3
Metodologia

O principal objetivo deste trabalho é estabelecer o padrão de comportamento dos clientes que abandonam uma empresa de telefonia fixa, por meio da identificação de suas características comuns, seus hábitos de consumo, os produtos que possuem, dentre outras variáveis que podem marcar o perfil dos clientes perdidos.

Com este foco, será definido um conjunto de variáveis de observação, bem como uma janela de tempo associada a este conjunto, o qual servirá de base de entrada para a geração de conhecimento sobre o comportamento dos clientes.

Para identificar a diferença entre o grupo de clientes que cancelam, do grupo de clientes que se mantêm ativos será utilizado o modelo de Regressão Logística Binária, que buscará conhecer o perfil dos clientes *churners*, possibilitando o desenvolvimento de ações de retenção que aumentem o *lifetime* dos mesmos.

Assim teremos as seguintes etapas a serem desenvolvidas:

1. Definição do universo e amostra;
2. Coleta dos dados e tratamento dos mesmos;
3. Identificação das variáveis a serem utilizadas;
4. Aplicação do Modelo de Regressão Logística para obtenção e análise dos resultados.

3.1.
Tipo de pesquisa

Segundo Yin (2005), a escolha do método de pesquisa a ser adotado depende de três fatores: o tipo de pesquisa que se faz, o controle sobre o comportamento do fenômeno pesquisado pelo observador e o foco da pesquisa no passado ou na atualidade. O método de pesquisa em si é dividido pelo autor em estudo de caso, pesquisa histórica, análise de arquivos, levantamento e experimento. Quanto ao propósito, divide-os em exploratórios, descritivos e explicativos.

Vergara (2007) apresenta uma taxionomia mais extensa, dividindo as pesquisas quanto aos fins e quanto aos meios. Quanto aos fins, as pesquisas podem ser exploratórias, descritivas, explicativas, metodológicas, aplicadas e intervencionistas. Quanto aos meios, podem ser pesquisa de campo, de laboratório, telematizadas, documentais, bibliográficas, experimentais, *ex post facto*, participantes, pesquisa-ação e estudo de caso.

A pesquisa descritiva apresenta características de determinada população ou determinado fenômeno. Pode também estabelecer correlações entre variáveis e definir sua natureza. Entretanto, não tem compromisso de explicar os fenômenos que descreve, embora sirva de base para tal explicação (VERGARA, 2007).

Assim, a presente pesquisa pode ser classificada, quanto aos fins, como uma pesquisa de caráter descritivo, pois teve como propósito obter informações sobre determinada população, analisar as variáveis e identificar características que representam o perfil de clientes mais propensos a se tornarem *chuners*. Com este objetivo, foi realizado um estudo empírico, fazendo a aplicação de um modelo estatístico, análise e interpretação dos resultados obtidos.

Quanto aos meios, este trabalho pode ser classificado como uma pesquisa de campo telematizada, documental, bibliográfica e experimental, devido à pesquisa que foi realizada em publicações e sites da internet sobre o assunto e empresas de telecomunicações, como também pela aplicação do Modelo de Regressão Logística sobre a base de dados de clientes de telefonia fixa.

> "Pesquisa experimental é investigação empírica na qual o pesquisador manipula e controla variáveis independentes e observa as variações que tal manipulação e controle produzem em variáveis dependentes". (VERGARA, 2007, p. 48).

Desta forma, esta é uma pesquisa empírica, na media em que objetiva determinar a influência de variáveis demográficas, comportamentais, dos dados históricos e de relação dos clientes com a empresa, sobre o risco de cancelamento, buscando definir, a partir de uma base de dados já existente, o perfil do cliente que abandona uma empresa de telefonia fixa.

O trabalho empírico pode ser subdividido em pesquisa qualitativa ou quantitativa, sendo que as duas perspectivas visam testar uma teoria ou hipótese (CRESWELL *apud* KARAM, 2006).

Dentro deste escopo o presente exercício, para cumprimento de seus objetivos, pode ser considerado como quantitativo pela aplicação de um modelo estatístico – regressão logística binária – sobre a base de dados, para prever o risco de cancelamento de clientes.

3.2. Universo e amostra

Para este estudo foi utilizada uma base de dados real de uma grande empresa de telefonia.

O universo da pesquisa foi constituído a partir da seguinte critério de seleção, disponível para o levantamento dos dados:

- Terminais de clientes Pessoa Física;
- Terminais ativos em agosto de 2008;
- *Churn* voluntário dos meses de junho, julho e agosto de 2008.

