Relatórios de Pesquisa em Engenharia de Produção V. 9 n. 06

UM MODELO PARA CLASSIFICAÇÃO DE CLIENTES DE OPERADORAS DE PLANOS DE SAÚDE

Vinícius de Paula Mendes (*vinicius@frg.com.br*)
Real Grandeza Fundação de Previdência e Assistência Social

Annibal Parracho Sant'Anna (*tppaps@vm.uff.br*)
Universidade Federal Fluminense

Resumo

Este trabalho desenvolve um modelo estatístico para o cancelamento dos clientes assinantes de planos de saúde. O modelo empregado é um modelo de regressão logística tendo como variáveis explicativas variáveis transacionais, demográficas e dados sobre o histórico de eventos. O objetivo é definir o perfil dos clientes com maior risco de cancelamento. Os resultados do ajustamento do modelo indicam que o perfil do assinante com maior risco de cancelamento da sua assinatura inclui: alto tempo entre exames e consultas, valores baixos de mensalidades atrasos nos pagamentos, menor uso de serviços opcionais, recência do vínculo e pouca utilização do plano. Os resultados da análise são, finalmente, confirmados empregando-se, para obtenção de classes de clientes com probabilidades semelhantes de evasão, um pareamento (matching) baseado em escores de propensão.

Palavras chave: Marketing de Relacionamento, Regressão Logística, Risco de cancelamento (Churn), Propensity Score Matching

Versão final recebida em 14/11/2009 - Publicado em 17/12/2009

1. INTRODUÇÃO

A grande maioria das empresas está focada em medir o retorno dos projetos, para investir recursos que maximizem os resultados e tragam retorno financeiro. Ao mesmo tempo, são realizadas pesquisas para identificar os clientes de maior valor e os clientes com maior risco de desligamento ou cancelamento (Churn).

A partir dessas pesquisas, algumas estratégias são selecionadas. Uma delas é o programa de relacionamento com clientes (CRM – Customer Relationship Management), que têm por objetivo aumentar o valor do cliente. Objetivos da diferenciação de clientes são encontrar os clientes de maior valor e os clientes de maior potencial de se tornar Churn.

Administradoras de planos de saúde, por exemplo, procuram formas de estimar o valor do cliente não pautando somente no valor de receita gasto por cliente ao longo do relacionamento. A sua disposição ao desligamento, ou a propensão a se tornar Churn é uma das avaliações possíveis de serem adotadas para medir a eficiência de um programa de relacionamento. E é nesse aspecto de propensão ao Churn, que esse estudo se concentrará.

A Seção 2 deste artigo apresenta o conjunto de dados analisado. A Seção sumariza a metodologia empregada baseada nos modelos de regressão logística. A Seção 4 apresenta os resultados do ajustamento. A Seção 5 discute esses resultados. A Seção 6 utiliza o recurso do pareamento para reavaliar o efeito das características estudadas.

2. OS DADOS

O banco de dados utilizado nesse estudo pertence a uma grande operadora de planos de saúde no Estado do Rio de Janeiro. Para a análise, foi preciso efetuar alguns filtros até que chegasse à base final. Entre os filtros, o principal deles foi o de só selecionar clientes (ativos ou inativos) que começaram a usar o plano de saúde dessa operadora a partir de 1999. A base final utilizada continha **130.552** assinaturas (domicílios), sendo **63.849 (48,9%)** domicílios ativos no plano de saúde e **66.703 (51,1%)** domicílios inativos no plano.

O filtro na base de dados original é justificado de acordo com a pauta desse estudo, que visa caracterizar clientes e suas possíveis chances de se tornar Churn. Visto isso, os planos vigentes e criados antes de 1999 são caracterizados por possuírem indivíduos, ou titulares nos domicílios, mais idosos e consequentemente menos propensos ao Churn. Pois esses

mesmos planos, adquiridos antes de 1999 trazem benefícios improváveis de serem conquistados novamente.

Além da variável principal, que distingue se um cliente é Ativo ou não (Cancelado ou não), outras variáveis importantes foram consideradas. Para decidir que variáveis poderiam ser consideradas no estudo, primeiramente se fez um levantamento empírico das informações contidas no banco de dados até a data de 31/07/2007 (última data disponível de atualização do banco). Para melhor identificação do modelo, algumas informações precisaram ser codificadas e recodificadas. Abaixo, seguem as variáveis analisadas, separadas por natureza Demográfica, Geográfica, Cadastral e Transacional:

Quadro 1: Descrição das Variáveis, presente no banco de dados

Natureza	Variável	Descrição
Demográficas	**Sexo**	Sexo do cliente
Demográficas	**Estado Civil**	Estado Civil do cliente
Demográficas	**Idade**	Idade do cliente
Geográficas	**Bairro**	Bairro onde se encontra o domicílio do cliente
Geográficas	**Área Administrativa**	Classifica dos bairros, segundo área administrativa - IBGE
Geográficas	**Área de Rendimento**	Classifica as áreas administrativas, segundo participação de receita com a empresa no rendimento do domicílio
Cadastral	**Data de inclusão**	Data de início do relacionamento
Cadastral	**Agregados**	Quantidade de agregados associados ao plano
Cadastral	**Dependentes**	Quantidade de dependentes associados ao plano
Cadastral	**Rede de Produtos**	Grupo de produtos onde se classifica o plano do associado
Cadastral	**Opcional Odontologia**	Identifica se o cliente possui opcional de odontologia
Cadastral	**Opcional Emergências**	Identifica se o cliente possui opcional de emergências
Cadastral	**Opcional Viagens**	Identifica se o cliente possui opcional de viagem
Cadastral	**Opcional Air**	Identifica se o cliente possui opcional de vôo
Cadastral	**Flag DCC**	Flag que identifica se o cliente optou por pagar o plano em débito em conta
Transacional	**Tempo de Consulta**	Tempo, em dias, desde a última consulta
Transacional	**Tempo de Exame**	Tempo, em dias, desde o último exame
Transacional	**Flag de Atraso**	Flag que identifica se o cliente atrasou o pagamento nos últimos 12 meses
Transacional	**Valor T1**	Último valor pago à empresa no consolidado do mês
Transacional	**Valor T2**	Penúltimo valor pago à empresa no consolidado do mês
Transacional	**Valor T3**	Antepenúltimo valor pago à empresa no consolidado do mês
Transacional	**Segmento de Utilização**	Variável recodificada que identifica o perfil de uso do plano com a empresa
Transacional	**Internação**	Quantidade de Internações nos últimos 12 meses
Transacional	**Cirurgia**	Quantidade de Cirurgias nos últimos 12 meses

Nesse banco de dados de 130.552 clientes analisados, clientes que possuíam planos individuais (com ou sem dependentes), as variáveis significativas consideradas para determinar o risco de cancelamento (Churn), no modelo de Regressão Logística, foram treze ao todo (em ordem de importância): Tempo de Exame, Segmento de Utilização, Flag de

Atraso, Tempo de Consulta, Ano de Inclusão, Valor T3, Faixa Etária, Área de Rendimento, Valor T2, Rede de Produto, Flag de Opcional, Sexo e Estado Civil.

A tabela de classificação mostrou que a taxa de acerto geral do modelo de Regressão Logística é acima de 85%, e que as taxas de acerto dos grupos individuais são altas e indicam uma consistência na previsão de qualquer um dos dois grupos. O grupo que cancela plano de saúde apresentou taxa de acerto acima de 80%, enquanto o grupo que não cancela, tem taxa de acerto acima de 90%. Esses números oferecem sustentação ao uso dos modelos de Regressão Logística para se estimar a probabilidade de um cliente se tornar Inativo.

3. ANÁLISE DE REGRESSÃO LOGÍSTICA

A Regressão Logística reescreve o modelo clássico de regressão linear de modo a confirmar o valor da variável resposta para a faixa de 0 (zero) a 1 (um), ao mesmo tempo em que as variáveis independentes possam variar continuamente. Isto é obtido pela equação abaixo:

$$\hat{Y}_i = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k}}$$

onde X1, X2,..., Xk representam as K variáveis explicativas e os β's, seus respectivos parâmetros, estimados através do método da Máxima-Verossimilhança. (β0 representa o parâmetro do intercepto). E representa a probabilidade estimada para cada indivíduo i.

Assim, a Regressão Logística é aplicada a uma variável dependente dicotômica, onde a variável dependente não representa os valores de dados brutos, mas representa a probabilidade do evento estudado ocorrer.

Nos modelos Logit, a principal suposição é a de que o LN(Odds), ou seja, logaritmo da razão entre as probabilidades de ocorrência e não ocorrência do evento é linear. Sendo assim, a equação da função Logit que descreve uma relação linear na Regressão Logística é:

$$\ln\left[\frac{\hat{Y}_i}{1-\hat{Y}_i}\right] = \beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k$$

onde os termos da direita são os termos padrão para as variáveis independentes e o intercepto numa equação de Regressão linear. E do lado esquerdo está o ln(Odds) é chamada de Logit.

