema. Por esta razão, julgou-se oportuno apresentar as três técnicas para utilização em um mesmo banco de dados e comparar seus aspectos positivos e negativos.

Por questão de economia de tempo e custo, este trabalho é desenvolvido por meio de dados secundários de clientes, fornecidos por um grande banco varejista brasileiro.

1.4 DELIMITAÇÃO DO TRABALHO

Nesse trabalho são construídos modelos de *credit scoring* baseados numa amostra de 20.000 clientes que obtiveram empréstimo de crédito pessoal em um grande banco de varejo que atua no mercado brasileiro. A amostra foi coletada em fevereiro de 2004 e refere-se aos empréstimos concedidos entre agosto de 2002 e fevereiro de 2003; apenas os contratos considerados bons ou

maus pela instituição foram selecionados para o trabalho; clientes cuja classificação era indeterminada não foram focalizados.

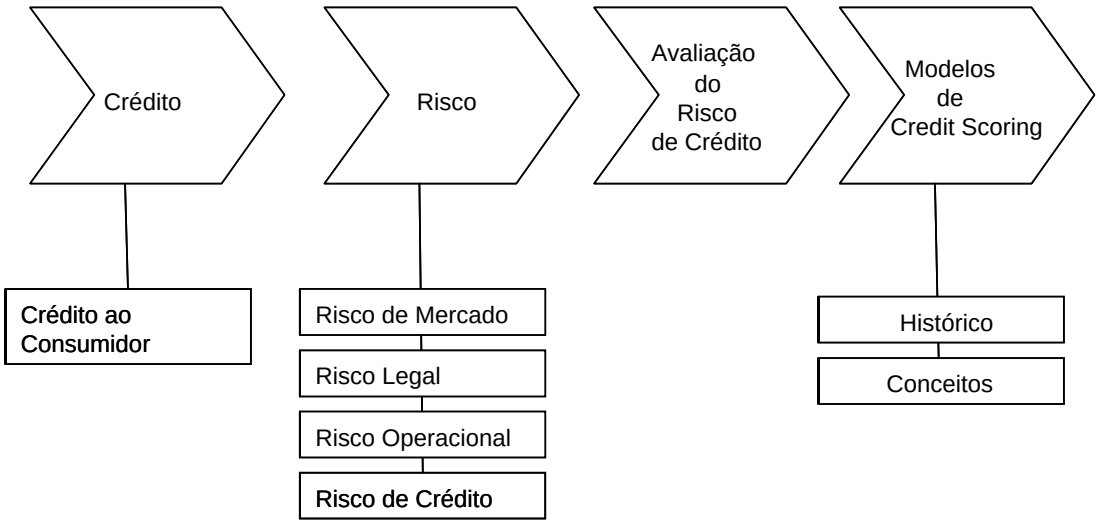
1.5 ORGANIZAÇÃO DO ESTUDO

Essa dissertação está estruturada em seis capítulos. Após este capítulo introdutório, o Capítulo 2 apresenta a fundamentação teórica, contendo os conceitos de crédito, risco e modelos de *credit scoring*. No Capítulo 3 são descritas as particularidades deste estudo, com a explicação do problema estudado. Na seqüência, o Capítulo 4 permite uma visão geral das técnicas adotadas neste estudo. O Capítulo 5 ilustra uma visão mais detalhada das técnicas e a forma como elas foram adotadas; este capítulo também aborda os resultados obtidos e a comparação entre as técnicas. Finalmente, o Capítulo 6 traz as conclusões advindas deste estudo, bem como recomendações para futuros estudos.

CAPÍTULO 2- FUNDAMENTAÇÃO TEÓRICA

Nesse capítulo serão apresentados conceitos teóricos que darão sustentação ao desenvolvimento do tema deste trabalho, os quais são associados conforme a figura 1, a seguir.

Figura 1: Encadeamento da teoria



Fonte: o Autor

2.1 CRÉDITO

Crédito, por definição, é “todo ato de vontade ou disposição de alguém de destacar ou ceder, temporariamente, parte do seu patrimônio a um terceiro, com a expectativa de que esta parcela volte a sua posse integralmente, após decorrido o tempo estipulado” (SCHRICKEL, 1995, p. 25).

Patrimônio pode ser entendido como dinheiro no caso de empréstimo monetário ou bens, empréstimo para uso ou venda com pagamento parcelado, ou a prazo.

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

Como se trata de um ato de vontade, cabe ao cedente do patrimônio a decisão de cedê-lo ou não, tendo o direito de recusar se achar conveniente.

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

2.1.1 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. Por outro lado, o crédito ao consumidor injeta recursos na economia, permitindo a produção e a expansão econômica de um país, trazendo desenvolvimento para a nação (LEWIS, 1992, p. 2).

“Nos dias atuais, crédito ao consumidor é um negócio essencial. O maior desafio desta indústria é tornar o crédito largamente disponível; assim tantas pessoas quanto possíveis terão a oportunidade de utilizar essa poderosa ferramenta” (LEWIS, 1992, p. 2). 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.

2.2 RISCO

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

2.2.1 Principais Tipos de Risco

A classificação dos riscos não obedece à norma absoluta; guarda relação com o processo de gerenciamento de risco de cada instituição. Porém, no ambiente das instituições financeiras brasileiras já é comum falar-se em risco de mercado, risco operacional e risco de crédito (BERGAMINI JR., 1997, p. 99). O risco legal ainda não é conceito bem definido; mesmo assim, será adotada a divisão dos quatro grupos: risco de mercado, risco legal, risco operacional e risco de crédito.

Como o objetivo desse estudo é a obtenção de um modelo para a previsão de risco de crédito, a apresentação desse risco será mais detalhada.

2.2.2 Risco de Mercado

> O risco de mercado, pode-se dizer, surge como conseqüência do crédito, e se potencializa pela sofisticação e complexidade dos produtos financeiros oferecidos e pela diversidade e instabilidade dos mercados de atuação, o que pode levar os bancos, em suas transações de intermediação financeira, a manterem posições passivas e ativas não coincidentes, em taxas, prazos ou moedas, levando-os a sofrer prejuízos em função deste descasamento (FIGUEIREDO, 2001, p. 10).

Este tipo de risco se apresenta de várias maneiras, tais como:

- Risco de taxa de juros: representa a possibilidade de perda financeira em função de variações das taxas de juros - flutuação das taxas de juros sobre as aplicações e captações, no mercado financeiro, em função das políticas macroeconômicas e turbulências do mercado;
- Risco de taxa de câmbio: representa a possibilidade de perda financeira em decorrência de variações na taxa de câmbio como descasamento em carteira indexada a alguma moeda estrangeira;
- Risco de liquidez: representa a possibilidade de o banco não ter condições de cumprir suas obrigações financeiras, seja por substanciais desencaixes no curto prazo, escassez de recursos ou, ainda, pela incapacidade de se desfazer, rapidamente, de uma posição, devido às condições de mercado;
- Risco de ações: possibilidade de perdas em função de mudanças no valor de mercado das ações componentes de uma carteira.

2.2.3 Risco Legal

O risco legal faz parte das exposições a riscos das instituições financeiras; porém, não existe ainda uniformização quanto a conceito e abrangência. O risco legal está relacionado a possíveis perdas quando um contrato não pode ser legalmente amparado (BERGAMINI JR., 1997, p. 98). Podem-se incluir aqui riscos de perdas por documentação insuficiente, insolvência, ilegalidade, falta de representatividade e/ou autoridade por parte de um negociador.

2.2.4 Risco Operacional

Os estudos sobre risco operacional estão em estágio inicial. O risco operacional está relacionado a possíveis perdas como resultado de sistemas e/ou controles inadequados, falhas de gerenciamento e erros humanos.

O risco operacional pode ser dividido em três grandes áreas (DUARTE JR., 1996, p. 28):

- Risco organizacional: está relacionado com uma organização ineficiente, administração inconsistente e sem objetivos de longo prazo bem definidos, fluxo de informações

internas e externas deficientes, responsabilidades mal definidas, fraudes, acesso a informações internas por parte de concorrentes;

- Risco de operações: está relacionado com problemas como *overloads* de sistemas (telefonia, elétrico, computacional etc.), processamento e armazenamento de dados passíveis de fraudes e erros, confirmações incorretas ou sem verificação criteriosa etc;

- Risco de pessoal: está relacionado a problemas como empregados não-qualificados e/ou pouco motivados, personalidade fraca, falsa ambição etc.

2.2.5 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". Mais especificamente enfocado para uma instituição financeira,

> Risco de Crédito define-se como a medida numérica da incerteza com relação ao recebimento futuro de um valor contratado (ou compromissado), a ser pago por um tomador de um empréstimo, contraparte de um contrato ou emissor de um título carregado nos estoques da instituição, descontadas as expectativas de recuperação e realização de garantias (DUARTE JR. *et al*, 1999, p. 67).

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

- Risco de inadimplência: risco do não-pagamento, por parte do tomador, de uma operação de crédito - empréstimo, financiamento, adiantamentos, operações de *leasing* - ou ainda a possibilidade de uma contraparte de um contrato ou emissor de um título não honrar seu crédito;

- Risco de degradação de garantia: risco de perdas em função das garantias oferecidas por um tomador deixarem de cobrir o valor de suas obrigações junto à instituição em função de desvalorização do bem no mercado, dilapidação do patrimônio empenhado pelo tomador;

- Risco de concentração de crédito: possibilidade de perdas em função da concentração de empréstimos e financiamentos em poucos setores da economia, classes de ativos, ou empréstimos elevados para um único cliente ou grupo econômico;

- Risco de degradação de crédito: perda pela queda na qualidade creditícia do tomador de crédito, emissor de um título ou contraparte de uma transação, ocasionando uma diminuição no valor de suas obrigações. Este risco pode acontecer em uma transação do tipo de aquisição de ações ou de títulos soberanos que podem perder valor;

- Risco soberano: risco de perdas envolvendo transações internacionais - aquisição de títulos, operações de câmbio - quando o tomador de um empréstimo ou emissor de um título não pode honrar seu compromisso por restrições do país sede.

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.
Na análise de crédito existem dois fatores cruciais a serem analisados:

1. Qual o risco que o solicitante de crédito apresenta;

2. Qual o risco máximo que a instituição pode aceitar.

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. A próxima seção focalizará esses aspectos.

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, 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 julgamento, o analista avalia a solicitação de empréstimo mediante ficha cadastral e/ou entrevista. Para este tipo de avaliação existem 4 "Cs" largamente mencionados na literatura pesquisada que devem ser considerados (SANTI FILHO, 1997; SCHRICKEL, 1995) [1]:

- Caráter: refere-se à intenção de pagar. O avaliador deve levar em consideração o cadastro do cliente, levantando informações sobre empréstimos anteriores, atuação na praça, existência de restrições;

- Capacidade: refere-se à habilidade de pagar. É considerado o aspecto mais subjetivo do risco, pois depende mais da percepção do analista do que da análise de dados cadastrais;

- Capital: refere-se ao potencial de "produzir" dinheiro. No caso de análise para pessoa física, o avaliador deve levar em consideração a renda do indivíduo e seu patrimônio para entender se ele possui meios de quitar o empréstimo;

- Condições: referem-se ao micro e macrocenário em que o tomador está inserido. Esse último aspecto foge do controle do tomador e requer a análise dos fatores externos que afetam a economia como planos de ajuste da economia, bolsas de valores em queda (ou em alta), entre outros.

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).

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

2.4 MODELOS DE *CREDIT SCORING*

2.4.1 Histórico

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

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). A busca por novas técnicas cresceu cada vez mais e métodos estatísticos que auxiliam na tomada de decisão foram introduzidos nas áreas estratégicas das empresas. 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.

Nos anos setenta, as maiores empresas de cartão de crédito, Visa e Mastercard, introduziram modelos nos seus negócios. Com isso, conseguiram diminuir suas taxas, aumentar sua carteira de clientes e tornaram-se mais competitivas. A General Motors também iniciou a utilização desta ferramenta na mesma época para o financiamento de veículos. 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. Em estudo de Matias e Siqueira (1996) sobre insolvência de bancos, há o comentário (p. 19):

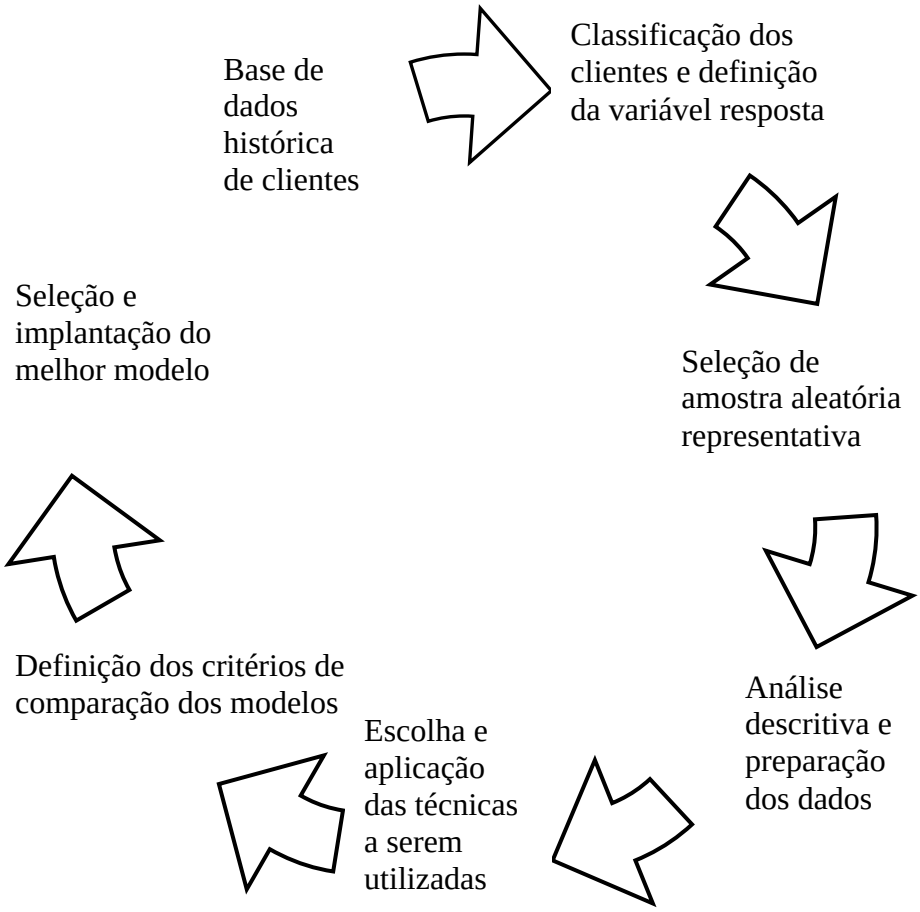
> Com a efetiva implantação do novo padrão monetário no Brasil, em 1º de julho de 1994, dentro do rígido controle de emissão monetária, os índices de inflação nominal apresentaram sensível queda – da ordem de 30% ao mês para cerca de 1,5% ao mês. Em decorrência, os ganhos do sistema bancário decorrentes de *floating* foram reduzidos de U$ 10 bilhões para menos de U$ 500 milhões, obrigando os bancos a buscarem fontes alternativas de receita. Já no segundo semestre de 1994, os bancos expandiram suas carteiras de empréstimos, principalmente para pessoas físicas, comércio e micro e pequenas empresas. Em um primeiro momento, alguns bancos não conseguiram se adaptar. Outros, com a maior concessão de crédito efetiva sem adequados padrões de crédito, deterioraram seus ativos com a crise da inadimplência e insolvência que sucedeu.

