Algoritmo CHAID aplicado à análise de risco de inadimplência no setor imobiliário

Miriam Rodrigues Silvestre (miriam@fct.unesp.br)
Dep. Matemática, Estatística e Computação (DMEC), FCT, Unesp Univ Estadual Paulista
Tamara Aparecida Paccas Silva (tamara_paccas@hotmail.com)
Estatístico formado pela Unesp Univ Estadual Paulista
Cláudio Sá Rodrigues de Lima (claudio2.lima@citi.com)
Banco Citibank S/A

INTRODUÇÃO

A inadimplência em relação aos pagamentos de aluguéis de imóveis é uma grande preocupação de empresas do ramo imobiliário. Desta forma, é uma vantagem para uma empresa imobiliária utilizar modelos estatísticos que consigam realizar previsões com relação à classificação de um novo cliente, antes mesmo que o contrato seja assinado, possibilitando à mesma a recusa de um cliente que o modelo tenha classificado como futuro inadimplente. Pode-se definir um cliente como inadimplente quando ele não cumpre os compromissos, muitas vezes pagando com demasiado atraso, ou até mesmo não pagando os aluguéis contratados; em contrapartida, um cliente é considerado adimplente quando paga seus aluguéis corretamente, de forma a cumprir os compromissos assumidos perante o contrato de aluguel.

Rosa (2000) empregou o algoritmo CHAID para construir modelos de concessão de crédito a clientes de instituições financeiras que desejavam fazer contratos de financiamento de compra de veículos. Os resultados obtidos com o método CHAID para a medida mais importante do ponto de vista de concessão de crédito, ou seja, o percentual de classificação correta de maus-pagadores, foi de 72,6%, levemente inferiores aos 73,3% obtidos pelo algoritmo REAL, e superiores aos 69,2% obtidos com Regressão Logística. Esses resultados podem ser encontrados na Tabela 4.11 de ROSA (2000, p. 50).

Neste trabalho foram construídos modelos estatísticos para a previsão de inadimplência, utilizando o algoritmo CHAID e CHAID-Exaustivo implementados no software Clementine, para dados referentes a contratos de aluguéis de imóveis residenciais de uma empresa imobiliária de Presidente Prudente (SP).

Os dados foram divididos em dois conjuntos (treinamento e teste), e os resultados obtidos mostraram melhor resultado para o algoritmo CHAID, com uma taxa de 90% de classificação correta dos clientes no conjunto de treinamento e 80% no conjunto de teste. O

modelo pode ser considerado adequado pois superou a capacidade preditiva de proporção por chances de 76,89%.

ALGORITMOS CHAID e CHAID-EXAUSTIVO

O algoritmo CHAID (*Chi-Squared Automatic Detection*) foi desenvolvido por Kaas, em 1980. Todo o procedimento é baseado na construção de tabelas de contingência *r*x*c*. Nas linhas da tabela encontram-se as *r* categorias da variável dependente e nas colunas as *c* categorias da variável independente em estudo. Em cada passo do algoritmo é avaliado se uma redução no número de categorias da variável independente será significativa ou não considerando-se a variável dependente. O objetivo final é construir uma árvore de classificação contendo somente as variáveis mais importantes para a classificação, juntamente com suas categorias mais significativas para a variável resposta. Os passos do algoritmo CHAID podem ser encontrados em Santos; Oliveira (2007). O algoritmo CHAID investiga todas as variáveis, após ser encontrada uma partição significante, o algoritmo não continua a agrupar as demais categorias, isto pode impedir que uma divisão melhor seja encontrada para aquela variável.

No algoritmo CHAID-Exaustivo, todos os agrupamentos são examinados, e seleciona-se aquele que apresente a maior associação com a variável dependente. Isto é feito para todas as variáveis independentes. O procedimento continua até que o critério de parada seja alcançado. Tal critério é o mesmo que o definido para o algoritmo CHAID.

