3
Metodologia

3.1
Tipo de Pesquisa

O principal objetivo de uma pesquisa é contribuir para o conhecimento, endereçando algumas das muitas perguntas não respondidas. Para fazê-lo de forma satisfatória e para avaliar se houve contribuição significante para a coleção do conhecimento, um pesquisador deve cumprir o método científico (Remenyi *et al*., 1995).

O método científico é um conjunto de regras informal, mas rigoroso, que foi desenvolvido para garantir a integridade, a confiabilidade e a reproducibilidade do trabalho (Remenyi *et al.*, 1995).

A escolha da metodologia está intimamente ligada à forma de observação do objeto de estudo. Segundo Remenyi *et al*. (1995), a essência do conhecimento científico é que ele é derivado das observações realizadas no mundo. Para o autor, tudo o que não pode ser observado, direta ou indiretamente, através de seus efeitos ou conseqüências, está fora do domínio da ciência.

De acordo com Remenyi (2002), a pesquisa científica pode ser classificada em dois grupos distintos: o teórico e o empírico. Para Creswell (1994), os trabalhos podem ser considerados teóricos quando não se fundamentam em dados empíricos produzidos ou analisados diretamente pelo pesquisador e/ou se propõem a discutir conceitos teóricos e/ou propor uma nova teoria em administração. Já os trabalhos empíricos utilizam dados coletados no campo ou em bases já existentes, e que utilizarão esses dados para um experimento ou hipótese. O empírico pode ser subdividido pelo tipo de pesquisa qualitativa e quantitativa, sendo que as duas perspectivas se propõe a testar uma teoria ou hipótese.

Seguindo o critério do autor, este trabalho pode ser classificado como empírico, com aplicação da metodologia quantitativa. Esta pesquisa é empírica porque se propõe a determinar a influência das variáveis demográficas, dos dados

do histórico de eventos e de transação dos clientes sobre o risco de cancelamento, utilizando uma base de dados já existente para definir o perfil do cliente que cancela sua assinatura de jornal. Quantitativa, porque a partir desta base de dados foi aplicado um modelo de regressão logística binário para prever o risco de desligamento dos clientes assinantes de jornais.

Este estudo se propõe a responder à seguinte pergunta: é possível definir o perfil do cliente que cancela uma assinatura de jornal?

Remenyi (2002) afirma ainda que na escolha da metodologia, a revisão da literatura deve ser analisada junto com alguns fatores. Esta revisão de literatura deve revelar um modelo de pesquisa e uma metodologia adequada e que já tenha sido aplicada antes a esse tipo de questão. Essa opção é consistente com a estrutura adotada nesta fundamentação, que serve de base para a elaboração de um modelo teórico antes de definir as variáveis que devem compor o modelo matemático de previsão de cancelamento.

3.2
Universo e Amostra

O universo da pesquisa é composto pelos clientes, ativos e inativos, registrados no banco de dados da empresa editora de jornal. A base de dados possui todas as assinaturas que estejam ativas, incluindo as vendidas desde o momento em que a empresa lançou o serviço de assinaturas em abril de 1976. Para as assinaturas inativas existe registro das assinaturas canceladas desde julho de 1995.

Os consumidores de assinaturas podem ser do tipo pessoa física e pessoa jurídica, mas a amostra foi selecionada considerando apenas a população de clientes do tipo pessoa física.

Foi definida uma janela de análise. Foram analisadas assinaturas canceladas entre junho de 2004 e maio de 2005 e assinaturas ativas em maio de 2005. De um total de 230.858 assinaturas ativas e 105.524 assinaturas canceladas extraiu-se duas amostras aleatórias simples, uma de treinamento com 35.549 casos e outra amostra para validação com 4.796 casos.

3.3
Coleta dos Dados

A coleta dos dados foi realizada da seguinte maneira:

a. pesquisa bibliográfica em livros, dissertações, *papers* e periódicos, para conceituar marketing de relacionamento, satisfação, lealdade e retenção de clientes, variáveis transacionais, segmentação de mercado, variáveis geográficas, demográficas, psicográficas e comportamentais, além de métodos estatísticos para tratamento e modelagem dos dados do histórico de eventos dos assinantes. Além de levantar as variáveis do modelo teórico de previsão de cancelamento;

b. pesquisa documental na base de dados fornecida pela empresa editora de jornais observando três grupos de variáveis: histórico de eventos, dados transacionais e variáveis demográficas. A unidade de medida do tempo empregada é o mês que é uma variável contínua e a unidade de análise é o assinante pessoa física.

3.4
Limitações do Método

Uma grande limitação deste estudo refere-se à coleta dos dados. As variáveis selecionadas estavam pré-definidas no banco de dados. Com isso, variáveis importantes dos clientes podem ter sido desconsideradas.

Outra restrição é o pequeno número de trabalhos publicados sobre a utilização de modelos de regressão logística para prever o risco de cancelamento de clientes.

O modelo desenvolvido na dissertação é único para a empresa analisada, portanto, este modelo não pode ser generalizado para outras empresas deste ou de outro setor.