Na Regressão Logística há um relacionamento linear com as variáveis independentes, mas é linear nas probabilidades de LOG e não nas probabilidades originais. Como o objeto de estudo é a probabilidade de ocorrência de um evento (o indivíduo se tornar Churn), a equação Logit pode ser transformada numa equação na probabilidade (Paula, 2004).

Diferente da Regressão linear clássica, os erros desse modelo não seguem uma distribuição normal, mas sim a de Bernoulli. Na Regressão Logística usa-se o método da máxima verossimilhança para se estimar os valores dos parâmetros $\beta 0$, ..., βk, que maximizem a probabilidade de se obter o conjunto observado de dados (Hosmer e Lemeshow, 1989).

Os parâmetros da Regressão Logística podem ser estimados de forma bem semelhante à Regressão linear múltipla pelo fato de que um modelo de base é primeiro estimado visando a fornecer um padrão para comparação. Na Regressão linear múltipla, a média é utilizada para estabelecer o modelo base e calcular a soma total dos quadrados dos afastamentos. Na Regressão Logística, o mesmo processo é utilizado, com a média usada no modelo estimado não para estabelecer a soma dos quadrados, mas para estabelecer o valor de probabilidade log.

4. MODELO DE REGRESSÃO

Segundo Hosmer e Lemeshow (2000), o objetivo da Regressão Logística é achar o melhor relacionamento entre a variável resposta (variável dependente) e um conjunto de variáveis explicativas ou preditivas, sendo o modelo final aquele que apresentar o melhor ajuste e for naturalmente razoável de se explicar.

Foram analisados contratos vigentes entre Agosto de 2006 a Julho de 2007. A base de dados da administradora de planos de saúde possui todas as informações dos clientes ativos e inativos, a partir de 1999.

Lu (2001) propõe que consumidores com menos de três meses de relacionamento sejam excluídos da análise, em função do baixo tempo de relacionamento. Devido à atuação do Marketing de Relacionamento da empresa concentrar-se no Estado do Rio de Janeiro, somente clientes residentes neste Estado foram considerados.

Outra seleção dos dados, como registrado na Tabela 2 a seguir, trata da exclusão de clientes, com data de inclusão anterior a 1999, tendo em vista a diferente legislação que regulamenta os planos de saúde anteriores a esta data.

E por fim, faz-se necessário o uso do filtro, talvez o mais importante de todos, de exclusão dos clientes assegurados por contratos empresariais, uma vez que quem determina a permanência ou não no plano não é o cliente, e sim o empregador.

Utilizando-se os filtros anteriormente propostos, e considerando-se somente domicílios, obteve-se um total **130.552** clientes dentre ativos e inativos, com data base final de *Julho de 2007*, sendo **63.849 (48,9%)** domicílios ativos e **66.703 (51,1%)** domicílios inativos. O Quadro 2 lista as variáveis finais que serão investigadas para obtenção do Escore de Churn.

Quadro 2: Descrição das variáveis utilizadas na Regressão Logística

Natureza	Variável	Descrição
Demográficas	**Sexo**	Sexo do cliente
Demográficas	**Estado Civil**	Estado Civil do cliente
Demográficas	**Idade**	Idade do cliente
Demográficas	**Faixa Etária**	Recodifica a idade em grupos de idade
Geográficas	**Bairro**	Bairro onde se encontra o domicílio do cliente
Geográficas	**Área Administrativa**	Classifica dos bairros, segundo área administrativa - IBGE
Geográficas	**Área de Rendimento**	Classifica as áreas administrativas, segundo participação de receita com a empresa no rendimento do domicílio
Cadastral	**Data de inclusão**	Data de início do relacionamento
Cadastral	**Safra de Entrada**	Recodificação na data de inclusão, considerando o ano de entrada
Cadastral	**Relacionamento**	Diferença, em dias, entre a atulização do banco (30/7/07) e a sua inclusão
Cadastral	**Agregados**	Quantidade de agregados associados ao plano
Cadastral	**Dependentes**	Quantidade de dependentes associados ao plano
Cadastral	**Flag Dependente /Agregado**	Variável recodificada que identifica se há pelo menos 1 agregado ou dependente associado ao plano
Cadastral	**Rede de Produtos**	Grupo de produtos onde se classifica o plano do associado
Cadastral	**Opcional Odontologia**	Identifica se o cliente possui opcional de odontologia
Cadastral	**Opcional Emergências**	Identifica se o cliente possui opcional de emergências
Cadastral	**Opcional Viagens**	Identifica se o cliente possui opcional de viagem
Cadastral	**Opcional Air**	Identifica se o cliente possui opcional de vôo
Cadastral	**Flag de Opcional**	Variável recodificada que identifica se há pelo menos 1 opcional
Cadastral	**Flag DCC**	Flag que identifica se o cliente optou por pagar o plano em débito em conta
Transacional	**Tempo de Consulta**	Tempo, em dias, desde a última consulta
Transacional	**Tempo de Exame**	Tempo, em dias, desde o último exame
Transacional	**Flag de Atraso**	Flag que identifica se o cliente atrasou o pagamento nos últimos 12 meses
Transacional	**Valor T1**	Último valor pago à empresa no consolidado do mês
Transacional	**Valor T2**	Penúltimo valor pago à empresa no consolidado do mês
Transacional	**Valor T3**	Antepenúltimo valor pago à empresa no consolidado do mês
Transacional	**Segmento de Utilização**	Variável recodificada que identifica o perfil de uso do plano com a empresa
Transacional	**Internação**	Quantidade de Internações nos últimos 12 meses
Transacional	**Cirurgia**	Quantidade de Cirurgias nos últimos 12 meses
Transacional	**Flag de Internação /Cirurgia**	Variável recodificada que identifica se há pelo menos 1 internação ou cirurgia nos últimos 12 meses

b

maximiza a precisão do modelo, foi adotado o método de iteração *stepwise forward*, que através da estatística de Wald encontrou o modelo parcimonioso com 13 variáveis. Em um segundo momento, utilizando as 13 variáveis finais, o modelo de Churn foi novamente validado.

O ponto de corte utilizado para classificação foi o de probabilidade igual **0.5**, pois define probabilidade de cancelamento igual para os dois grupos (ativos e inativos). Como a amostra possui porcentagem de clientes inativos muito próxima a 50%, foi possível adotar tal corte.

O resultado do modelo inicial apresenta a tabela de classificação considerando o modelo com apenas uma constante, ou seja, se arbitrariamente todas as assinaturas fossem consideradas canceladas, a taxa de acerto seria de **51,1%**. O modelo de Regressão Logística que irá estimar o risco de cancelamento de clientes precisa ser mais assertivo na classificação dos clientes.

Tabela 1: Modelo apenas com a constante

Tabela de Classificação		Predito		Percentual
		Não Churn	Churn	
Observado	Não Churn	0	63849	0%
	Churn	0	66703	100%
Percentual Médio				51,1%

A primeira variável incluída no modelo é a que tiver a estatística de pontuação mais alta, estatística Wald. No caso, a variável Tempo desde último exame é selecionada a compor o modelo. Em segundo lugar, a variável Segmento de uso do plano foi incorporada ao modelo. E em seguida, a variável dicotômica que indica Atraso no pagamento. Essas três variáveis contribuem com **70,5%** do poder explicativo do modelo.

O modelo de Regressão Logística final, para estimar o Churn individual, ficou ajustado da seguinte forma:

$$\hat{Y}_i = \frac{e^{\hat{\beta}_0 + \sum_{i=1}^{13} \hat{\beta}_i X_i}}{1 + e^{\hat{\beta}_0 + \sum_{i=1}^{13} \hat{\beta}_i X_i}}$$

onde X_1, X_2,..., X_{12} representam as variáveis explicativas e os β's, seus respectivos parâmetros, estimados por Máxima-Verossimilhança. (β_0 representa o parâmetro do intercepto). E representa a probabilidade estimada de cada indivíduo **i** se tornar Churn.

X_1 – Tempo de Exame (*tempo desde o último, em dias*),

X_2 – Segmento de Utilização,

X_3 – Flag de Atraso,

X_4 – Tempo de Consulta (*tempo desde a última, em dias*),

X_5 – Ano de Inclusão,

X_6 – Valor em T3 (*valor pago no antepenúltimo mês*),

X_7 – Faixa Etária,

X_8 – Área de Rendimento,

X_9 – Valor em T2 (*valor pago no penúltimo mês*),

X_{10} – Rede de Produto,

X_{11} – Opcional,

X_{12} – Sexo,

X_{13} – Estado Civil.

