Business Intelligence

PUC RIO

Pedro Henrique Gonçalves Nobre

Aplicação de modelos de data mining para previsão de churn em telecomunicações

Monografia de Final de Curso

16/12/2016

Monografia apresentada ao Departamento de Engenharia Elétrica da PUC/Rio como parte dos requisitos para a obtenção do título de Especialização em Business Intelligence.

Orientadores:

Leonardo Mendoza

RESUMO

A dissertação presente discorre sobre a previsão de churn em clientes pós-pagos de uma operadora de telecomunicações móveis. O Churn representa o momento onde ocorre a troca, por parte do usuário, de uma empresa por outra, ou o downgrade para os planos pré-pagos assim como o cancelamento total da linha.

O objetivo será conseguir, observando o comportamento dos clientes nos meses anteriores, prever os clientes que deixarão a empresa no mês seguinte, a modelagem tentará minimizar os falsos positivos para que não se faça ofertas de retenção para clientes sem necessidade.

Dados reais de uma operadora de telecomunicações móveis serão utilizados nesse trabalho. O nome da operadora não será revelado.

De nada adianta investir em aquisição de novos clientes se a fidelidade não for trabalhada a cada dia. Para isso, o segredo é conhecer o cliente, entender seu comportamento de uso e antecipar suas expectativas e necessidades. O custo de adquirir um novo cliente é muito superior ao custo de manter um cliente existente, por isso, antecipar o churn é significativo no campo das telecomunicações móveis.

A eficácia da previsão será avaliada de acordo com a acurácia do modelo em relação ao churn, ou seja, o total de acertos de churners sobre o total de clientes que o modelo classificou como churn.

ABSTRACT

This dissertation elaborates on the churn prediction within the postpaid users of mobile telecommunications companies. The churn represents the moment when the user switches from a company to another, from a postpaid to a pre-payed plan or even requests for a cancelation of his current plan.

By observing the behaviour pattern of the costumers in previous months, the idea is to predict and identify the ones who might decide to leave. This prediction will try to minimize false positives so special offers won't be made to clients unnecessarily.

There is no point in investing in acquirement of new clients, if no actions are taken to make sure they stay loyal to your business. With this in mind, the secret is to know the costumer, understand their behaviour and anticipate their expectations and necessities. The costs of capturing new clients are higher than the ones for keeping them satisfied, so taking that into consideration, anticipating the churn is an action that should be considered crucial in the mobile telecommunications industry.

In execution process of this dissertation, real data from a Brazilian telecommunications company was used. The name of this company will remain anonymous for legal purposes.

The efficacy of the prediction will be evaluated according to the accuracy of the model in relation to the churn, that is, the total churners hits on the total customers that the model classified as churn.

SUMÁRIO

1 INTRODUÇÃO

Churn é definido como o abandono de um cliente de um serviço para outro ou o cancelamento total de sua linha, e pode ter razões diferentes.

As motivações mais comuns para o abandono do serviço (churn) são: tarifas mais vantajosas na concorrência, má qualidade do serviço, problemas no atendimento, entre outros.

Os dados utilizados para construir o modelo serão fornecidos por uma operadora de telecomunicações, o objetivo principal é prever, a partir dos dados de determinado mês, quem serão os clientes churn no mês seguinte. A solução que será proposta na elaboração do projeto e será testada e avaliada ao fim do projeto.

O mercado de telecomunicações no Brasil busca aprimorar seus serviços a fim de consolidar seus clientes e converter novos usuários. Modelos de previsão são utilizados com frequência com o intuito de solucionar problemas de inadimplência, orçamento, churn, etc. Os modelos mais utilizados são as regressões, essas trazem resultados satisfatórios e são relativamente simples de serem aplicados apesar de exigirem certo conhecimento.

Este projeto busca apresentar algumas soluções para o problema de previsão do churn, baseados no conhecimento adquirido durante a realização do curso de Business Intelligence. Apesar de não haver modelo preciso, buscamos encontrar o padrão que melhor se encaixe neste caso específico. Os casos são singulares, por isso, se faz necessário analisar de forma crítica, objetivando o melhor resultado possível.

A abordagem proposta ao problema terá como objetivo cumprir os requisitos desejáveis em um sistema de previsão de churn. Segundo Balle, Casas, & Catarineu (2011) os requisitos são os seguintes:

- Precisão: na avaliação do classificador este deve ter uma medida de recall elevado (pelo menos todos os churners são identificados) e precisão relativamente elevada (não haver muitos falsos positivos).

- Desempenho: a rapidez com que o modelo pode ser executado com novos dados é essencial para poderem ser tomadas as decisões certas no tempo certo.

- Flexibilidade: o modelo tem que conseguir manter-se com bons índices de previsão com a previsível alteração nos padrões dos clientes que serão introduzidos cada vez que for necessário fazer uma previsão de churn.