3.5
Tratamento dos Dados

Foi escolhida uma amostra aleatória simples com 40.345 assinaturas de pessoas físicas de uma população de 1.765.221 assinaturas. Seguindo a proporção populacional, esta amostra é composta por 69% de assinantes ativos e 31% de inativos. Desta amostra de 40.345 assinaturas foi extraída uma nova amostra aleatória com 4.796 assinaturas, 35.549 assinaturas foram utilizadas para o desenvolvimento e aplicação do modelo e 4.796 assinaturas utilizadas para a validação do modelo que identifica o risco de cancelamento.

A técnica utilizada para indicar o risco de cancelamento dos assinantes deste jornal é a regressão logística binária que determina a probabilidade de uma assinatura pertencer ao grupo dos que cancelam ou ao grupo dos que não cancelam.

Tanto a variável dependente quanto as variáveis independentes são categóricas.

O tratamento estatístico dos dados e o desenvolvimento do modelo foram realizados com o apoio do SPSS 13.0.

3.5.1
Problema e Objetivo

Cada vez as empresas estão mais preocupadas em otimizar seu investimento em marketing, direcionando ações diferenciadas para públicos específicos. A segmentação passou a ser uma prática comum para se conhecer os clientes e implementar estratégias mais adequadas e, conseqüentemente, mais eficazes.

Diante disso, identificar o público mais propenso ao cancelamento torna-se necessário para se conhecer o tamanho do risco de perda na carteira de clientes de uma empresa e também para desenvolver ações de retenção específicas para este público.

Este estudo propõe um modelo de regressão logística que diferencia o grupo de clientes que cancela do grupo que não cancela. Sendo possível assim, conhecer o perfil dos clientes canceladores e desenvolver ações de retenção que aumentem

o tempo de permanência destes clientes. Diante disso, o objetivo é definir o perfil do cliente que cancela a assinatura do jornal.

3.5.2
O Método de Regressão Logística

Na literatura, existem duas técnicas com a capacidade de separar dois grupos ou alocar um novo elemento em um desses grupos. Esta é uma situação enfrentada por muitos pesquisadores. Ambas as técnicas, Análise Discriminante e Regressão Logística, se enquadram na classe de métodos estatísticos multivariados de dependência, pois relacionam um conjunto de variáveis independentes com uma variável dependente categórica (Sharma, 1996; Hair *et al*., 1998; Morgan e Griego, 1998).

As técnicas de discriminação buscam uma função ou conjunto de funções que discrimine os grupos definidos pela variável categórica, visando a minimizar erros de classificação. Em contexto no qual o conjunto de variáveis independentes possui comportamento probabilístico de normalidade multivariada, a análise discriminante é adequada, porque minimiza os erros de classificação (Sharma, 1996; Hair *et al*., 1998). Portanto a suposição de normalidade multivariada é necessária para que os resultados da análise discriminante sejam satisfatórios.

O modelo logístico é mais comumente utilizado porque a análise discriminante impõe às variáveis independentes condições como: serem normalmente distribuídas e terem suas matrizes de variância-covariância iguais entre os dois grupos de classificação. Outro ponto que é motivo de crítica quanto ao método da análise discriminante é que o resultado da expressão discriminante fornece um *score* que possui pouca interpretação intuitiva (Castro, 2003). Segundo Ohlson (1980, p.112), este *score* é basicamente um dispositivo (discriminatório) de classificação ordinal, não tendo embutido nenhum aspecto probabilístico. Além de não depender da exigência de normalidade das variáveis independentes e da igualdade de matrizes de covariância, a regressão logística é semelhante a uma regressão múltipla, pois possui o poder de incorporar efeitos não lineares (Hair, 1998). A possibilidade de interpretação direta dos coeficientes como medidas de associação é uma das grandes vantagens da regressão logística;

a interpretação destes coeficientes pode ser estendida para qualquer problema prático (Paula, 1999)

Hair (1998) afirma que existem algumas razões pelas quais a regressão logística representa uma alternativa atraente à análise discriminante sempre que a variável dependente tiver somente duas categorias. Em primeiro lugar, a regressão logística é menos afetada pelas desigualdades variância/covariância dentre os grupos, um pressuposto básico da análise discriminante. Em segundo lugar, a regressão logística pode cuidar facilmente de variáveis categóricas independentes, enquanto que na análise discriminante o uso de variáveis *dummy* criou problemas com as igualdades variância/covariância. Finalmente, a regressão logística é similar à regressão múltipla em termos de sua interpretação e nas medidas de diagnóstico direcionadas a casos disponíveis para o exame de resíduos.

Segundo Hosmer e Lemeshow (1989), a técnica de regressão logística tornou-se um método padrão de análise de regressão para variáveis medidas de forma dicotômica, especialmente nas áreas de ciência da saúde. Muitas situações em análise de dados envolvem prever o valor de uma variável de resultado categórico. Essas situações incluem aplicações na medicina prevendo o estado de saúde de um paciente, na pesquisa de marketing prevendo se uma pessoa irá comprar o produto ou em escolas prevendo o êxito de um aluno. A regressão logística é uma técnica que pode ser muito útil nessas situações ou em muitas outras. Mesmo quando a resposta de interesse não é originalmente do tipo binário, alguns pesquisadores têm dicotomizado a resposta de modo que a probabilidade de sucesso possa ser modelada através da regressão logística (Paula, 1999). O mesmo modelo pode ser utilizado com enfoque discriminatório, conforme descrevem Krzanowski (1988) e McLachlan (1992). Esses autores defendem o modelo logístico de discriminação como um método utilizado de forma mais abrangente, pois não faz suposições quanto à forma funcional das variáveis independentes, e o número de parâmetros envolvidos no processo de estimação provavelmente será menor.