Este texto mostra quão despreparados estavam os bancos da época para a concessão de crédito. A estabilidade da moeda e o aumento dos empréstimos ao consumidor proporcionaram condições favoráveis para que modelos de análise de crédito fossem efetivamente implantados no Brasil.

2.4.2 Conceitos

Neste tópico serão explicados os passos para a construção de um modelo de *credit scoring*. A figura 2 sintetiza estes passos.

Figura 2: Ciclo de desenvolvimento de um modelo



F o n t e : o A u t o r

Como já mencionado, a função de um modelo de *credit scoring* é mensurar o risco, sendo, portanto, uma ferramenta que auxilia na decisão de fornecer ou não o crédito para o solicitante. Informações cadastrais, bem como comportamentos anteriores dos clientes, são levados em

consideração para a modelagem e o resultado do modelo é identificar perfis de clientes que sejam atrativos para a empresa conceder o crédito.
Existem alguns 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

Neste ponto são definidos quais são os clientes considerados bons e quais os clientes considerados maus pela instituição. Cabe observar que cada instituição tem sua própria política de crédito e estes conceitos de bons e maus podem mudar dependendo da instituição. Na realidade, nessa classificação, além de clientes bons e maus, também existem os clientes excluídos, aqueles que possuem características peculiares e que não devem ser considerados (por exemplo, trabalha na instituição) e os clientes indeterminados, 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

Com a base e a variável resposta definidas, selecionam-se amostras representativas de clientes bons e maus. É importante que as amostras de bons e maus clientes tenham o mesmo tamanho para se evitar qualquer 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, outra para validação do modelo e a terceira para teste do modelo. No capítulo 3, seção 3.3, serão detalhadas as funções de cada uma das três amostras.

4. Análise descritiva e preparação dos dados

Consiste em analisar segundo critérios estatísticos cada variável a ser utilizada no modelo. Este tópico será abordado mais detalhadamente posteriormente.

5. Escolha e aplicação das técnicas a serem utilizadas para a construção do modelo

Existem diversas técnicas utilizadas para construção de modelos, algumas com maior ou menor complexidade. Neste trabalho serão utilizadas Regressão Logística, Redes Neurais e Algoritmos Genéticos. 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 existe um método claramente melhor que os demais, tudo dependendo de como a técnica escolhida 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). Estes critérios serão explicados no capítulo 4, seção 4.4.

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.

CAPÍTULO 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. A instituição gostaria que todos os clientes fossem classificados como bons ou maus pagadores, para poder estimar a distribuição de perdas de sua carteira de crédito, obter um *credit rating* e direcionar o gerenciamento das operações de acordo com o risco de inadimplência dos contratantes. Para viabilizar este 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. 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.

O crédito pessoal é uma operação de crédito ao consumidor rápida e prática. 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.

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 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. A partir desta estrutura pode-se acompanhar o andamento do contrato e precisar em que momento o cliente deixou de pagar uma ou mais parcelas.

No universo da instituição estudada, a proporção de bons contra maus é de 85% *versus* 15%; neste trabalho, optou-se pela alternativa de uma amostra igualitária, por se acreditar que desta forma a avaliação da qualidade do ajuste é mais precisa, evitando-se o problema de acertos de classificação a *posteriori* automáticos no grupo majoritário, independentemente do poder de aderência do modelo aos dados. Outra alternativa seria extrair uma amostra aleatória do universo e posteriormente ponderar os pesos de bons e maus de acordo com sua proporção na amostra; esta segunda alternativa é utilizada em Rosa (2000).

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 tem como função 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*)[2] do modelo. Nos modelos de regressão logística e algoritmos genéticos

[2] Superajuste ou *overfitting* é um fenômeno presente nas redes neurais quando o modelo fica "superajustado" aos dados de desenvolvimento; entretanto, o modelo não será bom em outros dados. A amostra de validação é uma solução para se evitar o superajuste. Ohtoshi (2003, p. 47) explica: "Quando o treinamento progride, o erro no treinamento naturalmente cai e a função de erro diminui. De fato, se o erro na amostra de validação pára de cair, isto indica que a rede está começando a iniciar um superajuste. Quando o superajuste ocorre na amostra de treinamento, é aconselhável diminuir o número de camadas escondidas ou de unidades da rede".

a amostra de validação terá o mesmo papel da amostra de teste, ou seja, avaliar a predição do modelo.

3.4 AS VARIÁVEIS

As variáveis explanatórias disponíveis contêm características que podem ser 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. A tabela 1 descreve as variáveis e suas respectivas escalas.

Tabela 1: Variáveis disponibilizadas para este estudo

Variável	**Escala**
Sexo	Nominal
Estado Civil	Nominal
Fone Residencial	Nominal
Fone Comercial	Nominal
Tempo no Emprego Atual	Razão
Salário do Cliente	Razão
Quantidade de Parcelas a Serem Quitadas	Razão
Primeira Aquisição	Nominal
Tempo na Residência Atual	Razão
Valor da Parcela	Razão
Valor Total do Empréstimo	Razão
Tipo de Crédito	Nominal
Idade	Razão
CEP Residencial	Nominal
CEP Comercial	Nominal
Código de Profissão	Nominal
Nome da Profissão	Nominal
Salário do Cônjuge	Razão
Tipo de Cliente - Bom (máximo 20 dias de atraso) ou Mau (acima de 60 dias de atraso)	Nominal

3.5 DEFINIÇÃO DA VARIÁVEL RESPOSTA

Para o desenvolvimento de um modelo de c*redit scoring* é preciso definir, num primeiro momento, o que a instituição financeira considera como um bom e mau pagador. Esta definição, da Variável Resposta, também denominada de Definição de *Performance*, está diretamente ligada à política de crédito da instituição. 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. A mensuração do atraso é calculada por meio da parcela paga com maior atraso pelo cliente; por exemplo, um cliente que atrasou três parcelas por vinte dias consecutivos ainda assim é considerado um bom cliente, ao passo que um cliente que tenha atrasado uma parcela por sessenta dias é considerado mau.

Os clientes que apresentam atrasos no intervalo entre bons e maus foram definidos como indeterminados.

Pode-se destacar a existência de um grupo de clientes que não faz parte do estudo, pois as informações relativas a ele não são armazenadas pela instituição. Trata-se do grupo de clientes recusados pela instituição antes mesmo de terem suas propostas cadastradas.

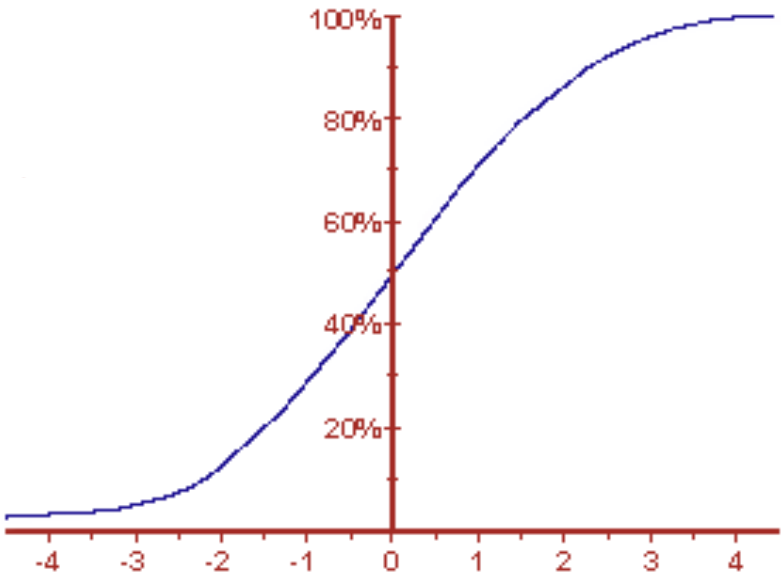
Vale ressaltar que a Definição de *Performance* pode variar de uma instituição para outra.

Da Definição de *Performance* resultam quatro classificações: bons, maus, indefinidos e recusados. No entanto, somente duas delas, Bons e Maus, são utilizadas para a construção da variável resposta, pois 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 separadamente pelo analista de crédito com base em análise qualitativa (Capítulo 2, seção 2.3); a decisão de aceitar ou não estes clientes depende da política mais ou menos conservadora adotada pela instituição financeira.

CAPÍTULO 4- TÉCNICAS UTILIZADAS

4.1 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, pois não pressupõe que os dados de entrada tenham distribuição Normal, embora seja desejável que as variáveis tenham essa distribuição (HAIR *et al*, 1998, p. 231). A regressão logística prediz a probabilidade de um evento ocorrer, a qual pode estar entre 0 e 1. A relação entre as variáveis independentes e a variável dependente se assemelha a uma curva em forma de S conforme ilustra a figura 3, a seguir.



Fonte: Adaptado pelo autor de SHARMA (1996, p. 320)

4.1.1 Histórico

Segundo Lima (2002, p. 77), a função logística surgiu em 1845, ligada a problemas de crescimento demográfico, problemas em que, até os dias de hoje, essa função é utilizada. Na década de 30, esta metodologia passou a ser aplicada no âmbito da biologia, e posteriormente nas áreas relacionadas a problemas econômicos e sociais. Paula (2002, p. 118) aponta que, apesar de o modelo de regressão logística ser conhecido desde os anos 50, foi devido a trabalhos do estatístico David Cox, na década de 70, que esta técnica tornou-se bastante popular entre os usuários de Estatística.

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

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

4.1.2 Conceitos

Nos modelos de regressão logística, a variável dependente é, em geral, uma variável binária[3] (nominal ou ordinal) e as variáveis independentes podem ser categóricas (desde que dicotomizadas após transformação) ou contínuas.

[3] Na maioria dos casos apresentados na literatura estudada, a regressão logística é apresentada com variável resposta binária. Entretanto, há o caso em que a variável resposta é múltipla, ou seja, com mais de duas categorias (Desai *et al*, 1997); inclusive, alguns *softwares* como o SPSS v.12.0 apresentam a opção de utilização de variável resposta múltipla.

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

A variável dependente binária Y pode assumir os valores:

$$Yi = \begin{cases} 1 & \text{Se o i-ésimo indivíduo pertence à categoria dos bons} \\ 0 & \text{Se o i-ésimo indivíduo pertence à categoria dos maus} \end{cases}$$

E 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 , 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):

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

4.1.2.1 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).

4.1.3 Pontos Fortes e Fracos da Aplicação de Regressão Logística

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

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

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

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

4.2 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.

Segundo Haykin (1999, p. 28):

> Uma rede neural é um processador maciçamente paralelamente (sic) distribuído constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para uso. Ela se assemelha ao cérebro em dois aspectos: 1) O conhecimento é adquirido pela rede por meio de um processo de aprendizagem; 2) Forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.

4.2.1 Histórico

Segundo vários autores, entre eles, Marks e Schnabl (1997, p. 3); Haykin (1999, p. 63) e Fausett (1994, p. 22), o primeiro modelo de rede neural surgiu com o trabalho de McCulloch e Pitts. Warren McCulloch foi um psiquiatra e neuroanatomista que estudava uma representação para o sistema nervoso. Em 1942, ele se associou com o matemático Walter Pitts e no ano seguinte eles publicaram um artigo que propunha um modelo matemático para uma rede neural, artigo este que até hoje é uma referência no estudo de redes neurais (HAYKIN, 1999, p. 63). Um segundo trabalho importante foi publicado por Hebb em 1949, no qual foram propostas as primeiras regras de aprendizado para redes neurais artificiais; este trabalho também inspirou muitos estudiosos em pesquisas posteriores.

Durante as décadas de 50 e 60 houve muitas pesquisas e estudos que permitiram avançar muito no campo das redes neurais. Fausett (1994, p. 23) chama este período de “anos dourados das redes neurais”. Estudos mostraram que a nova metodologia seria muito promissora; foram

propostos novos tipos de rede, novas regras de aprendizado e as redes foram ficando mais complexas.

Na década de 70, contudo, houve uma desaceleração nas pesquisas, conforme apontam Hair *et al* (1998, p. 545): "(...) no final dos anos 1960, pesquisas demonstraram que as redes neurais daquela época eram realmente muito limitadas e a área em si sofreu um geral retrocesso".

Foi somente nos anos 80 que, com o maior poder computacional, as redes neurais voltaram a ser largamente estudadas e aplicadas. Fausett (1994, p. 25) destaca o desenvolvimento do algoritmo *backpropagation* (retropropagação) como um divisor de águas para a popularidade das redes neurais. Até os dias atuais as redes neurais vêm sendo largamente empregadas e estudadas, sendo utilizadas em diferentes áreas de conhecimento como medicina, biologia, economia, administração e engenharia.

4.2.2 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):

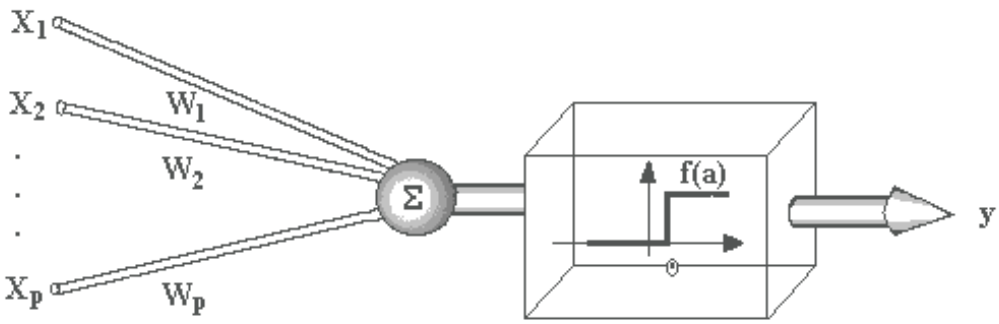
1. O processamento das informações ocorre dentro dos chamados neurônios;
2. Os estímulos são transmitidos pelos neurônios por meio de conexões;
3. Cada conexão tem associada a si um peso, que, numa rede neural padrão, multiplica-se ao estímulo recebido;
4. Cada neurônio contribui para a função de ativação (geralmente não linear) para determinar o estímulo de saída (resposta da rede).

O mencionado modelo pioneiro de McCulloch e Pitts de 1943 (figura 4), 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.

Figura 4: O modelo de McCullock e Pitts



Fonte: Tatibana e Kaetsu (S.d.)

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* $a < k$

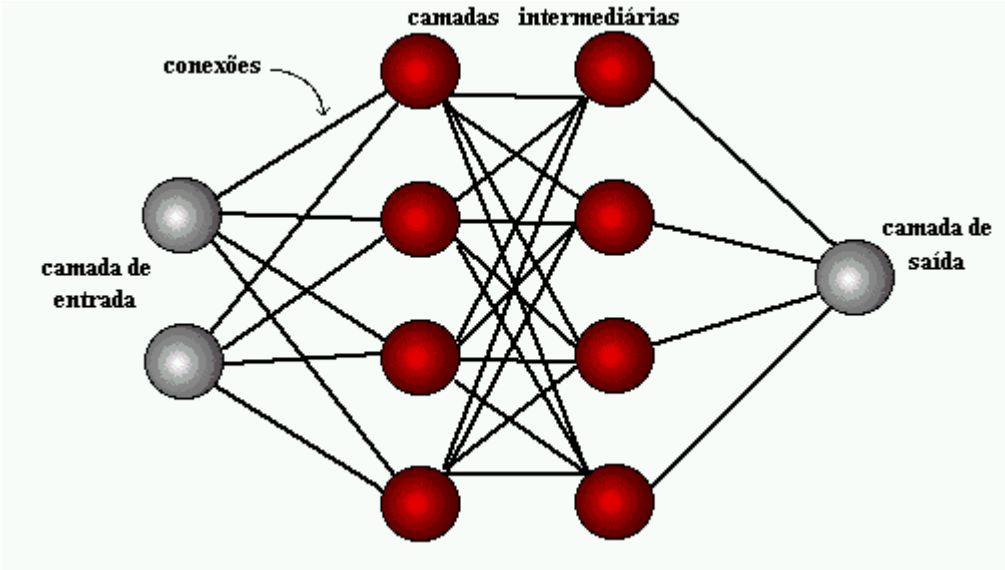
Na definição de um modelo de redes neurais três características devem ser observadas: a forma que a rede tem, chamada arquitetura; o método para determinação dos pesos, chamado algoritmo de aprendizado; e a função de ativação. Os próximos tópicos explicarão estas características.

