INSTITUTO FEDERAL DE SANTA CATARINA

LAYSSA ALVES PACHECO

Uso de aprendizagem de máquina para análise de rotatividade de cliente no ramo de telefonia móvel

São José - SC

Julho/2019

USO DE APRENDIZAGEM DE MÁQUINA PARA ANÁLISE DE ROTATIVIDADE DE CLIENTE NO RAMO DE TELEFONIA MÓVEL

Trabalho de conclusão de curso apresentado à Coordenadoria do Curso de Engenharia de Telecomunicações do campus São José do Instituto Federal de Santa Catarina para a obtenção do diploma de Engenheiro de Telecomunicações.

Orientador: Mario de Noronha Neto

Coorientador: Ramon Mayor Martins

São José - SC

Julho/2019

Layssa Alves Pacheco
Uso de aprendizagem de máquina para análise de rotatividade de cliente no ramo de telefonia móvel/ Layssa Alves Pacheco. – São José - SC, Julho/2019-
27 p. : il. (algumas color.) ; 30 cm.

Orientador: Mario de Noronha Neto

Coorientador: Ramon Mayor Martins

Monografia (Graduação) – Instituto Federal de Santa Catarina – IFSC
Campus São José
Engenharia de Telecomunicações, Julho/2019.

I. Orientador. II. Coorientador. III. Local. IV. Título

RESUMO

A rotatividade de clientes nos modelos de negócio baseados na prestação de serviços é uma avaliação delicada da saúde do negócio. A perda de clientes para a concorrência influencia no faturamento e nas decisões de investimento, visto que a aquisição de novos clientes é mais onerosa e custosa do que a manutenção de atuais clientes. Na indústria de telecomunicações, em que o número de clientes é alto e o índice de rotatividade chega a 2,2%, o acompanhamento dessa métrica para escolhas de medidas preventivas é essencial. Nesse contexto, o presente trabalho visa, a partir da técnica de aprendizagem de máquina, traçar o perfil dos clientes das companhias de telefonia móvel com maior tendência à rotatividade, assim como definir quais as principais características de comportamento desse grupo de clientes.

Palavras-chave: Rotatividade de clientes. Aprendizado de máquina. Telecomunicações. Análise de dados.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

SUMÁRIO

1 INTRODUÇÃO

Com o desenvolvimento de tecnologias, e maior acesso à informação, os clientes estão cada vez mais exigentes com relação aos produtos que compram e os serviços que contratam. Empresas com modelos de negócio baseados em prestação de serviços por assinatura são uma das mais afetadas por esse cenário. Isso porque a satisfação dos clientes está intimamente relacionada à retenção da assinatura e, consequentemente, ao faturamento da empresa.

As operadoras de telefonia móvel é um exemplo nesse mercado, pois precisam se reinventar constantemente devido aos avanços da tecnologia, o crescimento da concorrência e volume da base de clientes (FERREIRA, 2005). O direto à portabilidade oferece ao consumidor mais liberdade e poder de escolha, de forma a garantir que a escolha de um fornecedor seja por motivos relacionados aos índices de satisfação do cliente ao invés de dificuldades operacionais, por exemplo (DALVI et al., 2016) (A; NESTOR, 2018).

Diante desse contexto existem duas métricas complementares que são importantes para qualificar e quantificar a saúde desse tipo de modelo de negócio: o índice de retenção de cliente e a taxa de rotatividade de clientes, sendo a última também chamada de *churn*. As medidas são opostas entre si, de forma que a primeira está relacionada com a quantidade de clientes que permanecem consumidores do serviço após um determinado período, enquanto que o *churn* corresponde a quantidade de clientes que cancelaram o serviço em relação a base total de clientes.

Segundo Dalvi et al. (2016), a aquisição de um novo cliente é mais caro do que a manutenção de um atual. Com isso, além do *churn* ser responsável pela redução do fluxo de receita devido ao cancelamento de um serviço, também pode resultar em custo extra caso esse cliente decida retomar a assinatura. No setor de telefonia, em que número de cliente chega a bilhões de contratos e a média de *churn* é avaliada em 2,2% ao mês, a receita perdida pela não gestão da satisfação dos cliente é alta. Por isso é tão importante entender as características que definem um cliente como possível *churn* para as tomadas de decisão de investimento e faturamento (FERREIRA, 2005).

Entender e definir a satisfação do cliente é uma análise complexa e subjetiva, pois envolve critérios relacionados a Qualidade da Experiência (QoE), Qualidade de Serviço (QoS), qualidade do atendimento e da gestão do relacionamento cliente-empresa, identificação com cultura e valores, boa relação preço e serviço prestado, dentro outros índice. Na perspectiva do empresário, é notório que dependendo do perfil do consumidor cada critério tem um peso e importância diferente (A; NESTOR, 2018). Por isso, do ponto de vista estratégico é necessário clareza com relação aos parâmetros prioritários e o impacto deles nos negócios.

Nessa situação, a fim de reduzir custos e otimizar tempo, o uso de inteligência de dados é uma alternativa para mapear os principais comportamentos e os seus impactos nos negócios. Metodologias baseadas em aprendizado de máquina são capazes de detectar padrões de risco com precisão, assim como definir os principais motivos de retenção (LU et al., 2012). Dessa maneira, é possível obter dados confiáveis para a tomada de decisões estratégicas das empresas, assim como segmentar os clientes nas ações de marketing e de gestão de clientes, o que é comprovadamente eficaz na redução de custos, retenção dos clientes (MISHRA; REDDY, 2017) e consequentemente aumento de receita (FERREIRA, 2005).

