CLASSIFICAÇÃO DE *CHURN* NO SEGURO AUTOMÓVEL

POR

RUI MARCO CACHINHO ALMEIDA

RELATÓRIO DE PROJECTO APRESENTADO COMO REQUISITO PARCIAL
PARA A OBTENÇÃO DO GRAU DE MESTRE EM
ESTATÍSTICA E GESTÃO DE INFORMAÇÃO

PELO

INSTITUTO SUPERIOR DE ESTATÍSTICA E GESTÃO DE INFORMAÇÃO
DA
UNIVERSIDADE NOVA DE LISBOA

CLASSIFICAÇÃO DE *CHURN* NO SEGURO AUTOMÓVEL

POR

RUI MARCO CACHINHO ALMEIDA

RELATÓRIO DE PROJECTO APRESENTADO COMO REQUISITO PARCIAL
PARA A OBTENÇÃO DO GRAU DE MESTRE EM
ESTATÍSTICA E GESTÃO DE INFORMAÇÃO

PELO

INSTITUTO SUPERIOR DE ESTATÍSTICA E GESTÃO DE INFORMAÇÃO

DA

UNIVERSIDADE NOVA DE LISBOA

RELATÓRIO DE PROJECTO ORIENTADO POR
PROFESSOR DOUTOR VICTOR LOBO

NOVEMBRO DE 2010

"Deus quer, o homem sonha, a obra nasce."
Fernando Pessoa

Agradeço ao, meu orientador Professor Doutor Victor Lobo pela sua disponibilidade, experiência, sugestões e por todo o seu apoio, nos bons e maus momentos.

À minha família, por me aturar há tanto tempo e sempre com um sorriso no rosto.

Aos meus colegas e amigos pelo entusiasmo, curiosidade, infinita amizade, paciência, revisões, críticas, colaborações, momentos de diversão, etc.

À Marktest, nomeadamente à Drª Bárbara Gomes e à Drª Catarina Santos, por terem autorizado a utilização dos dados que serviram de base para este relatório de projecto.

É muito bom chegar ao fim e poder dizer que tive prazer e me diverti com a execução deste projecto. Sem todos vocês isso teria sido impossível. Obrigada a todos!

ÍNDICE

RESUMO.....VIII
GLOSSÁRIO.....IX

I INTRODUÇÃO..... 1

1.1 Objectivos.....2
1.2 Organização do trabalho.....2

II ENQUADRAMENTO..... 4

1. CHURN.....5
1.1 Gestão de Churn.....9
1.2 Churn no Sector Financeiro.....12
2. DATA MINING.....13
2.1 Árvores de Decisão.....14
a) Algoritmos de indução.....16
2.2 Redes Neuronais.....19
a) Elementos constitutivos.....19
b) Arquitectura.....21
c) Perceptrão Multicamada.....23
3. NÃO RESPOSTAS.....25
3.1 Tratamento de Não Respostas.....27
a) Eliminação.....27
b) Imputação.....28
4. PROBLEMA E OBJECTIVOS.....32

III MÉTODO.....34

1. CARACTERIZAÇÃO DA BASE DE DADOS.....35
1.1 O Barómetro de Seguros.....35
1.2 Estudo de Transferências de Seguros.....36
1.3 Base de Dados de Treino.....37
1.4 Base de Dados de Classificação.....38
1.5 Análise descritiva dos dados.....39
2. METODOLOGIA.....46
3. FERRAMENTAS.....47

IV ANÁLISE DE DADOS E RESULTADOS.....48

1. MODELOS.....54
1.1 Árvore de Decisão.....54
1.2 Rede Neuronal.....62
1.3 Probit.....67
1.4 Simulação de Monte Carlo.....74
2. SCORING.....76
3. RATING.....78

V DISCUSSÃO.....80

VI REFERÊNCIAS BIBLIOGRÁFICAS.....85

VII ANEXOS..... 88

Índice de Anexos

ANEXO A: Listagem de variáveis/indicadores – Estudo de Barómetro de Seguros.....89
ANEXO B: Listagem de variáveis/indicadores – Estudo de Transferências de Seguros.....96
ANEXO C: Autorização formal para uso dos dados.....98
ANEXO D: Output de Análise Descritiva.....100
ANEXO E: Listagem de Variáveis/Indicadores Criados.....101
ANEXO F: Árvore de Decisão.....104
ANEXO G: Diagrama SAS.....106
ANEXO H: Output GRETL.....108
ANEXO I: Simulação de Monte Carlo.....109

Índice de Ilustrações

Ilustração 1 – Exemplo de uma Árvore de Decisão 15
Ilustração 2 - Estrutura de um Neurónio Natural típico 20
Ilustração 3 - Estrutura de um Neurónio Artificial típico.....20
Ilustração 4 - Perceptrão Multicamada 24
Ilustração 5 - Árvore de Decisão (Três primeiros níveis) 59

Índice de Tabelas

Tabela 1 - Variáveis adicionais.....45
Tabela 2 - Varáveis incluídas no modelo de Árvore de Decisão 54
Tabela 3 - Taxas de Erro da Árvore de Decisão nos vários conjuntos de dados. 55
Tabela 4 - Varáveis incluídas no modelo de Rede Neuronal 62
Tabela 5 - Taxas de Erro da Rede Neuronal nos vários conjuntos de dados. 63
Tabela 6 - Varáveis incluídas no modelo Probit 67
Tabela 7 - Taxas de Erro do Probit nos vários conjuntos de dados. 68
Tabela 8 – Coeficientes β e níveis de significância do modelo Probit 71
Tabela 9 – Resultados da simulação de Monte Carlo. 74
Tabela 10 – Resultados das classificações efectuadas. 77
Tabela 11 – Rating de Risco de Churn. 78

Índice de Gráficos

Gráfico 1 - Impacto de diferentes Taxas Retenção nos resultados financeiros de uma companhia.....8
Gráfico 2 - Mudança de seguradora.....39
Gráfico 3 - Género do inquirido.....40
Gráfico 4 - Idade do inquirido.....40
Gráfico 5 - Número de seguros automóvel.....41
Gráfico 6 - Coberturas garantidas.....42
Gráfico 7 - Participação de sinistro.....42
Gráfico 8 - deixou de trabalhar com alguma seguradora.....43
Gráfico 9 - Prémio anual do Seguro Automóvel.....44
Gráfico 10 - Distribuição dos casos positivos e negativos na Base de Dados após Boosting.....51
Gráfico 11 - Matriz de Confusão da Árvore de Decisão.....56
Gráfico 12 - Lift da Árvore de Decisão.....57
Gráfico 13 - Ganhos Cumulativos da Árvore de decisão.....58
Gráfico 14 - Matriz de Confusão da Rede Neuronal.....64
Gráfico 15 - Lift da Rede Neuronal.....65
Gráfico 16 - Ganhos Cumulativos da Rede Neuronal.....66
Gráfico 17 - Matriz de Confusão do Probit.....69
Gráfico 18 - Lift do Probit.....70
Gráfico 19 - Ganhos Cumulativos do Probit.....71

Resumo

Actualmente assistimos à emergência de mercados cada vez mais saturados e competitivos, como acontece por exemplo no sector dos seguros. Os consumidores ganham cada vez mais poder de escolha. Tal facto tem contribuído para colocar o *Churn*, e a sua gestão, como um assunto central para a sobrevivência de muitas companhias. (e.g. Hadden, Tiwari, Roy, & Ruta, 2005).

Partindo de dados provenientes de uma sondagem, que visa avaliar os comportamentos e percepções dos portugueses face aos seguros, pretende-mos desenvolver modelos de classificação de probabilidade de C*hurn* para os possuidores de Seguro Automóvel.

Os primeiros modelos desenvolvidos apresentaram uma Taxa de Erro baixa (cerca de 20%) e idêntica à percentagem de *Churners* presente na Base de Dados. Não apresentavam qualquer ganho ao nível das curvas de *Lift,* o que aponta para um comportamento idêntico a uma decisão aleatória.

Ao aplicar a técnica de *Boosting* - de modo a equilibramos os casos negativos e positivos - foi-nos possível desenvolver três modelos: uma Árvore de Decisão; uma Rede Neuronal e um Probit (Taxas de Erro na ordem dos 36-45%). Estes modelos apresentam ganhos razoáveis ao nível das curvas de *Lift* e das matrizes de confusão, classificando grande parte dos *Churners* como potenciais *Churners* e apresentando um número reduzido de falsos positivos. Isto sugere que estes modelos podem ser bastante úteis no contexto de campanhas de retenção de clientes.

São identificados alguns grupos de risco, como por exemplo: os inquiridos que têm pouco contacto com a companhia e idade até 55 anos. Apresentamos igualmente sugestões para investigação futura.

Palavras-chave: Classificação; *Churn*; Seguro Automóvel; Rede Neuronal; Árvore de Decisão

Glossário

i.e.	isto é
c.f.	confronte com
e.g.	por exemplo
p. ou *pág.*	página
etc.	por aí adiante
et al.	e outros

I Introdução

Na actualidade, e à medida que os mercados vão ficando cada vez mais saturados, os consumidores vão adquirindo mais poder de escolha relativamente aos seus fornecedores (Hadden et al., 2005). Desta forma a questão do *Churn* - movimento de clientes entre fornecedores de um dado serviço (e.g. Hadden et al., 2005) - tem assumido um papel decisivo na gestão das companhias, tendo sido demonstrado pela investigação que é mais caro adquirir novos clientes do que reter os clientes actuais (Jacob, 1994).

Se numa abordagem mais reactiva à problemática do *Churn*, compreender porque o cliente abandona a companhia pode ser suficiente, no âmbito de abordagens mais pró-activas é necessário identificar os clientes que potencialmente podem vir a abandonar a companhia (Burez & Van den Poel, 2007).

O problema da identificação dos clientes que potencialmente podem vir a abandonar uma companhia tem sido conceptualizado como um problema de classificação (Andrade, 2007; Au & Ma, 2003; Burez & Van den Poel, 2007; Garcia, 2003; Hung, Yen, & Wag, 2006). Muitas empresas têm procurado dar-lhe resposta aplicando técnicas de *Data Mining*. Entre as técnicas mais usadas neste contexto temos as Árvores de Decisão e as Redes Neuronais (Hadden et al., 2005).

Entre as indústrias com mais necessidades nesta área temos, entre outras, as Telecomunicações (Wei & Chiu, 2002), a Banca e os Seguros (Morik & Kopcke, 2004).

1.1 Objectivos

Este trabalho foi realizado no âmbito de um projecto de desenvolvimento de um novo produto para a empresa de sondagens e estudos de opinião *Marktest.*

Partindo de uma Base de Dados proveniente de uma sondagem designada por "Barómetro de Seguros" (Marktest, 2006) o objectivo do presente trabalho é desenvolver modelos de classificação de probabilidade de *Churn* para os possuidores de Seguro Automóvel.

Estes modelos serão um complemento ao portfólio de produtos que a *Marktest* já disponibiliza aos clientes do sector segurador.

1.2 Organização do trabalho

Este trabalho visa estudar a problemática do *Churn* no Seguro Automóvel. De seguida apresentamos um pequeno resumo da sua orientação e organização.

Neste primeiro capítulo é apresentado o tema em estudo, bem como os seus objectivos.

No segundo capítulo é feito um enquadramento mais abrangente e profundo do tema do *Churn*. Numa primeira parte é apresentada uma definição de *Churn (e.g. Andrade, 2007; Au & Ma, 2003; Lejeune, 2001)*, é discutido o seu impacto económico para as companhias (e.g. Van den Poel & Larivière, 2004) e são apresentadas estratégias para lidar com este problema (e.g. Burez & Van den Poel, 2007). Numa segunda parte é introduzida a tecnologia de *Data Mining* enquanto ferramenta para responder a problemas de classificação de probabilidade de *Churn (Hadden et al., 2005)*. São descritas com mais pormenor as técnicas de Árvore de Decisão e Redes Neuronais. Num terceiro ponto centramos a

nossa atenção nas Não Respostas: São discutidos os vários tipos de Não Respostas que podemos encontrar e as formas de tratamento das mesmas (Batista & Monard, 2003; Graham, 2009; Little & Rubin, 1987; Roth, 1994; Schafer & Graham, 2002).

O terceiro capítulo apresenta-nos a metodologia adoptada para este projecto. Começamos por caracterizar as nossas Bases de Dados (Barómetro de Seguros e Transferências de Seguros), sendo feita igualmente a análise descritiva das mesmas. Neste capítulo relatamos os procedimentos executados para a realização deste trabalho e indicamos as Ferramentas Informáticas utilizadas.

O quarto capítulo é dedicado à Análise dos nossos dados, sendo apresentados os modelos desenvolvidos e as suas características: Taxas de Erro, Matrizes de Confusão, Gráficos de *Lift*, Gráficos de Ganhos Cumulativos, etc.

Por fim no último capítulo discutimos os resultados obtidos e apresentamos sugestões para a investigação futura.

II Enquadramento

Neste segundo capítulo fazemos um enquadramento do tema deste trabalho: o *Churn*.

Na primeira parte é feita uma revisão bibliográfica sobre o conceito de *Churn*, o seu impacto económico para as companhias e as estratégias que se podem adoptar para lidar com este fenómeno. A pertinência da classificação do *Churn* fica explicitada no contexto de uma gestão pró-activa do mesmo. É focado o caso específico do sector financeiro.

Na segunda parte introduzimos a temática do *Data Mining* como uma tecnologia capaz de dar resposta ao problema da classificação de probabilidade de *Churn*. Descrevemos ainda os algoritmos mais usados na literatura para resolver este problema: as Árvores de Decisão e as Redes Neuronais.

Dado que os nossos dados provêm de uma sondagem abordamos, na terceira parte, a questão das Não Respostas (ou *Missing Values*). Fazemos uma breve revisão de literatura sobre os diferentes tipos de Não respostas com que nos podemos deparar e sobre as técnicas de tratamento que temos actualmente ao nosso dispor.

Por último, na quarta parte, definimos o problema e os objectivos que nos propomos atingir com a realização deste projecto.

1. Churn

É comum dizer-se que os clientes são os activos mais valiosos de uma companhia, no entanto nem sempre é fácil conquistar novos clientes ou manter os nossos actuais clientes.

O termo *Churn* tem origem na expressão inglesa "*Change and turn*" e é vulgarmente usado na indústria e na literatura para reflectir a descontinuação de um contrato (Lazarov & Capota, 2007). Define-se como o movimento de clientes entre vários fornecedores de um dado serviço (Au & Ma, 2003; Hadden et al., 2005; Hongxia, Min, & Jianxia, 2009; Hung et al., 2006).

Pode ser pensado como uma medida de *Turnover*, ou de infidelidade, de uma base de clientes (Andrade, 2007; Lejeune, 2001). A sua operacionalização por ser feita através da percentagem de clientes que uma empresa perde num período de tempo específico, geralmente um ano (Lejeune, 2001).

O *Churn* está hoje no centro das preocupações das companhias pois à medida que os mercados vão ficando cada vez mais saturados, os consumidores vão adquirindo mais poder de escolha relativamente aos seus fornecedores (Hadden *et al.*, 2005). É justamente em mercados com estas características (elevada maturidade) que existe maior mudança de fornecedores. Os clientes, mudam em busca de melhores preços e melhores serviços (Hongxia *et al.*, 2009; Lejeune, 2001).

É comum distinguir-se, na literatura, dois tipos de *Churn*: o involuntário e o voluntário.

O *Churn* involuntário ocorre quando, por exemplo, um cliente deixa de pagar pelos serviços fornecidos e a companhia cancela o fornecimento dos mesmos (Hadden *et al.*, 2005).

O *Churn* voluntário refere-se às situações em que o cliente decide, de forma consciente, abandonar os serviços de uma companhia (Hadden *et al.*, 2005). Este divide-se ainda em *Churn* acidental e *Churn* deliberado.

No conceito de *Churn* acidental acidental integram-se, por exemplo, as situações em que as circunstâncias se alteram, fazendo com que o cliente não consiga manter o serviço (desemprego, mudança para uma região onde o serviço não está disponível, etc.). Este representa apenas uma pequena parte das taxas de *Churn* das companhias (Hadden et al., 2005; Lazarov & Capota, 2007).

O conceito de *Churn* deliberado refere-se às situações em que o cliente decide mudar de fornecedor. Entre as razões mais comuns para o *Churn* deliberado encontram-se factores relacionados com as relações entre a companhia e o cliente, factores estes que podem ser controlados pela companhia. Temos como exemplo: a clareza de facturação e o serviço pós-venda, etc. Este é o tipo de *Churn* que as companhias mais necessitam de combater (Hadden et al., 2005; Hadden, Tiwari, Roy, & Ruta, 2006; Lazarov & Capota, 2007).

A retenção de clientes tem um grande valor económico para as companhias (Buckinx & Van den Poel, 2005; Jacob, 1994) pois é mais rentável manter os actuais clientes do que adquirir novos. Ou seja, é mais barato manter os clientes actuais do que renovar continuamente a base de clientes (Burez & Van den Poel, 2007).

O valor económico da retenção de uma base de clientes é amplamente reconhecido. Os seus principais impactos podem sistematizar-se da seguinte forma (Van den Poel & Larivière, 2004):

- A forte retenção de uma base de clientes faz diminuir a necessidade de recrutar novos, e potencialmente perigosos, clientes.

- Permite à companhia concentra-se nas necessidades da sua base de clientes actual construindo com ela relações de longo prazo.

- Clientes com maior antiguidade tendem a gastar mais dinheiro.

- Quando satisfeitos, os clientes podem proferir boas referências da companhia através do *Word-of-Mouth*.

- Os clientes mais antigos tornam-se menos onerosos, devido ao conhecimento que a companhia adquire do seu comportamento e das suas necessidades.

- Tendem a ser menos sensíveis às estratégias de *Marketing* e aos apelos da concorrência.

- Perder clientes não representa apenas um custo de oportunidade apenas devido ao decréscimo das vendas, mas também devido a um aumento da necessidade de recrutar novos clientes, o que é seis a sete vezes mais caro que a retenção.

- As pessoas tendem a partilhar mais as experiências negativas do que positivas, o que pode resultar numa imagem negativa da companhia junto de potenciais novos clientes.

A título de exemplo, e para demonstrar o impacto financeiro da retenção de clientes podemos analisar uma simulação proposta por Van den Poel e Larivière (2004): Suponhamos que uma companhia tem um milhão de clientes, que a sua taxa de *Churn* é de 7% e que cada cliente tem uma contribuição líquida de 50€ por ano.

Podemos então calcular os resultados para a taxa de retenção, inicial, de 93% (*Churn* de 7%), e para as taxas retenção alternativas de 75% e 94% ao longo de 25 anos. Os resultados estão sistematizados no gráfico abaixo (Van den Poel & Larivière, 2004).

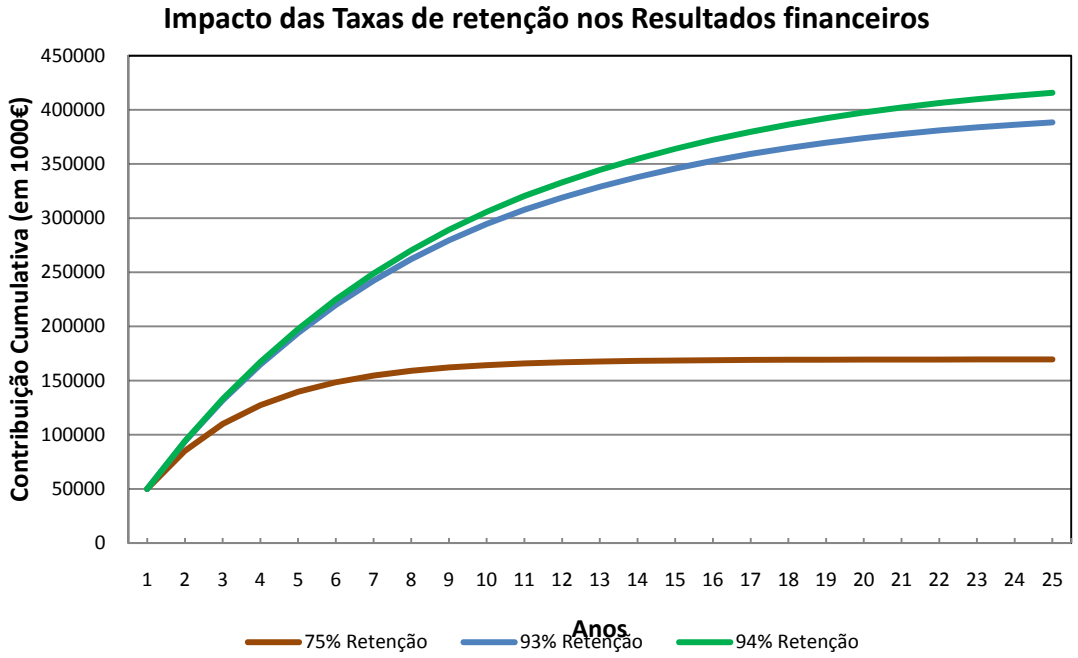


Gráfico 1 - Impacto de diferentes Taxas Retenção nos resultados financeiros de uma companhia

(Fonte: Van den Poel e Larivière (2004))

O aumento da taxa de retenção de clientes em 1p.p. - de 93% para 94% - aumentaria os resultados da companhia dos 392.2 Milhões de Euros para 419.7 Milhões de Euros. Assim, num período de 25 anos os resultados da companhia aumentariam em 27.5 Milhões de Euros (considerando um desconto de 6%).

As companhias têm assim necessidade de procurar meios que permitam aumentar o nível de retenção dos seus clientes, muitas encaram a retenção de clientes como a melhor estratégia para sobreviver (Wei & Chiu, 2002). A retenção de clientes tem-se tornado assim uma das pedras basilares das estratégias de *Marketing* (Ganesh, Arnold, & Reynolds, 2000) que se centram cada vez mais nos clientes e menos nos produtos (Hadden et al., 2006).

1.1 Gestão de Churn

O termo *Gestão de Churn* refere-se às estratégias e aos esforços que uma determinada companhia faz para reter os seus clientes mais rentáveis (Hung *et al.*, 2006).

As estratégias de gestão de *Churn* podem ser apresentadas em duas categorias (Burez & Van den Poel, 2007): as Não Dirigidas (*Untargeted)* e as Dirigidas (*Targeted)*.

As estratégias Não dirigidas (*Untargeted*) baseiam-se na publicidade em massa para aumentar os níveis de lealdade da marca (Burez & Van den Poel, 2007).

Estas estratégias apresentam um enorme inconveniente. Na generalidade dos casos não faz sentido aplicar estratégias de gestão de *Churn* a toda sua base de clientes, já que (Hadden *et al.*, 2005):

- Não vale a pena reter todos os clientes,
- Estas estratégias são dispendiosas, aplica-las a clientes que não têm intenção de abandonar é desperdiçar recursos.

Não admira então que falar de *Gestão de Churn,* para muitos autores (Au & Ma, 2003; Hung *et al.*, 2006), implique que se seja capaz de prever a intenção de um dado cliente em trocar os produtos e serviços de uma companhia pelos dos seus concorrentes.

As estratégias Dirigidas (*Targeted*) são focalizadas nos clientes que têm maior probabilidade de abandonar a companhia (Burez & Van den Poel, 2007). Assentam em oferecer a estes clientes algum incentivo para não a abandonarem. Existem dois tipos de estratégias *Targeted*: as reactivas e as pró-activas.

Nas abordagens reactivas a companhia espera até que o cliente a contacte para cancelar o serviço. Nessa altura a companhia apresenta-lhe um incentivo, por exemplo um desconto, para que ele se mantenha como cliente.

Nas abordagens pró-activas a companhia tenta identificar antecipadamente os clientes com maior probabilidade de a vir a abandonar num futuro próximo. Não se trata de caracterizar o *Churn*, ou de acompanhar a sua evolução ao longo do tempo, mas de identificar os clientes em risco de abandonar a companhia antes que tal ocorra (Morik & Kopcke, 2004).

Neste contexto compreender porque os clientes saem deve ser encarado apenas como um primeiro passo para a construção de estratégias de retenção eficazes, o segundo é a identificação antecipada dos clientes com elevados riscos de abandono (Au & Ma, 2003).

A companhia pode então centrar-se nestes clientes e, por exemplo, desenvolver incentivos específicos para que os clientes não a abandonem (Burez & Van den Poel, 2007).

As abordagens pró-activas podem ter enormes vantagens: por exemplo têm menores custos ao nível dos incentivos, pois geralmente não necessitam de ser tão elevados como os associados às abordagens reactivas. Contudo podem ser desperdiçadoras, se a identificação dos potenciais *Churner*s não for precisa, as companhias podem ser levadas a gastar recursos com clientes que não tinham intenção de abandonar os seus serviços (Burez & Van den Poel, 2007).

Dado que as abordagens pró-activas baseiam-se na ideia que a companhia consegue identificar antecipadamente os clientes que possuem um maior risco de *Churn (Morik & Kopcke, 2004)* estas necessitam – implicitamente - de se dotarem de ferramentas que lhes

permitam determinar com rigor a probabilidade de um dado cliente se tornar *Churner* (i.e. os clientes que têm maior probabilidade de abandonarem a companhia).

Estas ferramentas são, no essencial e na actualidade, modelos estatísticos como Árvores de Decisão, Redes Neuronais ou Modelos Econométricos (e.g. Hadden et al., 2005). Estes modelos são tão mais valiosos para a gestão pró-activa de *Churn* quanto: (a) mais *Churners* forem capazes de identificar e (b) menos Falsos Positivos gerarem (clientes que não tem intenção de abandonar a companhia identificados erroneamente como *Churners* pelos modelos).

Um modelo ideal seria capaz de identificar a totalidade dos *Churners* sem assinalar qualquer Falso Positivo. Na posse de uma ferramenta destas as companhias poderiam dirigir os seus recursos e orçamentos para os clientes identificados como *Churners*, conseguindo desta forma optimizar as suas estratégias de gestão pró-activa de *Churn*.

1.2 Churn no Sector Financeiro

Como exemplos de indústrias com elevadas necessidades nesta área temos as Telecomunicações (Wei & Chiu, 2002), a Banca e os Seguros (Morik & Kopcke, 2004).

É amplamente sabido que as companhias do sector financeiro (e.g. Banca, Seguros) detêm uma enorme quantidade de dados ao seu dispor (e.g. Morik & Kopcke, 2004; Staudt, Kietz, & Reimer, 1998). Estes têm tradicionalmente sido usados para várias tarefas como por exemplo análise de risco, seguros de saúde, credit scoring, etc. (e.g. Kahane, Levin, & Zahavi, 2005; Morik & Kopcke, 2004)

Recentemente as companhias do sector financeiro começaram a capitalizar estes dados no âmbito das suas políticas de *Customer Relationship Management* - CRM (Staudt et al., 1998). Um caso especial do uso destes dados para CRM é a gestão de *Churn (Morik & Kopcke, 2004)*. A relevância deste tipo de projectos vem, como já referimos, do impacto económico que a diminuição da taxa de *Churn* pode ter a longo prazo nos resultados financeiros (Van den Poel & Larivière, 2004).

As Bases de Dados destas companhias contêm, potencialmente, uma grande quantidade de conhecimento relevante para a gestão de *Churn* escondida. A extracção de conhecimento desta enorme massa de dados necessita de algoritmos eficientes. A utilização de técnicas de *Data Mining* surgiu como uma solução para esta tarefa e popularizou-se assim no sector financeiro (Morik & Kopcke, 2004).

2. *Data Mining*

O *Data Mining* procura a detecção automática de padrões, previamente desconhecidos e potencialmente úteis, que não se encontram explicitamente descriminados numa Base de Dados (Cortes, 2005; Hongxia *et al.*, 2009; Staudt *et al.*, 1998; Weiss & Indurkhya, 1998).

O *Data Mining* oferece-nos actualmente uma enorme quantidade de algoritmos para extrair conhecimento de Bases de Dados. Esses algoritmos são usados para resolver problemas como: descoberta de afinidades, *clustering*, predição, estimação de propriedades, descrição ou classificação de entidades (Bação, 2006; Cortes, 2005). A classificação de entidades é das tarefas mais comuns na área de apoio à tomada de decisão nas companhias (e.g. Cortes, 2005).

Na maior parte da investigação sobre *Churn* procura-se identificar e caracterizar os clientes que podem vir a abandonar a companhia, i.e. potenciais *Churner*s. Pretende-se, deste modo, a poder estabelecer medidas, de forma pró-activa, no sentido de evitar que no futuro clientes com características semelhantes possam vir a ter um comportamento de *Churn* (Hongxia *et al.*, 2009).

