Além da Conexão: Combinando Múltiplas Fontes de Dados para Entender e Prever Evasão de Internet Residencial

Vitor F. Zanotelli[1], Wadham Bottacin[1], Matheus S. De Martin[1], Pedro de Morais[1], Giovanni Comarela[1], Rodolfo Villaça[1], Vinícius F. S. Mota[1], Antonio A. de A. Rocha[2]

[1]Universidade Federal do Espírito Santo

[2]Universidade Federal Fluminense

{vitor.zanotelli, wadham.bottacin, matheus.martin, pedro.i.morais}@edu.ufes.br
{gc, rodolfo.villaca, vinicius.mota}@inf.ufes.br, arocha@ic.uff.br

Abstract. *User retention is an increasing concern among residential Internet service providers due to high competition. This paper proposes to leverage data sources of a multinational telecommunications company for the creation of machine learning models aimed at predicting customer churn. An initial analysis of the data is performed, and different classification models are compared, achieving promising results. The most influential characteristics of a customer's decision to leave are also identified, enabling the use of the proposed solution in strategies to mitigate the problem of customer churn.*

Resumo. *A retenção de usuários é uma preocupação crescente entre os provedores de acesso à Internet residencial. Nesse contexto, este trabalho explora dados internos de uma multinacional de telecomunicações e da Anatel para treinamento de modelos de aprendizado de máquina visando a previsão de evasão de seus clientes. Uma análise inicial mostra o desempenho desses modelos, quem alcançaram resultados de acurácia próxima aos 80%, e precisão e* recall *na faixa dos 70%. Também são identificadas as características mais influentes na decisão de saída de um cliente, viabilizando a implementação da abordagem proposta em estratégias para mitigação do problema de evasão.*

1. Introdução

O acesso à Internet vem se tornando cada vez mais indispensável para a participação do indivíduo na vida moderna. Nesse contexto, nota-se uma demanda crescente pelo acesso residencial à rede, criando um mercado amplo e competitivo onde empresas provedoras de acesso à Internet oferecem seus serviços e disputam a fidelização dos seus clientes.

Em face dessa alta demanda por conectividade, o mercado de provimento de acesso residencial à Internet é altamente competitivo no Brasil. Além das grandes empresas provedoras bem conhecidas no mercado, há ainda um número expressivo de prestadoras de pequeno porte espalhadas pelo país, que representavam 48% do mercado de Internet fixa até março/2022 [Anatel 2022]. Dada essa competitividade, as empresas têm direcionado seus esforços para a fidelização dos seus clientes [Yucesan et al. 2022]. Para elas, é importante não só adquirir clientes novos, mas a retenção de seus atuais clientes se torna uma estratégia importante para garantir sua participação no mercado. O sucesso

dessa estratégia, em geral, é medido por uma métrica chamada *churn*, que representa o total de clientes que cancelam o serviço em um certo período. Assim, a identificação ágil de clientes propensos a cancelar produtos ou serviços passa a ser uma tarefa relevante e desafiadora para essas empresas, uma vez que tal identificação permite a adoção de medidas preventivas para mitigar eventos de *churn* [Pimentel and Goldschmidt 2019].

Os fatores que podem ocasionar a decisão de um cliente em rescindir o contrato com um provedor de acesso são diversos. Problemas relacionados à qualidade do serviço ou experiência podem gerar insatisfação e, quando não solucionados, podem ser motivadores para a migração para outra provedora. A busca pelo melhor custo-benefício, nesse caso o menor preço pela maior banda ou equivalente, também é um motivo comum para essa migração. A satisfação do cliente com o atendimento também pode ser um fator importante para a retenção de clientes. Esses são exemplos de situações em que uma empresa pode influenciar no resultado do *churn*, seja por políticas internas que promovem a execução de um serviço de qualidade, por estratégias de *marketing* ou através do oferecimento de descontos aos novos e antigos clientes. Outros motivos, no entanto, fogem da área de influência da empresa, como, por exemplo, quando um usuário se muda para o exterior ou para alguma cidade onde a empresa não oferece seus serviços.

Assim, é importante que as empresas provedoras sejam capazes de correlacionar eventos de *churn* com os motivos que os ocasionam. Para isso, é necessário implantar estratégias de monitoramento tanto sobre a qualidade do serviço provido quanto ao perfil, nível de satisfação e reclamações dos clientes. Exemplos de informações relevantes são: métricas de desempenho de rede, QoE, tipo e histórico de defeitos em equipamentos, infraestrutura da rede de acesso, planos de dados contratados e histórico de reclamações. Em especial, as informações sobre as reclamações podem ser internas (i.e., reportadas diretamente do cliente para a empresa) ou coletadas de fontes externas, como redes sociais ou na própria Anatel (Agência Nacional de Telecomunicações).

Neste contexto, este trabalho visa correlacionar eventos de *churn* em uma grande empresa provedora de acesso residencial à Internet com uma das dimensões listadas no parágrafo anterior. O interesse é saber se possível prever, utilizando técnicas de aprendizado de máquina: *i)* quais clientes tentarão rescindir o contrato; *ii)* e quais clientes, de fato, deixarão a empresa. São propostos dois tipos de modelos, o primeiro é treinado usando a base interna de dados da empresa, composta por dados numéricos e categóricos referentes a informações do cliente, plano e reclamações internas. Nesse caso, são comparados os resultados alcançados pelos algoritmos de redes neurais, árvores de decisão, floresta aleatória e *gradient boost*. O segundo modelo possui como foco explorar a informação textual das reclamações registradas na Anatel, onde o texto das reclamações são transformados em características utilizando aprendizado profundo.

