UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
MESTRADO PROFISSIONALIZANTE "MODELAGEM MATEMÁTICA EM FINANÇAS"

Modelos de Credit Scoring:

Regressão Logística e

Redes Neurais

Marcos Hissashi Iguti

Orientado por:

Professor Dr. Gerson Francisco

São Paulo

2005

FICHA CATALOGRÁFICA

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

Iguti, Marcos Hissashi
Credit scoring, um comparativo entre regressão logística e redes neurais / Marcos Hissashi Iguti - São Paulo, 2005.
60 p.

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

1. Crédito 2. Crédito direto ao consumidor 3. Redes neurais I. Universidade de São Paulo. Faculdade de Economia, Administração e Contabilidade. II. Universidade de São Paulo. Instituto de Matemática e Estatística. III. Título.

CDD – 332.7

Modelos de Credit Scoring:

Regressão Logística e

Redes Neurais

Marcos Hissashi Iguti

Dissertação apresentada à Faculdade de Economia, Administração e Contabilidade e ao Instituto de Matemática e Estatística da Universidade de São Paulo para obtenção do Título de Mestre.

Agradecimento

Em primeiro lugar, gostaria de agradecer aos meus pais, Ariosto e Irene, e a minha esposa, Érika, que sempre me apoiaram e incentivaram meus estudos. Gostaria, também, de agradecer ao meu orientador, Prof. Dr. Gerson Francisco, que me inspirou e me encorajou a escolher esse tema ainda pouco explorado, e ao Dr. Fernando Ferreira, que muito me auxiliou devido à sua experiência em inteligência artificial e programação.

Abstract

Credit scoring is a method of evaluating the credit risk of loan applications. It became a popular tool of banks and credit card issuers that lend money directly to consumers, where huge volumes of transactions made high speed and high quality standards an important requirement. In this work, we present and compare two strategies used to develop credit scoring models, the logistic regression and the neural networks.

Sumário

Notações

ma	número de maus clientes aprovados pelo modelo
br	número de bons clientes reprovados pelo modelo
mr	número de maus clientes reprovados pelo modelo
ba	número de bons clientes aprovados pelo
erro M	erro percentual cometido pelo modelo, dentre a população de maus clientes
erro B	erro percentual cometido pelo modelo, dentre a população de bons clientes

Abreviações

AG	algoritmos genéticos
AIC	Akaike information criterion
GLM	generalized linear model
MLP	multilayer perceptron
RL	regressão logística
RNA	redes neurais artificiais
WOE	weights of evidence

1 Introdução

A concessão de crédito tem um papel fundamental na economia de um país. Os investimentos, a produção industrial, a taxa de desemprego e o crescimento do PIB são algumas das variáveis fortemente influenciadas pelo volume de empréstimos e crédito oferecidos pelas instituições financeiras (Dornbush et al., 1991).

No Brasil, até meados da década de 90, o volume de recursos destinado a concessão de crédito era bastante reduzido, os grandes bancos se limitavam a conceder crédito somente para satisfazer as exigências legais. Porém, a partir da implantação do Plano Real, quando iniciou-se um processo de estabilização da economia, ocorreu um rápido crescimento da indústria de crédito.

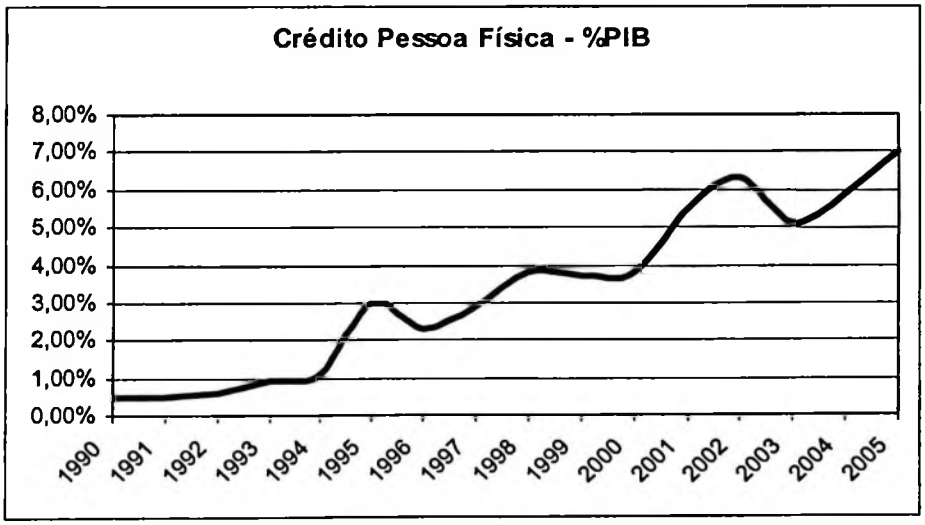


Figura 1.1: *Operações de Crédito (Pessoa Física)*[1]

As instituições financeiras (bancos, financeiras) nacionais e estrangeiras, têm feito maciços investimentos, apostando nesse segmento bastante promissor. Em pouco tempo, o mercado se mostrou extremamente competitivo e ser eficiente tornou-se fundamental. Os

[1] Banco Central do Brasil. Séries temporais. www.bcb.gov.br/?serietemp

modelos de *credit scoring* ganharam importância neste contexto, pois automatizam o processo de avaliação de crédito, tornando-o rápido, barato e padronizado.

O *credit scoring* é uma técnica que auxilia as instituições financeiras a avaliarem os clientes proponentes por crédito. Com base nele, decide-se por conceder ou não, qual o prazo máximo e volume que pode ser emprestado. É possível, também, definir estratégias de forma a maximizar a rentabilidade das financeiras.

Os modelos de *credit scoring* procuram achar uma relação entre as características do cliente e sua possível inadimplência. O sistema atribui uma nota (*score,* em inglês) ao cliente, que o banco utiliza para classificá-lo em termos de risco. O desenvolvimento de um modelo requer uma análise de um banco de dados com o histórico dos empréstimos anteriormente concedidos. Características como salário, tempo de serviço, idade, valor do financiamento e inúmeras outras, são exemplos de informações que, freqüentemente, se utilizam nos modelos. Com base nelas, determina-se quais características mais influenciam no risco de crédito de um cliente.

Inúmeras técnicas estatísticas são utilizadas no desenvolvimento de modelos de *credit scoring*, dentre as principais estão: a regressão logística (Cox, 1970), a análise discriminante (Hand et al., 1998), as redes neurais artificiais (West, 2000), os algoritmos genéticos (Holland, 1975), a programação matemática (Hand, 1981) e as árvores de decisão (Arminger et al., 1997).

O objetivo deste trabalho é comparar dois modelos de *credit scoring*: a regressão logística com as redes neurais artificiais. A regressão logística é um método linear e, provavelmente, o mais comumente usado nos modelos de *credit scoring*. Por outro lado, as redes neurais são técnicas computacionais não-lineares e relativamente recentes, baseadas no funcionamento do cérebro humano.

O trabalho foi organizado da seguinte forma: no Capítulo 2, descrevemos a base de dados, cedida por uma instituição financeira. A metodologia e as estratégias adotadas foram discutidas no Capítulo 3. No Capítulo 4, apresentamos a aplicação prática dos modelos discutidos sobre a base de dados. No Capítulo 5, fazemos uma análise dos resultados e sugerimos alternativas para melhoria do desempenho dos modelos. Por fim, as conclusões são discutidas no Capítulo 6.

1.1 História

A história do *credit scoring* é bastante recente. Os primeiros estudos voltados para identificação de grupos dentro de uma população foram introduzidos na década de 30. Em 1936, Fisher (Fisher, 1936) desenvolveu a análise discriminante, uma técnica estatística onde, a partir das características disponíveis de um indivíduo, cria-se uma regra de classificação que permite inferir a que população ele pertence. Na época, Fisher procurava diferenciar duas variedades de um mesmo tipo de planta através de suas medidas. Seu estudo foi base para diversos outros que se desenvolveram posteriormente.

Até o início do século XX, a avaliação dos clientes se baseava exclusivamente no julgamento dos analistas, o que tornava o processo lento e subjetivo. Nessa mesma época, nos EUA, algumas financeiras e lojas começaram a ter dificuldades na administração de créditos, pois os analistas estavam sendo convocados a participar da 2° Guerra Mundial. Isto impulsionou a demanda por um sistema padronizado de avaliação (Johnson, 1992).

Em 1941, Durand desenvolveu uma pesquisa onde reconhecia que seria possível utilizar a técnica da análise discriminante para classificar os bons e os maus empréstimos. Porém, seu estudo não foi utilizado na prática para a predição de desempenho de créditos concedidos. Somente em 1958, Bill Fair e Earl Isaac[1] introduziram o primeiro modelo de credit scoring para a *American Investments*.

A demanda por cartões de crédito no início dos anos 60 fizeram com que os bancos e as administradoras de cartões compreendessem o real valor das técnicas de *credit scoring*. A quantidade diária de pessoas que solicitavam um cartão tornou inviável a forma "manual" de avaliação. Mas, com o aprimoramento dos computadores, a implementação dos modelos tornou-se possível.

Inicialmente, a substituição dos analistas de crédito pelos modelos estatísticos foi vista com certa desconfiança, pois se pensava que esse tipo de avaliação requeria conhecimento, experiência e intuição, sendo uma tarefa exclusiva de um cérebro humano. Porém, os excelentes resultados alcançados pelos modelos derrubaram os paradigmas. As

[1] http://www.fairisaac.com/Fairisaac/Company/Milestones/

empresas que adotaram os modelos automatizados tiveram um queda de 50% nos índices de inadimplência (Myers et al., 1963).

1.2 Credit Scoring na atualidade

O aprimoramento dos modelos de *credit scoring*, aliado às vantagens que eles proporcionam com relação a custo e velocidade das avaliações, consolidaram sua utilização entre os bancos, financeiras e empresas de cartões de crédito.

Inicialmente, o *credit scoring* foi desenvolvido para avaliação de clientes de cartões de crédito e concessão de empréstimo para pessoas físicas. Porém, sua aplicação não ficou restrita somente nesse setor, descobriu-se diversas outras áreas de atuação (Thomas, 2002). Na área financeira foram criados modelos para avaliação de financiamento de imóveis, empréstimo para pequenas empresas, pedidos por extensão de crédito, prevenção de fraudes e muitas outras. Diferentes áreas, não diretamente ligadas ao crédito, também desenvolveram modelos baseados no *credit scoring*. Algumas empresas utilizam modelos para auxiliar nas decisões de marketing, por exemplo, para prever se um determinado cliente é ou não um potencial consumidor de um produto ou se ele é ou não leal à marca, isto é, trocaria por um produto concorrente. Nos Estados Unidos, ao invés da receita inspecionar toda a população, foi desenvolvido um modelo que seleciona as pessoas que terão seus impostos auditados. E até mesmo em prisões, os modelos têm sido utilizados para definir, entre a população carcerária, aqueles que terão direito a liberdade condicional.

Como se pode observar, o leque de aplicações de modelos como o *credit scoring* é bastante extenso, sendo por isso, objeto de muita pesquisa na atualidade.

2 Descrição do estudo

2.1 Descrição dos dados

Nesta seção descrevemos o conjunto de dados que foi utilizado e mostramos algumas de suas características.

Uma instituição financeira de pequeno porte, especializada em financiamento de automóveis, cedeu para uso na presente dissertação, seu banco de dados com informações cadastrais de seus clientes, juntamente com informações sobre o financiamento e a situação do pagamento das prestações.

Para a realização do estudo comparativo, dispusemos de uma amostra de 1.500 clientes que financiaram seus bens entre 1995 e 1996. As variáveis disponíveis para a sua elaboração foram:

1. Data do financiamento
2. Valor do financiamento
3. Valor das prestações
4. Quantidade de prestações
5. Data de nascimento
6. Data do cadastramento
7. Tempo de residência do cliente
8. Atividade da empresa (21 categorias)
9. Data de admissão
10. Salário, em número de salário mínimos
11. Maior atraso observado das prestações
12. Atividade da empresa
    a. Público
    b. Indústria
    c. Outros

13. Sexo
    a. Masculino
    b. Feminino
14. Estado civil
    a. Solteiro
    b. Casado
    c. Divorciado
    d. Viúvo
    e. Outros
15. Naturalidade
    a. São Paulo
    b. Grande São Paulo
    c. Interior Paulista
    d. Litoral Paulista
    e. Estados do Norte ou Nordeste
    f. Estados do Sul
    g. Grande Sudeste
    h. Grande Centro Oeste
    i. Estrangeiro
16. Tipo de residência
    a. Própria
    b. Alugada
    c. Moradia dos pais
    d. Pensão / Hotel
    e. Funcional
    f. Outros
17. Cheque especial
    a. Possui
    b. Não possui

18. Cartão de crédito
    a. Possui
    b. Não possui

2.2 Definição da variável resposta

O primeiro passo no desenvolvimento de um modelo de *credit scoring* é definir o critério que diferencie os "bons" dos "maus" pagadores. Existem inúmeras formas para se fazer isso, porém o critério mais utilizado é considerar como maus clientes, aqueles que apresentaram atraso em três ou mais prestações consecutivamente, ou seja, 90 dias (Thomas et al., 2000). Os demais poderiam ser classificados como bons clientes, ou ainda, existiria a opção de se criar uma classe intermediária. Por exemplo, seriam classificados como indefinidos aqueles com atraso nas prestações entre 60 e 90 dias, e, somente os que nunca atrasaram mais que 60 dias, seriam considerados como bons.