O problema do *Churn* é tradicionalmente conceptualizado como um problema de classificação (e.g. Andrade, 2007; Au & Ma, 2003; Burez & Van den Poel, 2007; Garcia, 2003; Hung *et al.*, 2006). Este é resolvido na maioria dos casos com recurso a técnicas de *Data Mining*. A análise da literatura sobre as técnicas de *Data Mining* mais usadas para classificação de *Churn* demonstra que as técnicas de Árvores de Decisão e Redes Neuronais dominam a investigação nesta área (Hadden *et al.*, 2005).

De seguida iremos fazer uma breve apresentação destas técnicas.

2.1 Árvores de Decisão

As Árvores de Decisão merecem actualmente uma grande atenção da comunidade científica de várias áreas do conhecimento (Geurts, Irrthum, & Wehenkel, 2009; Paula, 2002).

As Árvores de Decisão podem ser caracterizadas como representações do conhecimento e como robustas ferramentas de classificação (Bação, 2006; Lemos, 2003; Quinlan, 1986). A simplicidade está na origem da sua popularidade como iremos ver mais adiante (Cortes, 2005).

As Árvores de Decisão são constituídas por (Carvalho, 2005; Garcia, 2003; Paula, 2002):

- nós que representam os atributos (i.e. variáveis independentes),
- ramos, provenientes dos nós, que recebem os valores possíveis para as variáveis consideradas, e por
- nós folha (ou simplesmente folhas), que representam as diferentes classes da variável dependente de um conjunto de dados de treino, assim cada folha fica associada a uma determinada classe.

Ao contrário das Árvores na natureza – que crescem da raíz para as folhas - as Árvores de Decisão são construídas numa metodologia *top-down* (Bação, 2006; Quinlan, 1986). Representa-se no topo o nó de raiz da Árvore, deste crescem ramos que dão origem a novos nós, e assim sucessivamente até se chegar a um nó folha.

Nos ramos superiores da análise ficam, desta forma, representadas as variáveis que nos trazem maiores ganhos de informação e que contribuem para uma convergência mais rápida para um nó folha e, consequentemente, para uma conclusão (Bação, 2006; Cortes, 2005).

Analisando uma Árvore de Decisão é intuitivo que cada percurso na Árvore - da raiz a uma dada folha - corresponde a uma regra de classificação (Lemos, 2003). Estas regras são facilmente traduzíveis em acção, em linguagem de programação ou de Base de Dados (e.g. linguagem SQL).

Na ilustração seguinte (Garcia, 2003) é mostrado, de uma forma simples, uma Árvore de Decisão construída com base em três atributos.

Neste exemplo os nós são representados pelos atributos *a1*, *a3* e *a4* que se encontram apresentados na Árvore de acordo com o seu valor informativo. Cada círculo no final dos ramos da Árvore indica a classe associada aos nós folha.

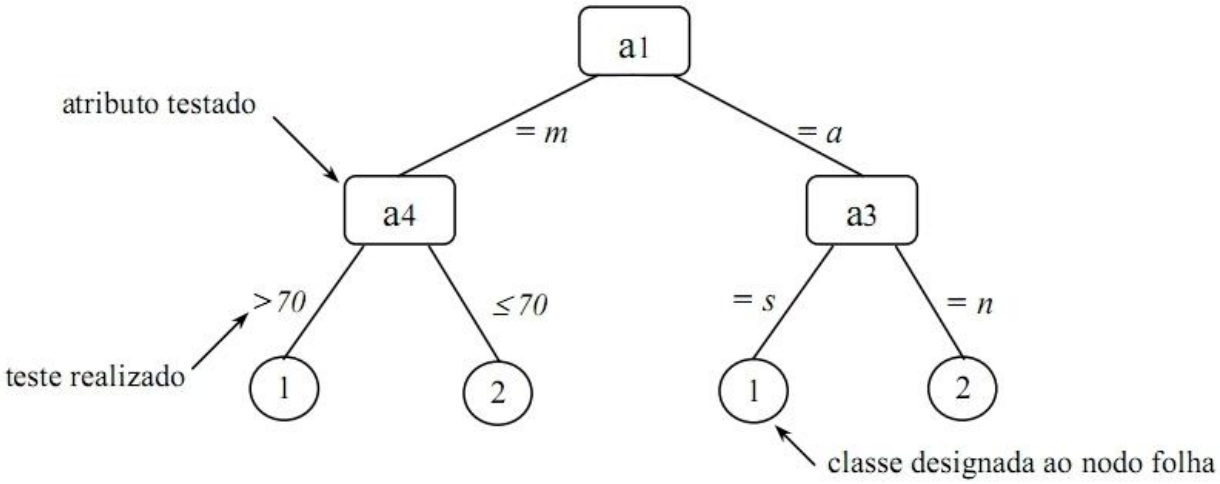


Ilustração 1 – Exemplo de uma Árvore de Decisão

(Fonte: Garcia (2003))

A classificação é representada pelo percurso que vai do nó raiz e se estende até ao nó folha. Por exemplo, num caso em que o atributo *a1* assumisse o valor *m* e o atributo *a4* assumisse um valor maior que 70 seria classificado na *classe 1*. Deste exemplo de classificação poderíamos extrair a seguinte regra: "*SE a1 = m E SE a4 > 70 ENTÃO classe 1*".

a) Algoritmos de indução

O objectivo de uma Árvore de Decisão é operar partições sucessivas num conjunto de dados de treino, até que cada um dos subconjuntos obtidos contenha casos de uma única classe (Petermann, 2006). O modelo obtido servirá para futuras classificações.

Esta definição implica que para induzir Árvores de Decisão é necessário um conjunto de dados de treino previamente classificados (i.e. registo dos quais se conhece a classe da variável dependente, sendo na maior parte dos casos dados históricos). O conjunto de treino deve ter exemplos negativos e positivos de uma classe (Lemos, 2003).

Existem descritos na literatura diversos algoritmos para induzir Árvores de Decisão, entre os mais populares estão os algoritmos CART, CHAID, ID3 ou C4.5.

O Algoritmo CART

O algoritmo CART – *Classification and Regression Trees* – foi durante muito tempo um dos algoritmos de indução de Árvores de Decisão mais populares.

O CART parte de um conjunto de dados pré-classificados em função de uma variável dependente de interesse e induz Árvores binárias (Cortes, 2005). Em cada partição procura isolar o maior número de casos possível que conduzem a um mesmo resultado (o mesmo valor na variável dependente) e que partilham os mesmos valores nos atributos analisados nos níveis anteriores (Cortes, 2005).

A questão é saber qual das variáveis de *input* produz a melhor partição, i.e. a que produz a melhor separação entre classes da variável dependente (Bação, 2006). A medida utilizada pelo CART é a "diversidade" dos valores da variável dependente em cada nó da Árvore.

Para escolher a melhor partição num determinado nó o CART testa todas as variáveis independentes, uma de cada vez, e avalia-as relativamente à diversidade. Finalmente escolhe aquela partição que apresenta uma menor "diversidade" (Bação, 2006).

O Algoritmo CHAID

A maior diferença entre o algoritmo CART e CHAID – *Chi-Squared Automatic Interaction Detector* - está na forma como este tenta travar o crescimento desmesurado da Árvore (Cortes, 2005). O CHAID aplica o teste do Qui-Quadrado, χ^2, para identificar as variáveis independentes a serem usadas para formar as partições da Árvore.

O Algoritmo ID3

O ID3 – *Iteractive Dichtomizer* – foi proposto por Ross Quinlan em 1979 e posteriormente melhorado pelo mesmo autor dando origem ao algoritmo C4.5 em 1995 (Cortes, 2005).

O ID3 suporta Árvores de Decisão com partições não binárias, i.e. um nó pode ter mais de dois ramos (Quinlan, 1986), e introduz os conceitos de "ganho de informação" e "entropia" (Cortes, 2005).

A "entropia" é uma medida que indica a homogeneidade do conjunto de treino utilizados face à função objectivo do problema em estudo (Carvalho, 2005; Cortes, 2005), permite caracterizar a pureza ou impureza do conjunto de dados (Lemos, 2003). Assim uma entropia elevada indica uma distribuição quase uniforme da variável dependente pelo espaço de *input* enquanto uma entropia baixa indica a existência de um valor predominante.

O "ganho de informação" é uma medida de quanto uma dada variável irá separar os casos de treino de acordo com as classes da variável

dependente. É a redução da entropia provocada pela partição dos casos de treino segundo uma dada variável (Carvalho, 2005; Lemos, 2003).

Como principais características deste algoritmo temos (Petermann, 2006):

- espaço de busca de hipóteses (Árvores) é completo, não havendo o risco de a melhor Árvore não se encontrar nesse espaço,

- não realiza *backtracking* na busca pela melhor Árvore; ou seja, uma vez escolhido um atributo num dado nível da Árvore, ele nunca volta a esse nível para alterar essa escolha. Por isso não existe o risco de a solução encontrada ser um óptimo local.

- usa todos os casos do conjunto de treino em cada passo da pesquisa devido à selecção de atributos baseada em medidas estatísticas. Com isso o algoritmo é menos sensível a casos erroneamente classificados e a Não respostas (*Missing Values)*.

O Algoritmo C4.5

O algoritmo C4.5 é, como já referimos, uma evolução natural do algoritmo ID3, tendo sido proposto pelo mesmo autor.

Tal como no algoritmo ID3 o C4.5 permite que em cada nó da Árvore sejam gerados mais do que dois sub-ramos. As Árvores geradas por este algoritmo têm frequentemente mais ramos nos primeiros nós da Árvore do que nos restantes contribuindo assim para uma convergência – chegada a uma folha terminal – mais rápida (Cortes, 2005).

Entre as melhorias introduzidas pelo C4.5 destacam-se, entre outras: capacidade de trabalhar com atributos contínuos, capacidade de trabalhar com valores ausentes e melhor desempenho computacional (Petermann, 2006).

2.2 Redes Neuronais

As redes neuronais têm, nos últimos anos ganho bastante popularidade no seio das soluções de suporte à decisão. Estas simulam o comportamento do cérebro humano (Cheng & Titterington, 1994; Cortes, 2005).

As Redes Neuronais oferecem uma arquitectura bastante potente, com elevadas capacidades de aprendizagem e de representação dos padrões de dados (Petermann, 2006). Apesar de serem um instrumento poderoso, as Redes Neuronais são também uma das ferramentas de *Data Mining* mais difíceis de interpretar para os analistas, com elas o processo de *Data Mining* assemelha-se muitas vezes a uma "Caixa Negra".

a) Elementos constitutivos

Tal como o nosso cérebro é constituído por vários neurónios e ligações entre eles, uma rede neuronal é constituída por várias unidades de processamento designadas por neurónios artificiais e "ligações" (pesos) entre elas (Bação, 2006; Gardner & Droling, 1998). A anatomia e o funcionamento do neurónio artificial são muito semelhantes à de um neurónio natural.

A anatomia de um neurónio natural está representada na figura seguinte:

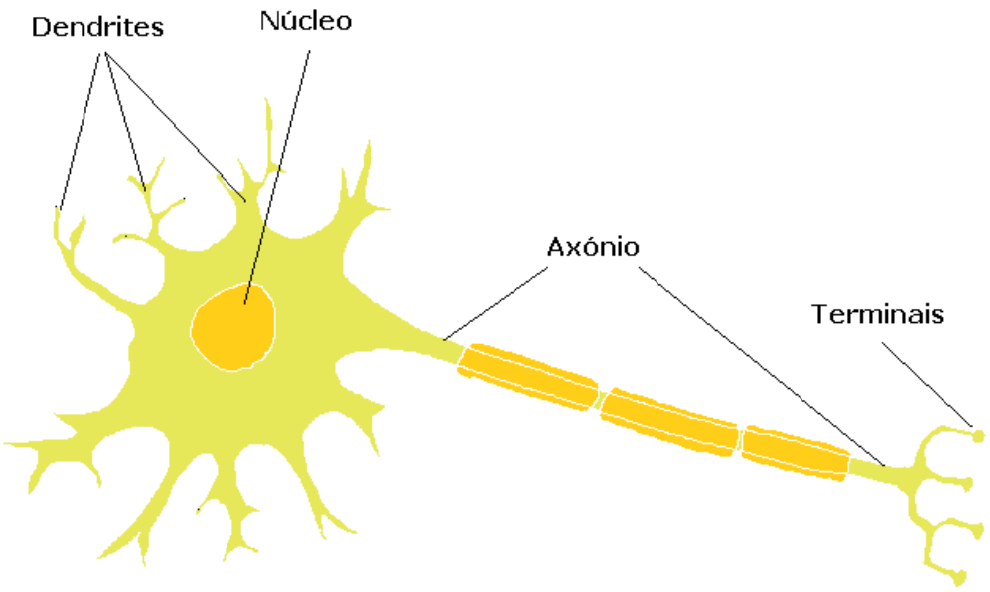


Ilustração 2 - Estrutura de um Neurónio Natural típico

(Fonte: www.schizophrenia.com/sznews/archives/005490.html acedido a 22 de Maio de 2010)

Um Neurónio natural recebe informação dos seus vizinhos através das suas Dendrites. A informação é processada no Núcleo e o *output* é transportado através do Axónio para os Terminais do Neurónio, de onde é passada aos neurónios seguintes caso o valor da estimulação exceda um determinado valor limiar.

A anatomia de um neurónio artificial pode ser vista na figura seguinte:

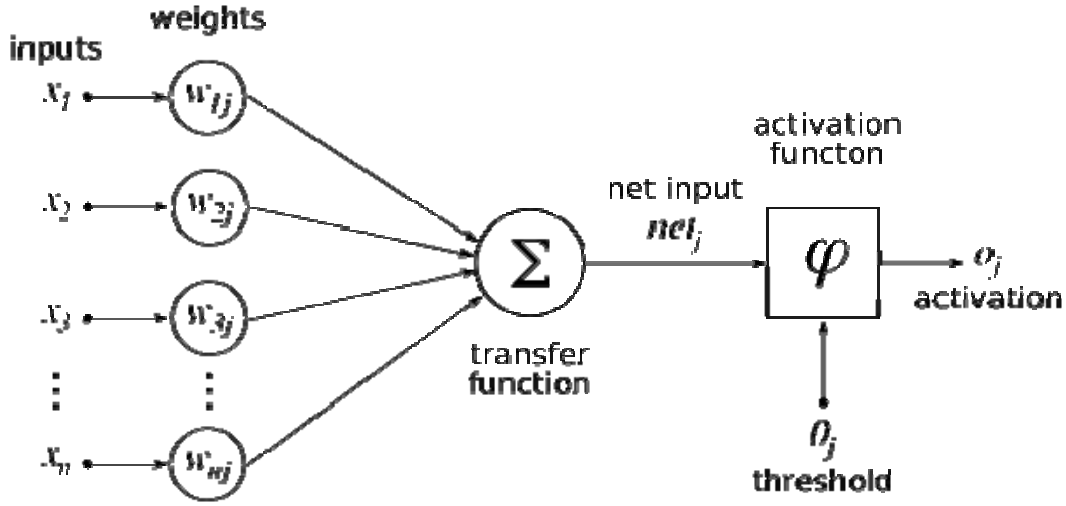


Ilustração 3 - Estrutura de um Neurónio Artificial típico

(Fonte: http://en.labs.wikimedia.org/wiki/Artificial_Neural_Networks acedido a 10 de Junho de 2010)

O neurónio artificial recebe informação externa à rede através dos pesos de recepção (*input*). A informação assim recebida é processada por uma função de combinação, sendo a soma ponderada a mais utilizada para "processar" estes valores.

O valor obtido é comparado com um determinado valor limiar (*Bias*) por uma função designada por função de activação. Se o *input* excede o valor limiar, o neurónio é activado, caso contrário será inibido. Em caso de activação o neurónio envia um *output*, através dos seus pesos de envio, para os neurónios a ele conectados.

b) Arquitectura

Numa rede neuronal artificial os neurónios estão geralmente organizados por camadas. Usualmente trabalha-se com três camadas (Bação, 2006; Petermann, 2006):

- A *camada de input* tem como papel apresentar os valores dos *inputs* à rede. Não realiza qualquer tipo de processamento nem possui qualquer tipo de função de activação.

- A(s) *camada(s) escondida(s)* é a camada que executa a maior parte do processamento e que extrai as características dos dados. Há arquitecturas que assentam em estruturas com várias camadas escondidas, outras não consideram nenhuma camada escondida, estando a camada de *input* directamente ligada à camada de *output*.

- A *camada de output* tem funções de processamento e produz os *outputs* da rede.

Geralmente cada neurónio de uma camada recebe *inputs* de todos os neurónios da camada anterior, com a excepção da camada de *input*.

Cada neurónio envia também outputs para todos os neurónios da camada seguinte, com a excepção da camada de output.

Na literatura estão descritos vários tipos de Redes Neuronais, o que difere entre eles são as ligações formadas e o algoritmo de treino. De uma forma geral os elementos que constituem uma rede e estão sujeitos a modificações são os seguintes (Petermann, 2006):

- Forma de ligação entre as camadas,
- Número de camadas escondidas,
- Quantidade de neurónios em cada camada,
- Função de activação,
- Algoritmo de aprendizagem.

Algoritmo de aprendizagem

Também na forma como aprendem as Redes Neuronais buscam inspiração no cérebro Humano. Para aprender o cérebro humano vai alterando a sua estrutura de conexões entre os neurónios com base na experiência (Bação, 2006).

As redes artificiais embora não consigam alterar a sua estrutura, alteram os pesos das conexões entre os neurónios. Essa alteração é feita, na grande maioria das redes, através do algoritmo de *Retro-propagação do Erro* ou *Backpropagation* (Bação, 2006; Gardner & Droling, 1998).

Este algoritmo compara as previsões da rede com os verdadeiros valores. O erro assim obtido, calculado ao nível da camada de *output*, é depois propagado para as camadas inferiores da rede até à camada de *input* (Bação, 2006).

Tal como descrito na literatura o algoritmo e Retro-propagação do erro funciona da seguinte forma (Gardner & Droling, 1998):

- a rede é inicializada com pesos aleatórios,

- o primeiro registo é apresentado à rede e é calculado o respectivo output,
- é calculado o erro pela diferença entre as previsões da rede e os verdadeiros valores (target),
- o erro é propagado para as camadas inferiores da rede,
- os pesos são alterados de modo a minimizar o erro da rede.

Os novos pesos são resultantes da soma do antigo peso à ponderação Delta. A ponderação Delta obtém-se pela multiplicação do Erro pela Taxa de Aprendizagem (Bação, 2006).

Nesse processo o algoritmo tenta corrigir os pesos que mais contribuíram para o erro obtido. Este processo é iterativo e termina quando os pesos convergem.

c) Perceptrão Multicamada

Existem, muitas arquitecturas de Redes Neuronais: Redes de Hopfield, Self-Organizing Map, Perceptrão, etc.

Sem dúvida que a Redes mais usadas em problemas de classificação de probabilidade de *Churn* são os Perceptrão Multicamada (Multilayer Perceptron - MLP).

Os MLP são uma extensão do modelo do Perceptrão. São constituídos por três ou mais camadas de neurónios: Camada de *Input*, uma ou mais Camadas Escondidas e uma Camada de *Output*.

Os MPL são redes de tipo *feed-forward* (Gardner & Droling, 1998) i.e. a informação circula num único sentido: da camada de *input* para a camada de *output*. São redes de aprendizagem supervisionada, geralmente utilizam como algoritmo de aprendizagem o algoritmo *Retro-propagação do erro* (*Backpropagation).*

Na figura seguinte podemos ver a representação gráfica de uma rede MLP.

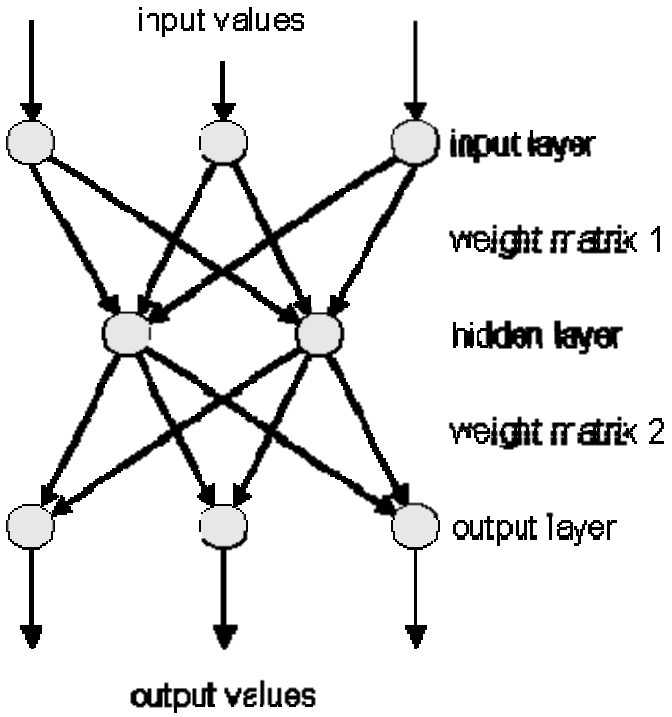


Ilustração 4 - Perceptrão Multicamada

(Fonte: www.fynu.ucl.ac.be/users/c.delaere/level2/MLP/ acedido a 10 de Junho de 2010)

3. Não Respostas

Antes de passamos à apresentação do nosso problema e o objectivo do nosso trabalho optamos por dedicar um capítulo às Não Respostas. Esta opção prende-se com a necessidade de percebemos o impacto que estes podem ter no desenvolvimento e comportamento de modelos de *Data Mining*.

A existência de registos com valores em falta/Não Respostas é transversal ao trabalho científico e de análise estatística. Podemos esperar a existência de Não Respostas sempre que estão envolvidos participantes humanos, como acontece nas Sondagens (Costa, 2000).

As Não Respostas são um desafio para os investigadores e analistas desde o início da pesquisa de campo (Graham, 2009), até porque, os métodos estatísticos convencionais foram concebidos para trabalhar com matrizes de dados completas e não funcionam adequadamente na presença de Não Respostas, pois estes não podem, por exemplo, ser divididos ou multiplicados (Little & Rubin, 1987; Weiss & Indurkhya, 1998).

Este problema pode ter várias origens, desde a avaria de equipamentos, a falhas de registo. No caso de dados que, como os nossos, são provenientes de Sondagens as Não Respostas têm essencialmente origem em recusas por parte dos inquiridos em responder a algumas questões (Costa, 2000).

Na literatura sobre Não Respostas (e.g. Batista & Monard, 2003; Costa, 2000; Graham, 2009; Schafer & Graham, 2002) é comum encontrar as mesmas conceptualizadas em três categorias:

1 – *Missing Completely at Random* (MCAR). Se os casos com valores em falta podem ser pensados como uma amostra aleatória de todos os

registos, então estamos na presença de MCARs. Ocorrem quando a probabilidade de se obter uma Não Resposta para um dado campo da Base de Dados não depende dos valores obtidos nos restantes campos ou do campo com o valor em falta. Ou seja, não depende dos valores conhecidos nem dos valores desconhecidos (Graham, 2009).

As Não Respostas do tipo MCAR são os que nos garantem um maior nível de aleatoriedade. Podem, por isso, ser aplicadas várias técnicas de tratamento sem que sejam introduzidos enviesamentos consideráveis nos dados (Batista & Monard, 2003).

O principal problema com que nos deparamos na presença de MCARs é que os modelos estatísticos perdem potência, contudo a estimação dos parâmetros mantem-se não enviesada (Graham, 2009).

2 – Missing at Random (MAR). Neste caso a probabilidade de ocorrer uma Não Resposta poderá depender dos valores dos restantes campos, mas não do campo com os valores em falta (Batista & Monard, 2003). Ou seja, poderá depender dos valores conhecidos, mas não dos desconhecidos (Graham, 2009).

Estas são Não Respostas condicionais, uma vez controladas pelos dados de que dispomos, quaisquer Não Respostas que continuem a existir são puramente aleatórias (Graham, 2009). Também nestes casos a estimação dos parâmetros não fica afectada, e a análise mantêm-se não enviesada (Graham, 2009).

3 –Missing Not at Random (MAR). Neste caso a probabilidade de ocorrer um Valor em Falta depende do campo com o valor em falta (Batista & Monard, 2003). Este tipo de Não Respostas levam geralmente à obtenção de parâmetros enviesados para os modelos estatísticos que se desejam obter (Graham, 2009).

3.1 Tratamento de Não Respostas

Como já vimos a existência de Não Respostas pode influenciar de forma bastante acentuada os resultados dos modelos de *Data Mining*. É por isso importante aplicar métodos e técnicas de tratamento adequadas a cada matriz de dados.

Podemos encontrar na literatura muitos métodos para tratar Não Respostas e muitas formas de categorizar as mesmas (e.g. Batista & Monard, 2003; Graham, 2009; Little & Rubin, 1987; Roth, 1994; Schafer & Graham, 2002).

Embora não seja objectivo do presente trabalho efectuar um levantamento exaustivo das técnicas de tratamento de Não Respostas apontamos algumas das mais usadas e implementadas nos pacotes estatísticos mais populares.

a) Eliminação

- *Listwise deletion*: É uma das técnicas mais antigas e das mais usadas, encontra-se implementada em quase todos os pacotes estatísticos. Consiste em eliminar todos os casos para os quais não temos informação completa ou seja, elimina todos os casos onde existem Não Respostas.

Um dos problemas desta técnica é que leva frequentemente à estimação enviesada dos parâmetros de interesse. Note-se que os casos completos podem não ser representativos de toda a população (Schafer & Graham, 2002). A literatura sugere que o uso da técnica *listwise deletion* leva a uma perda de potência dos modelos estatísticos motivada pela diminuição da amostra em estudo (e.g. Graham, 2009; Roth, 1994; Schafer & Graham, 2002). No caso da percentagem de casos com Não Respostas ser pequena (menos de 5%), o enviesamento e a perda de potência são mínimos (Schafer & Graham, 2002).

- *Pairwise deletion*: Esta técnica é bastante semelhante à anterior. A principal diferença da mesma prende-se com a forma como elimina os casos com Não respostas.

Se no caso anterior se eliminavam todos os casos com Não Respostas da nossa análise, neste caso apenas se eliminam os casos em que não temos informação para cada análise (Graham, 2009; Little & Rubin, 1987; Schafer & Graham, 2002). Por exemplo se tivemos a calcular uma matriz de correlações entre quatro variáveis, cada uma dessas correlações será calculada com base nos casos em que temos informação para as duas variáveis em questão. No caso da técnica *listwise deletion* usar-se-iam apenas os casos para os quais temos informação para todas as quatro variáveis.

Ao usar esta técnica corremos o risco de introduzir algum enviesamento na nossa análise devido à utilização de um número diferente de dados para o cálculo de cada parâmetro (Schafer & Graham, 2002). Por outro lado esta técnica, devido à mesma razão, torna complicado o cálculo do erro padrão (Graham, 2009; Schafer & Graham, 2002).

b) Imputação

Nesta categoria encontramos uma série de métodos destinados a substituir os Valores em Falta por valores estimados plausíveis.

A imputação traz-nos grandes vantagens relativamente às técnicas de eliminação, nomeadamente (Schafer & Graham, 2002):

- Permite que não sejam sacrificados casos, evitando assim a perda de poder estatísticos das análises.

- se a informação dos restantes casos permitir estimar os dados em falta, então estas técnicas podem fazer uso dessa informação afim de produzir uma matriz de dados completa que pode ser, posteriormente, analisada com recursos a métodos *standard*.

Entre as técnicas de imputação destacamos:

– Substituição pela média: Consiste em substituir cada valor em falta pela média do respectivo atributo.

Ao aplicar esta técnica a média da variável mantem-se, contudo muitos autores apontam que este método reduz a variância e distorce as co-variâncias e inter-correlações entre as variáveis (e.g. Schafer & Graham, 2002).

– *Hot-Deck e Cold-Deck*: Mais desejável que manter a média de uma dada variável seria o de manter a sua distribuição. Na literatura das Sondagens a técnica *Hot-Deck* tornou-se muito popular por ser mais efectiva a preservar a distribuição das variáveis (Roth, 1994; Schafer & Graham, 2002). A ideia básica desta técnica é a de substituir os valores em falta pelos valores obtidos em registos semelhantes na Base de Dados. Na técnica de *Cold-Deck* os valores a imputar são obtidos de outra Base de Dados que não aquela onde os mesmos irão ser imputados (Roth, 1994), por exemplo uma vaga anterior da sondagem (Little & Rubin, 1987).

