

**MODELOS DE “CREDIT SCORING”:  
REGRESSÃO LOGÍSTICA, CHAID E REAL**

**Paulo de Tarso Marques Rosa**

DISSERTAÇÃO APRESENTADA AO  
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA  
DA UNIVERSIDADE DE SÃO PAULO  
PARA OBTENÇÃO DO GRAU DE  
MESTRE EM ESTATÍSTICA

Área de concentração : **Estatística**

Orientador: **Prof. Dr. Carlos Alberto de Bragança Pereira**

Durante o curso de Mestrado e elaboração desse  
trabalho o autor recebeu apoio financeiro do CNPq

- São Paulo, outubro de 2000 -

## **MODELOS DE “CREDIT SCORING”: REGRESSÃO LOGÍSTICA, CHAID E REAL**

**PAULO DE TARSO MARQUES ROSA**

*Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Paulo de Tarso Marques Rosa e aprovada pela comissão julgadora.*

Banca examinadora:

- PROF. DR. CARLOS ALBERTO DE BRAGANÇA PEREIRA (ORIENTADOR) - IME-USP
- PROF. DR. JÚLIO MICHAEL STERN - IME-USP
- PROF. DR. JOSÉ AFONSO MAZZON - FEA-USP

- São Paulo, outubro de 2000 -

# ÍNDICE

<b>1</b>	<b>Introdução</b>	<b>1</b>
<b>2</b>	<b>Descrição do Estudo</b>	<b>3</b>
2.1	O Produto .....	3
2.2	Os Dados .....	4
2.3	As Variáveis .....	4
2.3.1	Variáveis Cadastrais .....	5
2.3.2	Variáveis de Utilização e Restrição .....	6
2.4	Definição da Variável Resposta .....	7
2.5	Quantidade de Parcelas dos Contratos .....	10
2.6	Amostras de Desenvolvimento e Validação .....	11
2.7	Tratamento das Variáveis .....	13
2.7.1	Valores Faltantes .....	14
2.7.2	Categorização de Variáveis .....	14
<b>3</b>	<b>Metodologia</b>	<b>18</b>
3.1	Régressão Logística Múltipla .....	18
3.2	CHAID .....	21
3.2.1	Método para Variáveis Dependentes Nominais .....	22
3.2.2	Tipos de Variáveis Independentes .....	24
3.2.3	Teste da Hipótese de Independência .....	25
3.2.4	Nível Descritivo Ajustado de Bonferroni .....	26
3.3	REAL .....	27
3.3.1	Descrição do Método de Construção da Árvore de Classificação ....	28
3.3.2	Função de Convicção .....	29
3.3.3	Função de Perda .....	30
3.3.4	Procedimento de Discretização das Variáveis Preditoras .....	31
3.3.5	Critérios de Parada .....	32
3.4	Medidas de Avaliação dos Modelos .....	32

<b>4 Aplicação</b>	<b>34</b>
4.1 Variáveis Preditoras .....	34
4.1.1 Variáveis Preditoras na Regressão Logística Múltipla .....	35
4.1.2 Variáveis Preditoras no CHAID .....	36
4.1.3 Variáveis Preditoras no REAL .....	36
4.2 Aplicação das Técnicas .....	36
4.2.1 Aplicação da Regressão Logística Múltipla .....	37
4.2.2 Aplicação do CHAID .....	42
4.2.3 Aplicação do REAL .....	47
4.3 Comparações .....	50
<b>5 Conclusão</b>	<b>54</b>
<b>A Tabelas de “Weights of Evidence”</b>	<b>56</b>
<b>B Tabelas de Medidas de Acerto para o REAL com Parâmetros (<math>r, v_c</math>)</b>	<b>61</b>
<b>Bibliografia</b>	<b>67</b>

## CAPÍTULO 1

### Introdução

A concessão de empréstimos a clientes é uma atividade financeira que vem crescendo a cada dia no Brasil. Durante o período inflacionário, as instituições financeiras obtinham grande parte de seus lucros com operações financeiras relacionadas à desvalorização da moeda e a carteira de crédito destas instituições continha um número de contratos suficientes somente para satisfazer as exigências legais. Após o fim da inflação, percebeu-se a necessidade de se aumentar as alternativas de investimento para substituir a rentabilidade do período inflacionário. Desde então, as instituições têm se preocupado em aumentar suas carteiras de crédito, mantendo, porém, a qualidade na concessão.

Em um primeiro instante, as instituições financeiras do país não foram capazes de gerenciar a quantidade e a qualidade dos empréstimos ao mesmo tempo, já que não possuíam experiência suficiente em concessão de crédito. A qualidade da concessão estava atrelada ao fato de uma proposta de solicitação de crédito ser avaliada por um ou mais analistas de crédito que apresentavam um parecer favorável ou desfavorável à operação. Esse sistema apresenta boa qualidade na concessão, porém, é inviável quando o número de propostas é muito grande. Surgiu, assim, a pergunta chave: como conceder crédito para um grande número de clientes de forma padronizada e mantendo uma qualidade aceitável?

Desde então, iniciou-se o processo de elaboração e aplicação de Modelos de “Credit Scoring”, utilizados já há alguns anos nos EUA e Inglaterra dentre outros países. Esses modelos permitem que a concessão do crédito não dependa da subjetividade de cada analista, e possa ser processada de forma padronizada, controlando, assim, os riscos e erros associados aos modelos.

Neste trabalho são descritas três aplicações de ferramentas estatísticas na construção de Modelos de “Credit Scoring”: Análise de Regressão Logística; Árvore de Classificação CHAID (“Chi-squared Automatic Interaction Detection”) e Árvore de Classificação **REAL** (“Real Attribute Learnig Algorithm”).

Um conjunto de dados reais, fornecido por uma instituição financeira brasileira, foi utilizado para a aplicação das técnicas abordadas.

O trabalho está organizado da seguinte forma: no Capítulo 2, é apresentada a descrição do estudo abordando as características relacionadas ao produto, às variáveis e à amostra de trabalho; no Capítulo 3, são apresentadas as técnicas de classificação; no Capítulo 4, são feitas as aplicações e comparações entre os modelos e no Capítulo 5, estão as conclusões do trabalho e sugestões para novos estudos.

## CAPÍTULO 2

### Descrição do estudo

Uma instituição financeira deseja conceder empréstimos para financiamentos de compra de veículos a seus clientes e, para isso, necessita de uma ferramenta que avalie o grau de risco associado a cada financiamento. A instituição gostaria que todos os seus clientes fossem classificados como bons e maus pagadores, pois assim, poderia direcionar seus esforços para conceder produtos de crédito aos clientes com menor risco de inadimplência.

Para viabilizar a elaboração do estudo, foram disponibilizados dados referentes a clientes que contrataram um produto de crédito no passado, além das informações sobre os pagamentos mensais efetuados por esses clientes.

#### 2.1 O Produto

Trata-se especificamente de um produto de crédito para clientes que desejam comprar veículos através de financiamento. No início das negociações entre o cliente e a instituição, definem-se a taxa de juros a ser praticada e o prazo de pagamento das parcelas do contrato. Dependendo do contrato, o prazo pode chegar a 48 meses. Após a verificação da documentação de compra do veículo e o acerto quanto à taxa de juros e prazo, firma-se um contrato entre as partes. O valor contratado pelo cliente é disponibilizado em sua conta corrente e o mesmo passa a ter obrigações mensais para com a instituição até o fim do contrato.

Uma característica importante desse produto é que a garantia da operação de crédito pode ser o próprio veículo financiado, ou seja, pode-se negociar taxas de juros menores,

uma vez que, caso o cliente não cumpra com suas obrigações a instituição pode retomar o veículo por vias judiciais.

## 2.2 Os Dados

Para a realização do estudo tem-se à disposição uma amostra de 33.691 clientes que contrataram o produto no período de Julho de 1996 a Junho de 1997. Trata-se de uma base de dados histórica que contém informações mensais de utilização do produto, ou seja, a partir da estrutura da base de dados, pode-se verificar em que momento o cliente deixou de pagar uma ou mais parcelas do seu contrato. Essas informações históricas estão disponíveis mês a mês desde a data de contratação do produto até o mês de Junho de 1998.

## 2.3 As Variáveis

As variáveis disponíveis podem ser divididas em dois grupos: *Variáveis Cadastrais* e *Variáveis de Utilização e Restrição*. As *Variáveis Cadastrais* são aquelas relacionadas ao cliente ou à conta corrente do cliente na instituição. Já as *Variáveis de Utilização e Restrição* são aquelas relativas aos produtos que o cliente tem na instituição ou relativas às restrições de crédito que o cliente tem no mercado. Denominam-se “restrições de crédito” os apontamentos cadastrais negativos do cliente, existentes no mercado, por algum problema de crédito. As *Variáveis Cadastrais* foram coletadas apenas no momento em que o cliente contratou o produto, ao passo que algumas *Variáveis de Utilização e Restrição* foram coletadas no momento da contratação do produto e outras desde a contratação até o final do contrato, com periodicidade mensal.

### 2.3.1 Variáveis Cadastrais

As Variáveis Cadastrais disponíveis para a elaboração do estudo são:

- *Idade do Cliente* (em anos);
- *Código do Estado Civil* ( 1 = Casado com Separação de Bens;  
2 = Casado com Comunhão de Bens;  
3 = Casado com Comunhão Parcial de Bens;  
4 = Solteiro;  
5 = Desquitado ou Divorciado;  
6 = Viúvo;  
7 = Marital );
- *Quantidade de Dependentes Menores de Idade*;
- *Quantidade de Dependentes Maiores de Idade*;
- *Quantidade de Dependentes* (soma das duas anteriores);
- *Código do Tipo de Residência* ( 1 = Própria;  
2 = Financiada;  
3 = Alugada;  
4 = Com os Pais;  
5 = Outro Tipo);
- *Indicador de Telefone Residencial* (sim ou não);
- *Indicador de Telefone Comercial* (sim ou não);
- *Grupo de Profissão* ( 1 = Profissionais Liberais;  
2 = Administrativo/Serviços/Indústria;  
3 = Comércio;  
4 = Proprietários;  
5 = Aposentados;  
6 = Outros);
- *Tempo no Atual Emprego* (meses);

- *Indicador de Posse de Cartão de Crédito* (sim ou não);
- *Idade do Veículo Próprio* (anos);
- *Tempo como Cliente da Instituição* (meses);
- *Ano de Fabricação do Veículo Financiado* (data).

### **2.3.2 Variáveis de Utilização e Restrição**

Variáveis coletadas na contratação do produto:

- *Data da Contratação do Produto*;
- *Data do Vencimento do Contrato*;
- *Valor do Contrato* (em Reais);
- *Valor das Parcelas do Contrato* (em Reais);
- *Quantidade de Avalistas*;
- *Quantidade de Parcelas do Contrato*;
- *Indicador de Apontamento Cadastral\** (sim ou não);
- *Indicador de Contrato de Crediário nos Últimos 2 anos* (sim ou não).

Variáveis coletadas mês a mês desde a contratação do produto até Junho de 1998:

- *Quantidade de Parcelas Pagas*;
- *Quantidade de Parcelas Restantes a Serem Pagas*;
- *Quantidade de Parcelas Vencidas (Não-Pagas)*;
- *Valor Total das Parcelas Restantes a Serem Pagas* (em Reais);
- *Valor Total das Parcelas Vencidas* (em Reais);
- *Quantidade de Dias de Atraso do Contrato*;
- *Quantidade de Apontamentos Cadastrais\**.

\* Apontamentos Cadastrais são as restrições de crédito que o cliente tem no mercado ou na própria instituição.

O conjunto de variáveis preditoras do estudo é composto por todas as 14 variáveis cadastrais e 3 variáveis de utilização: *Quantidade de Avalistas*, *Indicador de Apontamento Cadastral*, *Indicador de Contrato de Crediário nos Últimos 2 anos*. As demais variáveis foram utilizadas para a definição da variável resposta do estudo.

## 2.4 Definição da Variável Resposta (Definição de Performance)

O primeiro passo no desenvolvimento de Modelos de “Credit Scoring” é definir o que é um bom e um mau pagador. Esta definição, denominada Definição de Performance, está diretamente ligada à política de crédito da instituição financeira. Nesta etapa, os clientes são classificados em grupos de performance, de acordo com o comportamento na utilização do produto. Neste estudo não são discutidas as diversas formas de estabelecimento de uma Definição de Performance, e será utilizada, portanto, a política de crédito sugerida pela instituição que forneceu os dados. Sugestões sobre regras para definição de performance podem ser encontradas em McCahill (1998), Leonard (1998) e Lewis (1994).

A Definição de Performance adotada para o estudo foi:

**Clientes Excluídos:** são aqueles que a instituição não pretende avaliar o risco de crédito através de modelos de “Credit Scoring”. Para esses clientes, são utilizadas outras metodologias de análise de crédito.

- Funcionários da Instituição;
- Clientes com Contrato de Veículos Pesados (que não sejam automóveis).

**Clientes Ruins:** são aqueles que durante o período de utilização do produto se mostraram como de alto risco de inadimplência.

- Clientes com Ação Legal Tomada ou com Veículo Reapropriado;
- Clientes com 3 ou mais parcelas consecutivas vencidas;

- Clientes com 2 vezes ou mais 2 parcelas consecutivas vencidas;
- Clientes com 4 vezes ou mais 1 parcela vencida.

**Clientes Indeterminados:** são aqueles cujos comportamentos na utilização do produto não permitem distingui-los entre bons e maus pagadores.

- Clientes com 1 vez 2 parcelas consecutivas vencidas;
- Clientes com 3 vezes 1 parcela vencida.

**Clientes Bons:** são aqueles que utilizaram o produto de forma adequada e apresentam um baixo risco de inadimplência.

- Clientes com somente 2 vezes 1 parcela vencida;
- Clientes com somente 1 vez 1 parcela vencida;
- Clientes que nunca foram inadimplentes.

Deve-se notar que existe um grupo de clientes que não faz parte do estudo, pois, as informações relativas a eles não são armazenadas pela instituição. Trata-se do grupo de clientes reprovados pela instituição antes mesmo de terem suas propostas cadastradas. Esse conjunto pode ser composto pelos piores clientes, por motivos diversos, e são recusados de imediato devido às suas condições financeiras. Em geral, as instituições financeiras não se preocupam em armazenar informações cadastrais desse grupo (clientes rejeitados). Em Lewis (1994), Hoyland (1997) e Hand (1998) são discutidas algumas possibilidades de análise quando a instituição dispõe das informações dos clientes rejeitados.

Para se aplicar a Definição de Performance a um cliente, utiliza-se um procedimento hierárquico, ou seja, verifica-se, inicialmente, se o cliente encaixa na categoria de Clientes Excluídos, caso não esteja nesta categoria, verifica-se na seguinte (Clientes Ruins), e assim por diante. Desta forma, não há possibilidade de um cliente ser classificado em mais de uma categoria.

Vale ressaltar que a Definição de Performance pode variar de uma instituição para outra, de acordo com o grau de risco que se deseja trabalhar.

Apesar de a Definição de Performance resultar em quatro classificações de clientes (excluídos, ruins, indeterminados e bons), somente duas delas são utilizadas para a construção da variável resposta: ruins e bons.

Deve-se atentar para o fato de que a Definição de Performance está diretamente ligada ao bom poder de discriminação dos modelos finais, já que, ao não se definir de forma adequada os grupos de bons e maus pagadores, não há como os modelos finais discriminá-los satisfatoriamente.

A distribuição de Performance dos clientes, na amostra de estudo, está descrita na Tabela 2.1.

