UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ADMINISTRAÇÃO
DEPARTAMENTO DE CIÊNCIAS ADMINISTRATIVAS

LAUREN PAESE MARTINS DA SILVA

O MERCADO DE PRODUTOS DIGITAIS: UM ESTUDO DE *CHURN* DE MIGRADOS DE ASSINATURA DE JORNAL

PORTO ALEGRE
2017

LAUREN PAESE MARTINS DA SILVA

O MERCADO DE PRODUTOS DIGITAIS: UM ESTUDO DE *CHURN* DE MIGRADOS DE ASSINATURA DE JORNAL

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciências Administrativas da Escola de Administração da Universidade Federal do Rio Grande do Sul, como requisito parcial para a obtenção do grau de Bacharel em Administração.

Orientador: Prof. Dr. Vinícius Andrade Brei

PORTO ALEGRE

2017

LAUREN PAESE MARTINS DA SILVA

O MERCADO DE PRODUTOS DIGITAIS: UM ESTUDO DE *CHURN* DE MIGRADOS DE ASSINATURA DE JORNAL

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciências Administrativas da Escola de Administração da Universidade Federal do Rio Grande do Sul, como requisito parcial para a obtenção do grau de Bacharel em Administração.

BANCA EXAMINADORA

_____

Prof. Dr. Vinícius Andrade Brei – Orientador

_____

Profa. Ma. Carla Freitas Silveira Netto – Avaliadora

PORTO ALEGRE

2017

SUMÁRIO

RESUMO

Esta pesquisa buscou desenvolver um modelo de propensão a *churn* de assinantes migrados do produto impresso para o produto digital de uma empresa jornalística de Porto Alegre, Rio Grande do Sul. O estudo tem como objetivo identificar e entender o perfil de assinantes migrados – dados comportamentais e demográficos – para, por meio da aplicação, validação e comparação de dois métodos, construir um modelo eficaz de previsão de *churn*. Na construção do modelo foram utilizados os métodos de regressão logística e árvore de decisão, devido à praticidade de sua aplicação, capacidade de explicação da *target*, o *churn* e facilidade de compreensão. A base utilizada para a construção do modelo é composta de doze variáveis, das quais, na regressão logística, oito foram significativas; na elaboração da árvore de decisão, somente duas. Dentre os dois métodos utilizados, a árvore de decisão apresenta melhor acurácia.

Palavras-chave: *Churn*. Modelo preditivo. Assinatura.

ABSTRACT

This research aimed to develop the construction of a *churn* propensity model of migrated subscribers of the printed product to the digital product of a newspaper company in Porto Alegre, Rio Grande do Sul. The study pursues to identify and understand the profile of migrated subscribers - behavioral and demographic data – in order to construct an effective *churn* prediction model through the application, validation and comparison of two methods. In the construction of the model were used the logistic regression and decision tree methods, due to the practicality of its application, ability to explain the target chosen and ease of understanding. The base used for the construction of the model is composed of 12 variables, of which - in the logistic regression - 8 were significant and - in the elaboration of the decision tree - only 2. Among the two methods used, the decision tree presents a better accuracy.

Keywords: *Churn*. Predictive modelling. Signature.

INTRODUÇÃO

O contexto formado pelo aumento da competição suscitou a busca de vantagem competitiva pelas empresas por meio da fidelização de seus clientes, já que os mercados nos quais essas empresas competem tornaram-se saturados. Quando os produtos e serviços são de grande competitividade, a aquisição de novos clientes acaba ocorrendo às custas dos concorrentes, o que torna aqueles suscetíveis à mudança de escolha de serviço ou produto com o aparecimento de ofertas mais atrativas (KAMAKURA et al., 2003).

Sendo assim, as empresas devem manter seus clientes para se manter rentáveis. O campo acadêmico de gerenciamento de marketing considera o gerenciamento de relacionamento com o cliente como o mais apropriado para alcançar isso (LOOTS; GROBLER, 2014). O objetivo do marketing de relacionamento é a construção e manutenção de uma base de clientes comprometidos, que sejam rentáveis para a organização (PEPPERS; ROGERS, 2011). Para o estabelecimento do marketing de relacionamento, é necessário que haja, anteriormente, a aquisição de clientes com perfil passível de relacionamento de longo prazo com a empresa, pois assim a fidelização é facilitada, e a suscetibilidade a ofertas de concorrentes, amenizada.

Dentro do marketing de relacionamento, a retenção de clientes é de grande importância, pois com o crescente nível de competitividade no mercado, o grande desafio passa a ser o de reconhecer os clientes, mostrando a eles o quanto a empresa os estima por terem lhe conferido sua preferência (MILAN; TONI, 2012). Mas isso vai além da preocupação com a satisfação do cliente, que tem um papel essencial no desempenho das organizações (OLIVER, 2010), caracterizando-se como um novo modo de pensar a respeito do que gera receitas e lucro e sobre a forma de como se deveria fazer negócios (VAVRA; PRUDEN, 1995). Tendo em vista, então, a retenção de clientes, conceitos como o de *churn* são de grande importância, pois possuem relação direta com a retenção.

O *churn* do consumidor é definido pela propensão do cliente/consumidor de cessar a realização de negócios com uma empresa em um determinado período de tempo; tornou-se um dos principais desafios para as empresas em todo o mundo (CHANDAR et al. 2006**,** apud MA; TAN; SHU, 2015). O *churn* é diretamente relacionado à retenção de clientes e funciona como um indicador da eficiência na

retenção de uma empresa. Reduzir o *churn* de clientes, orientando campanhas de marketing especificamente para clientes com maior probabilidade de cancelamento, provou ser rentável às empresas. Objetivando aumentar a eficiência dessas campanhas, um modelo de predição é necessário, para que a identificação desses clientes seja possível (VERBRAKEN; VERBEKE; BAESENS, 2014).

Este estudo visa entender o perfil dos clientes migrados de assinatura impressa para assinatura digital, isto é, assinantes que antes recebiam o jornal impresso em sua casa ou trabalho e que optaram pelo produto digital da assinatura (acesso ilimitado ao site, aplicativos e jornal digital), ativos ou não na carteira, tendo em vista variáveis de perfil, relacionais e demográficas, de forma a analisar os desdobramentos das ações de migração na carteira a longo prazo. A migração do produto impresso para o digital vem ocorrendo com cada vez mais frequência e faz parte de uma estratégia da empresa, para redução de gastos com logística e maior aderência da mesma ao mercado digital. Em um primeiro momento a retenção dos assinantes migrados e manutenção dos mesmos como ativos na base se mostra relevante para a empresa. É necessário reduzir ao máximo o *churn* desse público; para tanto, é importante compreender e, principalmente, prever fatores que possam levar esse público a cancelar a assinatura após a migração. Neste sentido, a pergunta que esta pesquisa propõe é: qual o perfil de assinantes que tem a maior probabilidade de *churn* após migrar do produto impresso para a assinatura digital?

A utilização do marketing de relacionamento – aliado ao uso de dados – para prever o comportamento de clientes e, a partir disso, promover campanhas e desenvolver estratégias mostra-se relevante, principalmente levando em consideração o atual contexto mercadológico, no qual empresas de diversos setores percebem sua atuação em mercados saturados. Tendo em vista esse contexto, a aquisição de novos clientes torna-se mais difícil; em contrapartida, a retenção da carteira de clientes mostra-se cada vez mais fundamental para as empresas.

Portanto, este estudo apresenta relevância, pois se utiliza de informações já obtidas pela empresa para entender melhor o comportamento do assinante migrado do impresso para o digital. Assim, é possível criar um modelo que possibilite a antecipação da ocorrência do cancelamento, permitindo a ação por parte da empresa por meio de campanhas de retenção.

Esta pesquisa se aprofundará em temas relevantes para a construção do modelo preditivo: o conceito e gestão do *churn* e a escolha do método para a

construção do modelo. Posteriormente, abordará a escolha das variáveis importantes para a predição do cancelamento, o desenvolvimento do modelo, resultados, validação, limitações e conclusão.

1 OBJETIVOS

1.1 OBJETIVO GERAL

Elaborar um modelo de predição de *churn* para os clientes que migraram de assinatura de jornal impresso para assinatura digital.

1.2 OBJETIVOS ESPECÍFICOS

Entender o perfil dos clientes migrados ativos e cancelados.

Analisar variáveis comportamentais e demográficas dos clientes migrados ativos e cancelados.

Testar, validar e comparar dois métodos para construção de um modelo preditivo de *churn* para clientes migrados ativos e cancelados.

2 REFERENCIAL TEÓRICO

Este capítulo, com o intuito de fundamentar teoricamente a pesquisa, apresentará os conceitos envolvidos na composição do estudo, para que seja viável determinar os argumentos que a estruturam. Com o propósito de atender ao problema de pesquisa e aos objetivos do estudo, o capítulo será dividido em duas subseções. A primeira subseção abordará o conceito de *churn*; a segunda, o modelo preditivo.

2.1 *CHURN*

No mercado competitivo atual, muitas empresas percebem a importância de uma estratégia de negócios orientada para o cliente na manutenção da vantagem competitiva e de um nível de lucro estável. As empresas baseiam-se, principalmente, no rendimento que vem da sua base de clientes. No entanto, reter e manter clientes na carteira é uma tarefa difícil e dispendiosa para o marketing.

Dados os altos custos na aquisição de novos clientes – com a criação de novas contas, propagandas e promoções –, muitas vezes os custos relacionados à aquisição de novos clientes superam os custos de retenção (KEAVENEY, 1995). Em vista disso, torna-se uma crença de toda a indústria que a melhor estratégia de marketing para o futuro é a manutenção dos clientes existentes – aliado ao combate ao *churn* (KIM; YOON, 2004).

De forma alarmante, Ahmad e Buttle (2002) observam que a maior proporção de perda de clientes advém da leva mais recente de clientes. Portanto, a maioria das organizações enfrenta a perspectiva de ter de adquirir um número maior de clientes, pois já conta com a perda de parte dos clientes adquiridos. A realidade é que muitos destes têm alta propensão a cancelarem logo após a aquisição. O *churn*, dessa forma, é um problema, não só pela falta de envolvimento entre cliente e empresa, mas também em termos práticos, dado o elevado custo de aquisição de clientes contra os custos de retenção (REICHHELD; SASSER, 1990).

