UNIVERSIDADE PRESBITERIANA MACKENZIE
CENTRO DE CIÊNCIAS SOCIAIS E APLICADAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO DE EMPRESAS

MODELOS DE CLASSIFICAÇÃO DE RISCO DE CRÉDITO PARA FINANCIAMENTOS IMOBILIÁRIOS: REGRESSÃO LOGÍSTICA, ANÁLISE DISCRIMINANTE, ÁRVORES DE DECISÃO, BAGGING E BOOSTING

NEILSON SOARES LOPES

São Paulo
2011

NEILSON SOARES LOPES

MODELOS DE CLASSIFICAÇÃO DE RISCO DE CRÉDITO PARA FINANCIAMENTOS IMOBILIÁRIOS: REGRESSÃO LOGÍSTICA, ANÁLISE DISCRIMINANTE, ÁRVORES DE DECISÃO, BAGGING E BOOSTING

Dissertação apresentada ao Programa de Pós-Graduação em Administração de Empresas da Universidade Presbiteriana Mackenzie para a obtenção do título de Mestre em Administração de Empresas.

Orientador: Professor Doutor Herbert Kimura

São Paulo

2011

L864m Lopes, Neilson Soares.

Modelos de classificação de risco de crédito para financiamentos imobiliários: regressão logística, análise discriminante, árvores de decisão, Bagging e Boosting / Neilson Soares Lopes - 2011.

87 f. : il. ; 30 cm

Dissertação (Mestrado em Administração de Empresas) – Universidade Presbiteriana Mackenzie, São Paulo, 2011.

Bibliografia: f. 75-80.

1. Risco de crédito. 2. Análise discriminante. 3. Regressão logística. 4. Bagging. 5. Boosting. 6. Arcing. 7. Financiamento imobiliário. I. Título.

CDD 342.1278

REITOR DA UNIVERSIDADE PRESBITERIANA MACKENZIE

Benedito Guimarães Aguiar Neto

DECANO DE PESQUISA E PÓS-GRADUAÇÃO

Moises Ari Zilber

COORDENADOR DE PÓS-GRADUAÇÃO

Diana Luz Pessoa de Barros

COORDENADORA DO PROGRAMA EM ADMINISTRAÇÃO DE EMPRESAS

Darcy Mitiko Mori Hanashiro

DIRETOR DO CENTRO DE CIÊNCIAS SOCIAIS E APLICADAS

Sérgio Lex

NEILSON SOARES LOPES

Dissertação apresentada ao Programa de Pós-Graduação em Administração de Empresas da Universidade Presbiteriana Mackenzie para a obtenção do título de Mestre em Administração de Empresas.

MODELOS DE CLASSIFICAÇÃO DE RISCO DE CRÉDITO PARA FINANCIAMENTOS IMOBILIÁRIOS: REGRESSÃO LOGÍSTICA, ANÁLISE DISCRIMINANTE, ÁRVORES DE DECISÃO, BAGGING E BOOSTING

BANCA EXAMINADORA

_____

Prof.Dr. Herbert Kimura

Universidade Presbiteriana Mackenzie

_____

Prof.Dr. Leonardo Fernando Cruz Basso

Universidade Presbiteriana Mackenzie

_____

Prof.Dr. Eduardo Kazuo Kayo

Universidade de São Paulo

São Paulo, 08 de Agosto de 2011.

AGRADECIMENTOS

Agradeço ao meu orientador, Prof. Dr. Herbert Kimura, pelo empenho em buscar sempre a excelência no resultado deste trabalho, além do incentivo e parceria que demonstrou nestes dois anos em que aconteceram tantas mudanças em minha vida.

Agradeço ao Prof. Dr. Leonardo Fernando Cruz Basso pelo interesse que demonstrou por minha pesquisa, pois suas sugestões me mostraram caminhos que enriqueceram a construção do conhecimento que adquiri.

Agradeço também ao Prof. Dr. Luiz Carlos Jacob Perera que me apresentou ao CART da Salford Systems, software que contribuiu para os resultados deste trabalho.

Aos meus chefes e amigos, Max Freddy e Cláudia, agradeço pela confiança que sempre depositaram em meu trabalho e pelo tempo que me permitiram dedicar a este projeto tão especial.

Encerro agradecendo à minha mulher, Cyntia, e meus filhos Erik e Íris, por estarem sempre ao meu lado, e cuja força me conduziu até aqui.

RESUMO

Neste estudo foram aplicadas as técnicas paramétricas tradicionais de análise discriminante e regressão logística para análise de crédito de operações de financiamento imobiliário. Foi comparada a taxa de acertos destes métodos com as técnicas não-paramétricas baseadas em árvores de classificação, além dos métodos de meta-aprendizagem BAGGING e BOOSTING, que combinam classificadores para obter uma melhor precisão nos algoritmos.

Em um contexto de alto déficit de moradias, em especial no caso brasileiro, o financiamento de imóveis ainda pode ser bastante fomentado. Os impactos de um crescimento sustentável no crédito imobiliário trazem benefícios não só econômicos como sociais. A moradia é, para grande parte dos indivíduos, a maior fonte de despesas e o ativo mais valioso que terão durante sua vida.

Ao final do estudo, concluiu-se que as técnicas computacionais de árvores de decisão se mostram mais efetivas para a predição de maus pagadores (94,2% de acerto), seguida do BAGGING (80,7%) e do BOOSTING (ou ARCING, 75,2%). Para a predição de maus pagadores em financiamentos imobiliários, as técnicas de regressão logística e análise discriminante apresentaram os piores resultados (74,6% e 70,7%, respectivamente). Para os bons pagadores, a árvore de decisão também apresentou o melhor poder preditivo (75,8%), seguida da análise discriminante (75,3%) e do BOOSTING (72,9%). Para os bons pagadores de financiamentos imobiliários, BAGGING e regressão logística apresentaram os piores resultados (72,1% e 71,7%, respectivamente).

A regressão logística mostra que, para um tomador com crédito consignado, a chance se ser um mau pagador é 2,19 maior do que se este tomador não tivesse tal modalidade de empréstimo. A presença de crédito consignado entre as operações dos tomadores de financiamento imobiliário também apresenta relevância na análise discriminante.

Palavras-chave: Risco de Crédito, Análise Discriminante, Regressão Logística, BAGGING, Boosting, Arcing, Financiamento Imobiliário.

ABSTRACT

This study applied the techniques of traditional parametric discriminant analysis and logistic regression analysis of credit real estate financing transactions where borrowers may or may not have a payroll loan transaction. It was the hit rate compared these methods with the non-parametric techniques based on classification trees, and the methods of meta-learning bagging and boosting that combine classifiers for improved accuracy in the algorithms.

In a context of high housing deficit, especially in Brazil, the financing of real estate can still be very encouraged. The impacts of sustainable growth in the mortgage not only bring economic benefits and social. The house is, for most individuals, the largest source of expenditure and the most valuable asset that will have during her lifetime.

At the end of the study concluded that the computational techniques of decision trees are more effective for the prediction of payers (94.2% correct), followed by bagging (80.7%) and boosting (or arcing , 75.2%). For the prediction of bad debtors in mortgages, the techniques of logistic regression and discriminant analysis showed the worst results (74.6% and 70.7%, respectively). For the good payers, the decision tree also showed the best predictive power (75.8%), followed by discriminant analysis (75.3%) and boosting (72.9%). For the good paying mortgages, bagging and logistic regression showed the worst results (72.1% and 71.7%, respectively).

Logistic regression shows that for a borrower with payroll loans, the chance to be a bad credit is 2.19 higher than if the borrower does not have such type of loan.

The presence of credit between the payroll operations of mortgage borrowers also has relevance in the discriminant analysis.

Keywords: Credit Risk, Discriminant Analisys, Logistic Regression, BAGGING, Boosting, Arcing, Real Estate Financing.

LISTA DE GRÁFICOS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Neste estudo pretendeu-se responder às seguintes questões: as técnicas computacionais de árvore de decisão, além das técnicas de aprendizagem de máquina BAGGING e BOOSTING, são mais efetivas que as técnicas tradicionais análise discriminante e regressão logística na classificação de risco de crédito de financiamentos imobiliários? A presença de um crédito consignado interfere no risco de crédito dos financiamentos imobiliários?

Este estudo traz uma pesquisa quantitativa, com utilização de técnicas estatísticas para classificação das observações. Foram aplicadas as técnicas paramétricas tradicionais para análise de crédito: a análise discriminante e a regressão logística bem como as técnicas não-paramétricas árvore de classificação, BAGGING (BREIMAN, 1996b) e BOOSTING (MITCHELL, 1997). Foi comparada a taxa de acertos dos métodos citados.

A análise discriminante linear multivariada tem como base a obtenção de uma função discriminante que associa variáveis independentes a um escore indicativo do grupo mais provável no qual um indivíduo ou empresa pode ser classificado.

A regressão logística estabelece que o *logit* é uma função linear das variáveis explicativas X_i. No entanto, ao invés do escore Z obtido na análise discriminante e representativo da proximidade de uma observação em relação a um grupo, na regressão logística estima-se uma relação linear entre variáveis explicativas e a razão entre probabilidades de uma observação pertencer a um e a outro grupo. No caso da análise de crédito usando a equação da regressão logística descrita anteriormente para discriminação entre grupos de adimplentes ou inadimplentes, quanto maior o valor do *logit*, mais provável o indivíduo ser inadimplente.

A Árvore de classificação é uma técnica menos tradicional para discriminação entre grupos. Esta modelagem não-paramétrica utiliza algoritmos de partição recursiva (THOMAS, EDELMAN e CROOK, 2002). De acordo com Feldesman (2002), as árvores de classificação possuem diversas vantagens frente a modelos

paramétricos: não necessitam de transformações de dados como é o caso da função *logit* na análise de regressão logística, observações com valores faltantes não exigem tratamento especial, o sucesso na classificação não depende de premissas de normalidade de variáveis ou de igualdade de matrizes de variância-covariância entre grupos como é o caso da análise discriminante.

Breiman (1996b) descreve o BAGGING como um método para gerar várias versões aleatórias de um preditor, utilizando-se o mesmo algoritmo, na intenção de obter um indicador agregado. As médias de agregação sobre as versões predizem um resultado numérico por meio de um voto majoritário na previsão de uma classe. São formadas outras versões replicadas, buscando o aprendizado conjunto e utilizando esses conjuntos como novo aprendizado. O resultado traz ganhos substanciais em termos de precisão. Para Breiman (1996b), o método BAGGING melhora uma estimativa instável, além de reduzir a variância para um processo de base de dados, como árvores de decisão ou métodos que fazem seleção de variáveis e encaixe em um modelo linear.

Segundo Freund e Schapire (1997), o método BOOSTING combina iterativamente classificadores redefinindo probabilidades que aumentam a cada instância para os mais difíceis de serem classificados. Embora utilize a combinação de classificadores, assim como o BAGGING, neste as classificações incorretas recebem maior probabilidade na nova etapa, com o objetivo de se obter classificadores mais diversificados. Em Freund e Schapire (1999) o BOOSTING é descrito como um método poderoso na combinação de classificadores de base múltipla para formar um conjunto cujo desempenho que pode ser significativamente melhor do que de qualquer um dos classificadores-base.

A partir de uma ampla amostra contemplando dados de operações fornecidas por uma instituição brasileira que atua nos segmentos de financiamento imobiliário e crédito consignado, o estudo verificou quais variáveis e técnicas melhor discriminam os bons e maus pagadores. Foram definidos como maus pagadores os tomadores com operações onde o atraso é superior a 59 dias.

Considerando o uso de uma base de dados com um número elevado de observações, com um acompanhamento ao longo do tempo, foram estudados comportamentos temporais dos modelos de análise de crédito. Com isso, foram conduzidas análises de estabilidade dos modelos que, comumente, não são empreendidas em pesquisa acadêmicas sobre crédito.

Wagner (2004) mostra que Bancos que visam crescimento da carteira de financiamentos imobiliários devem ter mecanismos de controle de riscos eficientes. Apesar da crescente integração global do setor financeiro, o mercado de habitação é predominantemente nacional. Na maioria dos países, um número pequeno de bancos comerciais domina os empréstimos hipotecários. Apesar disso, o Estado e os bancos cooperativos e regionais ainda estão ativos em alguns países onde eles têm tradicionalmente desempenhado um papel importante, como na Alemanha, Espanha e Suíça. No Japão, dois grandes bancos privados dominam esta modalidade de financiamento. Nos Estados Unidos, o mercado de crédito hipotecário é extremamente competitivo. No México o maior credor é o Governo, por meio de agências de habitação, que originaram mais de 70% dos empréstimos em 2004 (BIS, 2006).

Para o BIS (2006), a evolução do mercado de financiamentos habitacionais é fruto de tendências macroeconômicas, tanto em países industrializados como em emergentes, que promoveram uma mudança estrutural que conduziram à redução observada nos níveis de risco e taxas de juros nominais de longo prazo. Alguns dos fatores que levaram a este cenário é a redução no nível e na volatilidade da inflação em todo o Mundo, desde 1970. Este fenômeno pode ter sido motivado pelo maior empenho das autoridades monetárias para a estabilidade de preços.

O resultado mais direto do declínio da taxa de juros é o acesso de mais famílias ao crédito. Além disso, o baixo nível risco para juros de longo prazo pode ter estimulado a busca por rentabilidade por parte das Instituições financeiras. Para o BIS (2006) os tomadores de financiamento imobiliário tendem a decidir sobre seus financiamentos com base na dívida nominal (valor principal + juros).

As evidências disponíveis sugerem que o declínio nas taxas de juros nominal tem estimulado tanto a procura quanto a oferta de financiamentos imobiliários. Observa-se que a relação entre o crescimento macroeconômico e do mercado imobiliário não parece ser unilateral, ou seja, há substancial evidência empírica de que os efeitos do desenvolvimento da habitação podem ter influenciado a evolução econômica em diversos países. A elevação do valor dos imóveis provocou o aumento no nível de riqueza das famílias. Na maioria dos países, as famílias são os principais proprietários do ativo habitacional, seja de suas próprias casas, seja de imóveis alugados. Assim, uma subida dos preços imobiliários expande o valor dos seus ativos em relação a suas responsabilidades, o que permite ampliar a capacidade para absorver novos empréstimos.

Ainda segundo o BIS (2006) a tecnologia tem permitido às Instituições financeiras utilizar-se de forma mais eficiente suas informações sobre os tomadores de financiamentos. O agrupamento em perfis de risco específicos, por meio da pontuação de crédito é uma destas maneiras.

Nos mercados de financiamento de habitação, particularmente nos Estados Unidos, houve um aumento da utilização de modelos de classificação de crédito. Fornecedores de serviço como as agências de classificação adaptaram seus relatórios, de modo que trouxessem em suas avaliações o tipo de tomador considerando, inclusive, suas modalidades de empréstimos (hipotecas, cartões de crédito, financiamentos de veículos, empréstimos estudantis, etc.). Também o acompanhamento histórico da transformação de um tomador de crédito permite uma medição mais precisa e eficaz do seu risco de crédito. Um dos benefícios decorrentes é a melhor visão dos credores e investidores na apuração de seus modelos estatísticos de inadimplência, melhorando, assim, tanto o apreçamento dos ativos quanto o gerenciamento de seu risco.

Um risco associado financiamento imobiliário é que as famílias tendem a se endividar acima de sua capacidade, pois são limitadas na avaliação da sua dívida de longo prazo, por exemplo, ignorando a volatilidade dos juros em financiamentos de taxa ajustável. Outro aspecto destacado pelo BIS (2006) é que os credores podem

não fornecer a orientação adequada para as famílias, particularmente às menos experientes.

Não é incomum indivíduos que possuem um financiamento para aquisição de um imóvel contraírem empréstimos consignados visando obter recursos para cobrir outras necessidades de caixa. Desta forma, como estes tomadores podem comprometer grande parte de sua renda com pagamento de prestações de empréstimos, torna-se relevante avaliar potenciais impactos do crédito consignado na capacidade de pagamento do financiamento imobiliário. Neste contexto, esta pesquisa também investigou se o crédito consignado tem influência no risco de crédito de financiamentos imobiliários.

O Banco Central do Brasil apresentou em seu Relatório de Economia Bancária (BRASIL, 2007) que o empréstimo pessoal com consignação em folha de pagamento, denominado crédito consignado, é uma das inovações de maior sucesso no mercado financeiro brasileiro nos últimos anos. Embora o crédito consignado já fosse utilizado por instituições financeiras em operações com trabalhadores assalariados, esta modalidade de empréstimo popularizou-se a partir da regulamentação por meio da Lei 10.820 e do Decreto 4.840. Em setembro de 2009, alcançou um volume de R$ 100,5 bilhões, constituindo um crescimento de 2,1% no mês e 32% em doze meses. As operações em que a mensalidade é descontada em folha de pagamento trazem outra particularidade: o prazo mais dilatado em relação às demais operações não-consignadas.

As Instituições Financeiras, apesar de buscarem aumentar o número de clientes e o volume de operações, procuram redirecionar os recursos para ativos mais rentáveis. Contudo, as alterações na alocação entre ativos mostram que cresce o crédito consignado em substituição às modalidades mais rentáveis, como cheque especial e cartão de crédito, reduzindo a margem de intermediação (BRASIL, 2009). Neste sentido, a gestão eficiente do risco destas operações é de grande relevância.

O empréstimo consignado tem seu risco de crédito mitigado pelo desconto automático das parcelas em folha de pagamento, podendo, portanto, oferecer menores taxas de juros aos tomadores de recursos. Considerando as características

da operação, muitas vezes, o crédito consignado é concedido até mesmo a pessoas com restrições cadastrais. Desta forma, há situações em que tomadores alavancam sua posição contraindo empréstimos consignados que colocam em risco sua capacidade de pagamento. De acordo com Cavallazzi (2006), pesquisa realizada em 2005 no Estado do Rio de Janeiro mostra que dentre 80 endividados selecionados, 39% comprometeram mais de 60% da renda para o pagamento do empréstimo consignado. Portanto, apesar de o crédito consignado ter um risco menor para a Instituição financeira, pode ocasionar um sobreendividamento do tomador e aumentar o risco de crédito de outra operação previamente contratada.

Este estudo está dividido em sete seções, onde a segunda apresenta o levantamento bibliográfico, abordando os conceitos iniciais, a relevância do crédito imobiliário e um panorama do crédito consignado no Brasil. A terceira descreve a metodologia utilizada na pesquisa, abordando as técnicas quantitativas, a amostra utilizada, o tratamento dado às variáveis e os estudos referenciais. A quarta seção fala das exclusões do trabalho e a quinta mostra os resultados obtidos. A sexta seção relata as conclusões e a sétima traz as referências bibliográficas utilizadas.