A Tabela 2 contém as variáveis presentes no modelo ajustado e seus respectivos parâmetros. Adiante se discutirá o ajuste do modelo e suas interpretações.

Tabela 2: Estimação do Modelo de Churn

Variáveis no Modelo	β	Desvio Padrão	Estatística Wald	Graus de Liberdade	P-Valor	EXP (β)	Intervalo de Confiança 95% - EXP (β	
							Limite Inferior	Limite Superior
Ano de Inclusao - 2007			594,443	8	0,000			
Ano de Inclusao - 1999	0,508	0,090	31,857	1	0,000	1,662	1,393	1,98
Ano de Inclusao - 2000	0,412	0,088	22,037	1	0,000	1,510	1,271	1,79
Ano de Inclusao - 2001	0,161	0,041	15,934	1	0,000	1,175	1,085	1,27
Ano de Inclusao - 2002	-0,006	0,037	0,023	1	0,880	0,994	0,926	1,06
Ano de Inclusao - 2003	0,279	0,035	63,964	1	0,000	1,322	1,235	1,41
Ano de Inclusao - 2004	0,072	0,028	6,798	1	0,009	1,074	1,018	1,13
Ano de Inclusao - 2005	-0,180	0,028	42,726	1	0,000	0,835	0,792	0,88
Ano de Inclusao - 2006	-0,602	0,030	396,772	1	0,000	0,548	0,516	0,58
Sexo - Masculino	-0,218	0,019	135,151	1	0,000	0,804	0,775	0,83
Faixa Etária - 59 anos ou mais			510,154	9	0,000			
Faixa Etária - até 18 anos	-0,484	0,024	394,199	1	0,000	0,616	0,587	0,64
Faixa Etária - de 19 a 23 anos	-0,180	0,035	25,820	1	0,000	0,835	0,779	0,89
Faixa Etária - de 24 a 28 anos	-0,037	0,026	1,978	1	0,160	0,964	0,916	1,01
Faixa Etária - de 29 a 33 anos	0,020	0,026	0,580	1	0,446	1,020	0,970	1,07
Faixa Etária - de 34 a 38 anos	-0,060	0,027	4,838	1	0,028	0,942	0,893	0,99
Faixa Etária - de 39 a 43 anos	0,083	0,030	7,868	1	0,005	1,087	1,025	1,15
Faixa Etária - de 44 a 48 anos	0,126	0,036	12,004	1	0,001	1,134	1,056	1,21
Faixa Etária - de 49 a 53 anos	0,363	0,046	62,716	1	0,000	1,437	1,314	1,57
Faixa Etária - de 54 a 58 anos	0,185	0,060	9,414	1	0,002	1,203	1,069	1,35
Rede de Produto - Personal			114,340	5	0,000			
Rede de Produto - Alfa	0,105	0,022	23,574	1	0,000	1,110	1,064	1,15
Rede de Produto - Beta	0,161	0,026	39,928	1	0,000	1,175	1,118	1,23
Rede de Produto - Delta	0,187	0,027	48,326	1	0,000	1,206	1,144	1,27
Rede de Produto - N/D	-0,436	0,049	78,862	1	0,000	0,647	0,587	0,71
Rede de Produto - Omega	-0,071	0,055	1,648	1	0,199	0,931	0,835	1,03
Área de Rendimento - A			135,320	4	0,000			
Área de Rendimento - B	0,092	0,019	23,507	1	0,000	1,096	1,056	1,13
Área de Rendimento - C	0,164	0,020	66,802	1	0,000	1,178	1,133	1,22
Área de Rendimento - Outros RJ	-0,485	0,046	112,953	1	0,000	0,616	0,563	0,67
Área de Rendimento - Outros	0,169	0,042	16,133	1	0,000	1,184	1,090	1,28
Estado Civil - Solteiro			59,111	2	0,000			
Estado Civil - Casado	0,053	0,019	8,194	1	0,004	1,055	1,017	1,09
Estado Civil - Outros	0,071	0,029	5,966	1	0,015	1,074	1,014	1,13
Seg de Utilização - Sem Segmento			3920,667	4	0,000			
Seg de Utilização - Cirurgia/Internação	0,768	0,024	1041,942	1	0,000	2,156	2,058	2,25
Seg de Utilização - Consulta/Exame Alta	0,481	0,021	536,463	1	0,000	1,617	1,553	1,68
Seg de Utilização - Consulta/Exame Baixa	-0,033	0,016	4,292	1	0,038	0,967	0,937	0,99
Seg de Utilização - Consulta/Exame Média	0,454	0,016	807,111	1	0,000	1,574	1,525	1,62
Flag de Atraso	1,804	0,027	4456,581	1	0,000	6,075	5,762	6,40
Flag Opcional	-0,314	0,021	225,488	1	0,000	0,731	0,701	0,76
Tempo de Consulta	0,074	0,002	2078,837	1	0,000	1,076	1,073	1,08
Tempo de Exame	0,174	0,002	8059,072	1	0,000	1,189	1,185	1,19
Valor T2	-0,001	0,000	320,298	1	0,000	0,999	0,999	0,99
Valor T3	-0,002	0,000	1167,411	1	0,000	0,998	0,998	0,99
Constante	-1,546	0,048	1016,635	1	0,000	0,213		

No modelo final encontram-se variáveis contínuas (Valor T2, Valor T3, Tempo de Consulta e Tampo de Exame), variáveis categóricas (Ano de Inclusão, Faixa Etária, Rede de Produto, Área de Rendimento, Estado Civil e Segmento de Utilização) e variáveis binárias (Sexo, Flag de Atraso e Flag Opcional). A interpretação dos parâmetros **β**'s acontece na seguinte forma:

- Variáveis Contínuas: Sendo **β** positivo, a medida que é acrescido um valor na variável contínua, aumenta-se a probabilidade de Churn. Sendo **β** negativo, diminui-se a probabilidade de Churn. A real relação pode ser vista através da Razão de Chance, ou EXP (**β**). Exemplificando, a razão de chance da variável Tempo de Exame é de 1,189. Isto significa dizer que, aumentando-se o Tempo de Exame em 1 (um) dia, a chance de se tornar Inativo aumenta na razão de 1,189.

- Variáveis Categóricas: É escolhida arbitrariamente uma categoria da variável como referência e, a partir dessa referência, se comparam as outras categorias da variável. Em outras palavras, por exemplo, em Área de Rendimento, a categoria de referência escolhida foi "A". Significa então que morar numa área de Rendimento "B" ou "C" aumenta a chance de se tornar Churn em relação a quem reside numa área de Rendimento "A".

- Variáveis Binárias: A presença da variável incide positiva ou negativamente na probabilidade de Churn (de acordo com o **β**). Exemplo, ter o Flag Opcional diminui a chance do indivíduo se tornar inativo (**β** negativo e EXP(**β**) menor que 1).

Além das interpretações dos parâmetros, se faz necessário analisar algumas estatísticas de validação, nos modelos de Regressão Logística.

A começar pela *Deviance*. A estatística de probabilidade *-2log likelihood (Deviance)* diminuiu à medida que foi incluída uma variável no modelo, indicando melhora. Em contrapartida, as medidas *pseudo* R^2 aumentaram à medida que previsores foram adicionados. O *pseudo* R^2 *de Nagelkerke* no último passo ficou em **0.695** (Tabela 6), considerado pela literatura como um excelente poder de explicação do modelo.

Tabela 3: Máxima verossimilhança e R^2

	- 2 LOG Verossimilhança	R^2 de Cox e	R^2 de Nagelkerke
Modelo	84682,377	0,522	0,6950

A medida Hosmer e Lemeshow de ajuste geral tem um teste estatístico que indica que não houve diferença estatisticamente significativa entre as classificações observadas e previstas para todos os modelos com duas ou mais variáveis. O valor Hosmer e Lemeshow mede a correspondência dos valores efetivos e previstos da variável dependente. Neste caso, o melhor ajuste do modelo é indicado por uma diferença menor na classificação observada e prevista. Um bom ajuste de modelo é indicado por um valor Qui-quadrado (Chi-square) não significante (Hair et al., 2005).

O SPSS 13.0 © utiliza algoritmos iterativos que buscam um subconjunto de variáveis para maximizar a probabilidade. Não é o mesmo que maximizar a precisão da estimativa. Então, pode haver problemas na utilização de métodos de regressão quando o objetivo da análise é a precisão da estimativa.

Tabela 4: Teste de Hosmer e Lemeshow

Teste de Hosmer e Lemeshow

	Chi-Quadrado	graus de liberdade	P-valor
Modelo	9773,344	8	0,0000

A medida Hosmer e Lemeshow, apresentada no modelo (Tabela 4), é significativa a muito altos níveis de significância (*Sig*). Essa medida indica a ausência de diferença significativa na distribuição de valores dependentes efetivos e previstos. Essas medidas combinadas sugerem a aceitação do modelo do último passo como um modelo significante de Regressão Logística.