O fenômeno do *churn* de consumidores é mais comumente observado nos mercados de serviços ao consumidor voláteis, tais como telefones móveis (ARCHAUX et al., 2004; WEI; CHIU, 2002), seguro (MORIK; KOPCKE, 2004), serviços de assinatura (COUSSMENT; DEN; POEL, 2009) e banca (LARIVIERE; POEL, 2004). Um cliente é propenso a cancelar o serviço (ou produto) quando ele/ela interrompe o

uso desse tal serviço, embora possa continuar a usar outros produtos ou serviços. O *churn* dos consumidores é um problema sério para muitas empresas, por diversas razões: a aquisição de um novo cliente é muito mais difícil e mais cara; processos relacionados ao término de um serviço para o consumidor são elevados; a perda de um cliente leva à perda de receita, bem como impacto negativo sobre a linha de fundo; a perda de clientes afeta negativamente muitas funções dentro de uma organização, pois clientes cancelados podem afetar o valor percebido da marca, além de poderem influenciar clientes prospectados.

Identifica-se, na literatura, dois tipos de *churn*: o involuntário e o voluntário. O *churn* involuntário acontece, por exemplo, quando um cliente para de pagar o serviço comprado e tem o fornecimento cancelado pela empresa (HADDEN et al., 2005). Cister (2005), explica que as razões involuntárias são consequência de uma ação da própria empresa, que por algum motivo (ex.: fraude, falta de pagamentos, falta de utilização do serviço) viu-se obrigada a romper sua relação com o cliente. Já o *churn* voluntário ocorre quando, de forma racional, o cliente resolve romper com uma companhia e não mais utilizar seus serviços (HADDEN et al., 2005). É dividido em dois tipos: acidental e o deliberado.

No *churn* acidental, inúmeras situações fazem com que o cliente não seja capaz de manter o serviço (ex.: desemprego, mudança para uma região onde o serviço não é ofertado); entretanto, reflete uma pequena parcela da taxa de ruptura das organizações. Já o *churn* deliberado acontece em situações em que o cliente opta por mudar de fornecedor (HADDEN et al., 2005). Para este tipo, as causas mais comuns são encontradas em fatores associados às relações entre empresa e cliente e que podem ser administrados pela própria empresa (por exemplo, clareza de faturação e serviço pós-venda).

As estratégias e os esforços de uma empresa para manter seus clientes mais lucrativos recebem o nome de gestão de *churn* (HUNG; YEN; WANG, 2006). Tais estratégias são frequentemente apresentadas dentro de duas categorias: as não dirigidas (*untargeted*) e as dirigidas (*targeted*). As estratégias não dirigidas sustentam-se na publicidade em massa, com o objetivo de ampliar os níveis de lealdade da marca, enquanto as estratégias dirigidas empreendem esforços em clientes que possuem maior probabilidade de romper com a empresa. Por meio das estratégias dirigidas, a empresa enfatiza clientes com maior risco de abandoná-la e oferece

incentivos para a não efetivação da ruptura (BUREZ; POEL, 2008). As organizações buscam, com mais vigor, combater esse tipo de *churn* (HADDEN et al., 2005).

Burez e Poel (2008) indicam que existem dois tipos de abordagens direcionadas para a gestão do *churn* de clientes: as reativas e as proativas. Quando uma empresa adota uma abordagem reativa, ela aguarda até que os clientes peçam à empresa para cancelar a sua relação de serviço. Nesta situação, a empresa vai oferecer ao cliente um incentivo para ficar. Por outro lado, quando uma empresa adota uma abordagem proativa, tenta identificar os clientes que estão propensos a cancelar antes que o cancelamento ocorra. A empresa, então, oferece incentivos para estes clientes, objetivando a permanência na carteira. Estratégias proativas e direcionadas são potencialmente vantajosas para as empresas, por proporcionarem custos menores de incentivo aos clientes. No entanto, a lógica proativa só apresenta eficiência se as previsões de *churn* forem acuradas e precisas, já que os incentivos só devem ser oferecidos a clientes com alta propensão de cancelar (TSAI; LU, 2009).

A abordagem proativa, portanto, é uma ferramenta relevante para as empresas na retenção e relacionamento com seus clientes. Os modelos de *churn* visam identificar sinais de *churn* precoce e reconhecer clientes com maior probabilidade de sair voluntariamente (VAFEIADIS et al., 2015). Para tanto, é importante construir um modelo de previsão de *churn* de clientes o mais exato possível (BUREZ; POEL, 2008; LARIVIERE; POEL, 2004).

2.2 MODELO PREDITIVO

O *churn* de clientes é um problema notório para a maioria das empresas, pois a perda de um cliente afeta a imagem de marca das receitas, tornando a aquisição de novos clientes difícil. Modelos preditivos confiáveis para a rotatividade de clientes podem ser úteis na concepção de planos de retenção.

Do ponto de vista analítico, gestão de *churn* consiste em (1) prever quais clientes estão propensos a cancelar e (2) avaliar qual ação é mais eficaz em manter esses clientes (HUNG; IENES; WANG, 2006). Estes estudos de previsão de *churn* geralmente usam duas estratégias para melhorar o desempenho do modelo: uma estratégia baseada em algoritmo e uma estratégia de base de dados (BAECKE; POEL, 2010). A primeira consiste na avaliação de vários algoritmos em dados

fornecidos e na melhoria ou invenção de algoritmos. A última consiste em aumentar a base de dados existente com novas fontes de dados.

Modelos de previsão de *churn* de clientes buscam detectar clientes com alta propensão de cancelamento. A acurácia preditiva, compreensibilidade e justificabilidade são os três aspectos fundamentais de um modelo de previsão de *churn*. Um modelo preciso permite o direcionamento correto de *churners* futuros em uma campanha de marketing de retenção, enquanto um conjunto de regras compreensíveis e intuitivas permite identificar os principais fatores responsáveis pelo *churn* dos consumidores, além de desenvolver uma estratégia de retenção eficaz, em conformidade com o conhecimento da área (VERBEKE et al., 2011).

O formato de previsão refere-se à aprendizagem da relação entre os dados que são observados em um período (ou janela) que termina antes de um determinado ponto no tempo e os dados que são observados em um período que se inicia após este mesmo ponto no tempo. O período anterior é chamado independente, preditor ou período de motivos, ou histórico de eventos do cliente; o último é chamado período dependente ou resposta. Entre esses dois, pode haver uma lacuna, chamada às vezes período de retenção, que se destina a servir como um período operacional ou de prática para organizar as ações reais de retenção e relacionamento antes que os clientes apresentem o comportamento de resposta focal (WEI; CHIU, 2002).

Existem diversas técnicas que podem ser utilizadas na construção de um modelo de *churn* preditivo. As técnicas escolhidas para a construção do quadro foram: árvore de decisão (HUNG; YEN; WANG, 2006; VERBEKE et al., 2011; WEI; CHIU, 2002; XIE et al., 2009); random forest (BUREZ; POEL, 2008); análise de cluster (QIAN; JIANG; TSUI, 2006); redes neurais (XIE et al., 2009; YU et al., 2011); support vector machine (COUSSEMENT; BENOIT; POEL, 2009); regressão logística (BOTELHO; TOSTES, 2010; COUSSEMENT; BENOIT; POEL, 2009; GUANGLI et al., 2011); e redes bayesianas (TSAI; LU, 2009).

Tabela 1 – Síntese das técnicas

Método	Definição	Ponto positivo	Ponto negativo
Árvore de Decisão	Com base nos registros do conjunto de treinamento, uma árvore é montada; a partir dela, pode-se classificar a amostra desconhecida sem necessariamente testar todos os valores dos seus atributos. Toda informação sobre cada objeto (caso) a ser classificado deve poder ser expressa em termos de uma coleção fixa de propriedades ou atributos. (YU et al., 2010)	Rápido uso computacional. Força de interpretabilidade (ZHAO; ZHANG, 2008)	A topologia e a qualidade da árvore estão diretamente ligados à correta escolha do algoritmo. Um mesmo conjunto de dados pode gerar distintas árvores.
Randon Forest	Combinação de preditores de árvores de modo que cada árvore depende dos valores de um vetor aleatório amostrado independentemente e com a mesma distribuição para todas as árvores na floresta (BREIMAN, 2001).	Alta precisão de classificação. Um novo método para determinar a importância variável. Capacidade de modelar interações complexas entre variáveis preditoras. Flexibilidade para realizar vários tipos de análise de dados estatísticos, incluindo regressão, classificação, análise de sobrevivência e aprendizagem não supervisionada; um algoritmo para imputar valores faltantes (JING et al., 2017).	O erro de generalização para as florestas converge em um limite quando o número de árvores na floresta torna-se muito grande (LIN et al., 2017).
Análise de Cluster	Na análise de agrupamentos (*cluster analysis*) a similaridade entre duas amostras pode ser expressa como uma função da distância entre os dois pontos representativos destas amostras no espaço n-dimensional. Quanto menor a distância entre os pontos, maior a semelhança entre as amostras.	Permite a redução da dimensionalidade dos pontos representativos das amostras pois, embora a informação estatística presente nas n-variáveis originais seja a mesma dos componentes principais, é comum obter em apenas 2 ou 3 das primeiras componentes principais mais que 90% desta informação.	Exige o mínimo de conhecimento para determinação dos parâmetros de entrada. Em geral, algoritmo algum atende a todos os requisitos para uma análise de cluster criteriosa e, por isso, é importante entender as características de cada algoritmo para a escolha de um método adequado a cada tipo de dado ou problema.