– Imputação por regressão: As varáveis numa Base de Dados possuem geralmente relações entre si, que podem servir de base para estimar valores a imputar.

Este método consiste em criar modelos para estimar os valores que vão substituir as Não Respostas. Os modelos são criados com base nas observações sobre as quais temos informação, ou seja para as quais não

temos Não Respostas servindo a variável com Não Respostas como variável de dependente (Batista & Monard, 2003). Após o modelo estar terminado é usado para prever os valores em falta da nossa variável de interesse (Schafer & Graham, 2002).

Uma importante limitação desta abordagem relaciona-se com o comportamento dos dados gerados. Geralmente as estimativas obtidas por estes métodos são melhor comportadas que os valores "reais". Naturalmente se usamos um dados conjunto de atributos para estimar um dado valor, ele será mais consistente com esse mesmo conjunto de atributos do que aquilo que o valor real (desconhecido) seria. Outra limitação importante é que este método exige a presença de relações entre os vários atributos da Base de Dados e o atributo com valores falta. Caso tal não aconteça estes modelos não serão capazes de gerar previsões plausíveis dos valores em falta, podendo até ter um efeito nefasto (Batista & Monard, 2003).

- Imputação Múltipla (Multiple Imputation - MI): Num cenário em que usamos uma regressão para imputar valores às Não Respostas, como descrito anteriormente, perdemos variância ao nível do erro (Graham, 2009; Graham, Hofer, Donaldson, McKinnon, & Schafer, 1997) já que neste cenário os valores preditos pela regressão estão sempre assentes sobre a recta enquanto os valores reais se afastam sempre, ainda que ligeiramente, da mesma (termo residual).

De forma a restaurar esta perda de variância esta técnica vai, num primeiro passo, adicionar um termo residual aleatório aos valores a imputar (Graham, 2009). Se assumimos que a distribuição do termo residual dos dados existentes também descreve a distribuição do termo residual das Não Respostas então podemos seleccionar aleatoriamente um elemento da distribuição de termos residuais dos dados existentes e adicioná-lo a cada valor imputado (Graham *et al.*, 1997).

De notar que também existe perda de variância relacionada com o facto de cada valor a imputar ser oriundo de uma única equação de regressão (Graham *et al.*, 1997). Uma forma de ultrapassar esta limitação seria obter múltiplas amostras da população o que não é possível. Para simular tal procedimento podem-se aplicar técnicas de *bootstraping* à Base de Dados (Graham, 2009; Graham *et al.*, 1997).

Um dos aspectos mais importantes nos modelos de *Data Mining* é a qualidade dos dados. No mundo real a qualidade dos dados é frequentemente ameaçada pela existência de elevadas taxas de Não Respostas (*Missing Values*).

A forma como se lida com as Não Respostas influencia de forma muito acentuada os resultados obtidos pelos algoritmos. As Não Respostas devem ser alvo de um tratamento cuidado, caso contrário o conhecimento produzido poderá ser afectado de enviesamento (Batista & Monard, 2003).

4. Problema e Objectivos

Este trabalho foi realizado no âmbito de um projecto de desenvolvimento de um novo produto para a empresa de Sondagens e estudos de opinião *Marktest*. Este projecto relaciona-se com a crescente necessidade das companhias em assentarem as suas tomadas de decisão e acções em dados objectivos. Assim estas procuram, cada vez mais, capitalizar os dados de que dispõem de modo a gerar o máximo valor para o seu negócio e reduzir o risco.

Recorrendo a dados dos seus estudos regulares "Barómetro de Seguros" e "Transferências de Seguros", a *Marktest* pretende gerar modelos estatísticos que permitam aos seus clientes ter *insights* sobre os comportamentos e percepções dos segurados Portugueses e, dessa forma, tomarem decisões mais acertadas.

Estando consciente da crescente necessidade que as Companhias têm de fazer uma gestão pró-activa do *Churn*, o novo produto, a disponibilizar aos clientes da *Marktest* - para uso efectivo em tomada de decisão - consiste no desenvolvimento de modelos de classificação de probabilidade de *Churn* para os possuidores de Seguro Automóvel. Este modelo será um complemento ao portfólio de produtos que a *Marktest* oferece aos seus clientes deste sector de actividade.

De modo a ter a máxima utilidade e valor possível para as seguradoras, e em linha com a literatura (e.g. Hadden et al., 2005), pretende-se que os modelos sejam capazes de identificar como potenciais *Churners* grande parte dos mesmos e que simultaneamente sejam capazes de gerar um número reduzido de Falsos Positivos.

No contexto específico deste projecto definimos *Churn* como a mudança de seguradora onde possui o seguro do, sendo considerado o automóvel conduzido com mais frequência no último ano. Neste sentido são

Churners todos os inquiridos que declararam ter mudado a apólice referente ao seguro do automóvel que conduzem com mais frequência de companhia durante o último ano.

Neste segundo capítulo fizemos um enquadramento do tema em estudo – o *Churn*. Apresentámos uma definição para o mesmo, o seu impacto económico para as companhias e apontámos a pertinência da classificação de probabilidade de *Churn* no contexto de uma gestão pró-activa do mesmo.

Posteriormente descrevemos os algoritmos mais usados em problemas de classificação de probabilidade de *Churn*: as Árvores de Decisão e as Redes Neuronais. Abordamos igualmente a questão das Não Respostas concluindo que a forma como lidamos com elas pode influenciar os nossos resultados. Foram, ainda, revelados o problema e os objectivos.

No próximo capítulo iremos abordar as questões metodológicas deste estudo: caracterizamos as nossas bases de dados e fazemos uma breve análise descritiva das mesmas, relatamos os procedimentos executados para a realização deste trabalho e listamos as Ferramentas informáticas utilizadas nas várias fases da realização deste trabalho.

III Método

Este capítulo apresenta a metodologia seguida, assim como as Ferramentas informáticas em que esta se apoiou para dar resposta aos objectivos traçados.

No primeiro ponto são apresentadas as Bases de Dados de que dispomos para a realização deste projecto. É igualmente efectuada uma pequena introdução metodológica às sondagens que estão na origem dos nossos dados.

No segundo são apresentadas as diferentes fases metodológicas seguidas ao longo da realização deste projecto. Esta tem como pontos principais o levantamento de necessidades, a identificação dos dados a utilizar, a definição dos objectivos do projecto, obtenção de autorização formal para a utilização dos dados, extracção dos dados e desenvolvimento dos modelos.

No terceiro ponto identificamos as Ferramentas informáticas que utilizamos para executar as várias tarefas inerentes a este projecto.

1. Caracterização da Base de Dados

Os dados de que dispomos para o nosso projecto são originários de duas sondagens sobre seguros, realizados a nível nacional (Portugal continental).

Um dos inquéritos – designado por "Barómetro de Seguros" - visa avaliar os comportamentos e percepções dos Portugueses face aos seguros, o outro – designado por "Transferências de Seguros" - tem como objectivo avaliar/quantificar as taxas de transferências inter-companhias/*Churn* existentes no Seguro Automóvel e os motivos associados a essa mudança.

1.1 O Barómetro de Seguros

Os dados do "Barómetro de Seguros" são recolhidos mensalmente (10 meses por ano). São recolhidas cerca de 1.000 entrevistas por mês através do método de entrevista telefónica assistida por computador – CATI - resultando numa amostra de cerca de 10 mil entrevistas por ano.

Os lares a inquirir são seleccionados de forma aleatória. A selecção do entrevistado, um em cada lar, é efectuada através do método de quotas, tendo em consideração as variáveis sexo, idade e distrito de residência. O Barómetro de Seguros consiste em indicadores respeitantes a:

- Notoriedade de Companhias de Seguros,
- Imagem das Companhias de Seguros,
- Seguro de Saúde,
- Seguro de Habitação,
- Seguro de Automóvel,
- Seguro de Acidentes Pessoais,
- Seguro de Acidentes de Trabalho,
- Seguro de Animais Domésticos,
- Seguros do Ramo Vida (Risco, Misto, PPR/E e Capitalização).

No que respeita ao Seguro Automóvel são recolhidos dados diversos entre os quais: o número de apólices, as companhias a que pertencem as apólices e companhia a que pertence a apólice do veículo que utiliza com mais frequência.

Posteriormente é pedido ao inquirido para se centrar na apólice do veículo que utiliza com mais frequência e responder a várias questões, entre as quais algumas respeitantes a características da apólice, nomeadamente: capital de responsabilidade civil, coberturas extra, prémio anual, número de sinistros, tempo de posse desta apólice, nome em que está a apólice (próprio, familiares, companhia), etc. Dispomos ainda de variáveis sociodemográficas.

Uma listagem exaustiva das variáveis disponíveis nesta Base de Dados encontra-se no Anexo A.

Os dados de que dispomos são respeitantes ao ano de 2006. A Base de Dados conta com cerca de 6.800 entrevistas respeitantes a possuidores/beneficiários de Seguro Automóvel.

1.2 Estudo de Transferências de Seguros

Anualmente uma amostra de entrevistados do Barómetro de Seguros, do ano anterior, volta a ser contactada para uma segunda entrevista. Nesta procura-se aferir se o inquirido mudou ou não a companhia onde possui o Seguro do Automóvel que utilizou com mais frequência durante esse período. Esta Base de Dados é constituída por cerca de 2.5 mil registos. Dispomos dos dados de 2007, ou seja, dos mesmos inquiridos do "Barómetro de Seguros" de 2006.

A principal variável disponível prende-se com a mudança (*Churner*), ou não (Não *Churner*), da companhia onde possui o Seguro do Automóvel que utiliza com mais frequência (variável binária) durante o último ano.

Consideram-se *Churners* os inquiridos que declaram ter mudado de seguradora, no que respeita ao seguro do automóvel que conduzem com mais frequência, no último ano (de 2006 para 2007).

Uma listagem exaustiva das variáveis constantes nesta Base de Dados pode ser encontrada no Anexo B.

A Marktest autorizou, formalmente, a utilização dos dados do estudo "Barómetro Seguros" de 2006 e "Transferências de Seguros" do ano de 2007 para a realização do presente projecto (c.f. Anexo C).

1.3 Base de Dados de Treino

A Base de Dados de treino para os nossos modelos de *Churn* resulta da união das duas Bases de Dados citadas anteriormente ("Barómetro de Seguros" e "Transferências de Seguros"). A ligação destas Bases de Dados foi feita em Microsoft Access fazendo corresponder a chave primária da Base de Dados do "Barómetro de Seguros" (Número de Entrevista) com a respectiva chave estrangeira na Base de Dados de "Transferências de Seguros".

Devido à natureza dos dados, do problema e do objectivo do nosso projecto decidimos usar unicamente os dados respeitantes às apólices de particulares, que habitualmente tomam as decisões respeitantes ao Seguro Automóvel e cuja apólice está em nome próprio.

A utilização destes filtros destina-se a garantir que as características do entrevistado são as mesmas do tomador de seguro. Desta forma procuramos assegurar que a ocorrência, ou não, de *Churn*, dependa das

características, percepções do entrevistado, bem como da sua relação com a seguradora e não das características e percepções de terceiros (e.g. pais, sogros, filhos, marido, entidade patronal, etc.).

Após a aplicação destes filtros ficamos com uma Base de Dados de treino com 1.661 registos.

1.4 Base de Dados de Classificação

Dado que apenas uma amostra dos inquiridos do "Barómetro de Seguros" é recontactada no âmbito do estudo de "Transferências de Seguros" reunimos os inquiridos que não foram recontactados para o estudo "Transferências de Seguros", que possuem/beneficiam de apólices de particulares, que habitualmente tomam decisões respeitantes ao Seguro Automóvel e cujo seguro está em nome próprio numa quarta Base de Dados. Designamos essa Base de Dados por Base de Dados de classificação. Esta é constituída por 2.573 registos.

1.5 Análise descritiva dos dados

Segue-se uma breve caracterização das principais variáveis através de uma análise descritiva e gráfica (c.f. Anexo D).

MUDANÇA DE SEGURADORA – o inquirido mudou ou não de seguradora onde possui o seguro do automóvel que utiliza com mais frequência.

Casos válidos: 1.661

Cerca de 4/5 (80%) dos inquiridos não mudou de seguradora onde possui o seguro do automóvel que utiliza com mais frequência.

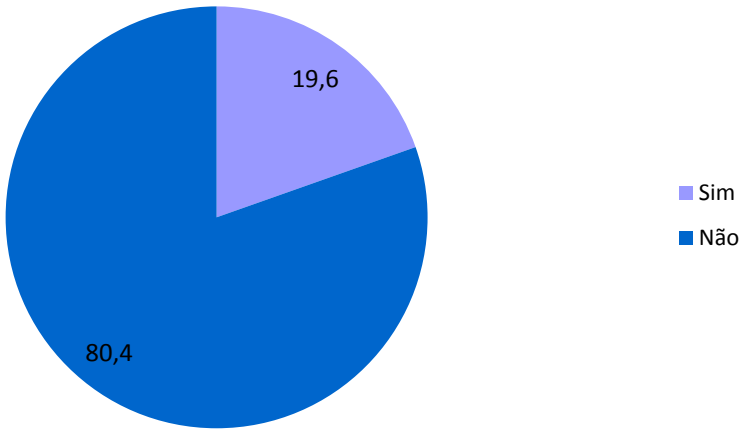


Gráfico 2 - Mudança de seguradora

GÉNERO – Género do inquirido

A Base de Dados é, na sua maioria, constituída por homens. Entre os *Churner*s o género masculino ganha ainda mais predominância.

Casos válidos: 1.661

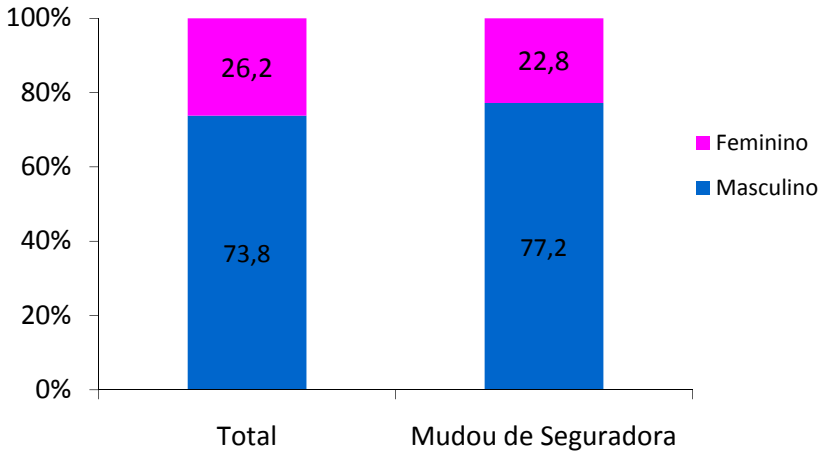


Gráfico 3 - Género do inquirido

IDADE – Idade do inquirido

Ao nível etário a nossa Base de Dados é constituída por indivíduos com idades compeendidas entre os 18 e 79 anos.

Casos válidos: 1.661

Em termos médios os inquridos apresentam uma idade de 47,22 anos (d.p.=14,7). Entre os *Churner*s encontramos uma idade média de 45,4 anos (d.p.=14,3).

O gráfico abaixo sugere que o comportamento de *Churn* tem maior probabilidade de ocorrer nas faxas etárias mais jovens.

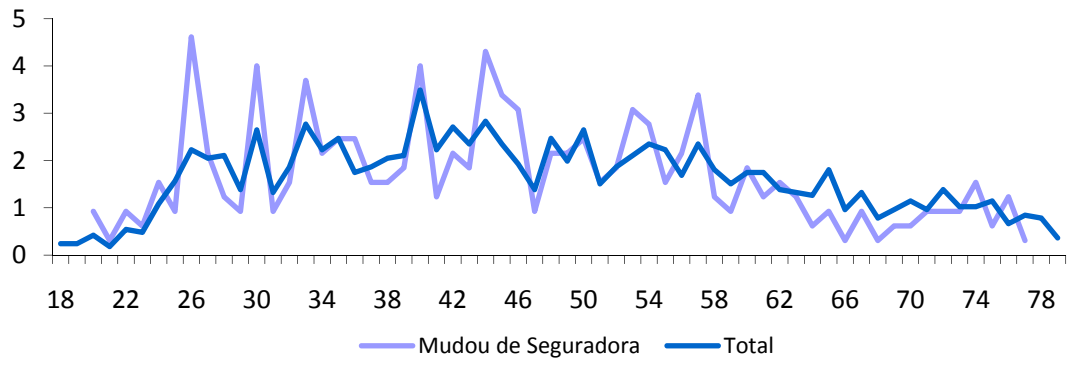


Gráfico 4 - Idade do inquirido

NÚMERO DE SEGUROS AUTOMÓVEL – Número de Seguros automóvel que o inquirido possui

Casos válidos: 1.661

No que respeita ao Seguro Automóvel, 2/3 dos inquiridos detêm apenas uma apólice (Média=1,45; d.p.=0,75). Os *Churner*s possuem um número de apólises idêntico (Média=1,49; d.p.=0,78).

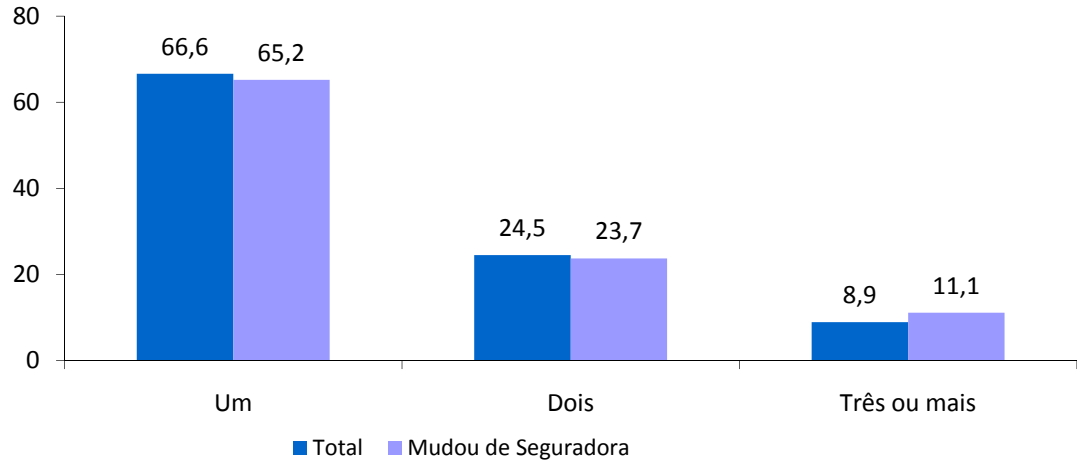


Gráfico 5 - Número de seguros automóvel

COBERTURAS – Coberturas garantidas pelo Seguro Automóvel

Cerca de metade das apólices em análise dizem apenas respeito às coberturas minímas obrigatórias, quer num grupo quer no outro.

Casos válidos: 1.630; Casos com Não Respostas*: 31*

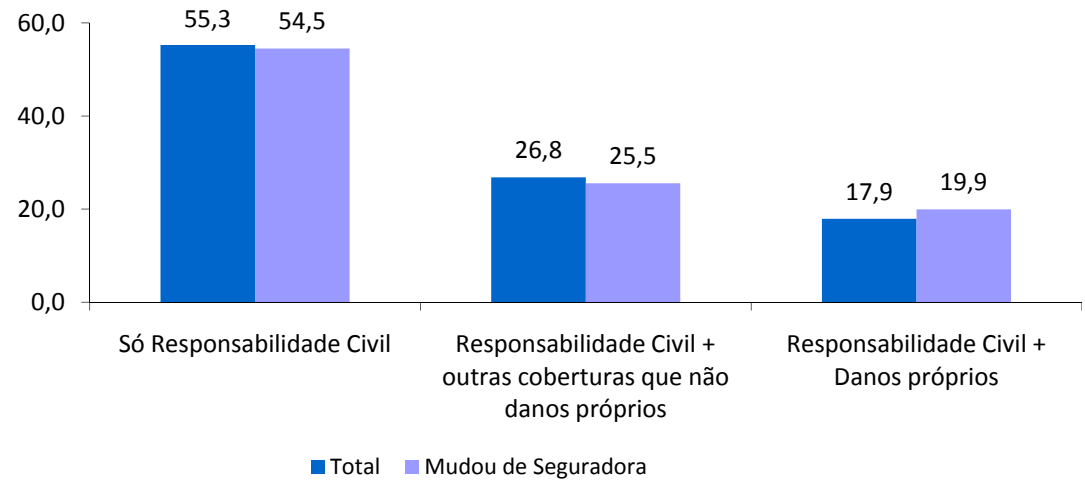


Gráfico 6 - Coberturas garantidas

PARTICIPAÇÃO DE SINISTRO – Participação de sinistros relativos à apólice em análise

Casos válidos: 1.543; Casos com Não Respostas*: 118*

Cerca de 75% dos inquiridos não participou qualquer sinistro relativo à apólice em análise.

É visível que existe uma percentagem ligeiramente superior de *Churner*s que participaram sinistros quando comparados com o total da nossa Base de Dados.

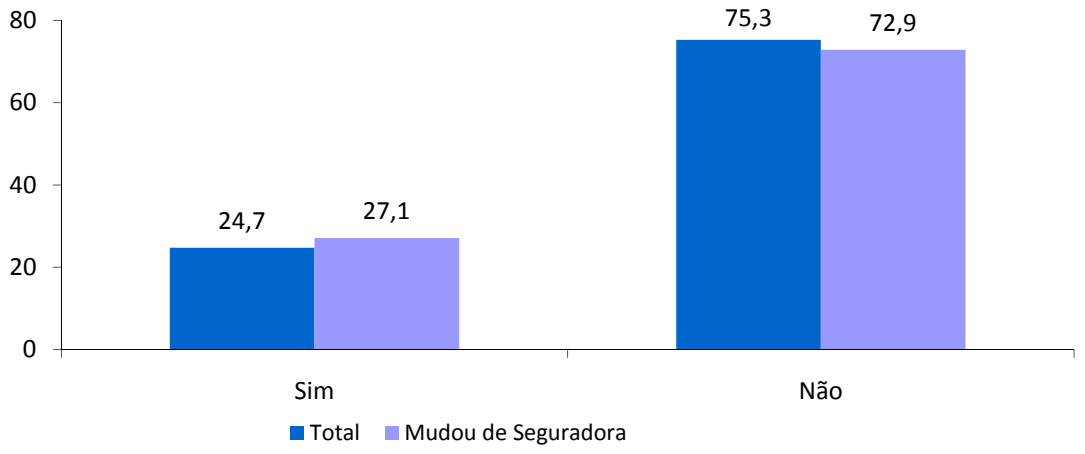


Gráfico 7 - Participação de sinistro

MUDOU ALGUMA SEGURADORA NO ÚLTIMO ANO – No último ano deixou de trabalhar com alguma Companhia de Seguros.

Casos válidos: 1.659; Casos com Não Respostas*: 2*

Mais de 90% dos inquiridos não abandonou qualquer Companhia ao longo dos último 12 meses.

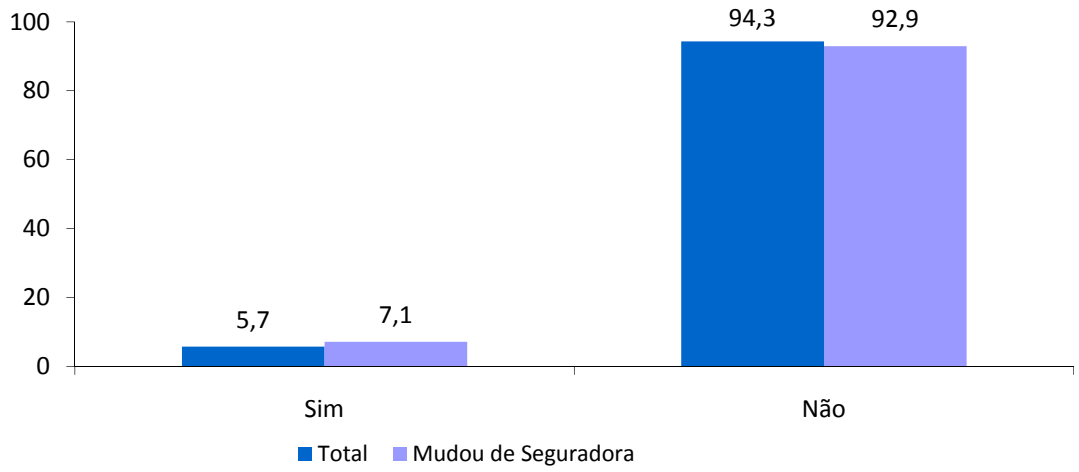


Gráfico 8 - deixou de trabalhar com alguma seguradora

PRÉMIO – Prémio anual do Seguro Automóvel.

Cerca de 40% dos inquiridos paga entre 101 a 250 euros pelo seu Seguro Automóvel. Os *Churner*s estão ligeiramente mais concentrados nas categorias mais elevadas do que a totalidade dos inquiridos.

Casos válidos: 1.361; Casos com Não Respostas*: 300*

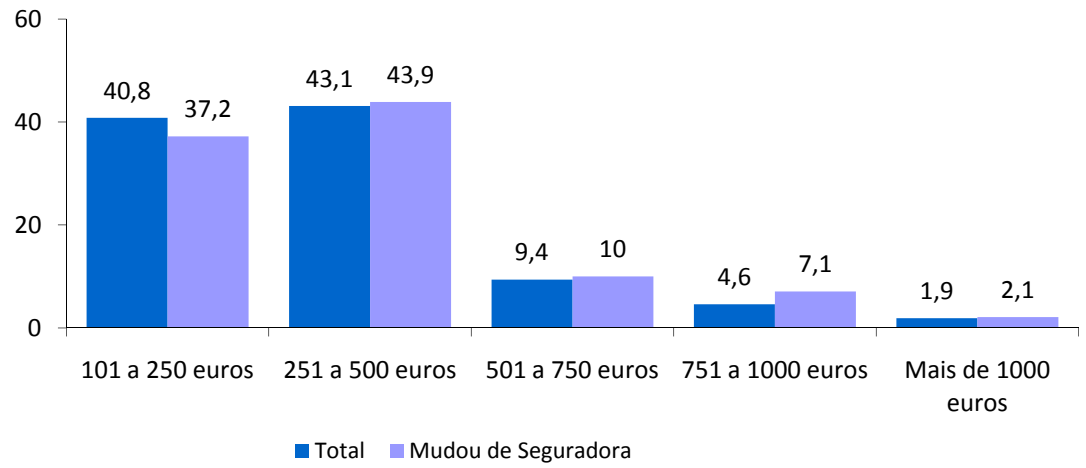


Gráfico 9 - Prémio anual do Seguro Automóvel

Para responder melhor aos objectivos propostos as variáveis constantes do Barómetro de Seguros foram enriquecidas. Através da transformação de algumas variáveis e do cálculo de novos indicadores, foi possível uma melhor interpretação do comportamento dos inquiridos.

Nesse sentido construímos novas variáveis, as principais encontram-se listadas na tabela seguinte:

Nome	Descrição	Tipo de variável
MAIS_CONTACTO	A seguradora do automóvel que utiliza com mais frequência é aquela com que tem mais contacto.	Binária (0-Não; 1-Sim)
PREMIO_250	Prémio do Seguro Automóvel.	Binária (0-Até 250 Euros; 1- Mais de 250 Euros)
COB_EXTRA	A apólice do inquirido inclui coberturas extra (seguro de viagem, danos próprios, etc.).	Binária (0-Não; 1-Sim)
N_AUTO	Número de Seguros Automóvel	Binária (0- Um Seguro Automóvel; 1- Mais de um Seguro)

Nome	Descrição	Tipo de variável
CLIENTE_EXCLUSIVO	Cliente Exclusivo ao nível do Seguro Automóvel.	Binária (0-Não é Cliente Exclusivo; 1- Cliente Exclusivo)
HABITACAO	Possui um seguro de habitação na mesma companhia que o seguro do automóvel que utiliza com mais frequência.	Binária (0-Não; 1-Sim)
TOTAL	Número de produtos diferentes que possui na mesma companhia que o seguro do automóvel que utiliza com mais frequência.	Numérica
VIDA	Número de produtos diferentes, do ramo vida, que possui na mesma companhia que o seguro do automóvel que utiliza com mais frequência.	Numérica

Tabela 1 - Variáveis adicionais

No Anexo E pode ser encontrada uma listagem mais exaustiva das várias Variáveis/Indicadores que criámos.