A tabela de classificação final, também utilizada quando se descreve um modelo na Técnica de Análise Discriminante (Hair et al., 2005), mostra taxas de acerto extremamente altas, de casos corretamente classificados para o modelo final proposto, de 13 variáveis. Como se verifica na Tabela 5.

Tabela 5: Classificação no modelo final

Tabela de Classificação		Predito		Percentual
		Não Churn	Churn	
Observado	Não Churn	58167	5682	91,1%
	Churn	10434	56269	84,4%
Percentual Médio				**87,7%**

A taxa de acerto geral foi de 87,7%. Além disso, as taxas de acerto de grupos individuais são consistentemente altas e não indicam um problema na previsão de qualquer um dos dois grupos. Apesar de altas, as taxas de acerto do grupo "não se torna Churn" é ainda maior que a taxa do grupo que "se torna Churn", 91,1% contra 84,4%.

O modelo inicial, que considerava apenas a constante, tinha uma taxa geral de acerto de 51,1%. O modelo completo com 13 variáveis aumenta 1,7 vezes a taxa de acerto na previsão.

Nos últimos passos do stepwise, a melhora no R^2 é pequena e a taxa de acerto geral do modelo não se altera tanto. Isto indica que as últimas variáveis inseridas no modelo não apresentam tanto peso para determinar o Escore de Churn. No entanto, tais variáveis permaneceram no modelo final para ajudar na definição do perfil dos domicílios que cancelam a assinatura (contrato) do plano de saúde. A Tabela 6 (abaixo) indica a Estatística de Wald que considera a importância de cada variável no modelo proposto.

O valor percentual foi calculado a partir dos valores absolutos de cada variável através da Estatística de Wald. De maneira percentualizada é possível comparar a importância das variáveis.

Tabela 6: Peso das variáveis no modelo final

	Estatística WALD	% Peso
Tempo de Exame	8.059,07	31,4%
Segmento de Utilização	6.310,48	24,6%
Flag Atraso	4.456,58	17,3%
Tempo de Consulta	2.078,84	8,1%
Ano Inclusão	1.174,45	4,6%
Valor em T3	1.167,41	4,5%
Faixa Etária	1.029,57	4,0%
Área de Rendimento	354,72	1,4%
Valor em T2	320,30	1,2%
Rede de Produto	306,68	1,2%
Opcional	225,49	0,9%
Sexo	135,15	0,5%
Estado Civil	73,27	0,3%

As 3 variáveis mais importantes para o modelo, ou seja, as que carregam a maior informação relacionada à probabilidade do indivíduo se desligar do plano de saúde são, respectivamente: o Tempo de Exame, a Segmentação de Utilização e o Flag de Atraso de Pagamento. Juntas, essas três variáveis carregam mais de 73% de toda variabilidade

explicada pelo modelo Logit. Tempo de consulta se apresenta como uma quarta variável (no nível de importância). Essas quatro variáveis ultrapassam os 80% da explicação do modelo probabilístico.

5. INTERPRETAÇÃO DOS RESULTADOS

Como será visto mais adiante, as principais variáveis explicativas do modelo, apresentam os sinais esperados, no que diz respeito à correlação com o Churn.

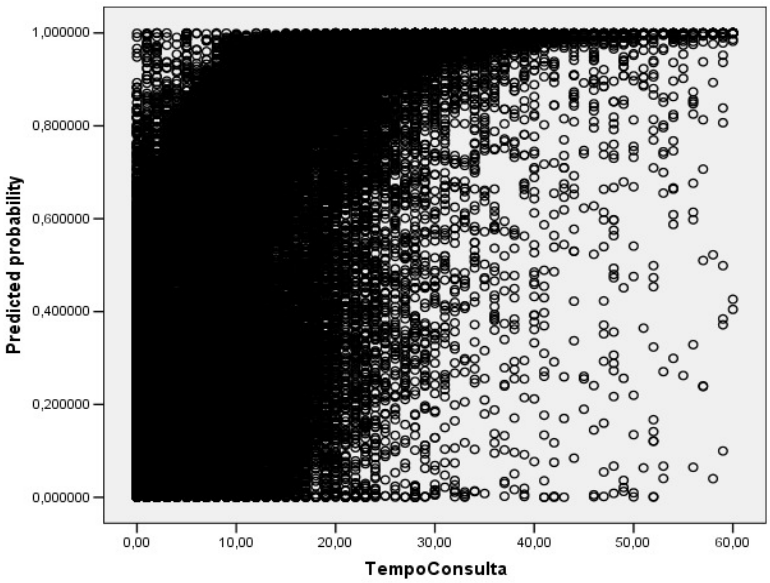
A. Tempo de Consulta – Esta variável representa, em dias, o tempo decorrido desde a última consulta. É de se esperar que pacientes com elevado tempo desde sua última consulta, tenham uma maior probabilidade de se tornar ex-cliente.

Tabela 7: Comparativo no Tempo de Consulta entre Churn e não-Churn

Tempo de Consulta	Mínimo	Máximo	Média	Mediana	Desvio Padrão
Cliente	0,00	101,00	5,25	2,00	8,46
Ex-Clilente	0,00	103,00	19,66	19,00	10,22
Total	**0,00**	**103,00**	**12,61**	**10,00**	**11,84**

Observando a Tabela 7 acima, é possível distinguir o grupo dos clientes que se tornaram Churn, (com tempo médio e tempo mediano de consulta acima de 19 dias); e o grupo dos clientes Não-Churn (com baixo tempo de consulta).

Figura 1. Dispersão entre o Tempo de Consulta e Probabilidade de Churn



A Figura 1 mostra através da Probabilidade predita (Predicted Probability) que quanto maior é o tempo de consulta do paciente, maior é sua chance de se tornar um Churn.

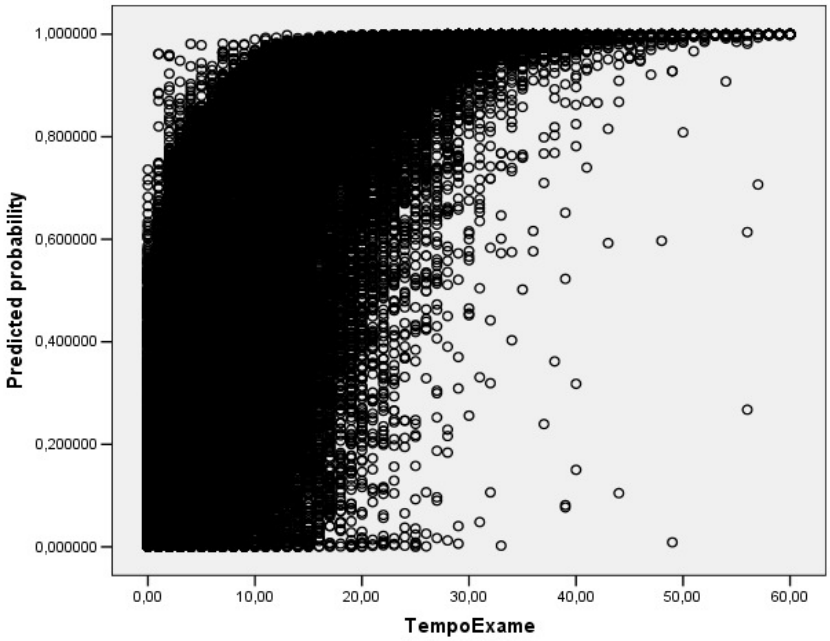
B. Tempo de Exame – Variável que representa, em dias, o tempo decorrido desde o último exame.

Tabela 8: Comparativo no Tempo de Exame entre Churn e não-Churn

Tempo de Exame	Mínimo	Máximo	Média	Mediana	Desvio Padrão
Cliente	0,00	99,00	4,65	3,00	7,05
Ex-Clilente	0,00	103,00	19,73	20,00	10,26
Total	**0,00**	**103,00**	**12,36**	**8,00**	**11,62**

O comportamento da variável Tempo de Exame se mostra similar ao comportamento do Tempo de consulta; ou seja, ex-clientes possuem um elevado nível em dias no tempo de exame, como se pode ver na Tabela 8.

Figura 2. Dispersão entre o Tempo de Exame e Probabilidade de Churn



Em outras palavras, à medida que se aumenta o Tempo de exame, aumenta a probabilidade de se tornar um ex-cliente. Pode-se ver isso na Figura 2.

C. Valor T2 – Representa o valor da mensalidade do plano, pago à empresa, no penúltimo mês. É razoável supor que clientes ativos tenham uma distribuição de valores superiores aos clientes inativos. A Tabela 9, abaixo, confirma essa suposição.