Continuação da Tabela 1

Método	Definição	Ponto positivo	Ponto negativo
Redes Neurais	A utilização de dados na criação de bases de conhecimento é o conceito-chave das redes neurais. Em vez de trabalhar com regras explícitas como os sistemas especialistas, as redes neurais utilizam critérios mais complexos e implícitos, baseados no aprendizado a partir de exemplos (TSAI; LU, 2009).	Aquisição de conhecimentos feita automaticamente a partir de exemplos coletados em bancos de dados.	Não trabalha com regras.
Support Vector Machine	Método muito efetivo para reconhecimento de padrões de propósito geral. Intuitivamente, o SVM aprende o limite entre as amostras pertencentes a duas classes, mapeando as amostras de entrada em um espaço de alta dimensão e buscando um hiperplano de separação. As máquinas de vetores de suporte (SVMs) usam um modelo linear para implementar limites de classes não-lineares através de vetores de entrada de mapeamento não-lineares em um espaço de grande dimensão (SHMILOVICI, 2005).	Modelo simples de ser analisado matematicamente. SVM pode servir como uma alternativa promissora, combinando os pontos fortes dos métodos estatísticos convencionais que são mais orientados pela teoria e fáceis de analisar, e mais métodos de aprendizagem mecânica, livres de distribuição e robustos. 1) existem apenas dois parâmetros livres a serem escolhidos, nomeadamente o limite superior eo parâmetro do kernel; (2) a solução de SVM é única, ótima e global, uma vez que o treinamento de uma SVM é feito resolvendo um problema quadrático linearmente constrangido; (3) As SVM são baseadas no princípio de minimização do risco estrutural (SRM).	Necessidade de transformação de dados que não são linearmente separáveis. Não há métodos para seleção da função kernel de transformação dos dados que não são linearmente separáveis, então a escolha se baseia em informações e tentativas empíricas.

Continuação da Tabela 1

Método	Definição	Ponto positivo	Ponto negativo
Regressão Logística	Relaciona um conjunto de variáveis independentes com uma variável dependente categórica (GUANGLI et al., 2011).	Método padrão para análise de variáveis dicotômicas. Não é necessário supor normalidade multivariada. É uma técnica mais genérica e mais robusta, pois sua aplicação é apropriada em grande variedade de situações;	Se faz necessário aplicar algum tipo de transformação para que o modelo obtenha as desejáveis propriedades dos modelos lineares. Na construção do modelo, algumas variáveis preditoras podem sofrer com a multicolinearidade; isto acontece quando, duas ou mais variáveis são correlacionadas e o uso delas no modelo pode prejudicar o processo de modelagem.
Redes Bayesianas	Rede Bayesiana é uma representação compacta de uma tabela de conjunção de probabilidades. Modelo gráfico que representa de forma simples as relações de causalidade das variáveis de um sistema (MENÊZES; FIRMINO; DROGUETT, 2005).	Representam tabelas conjunção de probabilidades de um domínio de forma compacta.	Para se construir uma rede cuja estrutura represente devidamente o domínio do problema, é necessário que para todo nó da rede esta propriedade seja atendida. A qualidade está diretamente ligada à escolha de pais, nós e a ordem destes.

Fonte: A autora (2017).

3 MÉTODO

3.1 TIPO DE PESQUISA

O problema de pesquisa proposto nesse estudo é a identificação do comportamento de clientes de uma assinatura de jornal, buscando verificar quais fatores podem fazê-lo cancelar a assinatura após a decisão pela troca do produto impresso para o produto digital. A partir dos fatores que influenciam a decisão do assinante, um modelo preditivo será construído, procurando prever o cancelamento da assinatura e possibilitando a ação da empresa, por meio de ações de marketing e relacionamento que tentem evitar esse cancelamento.

Por se tratar de uma modelagem de análise multivariada, que se propõe a identificar variáveis independentes que afetam a variável dependente a ser explicada – o cancelamento da assinatura –, os métodos utilizados para a construção do modelo foram a regressão logística e a árvore de decisão, que foram construídos de acordo com os estágios determinados por Hair et al. (2005), conforme ilustra a Tabela 2:

Tabela 2 - Métodos utilizados

Estágio Hair	Método: Regressão Logística	Método: Árvore de Decisão
Objetivos	Capítulo 2	Capítulo 2
Desenvolvimento do plano de análise	Item 4.2	Item 4.2
Pressupostos	Item 4.3	Item 4.4
Estimação do modelo	Item 5.1	Item 5.4
Interpretação dos Resultados	Item 5.2	Item 5.5
Validação dos Resultados	Item 5.3	Item 5.6

Fonte: A autora (2017).

3.2 DESENVOLVIMENTO DO PLANO DE ANÁLISE

Com a finalidade de atingir o objetivo proposto, que é classificar a propensão a cancelar de clientes (assinantes) migrados do produto impresso para produtos digitais, torna-se importante o conhecimento das variáveis presentes no banco de dados da empresa e a definição de um método adequado para o tratamento das variáveis e construção do modelo. Para abordar o fenômeno objeto deste estudo adotou-se uma abordagem substantiva, que implica o estudo de um determinado fenômeno em um contexto específico, ou seja, o estudo da propensão do cliente migrado do produto

impresso para o digital ao cancelamento de sua assinatura. Essa abordagem foi escolhida por permitir a explicação do fenômeno em um contexto concreto, possibilitando a previsão de fatos e proporcionando *insights* para políticas de ação.

Os procedimentos de modelagem – desenvolvimento e teste do modelo preditivo – foram realizados com dados primários de uma empresa jornalística do estado do Rio Grande do Sul. A variável dependente do modelo, isto é, o efeito observado como resultado da manipulação da variável independente, – propensão ao cancelamento da assinatura – será operacionalizada como uma variável dicotômica, ou seja, vai cancelar ou não vai cancelar. Já o conjunto de variáveis independentes, as condições ou causas para um determinado efeito ou consequência do modelo serão selecionadas de forma a apresentarem relação de causalidade com a propensão a *churn*, dada sua classificação.

Para o estudo, foram escolhidas variáveis relacionadas ao perfil da carteira de cliente, que consistem em onze características que foram avaliadas quanto a sua capacidade de predizer o cancelamento. O volume da amostra é 70% para a análise e 30% para aplicação do teste, com intuito de validar o modelo. A base utilizada tem um volume de 11.000 assinaturas, entre ativos e cancelados. A escolha final das onze variáveis que participariam da construção do modelo foi baseada na conversa com gestores de áreas chave da empresa, como *database* marketing e retenção, e no uso dessas variáveis para ações internas visando à retenção e o relacionamento com os assinantes.

Dado o propósito de explicar o cancelamento de migrados para o produto digital através de variáveis de comportamento e perfil, foram escolhidos dois métodos para a construção do modelo preditivo: a regressão logística e a árvore de decisão. A escolha da regressão logística deve-se à praticidade de aplicação do método e à capacidade de explicação da variável dependente. Segundo Hair et al. (2005), o uso da regressão logística é adequado quando se quer explicar um acontecimento através de apenas uma variável dependente – nesse caso o cancelamento – e diversas variáveis independentes e não métricas – no caso, as 12 variáveis escolhidas.

Já a escolha da árvore de decisão, deve-se a quatro motivos: facilidade de entendimento do método, fácil conversão em um conjunto de regras de produção, possibilidade de classificação de dados categóricos e numéricos – contanto que o atributo *output* seja categórico – e possibilidade de não haver pressupostos a priori sobre a natureza dos dados (ZHAO; ZHANG, 2008) Apesar de não ter um ótimo

desempenho na captura de relações complexas e não-lineares entre os atributos, no problema de *churn* dos clientes, a precisão de um DT pode ser alta, dependendo da forma dos dados (VAFEIADIS et al., 2015). Além disso, entre as técnicas de mineração de dados, a árvore de decisão é um dos métodos mais amplamente utilizados para a construção de modelos de classificação no mundo real por sua simplicidade e facilidade de interpretação (KIM, 2016).

A construção do modelo por meio de dois métodos possibilita, também, a comparação entre ambos e a identificação de fatores que favoreçam um e outro, como a interpretação de resultados e a eficácia na predição do cancelamento.

3.3 PRESSUPOSTOS DA REGRESSÃO LOGÍSTICA

A regressão logística é um método de associação de variáveis no qual se prediz a presença ou ausência de uma característica (no caso, o cancelamento da assinatura) por meio de um conjunto de variáveis preditoras ou explicativas (SOARES; SIQUEIRA, 1999). Este método se enquadra na classe de métodos estatísticos multivariados de dependência, pois relaciona um conjunto de variáveis independentes com uma variável dependente categórica (HAIR et al., 2005).

Esse método é aplicado para correlacionar modificações em variáveis independentes (chamadas "preditores") para o efeito consequente nas variáveis dependentes. O resultado da análise de regressão é uma função das variáveis independentes – função de regressão -, que é usada para prever os valores das variáveis dependentes (MOROTTI; GRANDI, 2017).

Na configuração da regressão logística, a resposta é binária e segue uma distribuição binômica em vez de normal. A resposta é transformada em uma probabilidade relativa de log ou escala logit antes de executar a regressão (BRIMACOMBE, 2016).

Trata-se de um problema multivariável, no qual é calculada a medida de associação de cada variável com o acontecimento que se deseja explicar. Na formulação do modelo preditivo foram consideradas variáveis independentes do perfil da carteira de assinantes, tais como gênero, idade, tempo de carteira, forma de pagamento, além de dados referentes ao uso do produto digital.

A variável cancelamento representa uma única relação multivariada com coeficientes de regressão que indicam o impacto relativo de cada variável preditora,

composta dos dados da base de clientes do jornal. A regressão logística prediz diretamente a probabilidade de ocorrência de um evento e os valores de probabilidade podem ser qualquer valor entre 0 e 1, mas o valor previsto deve ser limitado para cair dentro do intervalo de 0 e 1.

O método de regressão logística é comumente utilizado para analisar a probabilidade de ocorrência de variáveis binárias; isto é conseguido ajustando as probabilidades de log e as variáveis explicativas a um modelo linear (James et al., 2013), onde Y = (0, 1) é a variável binária, X = (X1, ..., Xn) são "n" variáveis explicativas, e β = (β0, ..., βn) são os coeficientes de regressão a serem estimados com base nos dados (ELÍO et al., 2017) . Nesse caso, de um modelo de previsão de *churn*, Y = 1, se o cliente não cancelar a assinatura; Y = 0, caso contrário.