4.2.2.1 Arquitetura

Como já mencionado, arquitetura refere-se ao formato da rede. Toda rede é dividida em camadas, usualmente classificadas em três grupos (conforme ilustra a figura 5, a seguir):

- 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.

Figura 5: Exemplo de uma rede neural

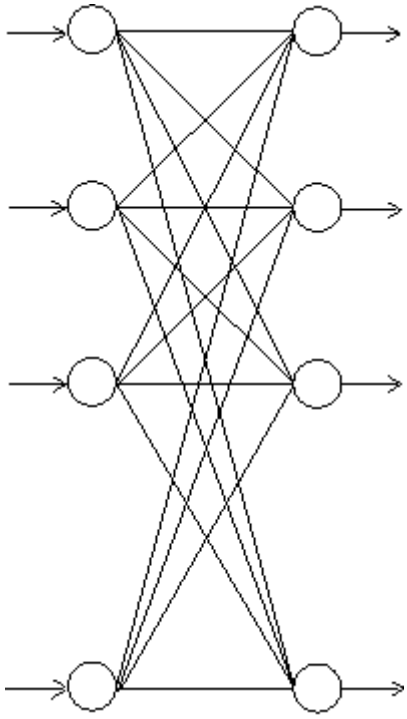


Fonte: Carvalho (S.d.)

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.

1. 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, como mostra a figura 6, a seguir. Algumas das redes que utilizam essa arquitetura são: Rede de Hebb, *perceptron*, ADALINE, entre outras.

Figura 6: Rede *Feedforward* com uma única camada

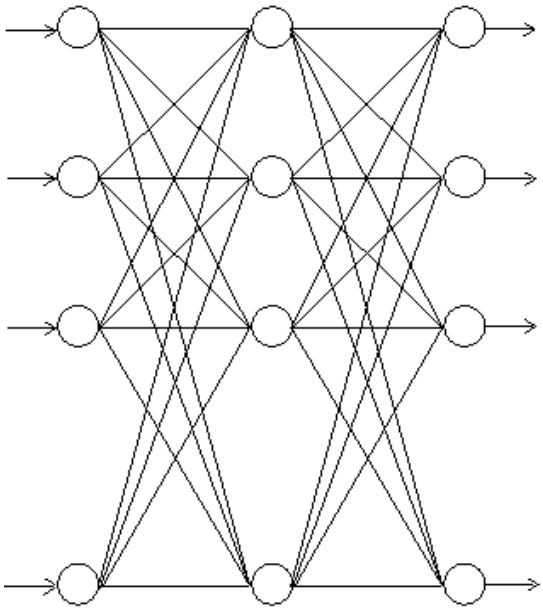


Camada de Entrada Camada de Saída

Fonte: Adaptado pelo autor de FAUSETT (1994, p. 13)

2. 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. A figura 7, a seguir, ajuda a entender melhor este conceito.

Figura 7: Rede *Feedforward* com múltiplas camadas

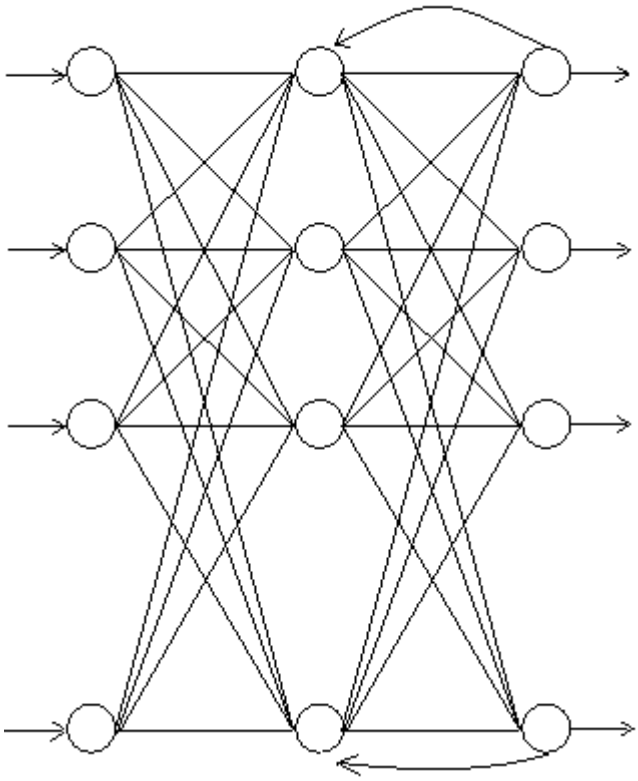


Camada de Entrada Camada Intermediária Camada de Saída

Fonte: Adaptado pelo autor de FAUSETT (1994, p. 13)

3. Redes Recorrentes: neste tipo de rede, a camada de saída possui ao menos uma ligação que realimenta a rede, como mostra a figura 8. As redes chamadas de BAM (*Bidirecional Associative Memory*) e ART1 e ART2 (*Adaptative Resonance Theory*) são redes recorrentes.

Figura 8: Rede Recorrente



Camada de Entrada Camada Intermediária Camada de Saída

Fonte: Adaptado pelo autor de HAYKIN (1999, p. 49)

4.2.2.2 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. Na seção 5.3.1 será detalhado qual algoritmo será utilizado neste trabalho, bem como as razões que levaram a esta escolha.

4.2.2.3 Funções de Ativação

Como já mencionado, 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: $f(x) = \begin{cases} 1 & Se\ x < k \\ 0 & Se\ x \geq k \end{cases}$

- Função Logística: $f(x) = \frac{1}{1+e^{(-\alpha x)}}$

- Função Tangente Hiperbólica: $f(x) = tanh(x)$

4.2.3 Pontos Fortes e Fracos das Redes Neurais

Berry e Linoff (1997, p. 331) apontam os seguintes pontos positivos na utilização de redes neurais:

- São versáteis: redes neurais podem ser usadas para a solução de diferentes tipos de problemas como previsão, agrupamento ou identificação de padrões;

- São capazes de identificar relações não-lineares entre as variáveis;

- São largamente utilizadas, estando disponíveis em vários *softwares.*

No tocante às desvantagens, os autores apontam (p. 333):

- Os resultados não são explicáveis: não são produzidas regras explícitas, a análise é feita dentro da rede e só o resultado é fornecido pela “caixa-preta”;

- A rede pode convergir para uma solução inferior: não há garantias de que a rede encontre a melhor solução possível; ela pode convergir para um máximo local[4].

4.3 ALGORITMOS GENÉTICOS

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

A idéia dos algoritmos genéticos se assemelha à evolução das espécies proposta por Darwin: os algoritmos vão evoluindo com o passar das gerações e os candidatos à solução do problema que se quer resolver “permanecem vivos” e se reproduzem (BACK *et al*, 1996).

4.3.1 Histórico

Bauer (1994, p. 11) assinala que no final dos anos 50 e começo dos anos 60 muitos biólogos começaram a experimentar simulações computacionais de sistemas genéticos. Particularmente importante foi o trabalho de Fraser de 1960 que iniciou o desenvolvimento mais profundo dos algoritmos genéticos.

Entretanto, foi John Holland quem começou a desenvolver as primeiras pesquisas no tema. Holland foi gradualmente refinando suas idéias e em 1975 publicou o seu livro *Adaptation in Natural and Artificial Systems*, hoje considerado a Bíblia de algoritmos genéticos. Desde então, estes algoritmos vêm sendo aplicados com sucesso nos mais diversos problemas de otimização e

[4] Nesta dissertação foi adotada uma amostra de validação para evitar este tipo de problema.

aprendizado de máquina. Nos anos 80, a aplicação do modelo de algoritmo genético de Holland por Axelrod (1987) ao dilema dos prisioneiros[5] popularizou ainda mais o uso desta técnica.

4.3.2 Conceitos

Segundo Picinini *et al* (2003, p. 464):

> Algoritmos evolutivos são métodos computacionais que permitem obter soluções em problemas para os quais não existem algoritmos exatos para solucioná-los, ou, se existem, a obtenção da solução requer elevado tempo de processamento. O algoritmo evolutivo mais conhecido é o algoritmo genético proposto por Holland.

O algoritmo é composto por uma população, que é representada por cromossomos, que nada mais são do que diversas soluções possíveis para o problema proposto. As soluções que são selecionadas para dar forma a soluções novas (a partir de um cruzamento) são selecionadas de acordo com a aptidão (*fitness*) dos cromossomos pais. Assim, quanto mais apropriado é o cromossomo, maior a possibilidade de ele se reproduzir. Esse processo é repetido até que a regra de parada seja satisfeita, ou seja, encontrar uma solução muito próxima da desejada.

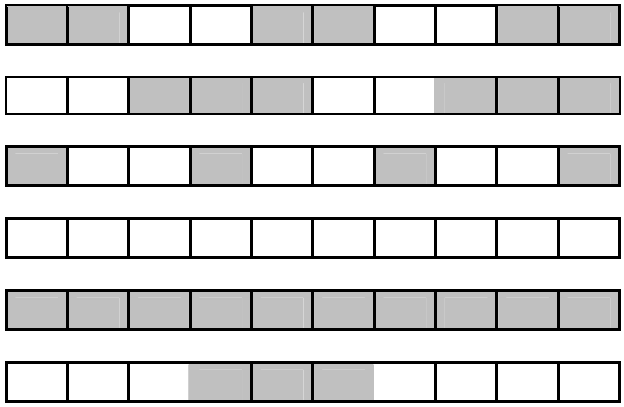
4.3.2.1 Fases de um algoritmo genético

Todo algoritmo genético passa pelas seguintes fases:

Início: primeiramente é gerada uma população formada por um conjunto aleatório de indivíduos (cromossomos) que podem ser vistos como possíveis soluções do problema, conforme a figura 9.

[5] O Dilema dos Prisioneiros descreve a situação em que dois prisioneiros estão presos em salas separadas, após cometerem um crime em que foram cúmplices. Como a polícia não tem provas suficientes para incriminá-los, é feita uma solicitação de confissão para cada um deles. Se ambos confessarem (ou colaborarem com a polícia), cada um será condenado a 5 anos de prisão. Se nenhum confessar, o julgamento será dificultado e eles provavelmente serão condenados a 2 anos de prisão. Por outro lado, se um dos prisioneiros confessar o crime, mas o outro não, aquele que confessou será condenado a apenas 1 ano de prisão, enquanto o outro será condenado a 10 anos. O dilema está em confessar ou não.

Figura 9: Cromossomos gerados aleatoriamente



Fonte: Adaptado pelo autor de LINGRAS (2001, p. 217)

Função de Aptidão (*Fitness*): uma função de aptidão é definida para avaliar a "qualidade" de cada um dos cromossomos.

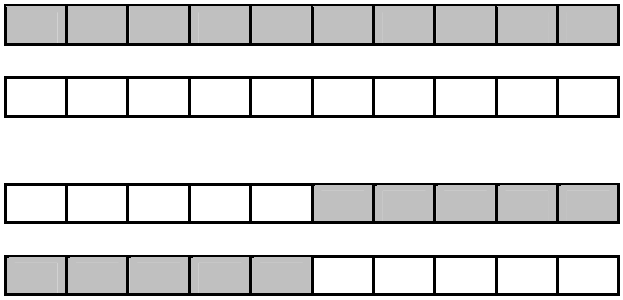
Seleção: de acordo com os resultados da função de aptidão, uma porcentagem dos mais adaptados é mantida, enquanto os outros são descartados (Darwinismo). A figura 10 ilustra esta fase.

Figura 10: Seleção dos Melhores

Fonte: Adaptado pelo autor de LINGRAS (2001, p. 217)

Cruzamento (*Cross-Over*): escolhem-se dois pais e baseando-se neles é gerado um "filho" baseado num critério específico de cruzamento. O mesmo critério é efetuado com o outro cromossomo, e o material dos dois cromossomos é trocado. Se nenhum cruzamento for executado, a prole é uma cópia exata dos pais. A figura 11 corresponde a esta fase.

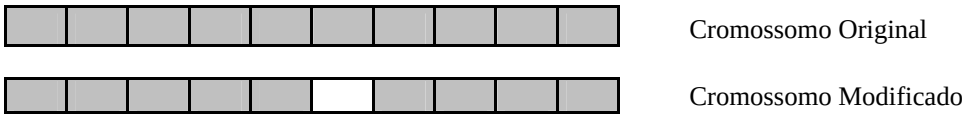
Figura 11: Cruzamento



Fonte: LINGRAS (2001, p. 217)

Mutação: atribui-se à população um percentual de mutação. A mutação é a alteração de algum dos genes do cromossomo (figura 12). O intuito da mutação é evitar que a população convirja para um máximo local. Assim, caso esta convergência ocorra, a mutação garante que a população irá "saltar" o ponto de mínimo local, tentando alcançar outros pontos de máximo.

Figura 12: Mutação



Fonte: LINGRAS (2001, p. 217)

Verificação do critério de parada: criada uma nova geração, verifica-se o critério de parada pré-estabelecido e retorna-se para a fase da função de aptidão, caso este critério não esteja satisfeito.

4.3.3 Pontos Fortes e Fracos dos Algoritmos Genéticos

Destacam-se os seguintes pontos positivos na utilização de algoritmos genéticos:

- Produzem resultados explicáveis diferentemente das redes neurais (BERRY; LINOFF, 1997, p. 357);
- São facilmente utilizáveis (BERRY; LINOFF, 1997, p. 357);

- Podem trabalhar com um grande conjunto de dados e variáveis (FENSTERSTOCK, 2005, p. 48).

Algumas das desvantagens apontadas na literatura são:

- Ainda são pouco utilizados para problemas de avaliação do risco de crédito (FENSTERSTOCK, 2005, p. 48);
- Necessitam de um grande esforço computacional (BERRY; LINOFF, 1997, p. 358);
- Estão disponíveis em poucos *softwares* (BERRY; LINOFF, 1997, p. 358).

4.4 CRITÉRIOS DE AVALIAÇÃO DE PERFORMANCE

Os critérios de avaliação de performance indicam quão adequado um modelo é. 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). Além das amostras, existem outros critérios que serão utilizados, apresentados nos tópicos seguintes.

4.4.1 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.

$$Tat = \frac{At}{N}$$

Tat...Taxa de acertos total

At...Indivíduos corretamente classificados

N...Número total de clientes

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

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

Tab...Taxa de acertos de clientes bons

Ab...Indivíduos bons corretamente classificados

Nb...Número total de clientes bons

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

Tam...Taxa de acertos de clientes maus

Am...Indivíduos maus corretamente classificados

Nm...Número total de clientes maus

Entretanto, existem casos de grandes diferenças entre a taxa de acerto de bons e maus clientes que podem distorcer a qualidade do modelo. Supondo-se que um modelo, aplicado a uma base com a mesma quantidade de bons e maus clientes, classificasse todos os clientes como bons, seria obtida uma taxa de acerto de 100% para os clientes bons e 0% para os clientes maus, perfazendo um total de 50%.