Aplicado o critério descrito anteriormente, a base de dados original passou a ter 9.075.436 registros, onde 234.144 eram de retiradas voluntárias e 8.841.292 de terminais ativos.

Para construção dos modelos, foram utilizadas duas amostras de proporções diferentes, após a aplicação da fundamentação de *Oversampling*, descrito mais à frente. A primeira com 68.360 terminais ativos e a segunda com 29.566 terminais retirados, extraídos através de seleção baseada em aleatoriedade simples.

No estudo apresentado nesse trabalho, cada terminal será entendido como um caso para o modelo; o que não traz prejuízo à análise já que a avaliação será feita para pessoa física, que na sua maioria possui apenas um terminal.

3.3. Coleta dos dados

A seguir, será apresentado como os dados foram coletados para execução deste trabalho:

1. pesquisa telematizada através de sites da internet, como TELECO e ANATEL para levantar informações sobre o setor de telecomunicações, principalmente no que diz respeito ao cenário atual, mercado e base de clientes;
2. pesquisa bibliográfica em livros, teses, dissertações, *papers*, revistas especializadas e periódicos para realizar o embasamento teórico aos assuntos abordados e necessários para realização do objetivo aqui proposto – Comportamento do Consumidor, Segmentação de Mercado, Relacionamento com Clientes, Retenção de Clientes e *Churn* na telefonia fixa. Como também, referencial teórico sobre métodos estatísticos para aplicação e modelagem de dados;

3. pesquisa documental, através da base de dados em estudo, constituída, como já citado, por clientes ativos e retirados (*churners*), e de uma diversidade de variáveis independentes, especificadas mais à frente.

3.4. Dados disponíveis e suas limitações

Uma das grandes limitações do estudo diz respeito à base de dados, principalmente com relação às variáveis:

1. Há uma limitação técnica com relação aos campos disponíveis para análise, devido à impossibilidade de utilizar algumas variáveis que talvez fossem relevantes para a análise, como por exemplo nível de satisfação do cliente (não disponível no banco de dados original);
2. Ainda são poucas as referências de trabalhos publicados sobre previsão do risco de cancelamento de clientes, utilizando regressão logística, que servissem de orientação para seleção das variáveis independentes.

Estas restrições com relação às informações disponíveis, com certeza, impactaram nos resultados obtidos.

Outro ponto importante é que os resultados deste trabalho são específicos para a empresa, portanto, este modelo não pode ser generalizado para outros setores.

3.5. Tratamento de dados

3.5.1. *Oversampling*

Segundo vários autores (BERRY *et al*, 2000; MOZER *et al*, 2000; YAN *et al*, 2001; ARCHAUX *et al*, 2004; AU *et al*, 2003), *apud* Ferreira (2005), para tratamento de base de dados, um dos problemas é referente às variáveis categóricas altamente desequilibradas em termos da proporção de cada classe existente.

Este é o caso da base original utilizada neste estudo. A mesma é constituída de 97,42% de terminais ativos e somente 2,58% de *churners*. Nestas situações, no momento da construção do modelo, devido à distribuição altamente desequilibrada entre as classes (ativos x *churners*), o mesmo

terminará enxergando somente uma delas, sendo incapaz de distinguir a classe de menor número de registros (FERREIRA, 2005). Segundo Ferreira (2005), este fato ocorre "porque o modelo reconhece que, se sua resposta for sempre dizer que todas as observações pertencem à classe com maior número de registros, ele acertará" 97,42% dos padrões.

Desta forma, o mesmo sugere, para evitar esse problema e facilitar a distinção entre as classes, a realização do procedimento conhecido como *oversampling*. Através deste processo, seleciona-se aleatoriamente um maior número de registros pertencentes à classe rara e um menor número de casos da classe comum, ajustando, desta forma, a proporção entre as mesmas.

No caso em questão, tratam-se de variáveis binárias, frequências de 20% à 30%, em geral geram bons resultados (BERRY *apud* FERREIRA, 2005). Seguindo esta orientação a base foi gerada da seguinte forma:

- **30,2%, 29.566 registros de *churners*; e**
- **69,8%, 68.360 registros de terminais ativos.**

3.5.2. Descrição das variáveis

As variáveis foram selecionadas seguindo os seguintes critérios:

- Com base na literatura estudada, apontando variáveis importantes para o processo de segmentação como geográficas, demográficas, psicográficas e comportamentais (KOTLER, 2000) – parte 2.4;
- Conforme características exemplificadas por Vavra (1993), utilizadas por empresas para o desenvolvimento de atividades de pós-marketing – parte 2.3.3; e
- Através da disponibilidade de informações da empresa em estudo.