Tabela 9: Comparativo no Penúltimo Valor Pago entre Churn e não-Churn

Valor T2	Mínimo	Máximo	Média	Mediana	Desvio Padrão
Cliente	0,00	9742,00	918,94	428,00	691,35
Ex-Clilente	0,00	7029,00	263,63	172,00	352,04
Total	**0,00**	**9742,00**	**437,40**	**264,00**	**573,26**

D. Valor T3 - Valor da mensalidade do plano, pago à empresa, no antepenúltimo mês. Clientes que não pagam esses valores (valor t2 e t3) tendem se tornar inativos. Tal afirmação se confirma através da variável dicotômica criada para verificação de inadimplência no pagamento (flag Atraso). Mais ainda, a Tabela 10 indica semelhança entre as variáveis Valor t2 e t3. Era de se esperar que os valores t1, t2 e t3 tivesse comportamentos semelhantes, visto que um assinante contrata um plano mensal num valor estipulado.

Tabela 10: Comparativo Antepenúltimo Valor Pago entre Churn e não-Churn

Valor T3	Mínimo	Máximo	Média	Mediana	Desvio Padrão
Cliente	0,00	9982,00	640,48	440,00	710,83
Ex-Clilente	0,00	8544,00	263,81	174,00	349,21
Total	**0,00**	**9982,00**	**448,03**	**266,00**	**587,26**

E. Flag Atraso – Indica se o cliente teve algum “não pagamento” nos últimos três meses. Onde 1 (um) indica atraso em um dos três possíveis pagamentos. Intuitivamente, tem-se que os clientes com atraso são os que mais se desligam do plano de saúde.

Tabela 11: Churn versus Atraso - Absoluto

		Churn		Total
		Não	Sim	
Atraso	Não	59.151	42.918	**102.069**
	Sim	4.698	23.785	**28.483**
Total		**63.849**	**66.703**	**130.552**

Tabela 12: Churn versus Atraso - Percentual

		Churn		Total
		Não	Sim	
Atraso	Não	45,3%	32,9%	**78,2%**
	Sim	3,6%	18,2%	**21,8%**
Total		**48,9%**	**51,1%**	**100,0%**

Nas tabelas 11 e 12 acima, se evidencia quem aqueles assinantes em atraso têm maior chance de se tornarem inativos.

F. Ano de Inclusão – Ano de entrada do participante na empresa (posterior a 1998). Proporcionalmente, clientes que entraram em 2004 e 2005 são mais propensos a se tornarem Churn, como se pode ver através das Tabelas 13 e 14 abaixo.

Tabela 13: Churn versus Ano de Inclusão - Absoluto

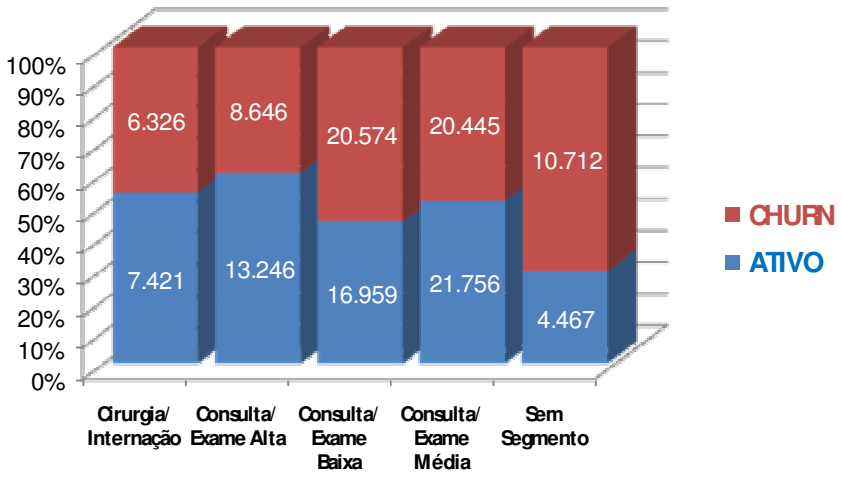
		Churn		Total
		Não	Sim	
Ano de Inclusão	1999	1.080	382	**1.462**
	2000	1.042	509	**1.551**
	2001	4.893	3.418	**8.311**
	2002	6.277	4.562	**10.839**
	2003	5.520	6.187	**11.707**
	2004	9.102	19.133	**28.235**
	2005	12.411	20.148	**32.559**
	2006	21.750	11.825	**33.575**
	2007	1.774	539	**2.313**
Total		**63.849**	**66.703**	**130.552**

Tabela 14: Churn versus Ano de Inclusão - Percentual

		Churn		Total
		Não	Sim	
Ano de Inclusão	1999	1,7%	0,6%	**1,1%**
	2000	1,6%	0,8%	**1,2%**
	2001	7,7%	5,1%	**6,4%**
	2002	9,8%	6,8%	**8,3%**
	2003	8,6%	9,3%	**9,0%**
	2004	14,3%	28,7%	**21,6%**
	2005	19,4%	30,2%	**24,9%**
	2006	34,1%	17,7%	**25,7%**
	2007	2,8%	0,8%	**1,8%**
Total		**100,0%**	**100,0%**	**100,0%**

G. Segmento de Utilização – Indica forma mais habitual com que os clientes utilizam os planos (Definição em Anexo 1). Dos clientes que não costumam utilizar, os considerados "Sem segmento", espera-se uma maior probabilidade de Churn. De fato isso ocorre, como mostra a Figura 3.

Figura 3. Distribuição entre os Clientes Ativos e Churn através do Segmento



H. Opcional – Indica a presença de algum tipo de opcional (Opcional Odontologia, Opcional p/ Emergências, Opcional para Viagens, Opcional Air) no plano original do paciente.

Tabela 15: Churn versus Opcional - Absoluto

		Churn		Total
		Não	Sim	
Opcional	Não	41.149	52.732	**93.881**
	Sim	22.700	13.971	**36.671**
Total		**63.849**	**66.703**	**130.552**

Tabela 16: Churn versus Opcional - Percentual

		Churn		Total
		Não	Sim	
Opcional	Não	31,5%	40,4%	**71,9%**
	Sim	17,4%	10,7%	**28,1%**
Total		**48,9%**	**51,1%**	**100,0%**

Os pacientes com opcionais têm mais chance de não permanecer ativo. Essa afirmação é corroborada através das Tabelas 15 e 16.

I. Faixa Etária – Quanto maior a faixa etária, maior é a chance de haver clientes ativos. Isso pode ser explicado pelo alto valor do plano, associado à faixa etária, mas principalmente à necessidade da idade. Clientes idosos necessitam recorrer mais ao plano se comparados aos mais jovens. Isso se evidencia nas Tabelas 17 e 18, abaixo.

Tabela 17: Churn versus Faixa Etária - Absoluto

.

		Churn		Total
		Não	Sim	
Faixa Etária	até 18 anos	18.669	19.075	**37.744**
	entre 19 e 23	3.174	4.286	**7.460**
	entre 24 e 28	7.037	9.720	**16.757**
	entre 29 e 33	6.888	9.678	**16.566**
	entre 34 e 38	6.318	7.785	**14.103**
	entre 39 e 43	5.148	6.057	**11.205**
	entre 44 e 48	3.639	3.519	**7.158**
	entre 49 e 53	2.624	2.290	**4.914**
	entre 54 e 58	2.100	1.380	**3.480**
	59 anos ou mais	8.252	2.913	**11.165**
Total		**63.849**	**66.703**	**130.552**

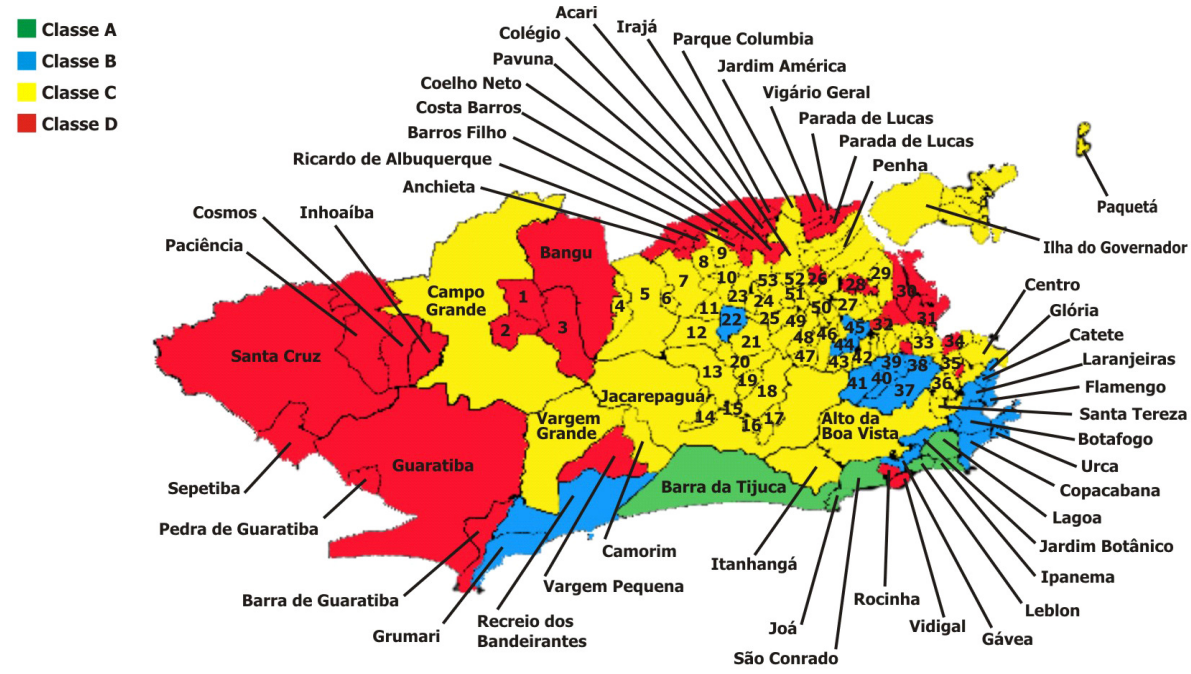
Tabela 18: Churn versus Faixa Etária - Percentual