- Escala: o modelo tem que reagir de forma aceitável ao aumento de dados com que poderá ser alimentado.

- Segmentação: esta característica prende-se com a capacidade de serem retirados dados concretos sobre o perfil de utilizadores mais propensos a deixarem o serviço e possivelmente incluir variáveis baseadas na experiência dos analistas que conhecem o negócio.

Este trabalho busca antecipar o churn de clientes pós-pagos de uma empresa de telecomunicações utilizando variáveis de cadastro, tráfego e faturamento. É desejado revelar, com a maior precisão possível, quais clientes deixarão a empresa no mês seguinte, através de modelos de classificação.

1.1 MOTIVAÇÃO

Dados simples e fartos das empresas de telecomunicações podem fornecer informações muito valiosas para a estratégia da empresa. Neste caso, utilizando apenas dados de cadastro, tráfego e faturamento dos clientes pretendemos prever a saída de clientes e, a partir dessa informação, atuar na retenção.

Uma previsão correta é essencial para a melhora nos indicadores de retenção que por consequência melhora o índice de churn, que é vital para as empresas de telecomunicações. Muitas vezes, é tarde demais para reverter o quadro quando o cliente resolve trocar de operadora ou fazer um downgrade para planos pré-pagos.

1.2 OBJETIVOS DO TRABALHO

A proposta deste trabalho é apresentar uma aplicação de algumas técnicas de data mining para identificar os clientes que resolvem deixar a operadora de forma voluntária, com intenção de antecipar possíveis clientes churn, pretendendo atuar na retenção dos mesmos a fim de reduzir possíveis prejuízos.

Com uma base que apresenta dados de faturamento, cadastro e consumo de clientes pós-pagos de uma empresa de telecomunicações, é possível classificar e perceber eventuais padrões e perfis de clientes propícios, ou não, ao churn, com o objetivo de fornecer subsídios para a tomada de decisão gerencial no momento da retenção.

O objetivo principal do presente trabalho é o desenvolvimento do modelo mais adequado capaz de assessorar a tomada de decisão no momento da construção de uma base de clientes propícios ao churn.

Os objetivos específicos propostos foram:

A) buscar o melhor entre os três modelos de classificação propostos com a utilização da ferramenta R;

B) fornecer informações relevantes ao final do processo alvo do estudo de caso;

C) Poder informar com a melhor acurácia possível quais são os clientes com possibilidade de saída para a tomada de decisão da área de retenção.

2 METODOLOGIAS

Neste projeto foram aplicados três modelos aos dados observados, buscando compará-los para identificar qual técnica classifica melhor o perfil do consumidor. Os modelos são: a regressão logística, árvore de classificação e random forest.

Os modelos de classificação descrevem o grupo ao qual o item pertence por meio do exame dos itens já classificados e pela inferência de um conjunto de regras. Exemplo: empresas de operadoras de cartões de crédito e companhias telefônicas preocupam-se com a perda de clientes regulares, a classificação pode ajudar a descobrir as características de clientes que provavelmente virão abandona-las e oferecer um modelo para ajudar os gerentes a prever quem são, de modo que se elabore antecipadamente campanhas especiais para reter esses clientes.

2.1 REGRESSÃO LOGÍSTICA

Em muitas situações práticas, pesquisadores desejam separar duas classes de objetos ou alocar um novo objeto em uma dessas classes. De todas as técnicas existentes para esta finalidade, a Regressão Logística é uma das mais encontradas na literatura. Esta técnica se enquadra na classe de métodos estatísticos multivariados de dependência, pois relaciona um conjunto de variáveis independentes com uma variável dependente categórica (Sharma, 1996; Hair et al., 1998; Morgan e Griego, 1998).

De forma geral, as técnicas de discriminação procuram encontrar uma função ou conjunto de funções que discrimine os grupos definidos pela variável categórica, visando a minimizar erros de classificação.

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. O mesmo modelo pode ser utilizado com enfoque discriminatório, conforme descrevem Krzanowski (1988) e McLachlan (1992). Esses autores argumentam que o modelo logístico de discriminação pode ser utilizado de forma bem mais geral, pois não faz suposições quanto à forma funcional das variáveis

independentes, e o número de parâmetros envolvidos no processo de estimação provavelmente será menor.

Algumas características da regressão logística são destacadas por Hair et al. (1998) em comparação com outros modelos:

. não é necessário supor normalidade multivariada;

. é uma técnica mais genérica e mais robusta, pois sua aplicação é apropriada em grande variedade de situações;

. é uma técnica similar à regressão linear múltipla.

