4
Metodologia

Este capítulo descreve a metodologia adotada na execução do trabalho de pesquisa: definição da variável alvo, delimitação da população, processo de seleção da amostra, técnicas e procedimentos empregados para o tratamento e análise dos dados e as limitações do método.

4.1.
Tipo de pesquisa

Esta pesquisa foi desenvolvida com a análise de dados reais de uma das quatro grandes operadoras do Brasil de Telefonia Móvel. O nome da operadora não é relevante para o estudo, portanto, será mantido em sigilo. A base de dados possui informações referentes à quatro meses, de outubro de 2013 a janeiro de 2014.

Para os testes e simulações realizados, foram utilizadas as ferramentas SAS Enterprise Miner, SAS Enterprise Guide e SPSS.

150 mil clientes foram estimulados à contratação de um plano de dados para uso no celular através de *Smart Message*. As características dos clientes que aceitaram ou não a oferta serão analisadas com a utilização de técnicas de mineração de dados, particularmente, árvores de decisão, de forma que seja possível identificar as características comuns de clientes que aceitaram ou não a oferta e usá-las para classificar outros clientes como mais ou menos propensos em futuras campanhas.

4.2.
Definição da variável alvo

É importante definir o conceito de propenso ou não propenso, pois ele será utilizado como alvo da modelagem. Serão classificados como PROPENSOS os clientes que foram estimulados via *Smart Message* para a contratação do serviço de internet para uso no celular e aceitaram a oferta, ou seja, ativaram o serviço em

até 10 dias após o envio da mensagem. Serão classificados como NÃO PROPENSOS os clientes que foram estimulados via *Smart Message* para a contratração do serviço de internet para uso no celular e não aceitaram oferta, ou seja, não ativaram o serviço em até 10 dias após o envio da mensagem.

4.3. Preparação dos dados

A seguir serão apresentadas algumas técnicas e práticas utilizadas na preparação de uma base de dados para a extração do conhecimento por meio da utilização de modelos de mineração de dados. Tais procedimentos visam garantir a consistência dos dados e uma representação dos mesmos que facilite a compreensão pelos modelos que serão aplicados. Outros procedimentos podem ser encontrados na literatura (BALLOU, TAYI, 1999; BERRY, LINOFF, 2000; PYLE, 1999) e podem ser aplicados de acordo com os dados disponíveis e em função da questão de negócios a ser solucionada. Os métodos abordados nesta seção serão utilizados na modelagem proposta para melhoria da efetividade de campanhas de *cross-selling*.

4.3.1. Tratamento e limpeza

A limpeza dos dados visa detectar e remover anomalias presentes na base de dados com o objetivo de melhorar a sua qualidade. A limpeza dos dados envolve a verificação da consistência das informações, a correção de possíveis erros e o preenchimento ou eliminação de valores nulos e redundantes.

A verificação das informações envolve a procura por valores que não deveriam existir na base por serem impossíveis na prática. Clientes com mais de 120 anos de vida, menos de 2 anos de idade ou relacionamento de 400 anos com a empresa por exemplo. Possivelmente, esses valores são decorrentes de erros de digitação ou de preenchimento de cadastros.

Depois de identificadas as inconsistências, elas devem ser corrigidas com valores possíveis (utilizando-se médias ou medianas da variável, por exemplo). Este processo busca evitar que os valores incorretos atrapalhem a compreensão dos dados pelos modelos, levando-os a conclusões errôneas.

A remoção de dados (inconsistentes, nulos ou duplicados) é feita para eliminar consultas desnecessárias que seriam executadas pelos modelos e que afetariam o seu desempenho.

A limpeza dos dados envolve também o tratamento de valores ausentes (*missing*). Se o número de observações ausentes for significativo, o desempenho dos modelos de análise de dados pode ser comprometido. Para lidar com valores ausentes, é possível ignorá-los ou preenchê-los. As duas alternativas apresentam vantagens e desvantagens.

Ignorar a descrição do indivíduo ou mesmo eliminar o descritor é indicado quando os dados são abundantes, mas pode ser impraticável se os dados são escassos ou contraindicado se o padrão possui informações importantes além das variáveis com valor ausente.

O preenchimento de um valor ausente pode ser feito manualmente, com uma constante global (não recomendado, pois o sistema pode identificar esse valor como alguma característica importante da variável se for muito freqüente), usando média, moda ou valor mais provável segundo algum modelo (regressão, regra de Bayes, árvores de decisão). Este procedimento salva o padrão da eliminação e aproveita todo o resto da sua informação, mas pode causar desempenho tendencioso na modelagem, principalmente se os valores ausentes forem muitos. Isso poderia levar o modelo a considerar certas estruturas de comportamento nos dados que não deveriam existir.

