6
Estudo de Caso

6.1.
Descrição

Para avaliar o desempenho do sistema de auxílio à retenção de clientes desenvolvido utilizou-se uma base de dados real de uma das quatro grandes operadoras do Brasil. O nome da operadora não é relevante e será mantido sob sigilo. A base de dados corresponde a dados de nove meses, de outubro de 2000 a julho de 2001.

Para os testes e simulações realizados, foram utilizadas as ferramentas SAS System, SAS Enterprise Miner, SAS Enterprise Guide, MATLAB 6.1 e ferramentas criadas pelo laboratório ICA (Inteligência Computacional Aplicada) da PUC-Rio: Rule Evolver [LOPE99-1] [LOPE99-2] e o sistema Neuro-Fuzzy Hierárquico BSP [GONÇ01] [SOUZ99].

6.2.
Dados disponíveis e suas limitações

Devido a sua própria natureza, o problema do *churn* é de importância estratégica para operadoras de telefonia celular, lidando diretamente com perda de clientes para um concorrente. Os dados referentes ao *churn* são altamente confidenciais, dado o óbvio risco de tais dados caírem em mãos de um concorrente, sendo, portanto, muito complicados de serem adquiridos por terceiros que não fazem parte da empresa.

A dificuldade na obtenção dos dados foi em todo momento a grande limitação deste trabalho, não permitindo um maior número de testes com o sistema desenvolvido, como era desejado. A única base de dados disponível para análise foi uma base de 100.000 clientes do final de 2000 até meados de 2001, quando a situação do mercado de telecomunicações móvel no Brasil era bem diferente da situação atual, após a entrada de mais concorrentes em 2002/2003, como detalha o Capítulo 4. Em 2000/2001 só havia duas operadoras operando no mercado local referente à base de dados. Devido à atual competição e de mudanças na regulamentação de um mercado que vai

caminhando para a maturidade, as taxas de churn mensal médio para as quatro principais operadoras do Brasil nos dias de hoje, conforme apresentado na Tabela 4.2, são bem superiores ao churn mensal de 1,245% encontrado na base analisada.

Por causa da impossibilidade de se obter mais dados, o fato de a taxa de *churn* da base ser de 1,245% significou que o número máximo da amostra de clientes *churners* disponível era de exatamente 1245 clientes (dado que a base possuía 100.000 registros). Os outros 98.775 correspondiam a clientes ativos. Como será visto a seguir, esse pequeno número de exemplos da classe *churner* limitou as possibilidades de testes mais profundos com a taxa de *oversampling* [ARCH04] [BERR00].

Outra grande limitação dos dados foi o pequeno número de variáveis disponíveis para a caracterização do cliente. Bases de *churn* apresentadas em outras pesquisas podem possuir 61, 143 ou até 251 varáveis de entrada relacionadas ao *churn* [ARCH04] [AU03] [MOZE00]. A base de dados disponível possuía 40 variáveis, dentre as quais algumas não eram relacionadas com o *churn*, como a chave numérica por número de cliente, ou não eram confiáveis (segundo especialistas da própria operadora), como a divisão em categorias dos clientes. Essa restrição em termos de informação disponível certamente teve seu impacto no desempenho do sistema.

A Tabela 6.1 apresenta a estrutura inicial da base de dados. É possível verificar a existência de dados de uso da rede, como QTD_CHAM_ENT e QTD_MIN_SAI; dados de relacionamento, como ANTIG_CONTA e ANTIG_APARELHO; dados de faturamento, como FATURA_MEDIA e FATURA_BASE; dados de atendimento ao cliente, como INF_FINAN e REC_ATEND; e dados demográficos, como SEXO e IDADE. Nenhum dado de mercado estava presente na base.

Sobre essa base de dados, com suas possibilidades e limitações, foi aplicado o sistema de retenção de clientes por mineração de dados proposto, visando a solução da questão do *churn*.

Tabela 6.1 – Estrutura da base de dados real para o estudo de caso.

Nome da Variável	Descrição
COD_CLIENTE	Chave numérica, única para cada cliente
ANTIG_CONTA	Idade do contrato, em meses
ANTIG_APARELHO	Idade do aparelho, em meses
FATURA_MEDIA	Fatura mensal média
CRESCI_FATURA	Crescimento da fatura
OSCI_FATURA	Oscilação da fatura
CRESC_CHAMADAS	Crescimento das chamadas
OSC_CHAMADAS	Oscilação das chamadas
SEXO	Sexo
IDADE	Idade
PLANO	Plano de assinatura
CATEGORIA	Categoria de cliente
TECNOLOGIA	Tecnologia, analógica ou digital
QTD_CHAM_ENT	Número de chamadas entrantes
QTD_TELS_ENT_DIF	Número de telefones diferentes entrantes
QTD_MIN_ENT	Minutos em ligações entrantes
QTD_MIN_SAI	Minutos em ligações iniciadas pelo cliente
QTD_CHAM_CXP	Número de chamadas para a Caixa Postal
PERC_MIN_SAI_VC1	% de min. em chamadas locais
PERC_MIN_SAI_VC2	% de min. em chamadas de longa distância
PERC_MIN_SAI_VC3	% de min. em chamadas internacionais
PERC_MIN_SAI_ROAM	% de min. em chamadas em roaming
PERC_MIN_SAI_OUTROS	% de min. em chamadas em outras situações
PERC_MIN_SAI_HN	% de min. em chamadas sob tarifa normal
PERC_MIN_SAI_HR	% de min. em chamadas sob tarifa reduzida
INF_ATEND	Número de chamadas para o Call Center por informações genéricas
INF_FINAN	Número de chamadas para o Call Center por informações financeiras
INF_TEC	Número de chamadas para o Call Center por informações tecnolócigas
REC_ATEND	Número de reclamações por razões genéricas
REC_FINAN	Número de reclamações por razões financeiras
REC_TEC	Número de reclamações por razões tecnológicas
FATURA_1AHEAD	Fatura mensal, 1 mês à frente do mês base
FATURA_BASE	Fatura mensal, mês base (maio 2001)
FATURA_1AGO	Fatura mensal, 1 mês antes do mês base
FATURA_2AGO	Fatura mensal, 2 mês antes do mês base
FATURA_3AGO	Fatura mensal, 3 mês antes do mês base
FATURA_4AGO	Fatura mensal, 4 mês antes do mês base
FATURA_5AGO	Fatura mensal, 5 mês antes do mês base
FATURA_6AGO	Fatura mensal, 6 mês antes do mês base
FATURA_7AGO	Fatura mensal, 7 mês antes do mês base