		Churn		Total
		Não	Sim	
Faixa Etária	até 18 anos	29,2%	28,6%	**28,9%**
	entre 19 e 23	5,0%	6,4%	**5,7%**
	entre 24 e 28	11,0%	14,6%	**12,8%**
	entre 29 e 33	10,8%	14,5%	**12,7%**
	entre 34 e 38	9,9%	11,7%	**10,8%**
	entre 39 e 43	8,1%	9,1%	**8,6%**
	entre 44 e 48	5,7%	5,3%	**5,5%**
	entre 49 e 53	4,1%	3,4%	**3,8%**
	entre 54 e 58	3,3%	2,1%	**2,7%**
	59 anos ou mais	12,9%	4,4%	**8,6%**
Total		**100,0%**	**100,0%**	**100,0%**

A definição das Faixas Etárias no estudo pautou-se nas faixas definidas originalmente na operadora de plano de saúde. Em virtude disso, só é possível notar diferenças significativas na última faixa de idade.

J. Área de Rendimento - Classificação de Áreas Administrativas do Estado do Rio de Janeiro, segundo participação de receita da área com a empresa. Essa variável detecta regiões, no Estado do Rio de Janeiro, com maior proporção de clientes com menor poder econômico. Vale lembrar que os bairros que correspondem a cada Área de Rendimento no Estado do Rio de Janeiro podem ser vistos na Figura 4.

Figura 4. Distribuição Econômica nos Bairros do Estado do Rio de Janeiro



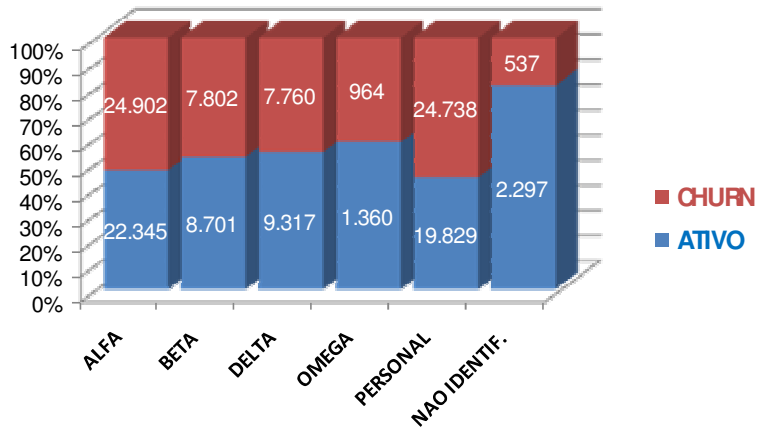
1.Santíssimo	10.Marechal Hermes	19.Pechincha	28.Complexo do Alemão	37.Tijuca	46.Engenho de Dentro
2.Senador Vasconcelos	11.Campo dos Afonsos	20.Tanque	29.Ramos	38.Maracanã	47.Água Santa
3.Senador Camará	12.Jardim Sulacap	21.Praça Seca	30.Maré	39.Vila Isabel	48.Pilares
4.Padre Miguel	13.Taquara	22.Vila Valqueire	31.Portuária	40.Andaraí	49.Quintino
5.Realengo	14.Curicica	23.Bento Ribeiro	32.Jacarezinho	41.Grajaú	50.Piedade
6.Magalhães Bastos	15.Cidade de Deus	24.Madureira	33.São Cristóvão	42.Engenho Novo	51.Tomás Coelho
7.Vila Militar	16.Gardênia Azul	25.Cascadura	34.Santo Cristo	43.Lins	52.Vicente de Carvalho
8.Deodoro	17.Anil	26.Vila Cosmos	35.Estácio	44.Méier	53.Vaz Lobo
9.Guadalupe	18.Freguesia	27.Inhaúma	36.Rio Comprido	45.Cachambi	

Fonte: IBGE (2004)

Apesar de ser significativa no modelo probabilístico de Churn, não há diferença significativa entre clientes e ex-clientes nas Áreas de Rendimento, ou seja, não se destaca nenhuma área em relação a outra (como visto na Tabela 2).

K. Rede de Produto - Grupo de produto onde se classificam os planos dos associados. Quanto melhor classificado é o plano, menos chance de se tornar inativo tem o cliente. Isso se deve a importância e os benefícios que esse plano lhe traz.

Figura 5. Distribuição entre os Clientes Ativos e Churn através do Plano



Personal é o plano mais básico, na rede de produtos, seguidos respectivamente pelos planos Alfa, Beta, Delta e Omega. Planos antigos ou mal especificados pela operadora foram alocados em Não identificados. As diferenças entre os planos (rede de produtos) citados são: a presença de privilégios, tais como uma rede de hospitais diferenciada, a opção de quarto ou enfermaria, entre outros. A Figura 5 ratifica a intuição de que quanto melhor é o plano do assinante, menos chance o mesmo tem de ser tornar Inativo.

6. O MÉTODO DE MATCHING

Matching (ou pareamento) é um método amplamente utilizado na literatura de avaliação de tratamentos. O método consiste basicamente em tomar como base as características das unidades tratadas e tentar encontrar unidades em um grupo de controle não experimental que possuam as mesmas características, previamente definidas no grupo de tratamento. Em seguida, o grupo de comparação é emparelhado ao grupo de tratamento através do *Propensity Escore* (Escore de propensão ou probabilidade predita de participação).

Segundo Hirano, Imbens e Ridder (2003) o efeito médio para uma subpopulação com um dado valor para as variáveis observáveis pode ser estimado simplesmente tirando a diferença entre as médias dos grupos de tratamento e controle nestas subpopulações. Para

que se possa obter uma estimativa não viesada do efeito do tratamento, tem de identificar um grupo de controle que seja o mais próximo possível do grupo de tratamento em termos das características gerais que são capturadas por uma variável designada de X. Rosenbaum e Rubin (1983) sustentam que ajustando as diferenças entre as unidades de tratamento e controle, apenas através do escore de propensão, todo o viés associado às diferenças nas variáveis prévias observáveis é removido.

A utilização do escore de propensão baseia-se em duas hipóteses-chave. A primeira é que a seleção nos observáveis requer que a participação no programa seja independente dos resultados, condicional nas covariáveis.

A segunda hipótese refere-se à existência de um suporte comum. Esta condição requer que existam unidades de ambos os grupos, tratamento e controle, para cada característica X para o qual se deseja comparar.

É necessário que $0 < P(X) < 1$. Isto assegura que para cada indivíduo tratado exista outro indivíduo não tratado pareado, com valores similares de X. Dessa forma, os indivíduos devem possuir uma probabilidade de serem participantes ou não participantes que se situe entre 0 e 1, não podendo ser esta igual aos extremos. Em outras palavras, a variável X precisa seguir uma distribuição de probabilidade.

O Escore de Propensão é definido por Rosenbaum e Rubin (1983) como a probabilidade condicional de receber um tratamento, dadas as características a priori:

$$p(X) \equiv Pr\{D = 1|X\} = E\{D|X\}.$$

onde D = {0, 1} é o indicador da exposição ao tratamento e X é o vetor multidimensional de características a priori.

O principal objetivo do PSM (*Propensity Score Matching*) é descobrir o efeito médio do tratamento (ATT), ou seja, responder à pergunta: o que aconteceria com o grupo de controle se recebesse o tratamento e o que aconteceria com o grupo de tratamento se ele não o recebesse?