A função é interpretada como a probabilidade de o cliente não cancelar a assinatura de acordo com as características que possui. Toda essa análise pode ser feita também como Y = 1 (se o cliente cancelar a assinatura; Y = 0, caso contrário), já que a interpretação é a mesma (BOTELHO; TOSTES, 2010).

A medida geral de quão bem o modelo se encaixa é dada pelo valor de verossimilhança. Um modelo bem ajustado terá um valor pequeno para -2LL. O teste qui-quadrado para a redução do valor de verossimilhança log fornece uma medida de melhora devido à introdução das variáveis independentes. Em seguida, o número de eventos factuais e preditos é comparado em cada classe com a estatística do qui-quadrado. Este teste fornece uma medida abrangente de precisão preditiva que não se baseia no valor de verossimilhança, mas na previsão real da variável dependente.

Uma das vantagens da regressão logística é que precisamos saber apenas se um evento ocorreu para, então, usar um valor dicotômico como nossa variável dependente. A partir deste valor dicotômico, o procedimento prediz sua estimativa da probabilidade de que o evento ocorrerá ou não. Se a probabilidade prevista é maior do que 50, então a predição é sim; caso contrário, não. O procedimento que calcula o coeficiente logístico compara a probabilidade de ocorrência de um evento com a probabilidade de não ocorrer. A regressão logística tem a vantagem de ser menos afetada quando os pressupostos básicos, particularmente a normalidade das variáveis, não são atendidos.

3.4 PRESSUPOSTOS DA ÁRVORE DE DECISÃO

A Árvore de Decisão (DT – Decision Tree) é um algoritmo de classificação supervisionada em que um dos resultados consiste em um conjunto de regras encadeadas do tipo “SE ENTÃO”, que formam uma estrutura hierárquica semelhante à de uma árvore. Trabalha em cima de uma base de dados e determina a classe, com base nos atributos de entrada. Usa uma estrutura de árvore de tipo fluxograma para segregar um conjunto de dados em várias classes predefinidas, fornecendo, assim, a descrição, categorização e generalização de conjuntos de dados dados (YU et al., 2010).

A árvore de decisão é um modelo representado graficamente por nós e ramos (WITTEN; FRANK, 2011). A indução tradicional de árvores de decisão emprega medidas heurísticas com base no espaço em perspectiva de atributos para selecionar um atributo de divisão ideal para a partição do nó de decisão para obter uma árvore melhorada (SUN; HU, 2017) .

A construção de um DT baseia-se principalmente em duas etapas: (1) uma fase de crescimento e (2) uma fase de poda (KARABADJI et al., 2017). Durante o processo de construção da árvore de classificação, as amostras em cada nó interior são divididas em subconjuntos com base em um atributo e este processo é repetido em cada subconjunto derivado de forma recursiva. A recursão é concluída quando um subconjunto em um nó tem o mesmo valor alvo, quando a divisão não melhora a previsão, ou quando a divisão é impossível devido a restrições definidas pelo usuário (KIM, 2016).

Após o crescimento da árvore cheia, as classes de saída das amostras são determinadas nos nós terminais. Cada nó terminal decide sobre um valor de destino com base na classe majoritária no nó terminal. Quando uma nova amostra é observada, ela é classificada como um dos nós terminais da árvore, dependendo das variáveis de entrada. Então, o alvo está previsto para ser a classe majoritária no nó da folha (KIM, 2016).

Uma árvore de decisão é uma árvore enraizada onde cada nó interno está associado a um teste e cada folha está associada a uma classe. Os ramos que saem de um nó interno estão associados aos possíveis resultados do teste correspondente ao nó (SAETTLER; LABER; PEREIRA, 2017).

A DT tem como objetivo tentar minimizar erros de classificação e o número de nós e folhas. Possui a vantagem de exibir o conjunto de regras gerado no treinamento, diferente de outros classificadores. Uma DT pode ser desenvolvida para trabalhar com dados nominais ou numéricos. Em ambos os casos, o seu algoritmo gera as regras baseando-se na análise de cada atributo, através dos seus valores e da sua relação com os demais parâmetros evolvidos. No caso de dados numéricos, os valores utilizados são discretizados (WITTEN; FRANK, 2011).

As árvores de decisão ganharam popularidade devido à facilidade de interpretação das regras descobertas. Árvore de decisão é uma técnica bem conhecida e teve muitas aplicações bem-sucedidas para problemas do mundo real (TSAI; LU, 2009).

4 ANÁLISE

4.1 ESTIMAÇÃO DO MODELO DE REGRESSÃO LOGÍSTICA

A base utilizada para a construção do modelo é composta 11.000 assinaturas, e foi dividida em treino – composta por 70% da base, e teste – que compreende os 30% restantes. Para a construção do modelo de regressão logística foram selecionadas onze variáveis independentes, além da *target*, que é o status da assinatura (ativa ou cancelada). As variáveis escolhidas para a estruturação do modelo seguem conforme a *codesheet* abaixo:

Tabela 3 - Codesheet

Variável	Significado da variável	Codificação na base	Tipo de escala
Canal de venda			
Tlmktint – Telemarketing interno	Canal de entrada do assinante na carteira	1	Discreta
Internet		2	
Centurion		3	
Explorer		4	
Outros – canais receptivos de vendas		5	
Forma pgto			
DEB CONTA	Forma de pagamento do assinante (registro mais recente)	1	Discreta
Cartaocredito		2	
FAT		3	
DOC		4	
DOC EMAIL		5	
CARNE		6	
Tempo de base			
	Tempo em meses do assinante na carteira do jornal		Contínua
Qtd. Incidências cobrança			
0	Quantidade de entradas em cobrança até fevereiro 2017	0	Discreta
1		1	
2		2	
3		3	
Gênero			
J	Gênero do assinante (J = pessoa jurídica)	1	Discreta
F		2	
M		3	
Idade			
	Idade do assinante		Contínua

Continuação da Tabela 3

Variável	Significado da variável	Codificação na base	Tipo de escala
Renda			
ATE 3SM	Renda estimada do assinante baseada no bairro de residência cadastrado em sua assinatura (SM – salários mínimos)	1	Discreta
DE 3SM ATE 4SM		2	
DE 4SM ATE 8SM		3	
DE 8SM ATE 14SM		4	
DE 14SM ATE 25SM		5	
ACIMA DE 25SM		6	
Sem informação		7	
Possui login acesso			
NÃO	Se possui login para acesso ao conteúdo digital	0	Discreta
SIM		1	
Navegação			
	Diferença de dias da última navegação até 09/05		Contínua
Região			
REGIAO1	Região de residência do assinante – região 1 (POA e proximidades), região 2 (campanha, extremo norte e sul).	1	Discreta
REGIAO2		2	
PR/SC/Correio		3	
Situação clube			
Sem clube	Situação do clube do assinante	1	Discreta
Cancelado		2	
Ativo		3	
Abonado		4	
Target			
ATIVO	Situação da assinatura	0	Discreta
CANCELADO		1	

Fonte: A autora (2017).

Para a escolha das variáveis foi determinante sua relevância no negócio, já que o objetivo dessa pesquisa é explicar o cancelamento de clientes migrados de modalidade impressa para modalidade digital. As variáveis escolhidas para a construção do modelo são utilizadas internamente na empresa para decisões de estratégia das áreas de uso e retenção, além de realização de estudos de carteira. Para tanto, foram escolhidas todas as variáveis comportamentais que constam no banco de dados, além de variáveis demográficas, que traçam o perfil da carteira do produto.

Atributos como tempo de base – que compreende a quanto tempo determinado assinante possui a assinatura – demonstram a valorização e o vínculo do cliente com o produto assinatura; a idade auxilia a identificar o público-alvo para cada modalidade

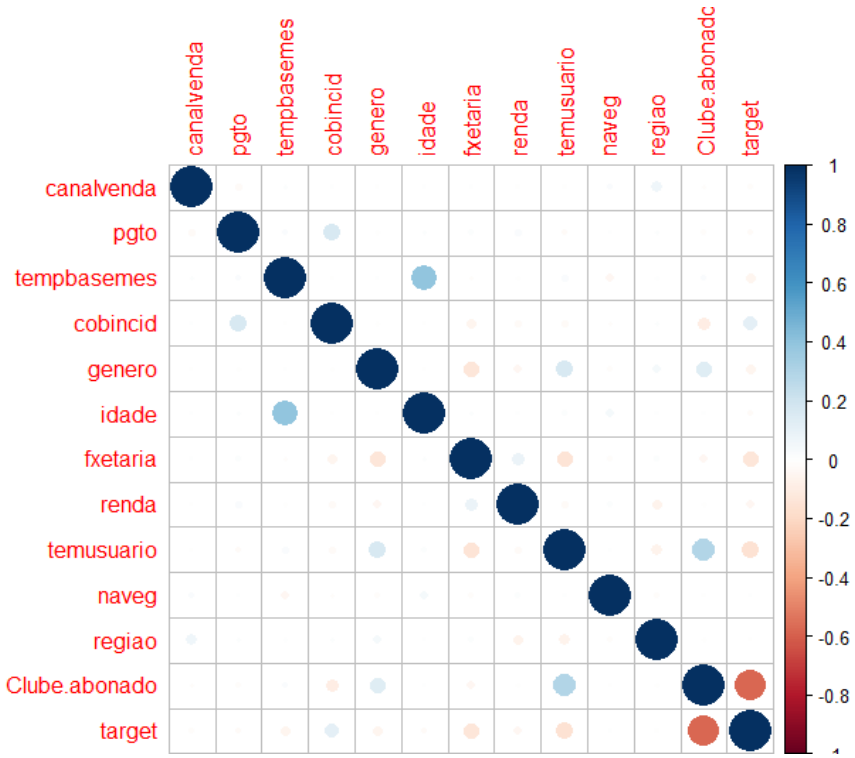
de assinatura (digital ou impressa), e as variáveis relacionadas diretamente com o digital – “possuir login” e navegação – ajudam na identificação da familiaridade da carteira com o ambiente digital, que é um fator relevante para o cancelamento. A variável navegação identifica a última data na qual o assinante se logou na plataforma digital do jornal. Já a quantidade de incidências em cobrança é relevante para a construção do modelo, pois a maior quantidade de incidências em cobrança pode significar previamente a intenção de cancelamento ou uma menor aderência e vínculo ao produto. A importância da variável “clube do assinante” está, mais uma vez, na agregação de valor ao produto, pois pode ser vista como um benefício além da assinatura.

