UTILIZAÇÃO DE DATAMINING NO COMÉRCIO ELETRÔNICO COMO FORMA DE CRIAÇÃO DE VALOR

Banca examinadora

Prof. Orientador: Wilton Oliveira Bussab

Prof. Abraham Sicsu Laredo

Prof. Rubens da Costa Santos

FUNDAÇÃO GETÚLIO VARGAS

ESCOLA DE ADMINISTRAÇÃO DE EMPRESAS DE SÃO PAULO

FÁBIO CORRÊA DA SILVA

UTILIZAÇÃO DE DATAMINING NO COMÉRCIO ELETRÔNICO COMO FORMA DE CRIAÇÃO DE VALOR

Dissertação apresentada ao Curso de Pós-Graduação da FGV/EAESP

Área de concentração: Operações e Tecnologia da Informação

como requisito para obtenção de título de Mestre em Admininstração

SUMÁRIO

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

CAPÍTULO 1 - INTRODUÇÃO

Resumo

Cada vez mais as empresas vêm enfrentando níveis de concorrência mais altos, menores margens nas vendas e perda de diferenciação de seus produtos e serviços. Por outro lado, o desenvolvimento tecnológico e particularmente a Internet vêm permitindo às empresas ter um contato próximo com seus clientes. Este mesmo desenvolvimento também permite que todos os dados, operações e transações empresas sejam armazenadas, fazendo com que maioria delas possuam grandes bancos de dados.

Este trabalho tem por objetivo:

- Entender quais as tendências do Marketing e como o surgimento de novos canais como a Internet podem alavancar a posição competitiva das empresas;
- Constatar a grande quantidade de dados que as empresas possuem, e qual o potencial de informações contida nos mesmos;
- Mostrar como Datamining pode auxiliar na descoberta de informações contidas em meio ao emaranhado de dados,
- Desenvolvimento de metodologia em um estudo de caso prático, onde se buscará determinar que clientes têm maiores probabilidades de compra, e quais seus padrões de consumoatravés de um estudo de caso, e como a empresa pode usufruir destas informações;
- Mostrar como as empresas podem usufruir das informações descobertas e quais são os valores econômico e estratégico.

Um Mundo em Transformação

Vê-se hoje nos negócios profundas transformações, onde pode-se destacar:

- Os produtos e serviços estão cada vez menos diferenciáveis, ocorrendo portanto uma brutal redução de margens e pouca fidelidade dos clientes. Dentro desta perspectiva torna-se uma questão de sobrevivência que as empresas conheçam cada vez melhor seus clientes, de modo que possam oferecer produtos e serviços quase que feitos sob medida, e concentrar seus esforços naqueles mais lucrativos.
- A maioria das organizações vêm nos dias de hoje acumulando uma quantidade expressiva de dados. O desenvolvimento tecnológico não só vem exponencialmente aumentando a capacidade operacional dos computadores (Lei de Moore[1]), como também vem diminuindo substancialmente o preço destes equipamentos. Este fato faz com que virtualmente toda e qualquer empresa tenha acesso a equipamentos e benefícios da tecnologia da informação. De fato, as grandes empresas têm gasto grandes quantias em sistemas ERP (SAP, Baan, Datasul, etc). Uma vez implementados, estes sistemas consolidam de forma estruturada toda a informação corporativa da empresa.
- A Internet, que além de trazer o cliente para dentro da empresa, torna possível o registro de todas as transações que realizam, consolidando-os em uma base de dados que posteriormente pode ser utilizada para pesquisa;
- Esta base de dados contem informações valiosas, estruturadas, muito dispersas e frequentemente pouco exploradas.

O valor destes dados só se torna útil quando podem ser transformados e entendidos, a saber:

- Informação: Arranjo lógico dos dados que permita a um especialista utilizá-lo em sua tomada de decisão. Por exemplo, a informação da venda de um produto específico por região pode auxiliar um profissional de marketing avaliar diferentes alternativas e

---

[1] Em 1975 Gordon Moore (fundador da Intel) previu que a capacidade computacional dos computadores dobraria a cada 18 meses. Desde então este fato se verifica, e ficou conhecido por "Lei de Moore"

decidir pela melhor estratégia para a divulgação do produto. A informação portanto é a *possibilidade de acesso aos fatos que se deseja conhecer a priori*;

- Conhecimento: Entendimento de como os dados e/ou informação relacionam-se entre si. Por exemplo, a análise de agrupamentos de grupos por afinidade (clusters) revela ao especialista os diferentes perfis de grupo. A análise dos carrinhos de compra de supermercado à luz da análise de agrupamentos permite ao gerente descobrir que o público da loja é basicamente composto por clientes do tipo "geração saúde", os que compram artigos finos, e os que compram artigos básicos, *sem que soubesse a priori que grupos seriam esses*. Esta *descoberta* permite ao gerente obter informações (porcentagem de clientes em cada grupo) para tomar decisões tais como o ajuste do mix de produtos a ser ofertado naquela loja. O conhecimento, portanto, é a *possibilidade de descoberta de padrões/informações não conhecidos a priori.*

Somente quando os dados são convertidos nestas duas formas é que permitem a sua plena assimilação, possibilitando a tomada de ações concretas dentro da a organização. Quanto maior o nível de informação e conhecimento, mais acertadas deverão ser as decisões tomadas, melhorando o desempenho da empresa. O desempenho da empresa é derivado da melhora do valor de agregado da informação ou conhecimento.

Neste ponto, algumas questões surgem:

- Como podem as organizações conhecer melhor seus consumidores?
- Como alocar recursos mais eficientemente?
- Qual o valor destas informações? Como impactam no resultado da empresa?

A ferramenta de datamining pode ajudar em muito na busca de soluções para estas questões. Por datamining entenda-se o conjunto de ferramentas estatísticas e de modelagem, que em linhas gerais facilita o acesso à informação e busca revelar padrões e relações ocultos nos bancos de dados das organizações. Por exemplo, a utilização de datamining em database marketing permite não só a redução de custos ao restringir o número dos destinatários de mala-direta, como também aumenta as chances retorno dos destinatários.

Todos estes passos serão abordados no estudo de caso com dados reais. O estudo de um caso terá por objetivo percorrer todos os passos necessários para um projeto de datamining, e trazer à tona toda a sorte de questionamentos e de decisões que ocorrem

durante o processo. O produto final deste caso será a apresentação de uma metodologia capaz de fazer previsões de compra de produtos e de descobrir padrões entre os clientes.

Por fim, a técnica datamining pode viabilizar o marketing one to one, maximizar as informações dos sistemas ERP e efetivamente utilizando plenamente a capacidade geradora de dados da Internet.

Tendências de Marketing

Kotler (1999) constata que a gerência de marketing aponta como problemas principais hoje em dia alguns pontos específicos a respeito de seus clientes, que podem ser resumidos por:

- Comoditização (perda da diferenciação dos produtos com relação aos concorrentes): Cada vez mais os fornecedores se equiparam, fazendo com que os compradores sejam cada vez mais sensíveis à marca do fornecedor, e mais sensíveis a preço, resultando numa menor fidelidade aos fornecedores;
- Sofisticação: Os compradores estão cada vez mais sofisticados, tendo altas expectativas quanto a serviços, atendimento e conveniência. Também, normalmente não têm tempo a perder.

Ao conduzir uma investigação junto aos gerentes acerca do desempenho de suas ferramentas de marketing, Kotler (1999) também descobre que:

- Os produtos vendidos de fato não diferiam muito dos de seus concorrentes, além de terem os preços rapidamente igualados pela concorrência, confirmando a comoditização;
- Elevação geral do custo de vendas (fornecimento gratuito de serviços, brindes caros, propaganda mais cara e menos eficaz), provavelmente fruto da maior sofisticação dos clientes.

Como reação a estas mudanças, o papel desempenhado pelo Marketing vem mudando ultimamente, conforme mostra a Tabela 1.

Tabela 1 – Mudanças no papel do Marketing

DE...	...PARA
Conquista de clientes	Manutenção dos clientes
Lucro em cada transação	Gerenciamento do valor do tempo de vida do cliente

DE...	...PARA
Vender produto	Compreender e atender as necessidades dos clientes
Planejamento de cada ferramenta de comunicação	Integração das ferramentas de comunicação

Devido a estas mudanças, as empresas necessitam cada vez mais:

- Conhecer seus clientes, como forma de tornar seus produtos, condições e serviços mais flexíveis;
- Identificar os clientes mais lucrativos, a fim de estabelecer diferentes níveis de atendimento;
- Inventar novas maneiras de criar e transmitir valor a seus clientes alvos, como forma de prestar melhores serviços e corresponder à crescente sofisticação de seus clientes
- Utilizar melhor os canais de distribuição, tornando o contato com o cliete mais efetivo e econômico.

Dentro desta realidade, Kotler mostra como o Marketing deve evoluir nos próximos anos:

- Conhecimento dos clientes. A maioria das empresas terá bancos de dados privados contendo informações sobre preferências, necessidades e exigências de seus clientes. Produtos poderão ter suas especificações determinadas on-line, e encomendado diretamente à empresa eletronicamente. O conjunto destas informações, consolidadas em uma base de dados, conterá implicitamente as preferências e os diferentes perfis de clientes;
- Carteira de clientes. Mais do que participação no mercado, as empresas tenderão a desenvolver a idéia de carteira de clientes, onde o alvo serão aqueles clientes que a empresa pode atender mais eficientemente (superando expectativas) e obter uma maior lucratividade;
- Retenção. O sucesso na superação de expectativas dos clientes a partir do melhor conhecimento dos mesmos, implicará em maior retenção de seus clientes. Ficará, portanto, ainda mais caro e difícil conquistar novos clientes;

Dado estes fatos, fica claro que somente aquelas empresas que conhecerem seus clientes serão capazes de sobreviver. Para que conheçam seus clientes, é necessário que lancem

mão de todas as informações disponíveis, e que possam também adquirir conhecimento destas informações.

Duas molas mestras serão responsáveis pela concretização destas tendências: Internet e comércio eletrônico.

Internet e Comércio Eletrônico

A grande força da Internet é sua abrangência. De acordo com dados obtidos pela Nielsen e Intelliquest, no começo de 1999 haviam 150 milhões de pessoas conectadas à rede no mundo (87 milhões nos Estados Unidos). Em 1996 eram 35 milhões, e em 1997 eram 58 milhões. Segundo o jornal Folha de São Paulo do dia 5 de Março de 2000, o Brasil possui 6 milhões de internautas e ocupa o sexto lugar no planeta em número de usuários de Internet.

Dados recentes sugerem que o número de pessoas utilizando a Internet para fins comerciais está crescendo mais rapidamente que o número de pessoas que entram na Internet. De acordo com dados obtidos pela Intelliquest, em 1998, 63% dos usuários americanos compraram através da Internet, e 81% de todos os usuários pretendiam comprar eletronicamente em 1999.

O potencial de consumo dos usuários de Internet é alto. Bernoff (1998) em artigo da Forrester Research é mostrado que o perfil do comprador através da Internet é bastante privilegiado:

- os 5% mais ricos da população americana compram atualmente pela Internet;
- 60% dos compradores têm renda anual superior a US$50 mil (o valor da mediana é superior a US$ 59 mil);

O crescimento esperado para o comércio eletrônico é gigantesco. Segundo um outro estudo realizado em 1998 pela Forrester Research (www.forrester.com), as previsões são de que o comércio varejista cresça de US$7.8 bilhões em 1998 para US$ 327 bilhões em 2002. Ou seja, espera-se que este montante em 2002 represente por volta de 10% de todo o mercado varejista.

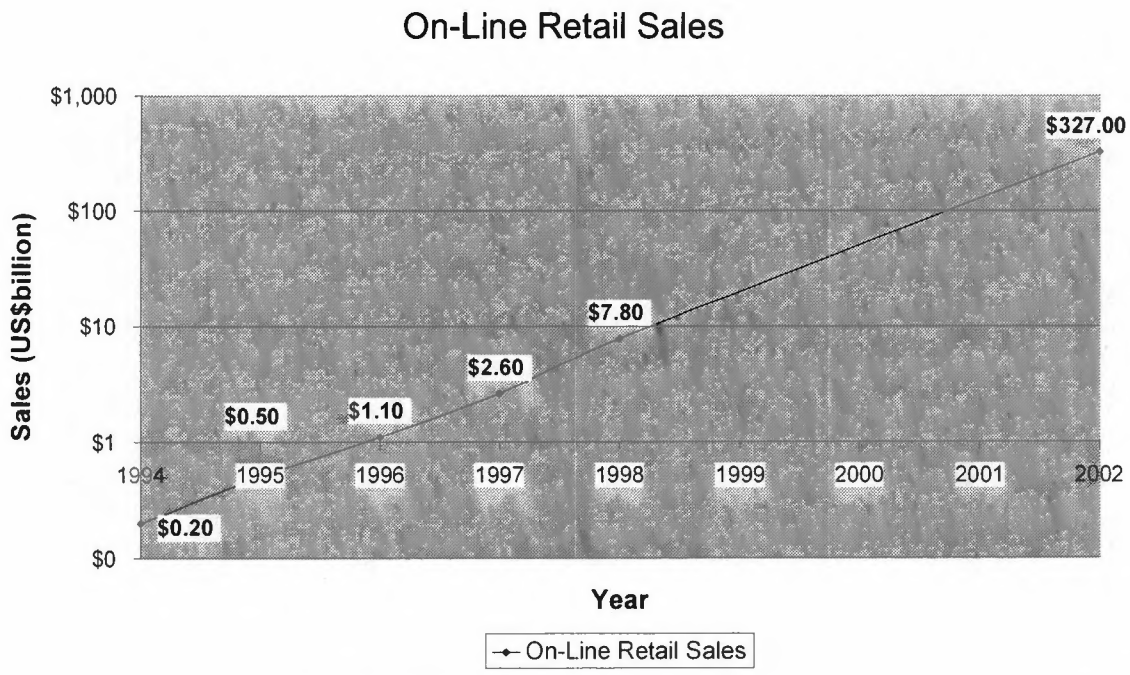


Figura 1 – Projeção de Crescimento do Comércio Eletrônico Business to Consumer (Varejo)

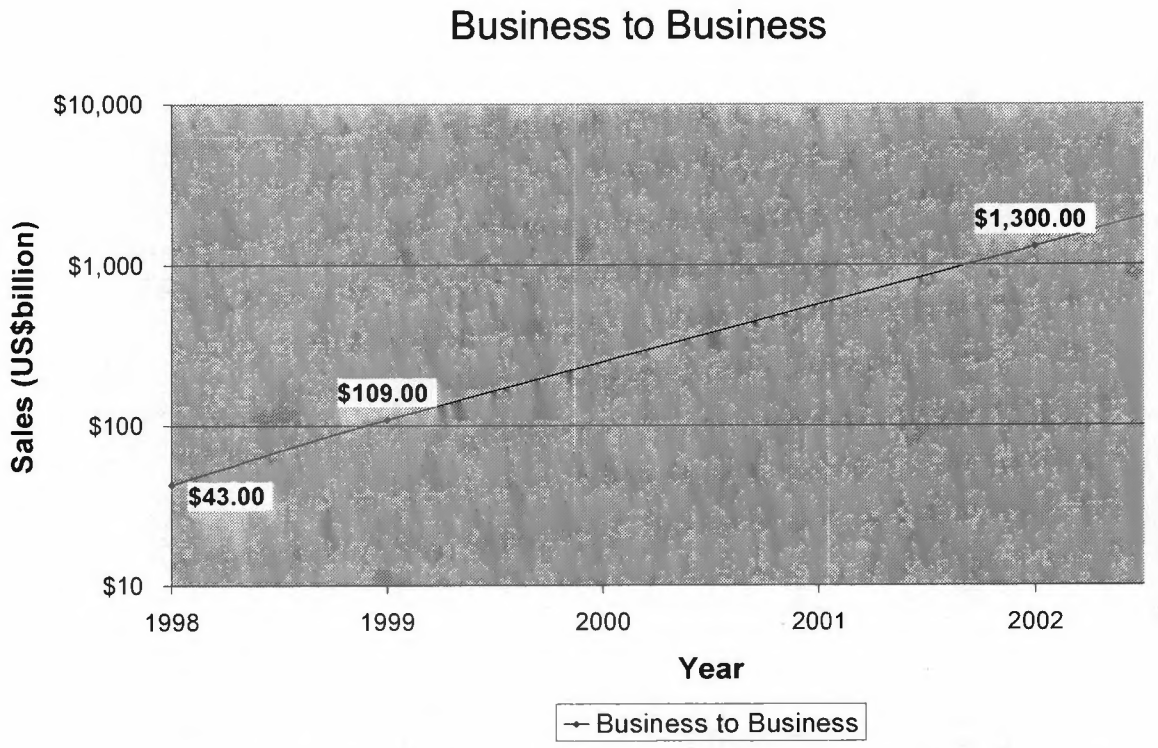


Figura 2 – Projeção de Crescimento do Comércio Eletrônico Business to Business (Corporativo)

Alternativamente, a Tabela 2 apresenta uma estimativa feita pelo Goldman Sachs do crescimento do comércio eletrônico nos EUA e América Latina.

Tabela 2 – Crescimento do Comércio Eletrônico nos EUA e América Latina

Ano	EUA			Brasil + Argentina + México		
	PIB	BtoB	% do PIB	PIB	BtoB	% do PIB
1997	7500			1400		
1998	7763	39	0.5%	1456		
1999	8034	114	1.4%	1514		
2000	8315	293	3.5%	1575	8	0.5%
2001	8606	521	6.1%	1638	23	1.4%
2002	8908	781	8.8%	1703	60	3.5%
2003	9219	1112	12.1%	1771	107	6.1%
2004	9542	1500	15.7%	1842	162	8.8%

Valores em bilhões de dólares
Fonte: Goldman Sachs

Previsões apontam que o comércio eletrônico corporativo responderá por aproximadamente 80% das transações da Internet.

Vale a pena lembrar que além das inúmeras oportunidades de vendas deste novo canal de vendas em função das taxas estonteantes de crescimento do comércio eletrônico (tanto o varejista como também o corporativo), os fornecedores passarão a ter contato direto com seus clientes. Terão, portanto, contabilizadas cada uma das transações realizadas com cada um de seus clientes.

Estes números apontam para inúmeras oportunidades e desafios:

- Somente as empresas que tiverem excelência no atendimento e serviço sobreviverão, caso contrário os compradores trocarão de fornecedores sem maiores problemas;
- Os compradores poderão comprar direto dos fornecedores. O fornecedor terá agora contato direto com o comprador ao invés de existir um intermediário, estreitando o relacionamento entre as partes e contribuindo para um melhor conhecimento dos clientes por parte dos fornecedores;
- Para que seja possível suportar as transações eletrônicas, as partes terão de investir em tecnologia, e naturalmente serão criadas bases de dados (enormes) com o histórico das transações de cada cliente. Estas bases de dados conterão importantes informações dos clientes que poderão ser processadas em busca de informações importantes como quais são os clientes mais rentáveis, os que compram mais, como compram, etc.;

- As bases de dados criadas existirão a um custo marginal muito baixo, dado que a existência delas será em função (pelo menos inicialmente) das transações comerciais;

Se estas oportunidades criadas pela Internet forem cruzadas com as previsões de Kotler discutidas anteriormente, notar-se-á que elas se superpõem. Isto mostra que a realidade da Internet e comércio eletrônico será um dos ambientes naturais para o futuro do Marketing.

Como tirar proveito destas oportunidades oriundas do comércio eletrônico? Como transformar os dados transacionais em informações gerenciais? Como utilizar estas informações? Que resultados podem ser obtidos?

Os capítulos seguintes terão por objetivo responder a estas questões. O Capítulo 3 (página 19) fará uma discussão de como são os sistemas de informação das organizações, assim como uma breve discussão de como métodos matemáticos e estatísticos podem ajudar a extrair conhecimento destas informações. A partir do Capítulo 4 (página 39), começará o estudo de um caso real onde serão desenvolvidos modelos de previsão de compras (Capítulo 5, página 43) e de modelos que ajudam a entender melhor os diferentes perfis de clientes através de suas transações (Capítulo 6, página 74). Estes modelos terão suas validades e poder de previsão estudados.

No Capítulo 7 (página 94) será feita uma discussão de como estes modelos podem ser aplicados no mundo real.

CAPÍTULO 3 - SISTEMAS DE INFORMAÇÃO E GERAÇÃO DE CONHECIMENTO

Nos dias de hoje adventos tecnológicos como sistemas ERP, database marketing e Internet tendem a gerar um volume de dados gigantesco. Com a queda dos custos dos computadores e programas, e o assombroso aumento do poder de processamento e estocagem dos dados, a maioria das empresas acabam por desenvolver uma volumosa base de dados. Uma vez que as empresas necessitam destes dados para empreender as atividades do dia a dia, o desenvolvimento de uma base de dados secundária [2]que possa dar suporte a outras áreas da empresa tende a ter um custo de desenvolvimento marginal, na maioria das vezes, muito baixo. O utilização das bases de dados secundárias para datamining justifica-se por geralmente conter informações preciosas para os negócios das empresas.

[2] A base de dados primária é desenvolvida com o objetivo específico de se estudar determinado fenômeno ou validar uma hipótese. A base de dados secundária refere-se a uma base de dados previamente existente que é utilizada para uma pesquisa, sem que tenha sido criada especificamente par aeste propósito.

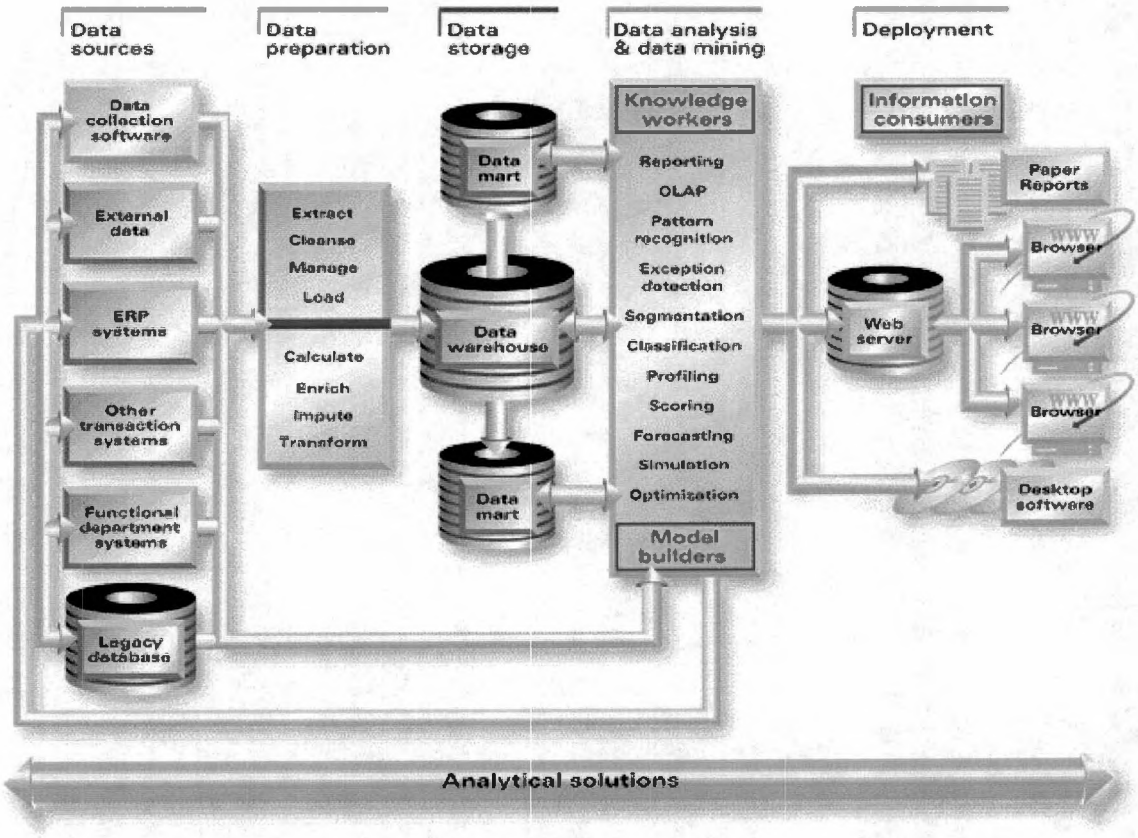


Figura 3 – Estrutura Genérica de Sistemas de Informação

A figura acima apresenta uma visão geral dos Sistemas de Informação de uma empresa., que em linhas gerais, pode ser assim descrita nas etapas seguintes.

Captação dos dados

Os dados são originam-se em sistemas transacionais (registro de compras no supermercados, emissão de bilhetes aéreos), em sistemas ERP que registram todas as transações operacionais das empresas, sistemas contábeis e outras fontes de dados. Estes dados são processados e armazenados em bancos de dados. A empresa pode ter acesso a estes dados por meio da geração de consultas ou geração de relatórios a partir do banco de dados. Normalmente este tipo de consulta exige um certo conhecimento técnico, pois deve-se ser capaz de digitar vários comandos em

linguagens específica de bancos de dados (como o SQL[3]), não sendo raro ter que consolidar previamente dados de várias bases de dados.

Banco de dados

Banco de dados é a estrutura sobre as quais todas as aplicações de sistemas de informações se sustentam. É responsável pelo armazenamento e recuperação dos dados de forma segura e rápida.

Os dados são armazenados em tabelas, que são a coleção de dados de um tópico específico (produtos, clientes, pedidos). As tabelas são organizadas em colunas (também denominadas campos ou ainda, variáveis), e em linhas (também denominadas registros).

Os campos contém informações comuns ao tópico da tabela. Por exemplo, se o tópico fosse Fornecedores, campos como Nome do Fornecedor, razão social, CGC, endereco, telefone, contato, serão informações comuns a todos os fornecedores.

Posteriormente durante as fases de modelagem deste trabalho passar-se á a referir aos campos por variáveis. A razão disto é que a modelagem encara os registros como eventos ou experiências distintas cujos campos são as variáveis de cada experiência. Entretanto do ponto de vista da estrutura de armazemamento da informação em tabelas, campos e variáveis são equivalentes.

Os registros são ocorrências ou eventos de uma tabela. Ou seja, ainda de acordo com o exemplo da tabela de Fornecedores, um registro seria o fornecedor A com CGC 999.999.999, endereço X, telefone (999) 999-9999, ... Outro registro seria uma ocorrência diferente, por exemplo o fornecedor B com seus respectivos dados cadastrais.

Os atuais bancos de dados relacionais permitem que duas tabelas que representam tópicos diferentes sejam relacionadas através de campos em comum. Por exemplo, poderia existir uma outra tabela cujo tópico Produtos que tivesse os campos Nome do Produto, Quantidade, Posição do estoque e Nome do Fornecedor. Esta tabela se

---

[3] SQL (Structured Query Language) – Linguagem utilizada para consultas, atualizações e

relacionaria com a de Fornecedores através do campo Nome de Fornecedor comum às duas tabelas. Esta característica permite também que novas representações de dados sejam feitas através da criação de novas tabelas que possuam registros em comum. Este recurso foi utilizado no estudo de caso deste trabalho, através do campo CUST_NUM comum às tabelas de Amostra do Arquivo de Pedidos e Amostra do Arquivo de dados Cadastrais (ver APÊNDICE A - e APÊNDICE B - páginas 98 e 99 respectivamente).

Um outro ponto importante é que com o relacionamento entre tabelas nos bancos de dados relacionais, torna-se possível representar a informação sem que haja redundância de dados. Ou seja, a mesma informação encontra-se em apenas um local do banco de dados, na tabela do tópico correspondente. A alteração do dado em somente um local fará com que todas as instâncias a ele referenciadas (através dos relacionamentos) sejam imediatamente atualizadas.

Datawarehouse

Nesta etapa a empresa pode se beneficiar, caso tenha implementado, de um sistema de datawarehouse. Este sistema nada mais é do que um gerenciador dos diferentes sistemas de informação e bancos de dados da empresa, integrando-os e fazendo com que pareçam uma base de dados única.

Este sistema é responsável pela:

- consolidação de todos os dados existentes: integra não só os bancos de dados de diversos tipos (hierárquicos, relacionais) e diferentes fabricantes, assim como diferentes bases de dados que a empresa possa possuir;
- resolução de problemas de integridade dos dados: verifica, filtra e valida os dados, eliminando por exemplo que campos como sexo tenham mais que dois valores possíveis (masculino e feminino), ou campos de estado da federação tenham siglas inexistentes;

---

gerenciamento de bancos de dados relacionais.

- espelhamento[4] e sincronização[5]: sincroniza os registros cujas bases de dados encontram-se fisicamente fragmentadas ou espelhadas, não permitindo que dois registros relativos à mesma informação possuam valores diferentes.

Datamarts

A utilização dos dados consolidados pelo data warehouse se dá através dos datamarts, que são uma representação dos dados para consultas específicas, que somente possui os dados relativos àquela tarefa. Por exemplo, as informações financeiras (receitas, empréstimos, etc.) da empresa estarão concentradas no datamart de finanças, cujo conteúdo será em sua maioria diferente do datamart de marketing (vendas, clientes,etc.).

As vantagens da utilização dos datamarts são:

- permitem ganhos de eficiência nas consultas, pois as buscas geradas por um comando de consulta passam a se restringir ao datamart em questão;
- a informação contida em um datamart já tem resolvidos todos os problemas de consolidação de diferentes bases de dados tais como espelhamento, sincronização e integridade;
- transparência e uniformidade na apresentação da informação;
- segurança e sigilo, uma vez que o usuário de um datamart não tem acesso (a menos que seja autorizado) às informações de outro datamart.

OLAP (On Line Analytical Processing)

Sistemas OLAP são sistemas que facilitam enormemente o acesso e visualização dos dados. Normalmente é implementado uma vez que toda a parte de consolidação de dados de um datawarehouse já se encontra implementado. A consulta a bancos de dados, embora não seja um problema tecnológico, normalmente envolve um

---

[4] Espelhamento de dados – técnica através da qual um determinado conjunto de dados é copiado localmente para ser utilizado em tarefas específicas, vizando a eficiência do sistema.

profissional técnico qualificado, pois exige a programação do mesmo. Quando se utiliza um sistema OLAP, as consultas e visualização de dados podem ser feitas muito facilmente pela gerência através da utilização de planilhas e softwares específicos, dispensando o envolvimento direto de um técnico especializado. Sistemas OLAP agregam valor às empresas ao oferecer a possibilidade de acesso aos fatos que se deseja conhecer a priori. Por exemplo, um executivo pode por ele mesmo descobrir que produtos a empresa vendeu por região a cada trimestre, e com um simples apertar de botões diferenciar aqueles que garantiram maior lucratividade daqueles não tão lucrativos, e ainda investigar as estruturas de custos dos menos rentáveis. É portanto muito útil como ferramenta de exploração da base de dados, permitindo ao usuário entendê-la e explorá-la. Note entretanto que o usuário precisa saber aquilo que procura para então ter acesso à informação desejada.

Em outras palavras, o usuário de OLAP pode gerar uma série de hipóteses na investigação de determinado fato, e confrontá-las com informações obtidas através da consulta ao banco de dados, tentando verificar se são ou não corretas. Entretanto, quando o número de variáveis é muito grande, fica cada vez mais difícil e demorado conceber e testar diferentes hipóteses.

Datamining

Segundo o Gartner Group, 1995, datamining é o "processo de descoberta de novas correlações, padrões e tendências significativas por análise minuciosa de grandes porções de dados armazenados em bancos de dados, usando tecnologias de reconhecimento de padrões e técnicas estatísticas e matemáticas".

De acordo com HAIR et al. (1998, página 674), o termo datamining trata-se da "procura por relação e padrões de dados em grandes bancos de dados. Como o próprio termo sugere, datamining tem uma orientação exploratória para a busca do conhecimento obscurecido pelos complexos padrões de associação e pelas grandes massas de dados.".

---

[5] Sincronização de dados: Técnica através da qual garante-se que os dados espelhados nas diferentes bases de dados possuam a mesma informação.

O Instituto SAS, que produz o software estatístico padrão dos mundos corporativo e acadêmico, define datamining como "Processo de seleção, exploração, modificação e modelagem de grandes bases de dados para a descoberta padrões previamente desconhecidos contidos nestes dados, com o fim de gerar vantagem competitiva nos negócios".