Nesse estudo, o PSM será utilizado para descobrir e emparelhar indivíduos com características parecidas, com resultados contrários, ou seja, indivíduos que se tornaram Churn (perdidos) versus indivíduos que não se tornaram. Fazendo esse emparelhamento é possível definir dois grupos similares para comparação de perfil.

Se a distribuição do tratamento fosse aleatória dentro de uma amostra (isto é, se o experimento fosse natural), essa pergunta teria uma resposta simples, a saber: bastaria testar

a diferença de médias da variável supostamente impactada pelo tratamento para os grupos de tratamento e de controle.

No modelo probabilístico, a probabilidade de um indivíduo se tornar Churn é regredida sobre seus supostos determinantes a fim de corrigir o viés de seleção na distribuição do tratamento. Nesse sentido, essa probabilidade é o escalar que se procura.

Depois, há o pareamento das probabilidades estimadas. Esse casamento é realizado da seguinte forma: seja $\hat{p}(X_i)$ a probabilidade de se tornar Churn do indivíduo i, que se tenha tornado efetivamente Churn. Se dentro de um raio (pequeno) partindo de $\hat{p}(X_i)$ existir pelo menos um $\hat{p}(X_j)$, em que j é um indivíduo que não se tornou Churn, os indivíduos i e j formarão um par tratamento-controle (Churn-não Churn). Assim, podem-se acompanhar esses grupos no tempo e efetuar o teste de médias a fim de calcular o efeito médio do "suposto tratamento" (neste estudo, o "tratamento" é o indivíduo se tornar Churn).

A metodologia de *Matching* consiste na escolha de um grupo de controle ideal a partir de uma amostra maior que a amostra do grupo de tratamento. O grupo de controle é "casado" com o grupo de tratamento a partir de um conjunto de características observadas ou utilizando um Escore de propensão (probabilidade de Churn dadas certas características). Quanto mais próximos esses Escores, melhor o "Matching". Conforme citou Ravallion (2003), um bom grupo de controle vem do mesmo ambiente econômico, social e cultural do grupo de tratamento. Nesse caso, a escolha do grupo de controle (indivíduos que não se tornaram Churn) se dará de forma aleatória.

Existem várias metodologias de aplicação do "Matching", as mais utilizadas na literatura e apresentadas por Becker e Ichino (2002) são:

- Vizinho mais próximo;
- Radius Matching;
- Kernel Matching e;
- Estratificação.

O método de Estratificação consiste na divisão das observações em intervalos de acordo com o Escore de propensão. Dentro de cada intervalo, a média do Escore de propensão das observações de controle e de tratamento deve ser igual. Para facilitar o processo, na prática, podem ser utilizados os mesmos blocos identificados pelo algoritmo da estimação do Escore de propensão. Em cada intervalo computa-se a diferença entre as médias da variável de

resultado das observações de controle e das observações de tratamento. O ATT final será a média do ATT de cada bloco ponderada pela distribuição das observações entre os blocos.

Uma das desvantagens dessa metodologia é que não são levadas em consideração observações que pertençam a um bloco onde apenas existam observações de controle ou apenas observações de tratamento. Uma maneira de resolver este problema seria utilizando a metodologia de "matching-Vizinho Mais Próximo", pois esta procura, para cada observação tratada, a observação de controle com o Escore de propensão a Churn mais próximo. Esta busca pode ser feita com ou sem reposição, visto que uma observação de controle pode ser o melhor "matching" para mais de uma observação tratada. Uma vez encontrados todos os "matching", a diferença dos resultados de cada grupo é computada e o ATT final será a média dessas diferenças.

Algumas vezes, a qualidade do "matching" pode não ser muito boa tendo em vista que o controle mais próximo de uma observação tratada pode estar bem distante em termos de Escore de propensão a Cancelamento. As metodologias "Radius matching" e "Kernel matching" têm uma solução para este problema.

No método de Radius matching, cada unidade tratada é "casada" com um controle pertencente a uma vizinhança predefinida com base no Escore de propensão do tratado. Se definirmos uma vizinhança restrita, existe uma grande chance de não encontrarmos controle dentro desta vizinhança para todas as observações tratadas.

Por outro lado, quanto menor a vizinhança, melhor será a qualidade do Matching, e mais próximos estarão os grupos de tratamento e de controle. O método de kernel leva em consideração todas as observações de tratamento e de controle, e essas são casadas (pareadas) de maneira ponderada.

Todos os controles são aproveitados e o peso utilizado para cada um é inversamente proporcional à distância do seu Escore de propensão e o Escore da observação tratada. Vale ressaltar que, em todos os métodos, a qualidade do Matching pode ser melhorada quando impomos uma região de suporte comum. A escolha do método a ser utilizado vai depender do tipo de dados que se tem disponível. É clara, a existência de um *trade-off* entre a qualidade do Matching e a quantidade de observações casadas.

A aplicação do PSM para encontrar indivíduos com potenciais a Churn apresenta vantagem metodológica sobre as demais alternativas, como, por exemplo, a de definir como potenciais a Churn indivíduos que não se tornaram Churn, mas apresentam $\hat{p}(X) > 0{,}5$.

A primeira vantagem é que a escolha de p(X) = 0,5 é arbitrária. A segunda vantagem é que o PSM possibilita a identificação dos indivíduos com Potencial a Churn "ocultos", o que o corte de probabilidade pode perder. Cabe salientar, entretanto, que a qualidade dessa classificação depende do modelo probabilístico.

Para aplicação do PSM foi utilizado um algoritmo desenvolvido por Raynald Levesque e adaptado ao SPSS 13.0 © por John Painter (Fev 2004). Este algoritmo possui uma limitação de seleção de amostra de até 20.000 indivíduos para classificação do PSM. Assim, foi selecionada uma amostra de 10.000 clientes Ativos (ou assinantes) e, através do algoritmo, foram encontrados 10.000 clientes Inativos ou Churn (um para cada indivíduo Ativo) para formarem os pares de probabilidade. Encontrados os públicos-alvos adequados, que possam ser testados numa ação de recuperação de clientes Inativos, passa-se a traçar um comparativo entre os dois públicos (10.000 ativos e 10.000 inativos) Abaixo, seguem comparações das características entre grupos. Cada Tabela representa o perfil dos clientes e ex-clientes selecionados através do PSM. Em cada tabela, podem-se notar diferenças entre os grupos. No final, se fará uma interpretação dos resultados.

Tabela 19: Churn PSM por Sexo

	Clientes		Ex-Clientes	
Sexo	Absoluto	%	Absoluto	%
Feminino	6.560	65,6%	5.880	58,8%
Masculino	3.440	34,4%	4.120	41,2%
Total	**10.000**	**100,0%**	**10.000**	**100,0%**

Tabela 20: Churn PSM por Faixa Etária

	Clientes		Ex-Clientes	
Faixa Etária	Absoluto	%	Absoluto	%
até 18 anos	2.390	23,9%	2.470	24,7%
entre 19 e 23	600	6,0%	690	6,9%
entre 24 e 28	1.540	15,4%	1.020	10,2%
entre 29 e 33	1.640	16,4%	1.240	12,4%
entre 34 e 38	1.330	13,3%	970	9,7%
entre 39 e 43	1.140	11,4%	780	7,8%
entre 44 e 48	480	4,8%	460	4,6%
entre 49 e 53	340	3,4%	480	4,8%
entre 54 e 58	200	2,0%	270	2,7%
59 anos ou mais	340	3,4%	1.620	16,2%
Total	**10.000**	**100,0%**	**10.000**	**100,0%**

Tabela 21: Churn PSM por Safra de Entrada (Ano de Inclusão)

	Clientes		Ex-Clientes	
Safra de Entrada	Absoluto	%	Absoluto	%
1999	40	0,4%	100	1,0%
2000	80	0,8%	100	1,0%
2001	430	4,3%	770	7,7%
2002	880	8,8%	620	6,2%
2003	1.340	13,4%	770	7,7%
2004	6.640	66,4%	1.290	12,9%
2005	590	5,9%	2.260	22,6%
2006	0	0,0%	3.790	37,9%
2007	0	0,0%	300	3,0%
Total	**10.000**	**100,0%**	**10.000**	**100,0%**

Tabela 22: Churn PSM por Tipo de Plano

	Clientes		Ex-Clientes	
Plano	Absoluto	%	Absoluto	%
ALFA	5.260	52,6%	2.910	29,1%
BETA	1.010	10,1%	1.350	13,5%
DELTA	1.060	10,6%	1.540	15,4%
OMEGA	150	1,5%	300	3,0%
PERSONAL	2.520	25,2%	3.770	37,7%
NAO IDENTIFICADO	0	0,0%	130	1,3%
Total	**10.000**	**100,0%**	**10.000**	**100,0%**

