INTERFACE TECNOLÓGICA

CHURN RATE: como reduzir em empresas de telecomunicações utilizando aprendizado de máquina

CHURN RATE: how to reduce in telecommunications companies using machine learning

Italo Giullian Carvalho de Albuquerque – italo.giullian@ufabc.edu.br
Universidade Federal do ABC – Santo André – São Paulo – Brasil

Matheus Henrique Bertuci - mateusbertuci@hotmail.com
Faculdade de Tecnologia de Catanduva – Catanduva – São Paulo – Brasil

Bruna Araujo Candeia - bruna.acandeia@gmail.com
Universidade Federal de Pernambuco – Recife – Pernambuco – Brasil

Natália de Oliveira Gomes - natalia.gomes@fachusc.com
Faculdade de Ciências Humanas do Sertão Central – Salgueiro – Pernambuco – Brasil

DOI: 10.31510/infa.v18i2.1183
Data de submissão: 16/08/2021
Data do aceite: 03/11/2021
Data da publicação: 30/12/2021

RESUMO

Churn rate corresponde a taxa de cancelamento quanto ao uso de produtos/serviços oferecidos por empresas. Empresas de telecomunicações, por exemplo, costumam apresentar valores elevados de *churn rate* devido à alta concorrência no setor. Sendo assim, fundamental que, as empresas tenham condições de prever quando e porque um cliente deixará de usar seus serviços, para que se possa realizar ações capazes de evitar ou ao menos minimizar o *churn*. O presente trabalho tem como premissa criar e comparar modelos de previsibilidade de *churn rate* a partir de dados de uma empresa de telecomunicações afim de munir a mesma com informações para a tomada de decisões mais assertivas em ações que minimizem os cancelamentos por parte dos clientes. Foram utilizados dados de uma operadora de telecomunicações com uma grande carteira de clientes. A base de dados foi tratada para implementação e realização de experimentos com algoritmos de aprendizado de máquina e técnicas de mineração de dados. Por fim, sete hipóteses foram testadas as quais demonstraram que: (H1) a maior taxa de rotatividade está em clientes do sexo feminino; (H2) clientes antigos tendem a fazer maior rotatividade; (H3) as mensalidades para quem possui streaming de vídeo são maiores; (H4) os custos tendem a abaixar conforme o passar do tempo de assinatura; (H5) clientes que não possuem dependentes tendem a *churnar* mais do que clientes que possuem dependentes; (H6) clientes que não possuem suporte técnico são mais propícios a *churn*; e (H7) clientes que possuem o serviço de *backup online* tendem a *churnar* menos.

Palavras-chave: Taxa de Rotatividade. Clientes. Dados.

INTERFACE TECNOLÓGICA

ABSTRACT

Churn rate corresponds to the cancellation fee related to the use of products/services offered by companies. Telecommunications companies, for example, tend to have high churn rate values due to the high competition in the sector. Therefore, it is essential that companies are able to predict when and why a customer will stop using their services, so that actions can be taken to avoid or at least minimize churn. The premise of this work is to create and compare churn rate predictability models from data from a telecommunications company in order to provide it with information for making more assertive decisions in actions that minimize customer cancellations. Data from a telecommunications operator with a large customer base were used. The database was treated to implement and carry out experiments with machine learning algorithms and data mining techniques. Finally, seven hypotheses were tested which showed that: (H1) the highest turnover rate is in female clients; (H2) old customers tend to have higher turnover; (H3) monthly fees for those who have video streaming are higher; (H4) costs tend to decrease as subscription time passes; (H5) clients who do not have dependents tend to churn more than clients who have dependents; (H6) customers who do not have technical support are more likely to churn; and (H7) customers who have the online backup service tend to churn less.

Keywords: Churn Rate. Customers. Data.

1 INTRODUÇÃO

Churn rate é um termo em inglês muito utilizado dentro dos departamentos de Sucesso do Cliente (*Customer Success*) das empresas para indicar a taxa de cancelamento dos clientes quanto ao uso dos produtos e/ou serviços oferecidos por elas. Empresas de telecomunicações, por exemplo, costumam apresentar valores elevados de *churn rate* devido à alta concorrência vivenciada diariamente no setor.