Ao contrário da análise discriminante, a regressão logística exige um número menor de pressupostos que são menos rígidos, são eles: as variáveis independentes devem ser intervalo, taxa ou dicotômicas; todos os previsores relevantes foram incluídos, nenhum previsor irrelevante foi incluído e a forma do relacionamento é linear; o valor esperado por termo de erro é zero; não há

autocorrelação; não há correlação entre o erro e as variáveis independentes; há uma ausência de multicolinearidade perfeita entre as variáveis independentes (SPSS, 2003).

Além disso, a regressão logística funciona quase tão bem quanto a análise discriminante mesmo quando os pressupostos desta são satisfeitos. Comparando as duas técnicas, Krzanowski (1988) diz que é consenso geral que a discriminação logística deve ser preferida, quando as distribuições são claramente não-normais. A mesma afirmação é sustentada por Press e Wilson (1978). Hair *et al*. (1998) apontam uma lista de motivos que levariam o pesquisador a optar pela regressão logística: não é necessário supor normalidade multivariada; é uma técnica mais genérica e mais robusta, pois sua aplicação é apropriada em muitas situações e é uma técnica similar à regressão linear múltipla.

Outra vantagem da regressão logística é sua abordagem probabilística, já que essa regressão estima a chance de ocorrer um certo evento a partir de uma série de variáveis independentes ou explanatórias. Segundo Hower e Lemeshow (1989), o objetivo da regressão logística é achar o melhor relacionamento entre a variável resposta (de saída ou dependente) e um conjunto de variáveis explanatórias ou preditivas, sendo o modelo final aquele que apresentar o melhor ajuste matemático e for naturalmente razoável de se explicar. A regressão logística é projetada para utilizar uma combinação de variáveis previsoras contínuas e categóricas para prever uma variável de resultado categórico ou dependente.

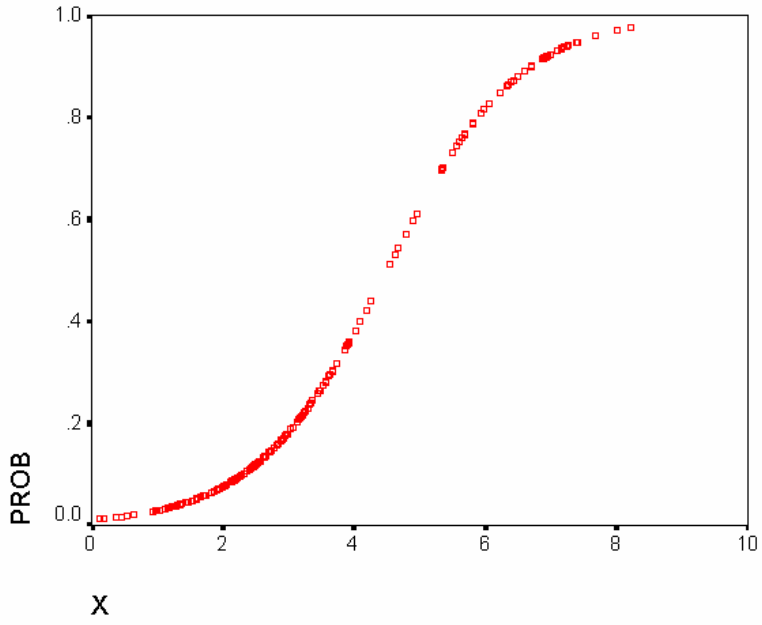
Hair (1999) defende que a regressão múltipla não pode ser usada para prever os valores de uma variável dependente dicotômica, pois, quando tentamos prever os valores de uma variável codificada, digamos, 0 ou 1, podemos considerar os valores estimados como sendo probabilidades, isto é, a probabilidade de obter um valor previsto de 1. Numa regressão múltipla com um ajuste de linha reta aos dados, este é freqüentemente o caso em que valores inferiores a 0 ou superiores a 1 são previstos. Ou seja, os modelos de regressão linear não são apropriados para se analisar a relação de diversos fatores e a probabilidade de ocorrência de um evento. Isto ocorre, porque o valor esperado da variável resposta, dado um conjunto de variáveis explicativas, pode assumir valores que vão de menos infinito até mais infinito e isto viola as leis de probabilidade. Além disso, um dos pressupostos da regressão é a homogeneidade da variância. Entretanto, para uma

variável dicotômica, o desvio médio e padrão são relacionados porque o desvio padrão está, onde *p* é a média da variável. Já que existe um relacionamento funcional entre o desvio padrão e a média, a homogeneidade de variância dentre os valores da variável dependente não pode ser satisfeita.