Foram, então consideradas, inicialmente, as seguintes variáveis:

- Variável Alvo ou Dependente:

No.	Descrição	Definição	Nome da variável
01	*Churn*	Informação que diferencia o terminal ativo do *chuner*.	FLAG_CHURN

Tabela 10 – Descrição da variável alvo ou dependente.

Onde:

"0" – terminal ativo;

"1" – *churner.*

- Variáveis Independentes:

No.	Descrição	Definição e Níveis	Nome da Variável
01	CEP	Identificação do endereço postal (rua) do terminal fixo.	CEP
02	Estado	Estado onde o terminal está / era instalado: RJ, MG, ES, BA, SE, AL, PE, CE, RN, PB, PI, MA, PA, AP, AM, RR.	UF

Tabela 11 – Descrição das variáveis independentes. (continua ...)

(continuação ...)

No.	Descrição	Definição e Níveis	Nome da Variável
03	Sexo	Identificação do gênero do titular do Terminal: M – masculino e F – feminino.	SEXO
04	Idade	Identificação da idade do titular do terminal.	IDADE
05	Quantidade terminais fixos residenciais	Quantidade de terminais fixos residenciais que a residência do titular possui (da mesma operadora).	QTD_FIXO_RES
06	Quantidade terminais fixos não residenciais	Quantidade de terminais fixos não residenciais que a residência do titular possui (da mesma operadora).	QTD_FIXO_NRES
07	Tempo primeiro terminal fixo	Tempo de planta do primeiro terminal fixo instalado na residência do titular (da mesma operadora) – em dias.	TEMPO_PRIM_FIXO
08	Tempo último terminal fixo	Tempo de planta do último terminal fixo instalado na residência do titular (da mesma operadora) – em dias.	TEMPO_ULT_FIXO
09	Quantidade de terminais móveis pré-pago	Quantidade de terminais móveis pré-pago que a residência do titular do terminal fixo possui da mesma operadora.	QTD_PRE_MOVEL
10	Quantidade de terminais móveis pós-pago	Quantidade de terminais móveis pós-pago que a residência do titular do terminal fixo possui da mesma operadora.	QTD_PÓS_MOVEL

11	Tempo do primeiro móvel	Tempo de planta do primeiro terminal móvel que a residência do titular do terminal fixo instalou da mesma operadora – em dias.	TEMPO_PRIM_MOVEL
12	Tempo do último móvel	Tempo de planta do último terminal móvel que a residência do titular do terminal fixo instalou da mesma operadora – em dias.	TEMPO_ULT_MOVEL
13	Quantidade de banda larga residencial	Quantidade de banda larga residencial que a residência do titular do terminal fixo possui da mesma operadora.	QTD_B_LARGA_RES
14	Quantidade de banda larga não residencial	Quantidade de banda larga não residencial que a residência do titular do terminal fixo possui da mesma operadora.	QTD_B_LARGA_NRES

Tabela 11 – Descrição das variáveis independentes. (continua ...)

(continuação ...)

No.	Descrição	Definição e Níveis	Nome da Variável
15	Tempo da primeira banda larga	Tempo de planta da primeira banda larga que a residência do titular do terminal fixo instalou da mesma operadora – em dias.	TEMPO_PRIM_B_LARGA
16	Tempo da última banda larga	Tempo de planta da última banda larga que a residência do titular do terminal fixo instalou da mesma operadora – em dias.	TEMPO_ULT_B_LARGA
17	Tempo de relacionamento com a empresa	Tempo de relacionamento do titular com a empresa, independente do produto ou serviço prestado – em dias.	TEMPO_RELAC_EMP
18	Quantidade terminais fixos	Quantidade de terminais fixos que a residência do titular possui (da mesma operadora) – residencial ou não residencial.	QTD_FIXO
19	Quantidade de banda larga	Quantidade de banda larga que a residência do titular do terminal fixo possui da mesma operadora – residencial ou não residencial.	QTD_B_LARGA
20	Grupo de produtos	Conjunto de produtos que a residência do titular do terminal fixo possui: 1 – fixo + móvel + banda larga; 2 – fixo + banda larga; 3 – fixo + móvel; e 4 – fixo.	GRUPO
21	Disponibilidade de banda larga na região	Indicativo de disponibilidade ou não de banda larga da operadora na região da residência do titular do terminal fixo: 0 – não possui disponibilidade; e 1 – possui disponibilidade.	POSSE_DISP_B_LARGA

22	Concorrência 1	Indicativo de atuação do principal concorrente de telefonia fixa para clientes B2C na região da residência do titular do terminal fixo: 0 – sem forte atuação (ou sem atuação) deste concorrente; e 1 – forte atuação deste concorrente.	CONCORR_N
23	Concorrência 2	Indicativo de atuação do segundo principal concorrente de telefonia fixa para clientes B2C na região da residência do titular do terminal fixo: 0 – sem forte atuação (ou sem atuação) deste concorrente; e 1 – forte atuação deste concorrente.	CONCORR_G
24	Quantidade de meses de utilização de internet	Quantidade de meses de utilização de internet nos últimos 6 meses pela residência do titular do terminal fixo.	QTD_MESES_INTERNET

Tabela 11 – Descrição das variáveis independentes. (continua ...)
(continuação ...)