Devido ao tamanho restrito da base de dados, foram criadas apenas duas classes. Definiu-se como maus clientes aqueles que apresentaram em seu histórico um atraso no pagamento das parcelas maior ou igual a 60 dias e os demais foram classificados como bons. Adotamos esse critério com a finalidade de aumentarmos a incidência de inadimplentes, caso contrário o número de observações seria muito pequeno, o que dificultaria a modelagem.

Tabela 2.1: *Tabela de classificação dos clientes*

Classificação	Qtd. Clientes	%
bons	1177	78,5%
maus	323	21,5%
Total	1500	100,0%

Segundo o critério adotado, das 1500 observações da base de dados, 1177 (78,5%) foram classificados como bons e o restante, 323 (21,5%), como maus clientes.

3 Metodologia

Neste capítulo serão apresentadas as técnicas estatísticas e critérios utilizados no desenvolvimento de modelos de *credit scoring*.

3.1 Tratamento das Variáveis

Existem dois tipos de variáveis, as quantitativas como por exemplo "valor do financiamento", "idade" e "salário" e as qualitativas como "estado civil", "sexo", e "naturalidade".

Um tratamento muito comum que se aplica nos dados quando se desenvolve modelos de *credit scoring,* é a categorização de variáveis quantitativas e também o agrupamento de categorias das variáveis qualitativas.

No caso de variáveis qualitativas, costuma-se fazer o agrupamento de categorias para se evitar que uma determinada classe apresente um número muito pequeno de observações, sendo dessa forma, pouco significativa, ou mesmo, para se eliminar parâmetros desnecessários.

As variáveis quantitativas muitas vezes não apresentam uma relação linear com o risco de crédito. Imagine, por exemplo, que se queira modelar a probabilidade de inadimplência em função do mês de contratação do empréstimo. A variável mês assumiria o valor 1 para expressar o mês de janeiro, 2 para fevereiro e assim por diante. Vamos supor, também, que exista um efeito sazonal da inadimplência, os meses de dezembro e janeiro apresentariam um maior índice, pois são meses em que se concentram gastos com o Natal, férias, tributos, matrícula da escola, etc. Por outro lado, os demais meses apresentariam um índice de inadimplência menor e constante. Portanto, teríamos o mês 1 com alto risco de inadimplência, de 2 a 11 o risco seria baixo e o mês 12 voltaria a ser alto, claramente uma relação não-linear. Uma possível solução para o problema, seria criar uma variável que assuma valor 0 para representar os meses de fevereiro a novembro, e valor 1 os meses janeiro e dezembro. Esse tipo de variável é denominada *dummy* (Gujarati, 1995).

Uma segunda alternativa ao tratamento de variáveis com comportamento não-linear, seria a transformação de variáveis, mas que tornaria o modelo de difícil interpretação, portanto, normalmente prefere-se categorizá-las. Mesmo nos casos em que essa relação seja linear, às vezes, obtém-se melhores resultados quando se efetua a categorização pois, assim, reduz-se a influência de valores discrepantes.

Existem inúmeros métodos para se fazer a categorização de uma variável. A forma mais comum consiste em utilizar como critério a medida de risco relativo (*RR)* ou *Weights of Evidence* (WOE) (Good, 1950). Esse estudo consiste em construir uma tabela similar a descrita abaixo.

Tabela 3.1: *Cálculo do WOE*

Categoria i	Num. Bons	Num. Maus	%Bons	%Maus	RR	WOE
Categoria 1	b1	m1	b1/b	m1/m	(b1/b)/(m1/m)	Log[(b1/b)/(m1/m)]
Categoria 2	b2	m2	b2/b	m2/m	(b2/b)/(m2/m)	Log[(b2/b)/(m2/m)]
Categoria 3	b3	m3	b3/b	m3/m	(b3/b)/(m3/m)	Log[(b3/b)/(m3/m)]
Categoria 4	b4	m4	b4/b	m4/m	(b4/b)/(m4/m)	Log[(b4/b)/(m4/m)]
Categoria 5	b5	m5	b5/b	m5/m	(b5/b)/(m5/m)	Log[(b5/b)/(m5/m)]
Total	b	m	1	1		

bi: número de clientes bons na categoria i

mi: número de clientes ruins na categoria i

O *RR* é o risco relativo de um cliente bom pertencer a categoria *i* em relação ao risco de um mau cliente ser dessa classe. Valores de *RR* maiores que 1 indicam que são categorias mais propensas de serem observadas entre os bons clientes que entre os maus e portanto possuem um risco de crédito menor. Por outro lado, valores de *RR* menores que 1, é uma evidência de que tal característica representa um maior risco de crédito. É comum

também, se utilizar o seu logaritmo, denominado *WOE* , com a vantagem de ser 0 o valor de referência. Nesse caso, teríamos:

- *WOE = 0* : significa que a razão entre bons e ruins é 1, ou seja, não há nenhum indício do cliente ser de maior ou de menor risco comparado à análise sem essa variável.
- *WOE > 0* : indica que o cliente apresenta menor risco de crédito, isso significa que a categoria apresenta algum poder de discriminação.
- *WOE < 0* : indica que o cliente apresenta maior risco de crédito, isso significa que a categoria apresenta algum poder de discriminação..

Citando um exemplo, a variável "estado civil" possui 5 categorias: "solteiro", "casado", "divorciado", "viúvo" e "outros". Calculou-se o índice *WOE* de cada uma delas conforme tabela abaixo.

Tabela 3.2: Exemplo de *cálculo do WOE*

Estado Civil	Num. Bons	Num. Maus	%Bons	%Maus	RR	WOE
Solteiros	286	92	0,24	0,28	0,85	-0,07
Casados	773	194	0,66	0,60	1,09	0,04
Divorciados	45	22	0,04	0,07	0,56	-0,25
Viúvos	26	3	0,02	0,01	2,38	0,38
Outros	47	12	0,04	0,04	1,07	0,03
Total	1177	323	1	1		

Duas categorias das cinco descritas na tabela chamam a atenção pelo valor do *WOE*. A categoria "divorciados" apresenta *WOE=-0,25*, indicando que é um tipo de cliente que apresenta um maior risco de crédito. No outro extremo, está a categoria "viúvos", com *WOE=0,38*, o que indica baixo risco de crédito.

As categorias "divorciados", "viúvos" e "outros" apresentam um número de incidência pequeno e, portanto, decidiu-se por reagrupar as 5 classes da seguinte forma:

- Categoria 1: "solteiros + divorciados", pois as duas categorias apresentavam WOE<0
- Categoria 2: "casados"
- Categoria 3: "viúvos + outros", pois as duas categorias apresentavam WOE>0

É importante ressaltar que existem inúmeras outras formas de se realizar a categorização, poderíamos por exemplo manter todas as cinco, ou então agrupá-las de forma a ter somente 2 classes.

A descrição das tabelas de categorização de todas as variáveis estão no Apêndice A.

3.2 Modelos

Iremos abordar nessa seção duas metodologias bastante utilizadas no desenvolvimento de modelos de *credit scoring*: regressão logística e redes neurais artificiais (RNA)

A regressão logística (Cox, 1970) é considerada um dos principais métodos de modelagem estatística de dados. Isso se deve, principalmente, a facilidade de interpretação dos parâmetros de um modelo e também pelo fato de estar implementada em quase todos os aplicativos estatísticos.

As redes neurais artificiais (Bishop, 1995) têm como principal propriedade o poder de modelar processos não lineares, permitindo que sejam capazes de reproduzir e detectar padrões complexos, que é o caso da inadimplência de clientes.

3.2.1 Regressão Logística

A regressão logística é um caso particular dos Modelos Lineares Generalizados (McCullagh e Nelder, 1989).

Ela é muito utilizado em modelagens em que a variável resposta é qualitativa com dois resultados possíveis, por exemplo sobrevivência de enxertos de ameixeira (sobrevive ou não sobrevive) ou desempenho de um empréstimo concedido (adimplente ou inadimplente). O uso da regressão logística pode ser tanto para fins descritivos, quando se quer explicar o relacionamento entre a variável resposta e as variáveis regressoras, como para fins preditivos.

O modelo de regressão logística pode ser expresso como:

$$\ln\left(\frac{\mu_i}{1-\mu_i}\right) = x_i^T\beta \quad \text{ou} \quad \mu_i = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}} \tag{3.1}$$

sendo que, μ_i é a probabilidade de sucesso, $x_i = (1, x_{i1}, x_{i2},..., x_{ip})^T$ é o vetor de dimensão $p+1$ de variáveis preditoras do indivíduo, i $(i=1,...,n)$, onde n é o tamanho da amostra, e $\beta=(\beta_0,\beta_1,....,\beta_p)^T$ é o vetor de parâmetros do modelo.

Uma propriedade interessante da regressão logística é a interpretação direta dos parâmetros. Por exemplo, no caso mais simples temos:

$$\ln\left(\frac{\mu(x_1)}{1-\mu(x_1)}\right) = \beta_0 + \beta_1 x_1 \tag{3.2}$$

$$\ln\left(\frac{\mu(x_1+1)}{1-\mu(x_1+1)}\right) = \beta_0 + \beta_1 (x_1+1) \tag{3.3}$$

Para facilitar a notação, denominamos:

$$chance(x_1) = \frac{\mu(x_1)}{1-\mu(x_1)} \text{ e } chance(x_1+1) = \frac{\mu(x_1+1)}{1-\mu(x_1+1)} \tag{3.4}$$

É fácil provar que:

$$\ln\left(\frac{chance(x_1+1)}{chance(x_1)}\right) = \beta_1 \tag{3.5}$$

Portanto, o aumento em uma unidade de x_l, aumenta-se o logaritmo da razão de chances (também chamado de logit de μ) em β_l unidades.

Os parâmetros do modelo são geralmente estimados maximizando-se a função de verossimilhança. A verossimilhança de uma amostra aleatória de variáveis binárias independentes de média μ_i é dada por:

$$L(Y/\beta)=\prod_{i=1}^{n}\mu_i^{y_i}(1-\mu_i)^{1-y_i} \qquad (3.6)$$

onde, $\mu_i=\left(\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right)$, na regressão logística. As soluções das equações são obtidas por métodos iterativos (Paula, 2000).

3.2.2 Redes Neurais

As redes neurais foram originalmente desenvolvidas com o intuito de modelar a comunicação e o processamento de informações do cérebro humano. Nele existem os dentritos que são dispositivos de entrada, conduzindo os sinais das extremidades para o corpo celular, e os axônios que são responsáveis por transmitir o sinal do corpo celular para suas extremidades. Analogamente ao cérebro, a rede neural consiste em uma entrada de sinais (dentritos), seguida do processamento dos mesmos (corpo celular) e transmissão de informação (axônios).

Um neurônio é uma unidade de processamento de informação que é fundamental para operações de uma rede neural. A figura 3.1 ilustra o modelo de um neurônio que é base para um projeto de redes neurais.

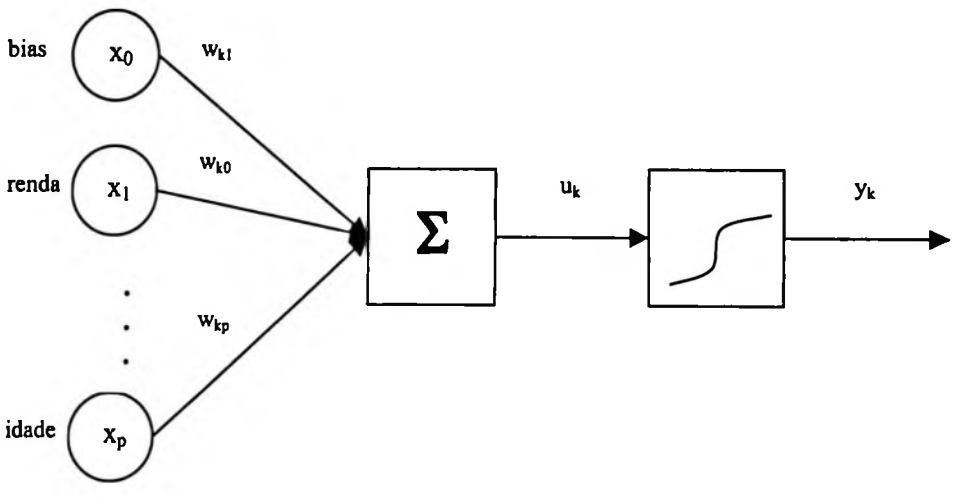


Figura 3.1: *Modelo de um neurônio*

Podemos representar um neurônio algebricamente como:

$$u_k = w_{k0}x_0 + w_{k1}x_1 + ... + w_{kp}x_p = \sum_{q=0}^{p} w_{kq}x_q \tag{3.7}$$

$$y_k = F(u_k) \tag{3.8}$$

Cada um dos $x_1,...,x_p$ é uma variável (por exemplo, característica do candidato ao crédito), assumindo um valor chamado sinal. Os pesos são representados pelos w_{kp}. Os índices k indica o neurônio em que o peso é aplicado e p indica a variável em questão. Numa rede neural de camada única, k=1 pois existe somente um neurônio.