Já a variável canal de venda – que se divide em telemarketing interno (canal ativo interno de vendas), centurion (venda ativa direta terceirizada), internet, explorer (empresa terceirizada de venda ativa) e outros (canais receptivos de vendas) - é importante, pois é a entrada do assinante na carteira; caso essa venda não tenha sido bem trabalhada, pode causar o descontentamento com o produto. Variáveis como renda, região, gênero e forma de pagamento, além de esboçarem o perfil da carteira, podem ser importantes para separarem comportamentos de cancelamento.

O primeiro procedimento realizado na base selecionada foi a verificação da multicolinearidade entre as variáveis escolhidas. A multicolinearidade em regressão é uma condição que ocorre quando algumas variáveis preditoras no modelo estão correlacionadas a outras variáveis preditoras; esse teste permite examinar a probabilidade estatística da existência de correlações significativas entre pelo menos algumas variáveis. Quando é forte mostra-se problemática, porque pode aumentar a variância dos coeficientes de regressão, tornando-os instáveis, pois este fator pode afetar o resultado da regressão logística.

O teste para multicolinearidade realizado para a base teve como resultado a seguinte matriz de correlação, através do coeficiente de Pearson:

Figura 1 – Coeficiente de Pearson



Fonte: A autora (2017).

A matriz de correlação apresenta as relações positivas e negativas das variáveis em relação entre si e com a target; nesse caso, o cancelamento da assinatura. Em azul estão representadas as correlações negativas e em vermelho as positivas. As correlações procuradas através da construção da matriz são as em vermelho, que indicam a relação da variável independente com a variável dependente.

Segundo Hair et al. (2005), é possível categorizar os resultados de significância da matriz usando a seguinte regra: ± 0.30 = mínimo, ± 0.40 = importante e ± .50 = praticamente significativo. De acordo com essa categorização, é possível afirmar que a variável clube abonado possui a maior significância, ou seja, tem a maior correlação com o cancelamento da assinatura. Além da variável clube, somente a variável “tem usuário” mostrou o mínimo de correlação, com significância de 0,3.

Após a realização dos testes para verificação de correlação e significância das variáveis da base, o modelo de regressão logística foi plotado no programa R, gerando resultados conforme a Tabela 5:

Tabela 5 – Modelo de regressão logística

	Estimate – LOG ODDS	Erro	z value	P – valor
(Intercept)	657,2	5,62	1,170	0,24201
Idade	-2,03	3,46	-0,585	0,55843
Gênero: mulher	1899	2,58	7.365	<0,01 ***
Gênero: homem	1537	2,56	6,013	<0,01 ***
3 a 4 salários mínimos	-690,4	4,90	-1,410	0,15842
4 a 8 salários mínimos	-1044	4,86	-2,148	<0,01 *
8 a 14 salários mínimos	-960	4,89	-1,963	<0,01 *
14 a 25 salários mínimos	-1383	4,94	-2,798	<0,01 **
Acima de 25 salários mínimos	-1553	5,04	-3,079	<0,01 **
Sem informação de renda	-936,9	4,96	-1,888	0,05909
Tempo ativo na base (em meses)	-1,48	5,35	-2,765	<0,01 **
Canal de venda internet	282,30	1,67	1,695	<0,01
Canal de venda centurion	173,40	1,30	1,334	0,18223
Canal de venda Explorer	-39,07	2,95	-0,132	0,89470
Outros canais de venda	-287,7	1,12	-2,571	<0,01 *
Forma de pgto: cartão de crédito	-101,4	9,20	-1,102	0,27061
Forma de pgto: Faturada	-1907	2,67	-0,007	0,99430
Forma de pgto: DOC	-689,9	9,14	-0,754	0,45055
Forma de pgto: DOC EMAIL	564,6	8,89	0,635	0,52552
Forma de pgto: CARNE	18550	1,08	0,002	0,99862
Incidências em cobrança	473,60	9,06	5,229	<0,01 ***
Possuir usuário de acesso	-730,6	1,48	-4,927	<0,01 ***
Navegação	-0,01	4,80	-1,583	0,11342
Região 2: interior RS	-175,4	1,45	-1,212	0,22563
PR/SC/Outros estados	-828,7	3,83	-2,165	<0,01 *
Clube cancelado	397,4	9,66	4,115	<0,01 ***
Clube ativo	-20180	2,63	-0,077	0,93875
Clube abonado	-20200	4,40	-0,046	0,96337

Fonte: A autora (2017).

A coluna Pr (> | z |) mostra os p-valores de duas colunas testando a hipótese nula de que o coeficiente é igual a zero (isto é, nenhum efeito significativo). O valor usual é 0,05, e, de acordo com essa medida, nenhum dos coeficientes tem um efeito significativo sobre o log-*odds ratio* da variável dependente. O valor z também testa o nulo que o coeficiente é igual a zero. A coluna Estimativa mostra os coeficientes na forma *log-odds*, mostrando se o efeito dos preditores é positivo ou negativo. Já a coluna p-valor mostra a significância de cada variável na predição do *churn* – a presença do asterisco representa a significância da variável. O maior nível de significância corresponde à presença de três asteriscos.

As estimativas do modelo de uma regressão logística são as estimativas de máxima verossimilhança alcançadas através de um processo iterativo. Por isso, é importante avaliar a qualidade do modelo logístico construído. Esse procedimento foi feito por meio do cálculo da medida pseudo R2, já que é aplicável a um amplo conjunto de variáveis dependentes limitadas e qualitativas e oferece uma valiosa ferramenta para avaliação de modelos (LAITILA, 1993).

O pseudo R2 representa o quanto da variação da variável é explicada pela variável dependente (HAIR et al., 2005), com significância de 1%. O valor obtido para o modelo foi de 0,514 – o que revela que o modelo representa pelo menos metade da variação da variável independente. Além disso, também é possível avaliar os coeficientes e a constante das variáveis independentes geradas no modelo quanto à sua significância estatística. Tal avaliação é dada através do teste de Wald, que é calculado tomando a razão do quadrado do coeficiente de regressão para o quadrado do erro padrão do coeficiente. A idéia é testar a hipótese de que o coeficiente de uma variável independente no modelo é significativamente diferente de zero. Se o teste não rejeitar a hipótese nula, isso sugere que a remoção da variável do modelo não prejudicará substancialmente o ajuste desse modelo.

Segundo os resultados do teste de Wald, que podem ser vistos na Tabela 6, conclui-se que as variáveis com maior significância para a construção do modelo, devido à significância do p-valor (menor que 0,01), são gênero, renda, incidências em cobrança, tem usuário, clube, tempo de base e canal venda.

Tabela 6 – Teste de Wald

	Wald	**df**	**P-value**
Idade	F = 0,3424356	1	P= 0,55845
Genero	F = 31,23585	2	P <0,01
Renda	F = 6,189936	6	P<0,01
Tempbasemes	F = 7,644786	1	P <0,01
Canalvenda	F = 3,531988	4	P <0,01
Pgto	F = 0,4398725	5	P = 0,8209
Cobincid	F = 27,34192	1	P <0,01
Temusuario	F = 24,27593	1	P <0,01
Naveg	F = 2,505976	1	P= 0,11348
Regiao	F = 2,985868	2	P <0,01
Clube.abonado	F = 5,646476	3	P <0,01

Fonte: A autora (2017).

As variáveis gênero, clube, região, pagamento e canal de venda foram transformadas em fatores, para que o resultado do modelo pudesse considerar as categorias de cada variável e não somente a variável como um todo.

4.2 INTERPRETAÇÃO DOS RESULTADOS DE REGRESSÃO LOGÍSTICA

Segundos os resultados da tabela descrita no item acima, nem todas as variáveis escolhidas para a construção do modelo foram significativas, ou seja, apresentam relevância para explicar o *churn* de migrados do produto impresso para o digital. Os coeficientes de regressão logística aparecem na primeira coluna e representam a alteração nas probabilidades log do resultado do modelo para um aumento de uma unidade na variável preditora e são referentes à base treino, que corresponde à 70% do total da base.

Portanto, tendo em vista o significado desse coeficiente, é possível afirmar o nível de relação de cada variável do modelo no que tange à explicação do cancelamento. Os coeficientes são difíceis de interpretar na sua forma original porque são expressos em termos de logaritmos quando usamos o logit como a medida dependente. Sendo assim, foi calculado o coeficiente logístico exponencial, que consiste na transformação (antilog) do coeficiente logístico original. Dentre as onze variáveis selecionadas – faixa etária, gênero, tempo de base, canal de venda, forma de pagamento, incidências em cobrança, possui usuário acesso, navegação, região e clube –, nove representaram significância na explicação do cancelamento da assinatura, sendo essas gênero, tempo de base, renda, clube, incidências em cobrança, canal de venda, região, ter login. A Tabela 7 ilustra a transformação:

Tabela 7 - Transformação

	ODDS RATIO	Limite inferior do intervalo de confiança 2.5 %	Limite superior do intervalo de confiança 97.5 %
(Intercept)	1,93	6,80	6,33
Idade	9,98	9,91	1,00
Gênero: mulher	6,68	4,05	1,12
Gênero: homem	4,65	2,84	7,75
3 a 4 salários mínimos	5,01	1,71	1,20
4 a 8 salários mínimos	3,52	1,21	8,39
8 a 14 salários mínimos	3,83	1,31	9,19
14 a 25 salários mínimos	2,51	8,50	6,10
Acima de 25 salários mínimos	2,12	7,06	5,26
Sem informação de renda	3,92	1,32	9,57
Tempo ativo na base (em meses)	9,99	9,97	1,00
Canal de venda internet	1,33	9,63	1,85
Canal de venda centurion	1,19	9,24	1,54
Canal de venda Explorer	9,62	5,48	1,75
Outros canais de venda	7,50	6,03	9,35
Forma de pgto: cartão de crédito	9,04	7,55	1,08
Forma de pgto.: Faturada	5,23	NA	1,06
Forma de pgto: DOC	5,02	6,50	2,84
Forma de pgto.: DOC EMAIL	1,76	3,46	1,34
Forma de pgto.: CARNE	1,13	0.000000e+00	NA
Incidências em cobrança	1,61	1,34	1,92
Possuir usuário de acesso	4,82	3,58	6,41
Navegação	1,00	1,00	1,00
Região 2: interior RS	8,39	6,34	1,12
PR/SC/Outros estados	4,37	2,05	9,34
Clube cancelado	1,49	1,23	1,80
Clube ativo	1,72	1,51	8,72
Clube abonado	1,69	1,79	5,27

Fonte: A autora (2017).