A realidade das empresas brasileiras em relação ao investimento e dedicação nesse tipo de inteligência nos negócios é recente. Silva, Silva e Gomes (2016) afirma que esse comportamento está em

ascensão na América Latina, mas que ainda é restrito a grande empresas. No Brasil, Silva, Silva e Gomes (2016) explica que 73% das empresas investem, mas com foco em desenvolvimento tecnológico e não para tomadas de decisão.

Com isso, o presente trabalho visa estudar e analisar a gama de consumidores de serviços de operadoras de telefonia móvel brasileira. Para isso, traçar-se-á o perfil dos clientes com relação a padrões de comportamento agrupando-os por similaridade. A partir disso, o objetivo é classificá-los de acordo com a probabilidade de *churn*.

1.1 Motivação

Entender da perspectiva de usuário as necessidades e expectativas de um cliente é de suma importância para qualquer negócio. Com isso, a principal motivação do presente trabalho é unir a inteligência de dados com o ramo dos negócios a fim de fomentar a tomada de decisões estratégicas através do uso de dados e tecnologias de análise.

1.2 Objetivos

O intuito do presente trabalho é entender os desafios do mercado brasileiro de operadoras de telefonia móvel diante do comportamento dos consumidores. Perante a literatura farda e baseada principalmente na realidade de outros países, faz-se necessário o uso de dados nacionais. Com isso, a pesquisa visa aplicar técnicas de aprendizado de máquina já referenciadas em outros estudos para análise de um banco de dados da Agência Nacional de Telecomunicações (ANATEL). O intuito é traçar os padrões de comportamento dos clientes de operadoras brasileiras e classificá-los entre clientes possíveis de *churn* ou não.

1.3 Organização do texto

O texto está organizado da seguinte forma:

- O Capítulo 2 inicia com um estudo da arte de seis artigos apresentando os principais resultados e conclusões da literatura a respeito do assunto. Em sequências há seções focadas em discutir com mais detalhes os temas *churn* e aprendizado de máquina, sendo que o último aborda as três principais técnicas citadas nas referências.
- O Capítulo 3 consiste na proposta de desenvolvimento do presente trabalho e está divido em seção 3.1, a qual define cada etapa de desenvolvimento, seção 3.2, o qual delimita os prazos para as entregas, e seção 3.3, que relaciona as expectativas com os resultados.

2 FUNDAMENTAÇÃO TEÓRICA

O presente capítulo inicia com apresentação de uma revisão geral da literatura com base em pesquisas relacionadas com tema do presente trabalho. O estudo do estado da arte tem o intuito de fundamentar o desenvolvimento do trabalho a partir dos resultados obtidos por outros pesquisados. Além disso, o capítulo aborda os principais conceitos relacionados a temática do trabalho. O estudo bibliográfico foca em explicar o *churn* assim como fundamentar o conhecimento a respeito do aprendizado de máquina.

2.1 Estado da arte

Ao longo da pesquisa bibliográfica é consenso entre os pesquisadores a importância da previsibilidade do *churn* na indústria, principalmente no ramo de telefonia móvel. De acordo com Dalvi et al. (2016), nos últimos anos a indústria de telecomunicações cresceu exponencialmente devido o aprimoramento das tecnologia móveis. Hoje é raro encontrar um pessoa que não tem alguma assinatura com operadoras de telefonia móvel. Com isso, a quantidade de usuários nas operadoras torna a saída de um cliente para o concorrente uma temática em questões financeiras e operacionais de suma importância.

Atualmente, segundo Mishra e Reddy (2017), o principal objetivo das empresas é reter os atuais clientes devido o custo e dificuldade de adquirir novos. Por isso, a relevância em entender os clientes e classificá-los entre possíveis *churn* e não *churn*. Lu et al. (2012) afirma que a indústria de telecomunicações tem uma taxa média mensal de *churn* de 2,2%, ou seja, a cada 50 assinantes ao menos um troca de operadora por mês. Se observarmos da perspectiva de Dalvi et al. (2016), o qual afirma que o ramo de telefonia móvel lida com uma quantidade de cliente em torno de bilhões usuários, o distrato mensal de 2,2% deles representa um recurso financeiro significativo que precisa ser previsto e administrado.

Quando o assunto é relacionado as operadoras de telefonia móvel, Mishra e Reddy (2017) afirma que os celulares tem um papel significativo na sociedade atual. Diante disso, da mudança do mercado e da demanda mundial pelo setor, causadas pelas saturação, competitividade acirrada e a portabilidade, as empresas precisaram e continuando precisando se reinventar, como complementa Dalvi et al. (2016). Isso promove o uso de artifícios como a variedade de ofertas de assinaturas com base em diferentes valores e serviços por parte das empresas, assim como a adoção do *Customer Relationship Management* (CRM).

Dentre as pesquisas bibliográficas, Dalvi et al. (2016) é o autor que mais debate sobre a importância da predição do *churn* em relação ao seus efeitos no mercado. Além disso, o autor ressalta a necessidade de um modelo capaz de fomentar os motivos de *churn* ao invés de apenas agrupar clientes entre os estado de possíveis *churn* e não *churn*. Isso porque, na visão de Dalvi et al. (2016), o marketing da indústria precisa se munir de dados para projetar políticas e táticas para estratégias de retenção. Apenas dessa maneira, de acordo com Dalvi et al. (2016), é possível reduzir custos e danos.

Sem uma lista de clientes identificados de acordo com os principais motivos que os levariam ao *churn*, para Dalvi et al. (2016), as ações de marketing seriam individuais e aleatórias, sem condizer com as estratégias da empresa. Isso porque, trabalhos individuais são custosos, demorados e não oferecem visibilidade do retorno de investimento. Enquanto que, ações de marketing estratégicas com grupos definidos e caracterizados são mais baratas e efetivas. Dessa maneira, aumenta-se a receita de permanência dos cliente e redução de gastos, principalmente na ótica de aquisição de novos clientes, como também concorda Mishra e Reddy (2017).