Radosavljevik, Putten e Larsen (2010), ao estudarem diversos cenários do *churn rate* em empresas de telecomunicações conseguiram elencar os principais motivos para cancelamento, sendo eles: má qualidade dos serviços prestados, pacotes de benefícios mais vantajosos e clientes com mais de um contato na mesma operadora.

A gestão do *churn rate* possibilita a aplicação de estratégias que contribuem para a retenção dos clientes. Sendo assim, é de grande valia a possibilidade de predizer quando o cliente irá parar de utilizar os produtos/serviços do qual faz uso, ou seja, *churnar* (MONARD; BARANAUSKAS, 2003).

Sendo assim, o presente trabalho tem como premissa a criação e comparação de modelos de previsibilidade de *churn rate* a partir de dados de uma empresa de telecomunicações afim de subsidiar a mesma com informações para a tomada de decisões mais assertivas em ações que minimizem os cancelamentos por parte dos clientes.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Ciência de dados

Segundo Porto e Ziviani (2014), a Ciência de Dados se baseia em teorias e técnicas derivadas de vários campos da ciência. E cada vez mais se mostra imprescindível nos mais diversos setores, tais como saúde, finanças, astronomia, bioinformática, segurança digital, entre outros. Além disso, através das diversas bases de dados, a Ciência de Dados permite a aplicação de técnicas com finalidade de descoberta de padrões e conhecimentos, gerando tendências comportamentais de informações sobre processos (HAN; KAMBER, 2006).

2.1.1 Pré-processamento de dados

O pré-processamento é uma etapa fundamental na Ciência de Dados, o qual corresponde a um conjunto de ações que envolvem preparação, organização e estruturação dos dados. É a etapa que precede a realização de análises e predições. Existem dois principais passos que envolvem este processo: limpeza de dados e transformação de dados (MONARD; BARANAUSKAS, 2003).

Segundo Gomes (2019), a limpeza de dados envolve o manuseio e/ou preenchimento de dados ausentes, redução de ruídos, identificação e remoção de valores desordenado, e a resolução de fragilidade. Já a transformação de dados é realizada para converter os dados originais em modelos mais apropriados e adequados para o processo de mineração.

Além das técnicas de mineração de dados, faz-se necessário utilizar softwares capazes de auxiliar durante o desenvolvimento da análise de dados, e de melhorar a visualização dos resultados extraídos. Com isso, a linguagem de programação *Python*, associada às suas bibliotecas, tem-se tornado cada vez mais utilizada para este propósito (BATISTA, 2003).

2.2 Aprendizado de máquina

O Aprendizado de Máquina é uma área da Inteligência Artificial que objetiva desenvolver técnicas computacionais sobre aprendizado e desenvolvimento de sistemas capazes de adquirir conhecimento de forma automática (MONARD; BARANAUSKAS, 2003). Sendo, o aprendizado por indução um dos mais utilizados Aprendizados de Máquina, uma vez que novos conhecimentos podem ser obtidos a partir de exemplos. No entanto, é um dos aprendizados mais provocantes, visto que o conhecimento obtido pode ultrapassar os

limites das premissas, e não existem garantias de que o conhecimento gerado seja verdadeiro (BATISTA, 2003).

Honda, Facure e Yaohao (2017), dividem o aprendizado indutivo em dois grupos: supervisionado e não-supervisionado. O aprendizado supervisionado é utilizado quando pretende-se prever uma variável específica a partir de uma lista de variáveis independentes. Os dados utilizados para treiná-lo incluem o resultado desejado, ou seja, inclui a variável específica resultante das variáveis independentes observadas. Já o aprendizado não-supervisionado é indicado quando se deseja adquirir uma representação informativa a partir de dados já existentes.

2.2.1 Classificação

Durante o processo de Aprendizado de Máquina é comum atribuir um rótulo para algum tipo de entrada, esse processo é denominado como classificação. Sistemas de classificação são usados, geralmente, quando as previsões são de natureza distinta e tem como objetivo identificar a qual categoria pertence uma determinada amostra do problema. E para tal processo utilizamos de classificadores estruturais os quais são baseados em regras ou distâncias ou, ainda, em redes neurais, como: *Decision Tree*, *Support Vector Machine* (SVM), *k-Nearest Neighbors* (KNN) e *Random Forest* (ROSSI, 2015).