**Tabela 2.1** Tabela de Definição de Performance

Definição de Performance	Categoria	N	%
Excluídos	Funcionários	2.565	7,6%
	Veículos Pesados	805	2,4%
	<i>Subtotal</i>	<b>3.370</b>	<b>10,0%</b>
Ruins	Ação Legal Tomada/ Veículo reapropriado	372	1,1%
	3 ou mais parcelas consecutivas vencidas	672	2,0%
	2 vezes ou mais 2 parcelas consecutivas vencidas	146	0,4%
	4 vezes ou mais 1 parcela vencida	199	0,6%
	<i>Subtotal</i>	<b>1.389</b>	<b>4,1%</b>
Indeterminados	1 vez 2 parcelas consecutivas vencidas	206	0,6%
	3 vezes 1 parcela vencida	105	0,3%
	<i>Subtotal</i>	<b>311</b>	<b>0,9%</b>
Bons	2 vezes 1 parcela vencida	261	0,8%
	1 vez 1 parcela vencida	740	2,2%
	Nunca Inadimplente	27.620	82,0%
	<i>Subtotal</i>	<b>28.621</b>	<b>85,0%</b>
<b>Total</b>	<b>Total</b>	<b>33.691</b>	<b>100,0%</b>

Pela Tabela 2.1, pode-se notar que o grupo de clientes, o qual a instituição não deseja avaliar pela ferramenta a ser desenvolvida (Clientes Excluídos), representa 10% do total dos clientes. Outro fato a ser ressaltado é que nessa amostra existe uma relação de aproximadamente 20 clientes bons para cada cliente ruim ( $28.621/1.389$ ).

## 2.5 Quantidade de Parcelas dos Contratos

Apesar de ser um produto desenvolvido para conter um prazo de financiamento de pelo menos 1 ano, podem existir contratos de somente 1 mês de duração. Ou seja, a quantidade de parcelas pode variar de um contrato para o outro. O problema de se considerar contratos de prazos muito distintos é que pode-se privilegiar, considerando-se a Definição de Performance, os clientes com contratos mais curtos, uma vez que esses usaram o produto por menos tempo e, assim, a chance desses se mostrarem ruins é menor do que a dos clientes com contratos mais longos. Uma alternativa para solucionar esse problema é excluir contratos muito curtos e muito longos, por serem atípicos. Com base na experiência de crédito, foram considerados somente os clientes cujos contratos tiveram duração de 12 a 36 meses. Desta forma, pretende-se trabalhar com um grupo de clientes mais homogêneo quanto ao tempo de utilização do produto. Considerando-se somente os clientes com prazo de contrato entre 12 e 36 meses, a Distribuição de Performance dos clientes é dada pela Tabela 2.2.

**Tabela 2.2** Tabela de Definição de Performance para Contratos com Duração entre 12 e 36 Meses.

Definição de Performance	Categoria	N	%
Excluídos	Funcionários	2.384	8,6%
	Veículos Pesados	661	2,4%
	<i>Subtotal</i>	<i>3.045</i>	<i>11,0%</i>
Ruins	Ação Legal Tomada/ Veículo reapropriado	313	1,1%
	3 ou mais parcelas consecutivas vencidas	651	2,4%
	2 vezes ou mais 2 parcelas consecutivas vencidas	133	0,5%
	4 vezes ou mais 1 parcela vencida	190	0,7%
	<i>Subtotal</i>	<i>1.287</i>	<i>4,7%</i>
Indeterminados	1 vez 2 parcelas consecutivas vencidas	198	0,7%
	3 vezes 1 parcela vencida	99	0,4%
	<i>Subtotal</i>	<i>297</i>	<i>1,1%</i>
Bons	2 vezes 1 parcela vencida	227	0,8%
	1 vez 1 parcela vencida	667	2,4%
	Nunca Inadimplente	22.091	80,0%
	<i>Subtotal</i>	<i>22.985</i>	<i>83,2%</i>
Total	<i>Total</i>	<i>27.614</i>	<i>100,0%</i>

## 2.6 Amostras de Desenvolvimento e Validação

Para este estudo, dispõem-se de todas as contratações do produto, realizadas no período de Julho de 1996 a Junho de 1997, totalizando 33.691 contratações. Com a exclusão dos contratos com menos de 12 ou mais de 36 meses de duração e daqueles que correspondem à categoria de Performance “Excluídos”, restam 24.569 contratações.

Como descrito na Seção 2.4, os clientes denominados indeterminados, representam o grupo de clientes cujo comportamento de crédito ainda não é suficientemente claro para indicá-los como bons ou maus pagadores. Duas alternativas podem ser adotadas para contornar esse problema:

- a) assumir que esse grupo apresenta características particulares distintas dos outros dois e construir um modelo capaz de discriminar os 3 grupos, ou;
- b) desconsiderar esse grupo, pela pequena quantidade de clientes nestas condições (297 casos), e construir um modelo que discrimine somente os clientes bons e ruins.

Na prática, os procedimentos usuais baseiam-se na alternativa (b), ou seja, simplesmente, desconsideram o grupo dos clientes classificados como indeterminados no processo de modelagem. Neste trabalho, essa alternativa também será utilizada. Sendo assim, tem-se uma amostra com 22.985 clientes bons e 1.287 clientes ruins, totalizando 24.272 clientes.

Em termos financeiros, o custo de se classificar inadequadamente um cliente ruim é muito maior do que o erro na classificação de um cliente bom. Isso se deve ao fato de que se a instituição concede o empréstimo a um cliente que quase certamente não irá pagar, é muito provável que a instituição perca integralmente o valor emprestado. Já no caso de a instituição não aprovar um crédito para um cliente bom, ela só estará deixando de ganhar com os encargos relativos ao empréstimo.

Por exemplo, supondo que todos os empréstimos de uma instituição sejam do valor de R\$ 1000,00 a serem pagos em 12 parcelas com uma taxa de juros de 10% no período inteiro. Supondo, também, que os clientes bons sempre pagam suas parcelas em dia e os ruins não chegam a pagar nenhuma das parcelas. A partir de um cálculo simples, seriam necessários 10 clientes bons para recuperar o prejuízo de haver aprovado um empréstimo para um cliente ruim.

Por esse motivo, uma característica muito comum na construção de Modelos de “Credit Scoring” é a utilização de amostras com a mesma quantidade de clientes bons e ruins. Determina-se uma quantidade satisfatória de clientes ruins e depois seleciona-se aleatoriamente a mesma quantidade de bons. Segundo os profissionais da área (ver Makuch, 1998), quando se utiliza uma amostra proporcional à população, como a quantidade de clientes bons é sempre muito maior do que a de ruins, o modelo final acaba sendo excelente para discriminar os clientes bons, porém, ineficiente para discriminar os ruins.

Neste trabalho, uma alternativa semelhante à apresentada acima é abordada. Em vez de se reduzir o número de clientes bons para equilibrar ao número de ruins, utiliza-se uma ponderação, de forma a fazer com que os clientes ruins tenham peso equivalente ao dos bons nas aplicações das técnicas. Desta forma, todos os clientes da amostra são considerados na análise.

Com o objetivo de se utilizar toda a amostra de clientes bons e ruins tanto para o desenvolvimento dos modelos quanto para a validação, utilizou-se o conceito da validação cruzada (ver Breiman et. al., 1984), onde os 24.272 clientes foram divididos, aleatoriamente, em 10 partições amostrais de tamanhos equivalentes. A partir dessas partições, foram construídos 10 conjuntos sendo cada um formado por 9 partições para desenvolvimento do modelo e 1 para validação. Aplicam-se as técnicas nos 10 grupos de desenvolvimento e avaliam-se os resultados nos respectivos grupos de validação.

As amostras de validação são utilizadas para verificar se o modelo estimado mantém seu poder de discriminação para amostras provindas da mesma população da amostra de desenvolvimento. Se o poder de discriminação variar muito de uma amostra para outra, pode significar que o modelo não é estável ou pode estar havendo uma superestimação (“overfitting”).

## 2.7 Tratamento das Variáveis

Com base na Definição de Performance, descrita na Seção 2.4, foi criada a variável resposta IDBR (Indicador de cliente Bom ou Ruim), que recebe os seguintes valores:

$$\text{IDBR} = \begin{cases} 0, & \text{se o cliente foi definido como RUIM;} \\ 1, & \text{se o cliente foi definido como BOM.} \end{cases}$$

Com relação às variáveis independentes, em geral, as informações que as instituições financeiras possuem sobre seus clientes apresentam as mais diversas formas, desde variáveis dicotômicas (ex.: sexo) até códigos (ex.: estado civil do cliente).

Uma característica muito comum nos dados dessas instituições é a grande freqüência de valores faltantes (“missing values”), principalmente, no caso de variáveis relativas às informações cadastrais do cliente.

Para analisar esse conjunto de variáveis, duas são as abordagens mais utilizadas na prática pelos profissionais da área. Os tratamentos das variáveis variam de acordo com a técnica de estimação dos modelos. Alguns profissionais usam as variáveis na sua forma natural, resultando em dois grupos: variáveis quantitativas e variáveis qualitativas. Esse procedimento pode comprometer a aplicação de algumas técnicas devido a suposições relativas a elas. Uma outra alternativa é a categorização das variáveis quantitativas e a utilização de somente variáveis categorizadas. Essa segunda abordagem é mais freqüente, pois apesar da clara perda de informação na categorização das variáveis, apresenta uma série de ganhos, que serão citados posteriormente. Neste trabalho, as variáveis serão tratadas de acordo com as restrições das técnicas que serão aplicadas.

### 2.7.1 Valores Faltantes (“Missing Values”)

A presença dos valores faltantes é muito comum nesta área do mercado financeiro, pois, a maior preocupação das instituições está voltada para a rentabilidade de seus produtos e não para o enriquecimento de suas bases de dados. O maior motivo da grande quantidade de “missings” deve-se ao mau preenchimento dos instrumentos cadastrais, como por exemplo: fichas cadastrais, fichas de abertura de conta corrente, fichas de proposta de negócio e etc. Já os dados relativos à utilização de produtos e apontamentos cadastrais dos clientes dificilmente apresentam problemas de valores faltantes.

Os “missing values” podem ser divididos, basicamente, em duas categorias: informação não concedida pelo cliente ou informação não solicitada pela instituição. No caso de a informação haver sido sonegada pelo cliente, pode mostrar indícios de que o cliente não as informou por serem negativas para o seu cadastro. Já no caso das informações faltantes por não questionamento da instituição, não se pode afirmar que isso esteja relacionado ao grau de risco de inadimplência do cliente. Portanto, a melhor alternativa de estudo seria a subdivisão do grupo de “missings”. Infelizmente, na base de dados em questão, não há como se diferenciar esses dois grupos citados. De qualquer forma, o conjunto de “missing values” será utilizado como uma categoria específica da variável, que pode representar menor ou maior risco frente às demais. Em nenhum momento do trabalho os clientes que apresentaram “missing values” para alguma variável foram excluídos da análise.

### 2.7.2 Categorização de Variáveis

Os ganhos relacionados à categorização das variáveis quantitativas podem ser atribuídos à padronização de resultados, estabilidade do modelo e transformação de variáveis (ver Gruenstein, 1998).

**Padronização de Resultados:** quando se constrói modelos para aplicação na área de crédito, deve-se primar pela simplicidade do uso e da implementação, tendo em vista que

nem sempre os profissionais que manusearão os modelos têm formação matemática-estatística. Desta forma, se as variáveis do modelo puderem ser mostradas em categorias, torna-se mais simples a implementação dos modelos e interpretação dos pesos relativos às categorias das variáveis.

***Estabilidade do Modelo:*** ao se trabalhar com variáveis quantitativas deve-se atentar para a eventualidade da aparição de um ou mais valores discrepantes (“outliers”), pois esses, podem afetar de forma assintosa os resultados. Já com a utilização das variáveis em categorias, esse problema é minimizado, melhorando, assim, a estabilidade do modelo.

***Transformação de Variáveis:*** quando uma variável independente de natureza quantitativa não apresenta relação direta (ex.: linear) com a variável resposta, um procedimento comum é a transformação da variável, na tentativa de verificar a relação entre ela e a resposta. As transformações usuais não são muito bem aceitas nas instituições financeiras, pois, podem dificultar a interpretação dos resultados. É difícil, por exemplo, interpretar a raiz quadrada da idade do cliente como indicador para a concessão ou não do crédito. A categorização das variáveis quantitativas, de certa forma, pode ser vista como uma transformação, pois há condições de serem agrupadas categorias com o mesmo comportamento frente a variável resposta.

Deve-se destacar que todo o processo de categorização de variáveis, será sempre guiado conforme a variável resposta (IDBR), ou seja, as categorias das variáveis independentes serão formadas de acordo com a relação com a variável IDBR. Inicialmente, são formadas categorias da variável preditora a partir da experiência dos analistas de crédito. Caso não seja uma variável de comportamento conhecido, o procedimento de divisão é feito a partir de percentis da distribuição da variável. Por exemplo, são estabelecidos os decis e verifica-se a freqüência de bons e maus clientes para cada categoria criada. Trata-se de um procedimento exploratório que visa identificar categorias semelhantes das variáveis com a relação a bons e maus clientes.

Após a criação das categorias, são calculadas 2 medidas:

- Razão de Bons e Ruins (%B/%R): proporção de bons na categoria sobre a proporção de ruins na categoria;
- “Weights of Evidence” (WOE) (Good, 1950): é o resultado do logaritmo natural da razão de bons e ruins.

O cálculo das medidas está exemplificado na Tabela 2.3:

**Tabela 2.3** Tabela de Exemplo para a Categorização de Variáveis

Categoria	Num. Bons	Num. Ruins	%Bons	%Ruins	%B/%R	WOE
Categ1	b <sub>1</sub>	r <sub>1</sub>	b <sub>1</sub> / b <sub>.</sub>	r <sub>1</sub> / r <sub>.</sub>	(b <sub>1</sub> / b <sub>.</sub> )/( r <sub>1</sub> / r <sub>.</sub> )	Ln [(b <sub>1</sub> / b <sub>.</sub> )/( r <sub>1</sub> / r <sub>.</sub> )]
Categ2	b <sub>2</sub>	r <sub>2</sub>	b <sub>2</sub> / b <sub>.</sub>	r <sub>2</sub> / r <sub>.</sub>	(b <sub>2</sub> / b <sub>.</sub> )/( r <sub>2</sub> / r <sub>.</sub> )	Ln [(b <sub>2</sub> / b <sub>.</sub> )/( r <sub>2</sub> / r <sub>.</sub> )]
Categ3	b <sub>3</sub>	r <sub>3</sub>	b <sub>3</sub> / b <sub>.</sub>	r <sub>3</sub> / r <sub>.</sub>	(b <sub>3</sub> / b <sub>.</sub> )/( r <sub>3</sub> / r <sub>.</sub> )	Ln [(b <sub>3</sub> / b <sub>.</sub> )/( r <sub>3</sub> / r <sub>.</sub> )]
Categ4	b <sub>4</sub>	r <sub>4</sub>	b <sub>4</sub> / b <sub>.</sub>	r <sub>4</sub> / r <sub>.</sub>	(b <sub>4</sub> / b <sub>.</sub> )/( r <sub>4</sub> / r <sub>.</sub> )	Ln [(b <sub>4</sub> / b <sub>.</sub> )/( r <sub>4</sub> / r <sub>.</sub> )]
Categ5	b <sub>5</sub>	r <sub>5</sub>	b <sub>5</sub> / b <sub>.</sub>	r <sub>5</sub> / r <sub>.</sub>	(b <sub>5</sub> / b <sub>.</sub> )/( r <sub>5</sub> / r <sub>.</sub> )	Ln [(b <sub>5</sub> / b <sub>.</sub> )/( r <sub>5</sub> / r <sub>.</sub> )]
Total	b <sub>.</sub>	r <sub>.</sub>	1	1	1	0

Onde:

b<sub>i</sub> : número de clientes bons na categoria i;

r<sub>i</sub> : número de clientes ruins na categoria i;

$$b_{\cdot} = \sum_{i=1}^s b_i \quad \text{e} \quad r_{\cdot} = \sum_{i=1}^s r_i$$

O WOE é uma medida descritiva que auxilia a identificação de categorias com alto ou baixo poder de discriminação, além de identificar aquelas categorias que discriminam melhor os bons clientes e aquelas que discriminam melhor os clientes ruins. Sejam os seguintes casos:

- **WOE = 0 (zero)**: isto significa que a razão entre bons e ruins é 1, indicando que se a variável assumir um valor dessa categoria, não há nenhum indício de o cliente ser de maior ou menor risco comparado à análise desconsiderando essa variável;

- **WOE > 0 (zero)**: positivo e quanto mais distante de zero, maiores são as chances de o cliente apresentar menos risco de crédito, isso significa que a categoria apresenta algum poder para discriminar clientes bons;
- **WOE < 0 (zero)**: negativo e quanto mais distante de zero, maiores são as chances de o cliente apresentar maior risco de crédito, isso significa que a categoria apresenta algum poder para discriminar clientes ruins.