2. LEVANTAMENTO BIBLIOGRÁFICO

2.1. Conceitos iniciais

De acordo com Lewis (1992), o crédito ao consumidor já era realizado há mais de 3.000 anos, desde os tempos dos babilônios. O crédito possibilita que indivíduos possam consumir ou ter acesso a bens, mesmo em situações nas quais enfrentem déficits de caixa. Neste contexto, o crédito estimula a economia, uma vez que disponibiliza recursos para que agentes possam satisfazer suas necessidades (BRIGHAM, GAPENSKI e EHRHARDT, 2001).

No entanto, operações de crédito possuem um risco inerente, no qual o tomador de recursos no presente pode não ter condições ou não ter interesse, no futuro, de pagar o empréstimo contraído. De fato, conforme Caouette, Altman e Narayanan (1998), se o crédito implica em expectativa, pelos credores, de recebimento de recursos em uma data futura, então o risco de crédito envolve a possibilidade de essa expectativa não se concretize devido ao inadimplemento da contraparte.

Assim, o risco de o financiador não receber o reembolso pelo seu empréstimo, seja em função de a contraparte não desejar ou não ser capaz de cumprir suas obrigações (JORION, 1998), está associado ao risco de crédito. Comumente considera-se como risco de crédito um evento de não-pagamento ou de inadimplência. Todavia, o risco de crédito pode advir também de uma deterioração da qualidade de crédito da contraparte, uma vez que o valor marcado a mercado dessa posição pode diminuir.

A expansão do crédito observado nos últimos anos no Brasil chama a atenção sobretudo no que diz respeito à qualidade das operações financeiras. Minsky (1982) concluiu em seus estudos que, quando uma economia entra em expansão sustentada por tempo prolongado, os organismos financeiros tendem a adotar maiores níveis de alavancagem, e esta postura ampliam sua exposição ao risco. Como conseqüência a esta maior alavancagem, as unidades econômicas ficam mais

vulneráveis a problemas de liquidez e a interdependência entre elas causa um efeito multiplicado no sistema econômico.

2.2. A relevância do crédito imobiliário

Entre 2002 e 2009 o valor financiado das operações de crédito imobiliário apresentou uma evolução de 851%. Se observada a quantidade de financiamentos, somente de 2008 a 2009 o aumento foi de 25% (ABECIP).

Atualmente existem dois sistemas em que são realizadas as operações de financiamento imobiliário no Brasil: o Sistema Financeiro de Habitação (SFH), criado na década de 60; e o Sistema Financeiro Imobiliário (SFI), criado em 1997 e introduzindo a alienação fiduciária em detrimento da hipoteca.

Segundo a ABECIP, os principais fatores de expansão do crédito imobiliário no Brasil estão associados à estabilidade econômica, recuperação da renda, redução da inadimplência e flexibilização das condições de financiamento.

Em relação aos contratos com mais de três parcelas de atraso em relação ao total, o estudo mostra que em 2003 este indicador era de 11,2%. Em 2006 reduziu para 6,3% e em 2009 apenas 2,6% dos contratos apresentam mais de três parcelas em atraso, com tendência à redução.

No que diz respeito ao déficit da habitação no Brasil, observa-se que há evidências de que a tendência de crescimento permanece. A ABECIP estima em 7,9 milhões de unidades este déficit. Mais de 80% deste déficit esta nas áreas urbanas (FJP, 2008).

O gráfico a seguir mostra, numa posição de 2004, um comparativo entre países no que diz respeito à proporção de financiamentos imobiliários, lastreados por recursos de poupança, em relação ao PIB. Observa-se que, em termos relativos, os financiamentos no Brasil são ainda incipientes. Contudo, o volume de operações com esta característica vem aumento no decorrer dos anos.

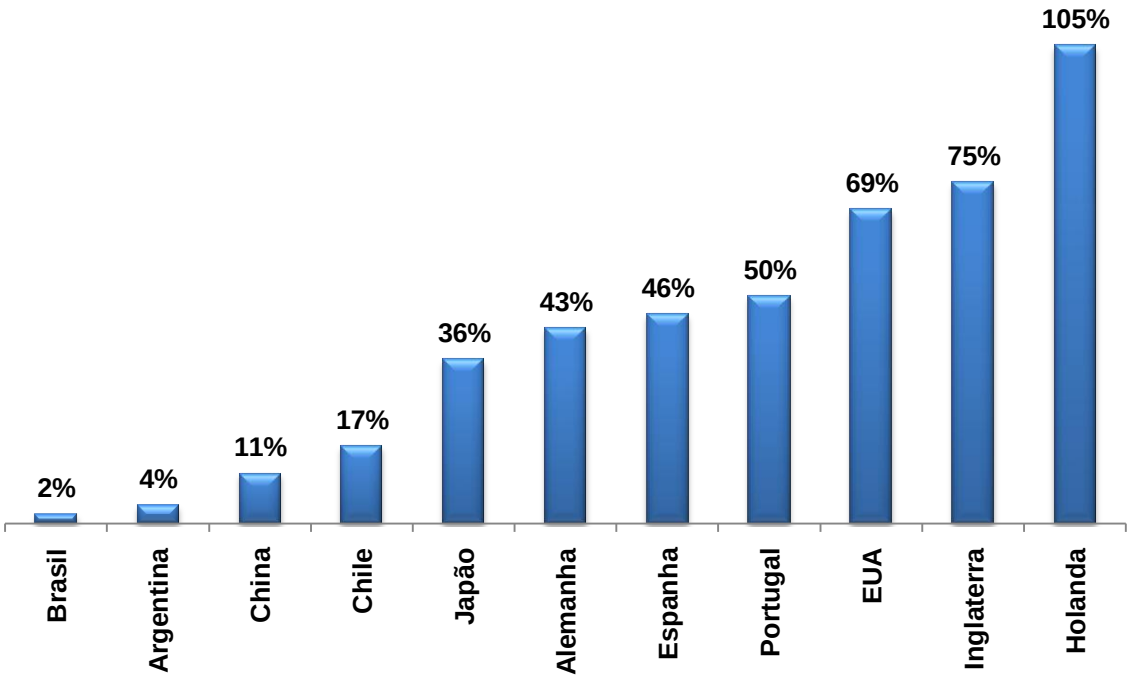

GRÁFICO 1: Financiamentos imobiliários em relação ao PIB
Fonte: ABECIP

Deve-se destacar que o impacto econômico e social das questões relacionadas à habitação, combinado com a tendência de crescimento observada, colocam a questão da moradia na pauta de governos e de órgãos nacionais e internacionais.

Dada esta situação, o déficit habitacional em países emergentes tende a se agravar, pois como o valor dos imóveis é muito superior à renda do indivíduo, existe a necessidade de que os empréstimos sejam de longo prazo, impondo uma maior complexidade na análise de riscos e, portanto, maior cautela na concessão de crédito.

2.3. O crédito consignado no Brasil

Dentre os diversos tipos de crédito ao consumidor, o empréstimo consignado é uma das modalidades em que o desconto das parcelas é feito diretamente na folha

de pagamento e, portanto, envolve um risco de crédito reduzido. Segundo Takeda e Bader (2005), a evolução do saldo de empréstimo consignado está fortemente relacionada a três períodos: (i) dezembro de 2003, com a promulgação da lei que dispõe sobre o crédito consignado em folha de pagamento, (ii) maio de 2004, com o início das operações consignadas para aposentados e pensionistas do INSS e (iii) dezembro de 2004, período a partir do qual diversas instituições já haviam firmado, contratualmente, convênios para a concessão desse tipo de crédito.

Através da Lei 8.112 de 1990 (BRASIL, 2004), os funcionários públicos já podiam realizar a operação de crédito com desconto de prestações em folha de pagamento, porém o mercado de crédito consignado popularizou-se somente com a Lei 10.820 de 2003 que propiciou acesso ao crédito consignado aos trabalhadores da iniciativa privada e aos aposentados e pensionistas do INSS.

Apesar de o crédito consignado permitir condições mais favoráveis de crédito, notadamente na forma de taxas de juros mais baixas, a facilidade de acesso a essa modalidade de operação pode induzir o tomador a alavancar-se demasiadamente, colocando em risco sua capacidade de pagamento de obrigações contratuais. Em pesquisa realizada pelo PROCON (2006), mais de 19% dos entrevistados afirmaram que seriam necessários cortes ou atrasos no pagamento de algum item essencial do orçamento doméstico em função da contratação de uma operação de empréstimo consignado (PINHEIRO, 2007). Neste contexto, se por um lado o crédito consignado pode estimular o consumo e aprimorar a qualidade de vida dos indivíduos, por outro lado, pode impor dificuldades financeiras devido à alta alavancagem e impactar outras obrigações contratuais do tomador.

2.4. Técnicas quantitativas

Para que a pesquisa tivesse maior aplicabilidade, foram desenvolvidos algoritmos e funções que ilustraram o uso de alguns dos modelos estudados. As análises foram realizadas com o auxílio do *software* SPSS (*Statistical Package for Social Sciences*), que permite grande flexibilidade no que diz respeito a volume de dados e testes estatísticos. As rotinas para árvores de classificação, BAGGING e

BOOSTING foram programadas no CART (*Classification and Regression Trees*) 6.0 da Salford Systems. O software idealizado por Dan Steinberg e Mikhail Golovnya busca padrões de relacionamento entre variáveis complexas para montar árvores de decisão de modelos preditivos de gestão de riscos, além de outras aplicações.

A pesquisa realizada neste trabalho foi do tipo quantitativo, com utilização de técnicas estatísticas para classificação de observações. Especificamente neste estudo, foram aplicadas as técnicas paramétricas tradicionais para análise de crédito: a análise discriminante e a regressão logística. Foi comparada a taxa de acertos dos métodos citados, antes e depois de aplicados os métodos meta-aprendizagem BAGGING (BREIMAN, 1996b) e BOOSTING (MITCHELL, 1997) que, combinam classificadores para obter uma melhor precisão nos algoritmos.

Para o caso específico de pesquisas sobre concessão de crédito, um problema comum envolve o estudo de variáveis que explicam diferenças entre grupos de clientes com qualidades de crédito distintas. Assim, do ponto de vista estatístico, a distinção entre possíveis grupos que seguem um determinado ranqueamento pode ser avaliada de modo simples e conveniente através de uma escala ordinal (O'CONNELL, 2006). De acordo com Stevens (1946), a principal característica de um dado ordinal é que podem representar categorias distinguíveis por diferenças em magnitude.

Considerando uma classificação entre bons ou maus pagadores ou entre *ratings* de crédito, pode-se inferir que variáveis com escala ordinal são especialmente úteis no estudo de potencial de inadimplência. Por exemplo, pode-se definir que o valor de variável é 0, se determinado cliente é bom pagador, ou seja, adimplente, e que o valor da variável é 1, se determinado cliente é mau pagador, isto é, inadimplente. Neste estudo foram classificados como maus pagadores os tomadores em que a operação de financiamento imobiliário apresenta atraso superior a 59 dias.

Em outro exemplo, no caso de *rating* de crédito, a variável ordinal pode assumir valores 1, 2, 3, 4 ou 5, representativos, por exemplo, de probabilidade de inadimplência baixíssima, baixa, média, alta ou altíssima. Apesar da facilidade de

definição de grupos através de variáveis ordinais, segundo Cliff (1996), a análise de variáveis dependentes ordinais implica em desafios na modelagem estatística, uma vez que diversas premissas de modelos usuais de regressão linear não são apropriadas para esses casos. No contexto de análise de crédito, técnicas paramétricas tradicionais para classificação de grupos de adimplentes e inadimplentes são a análise discriminante e a regressão logística.

Na etapa de aplicação das técnicas de construção dos modelos, a seleção e avaliação de variáveis independentes utilizadas são de fundamental importância para se prever, por exemplo, a multicolinearidade. As principais premissas da análise discriminante são a normalidade multivariada, homogeneidade de matrizes de variância e ausência de multicolinearidade. A regressão logística tem como principal premissa a ausência de multicolinearidade.

A escolha modelo proposto foi baseada na avaliação do percentual de acerto nas classificações. As observações foram cruzadas com as previsões. A quantidade de acertos a ser utilizada como referência foi, portanto, dividida pelo total de clientes que compuseram a base de dados analisada.

2.4.1. Análise discriminante

A análise discriminante linear multivariada tem como base a obtenção de uma função discriminante que associa variáveis independentes a um escore indicativo do grupo mais provável no qual um indivíduo ou empresa pode ser classificado. Supondo que a qualidade de crédito, tipificada em dois grupos, por exemplo, de adimplentes e inadimplentes, possa ser explicada por n variáveis independentes, a função discriminante tem a seguinte forma:

$$Z = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \cdots + b_n \cdot X_n$$

Onde,

Z representa o escore discriminante

b_i, para $i = 0, 1, \ldots, n$, constituem, respectivamente, o intercepto e coeficientes que ponderam as variáveis independentes X_i na função discriminante.

X_i, para $i = 1, \ldots, n$, representam os valores da i-ésima variável discriminatória, isto é, independente.

Com relação à função discriminante, o mecanismo da análise discriminante implica a obtenção do intercepto b_0 e dos coeficientes $b_1, \ldots, b_n$ que, ao serem multiplicados pelos valores das variáveis independentes, geram um escore Z, cuja variabilidade seja máxima entre os diferentes grupos e mínima dentro dos grupos (HAIR JR. *et al.*, 1998). Ou seja, na análise discriminante busca-se estimar os coeficientes b_i, $i = 0, 1, \ldots, n$, tais que os escores Z calculados para observações de um mesmo grupo sejam similares e os escores Z para observações de grupos diferentes sejam bastante distintos. Os centróides na análise discriminante indicam os valores médios dos escores das observações dos diferentes grupos. A proximidade de uma observação a um centróide implica em uma maior probabilidade de esta observação pertencer ao respectivo grupo do centróide. Desta forma, uma vez obtidos os coeficientes b_i, a partir de uma amostra é possível usar a função discriminante para estimar a qual grupo uma nova observação, com determinados valores de X_i, pertence.

A aplicação em crédito é imediata. A partir de variáveis independentes ou explicativas X_i, representativas de características do potencial tomador de recursos como, por exemplo, renda familiar, número de dependentes, idade, gênero, pode-se usar a função discriminante e obter o escore Z que pode indicar a classificação mais provável do indivíduo no grupo de adimplentes ou no grupo de inadimplentes.

Dentre as premissas da análise discriminante linear multivariada, as principais são: (i) nenhuma variável independente pode ser uma combinação linear das outras

variáveis discriminatórias, (ii) as matrizes de variâncias-covariâncias populacionais são iguais para os diferentes grupos (KLECKA, 1980) e (iii) que cada variável explicativa possui uma distribuição normal e, mais ainda, a distribuição conjunta dessas variáveis segue uma distribuição normal multivariada (HARRELL JR., 2001).

Assim, a despeito da simplicidade em sua formulação, a análise discriminante possui premissas que não são obedecidas na realidade. No entanto, é importante ressaltar que a análise discriminante tem sido uma das principais ferramentas estatísticas que subsidiam modelos de concessão de crédito.

2.4.2. Regressão logística

Considerando que a análise discriminante possui premissas restritivas que podem ser violadas na prática, diversos modelos de análise de concessão de crédito usam a regressão logística, que envolve o cálculo da função *logit* representativa do logaritmo neperiano da razão entre a probabilidade de se pertencer a um grupo e a probabilidade de se pertencer a outro grupo, conforme equação a seguir (O'CONNELL, 2006):

$$logit[\pi(X_1, \cdots, X_n)] = ln\left(\frac{\pi(X_1, \cdots, X_n)}{1 - \pi(X_1, \cdots, X_n)}\right) = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \cdots + b_n \cdot X_n$$

Onde,

$\pi(X_1, \cdots, X_n)$ representa a probabilidade, por exemplo, de um indivíduo ser inadimplente, dado seus valores das variáveis X_i.

Assim, a regressão logística estabelece que o *logit* é uma função linear das variáveis explicativas X_i. No entanto, ao invés do escore Z obtido na análise discriminante e representativo da proximidade de uma observação em relação a um grupo, na regressão logística estima-se uma relação linear entre variáveis explicativas e a razão entre probabilidades de uma observação pertencer a um e a outro grupo. No caso da análise de crédito usando a equação da regressão logística descrita anteriormente para discriminação entre grupos de adimplentes ou

inadimplentes, quanto maior o valor do *logit*, mais provável o indivíduo ser inadimplente. Com relação às premissas, a formulação da regressão logística permite maior flexibilidade em relação à análise discriminante. Em particular, as premissas da regressão logística envolvem a linearidade do relacionamento na equação e a ausência de iteração entre variáveis independentes (HARRELL JR., 2001). O relaxamento das premissas de normalidade de variáveis explicativas e de igualdade de matrizes de variâncias e covariâncias dá mais subsídios para se trabalhar com variáveis ordinais e até mesmo variáveis categóricas como explicativas do potencial de pagamento de um empréstimo. Deve-se ressaltar que dados de cadastros de crédito envolvem muitas vezes variáveis categóricas como, por exemplo, gênero e estado civil.

2.4.3. Árvores de classificação

Uma técnica menos tradicional para discriminação entre grupos que envolvem modelagem não-paramétrica e que foi investigada nesse estudo está associada às árvores de classificação, frequentemente denotada por algoritmos de partição recursiva (THOMAS, EDELMAN e CROOK, 2002). De acordo com Feldesman (2002), as árvores de classificação possuem diversas vantagens frente a modelos paramétricos: (i) não necessitam de transformações de dados como é o caso da função *logit* na análise de regressão logística, (ii) observações com valores faltantes não exigem tratamento especial, (iii) o sucesso na classificação não depende de premissas de normalidade de variáveis ou de igualdade de matrizes de variância-covariância entre grupos como é o caso da análise discriminante.

A idéia que envolve o algoritmo de partição recursiva é subdividir diversas vezes o conjunto de observações em dois de tal maneira que os subgrupos subseqüentes formados sejam cada vez mais homogêneos (THOMAS, EDELMAN e CROOK, 2002). No caso da análise de crédito, a primeira divisão ou corte distingue entre bons e maus pagadores, o segundo corte identifica a variável que diferencia bons e maus pagadores, o terceiro corte estabelece outra variável que diferencia o comportamento da variável anterior, e assim sucessivamente.

Para cada variável relevante, o algoritmo estabelece um valor de referência que serve para formar os subgrupos. Assim, por exemplo, se a variável diferenciadora X é contínua, então o algoritmo obtém um valor de corte k, tal que os dois subgrupos são formados em função de a observação ter um valor $X \leq k$ ou $X > k$. No caso de a variável diferenciadora X ser categórica, o algoritmo verifica todas as possíveis separações das categorias em duas e define uma medida que permite classificar os grupos (THOMAS, EDELMAN e CROOK, 2002). Portanto, a definição do valor de corte k é fundamental para o modelo de árvores de classificação. De acordo com Anderson (2007), as medidas mais usuais baseiam-se na estatística de Kolmogorov-Smirnov, no índice de impureza básico, no índice de Gini, no índice de entropia e na media da soma de quadrados.