No.	Descrição	Definição e Níveis	Nome da Variável
25	Quantidade de meses de tráfego de longa distância	Quantidade de meses de tráfego de longa distância nos últimos 6 meses da residência do titular do terminal fixo.	QTD_MESES_TRF_LD
26	Terminal fixo já retido	Indicativo que já havia tido a intenção de cancelamento do terminal e que o mesmo aceitou uma oferta diferenciada para se manter ativo nos últimos 6 meses: 0 – nunca recebeu oferta para ser retido; 1 – já recebeu oferta de retenção.	FLAG_RETIDO
27	Quantidade de solicitações realizadas no call center	Volume de solicitações realizadas no call center da empresa para o terminal fixo nos últimos 6 meses.	QTD_INSAPOIO_sum
28	Quantidade de meses de assinatura	Quantidade de meses de assinatura nos últimos 6 meses da residência do titular do terminal fixo.	QTD_MESES_ASSINAT
29	Quantidade de meses de receita total do(s) terminal(is) fixo(s)	Quantidade de meses de receita total de terminais fixos nos últimos 6 meses da residência do titular.	QTD_MESES_RECTTFIX
30	Quantidade de meses de receita de ligação para celular	Quantidade de meses de receita de ligação para celular gerada na residência por terminais fixos nos últimos 6 meses da residência do titular.	QTD_MESES_RECFATVC1

31	Quantidade de meses de utilização de ligação para celular	Quantidade de meses de utilização de ligação para celular gerada na residência por terminais fixos nos últimos 6 meses da residência do titular.	QTD_MESES_MINFATVC 1
32	Classificação de valor do cliente	Indicador de classificação do segmento do cliente por valor para a empresa: 7 – Diamante; 8 – Ouro; 9 – Prata; 10 – Bronze; 12 – Novos; 99 – Sem classificação.	NU_SEGMTO_ATENDTO
33	Tráfego internet médio	Volume de tráfego médio de utilização de internet realizado pela residência do titular, através da telefonia fixa da operadora – em minutos.	TRF_INTERNET_MEDIO

Tabela 11 – Descrição das variáveis independentes. (continua ...)

(continuação ...)

No.	Descrição	Definição e Níveis	Nome da Variável
34	Tráfego médio de consumo de ligação longa distância	Volume de tráfego médio de consumo de ligação de longa distância realizado pela residência do titular nos últimos 6 meses, através da telefonia fixa da operadora – em minutos.	TRF_LD_TOTAL_MEDIO
35	Receita média de utilização de longa distância	Receita média de consumo de ligação de longa distância realizado pela residência do titular nos últimos 6 meses, através da telefonia fixa da operadora.	REC_LD_TOTAL_MEDIA
36	Receita média de utilização serviços inteligentes	Receita média de consumo de serviços inteligentes (caixa postal, chamada em espera, ...) realizado pela residência do titular nos últimos 6 meses, através da telefonia fixa da operadora.	REC_SI_MEDIA
37	Receita média de banda larga	Receita média de banda larga realizada pela residência do titular nos últimos 6 meses, através de vínculo com a telefonia fixa da operadora.	REC_B_LARGAI_MEDIA
38	Receita média de utilização de serviços de consulta ao 102	Receita média de utilização de serviços de consulta ao 102 realizada pela residência do titular nos últimos 6 meses, através da telefonia fixa da operadora.	REC_102_MEDIA
39	Receita média de utilização do serviço IDTC	Receita média de utilização de serviço do Identificador de Chamadas realizada pela residência do titular nos últimos 6 meses, através da telefonia fixa da operadora.	REC_IDTC_MEDIA

40	Receita de assinatura média faturada	Receita de assinatura média faturada para residência do titular nos últimos 6 meses, pela utilização de telefonia fixa da operadora.	ASSINAT_FAT_MEDIA
41	Receita total média de telefonia fixa	Receita total média gerada pela utilização de telefonia fixa da operadora da residência do titular nos últimos 6 meses.	RECTOTFIX_MEDIA
42	Receita média de utilização de ligação para celular (fixo para móvel)	Receita média de utilização de ligação para celular gerada pela telefonia fixa da operadora da residência do titular nos últimos 6 meses.	RECFATVC1_MEDIA

Tabela 11 – Descrição das variáveis independentes. (continua ...)