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. 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.

*Ia= Tab*Tam*

Ia...Indicador de acertos

Tab...Taxa de acertos de clientes bons

Tam...Taxa de acertos de clientes maus

4.4.2 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.

Ambas as populações são divididas em intervalos iguais e para cada um é determinada a freqüência acumulada. Em cada intervalo calcula-se a diferença entre as freqüências acumuladas e o teste se dá focando a maior diferença entre elas. Matematicamente:

seja S_{n_1} (X) a função acumulada para a primeira amostra,

isto é, $S_{n_1}(X) = \frac{k}{n_1}$, onde

k= número de escores não superiores a X,

e seja S_{n_2} (X) a função acumulada para a segunda amostra,

isto é, $S_{n_2}(X) = \frac{k}{n_2}$, onde

k= número de escores não superiores a X.

A prova de Kolmogorov-Smirnov focaliza

$D = máx\,[S_{n_1}(X) - S_{n_2}(X)]$.

O exemplo apresentado na tabela 2, a seguir, foi adaptado de Lewis (1992, p. 144); o KS deste modelo hipotético é de 28%.

Tabela 2: Exemplo de cálculo no teste de Kolmogorov-Smirnov

	Número de clientes		Freqüência 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 pelo autor de LEWIS (1992, p. 144)

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). 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 àquele que possuir o maior valor no teste, pois este resultado indica uma separação maior entre bons e maus.

CAPÍTULO 5- APLICAÇÃO

Neste capítulo serão abordados os métodos de tratamento das variáveis, a aplicação das três 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; para o algoritmo genético foi utilizado um programa desenvolvido pelo autor em *Visual Basic*.

5.1 TRATAMENTO DAS VARIÁVEIS

Nesta seção são apresentados métodos de transformação de variáveis a serem utilizados nos três modelos. Quando determinada transformação for específica para alguma técnica, esta será explicitada.

Inicialmente, as variáveis quantitativas foram categorizadas. Rosa (2000, p. 14-15) aponta os ganhos obtidos com a categorização:

- Padronização dos Resultados: com a categorização das variáveis, os modelos tornam-se mais fáceis de serem implementados e fica mais simples a interpretação dos pesos relativos às categorias das variáveis;

- Estabilidade do Modelo: categorizando as variáveis quantitativas, o modelo fica menos suscetível a *outliers* (valores discrepantes). Com isso, a estabilidade do modelo é melhorada;

- Transformação das Variáveis: em estudos estatísticos, a transformação de uma variável é necessária por duas razões: quando uma variável independente quantitativa não apresenta relação linear com a variável resposta, ou na tentativa de obter a distribuição normal da variável (requisito desejável para a regressão logística, mas prescindível,

conforme já mencionado no capítulo 4, seção 4.1). Na prática, porém, algumas transformações podem ser de difícil compreensão. Por exemplo, torna-se complicado interpretar a relação entre bom ou mau pagador com a raiz quadrada da idade do cliente ou o logaritmo do salário. A categorização das variáveis quantitativas, por outro lado, é uma transformação de fácil compreensão e permite o agrupamento de variáveis de mesmo comportamento frente à variável resposta. Neste trabalho, como nenhuma das técnicas requer fortemente a condição da normalidade, não será utilizada nenhuma transformação com este intuito, apenas a categorização.

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 (TIPO). 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), conforme mostra a tabela 3, a seguir.

Tabela 3: Exemplo de cálculo do risco relativo

Variável	Número de Bons	Número de Maus	% Bons	% Maus	RR= %Bons / %Ruins
Decil1	b1	r1	b1/Tb	R1/Tr	(b1/Tb)/(r1/Tr)
Decil2	b2	r2	b2/Tb	R2/Tr	(b2/Tb)/(r2/Tr)
Decil3	b3	r3	b3/Tb	R3/Tr	(b3/Tb)/(r3/Tr)
Decil4	b4	r4	b4/Tb	R4/Tr	(b4/Tb)/(r4/Tr)
Decil5	b5	r5	b5/Tb	R5/Tr	(b5/Tb)/(r5/Tr)
Decil6	b6	r6	b6/Tb	R6/Tr	(b6/Tb)/(r6/Tr)
Decil7	b7	r7	b7/Tb	R7/Tr	(b7/Tb)/(r7/Tr)
Decil8	b8	r8	b8/Tb	R8/Tr	(b8/Tb)/(r8/Tr)
Decil9	b9	r9	b9/Tb	R9/Tr	(b9/Tb)/(r9/Tr)
Decil10	b10	r10	b10/Tb	r10/Tr	(b10/Tb)/(r10/Tr)
Total	Tb	Tr	1	1	1

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. O primeiro é evitar categorias com um número muito pequeno de observações, o que pode levar a estimativas pouco robustas dos parâmetros associados a elas. O segundo é a eliminação de parâmetros do modelo; se duas categorias apresentam risco próximo, é razoável agrupá-las numa única classe.

O RR, além de auxiliar no agrupamento das categorias, ajuda a entender se a categoria em questão está mais ligada a clientes bons ou ruins. Quando o resultado é muito acima de 1, significa que essa característica está mais ligada ao perfil de bom cliente; da mesma forma, para o resultado menor que 1 interpreta-se que a característica está relacionada aos maus clientes. No caso de a razão ser exatamente igual a 1, conclui-se que essa característica não discrimina bons e maus clientes. Esse método de agrupamento de categorias é explicado por Hand e Henley (1997, p. 527).

Ao trabalhar-se com as variáveis disponibilizadas, citadas no capítulo 3, 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[6]; em seguida, foi calculado o risco relativo de cada faixa (conforme tabela 3) e posteriormente houve o reagrupamento de acordo com risco relativo

[6] De acordo com o site dos correios, http://www.correios.com.br/servicos/cep/cep_estrutura.cfm, os cinco primeiros dígitos significam respectivamente Região, Sub-região, Setor, Sub-setor, Divisor de Sub-setor e os três últimos são Identificadores de Distribuição. Neste trabalho estão sendo utilizados os três primeiros dígitos, ou seja, região que, em geral, identifica o estado (ou grupo de estados); sub-região que, em geral, identifica o município (ou grupo de municípios) e setor.

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 descartada da análise por conter muitos dados faltantes (*missings*);

- Foram criadas duas novas variáveis, percentual do valor do empréstimo sobre o salário e percentual do valor da parcela sobre o salário. Ambas variáveis quantitativas, escala razão que foram categorizadas em faixas da mesma forma que as demais.

Após se aplicar esse método, obtiveram-se as categorias apresentadas na tabela 4. O cálculo do RR está apresentado no Apêndice A.

Tabela 4: Variáveis Categorizadas

Variável	Categoria	Nome da variável
Sexo	Masculino	V_SEXO_M
	Feminino	V_SEXO_F
Estado Civil	Casado	V_EST_C
	Solteiro	V_EST_S
	Outros	V_EST_O
Posse de Fone Residencial	Sim	V_FN_R_S
	Não	V_FN_R_N
Posse de Fone Comercial	Sim	V_FN_C_S
	Não	V_FN_C_N
Tempo no Emprego Atual	Até 24 meses	V_TP_E1
	De 25 a 72 meses	V_TP_E2
	De 73 a 127 meses	V_TP_E3
	Acima de 127 meses	V_TP_E4
Salário do Cliente	Até 650 reais	V_SAL_F1
	Acima de 650 a 950 reais	V_SAL_F2
	Acima de 950 a 1575 reais	V_SAL_F3
	Acima de 1575 a 2015 reais	V_SAL_F4
	Acima de 2015 a 3000 reais	V_SAL_F5
	Acima de 3000 reais	V_SAL_F6
Quantidade de Parcelas	Até 4	V_Q_PC_1
	5 ou 6	V_Q_PC_2
	7 a 9	V_Q_PC_3
	10 a 12	V_Q_PC_4
Primeira Aquisição	Sim	V_PR_AQ_S
	Não	V_PR_AQ_N
Tempo na Residência Atual	Até 12 meses	V_TP_R1
	De 13 a 24 meses	V_TP_R2
	De 25 a 120 meses	V_TP_R3
	Acima de 120 meses	V_TP_R4
Valor da Parcela	Até 125 reais	V_VL_PR1
	Acima de 125 a 160 reais	V_VL_PR2
	Acima de 160 a 260 reais	V_VL_PR3
	Acima de 260 reais	V_VL_PR4
Valor Total do Empréstimo	Até 300 reais	V_VL_EM1
	Acima de 300 a 400 reais	V_VL_EM2
	Acima de 400 a 500 reais	V_VL_EM3
	Acima de 500 a 800 reais	V_VL_EM4
	Acima de 800 a 1800 reais	V_VL_EM5
	Acima de 1800 reais	V_VL_EM6
Tipo de Crédito	Carnê	V_CRE_CN
	Cheque	V_CRE_CH
Idade	Até 25 anos	V_IDADE1
	De 26 a 40 anos	V_IDADE2
	De 41 a 58 anos	V_IDADE3
	Acima de 58 anos	V_IDADE4
Faixa de CEP Residencial	Faixa 1	V_CEP_F1
	Faixa 2	V_CEP_F2
	Faixa 3	V_CEP_F3
	Faixa 4	V_CEP_F4
	Faixa 5	V_CEP_F5
Faixa de CEP Comercial	Faixa 1	V_CEC_F1
	Faixa 2	V_CEC_F2
	Faixa 3	V_CEC_F3
	Faixa 4	V_CEC_F4
	Faixa 5	V_CEC_F5
Código de Profissão	Código 1	V_COD_P1
	Código 2	V_COD_P2
	Código 3	V_COD_P3
	Código 4	V_COD_P4
	Código 5	V_COD_P7
	Código 6	V_COD_P8
	Código 7	V_COD_P9
% Valor da Parcela / Salário	Até 10%	V_FX__P1
	Acima de 10 a 13,5%	V_FX__P2
	Acima de 13,5 a 16,5%	V_FX__P3
	Acima de 16,5 a 22,5%	V_FX__P4
	Acima de 22,5%	V_FX__P5
% Valor do Empréstimo / Salário	Até 28%	V_FX__E1
	Acima de 28 a 47,5%	V_FX__E2
	Acima de 47,5 a 65%	V_FX__E3
	Acima de 65%	V_FX__E4
Tipo de Cliente	Bom=1 Mau=0	TIPO

5.2 REGRESSÃO LOGÍSTICA

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

5.2.1 Modelo Implementado

Inicialmente, é interessante avaliar a relação logística entre cada variável independente e a variável dependente TIPO. As variáveis independentes focalizadas neste trabalho foram codificadas na forma de variáveis *dummies*. Para cada variável o número de categorias (k) determinou o número de variáveis *dummies* (k-1) incluídas no processamento da regressão logística.

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*.

Foram processados dois modelos *forward stepwise*: teste da razão de verossimilhança (LR - *likelihood-ratio test*) e a estatística de probabilidade condicional de máxima verossimilhança (COND - *conditional statistic*). Em modelos *forward stepwise* inicia-se apenas com o termo da constante, exceto quando se omite este parâmetro na especificação da modelagem, e em cada passo é introduzida a variável com o menor nível de significância para o escore estatístico, desde que este seja menor do que um valor de remoção (*cutoff*), definido como 0,05 neste trabalho. O processo continua até que nenhuma variável seja mais elegível para ser incluída e/ou haja convergência na comparação de estatísticas de qualidade da estimação em duas iterações sucessivas. Em ambos os métodos selecionados para processamento, a estatística de referência é a função de verossimilhança definida como a probabilidade de obter os resultados da amostra, dadas as estimativas dos parâmetros do modelo logístico. Como essa probabilidade é um valor

menor do que 1, convencionou-se usar a expressão –2LL (-2 multiplicado pelo logaritmo decimal da probabilidade – em inglês, *likelihood*). Assim, o resultado –2LL é uma medida da qualidade de ajuste do modelo estimado aos dados. Quanto menor o valor de -2LL, maior a qualidade do ajuste.

A tabela 5, com valores hipotéticos, ilustra a relação entre L e –2LL.

Tabela 5: Estatística –2LL

L	LL = log L	-2LL
1	0	0
0,7	-0,155	0,310
0,4	-0,398	0,796

Ambos os métodos verificam a mudança em –2LL assumida pelos modelos reduzido (só com uma constante incluída) e aquele com a consideração das variáveis já incorporadas. O método COND é computacionalmente menos intensivo por não requerer que o modelo seja reestimado sem cada uma das variáveis.

Foram realizadas duas simulações, uma para cada método. As variáveis foram selecionadas em cada passo, segundo estatísticas de escores. Idênticos resultados foram encontrados apesar de o modelo LR consumir tempo consideravelmente maior de processamento.

Das 53 variáveis independentes disponíveis, considerando-se k-1 *dummies* para cada variável de k níveis, foram incluídas 28 variáveis no modelo, a saber: V_Q_PC_1, V_PR_AQ_N, V_Q_PC_2, V_CRE_CN, V_TP_E1, V_IDADE2, V_VL_EM1, V_SEXO_M, V_IDADE1, V_Q_PC_3, V_TP_E2, V_CEP_F1, V_IDADE3, V_COD_P3, V_COD_P7, V_FX__E1, V_EST_S, V_TP_R2, V_VL_EM3, V_VL_EM2, V_TP_R3, V_FX__E3, V_CEC_F2, V_CEC_F3, V_COD_P1, V_COD_P8, V_VL_PR1, V_CEC_F1.

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

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

A expressão e^{Z} é denominada desigualdade.

Neste estudo, Z é a combinação linear das 28 variáveis independentes ponderadas pelos coeficientes logísticos:

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

5.2.2 Resultados

A tabela 6 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 6: 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		

- Coeficientes logísticos das variáveis independentes

Com variáveis categóricas, a avaliação do efeito de uma particular categoria deve ser feita em comparação com uma categoria de referência. O coeficiente para a categoria de referência é 0. Para exemplificação, será interpretado o coeficiente da variável V_Q_PC_1, sendo análogas as

considerações para as demais. A variável quantidade de parcelas tem 4 níveis. Portanto, devem ser consideradas 3 variáveis *dummies*. Todas as 3 foram incluídas no modelo *stepwise*. A variável V_Q_PC_1 representa a primeira faixa da escala ordinal para quantidade de parcelas, com os códigos 1 para o nível mais baixo e 0, caso contrário. Analogamente, a variável V_Q_PC_2 corresponde à segunda faixa, com os códigos 1 para o segundo nível e 0, caso contrário. A categoria referência é o nível mais alto, no caso a quarta faixa. O coeficiente logístico para V_Q_PC_1 é positivo, indicando que, comparada à mais alta faixa de número de parcelas, a faixa de valor baixo está associada ao aumento do log das desigualdades dos tipos de clientes. Em outras palavras, clientes com empréstimo bancário com menos parcelas (primeira faixa) têm maior probabilidade de serem bons clientes comparativamente àqueles com empréstimo a ser pago com número superior de parcelas (quarta faixa). O impacto na desigualdade é dado por Exp(B) = Exp(1,6733) = 5,3296. De fato, fixando-se um valor para todas as variáveis incluídas no modelo (zero, por exemplo) e variando-se apenas o número de parcelas, é possível comparar o impacto da primeira faixa em relação à quarta faixa. A desigualdade para o cliente com maior número de parcelas seria, neste exemplo, igual a 1,79822 e resultaria em 9,58405 para aquele com menor número. Logo, a desigualdade para o nível mais baixo é superior ao quíntuplo da usada como referência (impacto de 5,3296, aproximadamente). As probabilidades, dadas pela fórmula do modelo logístico, são, para os níveis alto e baixo de parcelas, respectivamente, iguais a 0,643 e 0,906.