As árvores de classificação possibilitam uma representação intuitiva e de fácil entendimento (BREIMAN *et al.*, 1984). Trabalhos que aplicam as árvores de classificação para análise de crédito, embora não tão difundidos quanto estudos que usam modelos paramétricos, podem ser encontrados em Coffman (1986) e Lemos, Steiner e Nievola (2005).

2.4.3.1. BAGGING

Breiman (1996b) descreve o BAGGING como um método para gerar várias versões aleatórias de um preditor, utilizando-se o mesmo algoritmo, na intenção de obter um indicador agregado. As médias de agregação sobre as versões predizem um resultado numérico por meio de um voto majoritário na previsão de uma classe. São formadas outras versões replicadas, buscando o aprendizado conjunto e utilizando esses conjuntos como novo aprendizado. O resultado traz ganhos substanciais em termos de precisão.

Para Breiman (1996b), o método BAGGING melhora uma estimativa instável, além de reduzir a variância para um processo de base de dados, como árvores de decisão ou métodos que fazem seleção de variáveis e encaixe em um modelo linear. É um método de execução relativamente simples. Posteriormente à sua invenção, Bühlmann e Yu (2002) mostraram que este método consiste numa uma operação de

alisamento que melhora o desempenho preditivo de árvores de regressão ou classificação. Em caso de árvores de decisão, Bühlmann e Yu (2002) confirmaram a proposição de Breiman (1996b) de que a técnica reduz a variância, reduzindo também o erro quadrado médio.

O método de Breiman (1996b) explora a instabilidade observada, que é o elemento fundamental, onde classificadores diferentes em termos de comportamento resultam em conjuntos com pequenas variações.

O algoritmo BAGGING é assim definido (BREIMAN, 1996b):

1. Construção de uma amostra (X_1^*, Y_1^*), ..., (X_n^*, Y_n^*), aleatoriamente, n vezes, a partir dos dados (X_1, Y_1), ..., (X_n, Y_n).

2. Calcule o estimador $\hat{g}^*(.)$ pelo princípio plug-in $\hat{g}^*(.) = h_n((X_1^*, Y_1^*), ..., (X_n^*, Y_n^*))(.)$.

3. Repita os dois primeiros passos M vezes, em torno de 50 a 100, resultando em $\hat{g}^{*k}(.)$ $(k = 1, \dots, M)$.

O estimador deduzido é: $\hat{g}_{bag}(.) = M^{-1}\sum_{k=1}^{M} \hat{g}^{*k}(.)$.

O estimador teórico é: $\hat{g}_{bag}(.) = E^*[\hat{g}^*(.)]$.

A quantidade teórica corresponde a $M \to \infty$: o número finito M, na prática é apurado pela aproximação de Monte Carlo mas, caso contrário, não deve ser utlizado como um parâmetro de ajuste para o BAGGING.

Sobre o desempenho do método BAGGING, Breiman (1996b) afirma que a variância do deve ser igual ou menor do que a variância do estimador original, e pode haver uma redução drástica se a estimativa original for instável.

2.4.3.2. BOOSTING

Segundo Freund e Schapire (1997), o método BOOSTING combina iterativamente classificadores redefinindo probabilidades que aumentam a cada instância para os mais difíceis de serem classificados. A redefinição tem por base o erro estatístico calculado para o conjunto observado. Embora utilize a combinação de classificadores, assim como o BAGGING, neste a interatividade da seleção de probabilidades não está presente. No BOOSTING, as classificações incorretas recebem maior probabilidade na nova etapa, com o objetivo de se obter classificadores mais diversificados.

Em Freund e Schapire (1999) o BOOSTING é descrito como um método poderoso na combinação de classificadores de base múltipla para formar um conjunto cujo desempenho que pode ser significativamente melhor do que de qualquer um dos classificadores-base. A forma mais utilizada para potencializar o algoritmo é chamado AdaBoost. Com a utilização do BOOSTING, os classificadores-base são colocados em seqüência no processamento, de forma que aqueles mal classificados na base anterior recebem maior peso no processamento seguinte.

O algoritmo AdaBoost, introduzido por Freund e Schapire (1997), resolveu muitas das dificuldades de ordem prática de algoritmos anteriores. O algoritmo tem como entrada um conjunto de dados $(X_1, Y_1), \ldots, (X_m, Y_m)$, onde cada x_i tem uma instância no espaço X e cada y_i tem um rótulo no conjunto Y. O AdaBoost requer um determinado algoritmo ou uma base de aprendizagem que foi replicada em uma série de rodadas $t = 1, \ldots, T$.

Uma das idéias principais do algoritmo é manter uma distribuição ou um conjunto de pesos sobre o conjunto base. Estes pesos sofrerão ajustes nos próximos processamentos, como dito anteriormente.

Deve-se encontrar a hipótese nula $h_t : X \rightarrow \{-1, +1\}$ apropriada para a distribuição D_t. a adequação da hipótese nula é medida pelo erro

$$\epsilon_t = Pr_{i \sim D_t}[h_t(x_i) \neq y_i] = \sum_{i:h_t(x_i) \neq y_i} D_t(i).$$

O erro descrito é medido em relação a D_t, algoritmo em que são distribuídos os pesos.

- Dado: $(x_i, y_i), \ldots, (x_m, y_m)$, onde $x_i \in X, y_i \in Y = \{-1; +1\}$
- Iniciar: $D_1(i) = \frac{1}{m}$.
- Para: $t = 1, \ldots, T$:
- Escolha: $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$.
- Atualize: $D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{se } h_t(x_i) = y_i \\ e^{\alpha_t} & h_t(x_i) \neq y_i \end{cases}$

$$= \frac{D_t(i)\, exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

Onde Z_t é um fator de normalização (escolhido de modo que D_{t+1} será uma distribuição).

A hipótese final será:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right).$$

3. METODOLOGIA

O propósito do estudo foi identificar se as técnicas computacionais de árvore de decisão e de aprendizagem de máquina BAGGING e BOOSTING são mais efetivas que as técnicas tradicionais análise discriminante e regressão logística na classificação de risco de crédito de financiamentos imobiliários e as possíveis influências do crédito consignado nestas operações. Uma vez discutidas as definições e o contexto em que estão inseridos o crédito consignado, o financiamento imobiliário e o risco de crédito implícito, foram avaliados os resultados comparativos dos acertos dos modelos, além das variáveis selecionadas com maior poder explicativo.

Foi utilizada uma base de dados de operações de crédito. Foi necessário o tratamento de inconsistências nas informações da base e definidos os critérios utilizados para a classificação dos tomadores de crédito como bons ou maus pagadores. Foram definidos como maus pagadores os tomadores com operações em atraso a mais de 59 dias.

A base de dados foi, então, dividida para que se obtenha uma massa de validação e testes. A base de validação conteve as mesmas variáveis, contudo as proporções de bons e maus pagadores apresentaram variação por se tratar de período de 1 semestre subsequente à base de testes.

Percorridas estas etapas, iniciou-se a seleção das variáveis para os modelos e a verificação de premissas necessárias à aplicação das técnicas estatísticas. Com base nestas atividades, foram aplicados os testes para seleção dos modelos, discutidos na seção 3.2.

Embora a pesquisa envolva dados de somente uma instituição financeira, sua representatividade no cenário nacional e, particularmente, sua forte atuação tanto em crédito consignado quanto em financiamento imobiliário, permite fazer inferências no que diz respeito às conclusões do trabalho.

Foram realizados pré-testes em base de dados defasada em 12 meses, que demonstram diferenças de comportamento de crédito do financiamento imobiliário e demais operações de crédito entre os tomadores que possuem ou não crédito consignado. As evidências do pré-teste sugeriram uma investigação mais profunda desse relacionamento.

Neste contexto, a pesquisa envolveu o cruzamento de variáveis ligadas ao crédito consignado e ao financiamento imobiliário, constituindo um estudo pouco explorado no âmbito acadêmico.

3.1. A amostra utilizada

Conforme já discutido, buscou-se identificar variáveis que permitam classificar grupos de adimplentes e inadimplentes. As amostras foram extraídas das bases de dados de 7.052.429 (junho de 2009) e 6.903.461 (dezembro de 2008) operações de crédito, cedidas por uma instituição brasileira de forte presença no Brasil, que atua nos segmentos de financiamento imobiliário e crédito consignado. Serão obedecidos critérios de confidencialidade impostos pela instituição no que concerne às informações estratégicas e ao sigilo de clientes.

3.1.1. O tratamento das variáveis

Dentro desse universo de operações foram marcadas as operações em situação de atraso no pagamento por mais de 59 dias, caracterizando-as como inadimplentes ou maus pagadores. Para a análise discriminante, o critério de ‘bons’ ou ‘maus’ pagadores foi estabelecido para investigar a acuidade dos modelos e a segregação dos diferentes tipos de tomadores.

Para potencializar os resultados aferidos, foram eliminados os clientes em que as informações cadastrais não se apresentaram de maneira completa (0,1%). As operações classificadas como ‘perdas’ ou ‘baixadas do ativo’ pela Instituição não foram objeto de análise e também foram excluídas da base. Da mesma forma, os

valores de limites não utilizados, por não serem tratados como operações de crédito pela Instituição não compuseram a base de análise. Carteiras de crédito cedidas pela Instituição Financeira a outras, com e sem coobrigação, não foram objetos de análise (0,1%). As operações de tomadores com renegociação de dívida foram igualmente excluídas da base (1,1%). Para alguns tomadores (0,03%) a variável dependente 'renda mensal' apresentou mais de um valor. Quando um destes valores foi zero, este foi desconsiderado. Entre os demais, foi utilizada a média aritmética como referência.

Foram mantidas na base apenas as operações de crédito de pessoas físicas, por se tratarem do foco do estudo. Foram selecionadas apenas as operações de financiamento imobiliário e marcadas aquelas que também apresentam crédito consignado para o mesmo tomador, utilizando-se períodos diferentes, de modo que se tivesse uma amostra de calibração (período de dezembro de 2008) e outra de validação (período de junho de 2009). As eventuais variações no cenário econômico do País, bem como as estratégias políticas de incentivo ao financiamento imobiliário não foram objetivo deste estudo e, portanto, não foram avaliados. Cabe ressaltar que as mudanças no perfil dos tomadores de crédito não foram avaliadas, assim como as estratégias da Instituição Financeira que cedeu a base de dados para a pesquisa.

Após o tratamento, as bases apresentaram as quantidades de operações a seguir descritas:

Tipo	Calibração	Validação
Bons pagadores	13.773	17.989
Maus pagadores	1.108	1.267
Total de operações	**14.881**	**19.256**

TABELA 1: Operações de financiamento imobiliário sem crédito consignado
Fonte: Elaborada pelo autor

Tipo	Calibração	Validação
Bons pagadores	3.006	3.562
Maus pagadores	158	176
Total de operações	**3.164**	**3.738**

TABELA 2: Operações de financiamento imobiliário com crédito consignado
Fonte: Elaborada pelo autor

3.1.1.1. As Variáveis independentes de cadastro

Neste estudo, as variáveis independentes selecionadas com a finalidade de classificar os tomadores, sob a ótica da capacidade de pagamento, foram:

Variável de cadastro	Descrição	Tipo	
Data de nascimento	Data de nascimento do tomador da operação.	Discreta	Informado
Sexo	Sexo do tomador da operação.	Nominal	Informado
Escolaridade	Nível de escolaridade do tomador da operação.	Ordinal	Informado
Estado civil	Estado civil do tomador da operação.	Nominal	Informado
Dependentes	Quantidade de dependentes do tomador da operação.	Discreta	Informado
Servidor público	Identifica se o tomador é ou não funcionário público.	Nominal	Informado
Idade	Idade em anos do tomador da operação.	Contínua	Calculado
Tipo de documento	Código que identifica se o tomador é pessoa física ou jurídica.	Nominal	Informado
Documento	Número do CPF criptografado do tomador.	Nominal	Informado
Bairro	Bairro de residência do tomador.	Nominal	Informado
Município	Município de residência do tomador.	Nominal	Informado
CEP	CEP da residência do tomador.	Discreta	Informado
Complemento de CEP	Complemento do CEP da residência do tomador.	Discreta	Informado
Unidade federativa	Estado de residência do tomador.	Nominal	Informado
Sistema	Identifica o sistema onde a operação foi realizada.	Nominal	Informado
Contrato	Código do contrato da operação.	Nominal	Informado
Subcontrato	Código do subcontrato da operação	Nominal	Informado
Renda	Renda mensal do tomador da operação.	Contínua	Informado
Tipo de tomador	Identifica se o tomador é policial militar, funcionário público, universitário, aposentado ou cliente comum (não é funcionário público, universitário, policial militar nem aposentado).	Nominal	Informado

TABELA 3: Variáveis independentes de cadastro
Fonte: Elaborada pelo autor

3.1.1.2. Variáveis independentes de crédito

As variáveis independentes específicas do crédito consignado, do financiamento imobiliário e das demais operações selecionadas para este estudo foram:

Variável de contrato	Descrição	Tipo	
Produto	Código do produto de crédito.	Nominal	Informado
Data de inicio	Data em que a operação foi realizada.	Nominal	Informado
Data de término	Data prevista para término da operação.	Nominal	Informado
Taxa	Taxa de juros contratada para a operação.	Contínua	Informado
Saldo	Saldo devedor da operação no mês de referência.	Contínua	Informado

Provisão	Provisão para devedores duvidosos da operação no mês de referência.	Contínua	Informado
Rubrica	Rubrica contábil em que a operação foi registrada.	Nominal	Informado
Total de parcelas	Total de parcelas mensais da operação.	Discreta	Informado
Parcelas em aberto	Quantidade de parcelas não-pagas no mês de referência.	Discreta	Informado
Data de referência	Último dia do mês de referência.	Nominal	Informado
Tem consignado	Identifica se o tomador do financiamento imobiliário tem também empréstimo consignado.	Nominal	Apurado
Status de pagamento	Identifica o estágio da dívida.	Nominal	Informado
Valor do contrato	Valor nominal da operação realizada.	Contínua	Informado
Tipo de garantia	Identifica se a garantia da operação é hipoteca ou alienação.	Nominal	Informado
Consignado - data de inicio	Data em que a operação de empréstimo consignado foi realizada.	Nominal	Informado
Consignado - data de término	Data prevista para término da operação de empréstimo consignado.	Nominal	Informado
Consignado - total de parcelas	Total de parcelas mensais da operação de empréstimo consignado.	Discreta	Informado
Consignado - total de parcelas em aberto	Quantidade de parcelas de empréstimo consignado não-pagas no mês de referência.	Discreta	Informado
Consignado – saldo	Saldo devedor da operação de empréstimo consignado no mês de referência.	Contínua	Informado
Consignado - status de pagamento	Identifica o estágio da dívida de empréstimo consignado.	Nominal	Informado
Consignado - valor do contrato	Valor nominal da operação de empréstimo consignado realizada.	Contínua	Informado
Consignado - contrato sobre renda	Indica o quanto o valor do contrato representa da renda do tomador.	Contínua	Calculado

TABELA 4: Variáveis independentes de crédito
Fonte: Elaborada pelo autor

3.1.1.3. Variável dependente

A variável dependente deste estudo foi ‘Default’ no financiamento imobiliário que representará o evento de não pagamento ou de pagamento em dia das obrigações contratuais, conforme segue:

Variável	Descrição	Tipo	
Default	Identifica as operações sem pagamento a mais de 59 dias, a partir do status de pagamento.	Nominal	Apurado

TABELA 5: Variável dependente
Fonte: Elaborada pelo autor

3.1.2. Descrição e comportamento individual das variáveis

Variável	Descrição	Bom pagador	Mau pagador
Região	Grande SP	91,7%	8,3%
	Litoral e interior SP	93,5%	6,5%
Tem crédito consignado	Não	92,6%	7,4%
	Sim	95,0%	5,0%
Garantia	Hipoteca	88,7%	11,3%
	Alienação	95,5%	4,5%
Sexo	Feminino	95,1%	4,9%
	Masculino	91,6%	8,4%
Escolaridade	Alfabetizado	33,3%	66,7%
	Analfabeto	81,8%	18,2%
	Não-informado	92,4%	7,6%
	Graduação completa	95,0%	5,0%
	Graduação sequencial completa	96,2%	3,8%
	Graduação incompleta	91,8%	8,2%
	Graduação sequencial incompleta	97,9%	2,1%
	Ensino fundamental completo	90,1%	9,9%
	Ensino fundamental incompleto	88,8%	11,2%
	Ensino médio completo	92,3%	7,7%
	Ensino médio incompleto	88,7%	11,3%
	Especialização/Mestrado/Doutorado	97,1%	2,9%
Estado civil	Não-informado	70,8%	29,2%
	Casado	92,6%	7,4%
	Divorciado	93,7%	6,3%
	Separado judicialmente consensual	95,5%	4,5%
	Separado judicialmente	90,4%	9,6%
	Solteiro	92,6%	7,4%
	Uniao estável	96,6%	3,4%
	Viuvo	94,4%	5,6%
Nº de dependentes	0	93,0%	7,0%
	1	93,5%	6,5%
	2	93,5%	6,5%
	3	91,2%	8,8%
	4	91,9%	8,1%
	5	89,2%	10,8%
	6	90,0%	10,0%
	7	50,0%	50,0%
	8	50,0%	50,0%
	12	100,0%	0,0%
	16	100,0%	0,0%
Servidor público	Não	89,5%	10,5%
	Sim	96,4%	3,6%
Parcelas do contrato	<= 60,00	96,1%	3,9%
	61,00 - 120,00	98,7%	1,3%
	121,00 - 180,00	91,7%	8,3%
	181,00 - 240,00	93,6%	6,4%
	241,00 - 300,00	86,0%	14,0%
	301,00+	57,7%	42,3%

Renda	<= 1000	91,1%	8,9%
	1001 - 2000	92,4%	7,6%
	2001 - 3000	94,1%	5,9%
	3001 - 4000	94,2%	5,8%
	4001 - 5000	95,1%	4,9%
	5001 - 10000	96,3%	3,7%
	10001+	95,4%	4,6%
Idade	<= 25,00	85,7%	14,3%
	26,00 - 35,00	95,7%	4,3%
	36,00 - 45,00	94,1%	5,9%
	46,00 - 55,00	91,4%	8,6%
	56,00 - 65,00	91,9%	8,1%
	66,00+	91,0%	9,0%

TABELA 6: Frequência das variáveis por grupo
Fonte: Elaborada pelo autor

3.2. Estudos referenciais

3.2.1. Seleção de variáveis

O risco de crédito é o maior risco de grandes bancos comerciais, portanto a medição e gestão de risco de crédito é sua tarefa central. Para Tang *et al.* (2009), criar modelos que possam medir eficazmente a inadimplência é essencial para esta gestão. A métrica do NPL (*non performing loan*) é uma informação que pode ajudar os bancos comerciais a administrar seu risco de crédito, com a finalidade de alocar mais recursos para ativos com valorização positiva e evitar o desperdício de dinheiro, reduzindo custos e melhorando a eficiência da gestão financeira. No entanto, a inadimplência nos bancos comerciais da China, onde o estudo foi realizado, tem suas características próprias. Os fatores que afetam o NPL também podem ser diferentes em outros Países. Portanto, segundo Tang *et al.* (2009) é essencial desenvolver modelos específicos de acordo com a característica de cada País.