Os resultados encontrados mostram que a identificação de usuários com alta probabilidade de rescisão de contrato, no cenário de Internet residencial no Brasil, é uma tarefa possível, utilizando tanto informações internas quanto externas. A importância das *features* também é discutida, destacando quais são mais importantes e, por consequência, identificando casos em que a empresa pode melhorar seu serviço para promover a retenção de clientes. Destaca-se a influência da retenção histórica dos planos e de alguns motivos relatados pelos clientes durante reclamações. Outra contribuição do trabalho é mostrar que apenas o conteúdo textual das reclamações realizadas à Anatel possui poder preditivo

equivalente às demais características utilizadas quando as técnicas de pré-processamento e normalização textual adequadas são utilizadas.

2. Trabalhos Relacionados

Algoritmos de aprendizado de máquina, mineração de dados e técnicas híbridas têm sido propostos para classificação de evasão (*churn*) ou retenção de clientes [Lu et al. 2014]. Esses modelos examinam os dados armazenados sobre os perfis dos clientes e suas interações com as organizações, com o propósito de prever duas categorias possíveis: *churn* (evasão) ou não *churn* (retenção) [Óskarsdóttir et al. 2017]. Com esse objetivo, os trabalhos utilizam uma combinação de informações que descrevem os perfis dos clientes, que podem ser categorizadas em três grupos distintos: *i*) dados do cliente, como idade, classe social, volume de gastos, renda potencial e endereço; *ii*) indicadores estatísticos, como duração do relacionamento, média de despesas e métricas de uso de rede; e iii) relacionamento com cliente, baseado em informações textuais extraídas das interações entre os clientes e as empresas [Pimentel and Goldschmidt 2019].

Parte considerável dos estudos relacionados à previsão de *churn* adota uma abordagem que consiste em construir modelos que combinam os dois primeiros grupos de informações. O modelo mais comum entre os estudos é o *Random Forest*, utilizado em diversos artigos ([Lu et al. 2014], [Wu et al. 2021], [Slof et al. 2021], [Bilal et al. 2022], [Caigny et al. 2020], [Stehani et al. 2020], [Bhuse et al. 2020]). Este modelo se destacou pela sua precisão e interpretabilidade, com o [Bilal et al. 2022] alcançando uma acurácia de 93.6% em um dos conjuntos de dados testados. Outro classificador mencionado é o *AdaBoost*, utilizado nos estudos [Wu et al. 2021] e [Bilal et al. 2022]. O *AdaBoost* mostrou-se eficaz no [Wu et al. 2021] alcançando uma pontuação *AUC* (*Area Under the Curve*) de 84%. A Regressão Logística também foi utilizada, com aplicações nos estudos [Wu et al. 2021], [Óskarsdóttir et al. 2017], [Bilal et al. 2022] e [Choudhari and Potey 2018]. O modelo híbrido de Árvore de Decisão com Regressão Logística no [Choudhari and Potey 2018] alcançou uma precisão de 97,18%.

Além disso, técnicas como *Naïve Bayes* e *Decision Tree* foram exploradas nos artigos [Lu et al. 2014], [Wu et al. 2021] e [Stehani et al. 2020], enquanto as Máquinas de Vetores de Suporte (*SVM*) foram aplicadas nos estudos [Wu et al. 2021], [Caigny et al. 2020] e [Stehani et al. 2020]. Por fim, métodos como *K-Nearest Neighbors* e *Deep Neural Networks* foram comparados em [Caigny et al. 2020] e [Bhuse et al. 2020], focando na otimização dos algoritmos com *Grid Search*.

O terceiro grupo foca na extração de informações das interações entre cliente e empresa, destacando a relevância dos dados de atendimento ao cliente, especialmente textuais, na previsão de evasão. Um estudo ([Lalwani et al. 2022]) evidenciou a eficácia de incluir dados textuais em modelos preditivos, com redes neurais convolucionais superando técnicas tradicionais de mineração de texto e alcançando uma *AUC* de 89.8%. No entanto, ressaltou-se que os dados textuais sozinhos não são tão eficientes quanto a combinação com dados estruturados. Em outra pesquisa ([Yucesan et al. 2022]), a análise de sentimentos e a detecção de padrões sequenciais foram integradas, resultando em um aumento na acurácia para 94,5%, significativamente superior aos 74,9% de modelos sem dados de sentimentos.

Estudos adicionais ([Ullah et al. 2019] e [Özköse et al. 2021]) exploraram o uso

de modelos baseados em *BERT* e outras técnicas de *deep learning*. O estudo [Ullah et al. 2019] utilizou um modelo que combina dados estruturados e textuais diversos, incluindo técnicas como *Term Importance* e *Phrase Embedding*, alcançando uma acurácia de 81.2%. Por outro lado, o estudo [Özköse et al. 2021] aplicou o modelo *BERT* para processar comunicações entre empresas e clientes, obtendo uma precisão de 91% e uma pontuação F1 de 90%. Essas pesquisas indicam que a combinação de dados textuais e estruturados, juntamente com o uso de técnicas de aprendizado de máquina como *BERT*, pode aumentar a precisão na previsão de *churn*.

O presente estudo avança na pesquisa de previsão de *churn* no setor de telecomunicações do Brasil com uma abordagem abrangente. Diferente de métodos que se limitam a um tipo de dado, foram combinados uma vasta gama de informações, incluindo características de clientes, indicadores estatísticos, reclamações e o aspecto temporal. Este trabalho é enriquecido pela parceria com uma grande operadora com longa história no Brasil, proporcionando um foco direcionado ao mercado nacional. Foram realizadas todas as análises e o pré-processamento em português, garantindo a relevância dos dados no contexto brasileiro. O estudo explora informações básicas do usuário, de evasão e até reclamações detalhadas na Anatel, agência reguladora independente, abrindo caminho para diversas análises.

3. Conjunto de Dados

3.1. Descrição

Os dados utilizados neste trabalho são referentes a uma grande empresa provedora de acesso residencial à Internet no Brasil[1]. A base de dados disponibilizada pela provedora contém vários conjuntos de dados referentes a departamentos distintos e até mesmo fornecidas por empresas terceirizadas. Para esse trabalho são usados os seguintes quatro conjuntos de informações: informações básicas de Usuário (`BSCS`); informações de reclamações à provedora (`Reclamações`); informações de reclamação na Anatel (`Anatel`); e informações sobre abertura e desfecho de processos de *churn* (`Churn`).