2. Metodologia

Numa fase exploratória, e no âmbito da apresentação de um projecto como requisito parcial para a obtenção do grau Mestre em Estatística e Gestão de Informação, abordamos a *Marktest* no sentido de aferir a sua disponibilidade para facultar dados para a realização do mesmo. A *Marktest* mostrou disponibilidade para colaborar connosco.

Numa primeira fase metodológica foi feito um levantamento dos vários produtos da *Marktest,* bem como das características e Metodologias dos mesmos. Tendo em atenção os interesses da *Marktest* e do Autor, ambas as partes concordaram em utilizar dados provenientes do Barómetro de Seguros da *Marktest*.

Numa segunda fase foi feito um levantamento dos indicadores presentes no Barómetro de Seguros e estudos associados (Transferências de Seguros). Foram apresentadas à *Marktest* várias propostas de estudos. Tendo em conta as necessidades dos seus clientes e o seu portfólio de produtos a *Marktest* optou pela proposta de desenvolvimento de modelos de classificação de probabilidade de *Churn* para o Seguro Automóvel.

Na terceira fase foi pedida autorização formal de acesso aos dados dos estudos "Barómetro de Seguros" e "Transferências de Seguros".

A *Marktest* deu autorização formal para a utilização dos dados do estudo "Barómetro Seguros" do ano de 2006 e "Transferências de Seguros" do ano de 2007 para a realização do presente projecto e a sua apresentação como requisito parcial para a obtenção do grau Mestre em Estatística e Gestão de Informação.

Numa fase posterior os dados foram extraídos do sistema informático próprio da Marktest e importados em Microsoft Excel. Após importação

com sucesso os dados foram tratados em Excel e Access de modo a serem utilizáveis em algoritmos de *Data Mining*.

Por último foram desenvolvidos os modelos de *Data Mining* para classificação de probabilidade de *Churn* no Seguro Automóvel.

3. Ferramentas

A realização deste projecto envolveu, nas suas diferentes fases e para as suas diferentes tarefas, a utilização de várias ferramentas informáticas, a saber:

- *Software* de Base de Dados – *Microsoft Access*,
- *Software* de análise estatística – *SAS Enterprise Guide e SPSS*,
- *Software* de *Data Mining* - *SAS Enterprise Miner*,
- *Software* de apoio ao cálculo e análise – *Microsoft Excel*,
- Software de Econometria - *GRETL*

Neste terceiro capítulo apresentámos a metodologia seguida. Começámos por caracterizar e descrever as nossas Bases de Dados e foi feita uma pequena introdução metodológica aos estudos que estão na origem destes dados. Posteriormente foram apresentadas as diferentes fases metodológicas seguidas e as aplicações informáticas usadas.

No próximo capítulo apresentaremos os nossos modelos e as suas características, Taxas de Erro, Matrizes de Confusão, Gráficos de *Lift*, Gráficos de Ganhos Cumulativos, etc. apresentaremos, ainda, os resultados de uma Simulação de Monte Carlo efectuada com todos os modelos desenvolvidos e uma proposta de *Scoring* e de *Rating* dos inquiridos que não foram incluídos na nossa Base de Dados (amostra) de treino.

IV Análise de dados e Resultados

Neste capítulo começamos, na primeira parte, por descrever o trabalho de junção das duas Bases de Dados de que disponhamos numa única Base de Dados. De seguida descrevemos a forma como os dados foram particionados no conjunto de treino, validação e teste. Discutimos os resultados das primeiras especificações dos modelos e a necessidade de fazer *Boosting* à nossa Base de Dados de Treino.

Posteriormente apresentamos as características (Taxas de Erro, Matrizes de Confusão, Gráficos de *Lift*, Gráficos de Ganhos Cumulativos, etc.) dos modelos desenvolvidos: Árvore de Decisão, Rede Neuronal e, adicionalmente, Probit.

Na terceira parte apresentaremos ainda os resultados de uma simulação de Monte Carlo efectuada com todos os modelos desenvolvidos. O Método de Monte Carlo (Metropolis & Ulam, 1949) consiste em gerar aleatoriamente um grande número de cenários prováveis (e.g. as *seeds* iniciais dos nossos modelos) com a finalidade de determinar as propriedades estatísticas das variáveis influenciadas por esses cenários (e.g. as taxas de erro dos nossos modelos).

Na quarta parte apresentamos uma proposta de *Scoring* dos inquiridos que não foram incluídos na nossa Base de Dados de Treino. São usados os três modelos desenvolvidos bem como a "união" (*ensemble*) destes.

Para fechar este capítulo apresentamos uma proposta de *Rating* dos inquiridos, em termos de Risco de *Churn* baseada na informação proveniente dos nossos modelos.

Os nossos dados dizem respeito a indivíduos com Seguro Automóvel particular, que habitualmente tomam as decisões respeitantes ao Seguro Automóvel e cujo seguro está em nome próprio (n=1661). Estes dados estavam inicialmente dispersos por duas Bases de Dados diferentes: uma referente ao estudo "Barómetro de Seguros" e outra ao estudo "Transferências de Seguros". As duas Bases de Dados foram unidas - fazendo corresponder a chave primária da primeira com a correspondente chave estrangeira da segunda – dando assim origem à nossa Base de Dados para treino dos modelos (Base de Dados de Treino). Esta Base de Dados foi ainda enriquecida com o cálculo de indicadores adicionais (c.f. Anexo E). Durante este processo muitas variáveis necessitaram de ser re-codificadas/transformadas de modo a serem passíveis de utilização em algoritmos de *Data Mining*.

Após este trabalho de preparação dos dados iniciamos o desenvolvimento dos modelos de *Churn*.

Como é habitual em *Data Mining*, a Base de Dados foi dividida em vários conjuntos de dados: Conjunto de Treino (60%) e Conjunto de Validação (40%). Nesta fase inicial do trabalho de modelação optamos por não considerar um Conjunto de Teste de modo a dispomos de mais observações/registos no Conjunto de Treino, numa fase mais avançada – e como descrito em pormenor mais à frente neste capítulo - iremos considerar também um Conjunto de Teste de modo a avaliar o desempenho dos modelos em dados desconhecidos.

A distribuição dos registos pelos dois conjuntos considerados foi estratificada de modo a obter-se uma proporção idêntica de *Churner*s e *Não Churners* em todos os conjuntos de dados. Esta distribuição foi efectuada recorrendo ao nó *Data Partition* do *SAS.*

As ferramentas de classificação mais utilizadas para resolver problemas semelhantes ao nosso são as Redes Neuronais e Árvores de Decisão (e.g.

Hadden et al., 2005). Centramos por isso a nossa atenção nestas ferramentas.

Como é comum em estudos de mercado, a nossa Base de Dados apresenta um número considerável de Não Respostas (*Missing Values)*. Na literatura (e.g. Batista & Monard, 2003; Schafer & Graham, 2002) existe evidência que a forma como se lida com as Não Respostas é determinante para os resultados dos algoritmos desenvolvidos. Assim quisemos igualmente ver o comportamento dos modelos desenvolvidos perante várias técnicas de tratamento de Não Respostas. Nomeadamente Substituição Baseada na Distribuição (*Hot-Deck*) e Substituição pelo Valor de Tendência Central (Média/Moda), todas elas técnicas disponíveis no nó de *Replacement* do SAS.

Após um amplo trabalho com as duas ferramentas os resultados obtidos foram fracos. Os melhores modelos a que chegámos, com cada uma das ferramentas, apresentavam taxas de erro na ordem dos 20%. Apesar de das taxas de erro serem baixas estas não reflectem necessariamente uma boa qualidade do modelo.

Note-se que estas taxas de erro são semelhantes à percentagem de *Churner*s presentes na nossa Base de Dados de Treino (cerca de 20%). A análise das matrizes de confusão sugere que os modelos têm tendência para classificar quase todos os inquiridos como Não *Churners*, acertando desta forma em cerca de 80% dos casos. Já a análise do gráfico de *Lift* revela que estes modelos têm curvas de *Lift* bastante fracas, na ordem de 1.

Este cenário sugere que os modelos nos trazem um ganho de conhecimento/informação nulo. A utilidade destes modelos para as companhias é muito reduzida uma vez que uma grande parte dos potenciais *Churners* não é devidamente identificado pelos modelos. No contexto de uma gestão pró-activa do *Churn* este facto não permitiria às

companhias, por exemplo, definir um grupo alvo para as suas campanhas de retenção.

Estes resultados mantiveram-se perante várias especificações dos modelos.

Perante este cenário, e com o objectivo de melhorar a qualidade do sinal presente nos dados, aplicamos à nossa Base de Dados a técnica de *Boosting*. Desta forma multiplicamos os casos negativos (Mudou de seguradora) até a sua frequência ser semelhante à dos casos positivos (Não Mudou de seguradora).

Foi assim necessário quadruplicar os casos negativos na nossa Base de Dados, dando origem a uma nova Base de Dados (que designamos por Base de Dados Boosted) com 2642 registos. A distribuição desta nova Base de Dados relativamente à mudança de seguradora é a seguinte:

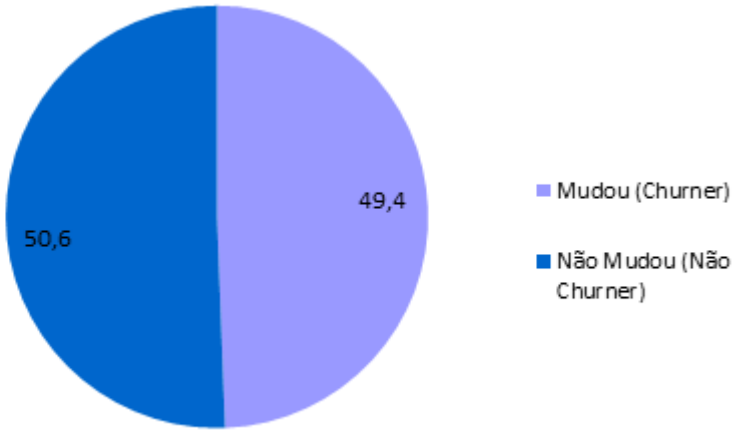


Gráfico 10 - Distribuição dos casos positivos e negativos na Base de Dados após *Boosting*

Sobre esta Base de Dados iniciamos novamente o trabalho de modelação de *Churn*.

No processo de desenvolvimento do modelo preditivo mais eficaz, foram feitas várias experiências, introduzindo e retirando algumas variáveis e

alterando alguns parâmetros dos algoritmos (técnicas de treino, critérios de decisão, etc.).

A título de exemplo na Árvore de Decisão experimentamos:

- *Splitting Criterion*: Chi-Square test; Entropy reduction; Gini reduction.
- *Observations required for a split search*: 15, 20 e 25.
- *Minimum number of observations on a leaf*: 5, 10 e 15.
- *Maximum depth of tree*: 5 a 9.
- *Model Assessment Measure*: Proportion misclassified; Total leaf impurity.

Na Rede Neuronal experimentamos, entre outros:

- *Hidden Units*: 3, 6 e 9.
- *Hidden Layers*: 1 e 2.
- *Activation Function*: Gauss; No Activation; Logistic; Exponential; Hyperbolic Tangent e Default
- *Combination Function*: Linear, Additive, Radial-General e Default.
- *Bias*: Bias e No Bias.
- *Training Technique*: Standard Backprop; Incremental Backprop; Quasi-Newton; Levenberg-Marquardt; Trust Region e Default.

No início do desenvolvimento dos modelos distribuímos os nossos dados da seguinte forma: Conjunto de Treino (60%, n=1585) e Conjunto de Validação (40%, n=1057). Nesta fase optamos por não considerar um Conjunto de Teste de modo a dispomos de mais observações/registos no Conjunto de Treino.

Numa fase final do desenvolvimento dos mesmos alterámos a distribuição dos dados pelos diferentes conjuntos para que pudéssemos aferir o desempenho dos modelos em dados desconhecidos. Introduzimos assim um Conjunto de Teste. A nova distribuição dos dados foi a

seguinte: Conjunto de Treino (60%, n=1585), Conjunto de Validação (20%, n=528) e Conjunto de Teste (20%, n=529). A distribuição dos registos pelos conjuntos foi estratificada de modo a obter-se uma proporção idêntica de *Churner*s e *Não Churners* em todos os conjuntos de dados.

A avaliação dos diferentes modelos preditivos, foi realizada através das taxas de erro e do rácio designado *Lift*, que mede a mudança na concentração de uma classe específica quando um modelo é utilizado para seleccionar uma amostra propositadamente enviesada a partir da população (Bação, 2006).

Desenvolvemos dois modelos: uma Rede Neuronal e uma Árvore de Decisão que se especificam a seguir. Adicionalmente aos objectivos deste projecto desenvolvemos igualmente um modelo Probit. No Anexo F encontra-se o Diagrama o *SAS* através do qual operacionalizamos as análises inerentes ao nosso projecto.

Assim procurámos desenvolver modelos que apresentassem uma taxa de erro o baixa possível e que, simultaneamente, nos garantissem matizes de confusão o mais "limpas" possível (i.e. em que a maioria dos *Churner*s fosse classificada como *Churner* e a maioria dos *Não Churners* fosse classificada como Não *Churner*) e elevados rácios de *Lift* (i.e. elevados ganhos relativamente a uma classificação aleatória).

1. Modelos

1.1 Árvore de Decisão

O melhor modelo de Árvore de Decisão que desenvolvemos para resolver o problema de classificação de probabilidade de *Churn* no Seguro Automóvel assenta nas seguintes variáveis:

Nome	Descrição
TOTAL	NÚMERO DE PRODUTOS DIFERENTES QUE POSSUI NA MESMA COMPANHIA QUE SEGURO DO AUTOMÓVEL QUE UTILIZA COM MAIS FREQUÊNCIA.
CLIENTE_EXCLUSIVO	CLIENTE EXCLUSIVO AO NÍVEL DO SEGURO AUTOMÓVEL BINÁRIA (0-NÃO É CLIENTE EXCLUSIVO; 1- CLIENTE EXCLUSIVO)
N_AUTO	NÚMERO DE SEGUROS AUTOMÓVEL (0- UM SEGURO AUTOMÓVEL; 1- MAIS DE UM SEGURO)
MUD_SEGURADORA	NO ÚLTIMO ANO DEIXOU DE TRABALHAR COM ALGUMA COMPANHIA DE SEGUROS (0-NÃO; 1-SIM)
PREMIO_250	PRÉMIO DO SEGURO AUTOMÓVEL (0-ATÉ 250 EUROS; 1- MAIS DE 250 EUROS)
MAIS_CONTACTO	A SEGURADORA DO AUTOMÓVEL QUE UTILIZA COM MAIS FREQUÊNCIA É AQUELA COM QUE TEM MAIS CONTACTO.
PARTICIPAÇÃO	PARTICIPAÇÃO DE SINISTROS RELATIVOS À APÓLICE EM ANÁLISE (0-NÃO; 1- SIM)
IDADE	IDADE DO INQUIRIDO
GÉNERO	GÉNERO DO INQUIRIDO

Tabela 2 - Varáveis incluídas no modelo de Árvore de Decisão

Este modelo foi desenvolvido com base no critério de decisão de redução da Entropia. Permitimos que cada nó desse origem apenas a dois ramos e que a Árvore tivesse uma profundidade máxima de seis níveis. A medida de avaliação do modelo (*model assessment value*) usada foi a proporção de observações mal classificadas.

A Árvore consegue prever correctamente cerca de 64.9% (100-35.1) dos casos no Conjunto de Treino.

Na tabela seguinte apresentamos o desempenho da nossa Árvore de Decisão nos três conjuntos de dados.

		Não Respostas	
Conjunto de dados	**Taxa de Erro (%)**	**Subst. Por valor baseado Distribuição (%)**	**Subst. por Tendência Central (%)**
Treino	35.1	35.4	35.1
Validação	35.4	36.2	35.6
Teste	43.9	43.9	43.5

Tabela 3 - Taxas de Erro da Árvore de Decisão nos vários conjuntos de dados.

No Conjunto de Treino e de Validação a Árvore revela taxas de erro idênticas. Contudo, no Conjunto de Teste o desempenho deste modelo sofre uma ligeira quebra, a sua Taxa de Erro aumenta cerca de 9p.p. face ao obtido no Conjunto de Treino. Isto pode indicar que a nossa Árvore de Decisão poderá apresentar alguns problemas de generalização dos seus resultados.

De notar que quando substituímos as Não Respostas pela Média ou Moda das respectivas variáveis a Taxa de Erro da nossa Árvore no Conjunto de Treino não se altera e no Conjunto de Validação aumenta duas décimas. Já no Conjunto de Teste o comportamento é o inverso, a aplicação desta técnica resultou numa diminuição da Taxa de Erro da Árvore em quatro décimas.

A análise da matriz de confusão da nossa Árvore revela que este modelo tem alguma tendência para obter falsos negativos, i.e. para classificar como *Não Churners* indivíduos que de facto o são. Os *Não Churners* são, no geral, identificados correctamente pela nossa Árvore, contudo existe alguma confusão no que respeita à classificação dos *Churner*s.

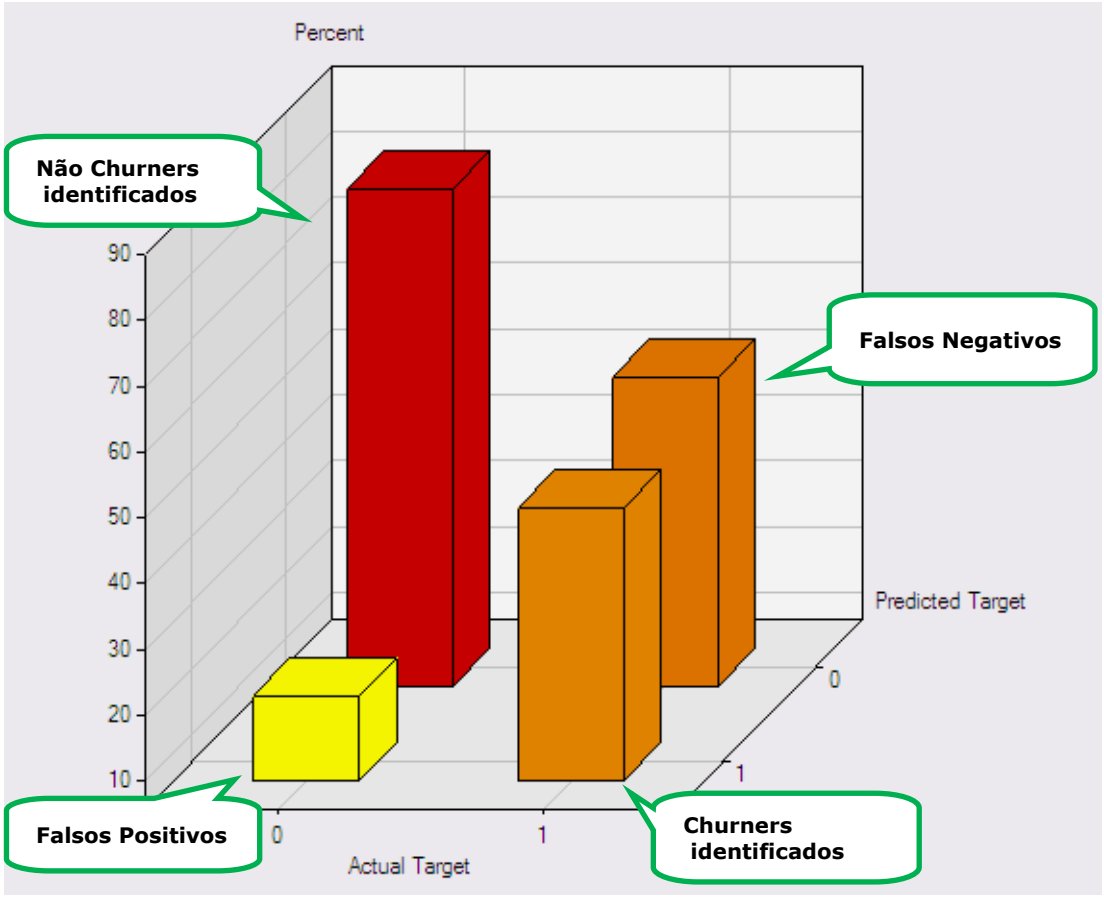


Gráfico 11 - Matriz de Confusão da Árvore de Decisão

Esta Árvore apresentou um *Lift* de, aproximadamente, 1.45 como podemos verificar no gráfico que se segue.

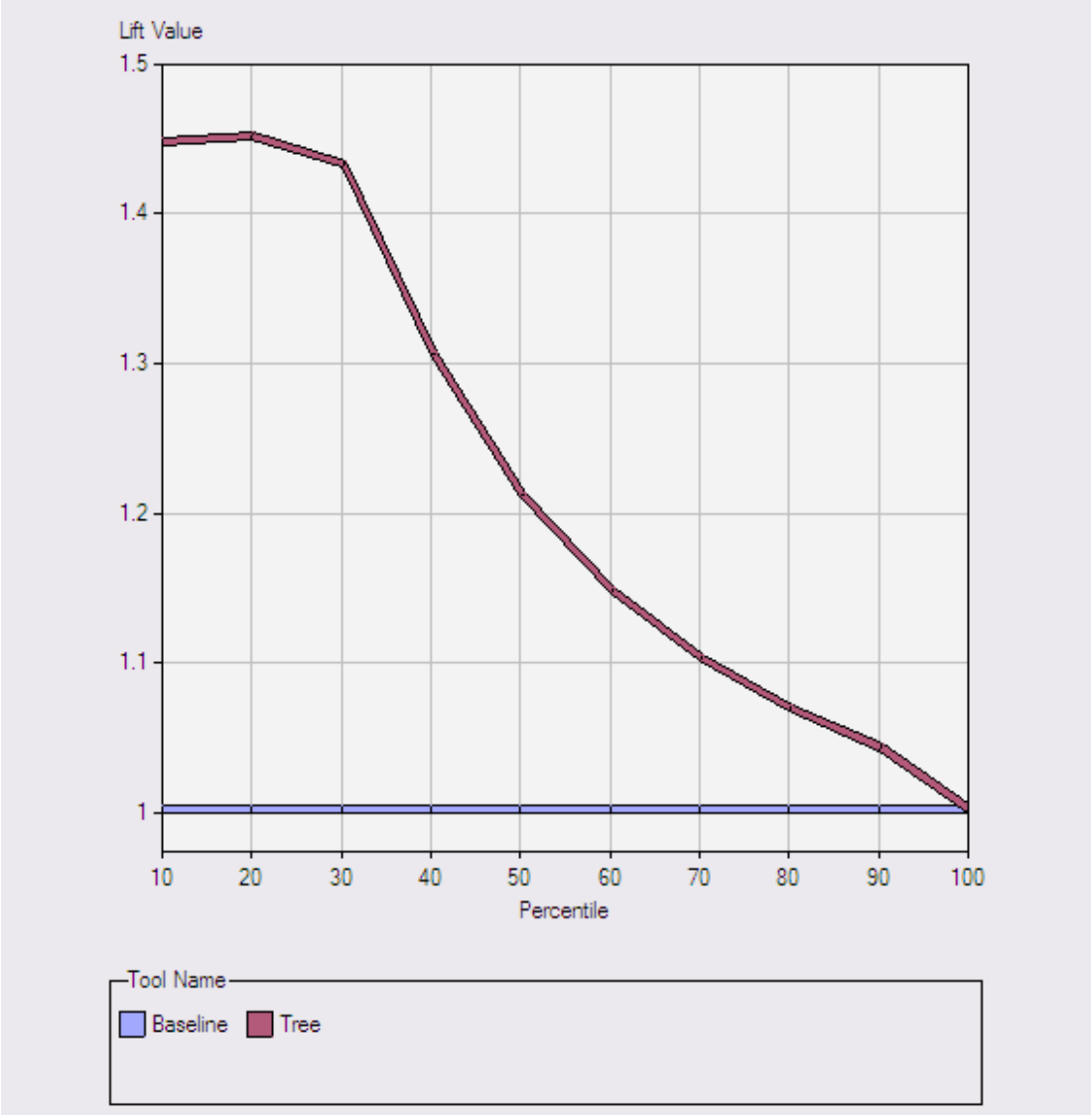


Gráfico 12 - *Lift* da Árvore de Decisão

Podemos fazer uma análise mais detalhada analisando o gráfico de ganhos cumulativos.

Nele podemos ver que entre os 10% de inquiridos que a nossa Árvore identifica como tendo maior probabilidade de abandonarem a companhia estão cerca de 14.5% dos *Churners*, o que se traduz num ganho de apenas 4,5p.p. face a uma decisão aleatória (i.e. sem modelo), a qual encontraria 10% dos *Churners*.

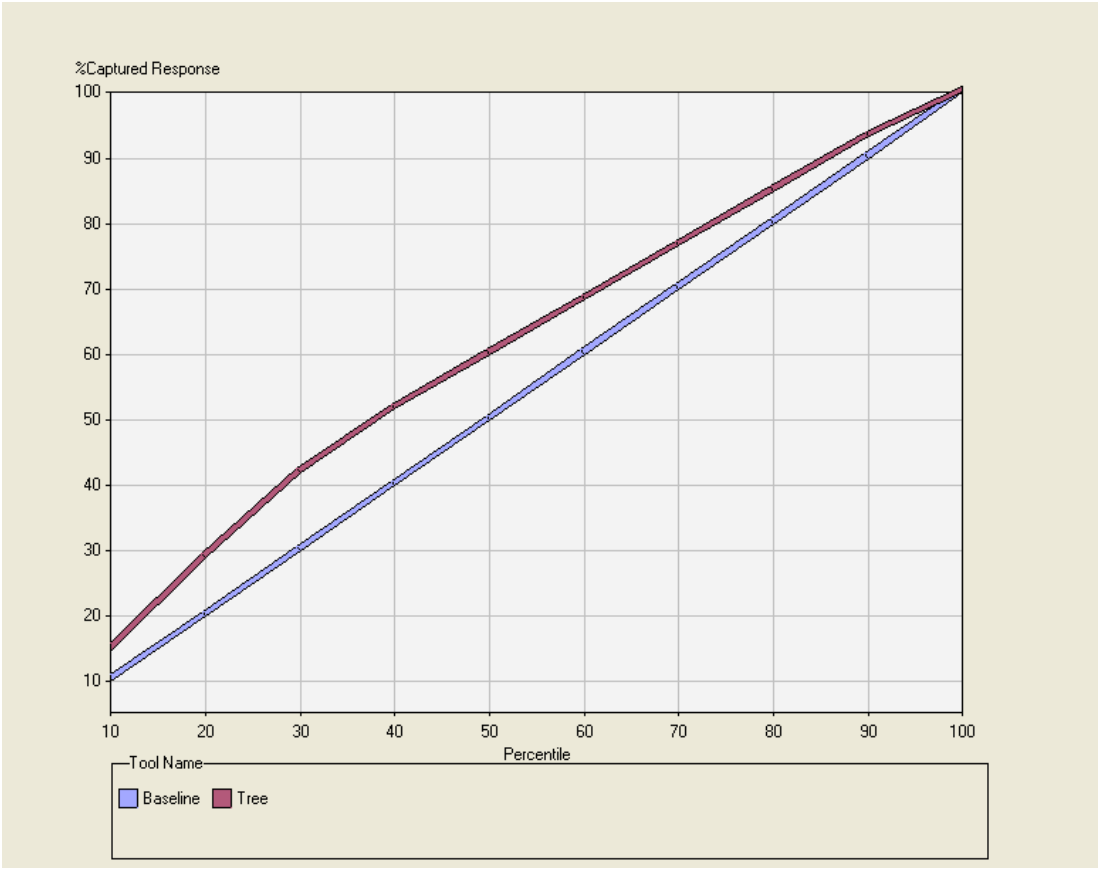


Gráfico 13 - Ganhos Cumulativos da Árvore de decisão

Entre os 30% de inquiridos que a nossa Árvore identifica como tendo maior probabilidade de abandonarem a companhia estão cerca de 41% dos *Churners*, entre os 40% com maior probabilidade de se tornarem *Churners* estão cerca de 50% dos mesmos.

No Gráfico 6 podemos ver a Árvore de Decisão que obtivemos. São apresentados os três primeiros níveis da mesma. No Anexo G encontra-se disponível um Gráfico que nos mostra os cinco primeiros níveis desta Árvore.