APLICAÇÃO DO ALGORITMO CHAID e CHAID-EXAUSTIVO A DADOS DE CLIENTES DE UMA IMOBILIÁRIA

Nesta seção será realizada a aplicação do algoritmo CHAID e CHAID Exaustivo aos dados cadastrais de clientes de uma imobiliária de Presidente Prudente (SP), que tenham efetivado contratos de aluguel de imóveis residenciais. Foram consultados os contratos de aluguel de imóveis realizados para um período de 12 meses e que tiveram seu vencimento até Fevereiro de 2008. Foram verificados os canhotos de todos os 12 aluguéis mensais, e anotadas informações sobre o pagamento com atraso ou não em cada mês. No total foram coletadas informações de 100 contratos. Após essa tarefa, foi possível classificar os clientes em uma de duas categorias: adimplente ou "bom-pagador" e inadimplente ou "mau-pagador". Para se classificar um cliente em uma dessas categorias, primeiramente foi necessário construir uma definição para as mesmas, levando-se em conta a experiência dos tomadores de decisão da imobiliária. Assim sendo, a variável dependente Y foi definida como sendo:

$$\text{"Mau-pagador", se} \begin{cases} \text{houve atraso no pagamento de 2 alugúéis consecutivos mais que uma vez.} \\ \text{ou} \\ \text{houve atraso no pagamento de 3 aluguéis consecutivos ao menos uma vez.} \\ \text{ou} \\ \text{o cliente desistiu do contrato antes do seu vencimento.} \end{cases}$$

"Bom – pagador", caso contrário. (1)

As variáveis independentes foram coletadas a partir do cadastro de clientes da empresa, tendo sido selecionadas 18 variáveis para compor o modelo. Uma das variáveis é denominada Caução, que pode ter como resposta sim ou não, e representa uma das condições exigidas pela empresa para a conclusão do contrato. O cliente opta pelo cadastro de um fiador ou pelo depósito antecipado, "caução", do valor de três aluguéis em uma conta-poupança. Ao final, se o contrato for cumprido corretamente pelo locatário, ele resgata o valor depositado, senão, o dinheiro fica em poder da empresa e do locador.

O conjunto de dados coletado apresentou 86% de adimplentes e 14% de inadimplentes. Para a elaboração do modelo CHAID, dividiu-se o conjunto de dados em duas partes sendo: 70% para treinamento, a qual foi utilizada para construir o modelo, e 30% para teste, destinada à avaliação da capacidade de generalização do modelo construído. Esta divisão respeitou os percentuais de cada classe, e encontra-se definida na Tabela 1, dada a seguir.

Tabela 1 – Partição do conjunto de dados

	Treinamento	Teste	**Total**
Adimplentes	60	26	**86**
Inadimplentes	10	4	**14**
Total	**70**	**30**	**100**

Inicialmente, definiram-se todos os parâmetros necessários para que fosse construído o modelo de classificação da variável dependente (Classificação do Locatário), utilizando o algoritmo CHAID pelo software Clementine da empresa SPSS (SPSS BRASIL, 2007). Foi utilizada a estatística Qui-Quadrado de Pearson e alfa=5%.

Para a construção do modelo inicial todas as 18 variáveis foram utilizadas. Assim, o Modelo 1 foi construído e calculou-se seu percentual de acerto no conjunto de treinamento. O primeiro modelo indicou que as variáveis mais importantes para a classificação da variável dependente eram em ordem decrescente de importância: Caução (pvalor=0,023), Outros Rendimentos do Locatário (pvalor=0,046) e Tipo de Imóvel (pvalor=0,026). Foi anotado o desempenho desse modelo para a classificação da variável dependente.

No Modelo 2, excluiu-se a variável mais importante para o Modelo 1, a variável Caução, este modelo indicou que a única variável a entrar no novo modelo seria a variável Motos (pvalor=0,036). Anotou-se seu desempenho de classificação.