Complementando a visão de Dalvi et al. (2016), Lopes (2007) aponta estudos que correlacionam

positivamente a relação entre índices de retenção e lucratividade, valor do cliente para a empresa em questões de lealdade e satisfação e aumento do valor da própria empresa. Ou seja, investigar e promover ações de retenção dos clientes é comprovadamente associado a diferentes fontes de lucratividade.

Para gerenciar os clientes existem plataformas de CRM, as quais já são usadas por diferentes setores de prestação de serviço. Dalvi et al. (2016) cita como exemplo de uso as operadoras de cartão de crédito que além de gerir os clientes também usam as informações do CRM para entender e prever comportamentos de inadimplência e *churn*. O CRM consegue produzir um grande banco de dados com ricas informações a respeito dos clientes e sua satisfação. Dessa forma, o CRM possibilita análises em diferentes níveis, como a previsibilidade de *churn* e segmentação de clientes. Além de facilitar o gerenciamento dos consumidores por si só, como o próprio nome da ferramenta já deixa claro. Além disso, A e Nestor (2018) apresenta, além dos pontos anteriores, a perspectiva de que o CRM pode ser usado para identificar oportunidades de melhoria da QoS, qualidade do atendimento e QoE do cliente.

A e Nestor (2018) alega que a retenção de clientes está intimamente relacionado com a QoE e que, portanto, este é o principal influenciador da receita impactada pelo *churn*. A QoE é medida de acordo com a nota atribuída pelo cliente com base na satisfação do serviço prestado. Com isso, A e Nestor (2018) defende o uso de aprendizado de máquina e inteligência artificial para encontrar os padrões de comportamento capazes de influenciar o *churn* e a satisfação dos clientes. A e Nestor (2018) ainda reforça a necessidade e aplica em sua pesquisa o pré-processamento dos dados CRM antes do uso.

Enquanto isso, a proposta desenvolvida por Dalvi et al. (2016) em sua pesquisa uniu mineração de dados com aprendizado de máquina ao utilizar as técnicas de regressão logística e árvore de decisão. A regressão logística é usada para entender até que ponto cada variável extraída afeta a decisão de *churn* do cliente. Enquanto que a árvore de decisão oferece uma visão gráfica geral dos dados disponíveis, a partir dos quais as regras podem ser geradas e as estratégias podem ser construídas para a retenção dos clientes. As principais variáveis avaliadas por Dalvi et al. (2016) foram estado, código de área, tipo de telefone móvel, frequência e período de uso, tempo e custo das chamadas e mensagens, ligações de curta ou longa distância, entre outras informações.

O modelo desenvolvido por Dalvi et al. (2016) é divido em três fases: visualização da análise de desempenho, teste e treinamento. A visualização da análise de desempenho consiste em gerar visualmente os resultados obtidos através das técnicas e ilustrá-los a partir da análise de matriz de confusão. O treinamento é a construção do modelo em si, no qual é feito o treinamento do algoritmo a partir das variáveis desejadas. Já o teste é composto pela construção da lista de clientes com alto risco de *churn* a partir do conjunto de dados de entrada, os quais precisam ter as mesmas variáveis usadas no banco de dados da etapa de treinamento. Após o desenvolvimento, a conclusão de Dalvi et al. (2016) é que o melhor modelo de previsiilidade de *churn* não é aquele com melhor precisão e exatidão estatística, mas o que fornece melhores informações para evitar ainda mais o comportamento de *churn* entre os clientes.

Outro detalhe é que Dalvi et al. (2016) justifica que não usa rede neural no processo devido limitações como o baixo desempenho com banco de dados muito grande, a necessidade de muito treinamento do modelo mesmo quando o bando de dados é pouco volumoso e a ausência de levantamento de dados referentes aos motivos de *churn*. Ou seja, a rede neural classificaria os clientes apenas em relação aos estados de possíveis *churn* e não *churn* e não traria informações úteis de ação com relação ao comportamento dos clientes, que do ponto de vista de Dalvi et al. (2016) é o mais importante. Em contrapartida, a pesquisa desenvolvida e defendida por Mishra e Reddy (2017) usou *Convolutional Neural Network* (CNN), que traduzido significa Rede Neural Convolucional. Além disso, Mishra e Reddy (2017) utilizou um banco de dados da *Center for Machine Learning and Intelligent Systems* (UCI), que é referência em estudos de aprendizado de máquina e possui um repositório de mais de 470 banco de dados para análises de

aprendizado de máquina.

O banco de dados escolhido por Mishra e Reddy (2017) já possui a informação de estado de *churn* dos clientes, o qual é definido entre 0 e 1, sendo 0 clientes não *churn* e 1 clientes *churn*. Com isso, Mishra e Reddy (2017) priorizou e conseguiu avaliar diferentes características da técnica de aprendizado de máquina, as quais o autor anterior não considerava de suma importância nesse tema de pesquisa. Mishra e Reddy (2017) obteve como resultado precisão de 91%, exatidão de 87%, taxa de erro de 13%, sensibilidade de 93% e medida-F1[1] de 92%. Características que são importantes para qualificar os resultados de uma análise como a do atual projeto.

Além disso, a pesquisa de Mishra e Reddy (2017) citou a possibilidade de agrupamento entre os clientes, principalmente em relação ao formato de pagamento. Dentre os pré-pagos e pós-pagos, de acordo com Mishra e Reddy (2017), o último é mais propensos ao *churn* com base em indicadores chave de desempenho nos registros de chamada. Dentre os dados mensurados estão a quantidade de chamadas e mensagens enviadas, o custo pago pelo cliente nos serviços e a frequência e periodicidade de uso ao longo do dia.