Na regressão logística, a probabilidade de ocorrência de um evento pode ser estimada diretamente. No caso da variável dependente *Y* assumir apenas dois possíveis estados (1 ou 0) e haver um conjunto de *p* variáveis independentes X_1 , X_2 , ... , X_p, o modelo de regressão logística pode ser escrito da seguinte forma:

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

onde,

$$g(x) = B_0 + B_1X_1 + \cdots + B_pX_p$$

Os coeficientes B_0, B_1, ..., B_p são estimados a partir do conjunto de dados, pelo método da máxima verossimilhança, que encontra uma combinação de coeficientes que maximiza a probabilidade de a amostra ter sido observada (Hosmer e Lemeshow, 1989). Considerando certa combinação de coeficientes B_0, B_1, ..., B_p e variando os valores de *X*, observa-se que a curva logística tem comportamento probabilístico no formato da letra *S*, o que é característica da regressão logística. Esse formato dá à regressão logística alto grau de generalidade, aliada a aspectos muito desejáveis:

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

b) Quando $g(x) \to +\infty$, então $P(Y = 1) \to 0$

Assim como podemos estimar diretamente a probabilidade de ocorrência de um evento, podemos estimar a probabilidade de não ocorrência por diferença:

$P(Y = 0) = 1 - P(Y = 1)$

Ao utilizarmos a 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:

$$\frac{P(Y=1)}{P(Y=0)} = e^{B_0+B_1X_1+\ldots+B_pX_p}$$

e, por consequência,

$$\ln\left[\frac{P(Y=1)}{P(Y=0)}\right] = B_0 + B_1X_1 + \ldots + B_pX_p$$

Para utilizar o modelo de regressão logística para discriminação de dois grupos, a regra de classificação é a seguinte:

. se P(Y=1) > 0,5 então classifica-se Y=1;

. em caso contrário classifica-se Y=0.

Podendo variar essa probabilidade para buscar melhores resultados de acordo com a necessidade do problema. No caso proposto, vamos variar esta probabilidade buscando a redução dos falsos positivos para que o objetivo seja alcançado.

2.2 ÁRVORE DE CLASSIFICAÇÃO

Árvores de decisão são modelos estatísticos que utilizam um treinamento supervisionado para a classificação e previsão de dados. Em outras palavras, em sua construção é utilizado um conjunto de treinamento formado por entradas e saídas. Estas últimas são as classes. Estes modelos utilizam a estratégia de dividir para conquistar: um problema complexo é decomposto em sub-problemas mais simples e recursivamente esta técnica é aplicada a cada sub-problema (Gama, 2004). As árvores de decisão estão entre os mais populares algoritmos de inferência e tem sido aplicado em várias áreas como, por exemplo, diagnóstico médico e risco de crédito (Mitchell, 1997), e deles pode-se extrair regras do tipo "se-então" que são facilmente compreendidas. A capacidade de discriminação de uma árvore vem da divisão do espaço definido pelos atributos em sub-espaços e a cada sub-espaço é associada uma classe.

A figura 1 representa uma árvore de decisão onde cada nó de decisão contém um teste para algum atributo, cada ramo descendente corresponde a um possível valor deste atributo, o conjunto de ramos são distintos, cada folha está associada a uma classe e, cada percurso da árvore, da raiz à folha corresponde uma regra de classificação. No espaço definido pelos atributos, cada folha corresponde a um hiper-retângulo onde a interseção destes é vazia e a união é todo o espaço (Gama, 2004).

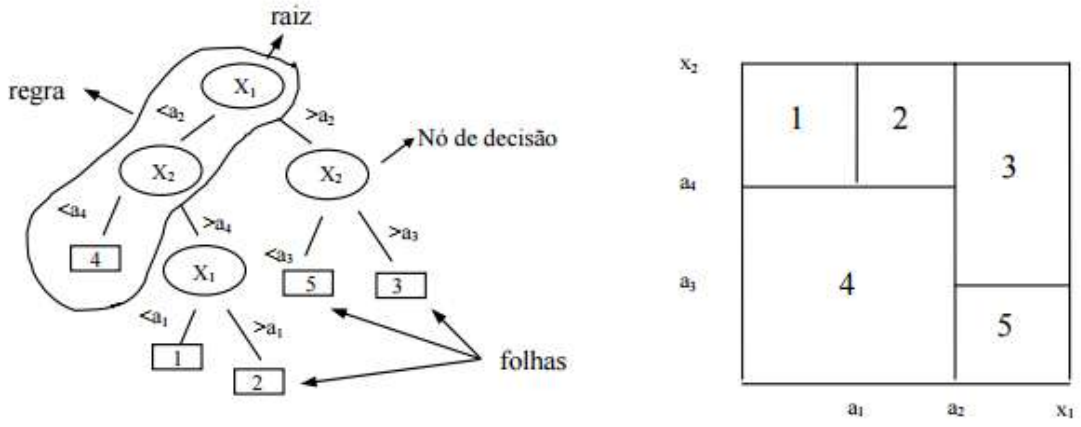

Figura 1 - Representação de uma árvore de decisão e sua respectiva representação no espaço (Gama, 2004)

O critério utilizado para realizar as partições é o da utilidade do atributo para a classificação. Aplica-se, por este critério, um determinado ganho de informação a cada atributo. O atributo escolhido como atributo teste para o corrente nó é aquele que possui o maior ganho de informação. A partir desta aplicação, inicia-se um novo processo de partição. Nos casos em que a árvore é usada para classificação, os critérios de partição mais conhecidos são baseados na entropia e índice Gini. (Onoda, 2001).

