4
Análise dos Resultados

4.1
Construção do Modelo de Regressão Logística

No SPSS 13.0, foi aplicado o modelo de regressão logística binário, método *stepwise foward*, para definir o modelo final que minimiza o número de variáveis e maximiza a precisão do modelo.

Um ponto importante é a definição do ponto de corte. Analisado o universo de assinaturas, nota-se que 69% das assinaturas existentes no banco de dados são ativas. Como a amostra respeita a proporção populacional, adotar esta taxa como ponto de corte parece ser o mais adequado. Quando não se conhece a proporção populacional, costuma-se usar o ponto de corte 0,5 que define probabilidades iguais para os dois grupos.

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 31%. O modelo de regressão logística que irá estimar o risco de cancelamento de clientes precisa ser mais assertivo na classificação dos clientes.

Classification Table[a,b]

			Predicted		
			STATUS		Percentage Correct
	Observed		CANCELAD	ATIVO	
Step 0	STATUS	CANCELAD	11057	0	100.0
		ATIVO	24492	0	.0
	Overall Percentage				31.1

a. Constant is included in the model.

b. The cut value is .690

A primeira variável a ser incluída no modelo será aquela que tiver a estatística de pontuação mais alta, estatística Wald, no caso a variável tempo de permanência é selecionada a compor o modelo. Em segundo lugar, a variável

forma de pagamento é incorporada ao modelo. E em seguida, o indicador de reclamação. Essas 3 variáveis contribuem com 92,1% do poder explanatório do modelo.

Na tabela abaixo, verificamos que a análise direcionada a passos utilizando a estatística Wald consumiu 11 passos até se obter o modelo final. Observando-se as significâncias estatísticas do modelo, constatamos que o coeficiente é significativo a cada passo.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	8848.068	5	.000
	Block	8848.068	5	.000
	Model	8848.068	5	.000
Step 2	Step	1499.479	2	.000
	Block	10347.547	7	.000
	Model	10347.547	7	.000
Step 3	Step	297.834	1	.000
	Block	10645.381	8	.000
	Model	10645.381	8	.000
Step 4	Step	272.349	3	.000
	Block	10917.730	11	.000
	Model	10917.730	11	.000
Step 5	Step	217.242	6	.000
	Block	11134.972	17	.000
	Model	11134.972	17	.000
Step 6	Step	229.119	9	.000
	Block	11364.091	26	.000
	Model	11364.091	26	.000
Step 7	Step	165.895	8	.000
	Block	11529.987	34	.000
	Model	11529.987	34	.000
Step 8	Step	148.315	6	.000
	Block	11678.302	40	.000
	Model	11678.302	40	.000
Step 9	Step	125.520	12	.000
	Block	11803.821	52	.000
	Model	11803.821	52	.000
Step 10	Step	49.970	1	.000
	Block	11853.791	53	.000
	Model	11853.791	53	.000
Step 11	Step	26.362	4	.000
	Block	11880.154	57	.000
	Model	11880.154	57	.000

Após diversas interações, o modelo final selecionou 11 das 14 variáveis incluídas inicialmente no modelo. Excluiu as variáveis: gênero, indicador de compra de produto agregado e indicador de compra de um anúncio de publicidade. As variáveis que resultaram do modelo final são: “tempo de permanência”, “forma de pagamento”, “indicador de reclamação”, “tipo de assinatura”, “fonte de venda”, “faixa etária”, “SD&W”, “LTV”, “região”, “indicador de participação em ações de fidelização” e “quantidade de produtos agregados comprados”.

4.2
Avaliação do Ajuste Geral

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	35227.478[a]	.220	.310
2	33727.998[a]	.253	.355
3	33430.165[a]	.259	.364
4	33157.816[b]	.264	.372
5	32940.573[c]	.269	.378
6	32711.454[c]	.274	.385
7	32545.559[c]	.277	.390
8	32397.244[c]	.280	.394
9	32271.724[d]	.283	.398
10	32221.754[d]	.284	.399
11	32195.392[d]	.284	.400

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

c. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

d. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

A cada passo, quando uma nova variável é incluída no modelo, a estatística de probabilidade – 2log diminui indicando uma melhora no modelo. Em contrapartida, as medidas pseudo R^2 aumentam à medida que previsores são adicionados. O pseudo R^2 de Nagelkerke no último passo aumentou em 30% o poder de explicação do modelo obtido no passo 1.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	4	1.000
2	139.350	7	.000
3	135.151	7	.000
4	187.903	7	.000
5	129.408	8	.000
6	134.869	8	.000
7	134.234	8	.000
8	120.247	8	.000
9	144.911	8	.000
10	142.347	8	.000
11	135.804	8	.000

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 *chi-quadrado* não significante (Hair, 1998).