Outra vantagem no uso do WOE na categorização das variáveis é a de que pode-se agrupar categorias com valores de WOE próximos, desde que exista uma interpretação plausível quanto à lógica de crédito. Esse procedimento auxilia na redução do número de categorias da variável tornando-a mais estável.

As tabelas contendo as categorias e respectivos “Weights of Evidence” para cada variável preditora, após o procedimento de categorização, encontram-se no Apêndice A (de A1 a A17).

## CAPÍTULO 3

### Metodologia

O objetivo deste Capítulo é descrever as técnicas estatísticas aplicadas ao problema. Outras metodologias e técnicas de construção de Modelos de “Credit Scoring” podem ser consultadas em Hoyland (1997), Hand e Henley (1997) e Thomas (1998). As aplicações das técnicas aos dados são apresentadas no Capítulo 4.

#### 3.1 Regressão Logística Múltipla

A Análise de Regressão Logística Múltipla (ver Hosmer e Lemeshow, 1989 e McCullagh e Nelder, 1989) para uma resposta binária é a técnica mais utilizada no desenvolvimento de Modelos de “Credit Scoring”. A grande extensão de seu uso deve-se a algumas vantagens oferecidas pela técnica:

- É a mais utilizada entre os profissionais da área (culturalmente difundida);
- Não apresenta problemas sérios de suposições, como, por exemplo, na Análise Discriminante, onde se pressupõe uma distribuição Normal Multivariada para as variáveis independentes (ver, Eisenbeis, 1977 e 1978);
- Apresenta facilidade computacional, uma vez que os pacotes estatísticos mais utilizados pelas instituições permitem o seu uso;
- É uma ferramenta poderosa para discriminação e é aplicável aos dados.

Vale ressaltar que o objetivo do estudo é classificar os clientes em 2 grupos distintos (bons e maus pagadores). Com o uso da técnica, pode-se atribuir a cada cliente

probabilidades de que ele pertença aos grupos de interesse. Desta forma, a Regressão Logística está sendo considerada como uma técnica de classificação onde o cliente será classificado no grupo que apresenta a maior probabilidade. A avaliação dos resultados do modelo também será feita enfocando esse objetivo.

Seja  $Y_i \in \{0,1\}$  a variável resposta ( $0 =$  “o  $i$ -ésimo cliente é ruim”,  $1 =$  “o  $i$ -ésimo cliente é bom”) e  $\mathbf{x} = (1, x_1, \dots, x_k)$  o vetor das  $k$  variáveis preditoras. Pode-se escrever o modelo de Regressão Logística Múltipla como um caso particular dos Modelos Lineares Generalizados (ver McCullagh e Nelder, 1989), com a seguinte função de ligação:

$$\ln\left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right) = \boldsymbol{\beta}^t \mathbf{x} ,$$

onde  $\boldsymbol{\beta}^t = (\beta_0, \beta_1, \dots, \beta_k)$  são os coeficientes que serão estimados e  $\pi(\mathbf{x}) = E[Y=1|\mathbf{x}] = P[Y=1|\mathbf{x}]$  a probabilidade de que o cliente seja bom dado as variáveis independentes. Essa probabilidade é então calculada como a Distribuição Logística,

$$\pi(\mathbf{x}) = \frac{\exp(\boldsymbol{\beta}^t \mathbf{x})}{1 + \exp(\boldsymbol{\beta}^t \mathbf{x})} .$$

O modelo pode ser escrito como:

$$Y = \pi(\mathbf{x}) + \varepsilon .$$

Tem-se que:

$$E[\varepsilon] = \{[1 - \pi(\mathbf{x})]\pi(\mathbf{x}) + [-\pi(\mathbf{x})][1 - \pi(\mathbf{x})]\} = 0 ,$$

e

$$\text{Var}[\varepsilon] = E[\varepsilon^2] - E^2[\varepsilon] = E[\varepsilon^2] = \pi(\mathbf{x}) [1 - \pi(\mathbf{x})].$$

O vetor de parâmetros  $\beta$  é estimado maximizando-se a função de verossimilhança em relação aos  $k+1$  elementos do vetor. As soluções das equações são obtidas por métodos iterativos. Em geral, o método utilizado é o de Mínimos Quadrados Reponderados. No passo  $(m+1)$  desse método tem-se:

$$\hat{\beta}^{(m+1)} = (\mathbf{X}^t \mathbf{V}^{(m)} \mathbf{X})^{-1} \mathbf{X}^t \mathbf{V}^{(m)} \mathbf{z}^{(m)},$$

onde,

$$m = 0, 1, \dots;$$

$\mathbf{X}_{(nx(k+1))}$ : é a matriz que contém as  $n$  observações para cada uma das  $k$  variáveis preditoras, sendo a primeira coluna da matriz composta por  $n$  repetições do valor 1;

$$\mathbf{V}_{(nxn)} = \text{diag}\{\pi_1(1-\pi_1), \pi_2(1-\pi_2), \dots, \pi_n(1-\pi_n)\};$$

$$\pi_i = P[Y_i=1 | \mathbf{x}_i];$$

$\mathbf{z} = (z_1, z_2, \dots, z_n)^t$ : vetor das variáveis independentes modificadas;

$$z_i = \eta_i + \frac{(y_i - \pi_i)}{[\pi_i(1-\pi_i)]};$$

$$\eta_i = \ln\left[\frac{\pi_i}{\pi_i(1-\pi_i)}\right];$$

$$i = 1, 2, \dots, n.$$

Assintoticamente, a variância de  $\hat{\beta}$  é dada por  $(\mathbf{X}^t \mathbf{V} \mathbf{X})^{-1}$ .

A estatística do Teste da Razão de Verossimilhanças é dada por,

$$D = -2 \sum_{i=1}^n \left[ y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i) \ln\left(1-\frac{\hat{\pi}_i}{y_i}\right) \right],$$

onde  $\hat{\pi}_i = \hat{\pi}(x_i)$  segue uma distribuição de Qui-Quadrado sob a hipótese nula  $((\beta_1, \dots, \beta_k) = (0))$ .

No caso do teste univariado sobre cada parâmetro ( $H_0: \beta_j = 0, j=0,1,\dots,k$ ), uma alternativa é a utilização do Teste de Wald (Rao, 1973), cuja estatística é dada por:

$$W_j = \frac{\hat{\beta}_j}{\hat{EP}(\hat{\beta}_j)},$$

onde  $\hat{EP}(\hat{\beta}_j)$  é o estimador do erro padrão de  $\hat{\beta}_j$ . Sob a hipótese nula,  $W_j$  segue uma distribuição Normal Padrão.

### 3.2 CHAID (“Chi-squared Automatic Interaction Detection”)

Pode-se classificar os modelos de segmentação em dois grupos de acordo com o tipo de segmentação. Se não há utilização de variável dependente (resposta), trata-se de um modelo desenvolvido a partir de análises de agrupamento e outras técnicas multivariadas, em contrapartida, a utilização de uma variável dependente define um segundo grupo de modelos de segmentação, do qual será feita a abordagem.

Os modelos de segmentação baseados em variáveis dependentes definem, a partir de combinações de variáveis independentes, segmentos homogêneos com relação a uma variável dependente.

Em 1963, Morgan e Sonquist propuseram o algoritmo AID (Automatic Interaction Detection), caracterizando-se pelo pioneirismo na detecção automática de interações entre variáveis (Morgan & Sonquist, 1963).

O CHAID é apresentado em Kass (1980) e representa uma evolução do algoritmo AID para o caso de variáveis dependentes categorizadas nominais. O método consiste em dividir a população em dois ou mais grupos baseados nas categorias da variável que melhor prediz a variável resposta. Então, quebram-se esses subgrupos em subgrupos menores baseados em outras variáveis preditoras. O processo de quebras continua até que não existam mais variáveis estatisticamente significantes que possam ser utilizadas para novas subdivisões, ou até que alguma outra regra de parada seja alcançada. Os subgrupos derivados do CHAID são mutuamente exclusivos, onde a soma dos elementos de cada subgrupo é igual ao total de elementos da população.

A aplicação do CHAID é altamente indicada quando o objetivo é o de produzir e analisar todas as classificações cruzadas de variáveis e suas categorias. O CHAID permite a automatização de grande parte do processo, rejeitando as classificações não significativas e permitindo que o foco seja dado à análise dos subgrupos que apresentam forte poder preditivo. Uma descrição mais detalhada do algoritmo pode ser encontrada em Magidson (1993).

### **3.2.1 Método para Variáveis Dependentes Nominais**

O método consiste de três estágios executados iterativamente para a determinação dos subgrupos finais: Agrupamento, Segmentação e Parada.

#### *Estágio 1: Agrupamento*

Para cada variável independente,  $X_1, X_2, \dots, X_k$ , são agrupadas as categorias consideradas não significativas, de acordo com os seguintes passos:

*Passo 1:* Constrói-se uma tabela cruzada entre cada variável independente  $X_j$  e a variável resposta  $Y$ ;

*Passo 2:* Para cada par de categorias escolhidas para serem agrupadas, calcula-se a estatística  $\chi^2$  (ver Seção 3.2.3) para testar a independência entre o par de categorias e a variável resposta Y. Calcula-se, então, o nível descritivo para cada um dos testes;

*Passo 3:* Verifica-se para cada variável  $X_j$  se há algum caso onde os testes elaborados no Passo 2 não foram estatisticamente significativos. Caso exista algum, esses pares de categorias devem ser agrupados em uma única categoria, formando uma nova categoria e deve-se seguir para o Passo 4. Se todos os testes forem estatisticamente significativos, deve-se seguir para o Passo 5;

*Passo 4:* Cada vez que 3 ou mais categorias iniciais das variáveis independentes são agrupadas, verifica-se se alguma delas não deve ser separada novamente, utilizando os mesmos testes  $\chi^2$  para comparar cada categoria isolada com as demais agrupadas. Caso alguma categoria deva ser desagrupada, retorna-se ao Passo 2;

*Passo 5:* Agrupa-se qualquer categoria que apresente freqüência abaixo da especificada inicialmente (tamanho mínimo de um subgrupo) com a categoria que apresenta a maior similaridade, ou seja, a que apresenta o menor valor da estatística  $\chi^2$ ;

*Passo 6:* Calcula-se o nível descritivo ajustado de Bonferroni (ver Seção 3.2.4) considerando as categorias agrupadas para cada variável independente.

### *Estágio 2: Segmentação*

Verifica-se qual variável independente apresenta o menor nível descritivo ajustado de Bonferroni quando testada a independência contra a variável resposta. Caso essa variável apresente um nível descritivo com valor abaixo do nível de significância determinado inicialmente, ela é utilizada para a segmentação quebrando o grupo inicial em subgrupos representados por suas categorias (agregadas). Então, são recalculadas as distribuições de freqüência em cada subgrupo. Se nenhuma variável independente

apresentar um nível descritivo com valor abaixo do nível de significância determinado inicialmente, não se efetua nenhuma quebra.

### *Estágio 3: Parada*

Os estágios anteriores são repetidos até que todos os subgrupos possíveis tenham sido analisados ou tenha sido atingido algum dos dois critérios de parada: tamanho mínimo de um subgrupo e nível descritivo com valor abaixo do nível de significância determinado inicialmente.

Deve-se ressaltar que o algoritmo não garante que a segmentação ótima seja encontrada, porém, apresenta bons resultados na prática e requer procedimentos computacionais simples (Kass, 1980).

### **3.2.2 Tipos de Variáveis Independentes**

O CHAID permite que três tipos de variáveis sejam definidas antes da execução do algoritmo: livre, flutuante e monótona. Essa definição afeta o procedimento de agrupamento de categorias e o cálculo dos níveis de significância feitos pelo algoritmo.

A variável monótona é aquela que apresenta as categorias em uma determinada ordem e somente poderão ser agrupadas as categorias vizinhas. A variável livre apresenta as categorias na forma nominal, indicando que se pode fazer qualquer tipo de agrupamento. Já a variável flutuante, assim como a variável monótona, apresenta categorias ordenadas, com exceção de uma delas. Essa última concepção é bastante útil quando se trabalha com categorias desconhecidas ou que representam valores faltantes.

As variáveis independentes que são dicotômicas não precisam ser definidas inicialmente, uma vez que o algoritmo faz um só tratamento para esse tipo de variável.

### 3.2.3 Teste da Hipótese de Independência

Considerando-se uma tabela Ix2, com  $n_{ij}$  ( $i = 1, \dots, I$  e  $j = 1, 2$ ) a freqüência para cada célula, denota-se A como a variável preditora (linha) com I categorias e B a variável resposta (coluna) com 2 categorias. O CHAID considera cada célula de freqüência na tabela Ix2 como proveniente do seguinte modelo log-linear saturado (Haberman, 1978):

$$H_1: \ln\left(\frac{N_{ij}}{Z_{ij}}\right) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}, \quad (3.1)$$

onde  $N_{ij}$  é a freqüência esperada para a célula  $(i, j)$ , sob as seguintes condições:

$$Z_{ij} = \frac{1}{W_{ij}},$$

onde  $W_{ij}$  é o peso amostral médio (caso seja usada alguma variável de ponderação) e os termos  $\lambda$  são parâmetros do modelo sujeitos às seguintes condições de identidade:

$$\sum_i \lambda_i^A = \sum_j \lambda_j^B = 0 \quad \text{e} \quad \sum_i \lambda_{ij}^{AB} = \sum_j \lambda_{ij}^{AB} = 0.$$

Sob essas condições, o nível descritivo antes de se agregarem as categorias é o nível descritivo associado à hipótese nula de independência contra a hipótese alternativa  $H_1$ , dada por (3.1):

$$H_0: \lambda_{ij}^{AB} = 0,$$

onde  $i = 1, 2, \dots, I$  e  $j = 1, 2$ .

No CHAID, as freqüências esperadas para as células são calculadas utilizando o algoritmo WLM (“Weighted Loglinear Modeling”) (ver Magidson, 1987).

Se nenhuma variável ponderadora é usada, cada  $Z_{ij}$  é igual a 1 e  $H_1$  pode ser representada na forma usual dos modelos log-lineares. Assim, o algoritmo converge logo

após a primeira iteração. Se for usada uma variável de ponderação, o algoritmo WLM requer algumas iterações para convergir. O critério de convergência pode ser controlado pelos parâmetros  $\varepsilon$  e  $maxit$  (número máximo de iterações), ocorrendo a parada quando: os parâmetros estimados mudarem em uma magnitude inferior a  $\varepsilon$  de uma iteração para outra; ou o número máximo de iterações atingir  $maxit$ .