Datamining refere-se portanto a um conjunto de ferramentas e técnicas que visam fazer aparecer padrões e relações a partir de grandes basses de dados. Este conjunto de ferramentas compõe-se de técnicas básicas tradicionais e antigas (Análise de dados Multivariados: Regressão Linear, Análise de Discriminante, Conjoint Analysis, Agrupamentos), assim como técnicas mais modernas como Redes Neurais e algoritmos genéticos.

Embora muitas destas técnicas sejam antigas, a utilização em grandes bases de dados não era viável tecnológica e economicamente em função da complexidade de processamento e memória requeridos pelos algoritmos. Somente com o aumento do poder computacional aliado ao baixo custo dos computadores recentes é que a prática de datamining começou a se tornar economicamente viável.

A motivação de uso no datamining é diferente tanto tecnicamente quanto como é encarada a sua utilização nos negócios.

Tradicionalmente os pesquisadores tratavam de problemas cuja formulação era bastante específica, normalmente na forma de uma hipótese (Qual o perfil dos consumidores do produto X? Como a compra de produtos Z está relacionada com a compra de produtos X?), onde era possível planejar um detalhado plano de pesquisa. O critério de validação dos resultados era baseado na validação ou não de hipóteses, que definiam em que direção outros experimentos preparados deveriam ou não ser conduzidos.

A motivação do datamining é exploratória, ou seja, é mais orientada à análise dos dados existentes no sentido que o pesquisador na maioria das vezes estará tratando de questões bem especificadas mas mais gerais (Que padrões de compra dos clientes de uma determinada loja, se existirem, estarão associados à compra do produto X?). Ou seja, sabe-se exatamente o que se procura, mas as hipóteses ou são desconhecidas, ou o número de hipóteses é proibitivamente grande para testá-las através de

experimentos. O critério de validação dos resultados é agora guiado pelo desenvolvimento modelos, e mensurando-se o grau de previsão com base em comparações com casos conhecidos através de uma amostra para validação.

A validação de resultados passa a desempenhar um papel fundamental, uma vez que fenômenos como overtraining (perda da capacidade de generalização do algoritmo quando aplicado num base de dados diferente da que foi treinado) tornam-se comuns. Deve-se lembrar que o teste de hipóteses ainda continua válido como forma de verificação se o grau de associação entre resultados obtidos e os esperados são consistentes, conforme será discutido no caso de estudo em Teste Qui-Quadrado (página 59).

A motivação para utilização nos negócios faz com que o enfoque da análise dos resultados seja direcionado muito mais para a utilidade dos mesmos do que para detalhes operacionais do processo algoritmico. Questões como "O que pode ser feito com os resultados obtidos?", "Que ações podem ser tomadas em finanças, marketing, vendas a partir dos resultados obtidos?", tornam-se mais importantes do que questões técnicas do tipo "Qual o melhor modelo?", "Sob que circunstâncias um modelo é melhor do que o outro?". Assim, não se espera que os modelos desenvolvidos sejam "perfeitos" e academicamente esmiuçados, mas sim que sejam úteis e confiáveis para o processo de tomada de decisão, e que auxiliem na descoberta de novas perspectivas do negócio. Os modelos devem de fato servir de suporte para aplicações do tipo:

- Mala direta e telemarketing: desenvolvimento de modelos para predição de resposta à mala direta, identificação dos clientes com potencial de maiores vendas, margens e receitas;
- Segmentação de mercado: identificação de segmentos de mercado menos óbvios;
- Programas de fidelidade: desenvolvimento de modelos de retenção;
- Identificação de Mix de produtos e promoções: podem ser desenvolvidos a partir do entendimento dos diferentes perfis de compra dos clientes;
- Descoberta de novas oportunidades no mercado.

Comparação de Datamining e OLAP

Datamining diferencia-se de OLAP na medida que mais do que verificar padrões hipotéticos, utiliza-se dos dados para trazer à tona tais padrões. Por exemplo, suponha um analista de risco que queira descobrir os fatores de risco para o não pagamento de crédito. A ferramenta de datamining, além de descobrir padrões triviais como credores com alto débito e com baixa renda e portanto potencialmente arriscados, poderia também descobrir um padrão menos intuitivo para o analista, que levasse em consideração outros fatores não triviais tais como idade, local de moradia e atividade profissional.

Embora diferentes, percebe-se que as duas ferramentas se complementam. A ferramenta OLAP oferece ao analista um entendimento geral que se revela importantíssimo para identificar as variáveis importantes, exceções e interações. Quanto mais o analista entender a base de dados, mais eficiente ele será no processo de descobrir novos padrões e interações. Portanto, a ferramenta de datamining não substitui analistas ou gerentes, mas constitui-se antes numa poderosa ferramenta de análise que pode contribuir em muito para o trabalho que eles realizam.

Note que datawarehouse, assim como OLAP não são requisitos para se realizar o datamining. De fato, existem muitos benefícios para a operação de datamining caso os dados tenham como origem um datawarehouse, porque espera-se que já tenham passado por operações em comum como filtragem, consolidação de dados provenientes de múltiplas fontes e preparação para manutenção futura, conforme mostra a Figura 4.

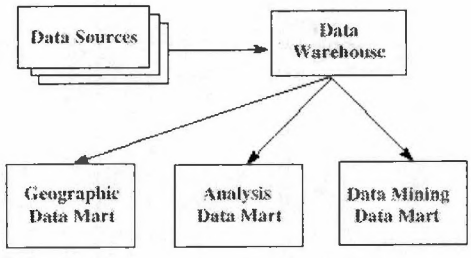


Figura 4 – Implantação do Datamining a partir de um Data Warehouse

Sistemas de datawarehouse como estes podem exigir investimentos de milhões de dólares, bem como alguns anos para implementação. Nestes casos, a garimpagem de dados por datamining diretamente nos dados das bases de dados pode ser realizada, necessitando para isso que as tarefas de integração, consolidação e integridade de dados de diversas fontes sejam realizadas. O produto final deste trabalho será um arquivo sobre o qual será utilizada a ferramenta de datamining, conforme ilustrado na Figura 5.

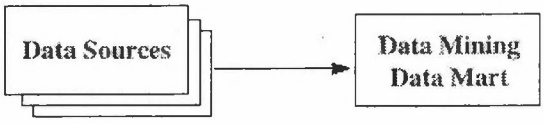


Figura 5 – Implantação do Datamining diretamente a partir dos dados

Ferramentas Estatísticas e Matemáticas Utilizados em Datamining

Datamining encontra padrões e relações através da construções de modelos.

Os modelos podem ser classificados conforme sua função e/ou motivação.

Classificação de modelos com relação a função

Modelos preditivos

Modelos preditivos usam resultados conhecidos para desenvolver modelos que predizem resultados para novos registros. Por exemplo, o histórico de pagamento das pessoas permite construir um modelo que, com base no perfil da atual clientela (renda annual, posse ou não de imóveis, histórico de crédito), ajuda a dizer se novos clientes serão ou não bons pagadores. Em modelos preditivos, os valores que se deseja estimar são chamados de variáveis alvo, resposta ou dependentes. Os valores utilizados para realizar a previsão são chamados de variáveis independentes ou preditores. Podemos citar classificação, regressão e séries temporais como técnicas que produzem modelos preditivos. Estes modelos são também chamados de modelos de aprendizado supervisionado, uma vez que são construídos a partir de dados que já foram classificados.

Modelos descritivos

Modelos descritivos descrevem padrões que ocorrem nos dados atuais, oferecendo diferentes perspectivas das interações das diferentes variáveis. São utilizados na tomada de decisões. Podemos citar análise de agrupamentos, associações e sequências como técnicas que produzem modelos descritivos. Estes modelos são também chamados de aprendizado não-supervisionado, uma vez que estes modelos não necessitam de resultados conhecidos para seu treinamento. Ao contrário, tais modelos é que provêem regras para classificação e identificação de grupos.

Comparação entre modelos preditivos e descritivos

A principal diferença entre os dois tipos de modelos é que enquanto o modelo preditivo oferece uma previsão explícita (lucratividade de um cliente, risco de pagamento de um empréstimo), o modelo descritivo oferece um maior esclarecimento sobre os padrões existentes, ou seja, oferece uma previsão implícita (clientes com determinado perfil normalmente têm uma determinada classificação de risco). O entendimento dos dados a partir com a utilização de modelos descritivos tendem a ajudar na elaboração de um modelo preditivo.

Classificação de modelos com relação à motivação

Usando o critério de fundamentação teórica ou empírica, as ferramentas de modelagem podem ser divididas em dois grupos:

Motivação Teórica

As de motivação teórica, também chamadas de teste de hipóteses, tentam confirmar ou refutar idéias pré-concebidas. Requerem que o pesquisador especifique o modelo e/ou coleta de dados tendo por base um conhecimento prévio, para então testar a validade ou não do modelo. Fazem parte deste conjunto análise de correlação, testes t, ANOVA, Análise de Regressão, Regressão Logística, Análise Discriminante, Previsão em séries temporais, etc.

Motivação a dados

As de motivação a dados, por sua vez, automaticamente criam modelos baseados nos padrões sugeridos pelos dados. Estes modelos também necessitam ser testados para

que sejam considerados válidos. Fazem parte deste conjunto: Análise de Agrupamentos, Análise Fatorial, Árvores de Decisão, Redes Neurais, etc.

As primeiras baseiam-se fortemente no conhecimento substantivo existente na área de aplicação. Nem sempre existem, principalmente em administração. As segundas podem ter sido geradas por associações espúrias, sem nenhuma sustentação teórica, porém são excelentes geradores de idéias a serem exploradas e validadas.

Descrição dos modelos

A modelagem é um processo interativo, onde o modelo final será normalmente uma combinação do conhecimento prévio e de novas informações descobertas.

A seguir são apresentadas características gerais dos diferentes modelos:

Modelos de Classificação

O objetivo destes modelos é identificar, dentre um número finito de classes existentes, aquela cujas caracterísitcas mais se assemelhem com as características de um determinado indivíduo (ou evento). Permite também predizer como novos indivíduos irão se comportar, baseando-se na idéia que dado que foram classificados em uma determinada classe, então estes indivíduos deverão apresentar o mesmo comportamento dos indivíduos daquela classe.

A criação de modelos de classificação depende da existência de casos anteriores já classificados, e indutivamente revelando o padrão preditivo. A classificação prévia pode ser obtida através de experiências já realizadas e coletando-se os resultados, através de registros pré-existentes em bancos de dados. Em alguns casos onde a classificação não é não pode ser inferida diretamente através de um exxperimento (por exemplo, o diagnóstico de doenças), é comum realizar a classificação prévia a partir da contratação de um expert.

Modelos de Regressão

Análise de regressão é uma técnica estatística que analisa a relação entre as variáveis dependentes e uma ou mais variáveis independentes. O objetivo desta análise é obter uma previsão da variável dependente a partir dos valores conhecidos das variáveis independentes. Os pesos das variáveis independentes são ajustados para que se

obtenha o máximo poder de previsão da variável dependente. Os pesos traduzem a contribuição relativa das variáveis independentes para a predição final. Contém, portanto, informações importantes de como cada variável independente se relaciona com a predição, facilitando a interpretação da influência de cada variável. Utiliza-se de valores pré-existentes para predizer quais serão os outros valores. É capaz de tratar problemas lineares e não-lineares.

Maiores detalhes sobre este tipo de análise serão discutidos no caso de estudo, em Nó de Regressão (página 61).

Séries Temporais

Similarmente à regressão, séries temporais utiliza-se de valores pré-existentes no tempo para predizer valores futuros. Baseia-se na premissa de que eventos que ocorreram no passado de alguma maneira podem explicar o futuro. Levam em consideração outros fatores tais como sazonalidade, hierarquia de períodos e aritmética de datas.

Análise de agrupamentos

A análise de agrupamentos divide o banco de dados em diferentes grupos, usando critérios (variáveis) previamente selecionados. O objetivo deste modelo é encontrar grupos cujos membros sejam bastante similares entre si e que os diferentes grupos formados sejam bastante diferentes entre si. Esta análise frequentemente exige a presença de um especialista, que dirá se os agrupamentos fazem ou não sentido. Uma vez obtidos os agrupamentos, eles podem ser utilizados para a classficação de novos registros.

Maiores detalhes sobre este tipo de análise serão discutidos no caso de estudo, em Nó Clustering (página 81).

Análise de Associação

Este modelo procura ocorrências que ocorrem juntas dado um evento. Fornece um conjunto de regras do tipo "Se A ocorrer, existe uma probabilidade x de B ocorrer também". Por exemplo, dado que um cliente comprou iogurte desnatado e queijo com baixo teor de gordura, então existe 85% de probabilidade de comprar também queijo desnatado.

Maiores detalhes sobre este tipo de análise serão discutidos no caso de estudo, em Nós de Associação (página 78).

Seqüência

Seqüências são muito parecidas com associações, exceto pelo fato que as ocorrências tomam lugar no tempo. Por exemplo, se a ação A subir mais que 12% e o índice BOVESPA cair, então a ação B tem 68% de probabilidade de aumentar nos próximos dois dias.

Maiores detalhes sobre este tipo de análise serão discutidos no caso de estudo, em Nós de Associação (página 78).

O Processo De Datamining

Aqui são apresentadas rapidamente as etapas típicas de um processo de datamining. Estas etapas são em termos gerais típicas de um processo de datamining, embora a linha divisória entre elas nem sempre seja clara e definida. É importante ressaltar que este processo é cíclico e iterativo, no sentido que todas as etapas são revistas e ajustadas várias vezes. Esta iteração tem como propósito melhorar o resultado final, procurando-se aumentar a sua utilidade prática e sua consistência a cada iteração.

Todas as etapas serão discutidas em detalhes durante o estudo de caso. Em função do ciclo do processo, somente a configuração final será mostrada, salvo comentários sobre caminhos previamente tomados que não deram certo e que levaram a modificações importantes no processo de datamining.

Identificação precisa do problema.

O primeiro passo é ter muito claro o objetivo a ser alcançado. Será este objetivo que norteará todo o processo, desde a escolha das bases de dados e variáveis até a validação do resultado final.

Suponha que a meta seja aumentar a eficiência de uma mala direta. Qual seria então o real objetivo, aumentar a quantidade de retornos, ou aumentar o valor potencial de cada retorno? O objetivo do problema deve ser capaz de responder a perguntas desta natureza. A enunciação efetiva do que se deseja descobrir é fundamental para o êxito

do processo, assim como deixar claro como deve ser medido o resultado obtido, para fins de comparação entre os modelos existentes e escolha do modelo final.

O objetivo do estudo de caso deste trabalho está definido no Capítulo 4 (página 39). É a partir daquela definição que se todos os passos posteriores foram tomados.

Coleta de dados

A partir da definição do problema, deve-se proceder à identificação das fontes de dados que contenham a informação desejada ou dados que se acredite possuir correlação com a informação desejada. Este processo normalmente envolve várias fontes de dados. Em alguns casos os dados disponíveis podem ou não conter a informação ou ser insuficientes, fazendo-se necessário recorrer a outras bases de dados (outros dados corporativos, bases de dados de terceiros, compra de dados de um bureau de dados).

A implantação de um datawarehouse nesta etapa pode agilizar bastante o processo, uma vez que problemas como integração, consolidação e integridade de dados de diversas fontes (ver Datawarehouse, página 22) já se encontrariam resolvidos. Caso contrário, é nesta etapa que estes problemas deveriam ser resolvidos.

Representação dos dados

A partir da clara enunciação do problema, deve-se partir para a correta representação dos dados. A representação deve ser eficaz de modo que cada registro contemple as informações desejadas na definição do problema.

Esta é a parte mais difícil do processo, pois caso não se consiga uma representação dos dados que capture em cada registro a informação que se deseja descobrir, todos os passos posteriores serão inúteis, não sendo raro ter que redefinir a representação caso os passos posteriores não produzam bons resultados.

A representação dos dados normalmente envolverá a manipulação e pré-processamento dos mesmos.

Por exemplo, um estudo realizado pelo autor na Universidade do Texas junto ao governo do Texas procurava determinar a existência ou não de fraudes de tiquetes de refeição doados à população de baixa renda. O tipo de fraude mais comum era o portador dos tiquetes trocá-los por dinheiro em estabelecimentos comerciais por uma

soma em dinheiro menor do que os tíquetes valiam. Algumas representações de dados foram tentadas sem sucesso. Entretanto, observou-se que os estabelecimentos comuns exibiam um histograma de número de compras realizadas por valor da transação de apenas um pico, enquanto os estabelecimentos suspeitos exibiam para o mesmo histograma dois picos. Este fato justifica-se porque os estabelecimentos suspeitos realizavam várias operações de compra ilegal de tiquetes de US$50.00 e pagavam US$30.00, portanto apresentando um pico adicional em US$30.00 em seu histograma. A solução encontrada e quqe deu resultados foi determinar os picos e vales do histograma de cada estabelecimento comercial (através de filtros ou redes neurais), e criar registros que tinham como campos as coordenadas (valor da transação x número de compras realizadas) para os três primeiros picos e vales. Est representação capturou a informação desejada em cada registro (indicadores de fraude nos máximos e mínimos) e com ela foi possível obter bons resultados.

No Capítulo 5 (*Representação dos dados*, página 44) e Capítulo 6 (Representação dos dados, página 74) será mostrado em detalhes o desenvolvimento da representação dos dados do caso objeto deste estudo.

Seleção Dos Dados

Muitas vezes se faz necessária uma seleção ou transformação na representação dos dados. Por exemplo, em modelos preditivos é necessário selecionar quais as variáveis dependentes e independentes, para que o algoritmo de modelagem seja capaz de funcionar corretamente.

As ferramentas atuais já possuem algoritmos específicos (normalmente lineares) que automaticamente fazem a seleção de variáveis, baseado na correlação com as variáveis dependentes. Entretanto, o entendimento dos dados é fundamental nesta etapa, uma vez que variáveis do tipo número identificador de registro podem ser descartadas previamente, por não carregarem nenhuma informação a não ser o número do respectivo registro. Ou ainda, variáveis que se sabe importantes podem ser incluídas, uma vez que muitas vezes seu relacionamento com a variável dependente pode ser não linear, portanto importante em modelagens não lineares como redes neurais.

O entendimento dos dados é também importante na seleção de registros. Tomando-se como exemplo os outliers, podem constituir casos de interesse prático na pesquisa, e a exclusão dos mesmos pode inibir a descoberta de informações importantes. Ou de outra maneira outliers podem ser dados entrados erroneamente. Somente um especialista que conheça os dados pode fazer o correto julgamento destas situações.

Transformação dos dados

Em algumas ocasiões pode ser vantajoso proceder à transformação de dados:

- Quando se deseja criar preditores mais adequados para o objetivo da pesquisa. Por exemplo, no caso em que se esteja desenvolvendo um modelo para inferir o risco financeiro de empresas, uma variável que fosse a razão entre débito e receita pode ser mais significativa que as duas variáveis isoladamente;
- Certos algoritmos exigem que os dados sejam ajustados em intervalos ou até mesmo em valores binários tais como positivo/falso. No caso de regressão linear, a variável dependente deve ser codificada em verdadeiro/falso. Em outros algoritmos lineares as variáveis dependentes podem precisar ser desmembradas em variáveis "dummy". Neste último caso, os sistemas atuais já possuem algoritmos que realizam esta tarefa automaticamente.

Hair (1998, página 23) mostra um diagrama-resumo com o formato das variáveis utilizados em cada modelo.

Construção De Modelos

De uma maneira geral o primeiro passo na construção de um modelo é a divisão da base de dados em dois ou três conjuntos de dados distintos, um para treinamento, outro para testes e (opcionalmente) outro para validação. O conjunto de dados para treinamento é utilizado para estimar e ajustar os parâmetros do modelo, enquanto o conjunto de testes é utilizado para verificar as predições obtidas pelo modelo treinado. Caso não fosse utilizado dois conjuntos distintos de dados, a acuidade do modelo tenderia a ser superestimada, uma vez que o modelo desenvolvido estaria trabalhando com os mesmos dados com que foi treinado. O terceiro conjunto de dados, conjunto de validação, justifica-se porque alguns algoritmos como o de redes neurais o conjunto de testes é utilizado ao mesmo tempo que o conjunto de

treinamento (o treinamento termina quando o nível de predição para de crescer com o número de iterações), precisando-se assim de um modelo final que dê uma medida independente de sua acuidade.

Um modelo é considerado treinado quando um ciclo completo de treinamento de de testes/validação é conlcuído. Os resultados devem ser então avaliados, para se certificar do poder de previsão, através de medidas como hit ratio, etc. Caso o modelo não tenha uma performance aceitável, todas as etapas anteriores devem ser revistas, e modificadas de maneira a se obter um melhor resultados.

Uma vez que exista um modelo treinado e que forneça resultados satisfatórios, ter-se-á uma estimativa de qual será sua taxa de acerto/desempenho em futuras bases de dados. Não significa entretanto que o modelo esteja correto, mas que se a mesma técnica for utilizada em uma sucessão dados de treinamento/testes similares no futuro, a acuidade média seria muito similar à obtida desta maneira. Portanto fica claro que um modelo deve sofrer retreinametos constantes, para que os parâmetros do modelo reflitam as mudanças que possam haver na evolução dos dados.

As ferramentas atuais de datamining permitem que diferentes modelos sejam testados ao mesmo tempo, conforme pode ser visto na Figura 8 (página 48), assim como a avaliação entre estes diferentes modelos (Figura 14, página 66).

Datamining na Indústria

Empresas que habitualmente lidam internamente com grandes bases de dados já utilizam datamining em suas operações.

Segundo Kotler (1997), desde 1993 a Amex (American Express) vem usando as informações de fatura de seus 35 milhões de associados para (entre outras aplicações) para enviar junto com a fatura ofertas específicas para clientes com determinado perfil. Na Bélgica foi colocado em teste um sistema que propõe ofertas a partir do código postal e do histórico de transações do cliente. Por exemplo, se é aberto um novo restaurante, aqueles clientes que possuam um perfil de jantarem fora e que morem nas imediações do restaurante podem receber uma oferta especial. Aplicações deste tipo têm claro o objetivo de estimular o consumo, aumentando a receita e promovendo a marca da Amex. Outras aplicações possíveis seriam detecção de

fraudes, análise de concessão de limites de despesas para o detentor do cartão. Em todos os casos, deve-se observar que as informações utilizadas pela Amex surgiram a partir de dados que já se encontravam em seu banco de dados, portanto o custo marginal de aquisição dos dados é virtualmente zero.

Um outro exemplo, citado por Kotler (Marketing no Século XXI, página 45), uma empresa líder no setor de vendas por catálogos com 2 milhões de clientes, convidou a IBM para fazer o datamining de seus dados. A IBM foi capaz de identificar 5000 micro-nichos. Por exemplo, descobriu-se que 850 clientes compraram camisa azul e gravata vermelha; provavelmente estes compradores também estariam interessados em um blazer azul, ou outras combinações de camisas e gravatas que combinassem com as anteriormente adquiridas. Aqui, claramente o objetivo é aumentar o índice de resposta dos clientes, ao oferecê-los ofertas sob medida. Acreditavam que pudessem chegar a índices de 10%, quando o tipicamente estes índices variam de 0.1 a 1%.

Datamining vem sendo empregado também em áreas pouco tradicionais no uso de tecnologia da informação. Por exemplo, os capitães e treinadores do time Toronto Raptors, da NBA (National Basketball Association - Liga de Basquete Americana), estão utilizando Datamining para escolher, a cada partida, os jogadores mais capacitados para enfrentar os diferentes adversários (Henry Baltazar, PC WeekLAbs, 16 de Março de 2000).

Aplicação de Datamining Em Comércio Eletrônico

Sob certos aspectos, a Internet se assemelha a venda por catálogos e varejo. Portanto é de se esperar que o sucesso obtido na aplicação de dataming nestes casos possa ser replicado na Internet.

Entretanto, certas caracterísitcas da Internet a fazem um meio muito mais eficiente e interativo que a venda por catálogos:

- Na Internet é possível acompanhar a ordem de compra dos produtos em um site, e sugerir ao final outros produtos que clientes do mesmo perfil geralmente também compram;
- O comprador pode configurar seu produto, e não apenas escolher um que já exista;

- A aquisição dos dados é imediata, gerando uma massa de dados crítica muito rapidamente e já estruturada para processamento;

Empresas como Dell (venda de computadores sob medida para clientes corporativos), Amazon (inicialmente realizava venda de livros, hoje já pode ser considerado um varejista), Yahoo (portal da Internet) e Charles Schwab (banco de investimentos) são exemplos de empresas bilionárias que começaram na Internet ou cujas vendas e negócios atualmente se dão quase que exclusivamente pela Internet.

A partir do próximo capítulo será investigado como um fornecedor de equipamentos pode se beneficiar dos resultados obtidos com o uso de datamining em seus registros de compras.

CAPÍTULO 4 – ESTUDO DE CASO

Este estudo de caso tem por objeto verificar que tipo de informações podem ser obtidas através da exploração da base de dados de um grande fabricante de computadores, e como estes resultados podem ser utilizados para a empresa. A pedido deste fabricante, não utilizaremos seu nome real e passaremos a designá-lo por X.

X é um grande fabricante de computadores, cujas vendas são efetuadas pela Internet. Os clientes de X colocam pedidos eletronicamente através da Internet ou através de sistemas EDI[6], especificando a configuração desejada. A partir destes pedidos, X constrói o computador de acordo com a configuração requisitada, e o entrega ao cliente. Com o tempo, X construiu uma grande base de dados em que se encontram registrados os pedidos colocados por cada cliente, assim como seus dados cadastrais. Estes serão os dados utilizados neste trabalho.

X tem basicamente dois objetivos:

1) saber quais clientes comprarão servidores no quadrimestre seguinte e
2) 2) conhecer quais os hábitos de compras dos clientes, particularmente quais são os padrões que levam à compra de um servidor.

A bases de dados de X utilizadas neste trabalho foram duas:

- Histórico dos pedidos colocados em dois anos de operações;
- Dados cadastrais dos clientes.

Vale a pena notar que:

- Os dados são todos provenientes de comércio eletrônico;
- as bases de dados utilizadas são operacionais, ou seja, é formada por dados que foram armazenados com o propósito de permitir as atividades de produção, comercialização, distribuição e faturamento da empresa. Portanto, não foram desenvolvidas (pelo menos inicialmente) com outro propósito que não fosse permitir X prestar seu serviço;

---

[6] Electronic Data Interchange, serviço eletrônico que implementa o padrão internacional EDIFACT que define sintaxes e comandos para transações eletrônicas entre empresas.

- o custo de coleta das informações para este estudo é muito pequeno, uma vez que a base de dados já existe;
- Embora se conheça claramente os objetivos do estudo, não se sabe a priori que informações podem ser encontradas, onde procurá-las e muito menos se elas existem ou não.

Arquivo do Histórico dos pedidos

Neste arquivo encontram-se os pedidos colocados eletronicamente ao longo de oito quadrimestres consecutivos (dois anos). É constituído por 830.139 registros e composto pelos campos mostrados na Tabela 3.

Tabela 3 – Descrição dos campos do arquivo de pedidos

CAMPO	TIPO	DESCRIÇÃO DO CAMPO
ORD_NUM	Char	Número do pedido
CUST_NUM	Char	Identificação do cliente para pedidos (Comum a esta base de dados e a base de dados cadastrais dos clientes)
TIE_NUM	Char	Tie number
CMPY	Char	Tipo de relação: 04 – Transacional 08 – Relacionamento
FY	Char	Ano fiscal
FQ	Char	Quadrimestre fiscal
FM	Char	Mês fiscal
NET_REV	Num	Receita líquida
PLP	Char	Linha de produto / Descrição / Núm. de referência: 2NT – Periférico – 1 3WS – Workstation – 2 4DM – Modelo 4DM – 3 4ES – Modelo 4ES – 4 4IN – Modelo 4IN – 5 4LT – Modelo 4LT – 6 4PD – Modelo 4PD – 7 4ST – Storage – 8 4SV – Servidor – 9
SYS_QTY	Num	System Quantity

Uma amostra destes dados encontra-se no APÊNDICE A - (página 98).

Dados Cadastrais dos Clientes

Este arquivo é composto de 139.333 registros e contém dados específicos de cada cliente, conforme mostra a Tabela 4.

Tabela 4 – Descrição dos campos do arquivo de clientes

CAMPO	TIPO	DESCRIÇÃO DO CAMPO
ABI_NUM	Char	Identificação do cliente (global)
CUST_NUM	Char	Identificação do cliente para pedidos (Comum a esta base de dados e a base de dados de pedidos)
ADSIZ	Char	Tamanho de anúncios do comprador num veículo de comunicação específico (Categorias: A, B, C, D)
CREDIT	Char	Qualidade do crédito, baseado em métodos indiretos (não se baseia no histórico de pagamento)
FAXPRES	Num	1 indica que a empresa possui fax
GROWCO	Char	Velocidade de crescimento do cliente: 1 indica o crescimento de 10% no número de empregados para companias com mais que 3 empregados
INDFRM	Char	1 significa que o anúncio no veículo de comunicação sai em nome da pessoa física, e 2 sai na pessoa jurídica
INDGROUP	Char	Tipo de indúsria a que o cliente pertence
LOCEMP	Num	Número de empregados locais
OFFSIZ	Char	Número de profissionais no negócio: A = 1 D = 4 B = 2 E = 5 – 9 C = 3 F = mais que 10
PACTEM	Num	Número de todos empregados que trabalham nas filiais da empresa
PDT	Num	Previsão de desktops – local
PDTP	Num	Previsão de desktops – matriz
PNB	Num	Previsão de notebooks – local
PNBP	Num	Previsão de notebooks – matriz
PORDER	Num	Previsão de gastos em PCs – local
PORDERP	Num	Previsão de gastos em PCs – matriz
PSRV	Num	Previsão de servidores – local
PSRVP	Num	Previsão de servidores – matriz
PWS	Num	Previsão de Workstations – local
PWSP	Num	Previsão de Workstations – matriz

CAMPO	TIPO	DESCRIÇÃO DO CAMPO
SIC	Char	Standard Industrial Classification - Padrao de classificação industrial dos Estados Unidos. O conteúdo em parenteses mostra como seria classificado o produto móveis segundo este padrão: 2 primeiros dígitos – Setor Industrial (Móveis) 3° dígito – Grupo Industrial (Móveis caseiros, escritório) 4° dígito – Produto (madeira, metal)
SSIC1	Char	Padrao de classificação industrial – Atividade Secundária
SLSVDT	Num	Modelagem desenvolvida internamente por X que busca prever o volume de vendas da filial do comprador
PACTSL	Num	Modelagem desenvolvida internamente por X que busca prever o volume total de vendas da matriz do comprador
STATE	Char	Estado onde o cliente se encontra
SUBNUM	Char	Número da subsidiária
WHITE	Char	Proporção de pessoas de decisão (colarinhos brancos): 1 indica que mais de 50% dos empregados são executivos

Uma amostra deste arquivo encontra-se no APÊNDICE B - (página 99).

O campo CUST_NUM é o mesmo nas duas bases de dados. É através deste campo que podem ser realizadas operações de merge.

CAPÍTULO 5 - DESENVOLVIMENTO DE MODELOS PARA A PREVISÃO DA COMPRA DE SERVIDORES

O objetivo é desenvolver modelos que façam a previsão de quais clientes comprarão servidores no próximo quadrimestre. Duas abordagens serão realizadas para este problema. A primeira delas consiste em desenvolver modelos levando-se em conta os dados cadastrais e também as compras anteriores dos clientes. A segunda abordagem, levará em conta somente os dados cadastrais dos clientes.

Premissas gerais dos modelos de previsão de vendas

Algumas premissas foram assumidas durante o desenvolvimento destes modelos. Estas premissas se aplicam a todos os modelos de previsão desenvolvidos a seguir. Premissas adicionais específicas de determinado modelo serão discutidas durante sua construção.