O conjunto `BSCS` registra as informações básicas do contrato de cada cliente da operadora. Parte das informações são relativas ao plano contratado, como: nome, descrição e *status* do plano, data de instalação e chaves identificadoras para uso interno da empresa. A outra parte contém informações relativas à localidade do cliente, como: unidade federativa, cidade, bairro e CEP. Num total, esse conjunto possui 14 colunas e 6.3 milhões de entradas, onde cada entrada representa um usuário contratando um serviço de Internet num instante no tempo.

As reclamações internas de clientes são registradas no conjunto `Reclamações`, podendo um mesmo usuário de serviço não apresentar nenhuma, uma ou múltiplas ocorrências ao longo do seu tempo de contrato. O conjunto apresenta um total de 29 colunas e aproximadamente 250 mil entradas, onde cada entrada representa uma ocorrência de reclamação por parte de um usuário. Dessas colunas, uma parte está relacionada a identificadores internos para a empresa registrar as ocorrências, o restante das colunas referem-se às informações técnicas de infraestrutura da empresa, *hardware* envolvido e,

[1]Por questões de sigilo, os dados utilizados neste trabalho são privados. No entanto, o código utilizado nas análises será disponibilizado quando este texto for publicado.

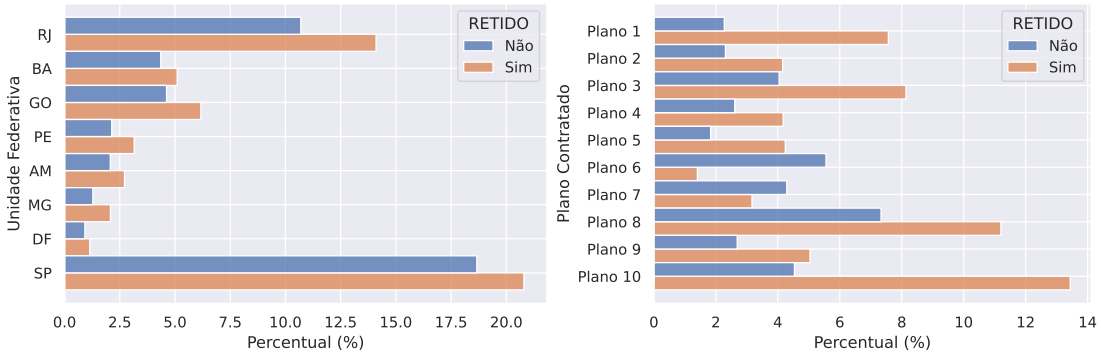

Figura 1. Distribuição dos usuários por UF, considerando a retenção após o processo de *churn* (Esquerda) e dos dez planos com mais usuários (Direita).

por fim, as informações de maior relevância para esse trabalho: os motivos registrados pelo usuário para a reclamação.

O conjunto `Anatel`, apresenta as reclamações realizadas à Agência Nacional de Telecomunicações sobre o serviço prestado pela provedora nos meses de maio, setembro e outubro de 2022. Há um total de 4.346 clientes que registraram reclamações na Anatel. Desses, 1.706 contactaram a empresa, **após** o registro da reclamação junto à Anatel, solicitando encerramento do serviço (i.e., iniciando um evento de *churn*). O conjunto apresenta 45 colunas, muitas dessas esparsas, e destas, são relevantes para este trabalho apenas 13, que representam: informação de localidade, *flag* de risco, tecnologias envolvidas, motivo e o texto livre escrito pelo cliente sobre a reclamação.

Por fim, o conjunto `Churn` registra os eventos de *churn* que ocorreram no período estudado. Um cliente, ao iniciar um evento sinaliza sua vontade de encerrar o serviço, e ao final do processo, em até 30 dias, decide se realizará o desligamento ou manterá o serviço. O conjunto apresenta 260 mil entradas, onde cada entrada é referente a um processo de *churn* de um cliente, sendo que um mesmo usuário pode aparecer uma ou mais vezes (e.g. iniciou o processo três vezes, mas apenas deixou o serviço na última). São 30 colunas, divididas em data de início e fim, informações internas e identificadores de contrato, motivo inicial e final para o desligamento e uma *flag* representando o estado final do processo: pertencem à classe 1 os usuários retidos e à classe 0 os usuários que deixam o serviço. Como o objetivo desse trabalho é estudar o fenômeno de *churn*, apenas os usuários na base `Churn` serão considerados nas análises seguintes.

Com relação ao recorte temporal, os dados fornecidos para `BSCS`, `Reclamações` e `Churn` são relativos a maio, junho, setembro e outubro de 2022. Já para `Anatel`, foram fornecidos, pela provedora, os dados referentes às reclamações feitas nos meses de maio, setembro e outubro de 2022.

3.2. Caracterização

Para a análise e a criação de modelos serão consideradas duas bases de dados unificadas, criadas a partir da junção dos conjuntos apresentados por meio de uma chave identificadora de cliente. A primeira base unificada é denominada `BaseInterna` e contém a interseção das bases: `BSCS`, `Reclamações` e `Churn`. Essa base é utilizada para analisar a saída de clientes utilizando informações internas a empresa. A segunda base, denominada `BaseExterna`, contém a interseção de `Anatel` e `Churn`. Essa base é utilizada para analisar o impacto das reclamações externas à empresa na evasão de clientes.