2.3 RANDOM FOREST

Random Forest trata-se de um algoritmo classificador que faz uso do método de árvores de decisão criada por Breiman (2001) possibilitando a mineração dos dados passados a mesma. Esta técnica possui uma ideia um pouco diferente dos algoritmos de árvores de decisão, a qual pertence, enquanto uma árvore possui o objetivo de construção total de uma estrutura a partir de uma base de dados o Random Forest tem o objetivo de efetuar a criação de várias árvores de decisão usando um subconjunto de atributos selecionados aleatoriamente a partir do conjunto original, contendo todos os atributos e que estes possuem um tipo de amostragem chamado de bootstrap, a qual é do tipo com reposição, possibilitando assim melhor análise dos dados. (NETO, 2014).

Com a quebra das massas de dados e construção de vários subconjuntos, uma árvore de decisão é construída. Com este procedimento então a construção das árvores ocorre pela seleção de atributos aleatoriamente a partir dos subconjuntos, onde os mesmos são aplicados nos nós de cada uma das árvores criadas. Uma Random Forest ou floresta aleatória é um conjunto dessas árvores de decisão. Após a criação dos conjuntos de árvores é possível efetuar a classificação de qual possui melhor ganho de conhecimento para a solução de determinado problema, para isto é necessário escolher um subconjunto de árvores de decisão que possui melhor lógica e vantagens para a tomada de decisão. Para cada subconjunto é dado um voto sobre qual classe o atributo chave deve pertencer, este voto possui um "peso" onde o mesmo é afetado pela igualdade entre as árvores, "sendo que quanto menor a similaridade entre duas árvores melhor, e pela força que cada árvore tem individualmente, ou seja, quanto mais precisa uma árvore for, melhor será sua nota." (NETO, 2014).

As Random Forests, conforme verificado anteriormente, possuem a característica de Dividir-para–Conquistar, e isto possibilita a mesma algumas características que se destacam referentes às outras técnicas, algumas delas são:

• Algoritmo mais poderoso do que comparado somente a uma árvore de decisão;

• Possui boa taxa de acerto quando testado em diferentes conjuntos de dados;

• Técnica exata;

• Evitam sobre ajuste (overfitting);

• Menos sensíveis a ruídos;

• Classificação aleatória das árvores sem intervenção humana.

A seguir é apresentado o funcionamento do método de classificação Random Forest através da Figura 2.

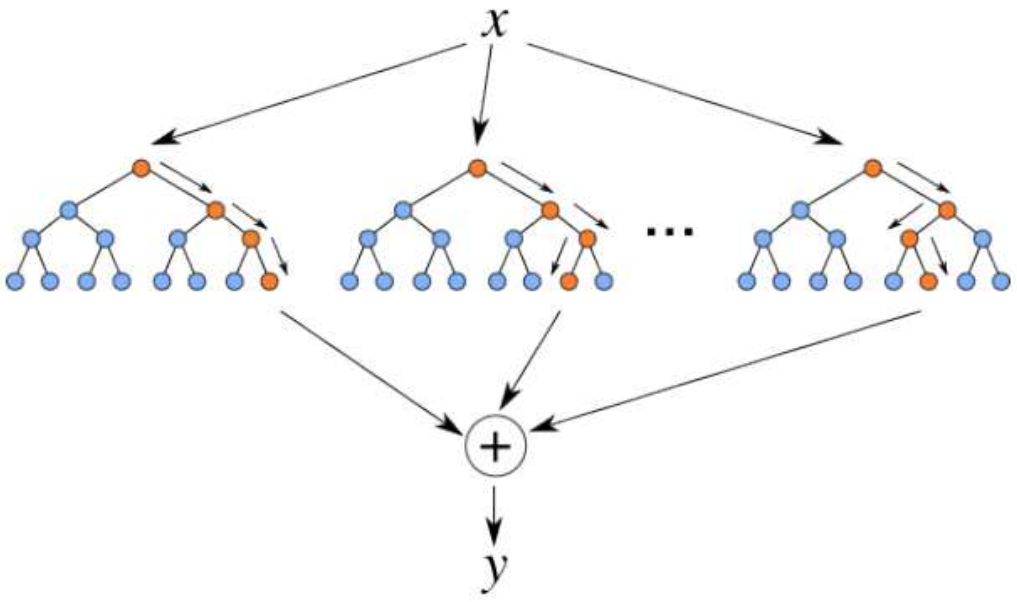

Figura 2 - Ilustração da lógica por trás do algoritmo Random Forest. (LORENZETT, 2016)

Na imagem anterior é possível verificar que partindo de um elemento X, no caso uma base de dados, gerou-se várias Random Forest, neste ponto cada uma gera várias regras e nelas a possibilidade de descoberta de novos padrões que poderão ser decisivos na tomada de decisão correta. Com as florestas criadas o próximo passo é calcular qual delas contém as regras mais exatas para a mineração.