No modelo do último passo, todas as medidas de ajuste melhoraram. O valor -2LL diminuiu para 32.195. Os valores R^2 variam de 0,310 para 0,400, indicando melhoria no modelo de 11 variáveis, embora seja um valor distante dos valores R^2 geralmente encontrados em regressão múltipla. A medida Hosmer e Lemeshow 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.

4.3
Precisão da Estimativa

Pode haver problemas na utilização de métodos de regressão logística direcionados a passos quando o objetivo da análise é a precisão da estimativa. Os algoritmos direcionados a passos buscam um subconjunto de variáveis que maximize a probabilidade, mas isto não é o mesmo que maximizar a precisão da estimativa (SPSS, 2003).

As matrizes de classificação, idênticas em natureza às utilizadas na análise discriminante (Hair, 1998), mostram taxas de acerto extremamente altas de casos corretamente classificados para o modelo de 11 variáveis. A taxa de acerto geral é de 75,3%, 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 que cancela é maior que a taxa do grupo que não cancela, 77,4% contra 74,3%.

O modelo inicial que considerava apenas a constante tinha uma taxa geral de acerto de 31,1%. O modelo completo com 11 variáveis aumenta 2,5 vezes a taxa de acerto na previsão.

A partir do passo 9, a melhora no R^2 é pequena e a taxa de acerto geral do modelo não se altera. Isto indica que as variáveis 10 e 11 poderiam ser descartadas do modelo final porque quanto menos variáveis um modelo tiver, menor o tempo de processamento. No entanto, as variáveis “indicador de participação em ações de fidelização” e “quantidade de produtos agregados comprados” foram mantidas neste modelo não para aumentar a precisão, mas para ajudar na definição do perfil dos clientes que cancelam a assinatura do jornal.

Classification Table[a]

			Predicted		
			STATUS		Percentage
	Observed		CANCELAD	ATIVO	Correct
Step 1	STATUS	CANCELAD	8519	2538	77.0
		ATIVO	7140	17352	70.8
	Overall Percentage				72.8
Step 2	STATUS	CANCELAD	8528	2529	77.1
		ATIVO	6633	17859	72.9
	Overall Percentage				74.2
Step 3	STATUS	CANCELAD	8464	2593	76.5
		ATIVO	6477	18015	73.6
	Overall Percentage				74.5
Step 4	STATUS	CANCELAD	8462	2595	76.5
		ATIVO	6407	18085	73.8
	Overall Percentage				74.7
Step 5	STATUS	CANCELAD	8634	2423	78.1
		ATIVO	6590	17902	73.1
	Overall Percentage				74.6
Step 6	STATUS	CANCELAD	8509	2548	77.0
		ATIVO	6371	18121	74.0
	Overall Percentage				74.9
Step 7	STATUS	CANCELAD	8560	2497	77.4
		ATIVO	6438	18054	73.7
	Overall Percentage				74.9
Step 8	STATUS	CANCELAD	8553	2504	77.4
		ATIVO	6347	18145	74.1
	Overall Percentage				75.1
Step 9	STATUS	CANCELAD	8549	2508	77.3
		ATIVO	6286	18206	74.3
	Overall Percentage				75.3
Step 10	STATUS	CANCELAD	8558	2499	77.4
		ATIVO	6275	18217	74.4
	Overall Percentage				75.3
Step 11	STATUS	CANCELAD	8554	2503	77.4
		ATIVO	6291	18201	74.3
	Overall Percentage				75.3

a. The cut value is .690

4.4
Validação do Modelo (*holdout sample*)

Para conseguir a eficiência classificatória do modelo, a amostra foi separada em duas partes: uma utilizada para estimação do modelo, e outra para testar a eficiência da classificação – *holdout sample* (Hair *et al*., 1998). A amostra utilizada para estimação, também chamada de amostra de treinamento, contou com 35.549 assinantes. O processo de escolha foi realizado no software SPSS através da geração de números aleatórios.