4.3.2. Oversampling

Muitos problemas que envolvem grandes bases de dados tratam de variáveis categóricas desequilibradas em termos da proporção de cada classe existente (BERRY, LINOFF, 2000; YAN *et al*., 2001). Por exemplo, uma base de telefonia celular pode possuir uma variável que denota se um cliente deixou ou não a empresa. Essa variável em geral possui algo em torno de 98% dos clientes como os que continuam na empresa e somente 2% dos clientes como os que terminaram sua relação com a operadora. O que acontece no momento da construção de qualquer modelo envolvendo uma variável deste tipo é que, dada essa distribuição desequilibrada entre as classes, o modelo enxerga apenas uma das classes, sendo incapaz de distinguir a classe de menor número de registros. Isso acontece porque

o modelo reconhece que, se sua resposta for sempre dizer que todas as observações pertencem à classe com maior número de registros, ele acertará 98% dos padrões.

Para evitar esse problema e facilitar a distinção de classes, é realizado um procedimento conhecido como *oversampling*. Com o uso do *oversampling* cria-se uma nova base de dados para a modelagem, selecionando-se geralmente todos os registros pertencentes à classe rara e realizando-se uma amostra aleatória das ocorrências da classe comum, ajustando a proporção entre as classes. No caso de variáveis binárias, em geral, se deseja que a classe rara corresponda a algo entre 10% e 40% da base para a modelagem. Freqüências entre 20% e 30% em geral dão bons resultados (BERRY, LINOFF, 2000). O procedimento de *oversampling* descrito é bastante similar para variáveis com mais de duas classes.

Entretanto, o *oversampling* possui limitações. Dado que só existe um pequeno número de observações da classe rara na base de dados, não é possível criar uma base de qualquer tamanho para a análise, mesmo que a base de dados original seja muito grande. Por exemplo, em uma base de 100.000 registros, se somente 2% pertencem a uma classe, isso significa que só estão disponíveis 2000 amostras desta classe rara. Sendo assim, é impossível construir uma base para modelagem com, por exemplo, 50.000 registros e uma freqüência maior do que 4% para a classe rara. É necessário, para atingir as proporções entre 10% e 40% mencionadas, que a base de dados seja bastante diminuída no que diz respeito às observações da classe comum.

4.3.3. Adição de variáveis derivadas

Na preparação dos dados, pode ser interessante a criação de novas variáveis a partir das variáveis existentes na base de dados. A criação de novas variáveis tem como objetivo enfatizar certos aspectos dos clientes que podem ajudar na modelagem. Existem muitas formas de se realizar esse incremento na informação presente na base de dados (MOZER *et al*., 2002; YAN, WOLNIEWICZ, DODIER, 2004).

Alguns exemplos da criação de variáveis derivadas são: o cálculo de razões entre variáveis (gasto médio por minuto de ligação, por exemplo), a criação de variáveis que demonstrem mais claramente variações de outras entradas ao longo do tempo (crescimento da fatura ao longo de alguns meses, por exemplo) e a criação de variáveis que agreguem diversas outras com pouca informação uma a uma (soma de todas as ligações para o Call Center, não importando sua natureza, por exemplo).

4.4. Universo e amostra

Inicialmente, foram selecionados 150 mil clientes para a ação, com as seguintes características:

- Clientes Pessoa Física
- Usuários de planos de voz pós-pagos
- *Opt in* (aceitam receber informações e ofertas da operadora)
- Clientes que não estejam inadimplentes há mais de 7 dias
- Usuários de *Webphone* e *Smartphone*
- Não usuários de internet no celular
- Não estão mais na Régua de Relacionamento de Boas Vindas (aproximadamente 3 meses após a data de ativação do contrato, onde o cliente recebe alguns informativos)

Dos quase 150 mil clientes abordados, apenas 1,05% aceitaram a oferta (um total de 1500 clientes). Como a distribuição das classes (quem aceitou e quem não aceitou) ficou desequilibrada, foi necessário o tratamento com o procedimento de *oversampling*, visto anteriormente. Assim, dois terços da amostra final (3000 clientes) foram escolhidos aleatoriamente da base de dados completa para compor uma amostra onde o grupo de clientes que aceitou a oferta representasse um terço (33%) do total. Após o tratamento e a limpeza dos dados, restaram 4500 clientes para serem trabalhados na modelagem.