A regressão logística foi desenvolvida nos anos 60 como uma solução para este problema. Segundo Hair (1999), a regressão logística é uma técnica robusta e bem adequada quando há violação do pressuposto de igualdade das matrizes de variância/covariância dentre os grupos.

Ao predizer o valor de uma variável numa escala de 0 a 1, faz sentido ajustar uma curva em forma de S aos dados.

GRÁFICO 4: Curva Logística



A função logística está limitada em 0 e 1, de modo que predições impossíveis não podem ocorrer. Há efetivamente uma família inteira de funções em forma de S, sendo o *probit* uma variante bem conhecida. Devido a várias considerações, a maior parte dos estatísticos concorda que a logística é um modelo opcional para a regressão com uma medida dependente dicotômica (SPSS, 2003).

O primeiro passo é verificar se os dados foram coletados em escala contínua ou discreta. Para situações tempo-discreto, Allison, P. (1984) sugere o uso do modelo *logit* e do método *maximum likelihood* para estimação dos parâmetros do

modelo. Se todas as variáveis explicativas forem categóricas, a estimação do modelo *logit* pode ser feita por método log-linear.

Para Moore (1994), a análise *logit* é um método que determina quais variáveis independentes devem ser incluídas no modelo para se prever adequadamente a variável dependente categórica. Como fator limitador para uso da análise *logit*, as variáveis dependentes e independentes devem ser categóricas, obrigatoriamente. Ott e Markewich *apud* Moore (1994) afimam que a regressão logística é o método mais apropriado de análise quando as variáveis independentes são contínuas.

A regressão logística reescreve o modelo clássico de regressão linear de modo a confirmar o valor da variável resposta para a faixa de 0 a 1, ao mesmo tempo em que as variáveis independentes possam variar continuamente. Isto é obtido pela equação abaixo, também conhecida como função logística.

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

que pode ser linearizada pela transformação:

$$g(x) = \ln \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 X$$

A regressão logística binária é a regressão 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. A equação geral para regressão logística é:

$$\ln(Odds) = \alpha + \mathrm{B}_1 X_1 + \mathrm{B}_2 X_2 + \ldots + \mathrm{B}_k X_k$$

onde os termos da direita são os termos padrão para as variáveis independentes e o intercepto numa equação de regressão. Entretanto, do lado esquerdo está o log natural da probabilidade e a quantidade ln(*Odds*) é chamada de *logit* e pode variar

de menos até mais infinito. Portanto retirando o problema de predição para fora dos limites da variável dependente. As probabilidades são relacionadas por:

$$Odds = \frac{prob}{1 - prob}$$

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, a equação logística pode ser transformada numa equação na probabilidade (Hair, 1999). Assume então esta forma:

$$prob(event) = \frac{1}{1 + e^{-(\alpha + B_1 X_1 + B_2 X_2 + \ldots + B_k X_k)}}$$

Diferente da regressão linear clássica, os erros desse modelo não seguem uma distribuição normal, mas sim a de Bernoulli. Assim, enquanto na regressão linear o método usado para estimar os coeficientes β_0, ..., β_n é o método dos mínimos quadrados, na regressão logística usa-se o método da máxima verossimilhança. Este último método encontra os valores dos parâmetros β_0, ..., β_n, que maximizem a probabilidade de se obter o conjunto observado de dados (Hosmer e Lemeshow, 1989).

Ao utilizarmos regressão logística, a principal suposição é a de que o logaritmo da razão entre as probabilidades de ocorrência e não ocorrência do evento é linear. Por essa razão, ao interpretar os coeficientes da regressão logística, interpreta-se e e não . Contudo, quando se utiliza o modelo logístico do ponto de vista de discriminação entre grupos, não há grande interesse na interpretação dos coeficientes (Garson, 2000).

A regressão logística é estimada de forma bem semelhante à regressão múltipla pelo fato de que um modelo de base é primeiro estimado visando a fornecer um padrão para comparação. Na regressão múltipla, a média é utilizada para estabelecer o modelo base e calcular a soma total dos quadrados. 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. Neste modelo, as correlações parciais para cada variável podem ser estabelecidas e a variável mais discriminante pode ser escolhida de acordo com o critério de seleção (Hair, 1998).

Para definir o ponto de corte é necessário conhecer a probabilidade *a priori* de um assinante cancelar a sua assinatura. Com isso, o ponto de corte para este estudo deve ser o valor que minimize os erros de classificação dos assinantes (erros Tipo I e Tipo II). O maior problema enfrentado pelos profissionais envolvidos é a obtenção do valor de corte. A grande questão é como obter um valor de corte confiável a ponto de evitar perdas para a empresa, tanto pela não classificação de risco para clientes que podem cancelar e, conseqüentemente, não serão impactados pelas ações de marketing, quanto pela atribuição de risco para clientes sem propensão ao cancelamento e que consomem parcelas importantes dos investimentos de marketing utilizados na retenção de clientes. Analisada a base de clientes dos últimos 12 meses percebe-se que a proporção entre ativos e cancelados é de 69% de clientes ativos. Conseqüentemente, define-se o ponto de corte com sendo 0,69.