A validação do modelo de regressão logística é obtida através da aplicação do modelo na amostra de validação (Hair, 1998). As taxas de acerto na amostra de validação são quase idênticas às taxas de acerto da amostra de treinamento. Isto leva à conclusão de que o modelo de regressão logística possui forte suporte empírico tanto na amostra de validação quanto na de treinamento.

Amostra de Validação – Tabela de Classificação

Classification Table[a]

Observed			Predicted		
			STATUS		Percentage
			CANCELAD	ATIVO	Correct
Step 11	STATUS	CANCELAD	1162	365	76.1
		ATIVO	809	2460	75.3
	Overall Percentage				75.5

a. The cut value is .690

Amostra de Treinamento – Tabela de Classificação

Classification Table[a]

Observed			Predicted		
			STATUS		Percentage
			CANCELAD	ATIVO	Correct
Step 11	STATUS	CANCELAD	8554	2503	77.4
		ATIVO	6291	18201	74.3
	Overall Percentage				75.3

a. The cut value is .690

4.5
Interpretação dos Resultados

O modelo de regressão logística selecionou as 11 variáveis que melhor explicam o cancelamento de uma assinatura de jornal. Gerado o modelo, atribui-se uma probabilidade de cancelamento a cada indivíduo da base em estudo. Todos os indivíduos foram classificados em dois grupos: o grupo que cancela e o grupo que não cancela.

O cliente pode se desligar da empresa motivado por diversos fatores. Alguns desses fatores podem ser descritos por meio de variáveis e utilizados para estabelecer uma relação entre cancelamento e a ocorrência ou não de situações relacionadas aos fatores em questão (Barros, 2002).

O modelo de regressão logística, utilizado neste trabalho, permite identificar o risco de cancelamento de clientes a partir de variáveis transacionais, demográficas e do histórico de eventos armazenados no banco de dados.

Ao desenvolver o modelo de regressão logística sobre a base de dados, foi possível determinar um modelo que contempla as seguintes variáveis: “tempo de permanência”, “forma de pagamento”, “indicador de reclamação”, “tipo de assinatura”, “fonte de venda”, “faixa etária”, “SD&W”, “LTV”, “região”, “indicador de participação em ações de fidelização” e “quantidade de produtos agregados comprados”.

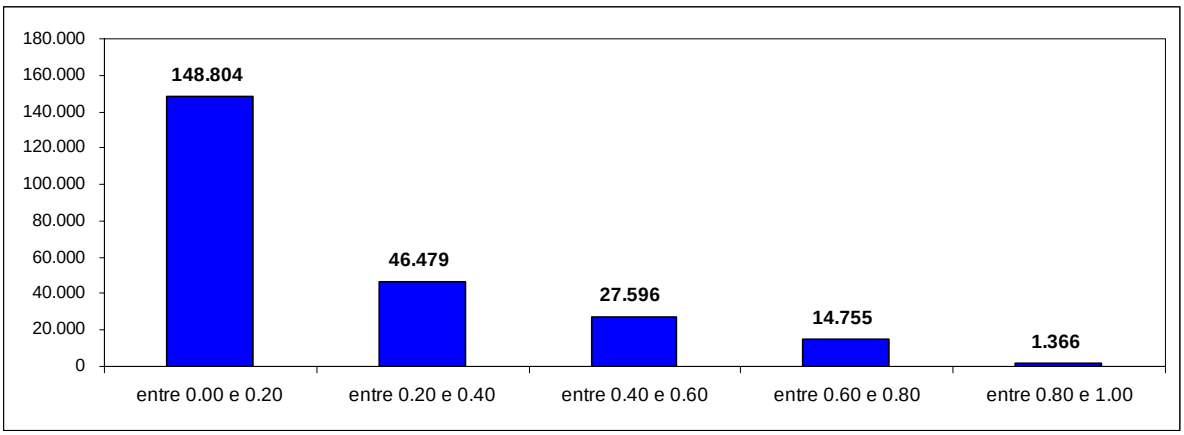
Essas variáveis aplicadas ao modelo de regressão logística geraram um *score* de probabilidade de cancelamento para cada um dos clientes da base de dados conforme o quadro a seguir.