(continuação ...)

No.	Descrição	Definição e Níveis	Nome da Variável
43	Tráfego médio de utilização de ligação para celular (fixo para móvel)	Tráfego médio de utilização de ligação para celular gerado pela telefonia fixa da operadora da residência do titular nos últimos 6 meses – em minutos.	TRF_FAT_VC1_MEDIO

Tabela 11 – Descrição das variáveis independentes.

Como pode ser visto na tabela 11, a estrutura inicial foi composta por uma variedade de 43 características sendo: geográficas (como UF e CEP); demográficas (como sexo e idade); comportamentais (tipo de serviço utilizado como longa distância e ligação para celular); utilização da rede (como tráfegos gerados); relacionais (como tempo de base e segmento); de receita (como receita de assinatura e receita total); como também de mercado (como concorrente 1 e concorrente 2).

3.5.3. Validação, exploração, limpeza de dados e transformações

Após a seleção da amostra e variáveis, que poderiam explicar o *churn* na telefonia fixa, foi iniciada a etapa de validação, exploração e limpeza de dados.

Para isto foi realizada uma análise descritiva das variáveis numéricas para avaliar a consistência da base de dados e buscar uma maior compreensão dos dados a serem tratados. Através da tabela 12, podem ser observados os resultados das estatísticas mínimo, máximo e desvio padrão destas variáveis.

	N	Mínimo	Máximo	Média	Desvio Padrão
IDADE	97.926	18	109	50	17
QTD_FIXO_RES	97.926	1	19	1	1
QTD_FIXO_NRES	97.926	-	4	0	0
TEMPO_PRIM._FIXO	97.926	121	17.543	2.847	2.523
TEMPO_ULT_FIXO	97.926	-	17.539	2.377	2.274
QTD_PRE_MOVEL	97.926	-	28	0	1
QTD_POS_MOVEL	97.926	-	10	0	0
TEMPO_PRIM_MOVEL	97.926	-	2.338	189	462
TEMPO_ULT_MOVEL	97.926	-	2.277	104	308
QTD_B_LARGA_RES	97.926	-	6	0	0
QTD_B_LARGA_NRES	97.926	-	3	0	0
TEMPO_PRIM_B_LARGA	97.926	-	2.643	89	290
TEMPO_ULT_B_LARGA	97.926	-	2.545	86	284
TEMPO_RELAC_EMP	97.926	121	17.543	2.867	2.508
QTD_FIXO	97.926	1	19	1	1
QTD_B_LARGA	97.926	-	6	0	0
QTD_MESES_INTERNET	97.926	-	6	1	2
QTD_MESES_TRF_LD	97.926	-	6	2	3
QTD_INSAPOIO_sum	97.926	-	141	0	1
QTD_MESES_ASSINAT	97.926	-	6	5	2
QTD_MESES_RECTTFIX	97.926	-	6	5	2
QTD_MESES_RECFATVC1	97.926	-	6	3	3
QTD_MESES_MINFATVC1	97.926	-	6	3	3
TRF_INTERNET_MEDIO	97.926	-	999	20	107
TRF_LD_TOTAL_MEDIO	97.926	-	998	24	58
REC_LD_TOTAL_MEDIA	97.926	-	896	11	30
REC_SI_MEDIA	97.926	-	150	2	5
REC_BLARGA_MEDIA	97.926	-	680	6	21
REC_102_MEDIA	97.926	-	108	0	1
REC_IDTC_MEDIA	97.926	-	132	2	5
ASSINAT_FAT_MEDIA	97.926	-	834	51	43
RECTOTFIX_MEDIA	97.926	-	999	92	85
RECFATVC1_MEDIA	97.926	-	989	19	41
TRF_FAT_VC1_MEDIO	97.926	-	883	14	30
Valid N (listwise)	97.926				

Tabela 12 – Estatísticas mínimo, máximo e desvio padrão das variáveis numéricas.

Através da análise exploratória da base não foram identificados casos de valores ausentes ou absurdos. Entretanto, há indicativos de *outliers* (na tabela 12), como por exemplo a variável REC_LD_TOTAL_MEDIA (receita total de longa distância média), apresenta valores máximos muito maiores do que sua média, como também desvio padrão alto.