O algoritmo *Random Forest* cria várias árvores de decisão, estabelecendo regras para tomada de decisão. Esse algoritmo apresenta uma estrutura semelhante a um fluxograma, com "nós" onde uma condição é verificada, e se atendida o fluxo segue por um caminho, caso contrário, por outro, sempre carregando ao próximo nó, até a finalização da árvore. Com os dados de treino, o algoritmo procura as melhores condições e onde colocar cada uma dentro do fluxo (ROSSI, 2015). O mesmo autor, ainda destaca que o KNN armazena conjunto de dados inteiro. As previsões são realizadas para uma nova instância observando todo o conjunto de treinamento para as K instâncias mais semelhantes e resumindo a variável de saída para essas instâncias de K.

O conceito original de SVM, consiste em um separador linear definido por hiperplano no espaço de dados de um problema de modo a determinar se os vetores de características fornecidos como entrada fazem parte de uma entre duas classes, positiva ou negativa. Esse separador é gerado a partir da distância máxima entre os pontos que fazem parte de cada uma das classes (ROSSI, 2015). As Neurais Artificiais (RNAs) consiste em técnicas

computacionais que alcançam determinado conhecimento através da experiência e apresentam um modelo inspirado na estrutura neural de organismos inteligentes. Assim, a RNA possui a capacidade de aprender por meio de exemplos (BRAGA et al., 2000).

3 MATERIAIS E MÉTODOS

Para realização desse trabalho, adotou-se uma abordagem quantitativa a qual se concentra na objetividade, uma vez que considera que a realidade há de ser entendida com base na análise de dados (FONSECA, 2002), e descritiva, que vai além do simples reconhecimento da existência de afinidade entre variáveis, e pretendem ajustar a natureza dessa relação (GIL, 2016).

3.1 Fonte de dados e tratamento dos dados

O cerne dessa pesquisa é analisar os dados referente aos clientes por meio do *churn rate* de uma empresa de telecomunicações. Os dados utilizados pela pesquisa consistem na base de dados *"Telco Customer Churn"* com dados de 7043 clientes e 21 recursos alocados, disponível no repositório de base aberta *Kaggle*.

O tratamento destes dados se deu através da verificação da presença de atributos e valores nulos, como também da formatação dos valores dos dados, ou seja, limpeza de dados e transformação de dados, a fim de evitar erros de processamento na etapa de análise como também durante a aplicação nos algoritmos de aprendizado de máquina. A Tabela 1 demonstra os valores iniciais do arquivo.

Tabela 1 - Amostra dos dados utilizados

ID	GÊNERO	ANTIGO	PARCEIRO	DEPENDENTES	POSSE	SERVICE TELEFONE	LINHAS MULTIPLAS	
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	
5575-GNVDE	Male	0	No	No	34	Yes	No	
3668-QPYBK	Male	0	No	No	2	Yes	No	
7795-CFOCW	Male	0	No	No	45	No	No phone service	
9237-HQITU	Female	0	No	No	2	Yes	No	→

ID	SERVICE INTERNET	SEGURANÇA ONLINE	PROTEÇÃO DISPOSITIVO	SUPORTE TÉCNICO	STREAMING TV	STREAMING VIDEO	…	CHURN
7590-VHVEG	DSL	No	No	No	No	No	…	No

5575-GNVDE	DSL	Yes	Yes	No	No	No	...	No
3668-QPYBK	DSL	Yes	No	No	No	No	...	Yes
7795-CFOCW	DSL	Yes	Yes	Yes	No	No	...	No
9237-HQITU	Fiber optic	No	No	No	No	No	...	Yes

Fonte: Autoria própria

3.2 Análise exploratória dos dados

Aos dados foram aplicadas técnicas de análise exploratória. Neste caso, atendendo ao número de variáveis que estão simultaneamente em análise, foram escolhidos os métodos de análise univariada, onde cada variável é tratada separadamente utilizando-se a estatística descritiva (MURTEIRA, 1993); de análise bivariada, para responder as hipóteses estabelecidas; e de análise multivariada, onde as relações se estabelecem entre mais de duas variáveis (REIS, 1997).

Além disso, foi utilizada a técnica *Min Max Scaler* para a re-escala dos dados, ou seja, os dados foram padronizados dentro de uma faixa baseado em alguns critérios e o *One Hot Encode* como método de codificação.