Em seguida, os valores u_k são submetidos a uma função, denominada função de transferência. O software utilizado neste estudo (*Netlab*), possui três diferentes tipos de funções implementadas: a linear, a logística e a *softmax* (Nabney, 2003). A função utilizada nos modelos aqui desenvolvidos foi a logística, pois são indicadas para problemas de classificação, que é o caso dos modelos de *credit score*.

A função logística é definida por:

$$F(u) = \frac{1}{1 + e^{-au}} \tag{3.9}$$

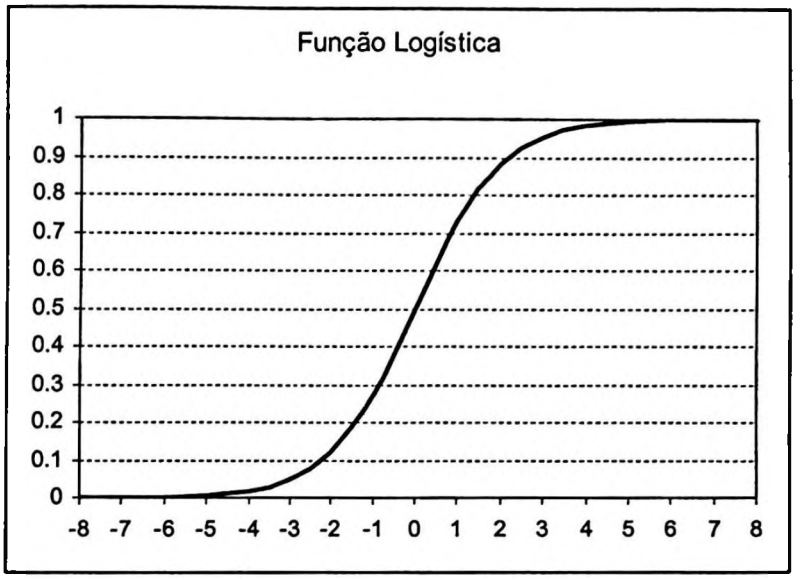


Figura 3.2: *Gráfico da função logística*

Um modelo de um único neurônio juntamente com uma função de transferência é denominado um *perceptron*. Demonstrou-se que nos casos lineares, o algoritmo de redes neurais de camada única converge quando estabelecidos pesos apropriados (Rosenblatt, 1958). Porém, Minsky e Papert (1969) concluíram que os *perceptrons* não são capazes de classificar casos não-lineares.

Em 1986, Rumelhart, Hinton e Williams (1986a, 1986b) demonstraram que as redes neurais podem classificar casos não-lineares usando *perceptrons* de múltiplas camadas (*multilayer perceptrons*) e função de transformação também não-linear.

Um MLP (*multilayer perceptron*) consiste em uma camada de entrada de sinais, uma de saída e de camadas intermediárias de neurônios (*hidden layer*). Cada neurônio da camada intermediária possui uma série de pesos distintos que são aplicados aos sinais de entrada, em seguida são somados, e aplicando-se, por fim, a função ativadora. As saídas dessa camada se tornam entradas para a camada seguinte e o processo se repete. Uma rede com três camadas (*three-layer network*) está ilustrada na figura 3.3.

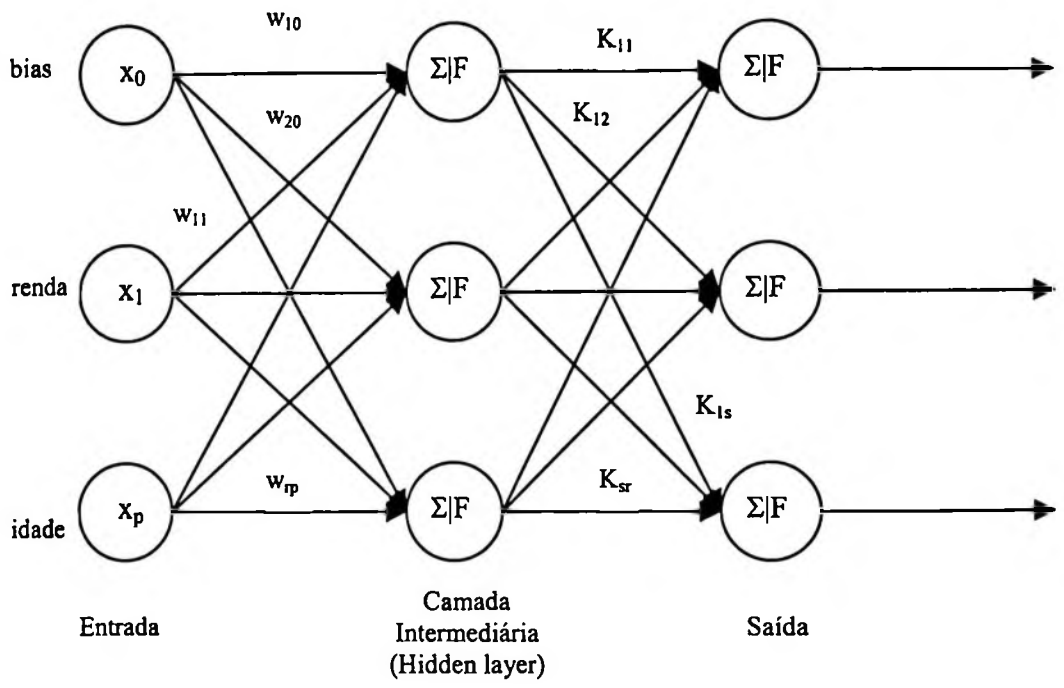


Figura 3.3. *A multilayer perceptron*

No exemplo acima temos p parâmetros de entrada, r neurônios na camada intermediária e s neurônios na camada de saída. Assim, pode-se representar algebricamente um *perceptron* de múltiplas camadas:

$$y_k = F_1\left(\sum_{q=0}^{p} w_{kq} x_q\right) \tag{3.10}$$

O índice 1 em F indica que é a primeira camada após a camada de input. Os y_k, $(k=1,...,r)$, são as saídas da primeira camada intermediária. A saída de uma camada se torna entrada da camada posterior, então:

$$z_v = F_2\left(\sum_{k=1}^{r} K_{vk} y_k\right) = F_2\left(\sum_{q=0}^{r} K_{vk}\left(F_1\left(\sum_{q=0}^{p} w_{kq} x_q\right)\right)\right) \tag{3.11}$$

onde z_v é a resposta do neurônio v na camada de saída, $(v=1,...,s)$, F_2 é a função ativadora e K_{vk} é o peso aplicado ao y_k layer que une o neurônio k da camada intermediária ao neurônio v na camada de saída.

3.3 Definição das variáveis

Um das maiores dificuldades no desenvolvimento de modelos é a seleção das variáveis. Normalmente, dispomos de um número muito grande de informações sobre os clientes e muitas delas podem não ser fundamentais para a modelagem, sendo somente uma fonte de ruído para o modelo, ou ainda, podem ser variáveis redundantes, que apresentem alta correlação com alguma outra. Somadas a isso, um modelo com menos variáveis torna a sua interpretação mais fácil.

Existem diversos procedimentos que podem ser utilizados na definição das variáveis explicativas mais relevantes (Neter et al., 1996). Os procedimentos mais conhecidos são maior R^2, menor s^2, *forward*, *backward*, *stepswise* e *AIC* (*Akaike Information Criterion*). Nenhum deles é comprovadamente superior aos demais, pois, na realidade, não existe um único grupo de variáveis que possa ser considerado como melhor. A seleção de variáveis é um processo pragmático e bastante subjetivo.

Para o estudo realizado, adotou-se o *AIC*. Este método, proposto por Akaike (1974), procura selecionar um modelo bem ajustado e com um número reduzido de parâmetros. Como o máximo do logaritmo da função verossimilhança $L(\beta)$ aumenta com o número de parâmetros do modelo, procura-se encontrar aquele que minimize o valor da função:

$$AIC = -2L(\beta) + 2p \tag{3.12}$$

onde, $L(\beta)$ é o logaritmo da função de verossimilhança e p é o número de parâmetros do modelo considerado.

3.4 Critérios para nota de corte

O *score* que divide a base de dados entre bons e maus clientes é chamado de nota de corte. Sua definição depende da estratégia e das características de cada instituição. Uma instituição especializada em financiamento de automóveis pode, por exemplo, ter menor aversão ao risco em relação a uma outra que financie viagens. Caso o cliente se torne

inadimplente a financeira pode recuperar o bem, diminuindo consideravelmente sua perda, diferentemente do que ocorre com os financiamento de viagens.

Portanto, um fator importante que deve ser levado em conta na escolha da nota de corte é a assimetria dos custos dos erros. A perda que se incorre ao se classificar inadequadamente um mau cliente é, normalmente, maior que a quantia que se deixa de ganhar com a reprovação de um bom cliente. Isso ocorre pois, quando a instituição concede empréstimo para um cliente e este deixa de pagar as prestações, a instituição pode ter perda do principal ou de parte dele. Por outro lado, o caso de não se aprovar o crédito para um bom cliente, a instituição só estaria deixando de ganhar com os encargos relativos ao empréstimo. Um critério freqüentemente utilizado é o de minimizar a perda média por cliente.

$$P_m = \frac{ma \cdot P_1 + br \cdot P_2}{T} \tag{3.13}$$

onde,

P_m : perda média por cliente

ma : número de maus clientes aprovados pelo modelo

br : número de bons clientes reprovados pelo modelo

T : total de clientes avaliados

P_1 : perda causada quando se aprova um mau cliente

P_2 : perda causada quando se reprova um bom cliente

A perda média mínima é graficamente identificada, traçando-se um gráfico de perda média em função da nota de corte. Observe um exemplo a seguir:

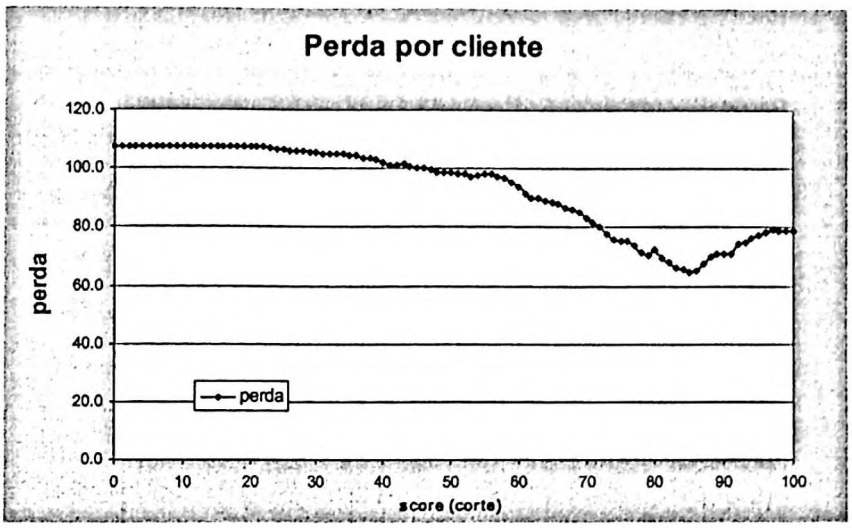


Figura 3.4. *Perda média por cliente (exemplo)*

No gráfico acima, observa-se que uma nota de corte muito baixa acarretará numa taxa de aprovação de maus cliente muito elevada, e, consequentemente, um grande prejuízo. A medida que se aumenta essa nota, ocorre uma diminuição do erro de aprovação de maus clientes mas também passa-se a reprovar os bons clientes. Existe portanto um ponto ótimo, onde a perda média por cliente é mínima. No exemplo acima, esse ponto está em 85.

Alternativamente, ao invés de se definir um único ponto, pode-se definir um intervalo onde os clientes que estiverem nela, são submetidos a uma avaliação mais rigorosa. Por exemplo, seriam aprovados os clientes com *score* maior que 90, reprovados os que apresentem *score* menor que 80, e aqueles que estiverem dentro desta faixa seriam submetidos a uma investigação mais minuciosa.