Após a convergência, calcula-se a estatística de Qui-Quadrado da Razão de Verossimilhanças, dada por:

$$\chi^2_{RV}(H_0) = 2 \sum_i \sum_j n_{ij} \ln \left( \frac{n_{ij}}{\hat{N}_{ij}} \right),$$

onde  $i = 1, \dots, I$  e  $j = 1, 2$ .

O nível descritivo é então obtido de uma distribuição Qui-Quadrado com  $(I-1)$  graus de liberdade.

### 3.2.4 Nível Descritivo Ajustado de Bonferroni

Quando duas ou mais categorias da variável independente são agrupadas, o CHAID estima um nível descritivo aplicando os passos precedentes para o nível descritivo não-ajustado e, então, aumenta esse nível multiplicando-o pelo multiplicador de Bonferroni  $M$  (ver Kass, 1980).

Por exemplo, suponha que uma variável independente A tenha 4 categorias e que deve ser dicotomizada. Suponha, também, que  $\alpha$  é o Erro do Tipo I desejado, associado ao teste de independência na tabela 4x2 de Ax B, sendo B a variável resposta. Considerando os testes para cada uma das 6 possíveis formas de se dicotomizar A, se os 6 testes forem independentes um dos outros, a probabilidade de se cometer um erro do tipo I em um ou mais desses testes é:

$$1 - (1 - \alpha)^6, \quad (3.2)$$

que é significativamente maior que  $\alpha$ . Quando  $\alpha$  tende a zero, a probabilidade dada pela Equação (3.2) tende a  $6\alpha$ . Generalizando, se  $M$  é igual ao número de maneiras de finalizar a variável com  $I' \leq I$  categorias, então, para  $\alpha$  pequeno tem-se:

$$1 - (1 - \alpha)^M = M\alpha.$$

Se forem realizados  $M$  testes, o nível descritivo deve ser  $\alpha/M$  ou menor, para ser considerado significante ao nível de significância  $\alpha$ .

Para uma variável preditora livre,  $M$  é dado por:

$$M = \sum_{i=0}^{I'-1} (-1)^i \frac{(I'-i)^I}{i!(I'-i)!}.$$

Para uma variável preditora monótona,  $M$  é dado por:

$$M = \binom{I-1}{I'-1}.$$

Para uma variável preditora flutuante,  $M$  é dado por:

$$M = \binom{I-1}{I'-1} \left( \frac{I'-1 + I'(I-I')}{I-1} \right).$$

### 3.3 REAL (“Real Attribute Learning Algorithm”)

Assim como o CHAID, o REAL é um modelo de segmentação baseado em variáveis dependentes. Esse algoritmo foi proposto por Stern et. al. (1998) e traz como grande vantagem frente ao CHAID a possibilidade do uso de variáveis independentes contínuas. O algoritmo foi construído com o objetivo de aplicação no mercado brasileiro de ações, de forma a ser uma ferramenta de suporte às estratégias de operações financeiras. A aplicação do REAL na área de crédito é pioneira, e assim como as demais técnicas, os

resultados de sua aplicação aos dados são apresentados no Capítulo 4. A comparação do REAL com outros métodos de árvores de classificação da mesma família pode ser encontrada no trabalho de Lauretto (1996).

### 3.3.1 Descrição do Método de Construção da Árvore de Classificação

O REAL é um algoritmo iterativo onde cada iteração principal corresponde à ramificação de um nó terminal da árvore. A classificação dos elementos que compõem este nó terminal se dá a partir da escolha da variável preditora e dos novos ramos gerados para cada intervalo de valores determinados da variável em questão. O processo de discretização ocorre quando valores adjacentes são agrupados em intervalos mutuamente excludentes. Apesar de o REAL ter sido criado para a aplicação em variáveis independentes contínuas, a utilização de variáveis categorizadas é possível, desde que haja uma ordenação dos valores das categorias das variáveis. No caso de variáveis nominais, pode-se utilizar a relação com a variável dependente para estabelecer a ordem das categorias (por exemplo: a proporção de clientes bons dentro das categorias da variável nominal). Dessa forma, o processo de discretização pode ser entendido como um processo de redução de categorias no caso das variáveis categorizadas.

Cada iteração principal do algoritmo corresponde aos seguintes passos:

*Passo 1:* Efetua-se a discretização (ou redução de categorias) de cada variável preditora e, em seguida, faz-se a avaliação de cada uma segundo uma determinada função de perda (ver Seção 3.3.3);

*Passo 2:* Seleciona-se a melhor variável preditora e a correspondente ramificação do nó;

*Passo 3:* Agrupam-se os intervalos que não atingiram um limite mínimo especificado de convicção (ver Seção 3.3.2).

### 3.3.2 Função de Convicção

Para um determinado nó da árvore, pode-se dividir os  $n$  clientes em 2 grupos:  $k$  clientes classificados erroneamente (no estudo: cliente bom classificado como ruim ou cliente ruim classificado como bom) e  $(n - k)$  clientes classificados corretamente. Define-se  $p$  como a probabilidade de um cliente ser classificado corretamente no nó e, consequentemente,  $q = (1 - p)$  como a probabilidade de erro na classificação. Assumindo que a distribuição a priori de  $q$  é uma Uniforme(0,1), tem-se que a probabilidade a posteriori de  $q$  dado as variáveis independentes é:

$$P(q | x) = \frac{P(x | q)f(q)}{\int_0^1 P(x | q)f(q)dq} = \frac{\binom{n}{k}q^k(1-q)^{n-k}}{\int_0^1 \binom{n}{k}q^k(1-q)^{n-k}dq}. \quad (3.3)$$

Desta forma, a partir de (3.3) a função de distribuição a posteriori  $D(c)$  pode ser escrita como:

$$D(c) = P(q \leq c | x) = \frac{\int_0^c \binom{n}{k}q^k(1-q)^{n-k}dq}{\int_0^1 \binom{n}{k}q^k(1-q)^{n-k}dq} = \frac{\int_0^c \binom{n}{k}q^k(1-q)^{n-k}dq}{\int_0^1 \binom{n}{k}q^k(1-q)^{n-k}dq} = \frac{\int_{q=0}^c \text{Bin}(n, k, q) dq}{\int_{q=0}^1 \text{Bin}(n, k, q) dq}, \quad (3.4)$$

onde  $\text{Bin}(n, k, q)$  é representação da distribuição Binomial.

A medida de convicção é então definida como:  $mc = 100*(1 - v)\%$ , onde:

$$v = \min \{ f(c) = 0 : c \in [0,1] \} \quad (3.5)$$

e  $f(c)$  é dada por:

$$f(c) = P(q \leq c) - (1 - g(c)) \quad (3.6)$$

e  $g(c)$  é uma bijeção monotonicamente crescente de  $[0,1]$  em si mesmo. A função  $g(c)$  adotada é:

$$g(c) = c^r, r \geq 1, \quad (3.7)$$

que é uma função convexa, onde  $r$  será denominado parâmetro de convexidade, e deverá ser fornecido pelo usuário. Substituindo (3.4) e (3.7) em (3.6) tem-se:

$$f(c) = \frac{\int_{q=0}^c \text{Bin}(n, k, q)}{\int_{q=0}^1 \text{Bin}(n, k, q)} - (1 - c^r), \quad (3.8)$$

e, finalmente,  $v$  é a raiz da função monotonicamente decrescente:

$$v = \min \{ f(c) = 0 : c \in [0,1] \}.$$

### 3.3.3 Função de Perda

Para avaliar qual é a melhor variável preditora que expandirá o novo nó, todas as variáveis são discretizadas em intervalos seguindo os procedimentos descritos na próxima seção. Feita a discretização de todas as variáveis, seleciona-se aquela que minimiza a função de perda ( $fp$ ). Essa função é baseada na medida de convicção  $v$ . Nesse caso, a função de perda é soma das medidas de convicção dos intervalos ponderada pelo número de clientes em cada intervalo:

$$fp = \sum_i n_i v_i. \quad (3.9)$$

### 3.3.4 Procedimento de Discretização das Variáveis Preditoras

O primeiro passo do procedimento de discretização consiste em ordenar os clientes com relação à variável preditora em estudo. Em seguida, agrupam-se todos os clientes que pertencem à mesma classe de valores da variável preditora.

Os passos seguintes consistem em agrupar intervalos adjacentes de maneira a diminuir a perda global dentro do nó. Ou seja, o ganho em se agrupar  $J$  intervalos adjacentes  $I_{h+1}, I_{h+2}, \dots, I_{h+J}$ , é o decréscimo relativo na função de perda, dado pela seguinte função de ganho:

$$fg(h,j) = \sum_j fp(n_j, k_j, r) - fp(n, k, r), \quad (3.10)$$

onde  $n = \sum_j n_j$  e  $k$  é o número de clientes em minoria no novo intervalo.

No final de cada passo são concatenados os intervalos que apresentam ganho máximo, ocorrendo a parada do procedimento quando as concatenações já não apresentarem ganho positivo.

Após o procedimento de discretização, cada intervalo obtido constituirá um novo nó, que deverá sofrer os mesmos passos descritos anteriormente.

No final do processo de discretização pode-se deparar com o fato de que um conjunto de intervalos (novos nós) adjacentes apresentariam melhores resultados se fossem expandidos conjuntamente, em vez da expansão isolada de cada um. Por esse motivo, o usuário deve especificar o parâmetro de entrada  $v_c$ , de forma que sejam reagrupados todos os intervalos adjacentes que não satisfaçam à desigualdade  $v < v_c$ , onde  $v_c \in [0,1]$ . Com o objetivo de prevenir a ocorrência de um “loop” infinito do algoritmo, utiliza-se como função de perda do novo intervalo a soma das funções de perdas dos intervalos que serão reagrupados. Esse procedimento visa reparar os possíveis erros de discretização prematura das variáveis preditoras em grupos pouco representativos.

### 3.3.5 Critérios de Parada

Interrompe-se a expansão de um determinado nó da árvore caso não exista nenhuma variável preditora cuja discretização diminua a função de perda ( $f_p$ ), por um limite de precisão numérica  $\epsilon > 0$ .

Outro fator que determina a parada da árvore é o parâmetro de convicção  $v_c$ , de tal forma que quanto maior for seu valor, maior é o número de reagrupamentos e menor é o tamanho da árvore resultante. Após a parada do algoritmo, os nós terminais são rotulados pela classe majoritária.

Diferentemente do CHAID, onde pode-se estabelecer a freqüência mínima em cada nó e o número máximo de camadas da árvore, no REAL estas especificações não estão implementadas como critérios de parada do algoritmo.

## 3.4 Medidas de Avaliação dos Modelos

As medidas de avaliação são utilizadas para certificar se os modelos apresentam resultados satisfatórios, além de servirem para comparação das técnicas utilizadas.

Quando constrói-se um modelo de discriminação, de início já se conhece o “status” do cliente (bom ou ruim). A partir das variáveis preditoras, pode-se estimar, pelo modelo, se o cliente é bom ou ruim, de acordo com uma determinada regra decisória. No caso da Regressão Logística, essa regra é determinada por um ponto de corte estabelecido (por exemplo: 0,5). Caso a estimativa da probabilidade esteja acima do ponto de corte o cliente é classificado como bom, caso contrário, como ruim. Na Tabela 3.1, estão representadas as freqüências para o cruzamento entre as classificações observadas e as estimadas, dado um determinado ponto de corte.

**Tabela 3.1** Tabela de Classificação.

		<i>Estimado</i>		
<i>Observado</i>		<i>Ruim</i>	<i>Bom</i>	<i>Total</i>
<i>Ruim</i>	$n_{11}$	$n_{12}$	$n_{1\cdot}$	
<i>Bom</i>	$n_{21}$	$n_{22}$	$n_{2\cdot}$	
<i>Total</i>	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot \cdot}$	

Onde:

$n_{11}$  : cliente *Ruim* classificado como *Ruim*;

$n_{12}$  : cliente *Ruim* classificado como *Bom*;

$n_{21}$  : cliente *Bom* classificado como *Ruim*;

$n_{22}$  : cliente *Bom* classificado como *Bom*;

$$n_{\cdot 1} = n_{11} + n_{21};$$

$$n_{\cdot 2} = n_{12} + n_{22};$$

$$n_{1\cdot} = n_{11} + n_{12};$$

$$n_{2\cdot} = n_{21} + n_{22};$$

$$n_{\cdot \cdot} = n_{11} + n_{21} + n_{12} + n_{22}.$$

Pode-se definir:

$TAT = (n_{11} + n_{22}) / (n_{\cdot \cdot})$  como a Taxa de Acerto Total;

$TAR = (n_{11}) / (n_{1\cdot})$  como a Taxa de Acerto dos Ruins;

$TAB = (n_{22}) / (n_{2\cdot})$  como a Taxa de Acerto dos Bons.

As três taxas de acerto, definidas acima, serão utilizadas para verificar a performance de cada técnica, além de servirem como medidas comparativas.

## CAPÍTULO 4

### Aplicação

Neste capítulo, as três técnicas descritas no Capítulo 3 são aplicadas aos dados. Um problema semelhante a este foi abordado por Arminger et. al. (1997), onde foram utilizadas e comparadas três técnicas de classificação distintas para um problema de “Credit Scoring” (Regressão Logística, Árvores de Decisão e Redes Neurais). Assim como no trabalho de Arminger et. al., o objetivo é comparar as técnicas do ponto de vista de acerto na classificação dos clientes. Outros resultados comparativos entre técnicas de classificação são apresentadas por Srinivasan e Kim (1987), abordando procedimentos não-paramétricos, e Kronborg et. al. (1998) aplicando Redes Neurais.

Inicialmente, são apresentadas as formas como as variáveis foram consideradas em cada uma das técnicas, seguidas pela apresentação dos resultados e por último a comparação entre os modelos.

As análises foram feitas nos “softwares” SPSS for Windows v. 8.0 (Regressão Logística), CHAID for Windows v. 6.0.1 (CHAID) e REAL (software desenvolvido pelos autores).

#### 4.1 Variáveis Preditoras

O CHAID, por sua definição, exige que as variáveis independentes (preditoras) estejam dispostas em categorias. Já o REAL, tem como exigência que as variáveis preditoras representem atributos reais. A Regressão Logística aceita os dois tipos de variáveis. Para facilitar a interpretação dos resultados finais e atender as restrições das

técnicas utilizadas, as variáveis foram tratadas de forma distintas para cada uma das técnicas.

#### 4.1.1 Variáveis Preditoras na Regressão Logística Múltipla

Mesmo com a possibilidade de se utilizar todas as variáveis na sua forma original, objetivando reproduzir a prática dos desenvolvedores de modelos do mercado, as variáveis contínuas foram categorizadas seguindo as regras descritas na Seção 2.7.2. As variáveis categorizadas foram: *Idade do Cliente*, *Tempo no Atual Emprego*, *Idade do Veículo Próprio*, *Tempo como Cliente da Instituição*, *Ano de Fabricação do Veículo Financiado*, *Quantidade de Dependentes Menores de Idade*, *Quantidade de Dependentes Maiores de Idade* e *Quantidade de Dependentes*. Desta forma, com todas as variáveis independentes apresentadas na forma de categorias, os valores para identificação dessas categorias são meramente códigos, sem significado numérico. Segundo sugestão de Hosmer e Lemeshow (1989), foram utilizadas variáveis dicotômicas (“dummy”) para representar as categorias das variáveis.