Entretanto, quando da avaliação e conferência destes valores na base original da companhia, os valores realmente existem. Optou-se, então, por não eliminá-los, pois os mesmos são provocados, na maioria dos casos, devido ao fato da ausência do serviço para muitos terminais (REC_LD_TOTAL_MEDIA, REC_SI_MEDIA, REC_BLARGA_MEDIA, ...).

As variáveis referentes às quantidades de terminais (fixo RES, fixo NRES, móvel pré-pago, móvel pós-pago, banda larga RES, banda larga NRES, fixo e banda larga), conforme quadros de 09 a 16 do anexo, tiveram seus valores agrupados, de acordo com a estruturação dos dados e conhecimento prático do negócio. A mesma alteração foi realizada para quantidade de *insapoios* (solicitações), tornando esta variável dicotômica. Desta forma, as mesmas ficaram com os seguintes níveis:

- Fixo RES: 0 - somente um terminal RES; 1 – dois ou mais terminais RES;
- Fixo NRES: 0 – nenhum terminal NRES; 1 – um ou mais terminais NRES;
- Móvel pré-pago: 0 – nenhum móvel pré; 1 – um móvel pré; 2 – dois ou mais móveis pré;
- Móvel pós-pago: 0 – nenhum móvel pós; 1 – um ou mais móveis pós;
- Banda larga RES: 0 – nenhuma banda larga RES; 1 – uma ou mais banda larga RES;
- Banda larga NRES: 0 – nenhuma banda larga NRES; 1 – uma ou mais banda larga NRES;
- Fixo: 0 – somente um terminal; 1 – dois ou mais terminais;
- Banda larga: 0 – nenhuma banda larga; 1 – uma ou mais banda larga;
- *Insapoio* (solicitação): 0 – sem *insapoio* (solicitação) e 1 – um ou mais *insapoios* (solicitação).

Foi realizada ainda a normalização das demais variáveis numéricas (por exemplo TEMPO_PRIM_FIXO e TEMPO_ULT_FIXO), através da criação do *zscore*, calculado pela fórmula:

- Normalização pelo desvio padrão: $y = \frac{x - \mu}{\sigma}$

3.5.4. Seleção das variáveis

Após o tratamento dos dados, buscou-se identificar as variáveis que possuíssem maior poder explicativo para o *churn* e eliminar possíveis multicolinearidades entre as mesmas.

Quando da avaliação da matriz de correlação (quadro 22 e 23 do anexo) das variáveis numéricas após a normalização das mesmas foram retiradas da aplicação do modelo as abaixo descritas, devido à alta e significativa correlação. Para esta exclusão foram selecionadas aquelas variáveis que apresentaram relação maior que 0,6 e significância menor que 0,01.

- Tempo de base do primeiro fixo;
- Tempo de base do último móvel;
- Receita média de banda larga;
- Tempo de base da última banda larga;
- Tráfego médio de longa distância;
- Receita média de IDTC;
- Receita média 102;
- Assinatura média;
- Tráfego médio de VC1.

Além das citadas acima também foi retirada a variável gênero, por apresentar resultados bem similares com relação ao cancelamento ou não.

3.5.5. Base selecionada

A base de dados final para a aplicação do modelo foi constituída, então, de 32 variáveis independentes, sendo 18 transformadas e 97.926 registros. Esta base foi dividida em dois conjuntos de dados, utilizados para treinamento do modelo e outra para validação da assertividade, nas mesmas proporções de 50%.

3.5.6. O modelo: regressão logística binária

O Modelo de Regressão Logística é uma técnica de análise multivariada de dependência cuja variável dependente é dicotômica, ou seja, uma variável nominal ou não métrica; permite ainda, estimar a probabilidade de ocorrência de um evento e identificar as variáveis independentes que contribuem efetivamente para a sua predição.

No caso em estudo a variável dependente Y_i está relacionada à situação do terminal *i*, sendo atribuída a esse terminal um Y_i igual a 0 (zero) caso esteja ativo e 1 (um) caso seja *churner*.

Através da regressão logística, pode-se estimar diretamente a probabilidade de ocorrência de um evento. O evento de interesse aqui é a previsão de *churn* de clientes de telefonia fixa, através das variáveis independentes listadas.

Segundo Hair *et al* (1995), os dois métodos estatísticos multivariados de dependência – Regressão Logística e Análise Discriminante – podem ser utilizados para separar dois grupos ou alocar um novo elemento em um desses grupos, pois relacionam um conjunto de variáveis independentes com uma variável dependente categórica.

