

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Inteligência Artificial em Análise de Churn em Telecomunicações

Agnaldo Sales Rodrigues de Sá

Trabalho de Conclusão de Curso

MBA em Inteligência Artificial e Big Data

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

**Inteligência Artificial em Análise de Churn em
Telecomunicações**

Agnaldo Sales Rodrigues de Sá

USP - São Carlos

2024

Agnaldo Sales Rodrigues de Sá

Inteligência Artificial em Análise de Churn em Telecomunicações

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo-ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Marcelo Garcia Manzato

USP — São Carlos
2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S163i Sales Rodrigues de Sá, Agnaldo
 Inteligência Artificial em Análise de Churn em
 Telecomunicações / Agnaldo Sales Rodrigues de Sá;
 orientador Marcelo Garcia Manzato. -- São Carlos,
 2024.
 20 p.

Trabalho de conclusão de curso (MBA em
Inteligência Artificial e Big Data) -- Instituto de
Ciências Matemáticas e de Computação, Universidade
de São Paulo, 2024.

1. Churn. 2. Retenção de Clientes. 3.
Inteligência Artificial. 4. Telecomunicações. 5.
Regressão. I. Garcia Manzato, Marcelo, orient. II.
Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

DEDICATÓRIA

*A minha esposa e meus filhos pela
compreensão, carinho e apoio
incontestável.*

RESUMO

Sá, A. S.R S. **Inteligência Artificial em Análise de Churn em Telecomunicações** 2024. 20f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

A indústria de telecomunicações desempenha um papel especial na conectividade global, enfrentando o desafio constante de segmentar, prever e reter clientes em um mercado altamente competitivo. A retenção de clientes é uma medida essencial de sucesso, com o fenômeno do "Churn"—quando um cliente encerra seu contrato—representando uma ameaça significativa à receita e à viabilidade das empresas. Dessa forma esse trabalho pode ser usado para mitigar o Churn usando um pipeline que inclui 4 dos melhores modelos de IA: Usando técnicas de aprendizado de máquina como Logistic Regression, Decision Tree, Random Forest, XGBoos, com esses modelos foram propostos para análise e previsão de churn em empresas de telecomunicações. O desempenho de vários modelos foram comparadas na base no conjunto de dados em estudo e feitas comparações sobre a eficácia dos algoritmos. Esses modelos são frequentemente eficazes em problemas de classificação e regressão, como a previsão de churn. Com a evolução da tecnologia e a crescente demanda do mercado, a inteligência artificial (IA) emerge como uma ferramenta poderosa para enfrentar os desafios do Churn. Esses 4 modelos de Machine Learning Preditiva tem o objetivo prever valores ou categorias de dados futuros com base em dados históricos e para ser usada para personalizar estratégias de retenção, segmentando clientes com base em seu comportamento e preferências individuais. Entretanto, o uso crescente de IA levanta preocupações éticas e de privacidade, exigindo que as empresas garantam transparência e segurança no manejo dos dados dos clientes. Em suma, a aplicação de IA na retenção de clientes está em constante evolução. Este estudo busca uma análise comparativa das abordagens de IA, com foco em modelos de regressão e classificação, para criar estratégias empresariais mais eficazes. A IA promete não apenas mitigar o Churn, mas também fortalecer a estabilidade e a lucratividade das empresas de telecomunicações. Os resultados mudam dependendo dos dados e da configuração específica do modelo, utilizou-se um Pipeline que combina pré-processamento e classificação em um fluxo de trabalho e o GridSearchC que busca pelos melhores parâmetros para o classificador usando validação cruzada, além disso aplicou-se o Undersampling que balanceia o conjunto de dados antes da busca por parâmetros. O **Melhor Modelo** foi o **Random Forest** que foi identificado como o melhor modelo, com a maior acurácia de 80.10%. Ele tem um bom equilíbrio entre precisão e recall, mas como os outros modelos, ainda apresenta desafios em capturar todos os casos de churn, refletido no recall relativamente baixo para a classe 1. Todos os modelos apresentam um desempenho significativamente melhor na classe 0 (não churn) do que na classe 1 (churn). Isso é comum em problemas de churn, onde a classe positiva (churn) é muitas vezes minoritária e mais difícil de prever corretamente.

Palavras-Chave: Churn; Retenção de Clientes; Inteligência Artificial; Telecomunicações.