Um exemplo seria o da variável *Quantidade de Avalistas*, com as categorias: *0, 1 e 2 ou mais*. Nesse caso a criação de 2 novas variáveis “dummy” (Var1 e Var2) seria suficiente para representar a variável preditora: Var1 = 1 se o cliente não possui avalistas e Var1 = 0 caso contrário; Var2 = 1 se o cliente possui exatamente 1 avalista e Var2 = 0 caso contrário. Sem a necessidade de criação de uma nova variável “dummy”, pode-se caracterizar o grupo de clientes com 2 avalistas ou mais da seguinte forma: Var1 = 0 e Var2 = 0. Essa última combinação é denominada referência. A Tabela 4.1 mostra a relação entre a variável preditora e as “dummy” criadas.

**Tabela 4.1** Exemplo de Criação de Variáveis “Dummy” para Representação de Variáveis Categorizadas.

<i>Quantidade de Avalistas</i>	<i>Variáveis "Dummy"</i>	
	<i>Var1</i>	<i>Var2</i>
<i>Nenhum</i>	1	0
<i>Um</i>	0	1
<i>Dois ou mais</i>	0	0

Para a Regressão Logística, todas as variáveis preditoras estão representadas pelas respectivas variáveis “dummy”. Na categorização das variáveis, os valores faltantes formam sempre uma categoria à parte.

#### **4.1.2 Variáveis Preditoras no CHAID**

Devido à restrição para aplicação da técnica, na qual todas as variáveis devem estar dispostas em categorias, as variáveis quantitativas foram categorizadas. Esse procedimento é o mesmo realizado para a Regressão Logística, diferenciando-se apenas pelo fato de que para a aplicação do CHAID as variáveis não foram dicotomizadas.

#### **4.1.3 Variáveis Preditoras no REAL**

Ao contrário do CHAID, para a aplicação do REAL há necessidade de que as variáveis apresentem no mínimo uma ordenação. Isto se deve ao fato de o algoritmo agrupar categorias ou valores vizinhos. No estudo, a maior parte das variáveis é originalmente categorizada e a maioria dessas é nominal. A alternativa encontrada para a utilização dessas variáveis foi a recodificação, baseada na ordenação fornecida pelos WOE (ver Seção 2.7.2). Com esse procedimento, tem-se a ordenação das categorias com relação à variável resposta. As demais variáveis assumem seus valores originais, havendo uma recodificação apenas para os valores faltantes, os quais assumem um valor negativo, com o objetivo de serem tratados como uma categoria de mais alto risco de inadimplência.

### **4.2 Aplicação das Técnicas**

As três técnicas foram aplicadas aos conjuntos de desenvolvimento descritos na Seção 2.6 e, em seguida, verificou-se a acurácia de cada técnica sobre os conjuntos de validação. Para a aplicação da Regressão Logística e do CHAID, a ponderação utilizada na

execução dos procedimentos foi: peso para cada cliente ruim igual 18 ( $\geq 22.985/1.287$ ) e peso para cada cliente bom igual a 1. Essa ponderação visa que a amostra represente uma proporção semelhante de bons e ruins e evita o privilégio de classificação correta para os bons como citado na Seção 2.6. No caso do REAL, não há como indicar uma variável de ponderação. Na Seção 4.2.3 estão descritos os procedimentos utilizados para tentar efetuar a ponderação desejada, apesar da restrição do algoritmo.

#### 4.2.1 Aplicação da Regressão Logística Múltipla

Foi gerada uma Análise de Regressão Logística considerando as 17 variáveis preditoras convertidas nas variáveis “dummy” equivalentes, no Pacote Estatístico SPSS. O método de seleção de variáveis utilizado foi o “Forward:LR” que baseia-se no Teste da Razão de Verossimilhanças (ver McCullagh e Nelder, 1989), com as probabilidades de entrada e saída, iguais a 0,01. Esses níveis são inferiores aos tradicionais (0,05 e 0,10) pelo fato de a amostra ser muito grande, podendo considerar significantes variáveis que são pouco importantes. O critério de seleção de variáveis também foi baseado na experiência de crédito e, sendo assim, variáveis consideradas significantes pelos testes estatísticos, porém, que apresentaram resultados sem interpretação lógica quanto a crédito, foram excluídas.

Gerando o modelo para todos os conjuntos de desenvolvimento as variáveis resultantes em todos eles foram: *Estado Civil, Tipo de Residência, Indicador de Apontamento Cadastral, Indicador de Telefone Comercial, Tempo como Cliente da Instituição, Grupo de Profissão, Idade do Cliente, Quantidade de Dependentes Menores de Idade e Indicador de Contratação de Crediário nos Últimos 2 anos.*

Para verificar o grau de acerto para cada um dos modelos gerados foram utilizadas as três medidas descritas na Seção 3.4: TAT, TAR e TAB. Considerando o ponto de corte como sendo a probabilidade 0,5, os valores das medidas para cada conjunto de desenvolvimento e, consequentemente, cada conjunto de validação, estão dispostos na Tabela 4.2.

**Tabela 4.2** Medidas de Classificação Correta para a Regressão Logística.

Conjunto	TAT	TAR	TAB
1	69,4%	68,7%	69,4%
2	68,9%	70,2%	68,9%
3	70,0%	68,7%	70,1%
4	70,3%	64,1%	70,6%
5	70,7%	65,5%	71,0%
6	70,1%	73,2%	69,9%
7	69,0%	71,4%	68,9%
8	69,9%	67,8%	70,1%
9	70,9%	68,3%	71,1%
10	69,6%	69,8%	69,5%

Analisando-se a Tabela 4.2, percebe-se que as taxas de acerto de classificação variam em torno de 70% quando considera-se um ponto de corte de 0,5. A baixa variabilidade dessas medidas indica que o conjunto de variáveis preditoras, selecionado pelo método da Regressão Logística, apresenta comportamento estável. Essas taxas de classificação são consideradas bastante razoáveis pelo mercado, quando se trata de um modelo de concessão de crédito. Modelos comportamentais que levam em consideração variáveis de utilização dos produtos tendem a apresentar melhores resultados.

Por apresentar a melhor TAR e uma das melhores TATs, o modelo relativo ao conjunto de desenvolvimento 6 foi escolhido como modelo final.

As variáveis selecionadas, as respectivas estimativas dos parâmetros, os erros padrões, as estatísticas de Wald, os graus de liberdade e os níveis descritivos, encontram-se dispostos na Tabela 4.3. As categorias com estimativa igual a zero representam as categorias das variáveis que não são representadas por uma única variável “dummy” (categoria de referência). Os valores foram acrescidos somente para fins de facilidade de visualização da variável por inteiro.

**Tabela 4.3** Estimativas dos Parâmetros, Erro Padrão, Estatística de Wald, Graus de Liberdade e Nível Descritivo, para o Modelo de Regressão Logística Final.

Variável	Coeficiente	Erro-Padrão	Wald	GL	Nível Descritivo
<b>Estado Civil</b>					
Missing	0,000	-	176,83	5	0,000
Casado	1,889	0,272	48,27	1	0,000
Solteiro	1,657	0,273	36,94	1	0,000
Desq./Div.	1,464	0,274	28,47	1	0,000
Viúvo	1,688	0,286	34,78	1	0,000
Marital	1,797	0,278	41,83	1	0,000
<b>Dep. Menores</b>			217,40	4	0,000
Missing	0,000	-	-	-	-
Nenhum	-2,673	0,286	87,13	1	0,000
1	-2,675	0,289	85,85	1	0,000
2	-2,645	0,289	83,81	1	0,000
3 ou mais	-3,186	0,291	119,89	1	0,000
<b>Tipo Residência</b>			495,18	5	0,000
Missing	0,000	-	-	-	-
Própria	0,655	0,113	33,46	1	0,000
Financiada	0,691	0,122	32,19	1	0,000
Alugada	0,014	0,115	0,02	1	0,901
Com os Pais	0,702	0,118	35,26	1	0,000
Outro	0,332	0,124	7,19	1	0,007
<b>Fone Com. (sim)</b>	0,651	0,035	355,69	1	0,000
<b>Profissão</b>			506,51	6	0,000
Missing	0,000	-	-	-	-
Prof. Liberal	-0,568	0,106	28,55	1	0,000
Adm/Serv/Ind	-0,259	0,108	5,77	1	0,016
Comércio	-0,857	0,109	62,45	1	0,000
Proprietários	-0,959	0,105	83,80	1	0,000
Aposentados	-0,314	0,127	6,10	1	0,014
Outros	-0,483	0,104	21,47	1	0,000
<b>Apto. Cadastral (sim)</b>	-1,361	0,023	3565,27	1	0,000
<b>Ind. Crediário (sim)</b>	-0,389	0,030	167,97	1	0,000
<b>Idade</b>			1001,55	5	0,000
Missing	0,000	-	-	-	-
Até 23 anos	3,683	0,135	747,21	1	0,000
de 24 a 33 anos	3,880	0,125	964,88	1	0,000
de 34 a 43 anos	3,637	0,124	861,35	1	0,000
de 44 a 53 anos	3,695	0,124	893,29	1	0,000
54 anos ou mais	3,765	0,126	896,79	1	0,000
<b>Tempo como Cliente</b>			1421,30	4	0,000
Missing	0,000	-	-	-	-
até 12 meses	-1,279	0,035	1309,47	1	0,000
de 13 a 36 meses	-0,562	0,036	249,43	1	0,000
de 37 a 60 meses	-0,340	0,037	85,57	1	0,000
de 61 a 96 meses	-0,307	0,036	70,83	1	0,000
(*) 97 anos ou mais	0,000	-	-	-	-
<b>Constante</b>	-1,972	0,078	637,52	1	0,000

(\*) categoria desconsiderada na construção da matriz X.

Um caso particular que ocorre neste estudo deve-se ao fato de que quando não há informação da idade do cliente, também não há especificação do tempo de conta do cliente com a instituição. Desta forma, ao se trabalhar com as variáveis dicotomizadas, verifica-se que uma das categorias da variável *Tempo como Cliente da Instituição* (Tradição do cliente) é combinação linear das demais categorias dessa variável e das categorias da variável *Idade do Cliente*. O SPSS desconsiderou uma das categorias da variável Tradição do Cliente ao montar a matriz X. A interpretação dos resultados pode ser feita da seguinte maneira: considerando todas as demais variáveis do modelo fixas, se o cliente não apresentar informação sobre sua idade (consequentemente, também não apresenta informação sobre tempo como cliente) não terá nenhum acréscimo em seu escore final, porém, para qualquer valor válido que apresente para a variável idade, terá um acréscimo positivo em seu escore (coeficiente de idade + coeficiente de tradição > 0, sempre que a idade não seja “missing”).

A interpretação da constante nesse modelo é a seguinte: se o cliente não possui informação sobre as variáveis *Estado Civil*, *Tipo de Residência*, *Tradição do Cliente*, *Grupo de Profissão*, *Idade do Cliente*, *Quantidade de Dependentes Menores de Idade* e, além disso, não possuir *Telefone Comercial*, *Apontamentos Cadastrais* e *não teve nenhum contrato de Crediário nos últimos 2 anos*, então, a chance desse cliente não se tornar inadimplente está relacionada ao valor da constante, ou seja:

$$P(\text{Bom}) = \{\exp(-1,972)/[1+\exp(-1,972)]\} = 0,143.$$

Analisando os coeficientes de cada uma das variáveis sem considerar a variação das demais pode-se destacar:

- *Estado Civil*: clientes casados (1,889) e em estado marital (1,797) destacam-se positivamente, enquanto os clientes desquitados ou divorciados (1,464) destacam-se negativamente quando comparados com os demais;
- *Quantidade de Dependentes Menores de Idade*: aparentemente, a grande diferença está entre o fato de o cliente possuir até 2 dependentes (não há grandes diferenças entre 0, 1 ou 2 dependentes com estimativas, -2,673, -2,675, -2,645, respectivamente) ou possuir 3 ou mais (-3,186). Destaca-se, também, que para esta variável a informação faltante parece não indicar maior risco de inadimplência do cliente;

- *Tipo de Residência*: o grande destaque para essa variável é a indicação de que clientes que habitam uma residência alugada (0,014) apresentam um risco de inadimplência maior que os demais;
- *Indicador de Telefone Comercial*: como esperado, clientes que possuem telefone comercial (0,651), apresentam risco menor do que aqueles que não possuem;
- *Grupo de Profissão*: os grupos de profissão que indicam maiores riscos são o de proprietários de seus negócios (-0,959) e o de trabalhadores do comércio (-0,857), enquanto os de menor risco concentram-se nos ramos administrativo, serviços e industrial (-0,259). Neste caso, os valores faltantes também não indicam um risco maior de inadimplência;
- *Indicador de Apontamento Cadastral*: clientes com algum histórico de inadimplência parecem apresentar maior risco de inadimplência futura (-1,361);
- *Indicador de Contratação de Crediário nos Últimos 2 anos*: há indícios de que clientes que já contrataram um produto de crédito nos últimos 2 anos apresentam um risco de inadimplência maior do que aqueles que não contrataram (-0,389);
- *Idade do Cliente*: o comportamento esperado era de que quanto maior fosse a idade menor seria o risco de inadimplência do cliente. Porém, na presença das demais variáveis o grupo de 24 a 33 anos destaca-se positivamente (3,880);
- *Tradição como Cliente*: percebe-se que quanto maior o tempo de relacionamento do cliente com a instituição menor é a chance dele se tornar inadimplente.

Aplicando-se o modelo final à amostra total de 24.272 clientes, chega-se à tabela de classificação disposta na Tabela 4.4.

**Tabela 4.4** Tabela de Classificação para o Modelo Final de Regressão Logística.

<i>Observado</i>	<i>Estimado</i>		
	<i>Ruim</i>	<i>Bom</i>	<i>Total</i>
<i>Ruim</i>	891	396	1.287
<i>Bom</i>	6.799	16.186	22.985
<i>Total</i>	7.690	16.582	24.272

As taxas de acerto, quando aplica-se o modelo para todos os clientes, são: TAT = 70,4%; TAR = 69,2% e TAB = 70,4%.

O SPSS possui um procedimento que permite ao usuário escolher qual das categorias da variável preditora deve ser a referência na construção das variáveis “dummy”. Outra opção importante refere-se ao método de seleção de variáveis, pois, quando há presença de variáveis “dummy”, se uma parte da variável é escolhida para fazer parte do modelo, automaticamente, todas as demais “dummy” relativas a essa variável preditora também são escolhidas. Esses procedimentos facilitam a estimação e a interpretação dos parâmetros estimados no modelo final.

#### **4.2.2 Aplicação do CHAID**

Para a aplicação do CHAID foram consideradas as mesmas 17 variáveis categorizadas utilizadas na Regressão Logística. As variáveis foram definidas inicialmente como livres, monótonas e flutuantes, conforme definição da Seção 3.2.2, da seguinte forma:

- Variáveis livres: *Estado Civil, Tipo de Residência, Grupo de Profissão;*
- Variáveis monótonas: *Indicador de Telefone Residencial, Indicador de Telefone Comercial, Indicador de Posse de Cartão de Crédito, Indicador de Apontamento Cadastral, Indicador de Crediário nos Últimos 2 Anos, Quantidade de Avalistas;*
- Variáveis flutuantes: *Idade do Cliente, Quantidade de Dependentes Menores, Quantidade de Dependentes Maiores, Quantidade Total de Dependentes, Tempo no Atual Emprego, Idade do Veículo Próprio, Tempo como Cliente da Instituição, Ano do Veículo Financiado.*

Os níveis de agrupamento e de significância foram especificados como  $p=0,01$ . Com o objetivo de impedir que a árvore resultante fosse composta de muitas camadas e ramos (com baixa freqüência), foi estabelecida uma freqüência mínima em cada ramo, equivalente a 2,5% do tamanho da amostra de desenvolvimento. Esse procedimento faz

com que as árvores encontradas sejam mais estáveis. A estatística de Qui-Quadrado escolhida foi a da Razão de Verossimilhanças (a outra opção dada pelo software é o Qui-Quadrado de Pearson). Para o algoritmo WLM (ver Seção 3.2.3) os valores adotados para  $\varepsilon$  e  $maxit$  foram 0 e 100, respectivamente.