4 Aplicação

Nesse estudo comparativo utilizamos 3 estratégias para ajustar os modelos:

- Estratégia 1: dados na forma natural
- Estratégia 2: dados parcialmente categorizados
- Estratégia 3: dados totalmente categorizados

Para cada uma das estratégias foram ajustados modelos em regressão logística e em redes neurais, saturados e não saturados. Os modelos não saturados, tiveram suas variáveis de entrada selecionadas com base com critério de Akaike (*AIC*, em inglês). No total, 12 modelos foram ajustados.

A categorização das variáveis foi feita utilizando como critério o *WOE*. Calculou-se o *WOE* de cada um dos atributos das variáveis qualitativas. Em seguida, as classes que apresentavam um risco de crédito semelhante e exibiam um número muito pequeno de observações foram agrupadas. No caso das quantitativas, a mesma técnica foi aplicada, procurou-se, dentre as diversas faixas de valores, as três que apresentavam o maior *WOE* em módulo. As tabelas contendo as categorias e respectivos *WOE* de cada variável encontram-se no Apêndice A

A resposta de uma modelagem logística é um valor entre 0 e 1. Convencionamos, trabalhar com *score* entre 0 e 100 e para isso multiplicou-se os resultados dos modelos por 100.

O ajuste dos modelos em redes neurais foi feito utilizando o software *Matlab 6.5* juntamente com as funções desenvolvidas pelo *Neural Computing Research Group*, da Aston University[1]. O *Splus 4.5*, acompanhado da biblioteca "*mass*", onde se encontram os algoritmos do *AIC*, foi o aplicativo usado no ajuste dos modelos em regressão logística.

[1] www.ncrg.aston.ac.uk

Os procedimentos de ajuste dos modelos serão descrito em detalhes somente para a estratégia 1. Para as demais estratégias, os procedimentos são exatamente os mesmos, e portanto, não serão explicitados.

4.1 Estratégia 1

Foram utilizadas nesta estratégia os dados na sua forma natural, isto é, nenhuma das variáveis quantitativas foi categorizada. As seguintes variáveis foram empregadas:

Variáveis quantitativas:

1. Valor de financiamento
2. Valor das prestações
3. Quantidade de prestações
4. Tempo de residência do cliente em número de anos
5. Idade
6. Tempo de serviço
7. Salário

Variáveis categóricas:

8. Tipo de Residência
   a. Própria
   b. Alugada, funcional ou dos pais
   c. Outros
9. Estado Civil
   a. Casado
   b. Solteiro ou desquitado
   c. Outros

10. Sexo
    a. Masculino
    b. Feminino
11. Naturalidade
    a. Cidade de São Paulo
    b. Estado de SP
    c. Região Norte, Sudeste, Centro-Oeste e estrangeiros
    d. Outros
12. Cheque especial
    a. Possui
    b. Não possui
13. Cartão de crédito
    a. Possui
    b. Não possui

4.1.1 Regressão Logística

Descrevemos nessa seção a seqüência de execução do ajuste de um modelo saturado e não saturado em regressão logística.

A base de dados foi aleatoriamente dividida em dois grupos:

- grupo 1 – 2/3 da base (1000 observações) para o ajuste do modelo
- grupo 2 - 1/3 da base (500 observações) para teste do modelo

Ajustou-se o modelo utilizando todas as variáveis disponíveis (modelo saturado), com os dados do grupo 1, com a auxílio do software *Splus 4.5*.

Em seguida foi definida a nota de corte. Para tanto, construiu-se um gráfico de perda média por cliente em função da nota de corte. Arbitrariamente, adotamos os seguintes critérios:

- perda de $500 quando o modelo aprova um mau cliente,
- perda de $100 quando o modelo reprova um bom cliente

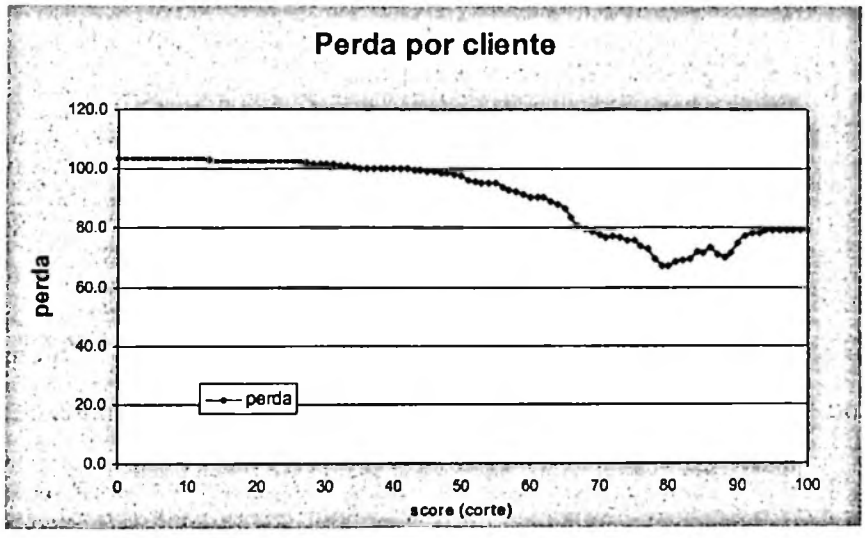


Figura 4.1: *Perda média por cliente*

A nota de corte é a mínima perda observada no gráfico. Neste caso o ponto de mínimo foi atingido com a nota de corte sendo 80.

Posteriormente, aplica-se o modelo ajustado sobre os dados de teste (grupo 2).

O desempenho do modelo é medido da seguinte forma:

$erroM = \frac{ma}{M_{total}}$, erro cometido entre os maus clientes

onde,

ma : número de maus clientes aprovados pelo modelo

M_{total} : número total de maus clientes

$erroB = \frac{br}{B_{total}}$, erro cometido entre os bons clientes

onde,

br : número de bons clientes reprovados pelo modelo

B_{total} : número total de bons clientes

O resultado do ajuste do modelo está demonstrado abaixo:

Tabela 4.1: *RL (saturado)*

observado	Predito		erro	erro total
	bom	Mau		
bom	236	147	38,4%	39,8%
mau	52	65	44,4%	

De um total de 383 bons clientes, o modelo aprovou 236 e reprovou 147 clientes, cometendo um erro de 38,4%. Dentre os maus clientes, o modelo reprovou 65 e aprovou 52, nesse caso a taxa de erro foi de 44,4%.

O passo seguinte foi ajustar novamente um modelo em regressão logística, porém dessa vez, selecionando as variáveis através do critério de Akaike. Pelo método, foram selecionadas 6 variáveis:

- Valor do financiamento
- Tempo de residência
- Sexo (dummy)
- Natural de São Paulo (dummy)
- Residência com os pais ou residência alugada (dummy)
- Servidor público (dummy)

Da mesma forma como foi feito anteriormente, definimos a nota de corte através do gráfico de perda média por cliente.

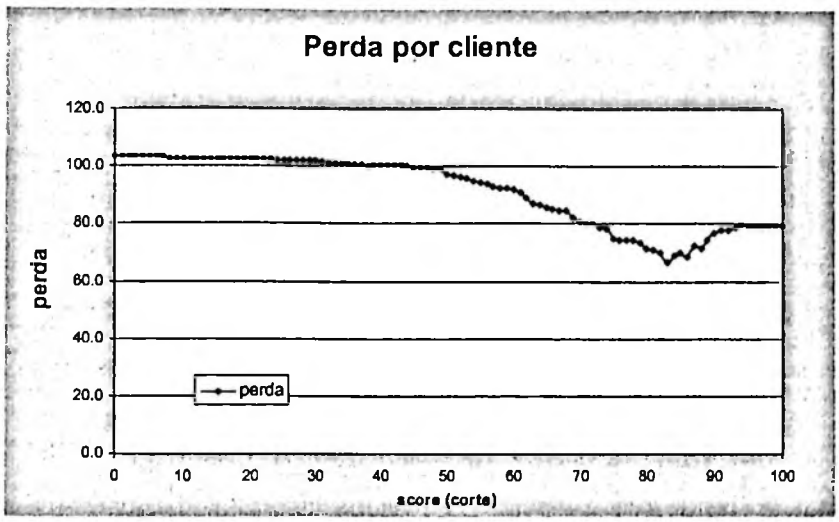


Figura 4.2: *Perda média por cliente*

Adotou-se nota de corte de 83, ponto de mínima perda por cliente, conforme gráfico acima. Em seguida, aplicou-se o modelo sobre o dados de teste.

Tabela 4.2: *RL (não saturado)*

observado	predito		erro	erro
	bom	mau		total
bom	174	209	54,6%	50,2%
mau	42	75	35,9%	

O modelo reprovou 209 clientes dentre os 383 bons clientes, cometendo um erro de 54,6%. No caso da amostra de maus clientes o erro foi de 35,9%.

4.1.2 Redes Neurais

Descrevemos nessa seção a seqüência de execução do ajuste de um modelo em redes neurais.

A base de dados foi aleatoriamente dividida em três grupos:

- grupo 1 - 20% da base (300 observações) para a determinação do número de neurônios

- grupo 2 - 50% da base (750 observações) para o treinamento da rede
- grupo 3 - 30% da base (450 observações) para teste do modelo.

A arquitetura utilizada foi a *Multi Layer Perceptron*, com a função logística e algoritmo de otimização *scaled gradient.*

O primeiro passo no desenvolvimento do modelo é definir o número de neurônios da camada interna. Utilizando-se a base de dados do grupo 1, ajustou-se modelos com número de neurônios variando de 1 até 30, determinando-se os erros quadráticos individualmente. O ajuste e o cálculo do erro são refeitos várias vezes em cada ponto, para que se chegue num erro quadrático médio, e assim, se tenha um gráfico com uma curva mais suave.

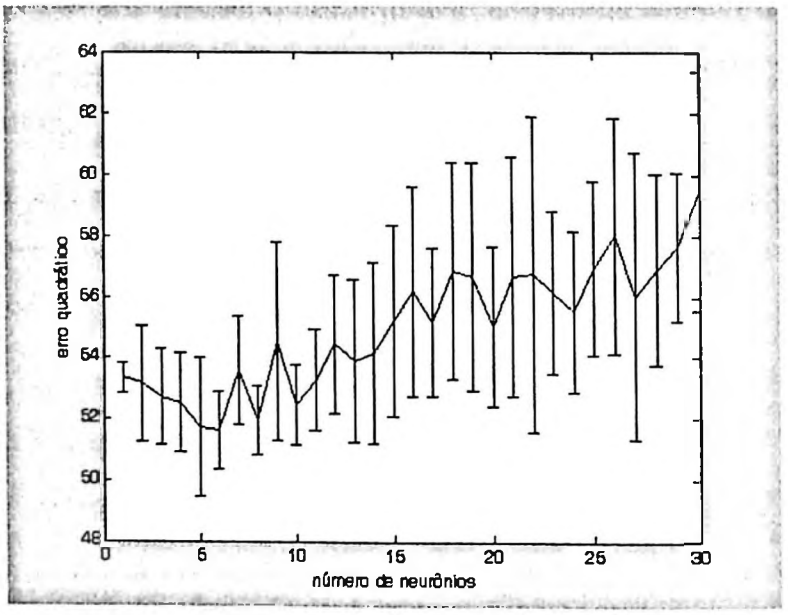


Figura 4.3: *Erro quadrático*

As barras verticais indicam o desvio padrão do erro quadrático de cada um dos pontos. O número de neurônios que será adotado no modelo, é o de menor erro observado. Conforme o gráfico acima, o mínimo é atingido no ponto 6.

Portanto, adotando-se o número de neurônio na camada intermediária sendo 6, ajusta-se um novo modelo com base nos dados do grupo 2 (treinamento). Traçamos, em seguida, o gráfico de perda por cliente.

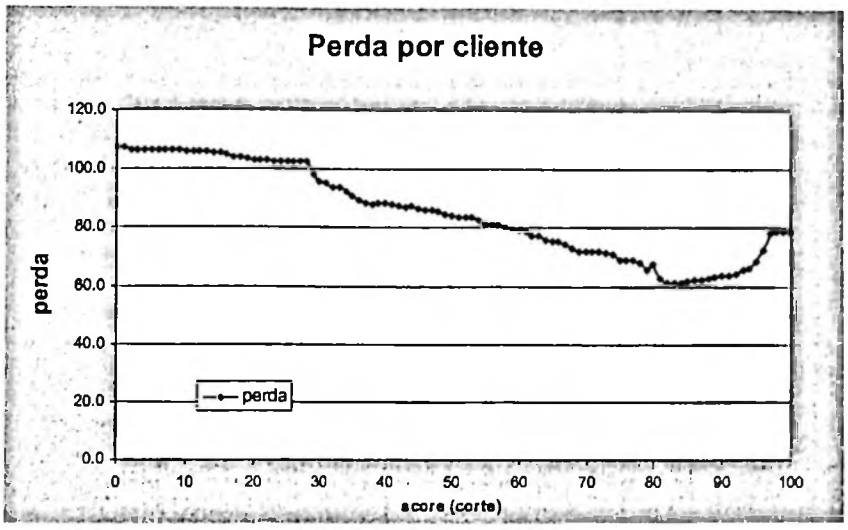


Figura 4.4: *Perda média por cliente*

A perda mínima é atingida no ponto onde a nota de corte é 82, conforme pode ser observada no gráfico acima.