Diferentemente, Lu et al. (2012) desenvolveu um algoritmo de aprendizado de máquina baseado em regressão logística com uso do *Adaptive Boosting* (AdaBoost), o qual é um algoritmo usado para aumentar a performance de outros algoritmos. As amostras consistiram em dois banco de dados diferentes, sendo um deles composto por clientes nos quais se deseja saber a probabilidade de *churn* e outro com casos já conhecidos.

Lu et al. (2012) trabalhou com dados de clientes reais de uma companhia e, dentre todas as referências bibliográficas, foi o único estudo que classificou o *churn* em dois tipos: involuntário e voluntário, os quais são melhores abordados na seção 2.2. Outro diferencial da pesquisa de Lu et al. (2012) é que o modelo de desenvolvimento considerou três momentos diferentes no estágio de relacionamento com o cliente. O primeiro momento, chamado de janela do comportamento, consistiu em 3 meses de extração de dados, no qual coletou-se todas as variáveis possíveis. O segundo momento, chamado de janela da implementação, é o período de um mês no qual a empresa realizou ações de marketing com os clientes e, portanto, os clientes que saíram da empresa nesse tempo foram removido do banco de dados. Por último, o terceiro momento, chamado de janela dos resultados, corresponde aos dois meses seguintes, consiste no período de predição do *churn*. Dessa maneira, Lu et al. (2012) trabalha com a previsibilidade de no mínimo dois meses.

Ao contrário de Mishra e Reddy (2017), o qual utilizou variáveis relacionados ao uso o celular do usuário, Lu et al. (2012) lida com fatores como tipo de plano, contrato e fatura como dados de entrada no algoritmo. Aliás, o banco de dados de referência possui 7190 clientes, sendo 678 *churn* e 6512 não *churn*, enquanto que o banco de dados da análise consiste em 700 amostras de clientes.

Nos resultados, Lu et al. (2012) dividiu o banco de dados de referência em dois *clusters*. O *cluster*-1 continha 76% da amostra de cliente e 6% deles foram classificados como *churn*. Já o *cluster*-2 continha 24% da amostra de cliente e 18% deles foram classificados como *churn*. Com isso, o *cluster*-1 foi definido com clientes de baixo risco enquanto os clientes do *cluster*-2 são os de alto risco. No banco de dados de análises, dos 700 clientes, 220 foram classificado na linha de alto risco e 478 na de baixo risco.

Ao final da pesquisa, a fim de aprimorar as ações de retenção, Lu et al. (2012) analisou os dados através de gráficos de *Receiver Operating Characteristic* (ROC) e a partir dos resultados buscou os 10% mais propensos ao *churn*. Com isso, Lu et al. (2012) encontrou que a proporção ideal para trabalhos de

[1] Medida-F1 é a métrica que relaciona sensibilidade com exatidão de forma a indicar a qualidade geral do modelo utilizado. Quanto maior o valor número obtido melhor o modelo usado. O resultado varia de 0 a 1.

retenção são amostras de 3,3% dos *top cluster*-1 e 31% dos *top cluster*-2. Pois, na perspectiva de Lu et al. (2012), a previsão de *churn* precisa gerar uma lista de priorizar clientes de contato, de forma a identificar o motivo do comportamento de um determinado cliente e possibilitar ações estratégicas de marketing.

2.2 *Churn*

O termo *churn* surgiu da junção das palavras da expressão inglesa "*Change and Turn*", que em tradução livre significa mudar e virar (ALMEIDA, 2010). A expressão, que em português é usada como rotatividade de clientes, é utilizada como referência ao processo em que uma empresa perde um cliente para uma empresa concorrente (FERREIRA, 2015).

Em questões numéricas, o *churn* é metrificado a partir da base total de clientes de uma empresa. De forma mais específica, a taxa de *churn* corresponde a porcentagem de clientes que fizeram o distrato no montante de clientes totais em um período de tempo específico. Além disso, o *churn* é classificado em dois tipos diferentes (LU et al., 2012) (FERREIRA, 2015): *churn* voluntário e *churn* involuntário, sendo que o primeiro é subdividido nas categorias de *churn* acidental e *churn* deliberado (ALMEIDA, 2010)(FERREIRA, 2015).

O *churn* involuntário é quando o distrato é feito por parte da empresa. Isso pode acontecer, como explica Ferreira (2015), por inadimplência ou por falta de uso do serviço, sendo que o último é caracterizado pela operadoras de telefonia móvel por clientes em formato pré-pago que não realizam recargas nos últimos três meses.

O *churn* voluntário é quando o distrato é feito por solicitação do cliente e pode ocorrer de duas maneiras:

- *Churn* acidental: quando o cliente, por motivos externos, precisa abandonar os serviços que usufrui. Em geral, esse tipo de *churn* não é da vontade do cliente e acontece por motivos como desemprego ou mudança de região (FERREIRA, 2015). A porcentagem do *churn* acidental é pequena, e insignificante, em relação a taxa de *churn* total da companhia (ALMEIDA, 2010).
- *Churn* deliberado: quando o cliente decide mudar a prestação de serviço para o concorrente. Esse caso, em específico, está relacionado a relação cliente e prestador de serviço, de forma que os fatores para impedir isso está intimamente ligado com as decisões de relacionamento com o cliente que a empresa define. Dessa maneira, o *churn* deliberado é o tipo de *churn* que as empresas mais se dedicam a combater (ALMEIDA, 2010). Dentro dessa classificação, Ferreira (2015) menciona dois subtipos de *churn*, os quais são:
    - *Churn* de canal: quando canais de distribuição e comercialização, os quais atendem os contratos dos clientes finais, decidem redirecionar o cliente de uma empresa para o concorrente ao final de um contrato, isso para que possam garantir uma nova comissão.
    - *Churn* promocional: quando um cliente deseja usufruir das vantagens de ser um novo cliente e migra para a empresa concorrente ao final de um contrato.