Entretanto, a Regressão Logística é mais comumente utilizada devido aos seguintes motivos:

a. A Regressão Logística, diferentemente da discriminante, não depende da exigência de normalidade das variáveis independentes e da igualdade de matrizes de covariância, ou seja, pode incorporar efeitos não lineares (HAIR *et al*, 1995);

b. A Regressão Logística pode utilizar variáveis categóricas independentes, enquanto que na Análise Discriminante o uso de variáveis *dummy* cria problemas com as igualdades variância/covariância (HAIR *et al*, 1995);

c. A expressão discriminante fornece um *score* que possui pouca interpretação intuitiva (KARAM, 2006);

d. Para Regressão Logística existe a possibilidade de interpretação direta dos coeficientes como medidas de associação, e a interpretação destes coeficientes pode ser entendida para qualquer problema prático (PAULA, 1999 *apud* KARAM, 2006);

e. A Regressão Logística apresenta uma abordagem probabilística, possibilitando estimar a chance de ocorrer um certo evento a partir de uma série de variáveis independentes ou explanatórias.

Assim foi feita a opção pela Regressão Logística, onde a probabilidade de ocorrência de um evento pode ser estimada diretamente. No caso da variável dependente Y assumir apenas dois possíveis estados (1 ou 0, "*sucesso"* ou "*fracasso"*) e haver um conjunto de i variáveis independentes x_1, x_2 , ... , x_i, o Modelo de Regressão Logística pode ser escrito da seguinte forma:

$$P(Y = 1) = \frac{1}{1 + e^{-g(x)}}$$

Onde: $g(x) = \beta_0 + \beta_1 x_1 + ... + \beta_i x_i$

Considerando certa combinação de coeficientes $\beta_0, \beta_1, ..., \beta_i$ e variando os valores de x, observa-se que a curva logística tem comportamento probabilístico no formato da letra *S* (figura 8), o que é característica da regressão logística. Esse formato dá à Regressão Logística um alto grau de generalidade, aliada a aspectos muito desejáveis:

a) Quando $g(x) \rightarrow + \infty$, então $P(Y = 1) \rightarrow 1$;

b) Quando $g(x) \rightarrow - \infty$, então $P(Y = 1) \rightarrow 0$.

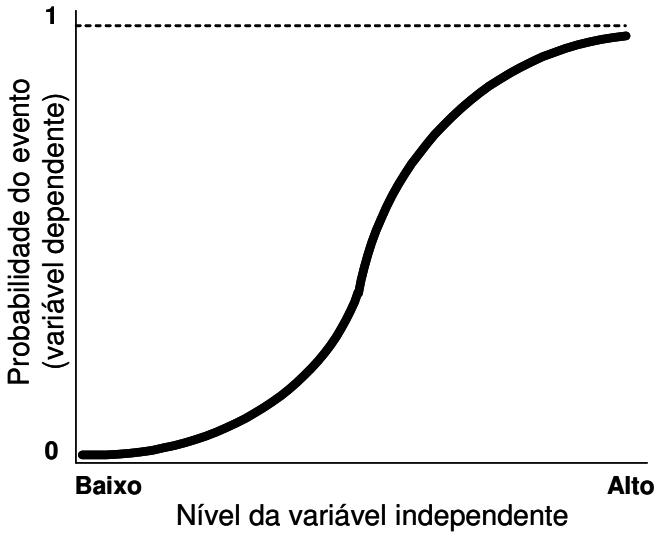

Figura 8 – Formato da Curva Logística.
Fonte: HAIR *et al*, 1995**.**

Desta forma, como se pode estimar diretamente a probabilidade de ocorrência de um evento, pode-se estimar a probabilidade de não ocorrência por diferença: $P\ (Y = 0) = 1 - P\ (Y = 1)$. Ao utilizar a regressão logística a principal suposição é que o logaritmo da razão entre as probabilidades de ocorrência e não ocorrência do evento é linear, onde a expressão de probabilidade é denominada *Odds*[2]:

$$Odds_i = \frac{P\ (Y = 1)}{P\ (Y = 0)} = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_i x_i}$$

A variável *Odds* resolve os problemas de estimativa de probabilidades entre 0 e 1, abaixo deste limite a solução é o logaritmo de *odds*, denominado de *logit* (HAIR *et al,* 1995)*:*

$$Logit_i = ln\left[\frac{P\ (Y = 1)}{P\ (Y = 0)}\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_i x_i$$

Por isto, ao interpretar os coeficientes da regressão logística, opta-se pela interpretação de e^{β} e não diretamente de β. Para utilizar o Modelo de Regressão Logística para discriminação de dois grupos, a regra de classificação é a seguinte:

- Se $P\ (Y = 1) > 0{,}5$ então classifica-se $Y = 1$;
- Em caso contrário, classifica-se $Y = 0$.