Aplicamos o modelo sobre os dados de teste, com os pesos, arquitetura e nota de corte anteriormente definidos. O seguinte resultado foi observado:

Tabela 4.3: *MLP (saturado)*

observado	predito		erro	erro total
	bom	mau		
Bom	155	197	56,0%	50,0%
Mau	28	70	28,6%	

De um total de 352 bons clientes, o modelo aprovou 155 e reprovou 197, cometendo um erro de 56,0%. Dentre os maus clientes, o modelo reprovou 70 e aprovou 28, nesse caso a taxa de erro foi de 28,6%.

O passo seguinte é ajustar um novo modelo com variáveis que foram definidas pelo critério de Akaike, quando se desenvolvia um modelo em regressão logística no item anterior.

Determina-se o número de neurônio da camada interna pelo gráfico de erro quadrático médio.

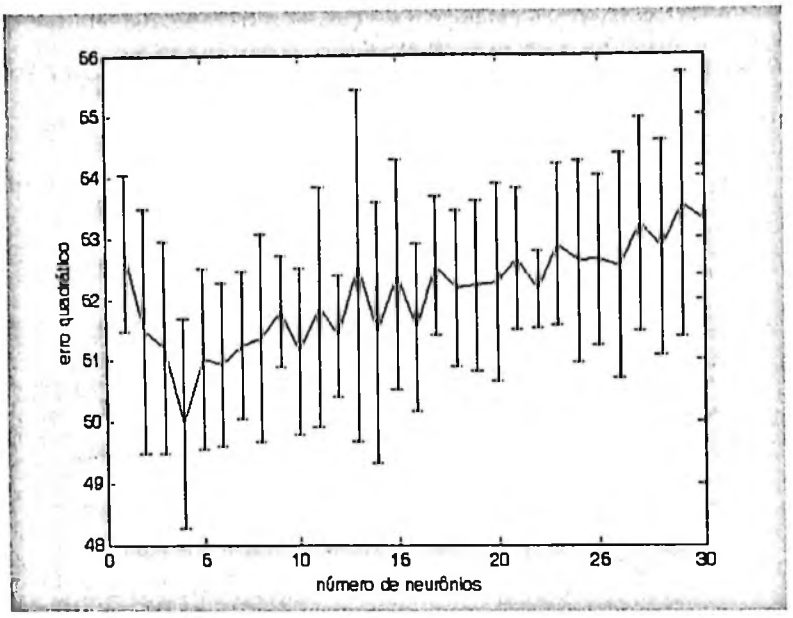


Figura 4.5: *Erro quadrático*

O menor erro quadrático foi atingido no modelo com 4 neurônios na camada intermediária. Traçamos em seguida, para se definir a nota de corte, o gráfico de perda média por clientes.

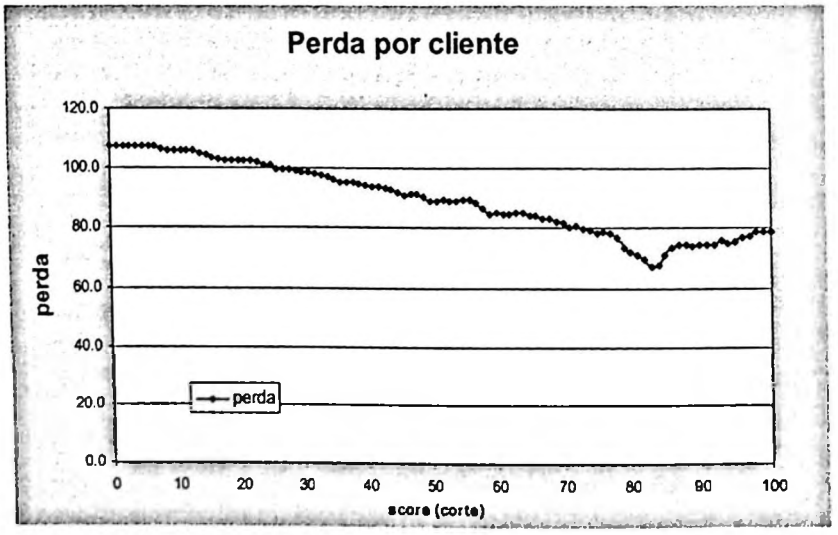


Figura 4.4: *Perda média por cliente*

O ponto de menor perda por cliente é atingido quando a nota de corte é 83. Com os parâmetros definidos, aplicamos o modelo nos dados de teste.

Tabela 4.7: *MLP (não saturado)*

observado	predito		erro	erro
	bom	mau		total
bom	187	165	46,9%	46,9%
mau	46	52	46,9%	

De um total de 352 bons clientes, o modelo aprovou 187 e reprovou 165, cometendo um erro de 46,9%. Dentre os maus clientes, o modelo reprovou 52 e aprovou 46, nesse caso a taxa de erro foi de 46,9%.

4.2 Estratégia 2

A estratégia 2 consiste utilizar o banco de dados na modelagem com as variáveis quantitativas parcialmente categorizadas. Decidiu-se por categorizar: "idade", "tempo de serviço" e "salário". A descrição completa das variáveis utilizadas encontra-se no apêndice B.

A seqüência de ajuste dos modelos segue exatamente a mesma metodologia da estratégia 1 que foi descrita no item anterior, e, portanto, não serão apresentadas. Os gráficos usados como apoio no desenvolvimento da modelagem estão no apêndice C.

Tabela 4.5: *RL (saturado)*

observado	Predito		Erro	erro
	bom	mau		total
Bom	180	203	53,0%	48,4%
Mau	39	78	33,3%	

Tabela 4.6: *RL (não saturado)*

observado	Predito		Erro	erro total
	bom	mau		
Bom	220	163	42,6%	43,6%
Mau	55	62	47,0%	

O modelo saturado em regressão logística (Tabela 4.5) aprovou 180 clientes e reprovou 203, entre os 383 bons clientes, cometendo um erro de 53,0%. Dentre os 117 maus clientes, foram corretamente reprovados 78, mas foram aprovados 39, um erro de 33,0%.

O modelo selecionado pelo critério de Akaike (tabela 4.6), reprovou 163 bons cliente e aprovou 55 maus clientes, cometendo erros de 42,6% e 47,0% respectivamente.

Tabela 4.7: *MLP (saturado)*

observado	Predito		erro	erro total
	bom	Mau		
bom	141	211	59,9%	54,0%
mau	32	66	32,7%	

Tabela 4.8: *MLP (não saturado)*

observado	Predito		erro	erro total
	bom	Mau		
bom	162	190	54,0%	48,9%
mau	30	68	30,6%	

Utilizando as técnicas de redes neurais, o modelo saturado (tabela 4.7) obteve erros entre os bons clientes de 59,9% e entre os maus de 32,7%. Em contrapartida, utilizando as variáveis escolhidas pelo critério de Akaike (tabela 4.8), os erro entre os bons e maus clientes foram 54,0% e 30,6% respectivamente.

4.3 Estratégia 3

Na terceira estratégia, resolveu-se por categorizar todas as variáveis. Da mesma forma, a seqüência de desenvolvimento e ajuste do modelo é o mesmo adotado no item 4.1. A descrição dos dados e gráficos usados estão nos apêndices B e C.

Tabela 4.9: *RL (saturado)*

Observado	predito		erro	erro
	bom	mau		total
bom	217	166	43,3%	42,8%
mau	48	69	41,0%	

Tabela 4.10: *RL (não saturado)*

observado	predito		erro	erro
	bom	Mau		total
bom	212	171	44,6%	44,6%
mau	52	65	44,4%	

O modelo saturado (tabela 4.9) apresentou uma taxa de erro de 43,3% entre os bons e de 41,0% entre os maus clientes, em contrapartida, o não saturado (tabela 4.10) obteve um erro de 44,6% e 44,4% respectivamente.

Tabela 4.11: *MLP (saturado)*

observado	predito		erro	erro
	bom	mau		total
bom	244	108	30,7%	35,8%
mau	53	45	54,1%	

Tabela 4.11: *MLP (não saturado)*

observado	predito		Erro	erro
	bom	mau		total
bom	141	211	59,9%	54,0%
mau	32	66	32,7%	

O modelo saturado (tabela 4.11) classificou erroneamente 211 dos 352 bons clientes, obtendo uma erro de 59,9%. Dentre os 98 maus clientes, 32 foram aprovados, incorrendo um erro de 32,7%.

5. Análise dos resultados

A tabela abaixo resume os resultado dos 12 modelos ajustados. A coluna "erros entre bons" apresenta em percentuais, os clientes que foram reprovados pelo modelo mas que deveriam ser classificados como bons. Por outro lado, a coluna "erros entre maus" mostra em percentuais os maus clientes aprovados pelo modelo.

Calculou-se, para auxiliar na comparação dos modelos, as distâncias de Mahalanobis, que é uma medida de quão distantes estão os bons dos maus clientes (vide Anexo A).

Tabela 5.1: *Resumo dos Resultados*

		Erro (entre bons)	Erro (entre maus)	Distância Mahalanobis
Estratégia 1	RL (saturado)	38,4%	44,4%	0,27
	RL (não saturado)	54.6%	35.9%	0,27
	MLP (saturado)	56,0%	28,6%	0,19
	MLP (não saturado)	46,9%	46,9%	0,12
Estratégia 2	RL (saturado)	53,0%	33,3%	0,49
	RL (não saturado)	42,6%	47,0%	0,50
	MLP (saturado)	59,9%	32,7%	0,26
	MLP (não saturado)	54,0%	30,6%	0,23
Estratégia 3	RL (saturado)	43,3%	41,0%	0,43
	RL (não saturado)	44,6%	44,4%	0,44
	MLP (saturado)	30,7%	54,1%	0,32
	MLP (não saturado)	59,9%	32,7%	0,31

Os modelos em regressão logística tiveram um melhor desempenho com a categorização de variáveis, isso pode ser constatado tanto pelos percentuais de erro como também pelas distâncias de Mahalanobis. O inverso pode ser observado para as redes neurais, as taxas de erro tiveram uma sensível piora com a categorização, apresentando em todos os modelos ajustados erros superiores a 50%.

Observa-se, também, que para os modelos em regressão logística, a escolha de variáveis pelo critério de Akaike, segregou melhor os bons e os maus cliente no caso das estratégias 2 e 3, onde foi introduzida a categorização de variáveis quantitativas. O mesmo não é válido para as redes neurais, na estratégia 1, onde seu desempenho foi melhor, a exclusão de variáveis por esse critério, fez a distância de Mahalanobis cair de 0,19 para 0,12.

Analisando-se o *score* dos clientes erroneamente avaliados pelo modelo, observamos que existem muitos que estão no limiar entre o bom e o mau pagador, ou seja, pertencem a uma região que pode ser considerada indefinida. Portanto, sugerimos que seja definido um intervalo onde os clientes que estiverem dentro dele, sejam submetidos a um analista de crédito para que seja feita avaliação mais profunda.

Para ilustrar, traçamos um gráfico do *score* gerado pelo modelo não saturado em regressão logística sobre os dados da estratégia 2, parcialmente categorizados:

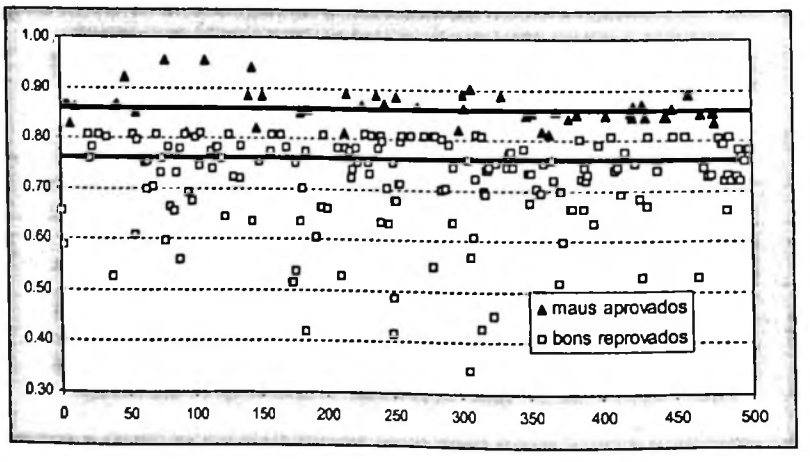


Figura 5.1: *Gráfico de score (RL Dados B)*

Definindo um intervalo entre 76 a 86, teríamos 196 clientes pertencentes a esse intervalo. Analisando-se 304 clientes restantes teríamos uma melhora significativa nos resultados.