Ye e Liu (2006) realizaram uma pesquisa empírica com os dados de grandes bancos. Os autores analisaram as características de inadimplência em termos de prazos de vencimento, forma de garantia, característica de tomador, entre outras. A

limitação dos autores, entretanto foi o pequeno tamanho da amostra, que poderia não refletir a representatividade desejada. Esta limitação foi superada neste estudo.

Altman *et al.* (2001) analisaram o impacto de várias suposições sobre a associação entre a probabilidade de inadimplência e a perda dos empréstimos bancários, procurando explicar, empiricamente, este relacionamento. O autor afirma que altos índices de inadimplência (NPL) têm forte correlação negativa com a taxa de recuperação de crédito.

Tang *et al.* (2009) também exploraram as variáveis associadas ao NPL dividindo-as em dois grupos: aquelas associadas ao contrato e as demais associadas ao tomador. No que diz respeito à classificação da dívida, o autor estudou sete grupos distintos: normal, dificuldades financeiras, deixando de negócios, fechando para baixo, o subprocesso de insolvência, falência e desconhecidos. Neste estudo, utilizou-se três grupos: normal (inclui atrasos até 14 dias), atrasos de 15 a 59 dias e superiores a 59 dias (NPL).

Para evitar a utilização de uma quantidade muito grande de variáveis, o que poderia interferir no efeito da discriminação, Tang *et al.* (2009) empregaram o método *stepwise* para selecionar as significativas. Uma das variáveis empregadas no modelo de Tang *et al.* (2009) foi a existência de garantia de hipoteca, que se mostrou significativa no método *stepwise*. Outra informação significativa foi a região geográfica dos tomadores dos empréstimos. Segundo os autores, as condições econômicas na região certamente influenciam na recuperação de inadimplência. De um modo geral, quanto mais desenvolvida a sua economia, maior é a probabilidade de recuperação na hipótese de não-pagamento.

Muitos fatores levam à inadimplência de hipotecas. As características regionais estão entre as mais importantes (QUERCIA e STEGMAN, 1992). O estudo de Lim *et al.* (2000), além de características regionais, utiliza renda familiar, nível de escolaridade e custo médio da habitação para prever inadimplência em financiamentos hipotecários.

Hand e Henley (1997) realizaram uma revisão dos problemas específicos que surgem no contexto de credit scoring para bancos e examinaram os métodos estatísticos usualmente aplicados em análises deste tipo.

Para este estudo, utilizaram como variáveis independentes o tempo de residência (0-1, 1-2, 3-4, mais de 5 anos), CEP, um indicador de posse de cartões de crédito, um indicador de poupança, faixa etária (18-25, 26-40, 41-55, mais de 55 anos), estado civil (casado, divorciado, solteiro, viúvo, outro), tempo de relacionamento com o banco em anos, tempo de serviço em anos.

3.2.2. Seleção das técnicas quantitativas

Segundo Press e Wilson (1978) uma análise ou classificação discriminante, classifica uma observação em uma das diversas populações que se pode identificar. Quando se relaciona as variáveis qualitativas com outras variáveis está se fazendo regressão logística. Os estimadores gerados por quaisquer destas técnicas geralmente são utilizadas na outra. Se as populações apresentam distribuições normais, com covariâncias idênticas, estimadores de análise discriminante são os preferidos. Na maioria das aplicações de análise discriminante, no entanto, pelo menos uma variável é qualitativa. Dependendo da não-normalidade identificada, prefere-se os modelos de regressão logística, com estimadores de máxima verossimilhança para resolver ambos os problemas.

3.2.2.1. Análise discriminante

A utilização de análise discriminante em estudos de desempenho em bancos é recorrentemente verificada. Taffler (1982) utilizou a análise discriminante para realizar estudos de previsão de falências no Reino Unido. Ele derivou os quatro indicadores financeiros que se mostraram mais significantes estatisticamente e atribuiu-lhes pesos diferentes, adequados à sua importância. Os escores encontrados representavam maior ou menor tendência ao default.

O trabalho pioneiro de Altman (1968), usando dados de sessenta e seis empresas americanas entre 1946 a 1965, deu início a diversos estudos que utilizam a análise discriminante para previsão de falência ou inadimplência. Trabalhos realizados por Kanitz (1978), Taffler (1982), Micha (1984), Altman (2005), Altman e Sabato (2007) são alguns exemplos de aplicação da análise discriminante para a análise de crédito em diversos contextos e países, incluindo Brasil.

Tang *et al.* (2009) utilizaram o método de análise discriminante para classificar a recuperação de crédito. Para os autores, a idéia básica da análise discriminante é supor que o sujeito da pesquisa pode ser classificado em várias categorias e existem dados de observação de amostras contidas em cada categoria. Partindo deste pressuposto, podemos estabelecer funções discriminantes de acordo com alguns critérios, e usá-la para classificar as amostras com a categoria desconhecida. Há diversas variações de análise discriminante, como Fisher, Sequencial, Bayesiana e Stepwise (BRUCH, 1975).

3.2.2.2. Regressão logística

Wiginton (1980) afirma que há décadas existe um interesse considerável na utilização de modelos quantitativos do comportamento do crédito ao consumo para as decisões de concessão de crédito. A maioria dos modelos são baseados no conceito de "scoring" pelo uso de pesos geralmente determinados como coeficientes estatisticamente significativos de algum modelo estatístico linear, em geral a análise discriminante. A estimativa de máxima verossimilhança dos modelos de regressão logística surge como uma alternativa que tem o objetivo de pontuar as classificações.

A aplicabilidade da regressão logística na análise de crédito é evidenciada pelos diversos estudos já conduzidos tanto no exterior quanto no Brasil como, por exemplo, os de Bensic, Sarlija e Zekic-Susac (2006, p.??) e inussi, Damacena e Ness Jr. (2002).

3.2.2.3. BAGGING e BOOSTING

Creamer e Freund (2004) utilizaram a abordagem BOOSTING em regressão logística para quantificar o risco de governança corporativa e Mercados latino-americanos.

O estudo objetivou a construção de um modelo preditivo para avaliar se o desempenho de uma companhia ou eficiência de um banco está acima ou abaixo da média de seus equivalentes de mercado, em função dos principais fatores de governança corporativa e de indicadores contábeis selecionados, tidos como importantes na avaliação das empresas.

A motivação foi a premissa de que os aspectos regulamentares dos países latino-americanos são incipientes, ocasionando problemas de agência que podem comprometer o desempenho das organizações estudadas.

O algorítmo utilizado foi o Adaboost (FREUND e SCHAPIRE, 1997), responsável pelo processo de ‘aprendizagem’ do modelo de previsão proposto. Outro objetivo foi avaliar o uso de Adaboost como uma ferramenta de previsão e interpretativa para o risco de governança das empresas.

Os resultados do Adaboost utilizado foram comparados com regressão logística simples e com o BAGGING. Foram realizados experimentos de validação cruzada em uma amostra de ADRs. A principal observação foi que, para um conjunto de dados uniformes (de mesma fonte e com empresas de características similares), os resultados do BOOSTING são similares àqueles das outras técnicas comparadas. A aplicação do BOOSTING se mostrou mais eficaz diante da não-uniforidade dos dados.

A aplicação do Adaboost possibilitou melhor identificação de relação entre as variáveis que determinam desepenho e eficiência das empresas analisadas na amostra.

Creamer e Freund (2004) observaram que, inicialmente, as variáveis de governança corporativa não parecem ser muito relevantes para predizer o desempenho corporativo. No entanto, quando os resultados dessas variáveis foram

interpretados em conjunto com as variáveis contábeis com o uso árvores de decisão, os efeitos da governança corporativa sobre o desempenho tornou-se evidente. Os recentes casos de falências dos EUA demonstram que, quando as empresas estão bem, as variáveis de governança corporativa não parecem ser relevantes. Contudo, em momentos de dificuldades financeiras, tais variáveis podem reprentar com eficácia a medida de desempenho e eficiência.

O estudo de Sabzevari *et al.* (2009) analisou o desempenho de diferentes modelos de escoragem para crédito e uma base de dados de operações bancárias. O objetivo foi comparar as aborgens tradicionais (probit e regressão logística) com modelos de mineração de dados (árvores de classificação e BAGGING).

Sabzevari *et al.* (2009) relatam que os métodos de pontuação são amplamente utilizados para operações de empréstimos ao consumidor e está se tornando cada vez mais comum no processo de avaliação de financiamentos imobiliários, garantindo um melhor fluxo de caixa às Instituições que emprestam e significativa redução do risco de crédito. As técnicas de modelagem analisadas foram desenvolvidas com a finalidade de potencializar a utlização das técnicas de escoragem. Segundo Sabzevari *et al.* (2009), os métodos estatísticos tradicionais *Probit* e *Logit*, além dos métodos de mineração de dados (árvores de decisão, redes neurais e BAGGING, por exemplo, devido às suas características de associação de memória e capacidade de generalização) estão se tornando muito populares em modelos de *credit scoring*.

No entanto, o Data Mining também está sendo criticado por seu longo processo de formação da base de dados, capacidade limitada de identificar a importância relativa de variáveis e algumas dificuldades de interpretação dos resultados. Um dos desafios que os pesquisadores enfrentam quando usam algoritmos de classificação para a construção de modelos de escoragem de crédito é disponibilidade de informação. As bases de dados utilizadas na modelagem não só contêm muitas observações, como também um grande número de recursos. Algumas das características podem ser irrelevantes para classificar o risco de crédito, alguns dados podem ser redundantes devido à sua alta correlação. Estes

problemas, segundo os autores, aumentam o esforço computacional e reduzem substancialmente a precisão dos modelos.

Para demonstrar a eficácia do credit scoring, foram utilizadas operações de clientes bancários. A principal observação mostra que a abordagem baseada no BAGGING e os modelos de *logit* apresentam um desempenho melhor, em relação às demais abordagens de pontuação. Em especial, o BAGGING é demonstrado com melhor desempenho se combinado à regressão logística, aumentando a precisão dos modelos.

O estudo de Zhang *et al.* (2010) propõem um modelo de avaliação de crédito no mercado consumidor chinês, mais especificamente para cartões de crédito. Neste modelo, utiliza o método de árvore de decisão BAGGING. A proposição é compor um modelo de classificadores agregados que combinam atributos preditivos. Tais classificadores, que possuem os mesmos atributos, são “treinados” a cada amostra extraída.

Zhang *et al.* (2010) mostram que o modelo foi testado em duas bases de dados de crédito do UCI Machine Learning Repository, e que os resultados das análises mostram que o desempenho do método proposto revela precisão na previsão de inadimplência.

Tanto as técnicas estatísticas quanto de inteligência artificial têm sido exploradas para a atividade financeira de credit scoring (WANG *et al.*, 2010). Embora não haja conclusões consistentes sobre a melhor técnica, estudos recentes sugerem que a combinação de classificadores múltiplos, ou seja, o aprendizado conjunto, pode ter um melhor desempenho que qualquer delas separadamente.

Wang *et al.* (2010) avalia comparativamente o desempenho de três métodos: BAGGING, BOOSTING e stacking, com base na base de quatro classificadores: regressão logística, árvore de decisão, redes neurais artificiais e máquinas de suporte vetorial. Os resultados experimentais revelam que os três métodos conjuntos podem melhorar substancialmente a aprendizagem individual da base de dados. Em particular, o BAGGING apresenta a melhor performance preditiva.

3.3. Variáveis selecionadas

A seleção das variáveis para desenvolvimento dos modelos levou em consideração os referenciais teóricos de estudos realizados, citados na seção 3.3; a significância estatística apresentada no Teste t de igualdade de média entre grupos de bons e maus pagadores; a estatística Levene para verificação da igualdade de variância; a correlação entre os preditores e os escores discriminantes; e a obtenção de melhores percentuais de acertos nos modelos de previsão.

A tabela 7, a seguir, apresenta dois testes estatísticos com suas respectivas significâncias, utilizadas como um dos critérios de seleção de variáveis. As interpretações são as seguintes:

- Os modelos de análise discriminante partem da premissa de que as variâncias são iguais. A significância do Teste Levene testou esta igualdade a uma significância de 5%. A 'escolaridade', os estados civis 'divorciado', 'solteiro' e 'viúvo', e os tipos de clientes 'universitário', 'diretor de cartório' e 'diretor de estatal' apresentaram significância maior que o referencial de 5%.

- O Teste t verificou a igualdade de médias entre os grupos de bons e maus pagadores a uma significância de 5%. A hipótese nula de médias iguais não pode ser rejeitada nos estados civis 'divorciado', 'separação judicial', 'solteiro' e 'viúvo', e nos tipos de clientes 'universitário', 'procurador', 'promotor', 'diretor de cartório', 'delegado' e 'diretor de estatal'.

Variável	Teste Levene para igualdade de variância	Teste t para igualdade de média
Nº de parcelas	0,000	0,000
Tem consignado? (0/1 ; s/n)	0,000	0,000
Valor do contrato	0,000	0,000
Renda	0,000	0,000
Garantia de alienação (0/1 ; s/n)	0,000	0,000
Sexo	0,000	0,000
Nº de dependentes	0,001	0,130
Funcionário público (0/1 ; s/n)	0,000	0,000
Idade (0/1 ; s/n)	0,007	0,000
Escolaridade	0,741	0,000
Casado (0/1 ; s/n)	0,000	0,024
Divorciado (0/1 ; s/n)	0,081	0,386
Separação judicial cons. (0/1 ; s/n)	0,000	0,016

Separação judicial (0/1 ; s/n)	0,000	0,056
Solteiro (0/1 ; s/n)	0,033	0,277
União estável (0/1 ; s/n)	0,000	0,000
Viúvo (0/1 ; s/n)	0,006	0,168
Universitário (0/1 ; s/n)	0,120	0,437
Funcionário público geral (0/1 ; s/n)	0,000	0,000
Procurador (0/1 ; s/n)	0,016	0,231
Juiz (0/1 ; s/n)	0,000	0,017
Promotor (0/1 ; s/n)	0,012	0,207
Diretor de cartório (0/1 ; s/n)	0,178	0,501
Delegado (0/1 ; s/n)	0,001	0,094
Diretor de estatal (0/1 ; s/n)	0,219	0,539
Policial com patente (0/1 ; s/n)	0,000	0,016
Policial militar (0/1 ; s/n)	0,000	0,000
Cliente comum (0/1 ; s/n)	0,000	0,000
Funcionário público sem folha (0/1 ; s/n)	0,000	0,042
Aposentado (0/1 ; s/n)	0,000	0,000

TABELA 7: Testes estatísticos
Fonte: Elaborada pelo autor

A partir da verificação dos critérios estabelecidos, foram selecionados os grupos de variáveis que segue:

Variável	Teste T	Estatística Levene*	Estudos referencias	Melhoria na previsão	Correlações*, **
Nº de parcelas	recomendada	recomendada	não-identificado	recomendada	alta
Tem consignado?	recomendada	recomendada	não-identificado	recomendada	fraca
Renda	recomendada	recomendada	Lim et al. (2000)	recomendada	média
Garantia de alienação	recomendada	recomendada	Ye e Liu (2006) Tang et al. (2009)	recomendada	média
Sexo	recomendada	recomendada	Krivo et al. (1998)	recomendada	média
Nº de dependentes	não-recomendada	recomendada	não-identificado	recomendada	fraca
Idade	recomendada	recomendada	Hand e Henley (1997)	recomendada	média
Escolaridade	recomendada	não-recomendada	Lim et al. (2000)	recomendada	fraca
Estado civil	recomendada	recomendada	Hand e Henley (1997)	recomendada	fraca
Tipo de tomador	recomendada	recomendada	Ye e Liu (2006)	recomendada	média

* Premissa apenas para análise discriminante.
** Alta: >=50%; Média: >=15% e <50%; Fraca: <15%

TABELA 8: Aspectos considerados para seleção de variáveis
Fonte: Elaborada pelo autor

A tabela a seguir apresenta as estatísticas descritivas das variáveis selecionadas. Maiores quantidade de parcelas mostram, em média, maior propensão ao não-pagamento, assim como as menores médias de renda. Com relação à garantia de alienação, quando ausente (zero no caso de a garantia ser hipoteca), sugere maior propensão ao inadimplemento, assim como a média de quantidade de dependentes maior. Também os tomadores com menores médias de idade aparentam ser melhores pagadores, assim como aqueles com melhor escolaridade.

Variável		Bom pagador				Mau pagador			
		Média	Desvio-padrão	Mínimo	Máximo	Média	Desvio-padrão	Mínimo	Máximo
Nº de parcelas		182,1	80,4	8,0	360,0	239,9	67,1	54,0	360,0
Tem consignado?		0,2	0,4	0	1,0	0,1	0,3	0	1,0
Renda		2.630,8	3.763,2	0	116.000,0	2.020,5	2.923,5	0	32.325,2
Garantia de alienação		0,7	0,5	0	1,0	0,4	0,5	0	1,0
Sexo		0,6	0,5	0	1,0	0,7	0,5	0	1,0
Nº de dependentes		0,7	1,1	0	16,0	0,7	1,2	0	8,0
Idade		46,6	10,9	1,0	85,0	49,1	10,7	23,0	86,0
Escolaridade		6,5	3,8	0	10,0	6,0	3,8	0	10,0
Estado civil	Solteiro	0,2	0,4	0	1,0	0,2	0,4	0	1,0
	União estável	0,1	0,3	0	1,0	0,0	0,2	0	1,0
Tipo de tomador	Policial militar	0,0	0,1	0	1,0	0	0,1	0	1,0
	Funcionário público sem folha	0,0	0,2	0	1,0	0,0	0,2	0	1,0
	Cliente comum	0,5	0,5	0	1,0	0,7	0,4	0	1,0
	Aposentado	0,0	0,2	0	1,0	0,1	0,2	0	1,0

TABELA 9: Estatística descritiva das variáveis
Fonte: Elaborada pelo autor

4. EXCLUSÕES

Este estudo não teve por objetivo discutir os custos das eventuais classificações incorretas, sejam elas de bons ou maus pagadores. Entende-se que a detecção de bons ou maus pagadores, e sua priorização na escolha das variáveis e parâmetros para a formulação dos modelos de concessão de crédito, passa pela questão estratégica e situacional, variando de instituição para instituição.

Com relação ao ponto de corte definido para os escores produzidos, entende-se que a decisão sobre este tipo de limite apresenta características estratégicas em termos de exposição a risco de crédito, e não acadêmicas não sendo, portanto, objeto deste estudo.