O modelo de regressão logística minimiza o número de variáveis para que o modelo resultante seja mais facilmente generalizado e mais estável numericamente, dado que quanto mais variáveis são incluídas no modelo, mais ele se torna dependente dos dados. O uso da técnica *stepwise* na regressão logística é o processo de inclusão ou exclusão de variáveis do modelo, baseado em critérios tais com a estatística G e o teste Wald.

Existem os métodos *enter, backward* e *foward*. O método *enter* incorpora ao modelo todas as variáveis e deve ser utilizado principalmente quando se tem certeza de que todas as variáveis são necessárias para se estimar os parâmetros do modelo. O método *backward* caracteriza-se por incorporar todas as variáveis e após percorrer várias etapas, uma variável por vez pode ser eliminada. Se em uma determinada etapa não houver eliminação de alguma variável, o processo é então interrompido e as variáveis restantes definem o modelo final. Numa dada etapa, temos um determinado modelo que denominamos modelo completo da etapa e são investigadas as contribuições individuais das variáveis a esse modelo. A variável de pior desempenho é eliminada comparando-se o modelo completo com o modelo reduzido, pela retirada de tal variável (Charnet *et al*, 2000).

O método *foward* caracteriza-se por considerar a variável de maior coeficiente de correlação amostral observado com a variável resposta. A cada etapa, uma variável pode vir a ser incorporada. Se em uma etapa não houver uma inclusão, o processo é interrompido e as variáveis selecionadas até esta etapa definem o modelo final (Charnet *et al*., 2000).

Em uma determinada etapa chega-se a um modelo definido como modelo reduzido. O modelo reduzido de cada etapa é comparado ao modelo em que uma nova variável é acrescentada. Existindo um modelo de melhor desempenho, a correspondente variável é incorporada ao elenco de variáveis já escolhidas. Enquanto em uma etapa do procedimento *backward* comparam-se vários modelos reduzidos com um único modelo completo devido ao objetivo de eliminar uma variável, em uma etapa do procedimento *forward* as comparações são feitas entre vários modelos completos e um único modelo reduzido, graças ao objetivo de incorporar uma variável (Charnet *et al*., 2000).

Os três métodos diferem pela forma que utilizam para selecionar uma variável. Os métodos direcionados a passos utilizam a estatística Wald, a mudança na probabilidade ou a estatística condicional como método de escolha para a eliminação ou inclusão de variáveis. (SPSS, 2003). O método escolhido foi o *foward conditional* que utiliza a estatística condicional para incluir variáveis ao modelo é menos intensivo computacionalmente.

Para a obtenção do modelo final de regressão logística, após estimar os coeficientes da equação de regressão, é necessário verificar se cada variável é significantemente relacionada com a variável resposta do modelo. Isto é realizado geralmente através da formulação de testes de hipóteses estatísticas, que avaliam o modelo com a variável e sem a variável (Hower e Lemeshow, 1989).

Há dois testes estatísticos para a significância do modelo final. Primeiro, um teste *chi-quadrado* para mudança no valor –2LL do modelo base é comparável com o teste F geral em regressão múltipla. Além disso, a medida Hosmer e Lemeshow de ajuste geral tem um teste estatístico que indica se houve ou não diferença estatisticamente significativa entre as classificações observadas e previstas. Estas duas medidas, em combinação, fornecem suporte para que se aceite o modelo de regressão logística como significante. Estes testes asseguram a evidência de significância estatística das variáveis, devendo se considerar outros

relevantes fatores como a importância da variável em relação ao evento modelado e a influência conjunta de outras importantes variáveis (Hair, 1998).

Segundo Hair (1998), o ajuste geral do modelo pode ser avaliado utilizando-se algumas medidas como o -2LL. Se no modelo avaliado houver um decréscimo no valor –2LL comparado ao modelo base, existe melhora no modelo, pois valores menores da medida –2LL indicam o melhor ajuste do modelo. A seguir, as medidas de adequação de ajuste comparam as probabilidades estimadas com as probabilidades observadas sendo que os valores mais altos indicam um ajuste melhor. Existem ainda três medidas comparáveis ao R^2 no modelo de regressão múltipla. A medida R^2 Cox e Snell opera da mesma forma, com valores mais altos indicando maior ajuste do modelo. Entretanto, esta medida está limitada pelo fato de que não consegue alcançar o valor máximo de 1, de modo que Nagelkerke propôs uma modificação que tem o alcance de 0 para 1. A terceira medida é a medida R^2 "pseudo" com base na melhoria do valor –2LL. O pseudo R^2 é calculado como:

$$R^2_{\log it} = \frac{2LL_{base} - \left(-2LL_{\mathrm{mod}elo}\right)}{-2LL_{base}}$$

A medida final do ajuste do modelo é o valor Hosmer e Lemeshow, que 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.

A medida Hosmer e Lemeshow ainda mostra a não-significância, indicando a ausência de diferença na distribuição de valores dependentes efetivos e previstos.

Finalmente, as matrizes de classificação, idênticas em natureza às utilizadas na análise discriminante, mostram se as taxas de acerto são altas ou baixas para os casos corretamente classificados no modelo.