- Não foram levadas em consideração as mudanças ocorridas na designação genérica dos produtos durante os dois anos dos dados disponíveis. Por exemplo, dentro da designação *servidores*, não se fez nenhuma distinção entre eles, mesmo quando modelos possam ter entrado ou saído de linha, muito provavelmente com especificações bem diferentes. Portanto, manteve-se a designação genérica;
- Não foram levadas em consideração as mudanças ocorridas no mix de produtos. Por exemplo, é de se esperar que os servidores de dois anos atrás atendiam menos estações que os servidores mais recentes, dado o aumento da capacidade verificadas nestes últimos. Portanto, não se levou em consideração o provável aumento da relação estação/servidor;
- Assumiu-se que os dados cadastrais se mantiveram estáticos durante este período, o que pode ser particularmente falso para aquelas empresas que apresentaram grande crescimento durante aquele período;

Previsão De Vendas A Partir Dos Dados Cadastrais E Do Histórico De Compras

Esta abordagem busca desenvolver um modelo que determine quais clientes deverão ou não comprar um servidor no próximo quadrimestre, com base tanto nos dados cadastrais como também no histórico de compras dos clientes.

Espera-se que este modelo seja capaz de se beneficiar de duas características do histórico de compras: 1) da causalidade nas compras (clientes compram produto A para então comprarem o produto B) e 2) das compras casadas (clientes que compram A também compram B).

Representação dos dados

Os dados originais da base de dados de pedidos não configuram uma boa representação para o treinamento dos modelos, porque, embora as datas de compra estejam presentes em cada entrada, a informação do histórico de compras encontra-se dispersa em várias entradas. Isto ocorre porque cada entrada corresponde a uma transação de compra de um produto do cliente. Portanto, o primeiro passo é desenvolver uma maneira de se representar cada histórico de compras de um cliente em uma entrada. Isto feito, deve-se associar os dados cadastrais do respectivo cliente.

Como já foi dito, a representação dos dados deverá ser capaz de não perder as informações da ordem das compras e das compras casadas.

A primeira representação tentada foi montar entradas que representassem, para cada cliente, a quantidade comprada de cada equipamento por quadrimestre. Tipicamente as entradas tinham o formato mostrado na Tabela 5.

Tabela 5 – Estrutura inicial dos dados para representação dos dados

	# Prod.	# Trimestres	Campos
Número do cliente	-	-	1
Quantidade do Produto 1 no Quadrimestre 1	1	1	1
Quantidade do Produto 2 no Quadrimestre 1	8	1	8
...			
Quantidade do Produto 9 no Quadrimestre 1			
Quantidade do Produto 1 no Quadrimestre 2	9	1	9
...			
Quantidade do Produto 9 no Quadrimestre 2			
Quantidade do Produto 1 no Quadrimestre 3	9	6	54
...			
Objetivo: Quantidade do Produto 9 no Quadrimestre 8			
Firmograficos	-	-	30
Total de campos por entrada			103

Notou-se, entretanto, que a maioria dos 72 campos relativos à compras de equipamentos por trimestre eram nulos, pois tipicamente a cada trimestre o cliente compra uma ou duas categorias de equipamentos, e faz uma a duas compras por ano. Portanto, somente dois a quatro destes 72 campos eram não nulos a cada entrada.

De acordo com a experiência de profissionais de mercado, suspeita-se que representações com muitos campos nulos ou com valor zero dificultam o desenvolvimento de bons modelos. Mesmo assim, tentou-se criar modelos com Redes Neurais, Regressão Linear e Árvores de Decisão para esta representação. Como era de se esperar, não foram obtidos bons resultados com estes modelos, uma vez que as predições resultantes eram quase aleatórias. No caso específico de modelagem com redes neurais, não foi possível treiná-las.

A segunda tentativa optou pela redução do número de campos não nulos. A solução adotada foi representar, por cliente, as datas da última (variáveis LAST) e penúltima (variáveis PENU) compras de cada equipamento, assim como a quantidade comprada de cada equipamento (variáveis OWNED) pelo cliente nos últimos 7 trimestres. Portanto as variáveis LAST e PENU podem ter valores de 1 a 7, correspondentes ao trimestre em que

ocorreram as compras. Caso a empresa não tenha comprado nenhum equipamento neste período, foi escolhido o valor 10 para estas variáveis[7]. A variável resposta é CURYN_9, que vale 1 se pelo menos um servidor foi comprado no oitavo trimestre, e 0 caso contrário. A cada cliente também foram adicionados os dados cadastrais. Deve-se notar que esta representação pode perder alguma informação sobre o histórico de vendas, uma vez que não permite a representação de um cliente que compre o mesmo produto em mais de duas ocasiões em um intervalo de sete trimestres. Porém, constatou-se que eram poucas as ocorrências deste tipo. Com o intuito de avançar no estudo, optou-se por se considerar somente as datas das duas últimas compras de cada produto, e caso não se obtivesse bons resultados, aumentar de duas para três datas de últimas compras.

Outra importante suposição refere-se aos diferentes timings do mesmo padrão de compra ao longo do tempo. Por exemplo, partamos da hipótese que exista um padrão para a seguinte seqüência de compras: laptop => notebook & laptop => servidor, e que o codigo do laptop seja 1, notebook 2 e servidor 9. Considere agora que os clientes A e B compraram os equipamentos nesta sequencia, onde A realiza suas compras nos trimestres de 4 a 6, e B realiza suas compras nos trimestres de 6 a 8. Portanto, embora o padrão de compras de A e B seja exatamente o mesmo (apenas começando em momentos diferentes), a representação das duas compras ficaria conforme a Tabela 6.

Tabela 6 – Representações de um mesmo padrão de compras deslocado no tempo

Cliente A (Trimestres 4 a 6)	PENU_1 = 4	LAST_1 = 5	LAST_2 = 5	CURYN_9 = 0
Cliente B (Trimestres 6 a 8)	PENU_1 = 6	LAST_1 = 7	LAST_2 = 7	CURYN_9 = 1

Portanto, o mesmo padrão de compra teria duas representações diferentes.

Como o número de amostras é muito grande, assumiu-se que todas as representações relevantes, mesmo que deslocadas no tempo, estariam representadas.

A ferramenta estatística utilizada para gerar a representação do histórico de compras foi o pacote computacional SAS. A partir do arquivo de pedidos dos clientes, para cada cliente foi gerada uma entrada com as variáveis LAST, PENU e OWNED. Depois, através da

[7] Numa primeira tentativa foi escolhido o valor 99, que não permitiu o correto desenvolvimento dos modelos por polarizar os dados para grandes números e diminuir a significância dos números de 1 a 7.

instrução *merge*, este resultado foi mesclado com o arquivo de dados cadastrais para se adicionar os dados cadastrais a cada entrada.

O código utilizado para a geração desta representação encontra-se no APÊNDICE C - (página 100), e uma amostra dos dados gerados encontra-se no APÊNDICE D - (página 104).

Enterprise Miner

A ferramenta utlizada para realizar o processamento dos dados e desenvolvimento dos modelos de Regressão Linear, Redes Neurais e Árvores foi o Enterprise Miner da SAS, mais conhecido como E-Miner. Além de trabalhar de modo integrado ao SAS, este sistema reúne facilidades como programação visual dos processos desejados, e uma vez que se especifica apenas o nome do arquivo de entrada, o sistema gera automaticamente os nomes de todos os outros arquivos intermediários. Assim, torna-se muito simples para o pesquisador tentar diferentes modelos, tornando rápido e eficiente o refinamento dos mesmos.

Existem também outros softwares no mercado similares, como o Clementine que é do mesmo fabricante do SSPS, entre outros.

Mapa geral da Metodologia Aplicada para o desenvolvimento do Modelo de Previsão

Na Figura 6 encontram-se representados os procedimentos para o desenvolvimento do modelo de previsão.

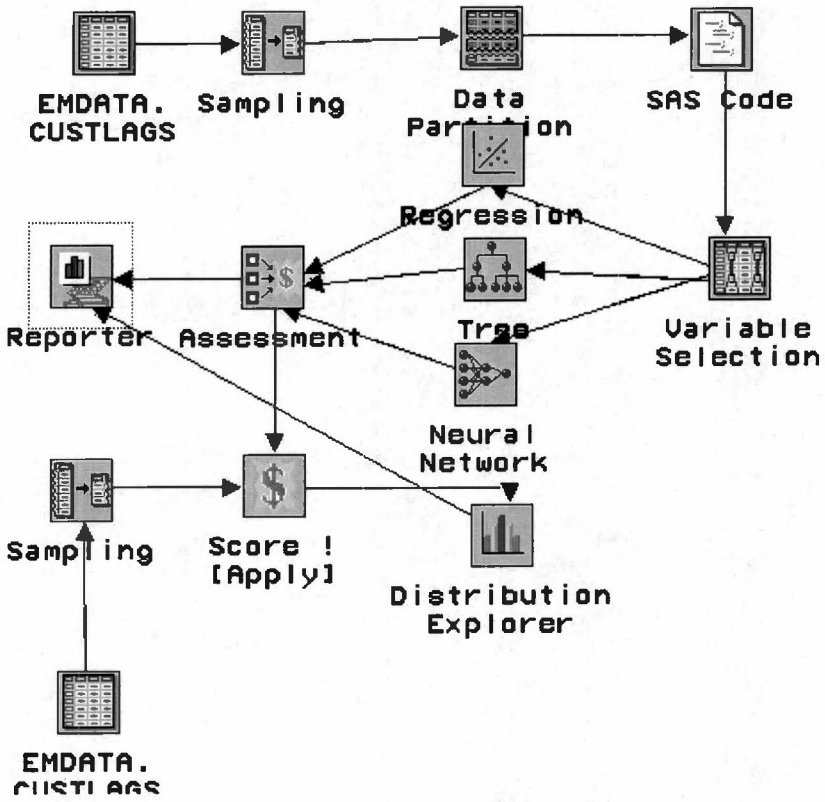


Figura 6 - Mapa da Metodologia para Modelos de Previsão

A estrutura acima representa graficamente o processamento dos dados na forma de um fluxograma. Cada nó executa uma função específica, e o processamento ocorre na direção indicada pelas setas entre os nós. Antes de entrar em detalhes do que é feito em cada nó, será dada uma explicação geral do que fazem e como se inter-relacionam entre si.

Em linhas gerais, o nó *EMDATA.CUSTLAGS* encarrega-se dos procedimentos de acesso e leitura da base de dados CUSTLAGS (APÊNDICE D -, página 104). No nó *SAMPLING* pode-se (ou não) fazer uma amostragem dos dados, com o intuito de acelerar o proceso de datamining nos modelos iniciais. Esta amostra é dividida em três segmentos de dados diferentes no nó *Data Partition*, um para o treinamento dos modelos, o segundo para teste dos modelos e o último para a validação dos dados. Deve-se notar que a partir deste ponto estas três bases de dados sofrerão o mesmo processamento nos nós subsequentes automaticamente. O nó *SAS Code* depura os dados, retirando os dados onde o campo CURYN_9 apresenta-se em branco. O nó *Variable Selection* permite uma depuração mais refinada, através da exclusão manual de variáveis assim como a exclusão automática com base em critérios estatísticos. É também neste nó que a variável CURYN_9 é definida

como objetivo (target) para os nós subsequentes de análise, Árvore (*Tree*), Regressão (*Regression*) e Redes Neurais (*Neural Networks*). São nestes nós que os respectivos modelos são treinados e validados. Para cada nó, os dados da partição de treinamento são utilizados para o treinamento (estimação de parâmetros) do modelo, que depois são testados e validados pelas partições correspondentes (estas três partições são provenientes do nó *Data Partition*). Os resultados e desempenho de cada modelo são depois comparados no nó *Assessment*. Após este nó existem duas setas, uma que vai para *Reporter* e outra que vai para *Apply!*. O nó *Apply!* permite aplicar o modelo treinado em uma nova base de dados, que neste caso é uma outra amostra da mesma base de dados da entrada (CUSTLAGS). A razão para isso será discutida mais adiante. O nó *Distribution Explorer* permite ver o resultado da aplicação do modelo na nova base de dados. Finalmente, o nó *Reporter* registra todas as etapas dos processos acima, gerando um relatório em formato HTML[8] que pode ser acessado em qualquer navegador Internet.

A seguir serão apresentados maiores detalhes de cada nó.

Nó EMDATA.CUSTLAGS

Este nó é responsável pela importação dos dados. Cuida dos procedimentos operacionais de acesso a arquivos na rede de computadores e da importação de bases de dados através de ODBC. Oferece informações como o número de entradas e valores máximos e mínimos de cada variável. Permite também a especificação de cada campo, ou seja, permite também determinar quais são os tipos (intervalo, ordinal, texto, etc) das variáveis. Uma amostra da saída deste nó encontra-se no APÊNDICE E -(página 105).

Nó SAMPLING

Foi feita uma amostragem de 10% da base de dados CUSTLAGS. Primeiramente tentou-se uma amostragem aleatória simples, mas não se obteve um bom modelo já que a variável CURYN_9 tinha muito mais valores 0 do que 1. Este fato decorre que a maioria dos clientes não havia comprado um servidor no trimestre corrente. Para resolver este problema, decidiu-se por um plano amostral estratificado usando a variável CURYN_9

[8] Hypertext Markup Language – Padrão utilizado nos navegadores de Internet

como critério de estratificação. Desta maneira, foram selecionados aleatoriamente duas amostras de igual tamanho para clientes compradores (CURYN_9 = 1) e não compradores (CURYN_9 = 0) de servidores no trimestre corrente. *Verificou-se assim que é muito importante o uso de planos amostrais convenientes nesta etapa do processo.*

Para verificar possíveis consequências deste procedimento, incluiu-se no final do processo uma nova amostra aleatória de CUSTLAGS, para verificar como o modelo se comporta nas condições originais da base de dados.

Nó Data Partition

Este nó dividiu a base de dados CUSTLAGS original em três segmentos: 40% das entradas em Treinamento (3326 entradas), 30% em Validação (2495 entradas) e 30 % em Teste (2495 entradas).

O segmento de Treinamento é utilizado para a modelagem preliminar dos dados segundo cada uma das técnicas.

O segmento de Validação serve para conferir a validade do modelo no nó de *Assessment*. Nos nós de Regressão Linear, Árvore e Redes Neurais, esta base de dados é também utilizada para ajuste fino dos modelos.

O segmento de Teste é utilizado para a obtenção de uma estimativa final não viesada do erro de generalização dos modelos.

Nó SAS Code

Este nó permite processamentos específicos na base de dados, através de comandos SAS. Neste caso, este nó apenas apaga da base de dados as entradas cuja variável objetivo (CURYN_9) esteja em branco.

O código deste nó encontra-se no APÊNDICE F -(página 107).

Seguindo a premissa que o E-Miner cuida de todos os arquivos intermediários, deve-se notar que as variáveis `&_TRA`, `&_TRAIN`, `&_TES`, `&_TEST`, `&_VAL` e `&_VALID` são

na realidade macros[9] que o E-Miner posteriormente desdobra em nomes de arquivos intermediários. Esta facilidade permite que o mesmo código seja reaproveitado inúmeras vezes, sem que haja necessidade reescrevê-lo a cada uso.

Nó Variable Selection

Este nó tem duas funções básicas: 1) selecionar que variáveis entrarão ou não no treinamento e qual papel possuirão, e 2) converte variáveis para formatos mais eficientes.

As variáveis CURAMT_X não foram utilizadas porque esta variável refere-se a quanto foi comprado do produto X no trimestre corrente, um dado que não estaria disponível no mundo real. As variáveis CURYN_X (com exceção da variável CURYN_9) também não foram utilizadas pelo mesmo motivo. A variável objetivo CURYN_9 foi designada como tal ao declará-la como *target* no sistema. Todas as outras variáveis foram declaradas como *input*.

O E-Miner então passa à fase de seleção automática ou manual das variáveis, e posteriormente muda a representação delas para ganhos de eficiência.

A primeira seleção automática realizada foi a eliminação das variáveis que tivessem mais que 50% de valores em branco.

A segunda seleção foi eliminar variáveis que tivessem R^2 menor que 0.005. R^2 é o coeficiente de determinação, que representa a proporção da soma dos quadrados da variável objetivo que são atribuídos à variações da variável de entrada. Em outras palavras, um R^2 de 0.4 significa que 40% da variação da variável objetivo é explicada pela variação da variável preditora ou independente (Notar que R^2 não mede a interação entre as variáveis independentes). Portanto, ao eliminar variáveis que possuam R^2 menor que 0.005, eliminam-se as variáveis cujas variações com pouca relação com a variável resposta.

Após esta etapa, o E-Miner ainda aprimora a seleção de variáveis através do método Forward Stepwise. Este método seleciona entre as variáveis rstantes, aquela que melhor explica as variações da variável preditora, ou seja, aquela com maior R^2. A cada iteração

---

[9] Macros: Identificadores dentro do código que têm seus nomes substituidos pelas variáveis definitivas

do processo, é adicionada uma nova variável, que será aquela que der a maior contribuição marginal para R^2. O processo só termina quando a contribuição marginal for menor que um valor pré-determinado, que em nosso caso foi R^2 maior que 0.0005.

Após esta seleção, o E-Miner procede à redefinição das variáveis, que tem por objetivo encontrar uma representação mais simples sem que se perca informações. Por exemplo, suponha uma variável de classe que possua quatro valores distintos, e que explique 20% do resultado da variável objetivo. O E-Miner ordena então os valores da variável independente pela média das predições.

Tabela 7 – Ordenação dos valores da variável independente pela média das predições

Média das predições	Valores da Variável classe
90	C
85	B
50	D
42	A

O algoritmo combinará os dois primeiros níveis (C e B) em apenas um caso a redução em R^2 for menor ou igual a 5% (ou seja, se R^2 cair para até 19%). Caso esta condição se verifique, O E-Miner agrupará os níveis C e B em apenas um, e repetirá o processo tentando agrupar o nível D ao grupo formado. Caso a condição não se verifique, o E-Miner tentará agrupar os níveis B e D, e assim sucessivamente.

Nota-se assim que, embora automático, o método de seleção e redefinição de variáveis é bastante poderoso no sentido de escolher as variáveis que de fato mais influem no modelo e na simplificação das mesmas.

No APÊNDICE G -(página 108) pode-se ver como o E-Miner redefiniu as variáveis, e no APÊNDICE H - (página 112) fica claro que a redefinição manteve o mesmo valor de R^2 para as novas variáveis, porém reduziu significativamente o número de graus de liberdade (DF – Degree of Freedom), indicando que houve uma simplificação significativa na representação da informação.

---

durante a compilação logo antes da execução do programa.

Nós de modelagem

Os nós Redes Neurais (Neural Network), Regressão (Regression) e de Árvore são nós preditivos, no sentido que precisam ser treinados com um conjunto de dados, e checados quanto à performance com outro conjunto de dados cujas respostas esperadas são conhecidas. A utilização do modelo só ocorre quando a checagem garante que as respostas do sistema seja em alguma medida similares às respostas esperadas.

O treinamento consiste em apresentar as variáveis de entrada, e comparar a saída do modelo com a saída esperada do conjunto de treinamentos. Ocorre então uma reparametrização do sistema de modo a aproximar a resposta ajustada à resposta esperada, e então novas entradas são apresentadas e checadas com uma nova resposta esperada. O processo prossegue desta maneira até que haja convergência das respostas dentro de determinados parâmetros estatísticos ou um certo número de iterações seja atingido.

Após esta etapa de treinamento, procede-se à checagem, que consiste em mostrar uma nova base de dados, e comparar as saídas com as respostas esperadas, sem no entanto haver reparametrizações. Caso não ocorram boas predições, o modelo pode ser retreinado com uma outra configuração que normalmente implica numa maior ou menor complexidade.

A complexidade de qualquer modelo preditivo está relacionado tanto com o número de parâmetros internos do modelo como também em alguns casos com a arquitetura interna dos mesmos. O aumento de complexidade dos três modelos usados neste trabalho se dão da seguinte maneira:

- na regressão linear, acontece pela adição e interações e de termos poinomiais;
- em redes neurais, pode-se adicionar mais células intermediárias e também novas camadas;
- em árvores de decisão, pode-se aumentar o número final de ramificações.

A seguir serão apresentados as três diferentes modelagens. Será realizada uma breve discusão para cada um dos nós e uma investigação das informações que ajudem a entender os dados e as relações entre as variáveis.

Não é o objetivo deste trabalho uma extensa discussão dos mecanismos internos e algoritmos utilizados na obtenção dos resultados. As ferramentas atuais de Datamining

(entre elas o E-Miner) já cuidam da maioria dos procedimentos e detalhes de cada modelagem, tentando na maioria das vezes diferentes configurações de um mesmo algoritmo e escolhendo aquela que produz os melhores resultados. Um fato empírico que reforça esta posição foi que durante a realização do processamento de dados deste trabalho por muitas vezes tentou-se variar os parâmetros sugeridos pelo E-Miner, e os resultados pouco diferiram um dos outros, fornecendo basicamente os mesmos resultados.

Entretanto, em cada modelo haverá a preocupação de se extrair a maior quantidade possível de informações sobre os dados a partir da análise dos parâmetros e indicadores relevantes de cada modelo.

Para fins de comparação entre os modelos, foi considerado importante julgar qualitativamente a consistência dos resultados obtidos, dentro de uma metodologia que permita compará-los nas mesmas bases e pelos mesmos critérios. A escolha mais natural para uma medida extensiva a todos os modelos é a chamada matriz de confusão, que indica as taxas de classificação adequadas e inadequadas. Esta escolha justifica-se porque 1) os resultados de todos os modelos utilizados podem ser configurados desta maneira, e 2) o tipo de informação que consta nas matrizes de confusão são em última análise o que o usuário da informação de datamining necessita para utilizar no mundo real.

Matriz de Confusão

A matriz de confusão é formada pelo cruzamento dos dados esperados e previstos, permitindo um rápido entendimento do poder de previsão de determinado modelo.

Um exemplo de matriz de confusão é mostrado na Tabela 8.

Tabela 8 – Exemplo Matriz da Confusão

TARGET	OUTPUT		
Frequency Percent Row Pct Col Pct	0	1	Total
0	636 38.20 75.99 76.08	201 12.07 24.01 24.25	837 50.27
1	200 12.01 24.15 23.92	628 37.72 75.85 75.75	828 49.73
Total	836 50.21	829 49.79	1665 100.00

No exemplo acima, as linhas referem-se aos valores reais da variável resposta, e as colunas referem-se às predições da mesma variável. Observa-se que de 1665 pedidos, foram previstas 836 não compras de servidores, onde de fato 636 não compraram e 200 compraram. Dos 829 pedidos restantes, verifica-se que 628 dos que foram previstos corretamente como compradores, e 201 não compraram. Os números abaixo dos mencionados mostram as porcentagens correspondentes.

A matriz de confusão deste exemplo pode ser representada graficamente conforme Figura 7.

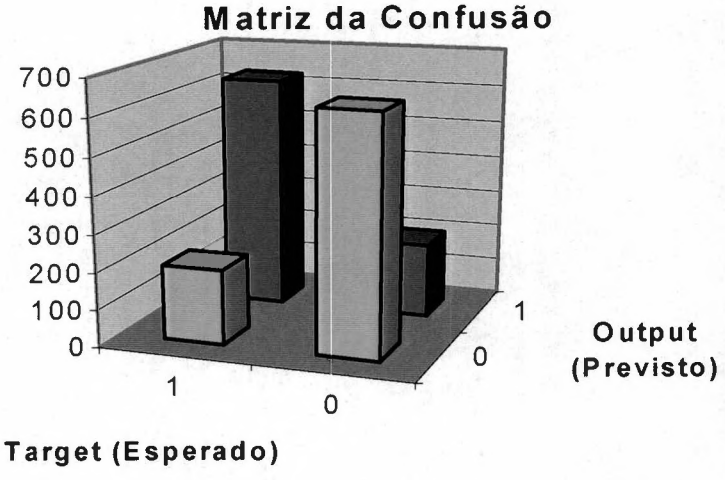


Figura 7 – Representação Gráfica da Matriz da Confusão

Por meio da análise da Figura 7 podemos qualitativamente observar que um modelo que gerasse estes resultados seria capaz de realizar boas previsões, já que acerta 76.08% para o evento 0, e acerta 75.75% das vezes para o evento 1 (olhar também na Tabela 8). Desta forma, bons modelos tendem a ter alta concentração de ocorrências na diagonal principal (diagonal cujos valores Esperados e Previstos são iguais), e baixa concentração de ocorrências nas outras células (cujos valores Esperados e Previstos são diferentes).

Mais adiante serão mostradas medidas estatísticas que permitem medir quantitativamente a eficiência de cada modelo permtindo compará-los.

Nó Redes Neurais

O nó de Redes Neurais (Neural Network) implementa o algoritmo de back-propagation de redes neurais. Este algoritmo encontra-se descrito em detalhes em Beale (1990).

Redes neurais podem ser utilizadas em problemas de classificação (para variáveis-objetivo de categoria), ou em problemas de regressão (para variáveis-objetivo contínuas).

Redes neurais possuem uma arquitetura de camadas formadas por células, onde a primeira camada é constituída das variáveis preditoras. A saída das células da camada de entrada são conectadas a uma camada interna, que por sua vez podem ser conectadas a outra camada interna ou a uma camada de saída. A camada de saída corresponde a uma ou mais variáveis resposta.

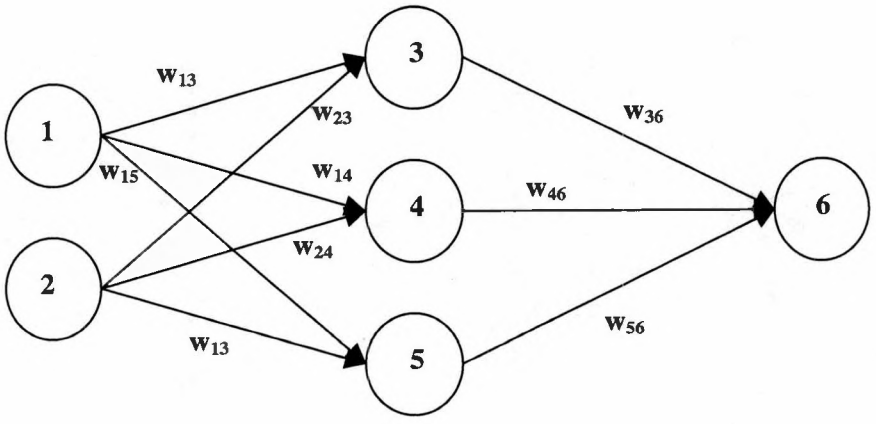


Figura 8 – Rede Neural com uma camada intemediária

Cada célula tem como entrada todas as saídas das células da camada anterior, porém uma só saída, que por sua vez é conectada a todas as células da camada posterior. O valor da saída de cada célula é obtido primeiramente fazendo-se o somatório das entradas ponderadas pelos respectivos pesos w_{xy}. Uma vez obtido o somatório, é aplicada uma função de ativação que normalmente que tem por objetivo limitar a grandeza dos resultados.

Embora há muitas implementações diferentes de redes neurais, a mais comum e que foi utilizada neste trabalho é a do tipo feed-forward backpropagation, com função de ativação sigmóide, que busca reduzir o erro do resultado previsto a cada evento. Inicialmente produzem-se aleatoriamente pesos w_{xy} iniciais. A partir dos dados de entrada na primeira camada, são calculados os valores das células das camadas intermediárias sucessivamente, até que os valores das células da camada de saída sejam calculados. O erro é computado subtraindo estes valores dos valores esperados. Os pesos w_{xy} são ajustados de maneira a minimizar este erro. Produz-se um novo conjunto de dados de entrada e o processo recomeça. Quando o erro fica menor que um determinado valor, diz-se que a rede está treinada, e pronta para ser utilizada.

Há vários problemas práticos de implementação, tais como:

- overfitting de dados: é a perda de generalização do modelo obtido na amostra de treinamento, ou seja, o modelo é capaz de classificar corretamente apenas os dados da amostra de treinamento, mas não tem sucesso ao trabalhar com dados similares de outra amostra. O E-Miner resolve este problema ao avaliar periodicamente a redução do erro na amostra Teste (e não na amostra de treinamento);
- soluções de mínimos locais: muitas vezes a rede treinada não tem um bom desempenho, porque encontrou uma solução de mínimo local, sendo necessário treiná-la de novo com diferentes pesos iniciais. Existem alguns algoritmos que evitam soluções de mínimos locais;
- o tempo de treinamento pode ser bastante demorado quando o número de células é grande, sendo certamente maior quando comparado a outros modelos;

- não é claro qual a melhor arquitetura de rede para um problema, ou seja, quantas camadas deve possuir e quantas células deve haver em cada uma das camadas, podendo exigir o teste de algumas configurações preliminares.

Este modelo é o de mais difícil interpretação, e seus parâmetros pouco ajudam a entender a inter-relação das variáveis. Tem como vantagem tratar de problemas com solução não linear.

Uma abordagem mais detalhada sobre este método pode ser obtida em Beale (1990).

Abaixo são mostradas a matriz de confusão deste modelo e sua representação gráfica.

Tabela 9 – Matriz de Confusão a partir do Modelo de Redes Neurais

Real * Previsto Crosstabulation

			Previsto		
			.00	1.00	Total
Real	.00	Count	648	189	837
		Expected Count	477.1	359.9	837.0
	1.00	Count	301	527	828
		Expected Count	471.9	356.1	828.0
Total		Count	949	716	1665
		Expected Count	949.0	716.0	1665.0

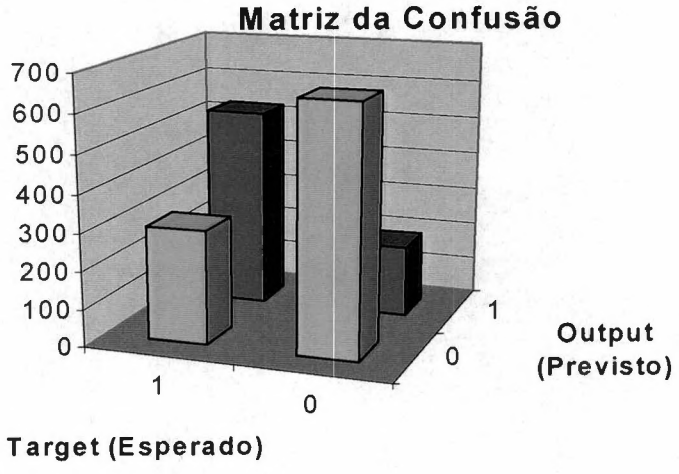


Figura 9 – Representação Gráfica da Matriz da Confusão para Redes Neurais

Neste ponto algumas questões são levantadas:

- Este é um bom ou mau modelo?
- Estes resultados indicam que grau de dependência entre o que era esperado e os resultados que de fato foram conseguidos?

Teste Qui-Quadrado

O teste do Qui-Quadrado endereça este tipo de questões. Para que fosse realizado, utilizou-se o programa SPSS, que tem um conjunto de ferramentas bastante completo para este fim de análise. Como o SPSS não lê a Matriz da Confusão diretamente, foi preciso gerar numa planilha eletrônica uma tabela com duas colunas, Real e Previsto, que simulasse os resultados do E-Miner.

Na Tabela 10 seguem os resultados do teste Qui-Quadrado gerados no SPSS.

Tabela 10 - Teste Qui-Quadrado para Modelo de Redes Neurais

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	286.398[b]	1	.000		
Continuity Correction[a]	284.725	1	.000		
Likelihood Ratio	295.908	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	286.226	1	.000		
N of Valid Cases	1665				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 356.06.

A fim de se determinar se a predição tem relação com os valores verdadeiros da variável dependente, deve-se testar a dependência estatística entre elas. A hipótese nula de independência entre as variáveis é rejeitada, dado que o valor do Qui-Quadrado é 286.4, portanto muito maior que o número de graus de liberdade que é igual a 1. O nível de significância (probabilidade de erro na rejeição da hipótese nula) é menor que 0.0005, indicando portanto que o modelo é estatisticamente consistente e confiável.

Segundo o Manual do SPSS (SPSS Base 9.0 – Applications Guide, Página 67), para que o teste do Qui-Quadrado tenha consistência estatística deve-se ainda observar duas condições:

- Em linhas gerais, deve-se se certificar que nenhuma célula da Matriz da Confusão possua valor esperado[10] menor que 1;
- Não mais que 20% das células pode ter valor esperado menor que 5 (no presente caso em que a matriz é 2 x 2, nenhuma célula pode ter valor esperado menor que 5).