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. Por exemplo, para a variável v_pr_aq, um cliente na situação de ter o primeiro empréstimo concedido, em comparação a um cliente experiente na obtenção de empréstimos, tem menor probabilidade de se comportar como bom solicitante de apoio financeiro.

- Coeficiente de correlação parcial

Trata-se de uma medida da força de relação entre a variável dependente e uma variável independente, mantendo-se constantes os efeitos das outras variáveis independentes. O sinal desta

estatística é o mesmo do coeficiente logístico e a sua magnitude indica a contribuição da variável no modelo preditivo. 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.

- Teste de significância de cada variável

A estatística de Wald é definida como o quadrado da razão entre o coeficiente logístico estimado e o seu erro padrão. Por meio desta estatística, que tem distribuição Qui-quadrado, testa-se a seguinte hipótese estatística para cada variável independente:

H_0 : o coeficiente logístico é igual a zero.

Pela tabela 6, constata-se que os coeficientes de todas as variáveis incluídas no modelo logístico são estatisticamente diferentes de zero. Assim, de acordo com os níveis descritivos do teste, todas se mostraram relevantes para a discriminação entre os bons e maus clientes.

- Teste de significância do modelo

Há dois testes estatísticos para se avaliar a significância do modelo final: teste Qui-quadrado da mudança no valor de –2LL e o teste de Hosmer e Lemeshow.

A tabela 7 apresenta o valor inicial de –2LL, considerando-se apenas a constante no modelo, o seu valor final, a diferença "*improvement*" e o nível descritivo para se medir a sua significância.

Tabela 7: 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

Este teste Qui-quadrado testa a hipótese estatística de que os coeficientes para todos os termos no modelo final, exceto a constante, são iguais a zero. Este teste é comparável ao teste F da técnica de regressão múltipla. O valor Qui-quadrado é a diferença entre os dois valores de –2LL.

Espera-se que a inclusão de variáveis independentes contribua significantemente para a redução da estatística –2LL.

No modelo de 28 variáveis, constatou-se que 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. Portanto, trata-se de um teste do ajuste do modelo aos dados. A tabela 8 apresenta os resultados deste teste para este trabalho.

Tabela 8: Teste de Hosmer e Lemeshow

	Grupo = maus clientes		Grupo = bons clientes		
Grupos	Observado	Esperado	Observado	Esperado	Total
1	690	687,497	110	112,503	800
2	599	605,544	201	194,456	800
3	539	549,053	262	251,947	801
4	502	490,734	298	309,266	800
5	428	436,455	373	364,545	801
6	395	381,757	406	419,243	801
7	327	323,942	473	476,058	800
8	257	259,166	543	540,834	800
9	181	178,014	620	622,986	801
10	82	87,889	714	708,111	796

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.

Para se entender a elaboração da tabela 8 e o valor obtido para a estatística Qui-quadrado, será feita uma breve descrição dos passos inerentes a este teste.

Inicialmente os dados foram classificados em ordem crescente do valor obtido para a probabilidade prevista pelo modelo, conforme fórmula do modelo logístico. Não será exibida esta ordenação devido à grande magnitude da amostra (8000 casos). Foram, então, formados 10 blocos, sendo que o tamanho de cada bloco deve ser menor ou igual a M, como segue:

$M = 0{,}1.\ N + 0{,}5$

Nesta fórmula, N é o número de observações utilizadas, no caso 8000.

Assim, o valor máximo de M é 800,5 ou 801. Esta condição foi atendida, conforme revela a tabela 8. Além disso, devem ser formados aglomerados de observações com valores similares das variáveis preditoras e tais aglomerados não podem ser repartidos para alocação dos elementos em diferentes grupos. Assim que um grupo é completado, inicia-se a formação do próximo.

Considerando-se o evento bom cliente, código 1 da variável binária dependente, é, então, construída a tabela com a probabilidade média de ocorrência deste evento em cada um dos 10 blocos construídos. Esta probabilidade será a média das probabilidades, segundo a fórmula do modelo logístico, de todas as observações dentro de cada bloco. A freqüência esperada de elementos em cada bloco será o produto desta probabilidade média pelo número de observações pertencentes ao bloco. Esta freqüência esperada é, então, comparada com a freqüência observada no bloco.

A estatística Qui-quadrado é, então, calculada pela expressão:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

A partir desta fórmula foi obtido o resultado 3,4307, referente à estatística Qui-quadrado deste teste, que conduziu à não rejeição da hipótese nula, resultado favorável para os objetivos deste estudo.

A seção 5.5 apresentará os resultados de classificação obtidos pelo modelo de regressão logística e a comparação com os demais modelos.

5.3 REDE NEURAL

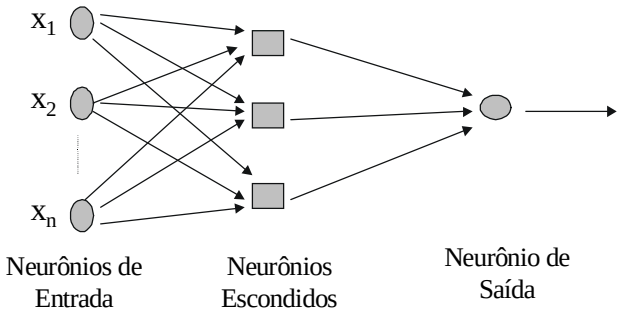
Neste trabalho, como já foi mencionado anteriormente, será utilizada uma rede com aprendizado supervisionado, pois já se conhece previamente se o cliente em questão é bom ou mau. Segundo Potts (1998, p. 44), a estrutura de rede neural mais utilizado para este tipo de problema é *multilayer perceptron* (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 trabalho também 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).

5.3.1 Modelo Implementado

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 neste trabalho. A figura 13, a seguir, ilustra o modelo.

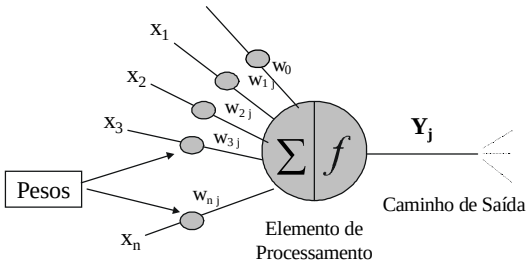
Figura 13: Modelo de rede neural artificial utilizado neste trabalho



Fonte: ARRAES *et al.* (2001)

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 figura 14, a seguir, resume a função computacional de um neurônio.

Figura 14: Função computacional do neurônio



Fonte: ARRAES *et al.* (2001)

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.

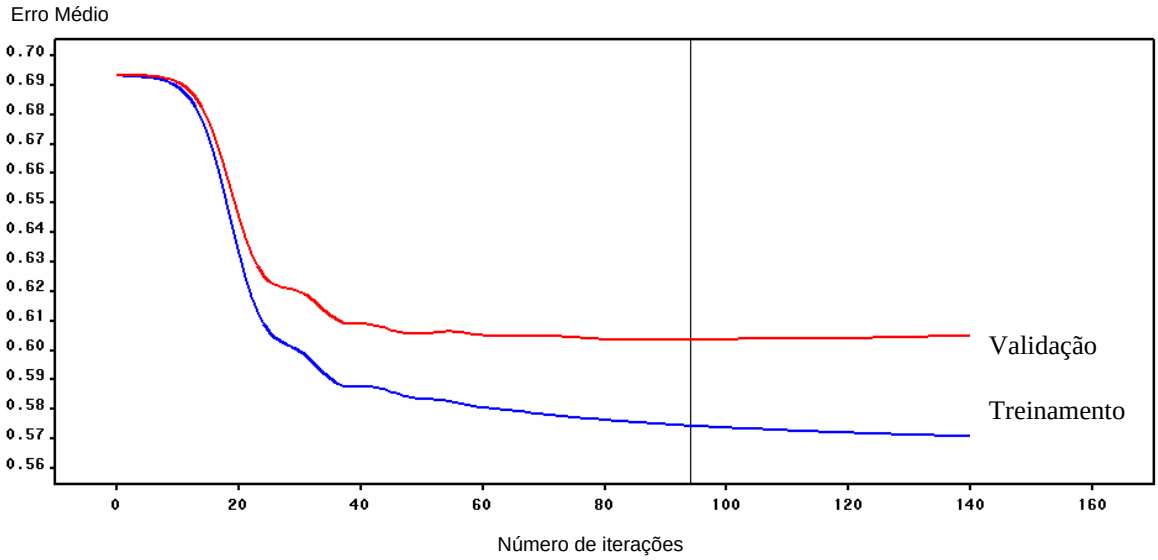
O treinamento da rede consiste em encontrar o conjunto de pesos W_i que minimiza uma função de erro. Neste trabalho, será utilizado para o treinamento o algoritmo *Back propagation*. Neste algoritmo a rede opera em uma seqüência de dois passos. Primeiro, um padrão é apresentado à camada de entrada da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

À luz do modelo de redes neurais adotado neste estudo foram processados os dados, cujas análises são apresentadas a seguir.

5.3.2 Resultados

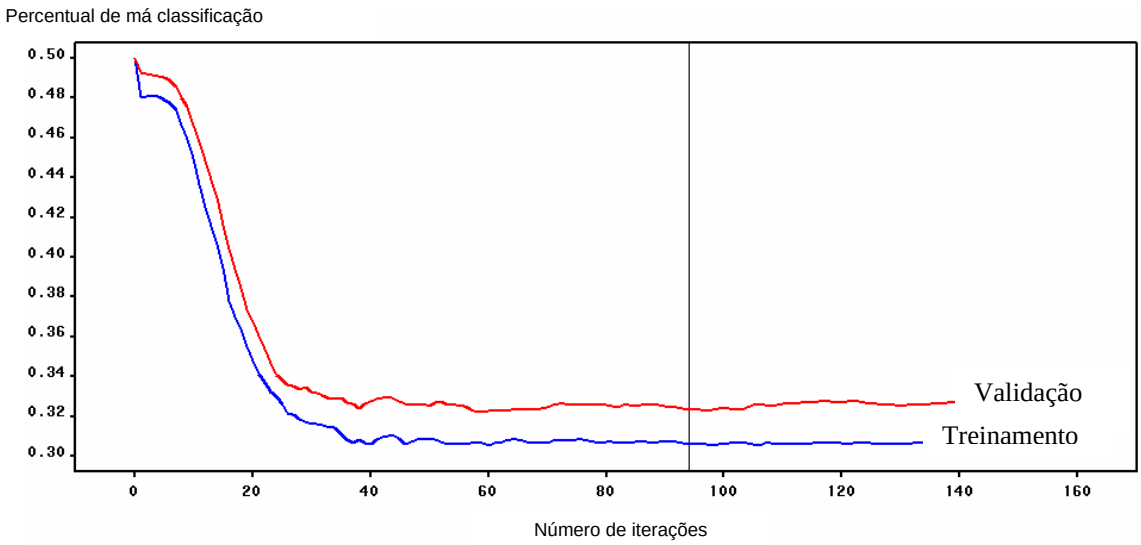
O erro médio do conjunto de dados de validação foi o critério de parada adotado neste modelo. Esse erro é calculado por intermédio do módulo da diferença entre o valor que a rede localizou e o esperado; calcula-se a sua média para os 8000 casos (amostra de treinamento) ou 6000 casos (amostra de validação). A figura 15 apresenta a curva de erro com diminuição progressiva até sua estabilização. O processamento detectou que a estabilidade do modelo ocorreu após a nonagésima quarta iteração, que é o ponto marcado pela linha vertical. 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.

Figura 15: Curva de erro médio



Na figura 16 é mostrada a evolução da má classificação de ambas as amostras. Inicialmente, a má classificação é de 50%, pois a alocação de um indivíduo como bom ou mau cliente é aleatória; com o aumento das iterações, é atingido o melhor resultado de 30,6% de erro para a amostra de treino e 32,3% para a amostra de validação. Na seção 5.5 serão mostrados os resultados com mais detalhes.

Figura 16: Curva de erro de classificação



Na tabela 9, estão algumas das estatísticas da rede adotada. Além da classificação incorreta e o erro médio, são apresentados ainda o erro quadrático e os graus de liberdade. O erro quadrático médio calcula-se pela média dos quadrados das diferenças entre o observado e o obtido pela rede. O número de graus de liberdade do modelo refere-se ao número de pesos estimados, à conexão de cada um dos atributos aos neurônios da camada intermediária e às ligações da camada intermediária com a saída.

Tabela 9: 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 ALGORITMOS GENÉTICOS

Na literatura consultada, foram encontradas duas maneiras de lidar com este tipo de problema por meio de algoritmos genéticos. A primeira, adotada por Chen *et al* (2002) e Fidelis *et al* (2000), soluciona o problema por meio de uma seqüência de regras tal qual uma árvore de decisão, ou seja, uma série de regras encadeadas que determinam se o cliente é bom ou mau, dependendo do caminho (ou galho da árvore) percorrido.

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

5.4.1 Modelo Implementado

O algoritmo implementado foi similar ao apresentado em Picinini *et al* (2003). Cada uma das 71 categorias de variável (seção 5.1) recebeu um peso aleatório inicial. A esses setenta e um coeficientes foi introduzido mais um, uma constante aditiva incorporada à equação linear. O valor de escore do cliente é dado por:

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

S_j = Escore obtido pelo cliente j

w_i = Peso relativo à categoria i

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

Para se definir se o cliente é bom ou mau foi utilizada a seguinte regra[7]:

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

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

Assim sendo, o problema que o algoritmo deve solucionar é encontrar o vetor $W=[w_1, w_2, ..., w_{72}]$ que resulte em um critério de classificação com uma boa taxa de acertos na predição do desempenho de pagamento do crédito.

Seguindo as fases de um algoritmo genético, conforme apresentado na seção 4.3.3, têm-se:

Início: foi gerada uma população de 200 indivíduos, com cada cromossomo contendo 72 genes. O peso inicial w_i de cada um dos genes foi gerado aleatoriamente no intervalo [-1,1] (Picinini *et al*, 2003, p. 464).

Função de Aptidão (*Fitness*): cada cliente foi associado ao cálculo de um escore e classificado como bom ou mau. Comparando-se com a informação já conhecida a priori sobre a natureza do cliente, pode-se calcular a precisão de cada cromossomo. O indicador de acertos (Ia), apresentado na seção 4.4.1 será a função de aptidão, ou seja, quanto maior o indicador, melhor será o cromossomo.

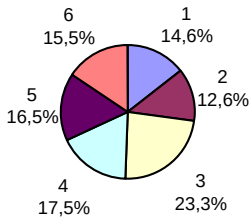
Seleção: neste trabalho foi utilizado um elitismo de 10%, ou seja, para cada nova geração, os vinte melhores cromossomos são mantidos, enquanto os outros cento e oitenta são formados por meio de cruzamento e mutação.

[7] A escolha do ponto de corte é arbitrária, e não tem grande impacto no resultado final. O ponto de corte é o que vai direcionar o cálculo dos parâmetros; se o ponto de corte for diferente, o algoritmo vai recalcular os parâmetros até atingir o resultado ótimo. Para facilitar o entendimento do processo e, ao mesmo tempo, torná-lo similar às outras técnicas conhecidas, como, por exemplo, a regressão logística, foi adotado zero como ponto de corte, da mesma forma que foi feito por Picinini *et al* (2003, p. 464).