Além disso, os principais motivos de *churn* citados por Ferreira (2015) são: a busca dos clientes por tecnologias mais avançadas, acesso a preços melhores, qualidade de serviço superior. Para mais, Ferreira (2015) menciona a influência de fatores psicológicos e sociais, como a influência de familiares e/ou amigos e associação a marca da empresa concorrente, principalmente em relação a cultura e/ou valores.

A avaliação do *churn*, tanto na saúde de um negócio como em questões financeiras e operacionais, se tornar mais preocupantes quanto mais maduro é o mercado. Isso, porque, como descreve Almeida

(2010) na sua pesquisa bibliográfica, quanto maior a maturidade, fluidez e saturação do mercado, mais os consumidores são empoderados para escolher e, consequentemente, mais susceptível a mudança os usuários estão, ameaçando assim o lucro das empresas (FERREIRA, 2015).

Diante desse cenário, Almeida (2010) cita o fato de que as estratégias de retenção de clientes e, consequentemente, ações para impedir o *churn* são as novas tendências de sobrevivência das companhias no mercado em futuro próximo. Por isso é tão relevante o assunto e o investimento da indústria, seja nos departamentos de tecnologia e inteligência de negócio quanto nos projetos de marketing, visto que o mercado tende a ser cada vez mais centrado no cliente ao invés de no produto.

Para simplificar a compreensão do impacto do *churn* na receita e sobrevivência de um negócio, Almeida (2010) apresenta um exemplo simples. O intuito é analisar a receita dos próximos 25 anos de uma empresa fictícia no qual cada cliente tem um contribuição líquida de 50 euros por ano. A receita é avaliada comparando três diferentes taxas anuais de *churn*, sendo tais 6%, 7% e 25%. O resultado pode ser visto na Figura 1.

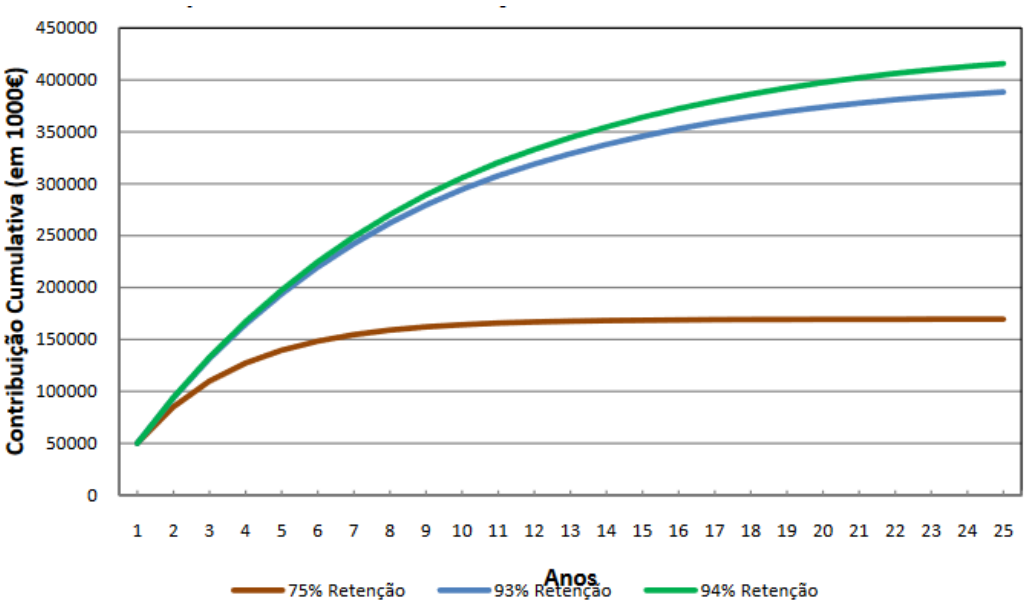

Figura 1 – Impacto de diferentes Taxas Retenção nos resultados financeiros de uma companhia

A empresa com *churn* de 6% tem acúmulo de mais de 400 milhões em 25 anos, enquanto que a de 7% tem valor abaixo disso, mas mais próximo do que a empresa com 25% de *churn*, a qual não chega a 200 milhões. Essa simulação avaliar apenas o *churn* como influência a fim de destacar o impacto do *churn* a longo prazo no faturamento de uma empresa.

Diante desse cenário, Almeida (2010) apresenta duas categorias para as táticas de gestão do *churn*: a gestão não dirigida e a gestão dirigida, sendo a última subdivida em dois tipos. A gestão não dirigida do *churn* consiste em publicidade em massa, ou seja, sem segmentação de clientes e com o intuito de fortalecer a marca da companhia. Ações como essas não são efetivas com relação ao *churn*, visto que não são focadas nas necessidades dos clientes tendenciosos ao cancelamento de contrato e, como já mencionado ao longo desse trabalho, representar perda de receita.

A gestão dirigida do *churn* baseia-se na publicidade segmentada para clientes classificados como tendenciosos ao *churn*. Nessa perspectiva, existem dois tipos: estratégias reativas e estratégias pró-ativas. As estratégias reativas são aquelas em que o cliente comunica o desejo de *churn* à empresa e, em contrapartida, recebe algum inentivo para não abandonar os serviços da empresa, o qual, em geral, é um desconto de

assinatura. Já as estratégias pró-ativas, são aquelas em que a empresa detecta antecipadamente os clientes de alto risco ao *churn* e de antemão promovem ações de retenção.