Com a escolha feita é aplicado na base de dados as mesmas e assim chegando a um resultado Y.

2.4 CRITÉRIO DE INFORMAÇÃO DE AKAIKE

O Critério de Informação de Akaike (AIC) admite a existência de um modelo "real" que descreve os dados que é desconhecido, e tenta escolher dentre um grupo de modelos avaliados, o que minimiza a divergência de Kullback-Leibler (K-L). O valor de K-L para um modelo f com parâmetros θ, em relação ao modelo "real" representado por g é $I(g,f\theta) = \int g(y) Ln\left(\frac{g(y)}{f(y \mid \theta)}\right) dy$.

Esta divergência está relacionada à informação perdida por se usar um modelo aproximado e não o "real". A estimativa do AIC para um determinado modelo é dada por: $AIC = -2L + 2k$ em que, L o MLFV do modelo com os parâmetros θ e k o número de parâmetros. O modelo com menor valor e AIC é considerado o modelo de melhor ajuste.

Este critério foi utilizado para definir as variáveis de input dos modelos, comparando modelos com todas as combinações de variáveis e encontrando o melhor ajuste através do menor AIC.

3 SOFTWARE R

O R é uma linguagem e ambiente para computação estatística e gráfica. É um projeto GNU (General Public License da Free Software Foundation) que é similar à linguagem e ambiente S, que foi desenvolvida nos Laboratórios Bell (antiga AT & T, agora Lucent Technologies) por Rick Becker, John Chambers e Allan Wilks, e também forma a base dos sistemas S-Plus. O R pode ser considerado como uma aplicação diferente do S. Há algumas diferenças importantes, mas muitos códigos escritos para o S são executados de maneira inalterada pelo R.

O R fornece uma ampla variedade de estatística (modelagem linear e não-linear, testes estatísticos clássicos, análise de séries temporais, testes paramétricos e não-paramétricos, suavização, etc.) e técnicas gráficas e é altamente extensível. A linguagem S é frequentemente o veículo de escolha para a pesquisa na metodologia estatística, e o R oferece o código aberto como uma rota para a participação nessa atividade.

Um dos pontos fortes do R é a facilidade com que pode ser produzida a edição, com qualidade, de gráficos, incluindo símbolos e fórmulas matemáticas, quando necessário. Grandes cuidados foram tomados na elaboração dos padrões durante a escolha do projeto gráfico, mas o usuário mantém o controle total.

O R está disponível como software livre, nos termos da GNU na forma de código aberto. Pode ser compilado e "roda" em um grande número de plataformas UNIX e sistemas semelhantes (incluindo FreeBSD e Linux), Windows e MacOS.

O R é um conjunto integrado de facilidades em software para manipulação de dados, cálculo e visualização gráfica. Inclui:

Um eficaz tratamento de dados e facilidade de armazenagem; Um conjunto de operadores de cálculos sobre arrays, em especial matrizes; Uma ampla, coerente e integrada coleção de ferramentas intermediárias para análise de dados; Facilidade gráfica para análise e visualização de dados quer na tela ou impresso;

Bem desenvolvida, simples e eficaz linguagem de programação que inclui condicionantes, loops, funções recursivas definidas pelo usuário e facilidades de entrada e de saída.

4 RESULTADOS

Para a realização desse estudo foi utilizada uma amostra do banco de dados, referente ao primeiro trimestre de 2016, de uma empresa de telecomunicações. A carteira de clientes em estudo possui mais de 1MM de clientes e tem cerca de 40.000 retiradas voluntárias mensalmente. A partir das respostas do melhor modelo, pretende-se realizar a previsão para fornecer uma base de clientes à área de retenção com os possíveis churners com o menor número de falsos positivos possível para que sejam realizadas ações para evitar o desligamento do cliente. A escolha das variáveis iniciais da base foram feitas a partir de especialistas que julgaram ser informações relevantes a respeito do comportamento do cliente.

Para a fase de treinamento, aplicamos os modelos de regressão logística, árvore de decisão e random forest em uma amostra de 57.740 clientes do primeiro trimestre de 2016, onde foi considerado churn o cliente que saiu no mês seguinte ao mês de registro das variáveis do modelo. Foram consideradas variáveis de cadastro, tráfego, faturamento e indicadores construídos a partir delas para classificá-los como churn ou não.