A rentabilidade das operações não foi considerada e não foi critério de seleção de quaisquer dos parâmetros utilizados.

Não figuram entre as operações de crédito analisadas aquelas rejeitadas pela Instituição, fato que cria um viés natural no que diz respeito à ausência potencial de maus pagadores.

5. RESULTADOS OBTIDOS

Selecionadas as variáveis com maior poder explicativo e executados os modelos, foram obtidos os resultados apresentados nas subseções a seguir, assim como estão descritos os testes realizados para verificar a acuidade a precisão das estimações.

5.1. Análise discriminante

Para verificar a premissa de igualdade da matriz de covariância foi utilizado o teste F. A significância apresentada foi menor que 1%, sugerindo rejeição da hipótese nula de igualdade entre as matrizes. A premissa é de que sejam iguais, então, neste modelo, a premissa não seria atendida. Contudo, o problema é amenizado pela grande quantidade de observações da amostra utilizada.

Foi observada a correlação entre os coeficientes discriminantes e a previsão de mau pagador proposta pelo modelo por meio da estatística de correlação canônica Eigenvalue. Tal estatística apresentou o valor de 27,4%, comparativamente a melhor entre os modelos de análise discriminante testados.

A estatística Lambda de Wilks apresentou significância inferior a 1%, sugerindo a rejeição da hipótese nula de a média entre os adimplentes e inadimplentes ser igual. Esta estatística mostra a proporção da variância não explicada pelo modelo, e apresentou o valor de 0,925.

A tabela 10 mostra o impacto relativo das variáveis selecionadas no poder explicativo do modelo:

Impacto relativo dos coeficientes da função discriminante	
Nº de parcelas	0,722
Tem consignado? (0/1 ; s/n)	0,097
Renda	-0,002
Garantia de alienação (0/1 ; s/n)	-0,395
Sexo	0,081

Nº de dependentes	0,075
Idade	0,190
Escolaridade	0,209
Solteiro	0,047
União estável	-0,075
Funcionário público sem folha de pgto	0,158
Policial militar	-0,034
Cliente comum	0,622
Aposentado	0,127

TABELA 10: Impacto relativo das variáveis do modelo
Fonte: Elaborada pelo autor

Os coeficientes discriminantes apresentados mostram que a quantidade de parcelas dos contratos de financiamento imobiliário contribui de forma mais forte para o modelo. A variável *dummy* de tipo de cliente comum também apresenta contribuição forte em relação às demais.

Também foi calculada a correlação entre os coeficientes discriminantes e os preditores utilizados. Novamente, a quantidade de parcelas apresentou a maior correlação (65%), seguida por tipo de cliente comum (57%) e garantia de alienação (-45%).

O nível geral de acerto do modelo de análise discriminante foi 74,9%. Para o grupo de bons pagadores, o modelo previu corretamente 75,3% dos eventos. Já para os maus pagadores o nível de acerto foi menor, representando 70,7% dos casos.

Para um tomador com crédito consignado, a correlação entre a variável e o score discriminante é -12,8%. Representou um poder explicativo intermediário com relação às demais variáveis utilizadas, ou seja, não é a que mais explicou nem a que menos explicou.

5.2. Regressão logística

O teste Qui-quadrado do modelo apresentou significância abaixo de 1%, sugerindo rejeição da hipótese nula de que os parâmetros são iguais a zero. Desta forma, ao menos um dos parâmetros do modelo é explicativo.

Qui-quadrado é um teste de hipóteses utilizado para encontrar um valor de dispersão para duas variáveis nominais, comparando proporções e as possíveis divergências entre frequência observada e esperada. Verifica o relacionamento entre variáveis qualitativas. O teste é do tipo não-paramétrico, portanto não depende de parâmetros populacionais. Se a hipótese nula não fosse rejeitada, significaria que os bons e maus pagadores se comportam de forma semelhante.

A estatística de Nagelkerke apresentou o valor de 18%. O objetivo desta estatística é explicar o poder explanatório do modelo. Em Regressão Logística, é similar ao Coeficiente de Determinação em Regressão Linear (HAIR JR. *et al.*, 1998).

Outro teste verificado foi o Hosmer/Lemeshow, que utiliza-se de distribuição Qui-quadrado para examinar o ajuste entre eventos esperados e observados (HAIR JR. *et al.*, 1998). Foi desenvolvido especificamente para avaliação de ajustes em Regressão Logística. Neste caso, a hipótese nula (observado = predito) foi rejeitada com significância inferior a 1%. Contudo trata-se de uma verificação que, isoladamente, não é conclusiva.

A tabela a seguir mostra a razão que mostra os efeitos de cada variável selecionada em termos de probabilidades:

Variável	Significância	Exp(B)
Nº de parcelas	,000	1,009
Tem consignado? (0/1 ; s/n)	,000	2,188
Renda	,327	1,000
Garantia de alienação (0/1 ; s/n)	,000	,572
Sexo	,079	1,134
Nº de dependentes	,002	1,091
Idade	,000	1,013
Escolaridade	,000	1,059
Solteiro	,080	1,146

União estável	,009	,656
Funcionário público sem folha de pgto	,000	3,935
Policial militar	,154	,481
Cliente comum	,000	6,326
Aposentado	,000	3,899

TABELA 11: Regressão logística – significância e fatores exponenciais
Fonte: Elaborada pelo autor

Destacam-se três variáveis que classificam tipos de tomadores: cliente comum (não é funcionário público, policial militar nem aposentado), funcionário público sem folha de pagamento (não recebe seu pagamento pela Instituição) e aposentado. Tais variáveis foram obtidas por variáveis *dummies*. A interpretação do fator exponencial mostra que, quando se trata de um cliente comum, a chance de o tomador se tornar inadimplente (ou de ser um mau pagador), é 6,33 maior do que se não for um cliente comum.

Os fatores exponenciais menores que 1, representam redução de probabilidade de maus pagadores. Por exemplo, a presença de garantia de alienação (em detrimento de hipotecas, que são os dois tipos de garantias para financiamentos imobiliários), reduz a probabilidade de que um tomador seja mau pagador. O mesmo ocorre quando o tomador é um policial militar. Como este tipo de tomador, em geral, recebe seu salário pela Instituição, faz sentido a redução de probabilidade de inadimplência.

O nível geral de acerto do modelo de regressão logística foi 71,9%. Para o grupo de bons pagadores, o modelo previu corretamente 71,7% dos eventos. Já para os maus pagadores o nível de acerto foi maior, representando 74,6% dos casos.

Para um tomador com crédito consignado, a chance se ser um mau pagador é 2,19 maior do que se este tomador não tivesse um crédito consignado.

5.3. Árvore de decisão

As árvores de decisão foram construídas utilizando o recurso de partição recursiva binária. O termo “binário” indica que as variáveis são divididas são divididas em duas quando é identificada uma diferença de comportamento que possa aumentar o poder preditivo (BREIMAN *et al.*, 1984). Estas divisões, chamadas “nós” se repetem enquanto for identificada uma quebra que conduza a uma melhor predição. A decisão sobre a melhor árvore a ser utilizada passa pelo critério de esforço de processamento além, evidentemente, do poder preditivo. Árvores com menores quantidades de nós, portanto mais simples sob a ótica computacional, requerem menor tempo de processamento. Contudo, árvores mais complexas tendem a ter melhor poder preditivo, embora requeiram maior esforço computacional.

Para a verificação do poder explicativo das variáveis selecionadas para a estruturação da árvore de decisão, utilizou-se a tabela a seguir:

Importância relativa	8 nós		Melhor custo (357 nós)		1.097 nós	
	classificação	score relativo	classificação	score relativo	classificação	score relativo
Servidor público	1	100	1	100	2	99,34
Parcelas do contrato	2	84,66	2	97,53	1	100
Garantia	3	74,85	3	78,55	5	74,52
Renda	4	40,87	4	72,46	3	92,29
Escolaridade	5	29,72	5	60,59	4	79,21
Idade	6	6,26	6	34,02	6	54,74
Nº de dependentes	7	2,88	7	25,74	7	44,09
Sexo	8	1,86	9	11,16	8	22,36
Tem crédito consignado	9	1,85	10	9,4	10	12,47
Região	10	0,93	8	11,39	9	16,97

TABELA 12: Árvore de decisão – importância relativa das variáveis
Fonte: Elaborada pelo autor

Observa-se na tabela 12 que as variáveis “servidor público”, que indica se o tomador da operação é ou não servidor público, e “parcelas do contrato”, são aquelas que têm maior poder preditivo relativo tanto na árvore com 8 nós, portanto mais simples, quanto na árvore de menor custo relativo (357 nós) e também na

árvore com 1.097 nós, que foi a mais complexa indicada pelo CART. As variáveis “renda” e “garantia” também figuram entre as mais importantes nas 3 observações.

A presença do crédito consignado entre as operações dos tomadores apresentou uma das menores representatividades para todos os cenários explorados para o poder preditivo do modelo.

A Árvore a seguir foi construída com a mesma base de dados utilizada nas técnicas de regressão logística e análise discriminante. A figura foi extraída do CART após o processamento. A linha verde na altura dos 357 nós representa a Árvore de menor custo (precisão versus processamento) atribuído pelo CART.

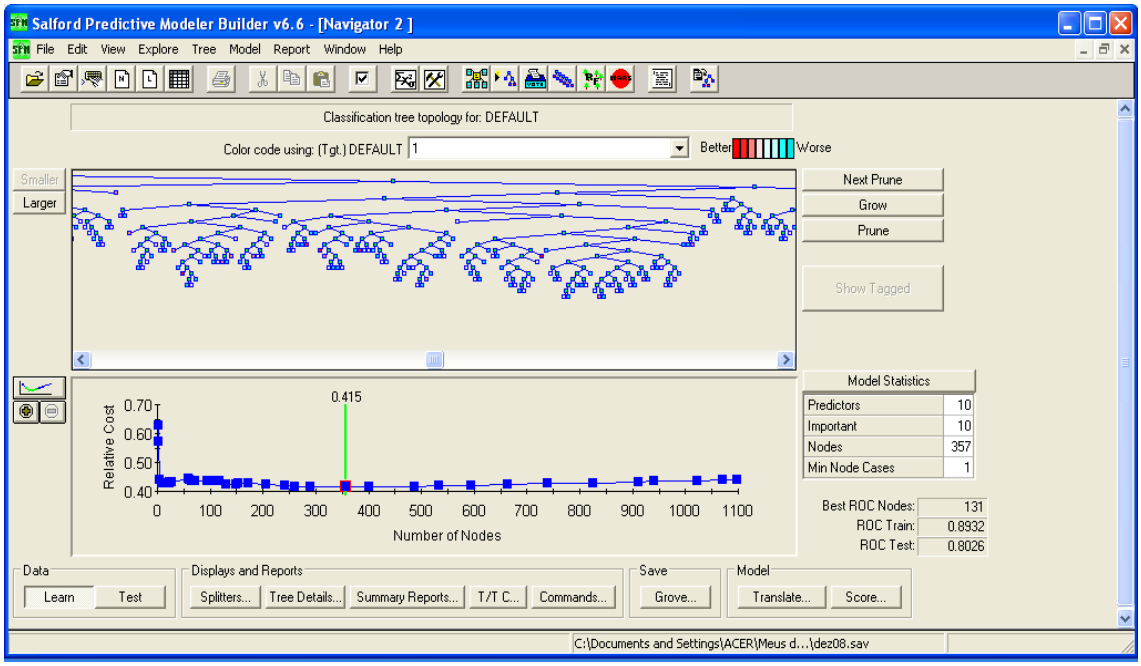

GRÁFICO 2: Árvore de decisão – seleção de nós
Fonte: CART

Como pode ser observada na tabela a seguir, a predição evolui gradativamente na medida em que é intensificada a complexidade da árvore produzida pelo CART.

Parâmetros	Base de desenvolvimento				
Nº de nós	% bons	% maus	qt total	% do modelo	Custo
4	68,2%	83,0%	18.045	69,2%	0,439
24	66,0%	89,0%	18.045	67,6%	0,432
71	71,0%	87,8%	18.045	72,2%	0,435
149	73,7%	89,3%	18.045	74,8%	0,426
260	74,6%	92,5%	18.045	75,9%	0,417
357	75,8%	**94,2%**	18.045	77,1%	0,415
488	78,2%	95,2%	18.045	79,4%	0,416
826	80,7%	98,5%	18.045	81,9%	0,429
940	81,7%	98,7%	18.045	82,9%	0,436
1097	81,9%	99,1%	18.045	83,1%	0,439

TABELA 13: Árvore de decisão – resultados por nós
Fonte: Elaborada pelo autor

O campo marcado, com 357 nós, é indicado pelo CART como sendo a configuração de menor custo relativo, conforme também pode ser verificado no gráfico 2. Isto significa que o esforço de processamento ponderado pelo nível de acerto do modelo tem melhor equacionamento neste nível de detalhamento. O processamento foi reproduzido na base de testes e apresentou resultados semelhantes.

A seguir são dispostos os gráficos de maus e bons pagadores, além do resultado agregado dos modelos, evidenciando os comportamentos preditivos da base de desenvolvimento aplicados à base de testes.

Com relação aos maus pagadores, verifica-se no gráfico 3 que há uma tendência de maiores acertos nas maiores quantidade de iterações de Árvores.

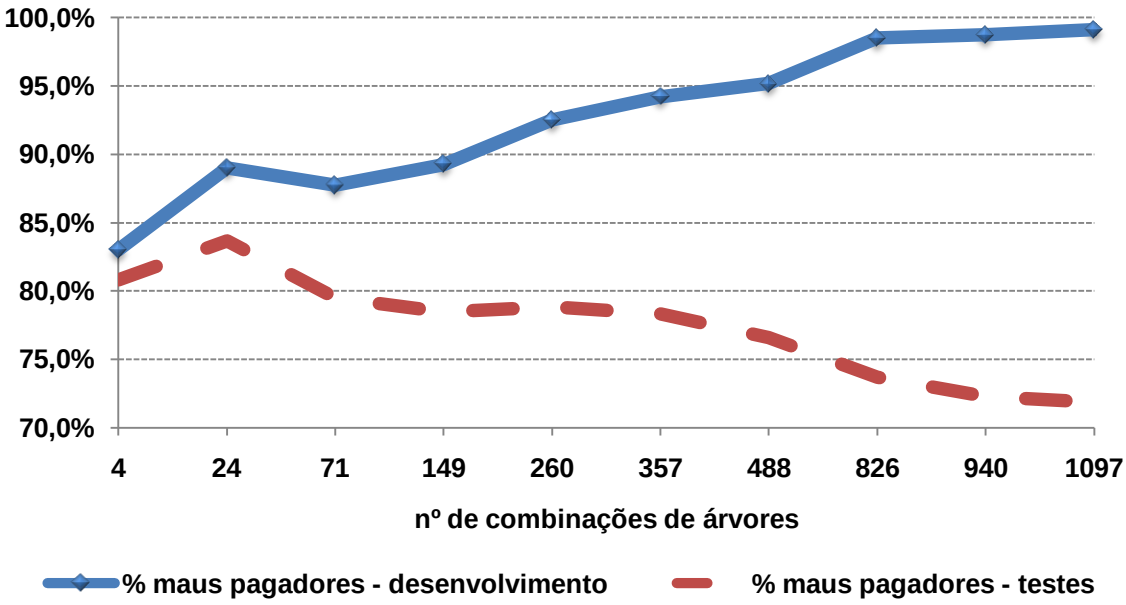

GRÁFICO 3: Árvore de decisão – maus pagadores
Fonte: Elaborado pelo autor

Contudo, na base de dados de testes, observa-se que a partir de 24 iterações a tendência se mostra invertida, ou seja, maiores quantidades de iterações levam à redução no número de acertos.

Para os bons pagadores, as Árvores de Decisão a partir apresentam melhora no poder preditivo a partir de 24 iterações, tendência esta que se confirma na base de testes, conforme mostrado no gráfico a seguir:

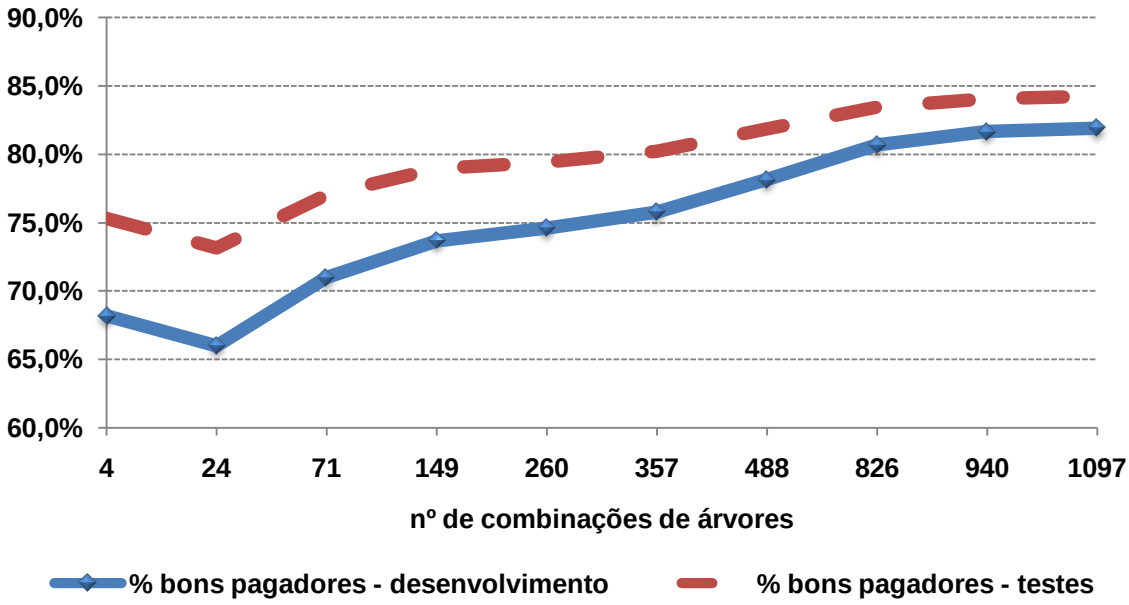

GRÁFICO 4: Árvore de decisão – bons pagadores
Fonte: Elaborado pelo autor

O CART construiu o número máximo de 1097 árvores, sugerindo que, a partir desta quantidade, não há ganhos em termos de predição.

No gráfico a seguir, que mostra o comportamento preditivo do modelo consolidado, observa-se a mesma tendência dos bons pagadores, ou seja, melhor poder preditivo a partir de 24 iterações. A validação dos resultados na base de testes mostra maior convergência com os resultados da base de desenvolvimento na medida em que se aumenta a quantidade de iterações.