3.3 Modelos de aprendizado de máquina

A solução proposta dividiu os dados em dois grupos: o grupo de treinamento e o grupo de teste. O grupo de treinamento consiste em 70% do conjunto de dados e visa treinar o algoritmo. O grupo de teste contém 30% do conjunto de dados e é usado para testar os algoritmos. Foi utilizado os algoritmos de *Random Forest*, *Decision Tree*, Regressão Logística, SVM, XGBoost, além de uma Rede Neural com 26 neurônios e 50 *epochs*, ou seja, treinado por 50 vezes, utilizando funções de ativação ReLU e Sigmóide, conforme mostrado na Figura 1.

Figura 1 – Treinamento do modelo de rede neural

```
[ ]  # Rede Neural utilizando TensorFlow
     rede_neural = Sequential()

     rede_neural.add(Dense(units=26, activation='relu', input_shape=(39, )))
     rede_neural.add(Dense(units=26, activation='relu'))
     rede_neural.add(Dense(units=26, activation='relu'))
     rede_neural.add(Dense(units=1, activation='sigmoid'))
```

```
rede_neural.summary()
```

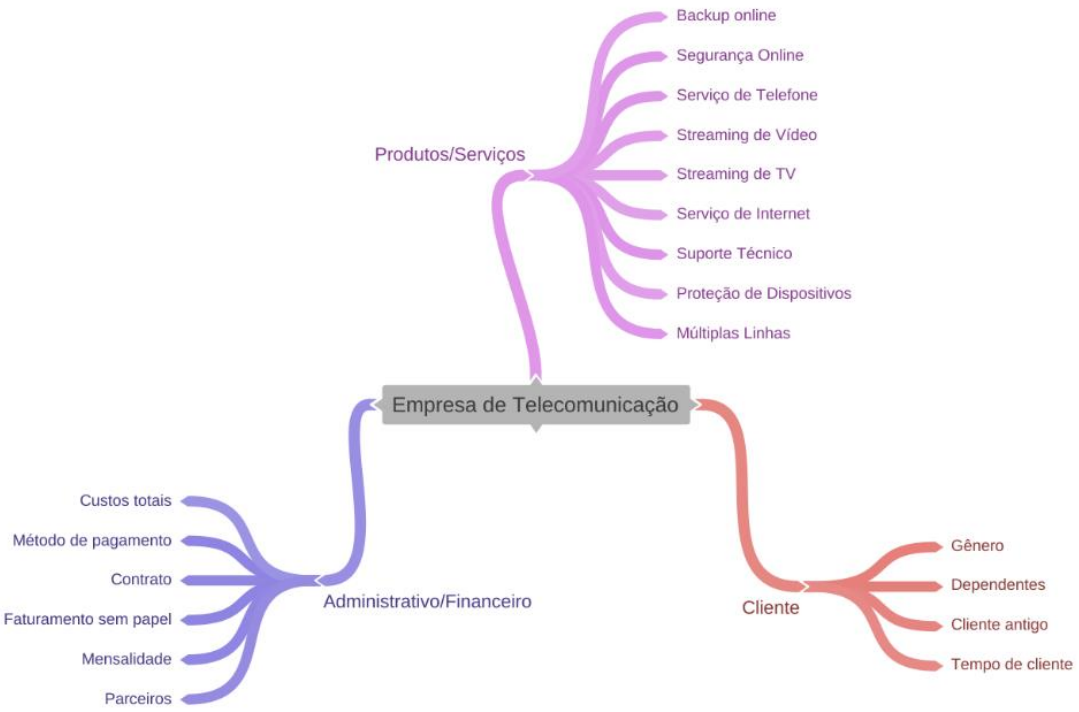
```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 26)                1040
_________________________________________________________________
dense_1 (Dense)              (None, 26)                702
_________________________________________________________________
dense_2 (Dense)              (None, 26)                702
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 27
=================================================================
Total params: 2,471
Trainable params: 2,471
Non-trainable params: 0
```

Fonte: Autoria própria.

3.4 Construção das hipóteses

A partir da análise da base de dados foi possível dividir os atributos em três universos: Cliente; Produtos/Serviços; e Administrativo/Financeiro, como demonstrado na Figura 2.

Figura 2 - Mapa Mental dos atributos



Fonte: Autoria própria.