Tabela 5.2: *RL (não saturado)*

observado	predito		erro	erro total
	bom	mau		
Bom	146	93	38,9%	37,8%
Mau	22	43	33,8%	

Os erros entre os bons clientes caíram de 42,6% para 38,9% e entre os maus de 47,0% para 33,8%.

Utilizando o mesmo critério para o modelo em redes neurais saturado, com os dados da estratégia 1, e definindo um intervalo entre 78 a 88, teríamos 144 clientes pertencentes a esse intervalo. Os resultados seriam:

Tabela 5.3: *MLP (saturado)*

observado	predito		erro	erro total
	bom	mau		
Bom	129	115	47.1%	44.4%
Mau	21	41	33.9%	

O erro entre os bons clientes reduziu de 56,0% para 47,1%, por outro lado, entre os maus houve uma piora, passando dos 28,6% para 33,9, não sendo tão eficiente quanto nos modelos em regressão logística.

6. Conclusão

Neste trabalho, utilizamos a regressão logística e as redes neurais para o desenvolvimento de modelos de *credit scoring*, duas técnicas largamente utilizadas. Aplicamos ambas as técnicas em dados parcialmente e totalmente categorizados, em modelos saturados e não saturados.

Os modelos em redes neurais ajustados com as variáveis na forma contínuas, isto é, sem categorizá-las , geraram modelos mais robustos. Isso ocorre porque muitas das variáveis contínuas apresentam um comportamento não linear em relação ao risco. Uma variável como renda mensal, por exemplo, pode ter uma contribuição positiva para o *score* para valores até 5 salários, negativa entre 6 e 15 salários e voltar a ser positiva para valores maiores que 16 salários mínimos. Dessa forma, quando categorizamos uma variável dessa natureza, estamos excluindo informações que poderiam ser importantes para que as redes identifiquem os padrões.

De forma geral, as duas as técnicas utilizadas se mostraram ferramentas muito úteis na avaliação de risco de crédito concedido às pessoas físicas. Ambas geraram modelos com resultados muito semelhantes, com uma pequena vantagem para a regressão logística, quando confrontados os resultados.

A principal dificuldade enfrentada na implementação dos modelos, foi a ausência de uma base de dados com um número maior de observações. Normalmente, recomenda-se que os dados utilizados para o ajuste de modelos sejam dividido na proporção de 50% de bons e 50% de maus clientes. Além disso tivemos que "artificialmente aumentar" a incidência de maus clientes na base, mudando o critério, que normalmente os definem, de 3 para 2 parcelas consecutivas de atraso.

Para estudos futuros, sugerimos testar a arquitetura *RBF (Radial Basis Function)* (Haykin, 1999). Uma alternativa ao critério de Akaike, também poderia ser implementada na escolha de variáveis para o modelo, como por exemplo os algoritmos genéticos.

Apêndice

APÊNDICE A

Tabelas - "Weights of Evidence"

Tabela A1: Tabela de WOE para a variável *Idade do Cliente*

Idade	Núm. Bons	Núm. Maus	%Bons	%Ruins	%B/%R	WOE
18< i < 28	214	55	18.2%	17.0%	106.8%	0.028
28 < i < 45	655	210	55.6%	65.0%	85.6%	-0.068
i > 45	308	58	26.2%	18.0%	145.7%	0.164
Total	1177	323	100.0%	100.0%		

Tabela A2: Tabela de WOE para a variável *Tipo de residência*

residência	Núm. Bons	Núm. Maus	%Bons	%Ruins	%B/%R	WOE
própria	803	197	68.2%	61.0%	111.9%	0.049
alugada + pais + func.	215	89	18.3%	27.6%	66.3%	-0.179
Outros	159	37	13.5%	11.5%	117.9%	0.072
Total	1177	323	100.0%	100.0%		

Tabela A3: Tabela de WOE para a variável *Estado Civil*

estado civil	Núm. Bons	Núm. Maus	%Bons	%Ruins	%B/%R	WOE
solt. + desq.	331	114	28.1%	35.3%	79.7%	-0.099
casado	773	194	65.7%	60.1%	109.3%	0.039
Outros	73	15	6.2%	4.6%	133.6%	0.126
Total	1177	323	100.0%	100.0%		

Tabela A4: Tabela de WOE para a variável *Tempo de Serviço*

T. Serviço	Núm. Bons	Núm. Maus	%Bons	%Ruins	%B/%R	WOE
0 < t < 5	404	130	34.3%	40.2%	85.3%	-0.069
5 < t < 10	415	115	35.3%	35.6%	99.0%	-0.004
t > 10	358	78	30.4%	24.1%	126.0%	0.100
Total	1177	323	100.0%	100.0%		

Tabela A5: Tabela de WOE para a variável *Salário do Cliente*

Salário	Núm. Bons	Núm. Maus	%Bons	%Ruins	%B/%R	WOE
0 < t < 10	466	89	39.6%	27.6%	143.7%	0.157
10 < t < 30	480	130	40.8%	40.2%	101.3%	0.006
t > 30	231	104	19.6%	32.2%	61.0%	-0.215
Total	1177	323	100.0%	100.0%		

Tabela A6: Tabela de WOE para a variável *Tempo de Residência*

T. Resid.	Núm. Bons	Núm. Maus	%Bons	%Ruins	%B/%R	WOE
0 < t < 3	607	135	51.6%	41.8%	123.4%	0.091
3 < t < 12	390	114	33.1%	35.3%	93.9%	0.027
t > 12	180	74	15.3%	22.9%	66.8%	-0.176
Total	1177	323	100.0%	100.0%		

Tabela A7: Tabela de WOE para a variável *Valor Financiado*

V. Financ.	Núm. Bons	Núm. Maus	%Bons	%Ruins	%B/%R	WOE
0 < v < 1100	439	90	37.3%	27.9%	133.9%	0.127
1100 < v < 3000	610	173	51.8%	53.6%	96.8%	-0.014
v > 3000	128	60	10.9%	18.6%	58.5%	-0.233
Total	1177	323	100.0%	100.0%		

Tabela A8: Tabela de WOE para a variável *Valor das Prestações*

V. Prest.	Núm. Bons	Núm. Maus	%Bons	%Ruins	%B/%R	WOE
0 < v < 300	453	95	38.5%	29.4%	130.9%	0.117
300 < v < 800	592	159	50.3%	49.2%	102.2%	0.009
v > 800	132	69	11.2%	21.4%	52.5%	-0.280
Total	1177	323	100.0%	100.0%		

Tabela A9: Tabela de WOE para a variável *Quantidade de Prestações*

Qtd. Prest.	Núm. Bons	Núm. Maus	%Bons	%Ruins	%B/%R	WOE
0 < q < 5	402	103	34.2%	31.9%	107.1%	0.030
5 < q < 10	623	179	52.9%	55.4%	95.5%	-0.020
q > 10	152	41	12.9%	12.7%	101.7%	0.007
Total	1177	323	100.0%	100.0%		

Tabela A10: Tabela de WOE para a variável *Naturalidade do Cliente*

Naturalidade	Núm. Bons	Núm. Maus	%Bons	%Ruins	%B/%R	WOE
São Paulo	293	90	24.9%	27.9%	89.3%	-0.049
Estado SP	378	93	32.1%	28.8%	111.5%	0.047
Sul + Indef.	167	37	14.2%	11.5%	123.9%	0.093
Outros	339	103	28.8%	31.9%	90.3%	-0.044
Total	1177	323	100.0%	100.0%		

Tabela A11: Tabela de WOE para a variável *Atividade da Empresa*

Ativ. Empresa	Núm. Bons	Núm. Mau	%Bons	%Ruins	%B/%R	WOE
Público	157	29	13.3%	9.0%	148.6%	0.172
Indústria	230	56	19.5%	17.3%	112.7%	0.052
Outros	790	238	67.1%	73.7%	91.1%	-0.041
Total	1177	323	100.0%	100.0%		

Tabela A12: Tabela de WOE para a variável *Sexo*

Sexo	Núm. Bons	Núm. Maus	%Bons	%Ruins	%B/%R	WOE
Masculino	980	256	83.3%	79.3%	105.1%	0.021
Feminino	197	67	16.7%	20.7%	80.7%	-0.093
Total	1177	323	100.0%	100.0%		

APÊNDICE B

Descrição das variáveis utilizadas nas estratégias adotadas

ESTRATÉGIA 1

Modelo Saturado

Variáveis quantitativas:

1. Valor do financiamento
2. Valor das prestações
3. Quantidade de prestações
4. Tempo de residência do cliente em número de anos
5. Idade
6. Tempo de serviço
7. Salário

Variáveis categóricas:

8. Atividade da empresa
    a. Público
    b. Indústria
    c. Outros
9. Tipo de Residência
    a. Própria
    b. Alugada, funcional ou dos pais
    c. Outros
10. Estado Civil
    a. Casado
    b. Solteiro ou desquitado
    c. Outros
11. Sexo
    a. Masculino
    b. Feminino

12. Naturalidade
    a. Cidade de São Paulo
    b. Estado de SP
    c. Região Norte, Sudeste, Centro-Oeste e estrangeiros
    d. Outros
13. Cheque especial
    a. Possui
    b. Não possui
14. Cartão de crédito
    a. Possui
    b. Não possui

Modelo não saturado (definito pelo AIC)

Variáveis quantitativas:

1. Valor do financiamento
2. Tempo de residência do cliente em número de anos

Variáveis categóricas:

3. Sexo
    a. Outros
    b. Feminino
4. Naturalidade
    a. Estado de SP
    b. Outros
5. Tipo de Residência
    a. Alugada, funcional ou dos pais
    b. Masculino
6. Atividade da empresa
    a. Público
    b. Outros

ESTRATÉGIA 2 (variáveis contínuas parcialmente categorizadas)

Modelo Saturado

Variáveis contínuas:

1. Valor de financiamento
2. Valor das prestações
3. Quantidade de prestações
4. Tempo de residência do cliente em número de anos

Variáveis categóricas:

5. Atividade da empresa
    a. Público
    b. Indústria
    c. Outros
6. Idade
    a. 18 a 28 anos
    b. 28 a 45 anos
    c. Maiores de 45
7. Tipo de Residência
    a. Própria
    b. Alugada, funcional ou dos pais
    c. Outros
8. Estado Civil
    a. Casado
    b. Solteiro ou desquitado
    c. Outros
9. Tempo de Serviço
    a. 0 a 5 anos
    b. 5 a 10 anos
    c. Mais de 10 anos

10. Número de salários mínimos
    a. 0 a 10 salários
    b. 10 a 22 salários
    c. Maior de 22 salários
11. Sexo
    a. Masculino
    b. Feminino
12. Naturalidade
    a. Cidade de São Paulo
    b. Estado de SP
    c. Região Norte, Sudeste, Centro-Oeste e estrangeiros
    d. Outros
13. Cheque especial
    a. Possui
    b. Não possui
14. Cartão de crédito
    a. Possui
    b. Não possui

Modelo não saturado (definito pelo AIC)

Variáveis contínuas:

1. Valor de financiamento
2. Tempo de residência do cliente em número de anos

Variáveis categóricas:

3. Idade
    a. 28 a 45 anos
    b. Outros
4. Tempo de serviço
    a. 0 a 5 anos
    b. Outros

5. Naturalidade
    a. Estado de SP
    b. Outros
6. Número de salários mínimos
    a. 0 a 10 salários
    b. Outros

ESTRATÉGIA 3 (todas as variáveis contínuas categorizadas)

Modelo Saturado

Variáveis categóricas:

1. Valor de financiamento
    a. Valores entre $0 e $1100
    b. Valores entre $1100 e $3000
    c. Valores maiores ou iguais a $3000
2. Valor das prestações
    a. Valores entre $0 e $300
    b. Valores entre $300 e $800
    c. Valores maiores ou iguais a $800
3. Quantidade de prestações
    a. Quantidade de prestações entre 0 e 5
    b. Quantidade de prestações entre 5 e 10
    c. Quantidade de prestações maiores ou iguais a 10
4. Tempo de residência do cliente em número de anos
    a. Tempo entre 0 e 3 anos
    b. Tempo entre 3 e 12 anos
    c. Tempo maior ou igual a 12 anos
5. Idade
    a. 18 a 28 anos
    b. 28 a 45 anos
    c. Maiores de 45