Pela observação *b* na Tabela 10 fica claro que estas duas condições são atendidas.

As outras estatísticas mostradas (Continuity Correction, Likelyhood Ratio e Linear by Linear Association) são variações de Teste do Qui-Quadrado. A informação mais importante destas medidas é o nível de significância (todos eles menores que 0.0005), e tendem a ser redundantes para grandes amostras de dados.

Outras medidas, específicas para dados nominais, também confirmam o teste do Qui-Quadrado, e são mostradas na Tabela 11.

Tabela 11 – Medidas de correlação para o Modelo de Redes Neurais

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Phi	.415			.000
	Cramer's V	.415			.000
	Contingency Coefficient	.383			.000
Interval by Interval	Pearson's R	.415	.022	18.587	.000[c]
Ordinal by Ordinal	Spearman Correlation	.415	.022	18.587	.000[c]
N of Valid Cases		1665			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

[10] O valor esperado de uma célula da matriz de confusão é obtido multiplicando-se o total da linha pelo total da coluna da respectiva célula, dividindo-se então pelo número total de elementos da matriz

As medidas Phi, Cramer's e Pearson's R tornam-se a mesma medida quando a matriz é 2 x 2. Estas medidas variam de 0 a valores próximos de 1, e dão uma medida de quão fortemente as duas variáveis estão correlatadas.

Uma outra medida bastante usada é o *hit ratio,* que é o número de ocorrências classificadas corretamente (ou seja, que estão na diagonal principal) dividido pelo número total de amostras. Portanto, o hit ratio mede a razão dos acertos (0=>0 e 1=>1) em relação ao número total de amostras. Neste caso, o hit ratio vale (648 + 527) / 1665 = 70.57%. Esta porcentagem deve ser comparada com a probabilidade a priori (definida como a razão entre o número de equipamentos vendidos e o número total de pedidos), que é aproximadamente 50%, dado que na amostra presente metade das pessoas compraram um servidor.

Concluindo, as medidas estatísticas mostram que o modelo é consistente, e os resultados produzidos razoavelmente correlatados com os valores esperados. O *hit ratio* de 70.57% também deixa claro o poder de previsão deste modelo baseado nas redes neurais.

Nó de Regressão

Regressão Linear utiliza valores existentes para desenvolver um modelo capaz de fazer predições. Este modelo estima a variável dependente através da combinação linear das variáveis independentes:

$$y = \alpha_0 + \sum \alpha_I x_I + \varepsilon$$

O treinamento do modelo consiste na escolha dos parâmetros α_I que minimizem o erro ε.

Embora normalmente este método seja aplicado para variáveis dependentes métricas, também pode ser adaptado para variáveis dummy, como no estudo deste caso. O E-Miner cuida automaticamente da implementação e ajustes dos dados.

Abaixo são apresentados os T-Scores resultantes. T-Score é igual ao parâmetro estimado dividido pelo seu standard error. Os valores de T-Score podem ser comparados diretamente para mostrar a influência relativa de cada um dos parâmetros na variável dependente, ou seja, é uma comparação livre de fatores de escala. São portanto um indicador da importância de cada variável independente na predição da variável dependente.

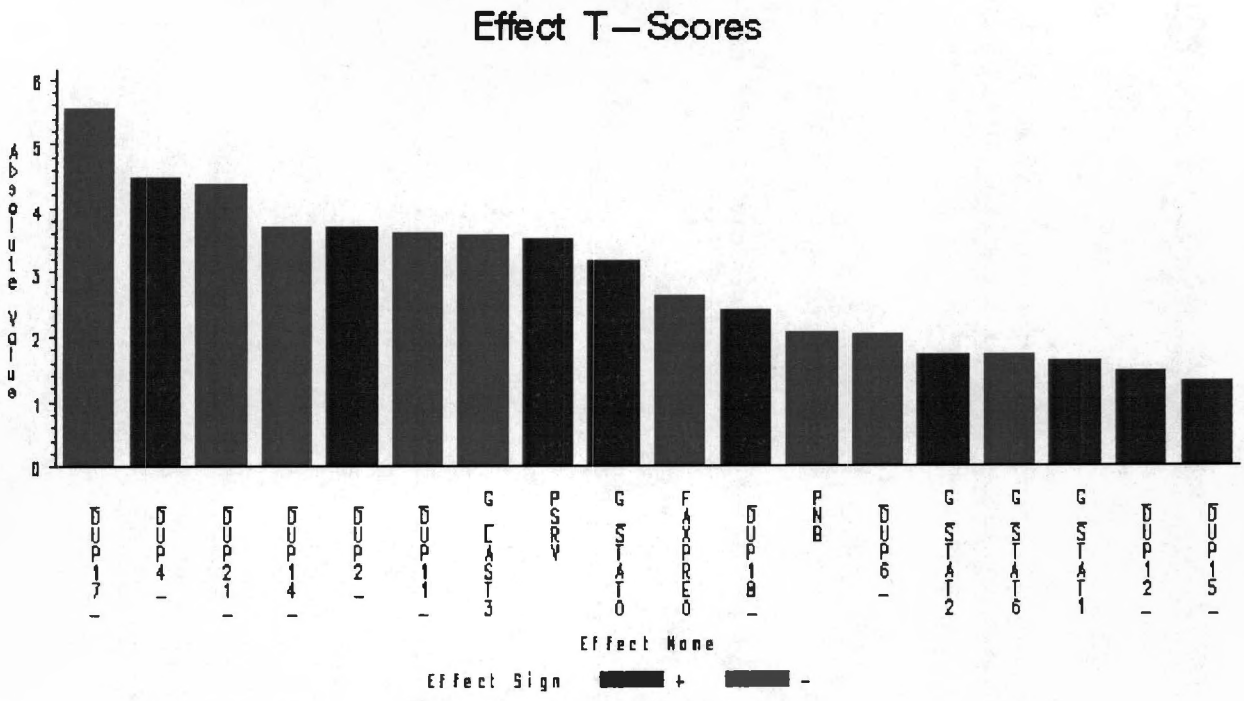


Figura 10 – Principais T-Scores da Regressão

Observando-se a Figura 10 e consultando-se o APÊNDICE I - (página 115), percebe-se que os parâmetros mais importantes para a modelagem dizem respeito ao histórico de compras do cliente (com exceção da variável PSRV).

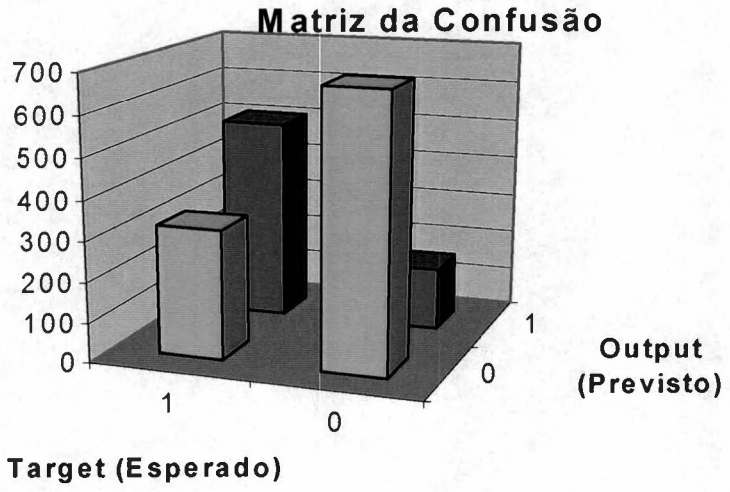
No APÊNDICE J - (página 116) é apresentado um resumo dos resultados estatísticos obtidos na regressão.

Tabela 12 - Matriz de Confusão a partir do Modelo de Regressão

Real * Previsto Crosstabulation - Regressão

			Previsto		
			.00	1.00	Total
Real	.00	Count	681	156	837
		Expected Count	504.7	332.3	837.0
	1.00	Count	323	505	828
		Expected Count	499.3	328.7	828.0
Total		Count	1004	661	1665
		Expected Count	1004.0	661.0	1665.0

Figura 11 - Representação Gráfica da Matriz da Confusão para Regressão



O teste do Qui-Quadrado e as medidas de correlação são satisfatórios para este modelo (ver APÊNDICE K -, página 117). O hit ratio é de 71.23%.

Em linhas gerais os dois modelos (Redes Neurais e Regressão) são equivalentes. Entretanto, como vimos, o modelo de regressão dá maiores informações acerca das variáveis e da relação entre elas.

Nó Árvore

Árvores de decisão representam através das ramificações de uma árvore as regras que levam a uma determinada classe da variável resposta. Cada ramificação possui uma decisão, que tenta dividir os dados de maneira a maximizar a distância (ou seja, a diferença entre os elementos) entre os grupos resultantes. Cada novo galho, por sua vez, dará origem a uma nova ramificação, o processo pára quando o conjunto de ramos não contiver mais elementos heterogêneos.

Os algoritmos das Árvores de Decisão são bastante rápidos, dado que percorrem as bases de dados com muitas variáveis tipicamente não mais que três vezes por ramificação. Comportam eficientemente árvores com grande número de preditores, fazendo com que sejam aplicáveis a grandes bases de dados com grande número de variáveis.

Caso não haja uma restrição para evitar o crescimento dos ramos da árvore, ela pode tornar-se muito grande, dificultando a interpretação do resultado, além de poder produzir efeito de overfitting mencionado em redes neurais. O estabelecimento de regras de parada

são portanto um ponto importante desta metodologia. O E-Miner cuida destes problemas automaticamente por meio de regras padrão, ou então ajustadaspor escolha do usuário.

Maiores detalhes de como são definidas as distâncias e de como os conjuntos podem ser particionados podem ser encontrados no Capítulo 2 do livro Jackson (1983).

A seguir encontra-se a árvore de decisão obtida para o estudo de caso:

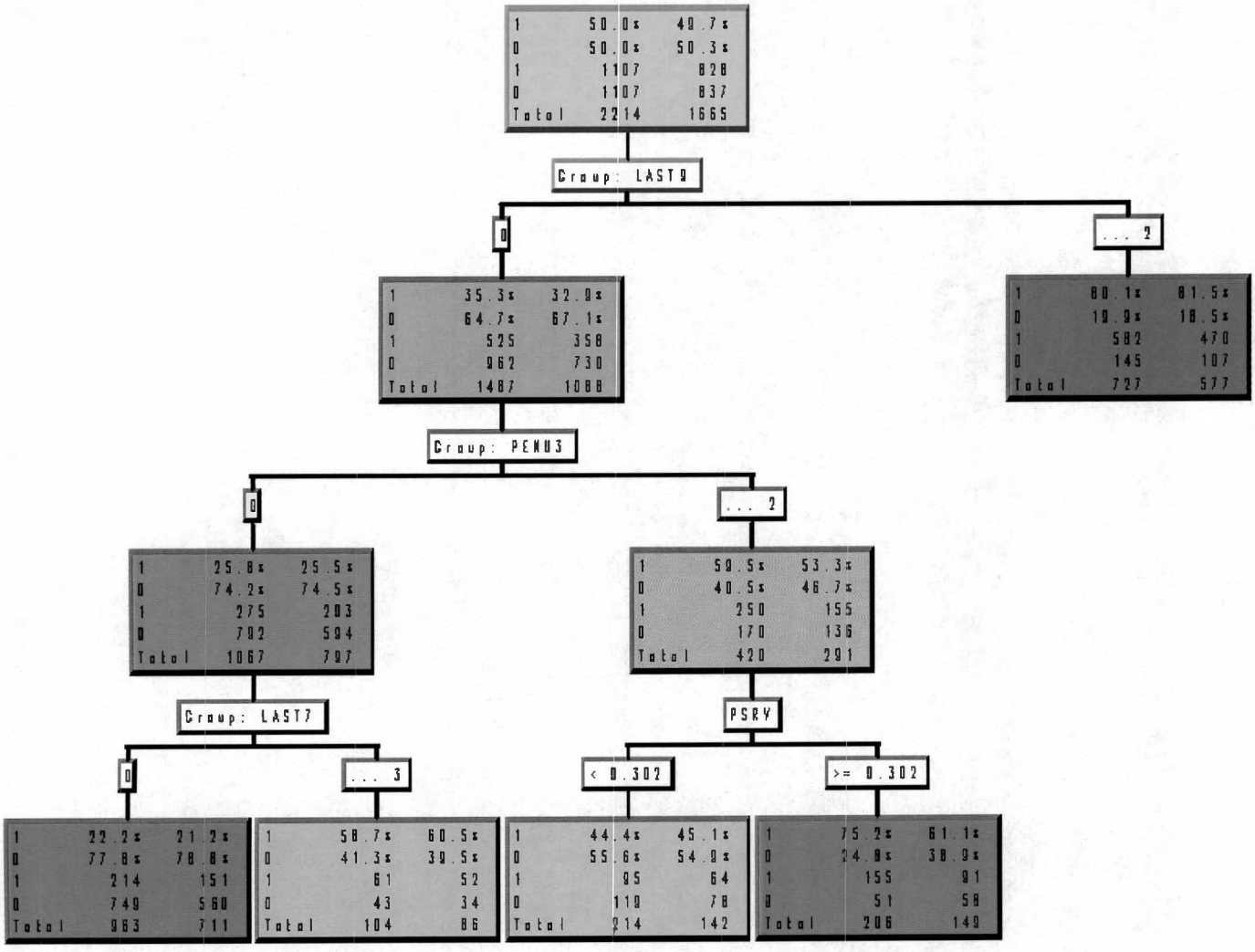


Figura 12 – Árvore de Decisão

As regras textuais para a árvore encontram-se no APÊNDICE K - (página 117).

Observa-se que as principais variáveis que distiguem um grupo do outro referem-se ao histórico de compra do cliente (com exceção de PSRV), e são basicamente as mesmas do modelo de Regressão Linear.

Tabela 13 - Matriz de Confusão a partir do Modelo de Árvore

Real * Previsto Crosstabulation - Árvore

			Previsto		
			.00	1.00	Total
Real	.00	Count	636	201	837
		Expected Count	420.3	416.7	837.0
	1.00	Count	200	628	828
		Expected Count	415.7	412.3	828.0
Total		Count	836	829	1665
		Expected Count	836.0	829.0	1665.0

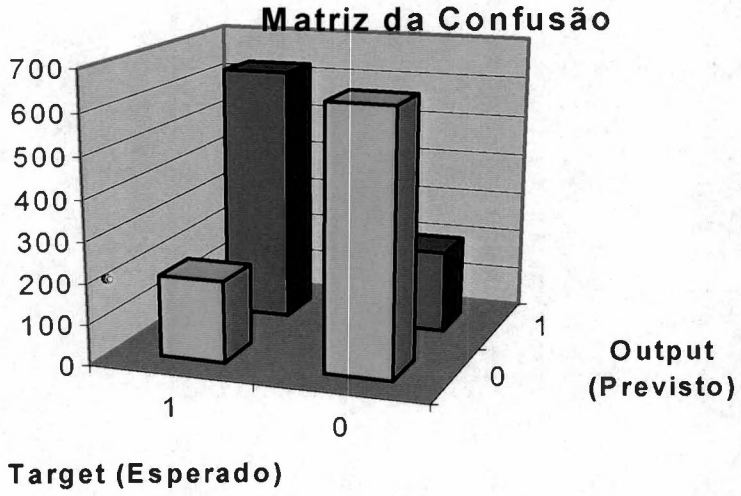


Figura 13 - Representação Gráfica da Matriz da Confusão para Árvore

O teste do Qui-Quadrado e as medidas de correlação são satisfatórios para este modelo (ver APÊNDICE K -, página 117). O hit ratio é 75.91%.

Comparação entre os nós

Um resumo dos resultados gerais obtidos é mostrado na Tabela 14.

	Hit Ratio	Kappa
Redes Neurais	70,57%	0,489
Regressão	71,23%	0,493
Árvore	75,91%	0,518

Tabela 14 – Comparação geral entre os modelos

Kappa é a medida estatística que mede o consenso entre a avaliação de dois juízes independentes. O valor 1 indica consenso total, e 0 significa que não há consenso.

Percebe-se então que embora os resultados sejam quase equivalentes, o modelo Árvore teve um desempenho geral ligeiramente melhor que os outros dois modelos.

Um outro método de comparação é verificar a performance dos modelos a cada percentil, conforme mostra a Figura 14.

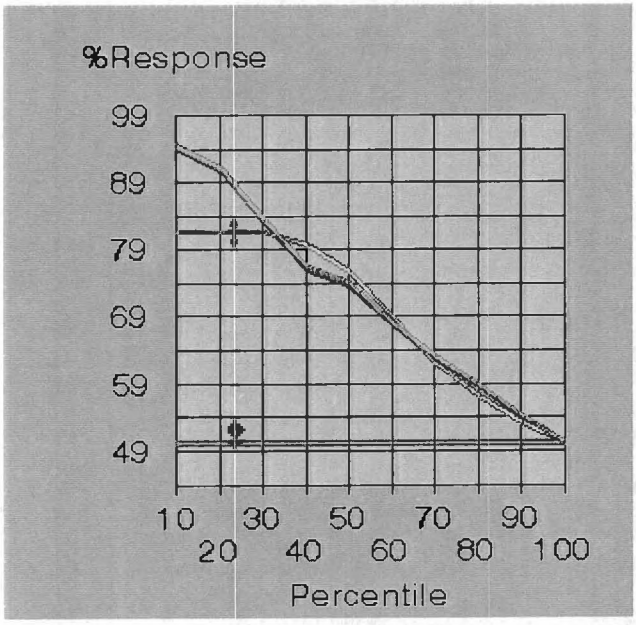


Figura 14 – Comparação da performance entre nós através da Resposta Acumulativa

A Figura 14 compara a performance dos diferentes modelos, através da resposta acumulativa. O eixo horizontal corresponde ao percentil das amostras, baseado na probabilidade a posteriori. Os resultados obtidos nos modelos são ordenados da esquerda para a direita, onde os elementos da esquerda são aqueles que têm a maior probabilidade de compra de servidores 4SV. Em outras palavras, suponha que serão enviadas cartas para os possíveis compradores, convidando-os a comprar servidores 4SV. A fim de maximizar o esforço desta campanha, serão mandadas cartas apenas para o primeiro percentil dos compradores que o modelo predisse que têm maior probabilidade de compra. Se olharmos a Figura 14, perceberemos que aproximadamente 94% dos compradores previstos pelos modelos de Redes Neurais e Regressão irão comprar, enquanto somente 82% dos compradores previstos pelo modelo de árvore o irão. Se entretanto forem enviadas cartas

para os 20% de compradores mais cotados, observa-se que 89% dos compradores previstos pelos modelos de Redes Neurais e Regressão irão comprar, enquanto somente 82% dos compradores previstos pelo modelo de Árvore o irão fazer. Portanto até aqui, os modelos de Redes Neurais e Regressão possuem melhor performance que o de Árvore. Deve-se notar entretanto que se enviarmos cartas para 35% a 50% dos compradores com maior probabilidade, o modelo de Árvore apresentará uma performance ligeiramente melhor que os outros dois. A partir de 50%, a performance dos modelos é idêntica. Como era de se esperar, se enviarmos cartas para 100% dos compradores, a probabilidade de compra torna-se igual a probabilidade a priori, sendo esta a razão que as linhas convergem para a linha de base (Baseline), que nada mais é que a probabilidade a priori.

Percebe-se então que a Resposta Acumulativa é uma medida da eficácia de previsão dos modelos. Maus modelos tendem a ter seu gráfico muito próximo da linha de base.

Embora o modelo Árvore tenha tido um desempenho geral melhor que os de Regressão e Redes Neurais, estes últimos se mostraram melhores nos primeiros percentis. Em certas aplicações (por exemplo, a otimização do envio de mala direta por meio da redução de envio de correspondências) sob certas condições (o número de correspondências enviadas deveria ser menor ou igual à zona onde o desempenho de um modelo for melhor que os outros), a utilização dos dois últimos modelos daria melhores resultados

Aplicação do melhor modelo para base aleatória do Servidor 4SV

De posse do modelo treinado, não deve-se esquecer que a a base de dados trabalhada até o momento possui aproximadamente o mesmo número de eventos de compras e não-compras. Conforme discutido em Nó SAMPLING (página 49), isto foi feito para que fosse possível realizar o treinamento dos modelos. Para realizar a avaliação de performance dos modelos já treinados em uma base de dados com as mesmas características da original, foi feita uma amostragem aleatória dos dados originais, onde se preservou as proporções dos casos existentes. Olhando-se a Figura 6 (página 48), a amostragem ocorre no nó SAMPLING localizado na parte inferior da figura, e a aplicação do modelo ocorre no nó Score.

Os resultados obtidos para servidores 4SV encontram-se resumidos na Tabela 15.

Tabela 15 – Matriz da Confusão para amostra aleatória de servidores 4SV

c4SV_R * c4SV_P Crosstabulation

Count

		c4SV_P		
		.00	1.00	Total
c4SV_R	.00	10925	2723	13648
	1.00	78	207	285
Total		11003	2930	13933

Tabela 16 – Teste Qui-Quadrado para Amostra aleatória de servidores 4SV

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	466.520[b]	1	.000		
Continuity Correction[a]	463.353	1	.000		
Likelihood Ratio	357.502	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	466.487	1	.000		
N of Valid Cases	13933				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 59.93.

Conforme discussão na página 59, podemos perceber que o resultado obtido descarta a hipótese de independência das variáveis, e é estatisticamente consistente ao satisfazer os critérios mínimos de contagem esperada.

O Hit-ratio é de 79.9%, consideravelmente alto se compararmos à probabilidade a priori de 2.0% ((78 +207) / 13933). Isto indica a capacidade de previsão do modelo.

Outras duas medida de interesse é a razão de acertos (tomando-se como evento a compra do equipamento, é a razão de quantas compras previstas realmente ocorreram pelo número total de compras previstas) e o lift (a razão entre a razão de acertos e a probabilidade a priori). Estas medidas são interessantes porque têm íntima correlação com o que realmente acontece no mundo real. Por exemplo, um gerente de marketing, após descobrir através de seus modelos quais seriam os clientes mais propensos a comprar servidores, poderia

enviar cartas para os mesmos. A porcentagem dos clientes que ao receberem a carta comprassem de fato o servidor seria a razão de acertos. O lift mede a vantagem da utilização do modelo se comparado à não utilização do mesmo.

Resultados Gerais Para Os Outros Equipamentos

Tabela 17 – Resultados gerais para os outros equipamentos

Produto	Modelo	Total	0=>0	0=>1	1=>0	1=>1	Priori	Razão de Acertos	Lift	Hit Ratio
2NT	Nenhum									
3WS	NN	13933	11874	1986	21	52	0.5%	2.6%	4.9	85.6%
4DM	NN	2000	1468	366	54	112	8.3%	23.4%	2.8	79.0%
4ES	NN	2000	1897	101	0	2	0.1%	1.9%	19.4	95.0%
4IN	REG	13933	10505	2957	158	313	3.4%	9.6%	2.8	77.6%
4LT	NN	2000	1826	149	7	18	1.3%	10.8%	8.6	92.2%
4PD	REG	2000	1284	60	145	511	32.8%	89.5%	2.7	89.8%
4ST	REG	13933	11569	2354	1	9	0.1%	0.4%	5.3	83.1%
4SV	REG	13933	10925	2723	78	207	2.0%	7.1%	3.5	79.9%

A Tabela 17 mostra para os outros equipamentos qual o modelo que deu o melhor resultado, assim como os dados das matrizes de confusão, probabilidade a priori, Razão de acertos, Lift e hit ratio.

Nota-se que de uma maneira geral os modelos apresentam bons resultados, já que produzem valores de lift tipicamente maiores que 2.5. Isto significa que há ganho real na capacidade de previsão dos mesmos.

Chama-se atenção para o fato de que os equipamentos 4ES, 4LT e 4ST não possuem os requerimentos de valores mínimos para as contagens esperadas, o que coloca em dúvida a consistência de que as previsões têm dependência estatística com os resultados. Este fato se deve sobretudo ao baixo número de eventos e compras destes equipamentos observados na amostra.

Avaliando O Valor Do Modelo

Até agora foi feita uma avaliação da capacidade de predição dos modelos, sem levar em consideração o impacto econômico. A avaliação deste impacto é muito importante, porque é ela que mede o ganho/perda que a empresa terá com os resultados obtidos anteriormente.

Matriz Lucro/Perda

O ganho ou perda que a utilização de um determinado modelo gera pode ser avaliada por meio de uma matriz Lucro/Perda. Esta matriz assemelha-se com a matriz de confusão, porém seu valor refere-se ao ganho/perda de um determinado resultado.

Por exemplo, suponha o caso dos servidores 4SV, cuja matriz de confusão encontra-se na Tabela 15 (página 68). Digamos que custe $10.000,00, e que a margem de venda seja de 30%. Logo, a venda de cada equipamento gera uma margem de $3000,00. Digamos também que o custo de marketing tradicional (revistas, jornais, maladireta, etc) seja de $50,00 por servidor. O universo de compradores é de 13933 compradores, dos quais somente 285 de fato comprarão. Portanto, X gastaria $696.650,00 para promover as vendas dos servidores 4SV, e a margem ganha seria de $855.000,00, tendo portanto um resultado líquido de $158.350,00.

Suponha agora que X tenha utilizado o modelo desenvolvido para os servidores 4SV. Em vez de apenas enviar mala direta e anúncios, X irá centrar seus esforços nos clientes que o modelo previu que irão comprar. Suponha também que os gastos de marketing com estes clientes serão o dobro ($100,00), uma vez que X espera investir mais nestes clientes. Temos então 4 alternativas:

- Modelo previu que cliente não iria comprar e de fato não comprou: Neste caso, o custo de divulgação é zero, e X fez uma economia de $50,00 por cliente;
- Modelo previu que cliente iria comprar quando na verdade não comprou: Neste caso, o custo de divulgação é $100,00, a margem ganha é zero;
- Modelo previu que cliente não iria comprar quando de fato comprou: Neste caso, o custo de divulgação é zero, e X perdeu a oportunidade de ganhar $3000,00 por cliente;
- Modelo previu que cliente iria comprar e de fato comprou: Neste caso, o custo de divulgação é $100,00, e X ganhou $3000,00 por cliente;

Estes resultados encontram-se na Tabela 18.

Tabela 18 – Tabela Lucro/Perda para servidores 4SV

		Previsto	
		0	1
Real	0	Custo de Divulgação: $0,00 Economia: $50.00 **Líquido: $50,00**	Custo de Divulgação: $100,00 Receita: $0.00 **Líquido: - $100,00**
	1	Custo de Divulgação: $0,00 Custo da Oportunidade: $3000,00 **Líquido: - $3000,00**	Custo de Divulgação: $100,00 Receita: $3000,00 **Líquido: $2900,00**

Se multiplicarmos as células desta matriz com as respectivas células da Tabela 15 (página 68), teremos o resultado líquido do uso do modelo, que será de $640.250,00. Este valor é portanto bastante superior ao resultado obtido tradicionalmente por meio de mala direta e revistas.

Este exemplo deixa claro que, neste caso, a utilização de modelos de previsão desenvolvidos e a concentração de esforços nos clientes com maior probabilidade de compra mais que se justificam financeiramente.

Previsão De Vendas A Partir Dos Dados Cadastrais

A motivação para o desenvolvimento deste modelo é determinar se os novos clientes (portanto aqueles que ainda não possuem histórico de vendas) têm tendência ou não de comprar um servidor no futuro.

Foi criada uma base de dados que relacionava os dados cadastrais com a variável SRV_YES, que vale 1 se o cliente comprou ou 0 se o cliente não comprou um servidor 4SV nos últimos dois anos. Deve-se notar que na realidade este modelo prediz se um cliente comprará ou não um servidor, *mas não especificamente se será no próximo trimestre.*

A estrutura da base de dados encontra-se no APÊNDICE M - (página 120).

.O mapa para esta metodologia encontra-se na Figura 15.

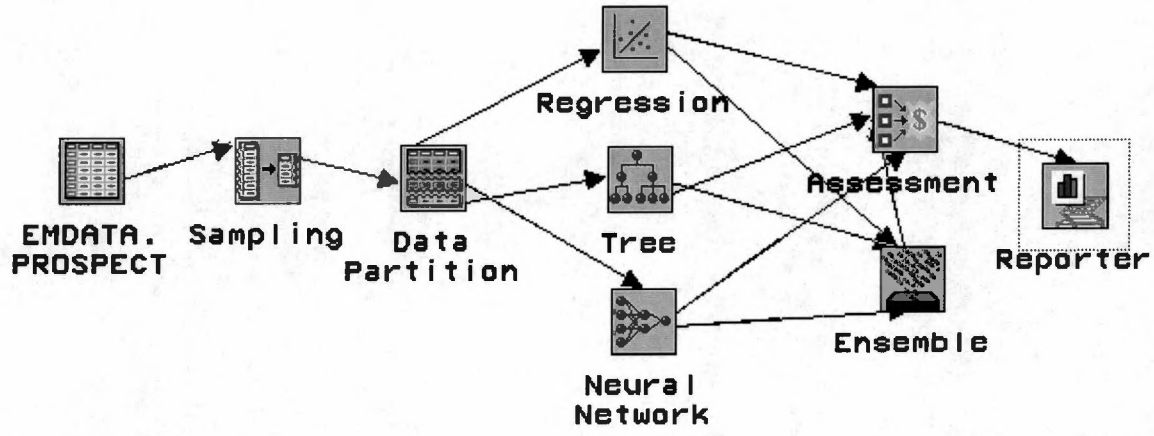


Figura 15 – Mapa da metodologia de previsão utilizando-se somente dados cadastrais

Os nós do mapa acima são basicamente os já discutidos no modelo anterior, com excessão do nó Ensemble, cuja modelagem resultante é a combinação dos nós de Regressão, Árvore e redes Neurais.

O melhor resultado obtido foi por meio do nó Árvore, uma vez que os outros dois (Redes Neurais e Regressão) previram que não haveria vendas para nenhum dos clientes. A matriz de confusão para o modelo árvoreencontra-se na Figura 16.

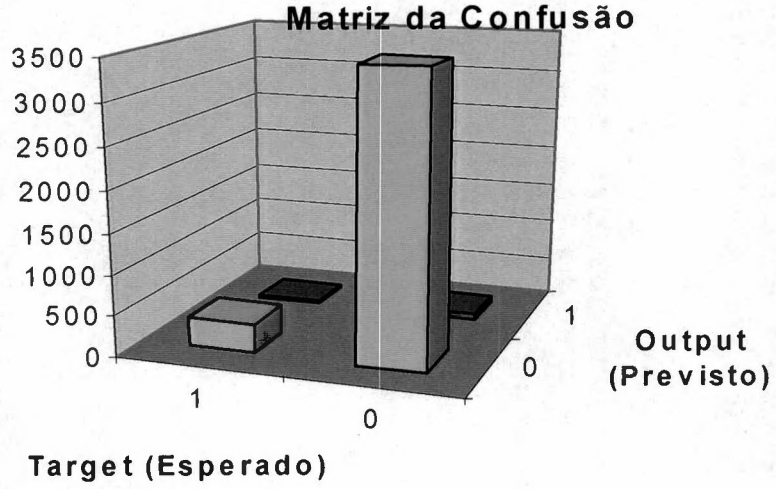


Figura 16 – Matriz de Confusão para o modelo cadastral

Percebe-se portanto que este modelo não deu bons resultados. Uma possível linha de investigação seria rever a hipótese de que os dados cadastrais pouco se alteram ao longo do tempo (ver Premissas gerais dos modelos de previsão de vendas, página 43), e repetir a experiência com dados atualizados a cada semestre. Por absoluta falta de atualização dos dados cadastrais ao longo do tempo, não houve interesse em dar prosseguimento a este modelo.

CAPÍTULO 6 - DESENVOLVIMENTO DE MODELOS PARA DESCOBERTA DE PADRÕES DE COMPRAS

A motivação para este tipo de modelagem é completamente diferente das modelagens anteriores. Enquanto aquelas tinham por objetivo apenas determinar que clientes comprariam ou não servidores e com que probabilidade, a partir de agora a preocupação é entender como e quando os clientes compram servidores. Ou seja, busca-se determinar e reconhecer as relações de causalidade entre compras anteriores e também de outros produtos.