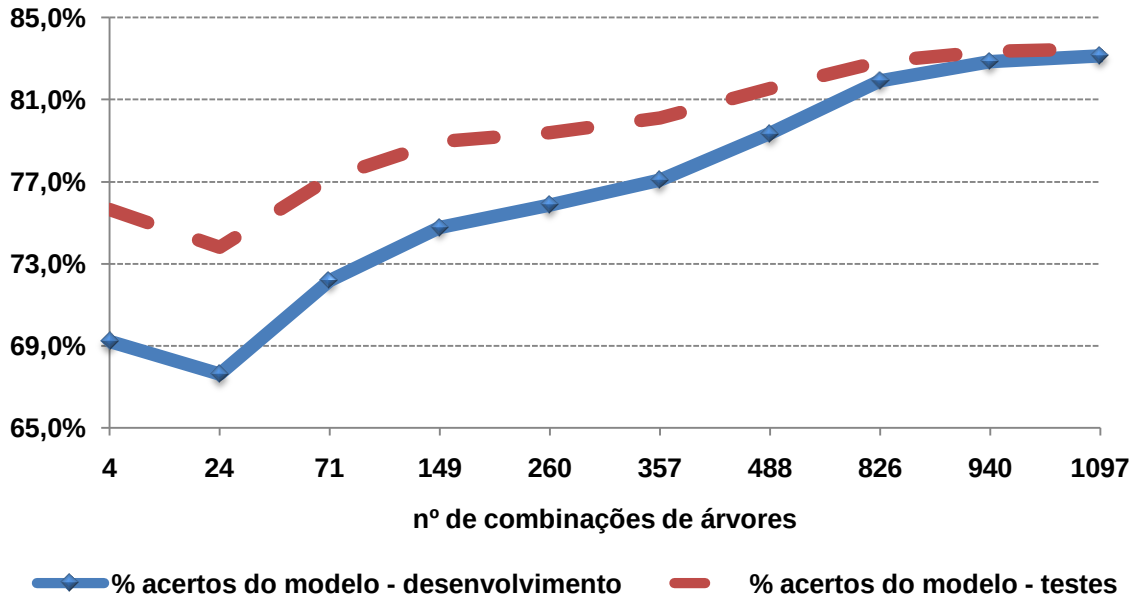

GRÁFICO 5: Árvore de decisão – modelo consolidado
Fonte: Elaborado pelo autor

Para efeitos comparativos, apresentados na seção 5.4, foi considerada a opção de 357 nós, onde se previu corretamente 77,1% das observações. Dos bons pagadores, 75,8% foram previstos corretamente enquanto dos maus pagadores se previu 94,2%.

5.3.1. BAGGING

A combinação de árvores construídas na técnica BAGGING é que construirá o voto majoritário que é a principal característica de classificação. Para Breiman *et al.*(1984), o BAGGING normalmente apresenta bons resultando utilizando por volta de 100 árvores. Este estudo partiu da comparação dos resultados de 10 árvores combinadas para se aferir qual combinação traria a melhor predição. Foram processadas novas combinações para a verificação da curva ótima de aproveitamento.

Também foi testado o poder preditivo dos modelos diante da variação de técnicas disponíveis na ferramenta de construção de árvores de decisão CART. Uma

delas foi a validação cruzada. Não é recomendada uma quantidade menor que 10 para a validação cruzada (BREIMAN *et al.*, 1984). Neste estudo utilizou-se a comparação entre 0, 5, 10, 20, 40, 50, 70 e 100 validações cruzadas.

Nº de validações cruzadas (10 árvores combinadas)	% bons	% maus	qt total	% do modelo
0	88,29%	37,10%	1.815	84,79%
5	87,17%	40,32%	1.815	83,97%
10	87,64%	37,90%	1.815	84,24%
20	88,05%	38,71%	1.815	84,68%
40	87,88%	38,71%	1.815	84,52%
50	87,70%	**40,32%**	1.815	84,46%
70	87,58%	39,52%	1.815	84,30%
100	88,11%	38,71%	1.815	84,74%

TABELA 14: Árvore de decisão – análise do método de validação cruzada
Fonte: Elaborada pelo autor

Como pode ser observado na tabela 14, aplicando-se 50 validações cruzadas se obtém um maior poder preditivo para maus pagadores. Contudo, o maior poder preditivo agregado do modelo foi obtido sem a validação cruzada, no denominado "recurso exploratório". Como esta opção oferece também menor esforço de processamento, esta foi a referência utilizada para se buscar qual a melhor quantidade de combinações de árvores para se predizer comportamento de crédito. Foi utilizada para validação uma base de dados com as mesmas variáveis com a situação das operações de crédito 6 meses depois.

Foi testada, em seguida, a melhor resposta em termos comparativos de predição fixando-se a proporção de 7% de maus pagadores e também com o balanceamento de 50% de bons e maus pagadores. Como pode ser observada na tabela 16, a predição de maus pagadores foi fortemente melhorada na base balanceada:

Nº de combinações de árvores (sem validação cruzada)	% bons	% maus	qt total	% do modelo
10	88,29%	37,10%	1815	84,79%
500	86,69%	35,48%	1815	83,20%

TABELA 15: BAGGING - matriz de confusão da amostra desbalanceada - base de desenvolvimento
Fonte: Elaborada pelo autor

Nº de combinações de árvores (sem validação cruzada)	% bons	% maus	qt total	% do modelo
10	75,19%	73,10%	274	74,09%
500	72,09%	80,69%	274	76,64%

TABELA 16: BAGGING - matriz de confusão da amostra balanceada - base de desenvolvimento
Fonte: Elaborada pelo autor

Verificou-se que utilizando 100 combinações de árvores para a aplicação do BAGGING se conseguiu a melhor predição de maus pagadores (80,69%), conforme se pode observar No gráfico a seguir:

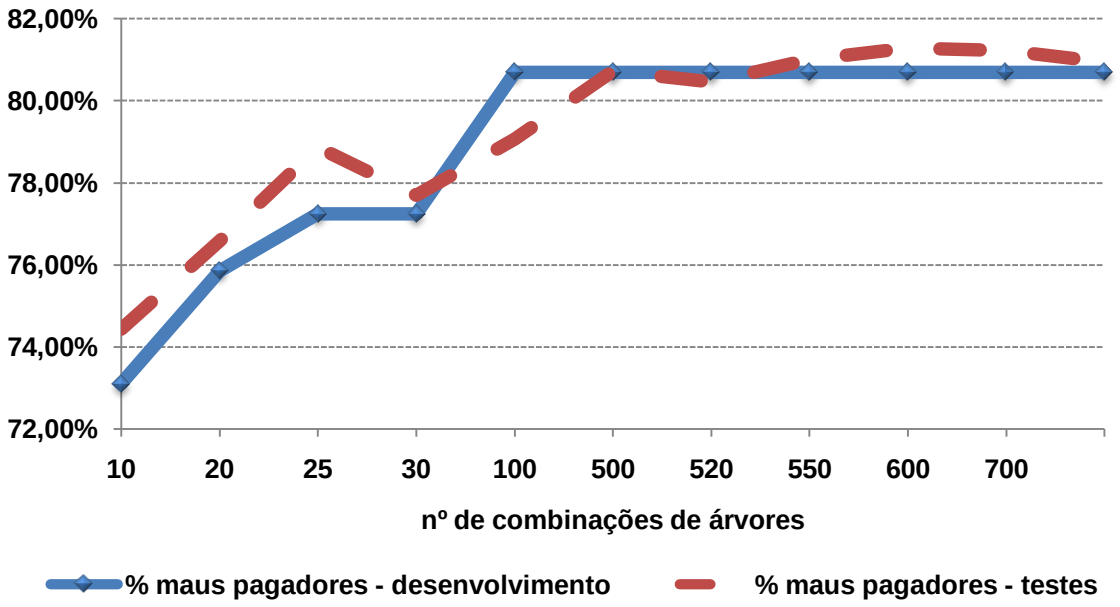

GRÁFICO 6: BAGGING – maus pagadores
Fonte: Elaborado pelo autor

A partir de 100 iterações de árvores, não há ganho significativo no poder preditivo de maus pagadores.

Quando reproduzido na base de testes, representado pela linha vermelha, o modelo se comportou de maneira similar.

Com relação aos bons pagadores, o nível de acertos fica em 75,97%, com 20 combinações de árvores, reduzindo a partir desta quantidade até chegar a 100 combinações. A partir disso, apresenta leve melhora na predição e tendência à estabilidade, conforme segue:

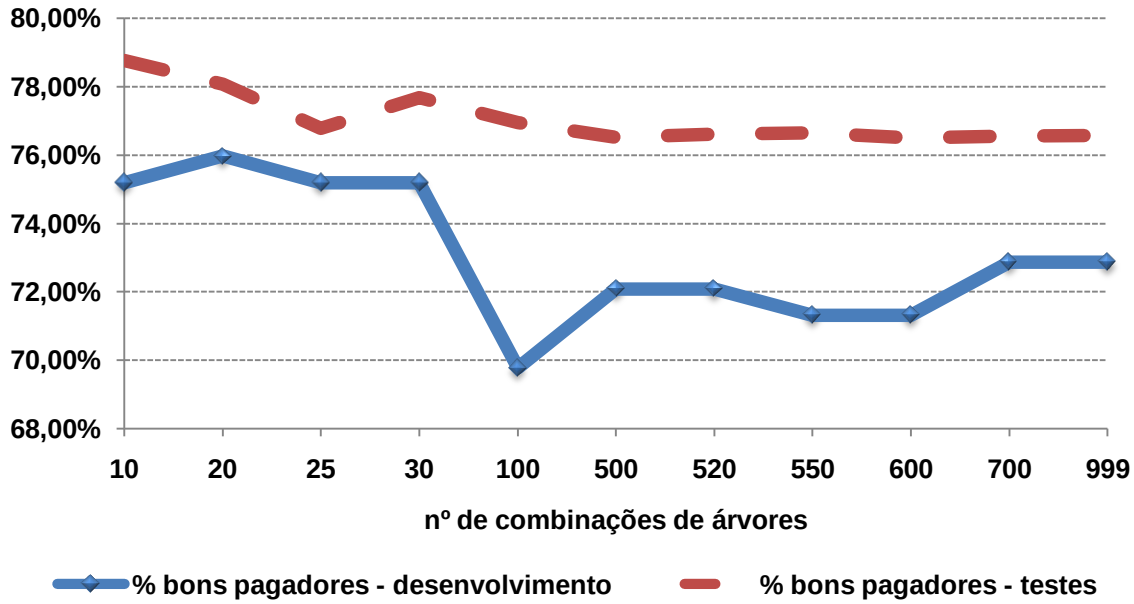

GRÁFICO 7: BAGGING – bons pagadores
Fonte: Elaborado pelo autor

A base de testes apresentou melhor poder preditivo para bons pagadores com 10 iterações de árvores. Embora tenha sido mais estável que a base de desenvolvimento, em termos de predição, apresentou tendência de piora no seu poder preditivo na medida em que se aumentou a quantidade de iterações.

O poder de predição do modelo consolidado não apresenta ganho relevante entre 700 e 999 iterações, sendo mais efetivo no nível entre 500 e 520 iterações. O gráfico a seguir mostra os resultados para o modelo consolidado:

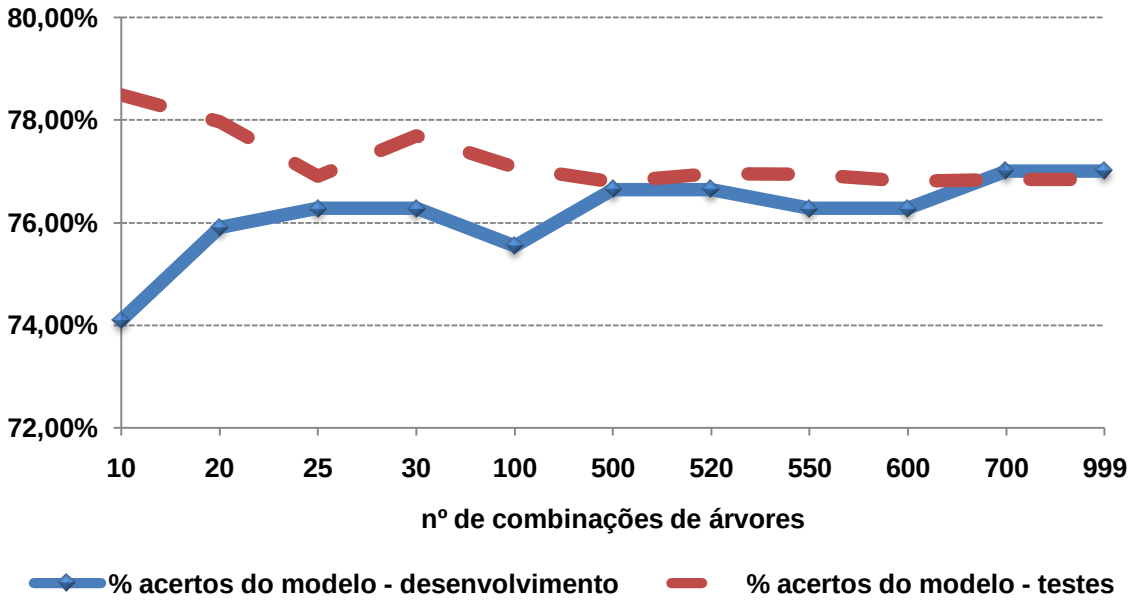

GRÁFICO 8: BAGGING – modelo consolidado
Fonte: Elaborado pelo autor

A seleção do modelo com a melhor combinação de Árvores para a comparação com as demais técnicas foi feita com base no modelo consolidado com 500 iterações.

5.3.2. BOOSTING

Breiman (1996a) promoveu uma mudança simples no algorítmo BOOSTING introduzido por Freund e Schapire (1997), originando a técnica "ARCING" (*Adaptative Resampling and Combining* ou Reamostragem Combinada Adaptativa) utilizada pelo CART. No ARCING, a probabilidade aumenta na medida em que a freqüência de erro na classificação de determinada observação aumenta. Assim, não se verifica probabilidade constante a cada reamostragem.

Para Breiman (1996a), o ARCING normalmente apresenta bons resultados utilizando por volta de 250 árvores.

O parâmetro "exponent" presente nos modelos ARCING do CART representa a seleção da configuração da força do modelo. Esta força define o peso colocado na

reamostragem dos casos selecionados que foram anteriormente classificados erroneamente; uma espécie de viés de preferência em detrimento dos casos selecionados que foram anteriormente classificados corretamente. Breiman *et al.*(1984) descobriu que uma potência 4 funciona bem, enquanto as configurações de um ou dois dão resultados praticamente idênticos ao BAGGING. Definindo um peso maior do que 4, seria difícil localizar uma amostra grande o suficiente para preencher a amostra de desenvolvimento se apenas uma pequena fração dos dados é erroneamente classificada. Dietterich (1998) relatou que, se a variável dependente é registrada erroneamente e, em seguida, é aplicado o ARCING progressivamente, as novas árvores renderiam modelos preditivos ruins.

Neste estudo foram testados outros parâmetros de peso, concluindo-se que pesos menores que 4 agregavam maior poder preditivo aos maus pagadores, com baixas variações no poder preditivo total do modelo. A seguir uma visão sumarizada destas observações, comparando-se os testes em amostras desbalanceadas nas quantidades de bons e maus pagadores:

Exponent	Combinação de árvores	% bons	% maus	% modelo
0,1	10	88,6%	34,7%	84,9%
0,5	10	89,0%	35,5%	85,3%
1,0	10	89,5%	34,7%	85,8%
2,0	10	91,7%	29,8%	87,4%
4,0	10	95,5%	15,3%	90,0%

TABELA 17: ARCING - Seleção de parâmetro "Exponent" - amostra desbalanceada
Fonte: Elaborada pelo autor

Na tabela a seguir, os resultados comparativos para a escolha do parâmetro "*exponent*" em uma amostra balanceada. Neste caso, considerou-se uma ponderação entre os resultados de maus pagadores e resultado agregado do modelo para a escolha do "*exponent*" 0,5, utilizado como referência nos demais processamentos:

Exponent	Combinação de árvores	% bons	% maus	% modelo
0,1	10	74,4%	70,3%	72,3%
0,5	10	77,5%	**68,3%**	72,6%
1,0	10	79,1%	67,6%	73,0%
2,0	10	79,8%	68,3%	73,7%
4,0	10	77,5%	63,5%	70,1%

TABELA 18: ARCING - Seleção de parâmetro "Exponent" - amostra balanceada
Fonte: Elaborada pelo autor

As tabelas a seguir apresentam os resultados dos modelos ARCING em amostras balanceadas e não balanceadas:

Nº de combinações de árvores (sem validação cruzada)	% bons	% maus	qt total	% do modelo
10	89,0%	35,5%	1815	85,3%
500	89,8%	33,1%	1815	86,0%

TABELA 19: ARCING – matriz de confusão da amostra desbalanceada
Fonte: Elaborada pelo autor

Nº de combinações de árvores (sem validação cruzada)	% bons	% maus	qt total	% do modelo
10	77,52%	68,28%	274	72,63%
500	75,19%	72,41%	274	73,72%

TABELA 20: ARCING – matriz de confusão da amostra balanceada
Fonte: Elaborada pelo autor

Conforme se pode observar também nos modelos BAGGING, a amostra balanceada no ARCING apresentou melhores resultados em termos de predição de maus pagadores que a amostra desbalanceada, em termos de proporção de bons e maus pagadores, embora o resultado agregado do modelo tenha se apresentado melhor na amostra desbalanceada.

Para verificação do poder preditivo dos modelos ARCING, foram testadas combinações de árvores de decisão até o limite proporcionado pelo software CART. Os resultados estão nos gráficos a seguir. O parâmetro de peso com que as

observações classificadas incorretamente retornam no próximo processamento, conforme descrito anteriormente nesta seção, foi 0,5:

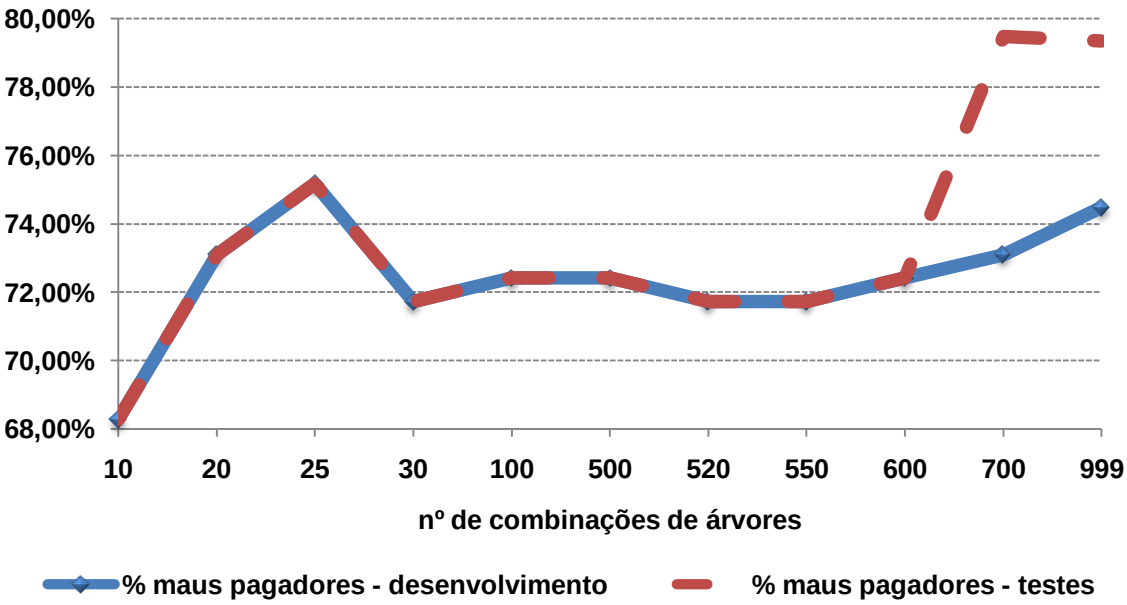

GRÁFICO 9: BOOSTING – maus pagadores
Fonte: Elaborado pelo autor

A observação mostra que, a partir de 30 combinações de árvores o poder preditivo para maus pagadores é prejudicado. Para os maus pagadores, há um ponto de maior predição que corresponde a 25 combinações ou iterações de árvores (75,17%). A partir de 520 iterações há uma tendência de ganho no poder preditivo, contudo o limite de iterações disponível no CART (999) não permitiu que se superasse o poder preditivo das 25 iterações.