Cruzamento (*Cross-Over*): para a escolha dos pais para o cruzamento, foi utilizado o método conhecido como roleta (*roulette wheel*) para seleção dentre os vinte cromossomos que foram mantidos (CHEN; HUANG, 2003, p. 436-437). Neste método, cada indivíduo recebe uma probabilidade de ser sorteado de acordo com seu valor de função de aptidão, conforme mostra a tabela 10, a seguir.

Tabela 10: Exemplo de Seleção de Pais via Roleta

Cromossomo	*Fitness*	%	% Acum.
Cromossomo1	15	14,6%	14,6%
Cromossomo2	13	12,6%	27,2%
Cromossomo3	24	23,3%	50,5%
Cromossomo4	18	17,5%	68,0%
Cromossomo5	17	16,5%	84,5%
Cromossomo6	16	15,5%	100,0%
Total	103	100,0%	



Neste exemplo, o Cromossomo3 tem 23,3% de chance de ser selecionado como pai. Sua chance é maior, pois é o cromossomo com maior valor de *fitness* (aptidão). Inclusive ele pode ser selecionado duas vezes, o que fará com que seu filho seja uma reprodução exata de si mesmo.

Para o processo de troca de material genético, foi utilizado um método conhecido como cruzamento uniforme (PAPPA, 2002, p. 22). Neste tipo de cruzamento, cada gene do cromossomo filho é escolhido aleatoriamente entre os genes de um dos pais, enquanto o segundo filho recebe os genes complementares do segundo pai, conforme mostra a figura 17.

Figura 17: Exemplo de Cruzamento Uniforme

Pai 1	1	1	1	1	1	1	1	1	1	1	1
Pai 2	2	2	2	2	2	2	2	2	2	2	2

Filho 1	1	2	1	1	2	1	2	2	1	1	2
Filho 2	2	1	2	2	1	2	1	1	2	2	1

Fonte: PAPPA (2002, p. 23)

Mutação: no processo de mutação, cada gene do cromossomo é avaliado independentemente. Cada gene de cada cromossomo tem probabilidade de 0,5% de sofrer mutação. Sempre que um gene for escolhido para a mutação, a alteração genética é realizada, adicionando-se um pequeno

valor escalar k neste gene. No experimento descrito, foi sorteado aleatoriamente um valor entre –0,05 e +0,05.

Verificação do critério de parada: como critério de parada, foi definido um número máximo de gerações igual a 600. Após as seiscentas iterações, o cromossomo com maior aptidão será a solução.
Os resultados obtidos para esta configuração de algoritmo são apresentados a seguir.

5.4.2 Resultados

O algoritmo foi executado três vezes conforme a configuração apontada na seção anterior. Aqui serão apresentados os resultados do algoritmo que obteve o maior Indicador de acertos (Ia).

Após a execução do algoritmo, as variáveis com peso muito pequeno foram descartadas. No trabalho de Picinini *et al* (2003, p. 464) os autores consideraram que as variáveis com peso inferior a 0,15 ou superior a -0,15 seriam descartadas por possuírem um peso não significativo para o modelo. Neste trabalho, depois de feita uma análise de sensibilidade, decidiu-se considerar como significativas para o modelo as variáveis com peso superior a 0,10 ou inferior a -0,10. Essa regra não foi aplicada para a constante, que se mostrou importante para o modelo mesmo com o valor abaixo do ponto de corte.

O peso das variáveis é apresentado na tabela 11. Nesta tabela foram separadas as variáveis que obtiveram peso negativo daquelas com peso positivo. O peso negativo indica que a variável tem uma relação maior com os clientes considerados maus (pois foi determinado na seção anterior que clientes com escore total negativo seriam considerados maus). O peso positivo, de forma inversa, indica relação com os clientes bons.

Tabela 11: Pesos finais das variáveis

Pesos Negativos

Variável	Peso (w)
V_FN_C_N	-0,98
V_CRE_CN	-0,98
V_IDADE2	-0,98
V_SAL_F1	-0,95
V_COD_P2	-0,91
V_Q_PC_4	-0,88
V_SAL_F4	-0,87
V_FX__P3	-0,8
V_CEP_F2	-0,79
V_VL_EM5	-0,76
V_Q_PC_3	-0,65
V_SAL_F3	-0,61
V_VL_EM4	-0,59
V_CEC_F2	-0,59
V_COD_P4	-0,56
V_TP_E1	-0,55
V_FN_R_S	-0,54
V_IDADE1	-0,54
V_CEC_F3	-0,5
V_TP_E2	-0,45
V_FX__P2	-0,45
V_CEP_F4	-0,44
V_FX__E1	-0,42
V_FX__E4	-0,39
V_VL_EM6	-0,28
V_CEP_F3	-0,28
V_PR_AQ_S	-0,28
V_CEP_F1	-0,23
V_CEC_F1	-0,22
V_CEC_F5	-0,21
V_TP_R2	-0,14
V_SAL_F2	-0,12
V_COD_P8	-0,12
Constante	-0,08

Pesos Positivos

Variável	Peso (w)
V_Q_PC_1	1,42
V_SEXO_F	0,97
V_COD_P7	0,95
V_FX__E3	0,95
V_EST_C	0,93
V_IDADE4	0,89
V_Q_PC_2	0,88
V_FX__P5	0,88
V_VL_EM1	0,83
V_CRE_CH	0,81
V_TP_R4	0,75
V_VL_EM2	0,59
V_EST_O	0,58
V_CEP_F5	0,57
V_TP_E4	0,56
V_FX__P1	0,55
V_SAL_F6	0,47
V_COD_P3	0,45
V_VL_PR4	0,41
V_TP_E3	0,39
V_TP_R3	0,39
V_VL_PR2	0,34
V_COD_P9	0,33
V_SEXO_M	0,29
V_VL_EM3	0,25
V_PR_AQ_N	0,24
V_TP_R1	0,19
V_EST_S	0,14
V_CEC_F4	0,13
V_COD_P1	0,13

Comparando-se estes resultados com os obtidos pela regressão logística, nota-se uma concordância nas variáveis com peso mais alto. Em ambos os modelos, a variável com maior peso negativo foi a variável V_CRE_CN e com maior peso positivo foi V_Q_PC1 (esta foi, em ambos os modelos, a variável com maior peso absoluto). Outras variáveis como V_TP_E1, V_IDADE2, V_Q_PC_2, V_VL_EM1, V_VL_EM2 também estão entre as variáveis com maior peso nos dois modelos, evidenciando que o resultado do algoritmo foi coerente.

5.5 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 a seguir. O detalhamento do cálculo do KS encontra-se no Apêndice B.

Tabela 12: 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	Mau	2111	889	70,4	Mau	2159	841	72,0
	Bom	1294	2706	67,7	Bom	1078	1922	64,1	Bom	1059	1941	64,7
	Total	4127	3873	69,2	Total	3189	2811	67,2	Total	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	Mau	2236	764	74,5	Mau	2255	745	75,2
	Bom	1430	2570	64,3	Bom	1177	1823	60,8	Bom	1193	1807	60,2
	Total	4409	3591	69,4	Total	3413	2587	67,7	Total	3448	2552	67,7

ALGORITMO GENÉTICO

		Treinamento				Validação				Teste		
		Predito				Predito				Predito		
		Mau	Bom	% Acerto		Mau	Bom	% Acerto		Mau	Bom	% Acerto
Observado	Mau	2692	1308	67,3	Mau	1946	1054	64,9	Mau	2063	937	68,8
	Bom	1284	2716	67,9	Bom	1043	1957	65,2	Bom	1073	1927	64,2
	Total	3976	4024	67,6	Total	2989	3011	65,1	Total	3136	2864	66,5

A tabela 12 mostra os resultados de classificação obtidos pelos três modelos. Todos eles 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 nos modelos de regressão logística e rede neural, e foram um pouco inferiores para o modelo de algoritmos genéticos. Outro resultado interessante é que, exceto para os algoritmos genéticos, 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 modelos logístico e redes neurais.

A tabela 13, a seguir, apresenta os resultados dos critérios Ia e KS que foram os escolhidos para comparar os modelos. Ressalte-se que os índices Ia são derivados dos resultados da tabela 12, conforme explanado no capítulo 4, seção 4.4.1.

Tabela 13: Índices de Comparação

Ia	Amostra		
	Treinamento	Validação	Teste
Regressão Logística	47,9	45,1	46,6
Rede Neural	47,9	45,3	45,3
Algoritmo Genético	45,7	42,3	44,2

KS	Amostra		
	Treinamento	Validação	Teste
Regressão Logística	38	35	37
Rede Neural	39	35	35
Algoritmo Genético	34	30	32

Os valores KS de todos os modelos podem ser considerados bons. Novamente, 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". Aqui novamente os modelos de regressão logística e redes neurais apresentam um resultado muito próximo, superior ao obtido pelo algoritmo genético.

Na escolha do modelo mais adequado para estes dados, analisando sob o prisma dos indicadores Ia e KS, foi eleito o modelo construído por **regressão logística**, pois, apesar de ter resultados

muito similares aos obtidos por redes neurais, este modelo apresentou melhores resultados na amostra de teste, sugerindo ser o mais adequado para a aplicação em outras bases de dados. Contudo, deve ser ressaltado, mais uma vez, que a adoção de qualquer um dos modelos traria bons resultados à instituição financeira.

CAPÍTULO 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 com o uso de Regressão Logística, Redes Neurais Artificiais e Algoritmos Genéticos.

No desenvolvimento de modelos de avaliação de crédito alguns cuidados devem ser tomados a fim de se garantir a qualidade do modelo, e a aplicabilidade posterior. Precauções na amostragem, definição clara nos critérios na classificação de clientes bons e maus e tratamento das variáveis da base de dados antes da aplicação das técnicas foram cuidados tomados neste estudo, visando otimizar resultados e minimizar erros.

Os três 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 apresentou resultados levemente superiores ao modelo construído por redes neurais e ambos mostraram-se superiores ao modelo baseado em algoritmos genéticos. O modelo proposto por este estudo para que a instituição pontue seus clientes é:

$$p = \frac{e^{Z}}{1+e^{Z}}$$, onde

p...probabilidade de o cliente ser considerado bom e

$Z = B_0 + B_1.X_1 + B_2.X_2 + ...+ B_{28}.X_{28}$, onde os valores de B_i e X_i encontram-se na tabela 6.

O percentual de acerto total para a amostra de teste foi para a regressão logística, redes neurais e algoritmos genéticos, respectivamente igual a 68,3; 67,7 e 66,5. Na literatura consultada, o percentual de acerto total flutua bastante, bem como o modelo mais adequado em cada banco de dados pode ser diferente do obtido neste estudo. A tabela 14, extraída do trabalho de Thomas (2000), mostra a variedade de resultados obtidos em outros trabalhos.

Tabela 14: Precisão da classificação dos modelos construídos para análise de crédito

	Regressão Linear	Regressão Logística	Árvores de Classificação	Programação Linear	Redes Neurais	Algoritmos Genéticos
Henley(1995)	56,6	**56,7**	56,2	-	-	-
Boyle (1992)	**77,5**	-	75	74,7	-	-
Srinivisan(1987)	87,5	89,3	**93,2**	86,1		
Yobas (1997)	**68,4**	-	62,3	-	62	64,5
Desai(1997)	66,5	**67,3**	**67,3**	-	64	-

Fonte: THOMAS (2000, p. 159)

A tabela 15, construída a partir da literatura pesquisada, é similar à tabela anterior e reforça a grande variedade de resultados. Note-se que, ao se analisarem as duas tabelas, os modelos apresentam uma precisão de classificação que varia de 56,2 a 93,2. Observa-se ainda que, excetuando-se a programação linear, todos os outros métodos apresentados, em ao menos um estudo, apresentaram a maior precisão.

Tabela 15: Precisão da classificação dos modelos construídos (literatura pesquisada) [8]

	Regressão Linear	Regressão Logística	Árvores de Classificação	Programação Linear	Redes Neurais	Algoritmos Genéticos	Análise Discriminante	REAL
Fritz e Hosemann (2000)			79,5		81,6	82,4	**82,7**	
Arraes *et al* (1999)		84,8			**85,4**			
Chen *et al* (2002)					91,9	**92,9**		
Nanda e Pendharkar (2001)						**65**	62,5	
Ohtoshi (2003)		83,5	73,9		**85**			83,1
Picinini *et al* (2003)		63,5			64,4	**67,5**		
Arminger *et al* (1997)		**67,6**	66,4		65,2			
Huang *et al* (2004)		77			**80**			
Semolini (2002)		**68,3**			67,4			
Rosa (2000)		70,4	66,6					**71,4**

[8] A metodologia REAL (*Real Attribute Learning Algorithm*), apresentada na tabela 15 é um modelo similar a uma árvore de classificação proposto por Stern *et al* (1998); mais detalhes podem ser encontrados em Rosa (2000) e Ohtoshi (2003).

Não foi objeto deste estudo uma abordagem mais profunda das técnicas focalizadas. As redes neurais e os algoritmos genéticos apresentam uma grande gama de estruturas e variações que podem (e devem) ser melhor exploradas. Os algoritmos genéticos, por serem um método bastante flexível e ainda não tanto pesquisado em problemas de concessão de crédito, podem ser aplicados de formas diversas a fim de otimizar o resultado obtido.

Técnicas novas neste tipo de problema, como análise de sobrevivência, também merecem atenção em estudos futuros.

BIBLIOGRAFIA

ABE, S. (1997) *Neural Networks and Fuzzy Systems,* Boston: Kluwer Academic Publishers.

ALMEIDA, F. C.; DUMONTIER, P. (1996) O Uso de Redes Neurais em Avaliação de Risco de Inadimplência, *Revista de Administração, São Paulo,* v. 31, n. 1, p. 52-63, São Paulo: Universidade de São Paulo.

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

ANDREEVA, G.; ANSELL, J.; CROOK, J. N. (2003) *Credit Scoring in the Context of the European Integration*, Edinburgh: Proceedings of Credit Scoring & Credit Control VIII Conference, September 2003, UEMS.

ARMINGER, G., ENACHE, D., BONNE T. (1997) Analyzing Credit Risk Data: A Comparison of Logistic Discrimination, Classification Trees and Feedforward Networks. *Computational Statistics*, v. 12, n. 2, p. 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.

AXELROD, R. (1987) The Evolution of Strategies in the Iterated Prisoner's Dilemma, In: *Genetic Algorithms and Simulated Annealing,* Londres: Pitman, p. 32-41.

BACK, B., LAITINEN, T., AND SERE, K. (1996) *Neural Networks and Genetic Algorithms for Bankruptcy Predictions*. Seul: Proceedings of the 3rd World Conference on Expert Systems, p. 123-130.

BARTH, N. (2004) *Análise Quantitativa de Informações Para Previsão de Inadimplência,* São Paulo: I Congresso Anual de Tecnologia da Informação.

BAUER, R. J. (1994) *Genetic Algorithms and Investment Strategies,* New York: John Wiley & Sons.

BERGAMINI, JR., S. (1997) Classificação de Riscos: O Modelo em Uso no BNDES, *Revista do BNDES,* v. 4, n. 8, p. 71-100, Rio de Janeiro: Banco Nacional de Desenvolvimento Econômico e Social.

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

BHATTACHARYYA, S. (2003) Evolutionary computation for database marketing *Journal of Database Management***,** v. 10, n. 4, p. 343-352, Londres: Henry Stewart Publications.