Na Tabela 4.5 estão apresentados os resultados da aplicação do CHAID para cada um dos conjuntos de desenvolvimento com respectiva aplicação no conjunto de validação. Diferentemente da Regressão Logística, onde para cada cliente estima-se a probabilidade dele ser bom, no CHAID, verifica-se a proporção de clientes bons dentro de cada nó e classifica-se o nó como bom se essa proporção for maior que um determinado ponto de corte, que no caso é 50%. Caso contrário, o nó é classificado como ruim.

**Tabela 4.5** Medidas de Classificação Correta para o CHAID.

Conjunto	TAT	TAR	TAB
1	68,2%	70,4%	66,1%
2	68,4%	71,0%	66,0%
3	68,8%	73,0%	64,9%
4	66,9%	67,2%	66,7%
5	70,9%	75,4%	66,0%
6	74,2%	80,5%	66,6%
7	69,8%	73,8%	65,8%
8	69,4%	72,0%	66,4%
9	69,9%	72,2%	67,6%
10	67,4%	69,0%	65,9%

Percebe-se, pela tabela, que a TAT gira em torno de 69%, a TAR apresenta valores acima de 70% e a TAB sempre apresenta valores abaixo de 68%.

Quanto às variáveis preditoras, nem sempre as mesmas variáveis compõem as árvores de classificação, quando mudam-se os conjuntos de desenvolvimento. A Tabela 4.6 mostra um resumo de como as variáveis preditoras são utilizadas nas árvores geradas pelos 10 conjuntos de desenvolvimento.

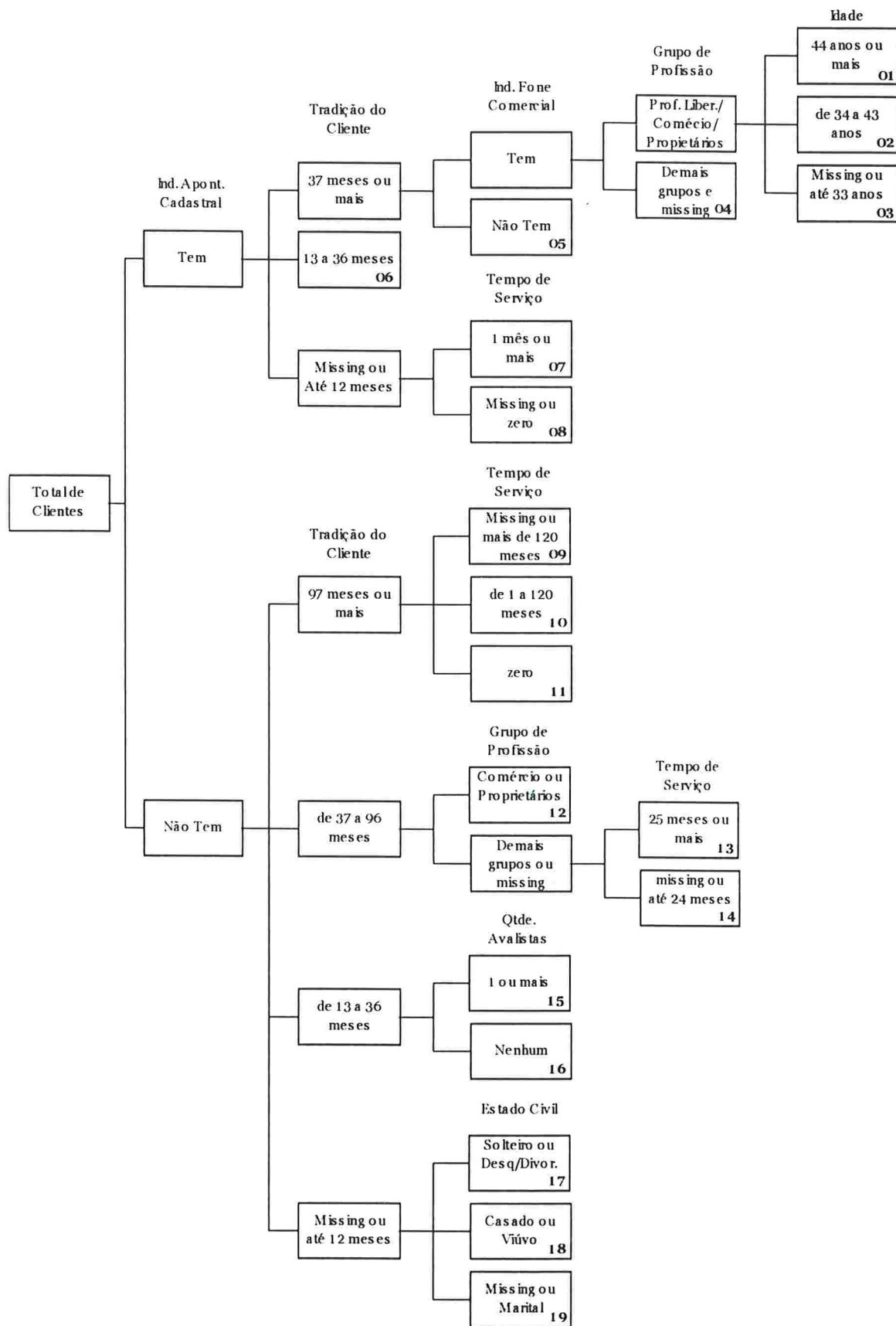
**Tabela 4.6** Freqüência de Utilização das Variáveis nas Gerações de Arvores CHAID.

Variável	Num. vezes utilizada
Ind. Apontamento Cadastral	10
Tempo como Cliente	10
Grupo de Profissão	10
Quantidade de Avalistas	9
Tempo no Emprego Atual	9
Estado Civil	8
Indicador de Crediário	5
Idade do Cliente	5
Ano do Veículo Financiado	4
Qtde. de Dependentes	2
Ind. Telefone Comercial	2
Tipo de Residência	1
Idade do Veículo Próprio	1

As únicas três variáveis que estão presentes em todas as árvores geradas são: *Indicador de Apontamento Cadastral*, *Tempo como Cliente da Instituição* e *Grupo de Profissão*.

Pela Tabela 4.5, pode-se notar que o conjunto de desenvolvimento 6 apresenta os melhores resultados de classificação e, portanto, a árvore por ele gerada será utilizada como a árvore final de classificação. A Figura 4.1 representa a árvore final de classificação. A numeração de 01 a 19, disposta na árvore, é utilizada para identificação dos nós.

**Figura 4.1** Árvore de Classificação Final do Algoritmo CHAID.



A proporção observada de clientes bons em cada nó é dada na Tabela 4.7. Pela tabela, nota-se que o nó com menor proporção de clientes bons é o de número 08 (11,0%) representado pelo conjunto de clientes com as seguintes características: possui algum tipo de apontamento cadastral, não possui informação de tempo como cliente ou é cliente da instituição há menos de um ano e está desempregado ou não informou o tempo no atual emprego. Por outro lado, o nó 09 é o que apresenta maior proporção de clientes bons (90,9%). Esse nó corresponde aos clientes que não possuem apontamento cadastral, possuem 97 meses ou mais de tempo de conta e estão empregados há mais de 12 meses, ou não informaram o tempo na profissão atual.

**Tabela 4.7** Proporção Observada de Clientes Bons para Cada Nó da Árvore.

Nó	% Bons
01	41,7%
02	32,3%
03	40,9%
04	48,9%
05	30,4%
06	34,4%
07	24,8%
08	11,0%
09	90,9%
10	73,9%
11	78,4%
12	56,3%
13	82,3%
14	72,0%
15	59,5%
16	67,0%
17	43,1%
18	54,8%
19	14,3%

Aplicando-se a árvore final à amostra total de 24.272 clientes, chega-se à tabela de classificação disposta na Tabela 4.8.

**Tabela 4.8** Tabela de Classificação para a Árvore de Classificação do CHAID.

<i>Observado</i>	<i>Estimado</i>		
	<i>Ruim</i>	<i>Bom</i>	<i>Total</i>
<i>Ruim</i>	935	352	1.287
<i>Bom</i>	7.758	15.227	22.985
<i>Total</i>	8.693	15.579	24.272

As taxas de acerto, quando aplica-se o modelo para todos os clientes, são: TAT = 66,6%; TAR = 72,6% e TAB = 66,2%.

#### 4.2.3 Aplicação do REAL

Para a aplicação do REAL, as 17 variáveis do estudo apresentam a ordenação, requerida, seguindo os procedimentos descritos na Seção 4.1.3. Para a obtenção dos resultados foi utilizado o software desenvolvido e cedido pelos autores.

Assim como nas aplicações das duas técnicas anteriores, trabalhou-se com os 10 conjuntos de desenvolvimento e respectivos conjuntos de validação. Por não haver no “software” nenhuma forma de definir pesos para os clientes, tentou-se, inicialmente, fazer com que para cada conjunto de desenvolvimento e validação a quantidade de clientes ruins fosse equivalente à quantidade de bons. Para isso, todas as informações de cada cliente ruim foram replicadas 17 vezes, obtendo-se um total de 18 réplicas de cada cliente ruim. Na Regressão Logística e no CHAID esse procedimento é equivalente ao de se definir um peso 18 para cada cliente ruim e um peso 1 para cada cliente bom.

Inicialmente, foram utilizados os pares de parâmetros de entrada ( $r, v_c$ ), de convexidade e de convicção, respectivamente, que levaram aos melhores resultados no trabalho de Lauretto (1996). Aplicando-se o algoritmo aos conjuntos de desenvolvimento (contendo as réplicas) e verificando-se os resultados no respectivo conjunto de validação, para nenhum par de parâmetros de entrada a TAR foi maior do que 20%, ao passo que a TAB sempre encontrava-se acima de 80%. Tais resultados mostram indícios de que o

procedimento de réplica dos clientes ruins, objetivando a ponderação, não faz com que o modelo final seja eficiente na classificação dos clientes ruins. Os resultados são muito semelhantes aos obtidos sem a utilização de réplicas, ou seja, para a amostra contendo 1.287 clientes bons e 22.985 clientes ruins.

Como o principal objetivo é classificar de forma satisfatória os clientes ruins (como discutido na Seção 2.6), foi proposta uma outra alternativa de amostragem. Construiu-se uma amostra contendo os 1.287 clientes ruins e 1.287 clientes bons selecionados aleatoriamente do total de 22.985. A partir dessa amostra de 2.574, repetiu-se o procedimento para obtenção de 10 conjuntos de desenvolvimento e respectivos conjuntos de validação. Nessa opção de amostragem, há perda de informação sobre as características dos clientes bons, pois, trabalha-se com apenas uma amostra desse grupo de clientes, porém, em contrapartida, são obtidas árvores com melhor poder de classificação dos clientes ruins.

Mantendo-se os mesmos conjuntos de desenvolvimento e validação, o algoritmo foi testado para os seguintes pares de parâmetros:  $(r, v_c) \in \{2; 2,5; 3; 3,5; 4\} \times \{0,3; 0,35; 0,4\}$ . Para cada par, gerou-se uma árvore sobre o conjunto de desenvolvimento e verificou-se o resultado das medidas TAT, TAR e TAB no respectivo conjunto de validação. Os resultados desse experimento podem ser consultados no Apêndice B (tabelas B1 a B15). Nessas tabelas, também estão dispostas a média de acerto para cada medida e o respectivo desvio padrão. Essas medidas podem ser utilizadas para verificar a performance e a estabilidade das árvores geradas. Deve-se ressaltar que as tabelas que consideram o parâmetro de convicção  $v_c = 0,3$  (Tabelas B1, B4, B7, B10 e B13) são idênticas, pois, todas as árvores geradas apresentam uma única camada com 2 nós, relativos à variável *Indicador de Apontamento Cadastral do Cliente*.

Ao se comparar os resultados das tabelas do Apêndice B, verifica-se que o par de parâmetros de entrada  $(r, v_c)$  que apresenta os melhores resultados é o par  $(2; 0,4)$ . Os resultados da aplicação do REAL com esse par de parâmetros de entrada nos conjuntos amostrais estão dispostos na Tabela 4.9.

**Tabela 4.9** Medidas de Classificação Correta para o REAL com Parâmetros (2 e 0,4).

Conjunto	TAT	TAR	TAB
1	74,3%	75,0%	73,6%
2	76,7%	73,5%	80,2%
3	75,9%	72,9%	78,4%
4	77,8%	80,2%	75,7%
5	71,6%	69,6%	73,8%
6	73,9%	73,0%	74,8%
7	77,0%	80,5%	73,9%
8	74,7%	76,7%	72,6%
9	75,1%	80,1%	69,4%
10	73,2%	74,0%	72,4%

Uma vez que a árvore gerada pelo conjunto de treinamento 7 é a que apresenta maior TAR e uma das maiores TAT, ela foi escolhida como a árvore final de classificação.

Grande parte das árvores geradas pelo REAL para os parâmetros especificados acima apresentam muitas camadas e uma extensa quantidade de nós. Desta forma, torna-se inviável a interpretação dos nós, como ocorrida no CHAID.

A árvore final de classificação possui 11 camadas e das 17 variáveis preditoras, somente 4 não foram utilizadas: *Indicador de Telefone Residencial*, *Indicador de Telefone Comercial*, *Indicador de Posse de Cartão de Crédito* e *Quantidade de Avalistas*.

Aplicando a árvore final de classificação à amostra total de 24.272 clientes, chega-se à tabela de classificação representada pela Tabela 4.10.

**Tabela 4.10** Tabela de Classificação para a Árvore de Classificação do REAL.

Observado	Estimado		
	Ruim	Bom	Total
<i>Ruim</i>	944	343	1.287
<i>Bom</i>	6.604	16.381	22.985
<i>Total</i>	7.548	16.724	24.272

As taxas de acerto, quando aplica-se o modelo para todos os clientes, são: TAT = 71,4%; TAR = 73,3% e TAB = 71,3%.

### 4.3 Comparações

As comparações entre as técnicas têm como finalidade identificar qual dos métodos apresenta melhores taxas de classificação correta e observar a concordância dos métodos quanto à classificação dos clientes. A verificação do método com maior poder de classificação será feita através da comparação das 3 medidas de acerto (TAT, TAR e TAB). Para analisar a concordância entre os métodos será usada a tabela de combinações de previsões, abordada por Arminger et. al. (1997).

Na Tabela 4.11 estão dispostas as medidas TAT, TAR e TAB, calculadas sobre as Tabelas 4.4, 4.8 e 4.10, representando as técnicas da Regressão Logística, CHAID e REAL, respectivamente.