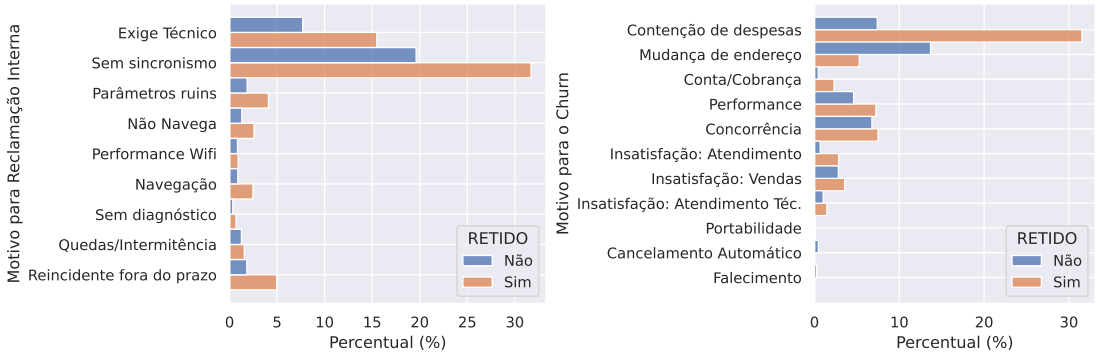

Figura 2. Distribuição dos motivos apresentados pelos usuários. Reclamações internas (Esquerda) e Reclamações para iniciar o *churn* (Direita).

Em relação a `BaseInterna`, após a interseção das bases, o conjunto de dados final possui 120 mil entradas de usuários que iniciaram o processo de *churn*. Desses, aproximadamente 55%, ou 66 mil, são usuários que decidiram não sair da empresa. Um mesmo usuário pode aparecer mais de uma vez na base e estará presente para cada vez que iniciar um processo de desligamento. Cerca de 90% das ocorrências de *churn* são a primeira tentativa de saída de um usuário, a recorrência é rara. Menos de 5% dos usuários abriram reclamações internas antes de tentarem, de fato, deixar o serviço. É bem comum que um mesmo usuário apareça mais de uma vez no `BSCS`, ou seja, usuários costumam mudar de plano ao longo do tempo. Em relação à importância do tempo de contrato de um cliente em um plano específico no *churn*, observou-se que evasões com menos de um ano de contrato são pouco frequentes (possivelmente devido a questões relativas à fidelização). Além disso, metade dos usuários possuem tempo de permanência no plano de até três anos, e valores acima de cinco anos são raros.

Analisando os dados, é possível identificar fatores correlacionados com o fenômeno de *churn*. A Figura 1 ilustra as unidades federativas em que a empresa é mais presente e os dez planos com o maior número de usuários. Em relação à localidade, nota-se que a presença da empresa em alguns estados é consideravelmente maior que em outros, sendo São Paulo e Rio de Janeiro os locais com o maior número de clientes. Também percebe-se, em todos os casos, que se mantêm o padrão global: o número de usuários retidos é maior que o de não retidos. No entanto, essa proporção varia entre os estados. Minas Gerais, por exemplo, apresenta o melhor resultado, com 62% de usuários retidos, e São Paulo apresenta o pior resultado, com 53% de retenção, estando abaixo da média nacional. A distribuição da retenção entre os planos também apresenta resultados similares[2], e é possível notar que existem diferenças tanto no volume de usuários por plano quanto em seu desempenho. Destaca-se, também, que embora a maioria dos planos tenha resultados positivos (maior retenção), foram identificados planos com resultados ótimos, tais como os planos 1, 3 e 10, onde a retenção é bem superior à média. Também foram identificados alguns planos péssimos, tais como os planos 6 e 7, onde o número de usuários retidos é menor que os não retidos.

O motivo que leva um usuário a pedir o desligamento do serviço também apresenta grande influência no resultado do processo de *churn*. A Figura 2 apresenta os motivos mais comuns apresentados nas reclamações internas e nos pedidos de desligamento. A exigência de um técnico para resolver algum problema e a falta de sincronismo são

[2]Os nomes foram substituídos pela sua posição em números de usuários.

Figura 3. Nuvem de palavras das reclamações de clientes na Anatel.

os motivos mais comuns para as reclamações internas, somando mais de 70% dos casos. Entre os motivos internos, nenhum apresenta casos em que a evasão é maior que a retenção, e alguns motivos apresentam resultados de retenção bem acima da média, com valores superiores a 60% e chegando até 70% como: exigir técnico, e parâmetros ruins. Os motivos que levam um cliente a iniciar o processo de desligamento também possuem alguns casos que merecem ser destacados. A contenção de despesas é um motivo facilmente contornável, com mais de 80% de retenção, enquanto as reclamações sobre a conta e cobrança também possuem alta retenção, com 84%. Motivos com maior desligamento também são encontrados, mas são motivos de ocorrência rara, como: mudança de endereço, cancelamento automático e falecimento. Nesses casos a retenção é de 29%, 2% e 3% respectivamente.

O conjunto unificado `BaseExterna` possui 1.706 entradas de dados. Dessas, 874 representam clientes que não foram retidos pela empresa e encerraram o contrato. O número de dados dessa base é consideravelmente menor que as demais, uma vez que o número de pessoas que buscam a Anatel é pequeno. A Figura 3 apresenta a nuvem de palavras gerada a partir de reclamações feitas por clientes da empresa à Anatel, e revela os principais assuntos que preocupam os consumidores da operadora de telecomunicações.

As palavras “cancelamento”, “Internet”, “plano” e “fatura” são as com maior destaque, sugerindo dificuldades dos clientes para cancelar serviços, entender ou contestar cobranças. Reclamações sobre “banda larga” e “internet” também são frequentes, apontando para insatisfação com o serviço de Internet fixa. As menções a “problema”, “serviço” e “atendente” enfatizam a insatisfação com o suporte oferecido aos clientes. Termos como “linha”, “contato”, “multa”, “prazo”, “técnico”, “reparo” e “instalação”, embora menos proeminentes, são igualmente importantes e geralmente se referem a questões contratuais e técnicas.