Faixa Probabilidade	Ativo	Cancelado	Total Geral	%canc.
entre 0.00 e 0.20	15,249	1,601	16,850	9%
entre 0.20 e 0.40	4,763	1,656	6,419	18%
entre 0.40 e 0.60	2,828	2,635	5,463	32%
entre 0.60 e 0.80	1,512	3,848	5,360	54%
entre 0.80 e 1.00	140	1,317	1,457	61%

Nota-se que o percentual de assinaturas canceladas cresce à medida que o *score* de probabilidade de cancelamento aumenta. Isto é coerente com a tabela de classificação que acerta em 75,3% dos casos.

Aplicado o *score* de probabilidade na base total de assinantes pode-se concluir que a maior parte dos clientes ativos apresenta baixa probabilidade de cancelar a assinatura do jornal, 62% da base de clientes tem até 20% de chance de cancelar a assinatura do jornal.

GRÁFICO 5: TOTAL DE ASSINATURAS X SCORE DE RISCO DE CANCELAMENTO.

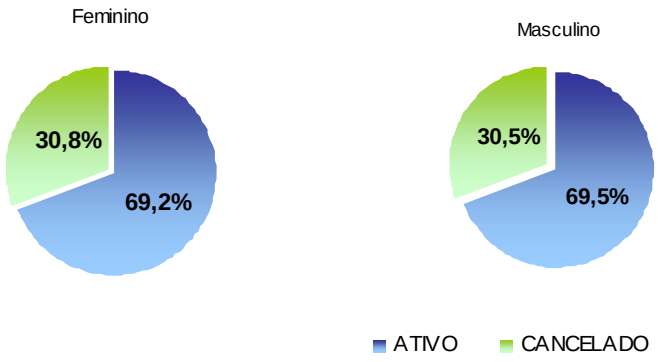


Na análise de regressão logística é possível identificar qual a mudança nas chances de um evento ocorrer dada a presença de um fator (variável categórica) ou alteração em uma variável contínua.

É possível descrever ainda o perfil dos clientes que cancelam, o que pode auxiliar na definição de ações profiláticas que ajudem a reduzir a perda de clientes.

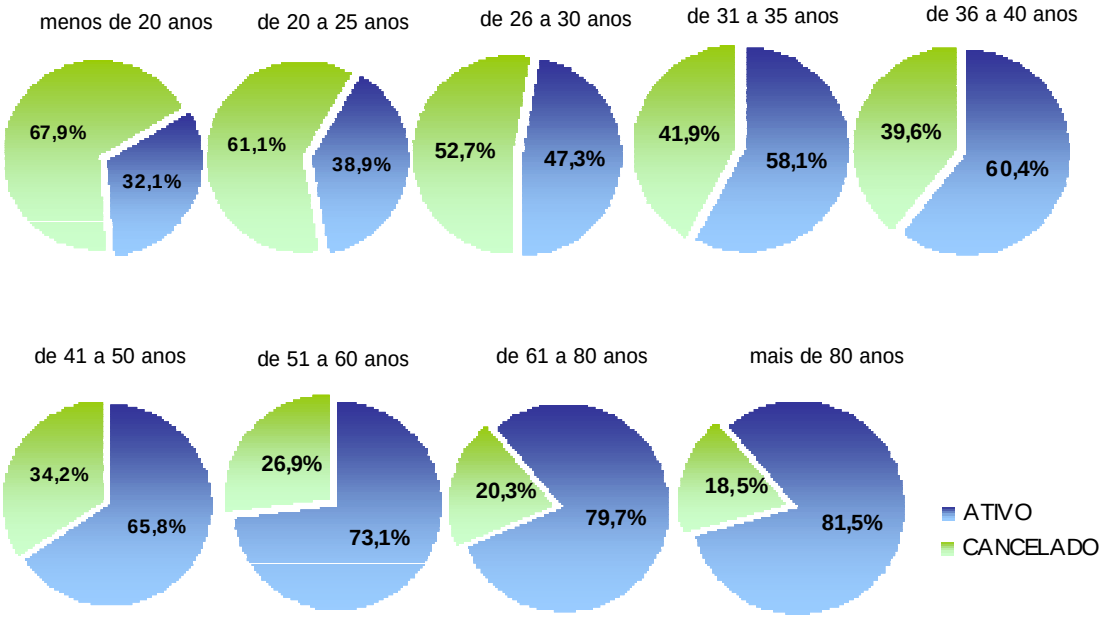
O gênero não discrimina o cancelamento. Não existe diferença no cancelamento entre homens e mulheres.