6. Tipo de Residência
    a. Própria
    b. Alugada, funcional ou dos pais
    c. Outros
7. Estado Civil
    a. Casado
    b. Solteiro ou desquitado
    c. Outros
8. Tempo de Serviço
    a. 0 a 5 anos
    b. 5 a 10 anos
    c. Mais de 10 anos
9. Atividade da empresa
    a. Público
    b. Indústria
    c. Outros
10. Número de salários mínimos
    a. 0 a 10 salários
    b. 10 a 22 salários
    c. Maior de 22 salários
11. Sexo
    a. Masculino
    b. Feminino
12. Naturalidade
    a. Cidade de São Paulo
    b. Estado de SP
    c. Região Norte, Sudeste, Centro-Oeste e estrangeiros
    d. Outros
13. Cheque especial
    a. Possui
    b. Não possui

14. Cartão de crédito
    a. Possui
    b. Não possui

Modelo não saturado (definito pelo AIC)

1. Valor de financiamento
    a. Valores entre $1100 e $3000
    b. Outros
2. Quantidade de prestações
    a. Quantidade de prestações entre 0 e 5
    b. Outros
3. Tempo de residência do cliente em número de anos
    c. Tempo maior ou igual a 12 anos
    d. Outros
4. Idade
    a. 28 a 45 anos
    b. Outros
5. Tipo de Residência
    a. Alugada, funcional ou dos pais
    b. Outros
6. Tempo de Serviço
    a. 0 a 5 anos
    b. Outros
7. Número de salários mínimos
    a. 0 a 10 salários
    b. 10 a 22 salários
    c. Maior de 22 salários
8. Naturalidade
    a. Estado de SP
    b. Outros

APÊNDICE C

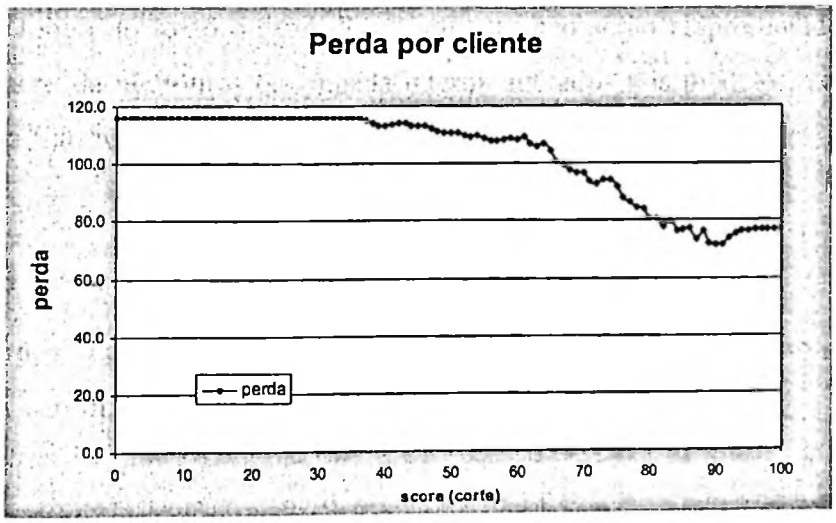
Neste apêndice, apresentaremos os gráficos utilizados no desenvolvimento dos modelos descritos no Capítulo 4. Somente serão exibidos os gráficos de perda média por cliente e os de erro quadrático das estratégias 2 e 3. Os gráficos da estratégia 1 foram mostrados ao longo do texto do trabalho e portanto não serão reapresentados. A nota de corte e o número de neurônios da camada interna adotados nos modelos, serão indicados abaixo de cada figura.

ESTRATÉGIA 2

Definimos como estratégia 2, utilizar o conjunto de dados parcialmente categorizados, conforme descrito no apêndice B.

Modelo em regressão logística com a utilização de todas as variáveis disponíveis (modelo saturado).

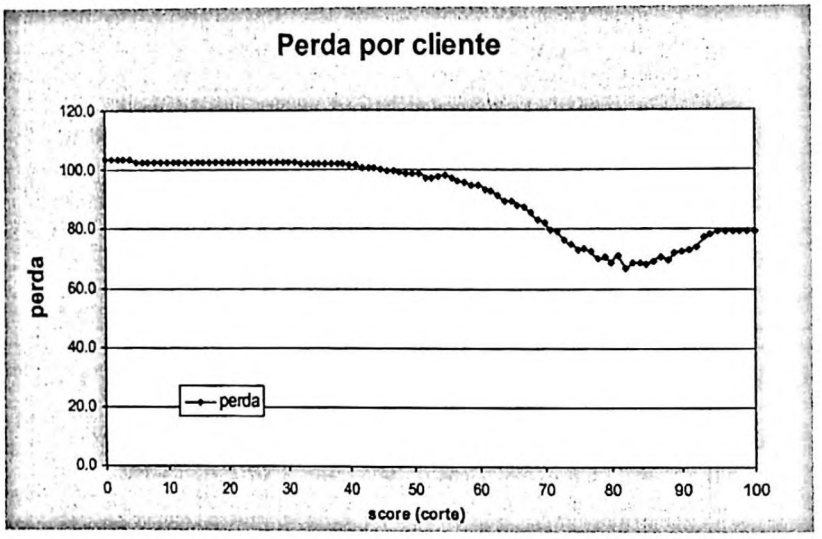
Gráfico C1: *Perda média por cliente*



Nota de corte = 92

Modelo em regressão logística com a definicão das variáveis preditoras através do critério de Akaike (modelo não saturado).

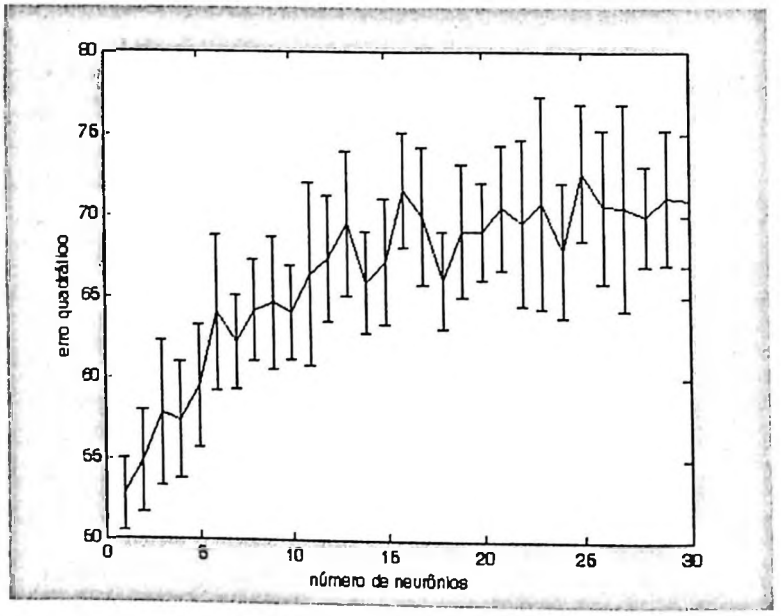
Gráfico C2: *Perda média por cliente*



Nota de corte = 82

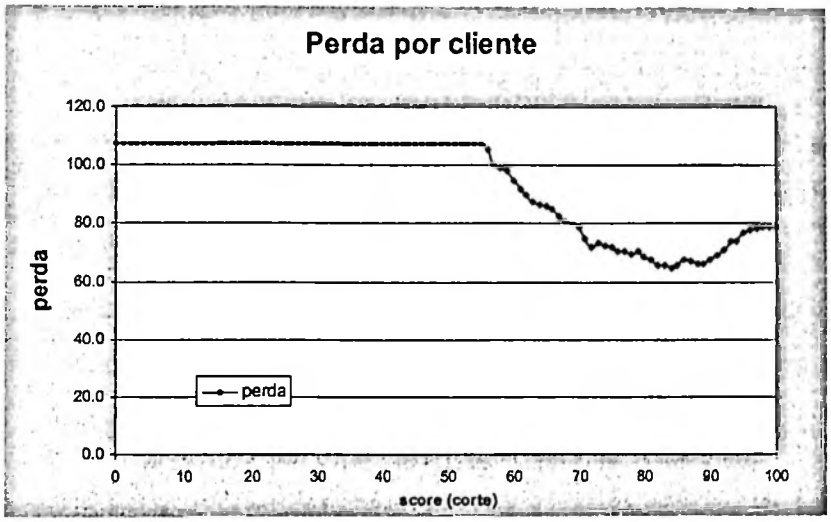
Modelo em redes neurais com a utilização de todas as variáveis disponíveis (modelo saturado).

Gráfico C3: *Erro quadrático*



Número de neurônios = 1

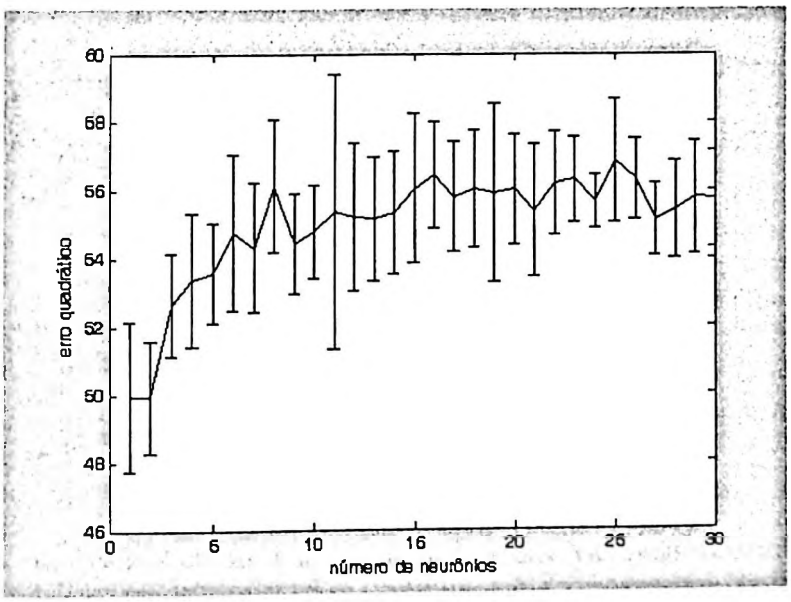
Gráfico C4: *Perda média por cliente*



Nota de corte = 84

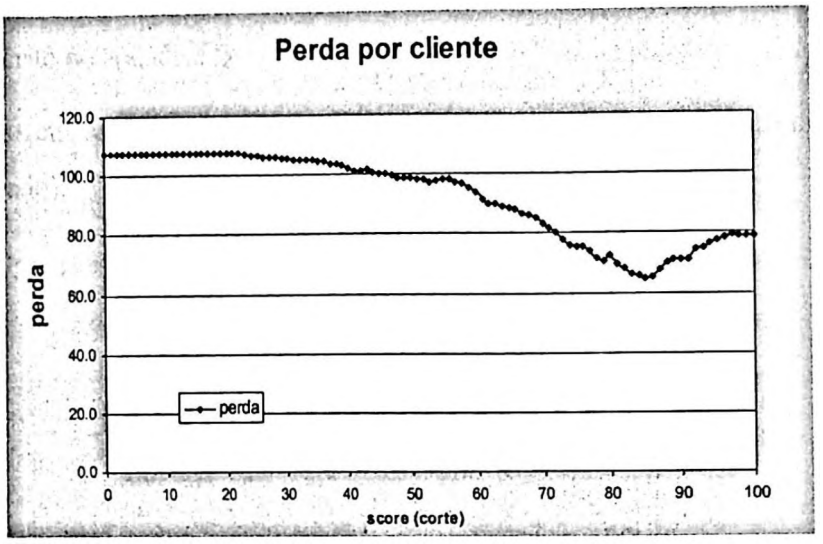
Modelo em redes neurias com a definicão das variáveis preditoras através do critério de Akaike (modelo não saturado).

Gráfico C5: *Erro quadrático*



Número de neurônios = 1

Gráfico C6: *Perda média por cliente*



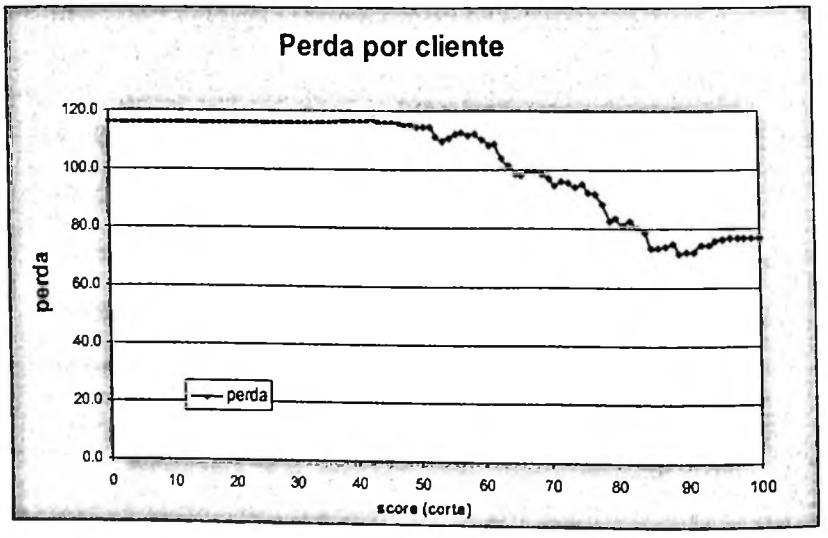
Nota de corte = 85

ESTRATÉGIA 3

Definimos como estratégia 3, utilizar o conjunto de dados totalmente categorizados, conforme descrito no apêndice B.