Representação dos dados

Para se buscar regras de associação, deve-se representar os dados de maneira que cada entrada seja a representação de um pedido feito por um cliente, de apenas um produto, e quando foi realizado. Portanto, caso o cliente compre dois produtos, o pedido deve ser desmembrado em dois. Também, a representação da época de compra deve ser representada em formato ordinal

A representação que atende aos requisitos acima foi feita a partir do Arquivo de Representação de Compras (ver APÊNDICE A -, página 98), onde se gerou uma lista de pedidos (ver APÊNDICE M -, página 120). O código SAS para a geração desta base de dados encontra-se no APÊNDICE O -, página 122).

Tabela 19 - Representação da Lista de pedidos – Entrada para Regras de Associação

CAMPO	TIPO	DESCRIÇÃO DO CAMPO
CUST_NUM	Char	Identificação do cliente para pedidos (Comum a esta base de dados e a base de dados cadastrais dos clientes)
PLP	Char	Linha de produto / Descrição / Núm. de referência: 2NT / Periférico / 1 3WS / Workstation / 2 4DM / Modelo 4DM / 3 4ES / Modelo 4ES / 4 4IN / Modelo 4IN / 5 4LT / Modelo 4LT / 6 4PD / Modelo 4PD / 7 4ST / Storage / 8 4SV / Servidor / 9
FY	Char	Ano fiscal
FQ	Char	Quadrimestre fiscal
YYQQ	Char	Representação em um mesmo campo do ano e do trimestre 983 = 1 984 = 2 991 = 3 992 = 4 993 = 5 994 = 6 001 = 7 002 = 8
QTR	Ordinal	Representação ordinal do trimestre da compra, onde o trimestre 1 corresponde ao terceiro trimestre de 1998, e assim sucessivamente. Pode assumir os valores de 1 a 8.
SEM	Ordinal	Representação ordinal do trimestre da compra, onde o semestre 1 corresponde ao segundo semestre de 1998, e assim sucessivamente. Pode assumir os valores de 1 a 4.
ANNUAL	Ordinal	Representação ordinal do ano da compra, onde o ano 1 corresponde ao ano que começa no segunde semestre de 1998 e vai até o segundo semestre de 1999, e assim sucessivamente. Pode assumir os valores de 1 a 2.

Regras de Associação

A descoberta por associação é a identificação de eventos que acontecem juntos numa determinada transação (basket analysis) ou ao longo do tempo (seqüências). Nos dois casos, o descobrimento das regras se dá através da análise da freqüêcia com que os eventos acontecem numa base de dados. Resultados típicos desta abordagem são regras do

tipo "se o item A é parte do evento X, então o item B também fará parte do evento X em Y% das vezes", sendo representada por A => B.

A regra acima pode ser decomposta em duas diferentes partes: a regra propriamente dita (Se A => B) e a estatística associada (ocorre em Y% dos casos).

A regra compõe-se do lado esquerdo ou "causa" (Se A...) e do lado direito ou "efeito" (...então B).

Com relação à estatística associada (Y% das vezes) a cada regra, existem duas importantes medidas, que são o percentual de suporte (suporte) e percentual de confiança (confiança):

- Suporte: porcentagem com que uma regra (lados direito e esquerdo) aparece na base de dados;
- Confiança: porcentagem de casos nos quais o lado direito da regra se verifica dado que o lado esquerdo ocorreu. É uma medida da força de associação entre os dois lados da regra.

Além destas medidas, que podem ser aplicadas tanto para associações e para sequências, existe ainda o "lift" que é específico para associações. O "lift é a divisão da confiança pela confiança esperada (percentual esperado a priori para o lado direito da regra), e significa quanto se ganha em poder de previsão com a regra.

A fim de clarificar os conceitos acima, a seguir é mostrado um exemplo prático:

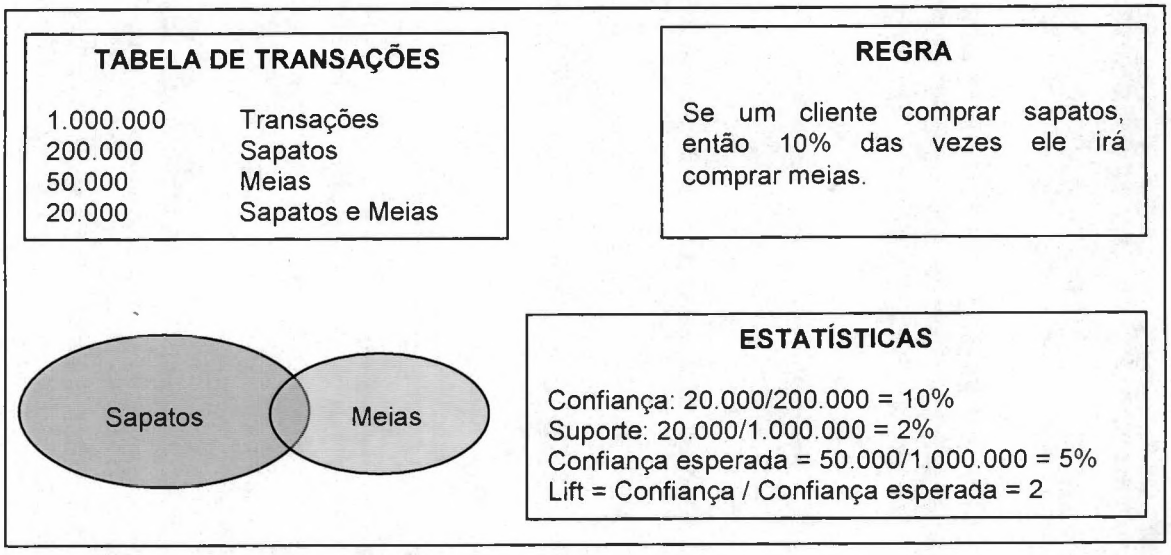


Figura 17 – Exemplo de regras de associação

Primeiro modelo de Regras de Associação

Abaixo segue a configuração do primeiro modelo tentado:

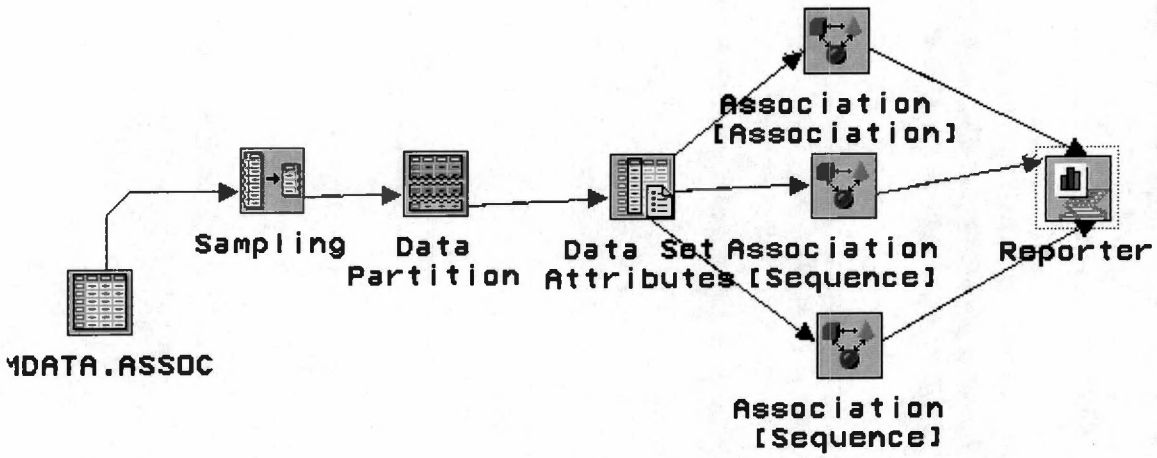


Figura 18 – Mapa do Primeiro modelo para Regras de Associação

A partir da entrada de dados *Assoc.sd2*, foi feita uma amostragem aleatória (Nó Sampling) como registro de 107.825 entradas. Logo após, foi feita a partição (Nó Data Partition) destes dados em três conjuntos: treinamento, validação e testes. Cada um destes conjuntos de dados possuem por volta de 45.000 registros.

Nó Data Set Attributes:

Este nó prepara os dados logicamente para posterior processamento, isto é, informa-se ao sistema qual o papel de cada dado.

Tabela 20 – Identificação do papel dos dados para extração de regras de associação

Name	Keep	New Model Role	Model Role	Measurement	New Measurement
CUST_NUM	Yes	id	id	interval	interval
PLP	Yes	target	target	nominal	nominal
FY	Yes	input	input	ordinal	ordinal
FQ	Yes	input	input	ordinal	ordinal
YYQQ	Yes	input	input	ordinal	ordinal
QTR	Yes	sequence	sequence	ordinal	ordinal
SEM	Yes	sequence	sequence	ordinal	ordinal
ANNUAL	Yes	sequence	sequence	ordinal	ordinal

Nós de Associação

Após o nó Data Set Attributes, aparecem três nós de Associação. O primeiro deles processa os dados sem levar em consideração a informação de quando os produtos foram comprados, isto é, enxerga as transações como se fossem cestas de mercadorias (Basket analysis). Os outros dois nós de associação tratam os dados como seqüências, sendo que no primeiro deles as compras são consolidadas a cada trimestre, e no segundo a cada semestre.

Modelo Associação

O primeiro nó tinha como critério só descobrir regras com suporte maior que 5% e confiança maior que 10%. Não foi encontrada nenhuma regra que satisfizesse estes requerimentos. Isto leva a crer que as "cestas de compras" são muito diversas entre si, não possuindo porcentualmente nenhum mix de produtos que se sobressaia sobre os demais. *Veremos no segundo modelo que isto não significa que não existam mix bem definidos, apenas são tantos e homogêneos que não se sobressaem, portanto não gerando um número crítico de supporte ou confiança.*

Modelo Sequencial – Trimestral

Neste modelo o fator tempo foi levado em consideração, em intervalos trimestrais. O critério é que a regra tenha pelo menos 2% de suporte. Os resultados obtidos foram:

Tabela 22 – Resultados de Sequências - Trimestral

NITEMS	COUNT	SUPPORT	CONF	RULE
2	1103	3.3	34.19	4DM ==> 4DM
2	940	2.81	29.14	4DM ==> 2NT
2	918	2.75	27.68	2NT ==> 2NT
2	761	2.28	22.94	2NT ==> 4DM

Resumidamente, podemos observar que clientes que compram os produtos 2NT ou 4DM tem probabilidade de aproximadamente 30% de comprar um dos dois produtos no próximo trimestre. Em qualquer dos casos, entretanto, este padrão só se verifica de 2 a 3% das transações.

Pode-se pensar em como explorar melhor esta informação. Talvez uma promoção dos dois produtos vendidos juntos possa encontrar interesse dos clientes. Mas somente 2 a 3% deles, o que não é uma porcentagem grande o suficiente para anunciar nos meios de comunicação. Mais, provavelmente os vendedores de X já perceberam este fato, e não deve ser nenhuma novidade para ninguém. Portanto, embora seja um resultado interessante, não teria muita aplicação prática.

Modelo Sequencial – Semestral

Este é exatamente o mesmo procedimento do caso anterior com exatamente os mesmos critérios, apenas o fator tempo foi considerado em intervalos semestrais. Os resultados obtidos foram:

Tabela 23 – Resultados de Sequências – Semestral

COUNT	SUPPORT	CONF	RULE
939	2.81	33.23	4DM ==> 4DM
816	2.44	28.87	4DM ==> 2NT
764	2.29	26.49	2NT ==> 2NT

Observa-se que pouco mudou. Os clientes que compram o produto 4DM tem probabilidade de aproximadamente 30% de comprar outro 4DM ou 2NT no próximo semestre, e aquele que comprou 2NT tem por volta de 30% de comprar outro 2NT. Em qualquer dos casos, entretanto, as regras obtidas só se verificam de 2 a 3% das transações. Este resultado também não parece ser de muita utilidade prática.

Pode-se pensar em como explorar melhor esta informação. Talvez uma promoção dos dois produtos vendidos juntos possa encontrar interesse dos clientes. Mas somente 2 a 3% deles, o que não é uma porcentagem grande o suficiente para anunciar nos meios de comunicação. Mais, provavelmente os vendedores de X já perceberam este fato, e não deve ser nenhuma novidade para ninguém. Portanto, embora seja um resultado interessante, não teria muita aplicação prática.

Segundo modelo de Regras de Associação

A motivação para este segundo modelo é tentar criar grupos de clientes que possuam hábitos de compras similares. A hipótese é que uma vez agrupados, estes hábitos de

compras devem resultar em regras com maior suporte, porque não estão mais misturados com outros hábitos de compras completamente diversos que apenas diluiriam o valor de suporte uns dos outros.

A maneira de se verificar esta hipótese é primeiramente classificar os clientes em diferentes agrupamentos, para só então tentar extrair regras de associação ou sequências.

A arquitetura de implementação desta idéia encontra-se abaixo:

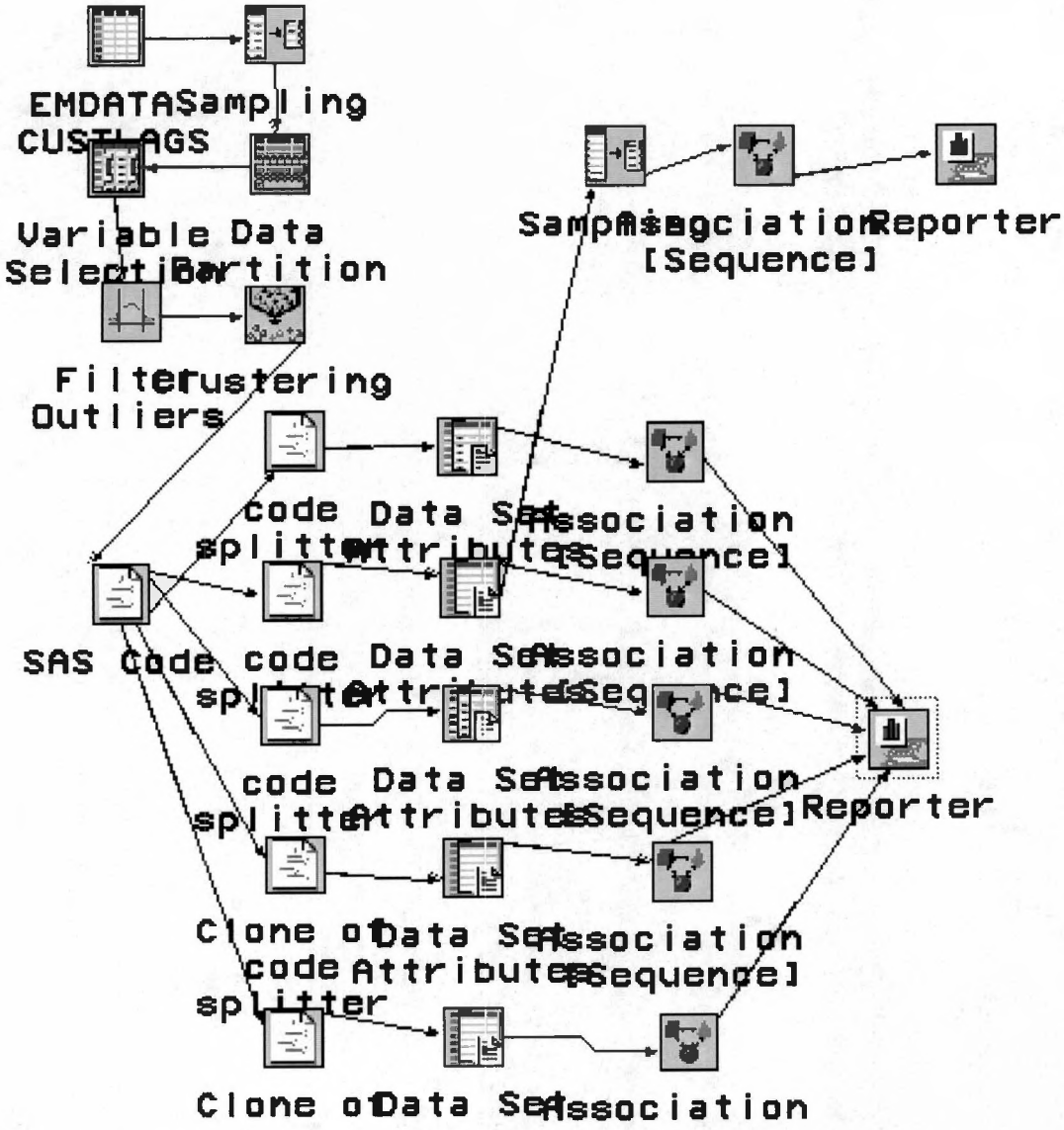


Figura 19 - Mapa do Segundo modelo para Regras de Associação

Preparação dos dados

Os três nós iniciais (emdata.custlags, sampling e data partition) são exatamente iguais aos casos estudados anteriormente (ver Nó EMDATA.CUSTLAGS, página 49). Como agora não existe uma variável objetivo, o nó variable selection não realizou seleção de variáveis com base em valores significativos de R^2 conforme havia sido feito anteriormente. Portanto, o nó variable selection eliminou apenas as variáveis em que mais de 50% dos registros eram valores em branco. O nó Filter Outliers a seguir eliminou apenas os registros que possuíam valores em branco.

A eliminação de valores em branco realizada em duas etapas (primeiro as variáveis e depois os registros) evitou que um grande número de registros seja eliminado, caso fosse somente aplicada a eliminação dos registros que contivessem campos em branco.

Nó Clustering

Este nó faz o agrupamento dos dados baseado na similaridade dos dados. Foi utilizado o critério de mínimos quadrados (least-squares) para julgar a similaridade.

O resultado geral da análise de agrupamentos encontra-se na Tabela 24.

Tabela 24 - Resultado da Análise de Agrupamentos

Frequency of Cluster	Root-Mean-Square Standard Deviation	Nearest Cluster	Distance to Nearest Cluster
660	1.299506215	2	5.66944046
3988	0.683561816	3	3.636362124
3533	0.925369199	2	3.636362124
1825	0.746001266	3	5.404995178
212	2.054534663	1	6.994114831

Onde:

- Cluster: identificação do agrupamento (cluster). Nota-se que foram encontrados 5 agrupamentos;
- Frequency of Clusters: número de elementos em cada agrupamento;
- RMS Standard Deviation (RMSSD): Desvio padrão dos elementos do cluster. Mede basicamente a variação dos elementos do agrupamento ao redor do respectivo centróide;
- Nearest Cluster: Identifica o agrupamento mais próximo;

- Distance to Nearest Cluster: apresenta a distância para o agrupamento mais próximo.

Nota-se que para todos os agrupamentos o RMSSD é pelo menos 3.4 vezes maior que a distâncias entre os agrupamentos mais próximos. Espera-se assim que os agrupamentos formados possuam elementos similares entre si, e que elementos de diferentes agrupamentos sejam significativamente diferentes entre si.

Na figura a seguir são mostradas as médias padronizads de cada variável por agrupamento:

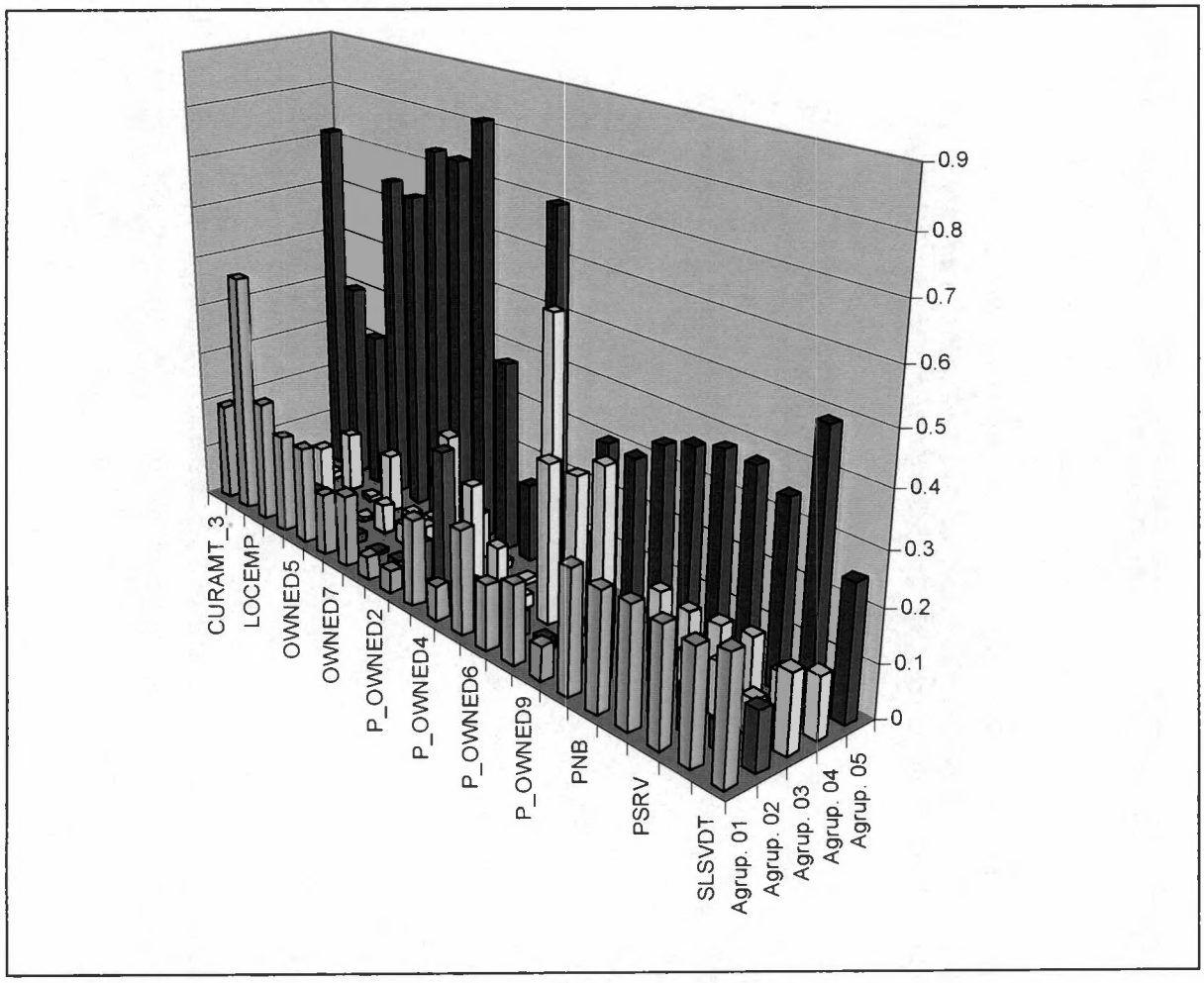


Figura 20 –Média padronizada das variáveis por agrupamento

Os dados da Figura 20 estão no APÊNDICE O - (página 122).

Observa-se que a composição dos valores médios padronizados em cada agrupamento é bastante distinto de agrupamento para agrupamento, dadas as diferenças de amplitude dos valores das variáveis e do não alinhamento das ocorrências dos vales e picos.

Uma breve descrição dos perfis dos agrupamentos seria:

- O agrupamento 1 realiza gastos maiores que a média para equipamentos 4PD (valores comparativamente altos para a variável CURAMT_7);
- O agrupamento 2 possui mais equipamentos 4DM que a média (valores comparativamente altos para a variável P_OWNED3);
- O agrupamento 3 possui respectivamente mais equipamentos 4LT, 4DM e servidores 4SV que a média (valores comparativamente altos para as variáveis P_OWNED6, P_OWNED7 e P_OWNED9);
- O agrupamento 4 possui mais equipamentos 4IN que a média (valores comparativamente altos para a variável P_OWNED5);
- O agrupamento 5 trata-se de grandes clientes (valores comparativamente altos para a maioria das variáveis).

O arquivo resultante deste procedimento adiciona uma nova variável, _SEGMENT_, que designa a que agrupamento pertence cada registro. Esta informação é utilizada a seguir para criar um conjunto de registros para cada agrupamento.

Nó SAS Code

Este nó permite que sejam dadas instruções na Linguagem SAS para processamento dos dados.

Neste caso, foi pedido um "merge" entre os arquivos de pedidos com o arquivo que possui a variável _SEGMENT_ (ver instruções no APÊNDICE O -, página 122). Desta maneira, obtém-se no arquivo de saída uma relação entre os pedidos e o agrupamento a que pertencem, através da identificação do cliente.

Nó Code Splitter

Nota-se que aparecem 5 nós Code Splitter. Cada nó destes seleciona apenas os pedidos de cada agrupamento para posterior procura por Regras de Associação (ver exemplo destes códigos no APÊNDICE Q -, página 124).

As seguir os pedidos de cada agrupamento foram processados em nós de Associação (ver Nós de Associação, página 78).

Será feita uma análise dos resultados gerais para cada agrupamento, e também uma análise direcionada para a compra de servidores (4SV) que é o objetivo de análise deste trabalho

Agrupamento 01

Este agrupamento possui 660 registros de vendas, e foram encontradas 4808 regras de Associação. Foram selecionadas as regras que possuem o maior valor de suporte e estão indicadas na Tabela 25.

Tabela 25 – Regras ordenadas por Suporte (> 35%) para o Agrupamento 1

COUNT	SUPPORT	CONF	RULE
421	63.79	72.71	4DM ==> 4DM
416	63.03	74.82	4IN ==> 4DM
402	60.91	69.43	4DM ==> 2NT
395	59.85	68.22	4DM ==> 4IN
386	58.48	78.14	2NT ==> 4DM
349	52.88	70.65	2NT ==> 2NT
333	50.45	59.89	4IN ==> 2NT
307	46.52	62.15	2NT ==> 4IN
291	44.09	50.26	4DM ==> 4DM & 2NT
281	42.58	67.87	4DM & 2NT ==> 4DM
279	42.27	56.48	2NT ==> 4DM & 2NT
278	42.12	67.15	4DM & 2NT ==> 2NT
266	40.3	71.12	4IN & 2NT ==> 4DM
266	40.3	47.84	4IN ==> 4DM & 2NT
265	40.15	45.77	4DM ==> 4IN & 2NT
249	37.73	59.14	4DM ==> 4DM ==> 4DM
247	37.42	42.66	4DM ==> 4DM & 4IN
240	36.36	60.76	4DM ==> 4IN ==> 4DM
240	36.36	57.97	4DM & 2NT ==> 4IN
237	35.91	66.95	4DM & 4IN ==> 4DM
237	35.91	58.96	4DM ==> 2NT ==> 4DM
236	35.76	42.45	4IN ==> 4IN
232	35.15	62.03	4IN & 2NT ==> 2NT

Nota-se que os pedidos deste agrupamento são na maioria de equipamentos 4DM e 2NT. Tipicamente 60% (suporte) dos compradores compram um destes equipamentos para logo no trimestre seguinte comprar um deles novamente. Quando compram um destes equipamentos terão por volta de 70% (confiança) de comprar outro deles.

Abaixo são mostradas as regras com maior valor de confiança (> 85%):

Tabela 26 - Regras ordenadas por Confiança (> 85%) para Agrupamento 1

COUNT	SUPPORT	CONF	RULE
15	2.27	93.75	2NT & 4PD ==> 4DM & 4IN & 2NT & 4PD ==> 4PD
13	1.97	92.86	4PD ==> 2NT & 4SV & 4PD ==> 2NT
25	3.79	92.59	4DM & 4LT ==> 4DM & 4IN & 2NT ==> 2NT
22	3.33	91.67	4DM & 2NT & 4LT ==> 4DM & 4IN & 2NT ==> 2NT
16	2.42	88.89	2NT & 4PD ==> 4DM & 4IN & 4PD ==> 4PD
15	2.27	88.24	4IN & 2NT & 4SV ==> 4IN ==> 4IN
14	2.12	87.5	3WS ==> 4DM & 4IN ==> 4DM
13	1.97	86.67	4DM & 4IN & 4LT ==> 4DM & 4IN & 2NT ==> 2NT
13	1.97	86.67	4DM & 4PD & 4LT ==> 4DM ==> 4DM
13	1.97	86.67	4DM & 4IN & 2NT & 4LT ==> 4DM & 4IN & 2NT ==> 2NT
35	5.3	85.37	2NT & 4SV ==> 4IN & 2NT ==> 2NT
23	3.48	85.19	4DM & 4LT ==> 4DM & 4IN & 2NT ==> 4DM & 2NT
23	3.48	85.19	4DM & 4LT ==> 4DM & 4IN & 2NT ==> 4DM
23	3.48	85.19	4DM & 2NT & 4LT ==> 4IN & 2NT ==> 2NT
23	3.48	85.19	2NT & 4SV ==> 4DM & 4IN & 2NT ==> 2NT
17	2.58	85	4DM & 2NT & 4LT ==> 4DM & 4LT ==> 2NT

A tabela com os maiores valores de confiança acima mostra que maioria das regras resultam na venda no último trimestre dos equipamentos 4DM ou 2NT, e as compras anteriores contém em sua totalidade um dos equipamentos citados. Portanto, esta análise não traz muita novidade à análise realizada com base nos valores de suporte. Entretanto, uma análise mais cuidadosa revela dois pontos interessantes.

O primeiro ponto vem da observação que o maior valor para confiança deste agrupamento resulta na venda para equipamentos 4PD. A filtragem de todas as regras com confiança maior que 75% para 4PD resulta em:

Tabela 27 - Regras para 4PD ordenadas por Confiança (> 75%) para Agrupamento 1

COUNT	SUPPORT	CONF	RULE
15	2.27	93.75	2NT & 4PD ==> 4DM & 4IN & 2NT & 4PD ==> 4PD
16	2.42	88.89	2NT & 4PD ==> 4DM & 4IN & 4PD ==> 4PD
16	2.42	84.21	4PD ==> 4DM & 4IN & 2NT & 4PD ==> 4PD
20	3.03	83.33	4PD ==> 4DM & 4IN & 4PD ==> 4PD
27	4.09	79.41	4PD ==> 4IN & 4PD ==> 4PD
21	3.18	77.78	2NT & 4PD ==> 4IN & 4PD ==> 4PD

Nota-se que embora o suporte destes casos seja pequeno, existe uma fortissima probabilidade de compra de 4PD para os clientes que satisfaçam o lado esquerdo das regras. Embora não justifique uma mala direta para estes compradores, uma aproximação mais direta (visita comercial, contato telefônico) para estes compradores tem grandes probabilidades de resultar na venda de 4PD

O segundo ponto diz respeito à compra do equipamento 4IN:

Tabela 28 - Regras para 4IN ordenadas por Confiança com Suporte > 40% para Agrupamento 1

COUNT	SUPPORT	CONF	RULE
395	59.85	68.22	4DM ==> 4IN
307	46.52	62.15	2NT ==> 4IN
265	40.15	45.77	4DM ==> 4IN & 2NT

Nota-se que embora em um primeiro momento não fosse aparente, existem eventos associados à compra do equipamento 4IN com consideráveis níveis de suporte e confiança. Portanto, compradores deste agrupamento também compram o equipamento 4IN.

Fica claramente identificado que os compradores deste agrupamento não só compram os equipamentos 4DM e 2NT sempre e repetidamente, mas também compram 4IN com uma frequência significativa.

Uma aplicação destes resultados no mundo real seria realizar uma mala direta para os clientes que compram 4DM e 2NT, propondo a venda de equipamentos 2NT, 4DM e uma condição especial para 4IN. Um contato telefônico para os compradores que satisfaçam o lado esquerdo das regras para 4PD teriam alta probabilidade de gerar vendas do equipamento.

Como o interesse deste estudo é a compra de servidores (4SV), observamos que não existem regras com suporte maior que 15%. Apresentamos abaixo as regras associadas com maior confiança:

Tabela 29 - Regras para 4SV ordenadas por Confiança (> 25%) para Agrupamento 1

COUNT	SUPPORT	CONF	RULE
15	2.27	31.25	4PD ==> 4DM & 4IN ==> 4SV
18	2.73	29.03	4PD ==> 2NT & 4PD ==> 4SV
23	3.48	27.38	4PD ==> 4PD ==> 4SV

Nota-se que os compradores têm pouca probabilidade de comprar servidores 4SV., possivelmente não justificando esforços de venda para este agrupamento.