Os coeficientes de regressão logística dão a alteração nas probabilidades log do resultado para um aumento de uma unidade na variável preditora. A significância das variáveis no cancelamento da assinatura acontece da seguinte forma, conforme o modelo: dentro da variável gênero, ser mulher aumenta em 6,68 vezes a probabilidade de cancelamento, enquanto ser homem aumenta essa probabilidade em 4,65 – a classificação pessoa jurídica não explica o cancelamento.

Em relação à variável renda, foram significativas quatro de sete classificações. As faixas de renda que explicam o cancelamento são de quatro a oito salários mínimos – pertencer a essa faixa de renda diminui a probabilidade de cancelamento em 3,52 – ; de oito a quatorze salários mínimos – essa faixa de renda diminui em 3,83 a probabilidade de cancelamento –; de 14 a 25 salários mínimos – a probabilidade de cancelamento nessa faixa de renda diminui em 2,51 –; e acima de 25 salários mínimos – que diminui a probabilidade de cancelamento em 2,12.

Já a variável tempo de base, que representa o tempo da assinatura em meses, explica a diminuição da probabilidade de cancelamento em 1,48 a cada aumento no tempo de base. A variável incidências em cobrança também se mostrou significativa, pois a cada aumento de entrada em cobrança a probabilidade de cancelamento aumenta 1,61 vezes.

A variável “tem usuário”, que expressa se o cliente possui acesso aos conteúdos digitais da assinatura, diminui a probabilidade de cancelamento em 4,82 se o cliente possuir acesso ao conteúdo. Já a variável região apresentou relevância no cancelamento apenas para a terceira classificação, ou seja, somente para clientes de outros estados (Paraná e Santa Catarina), e diminui em 4,37 caso o assinante pertença a essa região.

A variável canal de venda foi dividida em cinco classes; somente uma delas apresentou significância na explicação do cancelamento. Se a venda da assinatura foi realizada pelo canal “outros”, a probabilidade de cancelamento aumenta em 2,88. A última variável que apresentou significância no cancelamento foi a variável clube, que foi dividida em quatro classes: não possui clube, clube cancelado, clube ativo e clube abonado. Apenas a classificação clube cancelado se mostrou significativa e a probabilidade de cancelamento aumenta em 1,49 caso a assinatura tenha o clube cancelado.

4.3 VALIDAÇÃO DOS RESULTADOS DE REGRESSÃO LOGÍSTICA

A validação do modelo de regressão logística levou em consideração a base treino, que consiste nos 30% da base de assinantes. A validação tem como objetivo verificar a acurácia e precisão do método escolhido para a construção do modelo. Para a validação do modelo foi utilizada uma matriz de confusão. A matriz de confusão é um conceito da aprendizagem mecânica, que contém informações sobre

classificações reais e previstas feitas por um sistema de classificação. Uma matriz de confusão tem duas dimensões. Uma é indexada pela classe real de um objeto; a outra é indexada pela classe que o classificador prevê (DENG et al., 2016).

A matriz de confusão de um classificador indica o número de classificações corretas versus as previsões efetuadas para cada caso, sobre um conjunto de exemplos T. Uma matriz de confusão é uma matriz na qual a classe real ou um dado em teste é representado pela linha da matriz e a coluna da matriz de confusão representa a classificação desse dado particular. Os elementos diagonais indicam classificações corretas e, se a matriz não é normalizada, a soma da linha I é o número total de elementos da classe I que realmente apareceu no conjunto de dados.

Nesta matriz, as linhas representam os casos reais e as colunas as previsões efetuadas pelo modelo. O número de acertos para cada caso está indicado na diagonal principal da matriz. A matriz de confusão de um classificador ideal possui todos os restantes elementos iguais a zero.

A matriz de confusão gerada pelo modelo de regressão logística teve os resultados que podem ser visto na Tabela 8.

Tabela 8 – Matriz de confusão

		Classe predita		Acurácia 82%
Classe real		Não cancelamento	Cancelamento	Precisão
	Não cancelamento	869	25	97,2%
	Cancelamento	328	819	71,4%

Fonte: A autora (2017).

De forma geral, o modelo de regressão logística apresentou uma acurácia alta, de 82%. Portanto, é possível afirmar que este método foi bem escolhido para a construção do modelo. Em relação à precisão das variáveis, o modelo classificou de forma mais correta a variável de “não cancelamento”, com 97,2% de precisão.

4.4 ESTIMAÇÃO DO MODELO DE ÁRVORE DE DECISÃO

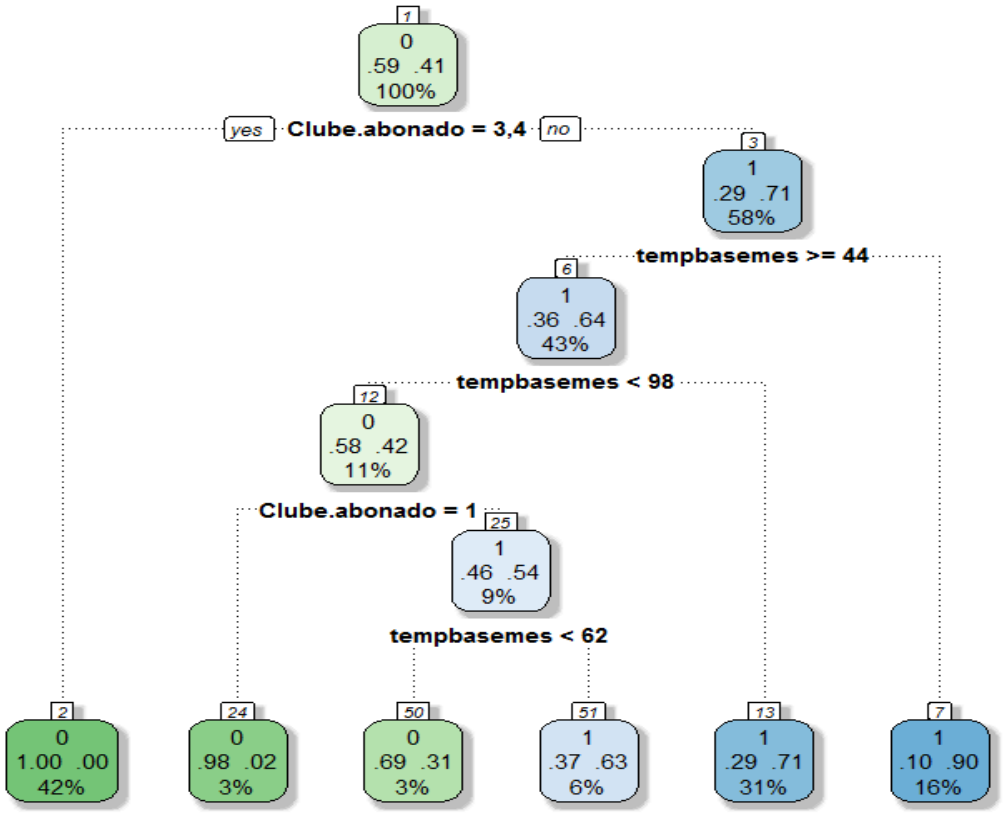
Para a construção da árvore de decisão, foram consideradas inicialmente todas as variáveis disponíveis. A árvore de decisão estimada foi capaz de escolher, dentre

todas as variáveis iniciais que poderiam ser úteis, as mais importantes para a predição.

A base de assinantes foi dividida em treino e teste, para que, após a construção da árvore, fosse feita a validação do modelo. A base treino representa 70% da amostra e a base teste, os 30% restantes. Este processo é feito para determinar o quanto o modelo é preciso na prática, ou seja, como seria o seu desempenho para um novo conjunto de dados. Assim, é possível garantir que o modelo realmente possui capacidade de generalização.

A árvore, resultante do modelo e escolha de variáveis, apresentou o resultado conforme a Figura 2:

Figura 2 – Árvore de decisão



Fonte: A autora (2017).

Legenda variáveis: clube.abonado 3 = Clube ativo clube.abonado 4 = Clube abonado clube.abonado 1= sem clube tempbasemes = tempo em meses que o assinante se encontra ativo na base.

4.5 INTERPRETAÇÃO DOS RESULTADOS DA ÁRVORE DE DECISÃO

Como um modelo lógico, a árvore de decisão mostra como o valor de uma variável de destino pode ser previsto usando os valores de um conjunto de variáveis preditoras (YU et al., 2010). Dado um conjunto de dados de treinamento, este algoritmo de aprendizado constrói uma árvore na qual cada nó é um atributo e os ramos dos nós são valores de atributo correspondentes (SARADHI; PALSHIKAR, 2011).

A árvore gerada levou em consideração as mesmas onze variáveis escolhidas para o modelo de regressão logística. Entretanto, é possível perceber que o método considerou como relevantes na explicação do cancelamento apenas duas delas: o clube – classificado em não possui clube, clube ativo, clube cancelado e clube abonado –; e tempo de base – variável contínua que consiste na quantidade de meses da assinatura. A árvore possui seis grupos de classificação, sendo que três deles caracterizam o cancelamento da assinatura.