BUGERA, V., KONNO, H., AND URYASEV. S (2002) Credit cards scoring with quadratic utility functions, *Journal of Multi-Criteria Decision Analysis*, v. 11, n. 4-5, p. 197-211, New York: John Wiley & Sons.

CANO, J. R. (2004) *Reducción de Datos basada en Selección Evolutiva de Instancias para Minería de Datos.* Tese de Doutorado. Departamento de Ciência da Computação e Inteligência Artificial, Universidade de Granada, Espanha.

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.

CARVALHO, A. P. L. F. [S.l., s.d.] *Redes Neurais Artificiais*, disponível em http://www.icmc.usp.br/~andre/research/neural/ acesso em 04/12/04

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

CHEN, M.-C.; HUANG, S.-H (2003) Credit scoring and rejected instances reassigning through evolutionary computation techniques, *Expert Systems with Applications*, v. 24, n. 4, p. 433-441 St. Louis :Elsevier Science.

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.

CZARN, A.; MACNISH C.; VIJAYAN, K. TURLACH, B.; GUPTA R. (2004) Statistical Exploratory Analysis of Genetic Algorithms. *IEEE Transactions on Evolutionary Computation* v. 8, n. 4, p. 405-421, Birmingham: IEEE Computational Intelligence Society.

DESAI V.S., CONVAY D.G., CROOK J.N., OVERSTREET G.A. (1997) Credit scoring models in the credit union environment using neural networks and genetic algorithms, *IMA J. Mathematics applied in Business and Industry*, v. 8, p. 323-346, Oxford: Oxford University Press.

DOBSON, A. (1990) *An Introduction to Generalized Linear Models*, Londres: Chapman & Hall.

DRYE T.; WETHERILL G.; PINNOCK A. (2001) When are customers in the market? Applying survival analysis to marketing challenges, *Journal of Targeting, Measurement and Analysis for Marketing,* v. 10, n. 2, p. 179-188, Londres: Henry Stewart Publications.

DUARTE, JR., A. M.; BASTOS, N. T.; PINHEIRO, F. P.; JORDÃO, M. R. (1999) Gerenciamento de Riscos Corporativos: Classificação, Definições e Exemplos, *Resenha BM&F*, n. 134, São Paulo: Bolsa de Mercadorias & Futuros

DUARTE, JR., A. M. (1996). Riscos: Definições, Tipos, Medição e Recomendações para seu Gerenciamento. *Resenha BM&F*, n. 114, p. 25-33 São Paulo: Bolsa de Mercadorias & Futuros

EMPRESA BRASILEIRA DE CORREIOS E TELEGRAFOS [S. l., s.d.] *Homepage da Estrutura do CEP* disponível http://www.correios.com.br/servicos/cep/cep_estrutura.cfm acesso em 07/03/05.

FAUSETT, L. (1994) *Fundamentals of Neural Networks,* Englewood-Cliffs: Prentice-Hall.

FENSTERSTOCK, F. (2005) Credit Scoring and the Next Step. *Business Credit,* v. 107, n. 3, p. 46-49, New York: National Association of Credit Management.

FIDELIS, M.V.; LOPES, H.S.; FREITAS, A.A. (2000) *Discovering comprehensible classification rules with a genetic algorithm*. La Jolla: Proceedings of Congress on Evolutionary Computation p. 805-810.

FIGUEIREDO, R. P. (2001) *Gestão de Riscos Operacionais em Instituições Financeiras – Uma Abordagem Qualitativa*, Dissertação de Mestrado. Belém: Universidade da Amazônia UNAMA.

FRANÇOIS, O.; LAVERGNE C. (2001) Design of evolutionary algorithms-A statistical perspective. *IEEE Transactions on evolutionary Computation* v. 5, n. 2, p. 129-148, Birmingham: IEEE Computational Intelligence Society.

FRITZ, S.; HOSEMANN, D. (2000) Restructuring the Credit Process: Behaviour Scoring for German Corporates *International Journal of Intelligent Systems in Accounting, Finance and Management***,** v. 9, n. 1, p. 9-21, Nottingham: John Wiley & Sons.

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

GOONATILAKE, S.; TRELEAVEN, P. C. (1995) *Intelligent Systems for Finance and Business,* New York: Wiley

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. (1998) *Análise Multivariada de Dados*, Porto Alegre: Bookman.

HALE, R.H. (1983) *Credit Analysis: A Complete Guide,* New York: John Wiley & Sons.

HALL, L.O.; OZYURT, I.B.; BEZDEK, J.C. (1999) Clustering with a genetically optimized approach *IEEE Transactions on evolutionary Computation,* v. 3, n. 2, p. 103-112, Birmingham: IEEE Computational Intelligence Society.

HAND, D. J.; HENLEY, W. E. (1997) Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of Royal Statistical Society: Series A,* n. 160, p. 523-541 Londres: Royal Statistical Society.

HARIK, G. R; LOBO, F. G; GOLDBERG, D. E. (1999) The compact genetic algorithm**.** *IEEE Transactions on Evolutionary Computation,* v. 3, n. 4, p. 287-297, Birmingham: IEEE Computational Intelligence 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,* v. 6, n. 3, p. 229-239, Londres: Henry Stewart Publications.

HAYKIN, S. (1999) *Redes Neurais Princípios e Prática,* Porto Alegre: Bookman.

HRUSCHKA, E. R. (2001) *Algoritmos Genéticos de Agrupamento para Extração de Regras de Redes Neurais* Tese de Doutorado. Departamento de Engenharia Civil Universidade de Federal do Rio de Janeiro UFRJ.

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*, v. 37 n. 4, p. 543-558, St. Louis :Elsevier Science

KIM, C. N. (2003) A Neural Network Approach to Compare Predictive Value of Accounting Versus Market Data *International Journal of Digital Management*, v. 3, Seul: Hanyang University, disponível em http://ijdm.digital.re.kr/past.html acesso em 29/01/05.

KISHORE, J. K.; PATNAIK, L. M.; MANI, V.; AGRAWAL, V. K. (2000) Application of genetic programming for multicategory pattern classification. *IEEE Transactions on evolutionary Computation,* v. 4, n. 3, p. 242-257, Birmingham: IEEE Computational Intelligence Society.

KNIGHT, K. (1990) Connectionist ideas and Algorithms. *Communications of the ACM* v. 33, n. 11, p. 59-74, New York: Association for Computing Machinery, Inc.

LEWIS, E. M. (1992) *An Introduction to Credit Scoring.* San Rafael: Fair Isaac and Co., Inc.

LIMA, J. (2002) *A Análise Econômico-Financeira de Empresas sob a Ótica da Estatística Multivariada* Dissertação de Mestrado, Curitiba: Universidade Federal do Paraná.

LINGRAS, P. (2001*)* Unsupervised Rough Set Classification using GAs *Journal of Intelligent Information Systems***.** v. 16, n. 3; p. 215-228, Boston: Kluwer Academic Publishers.

MAGYAR, G.; JOHNSSON M.; NEVALAINEN, O. (2000) An Adaptive Hybrid Genetic Algorithm for the Three-Matching Problem *IEEE Transactions on evolutionary Computation*, v. 4, n. 2, p. 135-146, Birmingham: IEEE Computational Intelligence Society.

MARKS, R.E.; AND SCHNABL, H. (1997) Genetic Algorithms and Neural Networks: a comparison based on the Repeated Prisoner's Dilemma, *Computational Techniques for Modelling Learning in Economics*, in the series Advances in Computational Economics, Dordrecht: Kluwer Academic Publishers, forthcoming. Australian Graduate School of Management Working Paper 97-014.

MARTINELI, E. (1999) *Extração de conhecimento de redes neurais artificiais* Dissertação de Mestrado. Instituto de Ciências Matemáticas e de Computação USP São Carlos.

MATIAS, A. B.; SIQUEIRA, J. O. (1996) *Risco Bancário: modelo de previsão de insolvência de bancos no Brasil.* Revista de Administração, São Paulo v. 31, n. 2, p. 19-28, São Paulo: Universidade de São Paulo.

NANDA, S.; PENDHARKAR, P. (2001) Linear models for minimizing misclassification costs in bankruptcy prediction *International Journal of Intelligent Systems in Accounting, Finance and Management,* v. 10, n. 3, p. 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.

OOGHE, H.; CLAUS, H.; SIERENS, N.; CAMERLYNCK, J. (2001) *International Comparison of Failure Prediction Models from Different Countries: An Empirical Analysis.* Working paper, Department of Corporate Finance, University of Ghent.

PAL, S. K.; WANG, P. P. (1996) *Genetic Algorithms for Pattern Recognition,* Boca Raton: CRC Press.

PAMPA QUISPE, N. R. (2003) *Técnicas e ferramentas para a extração inteligente e automática de conhecimento em banco de dados* Dissertação de Mestrado. Departamento de Engenharia Elétrica. Universidade Estadual de Campinas FEEC/UNICAMP.

PAPPA, G. L. (2002) *Seleção de Atributos Utilizando Algoritmos Genéticos Multiobjetivos* Dissertação de Mestrado. Departamento de Informática. Pontifícia Universidade do Paraná.

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, 2003, Bauru, SP. Anais do VI Simpósio Brasileiro de Automação Inteligente, p. 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.

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

SANTOS, J.O. (2000) *Análise de Crédito: Empresas e Pessoas Físicas,* São Paulo: Atlas.

SCARPEL, R. A., MILIONI A. Z. (2001). *Aplicação de Modelagem Econométrica À Análise Financeira de Empresas.* Revista de Administração v. 36, n. 11, p. 80-88, São Paulo: Universidade de São Paulo.

SCHRICKEL, W. K. (1995) *Análise de Crédito: Concessão e Gerência de Empréstimos,* São Paulo: Atlas.

SECURATO, J.R. (2002) *Crédito: Análise e Avaliação do Risco,* São Paulo: Saint Paul.

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.

SHARMA, S. (1996) *Applied Multivariate Techniques,* New York: John Wiley and Sons.

SIEGEL, S. (1975). *Estatística Não-Paramétrica Para as Ciências do Comportamento* São Paulo: Mc Graw-Hill.

STERN, J.M.; NAKANO, F.; LAURETTO, M.S.; RIBEIRO, C.O. (1998) *REAL: Algoritmo de Aprendizagem para Atributos Reais e Estratégias de Operação em Mercado,* Lisboa: Conferência Ibero-americana de Inteligência Artificial.

SILVA FILHO, D.; CARNEIRO, A.A.F.M. (2004) Dimensionamento evolutivo de usinas hidroelétricas. *SBA Controle & Automação*, v. 15, n. 4, p. 437-448, São José dos Campos: Sociedade Brasileira de Automática.

TATIBANA, C. Y.; KAETSU D. Y. [S. l., s.d.] *Homepage de Redes Neurais* disponível em http://www.din.uem.br/ia/neurais/ acesso em 04/12/04.

THOMAS, L. (2000) A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers, *International Journal of Forecasting*, v. 16, n. 2, p. 149-172, Londres: 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.

VASCONCELLOS, M. S. (2002) *Proposta de Método para Análise de Concessões de Crédito a Pessoas Físicas* Dissertação de Mestrado. Departamento de Economia Universidade de São Paulo FEA/USP.

ZERBINI, M. B. A. A. (2000) *Três Ensaios sobre Crédito* Tese de Doutorado. Departamento de Economia Universidade de São Paulo FEA/USP.

APÊNDICE A – CÁLCULO DO RISCO RELATIVO

SEXO	Bom	Mau	% Bom	% Mau	RR
Masculino	5528	5858	0,55	0,59	0,94
Feminino	4472	4142	0,45	0,41	1,08
Total	10000	10000	1	1	

ESTADO CIVIL	Bom	Mau	% Bom	% Mau	RR
Casado	4817	4189	0,48	0,42	1,15
Solteiro	3461	4284	0,35	0,43	0,81
Outros	1722	1527	0,17	0,15	1,13
Total	10000	10000	1	1	

PRIMEIRA AQUISIÇÃO	Bom	Mau	% Bom	% Mau	RR
Sim	4471	6480	0,45	0,65	0,69
Não	5529	3520	0,55	0,35	1,57
Total	10000	10000	1	1	

POSSE DE FONE COMERCIAL	Bom	Mau	% Bom	% Mau	RR
Sim	6980	7392	0,70	0,74	0,94
Não	3020	2608	0,30	0,26	1,16
Total	10000	10000	1	1	

TIPO DE CRÉDITO	Bom	Mau	% Bom	% Mau	RR
Carnê	917	2067	0,09	0,21	0,44
Cheque	9083	7933	0,91	0,79	1,14
Total	10000	10000	1	1	

POSSE DE FONE RESIDENCIAL	Bom	Mau	% Bom	% Mau	RR
Sim	9979	9957	1,00	1,00	1,00
Não	21	43	0,00	0,00	0,49
Total	10000	10000	1	1	

TEMPO DE RESIDÊNCIA	Bom	Mau	% Bom	% Mau	RR
Até 12 meses	659	850	0,07	0,09	0,78
De 13 a 24 meses	666	851	0,07	0,09	0,78
De 25 a 120 meses	3581	3717	0,36	0,37	0,96
Acima de 120 meses	5094	4582	0,51	0,46	1,11
Total	10000	10000	1	1	

VALOR DO EMPRÉSTIMO	Bom	Mau	% Bom	% Mau	RR
Até 300 reais	2083	1225	0,21	0,12	1,70
Acima de 300 a 400 reais	975	964	0,10	0,10	1,01
Acima de 400 a 500 reais	1521	1317	0,15	0,13	1,15
Acima de 500 a 800 reais	1826	2354	0,18	0,24	0,78
Acima de 800 a 1800 reais	2650	3154	0,27	0,32	0,84
Acima de 1800 reais	945	986	0,09	0,10	0,96
Total	10000	10000	1	1	

IDADE	Bom	Mau	% Bom	% Mau	RR
Até 25 anos	568	893	0,06	0,09	0,64
De 26 a 40 anos	3381	4215	0,34	0,42	0,80
De 41 a 58 anos	4182	3718	0,42	0,37	1,12
Acima de 58 anos	1869	1174	0,19	0,12	1,59
Total	10000	10000	1	1	

PARCELA	Bom	Mau	% Bom	% Mau	RR
Até 125 reais	2803	3118	0,28	0,31	0,90
Acima de 125 a 160 reais	2172	1909	0,22	0,19	1,14
Acima de 160 a 260 reais	2765	3119	0,28	0,31	0,89
Acima de 260 reais	2260	1854	0,23	0,19	1,22
Total	10000	10000	1	1	

TEMPO NO EMPREGO ATUAL	Bom	Mau	% Bom	% Mau	RR
Até 24 meses	1525	2580	0,15	0,26	0,59
De 25 a 72 meses	2926	3170	0,29	0,32	0,92
De 73 a 127 meses	2080	1778	0,21	0,18	1,17
Acima de 128 meses	3469	2472	0,35	0,25	1,40
Total	10000	10000	1	1	

% VALOR DA PARCELA/SALÁRIO	Bom	Mau	% Bom	% Mau	RR
Até 10%	2296	1667	0,23	0,17	1,38
Acima de 10 a 13,5%	2113	2035	0,21	0,20	1,04
Acima de 13,5 a 16,5%	1918	2046	0,19	0,20	0,94
Acima de 16,5 a 22,5%	2819	3629	0,28	0,36	0,78
Acima de 22,5%	854	623	0,09	0,06	1,37
Total	10000	10000	1	1	