6.3. Apresentação e discussão dos resultados

Seguindo a estrutura do sistema desenvolvido, apresentada no capítulo anterior, os passos iniciais são o estudo da estrutura de dados da empresa e a definição, com base no entendimento do negócio e do problema, das informações existentes que possam contribuir para uma melhor caracterização

do *churn*. Infelizmente, devido aos problemas de confidencialidade na aquisição de dados desta natureza, como destacado na seção anterior, essa etapa inicial não pode fazer parte do estudo de caso aqui apresentado. A obtenção dos dados foi feita através de contatos em uma operadora e tal contato forneceu toda a informação permitida e imediatamente disponível, conforme já detalhado. É importante ressaltar a importância desta primeira etapa para o melhor desempenho do sistema e, apesar dos dados disponíveis serem suficientes para o teste e validação do sistema, deve ficar claro que os resultados provavelmente poderiam ser melhores caso fosse permitida a realização desta etapa inicial com todo o cuidado que seria requisitado.

6.3.1. Validação, exploração e limpeza dos dados

Com os dados disponíveis, o sistema foi inicializado na etapa de validação, exploração e limpeza dos mesmos.

Nesta etapa um estudo exploratório dos dados foi realizado para se detectar impurezas, valores absurdos, ruído e buscar uma maior compreensão dos dados que iriam ser tratados. A Tabela 6.2 apresenta uma visão das principais estatísticas da base de dados inicial.

Ao avaliar-se com cuidado a Tabela 6.2, nota-se que algum esforço de limpeza de dados é essencial. Olhando-se para a variável IDADE, por exemplo, nota-se que seu valor máximo é de 357 anos e o seu valor mínimo é de zero, o que claramente denota valores absurdos. Além disso, essa mesma variável possui 30.071 valores ausentes. Outras variáveis, como, por exemplo, a FATURA_MEDIA e a QTD_MIN_SAI, apresentam valores máximos muito maiores do que suas médias, e desvios padrões bem altos. Investigando o histograma da distribuição de uma destas variáveis com cuidado nota-se que tais valores máximos são, na verdade, *outliers* bem isolados dos demais valores das variáveis. A Figura 6.1 ilustra a distribuição da variável FATURA_MEDIA original, antes de qualquer tratamento. Outras variáveis apresentaram problemas similares, com *outliers* e alguns com valores ausentes. Foram realizados filtros de *outliers* customizados para cada variável que apresentou esses valores, de modo a evitar os efeitos negativos que esses pudessem vir a ter no sistema. Tais filtros, feitos para variáveis não percentuais e não categóricas, foram colocados em torno do ponto que incluía 95% dos valores mais freqüentes de cada variável.

Tabela 6.2 – Estatísticas descritivas básicas das variáveis base de dados original.

Variável	Max	Média	Min	Missing	σ
ANTIG_CONTA	426,00	44,15	6,00	0,00	22,27
ANTIG_APARELHO	131,00	23,62	2,00	0,00	15,75
FATURA_MEDIA	4864,93	72,91	-7,19	1,00	65,49
CRESC_FATURA	43,91	0,02	-13,39	10,00	0,23
OSC_FATURA	58,07	0,25	-123,71	9,00	0,48
CRES_CHAMADAS	1,20	0,04	-1,20	1689,00	0,37
OSC_CHAMADAS	2,00	0,58	0,00	1688,00	0,40
IDADE	357,00	43,30	0,00	30071,00	13,31
QTD_TEL_DIF_ENT	633,00	27,60	0,00	4336,00	27,53
QTD_CHAM_ENT	2145,00	67,42	1,00	4336,00	78,97
QTD_MIN_ENT	2333,00	93,07	0,00	4336,00	115,15
QTD_MIN_SAI	7643,00	215,62	0,00	0,00	316,46
QTD_CHAM_CXP	181,00	1,27	0,00	0,00	4,16
PERC_MIN_SAI_VC1	1,00	0,79	0,00	0,00	0,26
PERC_MIN_SAI_VC2	1,00	0,05	0,00	0,00	0,10
PERC_MIN_SAI_VC3	1,00	0,02	0,00	0,00	0,06
PERC_MIN_SAI_ROAM	1,00	0,02	0,00	0,00	0,07
PERC_MIN_SAI_OUTROS	1,00	0,11	0,00	0,00	0,16
PERC_MIN_SAI_HN	1,00	0,58	0,00	0,00	0,23
PERC_MIN_SAI_HR	1,00	0,40	0,00	0,00	0,23
INF_ATEND	681,00	0,23	0,00	0,00	2,31
INF_FINAN	87,00	0,35	0,00	0,00	1,01
INF_TEC	297,00	0,19	0,00	0,00	1,15
REC_ATEND	3,00	0,00	0,00	0,00	0,04
REC_FINAN	6,00	0,01	0,00	0,00	0,13
REC_TEC	12,00	0,02	0,00	0,00	0,18
FATURA_1AHEAD	2962,16	73,27	-146,69	0,00	71,89
FATURA_BASE	2796,01	79,86	-23,16	0,00	77,66
FATURA_1AGO	3061,78	76,14	-110,59	2,00	70,26
FATURA_2AGO	12618,40	71,71	-85,93	2,00	85,63
FATURA_3AGO	4365,41	72,90	-110,56	1,00	72,53
FATURA_4AGO	2759,01	70,89	-51,07	1,00	69,66
FATURA_5AGO	2150,63	71,39	-84,54	10,00	68,39
FATURA_6AGO	3510,34	73,41	-36,03	1108,00	70,31
FATURA_7AGO	2279,53	66,82	-47,99	3611,00	64,25