Agrupamento 2

Este agrupamento possui 3988 registros de vendas. Para este agrupamento foram encontradas 55 regras de Associação. Foram selecionadas as regras que possuem o valor de suporte maior ou igual a 10%:

Tabela 30 - Regras ordenadas por Suporte (> 10%) para Agrupamento 02

COUNT	SUPPORT	CONF	RULE
1541	38.64	84.53	4DM ==> 4DM
992	24.87	54.42	4DM ==> 2NT
844	21.16	67.04	2NT ==> 4DM
741	18.58	40.65	4DM ==> 4DM & 2NT
662	16.6	52.58	2NT ==> 2NT
650	16.3	62.32	4DM & 2NT ==> 4DM
599	15.02	38.87	4DM ==> 4DM ==> 4DM
543	13.62	52.06	4DM & 2NT ==> 2NT
505	12.66	40.11	2NT ==> 4DM & 2NT
405	10.16	26.28	4DM ==> 4DM ==> 2NT
401	10.06	38.45	4DM & 2NT ==> 4DM & 2NT

Tabela 31 - Regras ordenadas por Confiança (> 50%) para Agrupamento 02

COUNT	SUPPORT	CONF	RULE
1541	38.64	84.53	4DM ==> 4DM
196	4.91	73.41	4SV ==> 4DM
110	2.76	68.32	2NT & 4SV ==> 4DM
137	3.44	67.82	4IN ==> 4DM
104	2.61	67.1	4DM & 4SV ==> 4DM
844	21.16	67.04	2NT ==> 4DM
650	16.3	62.32	4DM & 2NT ==> 4DM
992	24.87	54.42	4DM ==> 2NT
86	2.16	53.42	2NT & 4SV ==> 2NT
662	16.6	52.58	2NT ==> 2NT
543	13.62	52.06	4DM & 2NT ==> 2NT
139	3.49	52.06	4SV ==> 2NT
101	2.53	51.53	4SV ==> 4DM ==> 4DM
79	1.98	50.97	4DM & 4SV ==> 2NT
79	1.98	50.64	4DM ==> 4SV ==> 4DM

Observa-se que as transações que ocorrem neste agrupamento este agrupamento envolve quase que exclusivamente os equipamentos 2NT e 4DM. Os valores para suporte são significativamente menores que os do Agrupamento 1, indicando que os compradores deste cluster não compram os equipamentos citados com alta frequencia.

Encontram-se alguns valores de confiança razoáveis, mas que em geral não acrescentam muito às informações obtidas com a análise de suporte, posto que se tratam dos mesmos equipamentos do lado esquerdo das regras. com valores significativamente menores para suporte.

Somente dois casos parecem ser de interesse: compradores de servidores 4SV e 4IN compram 4DM 73.41% e 67.82% das vezes. Ainda assim estes valores de confiança não são tão altos como os encontrados no Agrupamento 1, de maneira que provavelmente não se justificam contatos como telefonemas para estes compradores.

Não existem regras para servidores 4SV com suporte maior que 4% ou confiança maior que 10%, indicando que compradores deste agrupamento normalmente não compram servidores.

Por ser este o agrupamento com maior número de registros, poder-se-ia argumentar como o número de registros pode afetar este resultado. Mais adiante em *Observação sobre tamanhos de Agrupamentos Diferentes* (página 92) será feita uma análise minuciosa a este respeito.

Agrupamento 3

Este agrupamento possui 3533 registros de vendas. Para este agrupamento foram encontradas 45 regras de Associação. Foram selecionadas as regras que possuem o valor de suporte maior ou igual a 5%:

Tabela 32 - Regras ordenadas por Suporte (> 5%) para Agrupamento 03

COUNT	SUPPORT	CONF	RULE
551	15.6	61.56	4DM ==> 2NT
535	15.14	51.44	2NT ==> 2NT
341	9.65	38.1	4DM ==> 4DM
281	7.95	27.02	2NT ==> 4DM
261	7.39	60	4SV ==> 2NT
249	7.05	62.72	4DM & 2NT ==> 2NT
193	5.46	36.07	2NT ==> 2NT ==> 2NT

Tabela 33 - Regras ordenadas por Confiança (> 50%) para Agrupamento 03

COUNT	SUPPORT	CONF	RULE
78	2.21	69.03	2NT & 4LT ==> 2NT
131	3.71	65.17	4LT ==> 2NT
70	1.98	64.22	4DM & 4SV ==> 2NT
249	7.05	62.72	4DM & 2NT ==> 2NT
551	15.6	61.56	4DM ==> 2NT
142	4.02	61.21	2NT & 4SV ==> 2NT
261	7.39	60	4SV ==> 2NT
139	3.93	59.4	4IN ==> 2NT
105	2.97	57.69	2NT & 4PD ==> 2NT
168	4.76	51.53	4PD ==> 2NT
535	15.14	51.44	2NT ==> 2NT

A análise deste agrupamento é muito similar ao anterior, sendo que os números de suporte e confiança são ainda menores. Portanto os compradores compram com menos frequência e mais aleatoriamente que os compradores do Agrupamento 2.

Não existem regras para servidores 4SV com suporte maior que 5% ou confiança maior que 17%, indicando que compradores deste agrupamento normalmente não compram servidores.

Agrupamento 4

Este agrupamento possui 1825 registros de vendas. Para este agrupamento foram encontradas 50 regras de Associação. Foram selecionadas as regras que possuem o valor de suporte maior ou igual a 10%, e valores de confiança maiores que 50%:

Tabela 34 – Regras ordenadas por Suporte (> 10%) para Agrupamento 04

COUNT	SUPPORT	CONF	RULE
396	21.7	56.49	4IN ==> 2NT
280	15.34	39.94	4IN ==> 4IN
272	14.9	49.54	2NT ==> 2NT
252	13.81	45.9	2NT ==> 4IN
217	11.89	75.09	4DM ==> 4IN
186	10.19	55.19	4IN & 2NT ==> 2NT

Observa-se que os valores de suporte não são muito significativos, portanto não se verifica uma tendência comum de compra. Entretanto a análise dos valores de confiança provê alguns pontos interessantes:

Tabela 35 - Regras ordenadas por Confiança (> 50%) para Agrupamento 4

COUNT	SUPPORT	CONF	RULE
217	11.89	75.09	4DM ==> 4IN
77	4.22	65.25	2NT & 4DM ==> 4IN
396	21.7	56.49	4IN ==> 2NT
54	2.96	55.67	4SV ==> 4IN
186	10.19	55.19	4IN & 2NT ==> 2NT
272	14.9	49.54	2NT ==> 2NT

Verifica-se na tabela acima que compradores que compram 4DM têm uma probabilidade de 75% de também comprarem 4IN.

Não existem regras para servidores 4SV com suporte maior que 3% ou confiança maior que 8%, indicando que compradores deste agrupamento normalmente não compram servidores.

Agrupamento 5

Este agrupamento possui 212 registros de vendas. Para este agrupamento foram encontradas mais de 65.000 regras de Associação. Foram selecionadas as regras que possuem o valor de suporte maior ou igual a 70%:

Tabela 36 - Regras ordenadas por Suporte (>= 70%) para Agrupamento 5

COUNT	SUPPORT	CONF	RULE
181	85.38	89.6	4DM ==> 2NT
181	85.38	89.6	4DM ==> 4SV
181	85.38	88.73	2NT ==> 2NT
180	84.91	89.11	4DM ==> 4DM
178	83.96	87.25	2NT ==> 4SV
177	83.49	83.49	4SV ==> 2NT
173	81.6	84.8	2NT ==> 4DM
165	77.83	77.83	4SV ==> 4DM
164	77.36	80.39	2NT ==> 4SV & 2NT
163	76.89	80.69	4DM ==> 4SV & 2NT
159	75	78.71	4DM ==> 2NT & 4DM
158	74.53	82.72	4SV & 2NT ==> 2NT
154	72.64	85.56	4IN ==> 2NT
153	72.17	75	2NT ==> 2NT & 4DM
152	71.7	86.36	2NT & 4DM ==> 2NT
152	71.7	83.98	2NT ==> 2NT ==> 2NT
151	71.23	71.23	4SV ==> 4SV
149	70.28	82.78	4IN ==> 4SV
147	69.34	83.52	2NT & 4DM ==> 4DM
147	69.34	83.52	2NT & 4DM ==> 4SV
147	69.34	81.67	4DM ==> 4DM ==> 2NT

Observa-se que predominantemente há transações com altos valores de suporte e confiança que resultam na compra de 2NT, 4DM e servidores 4SV. Isto significa que os compradores deste agrupamento compram uma grande quantidade e apresentam uma grande variedade de padrões de compra destes equipamentos. É de se esperar que este seja o conjunto dos grandes compradores.

Existem 95 regras com valor de confiança igual a 100% e baixo suporte, como pode ser observado na Tabela 37.

Tabela 37 – Amostra das 10 primeiras regras ordenadas por Confiança para o agrupamento 5

COUNT	SUPPORT	CONF	RULE
17	8.02	100	4SV & 4IN & 4PD ==> 2NT & 4LT ==> 2NT
16	7.55	100	4SV & 2NT & 4IN & 4PD ==> 2NT & 4LT ==> 2NT
9	4.25	100	2NT & 4DM & 3WS & 4ST ==> 2NT
9	4.25	100	4SV & 4DM & 3WS & 4ST ==> 4DM
9	4.25	100	4DM & 3WS & 4ST ==> 2NT
9	4.25	100	4SV & 4DM & 3WS & 4ST ==> 2NT
9	4.25	100	4SV & 2NT & 3WS & 4ST ==> 4DM
9	4.25	100	4DM & 3WS & 4ST ==> 4DM
9	4.25	100	3WS & 4ST ==> 4DM

Pode-se observar que são regras muito específicas, que em geral confirmam a observação de quem compra 4DM, 2NT ou 4SV antes tem alta probabilidade de comprar um destes equipamentos de novo.

Uma análise das regras ordenadas pelos valores de confiança mais rígida (com suporte >10%) revela basicamente o mesmo tipo de informação.

Tabela 38 - Regras ordenadas por Confiança (> 90%), com Suporte > 10% para Agrupamento 5

COUNT	SUPPORT	CONF	RULE
25	11.79	96.15	2NT & 4IN & 4PD ==> 2NT & 4LT ==> 2NT
30	14.15	93.75	2NT & 4PD ==> 4DM & 4LT ==> 2NT
29	13.68	93.55	2NT & 4IN ==> 4DM & 4IN & 4LT ==> 4DM
29	13.68	93.55	2NT & 4PD ==> 2NT & 4DM & 4LT ==> 2NT
27	12.74	93.1	4IN & 4PD ==> 2NT & 4LT ==> 2NT
26	12.26	92.86	2NT & 4DM & 4IN ==> 4DM & 4IN & 4LT ==> 4DM
26	12.26	92.86	2NT & 4DM & 4IN ==> 4DM & 4IN & 4LT ==> 2NT
26	12.26	92.86	2NT & 4IN ==> 2NT & 4DM & 4IN & 4LT ==> 2NT
26	12.26	92.86	2NT & 4IN ==> 2NT & 4DM & 4IN & 4LT ==> 4DM
25	11.79	92.59	2NT & 4IN & 4PD ==> 4LT ==> 2NT
61	28.77	92.42	2NT & 4LT ==> 4LT ==> 2NT
24	11.32	92.31	2NT & 4DM & 4IN ==> 2NT & 4DM & 4IN & 4LT ==> 2NT
24	11.32	92.31	2NT & 4DM & 4IN ==> 2NT & 4DM & 4IN & 4LT ==> 4DM
34	16.04	91.89	4SV & 4LT & 4PD ==> 2NT
22	10.38	91.67	4SV & 4DM & 4LT & 4PD ==> 2NT
30	14.15	90.91	4IN ==> 4DM & 4IN & 4LT ==> 4DM
59	27.83	90.77	2NT & 4LT ==> 2NT & 4LT ==> 2NT
49	23.11	90.74	4LT ==> 4DM & 4LT ==> 2NT
29	13.68	90.63	4DM ==> 4SV & 4LT & 4PD ==> 2NT
47	22.17	90.38	4LT ==> 2NT & 4DM & 4LT ==> 2NT
28	13.21	90.32	4SV & 4DM & 3WS ==> 2NT
28	13.21	90.32	4SV & 4DM & 3WS ==> 4DM
28	13.21	90.32	4DM & 4IN ==> 4DM & 4IN & 4LT ==> 4DM
28	13.21	90.32	2NT & 4IN ==> 4DM & 4IN & 4LT ==> 2NT
65	30.66	90.28	2NT ==> 2NT & 4DM & 4LT ==> 2NT
64	30.19	90.14	4LT ==> 4LT ==> 2NT

Padrões de compra de servidores 4SV do agrupamento 5

A fim de se avançar nesta análise, faz-se necessário focar em cada produto especificamente. De acordo com o objetivo deste estudo, será feita uma análise mais minuciosa para os servidores 4SV deste agrupamento.

Foram selecionadas as ocorrências de maior confiança que levam à compra de servidores 4SV. Esta seleção encontra-se na Tabela 39.

Tabela 39 - Regras para 4SV ordenadas por Confiança (Suporte > 60%) para Agrupamento 5

COUNT	SUPPORT	CONF	RULE
181	85.38	89.6	4DM ==> 4SV
178	83.96	87.25	2NT ==> 4SV
147	69.34	83.52	2NT & 4DM ==> 4SV
149	70.28	82.78	4IN ==> 4SV
163	76.89	80.69	4DM ==> 4SV & 2NT
164	77.36	80.39	2NT ==> 4SV & 2NT
140	66.04	77.35	2NT ==> 2NT ==> 4SV
136	64.15	77.27	2NT & 4DM ==> 4SV & 2NT
137	64.62	76.11	4DM ==> 4DM ==> 4SV
129	60.85	74.57	2NT ==> 4DM ==> 4SV
133	62.74	73.48	4DM ==> 2NT ==> 4SV
131	61.79	72.78	4IN ==> 4SV & 2NT
145	68.4	71.78	4DM ==> 4SV & 4DM
151	71.23	71.23	4SV ==> 4SV
128	60.38	70.72	2NT ==> 2NT ==> 4SV & 2NT
138	65.09	67.65	2NT ==> 4SV & 4DM
139	65.57	65.57	4SV ==> 4SV & 2NT
129	60.85	63.86	4DM ==> 4SV & 2NT & 4DM

A tabela anterior mostra mais uma vez que quem compra 4DM, 2NT ou servidor 4SV tem alta probabilidade de comprar outro servidor 4SV. Revela também que, neste agrupamento, a compra de 4IN é normalmente seguida pela compra de um servidor 4SV.

Um outro tipo de informação que se pode obter é comparar a compra de produtos isolados com a compra dos mesmos produtos combinados levam à aquisição de um servidor 4SV. Por exemplo, observando-se a Tabela 39, percebe-se que os padrões de compras 4DM→4SV, 2NT→4SV e 4DM&2NT→4SV têm respectivos suportes de 85,38%, 83,96% e 69,34%. Portanto, a compra de servidores 4SV ocorre 69,34% das vezes a partir compra dos equipamentos 4DM e 2NT em conjunto, 16,04% das vezes ocorre a partir da compras dos equipamentos 4DM sozinhos e 14,62% ocorre a partir da compra de equipamentos 2NT sozinhos. Este resultado é um indicativo que a oferta do mix 4DM e 2NT faz sentido e leva a compra de 4SV.

Observação sobre tamanhos de Agrupamentos Diferentes

Uma questão que pode ser levantada é com relação aos diferentes tamanhos de agrupamentos que foram obtidos, e ao número de regras obtidas em cada um deles.

Intuitivamente é de se esperar que para agrupamentos menores tende-se a obter um número maior de regras, porque as similaridades existentes em poucos elementos tenderiam a compor um percentual maior em função do baixo número de elementos. Estas

mesmas similaridades, num agrupamento maior, poderiam ter seu percentual diluído devido ao grande número de elementos. Esta hipótese explicaria porque no Agrupamento 2 com 3988 elementos foram encontradas apenas 55 regras, enquanto o Agrupamento 5 com 212 elementos foram encontradas mais de 65.000 regras.

A fim de testar esta hipótese, foi feita uma sub-amostragem do Agrupamento 2 (Ver Figura 19, página 80, a amostragem realizada a partir dos dados do Agrupamento 2), de forma que foram escolhidos ao acaso o mesmo número de elementos do Agrupamento 5 (212 elementos). Uma vez feita a amostragem, utilizou-se o nó *Association* para a procura por regras de associação.

Era então portanto de se esperar que o número de regras fosse grande, *mas nenhuma regra foi encontrada para esta amostra.* Isto é um indicativo que os padrões de compra no Agrupamento 2 não são de fato tão claros, não sendo possível capturá-los numa amostra menor.

Este fato nos faz suspeitar que talvez a hipótese anterior seja falsa. Outras simulações deveriam ser conduzidas para se verificar as suspeitas acima, podendo ser objeto de estudo de trabalhos posteriores.

CAPÍTULO 7 - CONCLUSÃO

Foi mostrado que técnicas de Datamining podem ajudar uma empresa a entender melhor os hábitos de compra dos consumidores, quem são eles, quais aqueles que tem maior probabilidade de compra. No estudo de caso, X pôde entender melhor o mercado em que atua, dando uma maior competitividade. a fazer previsões sobre vendas, perfis de clientes e a dinâmica de compra de produtos. Conforme discutido no Capítulo 2, estas características se alinham com os objetivos estratégicos das empresas atuais.

Entretanto, como os resultados obtidos com Datamining se justificam economicamente?

Valor do Conhecimento e Informação

Todas as empresas têm no conhecimento um elemento economicamente importante. É por meio do conhecimento e das informações que as empresas prosperam. O conhecimento (cliente, mercado, concorrentes, fornecedores e compradores) permite à empresa apontar as influências no lucro e prejuízo, o que agrega ou não valor a um produto ou serviço, e a tomar ações que de fato aumentem o valor para o acionista. À medida que a empresa é capaz de acessar a informação de uma maneira mais precisa, passa a ter melhores condições de ofertar melhores produtos e serviços mais adequados para seus clientes, escolher e diferenciar aqueles que são melhores. Assim, pode ofertar produtos melhores e serviços mais adequados, atender diferenciadamente os melhores clientes, tudo isto realizado mais eficientemente com os mesmos recursos.

Portanto, informação e conhecimento são atualmente vistos como capital intelectual (Stewart, 1997), tornando-se um ativo tão ou mais importante que ativos tradicionais como capital, fábricas, etc. De fato, existem muitas empresas de sucesso cujo principal ativo é a informação, tais como jornais, emissoras de televisão, database marketing, etc. Embora não sejam contabilizados, têm valor real. Esta é uma das razões para que o valor percebido pelo mercado livre de empresas de alta tecnologia seja maior que o book value.

Datamining e medidas de rendimento das empresas

A utilização de datamining, ferramenta que aumenta o poder de conhecimento e informação da empresa, afeta positivamente suas medidas de rendimento.

Uma pesquisa da Forrester Research Inc. mostrou que em 1999, 16% das grandes empresas tinham a expectativa de utilizar mas eficientemente as informações de clientes para a redução de custos, e previam que este número de empresas chegaria a 34% em 2001. Este resultado dá evidência à visão de Kotler sobre o futuro de marketing, ou seja, mostra a tendência das empresas conhecerem cada vez mais seus clientes como forma de aumentar sua competitividade e sobreviver num mundo com podutos cada vez menos diferenciados.

O ROI (Return on Investment) capta os ganhos em eficiência. Se a empresa é mais eficiente, então é por que ela consegue realizar mais com menos, ou seja, retorno é maior sem necessariamente aumentar o investimento.

Por exemplo, em Hammond (2000) é descrito um caso onde a empresa Outpost.com obteve um alto retorno sobre investimento com o projeto de datamining. Este projeto, em parceria com a empresa SAS, ajudou a Outpost.com a descobrir seus clientes mais rentáveis e os mais frequentes. Mais do que isso, num esforço de marketing direto que tinha como alvo clientes considerados inativos, 25% destes voltaram a realizar uma compra.

Uma outra medida, o índice q de Tobin, mede a razão do valor de um ativo pelo seu valor de reposição. Tipicamente, investimentos em tecnologia da informação tendem a elevar este índice. Datamining também eleva esta razão, pois um investimento em tecnologia da informação (marginalmente pequeno nas empresas de Internet e nas já informatizadas), pode agregar um valor desproporcionalmente maior. Tanto no estudo de caso como no exemplo da Outpost.com, verificou-se que o processo de datamining aumentou a previsibilidade, possibilitando com isso que os esforços de venda e marketing (investimentos) fossem direcionados para os clientes que tivessem maior probabilidade de resposta. As informações geradas a respeito dos hábitos de compra dos diferentes agrupamentos permitiam adequar o mix de produtos ofertados, assim como forneciam importantes informações sobre o desenvolvimento de novos produtos.

Fica clara a razão destes resultados por meio de uma análise análoga à realizada em Avaliando O Valor Do Modelo (página 69).

Portanto pode-se concluir que:

- Datamining está em termos de estratégia em sintonia com as tendências de Marketing;
- O desenvolvimento tecnológico tende cada vez mais a baratear as soluções de datamining tanto pelo ponto de vista de aplicações como também dos sistemas de infraestrutura de armazenamento de dados (condição primeira para exeistência de datamining), estendendo sua aplicação a um universo maior de empresas;
- O desenvolvimento tecnológico tende a tornar as soluções de datamining ainda mais poderosas;
- A Internet e o comércio eletrônico são palcos naturais para aplicações de datamining;
- A metodologia proposta neste trabalho para modelagem de previsões de venda e de identificação de hábitos de compra dos clientes mostrou-se eficiente e conveniente para aplicações práticas;
- Datamining gera real valor para as empresas, podendo impactar o resultado financeiro e aumentar as métricas de retorno do investimento.

Possibilidades De Aprofundamento Futuro

Dado o objetivo de dar uma visão geral de todas as etapas de realização de um projeto de dataming, este trabalho não se aprofundou em questões muito interessantes e importantes, tais como:

- Comparação de outras alternativas de representar o histórico de compras, e avaliação entre elas;
- Dentro da metodologia proposta, usar como critério de avaliação dos modelos a matriz lucro/perda diretamente durante as fases de treinamento dos mesmos;
- Investigar melhor a possibilidade de se construir modelos apenas com dados cadastrais;
- Experimentar alternativas para o processo de agrupamentos, a fim de produzir agrupamentos mais homogêneos. Procurar também desenvolver modelos que levem em consideração os gastos em moeda dos equipamentos;
- Impacto gerencial da adoção de datamining nas empresas: quais os fatores mais importantes, estudo de casos antes e depois da implementação, dificuldades, etc.

- Determinar os custos envolvidos e valor agregado gerado na adoção de Datamining em e-business para melhorar os modelos de avaliação de impacto econômico;
- Investigar o impacto do tamanho das amostras nos processos de agrupamento e descobrimento de regras de associação.

APÊNDICE A - AMOSTRA DO ARQUIVO DE PEDIDOS

CUST_NUM	CMPY	FY	FQ	FM	FW	ORD_NUM	TIE_NUM	PLP	NET_REV	SYS_QTY
4299	4	1999	4	11	9946	200350825	1	4IN	3475	1
5723	4	1999	2	6	9922	167568799	1	2NT	79.8	0
5723	4	1999	4	11	9945	198132763	3	2NT	16	0
5723	4	1999	4	11	9945	198132763	2	2NT	220	0
5723	4	1999	4	11	9945	198132763	4	2NT	13	0
5723	4	1999	4	12	9950	205650443	3	2NT	55	0
5723	4	2000	2	4	200015	237095963	1	2NT	39.8	0
5723	4	2000	2	5	200020	246754790	2	2NT	28.95	0
5723	4	1999	2	6	9922	167564863	1	3WS	5172	1
5723	4	1999	2	6	9926	172168882	1	3WS	5355	1
5723	4	1999	4	12	9950	207729211	1	3WS	5042	1
5723	4	2000	1	3	200012	232019919	1	3WS	7894	1
5723	4	1999	1	3	9910	153993951	1	4DM	6322	2
5723	4	1999	1	3	9913	156913485	1	4DM	3161	1
5723	4	1999	2	6	9922	167563634	1	4DM	2003	1
5723	4	1999	3	9	9935	185158532	1	4DM	1860	1
5723	4	1999	4	11	9945	198466955	1	4DM	4074	2
5723	4	1999	4	12	9950	205650427	1	4DM	2115	1
5723	4	2000	1	3	200009	226664480	1	4DM	2104	1
5723	4	2000	2	5	200018	244494555	1	4DM	4983	3

CUST_NUM	STATE	SIC	ADSIZ	CREDIT	POPCOD	INDFRM	OFFSIZ	SUBNUM	ULTNUM	SSIC1	LOCEMP	PACTEM	SLSVDT	PACTSL
4299	NY	382398		B	8	2				873106	75	.	10300	.
5639	NY	621101		A	9	2			5133186	509905	4500	8000	.	8500000
5723	VA	179977	A	B	7	2					20	.	1380	.
5767	GA	481304	D	A	1	2		5389390	420372682		150			.
11310	IL	811103	B	B	7	2	E				10	.	1150	.
11475	KY	738973		A	7	2		480366053	7519853	366998	100	.	.	.
11702	PA	353703		A	5	2			8771222	353503	100	145	.	24000
11788	WA	737103	A	B	9	2					12	.	1224	.
11953	TN	641112	A	B	8	2	A				10	.	860	.
12472	FL	873111	B	B	8	2				871110	11	.	1232	.

WHITE	GROWCO	PORDER	PDT	PNB	PSRV	PWS	PORDERP	PDTP	PNBP	PSRVP	PWSP	ABI_NUM	FAXPRES	
1		56775	22.275	4.275	1.2	0.9	.	.	.	.	.	117873067	1	
1		9076500	.	.	.	.	16136000	.	.	.	.	5133186	1	
		10040	3.9	0.92	0.24	0.04	.	.	.	.	.	436842165	1	
1		172500	62.25	13.5	2.1	7.8	.	.	.	.	.	5389390	1	
1	1	7510	3.29	0.34	0.22	0.02	.	.	.	.	.	424751253	1	
	1	48900	19.9	3.9	0.9	0.7	.	.	.	.	.	480366053	1	
		35100	15.3	2.1	0.6	0.7	50895	22.185	3.045	0.87	1.015	8771222	1	
1		31236	8.148	1.944	0.984	1.356	.	.	.	.	.	174498071	0	
1		8150	3.54	0.3	0.24	0.06	.	.	.	.	.	852391564	1	
1		20163	5.907	0.913	0.561	1.023	.	.	.	.	.	894431741	1	

APÊNDICE C - CÓDIGO SAS GERADOR DO HISTÓRICO DE COMPRAS

Abaixo é apresentado os códigos SAS utilizados para a geração do histórico de compras.

O primeiro deles, a partir do arquivo de pedidos, monta uma entrada por cliente que apresenta quantos produtos em cada trimestre foram comprados, criando uma base de dados chistory.sd2.

O segundo código utilizado, baseia-se na base de dados chistory e cria a base de dados com a representação do histórico de compras, a que se denominou custlags.sd2.

O código do mesmo encontra-se abaixo:

```
Title1 'Convert customer histories to lagged history format';
Title2 'with indicators for product mix';
Title3 '-------------------------------------------------------------------';

libname X_data  'D:\Datamining Class\fabio\emdata';

options pageno=1;

/******
proc contents data = X_data.chistory;


proc print data = X_data.chistory (obs=10) ;
        title4 'Printing 1st 10 obs for customer purchase histories';

*******/

data
run;
Title1 'Convert customer histories to lagged history format';
Title2 'with indicators for product mix';
Title3 '-------------------------------------------------------------------';

libname X_data  'D:\Datamining Class\fabio\emdata';

options pageno=1;

/***********************************************
proc contents data = X_data.chistory;


proc print data = X_data.chistory (obs=10) ;
        title4 'Printing 1st 10 obs for customer purchase histories';

run;
***********************************************/

data templags; **X_data.custlags;
        set X_data.chistory;
```

```
************************************************
        Enter data into 9-element vectors (one
        element per product type) for
        product MIX, current purchase (Y/N),
        current purchase amount, lag since last
        purchase, and lag since penultimate
        purchase.  Vectors are MIX_ CURYN_, CURAMT_,
        LAG1_, and LAG2_.  The elements of the
        vectors are encoded (1-9) as

        1 - 2NT 2 - 3WS 3 - 4DM
        4 - 4ES 5 - 4IN 6 - 4LT
        7 - 4PD 8 - 4ST 9 - 4SV

************************************************;

****    set CURYN_ and CURAMT_ vectors from Q8 data ***;

        array CURYN(9)  curyn_1-curyn_9;
        array CURAMT(9) curamt_1-curamt_9;
        array LAST(9)   LAST1-LAST9;
        array PENU(9)   PENU1-PENU9;
****        array P_LAST(9) P_LAST1-P_LAST9;
****        array P_PENU(9) P_PENU1-P_PENU9;
        array OWNED(9)  OWNED1-OWNED9;
        array P_OWNED(9)  P_OWNED1-P_OWNED9;

        curyn_1 = (Q82NT > 0);
        curyn_2 = (Q83WS > 0);
        curyn_3 = (Q84DM > 0);
        curyn_4 = (Q84ES > 0);
        curyn_5 = (Q84IN > 0);
        curyn_6 = (Q84LT > 0);
        curyn_7 = (Q84PD > 0);
        curyn_8 = (Q84ST > 0);
        curyn_9 = (Q84SV > 0);

%macro zero_or(x, v);

        if (&v > 0) then &x = &v;
        else &x = 0;
%mend zero_or;


%macro totprod(total, prod);

        %zero_or(temp1, Q1&prod);
        %zero_or(temp2, Q2&prod);
        %zero_or(temp3, Q3&prod);
        %zero_or(temp4, Q4&prod);
        %zero_or(temp5, Q5&prod);
        %zero_or(temp6, Q6&prod);
        %zero_or(temp7, Q7&prod);

        &total = temp1+temp2+temp3+temp4+temp5+temp6+temp7;

%mend totprod;

        %totprod(OWNED1, 2NT);
        %totprod(OWNED2, 3WS);
        %totprod(OWNED3, 4DM);
        %totprod(OWNED4, 4ES);
        %totprod(OWNED5, 4IN);
        %totprod(OWNED6, 4LT);
        %totprod(OWNED7, 4PD);
        %totprod(OWNED8, 4ST);
```

```
        %totprod(OWNED9, 4SV);

        totowned=OWNED1+OWNED2+OWNED3+OWNED4+OWNED5+OWNED6+OWNED7+OWNED8+OWNED9;
        P_OWNED1=OWNED1/totowned;
        P_OWNED2=OWNED2/totowned;
        P_OWNED3=OWNED3/totowned;
        P_OWNED4=OWNED4/totowned;
        P_OWNED5=OWNED5/totowned;
        P_OWNED6=OWNED6/totowned;
        P_OWNED7=OWNED7/totowned;
        P_OWNED8=OWNED8/totowned;
        P_OWNED9=OWNED9/totowned;

        %zero_or(curamt_1, Q82NT);
        %zero_or(curamt_2, Q83WS);
        %zero_or(curamt_3, Q84DM);
        %zero_or(curamt_4, Q84ES);
        %zero_or(curamt_5, Q84IN);
        %zero_or(curamt_6, Q84LT);
        %zero_or(curamt_7, Q84PD);
        %zero_or(curamt_8, Q84ST);
        %zero_or(curamt_9, Q84SV);


****    set LAST and PENU vectors from Q1-Q7 data ***;

/*        array q2NT(8) Q12NT Q22NT Q32NT Q42NT Q52NT Q62NT Q72NT Q82NT;

        i2NT = 0;       ** first or second appearance of 2NT? **;
        do i = 1 to 7;
                if ( q2NT(8-i) > 0 ) then do;
                        if (i2NT = 0) then do;
                                LAST(1) = i;
                                i2NT = 1;
                        end;
                        else PENU(1) = i;
                end;
        end;        /**/

%macro mlag(prod, prodnum);
          array q&prod(8) Q1&prod Q2&prod Q3&prod Q4&prod Q5&prod Q6&prod Q7&prod
Q8&prod;

        i&prod = 0;       ** first or second appearance of prod? **;
        LAST(&prodnum) = 10;
        PENU(&prodnum) = 10;
        do i = 1 to 7;
                if ( q&prod(8-i) > 0 ) then do;
                        if (i&prod = 0) then do;
                                LAST(&prodnum) = i;
                                i&prod = 1;
                        end;
                        else PENU(&prodnum) = i;
                end;

        end;
%mend  mlag;

%mlag (2NT, 1);
%mlag (3WS, 2);
%mlag (4DM, 3);
%mlag (4ES, 4);
%mlag (4IN, 5);
%mlag (4LT, 6);
%mlag (4PD, 7);
%mlag (4ST, 8);
```

```
%mlag (4SV, 9);


keep CUST_NUM CURAMT_1-CURAMT_9 CURYN_1-CURYN_9 OWNED1-OWNED9 P_OWNED1-P_OWNED9
LAST1-LAST9 PENU1-PENU9;


proc sort data=X_data.abi;
        by cust_num;

data X_data.custlags;
        merge templags X_data.abi;
        by cust_num;

        indgroup = substr(sic,1,2);
        if white  = ' ' then white  = '0';
        if growco = ' ' then growco = '0';
        if offsiz = ' ' then offsiz = '0';

        drop pdtp pnbp porderp psrvp pwsp sic ssic1 ultnum;


proc print data=X_data.custlags (obs=30);

run;
```