E levando em consideração os atributos existentes, o universo a qual o atributo está associado e a relação de um ou mais atributos em diferentes universos, 7 hipóteses puderam ser levantadas, sendo elas:

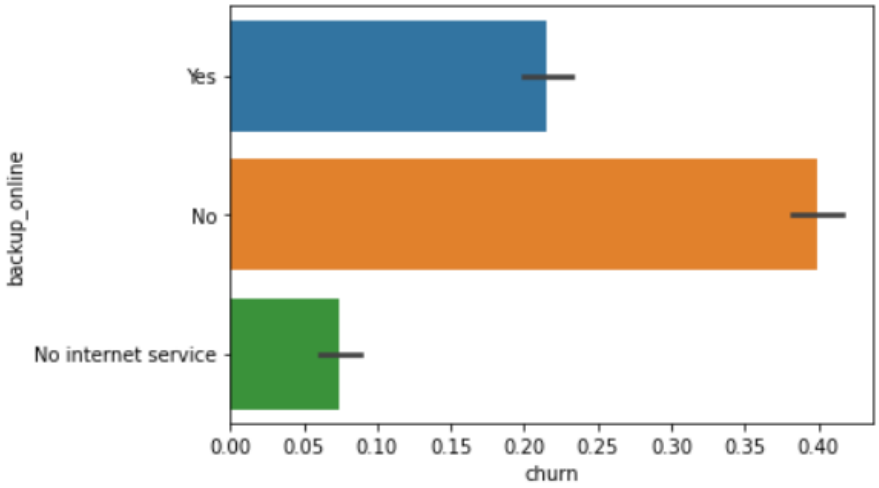
- (H1) Deveria haver mais *churns* entre os homens;
- (H2) Clientes mais antigos deveriam ter menos *churns*;
- (H3) As mensalidades devem ser maiores para quem utiliza streaming de vídeo;
- (H4) Quanto mais antigo o cliente for, menor deveria ser os custos totais do plano;
- (H5) Se tiver dependentes, a tendência é uma maior rotação;
- (H6) Se o plano do assinante possuir suporte técnico, a tendência é que o *churn* aconteça com menos frequência;
- (H7) Se o plano do assinante possuir *backup online*, a tendência é que o *churn* aconteça com menos frequência.

Das 7 hipóteses estabelecidas, a princípio, 3 foram trabalhadas mais profundamente pela pesquisa, são elas: (H5) se tiver dependentes, a tendência é uma menor rotação; (H6) se o plano do assinante possuir suporte técnico, a tendência é que o *churn* aconteça com menos frequência; e (H7) se o plano do assinante possuir *backup online*, a tendência é que o *churn* aconteça com menos frequência.

4 RESULTADOS E DISCUSSÕES

A primeira hipótese testada (H7) foi a de que os clientes que possuem o serviço de *backup online* tendem a *churnar* menos, visto que esse é um serviço essencial em caso de perda e roubo de telefone.

Gráfico 1 – Rotatividade por clientes que possuem Backup Online

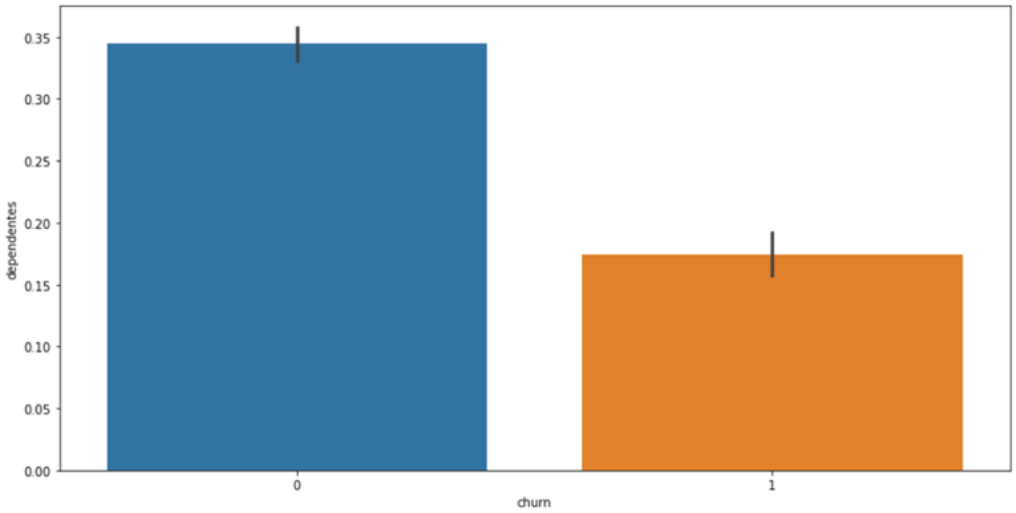


Fonte: Autoria própria.