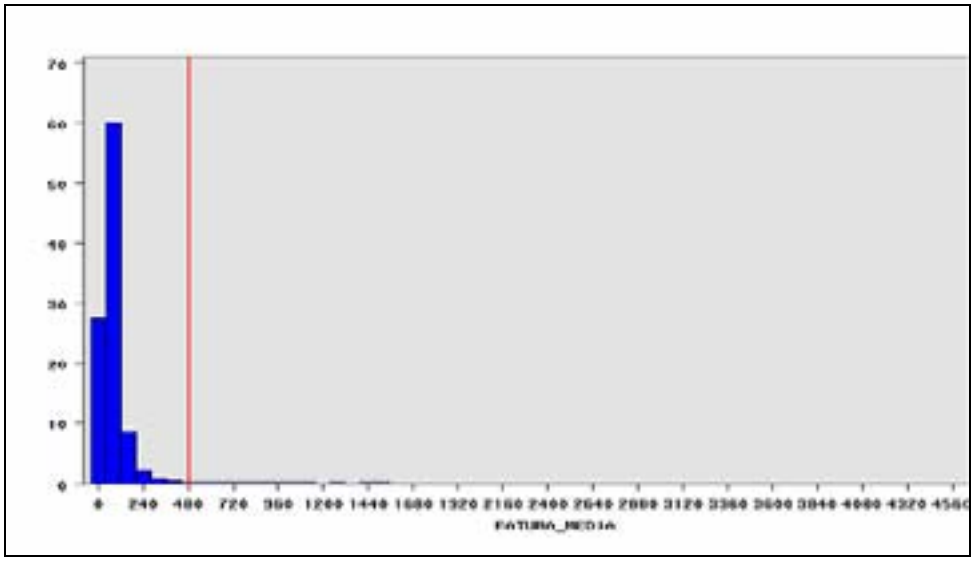

Figura 6.1 – Histograma da Distribuição da variável Fatura Média, a linha que corta o valor de R$480,00 no eixo *x* demonstra onde foi realizado o filtro de *outliers*.

Em relação aos registros com valores ausentes presentes na tabela, foi realizado um procedimento diferente. Sabe-se que o sistema possui uma etapa de *oversampling* a ser realizada em breve e que irá descartar a maior parte da tabela para obter uma base com proporção em torno de 30% de *churners*. Isso significava reduzir a base de 100.000 para em torno de 3.500 observações, visto que só existem 1,245% de *churners* na base original (1.245 registros). Logo, dado que, dos aproximadamente 99.000 clientes ativos, somente alguns (~2.500) fariam parte da base final, foi decidido pela eliminação de todos os registros que possuíssem valores *missing* antes do procedimento de *oversampling*, de modo que a base final não possuísse nenhum valor ausente.

6.3.2. Definição do alvo

Feita a validação, exploração e limpeza dos dados, o próximo passo foi definir a variável alvo segundo as regras de negócio da empresa. Uma janela de tempo de um mês foi selecionada, e o mês base de maio de 2001, escolhido, exatamente porque era o penúltimo mês disponível. Sendo o mês de julho de 2001 o último presente nos dados, este foi utilizado como o mês a ser previsto. Então, segundo a definição do que caracteriza o *churn* para a empresa, foi criada a variável alvo representando todos os clientes que encerraram todos os tipos de serviços com a operadora durante o mês seguinte ao mês base.

6.3.3.
Adição de Variáveis Derivadas e Transformações

Nas etapas de adição de variáveis derivadas e transformação de dados, as principais modificações nas bases de dados foram: a criação de uma variável que agregava todas as variáveis relacionadas à operação do *Call Center* da empresa (CALL_CENTER = INF_ATEND + INF_FINAN + INF_TEC + REC_ATEND + REC_FINAN + REC_TEC); o mapeamento em variáveis *dummy* das variáveis categóricas como SEXO, CATEGORIA e TECNOLOGIA; e a normalização por amplitude de todas as variáveis númericas segundo a Equação 3.2, de forma a evitar que a diferente variabilidade de cada entrada atrapalhasse os modelos.