A partir do modelo anterior, foi construído o Modelo 3, sem as variáveis Caução e Motos, restando como mais significativas as variáveis Outros Rendimentos do Locatário (pvalor=0,042) e Tipo de Imóvel (pvalor=0,005). Ainda tentou-se construir um novo modelo sem a variável Outros Rendimentos, mas este não conseguiu encontrar nenhuma variável significativa, ou seja, com pvalor<=0,05 para que fosse possível finalizar a construção do mesmo. Desta forma, o Modelo 3 foi considerado como a última possibilidade na construção de modelos.

Após a finalização destes modelos, iniciou-se a investigação do algoritmo CHAID-Exaustivo. Com a inclusão de todas as variáveis, o Modelo 4 foi formado somente pela variável Caução (pvalor=0,023). O Modelo 5, excluindo-se a variável Caução, indicou como mais importante a variável Motos (pvalor=0,036). O último Modelo 6 sem as variáveis Caução e Motos não pode ser finalizado pois não apresentou nenhuma variável significativa. A Tabela 2 apresenta um resumo geral contendo os resultados de classificação de todos os modelos construídos.

Tabela 2 – Resultados obtidos para as taxas de acerto (%) no conjunto de treinamento, segundo as variáveis utilizadas e o algoritmo aplicado.

Modelo Construído	Taxa Acerto (%) CHAID	Taxa Acerto (%) CHAID-Exaustivo
1. Todas as variáveis	85,71	85,71
2. Todas exceto Caução	85,71	85,71
3. Todas exceto Caução e Motos	**90,00**	-

Analisando-se a Tabela 2, observou-se que a maioria dos modelos construídos com os algoritmos CHAID e CHAID-Exaustivo, apresentaram taxas de acerto em percentual no conjunto de treinamento de 85,71%. A única exceção foi o Modelo 3 para o algoritmo CHAID, o qual apresentou um desempenho no conjunto de treinamento de 90%, por esse motivo, esse modelo foi escolhido como o melhor deles. A taxa de acerto obtida por esse modelo foi de 80,00% no conjunto de teste. Os resultados obtidos por esse modelo estão dispostos na Tabela 3.

Tabela 3 – Taxas de acerto em freqüência e percentual obtidas pelo Modelo CHAID

Conjunto de dados	Taxa de Acerto Total	Taxa de Acerto Bons-pagadores	Taxa de Acerto Maus-pagadores
Treinamento	63/70 = 90,00%	58/60 = 96,67%	5/10 = 50,0%
Teste	24/30 = 80,00%	24/26 = 92,31%	0/4 = 0,0%

Uma informação importante é saber se o modelo consegue predizer com capacidade maior do que simplesmente classificar os clientes por chances, ou seja, atribuir todos os clientes à classe com maior freqüência. Para tamanhos amostrais diferentes nas categorias de cada grupo, Hair et al. (2005, p. 226) sugere que se utilize o critério de chances proporcionais, o qual é obtido através da equação (2), que deverá ser calculado na amostra de teste:

$$C = p^2 + (1-p)^2. \qquad (2)$$

Na equação (2), considere *p*=proporção de adimplentes (bons-pagadores) na amostra, e (1-*p*)=proporção de inadimplentes (maus-pagadores) na amostra. Para o problema em estudo, tem-se: *p*=26/30=86,67%. Logo, a equação (2) resultará em:

$$C = (0{,}8667)^2 + (1-0{,}8667)^2 = 76{,}89\%.$$

Portanto, um modelo para ser considerado adequado deverá apresentar uma capacidade preditiva superior à proporção por chances de 76,89%. Analisando a Taxa de Acerto Total da Tabela 2, nota-se que o modelo construído com o algoritmo CHAID superam esse valor, e classificam corretamente pelo menos 80,00% dos clientes do conjunto de testes. Portanto, pode-se considerar que o modelo apresenta boa capacidade preditiva. A árvore gerada pelo Modelo 3 está disposta na Figura 1.

Note na Figura 1 que excluindo as variáveis Caução e Motos, o algoritmo apresenta como mais significativas para a classificação do locatário as variáveis: Outros Rendimentos do Locatário, com pvalor=0,0402, e Tipo de Imóvel, com pvalor=0,005.