Modelo em regressão logística com a utilização de todas as variáveis disponíveis (modelo saturado).

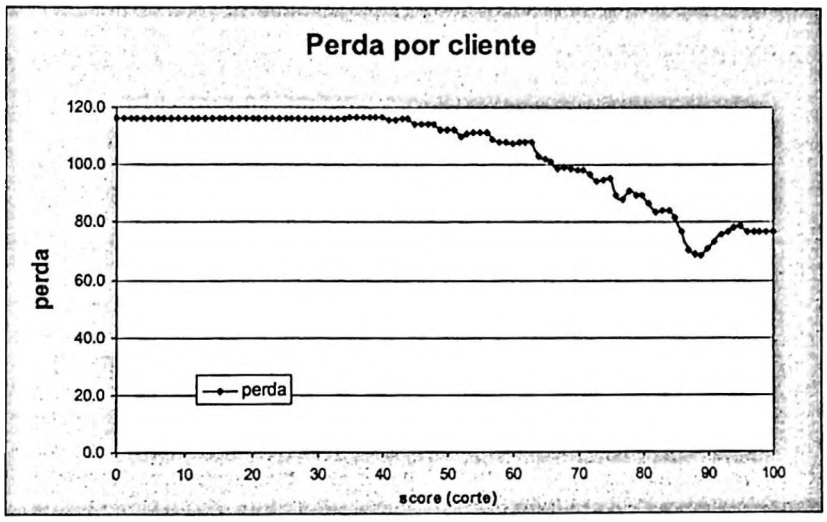
Gráfico C7: *Perda média por cliente*



Número de neurônios = 90

Modelo em regressão logística com a definicão das variáveis preditoras através do critério de Akaike (modelo não saturado).

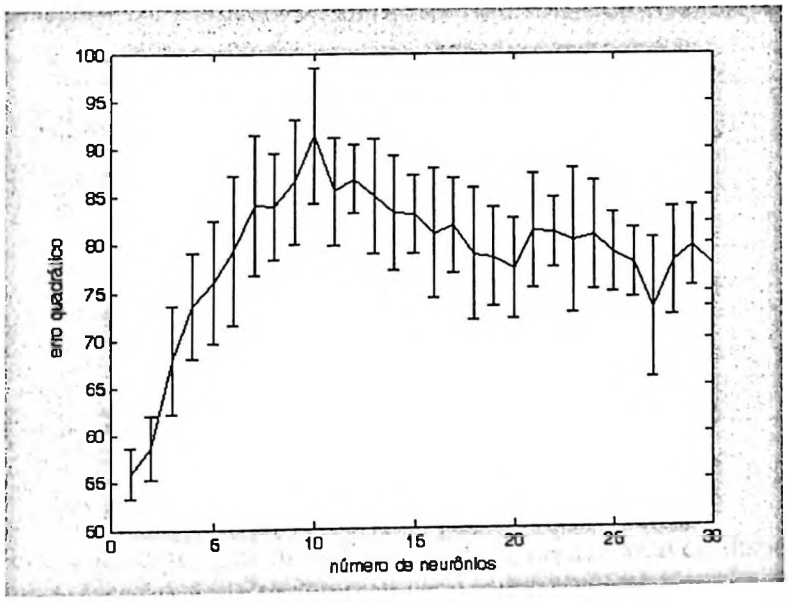
Gráfico C8: *Perda média por cliente*



Nota de corte = 89

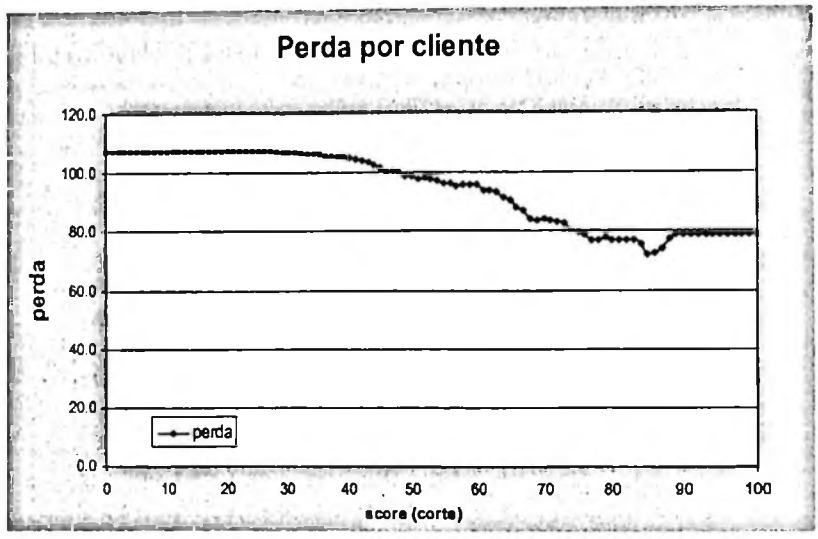
Modelo em redes neurais com a utilização de todas as variáveis disponíveis (modelo saturado).

Gráfico C9: *Erro quadrático*



Número de neurônios = 1

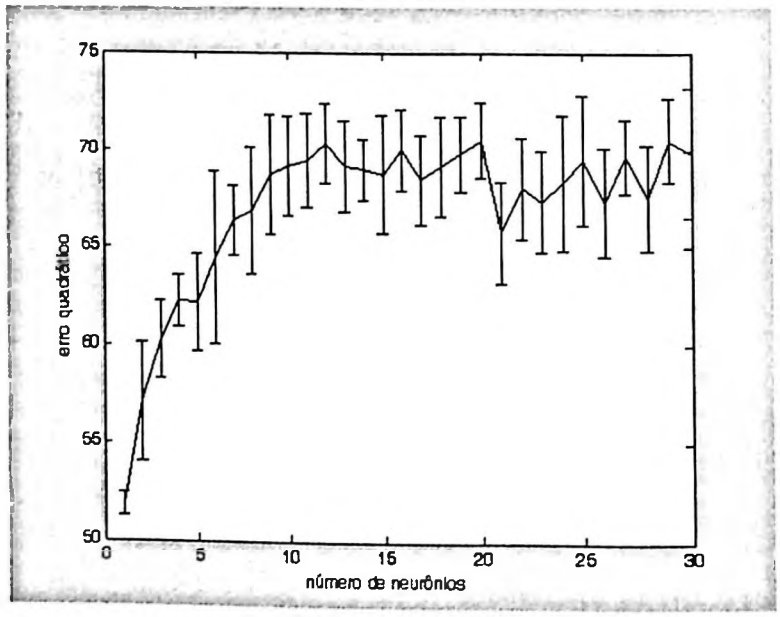
Gráfico C10: *Perda média por cliente*



Nota de corte = 85

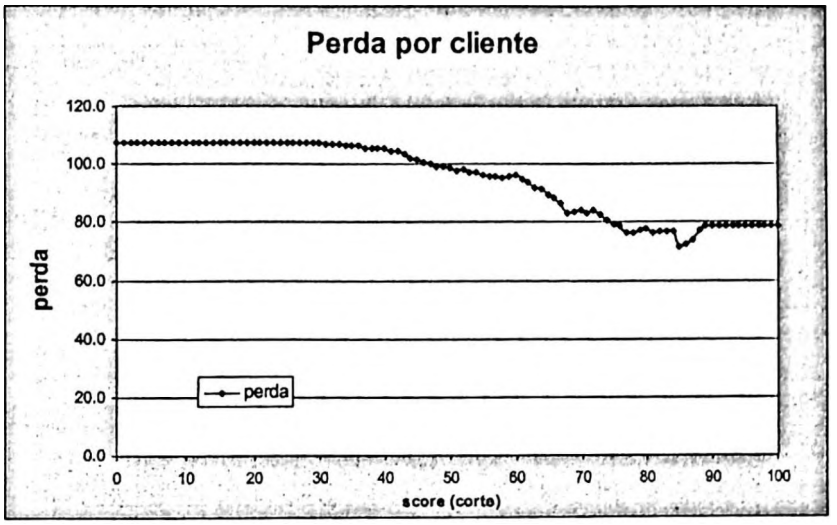
Modelo em redes neurias com a definicão das variáveis preditoras através do critério de Akaike (modelo não saturado).

Gráfico C11: *Erro quadrático*



Número de neurônios = 1

Gráfico C12: *Perda média por cliente*



Nota de corte = 85

Anexo

Distância de Mahalanobis

$$p_B(s) = \frac{n_B(s)}{n_B}$$

$$p_M(s) = \frac{n_M(s)}{n_M}$$

$$m_B = \sum s \bullet p_B(s)$$

$$m_M = \sum s \bullet p_M(s)$$

$$\sigma_B^2 = \left(\sum_s s^2 p_B(s) - m_B^2 \right)$$

$$\sigma_M^2 = \left(\sum_s s^2 p_M(s) - m_M^2 \right)$$

$$\sigma = \left(\frac{n_B \sigma_B^2 + n_M \sigma_M^2}{n} \right)^{1/2}$$

$$M = \frac{m_B - m_M}{\sigma}$$

Onde,

$n_B(s)$ e $n_M(s)$: número de bons e maus pagadores com score s

n_B e n_M: número de bons e de maus pagadores

s: score

M: Distância de Mahalanobis

Bibliografia

ARMINGER, G., ENACHE, D. and BONNE, T. (1997). *Analyzing credit risk data: a comparison of logistic discrimination, classification tree analysis and feedforward neural networks.* Computational Statistics, 12, 293-310.

BISHOP, C. M. (1995). *Neural networks for pattern recognition.* Oxford, Oxford University Press.

COX, D. R. (1970). *The analysis of binary data.* London, Methuen.

DORNBUSH, R., FISCHER, S. (1991). *Macroeconimia.* São Paulo, Makron.

FISHER, R.A. (1936), *The use of multiple measurement in taxonomic problem,* Annals of Eugenics, 7, 179-188.

GUJARATI, D. N. (1995). *Basic econometrics.* New Aster, Mc Graw Hill

HAYKIN, Simon (1999). *Redes neurais, princípios e prática.* 2. ed. Ontario, Bookman.

HAND, D. J. (1981). *Discrimination and classification.* Wiley, Chichester.

HAND, D. J. and Henley, D. J. (1997). *Statistical classification methods in consumer credit scoring: a review.* Journal of the Royal Statistical Society, Series A, 160, part 3, 523-541.

HOLLAND, J. H. (1975). *Adaptation in natural and artificial systems.* University of Michigan Press, Ann Arbor.

JOHNSON, R.W. (1992). *Legal, social and economics issues implementing scoring in the U.S., in credit scoring and credit control.* Oxford, Oxford University Press.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized linear models.* 2nd. Edition. London, Chapman and Hall.

MCCULLOCH, C. E. and SEARLE, S. R. (2001). *Generalized, linear, and mixed models.* New York, Wiley

MYERS, R. H. (1990). *Classical and Modern Regression with applications.* Belmont, Dusbury

MYERS, J. H., FORGY, E. W. (1963). *The development of numerical credit evaluation system.* J. Ameri. Statist. Assoc., 58, 799-806.

NABNEY, I. T. (2003). *Netlab, algorithms for pattern recognition.* London, Springer.

MINSKY, M. L. and PAPERT, S. A. (1969). *Perceptrons.* Cambrigde, MIT Press.

NETER, J. and WASSERMAN, W. (1974). *Applied linear statistical models.* Illinois, Richards.

PAULA, G. A. (2000). *Modelos de regressão com apoio computacional.* Instituto de Matemática e Estatística, Universidade de São Paulo.

ROSENBLATT, F. (1958). The percptron: *A probabilistic model for information storage and organization in the brain.* Psychological Rev, 65, 386-408.

RUMELHART, D. E., HINTON, G. E. and WILLIAMS, R. J. (1986a). *Learning representation by back-propagation errors.* Nature, 323, 533-536.

RUMELHART, D. E., HINTON, G. E. and WILLIAMS, R. J. (1986b). *Learning representation by error backpropagation, in Parallel distributed processing: explorations in the microstructure of cognition.* Cambrigde, MIT Press.

THOMAS, L. C., EDELMAN, D. B. and CROOK, J. N. (2002). *Credit scoring and its applications.* Philadelphia, Siam.

VERSTRAETEN, G. and VAN DEN POEL, D. (2003). *The impact of sample bias on consumer credit scoring performance and profitability.* Ghent, Belgium

WEST, D. (2000). *Neural network credit scoring problems.* Computers and Operational Research, 27, 1131-1152