APÊNDICE D - AMOSTRA DO ARQUIVO DE HISTORICO DE COMPRAS

CUST_NUM	CURYN_1	CURYN_2	CURYN_3	CURYN_4	CURYN_5	CURYN_6	CURYN_7	CURYN_8	CURYN_9	CURAMT_1	CURAMT_2	CURAMT_3	CURAMT_4	CURAMT_5	CURAMT_6	CURAMT_7	CURAMT_8	CURAMT_9
4299	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5539																		
5723	0	0	1	0	1	1	0	0	0	0	0	3	0	1	1	0	0	0
5767	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11310	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11475	0	1	1	0	0	0	0	0	1	0	5	4	0	0	0	0	0	5
11702	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11788																		
11963	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12472	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
12762	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12846	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13417	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

LAST1	LAST2	LAST3	LAST4	LAST5	LAST6	LAST7	LAST8	LAST9	PENU1	PENU2	PENU3	PENU4	PENU5	PENU6	PENU7	PENU8	PENU9	OWNED1
	10	10	10	10	2	10	10	10	10	10	10	10	10	10	10	10	10	0
10	1	1	10	10	10	2	10	10	2	10	4	5	10	10	10	10	10	0
10	10	2	10	10	10	7	1	10	1	10	10	4	10	10	10	7	10	0
10	10	10	10	10	10	10	10	10	5	10	10	10	10	10	10	10	10	0
10	1	1	10	4	10	1	2	1	2	10	7	10	10	10	10	10	5	0
10	10	10	10	10	10	1	10	10	10	10	10	10	10	10	10	3	10	0
10	10	10	10	1	10	4	10	10	10	10	10	10	2	10	10	10	10	0
10	10	7	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	0
10	10	10	10	5	10	10	10	10	10	10	10	10	10	10	10	10	10	0
10	10	4	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	0
10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	0

OWNED2	OWNED3	OWNED4	OWNED5	OWNED6	OWNED7	OWNED8	OWNED9	P_OWNED1	P_OWNED2	P_OWNED3	P_OWNED4	P_OWNED5	P_OWNED6	P_OWNED7	P_OWNED8	P_OWNED9	STATE	ADSIZ
0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	NY	
																	NY	
4	9	0	0	1	0	0	1	0	0.266666667	0.5	0	0	0.066666667	0	0	0.066666667	YA	A
0	3	0	0	1	17	0	1	0	0	0.136363636	0	0	0.045454546	0.772727273	0	0.045454546	GA	D
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	IL	B
2	59	0	1	0	12	1	21	0	0.020833333	0.614583333	0	0.010416667	0	0.125	0.010416667	0.21875	KY	
0	0	0	0	0	3	0	0	0	0	0	0	0	0	1	0	0	PA	
																	WA	A
0	0	0	2	0	1	0	0	0	0	0	0	0.666666667	0	0.333333333	0	0	TN	A
0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	FL	B
0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0	KY	A
0	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	IL	B
0	0	0	0	0	0	0	0										IL	A

CREDIT	POPCOD	INDFRM	OFFSIZ	SUBNUM	LOCEMP	PACTEM	SLSYDT	PACTSL	WHITE	GROWCO	PORDER	PDT	PNB	PSRV	PWS	ABI_NUM	FAXPRES	INDGROUP
B	8	2	0		75		10300			0	56775	22.275	4.275	1.2	0.9	117873067	1	36
A	9	2	0		4500	8000		8500000		1	0	9075500				5133186	1	62
B	7	2	0		20		1380			0	10040	3.9	0.92	0.24	0.04	436842165	1	17
A	1	2	0	5389390	150					1	172500	62.25	13.5	2.1	7.8	5389390	1	48
B	7	2	E		10		1150			1	7510	3.29	0.34	0.22	0.02	424751253	1	81
A	7	2	0	480366053	100					1	48900	19.9	3.9	0.9	0.7	480366053	1	73
A	5	2	0		100	145		24000		0	35100	16.3	2.1	0.6	0.7	87771222	1	35
B	9	2	0		12		1224			1	31236	8.148	1.944	0.964	1.356	174498071	0	73
B	8	2	A		10		660			1	8150	3.54	0.3	0.24	0.06	852391564	1	64
B	8	2		0	11		1232			1	20163	5.907	0.913	0.561	1.023	894431741	1	87
C	8	2		0	3		4500			1	3627	1.503	0.207	0.108	0.006	136492568	0	62
A	9	2	F		100		11500			1	88400	45.4	4.4	1.5	0.1	470693435	1	81
C	5	2	0		11		1353			1	28633	7.463	1.782	0.902	1.243	491645834	1	73

APÊNDICE E - RESULTADO DO NÓ EMDATA.CUSTLAGS

EMDATA.CUSTLAGS

Input Data Settings:

Source Data: EMDATA.CUSTLAGS (139,333 rows, 76 columns)
Output: EMDATA.VIEW_3UC
Description: EMDATA.CUSTLAGS
Role: RAW
Metadata Sample: EMPROJ.SMP_VIZW (2,000 rows)

INTERVAL VARIABLES

Name	Min	Max	Mean	Std Dev.	Missing %	Skewness	Kurtosis
CUST_NUM	17293.00	6806139.00	4578285.89	1647926.54	0.0000	-0.8450	-0.19
CURAMT_1	0.00	0.00	0.00	0.00	0.0300	.	.
CURAMT_2	0.00	7.00	0.01	0.23	0.0300	24.7296	668.11
CURAMT_3	0.00	34.00	0.29	1.59	0.0300	10.7747	159.84
CURAMT_4	0.00	6.00	0.00	0.14	0.0300	39.6009	1656.48
CURAMT_5	0.00	22.00	0.08	0.68	0.0300	21.3324	609.45
CURAMT_6	0.00	8.00	0.03	0.30	0.0300	16.7388	346.56
CURAMT_7	0.00	95.00	0.18	2.59	0.0300	28.0051	951.23
CURAMT_8	0.00	1.00	0.00	0.02	0.0300	44.0454	1940.00
CURAMT_9	0.00	9.00	0.03	0.28	0.0300	20.5203	579.06
OWNED1	0.00	0.00	0.00	0.00	0.0300	.	.
OWNED2	0.00	17.00	0.06	0.71	0.0300	18.7084	404.85
OWNED3	0.00	244.00	2.83	10.01	0.0300	14.5797	284.39
OWNED4	0.00	8.00	0.01	0.22	0.0300	28.2490	967.13
OWNED5	0.00	65.00	0.63	2.80	0.0300	15.3324	298.54
OWNED6	0.00	66.00	0.35	3.29	0.0300	16.3739	291.88
OWNED7	0.00	323.00	0.96	9.46	0.0300	24.1627	736.99
OWNED8	0.00	2.00	0.00	0.06	0.0300	23.3134	603.12
OWNED9	0.00	16.00	0.25	0.91	0.0300	9.3414	121.08
P_OWNED1	0.00	0.00	0.00	0.00	0.0715	.	.
P_OWNED2	0.00	1.00	0.01	0.07	0.0715	12.4263	164.56
P_OWNED3	0.00	1.00	0.63	0.43	0.0715	-0.5827	-1.45
P_OWNED4	0.00	1.00	0.00	0.04	0.0715	25.3825	692.80
P_OWNED5	0.00	1.00	0.20	0.37	0.0715	1.5358	0.62
P_OWNED6	0.00	1.00	0.04	0.18	0.0715	4.6093	20.55
P_OWNED7	0.00	1.00	0.05	0.20	0.0715	3.8330	13.64
P_OWNED8	0.00	0.33	0.00	0.01	0.0715	40.2622	1679.31
P_OWNED9	0.00	1.00	0.06	0.19	0.0715	3.9923	15.70
LOCEMP	1.00	2600.00	27.03	81.91	0.0115	18.4409	519.43
PACTEM	10.00	4120.00	305.11	759.44	0.9860	5.0372	26.10
SLSVDT	11.00	478400.00	4856.99	17184.45	0.0850	14.9564	344.72
PACTSL	191.00	498831.00	95299.39	149338.17	0.9885	1.9988	2.78
PORDER	77.00	1149200.00	15144.05	37569.00	0.0250	16.7082	445.77
PDT	0.04	150.72	5.45	9.44	0.0280	5.1638	46.22
PNB	0.00	42.00	0.98	2.18	0.0280	8.6494	116.90
PSRV	0.00	6.40	0.31	0.43	0.0280	4.1022	32.08
PWS	0.00	21.44	0.24	0.79	0.0280	12.9101	289.25

CLASS VARIABLES

Name	Values	Missing %	Order	Depends On
CURYN_1	1	0.0300	Ascending	
CURYN_2	2	0.0300	Ascending	
CURYN_3	2	0.0300	Ascending	
CURYN_4	2	0.0300	Ascending	
CURYN_5	2	0.0300	Ascending	
CURYN_6	2	0.0300	Ascending	
CURYN_7	2	0.0300	Ascending	
CURYN_8	2	0.0300	Ascending	
CURYN_9	2	0.0300	Descending	
LAST1	1	0.0300	Ascending	
LAST2	8	0.0300	Ascending	
LAST3	8	0.0300	Ascending	

LAST4	6	0.0300	Ascending
LAST5	8	0.0300	Ascending
LAST6	8	0.0300	Ascending
LAST7	8	0.0300	Ascending
LAST8	6	0.0300	Ascending
LAST9	8	0.0300	Ascending
PENU1	1	0.0300	Ascending
PENU2	7	0.0300	Ascending
PENU3	7	0.0300	Ascending
PENU4	3	0.0300	Ascending
PENU5	7	0.0300	Ascending
PENU6	7	0.0300	Ascending
PENU7	7	0.0300	Ascending
PENU8	1	0.0300	Ascending
PENU9	7	0.0300	Ascending
STATE	51	0.0000	Ascending
ADSIZ	4	0.1340	Ascending
CREDIT	6	0.0000	Ascending
POPCOD	6	0.0000	Ascending
INDFRM	2	0.0000	Ascending
OFFSIZ	7	0.0000	Ascending
SUBNUM	55	0.9715	Ascending
WHITE	2	0.0000	Ascending
GROWCO	2	0.0000	Ascending
ABI_NUM	2000	0.0000	Ascending
FAXPRES	2	0.0000	Ascending
INDGROUP	69	0.0000	Ascending

```
data &_TRA;
set &_TRAIN;
if CURYN_9 = . then delete;
run;

data &_TES;
set &_TEST;
if CURYN_9 = . then delete;
run;

data &_VAL;
set &_VALID;
if CURYN_9 = . then delete;
run;
```

APÊNDICE G - REDEFINIÇÃO DAS VARIÁVEIS

```
/*****************************************
 *
 * TOOL: Variable Selection
 * TYPE: EXPLORE
 * NODE : Variable Selection [T32V_3N4]
 *
 *****************************************/
 length _warn_ $ 4;
 label _warn_ = "Warnings";
 length _norm1 $ 16;
 length I_CURYN_ $ 1;
 _norm1 = upcase(trim(left(put(ADSIZ, $1.))));
 select(_norm1);
 when('') G_ADSIZ = 0;
 when('A') G_ADSIZ = 1;
 when('B') G_ADSIZ = 2;
 when('C') G_ADSIZ = 2;
 when('D') G_ADSIZ = 2;
 otherwise _warn_ = 'U';
 end;
 _norm1 = upcase(trim(left(put(CREDIT, $1.))));
 select(_norm1);
 when('A') G_CREDIT = 0;
 when('B') G_CREDIT = 1;
 when('C') G_CREDIT = 2;
 when('I') G_CREDIT = 2;
 when('P') G_CREDIT = 3;
 when('U') G_CREDIT = 2;
 otherwise _warn_ = 'U';
 end;
 _norm1     =     upcase(trim(left(put(INDGROUP,
$6.))));
 select(_norm1);
 when('07') G_INDGRO = 8;
 when('08') G_INDGRO = 8;
 when('13') G_INDGRO = 5;
 when('14') G_INDGRO = 0;
 when('15') G_INDGRO = 5;
 when('16') G_INDGRO = 7;
 when('17') G_INDGRO = 7;
 when('20') G_INDGRO = 3;
 when('22') G_INDGRO = 0;
 when('23') G_INDGRO = 5;
 when('24') G_INDGRO = 7;
 when('25') G_INDGRO = 8;
 when('26') G_INDGRO = 2;
 when('27') G_INDGRO = 4;
 when('28') G_INDGRO = 2;
 when('29') G_INDGRO = 4;
 when('30') G_INDGRO = 6;
 when('31') G_INDGRO = 0;
 when('32') G_INDGRO = 1;
 when('33') G_INDGRO = 2;
 when('34') G_INDGRO = 6;
 when('35') G_INDGRO = 6;
 when('36') G_INDGRO = 2;
 when('37') G_INDGRO = 5;
 when('38') G_INDGRO = 2;
 when('39') G_INDGRO = 7;
 when('40') G_INDGRO = 8;
 when('41') G_INDGRO = 5;
 when('42') G_INDGRO = 6;
 when('44') G_INDGRO = 8;
 when('45') G_INDGRO = 4;
 when('47') G_INDGRO = 5;
 when('48') G_INDGRO = 1;
 when('49') G_INDGRO = 3;
 when('50') G_INDGRO = 5;
 when('51') G_INDGRO = 4;
 when('52') G_INDGRO = 6;
 when('53') G_INDGRO = 4;
 when('54') G_INDGRO = 6;
 when('55') G_INDGRO = 7;
 when('56') G_INDGRO = 7;
 when('57') G_INDGRO = 2;
 when('58') G_INDGRO = 7;
 when('59') G_INDGRO = 6;
 when('60') G_INDGRO = 1;
 when('61') G_INDGRO = 3;
 when('62') G_INDGRO = 3;
 when('63') G_INDGRO = 5;
 when('64') G_INDGRO = 3;
 when('65') G_INDGRO = 5;
 when('67') G_INDGRO = 1;
 when('70') G_INDGRO = 2;
 when('72') G_INDGRO = 5;
 when('73') G_INDGRO = 2;
 when('75') G_INDGRO = 7;
 when('76') G_INDGRO = 6;
 when('78') G_INDGRO = 5;
 when('79') G_INDGRO = 6;
 when('80') G_INDGRO = 7;
 when('81') G_INDGRO = 5;
 when('82') G_INDGRO = 3;
 when('83') G_INDGRO = 4;
 when('84') G_INDGRO = 4;
 when('86') G_INDGRO = 5;
 when('87') G_INDGRO = 4;
 when('89') G_INDGRO = 6;
 when('91') G_INDGRO = 7;
 when('93') G_INDGRO = 0;
 when('94') G_INDGRO = 0;
 when('95') G_INDGRO = 0;
 when('96') G_INDGRO = 6;
 when('99') G_INDGRO = 3;
 otherwise _warn_ = 'U';
 end;
 _norm1      =      upcase(trim(left(put(LAST2,
BEST12.))));
 select(_norm1);
 when('.') G_LAST2 = 0;
 when('1') G_LAST2 = 1;
 when('2') G_LAST2 = 2;
 when('3') G_LAST2 = 2;
 when('4') G_LAST2 = 2;
 when('5') G_LAST2 = 2;
 when('6') G_LAST2 = 2;
 when('7') G_LAST2 = 2;
 otherwise _warn_ = 'U';
 end;
 _norm1      =      upcase(trim(left(put(LAST3,
BEST12.))));
 select(_norm1);
 when('.') G_LAST3 = 0;
 when('1') G_LAST3 = 1;
 when('2') G_LAST3 = 2;
 when('3') G_LAST3 = 2;
 when('4') G_LAST3 = 2;
 when('5') G_LAST3 = 3;
 when('6') G_LAST3 = 4;
 when('7') G_LAST3 = 4;
 otherwise _warn_ = 'U';
 end;
 _norm1      =      upcase(trim(left(put(LAST5,
BEST12.))));
 select(_norm1);
 when('.') G_LAST5 = 0;
 when('1') G_LAST5 = 1;
 when('2') G_LAST5 = 2;
 when('3') G_LAST5 = 2;
 when('4') G_LAST5 = 2;
 when('5') G_LAST5 = 3;
 when('6') G_LAST5 = 4;
 when('7') G_LAST5 = 4;
 otherwise _warn_ = 'U';
 end;
 _norm1      =      upcase(trim(left(put(LAST6,
BEST12.))));
 select(_norm1);
 when('.') G_LAST6 = 0;
 when('1') G_LAST6 = 1;
 when('2') G_LAST6 = 1;
 when('3') G_LAST6 = 1;
 when('4') G_LAST6 = 2;
 when('5') G_LAST6 = 3;
 when('6') G_LAST6 = 4;
 when('7') G_LAST6 = 4;
 otherwise _warn_ = 'U';
 end;
```

```
 _norm1      =      upcase(trim(left(put(LAST7,
BEST12.))));
 select(_norm1);
 when('.') G_LAST7 = 0;
 when('1') G_LAST7 = 1;
 when('2') G_LAST7 = 2;
 when('3') G_LAST7 = 2;
 when('4') G_LAST7 = 2;
 when('5') G_LAST7 = 2;
 when('6') G_LAST7 = 2;
 when('7') G_LAST7 = 3;
 otherwise _warn_ = 'U';
 end;
 _norm1      =      upcase(trim(left(put(LAST9,
BEST12.))));
 select(_norm1);
 when('.') G_LAST9 = 0;
 when('1') G_LAST9 = 1;
 when('2') G_LAST9 = 2;
 when('3') G_LAST9 = 2;
 when('4') G_LAST9 = 2;
 when('5') G_LAST9 = 2;
 when('6') G_LAST9 = 2;
 when('7') G_LAST9 = 2;
 otherwise _warn_ = 'U';
 end;
 _norm1      =      upcase(trim(left(put(OWNED3,
BEST12.))));
 select(_norm1);
 when('.') G_OWNED3 = 0;
 when('12') G_OWNED3 = 1;
 when('15') G_OWNED3 = 1;
 when('16') G_OWNED3 = 2;
 when('17') G_OWNED3 = 3;
 when('18') G_OWNED3 = 3;
 when('20') G_OWNED3 = 3;
 when('21') G_OWNED3 = 3;
 when('22') G_OWNED3 = 3;
 when('24') G_OWNED3 = 3;
 when('27') G_OWNED3 = 3;
 when('29') G_OWNED3 = 3;
 when('30') G_OWNED3 = 3;
 when('31') G_OWNED3 = 3;
 when('32') G_OWNED3 = 3;
 when('36') G_OWNED3 = 3;
 when('38') G_OWNED3 = 3;
 when('40') G_OWNED3 = 3;
 when('41') G_OWNED3 = 3;
 when('43') G_OWNED3 = 3;
 when('44') G_OWNED3 = 4;
 when('45') G_OWNED3 = 5;
 when('49') G_OWNED3 = 5;
 when('51') G_OWNED3 = 5;
 when('52') G_OWNED3 = 5;
 when('54') G_OWNED3 = 5;
 when('56') G_OWNED3 = 5;
 when('58') G_OWNED3 = 5;
 when('59') G_OWNED3 = 5;
 when('61') G_OWNED3 = 5;
 when('69') G_OWNED3 = 5;
 when('70') G_OWNED3 = 5;
 when('71') G_OWNED3 = 5;
 when('75') G_OWNED3 = 5;
 when('77') G_OWNED3 = 5;
 when('84') G_OWNED3 = 5;
 when('85') G_OWNED3 = 5;
 when('91') G_OWNED3 = 5;
 when('100') G_OWNED3 = 5;
 when('102') G_OWNED3 = 5;
 when('104') G_OWNED3 = 5;
 when('105') G_OWNED3 = 5;
 when('112') G_OWNED3 = 5;
 when('114') G_OWNED3 = 5;
 when('121') G_OWNED3 = 5;
 when('134') G_OWNED3 = 5;
 when('136') G_OWNED3 = 5;
 when('138') G_OWNED3 = 6;
 when('141') G_OWNED3 = 7;
 when('145') G_OWNED3 = 7;
 when('171') G_OWNED3 = 7;
 when('184') G_OWNED3 = 7;
 when('188') G_OWNED3 = 7;
 when('190') G_OWNED3 = 7;
 when('210') G_OWNED3 = 7;
 when('252') G_OWNED3 = 7;
 when('253') G_OWNED3 = 7;
 when('291') G_OWNED3 = 7;
 when('303') G_OWNED3 = 7;
 otherwise _warn_ = 'U';
 end;
 _norm1      =      upcase(trim(left(put(OWNED7,
BEST12.))));
 select(_norm1);
 when('.') G_OWNED7 = 0;
 when('23') G_OWNED7 = 1;
 when('50') G_OWNED7 = 1;
 when('52') G_OWNED7 = 1;
 when('67') G_OWNED7 = 1;
 when('82') G_OWNED7 = 1;
 when('108') G_OWNED7 = 1;
 when('114') G_OWNED7 = 1;
 when('128') G_OWNED7 = 1;
 when('144') G_OWNED7 = 1;
 when('148') G_OWNED7 = 1;
 when('154') G_OWNED7 = 1;
 when('158') G_OWNED7 = 1;
 when('168') G_OWNED7 = 1;
 when('172') G_OWNED7 = 1;
 when('188') G_OWNED7 = 1;
 when('191') G_OWNED7 = 1;
 when('192') G_OWNED7 = 1;
 when('255') G_OWNED7 = 1;
 when('355') G_OWNED7 = 1;
 otherwise _warn_ = 'U';
 end;
 _norm1      =      upcase(trim(left(put(PENU2,
BEST12.))));
 select(_norm1);
 when('.') G_PENU2 = 0;
 when('2') G_PENU2 = 1;
 when('3') G_PENU2 = 1;
 when('4') G_PENU2 = 1;
 when('5') G_PENU2 = 1;
 when('6') G_PENU2 = 2;
 when('7') G_PENU2 = 3;
 otherwise _warn_ = 'U';
 end;
 _norm1      =      upcase(trim(left(put(PENU3,
BEST12.))));
 select(_norm1);
 when('.') G_PENU3 = 0;
 when('2') G_PENU3 = 1;
 when('3') G_PENU3 = 1;
 when('4') G_PENU3 = 1;
 when('5') G_PENU3 = 1;
 when('6') G_PENU3 = 1;
 when('7') G_PENU3 = 2;
 otherwise _warn_ = 'U';
 end;
 _norm1      =      upcase(trim(left(put(PENU7,
BEST12.))));
 select(_norm1);
 when('.') G_PENU7 = 0;
 when('2') G_PENU7 = 1;
 when('3') G_PENU7 = 1;
 when('4') G_PENU7 = 1;
 when('5') G_PENU7 = 2;
 when('6') G_PENU7 = 2;
 when('7') G_PENU7 = 2;
 otherwise _warn_ = 'U';
 end;
 _norm1      =      upcase(trim(left(put(PENU9,
BEST12.))));
 select(_norm1);
 when('.') G_PENU9 = 0;
 when('2') G_PENU9 = 1;
 when('3') G_PENU9 = 1;
 when('4') G_PENU9 = 1;
 when('5') G_PENU9 = 1;
 when('6') G_PENU9 = 1;
 when('7') G_PENU9 = 1;
 otherwise _warn_ = 'U';
 end;
 _norm1 = upcase(trim(left(put(POPCOD, $1.))));
 select(_norm1);
 when('1') G_POPCOD = 2;
 when('5') G_POPCOD = 2;
 when('6') G_POPCOD = 0;
 when('7') G_POPCOD = 1;
 when('8') G_POPCOD = 1;
 when('9') G_POPCOD = 0;
 otherwise _warn_ = 'U';
```

```
end;
_norm1 = upcase(trim(left(put(STATE, $2.))));
select(_norm1);
when('AK') G_STATE = 7;
when('AL') G_STATE = 4;
when('AR') G_STATE = 7;
when('AZ') G_STATE = 4;
when('CA') G_STATE = 3;
when('CO') G_STATE = 3;
when('CT') G_STATE = 0;
when('DC') G_STATE = 1;
when('DE') G_STATE = 1;
when('FL') G_STATE = 6;
when('GA') G_STATE = 4;
when('HI') G_STATE = 6;
when('IA') G_STATE = 4;
when('ID') G_STATE = 8;
when('IL') G_STATE = 5;
when('IN') G_STATE = 1;
when('KS') G_STATE = 5;
when('KY') G_STATE = 6;
when('LA') G_STATE = 5;
when('MA') G_STATE = 0;
when('MD') G_STATE = 1;
when('ME') G_STATE = 0;
when('MI') G_STATE = 6;
when('MN') G_STATE = 5;
when('MO') G_STATE = 2;
when('MS') G_STATE = 7;
when('MT') G_STATE = 6;
when('NC') G_STATE = 3;
when('NE') G_STATE = 0;
when('NH') G_STATE = 7;
when('NJ') G_STATE = 2;
when('NM') G_STATE = 7;
when('NV') G_STATE = 6;
when('NY') G_STATE = 4;
when('OH') G_STATE = 4;
when('OK') G_STATE = 4;
when('OR') G_STATE = 6;
when('PA') G_STATE = 6;
when('RI') G_STATE = 0;
when('SC') G_STATE = 2;
when('SD') G_STATE = 7;
when('TN') G_STATE = 5;
when('TX') G_STATE = 4;
when('UT') G_STATE = 7;
when('VA') G_STATE = 0;
when('VT') G_STATE = 2;
when('WA') G_STATE = 5;
when('WI') G_STATE = 2;
when('WV') G_STATE = 7;
when('WY') G_STATE = 8;
otherwise _warn_ = 'U';
end;
_norm1 = upcase(trim(left(put(SUBNUM, $9.))));
select(_norm1);
when('') G_SUBNUM = 1;
when('001186881') G_SUBNUM = 0;
when('001197359') G_SUBNUM = 0;
when('001244631') G_SUBNUM = 0;
when('001274125') G_SUBNUM = 2;
when('001700459') G_SUBNUM = 0;
when('001740018') G_SUBNUM = 2;
when('004268314') G_SUBNUM = 0;
when('005004015') G_SUBNUM = 0;
when('005981147') G_SUBNUM = 2;
when('006094387') G_SUBNUM = 0;
when('006730030') G_SUBNUM = 0;
when('008268765') G_SUBNUM = 2;
when('008990905') G_SUBNUM = 0;
when('009056581') G_SUBNUM = 0;
when('009211293') G_SUBNUM = 2;
when('009376898') G_SUBNUM = 0;
when('150626182') G_SUBNUM = 0;
when('170384556') G_SUBNUM = 2;
when('200118719') G_SUBNUM = 0;
when('230051872') G_SUBNUM = 0;
when('316893593') G_SUBNUM = 0;
when('401325899') G_SUBNUM = 2;
when('402881080') G_SUBNUM = 2;
when('409213477') G_SUBNUM = 2;
when('413606534') G_SUBNUM = 0;
when('413652884') G_SUBNUM = 0;
when('415520295') G_SUBNUM = 2;
when('418937041') G_SUBNUM = 0;
when('436129233') G_SUBNUM = 0;
when('441411691') G_SUBNUM = 2;
when('441425196') G_SUBNUM = 2;
when('441426855') G_SUBNUM = 0;
when('441437126') G_SUBNUM = 2;
when('441446358') G_SUBNUM = 2;
when('443630645') G_SUBNUM = 0;
when('448236406') G_SUBNUM = 0;
when('450546908') G_SUBNUM = 2;
when('452339419') G_SUBNUM = 0;
when('453641532') G_SUBNUM = 0;
when('464860493') G_SUBNUM = 0;
when('465944437') G_SUBNUM = 0;
when('466067204') G_SUBNUM = 0;
when('470264193') G_SUBNUM = 0;
when('473708790') G_SUBNUM = 0;
when('479545089') G_SUBNUM = 0;
when('480366053') G_SUBNUM = 0;
when('488823360') G_SUBNUM = 0;
when('489160085') G_SUBNUM = 0;
when('490862919') G_SUBNUM = 0;
when('502242167') G_SUBNUM = 0;
when('502392913') G_SUBNUM = 2;
when('502643307') G_SUBNUM = 2;
when('620874917') G_SUBNUM = 0;
when('802881946') G_SUBNUM = 0;
when('808023998') G_SUBNUM = 0;
when('810144642') G_SUBNUM = 0;
when('813678323') G_SUBNUM = 0;
when('846821007') G_SUBNUM = 0;
when('854284486') G_SUBNUM = 0;
when('861517894') G_SUBNUM = 2;
when('878270487') G_SUBNUM = 0;
when('887627560') G_SUBNUM = 0;
when('890349129') G_SUBNUM = 2;
when('898277868') G_SUBNUM = 0;
when('900687427') G_SUBNUM = 0;
when('903407997') G_SUBNUM = 0;
when('905270096') G_SUBNUM = 2;
when('935816454') G_SUBNUM = 2;
when('936691344') G_SUBNUM = 0;
when('947230058') G_SUBNUM = 0;
when('960132934') G_SUBNUM = 0;
when('973958267') G_SUBNUM = 2;
when('979434404') G_SUBNUM = 2;
otherwise _warn_ = 'U';
end;
label G_ADSIZ = "Group: ADSIZ";
label G_CREDIT = "Group: CREDIT";
label G_INDGRO = "Group: INDGROUP";
label G_LAST2 = "Group: LAST2";
label G_LAST3 = "Group: LAST3";
label G_LAST5 = "Group: LAST5";
label G_LAST6 = "Group: LAST6";
label G_LAST7 = "Group: LAST7";
label G_LAST9 = "Group: LAST9";
label G_OWNED3 = "Group: OWNED3";
label G_OWNED7 = "Group: OWNED7";
label G_PENU2 = "Group: PENU2";
label G_PENU3 = "Group: PENU3";
label G_PENU7 = "Group: PENU7";
label G_PENU9 = "Group: PENU9";
label G_POPCOD = "Group: POPCOD";
label G_STATE = "Group: STATE";
label G_SUBNUM = "Group: SUBNUM";
_PVAL = 0.56212323901463;
select(G_LAST9);
when(0) _PVAL = _PVAL + 0.15448737592614;
when(1) _PVAL = _PVAL + 0.00708767785892;
when(2) _PVAL = _PVAL + 0;
otherwise;
end;
select(G_PENU3);
when(0) _PVAL = _PVAL + 0.16423846468521;
when(1) _PVAL = _PVAL + -0.00395114168637;
when(2) _PVAL = _PVAL + 0;
otherwise;
end;
select(G_PENU9);
when(0) _PVAL = _PVAL + 0.093154104441;
when(1) _PVAL = _PVAL + 0;
otherwise;
end;
select(G_LAST3);
when(0) _PVAL = _PVAL + 0.0280122345457;
when(1) _PVAL = _PVAL + -0.06739550621126;
```

```
when(2) _PVAL = _PVAL + 0.03982028326845;
when(3) _PVAL = _PVAL + 0.13154797106234;
when(4) _PVAL = _PVAL + 0;
otherwise;
end;
select(G_LAST7);
when(0) _PVAL = _PVAL + 0.19422392455504;
when(1) _PVAL = _PVAL + 0.01132894122099;
when(2) _PVAL = _PVAL + 0.0741093923674;
when(3) _PVAL = _PVAL + 0;
otherwise;
end;
if PSRV = . then _PVAL = _PVAL +
(0.5474283714419 * -0.05742269541369);
else _PVAL = _PVAL + (PSRV *
0.05742269541369);
select(G_INDGRO);
when(0) _PVAL = _PVAL + -0.80849762026103;
when(1) _PVAL = _PVAL + -0.38137002265222;
when(2) _PVAL = _PVAL + -0.31632028381019;
when(3) _PVAL = _PVAL + -0.24038454437569;
when(4) _PVAL = _PVAL + -0.25231194246192;
when(5) _PVAL = _PVAL + -0.25969237824948;
when(6) _PVAL = _PVAL + -0.1323084884387;
when(7) _PVAL = _PVAL + -0.17719817484178;
when(8) _PVAL = _PVAL + 0;
otherwise;
end;
select(G_LAST5);
when(0) _PVAL = _PVAL + -0.04043103748827;
when(1) _PVAL = _PVAL + -0.16619175054419;
when(2) _PVAL = _PVAL + -0.08918047105202;
when(3) _PVAL = _PVAL + -0.00253660262678;
when(4) _PVAL = _PVAL + 0;
otherwise;
end;
select(G_LAST6);
when(0) _PVAL = _PVAL + 0.1058006039096;
when(1) _PVAL = _PVAL + 0.07244686302087;
when(2) _PVAL = _PVAL + 0.06266852343089;
when(3) _PVAL = _PVAL + 0.14403500927347;
when(4) _PVAL = _PVAL + 0;
otherwise;
end;
select(G_PENU7);
when(0) _PVAL = _PVAL + -0.06022874728898;
when(1) _PVAL = _PVAL + -0.09253172560579;
when(2) _PVAL = _PVAL + 0;
otherwise;
end;
select(G_CREDIT);
when(0) _PVAL = _PVAL + -0.08936480735782;
when(1) _PVAL = _PVAL + -0.07281555539539;
when(2) _PVAL = _PVAL + -0.04232728857798;
when(3) _PVAL = _PVAL + 0;
otherwise;
end;
if PNB = . then _PVAL = _PVAL +
(1.94482967802146 * 0.00643089536848);
else _PVAL = _PVAL + (PNB * 0.00643089536848);
select(G_SUBNUM);
when(0) _PVAL = _PVAL + -0.53854331611914;
when(1) _PVAL = _PVAL + -0.36181947389694;
when(2) _PVAL = _PVAL + 0;
otherwise;
end;
_norm1 = upcase(trim(left(put(FAXPRES,
BEST12.))));
select(_norm1);
when('0') _PVAL = _PVAL + 0.05740051380307;
when('1') _PVAL = _PVAL + 0;
otherwise _warn_ = 'U';
end;
select(G_LAST2);
when(0) _PVAL = _PVAL + 0.05177279967804;
when(1) _PVAL = _PVAL + 0.1153626712614;
when(2) _PVAL = _PVAL + 0;
otherwise;
end;
select(G_OWNED3);
when(0) _PVAL = _PVAL + -0.09864459295558;
when(1) _PVAL = _PVAL + -0.45735811560594;
when(2) _PVAL = _PVAL + 0.72933014942199;
when(3) _PVAL = _PVAL + -0.20574432702156;
when(4) _PVAL = _PVAL + 0.51630317863175;
when(5) _PVAL = _PVAL + -0.14038163699148;
when(6) _PVAL = _PVAL + 0.84840445349298;
when(7) _PVAL = _PVAL + 0;
otherwise;
end;
select(G_STATE);
when(0) _PVAL = _PVAL + -0.1794182852008;
when(1) _PVAL = _PVAL + -0.14364951318855;
when(2) _PVAL = _PVAL + -0.14870017492936;
when(3) _PVAL = _PVAL + -0.0864966874609;
when(4) _PVAL = _PVAL + -0.09307356812613;
when(5) _PVAL = _PVAL + -0.06666720591524;
when(6) _PVAL = _PVAL + -0.03760662465782;
when(7) _PVAL = _PVAL + -0.04698537126562;
when(8) _PVAL = _PVAL + 0;
otherwise;
end;
_norm1 = upcase(trim(left(put(GROWCO, $1.))));
select(_norm1);
when('0') _PVAL = _PVAL + 0.04737895713057;
when('1') _PVAL = _PVAL + 0;
otherwise _warn_ = 'U';
end;
select(G_PENU2);
when(0) _PVAL = _PVAL + 0.13329045906046;
when(1) _PVAL = _PVAL + 0.11283674777319;
when(2) _PVAL = _PVAL + 0.36315796272438;
when(3) _PVAL = _PVAL + 0;
otherwise;
end;
select(G_ADSIZ);
when(0) _PVAL = _PVAL + -0.03360606412527;
when(1) _PVAL = _PVAL + -0.02682156886875;
when(2) _PVAL = _PVAL + 0;
otherwise;
end;
select(G_POPCOD);
when(0) _PVAL = _PVAL + -0.05125447331786;
when(1) _PVAL = _PVAL + -0.03355733014847;
when(2) _PVAL = _PVAL + 0;
otherwise;
end;
select(G_OWNED7);
when(0) _PVAL = _PVAL + 0.1841612013712;
when(1) _PVAL = _PVAL + 0;
otherwise;
end;
/* binary target */
if _warn_ ne 'U' then do;
_X = -(-3.41125213630808 + 6.48283453512645 *
_PVAL);
if (_X < -23) then _PVAL = 1;
else if (_X > 23) then _PVAL = 0;
else _PVAL = 1/(1+exp(_X));
if _PVAL < 0 then _PVAL = 1;
else if _PVAL > 1 then _PVAL = 0;
else _PVAL = 1 - _PVAL;
P_EVENT = _PVAL;
P_NEVENT = 1-P_EVENT;
if P_EVENT ge 0.5 then I_CURYN_ = '1';
else I_CURYN_ = '0';
end;
else do;
P_EVENT = 0.5;
P_NEVENT = 0.5;
I_CURYN_ = '1';
end;
label P_EVENT = "Predicted: EVENT for
CURYN_9";
label P_NEVENT = "Predicted: NO EVENT for
CURYN_9";
label I_CURYN_ = "Into: CURYN_9";
DROP _PVAL _X;
DROP _NORM1;
```