% VALOR DO EMPRÉSTIMO/SALÁRIO	Bom	Mau	% Bom	% Mau	RR
Até 28%	1551	452	0,16	0,05	3,43
Acima de 28 a 47,5%	2378	1645	0,24	0,16	1,45
Acima de 47,5 a 65%	1990	2178	0,20	0,22	0,91
Acima de 65%	4081	5725	0,41	0,57	0,71
Total	10000	10000	1	1	

CÓDIGO DE PROFISSÃO	Bom	Mau	% Bom	% Mau	RR
Código 1	976	910	0,10	0,09	1,07
Código 2	439	563	0,04	0,06	0,78
Código 3	1234	1037	0,12	0,10	1,19
Código 4	1101	1453	0,11	0,15	0,76
Código 5	842	750	0,08	0,08	1,12
Código 6	2315	2712	0,23	0,27	0,85
Código 7	3093	2575	0,31	0,26	1,20
Total	10000	10000	1	1	

CEP RESIDENCIAL	Bom	Mau	% Bom	% Mau	RR
Faixa 1	447	718	0,04	0,07	0,62
Faixa 2	1021	1267	0,10	0,13	0,81
Faixa 3	4719	4943	0,47	0,49	0,95
Faixa 4	1724	1542	0,17	0,15	1,12
Faixa 5	2089	1530	0,21	0,15	1,37
Total	10000	10000	1	1	

QUANTIDADE DE PARCELAS	Bom	Mau	% Bom	% Mau	RR
Até 4	2726	707	0,27	0,07	3,86
5 ou 6	2794	1997	0,28	0,20	1,40
7 a 9	2280	3841	0,23	0,38	0,59
10 a 12	2200	3455	0,22	0,35	0,64
Total	10000	10000	1	1	

CEP COMERCIAL	Bom	Mau	% Bom	% Mau	RR
Faixa 1	691	1070	0,07	0,11	0,65
Faixa 2	3279	3766	0,33	0,38	0,87
Faixa 3	2135	2041	0,21	0,20	1,05
Faixa 4	2334	1979	0,23	0,20	1,18
Faixa 5	1561	1144	0,16	0,11	1,36
Total	10000	10000	1	1	

SALÁRIO DO CLIENTE[9]	Bom	Mau	% Bom	% Mau	RR
Até 650 reais	1740	2185	0,17	0,22	0,80
Acima de 650 a 950 reais	1939	2145	0,19	0,21	0,90
Acima de 950 a 1575 reais	3033	2974	0,30	0,30	1,02
Acima de 1575 a 2015 reais	1032	955	0,10	0,10	1,08
Acima de 2015 a 3000 reais	1093	922	0,11	0,09	1,19
Acima de 3000 reais	1162	818	0,12	0,08	1,42
Total	9999	9999	1	1	

[9] Dois registros estavam com a variável salário em branco

APÊNDICE B – CÁLCULO DO KS

Para o cálculo dos valores de Kolmogorov-Smirnov, os escores obtidos foram padronizados no intervalo de 0 a 1. Em seguida foram definidos vinte intervalos de escore e calculado o KS para cada amostra em cada técnica. O valor do KS é o número em negrito na última coluna.

REGRESSÃO LOGÍSTICA - TREINAMENTO

	Número de clientes		Freqüência Acumulada		
Faixa de pontos	Bons	Maus	Bons	Maus	Diferença
0,95 |-- 1,00	12		0%	0%	0%
0,90 |-- 0,95	117	15	3%	0%	3%
0,85 |-- 0,90	256	32	10%	1%	8%
0,80 |-- 0,85	290	60	17%	3%	14%
0,75 |-- 0,80	388	113	27%	6%	21%
0,70 |-- 0,75	379	168	36%	10%	26%
0,65 |-- 0,70	384	183	46%	14%	31%
0,60 |-- 0,65	359	206	55%	19%	35%
0,55 |-- 0,60	332	245	63%	26%	37%
0,50 |-- 0,55	316	272	71%	32%	**38%**
0,45 |-- 0,50	280	299	78%	40%	38%
0,40 |-- 0,45	218	333	83%	48%	35%
0,35 |-- 0,40	212	323	89%	56%	32%
0,30 |-- 0,35	142	280	92%	63%	29%
0,25 |-- 0,30	117	270	95%	70%	25%
0,20 |-- 0,25	90	281	97%	77%	20%
0,15 |-- 0,20	47	321	98%	85%	13%
0,10 |-- 0,15	38	304	99%	93%	7%
0,05 |-- 0,10	23	266	100%	99%	1%
0,00 |-- 0,05		29	100%	100%	0%

REGRESSÃO LOGÍSTICA - VALIDAÇÃO

	Número de clientes		Freqüência Acumulada		
Faixa de pontos	Bons	Maus	Bons	Maus	Diferença
0,95 |-- 1,00	7	2	0%	0%	0%
0,90 |-- 0,95	77	12	3%	0%	2%
0,85 |-- 0,90	156	32	8%	2%	6%
0,80 |-- 0,85	227	57	16%	3%	12%
0,75 |-- 0,80	271	94	25%	7%	18%
0,70 |-- 0,75	287	115	34%	10%	24%
0,65 |-- 0,70	284	169	44%	16%	28%
0,60 |-- 0,65	293	183	53%	22%	31%
0,55 |-- 0,60	271	202	62%	29%	34%
0,50 |-- 0,55	238	212	70%	36%	34%
0,45 |-- 0,50	203	193	77%	42%	**35%**
0,40 |-- 0,45	166	217	83%	50%	33%
0,35 |-- 0,40	161	255	88%	58%	30%
0,30 |-- 0,35	115	216	92%	65%	27%
0,25 |-- 0,30	89	217	95%	73%	22%
0,20 |-- 0,25	59	243	97%	81%	16%
0,15 |-- 0,20	58	226	99%	88%	11%
0,10 |-- 0,15	27	202	100%	95%	5%
0,05 |-- 0,10	11	147	100%	100%	0%
0,00 |-- 0,05		6	100%	100%	0%

REGRESSÃO LOGÍSTICA - TESTE

	Número de clientes		Freqüência Acumulada		
Faixa de pontos	Bons	Maus	Bons	Maus	Diferença
0,95 |-- 1,00	7	2	0%	0%	0%
0,90 |-- 0,95	91	8	3%	0%	3%
0,85 |-- 0,90	183	34	9%	1%	8%
0,80 |-- 0,85	200	66	16%	4%	12%
0,75 |-- 0,80	278	113	25%	7%	18%
0,70 |-- 0,75	314	109	36%	11%	25%
0,65 |-- 0,70	296	142	46%	16%	30%
0,60 |-- 0,65	266	160	55%	21%	33%
0,55 |-- 0,60	272	216	64%	28%	35%
0,50 |-- 0,55	252	209	72%	35%	37%
0,45 |-- 0,50	239	238	80%	43%	**37%**
0,40 |-- 0,45	142	233	85%	51%	34%
0,35 |-- 0,40	153	250	90%	59%	30%
0,30 |-- 0,35	92	236	93%	67%	26%
0,25 |-- 0,30	74	215	95%	74%	21%
0,20 |-- 0,25	75	220	98%	82%	16%
0,15 |-- 0,20	39	200	99%	88%	11%
0,10 |-- 0,15	22	210	100%	95%	4%
0,05 |-- 0,10	5	133	100%	100%	0%
0,00 |-- 0,05		6	100%	100%	0%

REDE NEURAL - TREINAMENTO

	Número de clientes		Freqüência Acumulada		
Faixa de pontos	Bons	Maus	Bons	Maus	Diferença
0,95 |-- 1,00	181	7	5%	0%	4%
0,90 |-- 0,95	94	7	7%	0%	7%
0,85 |-- 0,90	62	7	8%	1%	8%
0,80 |-- 0,85	66	17	10%	1%	9%
0,75 |-- 0,80	145	46	14%	2%	12%
0,70 |-- 0,75	746	269	32%	9%	24%
0,65 |-- 0,70	805	382	52%	18%	34%
0,60 |-- 0,65	448	284	64%	25%	38%
0,55 |-- 0,60	233	213	70%	31%	**39%**
0,50 |-- 0,55	199	198	74%	36%	39%
0,45 |-- 0,50	183	184	79%	40%	39%
0,40 |-- 0,45	148	213	83%	46%	37%
0,35 |-- 0,40	146	220	86%	51%	35%
0,30 |-- 0,35	141	245	90%	57%	33%
0,25 |-- 0,30	115	270	93%	64%	29%
0,20 |-- 0,25	164	514	97%	77%	20%
0,15 |-- 0,20	92	544	99%	91%	9%
0,10 |-- 0,15	24	115	100%	93%	6%
0,05 |-- 0,10	4	92	100%	96%	4%
0,00 |-- 0,05	4	173	100%	100%	0%

REDE NEURAL - VALIDAÇÃO

	Número de clientes		Freqüência Acumulada		
Faixa de pontos	Bons	Maus	Bons	Maus	Diferença
0,95 |-- 1,00	117	28	4%	1%	3%
0,90 |-- 0,95	46	19	5%	2%	4%
0,85 |-- 0,90	49	9	7%	2%	5%
0,80 |-- 0,85	73	11	10%	2%	7%
0,75 |-- 0,80	101	32	13%	3%	10%
0,70 |-- 0,75	521	195	30%	10%	20%
0,65 |-- 0,70	628	344	51%	21%	30%
0,60 |-- 0,65	325	233	62%	29%	33%
0,55 |-- 0,60	196	153	69%	34%	34%
0,50 |-- 0,55	180	153	75%	39%	**35%**
0,45 |-- 0,50	115	161	78%	45%	34%
0,40 |-- 0,45	100	135	82%	49%	33%
0,35 |-- 0,40	118	147	86%	54%	32%
0,30 |-- 0,35	87	184	89%	60%	28%
0,25 |-- 0,30	103	207	92%	67%	25%
0,20 |-- 0,25	129	406	96%	81%	16%
0,15 |-- 0,20	75	368	99%	93%	6%
0,10 |-- 0,15	12	67	99%	95%	4%
0,05 |-- 0,10	15	57	100%	97%	3%
0,00 |-- 0,05	10	91	100%	100%	0%

REDE NEURAL - TESTE

	Número de clientes		Freqüência Acumulada		
Faixa de pontos	Bons	Maus	Bons	Maus	Diferença
0,95 |-- 1,00	112	23	4%	1%	3%
0,90 |-- 0,95	57	18	6%	1%	4%
0,85 |-- 0,90	54	14	7%	2%	6%
0,80 |-- 0,85	69	15	10%	2%	7%
0,75 |-- 0,80	99	26	13%	3%	10%
0,70 |-- 0,75	575	202	32%	10%	22%
0,65 |-- 0,70	623	312	53%	20%	33%
0,60 |-- 0,65	292	237	63%	28%	34%
0,55 |-- 0,60	211	183	70%	34%	35%
0,50 |-- 0,55	163	163	75%	40%	35%
0,45 |-- 0,50	144	153	80%	45%	**35%**
0,40 |-- 0,45	118	154	84%	50%	34%
0,35 |-- 0,40	98	150	87%	55%	32%
0,30 |-- 0,35	88	166	90%	61%	30%
0,25 |-- 0,30	95	216	93%	68%	26%
0,20 |-- 0,25	134	406	98%	81%	16%
0,15 |-- 0,20	45	348	99%	93%	6%
0,10 |-- 0,15	8	81	100%	96%	4%
0,05 |-- 0,10	11	53	100%	97%	3%
0,00 |-- 0,05	4	80	100%	100%	0%

ALGORITMO GENÉTICO - TREINAMENTO

	Número de clientes		Freqüência Acumulada		
Faixa de pontos	Bons	Maus	Bons	Maus	Diferença
0,95 |-- 1,00	6	0	0%	0%	0%
0,90 |-- 0,95	21	1	1%	0%	1%
0,85 |-- 0,90	58	4	2%	0%	2%
0,80 |-- 0,85	122	25	5%	1%	4%
0,75 |-- 0,80	216	60	11%	2%	8%
0,70 |-- 0,75	400	108	21%	5%	16%
0,65 |-- 0,70	457	183	32%	10%	22%
0,60 |-- 0,65	535	268	45%	16%	29%
0,55 |-- 0,60	560	352	59%	25%	**34%**
0,50 |-- 0,55	458	464	71%	37%	34%
0,45 |-- 0,50	400	529	81%	50%	31%
0,40 |-- 0,45	295	497	88%	62%	26%
0,35 |-- 0,40	207	434	93%	73%	20%
0,30 |-- 0,35	141	384	97%	83%	14%
0,25 |-- 0,30	72	275	99%	90%	9%
0,20 |-- 0,25	32	198	100%	95%	5%
0,15 |-- 0,20	14	125	100%	98%	2%
0,10 |-- 0,15	3	57	100%	99%	1%
0,05 |-- 0,10	3	30	100%	100%	0%
0,00 |-- 0,05	0	6	100%	100%	0%

ALGORITMO GENÉTICO - VALIDAÇÃO

	Número de clientes		Freqüência Acumulada		
Faixa de pontos	Bons	Maus	Bons	Maus	Diferença
0,95 |-- 1,00	2	0	0%	0%	0%
0,90 |-- 0,95	4	0	0%	0%	0%
0,85 |-- 0,90	14	0	1%	0%	1%
0,80 |-- 0,85	39	9	2%	0%	2%
0,75 |-- 0,80	86	23	5%	1%	4%
0,70 |-- 0,75	158	52	10%	3%	7%
0,65 |-- 0,70	284	93	20%	6%	14%
0,60 |-- 0,65	367	183	32%	12%	20%
0,55 |-- 0,60	417	253	46%	20%	25%
0,50 |-- 0,55	385	279	59%	30%	29%
0,45 |-- 0,50	381	353	71%	42%	**30%**
0,40 |-- 0,45	322	394	82%	55%	27%
0,35 |-- 0,40	244	358	90%	67%	24%
0,30 |-- 0,35	128	364	94%	79%	16%
0,25 |-- 0,30	93	269	97%	88%	10%
0,20 |-- 0,25	49	189	99%	94%	5%
0,15 |-- 0,20	21	111	100%	98%	2%
0,10 |-- 0,15	6	50	100%	99%	1%
0,05 |-- 0,10	0	19	100%	100%	0%
0,00 |-- 0,05	0	1	100%	100%	0%

ALGORITMO GENÉTICO - TESTE

	Número de clientes		Freqüência Acumulada		
Faixa de pontos	Bons	Maus	Bons	Maus	Diferença
0,95 |-- 1,00	3	0	0%	0%	0%
0,90 |-- 0,95	15	3	1%	0%	1%
0,85 |-- 0,90	44	5	2%	0%	2%
0,80 |-- 0,85	80	16	5%	1%	4%
0,75 |-- 0,80	189	59	11%	3%	8%
0,70 |-- 0,75	284	100	21%	6%	14%
0,65 |-- 0,70	348	131	32%	10%	22%
0,60 |-- 0,65	374	205	45%	17%	27%
0,55 |-- 0,60	417	303	58%	27%	31%
0,50 |-- 0,55	377	343	71%	39%	**32%**
0,45 |-- 0,50	310	380	81%	52%	30%
0,40 |-- 0,45	239	383	89%	64%	25%
0,35 |-- 0,40	139	349	94%	76%	18%
0,30 |-- 0,35	102	262	97%	85%	13%
0,25 |-- 0,30	43	194	99%	91%	8%
0,20 |-- 0,25	23	139	100%	96%	4%
0,15 |-- 0,20	7	77	100%	98%	2%
0,10 |-- 0,15	6	30	100%	99%	1%
0,05 |-- 0,10	0	12	100%	100%	0%
0,00 |-- 0,05	0	9	100%	100%	0%