6.3.4.
Seleção de variáveis

Na etapa de seleção de variáveis, antes de serem executados os métodos propostos (LSE, SIE e ANFIS), algumas variáveis foram eliminadas manualmente: COD_CLIENTE, por ser somente a chave única que define cada cliente; PERC_MIN_SAI_HR, por ser o complemento da variável PERC_MIN_SAI_HN; CATEGORIA, devido a especialistas da operadora afirmarem que os métodos utilizados na sua criação não forneciam uma segmentação de clientes confiável; TECNOLOGIA, por falta de variação, dado que 91% dos clientes na base original já possuíam a tecnologia digital; CRESC_CHAMADAS e OSC_CHAMADAS, por apresentarem alta e significativa correlação com as variáveis CRESC_FATURA e OSC_FATURA, respectivamente; SEXO, porque a sua distribuição tanto para a classe de churner quanto para a classe de ativo era praticamente idêntica; e as seis variáveis referentes ao serviço de *Call Center* por tipo de contato (INF_ATEND, INF_FINAN, INF_TEC, REC_ATEND, REC_FINAN, REC_TEC), por fornecerem pouca informação, sendo as variáveis INF nulas para pelo menos 80% da base original e as REC nulas para pelo menos 98% da base.

Após a retirada das variáveis acima descritas, a tabela original que possuía 40 variáveis, ficou reduzida a 28 (incluindo aqui a variável CALL_CENTER criada). Os métodos de seleção de variáveis estudados foram, então, aplicados a essa tabela reduzida, na busca das entradas que possuíssem maior poder explicativo com relação à variável alvo. Os resultados encontram-se nas Tabelas 6.3, 6.4 e 6.5.

Tabela 6.3 – Resultados da aplicação do método de seleção de variáveis LSE.

Rank	Importância	Descrição
1	0.3790	% de Min. em VC3
2	0.2560	% de Min. em Roaming
3	0.1840	% de Min. em VC2
4	0.0930	% de Min. em VC1
5	0.0300	Oscilação Fatura
6	0.0220	% de Min. no Horário N
7	*0.0030*	Cresc. da Fatura
8	*0.0020*	Antiguidade do Aparelho
9	*0.0010*	Quant. Chamadas CXP
10	*0.0010*	Quant. de Tel Dif. Ent.
11	*0.0010*	Antiguidade da Conta
12	*0.0005*	Fatura Média

Tabela 6.4 – Resultados da aplicação do método de seleção de variáveis SIE.

Rank	Importância	Descrição
1	0.5078	% de Min. em VC3
2	0.2412	Cresc. da Fatura
3	0.0867	% de Min. em Roaming
4	0.0553	% de Min. em VC2
5	0.0521	% de Min. em VC1
6	0.0465	Oscilação Fatura

Tabela 6.5 – Resultados da aplicação do método de seleção de variáveis ANFIS.

Rank	Descrição
1	Antiguidade da Conta
2	Antiguidade do Aparelho
3	Quant. de Tel. Dif. Ent.
4	Crescimento da Fatura
5	Quant. de Min. Ent.
6	Quant. de Chamadas Ent.
7	Call Center
8	Fatura Média
9	% Min. no Horário N
10	Oscilação Fatura

Avaliando-se os resultados apresentados nas Tabelas 6.3, 6.4 e 6.5, nota-se que existe uma semelhança grande entre os resultados dos métodos LSE e SIE, colocando como mais importantes na definição do *churn* as variáveis que denotam o perfil de uso da rede pelo assinante (PERC_MIN_SAI_VC1, PERC_MIN_SAI_VC2, PERC_MIN_SAI_VC3, PERC_MIN_SAI_ROAM) e o comportamento de sua fatura ao longo dos meses disponíveis (CRESC_FATURA, OSC_FATURA). É interessante notar que os resultados dos métodos LSE e SIE destoam completamente dos resultados apresentados pelo

método ANFIS, onde variáveis de relacionamento ocupam as primeiras posições. A causa disso reside no fato de o LSE e o SIE serem métodos independentes de qualquer modelo, enquanto o ANFIS depende do treinamento de um modelo adicional. Conforme discutido no capítulo 3, o uso de métodos de seleção de variáveis dependentes do modelo pode não ser conclusivo, uma vez que o resultado depende da parametrização do modelo, que pode não ter sido adequada. Deste modo, os resultados dos modelos LSE e SIE foram considerados mais confiáveis, mesmo porque eles corroboram um o outro.

Entretanto, além de fornecerem pistas de quais variáveis são mais importantes na caracterização do *churn*, o que pode levar, por exemplo, ao planejamento de incentivos voltados para minutos grátis ou tarifas especiais, os métodos de seleção de variáveis determinam quais as variáveis de entrada que deverão fazer da base final para a modelagem. Deste modo, foram realizados testes com bases de dados que possuíam também as entradas sugeridas pelo método de seleção de variáveis baseado no modelo ANFIS, tais como a QTD_TEL_DIF_ENT, ANTIG_CONTA, FATURA_MÉDIA e ANTIG_APARELHO. O desempenho dos modelos contendo também essas variáveis foi superior ao dos modelos onde só foram utilizadas variáveis definidas pelos métodos LSE e SIE.

Conclui-se que métodos de seleção de váriáveis podem ser de grande valor na identificação das variáveis relevantes, fornecendo também compreensão sobre os principais fatores que caracterizam uma variável alvo. De qualquer forma, nunca se deve esquecer que o conhecimento do negócio é essencial na etapa de seleção de variáveis, sendo ele crítico na definição de variáveis que possam contribuir para o desempenho dos modelos.