4. Modelos para predição de *churn*

As caracterizações apresentadas na seção anterior motivam a construção de um modelo de previsão de *churn* com base nas informações fornecidas pela provedora. Neste contexto, o objetivo desta seção é apresentar a metologia utilizada para realizar tal tarefa. Primeiro,

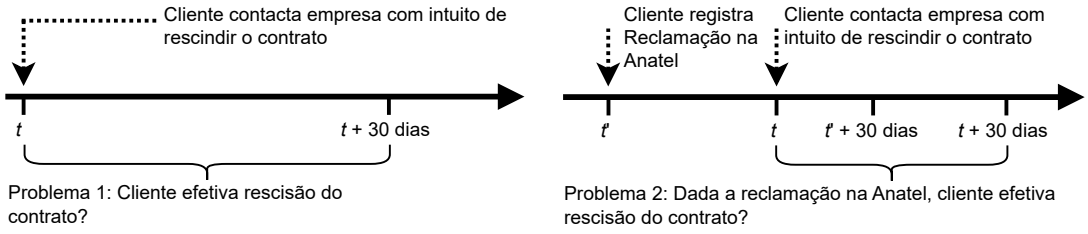

Figura 4. Ilustração dos problemas de predição de *churn*. Problema geral (Esquerda) e Problema para clientes que recorreram à Anatel (Direita).

é apresentada uma formalização do problema de pesquisa a ser resolvido. Em seguida, são apresentados as características extraídas dos dados, os algoritmos de aprendizado utilizados e as métricas utilizadas para validar a metodologia.

4.1. Formalização do Problema de Predição

O objetivo desta seção é formalizar o problema de previsão de *churn*. Dadas as características do conjunto de dados, a tarefa de predição foi decomposta em duas perguntas, uma de propósito mais geral, utilizando a `BaseInterna` e outra para o caso específico dos clientes que também fizeram uma reclamação junto à Anatel, utilizando a `BaseExterna`. A Figura 4 ilustra ambos os casos.

Problema de Predição 1: Dado que um cliente contactou a empresa com o intuito de rescindir o contrato no tempo t, o cliente rescindirá o contrato entre os tempos t e $t + W_1$?

Problema de Predição 2: Dado que um cliente registrou uma reclamação junto à Anatel no tempo t' **e** contactou a empresa com o intuito de rescindir o contrato no tempo $t \in (t', t' + W_2]$, o cliente rescindirá o contrato entre os tempos t e $t + W_1$?

Para abordar os problemas acima, decidiu-se utilizar uma abordagem baseada em aprendizado de máquina. A ideia é transformar os dados referentes a cada cliente em vetores de características (ou *features*) $\mathbf{x} \in \mathbb{R}^d$ e, então, construir um modelo que seja capaz de mapear tal vetor na resposta a ambos os problemas de predição. Com base em experimentos preliminares, nas características/limitações dos dados obtidos e também na natureza da aplicação estudada, os valores de W_1 e W_2 foram definidos em 30 dias.

4.2. Extração de Características

Os dados apresentados na Seção 3 relativos à `BaseInterna` serviram de base para a seleção de várias características (ou *features*) usadas para compôr a entrada dos modelos de aprendizado para o **Problema de Predição 1**. A Tabela 1 apresenta as principais características selecionadas, com uma breve descrição de como elas foram consideradas e tratadas. De forma geral, é importante mencionar que variáveis categóricas foram convertidas para um conjunto de variáveis binárias, usando as técnicas de *one-hot encoding* ou *target encoding* [Micci-Barreca 2001].

O processo de converter em características numéricas o conteúdo textual das reclamações dos clientes feitas junto à Anatel (i.e., `BaseExterna`), usadas no **Problema de Predição 2**, é uma tarefa mais delicada. Há uma variedade de técnicas clássicas para pré-processamento e tratamento de dados textuais. Ao leitor interessado, recomenda-se o texto apresentado em [Baeza-Yates and Ribeiro-Neto 2011]. Neste trabalho, no entanto, optou-se por uma abordagem baseada em aprendizado profundo. Mais especifi-

Tabela 1. Descrição das características não textuais utilizadas.

Característica	Descrição/Tratamento
`plano`	Variável categórica descrevendo o plano contratado pelo cliente. Foi convertida em um *score* usando *target encoding* (representando a fração de clientes retidos em cada plano) e em um peso (inteiro representando o número de clientes contratantes do plano).
`idade_plano`	tempo de contrato do cliente no plano atual.
`n_churn`	número de vezes que o cliente abriu processos de *churn* antes do processo atual.
`n_bscs`	número de planos de Internet diferentes que o cliente já teve com a empresa.
`n_sn`	número de vezes que o cliente fez reclamações diretamente à provedora, antes de abrir o processo de *churn*.
`cep`	CEP da residência do plano contratado. Essa informação foi convertida em três características: o próprio CEP, um *score* e um peso para cada CEP (sendo os últimos dois computados de forma similar a usada para a característica `plano`).
`motivo`	motivos indicados pelo cliente para insatisfação. Os motivos são indicadores categóricos provenientes de reclamações anteriores ou também de indicações feitas na abertura do processo de *churn*. Essa informação foi transformada em característica para o modelo via *one-hot encoding*.
`uf`	Unidade da federação da residência do plano contratado. Essa informação foi transformada em característica para o modelo via *one-hot encoding*.

camente, as reclamações foram transformadas em vetores numéricos (i.e., *embeddings*) utilizando o modelo pré-treinado BERTimbau [Souza et al. 2020].

Antes da aplicação do BERTimbau, verificou-se necessária uma etapa extra de normalização dos textos das reclamações. Essa necessidade foi constatada ao observar uma grande quantidade de reclamações fazendo uso de linguagem informal, contendo gírias e erros ortográficos. Assim, utilizando a biblioteca `enelvo` [Bertaglia and Nunes 2016], as seguintes transformações textuais foram realizadas: correção ortográfica; correção de abreviações; substituição de gírias; ajuste de pontuação; padronização de numerais; e unificação de variantes lexicais.