4.1 VARIÁVEIS UTILIZADAS

As variáveis de input dos modelos, foram indicadas por especialistas da empresa como sendo importantes sobre o comportamento dos clientes e estão segmentadas pela sua origem nas tabelas abaixo:

Cadastro		
Coluna	**Tipo**	**Descrição**
MSISDN	Num	Terminal do cliente
PLANO	Char(100)	Plano contratado
UF	Char(2)	Unidade federativa
IDADE_GROSS	Char(10)	Faixa de tempo na base
FLAG_PCT_DADOS	Char(3)	Indica se o cliente possui ou não pacote de dados
FLAG_PLANO	Char(2)	Indica se o plano contratado é 2G, 3G ou 4G
FLAG_CHIP	Char(2)	Indica se o chip do cliente é 2G, 3G ou 4G
FLAG_APARELHO	Char(2)	Indica se o aparelho do cliente é 2G, 3G ou 4G
FLAG_CITY_4G	Char(2)	Indica se a cidade do cliente possui rede 4G

Tabela 1 - Dicionário de dados da tabela Cadastro.

Tráfego		
Coluna	**Tipo**	**Descrição**
MINUTO_ORIGEM	Num	Quantidade de minutos de chamadas realizadas
MINUTO_DESTINO	Num	Quantidade de minutos de chamadas recebidas
_2GMB	Num	Consumo de dados em velocidade 2G
_3GMB	Num	Consumo de dados em velocidade 3G
_4GMB	Num	Consumo de dados em velocidade 4G

Tabela 2 - Dicionário de dados da tabela Tráfego.

Faturamento		
Coluna	**Tipo**	**Descrição**
ASSINA_PRINC	Num	Valor pago pelo cliente pela assinatura do plano
ASSINA_PCT_DADOS	Num	Valor pago pelo cliente pela assinatura do pacote de dados
ASSINA_PCT_SMS	Num	Valor pago pelo cliente pela assinatura do pacote de sms
ASSINA_LD	Num	Valor pago pelo cliente pela assinatura do pacote de longa distância
ASSINA_DEPENDENTE	Num	Valor pago pelo cliente pela assinatura dos dependentes
EXCED_LOCAL	Num	Valor pago pelo cliente pelos excedentes de minutos em ligações locais
EXCED_LD	Num	Valor pago pelo cliente pelos excedentes de minutos em ligações longa distância
EXCED_MSG	Num	Valor pago pelo cliente pelos excedentes em mensagens de texto
EXCED_DADOS	Num	Valor pago pelo cliente pelos excedentes em consumo de dados
EXCED_SVA	Num	Valor pago pelo cliente pelos excedentes em serviços de valor agregado
EXCED_OUTROS	Num	Valor pago pelo cliente por outras cobranças

Tabela 3 - Dicionário de dados da tabela Faturamento.

Campos Calculados		
Coluna	**Tipo**	**Descrição**
MAIOR_REDE_DADOS	Char(2)	Indica em qual velocidade o cliente utiliza mais a rede
FAIXA_DIAS_USO_DADOS	Char(10)	Indica quantos dias o cliente usa dados por mês
FAIXA_DIAS_USO_SMS	Char(10)	Indica quantos dias o cliente usa sms's por mês
FAIXA_MB	Char(10)	Indica faixa que o cliente está em relação ao uso de dados
INDICADOR	Char(20)	Indica o que o cliente utiliza do plano
FLAG_GERAL	Char(40)	Combinação plano, chip e aparelho

Tabela 4 - Dicionário de dados da tabela Campos Calculados.

A base foi dividida em 70% para treinamento e 30% para validação. A base de treinamento foi balanceada, ou seja, metade dos clientes eram churn e a outra metade não. A base de teste possui a mesma proporção da realidade da empresa, apenas cerca de 2,5% dos clientes se desligam voluntariamente da empresa mensalmente.

4.2 SELEÇÃO AUTOMÁTICA DE VARIÁVEIS

Como a seleção de todas as regressões possíveis necessita de um considerável esforço computacional, outros métodos foram desenvolvidos para selecionar o melhor subconjunto de variáveis sequencialmente, adicionando ou removendo variáveis em cada passo.

O critério para a adição ou remoção das variáveis é geralmente baseado na estatística F, comparando modelos com e sem as variáveis em questão. O AIC, assim como outros critérios, também pode ser utilizado na decisão de inserir e remover variáveis. Foram aplicadas técnicas de seleção de variáveis automáticas foward e backward que se baseiam no AIC para definir as variáveis que ficarão no modelo final.

Após a seleção automática sobraram 23 variáveis que foram input no modelo.