6.3.5. Oversampling

Foi realizado um procedimento de *oversampling* que reduziu a base de dados original com 100.000 registros para uma base de somente 3.501 observações, dentre as quais 1.001 (28,67%) pertenciam à classe *churner* e 2.500 pertenciam à classe ativo (71,33%). Dos 1245 *churners* iniciais, 244 foram eliminados no processo de limpeza dos dados; portanto a quantidade final de *churners* disponível foi de 1.001.

A taxa de *oversampling* foi mantida constante devido, principalmente, ao reduzido número de clientes da classe *churner* disponível: aumentá-la significaria reduzir demais o número de registros presentes na base final para modelagem.

Para evitar qualquer erro amostral causado pelo *oversampling* na base final, um procedimento semelhante ao da validação cruzada também foi empregado para a amostra de 2.500 clientes ativos que foi retirada da maior parte da população. Foram construídas 10 bases de dados de *oversampling*, cada uma com diferentes amostras aleatórias dos 2.500 clientes ativos. Obviamente os clientes *churners* foram mantidos constantes, uma vez que só existiam 1001 deles.

6.3.6. Criação das bases de dados

A base de dados final para a modelagem consistiu de 15 variáveis de entrada, além da variável alvo, e 3501 registros. As entradas selecionadas estão dispostas na Tabela 6.6. Elas foram escolhidas combinando-se as variáveis relevantes identificadas através de todos os métodos de seleção estudados.

Essa base foi particionada em conjuntos de treinamento, validação e teste nas proporções de 70%, 20% e 10%, respectivamente. Um procedimento de validação cruzada foi empregado para evitar o erro amostral na criação dos conjuntos e 10 bases, com diferentes amostras de treinamento, validação e teste, foram utilizadas para a modelagem de cada método.

Tabela 6.6 – Variáveis restantes para a modelagem após a seleção de variáveis

Nome da Variável	**Descrição**
ANTIG_CONTA	Idade do contrato, em meses
ANTIG_APARELHO	Idade do aparelho, em meses
FATURA_MEDIA	Fatura mensal média
CRESCI_FATURA	Crescimento da fatura
OSCI_FATURA	Oscilação da fatura
QTD_CHAM_ENT	Número de chamadas entrantes
QTD_TELS_ENT_DIF	Número de telefones diferentes entrantes
QTD_MIN_ENT	Minutos em ligações entrantes
QTD_CHAM_CXP	Número de chamadas para a Caixa Postal
PERC_MIN_SAI_VC1	% de min. em chamadas locais
PERC_MIN_SAI_VC2	% de min. em chamadas de longa distância
PERC_MIN_SAI_VC3	% de min. em chamadas internacionais
PERC_MIN_SAI_ROAM	% de min. em chamadas em roaming
PERC_MIN_SAI_HN	% de min. em chamadas sob tarifa normal
CALL_CENTER	Ligações para o Call Center

6.3.7. Modelagem

O treinamento e teste da capacidade de generalização dos modelos avaliados se deu através de um longo processo de otimização de parâmetros de cada um deles, na busca pelo que melhor descrevesse os dados e classificasse a variável alvo de forma mais correta.

O modelo do Classificador Bayesiano serviu como base de comparação para os demais, sendo as probabilidades *a priori* calculadas de acordo com as proporções das classes na base de dados.

Os modelos de redes neurais MLP (*Multi-Layer Perceptron*) e PNN (*Probabilistic Neural Networks*) apresentaram desempenho muito similar quando o número de neurônios na camada escondida era variado entre 10 e 30 e a taxa de decaimento dos pesos (*weight decay*) estava em 0.3. O algoritmo RProp [BISH96] foi utilizado no treinamento. As redes finais apresentaram 25 neurônios na camada escondida.

O modelo ideal de Árvore de Decisão encontrado utilizou o algoritmo C4.5, com o mínimo número de observações necessário para o *split* sendo 10, o número mínimo de observações em uma folha sendo 1 e a profundidade máxima da árvore sendo 8 níveis.

A melhor configuração de parâmetros para o Rule-Evolver foi: função de avaliação *Rule-Interest*; recompensa para acurácia e abrangência; operador de *don't care*; crossovers de um ponto, dois pontos e média inicializados com probabilidade de 20%, 20% e 30% e tendo no fim probabilidades de 10%; mutações simples, *don't care* e *creeps* inicializadas com probabilidade de 10% e terminando com probabilidades de 20%, 20% e 30% e a utilização de elitismo e normalização linear.

O melhor modelo NFHB, encontrado após inúmeros testes com sua taxa de decomposição, limitadora do crescimento da sua estrutura, possuía essa taxa no valor de 0.08.

O modelo de máquinas de vetor de suporte (SVM) final, após vários testes, possuía *kernel* gaussiano com parâmetro 1.5 e parâmetro de regularização *C* igual a 10.

Os resultados obtidos para cada um destes modelos, em comparação com os demais, estão presentes na Tabela 6.8 e serão discutidos agora.

Ao se analisarem os resultados da Tabela 6.8, primeiramente observa-se que a taxa de classificação correta encontrada nos melhores modelos está bem

próxima dos valores apresentados por [ARCH04] [AU03] [MOZE00] [YAN04], em torno dos 60-75% (Tabela 6.7). Tal semelhança demonstra a dificuldade de obtenção de uma alta taxa de classificação correta em problemas de churn.

Dentre os métodos avaliados nesta dissertação, o modelo com a melhor classificação de cliente ativo como cliente ativo corretamente é o modelo de SVM mas, ao mesmo tempo, esse modelo apresentou o maior erro ao classificar um *churner* como *churner*. O modelo com melhor desempenho em ambas as classes simultaneamente foi o modelo de redes neurais MLP, muito próximo ao das redes neurais PNN. O modelo baseado no Rule Evolver foi o modelo com pior classificação de ativo como ativo.