A validação do modelo de regressão logística pode ser obtida através do mesmo método utilizado na análise discriminante: criação de amostras de treinamento e validação (Hair, 1998). Se as taxas de acertos da amostra de treinamento e da amostra de validação forem similares, pode-se concluir que o

modelo tem suporte empírico no mesmo nível para explicar as variáveis dependentes.

3.5.3
Coleta e Análise dos Dados

A amostra inicial contava com 36.000 assinaturas. Após a análise da base de dados, foram excluídos os assinantes com as seguintes características:

a. assinaturas com forma de pagamento “não informado”;
b. assinaturas com tipo de assinatura “em branco”;
c. assinantes com idade inferior a 18 anos e superior a 90 anos.

Depois desse tratamento, a amostra ficou com 35.549 assinaturas, sendo 24.492 ativas e 11.057 canceladas. A base final para estudo contém as variáveis: status (ativa ou inativa), tipo de assinatura, fonte de venda, forma de pagamento, região, tempo de permanência, LTV, SD&W, gênero, faixa etária, indicativo de participação em ações de fidelização, publicação de anúncio nos últimos 12 meses, histórico de reclamação, compra de produto agregado, quantidade de produtos agregados comprados.

No banco de dados da empresa editora de periódicos a proporção é de 31% de inativos e 69% de ativos.

3.5.4
Definição das Variáveis

Em uma primeira etapa foram consideradas as seguintes variáveis:

Variáveis Transacionais		
Descrição	**Definição**	**Níveis**
Status do cliente	Indica se o cliente continua assinando o jornal ou se cancelou o serviço.	Ativo e Inativo
Tipo de Assinatura	Período pelo qual o cliente contratou a assinatura	Anual; boleto mensal; débito automático; semestral e trimestral.
Fonte de Venda	Representa o canal de vendas pelo qual o cliente adquiriu uma assinatura.	Agente; antigo (na migração do sistema os dados foram convertidos como antigo); Internet; outros; sem fonte; terceiros; telemarketing ativo; telemarketing receptivo.
Forma de pagamento	Meio de pagamento utilizado pelo cliente	Boleto bancário; cartão de crédito no débito automático; cartão de crédito parcelado; débito em conta corrente automático; débito em conta corrente parcelado; não informada.
Tempo de permanência	Quantidade de meses que o cliente ficou ativo no caso dos cancelados; ou que ainda está ativo, no caso dos atuais clientes	De 0 até 3 meses; de 3 até 6 meses; de 6 meses até 1 ano; de 1 ano até 2 anos; de 2 anos até 3 anos; mais de 3 anos.
Quantidade de Produto Agregado	Indica o volume de produtos comprados pelos clientes	Até de 3 produtos; entre 4 e 7 produtos; entre 8 e 10 produtos e mais de 10 produtos
LTV	*Score* de segmentação dos clientes por *life time value* que é igual ao tempo de permanência atual em meses+ tempo de vida futuro em meses (estimado por análise de sobrevivência) x margem (R$) da assinatura mensal do cliente.	Platina; Diamante; Ouro; Prata; Bronze; Lata; Indefinido.

Indicador de compra de produto agregado	Indica se o cliente já comprou ou não um produto agregado (livros, guias, CDs e outros colecionáveis)	Sim ou Não
Indicador de compra de anúncio	Indica se o cliente já comprou um anúncio nos classificados ou não	Sim ou Não
Indicador de Reclamação	Indica se o cliente já fez uma reclamação através do *call center* ou pela internet	Sim ou Não
Variáveis Geográficas		
Descrição	**Definição**	**Níveis**
Região	Bairro ou cidade onde o assinante recebe o jornal agrupado por similaridade nos custos de distribuição	Baixada, subúrbio, Barra, Zona Norte, Zona Sul, Brasília, interior do RJ, Niterói, outros estados, São Paulo/Belo Horizonte
Variáveis Demográficas		
Descrição	**Definição**	**Níveis**
Gênero	Indica o sexo do assinante; alguns assinantes não têm o campo preenchido	Feminino; masculino e indefinido
Faixa Etária	Agrupamento que indica a idade do assinante titular	menos de 20 anos; de 20 a 25 anos; de 26 a 30 anos; de 31 a 35 anos; de 36 a 40 anos; de 41 a 50 anos; de 51 a 60 anos; de 61 a 80 anos; mais de 80 anos; sem preenchimento
SD&W	*Score* de classificação sócio-demográfica utilizando variáveis do senso do IBGE para os setores censitários do Rio de Janeiro	Solidez; segurança; equilíbrio em conquista; estabilidade; sobrevivência; privação; penúria; subsistência e indefinido.
Variáveis de Fidelização		
Descrição	**Definição**	**Níveis**
Indicador de participação em ações de fidelização	Indica se alguma vez o cliente participou das ações de relacionamento tais como vantagens, descontos ou participação em eventos exclusivos.	Sim ou Não

A SD&W é uma empresa de consultoria que desenvolve um modelo que utiliza variáveis sócio- demográficas do censo para classificar a população em 9 segmentos distintos de acordo com as características destes indivíduos.

O objetivo deste modelo é desenvolver indicadores (ou scores) capazes de traduzir informações tais como classe de renda, instrução, população, moradia, infra-estrutura.