A variável mais importante na explicação do cancelamento foi o clube, especificamente as classificações três e quatro – clube ativo e clube abonado –, primeiro nó da árvore. Assinaturas com clube ativo e abonado tem 59% de probabilidade de pertencerem ao primeiro grupo de classificação da árvore, que consiste no não cancelamento da assinatura, representando 41% da amostra de treino.

Assinaturas que não possuem a classificação clube ativo e clube abonado entram no segundo nó da árvore – tempo de base maior ou igual a 44 meses –; esse público consiste em 58% da amostra treino. Esse nó novamente é dividido em SIM ou NÃO, isto é, possui tempo de base maior ou igual a 44 meses ou possui tempo de base inferior a 44 meses. O grupo que possui mais de 44 meses de tempo de base tem 29% de probabilidade de pertencer ao grupo que possui menos de 98 meses de tempo base (terceiro nó da árvore). Já o grupo que possui menos de 44 meses de tempo de base tem probabilidade de 71% de pertencer ao sexto grupo de classificação da árvore – consistindo no cancelamento da assinatura, representando 16% da base.

O terceiro nó da árvore corresponde ao tempo de base inferior a 98 meses, que representa 43% da base de treino. A parte que possui tempo de base inferior a 98 meses, com 36% de probabilidade, cai para o quarto nó da árvore – clube cancelado. Já a parcela que tem tempo de base maior que 98 meses, apresenta 64% de

probabilidade de pertencer à quinta classificação da árvore – que consiste no cancelamento da assinatura e representa 31% da base.

A parcela que possui tempo de base inferior a 98 meses e o clube cancelado tem 58% de probabilidade de pertencerem ao segundo grupo de classificação da árvore – de não cancelamento, que representam apenas 3% da base. Já a parcela que possui tempo de base inferior a 98 meses e não tem o clube cancelado, recaem no quinto nó da árvore – tempo de base menor que 62 meses.

As assinaturas que possuem menos de 62 meses de tempo de base tem 46% de probabilidade de pertencerem ao grupo 3 da árvore – de não cancelados. Já as assinaturas que possuem mais de 62 meses de tempo de base tem a probabilidade de 54% de pertencerem ao quarto grupo de classificação da árvore – constituído pelo cancelamento e representando 6% da base.

4.6 VALIDAÇÃO DOS RESULTADOS DA ÁRVORE DE DECISÃO

A validação da árvore de decisão levou em consideração a base treino, que consiste nos 30% da base de assinantes. Para validação do modelo foi utilizada a matriz de confusão. A matriz de confusão é um conceito da aprendizagem mecânica, que contém informações sobre classificações reais e previstas feitas por um sistema de classificação. Uma matriz de confusão tem duas dimensões, uma dimensão é indexada pela classe real de um objeto, a outra é indexada pela classe que o classificador prevê (DENG et al., 2016).

Nesta matriz as linhas representam os casos reais e as colunas as previsões efetuadas pelo modelo. O número de acertos para cada caso está indicado na diagonal principal da matriz. A matriz de confusão de um classificador ideal possui todos os restantes elementos iguais a zero.

A matriz de confusão gerada pelo modelo de árvore de decisão teve resultados conforme a Tabela 9.

Tabela 9 – Matriz de confusão

		Classe predita		Acurácia 86%
Classe real		Não cancelamento	Cancelamento	Precisão
	Não cancelamento	912	10	98,9%
	Cancelamento	285	834	74,5%

Fonte: A autora (2017).

A partir dos resultados obtidos através da matriz de confusão, é possível afirmar que a árvore de decisão é um método apropriado para a construção do modelo preditivo de *churn* de migrados, pois apresentou 86% de acurácia. Em relação à precisão das variáveis, o modelo classificou de forma mais correta a variável de “não cancelamento”, com 98,9% de precisão.

5 CONCLUSÃO

Conforme explicado anteriormente, a retenção de clientes é mais barata se comparada à aquisição; logo, ações que objetivam uma melhor gestão de *churn* são de extrema relevância para as empresas. A construção dos dois modelos de *churn* preditivo têm como objetivo prever o cancelamento de assinaturas e, portanto, auxiliar a empresa em ações de relacionamento e retenção que consigam manter o máximo de assinantes saudáveis na carteira – diminuindo, assim, a necessidade de aquisição de novos clientes.

A construção de modelos preditivos a partir de informações encontradas no banco de dados de uma empresa é uma forma de incrementar os recursos de retenção de clientes, focando na estratégia de manutenção da carteira, em vez de aquisição de novos clientes, conforme colocado por Kim e Yoon (2004). Além de ser uma estratégia de retenção, a utilização de métodos estatísticos possibilita um maior acerto nas ações, pois estarão baseadas em um modelo com boa acurácia, e não só na vivência de gestores e histórico de ações prévias utilizadas pela organização.

A partir do desenvolvimento do modelo preditivo é possível agir de forma a evitar o cancelamento de assinaturas através de ações dirigidas, ou “targeted”, isto é, desenvolver ações que impactem somente o público identificado nesse estudo como com maior propensão ao cancelamento. Para tanto, é preciso que os resultados obtidos nessa pesquisa sejam incorporados à base de dados da empresa, o que é feito através da criação de *scores*, que funciona da seguinte forma: para cada assinante presente na base de dados que seja um migrado do produto impresso para o digital será atribuído um número de 0 a 1 que compreende a probabilidade de cancelamento.

Através dessa implementação é possível a ação preventiva por parte da empresa, visando à manutenção desses assinantes na base. Essas ações, além de serem pautadas pelo *score* na base, podem consistir em ações relacionadas às variáveis mais relevantes identificadas na árvore de decisão: tempo de base e clube do assinante – gratificando assinantes antigos na base e ofertando e incentivando o uso do clube do assinante para assinantes migrados.

Após obtenção do resultado do modelo de predição de acordo com as variáveis explicitadas anteriormente, o modelo de regressão logística foi novamente rodado, mas composto apenas das variáveis que representaram significância no modelo

inicial, sendo elas: tempo de base, gênero, renda, região, “possui login”, clube do assinante, canal de venda e incidências em cobrança. O modelo resultante seguiu resultando nas mesmas classificações das variáveis como significativas conforme o modelo anterior (composto pelas 12 variáveis) e a mesma acurácia, de 82%.

Além da possibilidade de ações de marketing visando ao relacionamento e a retenção de assinantes a construção do modelo a partir do método da árvore de decisão, que apresentou a maior acurácia – de 86% –, permite que esse modelo seja aplicado a outras bases da empresa – como propensão a cancelamento de assinantes de impresso, ou propensão à compra, devido ao seu bom desempenho. Levando em consideração a aplicação dos dois métodos – regressão logística e árvore de decisão – para a construção do modelo preditivo de *churn* de migrados, é possível afirmar, por meio dos resultados da validação dos modelos, que a árvore de decisão é o método mais adequado, levando em consideração as variáveis escolhidas para o estudo. O método de árvore de decisão, além de mais visual e gerencial, obteve uma maior acurácia se comparada à da regressão logística, que foi de 82%. Apesar da definição de onze variáveis para a construção do modelo, o método da árvore de decisão considerou apenas cinco variáveis como significativas na explicação do cancelamento.

6 LIMITAÇÕES E SUGESTÃO DE PESQUISAS FUTURAS

O modelo construído levou em consideração onze variáveis de um banco de dados com diversas informações de perfil e comportamento. O resultado obtido é diretamente relacionado à escolha dessas variáveis, que foram escolhidas com base no uso das mesmas para ações de retenção e relacionamento, além de estudos internos e conversas com gestores das áreas de *database* marketing e retenção. Devido a limitações quanto às variáveis de comportamento no banco de dados – como dias navegados, frequência de uso – há, consequentemente, uma limitação no modelo construído.

A presença de poucas variáveis relacionadas ao comportamento dos assinantes, tanto de uso como de preferências e valorização do produto, poderia ser suprida pela realização de uma pesquisa com os clientes, visando à incorporação de variáveis relacionadas a hábito de leitura e uso do jornal na forma digital, que auxiliariam ainda mais na predição do cancelamento da assinatura. Além disso, conforme citado no item modelo preditivo, no referencial teórico dessa pesquisa existem diversos métodos para a construção de um modelo de predição, cada um com suas vantagens, desvantagens e adequações compatíveis às características das variáveis independentes e dependentes que se quer estudar. Em razão disso, a escolha dos dois métodos também limita o resultado da pesquisa, pois há a possibilidade de construção do modelo com base em outros métodos descritos neste referencial teórico.

REFERÊNCIAS

AHMAD, Rizal; BUTTLE, Francis. Customer retention management: a reflection of theory and pratice. **Marketing Intelligence & Planning**, v. 20, n. 3, p. 149-161, 2002.

ARCHAUX, Cédric et al. An SVM based *Churn* Detector in Prepaid Mobile Telephony. **IEEE Explore**, 2014.

BAECKE, P.; POEL, Dirk van den. Improving purchasing behavior predictions by data augmentation with situational variables. **International Journal of Information Technology & Decision Making**, v. 36, n. 3, p. 367-383.

BOTELHO, Delane; TOSTES, Frederico Damian. Modelagem de probabilidade de *churn*. **Revista de Admistração de Empresas**, São Paulo, v. 50, n. 4, p. 396-410, 2010.

BREIMAN, Leo. Random forests. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001. Disponível em: <https://link.springer.com/article/10.1023%2FA%3A1010933404324>. Acesso em: 25 jun. 2017.

BUCKNIX, Wolfgang; POEL, Dirk van den. Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. **European Journal of Operational Research**, v. 164, n. 1, p. 252–268, 2005.

BUREZ, Jonathan; POEL, Dirk van den. Separating financial from commercial customer *churn*: a modeling step towards resolving the conflict between the sales and credit department. **Expert Systems with Applications**, v. 35, n. 1-2, p. 497-514, 2008.

CISTER, Angela Maria. **Mineração de dados para a análise de atrito em telefonia móvel**. 2005. Tese (Doutorado em Engenharia Civil) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2005.