Aplicando uma análise bivariada, o Gráfico 1 mostra que de fato os clientes que possuem o serviço de *backup online* tendem a rotacionar menos. É possível que isso se de devido a grande preocupação dos clientes com os seus dados.

A segunda hipótese testada (H5) acredita que clientes com dependentes possuem maior tendência a *churnar*. O Gráfico 2, ilustra a segunda hipótese, e é possível perceber que a hipótese é falsa, uma vez que os clientes que não possuem dependentes tendem a rotacionar mais por empresas de telecomunicações do que clientes que possuem dependentes.

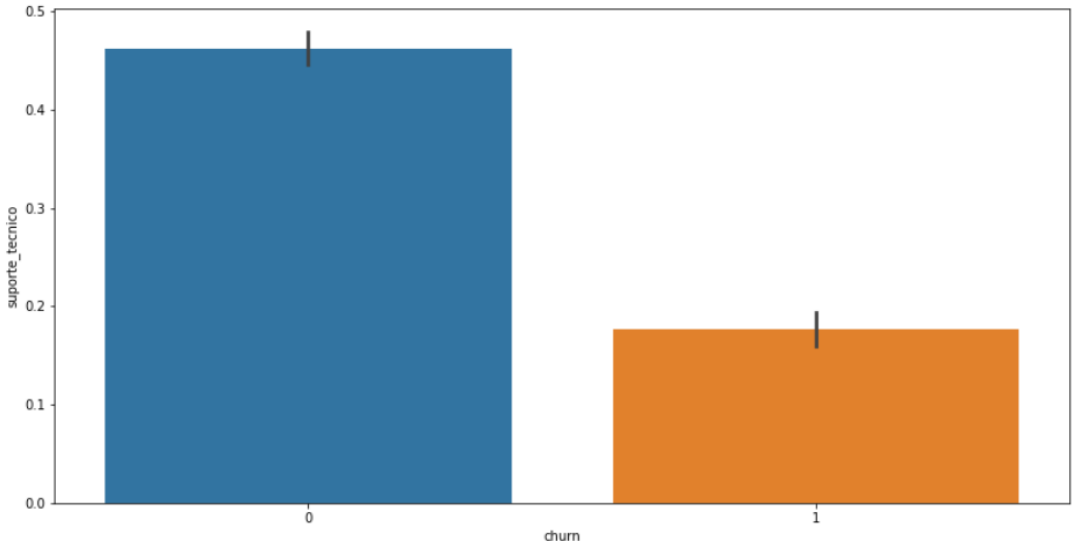
Gráfico 2 – Rotatividade por clientes que tenham dependentes



Fonte: Autoria própria.

A última hipótese testada (H6), verifica se a existência de suporte técnico ao plano contratado minimiza o *churn*, uma vez que o suporte pode vir a auxiliar o cliente em problemas que possam acontecer (Gráfico 3).

Gráfico 3 – Rotatividade por clientes que tenham suporte técnico no plano



Fonte: Autoria própria.

Essa hipótese é verdadeira. O gráfico demonstra a existência de uma maior rotatividade em clientes que não possuem suporte técnico no seu plano.

A tabela 2 apresenta todas as hipóteses testada e os resultados das mesmas.