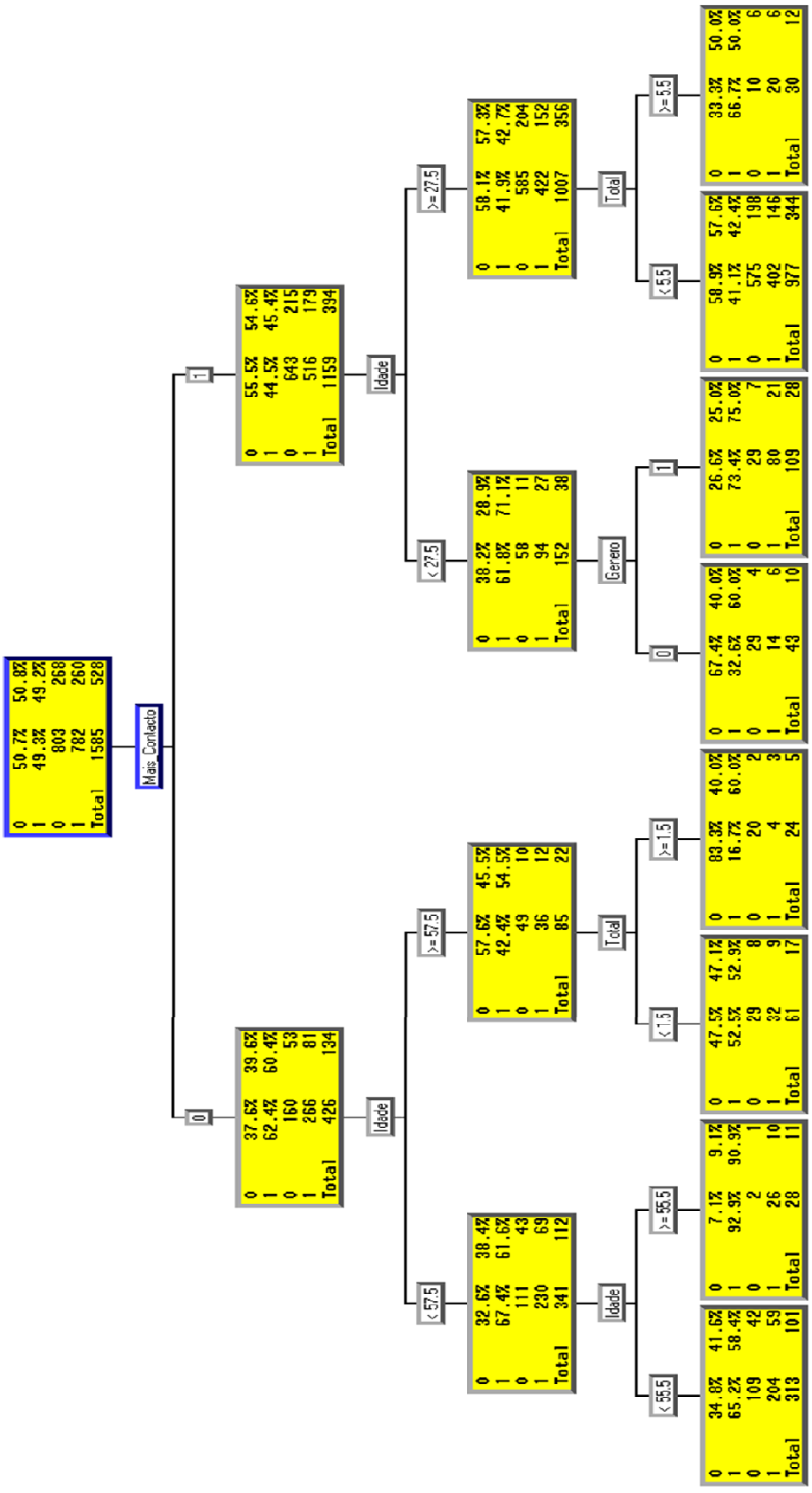


Ilustração 5 - Árvore de Decisão (Três primeiros níveis)

A Árvore inicia-se com uma partição baseada na variável que assinala se a seguradora onde possui o seguro do automóvel que utiliza com mais frequência corresponde ou não à companhia com a qual tem mais contacto, denotando a importância do contacto entre seguradoras e clientes para construir uma relação comercial duradoura e estável a longo prazo. A esta variável segue-se a IDADE, GÉNERO e NÚMERO TOTAL DE PRODUTOS.

Com base na nossa Árvore podemos extrair algumas regras de classificação de *Churn*. A título de exemplo deixa-mos as seguintes:

- SE a companhia onde possui o seguro do automóvel que utiliza com mais frequência Não corresponde à companhia com a qual tem mais contacto E tem menos de 55.5 anos ENTÃO é um potencial *Churner*.

- SE a companhia onde possui o seguro do automóvel que utiliza com mais frequência corresponde à companhia com a qual tem mais contacto E tem menos de 27.5 anos E é Homem ENTÃO é um potencial *Churner*.

De notar que algumas das regras que podemos extrair desta Árvore têm alguns aspectos que vão contra as nossas expectativas iniciais, por exemplo podemos analisar as seguintes regras:

- SE a companhia onde possui o seguro do automóvel que utiliza com mais frequência corresponde à companhia com a qual tem mais contacto E tem pelo menos 27.5 anos E possui até 5.5 produtos diferentes nesta Companhia ENTÃO Não é um potencial *Churner*.

- SE a companhia onde possui o seguro do automóvel que utiliza com mais frequência corresponde à companhia com a qual tem mais contacto E tem pelo menos 27.5 anos E possui pelo menos 5.5 produtos diferentes nesta companhia ENTÃO é um potencial *Churner*.

Do ponto de vista intuitivo seria de esperar que a Árvore produzisse a classificação inversa nestes dois casos. Assim os inquiridos que mais produtos têm teriam menor propensão ao *Churn*. Estes "falhas" na classificação podem estar relacionadas com a impureza da matriz de confusão que analisamos anteriormente. Contudo não nos podemos esquecer que estamos a analisar apenas os três primeiros níveis da Árvore, sendo possível que ao analisar mais níveis estas classificações se aproximassem mais das nossas expectativas. De notar também a reduzida base amostral de algumas destas células.

Uma outra leitura possível é que os indivíduos que possuem mais produtos têm um maior gasto com seguros, tendo assim necessidade de redimensionar a sua carteira de seguros, apresentando consequentemente uma maior probabilidade de se tornarem *Churners*.

1.2 Rede Neuronal

O nosso melhor modelo de Rede Neuronal assenta em 12 variáveis da nossa Base de Dados. São elas:

Nome	Descrição
VIDA	NÚMERO DE PRODUTOS DIFERENTES, DO RAMO VIDA, QUE POSSUI NA MESMA COMPANHIA QUE SEGURO DO AUTOMÓVEL QUE UTILIZA COM MAIS FREQUÊNCIA
TOTAL	NÚMERO DE PRODUTOS DIFERENTES QUE POSSUI NA MESMA COMPANHIA QUE SEGURO DO AUTOMÓVEL QUE UTILIZA COM MAIS FREQUÊNCIA
HABITAÇÃO	POSSUI UM SEGURO DE HABITAÇÃO NA MESMA COMPANHIA QUE SEGURO DO AUTOMÓVEL QUE UTILIZA COM MAIS FREQUÊNCIA (0-NÃO; 1-SIM)
CLIENTE_EXCLUSIVO	CLIENTE EXCLUSIVO AO NÍVEL DO SEGURO AUTOMÓVEL (0-NÃO É CLIENTE EXCLUSIVO; 1-CLIENTE EXCLUSIVO)
N_AUTO	NÚMERO DE SEGUROS AUTOMÓVEL (0- UM SEGURO AUTOMÓVEL; 1- MAIS DE UM SEGURO)
MUD_SEGURADORA	NO ÚLTIMO ANO DEIXOU DE TRABALHAR COM ALGUMA COMPANHIA DE SEGUROS (0-NÃO; 1-SIM)
MAIS_CONTACTO	A SEGURADORA DO AUTOMÓVEL QUE UTILIZA COM MAIS FREQUÊNCIA É AQUELA COM QUE TEM MAIS CONTACTO (0-NÃO; 1- SIM)
COB_EXTRA	A APÓLICE DO INQUIRIDO INCLUI COBERTURAS EXTRA (0-NÃO; 1- SIM)
PARTICIPAÇÃO	PARTICIPAÇÃO DE SINISTROS RELATIVOS À APÓLICE EM ANÁLISE (0-NÃO; 1- SIM)
RC_MAIS	O CAPITAL DE RESPONSABILIDADE CIVIL É SUPERIOR AO OBRIGATÓRIO (0-NÃO; 1- SIM)
IDADE	IDADE DO INQUIRIDO
GÉNERO	GÉNERO DO INQUIRIDO

Tabela 4 - Varáveis incluídas no modelo de Rede Neuronal

Como a nossa variável de interesse (*Target*) é binária optamos por usar como critério de selecção do modelo (*Model Selection Criteria*) a Taxa erro.

A Rede, tem uma arquitectura MLP com uma camada escondida constituída por seis neurónios. Usa uma função de activação Gaussiana e uma função de combinação Linear. Foi usada a técnica de treino *Quasi-Newton*, os pesos foram iniciados aleatoriamente.

Com estas especificações o modelo consegue prever correctamente cerca de 70.7% (100-29.3) dos casos no Conjunto de Treino.

O desempenho da nossa Rede nos vários conjuntos de dados pode ser visto na tabela seguinte.

		Não Respostas	
Conjunto de dados	**Taxa de Erro (%)**	**Subst. Por valor baseado Distribuição (%)**	**Subst. por Tendência Central (%)**
Treino	29.3	28.6	32.8
Validação	35.8	36.7	39.0
Teste	36.7	38.4	38.0

Tabela 5 - Taxas de Erro da Rede Neuronal nos vários conjuntos de dados.

As taxas de erro obtidas pela rede, são similares nos três conjuntos de dados variando em menos de 1p.p. entre o Conjunto de Validação e de Teste. No Conjunto de Treino é, naturalmente, mais baixa.

As técnicas de tratamento de Não Respostas aplicadas não resultaram em nenhum ganho ao nível das Taxas de Erro apresentadas em nenhum dos conjuntos de dados estudados.

A análise da matriz de confusão deste modelo revela que a mesma é relativamente "limpa", i.e. o modelo tende a classificar os *Churner*s como potenciais *Churner*s e os *Não Churners* como potencialmente *Não Churners*.

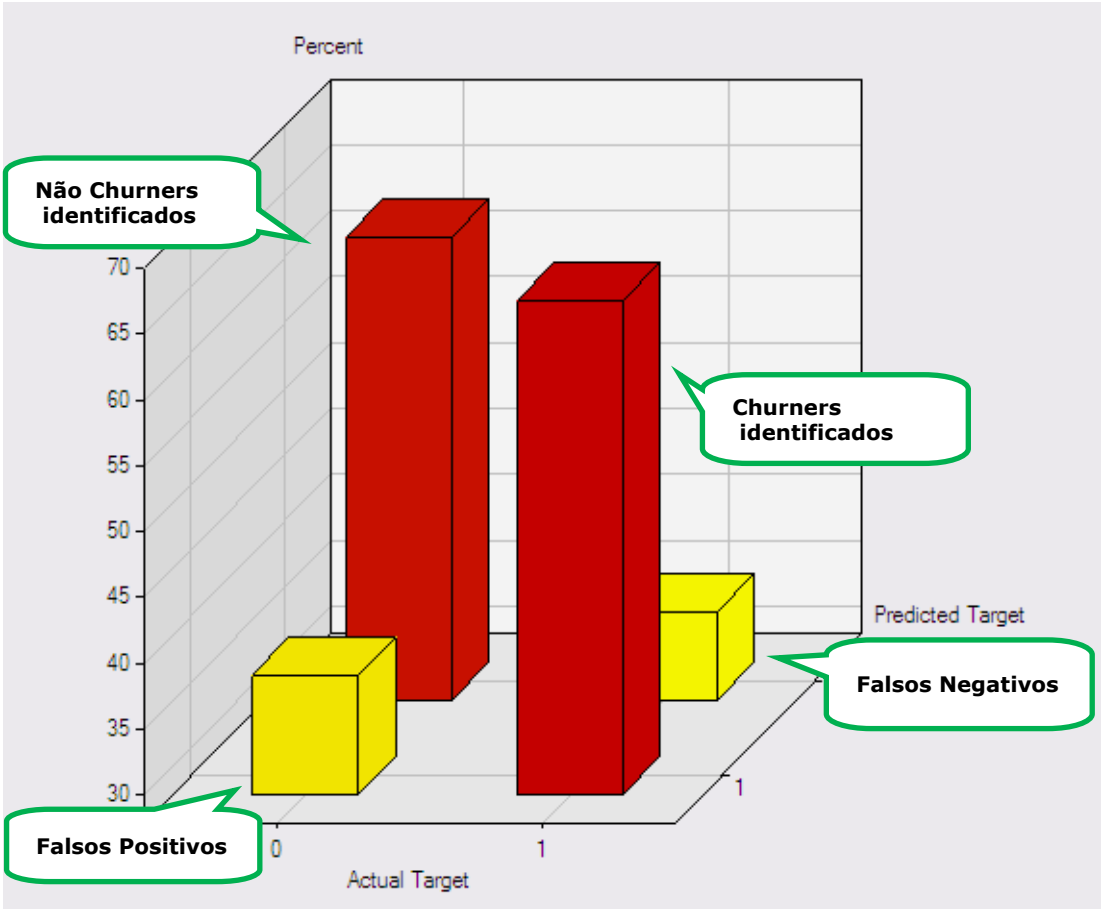


Gráfico 14 - Matriz de Confusão da Rede Neuronal.

Entre as classificações erradas destaca-se a classificação de *Não Churners* como *Churner*s. Este erro apresenta um valor ligeiramente superior à classificação errónea de *Churner*s como *Não Churners*.

Esta Rede apresentou um *Lift* de, aproximadamente, 1.6, como podemos ver no gráfico seguinte. Este valor de *Lift* é Ligeiramente superior ao obtido com a Árvore de Decisão.

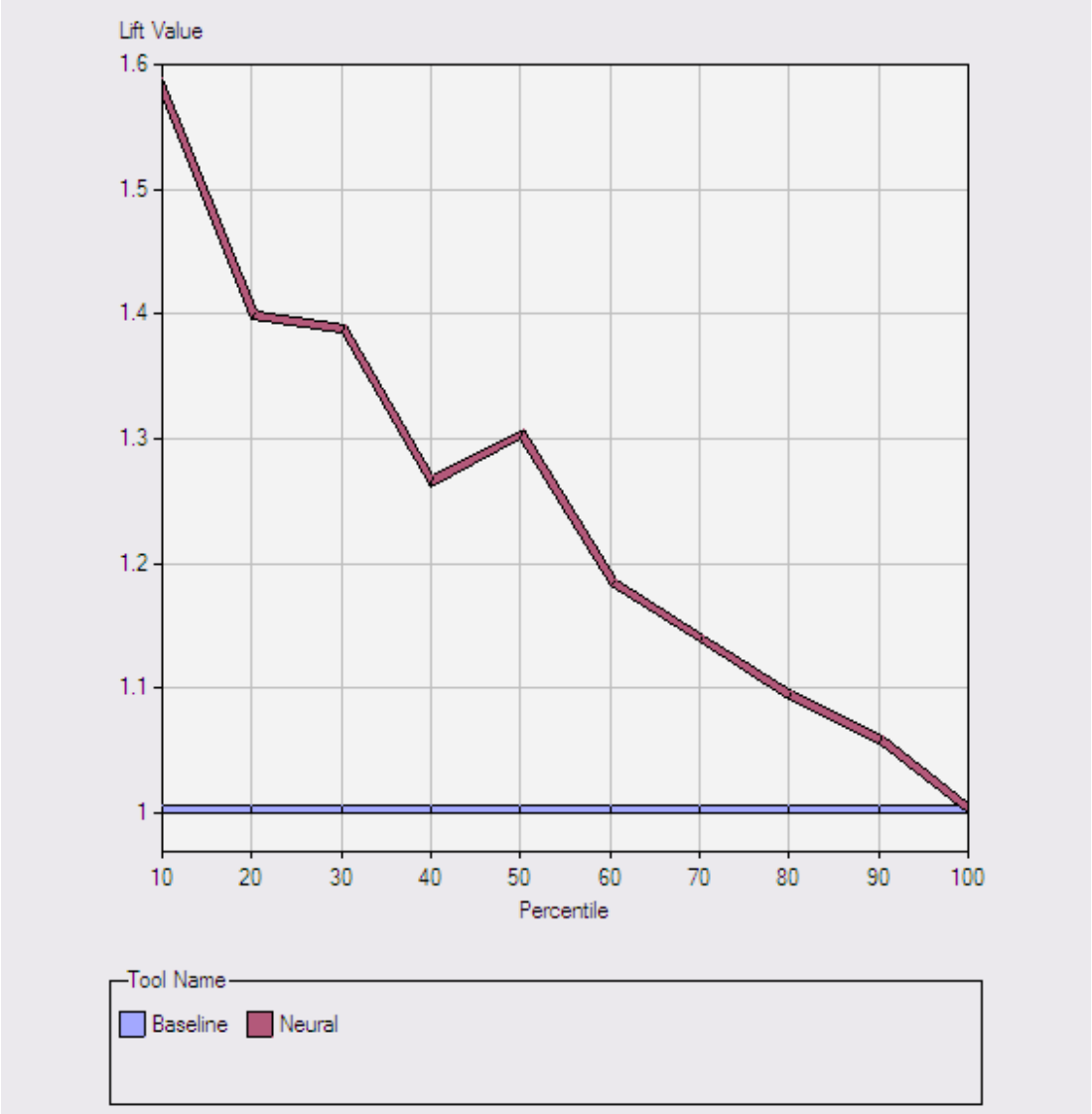


Gráfico 15 - *Lift* da Rede Neuronal

Tal como fizemos para a Árvore de Decisão procedemos à análise do gráfico de ganhos cumulativos.

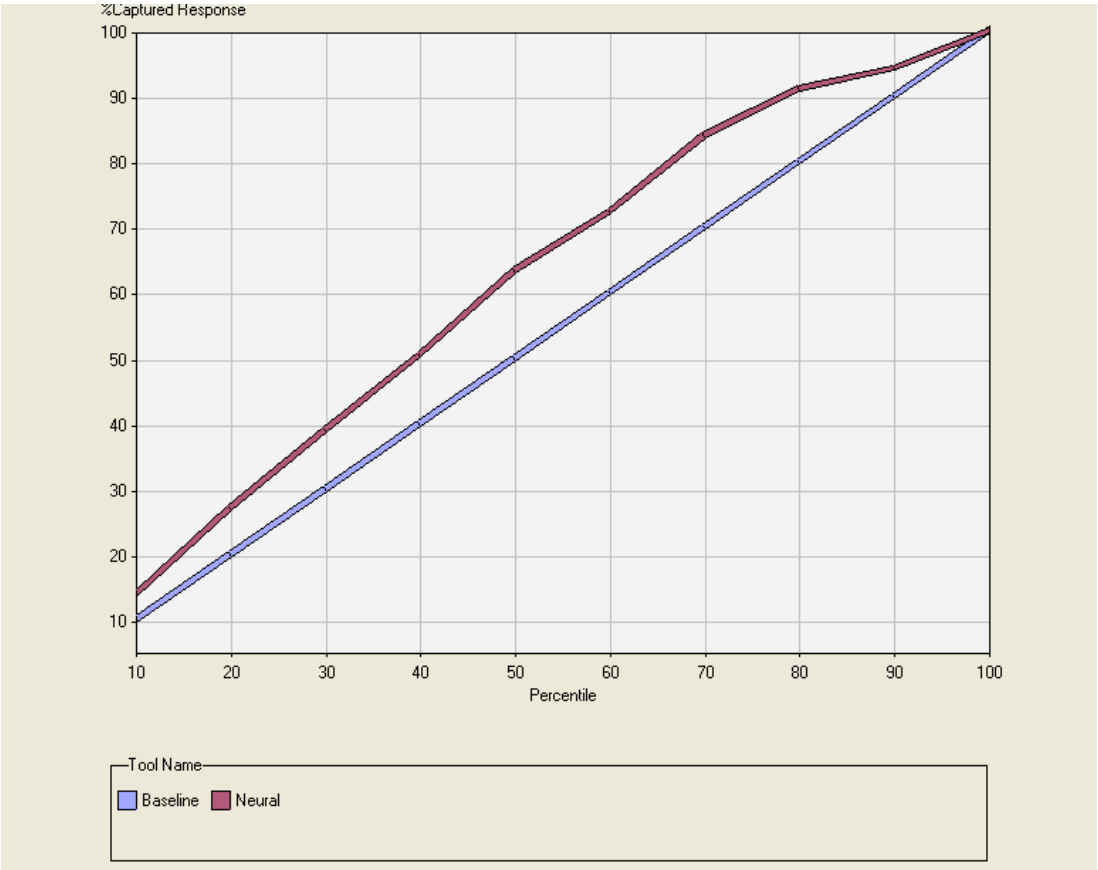


Gráfico 16 - Ganhos Cumulativos da Rede Neuronal

No caso da Rede podemos ver que entre os 10% de inquiridos que a Rede identifica como tendo maior probabilidade de abandonarem a companhia estão cerca de 16% dos *Churners*. A Rede produz assim um ganho de 6p.p. face a uma decisão aleatória, a qual encontraria 10% dos *Churners*. Este ganho ultrapassa em cerca de 2.5p.p. o ganho obtido com a Árvore de Decisão.

Entre os 30% de inquiridos que a Rede Neuronal identifica como tendo maior probabilidade de abandonarem a companhia estão cerca de 40% dos *Churners*; entre os 40% com maior probabilidade de se tornarem C*hurners* estão cerca de 51% dos mesmos.

1.3 Probit

Adicionalmente ao desenvolvimento dos dois modelos já apresentados resolvemos explorar o uso de mais um algoritmo de classificação: o Probit. O nosso interesse em avaliar o desempenho deste algoritmo prende-se com três níveis de razões: (a) o desempenho dos algoritmos de Árvores de Decisão e Redes Neuronais; (b) explorar as ferramentas de classificação disponíveis no *SAS* e (c) avaliar o desempenho do Probit *per si* de modo a avaliar a pertinência, ou não, de investir em *software* específico de *Data Mining*. O Probit está implementado de forma mais transversal (inclusive em *software* Livre e *Open Source*) que os primeiros, tornando-se por isso mais acessível e económico.

Para desenvolver o nosso modelo Probit procedemos como nos modelos anteriores e experimentamos várias especificações. O nosso melhor modelo Probit conta com a informação das seguintes variáveis:

Nome	Descrição
CLIENTE_EXCLUSIVO	CLIENTE EXCLUSIVO AO NÍVEL DO SEGURO AUTOMÓVEL (0-NÃO É CLIENTE EXCLUSIVO; 1-CLIENTE EXCLUSIVO)
MAIS_CONTACTO	A SEGURADORA DO AUTOMÓVEL QUE UTILIZA COM MAIS FREQUÊNCIA É AQUELA COM QUE TEM MAIS CONTACTO (0-NÃO; 1- SIM)
PARTICIPACAO	PARTICIPAÇÃO DE SINISTROS RELATIVOS À APÓLICE EM ANÁLISE (0-NÃO; 1- SIM)
IDADE	IDADE DO INQUIRIDO
GENERO	GÉNERO DO INQUIRIDO

Tabela 6 - Varáveis incluídas no modelo Probit

No desenvolvimento deste modelo tivemos em conta não só a taxa de erros, matrizes de confusão e *Lifts* que os vários modelos testados apresentaram como também a significância das variáveis envolvidas no modelo.

O nosso Probit consegue prever correctamente cerca de 60.6% (100-39.4) dos casos no Conjunto de Treino.

Na tabela seguinte apresentamos a Taxa de Erro desempenho do nosso Probit nos três conjuntos de dados.

		Não Respostas	
Conjunto de dados	**Taxa de Erro (%)**	**Subst. Por valor baseado Distribuição (%)**	**Subst. por Tendência Central (%)**
Treino	39.4	39.4	39.4
Validação	40.3	40.7	40.7
Teste	45.7	45.6	45.9

Tabela 7 - Taxas de Erro do Probit nos vários conjuntos de dados.

A Taxa de Erro do Probit é bastante semelhante no Conjunto de Treino e no de Validação, apresentando valores próximos de 40%. Já no Conjunto de Teste a Taxa de Erro do modelo sobe para cera de 46%.

As duas técnicas de tratamento de Não Respostas não alteraram o comportamento do Probit, produzindo apenas variações, na sua Taxa de Erro, na ordem das duas décimas.

A análise da matriz de confusão revela que o Probit consegue identificar os clientes que não são potenciais *Churner*s com alguma facilidade, pois a grande maioria dos clientes que não mudaram de companhia são classificados por este modelo como não sendo potenciais *Churner*s.

A maior dificuldade do modelo está em identificar os clientes que se podem tornar *Churner*s. Apesar de a maioria dos potenciais *Churner*s serem classificados correctamente pelo modelo, nota-se uma tendência para obter falsos negativos. Ou seja, tem alguma tendência para classificar clientes que são potenciais *Churner*s como não o sendo.

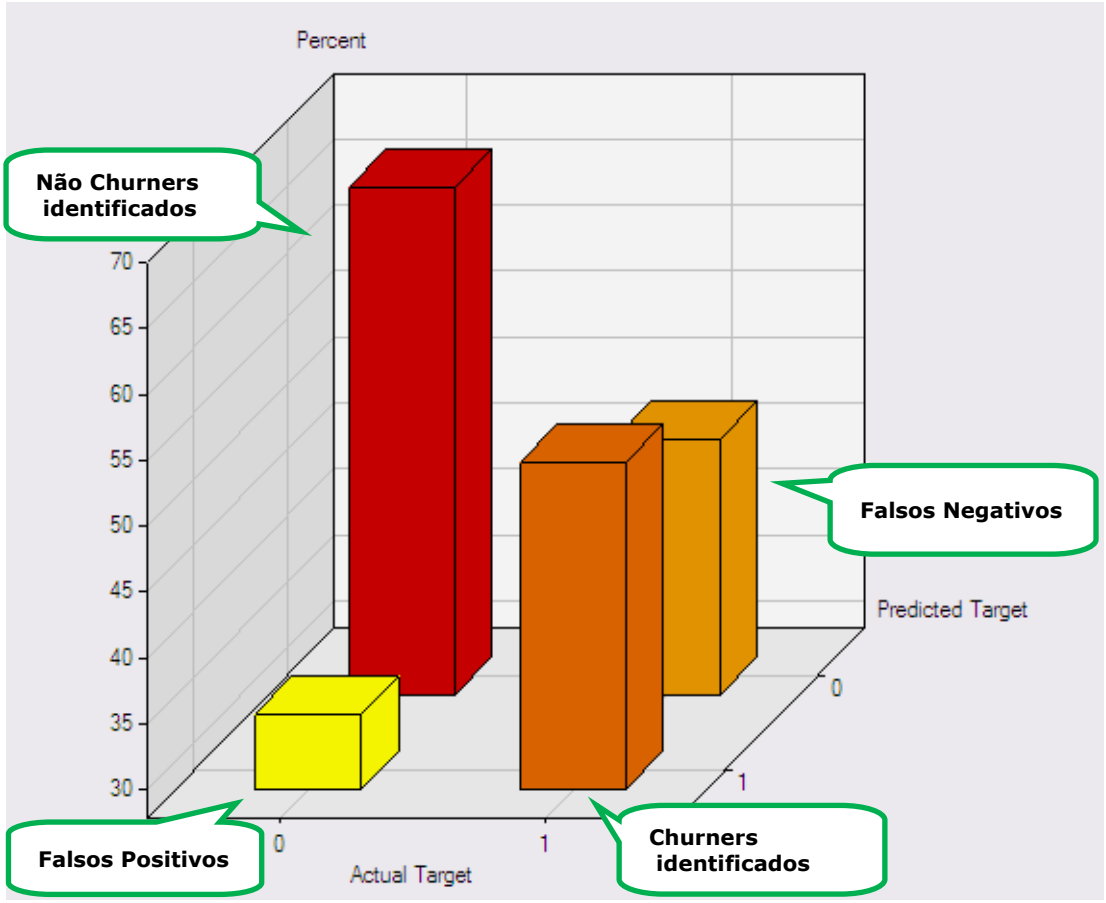


Gráfico 17 - Matriz de Confusão do Probit

O nosso Probit apresentou um *Lift* de, aproximadamente, 1.5 como mostra o gráfico seguinte.