Tabela 6.7 – Resultados obtidos na literatura para modelagem do *churn* através de métodos de mineração de dados.

Referência	**Tipo de Modelo Utilizado**	**taxa de classificação correta**
ARCH04	Redes Neurais	62%
AU03	Algoritmos Genéticos	75%
MOZE00	Redes Neurais	68%
	Árvores de Decisão	60%
YAN04	Máquinas de Vetor de Suporte	68%
	Redes Neurais	69%

Tabela 6.8 – Matrizes de confusão para os melhores modelos de cada família de métodos avaliados.

		Treinamento		**Validação**		**Teste**	
		predito		**predito**		**predito**	
Modelo	**real**	*ativo*	*churn*	*ativo*	*churn*	*ativo*	*churn*
Classificador Bayesiano	*ativo*	61%	39%	60%	40%	**57%**	43%
	churn	33%	66%	38%	62%	41%	**59%**
Rede Neural MLP	*ativo*	75%	25%	73%	27%	**70%**	30%
	churn	30%	70%	31%	69%	35%	**65%**
Rede Neural PNN	*ativo*	73%	27%	70%	30%	**68%**	32%
	churn	35%	65%	37%	63%	37%	**63%**
Árvore de Decisão	*ativo*	70%	30%	65%	35%	**63%**	37%
	churn	33%	67%	36%	64%	40%	**60%**
Rule-Evolver	*ativo*	59%	41%	50%	50%	**44%**	56%
	churn	28%	72%	38%	62%	40%	**60%**
NFHB	*ativo*	64%	36%	60%	40%	**59%**	41%
	churn	30%	70%	32%	68%	35%	**65%**
SVM	*ativo*	88%	12%	85%	15%	**84%**	16%
	churn	70%	30%	72%	28%	72%	**28%**

Apesar de não terem alcançado o melhor desempenho de classificação entre os modelos avaliados, alguns modelos merecem atenção especial: aqueles

modelos, que além de classificarem padrões, geram também regras lingüísticas sobre os clientes, definindo perfis de *churners*. São eles: o NFHB, as Árvores de Decisão e o Rule-Evolver. As regras criadas por tais modelos podem contribuir de forma significativa para a compreensão do cliente, como será visto adiante.

No entanto, antes de discutir alguns exemplos de regras encontradas, é importante familiarizar-se com os conceitos de acurácia e abrangência para que a avaliação das regras possa ser completa.

O conceito de acurácia de uma dada regra SE *C* ENTÃO *P*, onde *C* é o número total de observações que satisfaz a condição SE e *P* é o número total de observações que satisfazem a afirmativa ENTÃO, mede o grau de confiabilidade de uma regra. Como a Equação 6.1 demonstra, a acurácia de uma regra corresponde à porcentagem de padrões que satisfazem os requisitos de *C* e *P* dentre todas as observações que satisfazem a condição *C*.

$$acurácia = \frac{|C \,\&\, P|}{|C|} \qquad \textit{Equação 6.1}$$

Por outro lado, a abrangência de uma regra pode ser interpretada como o grupo de todos os padrões que satisfazem a regra. A Equação 6.2 apresenta o cálculo da abrangência como a razão entre as observações que satisfazem *C* e *P* por todas aquelas que representam *P*.

$$abrangência = \frac{|C \,\&\, P|}{|P|} \qquad \textit{Equação 6.2}$$

Com essas definições, algumas das regras geradas pelos métodos citados podem ser avaliadas, objetivando um melhor entendimento das razões que levam ao *churn*.

Abaixo está uma das melhores regras gerada pelo Rule-Evolver, com acurácia de 63% e abrangência de 71%. Os valores das variáveis presentes na regra, com exceção da entrada OSC_FATURA, estão normalizados por amplitude segundo a Equação 3.2. Foram deixados assim pois é interessante avaliá-los dessa forma.

SE

OSC_FATURA ≥ 0.04	**E**	OSC_FATURA ≤ 1.14
E FATURA_MEDIA ≥ 0.04	**E**	FATURA_MEDIA ≤ 0.48
E ANTIG_CONTA ≥ 0.06	**E**	ANTIG_CONTA ≤ 0.47
E ANTIG_APARELHO ≥ 0.09	**E**	ANTIG_APARELHO ≤ 0.75

ENTÃO

CLIENTE = **CHURNER**

Observando-se essa regra com cuidado, nota-se que ela pode prover bastante informação sobre um dos perfis de cliente *churner* típicos. A primeira linha da regra não é relevante, pois inclui quase todo o domínio da variável OSC_FATURA (que mede como o valor da fatura do cliente oscilou ao longo dos nove meses presentes na base de dados). No entanto, a segunda e terceira linha merecem atenção especial, pois elas apontam claramente na direção de que um cliente com altas chances de abandonar a empresa possui fatura mensal abaixo da média da população (FATURA_MEDIA ≤ 0.48) e possui um contrato relativamente novo com a empresa (ANTIG_CONTA ≤ 0.47). A quarta linha da regra que trata da variável ANTIG_APARELHO (a idade do aparelho do cliente), também traz pouca informação, dado que o intervalo de variação é grande demais.