Tabela 23: Churn PSM por Área de Rendimento

	Clientes		Ex-Clientes	
Area de Rendimento	Absoluto	%	Absoluto	%
A	1.160	11,6%	1.750	17,5%
B	4.380	43,8%	4.770	47,7%
C	3.990	39,9%	2.980	29,8%
OUT CIDADE DO RJ	60	0,6%	200	2,0%
Outros	410	4,1%	300	3,0%
Total	**10.000**	**100,0%**	**10.000**	**100,0%**

Tabela 24: Churn PSM por Segmento de Utilização

	Clientes		Ex-Clientes	
Segmento	Absoluto	%	Absoluto	%
Cirurgia/Internação	770	7,7%	910	9,1%
Consulta/Exame Alta	740	7,4%	720	7,2%
Consulta/Exame Baixa	4.000	40,0%	3.230	32,3%
Consulta/Exame Média	2.950	29,5%	2.150	21,5%
Sem Segmento	1.540	15,4%	2.990	29,9%
Total	**10.000**	**100,0%**	**10.000**	**100,0%**

Tabela 25: Churn PSM por Estado Civil

	Clientes		Ex-Clientes	
Estado Civil	Absoluto	%	Absoluto	%
Solteiro	6.710	67,1%	6.330	63,3%
Casado	2.920	29,2%	2.910	29,1%
Outros	370	3,7%	760	7,6%
Total	**10.000**	**100,0%**	**10.000**	**100,0%**

Tabela 26: Churn PSM por Flag de Atraso

	Clientes		Ex-Clientes	
Flag de Atraso	Absoluto	%	Absoluto	%
Sim	4.560	45,6%	7.820	78,2%
Não	5.440	54,4%	2.180	21,8%
Total	**10.000**	**100,0%**	**10.000**	**100,0%**

Tabela 27: Churn PSM por Tempo de Consulta, Tempo de Exame, e pelos Valores pagos da penúltima e antepenúltima parcela

EX-CLIENTE	Tempo de Consulta	Tempo de Exame	Valor T2	Valor T3
Média	34,77	35,48	135,10	120,91
Mediana	32,00	33,00	74,00	100,00
Desvio Padrão	9,75	9,21	201,78	182,53

CLIENTE	Tempo de Consulta	Tempo de Exame	Valor T2	Valor T3
Média	17,99	18,02	475,72	469,05
Mediana	10,00	12,00	297,50	306,00
Desvio Padrão	20,00	18,00	572,43	557,73

A análise comparativa da seleção através do PSM, resumida pelas Tabelas 22 a 30, acima, fornece indícios de que os grupos selecionados não apresentam características semelhantes, apesar de termos os grupos selecionados a partir dos pares de probabilidades aproximados.

Em algumas variáveis essas diferenças não ocorrem significativamente, tais como Idade, Estado Civil e Área de Rendimento. No restante, há significativas diferenças. Além disso, pode-se observar que o grupo selecionado de ex-clientes é caracterizado por possuir planos (rede de produtos) mais diferenciados, ter tempo de exame e de consulta menores em relação aos clientes.

Significa dizer que existem perfis diferentes de indivíduos ativos e inativos com propensão a se tornarem Churn semelhantes. Em virtude disso, é importante atestar esses diferentes perfis e traçar planejamentos distintos para cada grupo de clientes.

7. CONCLUSÃO

O principal objetivo deste estudo foi elaborar um modelo probabilístico de regressão logística que descrevesse, para cada indivíduo presente na base analisada de clientes da empresa, o risco de cancelamento no plano de saúde. Consequentemente, após a aplicação do modelo de Churn, foi possível traçar o perfil dos assinantes mais propensos a se tornar Churn, como também os menos propensos. Além de traçar perfil dos ex-assinantes.

É possível afirmar que o modelo proposto pode ser eficiente para determinação do risco de cancelamento de clientes, a partir de variáveis geográficas, demográficas e transacionais.

Nos 130.552 clientes analisados, as variáveis significativas, no modelo Logit, para determinar a propensão a Churn foram 13 (em ordem de importância): Tempo de Exame, Segmento de Utilização, Flag de Atraso, Tempo de Consulta, Ano de Inclusão, Valor T3, Faixa Etária, Área de Rendimento, Valor T2, Rede de Produto, Flag de Opcional, Sexo e Estado Civil.

A tabela de classificação mostrou que a taxa de acerto geral do modelo de Regressão Logística é de 87,7% e que as taxas de acerto dos grupos individuais são altas e indicam uma consistência na previsão de qualquer um dos dois grupos. O grupo que cancela apresentou taxa de acerto de 84,4% enquanto o grupo que não cancela tem taxa de acerto de 91,1%. O pseudo R^2 de Nagelkerke apresentou um poder de explicação de 0,695 e a medida

Hosmer e Lemeshow de ajuste geral através de um teste estatístico indica que não houve diferença estatisticamente significativa entre as classificações observadas e previstas para o modelo final. Além disso, o valor de ***-2LL*** aumentou a cada passo. A combinação dessas medidas de avaliação do ajuste e da precisão do modelo indica a aceitação deste como um modelo de regressão logística significante.

Assim, de acordo com as variáveis assumidas no modelo de regressão logística, conclui-se que o perfil do assinante com maior risco de cancelamento da sua assinatura é: aquele com maior tempo entre exames e consultas, com valores mais baixo de mensalidades e, consequentemente, rede de produtos inferiores, clientes que atrasam mais os pagamentos e têm menos opcionais, clientes mais novos e que pouco utilizam o plano.

Finalmente foi realizada uma comparação entre os clientes assinantes ativos e os clientes inativos utilizando o *Propensity Score Matching* numa amostra de 20.000 clientes. O objetivo desta comparação era caracterizar assinantes e ex-assinantes com probabilidades a Churn similares. Do ponto de vista mercadológico, a principal contribuição do PSM consiste no uso alternativo de seleção de público, o qual pode ser facilmente replicado para outros problemas de pesquisa. Esta análise permitiu afirmar que o modelo proposto pode ser eficiente para determinação do risco de cancelamento de clientes, a partir de variáveis geográficas, demográficas e transacionais.

REFERÊNCIAS BIBLIOGRÁFICAS

ANS – Agência Nacional de Saúde Suplementar, < http://www.ans.gov.br>, 2003.

BECKER S.O., ICHINO, A. Estimation of average treatment effects based on propensity score. Stata Journal, v2, p.358-377, 2002. Disponível em: www.sobecker.de/pscore.html.

BERGER, Paul D.; NASR, NadaI. Customer Lifetime Value: Marketing Models and Applications, Journal of Interactive Marketing, 12 (Winter), 17–30, 1998.

CISTER, Angelo M.; EBECKEN, Nelson F. F. - "CRM through DM: a case study" – Third International Conference on Data Mining- DATA MINING III, Bolongna, Italy-2002. . CISTER, Angelo M. Mineração de dados para a análise de atrito em telefonia móvel tese de doutorado - UFRJ - Eng. Civil, 2005.

GUPTA, Sunil; - LEHMANN, Donald R. Gerenciando Clientes como Investimentos: o valor estratégico dos clientes a longo prazo, ISBN: 85-363-0669-6: Bookman, 2006.

HAIR, Joseph F. Jr.; ANDERSON. Rolph E.; TATHAM Ronald L.; BLACK William C. Análise Multivariada de Dados, ISBN 85-363-0482-0: Bookman 2005.

HECKMAN J.; Ichimura H.; SMITH J.; TODD P.,Matching as an econometrics evaluation estimator – Review of Economics Studies,65, 261-294, 1998.

HIRANO, K., IMBENS, G. W., RIDDER, G. Efficient estimation of average treatment effects using the estimated propensity score. Cambridge, MA.: National Bureau of Economic Research, 2003. Disponível em: http://ideas.repec.org/s/nbr/nberwo.html.

HOSMER, David; LEMESHOW, Stanley. Applied Logistic Regression – 2º Edition. University of Massachusetts, Amherst, Massachusetts, 2000.

IBGE – Instituto Brasileiro de Geografia e Estatística, < http://www.ibge.gov.br>, 2004

KUMAR, V.; RAMANI, G.; BOHLING, T. Customer Lifetime Value Approaches and Best Practice Applications, Journal of Interactive Marketing, Vol 18. pag 60-72, 2004. .

LU, Junxiang. Predicting Customer Churn in the Telecommunication: An Aplication of Survival Analysis Modeling Using SAS. Kansas – USA, Sprint Communications Company, 2001. .

MENARD, SCOTT. Applied logistic regression analysis. 1995.

PAULA, G.A. Modelos de regressão com apoio computacional. IME - USP. 294p. Disponível em <http:// www.ime.usp.br/~giapaula>, 2004.

ROSENBAUM, P., Rubin, D. “The central role of the propensity score in observational studies for causal effects”. Biometrika 70, 41–55. Rubin, D., 1983.