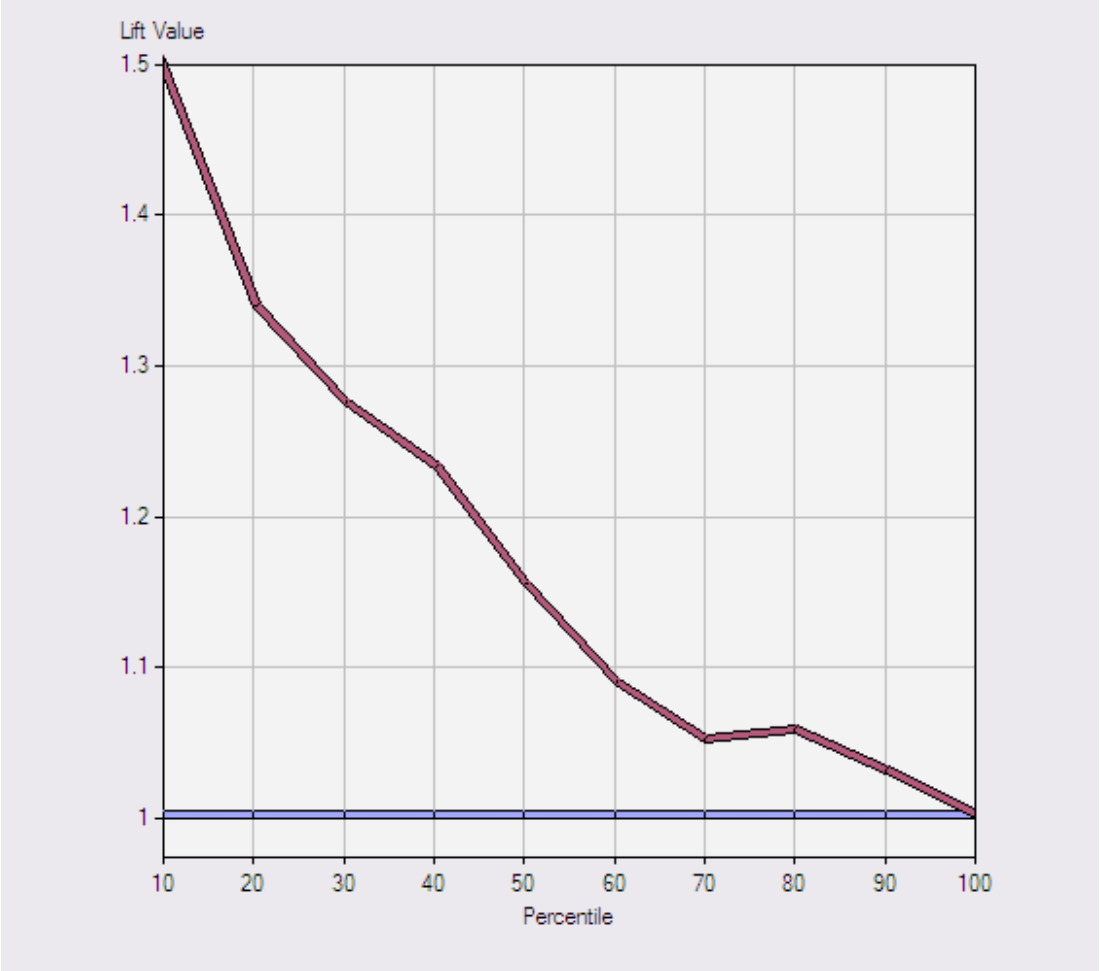


Gráfico 18 - *Lift* do Probit

A análise do gráfico de ganhos cumulativos revela que entre os 10% de inquiridos que este modelo identifica como tendo maior probabilidade de abandonarem a companhia estão cerca de 15% dos *Churners,* valor que garante um ganho de 5p.p. face a uma situação aleatória.

Entre os 30% de inquiridos que o Probit identifica como tendo maior probabilidade de abandonarem a companhia estão cerca de 40% dos *Churners*. 50% dos *Churners* estão entre os 40% de inquiridos que o Probit assinala como tendo maior probabilidade de se tornarem *Churners*. Valores muito semelhantes aos obtidos com a Árvore de Decisão.

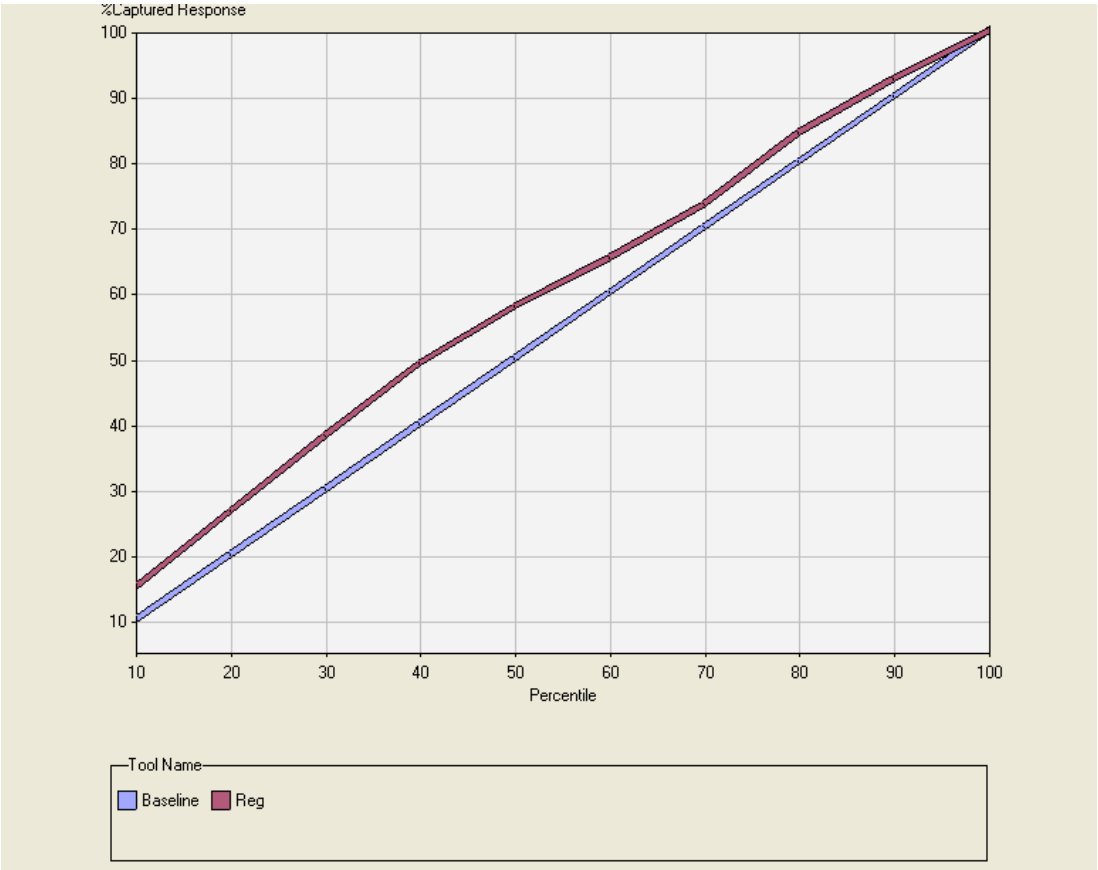


Gráfico 19 - Ganhos Cumulativos do Probit

Adicionalmente a esta análise replicamos este modelo Probit com o *software, open source, GRETL* (c.f. output no Anexo H). Obtivemos uma matriz de confusão semelhante á obtida no *SAS*.

Variável	Coeficiente	Declive Média	T-Rácio	Valor-P
Constante	0,715	4,781	4,781	<0,00001
Mais_Contacto	-0,373	-0,148	-4,941	<0,00001
Cliente_Exclusivo	-0,307	-0,121	-3,037	0,002
Habitação	-0,243	-0,096	-3,148	0,002
Género	0,241	0,096	3,148	0,002
Participação	0,226	0,090	2,985	0,003
Idade	-0,008	-0,003	-3,513	0,000

Tabela 8 – Coeficientes β e níveis de significância do modelo Probit

Analisando mais detalhadamente o modelo, começamos por notar que todas as variáveis são estatisticamente significativas e que os sinais apresentados pelos estimadores β vão ao encontro das nossas expectativas.

Das variáveis em análise, o GÉNERO e a PARTICIPAÇÃO DE SINISTROS têm um impacto positivo na probabilidade de o inquirido vir a ser um *Churner*. As restantes variáveis têm um impacto negativo nessa mesma probabilidade.

A variável que mais impacto tem na redução da probabilidade de *Churn* é a relacionada com o contacto que se tem com a companhia. Quando a companhia titular da apólice de Seguro Automóvel que o inquirido utiliza com mais frequência se torna a companhia preferencial (i.e. com a qual o inquirido tem mais contacto), a probabilidade de este abandonar a companhia decresce de forma muito significativa.

O mesmo acontece quando estamos perante um cliente que apenas possui Seguros automóvel numa companhia, i.e. é um cliente exclusivo ao nível do Seguro Automóvel. Não podemos deixar de notar o comportamento dos inquiridos que possuem Seguro HABITAÇÃO na mesma companhia do Seguro Automóvel. Estes têm uma probabilidade inferior de se tornarem *Churners* quando comparados com os restantes inquiridos. Podemos ainda dizer que a probabilidade de um cliente abandonar a companhia tende a diminuir com a IDADE.

Analisando as variáveis que têm um impacto positivo, o Probit diz-nos que os Homens (GÉNERO=1) têm uma maior probabilidade de se tornarem *Churners* quando comparados com as Mulheres. Os clientes que já tiveram necessidade de apresentar uma PARTICIPAÇÃO de sinistro à companhia também têm a probabilidade de se tornarem *Churners*.

Sendo o Probit um modelo não linear, os coeficientes β não correspondem aos efeitos marginais das variáveis na probabilidade de o inquirido ter um comportamento de *Churn* no futuro, não se podendo interpretar os valores numéricos destes coeficientes (Pinheiro, 2008).

Para podemos fazer tal interpretação, devemos primeiro obter os efeitos marginais, contudo nestes modelos, o efeito marginal não é constante para todos os valores de um dado regressor, ou seja, o efeito que uma variável tem sobre o *Churn* varia consoante o valor dessa mesma variável. Por isso, temos que especificar um ponto e calcular os efeitos marginais para o mesmo. De acordo com o procedimento usual, quando se trabalha com modelos Probit, forma calculados os efeitos marginais para o ponto definido pela média de todos os regressores (Pinheiro, 2008) e obtivemos os efeitos apresentados na tabela acima.

Feito este cálculo podemos dizer que quando a companhia titular da apólice de Seguro Automóvel que o inquirido conduz com mais frequência se torna a companhia preferencial (com a qual o inquirido tem mais contacto), a probabilidade de o inquirido abandonar essa mesma companhia decresce cerca de 14.8%, mantendo todas as outras variáveis na respectiva média. Este valor é ligeiramente superior ao que encontramos para o caso de CLIENTES EXCLUSIVOS, os quais têm uma probabilidade de se tornarem Churners inferior em 12.1% quando comparados com aqueles que distribuem os seus Seguros Automóvel por várias companhias, mantendo todas as outras variáveis na respectiva média.

1.4 Simulação de Monte Carlo

Depois de desenvolvemos os três modelos apresentados tivemos interesse em testar a sua estabilidade.

Para tal sujeitamos os nossos modelos a uma simulação de Monte Carlo. A nossa simulação de Monte Carlo (c.f. Anexo I) consistiu em correr diversas vezes os nossos modelos iniciando-os em *seeds* diferentes (determinadas aleatoriamente). Fizemos 31 repetições, depois computamos a Média e o Desvio-padrão das taxas de erro obtidas por cada um dos modelos no Conjunto de Teste. Obtivemos os resultados presentes na tabela 4.

Modelo	**Taxa de Erro no Conj. Teste (%)**	**Taxa de Erro Média (%) (n=31 repetições)**	**Desvio Padrão (n=31 repetições)**
Probit	45.7	41.9	1.85
Árvore de Decisão	43.9	40.0	1.60
Rede Neuronal	36.7	37.3	1.33

Tabela 9 – Resultados da simulação de Monte Carlo.

Esta simulação revela, de um ponto de vista geral, que os modelos são bastante robustos, sendo o seu comportamento bastante semelhante e consistente ao longo das repetições efectuadas. As taxas de erro médias obtidas nas repetições efectuadas são semelhantes as taxas de erro obtidas originalmente no Conjuntos de Teste, nunca se desviando destas mais de 4p.p. O Desvio-padrão desta taxa de erro é baixo nunca superando a barreira dos 2 pontos. o que sugere que os modelo são bastante estáveis.

Na nossa simulação de Monte Carlo a Rede Neuronal mostrou ser o modelo com uma Taxa de Erro Média mais baixa e mais próxima da obtida inicialmente no Conjunto de Teste, bem como o modelo mais estável, ou seja aquele que apresenta um menor Desvio-padrão de resultados entre os três modelos testados.

O Probit encontra-se no antípoda. Este revelou-se o modelo que apresenta uma maior taxa de erro, em termos médios, nas classificações que efectua sendo esta superior, em mais de 4p.p, à taxa de erro da Rede Neuronal. O Probit é igualmente o modelo com mais variabilidade nos seus resultados, tendo um Desvio-padrão de cerca de 1.85, valor ligeiramente superior ao dos restantes modelos.

A Árvore de Decisão apresentou um nível de consistência intermédio com um Desvio-Padrão de 1.60. Tal como o Probit a Taxa de Erro Média é mais baixa do que o valor obtido inicialmente no Conjunto de Teste afastando-se desta cerca de 4p.p.

2. Scoring

Por último tivémos interesse em observar o comportamento dos nossos modelos em situação de Classificação (*Scoring*).

Como já fizemos referência os nossos dados provêm de dois estudos designados por "Barómetro de Seguros" e "Transferências de Seguros". O Segundo estudo, "Transferências de Seguros", é feito por recontacto dos indivíduos que foram inquiridos no estudo "Barómetro de Seguros". São recontactados cerca de 2.500 inquiridos.

Para ver o comportamento dos nossos modelos em situação de classificação recorremos aos dados dos inquiridos que não foram recontactados no âmbito do estudo de "Transferências de Seguros".

Assim, foram utilizados todos os registos respeitantes a apólices de particulares, de inquiridos que habitualmente tomam decisões respeitantes ao Seguro Automóvel, cujo seguro está em nome próprio e que não foram não foram recontactados para o estudo "Transferências de Seguros", não fazendo por isso parte da Base de Dados de treino. Esta quarta Base de Dados, que designaremos por Base de Dados de Classificação é constituída por cerca de 2.573 registos.

Como estes registos não fazem parte da Base de Dados de Treino, não temos nenhuma informação sobre o seu comportamento de Mudança, ou não, de Companhia de Seguro Automóvel. É assim possível e pertinente aplicar os nossos modelos de classificação a estes dados com o objectivo de explorar melhor o comportamento dos nossos modelos.

Note-se que não temos forma de averiguar se as classificações produzidas são correctas ou não, esta simulação de *Scoring* tem como objectivo analisar mais profundamente o comportamento dos modelos.

A simulação de *Scoring* foi feita em vários passos. Primeiro procedemos à classificação da Base de Dados de Classificação com cada um dos modelos individualmente. Posteriormente recorremos ao nó de *Ensemble* do *SAS* para fazemos uma classificação única com base na informação fornecida pelos três modelos.

Na tabela seguinte apresentamos os resultados obtidos para cada uma das classificações efectuadas.

Modelo	**Potenciais *Churners* (%)**
Probit	43.1
Árvore de Decisão	28.8
Rede Neuronal	44.4
Ensemble	35.4
Base	**2573**

Tabela 10 – Resultados das classificações efectuadas.

A Rede Neuronal e o Probit classificaram um número semelhante de inquiridos como potenciais *Churners*. A Árvore de Decisão destaca-se dos restantes modelos por classificar um número inferior de inquiridos como potenciais *Churners.* Este facto pode fica-se a dever à tendência, já discutida, da Árvore de Decisão para classificar como *Não Churners* indivíduos que são efectivamente *Churners*. Com o nó de *Ensemble* do SAS classificamos cerca de um terço dos inquiridos com potenciais *Churners*.

3. Rating

Para complementar os objectivos a que nos propusemos, quisemos enriquecer o nosso projecto com a criação de um *Rating* onde pudéssemos classificar, com diferentes graus de risco, os inquiridos.

O *Rating* deveria usar toda a informação disponibilizada pelos modelos desenvolvidos. Nesse contexto optamos por construir um *Rating* baseado no número de "classificações positivas" que o indivíduo recebe tendo por base os três modelos desenvolvidos.

Desta forma obtivemos o seguinte *Rating*:

Número de Classificações como *Churner*	**n**	**%**	**Proposta de Nomenclatura**
Nenhuma	944	36.7	Não *Churner*/Cliente Fiel
Uma	687	26.7	Baixo Risco de *Churn*
Duas	522	20.3	Risco de *Churn* Considerável
Três	420	16.3	Risco de *Churn* Elevado
Base	2573		

Tabela 11 – Rating de Risco de *Churn*.

Cerca de 1/3 dos segurados inquiridos apresenta um risco muito baixo de mudar a companhia onde possui o seguro do automóvel que utiliza com mais frequência, pois não são assinalados por nenhum dos nossos modelos como potenciais *Churners*.

Este *Rating* permite ainda isolar um *Target* de 420 inquiridos, cerca de 16% da nossa amostra, que são assinalados por todos os modelos desenvolvidos como potenciais *Churner*s. Por isso consideramos que são um grupo com um risco de *Churn* elevado.

Neste capítulo descrevemos o trabalho de preparação da Base de Dados de Treino e a respectiva partição pelos conjuntos de Treino, Validação e Teste. Foi discutida a necessidade de fazer *Boosting* à Base de Dados de Treino. Seguidamente foram apresentados os nossos três modelos (Árvore de decisão, Rede Neuronal e Probit). Apresentamos ainda o resultado de uma simulação de Monte Carlo e o Scoring efectuado aos inquiridos que não fizeram parte da nossa Base de Dados de Treino. Para fechar este capítulo apresentámos uma proposta de *Rating* dos inquiridos, em termos de Risco de *Churn*.

No próximo capítulo iremos discutir os resultados obtidos e apresentar algumas propostas para investigação futura.

V Discussão

A execução deste projecto permitiu-nos explorar as nossas Bases de Dados de uma forma pouco habitual na indústria dos estudos de mercado e de opinião.

Conseguimos cumprir o objectivo a que nos propusemos com sucesso: desenvolvemos três modelos de classificação de probabilidade de *Churn* com base na informação das nossas Bases de Dados recorrendo a três algoritmos diferentes (Árvores de decisão, Rede Neuronal e Probit). Foi-nos igualmente possível classificar os inquiridos que não constavam da Base de Dados de Treino através da construção da Base de Dados de Classificação e da aplicação dos três modelos desenvolvidos.

Os modelos de Rede Neuronal e Árvore de Decisão apresentaram alguma supremacia face ao modelo Probit. A Rede Neuronal foi o modelo que apresentou a taxa de erro mais baixa entre os três modelos considerados.

Contribuições

Os modelos desenvolvidos inicialmente apresentaram um ganho nulo face a uma tomada de decisão aleatória. Para ultrapassar esta dificuldade foi necessária a aplicação da técnica de *Boosting* à nossa Base de Dados de Treino. Quadruplicamos os casos negativos (*Churner*s) de modo a obter uma frequência semelhante de *Churner*s e *Não Churners*. Foi sobre esta Base de Dados (Base de Dados *Boosted*) que desenvolvemos os nossos modelos finais.

Partimos assim de uma situação em que uma decisão aleatória levar-nos-ia a acertar em cerca de 50% dos casos. Os nossos modelos apresentam taxas de erro inferiores a este número, indicando algum ganho

relativamente a este valor - na ordem dos 3.6p.p., para o Probit, aos 12.7p.p., para a Rede Neuronal (Conjunto de Teste).

Este ganho ao nível da taxa de erro é modesto, contudo tal não significa que os modelos desenvolvidos não tenham valor prático e económico paras as companhias. O ganho obtido ao nível das taxas de erro, as matrizes de confusão mais "limpas" quando comparadas com as que seriam obtidas numa decisão aleatória e o *Lift* superior a 1 sugerem que a utilização dos nossos modelos traria potencialmente um elevado valor para as companhias.

O modelo que tem potencialmente mais valor para a optimização de uma campanha de retenção é a Rede Neuronal pois é o modelo com uma matriz de confusão mais "limpa". Ao investir os seus recursos (tempo, dinheiro, etc.) nos clientes identificados pela Rede Neuronal uma dada Companhia iria conseguir agir proactivamente junto da grande maioria dos clientes que estavam em risco de a abandonar, uma vez que grande parte dos potenciais *Churners* são identificados correctamente pela Rede Neuronal. Garantiria simultaneamente que poucos recursos seriam desperdiçados com clientes que não têm intenção de a abandonar, uma vez que o número de clientes erroneamente identificados como potenciais *Churners* (Falsos Positivos) é igualmente reduzido. Seria então de esperar que a campanha tivesse um retorno atractivo para a Companhia.

Note-se que o ganho ao nível das taxas de erro, apesar de pequeno, mostrou-se bastante consistente, pois na simulação de Monte Carlo os modelos apresentaram taxas de erro próximas às originais, o desvio-padrão das mesmas mostrou-se bastante reduzido. As taxas de erro também não sofreram alterações significativas com a introdução de várias técnicas de imputação de Não Respostas (*Missing Values),* o que sugere que as Não Respostas presentes na nossa base de dados são aleatórios e não introduziram enviesamentos na estimação dos modelos.

Grupos de Risco

Apesar dos modestos ganhos apresentados pelos nossos modelos estes permitem-nos identificar alguns grupos de inquiridos com maior risco de se tornarem *Churner*s do que outros. Por exemplo:

A nossa Árvore de decisão sugere que os inquiridos que têm pouco contacto com a companhia onde possuem o seguro do automóvel que utilizam com mais frequência e têm menos de 55.5 anos são potenciais *Churners*.

Um outro grupo de rico será o constituído pelos inquiridos que têm contacto com a companhia onde possuem o seguro do automóvel que utilizam com mais frequência, têm menos de 27.5 anos e são Homens.

Analisando mais detalhadamente estes grupos notamos que o contacto entre o inquirido e a seguradora parece ser um aspecto essencial surgindo como a variável que traz maiores ganhos de informação em dois modelos: Árvore de Decisão e Probit, fazendo igualmente parte da Rede Neuronal.

Surgem-nos igualmente variáveis sociodemográficas como a IDADE e o GÉNERO. Quer a Árvore de Decisão quer o Probit apontam que os mais Jovens e o Homens tendem a ter maior propensão para o *Churn*.

Por outro lado os modelos revelam que a posse de determinados produtos, como o Seguro Habitação, é vinculante. Os inquiridos que possuem Seguro Habitação na mesma companhia que o seguro do automóvel que utilizam com mais frequência têm menos probabilidade de se tornarem *Churners*.

Limitações

Estamos conscientes que o nosso projecto apresenta algumas limitações. Uma das principais relaciona-se com o nível de mensuração usado na maior parte das variáveis postas ao nosso dispor.

Grande parte das variáveis tem um nível de mensuração muito baixo. Tratam-se sobretudo de variáveis binárias, o que dificulta o trabalho de descriminação entre *Churner*s e *Não Churners* por parte dos modelos.

Outra limitação é que as Sondagens de onde provêm os nossos dados não foram desenhadas com o objectivo de desenvolver modelos de classificação de probabilidade de *Churn*, mas com o objectivo de avaliar os comportamentos e percepções dos Portugueses face aos seguros e estimar taxas de *Churn*. Nesse sentido estas Sondagens não possuem necessariamente um conjunto de variáveis discriminantes do ponto de vista da classificação de probabilidade de *Churn*.

Propostas para investigação futura

De notar que obtivemos resultados contra as nossas expectativas no que respeita à variável respeitante ao número total de produtos diferentes que o inquirido possui na seguradora onde possui o seguro do automóvel que utiliza com mais frequência. Assim os inquiridos que têm um maior número de produtos têm maior probabilidade de se tornarem *Churners*. Salientamos que estamos a analisar o número de produtos e não de apólices que o inquirido detém. Apesar de obtemos resultados contra-intuitivos com esta variável decidimos mante-la nos nossos modelos, como chamada de atenção a investigação futura para a importância da distribuição do portfólio de seguros que o inquirido faz pelas várias companhia a operar no mercado.

Sugerimos que em investigações futuras sejam feitas alterações ao questionário, que está na origem dos nossos dados, no sentido de

permitir determinar com mais rigor (i.e. ao nível da apólice) essa distribuição.

Contudo não consideramos que estejamos necessariamente na presença de uma incoerência. Uma outra leitura possível deste resultado é que os indivíduos que possuem mais produtos têm um maior gasto com seguros, tendo necessidade de redimensionar a sua carteira de seguros. Desta forma apresentam uma maior probabilidade de virem a ser *Churners*. A abordagem mais profunda deste ponto é pertinente para a investigação futura.

A nossa Rede Neuronal e a nossa Árvore incluem uma variável que diz respeito ao abandono de alguma seguradora ao longo do último ano. Este abandono não diz necessariamente respeito a um seguro automóvel, desta forma consideramos pertinente no futuro recolher informação que permita saber se o inquirido abandonou seguros automóvel em alguma companhia recentemente e/ou se o tenciona fazer num futuro próximo.

Outros aspectos que nos parecem importantes passam pela recolha, e posterior introdução nos modelos, de variáveis respeitantes à intenção de mudança de automóvel, uma vez que ao mudar de automóvel inquirido terá necessariamente de contratar um novo seguro e nesse processo escolher uma outra companhia de seguros.

Também relevante seria a introdução de métricas de satisfação, uma vez que a investigação tem encontrado uma associação elevada entre satisfação e lealdade (e.g. Coelho & Vilares, 2006). Apesar de dispormos de algumas métricas de satisfação elas dizem respeito a aspectos particulares (e.g. Grau de satisfação em relação ao Aconselhamento prestado na aquisição do seguro de automóvel) e são feitas apenas a uma pequena fracção da Base de dados (e.g. Indivíduos que subscreveram o seguro há menos de 1 ano e foram eles que decidiram fazer esse seguro) tornando difícil a sua utilização nos nossos modelos.

VI Referências Bibliográficas

Andrade, D. (2007). *Uma análise de cancelamentos em telefonia utilizando mineração de dados.* Unpublished Tese de Mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

Au, T., & Ma, G. (2003). Applying and Evaluating Models to Predict Customer Attrition Using Data Mining Techniques. *Journal of Comparative Intnational Management, 6*(1).

Bação, F. (2006). *Data Mining*.Unpublished manuscript, Lisboa.

Batista, G., & Monard, M. (2003). An analysis of four missing data treatment methods for Supervised Learning. *Applied Artificial Intelligence, 17*, 519-533.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research, 164*, 252–268.

Burez, J., & Van den Poel, D. (2007). CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications, 32*, 277–288.

Carvalho, D. (2005). *Árvore de decisão / algoritmo genético para tratar o problema de pequenos disjuntos em classificação de dados.* Unpublished Tese de Mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

Cheng, B., & Titterington, D. (1994). Neural Networks: A Review from a Statistical Perspective. *Statistical Science, 19*(1), 2-30.

Coelho, P., & Vilares, M. (2006). *Satisfação e Lealdade do Cliente: Metodologias de Avaliação, Gestão e Análise*. Lisboa: Escolar Editora.

Cortes, B. (2005). *Sistemas de suporte à decisão.* Lisboa: FCA.

Costa, A. (2000). *Técnicas de estimação no âmbito da pós-estratificação.* Unpublished Tese de Mestrado, Instituto superior de Estatística e Gestão de Informação - Universidade Nova de Lisboa, Lisboa.

Ganesh, J., Arnold, M., & Reynolds, K. (2000). Understanding the Customer Base of Service Providers: An Examination of the Differences Between Switchers and Stayer. *Journal of Marketing, 64*(3), 65-87.

Garcia, S. (2003). *O Uso de árvores de decisão na descoberta de conhecimento na área da saúde.* Unpublished Tese de Mestrado, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Gardner, M., & Droling, S. (1998). Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. *Atmospheric Environment, 32*(14/15), 2627-2636.

Geurts, P., Irrthum, A., & Wehenkel, L. (2009). Supervised learning with decision tree-based methods in computational and systems biology. *Molecular BioSystems, 5*, 1593-1605.