Avaliando uma das regras geradas pelo NFHB, que incorpora conceitos linguísticos como BAIXO e ALTO ao invés de intervalos de variação como o Rule-Evolver, tem-se que:

SE

ANTIG_CONTA é BAIXA

E FATURA_MÉDIA é BAIXA

E CALL_CENTER é ALTA

E QUANT_TELS_DIF_ENT é BAIXA

ENTÃO

CLIENTE = CHURNER

Essa regra ilustra que clientes churners geralmente possuem baixas faturas mensais médias e um relacionamento recente com a operadora, utilizam o serviço de *call center* com freqüência (talvez para reclamar da operadora), e possuem um baixo número de telefones diferentes ligando para eles (o que reduz o receio de trocar de operadora e perder o número de telefone utilizado).

Essas duas regras são alguns exemplos dentre as muitas geradas pelos dois métodos mencionados. As regras da Árvore de Decisão apresentaram resultados similares. Essas regras são bons exemplos do potencial que tais modelos possuem em gerar inteligência comportamental para a compreensão do *churn* e suas razões.

Esses resultados e a literatura citada demonstram quão complicada é uma base de dados de *churn* para tarefas de classificação de padrões. Entretanto, os resultados obtidos podem oferecer grandes vantagens competitivas para empresas que apliquem o sistema de retenção desenvolvido para guiar suas ações de combate ao *churn*, como a análise da lucratividade dos modelos demonstra a seguir.

6.3.8. Análise da lucratividade

Para se calcular a lucratividade de cada modelo, segundo o modelo de lucratividade de [MOZE00], é necessário saber: o tamanho do universo de clientes envolvidos na ação de retenção planejada; a taxa de retenção esperada com os incentivos; o custo dos incentivos; determinar uma janela de tempo na qual os incentivos poderiam fazer efeito; saber a fatura média do grupo que se pretende atingir com a ação; e o custo de aquisição de um novo cliente.

Para exemplificar, imagina-se o seguinte possível cenário: uma operadora com dois milhões de clientes, uma janela de seis meses para o incentivo fazer efeito, e um custo de aquisição de um novo cliente de R$300,00. Escolhendo um único cenário, com a taxa de retenção igual a 50%, fatura média igual a R$100,00 e custo do incentivo igual a R$20,00, alcançam-se os resultados da Tabela 6.9. Os resultados estão em lucro por cliente *churner*.

Tabela 6.9 – Lucratividade para os modelos segundo os resultados do conjunto de testes.

	Lucratividade	
Modelo	**churn mensal de 1,25%**	**churn mensal de 3,00%**
Classificador Bayesiano	- R$ 20,24	R$ 60,03
Rede Neural MLP	R$ 32,60	R$ 88,60
Rede Neural PNN	R$ 22,36	R$ 82,09
Árvore de Decisão	R$ 0,68	R$ 69,75
Rule-Evolver	- R$ 59,36	R$ 45,17
NFHB	- R$ 2,16	R$ 74,37
SVM	R$ 4,32	R$ 34,19

Da Tabela 6.9 observa-se que, com uma taxa de 1,25% de *churn* mensal (que é a taxa presente na base de dados), alguns modelos não conseguem atingir o patamar de geração de lucros e na verdade geram custos para a operadora (valores negativos na Tabela 6.9), principalmente devido à classificação errada de clientes ativos, o que provoca um gasto em incentivos para um número grande de clientes que não iriam abandonar a empresa. Os modelos de redes neurais são os únicos que apresentam desempenho razoavelmente bom sob uma taxa de *churn* de 1,25%. Supondo que os modelos possuíssem o mesmo desempenho apresentado para o caso de uma taxa de *churn* mensal de 3%, que, segundo a Tabela 4.2 é a taxa média de algumas operadoras nos dias de hoje, todos os modelos se justificariam em termos de retorno de investimento, especialmente as redes neurais e o NFHB, demonstrando o valor do sistema de retenção de clientes desenvolvido.

Apesar desta análise de um único cenário ser ilustrativa, crê-se ser mais interessante realizar uma análise dinâmica da lucratividade, variando-se a taxa de retenção esperada pela aplicação do sistema, a fatura média dos clientes a serem atingidos e o custo dos incentivos. Assim, é possível obter as Figuras 6.2, 6.3, 6.4 e 6.5. Essas figuras detalham a lucratividade possível que uma empresa pode ter ao aplicar o sistema proposto com o modelo, dependendo do nível de cliente que se deseja reter (fatura mensal do cliente, no eixo *y* dos gráficos), do preço do incentivo que se pode oferecer (eixo *x* dos gráficos) e da taxa de retenção esperada (uma para cada gráfico). As figuras correspondem à aplicação do modelo de redes neurais MLP, que apresentou o desempenho mais promissor na etapa anterior.

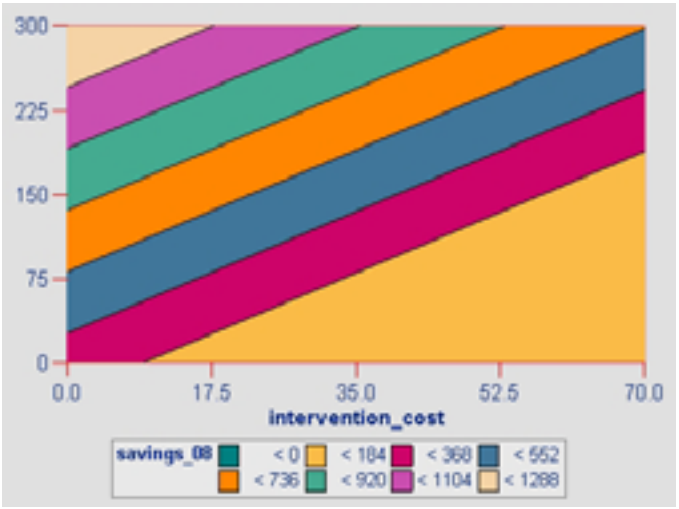

Figura 6.2 – Faixas de lucratividade para uma taxa de retenção de 80%.