Com isso, Almeida (2010) afirma que o aumento do valor econômico devido a retenção de clientes pelas empresas está relacionado aos seguintes fatores:

- Redução da necessidade de aquisição de novos clientes e, portanto, diminuição do valor investido no setor (ALMEIDA, 2010).
- Aumento na quantidade de boas referências devido a satisfação dos clientes e, com isso, construir uma forma mais barata de aquisição de clientes (ALMEIDA, 2010).
- Aumento da idade dos clientes na base de clientes e, por consequência, aumento do valor investido pelos clientes na empresa. Isso porque, é comprovado que clientes mais antigos, tendem a investir mais nos serviços que já utilizam (ALMEIDA, 2010).

2.3 Aprendizado de máquina

O aprendizado de máquina é um subgrupo da ciência da computação que se desenvolveu a partir de estudos de reconhecimento de padrões e teorias de aprendizagem computacional dentro da área de inteligência artificial (SCHNEIDER, 2016). Dessa forma, o aprendizado de máquina cresceu com o intuito de construir sistemas capazes de adquirir e reproduzir conhecimento de forma automatizada (MONARD; BARANAUSKAS, 2003).

Nesse contexto, Schneider (2016) explica que o aprendizado de máquina pode ser caracterizado pelo uso de modelos e algoritmos capazes de prever resultados futuros a partir de dados de entrada conhecidos. Não existe um algoritmo ou modelo ideal para todos os cenários, portanto, a escolha do que tem o melhor desempenho dependendo dos objetivos e do contexto de aplicação (MONARD; BARANAUSKAS, 2003).

Segundo (SCHNEIDER, 2016), em relação ao tratamento dos dados de entrada, o aprendizado de máquina é classificado em três tipos: aprendizagem supervisionada, aprendizagem não supervisionada e aprendizado por reforço.

O aprendizado de máquina supervisionado é o qual prevê uma variável dependente a partir de uma lista de variáveis independentes. Nesse contexto, o algoritmo é treinado com um banco de dados que possui a relação entre a variável dependente e as variáveis independentes, de forma que ao analisar o banco de dados que possui apenas as variáveis independentes o algoritmo seja capaz de discernir qual o comportamento ideal com relação a variável dependente (MONARD; BARANAUSKAS, 2003) (LAMFO, 2017).

O aprendizado de máquina não supervisionado é aquele no qual a partir de um banco de informações deseja-se agrupar os dados por padrões de semelhança, conseguindo assim gerar grupos semelhantes e definir características similares. Nesse caso não há históricos para fundamentação e previsão de resultados (MONARD; BARANAUSKAS, 2003) (LAMFO, 2017).

O aprendizado de máquina por reforço consiste em oferecer ao algoritmo informações de ambiente, ações e comportamento por reforço. Com isso, o modelo é capaz de decidir entre as ações a partir do ambiente com base nos diferentes comportamento por reforço, os quais nada mais são do comportamentos anteriores registrados. Esse é o modelo que mais trabalha com incertezas de cenários (MARTINS et al., 2003) (LAMFO, 2017).

De acordo com Schneider (2016), no tocante aos dados de saída, o aprendizado de máquina pode ser catalogado em:

- Associação: a partir de um conjunto de dados o algoritmo é capaz de afirmar qual a associação entre os dados a partir dos atributos (CÔRTES; PORCARO; LIFSCHITZ, 2002).
- Classificação: a partir da análise de um determinado atributo no conjunto de dados o algoritmo é capaz de definir em qual classe cada dado se encaixa (CÔRTES; PORCARO; LIFSCHITZ, 2002).
- Clusterização: a partir de um conjunto de dados o algoritmo é capaz de subdividir o conjunto em subconjunto de caraterísticas homogêneas (CÔRTES; PORCARO; LIFSCHITZ, 2002).
- Regressão: a partir de funções estatísticas o algoritmo é capaz determinar a probabilidade das entradas em relação as classes de saída (CÔRTES; PORCARO; LIFSCHITZ, 2002).
- Análise de sequências: a partir do dado o algoritmo é capaz de relacionar um comportamento futuro consequente do dado inicial (CÔRTES; PORCARO; LIFSCHITZ, 2002).
- Visualização: a partir da visualização dos dados consegue-se afirmar a respeito de características e comportamentos padrões (CÔRTES; PORCARO; LIFSCHITZ, 2002).

Na literatura, as classificações em relação aos dados de saída estão relacionados as técnicas de mineração de dados. A mineração de dados, diferente do aprendizado de máquina, visa descobrir resultados ainda não previstos (SCHNEIDER, 2016).

A partir disso, nas seções seguintes são apresentados três tipos de técnicas que foram abordadas em pesquisas similares ao objetivo do presente projeto, sendo tais: clusterização (subseção 2.3.1), regressão logística (subseção 2.3.2) e árvore de decisão (subseção 2.3.3).

2.3.1 Clusterização

A clusterização é uma técnica de aprendizado de máquina não supervisionado. Isso porque, diante de um conjunto de dados, a análise de *cluster* é capaz de agrupar os elementos por similaridade, sendo que os atributos responsáveis pela similaridade não são conhecidos previamente (DONI, 2004).

Os agrupamentos de saída da clusterização são chamados de *clusters*. Os elementos que compõe um *clusters* são definidos a fim de maximizar a homogeneidade entre si de forma simultânea em que maximizam a heterogeneidade entre os *clusters* (DONI, 2004).

Existem vários tipos de algoritmos de clusterização, contudo os mais usados são a clusterização hierárquica e particionada. O primeiro método processa os *clusters* a partir de *clusters* previamente formados, sendo tais classificados como aglomerados ou divisivos.