Como já citado acima, o foco deste estudo é estimar, a partir de uma série de variáveis, a probabilidade de ocorrência do evento *churn*. Então, partindo do valor dicotômico da variável dependente (ativo – "0" e *churner* – "1"), o modelo resultará na estimativa da probabilidade do evento ocorrer ou não. Para a probabilidade prevista maior que 0,5, então a previsão será sim, caso contrário será não. Desta forma, o processo que calcula o coeficiente logístico compara a probabilidade de um evento ocorrer com a probabilidade de ele não ocorrer. Assim:

[2] *Odds* - razão entre a probalibidade de um evento ocorrer sobre a probabilidade do mesmo não ocorrer, que é utilizado para mensurar a variável dependente em regressão logística (HAIR *et al*, 1995).

$$\frac{\textit{Prob (evento ocorrer)}}{\textit{Prob (evento não ocorrer)}} = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_i x_i}$$

Onde os coeficientes estimados $\beta_0, \beta_1, \ldots, \beta_i$ são na verdade, medidas das variações na proporção das probabilidades, chamada de razão de desigualdade. Coeficiente positivo indica aumento da probabilidade, o contrário (negativo) representa diminuição de probabilidade prevista.

Para comprovar a adequação do modelo, deve-se verificar a precisão dos parâmetros estimados, construir intervalos de confiança, testar hipóteses e por último realizar análise de diagnóstico e de resíduos.

Este processo, denominado inferência, é baseado na teoria de máxima verossimilhança. De acordo com esta teoria existem três estatísticas para testar hipóteses relativas aos β's: 1. Estatística de Verossimilhança; 2. Estatística Wald; e 3. Estatística *Score*.

Para testar as hipóteses:

$$H_0: \beta = \beta_0$$

$$H_A: \beta \neq \beta_0$$

A maximização da função de verossimilhança que calcula a probabilidade de que um evento ocorra é equivalente a minimizar a função logaritmo de verossimilhança (-2LL)[3]. O objetivo é verificar o poder de ajuste da equação, ou seja, verificar o quanto as variáveis independentes explicam a variável dependente (o poder de influência sobre a variável dependente). Um modelo com bom ajuste terá um valor baixo para -2LL, sendo que o valor mínimo é 0 (zero). Um modelo com ajuste perfeito terá como resposta um valor de verossimilhança igual a 1 (um) e, portanto, -2LL será igual a 0 (zero).

Segundo Zanini (2007), o valor da verossimilhança também pode ser comparado entre equações, onde a diferença representa a mudança no ajuste preditivo de uma equação para outra. Programas estatísticos têm testes estatísticos automáticos para a significância dessas diferenças. O teste qui-quadrado para a redução no valor do logaritmo da verossimilhança fornece uma medida de melhora devido à introdução da(s) variável(eis) independente(s).

---

[3] Minimização da função de verossimilhança significa -2 vezes o logaritmo do valor da verossimilhança e é chamada de -2LL, ou -2logverossimilhança. Um modelo bem ajustado terá um valor pequeno para -2LL.

Para o ajuste geral do modelo (medir seu poder de explicação), além do teste qui-quadrado, é utilizado, de forma similar à regressão linear, o coeficiente de explicação ou determinação "**R^2**", denominado para regressão logística como "Pseudo R^2" e calculado pela equação[4]:

$$R^2_{logit} = \frac{2LL_{base} - (-2LL_{modelo})}{-2LL_{base}}$$

Somado a isto, a medida Hosmer e Lemeshow de ajuste é um teste estatístico que tem a finalidade de avaliar a validade preditiva do Modelo de Regressão Logística, baseado, não no valor de verossimilhança, mas na visão real da variável dependente. Estas medidas combinadas darão suporte para que o aceite do modelo de regressão logística como significante, assegurando a evidência de significância estatística das variáveis relevantes, fatores como a importância da variável em relação ao evento modelado e a influência conjunta de outras importantes variáveis (HAIR *et al*, 1995).

Hair *et al* (1995), sugere as seguintes etapas para aplicação do modelo estatístico:

1. Objetivos da pesquisa;
2. Desenho da pesquisa;
3. Premissas estatísticas;
4. Construção do Modelo de Regressão Logística;
5. Avaliação do Ajuste Geral do Modelo;
6. Interpretação dos Resultados; e
7. Validação dos Resultados.

Como as três primeiras já foram apresentadas neste trabalho, as demais etapas (4 a 7) serão demonstradas a seguir.

[4] Um modelo com ajuste perfeito tem um valor de -2LL de 0 e um R^2_{logit} de 1 (HAIR et al, 1995).