GRÁFICO 6: SCORE DE RISCO DE CANCELAMENTO x GÊNERO



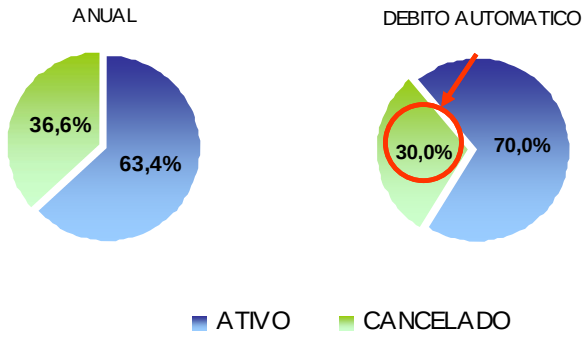
Os mais jovens cancelam mais que os clientes mais velhos. Assinantes com menos de 20 anos têm 2 vezes mais chances de cancelar do que um cliente com mais de 50 anos. O *score* de risco de cancelamento vai se reduzindo com o aumento da faixa etária dos clientes.

GRÁFICO 7: SCORE DE RISCO DE CANCELAMENTO x IDADE



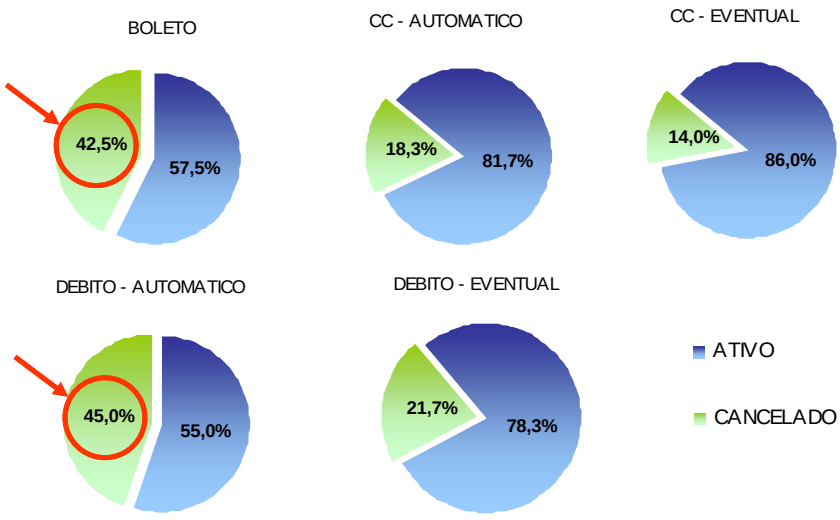
As assinaturas do tipo anual apresentam taxas de cancelamento mais elevadas que as assinaturas da modalidade débito automático. Uma assinatura anual tem 18% mais chance de cancelar do que uma assinatura débito automático.

GRÁFICO 8: SCORE DE RISCO DE CANCELAMENTO x TIPO DE ASSINATURA



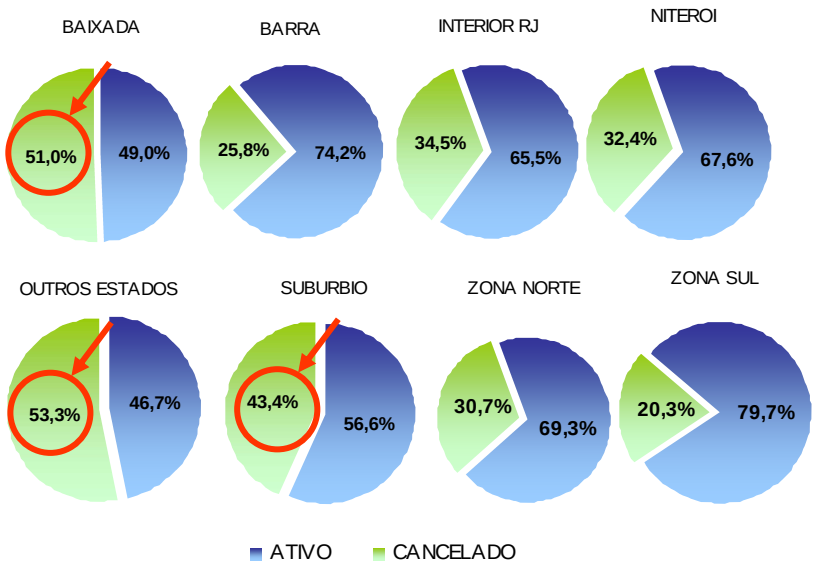
Forma de pagamento é a segunda variável mais importante do modelo e explica o cancelamento. Uma assinatura paga no débito em conta corrente ou boleto bancário tem 1,5 vezes mais chance de cancelar do que uma assinatura paga no cartão de crédito. As assinaturas pagas no cartão de crédito apresentam as menores chances de cancelamento.