APÊNDICE H - SELEÇÃO FINAL ATRAVÉS DO CRITÉRIO R^2

11:46 Thursday, December 16, 1999 80

R-Squares for Target variable: CURYN_9

Effect	DF	R2
Class: LAST9	7	0.1882
Group: LAST9	2	0.1850
Class: PENU3	6	0.1511
Group: PENU3	2	0.1508
Class: PENU9	6	0.1423
Group: PENU9	1	0.1421
Class: LAST3	7	0.1388
Group: LAST3	4	0.1381
Class: LAST7	7	0.1144
Group: LAST7	3	0.1127
AOV16: PSRV	11	0.1103
Var: PSRV	1	0.0986
Class: INDGROUP	71	0.0844
Group: INDGROUP	8	0.0829
AOV16: PDT	12	0.0826
Class: LAST5	7	0.0796
Group: LAST5	4	0.0787
Var: PDT	1	0.0768
Class: PENU5	6	0.0764
Class: LAST6	7	0.0762
Group: PENU5	2	0.0754
Class: PENU7	6	0.0753
Group: LAST6	4	0.0752
Group: PENU7	2	0.0738
Class: PENU6	6	0.0600
Group: PENU6	1	0.0593
AOV16: PNB	12	0.0568
Class: CREDIT	5	0.0501
Group: CREDIT	3	0.0499
Var: PNB	1	0.0448
Class: LAST4	7	0.0387
Group: LAST4	1	0.0383
AOV16: PWS	13	0.0370
Var: PWS	1	0.0365
Class: SUBNUM	73	0.0344
Group: SUBNUM	2	0.0344
Class: FAXPRES	1	0.0335
Class: LAST2	7	0.0331
Group: LAST2	2	0.0330
Class: OWNED3	58	0.0320
Group: OWNED3	7	0.0320
Class: OWNED1	1	0.0299
Class: STATE	49	0.0225
Group: STATE	8	0.0221
Class: LAST8	7	0.0178
Group: LAST8	1	0.0175
Class: GROWCO	1	0.0159
Class: PENU2	6	0.0149
Class: INDFRM	1	0.0148
Group: PENU2	3	0.0147
Var: SLSVDT	1	0.0143
Class: PENU4	6	0.0114
Group: PENU4	1	0.0114
AOV16: SLSVDT	8	0.0107
Class: ADSIZ	4	0.0102

Group: ADSIZ	2	0.0102	
Class: POPCOD	5	0.0095	
Class: OFFSIZ	6	0.0095	
Group: POPCOD	2	0.0095	
Group: OFFSIZ	2	0.0094	
Class: OWNED7	19	0.0087	
Group: OWNED7	1	0.0087	
Class: WHITE	1	0.0087	
Class: OWNED9	11	0.0054	
Group: OWNED9	1	0.0054	
Class: PENU8	4	0.0050	R2 < MINR2
Group: PENU8	1	0.0050	R2 < MINR2
Class: OWNED5	7	0.0032	R2 < MINR2
Group: OWNED5	1	0.0032	R2 < MINR2
AOV16: LOCEMP	4	0.0027	R2 < MINR2
Var: LOCEMP	1	0.0022	R2 < MINR2
Class: OWNED6	4	0.0018	R2 < MINR2
Group: OWNED6	1	0.0018	R2 < MINR2
Class: OWNED2	3	0.0014	R2 < MINR2
Group: OWNED2	1	0.0014	R2 < MINR2
AOV16: PORDER	4	0.0014	R2 < MINR2
AOV16: PACTSL	3	0.0014	R2 < MINR2
Var: PACTEM	1	0.0013	R2 < MINR2
Var: PACTSL	1	0.0006	R2 < MINR2
Var: PORDER	1	0.0006	R2 < MINR2
AOV16: PACTEM	2	0.0005	R2 < MINR2
Class: OWNED8	0	0.0000	R2 < MINR2
Class: PENU1	0	0.0000	R2 < MINR2
Class: OWNED4	0	0.0000	R2 < MINR2
Class: LAST1	0	0.0000	R2 < MINR2

SS and R2 portion for Effects chosen for target: CURYN_9

Effect	DF	R2	SS	EMS
Group: LAST9	2	0.1850	102.41160	0.2040201
Group: PENU3	2	0.0699	38.68944	0.1866903
Group: INDGROUP	8	0.0288	15.94625	0.1801239
Group: LAST7	3	0.0239	13.20234	0.1743632
Group: LAST3	4	0.0149	8.27122	0.1709112
Group: LAST5	4	0.0089	4.94598	0.1689649
Group: SUBNUM	2	0.0081	4.46252	0.1670798
Group: STATE	8	0.0076	4.17925	0.1657759
Class: FAXPRES	1	0.0071	3.93322	0.1640469
Group: OWNED3	7	0.0046	2.55881	0.1633975
Var: PSRV	1	0.0028	1.53012	0.1627680
Group: PENU9	1	0.0024	1.30404	0.1622420
Group: POPCOD	2	0.0017	0.95320	0.1619520
Group: CREDIT	3	0.0011	0.63507	0.1618831
Var: PNB	1	0.0012	0.64093	0.1616617
Group: LAST6	4	0.0010	0.54092	0.1617107
Class: GROWCO	1	0.0009	0.49948	0.1615542
Group: PENU2	3	0.0006	0.34679	0.1616182
Group: LAST2	2	0.0006	0.34562	0.1616078
Group: ADSIZ	2	0.0006	0.31992	0.1616093
Group: PENU7	2	0.0006	0.32916	0.1616066
Group: OWNED7	1	0.0009	0.51530	0.1614420

The final ANOVA table for target: CURYN_9

Effect	DF	R2	SS
Model	64	0.3732	206.56120
Error	2149		346.93880
Total	2213		553.50000

SS and R2 portion for Effects not chosen for target: CURYN_9

Effect		DF	R2	SS
Var:	PDT	1	0.0000	0.00023
Group:	PENU5	2	0.0001	0.05823
Group:	PENU6	1	0.0000	0.01739
Group:	LAST4	1	0.0001	0.08233
Var:	PWS	1	0.0004	0.19968
Class:	OWNED1	1	0.0002	0.09214
Group:	LAST8	1	0.0001	0.03551
Class:	INDFRM	1	0.0001	0.06487
Var:	SLSVDT	1	0.0003	0.18706
Group:	PENU4	1	0.0000	0.00404
Group:	OFFSIZ	2	0.0002	0.08540
Class:	WHITE	1	0.0003	0.17355
Group:	OWNED9	1	0.0001	0.05544

Estimating logistic

Iter	Alpha	Beta
0	-2.0928	4.1578
1	-3.0570	5.8764
2	-3.3829	6.4350
3	-3.4111	6.4825
4	-3.4113	6.4828

APÊNDICE I - T-SCORES DA REGRESSÃO LINEAR

Variable	Label	Parameter Estimates	Effect T-scores
INTERCEP	Intercept	5.1269	0.06499
PNB		-0.0540	-2.07340
PSRV		0.6416	3.53266
FAXPRE0	FAXPRES 0	-0.1748	-2.63213
G_INDG0	Group: INDGROUP 0	10.7370	0.11233
G_INDG1	Group: INDGROUP 1	0.0144	0.00120
G_INDG2	Group: INDGROUP 2	-0.8168	-0.06835
G_INDG3	Group: INDGROUP 3	-1.1119	-0.09304
G_INDG4	Group: INDGROUP 4	-1.1655	-0.09754
G_INDG5	Group: INDGROUP 5	-1.0607	-0.08877
G_INDG6	Group: INDGROUP 6	-1.7647	-0.14767
G_INDG7	Group: INDGROUP 7	-1.5215	-0.12732
DUP1	Group: LAST3 0	-0.0437	-0.34361
DUP2	Group: LAST3 1	0.5640	3.69449
G_LAST2	Group: LAST3 2	-0.0067	-0.06184
G_LAST3	Group: LAST3 3	-0.6829	-3.58878
DUP3	Group: LAST5 0	-0.1462	-1.23357
DUP4	Group: LAST5 1	0.8814	4.45716
DUP5	Group: LAST5 2	0.0761	0.53340
DUP6	Group: LAST5 3	-0.4824	-2.03892
DUP11	Group: LAST7 0	-1.1254	-3.60532
DUP12	Group: LAST7 1	0.5242	1.47530
DUP13	Group: LAST7 2	-0.3816	-1.12637
DUP14	Group: LAST9 0	-0.4397	-3.70824
DUP15	Group: LAST9 1	0.2393	1.29895
G_OWNE0	Group: OWNED3 0	-0.1527	-0.00307
G_OWNE1	Group: OWNED3 1	9.7671	0.07522
G_OWNE2	Group: OWNED3 2	-12.5658	-0.07179
G_OWNE3	Group: OWNED3 3	10.3554	0.10751
G_OWNE4	Group: OWNED3 4	-11.3419	-0.06480
G_OWNE5	Group: OWNED3 5	9.9069	0.12130
G_OWNE6	Group: OWNED3 6	-13.9499	-0.07970
G_PENU0	Group: PENU2 0	-1.9019	-0.04584
G_PENU1	Group: PENU2 1	-0.8141	-0.01962
G_PENU2	Group: PENU2 2	-4.6719	-0.11259
DUP17	Group: PENU3 0	-0.5877	-5.55192
DUP18	Group: PENU3 1	0.2648	2.41135
DUP21	Group: PENU9 0	-0.6963	-4.37698
G_STAT0	Group: STATE 0	0.6210	3.18010
G_STAT1	Group: STATE 1	0.3957	1.62550
G_STAT2	Group: STATE 2	0.3474	1.72207
G_STAT3	Group: STATE 3	0.1095	0.63906
G_STAT4	Group: STATE 4	0.0863	0.53805
G_STAT5	Group: STATE 5	-0.0333	-0.18342
G_STAT6	Group: STATE 6	-0.3027	-1.71108
G_STAT7	Group: STATE 7	-0.2846	-0.93853
G_SUBN0	Group: SUBNUM 0	12.1057	0.15152
G_SUBN1	Group: SUBNUM 1	-0.3806	-0.00876

APÊNDICE J - RESUMO DOS DADOS OBTIDOS POR REGRESSÃO LINEAR

Fit Statistic	Training	Validation	Test
Akaike's Information Criterion	2158.4878128	.	.
Average Squared Error	0.1528809489	0.1790114003	0.1752617748
Average Error Function	0.4657831556	0.5465445251	0.5364175522
Degrees of Freedom for Error	2166	.	.
Model Degrees of Freedom	48	.	.
Total Degrees of Freedom	2214	.	.
Divisor for ASE	4428	3330	3330
Error Function	2062.4878128	1819.9932687	1786.270449
Final Prediction Error	0.1596568359	.	.
Maximum Absolute Error	0.9968618244	0.9999949003	0.9999974532
Mean Square Error	0.1562688924	0.1790114003	0.1752617748
Sum of Frequencies	2214	1665	1665
Number of Estimate Weights	48	.	.
Root Average Sum of Squares	0.3909999347	0.4230973886	0.4186427771
Root Final Prediction Error	0.3995708146	.	.
Root Mean Squared Error	0.395308604	0.4230973886	0.4186427771
Schwarz's Bayesian Criterion	2432.2105063	.	.
Sum of Squared Errors	676.95684194	596.10796293	583.6217102
Sum of Case Weights Times Freq	4428	3330	3330
Misclassification Rate	0.22899729	0.2876876877	0.2786786787
Frequency of Classified Cases	2214	1665	1665
Frequency of Unclassified Cases	0	0	0
Total Profit for CURYN_9	1103.3840411	829.78068129	829.78068129
Average Profit for CURYN_9	0.4983667755	0.4983667755	0.4983667755

APÊNDICE K - TESTES DE QUI-QUADRADO E MEDIDAS DE CORRELAÇÃO PARA OS MODELOS DE REGRESSÃO E ÁRVORE

Teste Qui-Quadrado para Modelo de Regressão

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	311.882[b]	1	.000		
Continuity Correction[a]	310.115	1	.000		
Likelihood Ratio	324.415	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	311.694	1	.000		
N of Valid Cases	1665				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 328.71.

Medidas de correlação para o Modelo de Regressão

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Phi	.433			.000
	Cramer's V	.433			.000
	Contingency Coefficient	.397			.000
Interval by Interval	Pearson's R	.433	.022	19.578	.000[c]
Ordinal by Ordinal	Spearman Correlation	.433	.022	19.578	.000[c]
N of Valid Cases		1665			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

Teste Qui-Quadrado para Modelo de Árvore

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	447.290[b]	1	.000		
Continuity Correction[a]	445.220	1	.000		
Likelihood Ratio	469.838	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	447.022	1	.000		
N of Valid Cases	1665				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 412.26.

Medidas de correlação para o Modelo de Árvore

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Phi	.518			.000
	Cramer's V	.518			.000
	Contingency Coefficient	.460			.000
Interval by Interval	Pearson's R	.518	.021	24.715	.000[c]
Ordinal by Ordinal	Spearman Correlation	.518	.021	24.715	.000[c]
N of Valid Cases		1665			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

APÊNDICE L - REGRAS TEXTUAIS PARA ÄRVORE DE DECISÃO

```
IF  Group: LAST9 IS ONE OF: 1 2
 THEN
     NODE     :        3
     N        :      727
     1        :    80.1%
     0        :    19.9%

 IF  Group: LAST7 EQUALS 0
 AND Group: PENU3 EQUALS 0
 AND Group: LAST9 EQUALS 0
 THEN
     NODE     :        8
     N        :      963
     1        :    22.2%
     0        :    77.8%

 IF  Group: LAST7 IS ONE OF: 1 2 3
 AND Group: PENU3 EQUALS 0
 AND Group: LAST9 EQUALS 0
 THEN
     NODE     :        9
     N        :      104
     1        :    58.7%
     0        :    41.3%

 IF     0.302 <= PSRV
 AND Group: PENU3 IS ONE OF: 1 2
 AND Group: LAST9 EQUALS 0
 THEN
     NODE     :       11
     N        :      206
     1        :    75.2%
     0        :    24.8%

 IF  Group: LAST7 EQUALS 0
 AND PSRV <     0.302
 AND Group: PENU3 IS ONE OF: 1 2
 AND Group: LAST9 EQUALS 0
 THEN
     NODE     :       20
     N        :      193
     1        :    40.9%
     0        :    59.1%

 IF  Group: LAST7 IS ONE OF: 1 2 3
 AND PSRV <     0.302
 AND Group: PENU3 IS ONE OF: 1 2
 AND Group: LAST9 EQUALS 0
 THEN
     NODE     :       21
     N        :       21
     1        :    76.2%
     0        :    23.8%
```

APÊNDICE M - ESTRUTURA DA BASE DE DADOS PARA O MODELO APENAS COM DADOS CADASTRAIS

Model Name	Role	Measurement	Type	Variable Format	Label
ABI_NUM	rejected	nominal	char	$9.	
ADSIZ	input	nominal	char	$1.	
CREDIT	input	nominal	char	$1.	
FAXPRES	input	binary	num	BEST12.	
GROWCO	input	binary	char	$1.	
INDFRM	input	binary	char	$1.	
INDGROUP	input	nominal	char	$6.	
LOCEMP	input	interval	num	BEST12.	
OFFSIZ	input	nominal	char	$1.	
PACTEM	rejected	interval	num	BEST12.	
PACTSL	rejected	interval	num	BEST12.	
PDT	input	interval	num	BEST12.	
PNB	input	interval	num	BEST12.	
POPCOD	input	nominal	char	$1.	
PORDER	input	interval	num	BEST12.	
PSRV	input	interval	num	BEST12.	
PWS	input	interval	num	BEST12.	
SIC	rejected	nominal	char	$6.	
SLSVDT	input	interval	num	BEST12.	
SRV_QTY	rejected	interval	num	BEST12.	
SRV_REV	rejected	interval	num	BEST12.	
SRV_YES	target	binary	num	BEST12.	
SUBNUM	rejected	nominal	char	$9.	
WHITE	input	binary	char	$1.	
CMPY	input	nominal	char	$2.	
STATE	input	nominal	char	$2.	

APÊNDICE N - AMOSTRA DA LISTA DE PEDIDOS

CUST_NUM	PLP	FY	FQ	YYQQ	QTR	SEM	ANNUAL
4299	4IN	1999	4	994	6	3	2
5723	4DM	1999	1	991	3	2	1
5723	2NT	1999	2	992	4	2	1
5723	3WS	1999	2	992	4	2	1
5723	4DM	1999	2	992	4	2	1
5723	4DM	1999	3	993	5	3	2
5723	2NT	1999	4	994	6	3	2
5723	3WS	1999	4	994	6	3	2
5723	4DM	1999	4	994	6	3	2
5723	4LT	1999	4	994	6	3	2
5723	4SV	1999	4	994	6	3	2
5723	3WS	2000	1	1	7	4	2
5723	4DM	2000	1	1	7	4	2
5723	2NT	2000	2	2	8	4	2
5723	4DM	2000	2	2	8	4	2
5723	4IN	2000	2	2	8	4	2
5723	4LT	2000	2	2	8	4	2
5767	2NT	1998	3	983	1	1	1
5767	4LT	1998	3	983	1	1	1
5767	4PD	1998	3	983	1	1	1
5767	2NT	1999	1	991	3	2	1
5767	2NT	1999	2	992	4	2	1
5767	4DM	1999	2	992	4	2	1
5767	2NT	1999	3	993	5	3	2
5767	2NT	1999	4	994	6	3	2
5767	4DM	1999	4	994	6	3	2
5767	2NT	2000	1	1	7	4	2
5767	4PD	2000	1	1	7	4	2
5767	4SV	2000	1	1	7	4	2
11310	4SV	1999	1	991	3	2	1
11475	4DM	1998	3	983	1	1	1
11475	2NT	1998	4	984	2	1	1
11475	4DM	1998	4	984	2	1	1
11475	4DM	1999	1	991	3	2	1
11475	4SV	1999	1	991	3	2	1
11475	2NT	1999	2	992	4	2	1
11475	4DM	1999	2	992	4	2	1
11475	4IN	1999	2	992	4	2	1
11475	4SV	1999	2	992	4	2	1
11475	2NT	1999	3	993	5	3	2
11475	4DM	1999	3	993	5	3	2
11475	4SV	1999	3	993	5	3	2
11475	2NT	1999	4	994	6	3	2
11475	3WS	1999	4	994	6	3	2
11475	4DM	1999	4	994	6	3	2
11475	4ST	1999	4	994	6	3	2
11475	4SV	1999	4	994	6	3	2
11475	2NT	2000	1	1	7	4	2
11475	3WS	2000	1	1	7	4	2
11475	4DM	2000	1	1	7	4	2
11475	4PD	2000	1	1	7	4	2
11475	4SV	2000	1	1	7	4	2
11475	2NT	2000	2	2	8	4	2
11475	3WS	2000	2	2	8	4	2
11475	4DM	2000	2	2	8	4	2
11475	4SV	2000	2	2	8	4	2

APÊNDICE O - MÉDIA PADRONIZADA DAS VARIÁVEIS POR AGRUPAMENTO

Name	Agrup. 01	Agrup. 02	Agrup. 03	Agrup. 04	Agrup. 05
CURAMT_3	0.1937738	0.034907	0.0065417	0.0090902	0.7117017
CURAMT_7	0.4790874	0.0058861	0.0405262	0.0157843	0.3985117
LOCEMP	0.2399399	0.0975109	0.1087051	0.1207147	0.3135833
OWNED3	0.1930122	0.0612454	0.0266357	0.0091682	0.6500137
OWNED5	0.1905056	0.0046223	0.006377	0.1173076	0.6314221
OWNED6	0.1213572	0.0156215	0.0570427	0.0193124	0.7345303
OWNED7	0.1421874	0.0061074	0.0409641	0.0345898	0.7293654
OWNED9	0.0439016	0.0138719	0.0606422	0.019046	0.8136531
P_OWNED2	0.0427245	0.0760612	0.2658537	0.0916494	0.3732428
P_OWNED3	0.167561	0.2784763	0.1907717	0.0196587	0.1521063
P_OWNED4	0.0668777	0.052057	0.0954375	0.0095324	0.7025034
P_OWNED5	0.2001863	0.0077421	0.0267502	0.5422014	0.0997113
P_OWNED6	0.1244111	0.0470013	0.3054698	0.066821	0.3025714
P_OWNED7	0.1512456	0.0245901	0.3053353	0.0728264	0.2979458
P_OWNED9	0.066837	0.0320801	0.3482481	0.0505032	0.3448567
PDT	0.2355789	0.0779463	0.0926099	0.1217988	0.3648175
PNB	0.2252743	0.0841867	0.0895526	0.1126578	0.3818233
PORDER	0.2290122	0.0782545	0.0917351	0.1177449	0.3771342
PSRV	0.2218771	0.0873909	0.1039803	0.1266783	0.3456293
PWS	0.2160829	0.0532165	0.0678395	0.0889991	0.489573
SLSVDT	0.2353333	0.1099249	0.1483521	0.1170219	0.2529921

APÊNDICE P - CÓDIGO DE MERGER DOS PEDIDOS COM OS RESPECTIVOS AGRUPAMENTOS

```
data &_TRA;
        merge &_TRAIN X_data.assoc;
        by cust_num;
 if _SEGMNT_ = . then delete;
 keep cust_num PLP FY FQ QTR SEM ANNUAL _SEGMNT_
 run;


 data &_TES;
        merge &_TEST X_data.assoc;
        by cust_num;
 if _SEGMNT_ = . then delete;
 keep cust_num PLP FY FQ QTR SEM ANNUAL _SEGMNT_
 run;

 data &_VAL;
        merge &_VALID X_data.assoc;
        by cust_num;
 if _SEGMNT_ = . then delete;
 keep cust_num PLP FY FQ QTR SEM ANNUAL _SEGMNT_
 run;
```

```
data &_TRA (keep=cust_num PLP FY FQ QTR SEM ANNUAL _SEGMNT_);
 set &_TRAIN;
    if _SEGMNT_ ^= '1' then delete;
 run;

 data &_TES (keep=cust_num PLP FY FQ QTR SEM ANNUAL _SEGMNT_);
 set &_TEST;
    if _SEGMNT_ ^= '1' then delete;
 run;

 data &_VAL (keep=cust_num PLP FY FQ QTR SEM ANNUAL _SEGMNT_);
 set &_VALID;
    if _SEGMNT_ ^= '1' then delete;
 run;
```

Observação:

Os outros nós *Code Splitter* são idênticos a este, sendo a única diferença a mudança da instrução _SEGMENT_ ^= '1' para _SEGMENT_ ^= 'Y', onde Y é o número do agrupamento.

APÊNDICE R - CÓDIGO SAS GERADOR DA LISTA DE PEDIDOS

Este foi o código utilizado na criação da lista de pedidos para posterior busca de regras de associação. Uma amostra da saída resultante encontra-se no APÊNDICE M -, página 120. Uma amopstra dos dados de entrada (bsd_ut) encontra-se noAPÊNDICE A -, página98.

```
**--Creating summary dataset by Customer and Fiscal Quarter--**;
**--Whenever you use a BY statement, the data must be sorted in
    this order. Therefore, it is generally a good idea to always
    make sure your data is sorted in the appropriate fashion--**;

    proc sort data = X_data.bsd_ut;

      by cust_num PLP FY fq ;

    proc summary data = X_data.bsd_ut;

       var sys_qty net_rev ;

       by cust_num PLP FY fq ;

       output out=custsum sum=qty net ;

proc format;
        value yrqtr
                                  983 = 1   984 = 2
                991 = 3   992 = 4   993 = 5   994 = 6
                  1 = 7     2 = 8;

data X_data.assoc;
        set custsum;

        yyqq = 10*MOD(FY,100) + FQ;
        QTR1 = put(yyqq, yrqtr.);          ** DON'T FORGET THE . **;
        QTR = input(QTR1, best12.);

        if QTR in (1,2) then SEM = 1;
        if QTR in (3,4) then SEM = 2;
        if QTR in (5,6) then SEM = 3;
        if QTR in (7,8) then SEM = 4;

        if SEM in (1,2) then ANNUAL = 1;
        if SEM in (3,4) then ANNUAL = 2;

        keep cust_num yyqq PLP FY FQ QTR SEM ANNUAL;


proc print data =X_data.assoc (obs=30) ;
       title4 'X_ data:Summarized by customer, PLP, FY, FQ';

run;
```

BIBLIOGRAFIA

- Beale, R., Jackson, T. *"Neural Computing, An Introduction"* Bristol, Arrowsmith, 1990;
- Bernoff, J "Consumer eCommerce Readiness", www.forrester.com/ER/Research/Report/0,1338,3865,FF.html, 1998
- Edelstein, H. *"Mining for Gold"*, Information Week April 21, 1997, www.twocrows.com/iwk9704.html
- Hair et al, *"Multivariate Data Analysis"* Fifth Edition, Prentice Hall, 1998
- Hammond, M. , PC Week (14 de Março de 2000)
- Information Discovery *"Measuring the Dollar Value of Mined Information"*, www.datamine.aa.psiweb.com/infoval2.html, www.datamining.com, 1998
- Information Discovery *"OLAP and datamining, Bridging The Gap"*, www.datamine.aa.psiweb.com/bridge.html, www.datamining.com, 1996
- Jackson, B. B. *"Multivariate Data Anaysis"* Illinois, Richard Irwin, 1983
- Kotler, P. *"Marketing Management"* Ninth Edition, Prentice Hall, 1997
- Kotler, P. *"Marketing Para O Século XXI"* Terceira Edição, Futura, 1999
- Malhotra, N. *"Marketing Research, An Applied Orientation"* Second Edition, Prentice Hall, 1996
- Stewart, T. A. "*Intellectual Capital: The New Wealth of Organizations*" New York, Doubleday, 1997.
- Two Crows Corporation, *"Introduction to Data Mining and Knowledge Discovery"* Third Edition, www.twocrows.com, 1999

ÍNDICE