Inicialmente as seguintes variáveis do censo 2000 foram utilizadas: número de domicílios, população, renda domiciliar, grau de instrução do chefe de família, tipo de moradia (casa, apartamento/ favela ou não), condição de moradia (próprio, alugado, cedido), número de moradores no domicílio,faixa etária da população, condição na família (chefe, cônjuge, filhos), existência de empregados domésticos no domicílio, infra-estrutura (abastecimento de água, instalações sanitárias, coleta de lixo).

Através de análise fatorial foram criados fatores (ou indicadores) capazes de agregar informações correlacionadas, de maneira a não se perder o conteúdo das mesmas. Sendo assim, cada fator responde por uma parcela da variabilidade dos dados. O ideal é se obter o menor número de fatores com um maior percentual de variabilidade explicada.

Utilizando as 11 variáveis do censo, foram determinados 5 indicadores cuja nomenclatura está diretamente relacionada às características das variáveis que o compõem. São eles:

Nome	Componentes principais do indicador
Padrão de Vida (1)	▪ Domicílios com renda superior a 15 S.M. ▪ Chefe do Domicílio c/ superior completo ▪ Empregados no Domicílio ▪ Domicílios do tipo Apartamento
Composição Familiar Reduzida e Madura (2)	▪ Domicílios com até 2 moradores ▪ Domicílios alugados ▪ Moradores com 60 anos ou mais ▪ Ausência de filhos/ enteados
Ausência de infra-estrutura de desenvenvolvimento. (3)	▪ Ausência de Rede Geral ou Fossa Séptica ▪ Uso de vala (para esgoto) ▪ Lixo queimado ▪ Ausência de coleta de lixo
Adensamento Populacional (4)	▪ Número de domicílios ▪ População Total
Favelização (5)	▪ Proporção de domicílios do tipo Casa em aglomerado subnormal (favela)

Cada um destes indicadores foi calculado para a menor unidade de área disponível - o setor censitário (fonte: IBGE/2000). Em seguida, a técnica estatística utilizada foi a análise de clusters, que consiste na formação de grupos com maior homogeneidade interna e maior heterogeneidade entre si, segundo um determinado grupo de variáveis. Neste caso, as variáveis utilizadas são os indicadores.

Desta análise de cluster surgiram os grupos:

- SOLIDEZ (4% da População): Grupo que apresenta solidez seja ela conquistada através do tempo ou herdada.
    - É o mais Alto Padrão de Vida: Classe A (65%), chefe de família com nível superior, presença de empregados domésticos.
    - Vivem em apartamento próprio, com total infra-estrutura de desenvolvimento.
    - Família de tamanho tradicional, mas não há destaque para presença de filhos / crianças ou adolescentes
    - Somente Rio de Janeiro e Niterói.
    - Alguns bairros onde predomina: no RJ: parte da Urca, São Conrado, Joá, Leblon, Lagoa, parte da Barra da Tijuca. Em Niterói: parte de Icaraí, Santa Rosa, Ingá.
- SEGURANÇA (10%): Grupo baseado na segurança obtida através do desenvolvimento profissional.
    - Alto Padrão de Vida: Classe A/B (41% na Classe A), chefe de família com nível superior, alguma presença de empregados domésticos.
    - Vivem em apartamento, alguns alugados.
    - Composição familiar madura e reduzida, destaque para domicílios com até 2 moradores (cerca de metade dos domicílios).
    - Somente Rio de Janeiro e Niterói, com destaque para Niterói.

- Alguns bairros onde predomina: no RJ: Copacabana, Tijuca, Botafogo, Maracanã, parte do Grajaú. Em Niteró: parte de Icaraí, Santa Rosa, Ingá e Cubango.

- EQUILÍBRIO EM CONQUISTA (3%): Grupo de famílias em processo de conquista de equilíbrio financeiro, com filhos ainda pequenos/jovens.

  - Alto Padrão de Vida: Classe A/B (36% na Classe A), boa proporção de chefes de família com nível superior, alguma presença de empregados domésticos.

  - Vivem em casa (70%) própria. Em alguns casos, não contam com total infra-estrutura urbana de desenvolvimento (não têm rede geral de esgoto - usam fossa séptica; usam poço/ nascente para abastecimento de água)

  - Presença de filhos pequenos e/ou adolescentes.

  - Família de tamanho tradicional, em alguns casos com até mais de 4 moradores.

  - Rio de Janeiro e Niterói, sendo que cerca de 25% dos setores está em Niterói.

  - Alguns bairros no RJ: parte da Barra da Tijuca, parte do Itanhangá, parte de Jacarepaguá. Em Niterói: Itaipú, Jacaré, parte da Lagoa de Piratininga, parte de São Francisco.

- ESTABILIDADE (4%): Grupo de família madura e reduzida, com estabilidade calcada no emprego.

  - Bom Padrão de Vida: Classe A/B (55%), boa proporção de chefes de família com nível superior.

  - Vivem em apartamento próprio ou alugado (maior proporção de alugados entre os segmentos)

  - Moram sozinhos ou no máximo com mais uma pessoa (dois terços do segmento possuem até 2 moradores). Idade mais madura.