Analisando mais detalhadamente a Figura 1 pode-se concluir que se um cliente apresentar Outros Rendimentos <= 610,310, ele deverá ter 91,071% de chances de pertencer à classe b. Portanto, a regra de decisão é classificá-lo como pertencente à classe b, dos bons-pagadores. Já se seus Outros Rendimentos forem superiores a esse valor, e o Tipo de Imóvel que se pretende alugar for um Apartamento, o cliente terá 100,00% de chance de pertencer à classe b. Entretanto, se o Tipo de Imóvel for uma Residência, o mesmo cliente terá 71,429% de chance de pertencer à classe m, dos maus-pagadores.

Conclui-se que o algoritmo CHAID apresentou melhor resultado que o CHAID-Exaustivo, contrariando as expectativas. E o melhor modelo forneceu uma capacidade preditiva de 80,00%, superando em 3,11% a capacidade preditiva por chances proporcionais, o que indica que a técnica apresentou resultados satisfatórios, e que quando aplicados pelos tomadores de decisão da imobiliária, poderão garantir 80% de probabilidade de avaliar corretamente a potencialidade de um cliente vir a ser um bom ou mau-pagador. Desta forma, a imobiliária poderá recusar a finalização de contratos com clientes classificados como maus-pagadores, evitando assim, que tenha prejuízos com o não pagamento de aluguéis de imóveis.

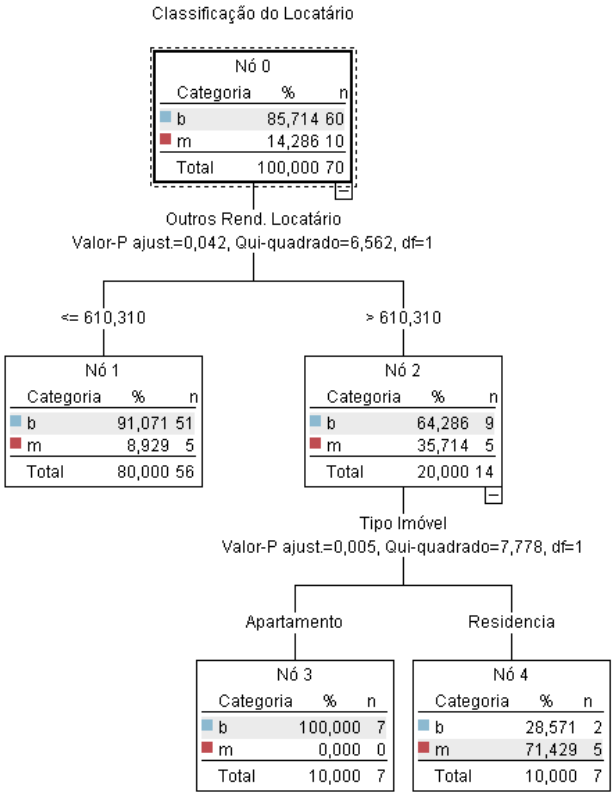

Figura 1 – Arvore de classificação CHAID gerada pelo software Clementine, utilizando-se todas as variáveis, exceto Caução e Motos

REFERÊNCIAS

KAAS, G. V. An exploratory technique for investigation large quantities for categorical data. Statistical, 29, 2 : 119-127, 1980.

HAIR Jr., J. F.; ANDERSON, R. E.; TATHAM, R. L., BLACK, W. C. Análise multivariada de dados. Tradução: Adonai Schlup Sant´Anna e Anselmo Chaves Neto. 5a. ed. Porto Alegre: Bookman, 2005.

SANTOS, A. R.; OLIVEIRA, L. A. Comparação entre os algoritmos CHAID, CHAID-Exaustivo, CART e QUEST para dados com variável resposta categórica nominal via simulação bootstrap. no. 180. Jul 2007. 40 f. Relatório Técnico do Departamento de Estatística – UFSCar.

SPSS BRASIL. São Paulo: 2007. Apresenta informações sobre o software SPSS CLEMENTINE. Disponível em: <http://www.spss.com.br/clementine/index.htm>. Acesso em: 30 mar. 2009.