4.3. Treinamento dos Modelos de Predição

Para a construção do modelo preditivo para o **Problema de Predição 1**, os dados referentes ao mês de maio foram usados para o treinamento dos modelos, incluindo a etapa de cálculos do *target score*, e o mês de junho de 2022 foi usado para teste. Ressalta-se que nenhuma informação do conjunto de teste foi utilizada para a preparação dos modelos.

Os algoritmos de aprendizado considerados foram[3]: *Decision Tree* (DT), *Random Forest* (RF), *eXtreme Gradient Boosting* (XGB) e *MultiLayer Perceptron* (MLP). Os hiperparâmetros dos modelos foram definidos via busca exaustiva, retendo uma parte do conjunto de treinamento para validação.

No caso do **Problema de Predição 2**, dado o pequeno número de instâncias, não foi possível fazer uma divisão temporal do conjunto de dados. Nesse caso, aleatoriamente, 80% dos dados foram reservados para treino do modelo e o restante para teste. Nenhuma informação do conjunto de teste foi utilizada durante o treinamento. Esse processo de divisão entre treino e teste foi repetido 10 vezes para o cálculo de médias e desvios padrão.

[3]Versões presentes nas bibliotecas `scikit-learn` (`https://scikit-learn.org`) e `xgboost` (`https://xgboost.ai`).

Tabela 2. Avaliação dos modelos para o Problema de Predição 1.

		Treino: 05/2022, Teste: 06/2022				Treino: 09/2022, Teste: 10/2022			
Modelo	**Classe**	***Accuracy***	***Precision***	***Recall***	F_1***-score***	***Accuracy***	***Precision***	***Recall***	F_1***-score***
DT	0	0,69	0,65	0,65	0,65	0,65	0,64	0,58	0,60
	1		0,72	0,73	0,72		0,66	0,71	0,68
RF	0	0,78	0,74	0,78	0,76	0,74	0,76	0,64	0,70
	1		0,81	0,78	0,79		0,72	0,82	0,77
XGB	0	0,78	0,74	0,77	0,75	0,75	0,77	0,67	0,72
	1		0,81	0,78	0,80		0,74	0,82	0,78
MLP	0	0,78	0,72	0,86	0,78	0,73	0,76	0,63	0,69
	1		0,87	0,73	0,79		0,72	0,82	0,77

Neste caso, dada a natureza dos vetores de características resultantes do BERTimbau, i.e., os *embeddings*, por simplicidade, apenas modelos baseados em redes neurais foram considerados, mais especificamente, uma MLP. Novamente, os hiperparâmetros da MLP foram escolhidos via busca exaustiva.

Para avaliar os modelos de predição de *churn*, foram utilizadas métricas clássicas de avaliação de modelos para problemas de classificação binária. Neste contexto, a classe 1 indica um cliente retido pela empresa, ou seja, ausência de *churn*. A classe 0 indica um cliente que rescindiu o contrato com a empresa. Especificamente, as métricas consideradas foram *precision*, *recall*, F_1*-score* e *Accuracy*. As definições das métricas são omitidas deste texto por questões de espaço, mas podem ser encontradas em [Zaki and Jr 2014].

5. Resultados

5.1. Problema de Predição 1 – Características Numéricas e Categóricas

Os resultados da avaliação dos modelos treinados, divididos por classe, são apresentados na Tabela 2. Primeiramente, pode-se perceber que a predição de *churn* é uma tarefa possível. Analisando os resultados, pode-se observar que a métrica de acurácia atinge valores relativamente altos, até 78% (lembrando que a base possui aproximadamente 55% de instâncias com rótulos positivos – vide Seção 3). Os resultados das métricas para cada classe são distintos, a precisão dos modelos para a Classe 1 (clientes retidos) é maior que a Classe 0 (clientes não retidos) para a maioria dos casos, variando entre 66% e 87% no primeiro caso e 64% e 76% no segundo. O *recall*, exceto pelo modelo MLP, apresenta também o maior valor para Classe 1, variando de 73% até 82%, enquanto para a Classe 0 este valor varia entre 65% e 86%.

Um segundo aspecto é relacionado à diferença de desempenho entre os algoritmos de aprendizado escolhidos. Em geral, os modelos têm desempenho similar, exceto a árvore de decisão (DT), que gerou resultados inferiores. Esse aspecto dos resultados indica que modelagem do problema (escolha e tratamento das características) é um fator mais importante para tratar a complexidade do problema de predição de *churn* do que a escolha dos algoritmos de aprendizado (desde que modelos adequados sejam utilizados).

Pode-se observar que os resultados da tabela também são diferentes quando conjuntos distintos de treino e teste são considerados. Tal diferença, apesar de pequena, não é negligível. Esse fato indica, novamente, a complexidade do problema estudado. Em outras palavras, evidencia-se que a natureza do fenômeno de *churn* muda com o tempo, e, por isso, os modelos devem ser avaliados e retreinados com certa periodicidade.