Variáveis após seleção automática		
Coluna	**Tipo**	**Fonte**
PLANO	Char(100)	Cadastro
UF	Char(2)	Cadastro
IDADE_GROSS	Char(10)	Cadastro
FLAG_APARELHO	Char(2)	Cadastro
FLAG_CITY_4G	Char(2)	Cadastro
FAIXA_DIAS_USO_DADOS	Char(10)	Cadastro
FAIXA_MB	Char(10)	Cadastro
INDICADOR	Char(20)	Cadastro
FLAG_GERAL	Char(40)	Cadastro
MINUTO_ORIGEM	Num	Tráfego
MINUTO_DESTINO	Num	Tráfego
_2GMB	Num	Tráfego
_3GMB	Num	Tráfego
ASSINA_PRINC	Num	Faturamento
ASSINA_PCT_DADOS	Num	Faturamento
ASSINA_PCT_SMS	Num	Faturamento
ASSINA_LD	Num	Faturamento
ASSINA_DEPENDENTE	Num	Faturamento
EXCED_LOCAL	Num	Faturamento
EXCED_LD	Num	Faturamento
EXCED_DADOS	Num	Faturamento
EXCED_SVA	Num	Faturamento
EXCED_OUTROS	Num	Faturamento

Tabela 5 - Variáveis de input nos modelos.

4.3 RESULTADOS DOS MODELOS - TREINAMENTO

A saída de todos os modelos aplicados é a probabilidade do indivíduo pertencer a classe churn, foi necessário variar o limite para os casos serem considerados churn para buscar uma boa acertividade na classe churn, ou seja, reduzir o número de falsos positivos para que se realizem ações apenas em clientes realmente propensos a deixar a companhia.

Os três modelos apresentaram bom desempenho no treinamento, a regressão logística conseguiu acertar 38.903 dos 47.104 casos da base de treinamento, com apenas 685 falsos positivos.

LIMITE	CHURN OBSERVADO	CHURN PREVISTO		ACURÁCIA	FALSOS POSITIVOS
		NÃO	SIM		
50%	NÃO	48,21%	3,27%	83,26%	8,54%
	SIM	13,47%	35,05%		
80%	NÃO	50,03%	1,45%	82,59%	4,28%
	SIM	15,96%	32,56%		
90%	NÃO	51,11%	0,37%	75,25%	1,52%
	SIM	24,38%	24,13%		
95%	NÃO	51,38%	0,11%	69,32%	0,59%
	SIM	30,57%	17,94%		

Tabela 6 - Resultados do treinamento da Regressão Logística.

O modelo de árvore de classificação conseguiu acertar 39.139 dos 47.104 casos da base de treinamento com uma melhor performance nos falsos positivos, apenas 318.

LIMITE	CHURN OBSERVADO	CHURN PREVISTO		ACURÁCIA	FALSOS POSITIVOS
		NÃO	SIM		
50%	NÃO	49,72%	1,77%	84,65%	4,81%
	SIM	13,58%	34,93%		
80%	NÃO	50,81%	0,68%	83,09%	2,05%
	SIM	16,23%	32,28%		
90%	NÃO	51,02%	0,47%	81,87%	1,49%
	SIM	17,67%	30,85%		
95%	NÃO	51,20%	0,29%	79,68%	1,00%
	SIM	20,04%	28,48%		

Tabela 7 - Resultados do treinamento da Árvore de Classificação.

O modelo Random Forest previu corretamente 40.906 casos dos 47.104 da base de treinamento sem nenhum falso positivo registrado.

LIMITE	CHURN OBSERVADO	CHURN PREVISTO		ACURÁCIA	FALSOS POSITIVOS
		NÃO	SIM		
50%	NÃO	51,38%	0,11%	99,72%	0,23%
	SIM	0,17%	48,35%		
80%	NÃO	51,49%	0,00%	86,84%	0,00%
	SIM	13,16%	35,36%		
90%	NÃO	51,49%	0,00%	84,89%	0,00%
	SIM	15,11%	33,40%		
95%	NÃO	51,49%	0,00%	82,55%	0,00%
	SIM	17,45%	31,06%		

Tabela 8 - Resultados do treinamento do modelo Random Forest.

Após o treinamento, devemos realizar a validação do modelo através de novos dados ainda desconhecidos do modelo. A base de teste possui 10.636 registros e tem a mesma proporção de churn real da empresa, portanto, a base de validação possui apenas 341 casos churn.

4.4 RESULTADOS DOS MODELOS - VALIDAÇÃO

Na validação, todos os modelos continuaram apresentando bom rendimento com boa acurácia e baixo índice de falsos positivos. Ao realizar os testes, percebemos que o limite em 95% modela melhor os interesses deste trabalho, minimizando os falsos positivos.

O modelo de regressão logística conseguiu acertar 121 dos 341 casos churn com 21 falsos positivos.

LIMITE	CHURN OBSERVADO	CHURN PREVISTO		ACURÁCIA	FALSOS POSITIVOS
		NÃO	SIM		
50%	NÃO	90,31%	6,49%	92,54%	74,35%
	SIM	0,97%	2,24%		
80%	NÃO	93,66%	3,13%	95,76%	59,89%
	SIM	1,11%	2,10%		
90%	NÃO	95,92%	0,87%	97,48%	35,91%
	SIM	1,65%	1,56%		
95%	NÃO	96,60%	0,20%	97,73%	14,79%
	SIM	2,07%	1,14%		

Tabela 9 - Resultados da avaliação do modelo Regressão Logística.