A clusterização hierárquica aglomeradora considera cada elemento do bando de dados de entrada como um *cluster*, com isso os aglomerá de acordo com os graus de semelhança. Enquanto que clusterização hierárquica divisora assume que o conjunto de dados de entrada é um *cluster* único que é particionado sucessivamente até o máximo de divisões (GOMES, 2011).

A clusterização particionada, diferentemente da hierárquica, opera através da recursividade dos *clusters* que se formam ao longo do processo. Esse método é capaz de determinar os *cluster* de uma só vez(GOMES, 2011).

2.3.2 Regressão logística

A regressão logística é uma técnica de aprendizado de máquina supervisionado, portanto a partir de variáveis independentes prevê-se variáveis dependentes após o treinamento do algoritmo. Mais especificamente, a regressão logística tem variáveis dependentes do tipo binárias ou categóricas (SCHNEIDER, 2016) e variáveis independentes caracterizadas como categóricas ou contínuas (GEVERT, 2009).

A análise de regressão logística identifica quais variáveis independentes influenciam no resultado da variável dependente e como a influncia acontece. Dessa forma, o modelo da regressão logística mede a relação entre as variáveis estimando a probabilidade entre elas, de maneira a restringir os resultados de saída para um intervalo de probabilidade entre 0 e 1. O cálculo é baseado na Equação 2.1.

$$h(x) = \theta(w^T x), \text{ no qual } \theta(s) = \frac{e^s}{1 + e^s} \tag{2.1}$$

As vantagens da regressão logística são: simplicidade de executar, facilidade de interpretação e nível de acurácia satisfatória considerando o grau de complexidade (SCHNEIDER, 2016).

2.3.3 Árvore de decisão

A árvore de decisão é um dos modelos mais conhecidos de aprendizado de máquina por ser intuitivo e de fácil compreensão e aplicação. Esse método se encaixa nos algoritmos de aprendizado de máquina supervisionado, assim como a regressão logística (NETO, 2018).

O intuito da árvore de decisão é dividir sucessivamente o conjunto de dados de entrada que se encontram na raiz de forma sucessiva até as folhas. Para isso, há os nós internos, os quais classificam os atributos e subdividem o conjunto em subconjuntos. As subdivisões em relação aos atributos continuam até que se atinjam as folhas, que são as classificações alvo (NETO, 2018).

Uma representação da árvore de decisão é apresentada na Figura 2, a qual apresenta um sistema para análise da compra de um computador.

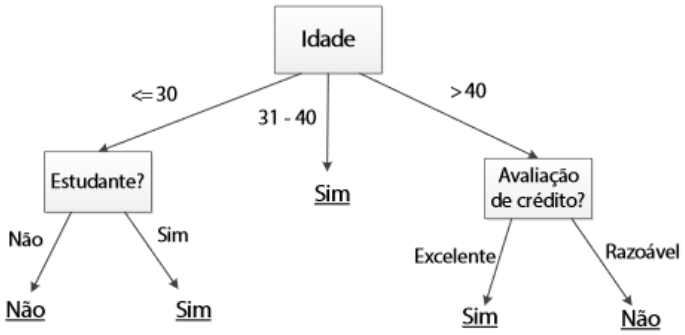

Figura 2 – Árvore de decisão para a avaliação de compra

3 PROPOSTA

O trabalho possui como objetivo traçar o perfil dos clientes de companhia telefônica classificando-os de acordo com características em comum. A partir desse resultado, o intuito é estabelecer quais são considerados possíveis *churn* ou não. Para isso, a presente pesquisa visa usar um banco de dados disponibilizado pela ANATEL que tem informações de mais de 120 mil usuários de operadoras telefônicas brasileiras.

Dentre os dados disponibilidades pela ANATEL há, principalmente, informações de localidade e satisfação de qualidade de serviço. A finalidade é manusear dados referentes a realidade brasileira através de técnicas específicas de aprendizado de máquina.

3.1 Metodologia

A metodologia do trabalho é dividida em três etapas principais, as quais são: a definição e levantamento dos requisitos, o desenvolvimento e testes, e, por fim, as análises e escrita do documento final. A seguir são detalhadas cada etapa.

3.1.1 Definição e levantamento dos requisitos

A etapa de definição e levantamento dos requisitos consiste em estudar e delimitar quais são as informações básicas necessárias dos usuários de telefonia móvel para o desenvolvimento da pesquisa. Para isso, essa etapa exige a leitura de diferentes estudos a respeito da análise de rotatividade de clientes de telefonia móvel.

Ainda nessa conjuntura, essa etapa prevê o estudo e definição da técnica de aprendizado de máquina mais adequada para o desenvolvimento do projeto. A perspectiva é que, assim como para o passo anterior, estude-se diferentes pesquisas a respeito do assunto e, diante do nosso cenário, se escolha a técnica que melhor se aplica.

As principais entregas dessa etapa são:

1. Definição do banco de dados a ser usado.
2. Definição da(s) técnica(s) de aprendizado de máquina a ser aplicada.
3. Listagem das principais características de comportamento a serem avaliadas no banco de dados.

3.1.2 Desenvolvimento e testes

A etapa de desenvolvimento e testes consiste basicamente em desenvolver a técnica de aprendizado de máquina escolhida na etapa anterior e adaptá-la às necessidades do trabalho. Em sequência, o objetivo é testar o que foi desenvolvido com o banco de dados com as informações coletadas.

As principais entregas dessa etapa são:

1. Lista de clientes agrupados por padrões de comportamento.
2. Lista de clientes agrupados através da classificação de *churn* ou não *churn*.

3.1.3 Análises e escrita do documento final

A etapa de análises e escrita do documento final consiste em avaliar os principais resultados obtidos e descrevê-los no documento final do trabalho. Nesse momento, pretende-se comparar os resultados obtidos com os esperados com base do estudo realizado na primeira etapa.