GRÁFICO 9: SCORE DE RISCO DE CANCELAMENTO x FORMA DE PAGAMENTO



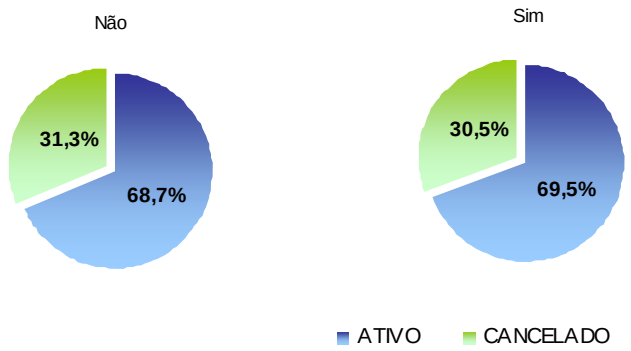
A Zona Sul é a região que menos incrementa o *score* de risco de cancelamento de uma assinatura, enquanto as assinaturas fora do estado aumentam o *score* de risco em 2,6 vezes. As regiões que apresentam o maior risco de cancelamento são Baixada, subúrbio e outros estados.

GRÁFICO 10: SCORE DE RISCO DE CANCELAMENTO x REGIÃO



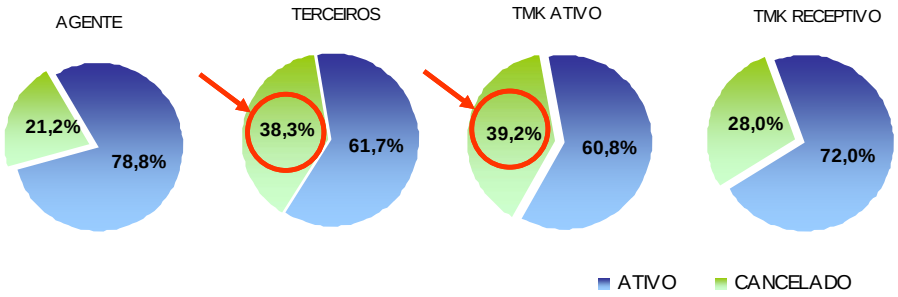
Os clientes que foram impactados por ações de fidelização não apresentam diferença no *score* de cancelamento comparado aos assinantes que nunca foram impactados por estas ações. Talvez seja preciso redefinir as ações de marketing de relacionamento para que contribuam de forma mais efetiva para a fidelização de clientes.

GRÁFICO 11: SCORE DE RISCO DE CANCELAMENTO x PARTICIPAÇÃO EM AÇÕES DE FIDELIZAÇÀO



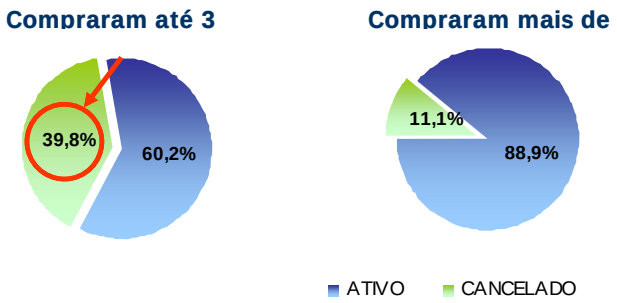
Dos clientes classificados no grupo dos que cancelam, 55% adquiriram sua assinatura através do telemarketing ativo, no grupo dos que não cancelam 38% fizeram a assinatura pelo canal telemarketing ativo.

GRÁFICO 12: SCORE DE RISCO DE CANCELAMENTO x CANAL DE VENDA



Assinantes que compram menos de 3 produtos agregados como livros, CDs, DVDs e guias apresentam um *score* de risco de cancelamento 3,6 vezes maior do que o de clientes que compram estes produtos da empresa. Dos clientes que compram mais de 3 produtos agregados, 11% cancelam a sua assinatura, enquanto a taxa de cancelamento da base é de 31%. Além disso, quanto mais produtos um cliente compra, menor o *score* de cancelamento.