Tabela 2 – Validação das hipóteses

HIPÓTESES	RESULTADO
(H1) Deveria haver mais *churns* entre os homens;	Falso. A maior taxa de rotatividade está em clientes do sexo feminino.
(H2) Clientes mais antigos deveriam ter menos *churns*;	Falso. Clientes antigos tendem a fazer maior rotatividade.
(H3) As mensalidades devem ser maiores para quem utiliza streaming de vídeo;	Verdade. As mensalidades para quem possui streaming de vídeo são maiores.
(H4) Quanto mais antigo o cliente for, menor deveria ser os custos totais do plano;	Verdade. Os custos tendem a abaixar conforme o passar de tempo de assinatura.
(H5) Se tiver dependentes, a tendência é uma maior rotação;	Falso. Quando não se tem dependentes o *churn* é maior.
(H6) Se o plano do assinante possuir suporte técnico, a tendência é que o *churn* aconteça com menos frequência;	Verdade. A rotatividade é maior para quem não possui suporte técnico no plano.
(H7) Se o plano do assinante possuir *backup online*, a tendência é que o *churn* aconteça com menos frequência.	Verdade. Os clientes que possuem o serviço de backup online tendem a rotacionar menos.

Fonte: Autoria própria.

5 CONSIDERAÇÕES FINAIS

O presente estudo, realizou a criação e comparação de modelos de previsibilidade de *churn rate* a partir de dados de uma empresa de telecomunicações afim de subsidiar a mesma com informações para a tomada de decisões mais assertivas em ações que minimizem os cancelamentos por parte dos clientes.

A partir da definição de um modelo de previsibilidade três hipóteses foram testadas: se o plano do assinante possuir *backup online*, a tendência é que o *churn* aconteça com menos frequência; se tiver dependentes, a tendência é uma menor rotação; e se o plano do assinante possuir suporte técnico, a tendência é que o *churn* aconteça com menos frequência.

Ao colocar em prática o conceito de previsão de *churn* em empresas de telecomunicações, é possível criar campanhas ou ações para reter esses clientes, evitando uma queda no número de conversões e uma baixa no faturamento da empresa. Essa aplicação mostra outro ponto positivo, que é a possibilidade da redução de custos e observação de novas

oportunidades em potenciais clientes, pois quando se consegue ter um uma visão global do comportamento de seus consumidores, torna-se mais fácil predizer suas necessidades e oferecer soluções que atendam essas demandas de forma assertiva. Logo, a aprendizagem de máquina é um aliado indispensável por ser capaz de auxiliar no processo de tomada de decisões, pois fornece previsões aproximadas de realidades futuras, que podem ser analisadas pelos gestores. Essas previsões podem ser utilizadas como estratégia empresarial, pois é possível enxergar uma mudança de comportamento de clientes em determinados tipos de serviços e produtos.

REFERÊNCIAS

Batista, G. (2003). **Pré-processamento de dados em aprendizado de máquina supervisionado**. In: Tese (Doutorado) – Curso de Instituto de Ciências Matemáticas e de Computação, São Carlos.

Braga, A. P. et al. **Redes neurais artificiais: teoria e aplicações**. Rio de Janeiro: LTC, 2000. 250p.

Fonseca, J. (2002). **Metodologia da pesquisa científica**. Fortaleza: UECE.

Gil, A. (2016). **Como elaborar projetos de pesquisa**. 4. ed. – São Paulo: Atlas.

Gomes, P. (2019). **Conheça as etapas do pré-processamento de dados**. Datageeks: https://www.datageeks.com.br/pre-processamento-de-dados/, abril.

Han, J., Kamber, M. (2006). **Data Mining: Concepts and Techniques.** 2. ed. – São Francisco: Elsevier.

Honda, H., Facure, M., Yaoha, P. (2017). **Os três tipos de aprendizagem de máquina.** https://lamfo-unb.github.io/2017/07/27/tres-tipos-am/, abril.

Monard, M., Baranauskas, J. (2003). **Sistemas Inteligentes Fundamentos e Aplicação**. 1. ed. – Barueri: Manole Ltda.

Murteira, B. (1993). **Análise Exploratória de Dados: estatística descritiva**. McGraw Hill.

Porto, F., Ziviani, (2014) “Seminário de Grandes Desafios da Computação no Brasil” – Rio de Janeiro.

Radosavljevik, D., Putten, P. Van Der, & Larsen, K. (2010). **The Impact of Experimental Setup in Prepaid Churn Prediction for Mobile Telecommunications**: What to Predict, for Whom and Does the Customer Experience. Trans. MLDM, 3(2), 80–99.

Reis, E. (1997). **Estatística multivariada aplicada.** Lisboa.

Rossi, R. (2015). **Classificação automática de textos por meio de aprendizado de máquina baseado em redes**. In: Tese (Doutorado) – Universidade de São Paulo, São Carlos.