As principais entregas dessa etapa são:

1. Documento final com análises e avaliações dos resultados.
2. Apresentação final do projeto.

3.2 Cronograma

A Tabela 1 apresenta o cronograma com as atividades previstas para o próximo semestre.

Tabela 1 – Cronograma das atividades previstas

Etapa	Meses					
	Jul	Ago	Set	Out	Nov	Dez
A1	√					
A2	√	√				
A3	√	√				
A4		√	√			
A5			√	√		
A6			√	√		
A7				√	√	
A8					√	√
A9						√

- **A1:** Correção e atualização do pré-projeto de acordo com as observações da banca avaliadora.
- **A2:** Estudo e definição do banco de dados a ser usado.
- **A3:** Estudo e definição da técnica de aprendizado de máquina a ser aplicada.
- **A4:** Desenvolvimento do algoritmo de aprendizado de máquina.
- **A5:** Testes com o algoritmo construído e o banco de dados escolhido.
- **A6:** Análise dos resultados e tempo para possíveis correções necessárias.
- **A7:** Escrita do documento final com as etapas de desenvolvimento.
- **A8:** Período de correção e revisão do documento final pelo orientador, co-orientador e banca.
- **A9:** Apresentação final.

3.3 Expectativa de resultados

Ao final do trabalho, objetiva-se a classificação dos perfis dos clientes com maior propensão a rotatividade entre as operadoras de telefonia móvel. Baseado nos resultado obtidos, tem-se a expectativa de discutir os principais gargalos e falhas das companhias que propiciam os clientes a esse comportamento, assim como entender e descrever as principais características que determinam a permanência de um cliente.

O desejo é identificar um paralelo de perfil de cliente entre os que tendem a sair e os que tendem a permanecer com uma operadora de telefonia móvel. De forma que, mediante esse cenário, a pesquisa

também seja capaz de comparar os resultados obtidos com os apresentados nas pesquisas usadas como base para as definições e levantamento de requisitos. O objetivo, nesse ponto, conseguir confrontar as semelhanças e diferenças entre as conclusões dos diferentes estudos e registrar para casos futuros.

REFERÊNCIAS

A, K. O.; NESTOR, D. J. Practical implementation of machine learning and predictive analytics in cellular network transactions in real time. In: . [S.l.: s.n.], 2018. Citado 2 vezes nas páginas 13 e 16.

ALMEIDA, R. M. C. *Classificação de churn no seguro automovel*. Dissertação (Dissertação de Mestrado) — Universidade Nova de Lisboa, 2010. Citado 3 vezes nas páginas 18, 19 e 20.

CÔRTES, S. da C.; PORCARO, R. M.; LIFSCHITZ, S. *Mineração de dados-funcionalidades, técnicas e abordagens*. [S.l.]: PUC, 2002. Citado na página 21.

DALVI, P. K. et al. Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. In: IEEE. *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*. [S.l.], 2016. p. 1–4. Citado 3 vezes nas páginas 13, 15 e 16.

DONI, M. V. Análise de cluster: métodos hierárquicos e de particionamento. 2004. Citado na página 21.

FERREIRA, A. L. L. *Modelo de identificação de churn rotacional nas comunicações moveis*. Dissertação (Dissertação de Mestrado) — Universidade do Porto, 2015. Citado 2 vezes nas páginas 18 e 19.

FERREIRA, J. B. Mineração de dados na retenção de clientes em telefonia celular. *Rio de Janeiro*, 2005. Citado na página 13.

GEVERT, V. G. Análise de crédito bancário com o uso de modelos de regressão logística, redes neurais e support vector machine. *Universidade Federal do Paraná*, 2009. Citado na página 22.

GOMES, B. M. V. *Previsão de churn em companhias de seguros*. Tese (Doutorado), 2011. Citado na página 21.

LAMFO. *Os Três Tipos de Aprendizado de Máquina*. 2017. Disponível em: <https://lamfo-unb.github.io/2017/07/27/tres-tipos-am/>. Acesso em: 27 jun. 2019. Citado na página 20.

LOPES, J. E. F. *Satisfação, lealdade e retenção: um pre-experimento aplicado à telefônia movel*. Dissertação (Dissertação de Mestrado) — Universidade de Uberlâdia, 2007. Citado na página 15.

LU, N. et al. A customer churn prediction model in telecom industry using boosting. *IEEE Transactions on Industrial Informatics*, IEEE, v. 10, n. 2, p. 1659–1665, 2012. Citado 4 vezes nas páginas 13, 15, 17 e 18.

MARTINS, P. S. et al. *Aprendizado de máquina para otimização de parâmetros em sistemas baseados em conhecimento*. Tese (Doutorado) — Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós . . . , 2003. Citado na página 20.

MISHRA, A.; REDDY, U. S. A novel approach for churn prediction using deep learning. In: IEEE. *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. [S.l.], 2017. p. 1–4. Citado 4 vezes nas páginas 13, 15, 16 e 17.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, v. 1, n. 1, p. 32, 2003. Citado na página 20.

NETO, M. V. d. G. O processo crisp-dm aplicado na construção de uma solução para analise de risco de credito. *Universidade Federal dePernambuco*, 2018. Citado na página 22.

SCHNEIDER, P. H. *Análise preditiva de Churn com ênfase em técnicas de Machine Learning: uma revisão*. Tese (Doutorado), 2016. Citado 3 vezes nas páginas 20, 21 e 22.

SILVA, R. A. D.; SILVA, F. C. A.; GOMES, C. F. S. O uso do business intelligence (bi) em sistema de apoio à tomada de decisão estratégica. *Revista GEINTEC-Gestão, Inovação e Tecnologias*, v. 6, n. 1, p. 2780–2798, 2016. Citado 2 vezes nas páginas 13 e 14.