COUSSEMENT, Kristof; BENOIT, Dries; POEL, Dirk van den. Improved marketing decision making in a customer *churn* prediction context using generalized additive models. **Expert Systems with Applications**, 2009. Disponível em: <http://doi.org/10.1016/j.eswa.2009.07.029>. Acesso em: 26 jun. 2017.

DENG, Xinyang et al. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. **Information Sciences**, v. 340-341, p. 250-261, 2016. Disponível em: <https://doi.org/10.1016/j.ins.2016.01.033>. Acesso em: 25 jun. 2017.

ELÍO, Javier et al. Logistic regression model for detecting radon prone areas in Ireland. **Science of The Total Environment**, v. 599, p. 1317-1329, 2017. Disponível em: <https://doi.org/10.1016/j.scitotenv.2017.05.071>. Acesso em: 26 jun. 2017.

GUANGLI, Nie et al. Credit card *churn* forecasting by logistic regression and decision tree. **Expert Systems with Applications**, v. 38, n. 12, p. 15273-15285, 2011.

Disponível em: <https://doi.org/10.1016/j.eswa.2011.06.028>. Acesso em: 26 jun. 2017.

HADDEN, John et al. Computer assisted customer *churn* management: State-of-the-art and future trends. **Computers & Operations Research**, v. 34, 2902-2917, 2005.

HAIR, Joseph et al. **Análise multivariada de dados**. Porto Alegre: Artmed, 2005.

HUNG, Shin-Yuan; YEN, David; WANG, Hsiu-Yu. Applying data mining to telecom *churn* management. **Expert Systems with Applications**, v. 31, n. 3, p. 515-524, 2006.

JING, Xia et al. Adjusted weight voting algorithm for random forests in handling missing values. **Pattern Recognition**, v. 69, p. 52-60, 2017. Disponível em: <https://doi.org/10.1016/j.patcog.2017.04.005>. Acesso em: 26 jun. 2017.

KAMAKURA, Wagner et al. Cross-selling through database marketing: a mixed data factor analyzer for data augmentation and rediction. **International Journal of Research in Marketing**, v. 20, 45-65, 2003. Disponível em: <https://doi.org/10.1016/S0167-8116(02)00121-0>. Acesso em: 25 jun. 2017.

KARABADJI, Nour El Islem et al. An evolutionary scheme for decision tree construction. **Knowledge-Based Systems**, v. 119, p. 166-177, 2017. Disponível em: <https://doi.org/10.1016/j.knosys.2016.12.011>. Acesso em: 26 jun. 2017.

KEAVENEY, Susan M. Customer switching behavior in service industries: An exploratory study. **Journal of Marketing**, v. 59, n. 2, p. 71-82, 1995. Disponível em: <http://psycnet.apa.org/doi/10.2307/1252074>. Acesso em: 25 jun. 2017.

KIM, Hee-Su; YOON, Choong-Han. Determinants of subscriber *churn* and customer loyalty in the Korean mobile telephony market. **Telecommunications Policy**, v. 28, n. 9/10, p. 751-765, 2004. Disponível em: <https://doi.org/10.1016/j.telpol.2004.05.013>. Acesso em: 25 jun. 2017.

KIM, Kyoungok. A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree. **Pattern Recognition**, v. 60, p. 157-163, 2016. Disponível em: <https://doi.org/10.1016/j.patcog.2016.04.016>. Acesso em: 26 jun. 2017.

LAITILA, Thomas. A pseudo-R2 measure for limited and qualitative dependent variable models. **Journal of Econometrics**, v. 56, n. 3, p. 341-355, 1993. Disponível em: <https://doi.org/10.1016/0304-4076(93)90125-O>. Acesso em: 25 jun. 2017.

LARIVIERE, Bart; POEL, Dirk van den. Investigating the role of product features in preventing customer *churn*, by using survival analysis and choice modeling: The case of financial services. **Expert Systems with Applications**, v. 27, n. 2, p. 277-285, 2004.

LIN, Lin et al. Random forests-based extreme learning machine ensemble for multi-regime time series prediction. **Expert Systems with Applications**, v. 83, p. 164-176,

2017. Disponível em: <https://doi.org/10.1016/j.eswa.2017.04.013>. Acesso em: 25 jun. 2017.

MA, Shaohui, TAN, Hui; SHU, Fang. When is the best time to reactivate your inactive customers? **Marketing Letters**, v. 26, n. 1, 81-98, 2015. Disponível em: <https://link.springer.com/article/10.1007%2Fs11002-013-9269-7>. Acesso em: 25 jun. 2017.

MENÊZES, Regilda da Costa Silva; FIRMINO, Paulo Renato; DROGUETT, Enrique López. Análise de confiabilidade humana via redes Bayesianas. In: SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, 37., 2005, Gramado. **Anais eletrônicos...** Gramado: SBPO, 2005. Disponível em: <http://www.din.uem.br/sbpo/sbpo2005/pdf/arq0176.pdf>. Acesso em: 30 out. 2016.

MILAN, Gabriel Sperandio; TONI, Deonir de. A construção de um modelo sobre a retenção de clientes e seus antecedentes em um ambiente de serviços. **Revista Eletrônica de Administração**, Porto Alegre, v. 18, n. 2, p. 433-467, 2012. Disponível em: <http://www.scielo.br/pdf/read/v18n2/a06v18n2.pdf>. Acesso em: 30 out. 2016.

MORIK, Katharina; KOPCKE, Hanna. Analysing customer *churn* in insurance data: a case study. In: BOULICAUT, Jean-Francois et al. (Orgs.). **Knowledge discovery in databases**: PKDD 2004. New York: Springer, 2004.

MOROTTI, Stefano; GRANDI, Eleonora. Logistic regression analysis of populations of electrophysiological models to assess proarrythmic risk. **MethodsX**, v. 4, p. 25-34, 2017. Disponível em: <https://doi.org/10.1016/j.mex.2016.12.002>. Acesso em: 26 jun. 2017.

OLIVER, Paulo Roberto da Costa. **Projetos de ECM/BPM**: os segredos da construção. 1. ed. São Paulo: Biblioteca24horas, 2010.

PEPPERS, Don; ROGERS, Martha. **Managing customer relationships**: a strategic framework. 2. ed. New York: Wiley, 2011.

QIAN, Zhiguang; JIANG, Wei; TSUI, Kwok-Leung. *Churn* detection via customer profile modelling. **International Journal of Production Research**, v. 44, n. 14, p. 2913-2933, 2006.

REICHHELD, Frederick; SASSER, W. Earl. Zero defections: quality comes to services. **Harvard Business Review**, v. 68, n. 5, 105-111, 1990.

SAETTLER, Aline; LABER, Eduardo; PEREIRA, Felipe de A. Mello. Decision tree classification with bounded number of errors. **Information Processing Letters**, v. 127, p. 27-31, 2017. Disponível em: <https://doi.org/10.1016/j.ipl.2017.06.011>. Acesso em: 26 jun. 2017.

SARADHI, Vijaya; PALSHIKAR, Girish. Employee *churn* prediction. **Expert Systems with Applications**, v. 38, n. 3, p. 1999-2006, 2011. Disponível em: <https://doi.org/10.1016/j.eswa.2010.07.134>. Acesso em: 26 jun. 2017.

SHMILOVICI, Armin. Support vector machines. In: MAIMON, Oded; ROKACH, Lior. (Org.). **Data mining and knowledge discovery handbook**. New York: Springer-Verlag, 2005. p. 257–276.

SOARES, José Francisco; SIQUEIRA, Arminda Lúcia. **Introdução à estatística médica**. Belo Horizonte: Departamento de Estatística/UFMG, 1999.

SUN, Huaining; HU, Xuegang. Attribute selection for decision tree learning with class constraint. **Chemometrics and Intelligent Laboratory Systems**, v. 163, p. 16-23, 2017. Disponível em: <https://doi.org/10.1016/j.chemolab.2017.02.004>. Acesso em: 26 jun. 2017.

TSAI, Chih-Fong; LU, Yu-Hsin. Customer *churn* prediction by hybrid neural networks. **Expert Systems with Applications**, v. 36, n. 10, p. 12547-12553, 2009. Disponível em: <https://doi.org/10.1016/j.eswa.2009.05.032>. Acesso em: 26 jun. 2017.

VAFEIADIS, Thanasis et al. A comparison of machine learning techniques for customer churn prediction. **Simulation Modelling Practice and Theory**, v. 55, p. 1-9, 2015. Disponível em: <https://doi.org/10.1016/j.simpat.2015.03.003>. Acesso em: 26 jun. 2017.

VAVRA, Terry; PRUDEN, Douglas. Using aftermarketing to maintain a customer base. **Discount Merchandiser**, v. 35, n. 5, 1995.

VERBEKE, Woute et al. Building comprehensible customer *churn* prediction models with advanced rule induction techniques. **Expert Systems with Applications**, v. 38, n. 3, p. 2354-2364, 2011.

VERBRAKEN, Thomas; VERBEKE, Wouter; BAESENS, Bart. Profit optimizing customer *churn* prediction with Bayesian network classifiers. **Intelligent Data Analysis**, v. 18, n. 1, p. 3-24, 2014. Disponível em: <https://doi.org/10.3233/IDA-130625>. Acesso em: 26 jun. 2017.

WEI, Chih-Ping; CHIU, I-Tang. Turning telecommunications call details to *churn* prediction: a data mining approach. **Expert Systems with Applications**, v. 23, p. 103-112, 2002.

WITTEN, Ian; FRANK, Eibe. **Data mining**: practical machine learning tools and techniques. São Francisco: The Morgan Kaufmann series in data management systems, 2011.

YU, Zhun et al. A decision tree method for building energy demand modeling. **Energy and Buildings**, v. 42, n. 10, p. 1637-1646, 2010. Disponível em: <https://doi.org/10.1016/j.enbuild.2010.04.006>. Acesso em: 26 jun. 2017.

ZHAO, Yongheng; ZHANG, Yanshia. Comparison of decision tree methods for finding active objects. **Advances in Space Research**, v. 41, n. 12, p. 1955-1959, 2008. Disponível em: <https://doi.org/10.1016/j.asr.2007.07.020>. Acesso em: 26 jun. 2017.