Os resultados também apresentam diferenças de desempenho dos modelos entre as

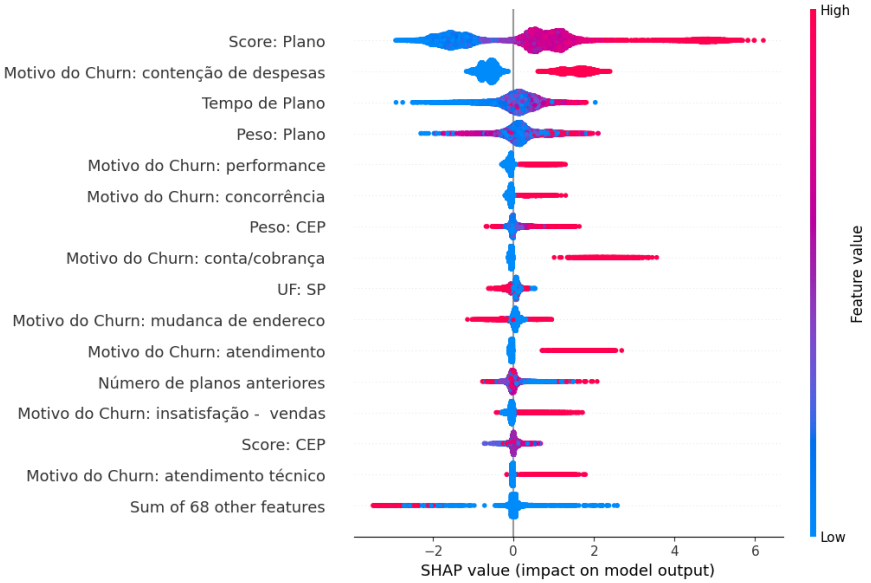

Figura 5. Análise da importância das características na saída do modelo utilizando a metodologia SHAP. Considerando o classificador XGB, conjunto de treino com dados de maio de 2022 e conjunto de teste com dados de junho de 2022. Descrição das características vide Tabela 1.

classes e entre as métricas de precisão e *recall*. Essas diferenças podem guiar/influenciar as decisões de uma empresa em relação as suas estratégias para retenção de clientes. Dependendo do custo de converter um cliente que não seria retido, a precisão ou o *recall* podem ser mais importantes. Se o custo de conversão para um cliente for baixo, por exemplo, dependendo apenas de algum processo automatizado, priorizar o *recall* pode ser interessante. Já em casos onde o custo é alto, como em ações que necessitam da intervenção de algum funcionário, priorizar a precisão pode ser de maior importância.

Do ponto de vista dos gestores de uma empresa, também é importante que, além de utilizar um modelo assertivo, seja possível entender quais são os fatores que de fato levam um cliente cancelar ou não um contrato. Esse entendimento permite que ações preventivas possam ser tomadas para diminuir o *churn* a longo prazo. Nesse sentido, a Figura 5 apresenta uma análise da importância de cada característica para a saída do modelo XGB[4], conforme a metodologia SHAP (*SHapley Additive exPlanations*) [Lundberg and Lee 2017]. Nota-se que as principais *features* responsáveis pelos resultados do modelo estão relacionadas com o plano e motivos para iniciar o processo de *churn*. O *score*, tempo e peso do plano aparecem em primeiro, terceiro e quarto lugares, respectivamente. Planos com *score* alto e clientes com muito tempo de plano possuem uma maior tendência de causar a previsão correta de retenção pelo modelo. Reclamações de performance da rede da operadora também se apresentam como sanáveis, possivelmente com o envio de técnicos para a solução do problema, ou até mesmo com a oferta de *upgrade* de velocidade. Dentre os motivos que apresentam maior relevância tem-se a contenção de despesas, a performance da rede e a concorrência. A contenção de despesas e reclamações de conta/cobrança são motivos que aparentam ser de fácil solução. Possivelmente, sendo o caso em que o cliente recebe alguma promoção ou redução em

[4]Resultados similares para os modelos MLP e RF e para o outro período de tempo analisado. Figuras não apresentadas por questão de espaço.

Tabela 3. Análise do impacto dos atributos textuais. Valores representam médias e desvios padrão (entre parênteses) de 10 repetições do experimento.

Características	Classe	*Accuracy*	*Precision*	*Recall*	F_1-*score*
Categóricas	0	0,68 (0,02)	0,71 (0,03)	0,78 (0,03)	0,75 (0,02)
	1		0,63 (0,03)	0,53 (0,04)	0,57 (0,03)
BERTimbau	0	0,64 (0,03)	0,66 (0,02)	0,79 (0,12)	0,71 (0,06)
	1		0,61 (0,06)	0,43 (0,11)	0,49 (0,04)
BERTimbau + Normalização	0	0,67 (0,02)	0,70 (0,03)	0,76 (0,07)	0,73 (0,02)
	1		0,62 (0,05)	0,54 (0,10)	0,57 (0,06)
BERTimbau + Normalização + Categóricas	0	0,69 (0,01)	0,74 (0,03)	0,74 (0,10)	0,73 (0,03)
	1		0,63 (0,05)	0,62 (0,12)	0,61 (0,01)

sua mensalidade como contraproposta para evitar a sua evasão. Informações relacionadas a localidade possuem um peso menor, mas também estão presentes nas *features* mais importantes. São relevantes o peso e o *score* do CEP, e se o usuário é de São Paulo.

5.2. Problema de Predição 2 – Dados de Reclamações na Anatel

Os resultados do segundo experimento, relativo a `BaseExterna` de dados, também mostram a existência de um sinal entre as *features* baseadas em informações externas e as decisões finais de saída ou não por parte dos clientes. Os resultados de predição por modelos e por classe são apresentados na Tabela 3. A tabela também apresenta quatro diferentes variações do modelo: a primeira considera apenas as variáveis categóricas presentes na `BaseExterna`; a segunda considera apenas as características obtidas via os *embeddings* do BERTimbau pré-treinado aplicado ao texto das reclamações; o terceiro considera os *embeddings* do BERTimbau nos textos normalizados; e, por fim, a quarta variação representa a junção da primeira e da terceira.

Primeiro, pode-se perceber que os resultados são ligeiramente inferiores aos apresentados na Tabela 2. Um segundo ponto importante é que a etapa de normalização dos textos contribui para uma melhoria significativa dos resultados, como evidenciado pelas Linhas 2 e 3 da Tabela 3. No entanto, os melhores resultados obtidos apenas com *features* textuais são equivalentes aos resultados obtidos apenas com variáveis categóricas e aos resultados com a combinação entre *features* categóricas e textuais.