Graham, J. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology, 60*, 549-576.

Graham, J., Hofer, S., Donaldson, S., McKinnon, D., & Schafer, J. (1997). Analysis with missing data in prevention research. In K. Bryant, M. Windle & S. West (Eds.), *The Science of Prevention: Methodological Advances From Alcohol and Substance Abuse Research* (Vol. 1, pp. 325-366). Washington: American Psychological Association.

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2005). Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research, 34*, 2902 – 2917.

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2006). Churn Prediction: Does Technology Matter? *International Journal of Intelligent Systems and Technologies, 1*(2), 104-110.

Hongxia, M., Min, Q., & Jianxia, W. (2009, 16-19 Agosto). *Analysis of the Business Customer Churn Based on Decision Tree Method.* Paper presented at the The Ninth Conference on Electronics Measurement & Instruments, Beijing, China.

Hung, S., Yen, D., & Wag, H. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications 31*, 515-524.

Jacob, R. (1994). Why some customers are more equal than others. *Fortune, 130*(6).

Kahane, Y., Levin, R., & Zahavi, J. (2005). Applying data Mining Technology for Insurance Rate Making - An Example of Automobile Insurance. *Asia-Pacific Journal of Risk and Insurance, 2*(1).

Lazarov, V., & Capota, M. (2007). Churn prediction, *Business Analytics Course*. Munique: TUM Computer Science.

Lejeune, M. (2001). Measuring the impact of data mining on churn management. *Internet Research, 11*(6), 375 -387.

Lemos, E. (2003). *Análise de crédito bancário com o uso de data mining: redes neurais e árvores de decisão.* Unpublished Tese de Mestrado, Universidade Federal do Paraná, Curitiba.

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data*. United States of America: Jonh Wiley & Sons.

Marktest. (2006). *Basef Seguros*. Retrieved 30 Outubro de 2010, from www.marktest.pt/mpt/

Metropolis, N., & Ulam, S. (1949). The monte carlo method. *Journal of the American Statistical Association, 44*(247), pp. 335-341.

Morik, K., & Kopcke, H. (2004). Analysing Customer Churn in Insurance Data – A Case Study. In J.-F. Boulicaut (Ed.), *Proceedings of the Eight European Conference on Principles and Practice of Knowledge Discovery in Databases* (Vol. 3202, pp. 325-336). Pisa, Itália: Springer.

Paula, M. (2002). *Indução automática de árvores de decisão.* Unpublished Tese de Mestrado, Universidade Federal de Santa Catarina, Florianópolis.

Petermann, R. (2006). *Modelo de mineração de dados para classificação de clientes em telecomunicações.* Universidade Católica do Rio Grande do Sul, Porto Alegre.

Pinheiro, M. (2008). *Iniciação à Econometria*.Unpublished manuscript, Lisboa.

Quinlan, J. (1986). Induction of decision Trees. *Machine Learning, 1*, 81-106.

Roth, P. (1994). Missing data: a conceptual review for applied psychologists. *Personnel Psychology, 47*(3), 537-560.

Schafer, J., & Graham, J. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods, 7*(2), 147-177.

Staudt, M., Kietz, J., & Reimer, U. (1998). A data mining support environment and its application to insurance data. In R. Agrawal & P. Stolorz (Eds.), *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (pp. 105-111). Zurich: AAAI Press.

Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for Financial services using proportional hazard models. *European Journal of Operational Research, 157*, 196–217.

Wei, C., & Chiu, I. (2002). Turning telecommunications call details to churn prediction: a data mining approach. *Expert Systems with Applications, 23*, 103-112.

Weiss, S., & Indurkhya, N. (1998). *Predictive data mining*. United States of America: Morgan Kaufmann Publishers.

ANEXOS

ANEXO A

Listagem de Variáveis/Indicadores – Estudo de Barómetro de Seguros

Variável	Base Amostral
Notoriedade	
Notoriedade espontânea de companhias - 1ª referência	Totalidade dos Indivíduos
Notoriedade espontânea de companhias - 2ª referência	Totalidade dos Indivíduos
Notoriedade espontânea de companhias - Outras referências	Totalidade dos Indivíduos
Notoriedade espontânea de companhias -Total	Totalidade dos Indivíduos
Notoriedade total de companhias (espontâneo+sugerido)	Totalidade dos Indivíduos
Imagem	
Melhor companhia	Totalidade dos Indivíduos
Razões porque considera a melhor companhia	Indivíduos que consideram uma companhia a melhor
Pior companhia	Totalidade dos Indivíduos
Razões porque considera a pior companhia	Indivíduos que consideram uma companhia a pior
Companhia que possui melhores seguros	Totalidade dos Indivíduos
Companhia que apresenta melhores preços	Totalidade dos Indivíduos
Companhia que é mais eficiente/trabalha melhor	Totalidade dos Indivíduos
Companhia que cumpre melhor o que promete	Totalidade dos Indivíduos
Companhia que informa com mais clareza	Totalidade dos Indivíduos
Companhia que é mais inovadora	Totalidade dos Indivíduos
Companhia que é mais sólida	Totalidade dos Indivíduos
Companhia que é mais qualificada tecnicamente	Totalidade dos Indivíduos
Companhia com que se identifica mais	Totalidade dos Indivíduos
Aspecto mais importante na escolha da companhia, caso fizesse um seguro	Totalidade dos Indivíduos
Companhia que escolheria se neste momento fizesse um seguro	Totalidade dos Indivíduos
Meios de Contacto	
Meios de contacto das companhias de seguros para contactar os clientes	Indivíduos que possuem pelo menos um seguro

Meios pelos quais gostaria de ser contactado pelas companhias de seguros	Indivíduos que possuem pelo menos um seguro
Habitualmente como ou onde obtém informações sobre seguros - Sugerido	Indivíduos que possuem pelo menos um seguro
Seguros na mesma Companhia	
Possuir todos os seguros na mesma companhia é ou não vantajoso	Indivíduos que possuem pelo menos um seguro
Vantagens de possuir todos os seguros na mesma companhia	Indivíduos que possuem pelo menos um seguro e consideram vantajoso possuir todos os seguros na mesma companhia
Desvantagens de possuir todos os seguros na mesma companhia	Indivíduos que possuem pelo menos um seguro e não consideram vantajoso possuir todos os seguros na mesma companhia
Companhia Principal	
Companhia principal/ com que tem mais contacto	Indivíduos que possuem pelo menos um seguro
Aspectos que mais agradam na companhia mais importante	Indivíduos que possuem pelo menos um seguro e identificam a companhia mais importante
Aspectos que mais desagradam na companhia mais importante	Indivíduos que possuem pelo menos um seguro e identificam a companhia mais importante
Grau de satisfação com a companhia mais importante	Indivíduos que possuem pelo menos um seguro e identificam a companhia mais importante
Grau de satisfação Global com a companhia mais importante	Indivíduos que possuem pelo menos um seguro e identificam a companhia mais importante
Posse de Seguros	
Posse dos Seguros	Totalide de Indivíduos
Número de seguros possuídos	Indivíduos que possuem pelo menos 1 seguro
Companhias onde possui seguros	Indivíduos que possuem pelo menos 1 seguro
Número de companhias onde possui seguro	Indivíduos que possuem pelo menos 1 seguro
Seguro de Saúde	
Posse/Benefício de seguro de Saúde	Totalidade dos indivíduos
Número de seguros de saúde que possui/beneficia	Indivíduos que possuem seguro de saúde
Principal motivo porque decidiu fazer um seguro de Saúde	Indivíduos que possuem seguro de saúde

Companhias onde possui/beneficia de seguro de saúde	Indivíduos que possuem seguro de saúde
Seguro Automóvel	
Posse/benefício de seguro Automóvel - Conduz regularmente	Totalidade dos Indivíduos
Número de seguros de automóvel que possui/beneficia	Indivíduos que possuem seguro automóvel
Local onde fez contrato do seguro de automóvel que utiliza com mais frequência	Indivíduos que possuem seguro de automóvel
Banco onde fez seguro de Automóvel que utiliza com mais frequência	Indivíduos que possuem seguro de automóvel feito através do banco
Tipo de veiculo que conduz regularmente	Indivíduos que possuem/beneficiam seguro de automóvel e conduzem regularmente
Companhias onde possui/beneficia de seguro de automóvel	Indivíduos que possuem seguro de automóvel
Companhia onde possui ou beneficia do seguro de Automóvel que utiliza com mais frequência	Indivíduos que possuem seguro de automóvel
Razões de escolha da companhia do seguro Automóvel	Indivíduos que possuem seguro de automóvel e identificam a companhia onde possuem o seg.automóvel de maior prémio
Nome em que está o seguro Automóvel que possui ou beneficia	Indivíduos que possuem seguro de automóvel
Coberturas garantidas pelo seguro Automóvel	Indivíduos que possuem seguro de automóvel
Capital Responsabilidade Civil do seguro Automóvel	Indivíduos que possuem seguro de automóvel
Quem decidiu fazer o contrato de seguro de Automóvel	Indivíduos que possuem seguro de automóvel
Tempo de posse do seguro Automóvel	Indivíduos que possuem seguro de automóvel
Grau de satisfação em relação ao Atendimento prestado na aquisição do seguro de automóvel, em termos de comportamento, disponibilidade e compreensão por parte das pessoas que trataram do seu caso	Indivíduos que subscreveram o seguro há menos de 1 ano e foram eles que decidiram fazer esse seguro
Grau de satisfação em relação ao Aconselhamento prestado na aquisição do seguro de automóvel: incluindo a adequação dos conselhos às suas necessidades e frequência com que o mantiveram informado	Indivíduos que subscreveram o seguro há menos de 1 ano e foram eles que decidiram fazer esse seguro
Grau de satisfação com a Eficácia, isto é, rapidez, clareza	Indivíduos que subscreveram o seguro

e qualidade de resposta, prestada na aquisição do seguro de automóvel	há menos de 1 ano e foram eles que decidiram fazer esse seguro
Prémio anual do seguro Automóvel	Indivíduos que possuem seguro de automóvel
Canal de contacto para tratar de algum assunto do seguro de Automóvel	Indivíduos que possuem seguro de automóvel
Grau de satisfação com o canal utilizado para o contacto do seguro Automóvel - Global	Indivíduos que possuem seguro de automóvel e já contactaram
Grau de satisfação com o canal utilizado para o contacto do seguro Automóvel - Companhia	Indivíduos que possuem seguro de automóvel e já contactaram a companhia quando querem tratar algum assunto
Grau de satisfação com o canal utilizado para o contacto do seguro Automóvel- Mediador	Indivíduos que possuem seguro de automóvel e já contactaram o mediador quando querem tratar algum assunto
Participação de algum sinistro deste seguro à companhia	Indivíduos que possuem seguro de automóvel
Número de participações de sinistros automóvel nos últimos 5 anos	Indivíduos que possuem seguro de automóvel e já fizeram participação de um sinistro de automóvel
Há quanto tempo fez a última participação de sinistro à sua companhia	Indivíduos que possuem seguro de automóvel e referiram o nº de participações de sinistro nos últimos 5 anos
Foi considerado culpado ou não em algum deles	Indivíduos que possuem seguro de automóvel e já fizeram participação de um sinistro de automóvel nos últimos 5 anos
Grau de satisfação em relação ao Atendimento prestado no processo de resolução de sinistro automóvel, quer ao nível de comportamento, de disponibilidade e compreensão por parte das pessoas que trataram do seu caso	Indivíduos que possuem seguro de automóvel e já fizeram participação de um sinistro de automóvel há menos de um ano
Grau de satisfação em relação ao Aconselhamento prestado no processo de resolução de sinistro automóvel: incluindo a adequação dos conselhos às suas necessidades e frequência com que o mantiveram informado	Indivíduos que possuem seguro de automóvel e já fizeram participação de um sinistro de automóvel há menos de um ano
Grau de satisfação em termos de Eficácia, isto é, de rapidez, de clareza e de qualidade de resposta, no processo de resolução de sinistro automóvel	Indivíduos que possuem seguro de automóvel e já fizeram participação de um sinistro de automóvel há menos de um ano

Classificação da companhia no processo de resolução do sinistro automóvel quanto ao ser...	Indivíduos que possuem seguro de automóvel e que já fizeram participação de um sinistro de automóvel há menos de um ano
Seguro de Acidentes Pessoais	
Posse/Benefício de seguros de Acidentes Pessoais	Totalidade dos indivíduos
Número de seguros de acidentes pessoais que possui/beneficia	Indivíduos que possuem seguro de acidentes pessoais
Companhias onde possui/beneficia de seguro de acidentes pessoais	Indivíduos que possuem seguro de acidentes pessoais
Ramo Vida	
Principal motivo porque decidiu fazer seguro Ramo Vida	Indivíduos que possuem pelo menos um seguro do ramo vida
Seguro Vida Puro Risco	
Posse/Benefício de seguro de Vida Puro Risco	Totalidade dos indivíduos
Número de seguros de vida puro risco que possui/beneficia	Indivíduos que possuem seguro de vida puro risco
Companhias onde possui/beneficia de seguro de vida puro risco	Indivíduos que possuem seguro de vida puro risco
Seguro Vida Misto	
Posse/Benefício de seguro Vida Misto	Totalidade dos indivíduos
Número de seguros de vida misto	Indivíduos que possuem seguro de vida misto
Companhias onde possui/beneficia de seguro de vida misto	Indivíduos que possuem seguro de vida misto
Seguro PPR/E	
Posse/Benefício de PPR/E	Totalidade dos indivíduos
Número de PPR/E	Indivíduos que possuem PPR/E
Companhias onde possui/beneficia PPR/E	Indivíduos que possuem PPR/E
Companhia onde possui ou beneficia do PPR/E que tem o maior prémio	Indivíduos que possuem PPR/E
Seguro de Capitalização	
Posse/Benefício de seguro de Capitalização	Totalidade dos indivíduos
Número de seguros de capitalização	Indivíduos que possuem seguro de capitalização
Companhias onde possui/beneficia de seguro de capitalização	Indivíduos que possuem seguro de capitalização
Seguro Acidentes de Trabalho	
Posse/benefício de Seguro de Acidentes de Trabalho	Totalidade dos indivíduos

Número de seguros de acidentes de trabalho	Indivíduos que possuem seguro de acidentes de trabalho
Companhias onde possui/beneficia de seguro de acidentes de trabalho	Indivíduos que possuem seguro de acidentes de trabalho
Seguro de Caça	
Posse/benefício de Seguro de Caça	
Companhias onde possui/beneficia de seguro de caça	Indivíduos que possuem seguro de caça
Seguro de Animais Domésticos	
Posse/Benefício de seguros de animais domésticos	Totalidade dos indivíduos
Número de seguros de animais domésticos	Indivíduos que possuem seguro de animais domésticos
Companhias onde possui/beneficia de seguro de animais domésticos	Indivíduos que possuem seguro de animais domésticos
Abandono	
No último ano deixou de trabalhar com alguma companhia	Indivíduos que possuem pelo menos 1 seguro
Razões porque deixou de trabalhar com a	Indivíduos que possuem pelo menos um seguro e deixaram de trabalhar com uma companhia no último ano
Perfil	
Género	Indivíduos que possuem pelo menos um seguro feito através de mediador
Idade	Indivíduos que possuem pelo menos um seguro feito através de mediador e mudariam de companhia se o mediador aconselhasse
Região	Indivíduos que possuem pelo menos um seguro feito através do mediador e não mudariam de companhia se o mediador aconselhasse
Ocupação	Totalidade dos entrevistados
Instrução	Totalidade dos Indivíduos
Classe Social	Indivíduos que possuem pelo menos um seguro

ANEXO B

Listagem de Variáveis/Indicadores – Estudo de Transferências de Seguros

Variável	Base Amostral
Companhia onde possui seg. Automóvel que utiliza com mais frequência	Indivíduos que residem em lares com rede fixa e possuem seguro de automóvel
Mudança de companhia ou entidade onde possuia o seguro de automóvel nos últimos 12 meses	Indivíduos que possuem seguro de automóvel e identificam a companhia do seguro de automóvel de maior prémio
Razões de mudança Cª onde possuia Seg. Automóvel utiliza com mais frequência	Indivíduos que residem em lares com rede fixa e que mudaram de companhia do seguro de automóvel utilizado com mais frequência, nos últimos 12 meses

ANEXO C

Autorização Formal para uso dos Dados

ANEXO D

Output de Análise Descritiva

```
FREQUENCIES
  VARIABLES=mud_seg
  /ORDER=  ANALYSIS .
```

Frequencies

[DataSet1] D:\My Documents\Tese de Mestrado\Experiencia Boosting\Transf_Automovel_Modelo_Boosting.sav

Statistics

Mudou de Seguradora

N	Valid	1661
	Missing	0

Mudou de Seguradora

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Não	1336	80,4	80,4	80,4
	Sim	325	19,6	19,6	100,0
	Total	1661	100,0	100,0	

```
CROSSTABS
  /TABLES=Genero  BY mud_seg
  /FORMAT= AVALUE TABLES
  /CELLS= COUNT ROW COLUMN
  /COUNT ROUND CELL .
```

Crosstabs

[DataSet1] D:\My Documents\Tese de Mestrado\Experiencia Boosting\Transf_Automovel_Modelo_Boosting.sav

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Género do Inquirido * Mudou de Seguradora	1661	100,0%	0	,0%	1661	100,0%

Género do Inquirido * Mudou de Seguradora Crosstabulation

			Mudou de Seguradora		
			Não	Sim	Total
Género do Inquirido	Feminino	Count	361	74	435
		% within Género do Inquirido	83,0%	17,0%	100,0%
		% within Mudou de Seguradora	27,0%	22,8%	26,2%
	Masculino	Count	975	251	1226
		% within Género do Inquirido	79,5%	20,5%	100,0%
		% within Mudou de Seguradora	73,0%	77,2%	73,8%
Total		Count	1336	325	1661
		% within Género do Inquirido	80,4%	19,6%	100,0%
		% within Mudou de Seguradora	100,0%	100,0%	100,0%

```
CROSSTABS
  /TABLES=Idade  BY mud_seg
  /FORMAT= AVALUE TABLES
  /CELLS= COUNT ROW COLUMN
  /COUNT ROUND CELL .
```

Crosstabs

[DataSet1] D:\My Documents\Tese de Mestrado\Experiencia Boosting\Transf_Automovel_Modelo_Boosting.sav

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Idade do Inquirido * Mudou de Seguradora	1661	100,0%	0	,0%	1661	100,0%

Idade do Inquirido * Mudou de Seguradora Crosstabulation

			Mudou de Seguradora		
			Não	Sim	Total
Idade do Inquirido	18	Count	3	1	4
		% within Idade do Inquirido	75,0%	25,0%	100,0%
		% within Mudou de Seguradora	,2%	,3%	,2%
	19	Count	4	0	4
		% within Idade do Inquirido	100,0%	,0%	100,0%
		% within Mudou de Seguradora	,3%	,0%	,2%
	20	Count	4	3	7
		% within Idade do Inquirido	57,1%	42,9%	100,0%
		% within Mudou de Seguradora	,3%	,9%	,4%
	21	Count	2	1	3
		% within Idade do Inquirido	66,7%	33,3%	100,0%
		% within Mudou de Seguradora	,1%	,3%	,2%
	22	Count	6	3	9
		% within Idade do Inquirido	66,7%	33,3%	100,0%
		% within Mudou de Seguradora	,4%	,9%	,5%
	23	Count	6	2	8
		% within Idade do Inquirido	75,0%	25,0%	100,0%
		% within Mudou de Seguradora	,4%	,6%	,5%
	24	Count	13	5	18
		% within Idade do Inquirido	72,2%	27,8%	100,0%
		% within Mudou de Seguradora	1,0%	1,5%	1,1%

Idade do Inquirido * Mudou de Seguradora Crosstabulation

			Mudou de Seguradora		
			Não	Sim	Total
Idade do Inquirido	25	Count	23	3	26
		% within Idade do Inquirido	88,5%	11,5%	100,0%
		% within Mudou de Seguradora	1,7%	,9%	1,6%
	26	Count	22	15	37
		% within Idade do Inquirido	59,5%	40,5%	100,0%
		% within Mudou de Seguradora	1,6%	4,6%	2,2%
	27	Count	27	7	34
		% within Idade do Inquirido	79,4%	20,6%	100,0%
		% within Mudou de Seguradora	2,0%	2,2%	2,0%
	28	Count	31	4	35
		% within Idade do Inquirido	88,6%	11,4%	100,0%
		% within Mudou de Seguradora	2,3%	1,2%	2,1%
	29	Count	20	3	23
		% within Idade do Inquirido	87,0%	13,0%	100,0%
		% within Mudou de Seguradora	1,5%	,9%	1,4%
	30	Count	31	13	44
		% within Idade do Inquirido	70,5%	29,5%	100,0%
		% within Mudou de Seguradora	2,3%	4,0%	2,6%
	31	Count	19	3	22
		% within Idade do Inquirido	86,4%	13,6%	100,0%
		% within Mudou de Seguradora	1,4%	,9%	1,3%
	32	Count	26	5	31
		% within Idade do Inquirido	83,9%	16,1%	100,0%
		% within Mudou de Seguradora	1,9%	1,5%	1,9%
	33	Count	34	12	46
		% within Idade do Inquirido	73,9%	26,1%	100,0%
		% within Mudou de Seguradora	2,5%	3,7%	2,8%
	34	Count	30	7	37
		% within Idade do Inquirido	81,1%	18,9%	100,0%
		% within Mudou de Seguradora	2,2%	2,2%	2,2%
	35	Count	33	8	41
		% within Idade do Inquirido	80,5%	19,5%	100,0%
		% within Mudou de Seguradora	2,5%	2,5%	2,5%

Idade do Inquirido * Mudou de Seguradora Crosstabulation

			Mudou de Seguradora		
			Não	Sim	Total
Idade do Inquirido	36	Count	21	8	29
		% within Idade do Inquirido	72,4%	27,6%	100,0%
		% within Mudou de Seguradora	1,6%	2,5%	1,7%
	37	Count	26	5	31
		% within Idade do Inquirido	83,9%	16,1%	100,0%
		% within Mudou de Seguradora	1,9%	1,5%	1,9%
	38	Count	29	5	34
		% within Idade do Inquirido	85,3%	14,7%	100,0%
		% within Mudou de Seguradora	2,2%	1,5%	2,0%
	39	Count	29	6	35
		% within Idade do Inquirido	82,9%	17,1%	100,0%
		% within Mudou de Seguradora	2,2%	1,8%	2,1%
	40	Count	45	13	58
		% within Idade do Inquirido	77,6%	22,4%	100,0%
		% within Mudou de Seguradora	3,4%	4,0%	3,5%
	41	Count	33	4	37
		% within Idade do Inquirido	89,2%	10,8%	100,0%
		% within Mudou de Seguradora	2,5%	1,2%	2,2%
	42	Count	38	7	45
		% within Idade do Inquirido	84,4%	15,6%	100,0%
		% within Mudou de Seguradora	2,8%	2,2%	2,7%
	43	Count	33	6	39
		% within Idade do Inquirido	84,6%	15,4%	100,0%
		% within Mudou de Seguradora	2,5%	1,8%	2,3%
	44	Count	33	14	47
		% within Idade do Inquirido	70,2%	29,8%	100,0%
		% within Mudou de Seguradora	2,5%	4,3%	2,8%
	45	Count	28	11	39
		% within Idade do Inquirido	71,8%	28,2%	100,0%
		% within Mudou de Seguradora	2,1%	3,4%	2,3%
	46	Count	22	10	32
		% within Idade do Inquirido	68,8%	31,3%	100,0%
		% within Mudou de Seguradora	1,6%	3,1%	1,9%

Idade do Inquirido * Mudou de Seguradora Crosstabulation

			Mudou de Seguradora		
			Não	Sim	Total
Idade do Inquirido	47	Count	20	3	23
		% within Idade do Inquirido	87,0%	13,0%	100,0%
		% within Mudou de Seguradora	1,5%	,9%	1,4%
	48	Count	34	7	41
		% within Idade do Inquirido	82,9%	17,1%	100,0%
		% within Mudou de Seguradora	2,5%	2,2%	2,5%
	49	Count	26	7	33
		% within Idade do Inquirido	78,8%	21,2%	100,0%
		% within Mudou de Seguradora	1,9%	2,2%	2,0%
	50	Count	36	8	44
		% within Idade do Inquirido	81,8%	18,2%	100,0%
		% within Mudou de Seguradora	2,7%	2,5%	2,6%
	51	Count	20	5	25
		% within Idade do Inquirido	80,0%	20,0%	100,0%
		% within Mudou de Seguradora	1,5%	1,5%	1,5%
	52	Count	25	6	31
		% within Idade do Inquirido	80,6%	19,4%	100,0%
		% within Mudou de Seguradora	1,9%	1,8%	1,9%
	53	Count	25	10	35
		% within Idade do Inquirido	71,4%	28,6%	100,0%
		% within Mudou de Seguradora	1,9%	3,1%	2,1%
	54	Count	30	9	39
		% within Idade do Inquirido	76,9%	23,1%	100,0%
		% within Mudou de Seguradora	2,2%	2,8%	2,3%
	55	Count	32	5	37
		% within Idade do Inquirido	86,5%	13,5%	100,0%
		% within Mudou de Seguradora	2,4%	1,5%	2,2%
	56	Count	21	7	28
		% within Idade do Inquirido	75,0%	25,0%	100,0%
		% within Mudou de Seguradora	1,6%	2,2%	1,7%
	57	Count	28	11	39
		% within Idade do Inquirido	71,8%	28,2%	100,0%
		% within Mudou de Seguradora	2,1%	3,4%	2,3%

Idade do Inquirido * Mudou de Seguradora Crosstabulation

			Mudou de Seguradora		
			Não	Sim	Total
Idade do Inquirido	58	Count	26	4	30
		% within Idade do Inquirido	86,7%	13,3%	100,0%
		% within Mudou de Seguradora	1,9%	1,2%	1,8%
	59	Count	22	3	25
		% within Idade do Inquirido	88,0%	12,0%	100,0%
		% within Mudou de Seguradora	1,6%	,9%	1,5%
	60	Count	23	6	29
		% within Idade do Inquirido	79,3%	20,7%	100,0%
		% within Mudou de Seguradora	1,7%	1,8%	1,7%
	61	Count	25	4	29
		% within Idade do Inquirido	86,2%	13,8%	100,0%
		% within Mudou de Seguradora	1,9%	1,2%	1,7%
	62	Count	18	5	23
		% within Idade do Inquirido	78,3%	21,7%	100,0%
		% within Mudou de Seguradora	1,3%	1,5%	1,4%
	63	Count	18	4	22
		% within Idade do Inquirido	81,8%	18,2%	100,0%
		% within Mudou de Seguradora	1,3%	1,2%	1,3%
	64	Count	19	2	21
		% within Idade do Inquirido	90,5%	9,5%	100,0%
		% within Mudou de Seguradora	1,4%	,6%	1,3%
	65	Count	27	3	30
		% within Idade do Inquirido	90,0%	10,0%	100,0%
		% within Mudou de Seguradora	2,0%	,9%	1,8%
	66	Count	15	1	16
		% within Idade do Inquirido	93,8%	6,3%	100,0%
		% within Mudou de Seguradora	1,1%	,3%	1,0%
	67	Count	19	3	22
		% within Idade do Inquirido	86,4%	13,6%	100,0%
		% within Mudou de Seguradora	1,4%	,9%	1,3%
	68	Count	12	1	13
		% within Idade do Inquirido	92,3%	7,7%	100,0%
		% within Mudou de Seguradora	,9%	,3%	,8%