**Tabela 4.11** Medidas de Classificação Correta para os Três Modelos.

Modelo	TAT	TAR	TAB
Regressão Logística	70,4%	69,2%	70,4%
CHAID	66,6%	72,6%	66,2%
REAL	71,4%	73,3%	71,3%

Comparando-se a TAT para os três modelos, nota-se uma leve vantagem do REAL (71,4%) sobre a Regressão Logística (70,4%), sendo que ambos apresentam resultados melhores do que o CHAID (66,6%). Ao se comparar a TAB os resultados são muito parecidos aos da comparação da TAT. Analisando a medida mais importante de acerto, do ponto de vista de concessão de crédito, nota-se que novamente o REAL apresenta o melhor desempenho (73,3%), seguido do CHAID (72,6%) e o pior desempenho é o da Regressão Logística (69,2%). Apesar das medidas apresentarem valores muito próximos, o REAL destaca-se frente às outras duas técnicas, no que diz respeito ao acerto na classificação dos clientes.

Deve-se ressaltar que as medidas de acerto para cada modelo estão baseadas em pontos de corte onde clientes com nota igual ou superior a esse valor são classificados como bons e, caso contrário, como ruins. Talvez para outros valores desses pontos de corte

essas relações entre as taxas de acerto não se repitam. De qualquer forma, os pontos de corte escolhidos são os utilizados usualmente pelo mercado.

Objetivando-se medir a concordância de classificação dos três modelos, foi construída a Tabela 4.12, que apresenta nas suas linhas as combinações das classificações previstas por cada técnica (esperada) e em suas colunas a verdadeira classificação dos clientes (observada). Nas células da tabela estão dispostas as freqüências dos clientes que se encaixam em cada combinação.

**Tabela 4.12** Tabela de Concordâncias de Classificação dos Clientes.

Combinação (esperado)			Observado		
Logística	CHAID	REAL	R	B	Total
R	R	R	810	5.042	5.852
R	R	B	34	1.062	1.096
R	B	R	22	163	185
R	B	B	35	756	791
B	R	R	83	1.036	1.119
B	R	B	8	618	626
B	B	R	29	363	392
B	B	B	266	13.945	14.211
Total			1.287	22.985	24.272

Pela Tabela 4.12, pode-se notar que há concordância entre as três técnicas, combinações (R, R, R) e (B, B, B), para 20.063 clientes ( $5.852+14.211$ ), que representam 82,7% do total de clientes. Analisando-se os demais 17,3% dos casos, pode-se verificar que as técnicas que mais apresentam concordâncias são o CHAID e o REAL com 1.910 casos ( $791+1.119$ ), representando 7,9% do total de clientes, ao passo que as duas que apresentam menor concordância são a Regressão Logística e o REAL, com 811 casos ( $185+626$ ), representando 3,3% do total de clientes.

Para uma análise mais apurada quanto à concordância de classificação das técnicas, associada aos acertos na classificação, foi criada a Tabela 4.13, produzida a partir da Tabela 4.12. Essa tabela apresenta nas colunas duas medidas: a primeira é a razão entre clientes

bons e ruins para cada combinação de classificação das técnicas e a segunda nada mais é que o percentual de clientes para cada combinação.

**Tabela 4.13** Tabela da Razão de Clientes Bons e Ruins e Percentual de Clientes para as Combinações de Classificação dos Três Modelos.

Logística	Combinação (esperado)			Observado	
	CHAID	REAL	B/R	%Clientes	
R	R	R	6,2	24,1%	
R	R	B	31,2	4,5%	
R	B	R	7,4	0,8%	
R	B	B	21,6	3,3%	
B	R	R	12,5	4,6%	
B	R	B	77,3	2,6%	
B	B	R	12,5	1,6%	
B	B	B	52,4	58,5%	
Total			17,9	100,0%	

Pela Tabela 4.13, percebe-se de forma mais simplificada em quais situações as concordâncias entre as técnicas levam a resultados de classificação com melhor performance. A medida base de comparação da tabela é a razão entre o total de clientes bons e ruins observada na amostra (17,9). Valores de B/R abaixo de 17,9 indicam, de maneira proporcional, uma presença maior de clientes ruins do que de clientes bons. Valores acima de 17,9 indicam interpretação contrária.

Verificando as combinações de classificação concordantes para as três técnicas, nota-se que ocorre o esperado, pois, para a combinação (R, R, R) o valor de B/R é 6,2, enquanto para (B, B, B) o valor da razão é 52,4. Analisando-se as situações onde há concordância entre duas técnicas que discordam de uma terceira, pode-se destacar a combinação (R, R, B), onde tanto a Regressão Logística quanto o CHAID, classificam os clientes como ruins, ao passo que o REAL classifica como bons. Nesse caso, mesmo sendo minoria, o REAL acerta mais, uma vez que o valor da razão B/R é 31,2. Na combinação inversa (B, B, R) o mesmo ocorre, pois, o valor de B/R é 12,5 e das três técnicas, somente o REAL classifica tais clientes como ruins. Nas demais combinações, a concordância entre

duas técnicas apresenta os melhores resultados. Novamente, os resultados mostram que o algoritmo REAL resulta em uma performance superior às outras duas técnicas.

Algumas alterações nos procedimentos de aplicação das três técnicas poderiam ser elaboradas para melhorar a performance de cada uma. Na Regressão Logística, a utilização de cada variável contínua na sua forma natural poderia aumentar o poder de classificação do modelo. Para o CHAID, a redução da freqüência mínima para cada nó da árvore poderia fazer com que a árvore apresentasse uma maior expansão, gerando melhores resultados. Na aplicação do REAL, um ajuste no “software”, permitindo que seja especificada uma variável de ponderação, permitiria o uso da amostra inteira na construção das árvores, o que poderia resultar em melhores desempenhos. Considerando a Regressão Logística e o CHAID, essas alterações afetariam diretamente a interpretação dos resultados e poderiam tornar as técnicas não atraentes do ponto de vista de aplicação no mercado de crédito.

## CAPÍTULO 5

### Conclusão

Neste trabalho foi descrito o problema de concessão de crédito a clientes de instituições financeiras que desejam contratar um produto de financiamento de compra de veículos. Foram apresentadas as características do produto e as variáveis disponíveis para o estudo, assim como métodos de tratamento dessas variáveis.

Três técnicas para classificação dos clientes foram consideradas, sendo que duas já são bastante utilizadas no mercado pelos profissionais da área (Regressão Logística e CHAID) e a outra é uma técnica mais recente que ainda não havia sido aplicada com esse objetivo (REAL).

Na aplicação descrita, a técnica da Regressão Logística apresentou as seguintes vantagens: flexibilidade de poder ser executada com qualquer tipo de variável; o resultado da análise ser uma probabilidade, o que permite a ordenação dos clientes quanto ao risco de inadimplência e a disponibilidade de execução nos principais pacotes estatísticos disponíveis no mercado. Como principal desvantagem pode-se citar o fato da interpretação dos parâmetros do modelo não ser trivial, como no caso da Regressão Linear.

O CHAID, como ferramenta de árvore de decisão, destaca-se por apresentar diversos mecanismos de interferência na árvore, que facilitam a manipulação pelo usuário. Tem como principal vantagem a interpretação do modelo, onde torna-se fácil para o usuário final compreender o método de classificação do algoritmo. Como desvantagens, pode-se ressaltar a restrição de trabalhar somente com variáveis categorizadas e atribuir somente aos nós (e não aos clientes) as chances de inadimplência.

O REAL por ser um algoritmo novo e na sua primeira versão, ainda impõe algumas restrições, tais como, não permitir que o usuário defina tamanhos mínimos para os nós terminais da árvore. Uma grande vantagem do algoritmo frente aos seus concorrentes é que

ele permite o uso de variáveis preditoras contínuas e utiliza procedimentos de discretização e agrupamento mais sofisticados. A principal desvantagem é a dificuldade da interpretação dos nós da árvore gerada, uma vez que as árvores finais apresentam muitas camadas e uma quantidade excessiva de nós.

Nas comparações dos modelos, os melhores resultados obtidos foram através do REAL. Com relação à classificação dos clientes, as três técnicas apresentaram um alto índice de concordância, enquanto que, na análise duas a duas, as técnicas de árvore de decisão foram mais concordantes.

No desenvolvimento de modelos de “Credit Scoring”, deve-se considerar os objetivos finais de suas aplicações. No contexto dos métodos abordados neste trabalho, se o objetivo for classificar os clientes da forma mais precisa, não importando a interpretação da regra de classificação, sugere-se o uso do REAL. Caso haja interesse na compreensão da ferramenta de classificação, ainda assim, atingindo-se bons resultados, sugere-se a utilização da Regressão Logística. Por último, caso o objetivo principal seja a simplicidade da regra de classificação, sem abdicar de resultados satisfatórios, sugere-se a aplicação do CHAID.

Como sugestão para trabalhos futuros, outros métodos como os de Redes Neurais poderiam ser comparados aos procedimentos usuais. Além disso, a Análise Discriminante, que é uma técnica muito difundida no mercado, mereceria uma análise detalhada referente aos problemas do seu uso, já que as suposições requeridas quase nunca são satisfeitas nesse tipo de estudo.

## APÊNDICE A

### Tabelas de “Weights of Evidence”

Após a categorização das variáveis de natureza quantitativa, o conjunto final das variáveis preditoras se mostrou da seguinte forma:

**Tabela A1** Tabela de “Weights of Evidence” para a Variável *Idade do Cliente*

<i>Idade</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>%B/%R</i>	<i>WOE</i>
Missing	117	208	325	1,3%	0,10	-2,307
até 23 anos	70	1041	1111	4,6%	0,83	-0,183
de 24 a 33 anos	349	6570	6919	28,5%	1,05	0,053
de 34 a 43 anos	426	7528	7954	32,8%	0,99	-0,011
de 44 a 53 anos	216	4855	5071	20,9%	1,26	0,230
54 anos ou mais	109	2783	2892	11,9%	1,43	0,357
Total	1287	22985	24272	100,0%	1,00	0,000

**Tabela A2** Tabela de “Weights of Evidence” para a Variável *Estado Civil*

<i>Estado Civil</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>%B/%R</i>	<i>WOE</i>
Missing	127	870	997	4,1%	0,38	-0,958
Casado	689	15041	15730	64,8%	1,22	0,201
Solteiro	297	4587	4884	20,1%	0,86	-0,145
Desq ou Div	116	1504	1620	6,7%	0,73	-0,320
Viuvo	20	388	408	1,7%	1,09	0,083
Marital	38	595	633	2,6%	0,88	-0,132
Total	1287	22985	24272	100,0%	1,00	0,000

**Tabela A3** Tabela de “Weights of Evidence” para a Variável  
*Quantidade de Dependentes Menores de Idade*

<i>Qtde. Dep. Menores</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>%B/%R</i>	<i>WOE</i>
Missing	123	852	975	4,0%	0,39	-0,947
0	693	13341	14034	57,8%	1,08	0,075
1	176	3673	3849	15,9%	1,17	0,156
2	189	3857	4046	16,7%	1,14	0,133
3 ou mais	106	1262	1368	5,6%	0,67	-0,406
Total	1287	22985	24272	100,0%	1,00	0,000

**Tabela A4** Tabela de “Weights of Evidence” para a Variável  
*Quantidade de Dependentes Maiores de Idade*

<i>Qtde. Dep. Maiores</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>%B/%R</i>	<i>WOE</i>
Missing	123	852	975	4,0%	0,39	-0,947
0	912	16075	16987	70,0%	0,99	-0,013
1	180	3999	4179	17,2%	1,24	0,218
2 ou mais	72	2059	2131	8,8%	1,60	0,471
Total	1287	22985	24272	100,0%	1,00	0,000

**Tabela A5** Tabela de “Weights of Evidence” para a Variável *Quantidade de Dependentes*

<i>Quantidade de Dependentes</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>%B/%R</i>	<i>WOE</i>
Missing	123	852	975	4,0%	0,39	-0,947
0	621	11410	12031	49,6%	1,03	0,028
1	141	2670	2811	11,6%	1,06	0,059
2	176	3901	4077	16,8%	1,24	0,216
3 ou mais	226	4152	4378	18,0%	1,03	0,028
Total	1287	22985	24272	100,0%	1,00	0,000

**Tabela A6** Tabela de “Weights of Evidence” para a Variável *Tipo de Residência*

<i>Tipo de Residência</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>%B/R</i>	<i>WOE</i>
Missing	156	1221	1377	5,7%	0,44	-0,825
Própria	648	14200	14848	61,2%	1,23	0,205
Financiada	66	1438	1504	6,2%	1,22	0,199
Alugada	263	3039	3302	13,6%	0,65	-0,435
Com os Pais	101	2163	2264	9,3%	1,20	0,182
Outros	53	924	977	4,0%	0,98	-0,024
Total	1287	22985	24272	100,0%	1,00	0,000

**Tabela A7** Tabela de “Weights of Evidence” para a Variável *Indicador de Telefone Residencial*

<i>Ind. Fone Residencial</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>%B/R</i>	<i>WOE</i>
Não Tem	153	1156	1309	5,4%	0,42	-0,860
Tem	1134	21829	22963	94,6%	1,08	0,075
Total	1287	22985	24272	100,0%	1,00	0,000

**Tabela A8** Tabela de “Weights of Evidence” para a Variável *Indicador de Telefone Comercial*

<i>Ind. Fone Comercial</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>%B/R</i>	<i>WOE</i>
Não Tem	349	3292	3641	15,0%	0,53	-0,638
Tem	938	19693	20631	85,0%	1,18	0,162
Total	1287	22985	24272	100,0%	1,00	0,000

**Tabela A9** Tabela de “Weights of Evidence” para a Variável *Grupo de Profissão*

<i>Grupo de Profissão</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>B/R</i>	<i>WOE</i>
Missing	155	1267	1422	5,9%	0,46	-0,782
Prof. Liberais	206	4540	4746	19,6%	1,23	0,210
Adm/Serv/Ind	109	3303	3412	14,1%	1,70	0,529
Comércio	130	1901	2031	8,4%	0,82	-0,200
Proprietários	407	6172	6579	27,1%	0,85	-0,164
Aposentados	21	657	678	2,8%	1,75	0,561
Outros	259	5145	5404	22,3%	1,11	0,106
Total	1287	22985	24272	100,0%	1,00	0,000

**Tabela A10** Tabela de “Weights of Evidence” para a Variável *Tempo no Atual Emprego*

<i>Tempo no emprego</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>%B/%R</i>	<i>WOE</i>
Missing	120	784	904	3,7%	0,37	-1,006
0	390	6224	6614	27,2%	0,89	-0,113
de 1 a 12 meses	80	1320	1400	5,8%	0,92	-0,079
de 13 a 24 meses	92	1489	1581	6,5%	0,91	-0,098
de 25 a 48 meses	151	2616	2767	11,4%	0,97	-0,030
de 49 a 120 meses	276	5761	6037	24,9%	1,17	0,156
121 meses ou mais	178	4791	4969	20,5%	1,51	0,410
Total	1287	22985	24272	100,0%	1,00	0,000

**Tabela A11** Tabela de “Weights of Evidence” para a Variável *Indicador de Cartão de Crédito*

<i>Ind. Cartão de Crédito</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>%B/%R</i>	<i>WOE</i>
Não Tem	422	4865	5287	21,8%	0,65	-0,438
Tem	865	18120	18985	78,2%	1,17	0,160
Total	1287	22985	24272	100,0%	1,00	0,000

**Tabela A12** Tabela de “Weights of Evidence” para a Variável *Idade do Veículo Próprio*

<i>Idade do veículo Próprio</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>%B/%R</i>	<i>WOE</i>
Missing / Não Possui	474	7032	7506	30,9%	0,83	-0,186
0Km	90	2065	2155	8,9%	1,28	0,251
de 1 a 5 anos	401	8201	8602	35,4%	1,15	0,136
mais de 5 anos	322	5687	6009	24,8%	0,99	-0,011
Total	1287	22985	24272	100,0%	1,00	0,000