Nesse sentido, é importante mencionar que os modelos da Tabela 3 foram treinados com um conjunto de dados significativamente menor do que os modelos da Tabela 2, com um conjunto de características categóricas diferentes (reduzido), e que a base de dados possui um grau de desbalanceamento diferente (vide Seção 3). Apesar dessas limitações, pode-se perceber que as *features* textuais competem em igualdade com as categóricas. Isso indica que o conteúdo das reclamações feitas à Anatel por si só é uma informação rica para o entendimento do fenômeno de *churn*. Além disso, na presença de um conjunto de dados maior, uma direção promissora para melhorar os resultados é a realização do processo de *fine-tuning* do modelo pré-treinado para o contexto de reclamações e Internet residencial. A obtenção de bases de dados maiores e a realização desse processo são alvo de trabalhos em andamento e futuros.

6. Conclusão

Este trabalho avaliou a aplicação de técnicas de aprendizado de máquina para a geração de modelos capazes de prever a evasão de clientes. Foram consideradas duas bases de dados: i) somente dados internos da operadora; e ii) além dos dados da base interna, foram

acrescentadas informações textuais das reclamações dos clientes registradas na Anatel. Adicionalmente, por meio dos modelos gerados, realizou-se um estudo sobre as características mais importantes que influenciam na solução deste problema.

Em resumo, a partir da avaliação dos modelos treinados neste artigo, pode-se afirmar que os resultados foram satisfatórios, e em se tratando da predição de comportamento humano, atingem métricas relativamente altas, com a acurácia próxima aos 80%, e precisão e *recall* na faixa dos 70%. A introdução de características textuais adiciona ganhos marginais, sendo importante para o desenvolvimento de trabalhos futuros. Outro resultado importante, observado a partir da avaliação em diferentes janelas de tempo, é que a natureza do fenômeno de *churn* muda frequentemente, e, por isso, os modelos devem ser avaliados e retreinados com alguma periodicidade. Para verificar a influência do tempo de forma quantitativa, fontes de dados com períodos mais longos são necessárias.

Como trabalhos futuros, pretende-se expandir a coleta dos dados , de modo a avaliar se isso pode melhorar o resultado dos modelos e gerar um maior entendimento das causas de evasão. A busca por informações distintas, tais como aquelas provenientes de citações à operadora nas redes sociais, também podem ajudar nesse contexto. Modelos mais sofisticados também podem ser testados, desde que seja possível a sua interpretação, pois o entendimento dos fatores envolvidos no *churn* é essencial para a criação de estratégias para tratar esse problema.

Agradecimentos

Este trabalho possui financiamento de: CNPq, CAPES (Código de Financiamento 001), FAPES (#2023/RWXSZ; #2022/ZQX6; #2022/NGKM5; #2021/GL60J) e Fapesp/MCTI/CGI.br (#2020/05182-3).

Referências

Anatel (2022). Infographic: Overview of telecommunication in brazil - march 2022. `https://www.gov.br/anatel/pt-br/dados/relatorios-de-acompanhamento/2022\#R2022\_8`. [Online; acessado em janeiro de 2024].

Baeza-Yates, R. and Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The Concepts and Technology behind Search*. Addison-Wesley Publishing Company, USA, 2nd edition.

Bertaglia, T. F. C. and Nunes, M. d. G. V. (2016). Exploring word embeddings for unsupervised textual user-generated content normalization. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 112–120.

Bhuse, P., Gandhi, A., Meswani, P., Muni, R., and Katre, N. (2020). Machine learning based telecom-customer churn prediction. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pages 1297–1301.

Bilal, F., Syed, Almazroi, A., Abdulwahab, Bashir, Saba, Khan, H., Farhan, Almazroi, A., and Abdulaleem (2022). An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry. *PeerJ Computer Science*, 8:e854.

Caigny, A. D., Coussement, K., Bock, K. W. D., and Lessmann, S. (2020). Incorporating textual information in customer churn prediction models based on a convolutional neural network. *International Journal of Forecasting*, 36(4):1563–1578.

Choudhari, A. S. and Potey, M. (2018). Predictive to prescriptive analysis for customer churn in telecom industry using hybrid data mining techniques. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–6.

Lalwani, P., Mishra, M.K., Chadha, and et al., J. (2022). Customer churn prediction system: a machine learning approach. *Computing*, 104(271–294).

Lu, N., Lin, H., Lu, J., and Zhang, G. (2014). A customer churn prediction model in telecom industry using boosting. *IEEE Transactions on Industrial Informatics*, 10(2):1659.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *SIGKDD Explor. Newsl.*

Pimentel, T. P. and Goldschmidt, R. R. (2019). Sequential sentiment pattern mining to predict churn in crm systems: A case study with telecom data. In *Proceedings of the XV Brazilian Symposium on Information Systems (SBSI'19)*, pages Article 11, 1–8, New York, NY, USA. Association for Computing Machinery.

Slof, D., Frasincar, F., and Matsiiako, V. (2021). A competing risks model based on latent dirichlet allocation for predicting churn reasons. *Decision Support Systems*, 146:113541.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Stehani, S., Karunya, N., Ranjan, D. R. J. B., Sumathipala, S., and Sandanayake, T. C. (2020). Customer churn reasoning in telecommunication domain. In *2020 International Conference on Image Processing and Robotics (ICIP)*, pages 1–5.

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., and Kim, S. W. (2019). A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE Access*, 7:60134–60149.

Wu, S., Yau, W. C., Ong, T. S., and Chong, S. C. (2021). Integrated churn prediction and customer segmentation framework for telco business. *IEEE Access*, 9:62118–62136.

Yucesan, M., Edwine, N., Wang, W., Song, W., and Ssebuggwawo, D. (2022). Detecting the risk of customer churn in telecom sector: A comparative study. *Mathematical Problems in Engineering*, 2022:8534739.

Zaki, M. J. and Jr, W. M. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, USA.

Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., and Vanthienen, J. (2017). Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. *Expert Systems With Applications*, 85:204–220.

Özköse, Y. E., Haznedaroğlu, A., and Arslan, L. M. (2021). Customer churn analysis with deep learning methods on unstructured data. In *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–5.