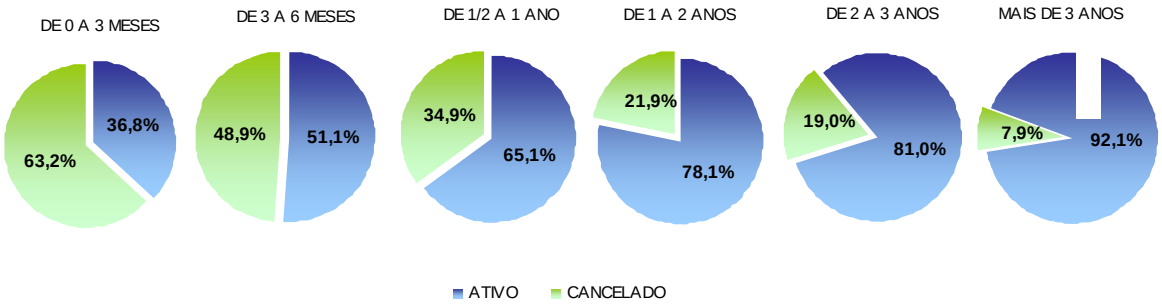
GRÁFICO 13: SCORE DE RISCO DE CANCELAMENTO x QTDE DE PRODUTOS AGREGADOS COMPRADOS



Apesar disso, clientes que fizeram um anúncio de publicidade apresentam taxas de cancelamento similares aos que não anunciaram.

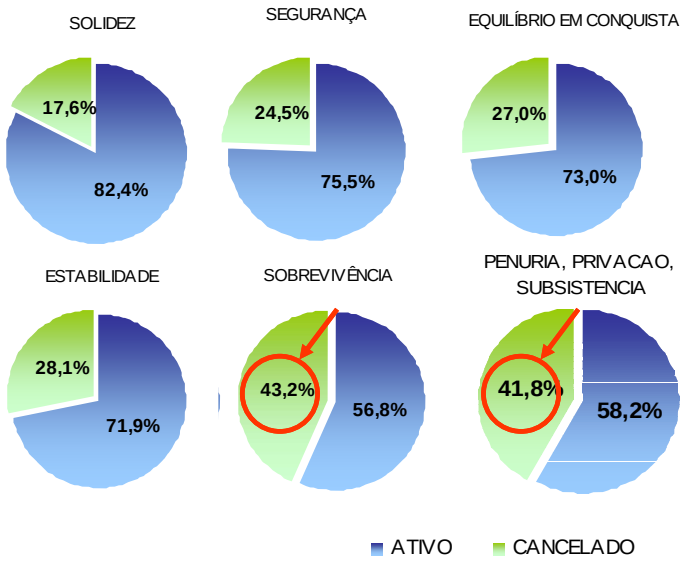
O risco de cancelamento é decrescente à medida que o tempo de permanência do cliente na carteira aumenta. Dos clientes que estão na carteira no máximo há 3 meses, 63% cancelam a sua assinatura. A taxa de cancelamento entre os clientes com mais de 3 anos na base é 8%.

GRÁFICO 14: SCORE DE RISCO DE CANCELAMENTO x TEMPO DE PERMANÊNCIA



Quanto menor o nível sócio demográfico do indivíduo, maior o *score* de risco de cancelamento. Os menores níveis de risco de desligamento estão entre os clientes dos segmentos SD&W "solidez", "segurança", "equilíbrio em conquista" e "estabilidade". Os assinantes dos segmentos sócio-demográficos "sobrevivência", "privação", "penúria" e "subsitência" têm risco de cancelamento 1,8 vez maior do que o risco dos demais segmentos.

GRÁFICO 15: SCORE DE RISCO DE CANCELAMENTO x SD&W



Os clientes que não fizeram reclamações apresentam taxas de cancelamento maiores que os clientes que entraram em contato com a empresa para reclamar. Clientes que não reclamam apresentam um risco de cancelamento 40% maior do que os clientes que contataram a empresa para apresentar a sua reclamação.

GRÁFICO 16: SCORE DE RISCO DE CANCELAMENTO x INDICADOR DE RECLAMAÇÃO