**Tabela A13** Tabela de “Weights of Evidence” para a Variável *Tempo como Cliente da Instituição*

<i>Tradição do Cliente</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>%B/%R</i>	<i>WOE</i>
Missing	117	208	325	1,3%	0,10	-2,307
até 12 meses	344	3352	3696	15,2%	0,55	-0,606
de 13 a 36 meses	239	4018	4257	17,5%	0,94	-0,060
de 37 a 60 meses	190	3906	4096	16,9%	1,15	0,141
de 61 a 96 meses	171	4062	4233	17,4%	1,33	0,285
97 meses ou mais	226	7439	7665	31,6%	1,84	0,611
Total	1287	22985	24272	100,0%	1,00	0,000

**Tabela A14** Tabela de “Weights of Evidence” para a Variável  
*Idade do Veículo Financiado*

<i>Idade do Veículo Financiado</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>%B/%R</i>	<i>WOE</i>
Missing	0	15	15	0,1%	-	-
6 anos ou mais	381	6672	7053	29,1%	0,98	-0,020
de 4 a 5 anos	363	6347	6710	27,6%	0,98	-0,021
Até 3 anos	543	9951	10494	43,2%	1,03	0,026
Total	1287	22985	24272	100,0%	1,00	0,000

**Tabela A15** Tabela de “Weights of Evidence” para a Variável *Quantidade de Avalistas*

<i>Quantidade de Avalistas</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>%B/%R</i>	<i>WOE</i>
0	756	13602	14358	59,2%	1,01	0,007
1	497	9020	9517	39,2%	1,02	0,016
2 ou mais	34	363	397	1,6%	0,60	-0,514
Total	1287	22985	24272	100,0%	1,00	0,000

**Tabela A16** Tabela de “Weights of Evidence” para a Variável  
*Indicador de Apontamento Cadastral*

<i>Ind. Apontamento Cadastral</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>%B/%R</i>	<i>WOE</i>
Não Tem	501	16287	16788	69,2%	1,82	0,599
Tem	786	6698	7484	30,8%	0,48	-0,740
Total	1287	22985	24272	100,0%	1,00	0,000

**Tabela A17** Tabela de “Weights of Evidence” para a Variável  
*Indicador de Contrato de Crediário nos Últimos 2 anos*

<i>Ind. Contrato de Crediário ult. 2 anos</i>	<i>B</i>	<i>R</i>	<i>Total</i>	<i>%Total</i>	<i>%B/%R</i>	<i>WOE</i>
Não Tem	1034	19282	20316	83,7%	1,04	0,043
Tem	253	3703	3956	16,3%	0,82	-0,199
Total	1287	22985	24272	100,0%	1,00	0,000

## APÊNDICE B

### Tabelas de Medidas de Acerto para o REAL com Parâmetros ( $r$ , $v_c$ )

As medidas de acerto consideradas são TAT (taxa de acerto total), TAR (taxa de acerto de ruins) e TAB (taxa de acerto de bons), definidas na Seção 3.4.

**Tabela B1** Medidas de Classificação Correta para o REAL com Parâmetros (2 e 0,3).

Conjunto	TAT	TAR	TAB
1	0,661	0,591	0,736
2	0,677	0,632	0,727
3	0,685	0,610	0,748
4	0,700	0,686	0,713
5	0,634	0,533	0,746
6	0,650	0,556	0,740
7	0,665	0,626	0,701
8	0,650	0,632	0,669
9	0,658	0,640	0,678
10	0,663	0,606	0,716
Média	0,664	0,611	0,718
DP	0,019	0,044	0,028

**Tabela B2** Medidas de Classificação Correta para o REAL com Parâmetros (2 e 0,35).

Partição	TAT	TAR	TAB
1	0,724	0,742	0,704
2	0,747	0,728	0,769
3	0,751	0,746	0,755
4	0,778	0,793	0,765
5	0,673	0,681	0,664
6	0,732	0,722	0,740
7	0,767	0,780	0,754
8	0,728	0,759	0,694
9	0,735	0,801	0,661
10	0,728	0,740	0,716
Média	0,736	0,749	0,722
DP	0,029	0,036	0,040

**Tabela B3** Medidas de Classificação Correta para o REAL com Parâmetros (2 e 0,4).

Conjunto	TAT	TAR	TAB
1	0,743	0,750	0,736
2	0,767	0,735	0,802
3	0,759	0,729	0,784
4	0,778	0,802	0,757
5	0,716	0,696	0,738
6	0,739	0,730	0,748
7	0,770	0,805	0,739
8	0,747	0,767	0,726
9	0,751	0,801	0,694
10	0,732	0,740	0,724
Média	0,750	0,756	0,745
DP	0,019	0,037	0,031

**Tabela B4** Medidas de Classificação Correta para o REAL com Parâmetros (2,5 e 0,3).

Conjunto	TAT	TAR	TAB
1	0,661	0,591	0,736
2	0,677	0,632	0,727
3	0,685	0,610	0,748
4	0,700	0,686	0,713
5	0,634	0,533	0,746
6	0,650	0,556	0,740
7	0,665	0,626	0,701
8	0,650	0,632	0,669
9	0,658	0,640	0,678
10	0,663	0,606	0,716
Média	0,664	0,611	0,718
DP	0,019	0,044	0,028

**Tabela B5** Medidas de Classificação Correta para o REAL com Parâmetros (2,5 e 0,35).

Partição	TAT	TAR	TAB
1	0,704	0,742	0,664
2	0,747	0,735	0,760
3	0,728	0,746	0,712
4	0,759	0,802	0,721
5	0,661	0,689	0,631
6	0,724	0,730	0,718
7	0,747	0,805	0,694
8	0,732	0,774	0,685
9	0,720	0,801	0,628
10	0,697	0,732	0,664
Média	0,722	0,756	0,688
DP	0,029	0,038	0,042

**Tabela B6** Medidas de Classificação Correta para o REAL com Parâmetros (2,5 e 0,4).

Conjunto	TAT	TAR	TAB
1	0,712	0,727	0,696
2	0,739	0,728	0,752
3	0,735	0,737	0,734
4	0,763	0,802	0,728
5	0,685	0,704	0,664
6	0,724	0,730	0,718
7	0,755	0,805	0,709
8	0,739	0,767	0,710
9	0,724	0,801	0,636
10	0,724	0,740	0,709
Média	0,730	0,754	0,706
DP	0,022	0,037	0,034

**Tabela B7** Medidas de Classificação Correta para o REAL com Parâmetros (3 e 0,3).

Conjunto	TAT	TAR	TAB
1	0,661	0,591	0,736
2	0,677	0,632	0,727
3	0,685	0,610	0,748
4	0,700	0,686	0,713
5	0,634	0,533	0,746
6	0,650	0,556	0,740
7	0,665	0,626	0,701
8	0,650	0,632	0,669
9	0,658	0,640	0,678
10	0,663	0,606	0,716
Média	0,664	0,611	0,718
DP	0,019	0,044	0,028

**Tabela B8** Medidas de Classificação Correta para o REAL com Parâmetros (3 e 0,35).

Conjunto	TAT	TAR	TAB
1	0,704	0,742	0,664
2	0,728	0,743	0,711
3	0,716	0,746	0,691
4	0,747	0,793	0,706
5	0,661	0,689	0,631
6	0,712	0,730	0,695
7	0,735	0,797	0,679
8	0,716	0,774	0,653
9	0,716	0,801	0,620
10	0,697	0,732	0,664
Média	0,713	0,755	0,671
DP	0,023	0,036	0,031

**Tabela B9** Medidas de Classificação Correta para o REAL com Parâmetros (3 e 0,4).

Conjunto	TAT	TAR	TAB
1	0,712	0,735	0,688
2	0,724	0,728	0,719
3	0,720	0,737	0,705
4	0,763	0,793	0,735
5	0,689	0,704	0,672
6	0,712	0,722	0,702
7	0,747	0,805	0,694
8	0,720	0,759	0,677
9	0,724	0,801	0,636
10	0,716	0,732	0,701
Média	0,723	0,752	0,693
DP	0,020	0,036	0,027

**Tabela B10** Medidas de Classificação Correta para o REAL com Parâmetros (3,5 e 0,3).

Conjunto	TAT	TAR	TAB
1	0,661	0,591	0,736
2	0,677	0,632	0,727
3	0,685	0,610	0,748
4	0,700	0,686	0,713
5	0,634	0,533	0,746
6	0,650	0,556	0,740
7	0,665	0,626	0,701
8	0,650	0,632	0,669
9	0,658	0,640	0,678
10	0,663	0,606	0,716
Média	0,664	0,611	0,718
DP	0,019	0,044	0,028

**Tabela B11** Medidas de Classificação Correta para o REAL com Parâmetros (3,5 e 0,35).

Conjunto	TAT	TAR	TAB
1	0,704	0,742	0,664
2	0,728	0,743	0,711
3	0,716	0,746	0,691
4	0,739	0,793	0,691
5	0,661	0,689	0,631
6	0,704	0,730	0,679
7	0,720	0,797	0,649
8	0,712	0,774	0,645
9	0,712	0,801	0,612
10	0,693	0,732	0,657
Média	0,709	0,755	0,663
DP	0,021	0,036	0,030

**Tabela B12** Medidas de Classificação Correta para o REAL com Parâmetros (3,5 e 0,4).

Conjunto	TAT	TAR	TAB
1	0,712	0,735	0,688
2	0,716	0,721	0,711
3	0,716	0,737	0,698
4	0,747	0,793	0,706
5	0,693	0,704	0,680
6	0,704	0,730	0,679
7	0,724	0,805	0,649
8	0,712	0,759	0,661
9	0,716	0,801	0,620
10	0,716	0,732	0,701
Média	0,716	0,752	0,679
DP	0,014	0,036	0,029

**Tabela B13** Medidas de Classificação Correta para o REAL com Parâmetros (4 e 0,3).

Conjunto	TAT	TAR	TAB
1	0,661	0,591	0,736
2	0,677	0,632	0,727
3	0,685	0,610	0,748
4	0,700	0,686	0,713
5	0,634	0,533	0,746
6	0,650	0,556	0,740
7	0,665	0,626	0,701
8	0,650	0,632	0,669
9	0,658	0,640	0,678
10	0,663	0,606	0,716
Média	0,664	0,611	0,718
DP	0,019	0,044	0,028

**Tabela B14** Medidas de Classificação Correta para o REAL com Parâmetros (4 e 0,35).

Conjunto	TAT	TAR	TAB
1	0,704	0,742	0,664
2	0,728	0,743	0,711
3	0,712	0,746	0,683
4	0,739	0,793	0,691
5	0,658	0,689	0,623
6	0,700	0,730	0,672
7	0,712	0,780	0,649
8	0,712	0,774	0,645
9	0,658	0,640	0,678
10	0,663	0,606	0,716
Média	0,699	0,724	0,673
DP	0,029	0,061	0,029

**Tabela B15** Medidas de Classificação Correta para o REAL com Parâmetros (4 e 0,4).

Conjunto	TAT	TAR	TAB
1	0,724	0,742	0,704
2	0,720	0,728	0,711
3	0,716	0,746	0,691
4	0,747	0,785	0,713
5	0,673	0,689	0,656
6	0,700	0,730	0,672
7	0,716	0,789	0,649
8	0,712	0,759	0,661
9	0,716	0,801	0,620
10	0,713	0,732	0,694
Média	0,714	0,750	0,677
DP	0,019	0,034	0,031

## BIBLIOGRAFIA

- Arminger, G., Enache, D., Bonne, T. (1997). Analysing Credit Risk Data: A Comparison of Logistic Discrimination, Classification Tree Analysis, and Feedforward Networks. *Computational Statistics*, **12**, 293-310.
- Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International, California.
- Eisenbeis, R. A. (1977). Pitfalls in the Application of Discriminant Analysis in Business, Finance, and Economics. *The Journal of Finance*, **Vol XXXII**, n.3, 875-900.
- Eisenbeis, R. A. (1978). Problems in Applying Discriminant Analysis in Credit Scoring Models. *Journal of Banking and Finance*, **2**, 205-219.
- Good, I. J. (1950). *Probability and the Weighing of Evidence*. Charles Griffin, London.
- Gruenstein, J. M. L. (1998). Optimal Use of Statistical Techniques in Model Building. In: *Credit Risk Modeling: Design and Application*. Mays, E. (ed). pp. 81-112. Amacon: New York.
- Haberman, S. (1978). *Analysis of Qualitative Data*. Vol. 1: Introductory Topics. New York: Academic Press.
- Hand, D. J. and Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of Royal Statistical Society A*, **160**, part.3, pp.523-541.

- Hand, D. J. (1998). Consumer Credit and Statistics. In: *Statistics in Finance*. Hand, D. J. and Jacka, S. D. (eds). pp. 69-81. Edward Arnold.
- Hosmer, Jr. D. W. and Lemeshow, S. (1989). *Applied Logistic Regression*. John Wiley and Sons: New York.
- Hoyland, C. (1997). *Data-Driven Decisions for Consumer Lending*. Lafferty Publications Ltd.: Dublin, Ireland.
- Kass, G.V. (1980). An Explanatory Technique for Investigating Large Quantiles of Categorical Data. *Applied Statistics*, v.29, n.2, p.119-127.
- Kronborg, D., Tjur, T. and Vincents, B. (1998). *Credit Scoring: Discussion of Methods and a Case Study*. Preprint n. 7/1998. Department of Management Science and Statistics, Copenhagen Business School, Denmark.
- Lauretto, M. S. (1996). *Árvores de Classificação para Escolha de Estratégias de Operações em Mercados de Capitais*. Dissertação de Mestrado. Instituto de Matemática e Estatística – Universidade de São Paulo – Brasil.
- Leonard, K. J. (1998). Credit Scoring and Quality Management. In: *Statistics in Finance*. Hand, D. J. and Jacka, S. D. (eds). pp. 105-126. Edward Arnold.
- Lewis, E. M. (1994). *An Introduction to Credit Scoring*. Fair, Isaac & Co., Inc.: California.
- Magidson, J. (1987). Weighted Log-Linear Modeling. In: Proceedings of the Social Statistics Section. *American Statistical Association*, pp.171-174.
- Magidson, J. (1993). The CHAID Approach to Segmentation Modeling. In: *Handbook of Marketing Research*. Bagozzi, R. (ed). pp. 118-159. Blackwell.

Makuch, W. M. (1998). The Basics of a Better Application Score. In: *Credit Risk Modeling: Design and Application*. Mays, E. (ed). pp. 59-80. Amacon: New York.

McCahill, L. J. (1998). Organizational Issues in Building and Maintaining Credit Risk Models. In: *Credit Risk Modeling: Design and Application*. Mays, E. (ed). pp. 13-22. Amacon: New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2ed. Chapman and Hall: London.

Morgan, J.N., Sonquist, J.A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, **58**, pp.415-435, June.

Rao, C. R. (1973). *Linear Statistical Inference and Its Application*. Second Edition. John Wiley and Sons: New York

SPSS for Windows: CHAID, Release 6.0, Magidson, J.: SPSS Inc. 1993.

SPSS for Windows: Advanced Statistics, Release 6.1, Norušis, M. J.: SPSS Inc. 1994.

Srinivasan, V. and Kim, Y. H. (1987). Credit Granting: A Comparative Analysis of Classification Procedures. *Journal of Finance*, v. **XLII**, n.3, pp.665-683.

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. In: *Proceedings of IBERAMIA 98 - Sixth Iberoamerican Conference on Artificial Intelligence*, Lisboa.

Thomas, L. C. (1998). Methodologies for Classifying Applicants for Credit. In: *Statistics in Finance*. Hand, D. J. and Jacka, S. D. (eds). pp. 83-103. Edward Arnold.