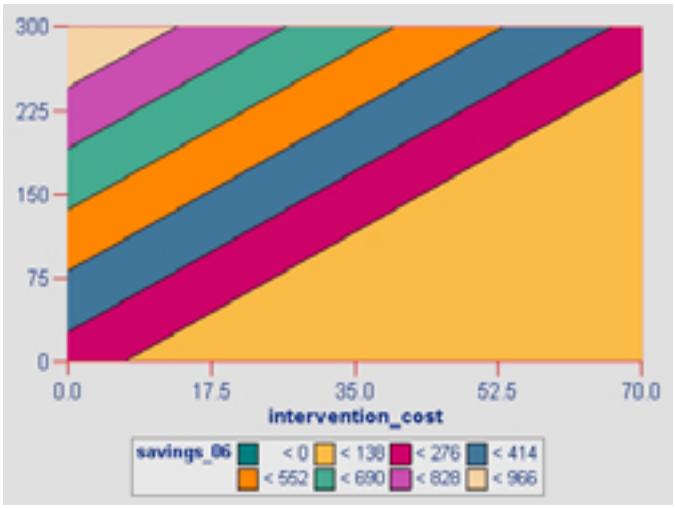

Figura 6.3 – Faixas de lucratividade para uma taxa de retenção de 60%.

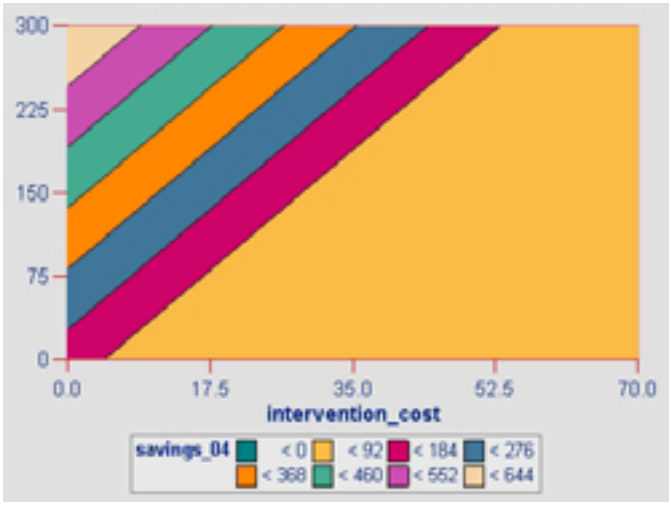

Figura 6.4 – Faixas de lucratividade para uma taxa de retenção de 40%.

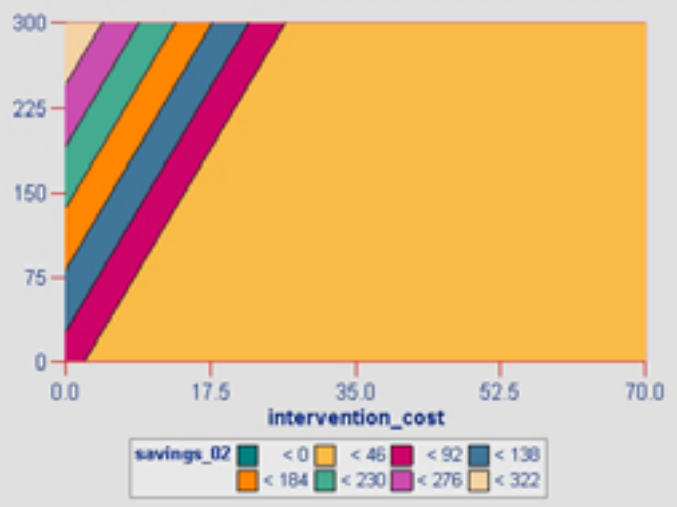

Figura 6.5 – Faixas de lucratividade para uma taxa de retenção de 20%.

As Figuras 6.2 a 6.5 demonstram a lucratividade advinda da aplicação do sistema de retenção de clientes. Na figura 6.2, por exemplo, correspondente a uma taxa de retenção de 80%, pode-se observar que, se um incentivo de R$70,00 for oferecido a clientes com faturas maiores do que R$200,00 (i.e. clientes valiosos), a empresa terá um lucro na faixa de R$368,00 por cliente *churner* por mês, comparado ao custo que ela teria se deixasse o cliente ir embora. Na medida que a taxa de retenção diminui, observa-se que somente incentivos muito baratos e para clientes de alto valor surtiriam algum efeito na geração de lucratividade para a empresa.

É importante concluir que o sistema de auxílio à retenção proposto deve ser empregado em clientes de alto valor que a empresa deseja manter. Não poder escolher clientes desta natureza para a base de dados utilizada neste estudo foi mais um grande limitador.

Baseado nos resultados obtidos até aqui, para a conclusão do sistema só falta implementá-lo em ambiente de produção voltado para a população de clientes que a empresa deseja manter e prosseguir com seu monitoramento até que alguma renovação seja necessária.

6.4. Resumo

O sistema de auxílio à retenção de clientes proposto, ao ser aplicado em uma base de dados real, demonstrou resultados interessantes que sem dúvida

podem ser de grande valia para a operação de empresas na indústria de telefonia celular.