Idade do Inquirido * Mudou de Seguradora Crosstabulation

			Mudou de Seguradora		
			Não	Sim	Total
Idade do Inquirido	69	Count	14	2	16
		% within Idade do Inquirido	87,5%	12,5%	100,0%
		% within Mudou de Seguradora	1,0%	,6%	1,0%
	70	Count	17	2	19
		% within Idade do Inquirido	89,5%	10,5%	100,0%
		% within Mudou de Seguradora	1,3%	,6%	1,1%
	71	Count	13	3	16
		% within Idade do Inquirido	81,3%	18,8%	100,0%
		% within Mudou de Seguradora	1,0%	,9%	1,0%
	72	Count	20	3	23
		% within Idade do Inquirido	87,0%	13,0%	100,0%
		% within Mudou de Seguradora	1,5%	,9%	1,4%
	73	Count	14	3	17
		% within Idade do Inquirido	82,4%	17,6%	100,0%
		% within Mudou de Seguradora	1,0%	,9%	1,0%
	74	Count	12	5	17
		% within Idade do Inquirido	70,6%	29,4%	100,0%
		% within Mudou de Seguradora	,9%	1,5%	1,0%
	75	Count	17	2	19
		% within Idade do Inquirido	89,5%	10,5%	100,0%
		% within Mudou de Seguradora	1,3%	,6%	1,1%
	76	Count	7	4	11
		% within Idade do Inquirido	63,6%	36,4%	100,0%
		% within Mudou de Seguradora	,5%	1,2%	,7%
	77	Count	13	1	14
		% within Idade do Inquirido	92,9%	7,1%	100,0%
		% within Mudou de Seguradora	1,0%	,3%	,8%
	78	Count	13	0	13
		% within Idade do Inquirido	100,0%	,0%	100,0%
		% within Mudou de Seguradora	1,0%	,0%	,8%
	79	Count	4	2	6
		% within Idade do Inquirido	66,7%	33,3%	100,0%
		% within Mudou de Seguradora	,3%	,6%	,4%
Total		Count	1336	325	1661
		% within Idade do Inquirido	80,4%	19,6%	100,0%
		% within Mudou de Seguradora	100,0%	100,0%	100,0%

```
MEANS
  TABLES=Idade  BY mud_seg
  /CELLS MEAN COUNT STDDEV  .
```

Means

```
[DataSet1] D:\My Documents\Tese de Mestrado\Experiencia Boosting\Transf_
Automovel_Modelo_Boosting.sav
```

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Idade do Inquirido * Mudou de Seguradora	1661	100,0%	0	,0%	1661	100,0%

Report

Idade do Inquirido

Mudou de Seguradora	Mean	N	Std. Deviation
Não	47,66	1336	14,781
Sim	45,42	325	14,315
Total	47,22	1661	14,713

```
CROSSTABS
  /TABLES=num_auto  BY mud_seg
  /FORMAT= AVALUE TABLES
  /CELLS= COUNT ROW COLUMN
  /COUNT ROUND CELL .
```

Crosstabs

```
[DataSet1] D:\My Documents\Tese de Mestrado\Experiencia Boosting\Transf_
Automovel_Modelo_Boosting.sav
```

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Número de Seguros Automóvel que possui * Mudou de Seguradora	1661	100,0%	0	,0%	1661	100,0%

Número de Seguros Automóvel que possui * Mudou de Seguradora Crosstabulation

			Mudou de Seguradora		
			Não	Sim	Total
Número de Seguros Automóvel que possui	1	Count	895	212	1107
		% within Número de Seguros Automóvel que possui	80,8%	19,2%	100,0%
		% within Mudou de Seguradora	67,0%	65,2%	66,6%
	2	Count	330	77	407
		% within Número de Seguros Automóvel que possui	81,1%	18,9%	100,0%
		% within Mudou de Seguradora	24,7%	23,7%	24,5%
	3	Count	81	27	108
		% within Número de Seguros Automóvel que possui	75,0%	25,0%	100,0%
		% within Mudou de Seguradora	6,1%	8,3%	6,5%
	4	Count	21	8	29
		% within Número de Seguros Automóvel que possui	72,4%	27,6%	100,0%
		% within Mudou de Seguradora	1,6%	2,5%	1,7%
	5	Count	9	1	10
		% within Número de Seguros Automóvel que possui	90,0%	10,0%	100,0%
		% within Mudou de Seguradora	,7%	,3%	,6%
Total		Count	1336	325	1661
		% within Número de Seguros Automóvel que possui	80,4%	19,6%	100,0%
		% within Mudou de Seguradora	100,0%	100,0%	100,0%

```
MEANS
  TABLES=num_auto  BY mud_seg
  /CELLS MEAN COUNT STDDEV  .
```

Means

[DataSet1] D:\My Documents\Tese de Mestrado\Experiencia Boosting\Transf_Automovel_Modelo_Boosting.sav

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Número de Seguros Automóvel que possui * Mudou de Seguradora	1661	100,0%	0	,0%	1661	100,0%

Report

Número de Seguros Automóvel que possui

Mudou de Seguradora	Mean	N	Std. Deviation
Não	1,44	1336	,737
Sim	1,49	325	,776
Total	1,45	1661	,745

```
CROSSTABS
  /TABLES=rc_mais  BY mud_seg
  /FORMAT= AVALUE TABLES
  /CELLS= COUNT ROW COLUMN
  /COUNT ROUND CELL .
```

Crosstabs

```
[DataSet1] D:\My Documents\Tese de Mestrado\Experiencia Boosting\Transf_
Automovel_Modelo_Boosting.sav
```

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Coberturas Extra * Mudou de Seguradora	1630	98,1%	31	1,9%	1661	100,0%

Coberturas Extra * Mudou de Seguradora Crosstabulation

			Mudou de Seguradora		
			Não	Sim	Total
Coberturas Extra	Não	Count	954	239	1193
		% within Coberturas Extra	80,0%	20,0%	100,0%
		% within Mudou de Seguradora	72,9%	74,5%	73,2%
	Sim	Count	355	82	437
		% within Coberturas Extra	81,2%	18,8%	100,0%
		% within Mudou de Seguradora	27,1%	25,5%	26,8%
Total		Count	1309	321	1630
		% within Coberturas Extra	80,3%	19,7%	100,0%
		% within Mudou de Seguradora	100,0%	100,0%	100,0%

```
CROSSTABS
  /TABLES=rc_danosproprios  BY mud_seg
  /FORMAT= AVALUE TABLES
  /CELLS= COUNT ROW COLUMN
  /COUNT ROUND CELL .
```

Crosstabs

```
[DataSet1] D:\My Documents\Tese de Mestrado\Experiencia Boosting\Transf_
Automovel_Modelo_Boosting.sav
```

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Danos próprios * Mudou de Seguradora	1630	98,1%	31	1,9%	1661	100,0%

Danos próprios * Mudou de Seguradora Crosstabulation

			Mudou de Seguradora		
			Não	Sim	Total
Danos próprios	Não	Count	1081	257	1338
		% within Danos próprios	80,8%	19,2%	100,0%
		% within Mudou de Seguradora	82,6%	80,1%	82,1%
	Sim	Count	228	64	292
		% within Danos próprios	78,1%	21,9%	100,0%
		% within Mudou de Seguradora	17,4%	19,9%	17,9%
Total		Count	1309	321	1630
		% within Danos próprios	80,3%	19,7%	100,0%
		% within Mudou de Seguradora	100,0%	100,0%	100,0%

Coberturas Contratadas

	Não Mudou de seguradora	Mudou de sguradora	Total
Só Responsabilidade Civil	726	175	901
Responsabilidade Civil + outras coberturas que não danos próprios	355	82	437
Responsabilidade Civil + Danos próprios	228	64	292
Total	1309	321	1630

	Não Mudou de seguradora	Mudou de sguradora	Total
Só Responsabilidade Civil	55,46218487	54,51713396	55,27607
Responsabilidade Civil + outras coberturas que não danos próprios	27,11993888	25,54517134	26,80982
Responsabilidade Civil + Danos próprios	17,41787624	19,9376947	17,91411
Total	100	100	100

```
CROSSTABS
  /TABLES=participacao  BY mud_seg
  /FORMAT= AVALUE TABLES
  /CELLS= COUNT ROW COLUMN
  /COUNT ROUND CELL .
```

Crosstabs

```
[DataSet1] D:\My Documents\Tese de Mestrado\Experiencia Boosting\Transf_
Automovel_Modelo_Boosting.sav
```

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Participação de sinistros * Mudou de Seguradora	1543	92,9%	118	7,1%	1661	100,0%

Participação de sinistros * Mudou de Seguradora Crosstabulation

			Mudou de Seguradora		
			Não	Sim	Total
Participação de sinistros	Não	Count	925	237	1162
		% within Participação de sinistros	79,6%	20,4%	100,0%
		% within Mudou de Seguradora	75,9%	72,9%	75,3%
	Sim	Count	293	88	381
		% within Participação de sinistros	76,9%	23,1%	100,0%
		% within Mudou de Seguradora	24,1%	27,1%	24,7%
Total		Count	1218	325	1543
		% within Participação de sinistros	78,9%	21,1%	100,0%
		% within Mudou de Seguradora	100,0%	100,0%	100,0%

```
CROSSTABS
  /TABLES=mud_seguradora  BY mud_seg
  /FORMAT= AVALUE TABLES
  /CELLS= COUNT ROW COLUMN
  /COUNT ROUND CELL .
```

Crosstabs

```
[DataSet1] D:\My Documents\Tese de Mestrado\Experiencia Boosting\Transf_
Automovel_Modelo_Boosting.sav
```

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Mudou alguma seguradora no último ano * Mudou de Seguradora	1659	99,9%	2	,1%	1661	100,0%

Mudou alguma seguradora no último ano * Mudou de Seguradora Crosstabulation

			Mudou de Seguradora		
			Não	Sim	Total
Mudou alguma seguradora no último ano	Não	Count	1262	302	1564
		% within Mudou alguma seguradora no último ano	80,7%	19,3%	100,0%
		% within Mudou de Seguradora	94,6%	92,9%	94,3%
	Sim	Count	72	23	95
		% within Mudou alguma seguradora no último ano	75,8%	24,2%	100,0%
		% within Mudou de Seguradora	5,4%	7,1%	5,7%
Total		Count	1334	325	1659
		% within Mudou alguma seguradora no último ano	80,4%	19,6%	100,0%
		% within Mudou de Seguradora	100,0%	100,0%	100,0%

```
CROSSTABS
  /TABLES=premio  BY mud_seg
  /FORMAT= AVALUE TABLES
  /CELLS= COUNT ROW COLUMN
  /COUNT ROUND CELL .
```

Crosstabs

```
[DataSet1] D:\My Documents\Tese de Mestrado\Experiencia Boosting\Transf_
Automovel_Modelo_Boosting.sav
```

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Prémio anual do seguro Automóvel * Mudou de Seguradora	1361	81,9%	300	18,1%	1661	100,0%

Prémio anual do seguro Automóvel * Mudou de Seguradora Crosstabulation

			Mudou de Seguradora		
			Não	Sim	Total
Prémio anual do seguro Automóvel	101 a 250 euros	Count	455	100	555
		% within Prémio anual do seguro Automóvel	82,0%	18,0%	100,0%
		% within Mudou de Seguradora	41,7%	37,2%	40,8%
	251 a 500 euros	Count	468	118	586
		% within Prémio anual do seguro Automóvel	79,9%	20,1%	100,0%
		% within Mudou de Seguradora	42,9%	43,9%	43,1%
	501 a 750 euros	Count	101	27	128
		% within Prémio anual do seguro Automóvel	78,9%	21,1%	100,0%
		% within Mudou de Seguradora	9,2%	10,0%	9,4%
	751 a 1000 euros	Count	44	19	63
		% within Prémio anual do seguro Automóvel	69,8%	30,2%	100,0%
		% within Mudou de Seguradora	4,0%	7,1%	4,6%
	Mais de 1000 euros	Count	24	5	29
		% within Prémio anual do seguro Automóvel	82,8%	17,2%	100,0%
		% within Mudou de Seguradora	2,2%	1,9%	2,1%
Total		Count	1092	269	1361
		% within Prémio anual do seguro Automóvel	80,2%	19,8%	100,0%
		% within Mudou de Seguradora	100,0%	100,0%	100,0%

ANEXO E

Listagem de Variáveis/Indicadores Criados

Nome	Descrição	Tipo de variável
MAIS_CONTACTO	A seguradora do automóvel que utiliza com mais frequência é aquela com que tem mais contacto. *Casos válidos: 1656; Casos com Missing Values: 5*	Binária (0-Não; 1-Sim)
PREMIO_250	Prémio do Seguro Automóvel. *Casos válidos: 1361; Casos com Missing Values: 300*	Binária (0-Até 250 Euros; 1- Mais de 250 Euros)
COB_EXTRA	A apólice do inquirido inclui coberturas extra (seguro de viagem, danos próprios, etc.). *Casos válidos: 1630; Casos com Missing Values: 31*	Binária (0-Não; 1-Sim)
N_AUTO	Número de Seguros Automóvel	Binária (0- Um Seguro Automóvel; 1- Mais de um Seguro)
CLIENTE_EXCLUSIVO	Cliente Exclusivo ao nível do Seguro Automóvel.	Binária (0-Não é Cliente Exclusivo; 1- Cliente Exclusivo)
SAUDE	Possui um seguro de saúde na mesma Companhia que seguro do automóvel que utiliza com mais frequência.	Binária (0-Não; 1-Sim)
HABITACAO	Possui um seguro de habitação na mesma Companhia que seguro do automóvel que utiliza com mais frequência.	Binária (0-Não; 1-Sim)
AP	Possui um seguro de Acidentes Pessoais na mesma Companhia que seguro do automóvel que utiliza com mais frequência.	Binária (0-Não; 1-Sim)
RC	Possui um seguro de Responsabilidade Civil na mesma Companhia que seguro do automóvel que utiliza com mais frequência.	Binária (0-Não; 1-Sim)

Nome	Descrição	Tipo de variável
PURO_RISCO	Possui um seguro de Vida Puro Risco na mesma Companhia que seguro do automóvel que utiliza com mais frequência.	Binária (0-Não; 1-Sim)
VIDA_MISTO	Possui um seguro de Vida Misto na mesma Companhia que seguro do automóvel que utiliza com mais frequência.	Binária (0-Não; 1-Sim)
PPRE	Possui um seguro de PPRE na mesma Companhia que seguro do automóvel que utiliza com mais frequência.	Binária (0-Não; 1-Sim)
CAPITALIZACAO	Possui um seguro de Capitalização na mesma Companhia que seguro do automóvel que utiliza com mais frequência.	Binária (0-Não; 1-Sim)
TRABALHO	Possui um seguro de Acidentes de Trabalho na mesma Companhia que seguro do automóvel que utiliza com mais frequência.	Binária (0-Não; 1-Sim)
CONDOMINIO	Possui um seguro de Condomínio na mesma Companhia que seguro do automóvel que utiliza com mais frequência.	Binária (0-Não; 1-Sim)
ANIMAIS	Possui um seguro de Animais domésticos na mesma Companhia que seguro do automóvel que utiliza com mais frequência.	Binária (0-Não; 1-Sim)
CACA	Possui um seguro de Caça na mesma Companhia que seguro do automóvel que utiliza com mais frequência.	Binária (0-Não; 1-Sim)
TOTAL	Número de produtos diferentes que possui na mesma Companhia que seguro do automóvel que utiliza com mais frequência.	Numérica
VIDA	Número de produtos diferentes, do ramo vida, que possui na mesma Companhia que seguro do automóvel que utiliza com mais frequência.	Numérica
OUTROS_PROD	Número de produtos diferentes, do ramo não vida, que possui na mesma Companhia que seguro do automóvel que utiliza com mais frequência.	Numérica
RC_MAIS	O capital de Responsabilidade Civil é superior ao obrigatório	Binária (0-Não; 1-Sim)

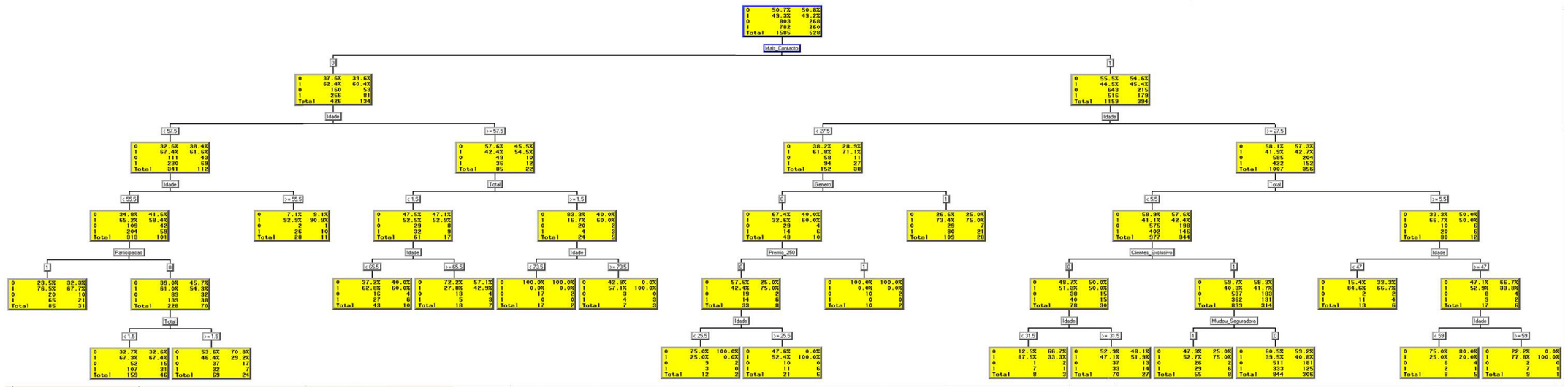
ANEXO F

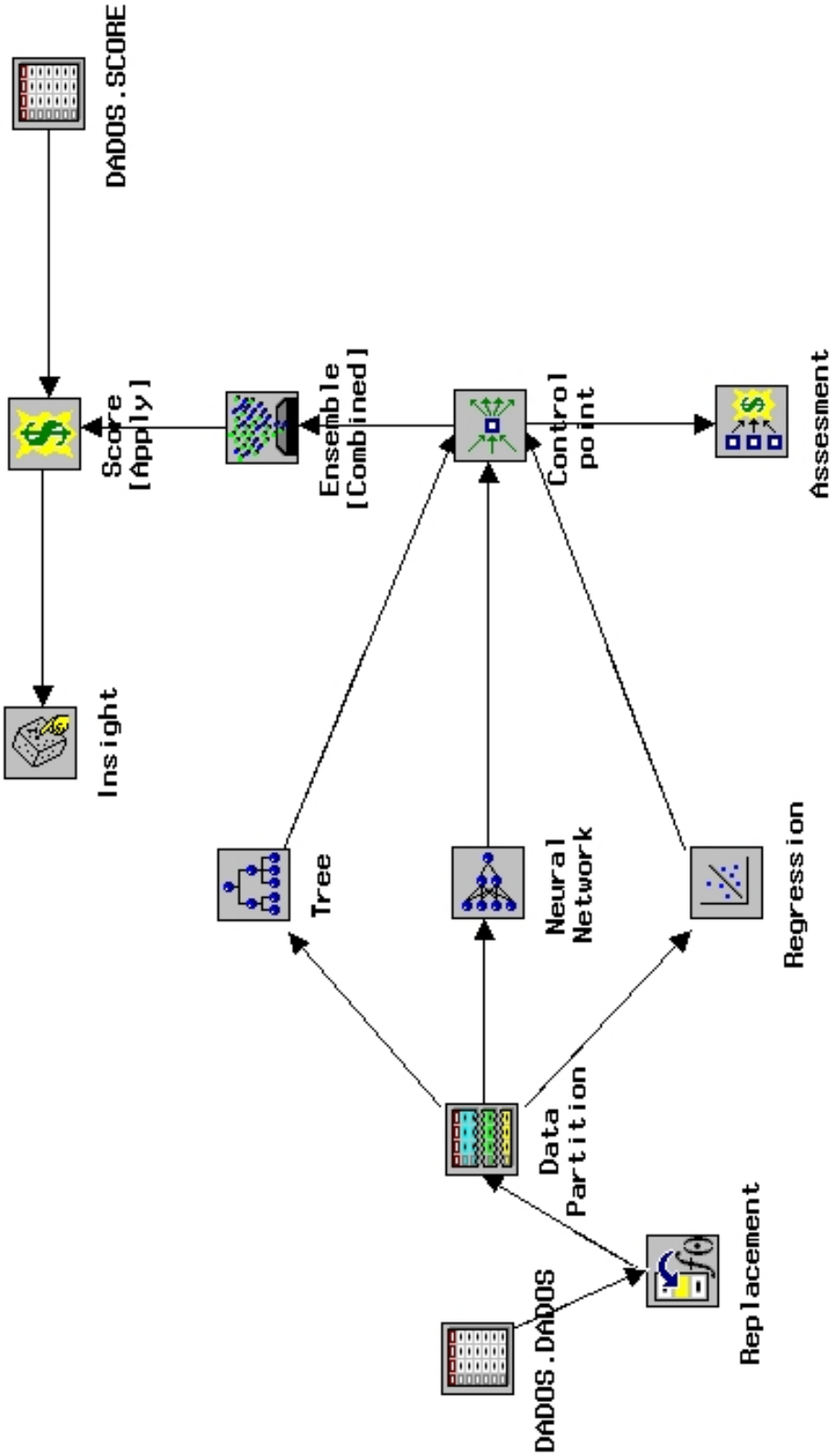
Árvore de Decisão

ANEXO G

Diagrama SAS

Árvore de Decisão – Cinco primeiros níveis




DADOS.SCORE
Score [Apply]
Ensemble [Combined]
Control point
Assesment
Insight
Tree
Neural Network
Regression
Data Partition
Replacement
DADOS.DADOS

ANEXO H

Output GRETL

Convergência atingida depois de 5 iterações

Modelo 2: Probit, usando as observações 1-1585 (n = 1582)
Observações omissas ou incompletas foram ignoradas: 3
Variável dependente: mud_seg

	coeficiente	erro padrão	rácio-t	valor p	
const	0,715101	0,149561	4,781	1,74e-06	***
Genero	0,240853	0,0765133	3,148	0,0016	***
Idade	-0,00787150	0,00224097	-3,513	0,0004	***
participacao	0,226129	0,0757498	2,985	0,0028	***
Mais_Contacto	-0,373249	0,0755454	-4,941	7,78e-07	***
habitacao	-0,242505	0,0770316	-3,148	0,0016	***
Cliente_Exclusi	-0,307073	0,101108	-3,037	0,0024	***

Média var. dependente	0,492415
D.P. var. dependente	0,398876
R-quadrado de McFadden	0,042321
R-quadrado ajustado	0,035937
Log. da verosimilhança	-1049,977
Critério de Akaike	2113,953
Critério de Schwarz	2151,518
Critério de Hannan-Quinn	2127,910

Número de casos 'correctamente preditos' = 960 (60,7%)
f(beta'x) na média das variáveis independentes = 0,399
Teste de razões de verosimilhanças: Qui-quadrado(6) = 92,8003 [0,0000]

		Predito 0	Predito 1
Actual	0	508	295
	1	327	452

Convergência atingida depois de 5 iterações

Modelo 2:
Probit, usando as observações 1-1585 (n = 1582)
Observações omissas ou incompletas foram ignoradas: 3
Variável dependente: mud_seg

	coeficiente	erro padrão	rácio-t	declive
const	0,715101	0,149561	4,781	
Genero	0,240853	0,0765133	3,148	0,0955489
participacao	0,226129	0,0757498	2,985	0,0899453
Mais_Contacto	-0,373249	0,0755454	-4,941	-0,147708
habitacao	-0,242505	0,0770316	-3,148	-0,0962117
Idade	-0,00787150	0,00224097	-3,513	-0,00313976
Cliente_Exclusi	-0,307073	0,101108	-3,037	-0,121455

Média var. dependente	0,492415
D.P. var. dependente	0,398876
R-quadrado de McFadden	0,042321
R-quadrado ajustado	0,035937
Log. da verosimilhança	-1049,977
Critério de Akaike	2113,953
Critério de Schwarz	2151,518
Critério de Hannan-Quinn	2127,910

Número de casos 'correctamente preditos' = 960 (60,7%)
f(beta'x) na média das variáveis independentes = 0,399
Teste de razões de verosimilhanças: Qui-quadrado(6) = 92,8003 [0,0000]

		Predito 0	Predito 1
Actual	0	508	295
	1	327	452

ANEXO I

Simulação de Monte Carlo

Original	Seed	Modelo	Teste
	12345	Probit	0,4575
		Árvore	0,4386
		Rede	0,3667
Repetição 1	**Seed**	**Modelo**	**Teste**
	3059	Probit	0,4329
		Árvore	0,4405
		Rede	0,3611
Repetição 2	**Seed**	**Modelo**	**Teste**
	3255	Probit	0,4499
		Árvore	0,4140
		Rede	0,3781
Repetição 3	**Seed**	**Modelo**	**Teste**
	7484	Probit	0,4367
		Árvore	0,4026
		Rede	0,3535
Repetição 4	**Seed**	**Modelo**	**Teste**
	7934	Probit	0,4019
		Árvore	0,3925
		Rede	0,3547
Repetição 5	**Seed**	**Modelo**	**Teste**
	4536	Probit	0,4348
		Árvore	0,4026
		Rede	0,3705
Repetição 6	**Seed**	**Modelo**	**Teste**
	2994	Probit	0,4057
		Árvore	0,3755
		Rede	0,3585
Repetição 7	**Seed**	**Modelo**	**Teste**
	8485	Probit	0,4019
		Árvore	0,3925
		Rede	0,3705
Repetição 8	**Seed**	**Modelo**	**Teste**
	1315	Probit	0,4083
		Árvore	0,3906
		Rede	0,3811

Repetição 9	**Seed**	**Modelo**	**Teste**
	5633	Probit	0,4113
		Árvore	0,3981
		Rede	0,4075

Repetição 10	**Seed**	**Modelo**	**Teste**
	5496	Probit	0,4453
		Árvore	0,4113
		Rede	0,3887

Repetição 11	**Seed**	**Modelo**	**Teste**
	2036	Probit	0,4461
		Árvore	0,4121
		Rede	0,3705

Repetição 12	**Seed**	**Modelo**	**Teste**
	656	Probit	0,4113
		Árvore	0,4151
		Rede	0,3566

Repetição 13	**Seed**	**Modelo**	**Teste**
	5493	Probit	0,4159
		Árvore	0,4197
		Rede	0,3800

Repetição 14	**Seed**	**Modelo**	**Teste**
	4970	Probit	0,3894
		Árvore	0,3648
		Rede	0,3611

Repetição 15	**Seed**	**Modelo**	**Teste**
	1401	Probit	0,4094
		Árvore	0,3943
		Rede	0,3642

Repetição 16	**Seed**	**Modelo**	**Teste**
	2203	Probit	0,4340
		Árvore	0,4226
		Rede	0,3566

Repetição 17	**Seed**	**Modelo**	**Teste**
	8848	Probit	0,4272
		Árvore	0,3837
		Rede	0,3894

Repetição 18	**Seed**	**Modelo**	**Teste**
	6136	Probit	0,4216
		Árvore	0,4121
		Rede	0,3705

Repetição 19	**Seed**	**Modelo**	**Teste**
	9645	Probit	0,3906
		Árvore	0,4038
		Rede	0,3717

Repetição 20	**Seed**	**Modelo**	**Teste**
	9866	Probit	0,4321
		Árvore	0,3811
		Rede	0,3731

Repetição 21	**Seed**	**Modelo**	**Teste**
	7203	Probit	0,4358
		Árvore	0,4189
		Rede	0,3906

Repetição 22	**Seed**	**Modelo**	**Teste**
	2505	Probit	0,3875
		Árvore	0,4026
		Rede	0,3913

Repetição 23	**Seed**	**Modelo**	**Teste**
	8010	Probit	0,4026
		Árvore	0,3856
		Rede	0,3819

Repetição 24	**Seed**	**Modelo**	**Teste**
	7990	Probit	0,4189
		Árvore	0,3774
		Rede	0,3585

Repetição 25	**Seed**	**Modelo**	**Teste**
	2146	Probit	0,3868
		Árvore	0,3943
		Rede	0,3811

Repetição 26	**Seed**	**Modelo**	**Teste**
	41	Probit	0,4405
		Árvore	0,4064
		Rede	0,3781

Repetição 27	**Seed**	**Modelo**	**Teste**
	3669	Probit	0,4415
		Árvore	0,4038
		Rede	0,3623

Repetição 28	**Seed**	**Modelo**	**Teste**
	3234	Probit	0,4216
		Árvore	0,4140
		Rede	0,3932

Repetição 29	**Seed**	**Modelo**	**Teste**
	2817	Probit	0,4057
		Árvore	0,3925
		Rede	0,3755

Repetição 30	**Seed**	**Modelo**	**Teste**
	9288	Probit	0,4057
		Árvore	0,3925
		Rede	0,3660

Repetição 31	**Seed**	**Modelo**	**Teste**
	6862	Probit	0,4226
		Árvore	0,3925
		Rede	0,3679

Modelo	**Média**	**D. Padrão Amostral**
Probit	0,4186	0,0185
Árvore	0,4003	0,0160
Rede	0,3730	0,0133