4.5.
Coleta e estudo dos dados

Para coletar e agregar os dados que serão utilizados na modelagem, é necessário entender a natureza desses dados e definir quais deles serão necessários para o modelo (BERRY, LINOFF, 2000). Em geral, os tipos de dados presentes nas empresas de telefonia celular, e que podem ser interessantes para modelagens de *cross-selling*, podem ser agrupados da seguinte forma (MOZER *et al*., 2000):

- Perfil de Uso do Cliente: detalhes sobre as chamadas feitas pelo usuário (data, tempo de duração, localidade, números de telefones diferentes discados, tipo de ligação), detalhes sobre chamadas perdidas (devido à falta de cobertura, por exemplo) e dados sobre a qualidade do serviço prestado (interferência, má cobertura);
- Faturamento: toda a informação financeira que aparece na conta do usuário (faturas mensais, valor da assinatura, cobranças por uso *roaming*, cobranças por minutos adicionais, entre outros);
- Atendimento ao Cliente: dados sobre os contatos feitos entre cliente e o serviço de atendimento ao cliente e suas resoluções;
- Relacionamento: detalhes sobre o relacionamento com o cliente (antiguidade da conta, aparelho possuído, tecnologia utilizada, situação de crédito, e outros);
- Demográfico: dados sobre a posição geográfica dos clientes e suas características como população (sexo, idade, estado civil, etc);

Apesar de sempre se desejar ter acesso à maior quantidade de informação possível, muitas vezes alguns dos dados desejados não são confiáveis ou não podem ser obtidos. Deve-se trabalhar com as possibilidades permitidas pela realidade, mesmo sabendo que o modelo poderia ter um desempenho superior caso certos dados existissem.

Uma particularidade da indústria de telefonia móvel é a necessidade de segmentação dos dados para a elaboração dos modelos. Isso ocorre porque existem vários tipos de clientes e cada um deles possui perfis muito diferentes, o que tornaria complicada qualquer tentativa de modelagem em conjunto (BERRY, LINOFF, 2000). Os tipos de clientes mais comuns são os pré-pagos, os pós-pagos e os clientes empresa (Pessoa Jurídica). Podem existir ainda outros tipos de

clientes, dependendo especificamente de cada operadora e de suas definições de negócio. Na exploração, validação e limpeza dos dados esse fato deve ser observado e os clientes pertencentes a diferentes tipos devem ser separados para a construção de modelagens específicas para cada categoria de consumidor.

Para este estudo especificamente, foram selecionadas vinte variáveis consideradas como relevantes para a análise com base no problema estudado. A Tabela 4.1 apresenta a estrutura inicial da base de dados.

COD_CLIENTE	Código de identificação de cada cliente
DDD	DDD do acesso
UF	UF do acesso
CIDADE	Cidade do acesso
MARCA_APARELHO	Marca do aparelho
MODELO_APARELHO	Modelo do aparelho
IN_TOTAL	Quantidade total de minutos utilizados
MIN_OUTGOING	Minutos utilizados para originar chamadas
MIN_LOCAL	Minutos utilizados para originar chamadas locais
MIN_ONNET	Minutos utilizados para originar chamadas locais para a mesma operadora
MIN_OFFNET_MOVEL	Minutos utilizados para originar chamadas locais para outras operadoras
MIN_FIXO	Minutos utilizados para originar chamadas para telefones fixos
MIN_LD	Minutos utilizados para originar chamadas de longa distância
MIN_INCOMING	Minutos utilizados em ligações recebidas
AGING	Tempo de base do cliente
SEGMENTO	Segmento do cliente
INFORMACAO	Quantidade de vezes que o cliente entrou em contato com o Call Center para solicitar uma informação
SOLICITACAO	Quantidade de vezes que o cliente entrou em contato com o Call Center para fazer alguma solicitação
RECLAMACAO	Quantidade de vezes que o cliente entrou em contato com o Call Center para fazer alguma reclamação
INTERACOES	Quantidade de vezes que o cliente teve alguma interação com o Call Center, seja para fazer uma solicitação, pedir uma informação ou fazer uma reclamação
PLANO_DE_VOZ	Plano de voz do cliente
VARIAVEL_ALVO	Indicação se o cliente aceitou ou não o pacote de dados oferecido

Tabela 4.1: Variáveis de Análise
Fonte: Própria