  - Rio de Janeiro e Niterói

- Alguns bairros no RJ: centro, parte da Ilha do Governador, Santa Tereza, Glória, parte de Copacabana, Flamengo. Em Niterói: Centro.

- SOBREVIVÊNCIA (49%): Grupo de sobrevivência relacionada ao emprego e escolaridade intermediária.
  - Padrão de Vida intermediário: Classe B/C (58%), chefe de família com nível de escolaridade intermediário.
  - Vivem em casa (a maioria) ou apartamento próprio, com boa infra-estrutura urbana.
  - Família de tamanho tradicional.
  - É o grupo mais freqüente em praticamente todos os municípios estudados.
  - Nova Iguaçu, Caxias. Alguns bairros no RJ: a maior parte da Zona Norte: Méier, Penha, São Cristóvão, Realengo, Marechal Hermes, Olaria. Em Niterói : Barreto, parte de Fonseca, Santa Bárbara.

- SUBSISTÊNCIA (15%): Grupo onde a subsistência está vinculada à proximidade do desenvolvimento urbano, mas sem usufruir completamente da infra-estrutura.
  - Baixo Padrão de Vida: Classe D/E (57%), baixo nível de escolaridade
  - Vivem em casa (a maioria) própria, com fraca infra-estrutura de desenvolvimento, embora não seja favela.
  - Família mais numerosa, com presença de filhos /jovens.
  - Nova Iguaçu, Caxias e Rio de Janeiro.
  - Alguns bairros no RJ: Vargem Grande, Guaratiba, parte de Campo Grande, parte de Bangu, Santa Cruz. Em Niterói: Ititioca, Maceió.

- PRIVAÇÃO (11%): Grupo com baixíssimo padrão de vida, porém com uso-fruto de uma infra-estrutura existente.
  - Baixíssimo Padrão de Vida: Classe D/E (63%), baixo nível de escolaridade do chefe de família

- Vivem em casa (a maioria) própria, em favela, porém desfrutam de razoável infra-estrutura urbana.
- Família numerosa, com presença de filhos /jovens.
- Rio de Janeiro (principalmente), Caxias e Niterói.
- Alguns bairros no RJ: parte da Rocinha, Complexo do Alemão, Vidigal, parte do Catumbi. Em Niterói: Morro do Estado.

- PENÚRIA (4%): Ausência de padrão de vida e de acesso à infra-estrutura de desenvolvimento.
  - Baixíssimo Padrão de Vida: Classe D/E (68%), baixo nível de escolaridade do chefe de família.
  - Vivem em casa (a maioria) própria, em favela, em precárias condições de infra-estrutura.
  - Família numerosa, com presença de filhos /jovens.
  - Rio de Janeiro (principalmente), Niterói e Caxias.
  - Alguns bairros no RJ: parte do Catumbi (Rua Itapiru), Tijuca (próximo Rua São Miguel), parte da Rocinha.

A empresa desenvolve uma série de ações de fidelização para os assinantes do jornal. Essas ações estão relacionadas a vantagens em estabelecimentos conveniados, por exemplo, a rede de farmácias Drogasmil e o Mc Donald's, além da participação em eventos exclusivos como a pré-estréia de filmes ou peças de teatro. Esses clientes são selecionados pela empresa e recebem uma carta para comunicar o benefício oferecido.

3.5.5
Descrição da Base de Dados

Após serem selecionadas as 14 variáveis que poderiam explicar o cancelamento de uma assinatura de jornal, estatísticas descritivas foram obtidas para avaliar a consistência da base de dados.

Não foram identificados padrões inusitados, *outliers*, problemas de dados ausentes e incoerências. Com isso, foi aplicado o modelo de regressão logística objetivando investigar a relação dessas variáveis com a probabilidade de cancelamento de uma assinatura.

Quanto ao tipo de assinatura, a base é composta por 88% das assinaturas em débito automático, 11% por assinaturas da modalidade anual e 1% em outras modalidades. Metade paga no cartão de crédito, 25% no débito em conta corrente e 25% através de boleto bancário.

Quase metade das assinaturas foram adquiridas através do telemarketing ativo – interno (37%) e terceirizado (7%) – e 18% através do telemarketing receptivo.

Dois terços dos assinantes estão localizados na Zona Sul, Barra, Niterói e Zona Norte do Rio de Janeiro, a distribuição geográfica dos assinantes ativos e dos clientes que cancelam está no Anexo B deste estudo. Sendo 52% da base classificada nos segmentos prioritários da SD&W solidez, segurança, equilíbrio em conquista e estabilidade. Dos clientes, 22% possuem LTV platina, diamante ou ouro, sendo que 36% da base é cliente há mais de 3 anos.

Da base analisada, 51% pertencem ao sexo masculino e 68% dos assinantes possuem mais de 40 anos de idade.

Participaram de uma ação de fidelização 73% dos clientes e 62% fizeram alguma reclamação sobre o produto ou os serviços da empresa.

Dos clientes assinantes, 35% compraram um outro produto da empresa. Destes, 86% compraram um produto agregado, 9% compraram um anúncio de publicidade e 5% compraram ambos.

As tabelas com as análises descritivas da base podem ser encontradas no Anexo A.