O modelo de árvore de classificação conseguiu acertar 201 dos 341 casos churn com 87 falsos positivos.

LIMITE	CHURN OBSERVADO	CHURN PREVISTO		ACURÁCIA	FALSOS POSITIVOS
		NÃO	SIM		
50%	NÃO	93,21%	3,58%	95,51%	60,96%
	SIM	0,91%	2,29%		
80%	NÃO	95,19%	1,61%	97,34%	42,75%
	SIM	1,05%	2,15%		
90%	NÃO	95,55%	1,24%	97,61%	37,61%
	SIM	1,15%	2,06%		
95%	NÃO	95,98%	0,82%	97,87%	30,21%
	SIM	1,32%	1,89%		

Tabela 10 - Resultados da avaliação do modelo Árvore de Classificação.

Por fim, o modelo random forest conseguiu acertar 191 dos 341 casos churn com apenas 34 falsos positivos.

LIMITE	CHURN OBSERVADO	CHURN PREVISTO		ACURÁCIA	FALSOS POSITIVOS
		NÃO	SIM		
50%	NÃO	92,48%	4,32%	94,80%	65,01%
	SIM	0,88%	2,32%		
80%	NÃO	95,76%	1,03%	97,96%	31,98%
	SIM	1,01%	2,20%		
90%	NÃO	96,17%	0,62%	98,20%	23,40%
	SIM	1,18%	2,03%		
95%	NÃO	96,47%	0,32%	98,27%	15,11%
	SIM	1,41%	1,80%		

Tabela 11 - Resultados da avaliação do modelo Random Forest.

5 CONCLUSÕES

Podemos concluir que todos os modelos foram capazes de identificar o perfil dos churners. O objetivo da criação de uma base de possíveis clientes churn no mês seguinte pode ser realizado através da modelagem proposta neste artigo, minimizando os falsos positivo, para que não se faça ofertas para clientes que podem não estar com intenção de realizar o desligamento.

A julgar pela base de validação proposta, o modelo que se ajusta melhor aos dados utilizados neste trabalho é o random forest com 150 árvores e considerando churn todos os clientes com probabilidade acima de 95% de pertencer a essa classe. Este modelo foi capaz de prever corretamente 56% dos 341 clientes churn, com apenas 15% de falsos positivos, ou seja, o modelo foi capaz de fornecer uma base de 225 clientes onde apenas 34 eram falsos positivos. Com o modelo proposto apoiando uma boa estratégia de retenção, pode-se chegar a uma excelente redução de churn na empresa.

6 REFERÊNCIAS BIBLIOGRÁFICAS

CORRAR, Luiz J.; PAULO, Edilson; DIAS FILHO, José Maria. *Análise Multivariada*. São Paulo: Atlas, 2012. 541 p.

GUEDES, M., Rafael, A., Guimarães, L.V *Sistema de identificação de íris utilizando Local Binary Pattern e Random Forest*, 2010.

HAIR, J. F. et al. *Multivariate data analysis*. 5. ed. New Jersey: Prentice-Hall, 1998.

HOSMER, D.; LEMESHOW, S. *Applied logistic regression*. New York: John Wiley & Sons, 1989.

LORENZETT, Cassio Dal Castel;TELÖCKEN, Alex Vinícios. *Estudo Comparativo entre os algoritmos de Mineração de Dados Random Forest e J48 na tomada de Decisão*. 2016. 10 p. Dissertação (Curso de Ciência da Computação)- Universidade De Cruz Alta (UNICRUZ), Rio Grande do Sul, 2016. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/spdc/2016/004.pdf>. Acesso em: 06 dez. 2016.

MCLACHLAN, G. *Discriminant analysis and statistical pattern recognition*. New York: John Wiley & Sons, 1992.

MINUSSI, João Alberto; DAMACENA, Cláudio; NESS JR, Walter Lee. *Um modelo de previsão de solvência utilizando regressão logística*. Rev. adm. contemp., Curitiba , v. 6, n. 3, p. 109-128, Dec. 2002 . Available from <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-65552002000300007&lng=en&nrm=iso>. access on 06 Dec. 2016. http://dx.doi.org/10.1590/S1415-65552002000300007.

MORGAN, G. A.; GRIEGO, O. V. *Easy use and interpretation of SPSS for Windows*: answering research questions with statistics. New Jersey: Lawrence Erlbaum, 1998.

RODRIGUES, Marco A.S. (2006), *Árvores de Classificação* , Monografias da SEIO. Depto Matemática da Univ. dos Açores: Ponta Delgada, www.uac.pt/~amendes.

SHARMA, S. *Applied multivariate techniques*. New York: John Wiley & Sons, 1996.

KRZANOWSKY, W. J. *Principles of multivariate analysis*. Oxford: Clarendon Press, 1988.