Com relação ao poder preditivo do modelo para bons pagadores, observou-se que entre 10 e 20 iteração esteve o maior número de acertos (77,52%). A partir de 100 iterações se observa tendência de queda na assertividade.

Quando reproduzido o modelo na base de testes são observados comportamentos preditivos semelhantes, apenas com descolamento da curva a partir de 600 iterações, quando se vê que o aumento na combinação de árvores traz melhor predição.

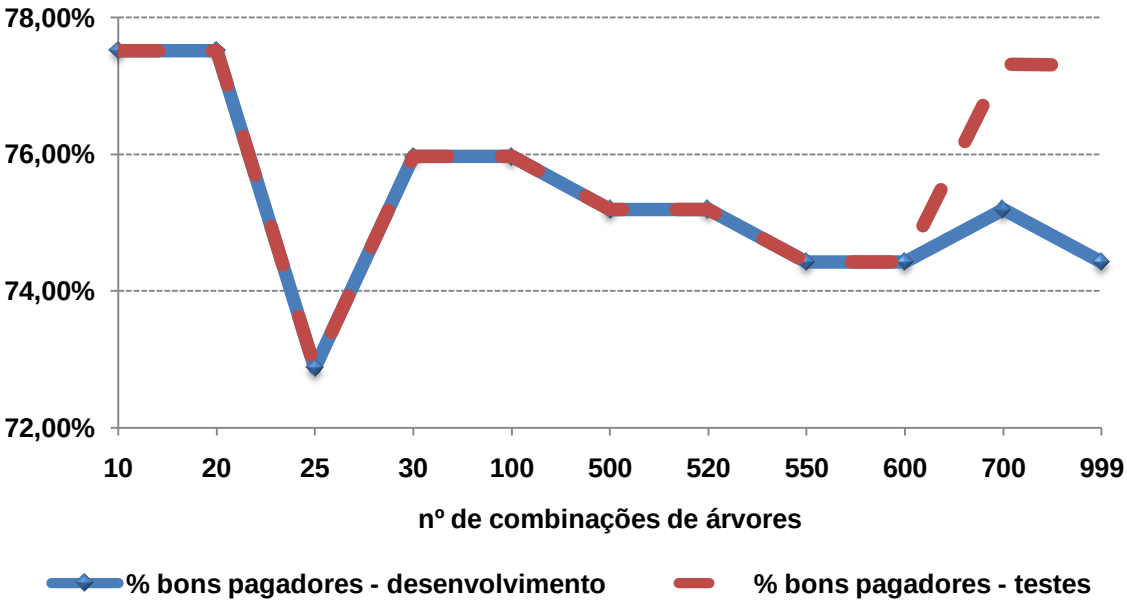

GRÁFICO 10: BOOSTING – bons pagadores
Fonte: Elaborado pelo autor

O modelo consolidado apresentou maior poder de predição com 20 iterações (75,18%), com tendência de redução na assertividade até 550 iterações. A partir daí a tendência é revertida, sugerindo que maior quantidade de iterações de Árvores aumenta o número de acertos do modelo. A limitação do CART em 999 árvores não permite inferir sobre a melhor quantidade de combinações a partir de 999, conforme se observa no gráfico 11, a seguir:

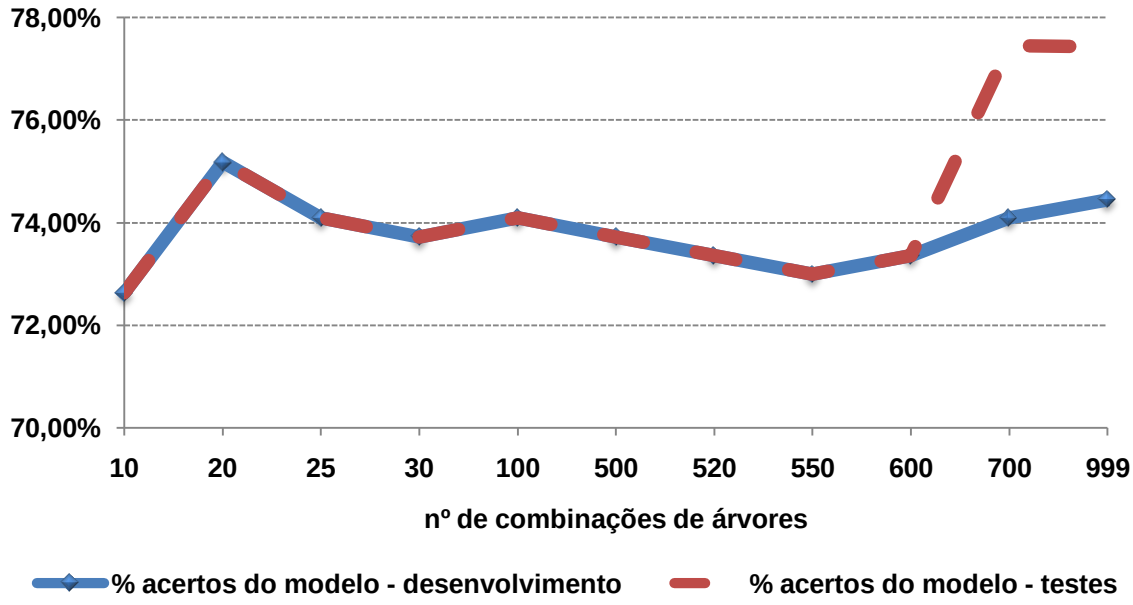

GRÁFICO 11: BOOSTING – modelo consolidado
Fonte: Elaborado pelo autor

Para efeito de comparação com as demais técnicas utilizou-se o melhor poder preditivo de maus pagadores, ou seja, 25 iterações.

5.4. Análise comparativa

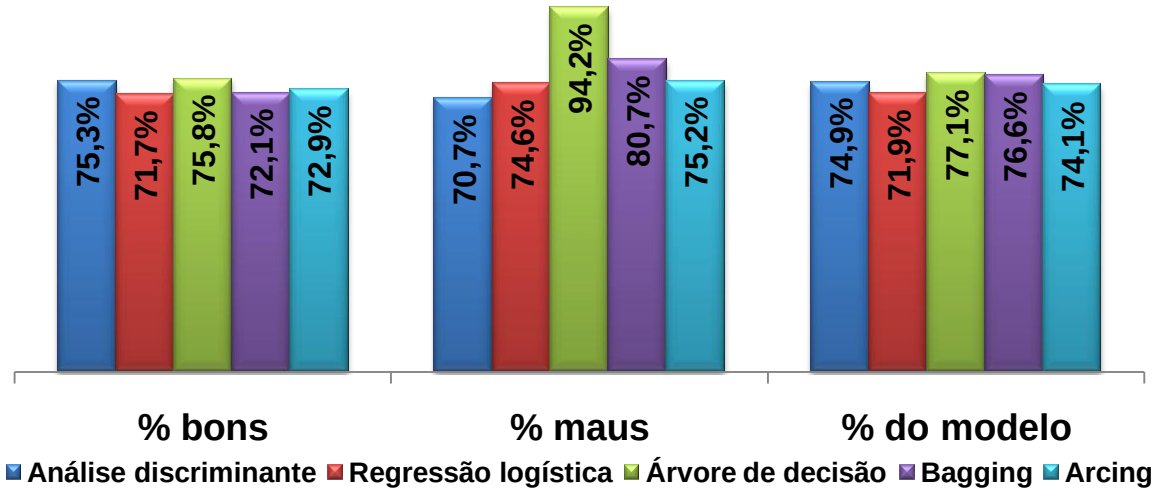

GRÁFICO 12: Comparativo de técnicas
Fonte: Elaborado pelo autor

Conforme definido na seção 5.3, a árvore de decisão utilizada na comparação apresentou 357 nós. O BAGGING utilizado conteve 500 iterações e o BOOSTING ou ARCING 25 iterações. O critério foi o melhor poder preditivo de maus pagadores, equacionado pelos bons pagadores e modelo consolidado. Nos modelos selecionados, foram utilizadas as informações de predição de bons, maus e agregado do modelo para se construir o gráfico 12.

Em termos comparativos, as árvores de decisão se mostram mais efetivas para a predição de maus pagadores (94,2% de acerto), seguida do BAGGING (80,7%) e do BOOSTING (ou ARCING, 75,2%). Para a predição de maus pagadores em financiamentos imobiliários, as técnicas de regressão logística e análise discriminante apresentaram os piores resultados (74,6% e 70,7%, respectivamente). Para os bons pagadores, a árvore de decisão também apresentou o melhor poder preditivo (75,8%), seguida da análise discriminante (75,3%) e do BOOSTING (72,9%). Para os bons pagadores de financiamentos imobiliários, BAGGING e regressão logística apresentaram os piores resultados (72,1% e 71,7%, respectivamente).

Em todas as observações, a técnica não paramétrica de Árvore Decisão obteve o melhor poder preditivo. Cabe ressaltar que as Árvores de Decisão utilizam o recurso de partição recursiva e realizam tantas quebras nas variáveis quanto possa melhorar seu poder preditivo, por vezes não atribuindo sentido a estas quebras. A elevada quantidade de nós (357) para as 10 variáveis selecionadas remete a esta conclusão.

Na tabela 21, a seguir, os resultados apresentam-se consistentes entre os modelos desenvolvidos aplicados à base de testes. A base de testes contém operações de financiamento imobiliário onde os tomadores podem ou não possuir crédito consignado.

Técnicas comparadas	Base de desenvolvimento		Base de testes	
	% bons	% maus	% bons	% maus
Análise discriminante	75,3%	70,7%	81,1%	67,7%
Regressão logística	71,7%	74,6%	70,6%	69,2%
Árvore de decisão	75,8%	94,2%	80,2%	78,3%
Bagging	72,1%	80,7%	76,5%	80,7%
Arcing	72,9%	75,2%	72,9%	75,2%

TABELA 21: Comparativo das técnicas – base de desenvolvimento e base de testes
Fonte: Elaborada pelo autor

6. CONCLUSÕES

Neste estudo foram aplicadas as técnicas paramétricas tradicionais de análise discriminante e regressão logística para análise de crédito de financiamentos imobiliários com e sem crédito consignado. Foi comparada a taxa de acertos dos métodos citados, com as técnicas não-paramétricas baseada em árvores de classificação, além dos métodos de meta-aprendizagem BAGGING e BOOSTING, que combinam classificadores para obter uma melhor precisão nos algoritmos.

As técnicas de Árvore de Decisão e BAGGING se mostraram consistentemente superiores às técnicas tradicionais de Análise discriminante e Regressão Logística no que diz respeito à predição de maus pagadores.

Ao final do estudo, concluiu-se que as técnicas computacionais de árvores de decisão se mostram mais efetivas para a predição de maus pagadores (94,2% de acerto), seguida do BAGGING (80,7%) e do BOOSTING (ou ARCING, 75,2%). Para a predição de maus pagadores em financiamentos imobiliários, as técnicas de regressão logística e análise discriminante apresentaram os piores resultados (74,6% e 70,7%, respectivamente). Para os bons pagadores, a árvore de decisão também apresentou o melhor poder preditivo (75,8%), seguida da análise discriminante (75,3%) e do BOOSTING (72,9%). Para os bons pagadores de financiamentos imobiliários, BAGGING e regressão logística apresentaram os piores resultados (72,1% e 71,7%, respectivamente).

Na análise discriminante verifica-se que as variáveis ‘quantidade de parcelas’ e ‘clientes comuns’, representam o maior impacto relativo na identificação de maus pagadores. Os ‘clientes comuns’ não são funcionários públicos nem recebem seu pagamento pela Instituição, dificultando a prioridade no débito das parcelas do financiamento. O aumento na quantidade de parcelas representa maior propensão à inadimplência. Maiores quantidades de parcelas tornam a operação mais suscetível às variações de cenários econômicos e também às eventualidades que podem ocorrer no planejamento financeiro dos tomadores, podendo estas serem as possíveis explicações para o poder das variáveis. O ‘tipo de garantia’, que nesta

base representa a alienação (em detrimento da hipoteca) também se mostra efetiva no poder discriminante. Neste caso, o aspecto jurídico da facilidade de recuperação da alienação, quando comparada à hipoteca, pode ser a explicação para a menor inadimplência. A ‘renda’, embora apontada em estudos como sendo uma variável explicativa relevante, neste caso, não se mostrou de grande poder discriminatório.

Na regressão logística, o fator exponencial mostra que, quando se trata de um ‘cliente comum’, a chance de o tomador ser um mau pagador é 6,33 vezes maior. Para um tomador com crédito consignado, a chance se ser um mau pagador é 2,19 maior do que se este tomador não tivesse tal modalidade de empréstimo. A presença de crédito consignado entre as operações dos tomadores de financiamento imobiliário também apresenta relevância na análise discriminante.

As técnicas não-paramétricas de Árvores de Decisão, incluindo BAGGING e BOOSTING que, nesse estudo, partiram de árvores, não permitem esta avaliação pormenorizada por variável inserida no modelo, visto que utilizam a partição recursiva até o limite do poder preditivo.

Em todas as técnicas comparadas, a base de testes apresentou resultados consistentes e compatíveis com a base de desenvolvimento. O período da base de desenvolvimento é dezembro de 2008 e da base de testes é junho de 2009. Cabe ressaltar que, entre este período e agosto de 2011, período da publicação deste estudo, ocorreram mudanças nas políticas governamentais brasileiras de incentivo ao crédito que podem ter alterado o contexto dos financiamentos imobiliários. Contudo, tais mudanças não foram objeto deste estudo.

A partir de uma ampla amostra contemplando dados de operações fornecidas por uma instituição brasileira que atua nos segmentos de financiamento imobiliário e crédito consignado, o estudo verificou quais variáveis e técnicas melhor discriminam os bons e maus pagadores. Com isso, foram conduzidas análises de estabilidade dos modelos que, comumente, não são empreendidas em pesquisa acadêmicas sobre crédito.

Sugere-se para os próximos estudos:

- A aferição de custo de não-concessão de empréstimos consignados para tomadores com financiamentos imobiliários;

- O estudo da relação entre o comprometimento da renda - causado pelo valor da parcela mensal paga - e o risco de crédito;

- O estudo aprofundado do impacto do crédito consignado no risco de crédito dos financiamentos imobiliários, com modelos específicos para tomadores com e sem estes tipos de operações;

- A utilização de modelos diferentes para objetivos diferentes, ou seja, modelos diferentes para a detecção de bons e maus pagadores;

- A verificação dos impactos das mudanças no cenário econômico brasileiro e das estratégias de incentivo governamental para financiamentos imobiliários, como o Programa Minha Casa, Minha Vida;

- A mediação do risco de crédito nas demais operações dos tomadores de financiamentos imobiliários.

REFERÊNCIAS BIBLIOGRÁFICAS

ABECIP. **Associação Brasileira das Entidades de Crédito Imobiliário e Poupança**. Disponível em:

<http://www.abecip.org.br/default.asp?resolucao=1024X600>. Acesso em: 20 set. 2010.

ALTMAN, E. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. **Journal of Finance.** Vol. 23 (4), 1968.

_____. An Emerging Market Credit Scoring System for Corporate Bonds. **Emerging Markets Review**. Vol. 6, 2005.

ALTMAN, E.; SABATO, G. Modeling Credit Risk for SMEs: Evidence from the US Market. **ABACUS**. Vol. 43 (3), 2007.

_____.; RESTI, A.; SIRONI, A. **Analyzing and explaining default recovery rates.** The International Swaps and Derivatives Association. London, 2001.

ANDERSON, R. The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation. **Oxford University Press**, Oxford, 2007.

BENSIC, M.; SARLIJA, N.;ZEKIC – SUSAC, M. Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. **Intelligent Systems in Accounting, Finance and Management.** Vol. 13 (3), 2006.

BIS – Bank for International Settlements – Housing finance in the global financial markets. **CGFS Papers** Nº 26, 2006. Disponível em: http://www.bis.org/publ/cgfs26.pdf. Acesso em: 22 out. 10.

BRASIL. Banco Central do Brasil. Relatório de Economia Bancária e Crédito – 2007; texto técnico: **Porque o volume de empréstimo consignado no setor privado é baixo? Qual a solução?** Por Tson Chu, Victorio; Lundberg, Eduardo Luis e Takeda Tony, 2007.

BRASIL. Banco Central do Brasil. **Relatório de Estabilidade Financeira**, Vol. 8 (2), 2009.

_____. Decreto n° 4.961, de 20 de janeiro de 2004. Regulamenta o art. 45 da Lei no 8.112, de 11 de dezembro de 1990, que dispõe sobre as consignações em folha de pagamento dos servidores públicos civis, dos aposentados e dos pensionistas da administração direta, autárquica e fundacional do Poder Executivo da União, e dá outras providências. Disponível em: <http://presidencia.gov.br>. Acesso em: 23 mai. 2010.

_____. Lei Federal n° 10.820, de 17 de dezembro de 2003. Dispõe sobre a autorização para desconto de prestações em folha de pagamento, e dá outras providências. Disponível em: <http://www.presidencia.gov.br>. Acesso em: 12 ago. 2010.

BREIMAN, L. **ARCING classifiers (Technical report)**. Berkley:Estatistics Departament, University of California, 1996a.

_____. BAGGING predictors. **Machine Learning**. Vol. 24 (2), 1996b.

_____.; FREIDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and Regression Trees.** Wadsworth, 1984.

BRIGHAM, E. F.; GAPENSKI, L. C.; EHRHARDT, M. C. **Administração financeira – teoria e prática**. São Paulo: Atlas, 2001.

BRUCH, l. **Discriminant Analysis**, Hafner Press, New York, NY, USA, 1975.

BÜHLMANN, P.; YU, B. **Analyzing BAGGING.** Annals of Statistics 30, 2002.

CAOUETTE J.; ALTMAN E. M.; NARAYANAN, P. **Managing Credit Risk: The Next Great Financial Challenge.** Wiley Frontiers in Finance, 1998.

CAVALLAZZI, R. L. **O perfil do superendividamento: referências no Brasil**. In: Rosângela Lunardelli Cavallazzi; Claudia Lima Marques. (Org.). Direitos do Consumidor Endividado. 1.ed. São Paulo: Editora revista dos Tribunais. Vol. 29, 2006.

CLIFF, N. **Ordinal methods for behavioral data analysis**. New Jersey: Lawrence Erlbaum Associates, 1996.

COFFMAN, J. Y. **The proper role of tree analysis in forecasting the risk behavior of borrowers, Management Decision Systems**. MDS Reports, 1986.

CREAMER, G., FREUND, Y.**Predicting performance and quantifying corporate governance risk for latin american adrs and banks**. In: I Proceedings of the financial engineering and applications conference, MIT-Cambridge, 2004.

DIETTERICH, T. An experimental comparison of three methods for constructing ensembles of decision trees: BAGGING, BOOSTING and Randomization. **Machine Learning**, 40, 1998.

FELDESMAN, M. R. Classification trees as an alternative to linear discriminant analysis. **American Journal of Physical Anthropology**. Vol 119 (3), 2002.

FJP. Fundação João Pinheiro. Centro de Estatística e Informações. **Déficit habitacional no Brasil 2006.** Convênio PNUD/Ministério das Cidades, Secretaria Nacional de Habitação, Brasília, 2008. Disponível em: < www.fjp.mg.gov.br > Acesso em: 30 set. 2008.

FREUND, Y.; SCHAPIRE, R. A decision-theoretic generalization of on-line learning and an application to BOOSTING. **Journal of Computer and System Sciences**. Vol. 55 (1), 1997.

_____.; _____. A short introduction to BOOSTING. **Journal of Japanese Society for Artificial Intelligence**. Vol. 14 (5), 1999.

HAIR, Jr., J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Multivariate data analysis.** 5.ed. New Jersey: Prentice Hall, 1998.

HAND, D. J.; HENLEY, W. E. Statistical Classification Methods in Consumer Credit Scoring: A ReviewAuthor (s). **Journal of the Royal Statistical Society. Series A** (Statistics in Society), vol. 160, n.3, pp. 523-541, 1997. Disponível em: http://www.jstor.org/stable/2983268. Acesso em: 14 dez. 2010.

HARRELL Jr. **Regression modeling strategies with applications to linear models, logistic regression and survival analysis**. New York: Springer-Verlag, 2001.

JORION, P. **Value at Risk: a nova fonte de referência para o controle do risco de mercado.** São Paulo: Bolsa de Mercadorias & Futuros, 1998.

KANITZ, S. C**. Como Prever Falências**. São Paulo: McGraw-Hill, 1978.

KLECKA, W. R**. Discriminant Analysis. Quantitative Applications in the Social Sciences.** USA: Sage Publications, 1980.

KRIVO, L. J., PETERSON, R. D., RIZZOR, H., REYNOLDS, J. R. **Race, Segregation, and the Concentration of Disadvantage**: 1980-1990. Social Problems 45(1):61-80, 1998.

LEMOS, E. P.; STEINER, M. T. A.; NIEVOLA, J. C. Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining. **Revista de Administração da USP**. Vol. 40 (3), 2005.

LEWIS, E. M. **An Introduction to Credit Scoring**. Athena Press, San Rafael, 1992.

LIM, T. S.; LOH, W. Y.; SHIH, Y. S. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. **Machine Learning.** Vol. 40, 2000.

MICHA, B. Analysis of business failures in France. **Journal of Banking and Finance**. Vol. 8, 1984.

MINSKY, H. **Can "it" happen again?** Essays on instability and finance. New York: M. E. Sharpe, 1982.

MINUSSI, J. A.; DAMACENA, C.; NESS JR, W. L. **Um Modelo de Previsão de Solvência Utilizando Regressão Logística.** Resumo. RAC. Vol. 6 (3), Set./Dez, 2002.

MITCHELL, T. M. **Machine Learning**. McGraw-Hill, 1997.

O'CONNELL, A. A. **Logistic regression models for ordinal response variables**. Sage University Paper Series on Quantitative Applications in the Social Sciences. London: Sage Publications, 2006.

PINHEIRO, M. A armadilha do crédito. **Carta Capital**. Vol. 438, 2007.

PRESS, S.; WILSON, S. Choosing Between Logistic Regression and Discriminant Analysis. **Journal of the American Statistical Association**. Vol. 73. 1978.

PROCON. Departamento de Equipe de Pesquisas. **Empréstimo consignado para aposentados e pensionistas do INSS**. São Paulo: 2006. Disponível em: <http://www.procon.sp.gov/br/pdf/empconsig.pdf> Acesso em: 20 out. 2009.

QUERCIA, R. G.; STEGMAN, M. A. Residential Mortgage Default: A Review of the Literature. **Journal of Housing Research**. Vol 3 (2), 1992.

SABZEVARI, H.; SOLEYMANI, M.; NOORBAKHSH, E. **A comparison between statistical and data mining methods for credit scoring in case of limited available data**, 2009. Disponível em: <http://www.crc.man.ed.ac.uk/conference/archive/2007/papers/sabzevari-et-al.pdf>. Acesso em: 07 dez. 2010.

STEVENS, S. S. **On the Theory of scales of measurement.** Science, 103 (2684), 1946.

TAFFLER, R.J. Forecasting company failure in the UK using discriminant analysis and financial ratio data. **Journal of the Royal Statistical Society**. Vol. 145, Part 3, 1982.

TANG, Y.; CHEN, H.; WANG, B.; CHEN, M.; CHEN, M.; YANG, X. Discriminant Analysis of Zero Recovery for China's NPL. **Journal of Applied Mathematics and Decision Sciences.** Article ID 594793. doi:10.1155/2009/594793, 2009.

TAKEDA, T.; BADER, F. L. C. **Consignação em folha de Pagamento: Fatores da Impulsão do Crédito.** In: BANCO CENTRAL DO BRASIL. Relatório de Economia Bancária e Crédito. 69-87, 2005. Disponível em: <http://www.bcb.gov.br/?RELECONCRED>. Acesso em: 23 jul. 2009.

THOMAS, L. C.; EDELMAN, D. B.; CROOK, J. N. **Credit scoring and its applications. USA:** Society for Industrial and Applied Mathematics, 2002.

WAGNER, H. The use of credit scoring in the mortgage industry. **Journal of Financial Services Marketing**; 9, 2; ABI/INFORM Global, 2004.

WANG, G.; HAO, J.; MA, J.; JIANG, H. **A comparative assessment of ensemble learning for credit scoring**. Expert Systems with Applications, vol.38, Issue 1, jan. 2011, p.223-230, ISSN 0957-4174, DOI: 10.1016/j.eswa.2010.06.048. Disponível em:<http://www.sciencedirect.com/science/article/B6V03-50G696B-F/2/ad95ad6a8b2b62a0cee230491cfafb2f>. Acesso em: 01 fev. 2011.

WIGINTON, J. C. A note on the comparison of *logit* and discriminant models of consumer credit behavior. **Journal of Financial and Quantitative Analysis**. Vol.15, 1980.

YE, X, K.; LIU, H. L. **Research on structure of loss given default of NPLs in commercial bank**. Contemporary Finance, Vol. 6, 2006.

ZHANG, D.; ZHOU, X.; LEUNG, S. C. H.; ZHENG, J. **Vertical BAGGING decision trees model for credit scoring.** Expert Systems with Applications, vol. 37, Issue 12, dez. 2010, p.7838-7843, ISSN 0957-4174, DOI: 10.1016/j.eswa.2010.04.054. Disponível em: <http://www.sciencedirect.com/science/article/B6V03-5017HGC7/2/cea5d55fb0225695cd86d8b1f897c402>. Acesso em: 12 dez. 2010.

APÊNDICE A: Algoritmo de Análise Discriminante em SPSS

```
DISCRIMINANT
/GROUPS=default(0 1)
/VARIABLES=
TOT_PARCELAS_CTR
tem_consignado_max
RENDA
ALIENAÇÃO
SEXO
QTDE_DEPENDENTES_PF
idade
Escolaridade
d5_estcivil
d6_estcivil
d15_profissão
d13_profissão
d14_profissão
d16_profissão
/ANALYSIS ALL
/SAVE=SCORES
/PRIORS equal
/STATISTICS=BOXM COEFF TABLE
/PLOT=SEPARATE
/CLASSIFY=NONMISSING POOLED.
```

APÊNDICE B: Algoritmo de Regressão Logística em SPSS

```
LOGISTIC REGRESSION VARIABLES default
/METHOD=enter
TOT_PARCELAS_CTR
tem_consignado_max
RENDA
ALIENAÇÃO
SEXO
QTDE_DEPENDENTES_PF
idade
Escolaridade
d5_estcivil
d6_estcivil
d15_profissão
d13_profissão
d14_profissão
d16_profissão
/SAVE=PRED PGROUP
/CLASSPLOT
/PRINT=GOODFIT
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.07).
```

APÊNDICE C: Algoritmo de Árvore de Decisão no CART

```
USE "C:\Documents and Settings\ACER\Meus documentos\Mestrado\mestrado resumido\cart\dez08.sav" ENCODING=SHIFTJIS
VARIABLES IN RECT FILE ARE:
REGIAO_CEP
TOT_FX_VCTO
TOT_PARCELAS_CTR
TEM_CONSIGNADO_MAX
VALOR_CONTRATO
RENDA
ALIENACAO
SEXO
DESCR_ESCOLARIDADE$
QTDE_DEPENDENTES_PF
PUBLICO
SERVIDOR_PUBLICO
IDADE
ESCOLARIDADE
DEFAULT
FAIXA_PARCELAS
FAIXA_RENDA
FAIXA_IDADE

 C:\Documents and Settings\ACER\Meus documentos\Mestrado\mestrado resumido\cart\dez08.sav: 18045 records.

Model reset.
>REM ***Setting General options
>LOPTIONS MEANS = NO, PREDICTIONS = NO, TIMING = NO, GAINS = NO, ROC = NO, PLOTS = NO
>FORMAT = 5
>REM***Setting CART options
>LOPTIONS, NOPRINT = NO, PS = NO
>BOPTIONS SURROGATES = 5 PRINT = 5, COMPETITORS = 5 CPRINT = 5, TREELIST = 10,
>BRIEF
>REM ***Setting MARS default options
>BOPTIONS PENALTY = 0.000000, SPEED = 4, INTERACTIONS = 1, MINSPAN = 0, BASIS = 15
>MARS SEED = 987654321
>BOPTIONS OLS  = YES
>PRINT = TERSE
>REM FOCUS CLASS = 1
>BOPTIONS MISSING = NO
>DISCRETE MISSING = MISSING
>CATEGORY
```

```
>AUXILIARY
>MODEL DEFAULT
Model reset: DEFAULT
>KEEP
>KEEP ALIENACAO, ESCOLARIDADE, FAIXA_IDADE, FAIXA_PARCELAS,
FAIXA_RENDA, PUBLICO,
> QTDE_DEPENDENTES_PF, REGIAO_CEP, SERVIDOR_PUBLICO, SEXO,
TEM_CONSIGNADO_MAX
>LOPTIONS UNS = NO
>CATEGORY DEFAULT
>AUXILIARY DESCR_ESCOLARIDADE$, IDADE, RENDA, TOT_FX_VCTO,
TOT_PARCELAS_CTR,
>VALOR_CONTRATO
>BOPTIONS
>FORCE
>DISALLOW
>ERROR FILE = "C:\Documents and Settings\ACER\Meus
documentos\Mestrado\mestrado resumido\cart\imobiliário jun09.sav"

 C:\Documents and Settings\ACER\Meus documentos\Mestrado\mestrado
resumido\cart\imobiliário jun09.sav: 22994 records.

 VARIABLES IN RECT FILE ARE:
TOT_FX_VCTO
TOT_PARCELAS_CTR
TEM_CONSIGNADO_MAX
VALOR_CONTRATO
RENDA
ALIENACAO
SEXO
DESCR_ESCOLARIDADE$
QTDE_DEPENDENTES_PF
PUBLICO
SERVIDOR_PUBLICO
IDADE
ESCOLARIDADE
DEFAULT
REGIAO_CEP
FAIXA_PARCELAS
FAIXA_RENDA
FAIXA_IDADE
>BOPTIONS SERULE = 0, IMPORTANCE = 1
>BOPTIONS SURROGATES = 5 PRINT = 5, COMPETITORS = 5 CPRINT = 5
>METHOD GINI POWER =  0.0000
>BUILD

Salford Predictive Modeler: CART(R) version 6.6.0.091
```

APÊNDICE D: Algorítmo de BAGGING no CART

```
USE "C:\Arquivos de programas\Salford Predictive Modeling\Predictive Miner\bin\dez08.sav" ENCODING=SHIFTJIS
 VARIABLES IN RECT FILE ARE:
REGIAO_CEP
TOT_FX_VCTO
TOT_PARCELAS_CTR
TEM_CONSIGNADO_MAX
VALOR_CONTRATO
RENDA
ALIENACAO
SEXO
QTDE_DEPENDENTES_PF
PUBLICO
SERVIDOR_PUBLICO
IDADE
ESCOLARIDADE
DEFAULT
FAIXA_PARCELAS
FAIXA_RENDA
FAIXA_IDADE

 C:\Arquivos de programas\Salford Predictive Modeling\Predictive Miner\bin\dez08.sav: 2532 records.

 Model reset.
>REM ***Setting General options
>LOPTIONS MEANS = NO, PREDICTIONS = NO, TIMING = NO, GAINS = NO, ROC = NO, PLOTS = NO
>FORMAT = 5
>REM***Setting CART options
>LOPTIONS, NOPRINT = NO, PS = NO
>BOPTIONS SURROGATES = 5 PRINT = 5, COMPETITORS = 5 CPRINT = 5, TREELIST = 10,
>        BRIEF
>REM ***Setting MARS default options
>BOPTIONS PENALTY = 0.000000, SPEED = 4, INTERACTIONS = 1, MINSPAN = 0, BASIS = 15
>MARS SEED = 987654321
>BOPTIONS OLS  = YES
>PRINT = TERSE
>REM FOCUS CLASS = 1
>CATEGORY
>AUXILIARY
>MODEL DEFAULT
 Model reset: DEFAULT
```

```
>KEEP
>KEEP ALIENACAO, ESCOLARIDADE, FAIXA_IDADE, FAIXA_PARCELAS,
FAIXA_RENDA,
> QTDE_DEPENDENTES_PF, REGIAO_CEP, SERVIDOR_PUBLICO, SEXO,
TEM_CONSIGNADO_MAX
>LOPTIONS UNS = NO
>CATEGORY DEFAULT
>AUXILIARY IDADE, PUBLICO, RENDA, TOT_FX_VCTO, TOT_PARCELAS_CTR,
VALOR_CONTRATO
>METHOD GINI POWER =  0.0000
>MOPTIONS CYCLES = 10, EXPLORE = YES,
>DETAILS = NONE, RTABLES = YES, TRIES = 3, ARC = NO, SETASIDE = PROP
= 0.100000
>COMBINE

 Salford Predictive Modeler: CART(R) version 6.6.0.091
```

APÊNDICE E: Algorítmo de ARCING no CART

USE "C:\Arquivos de programas\Salford Predictive Modeling\Predictive Miner\bin\dez08.sav" ENCODING=SHIFTJIS
VARIABLES IN RECT FILE ARE:
REGIAO_CEP
TOT_FX_VCTO
TOT_PARCELAS_CTR
TEM_CONSIGNADO_MAX
VALOR_CONTRATO
RENDA
ALIENACAO
SEXO
QTDE_DEPENDENTES_PF
PUBLICO
SERVIDOR_PUBLICO
IDADE
ESCOLARIDADE
DEFAULT
FAIXA_PARCELAS
FAIXA_RENDA
FAIXA_IDADE

C:\Arquivos de programas\Salford Predictive Modeling\Predictive Miner\bin\dez08.sav: 2532 records.

Model reset.
>REM ***Setting General options
>LOPTIONS MEANS = NO, PREDICTIONS = NO, TIMING = NO, GAINS = NO, ROC = NO, PLOTS = NO
>FORMAT = 5
>REM***Setting CART options
>LOPTIONS, NOPRINT = NO, PS = NO
>BOPTIONS SURROGATES = 5 PRINT = 5, COMPETITORS = 5 CPRINT = 5, TREELIST = 10,
>BRIEF
>REM ***Setting MARS default options
>BOPTIONS PENALTY = 0.000000, SPEED = 4, INTERACTIONS = 1, MINSPAN = 0, BASIS = 15
>MARS SEED = 987654321
>BOPTIONS OLS = YES
>PRINT = TERSE
>REM FOCUS CLASS = 1
>CATEGORY
>AUXILIARY
>MODEL DEFAULT

Model reset: DEFAULT
>KEEP
>KEEP ALIENACAO, ESCOLARIDADE, FAIXA_IDADE, FAIXA_PARCELAS, FAIXA_RENDA,
>QTDE_DEPENDENTES_PF, REGIAO_CEP, SERVIDOR_PUBLICO, SEXO, TEM_CONSIGNADO_MAX
>LOPTIONS UNS = NO
>CATEGORY DEFAULT
>AUXILIARY IDADE, PUBLICO, RENDA, TOT_FX_VCTO, TOT_PARCELAS_CTR, VALOR_CONTRATO
>ERROR FILE = "C:\Arquivos de programas\Salford Predictive Modeling\Predictive Miner\bin\jun09.sav"

C:\Arquivos de programas\Salford Predictive Modeling\Predictive Miner\bin\jun09.sav: 22994 records.

VARIABLES IN RECT FILE ARE:
TOT_FX_VCTO
TOT_PARCELAS_CTR
TEM_CONSIGNADO_MAX
VALOR_CONTRATO
RENDA
ALIENACAO
SEXO
DESCR_ESCOLARIDADE$
QTDE_DEPENDENTES_PF
PUBLICO
SERVIDOR_PUBLICO
IDADE
ESCOLARIDADE
DEFAULT
REGIAO_CEP
FAIXA_PARCELAS
FAIXA_RENDA
FAIXA_IDADE
>BOPTIONS SERULE = 0, IMPORTANCE = 1
>BOPTIONS SURROGATES = 5 PRINT = 5, COMPETITORS = 5 CPRINT = 5
>METHOD GINI POWER = 0.0000
>BATTERY
>BUILD

Salford Predictive Modeler: CART(R) version 6.6.0.091