ABSTRACT

Sá, A. S. R. S. **Artificial Intelligence in Churn Analysis in Telecommunications** 2024. 20f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

The telecommunications industry plays a special role in global connectivity, constantly facing the challenge of segmenting, predicting, and retaining customers in a highly competitive market. Customer retention is a critical measure of success, with the phenomenon of "Churn"—when a customer terminates their contract—representing a significant threat to the revenue and viability of companies. This work aims to mitigate churn by using a pipeline that includes four of the best AI models: Logistic Regression, Decision Tree, Random Forest, and XGBoost. These models were proposed for the analysis and prediction of churn in telecommunications companies. The performance of various models was compared based on the dataset under study, and comparisons were made regarding the effectiveness of the algorithms. These models are often effective in classification and regression problems, such as churn prediction. With the evolution of technology and growing market demand, artificial intelligence (AI) emerges as a powerful tool to tackle the challenges of churn. These four predictive Machine Learning models aim to forecast future values or categories based on historical data and are used to personalize retention strategies by segmenting customers based on their behavior and individual preferences. However, the increasing use of AI raises ethical and privacy concerns, requiring companies to ensure transparency and security in handling customer data. In summary, the application of AI in customer retention is constantly evolving. This study seeks a comparative analysis of AI approaches, focusing on regression and classification models, to create more effective business strategies. AI promises not only to mitigate churn but also to strengthen the stability and profitability of telecommunications companies. The results vary depending on the data and the specific model configuration; a pipeline was used that combines preprocessing and classification in a workflow, and GridSearchCV was applied to search for the best parameters for the classifier using cross-validation. Additionally, undersampling was applied to balance the dataset before the parameter search. The best model identified was Random Forest, with the highest accuracy of 80.10%. It has a good balance between precision and recall, but like the other models, it still faces challenges in capturing all churn cases, reflected in the relatively low recall for class 1. All models perform significantly better on class 0 (non-churn) than on class 1 (churn). This is common in churn problems, where the positive class (churn) is often a minority and harder to predict correctly.

Keywords: Churn; Churn; Customer Retention; Artificial Intelligence; Telecommunications.

LISTA DE FIGURAS

Figura 1. Exemplo de gráfico usado para explorar a distribuição de variáveis	24
Figura 2. Resultados sem HiperParâmetros	27
Figura 3. Resultados com HiperParametros	28

LISTA DE GRÁFICOS

Gráfico 1. Gráfico Acurácia dos 4 modelos de Comparação antes e depois dos hiperparâmetros	28
Gráfico 2. Desempenho do melhor modelo, Random Forest Com e Sem otimização de hiperparâmetros, para comparar a Acurácia, Precisão, Revocação e F1-score entre os dois modelos.	29
Gráfico 3. Desempenho do melhor modelo, Random Forest Com e Sem otimização de hiperparâmetros, para comparar a Acurácia, Precisão, Revocação e F1-score entre os dois modelos.	29
Gráfico 4. Matrizes de comparação entre antes e depois da hiperparametrização	31
Gráfico 5. Análise da accuracy, precisão, recall e socre	34

SUMÁRIO

1 INTRODUÇÃO	10
2 REFERENCIAL TEÓRICO	13
2.1 Churn em Telecomunicações	13
2.2 Aplicação da Inteligência Artificial na Previsão de Churn	14
2.2.1 <i>Modelos de Machine Learning</i>	15
2.3 Pré-processamento de Dados e Balanceamento de Classes	16
2.4 Implicações e Benefícios para as Empresas de Telecomunicações	18
3 METODOLOGIA	21
3.1 Área de Estudo	21
3.2.1 <i>Preparação dos Dados</i>	21
3.2.2 <i>Análise Exploratória dos Dados (EDA)</i>	21
3.2.3 <i>Divisão dos Dados</i>	22
3.2.4 <i>Pré-processamento dos Dados:</i>	22
3.2.5 <i>Treinamento e Avaliação dos Modelos:</i>	22
3.2.6 <i>Ajuste de Hiperparâmetros:</i>	23
4 RESULTADOS	24
4.1 Resultados Sem HiperParametros	24
4.2 Resultados Com HiperParametros	26
4.3 Comparação	27
4.4 Discussão	33
6 CONCLUSÃO	36
REFERÊNCIAS	37

1 INTRODUÇÃO

A indústria de telecomunicações, enquanto protagonista incontestável na conectividade global, tem como objetivo a segmentação predição e retenção de clientes que é uma peça vital no tabuleiro da competitividade: a imperativa expansão de sua base de clientes contra o desafio persistente de segmentação predizer e reter esses clientes em um cenário altamente competitivo. A retenção de clientes, é uma medida de sucesso no setor, emerge como uma fronteira crítica onde estratégias eficazes se transformam em vantagens competitivas sustentáveis. No centro deste desafio reside o “Churn”, um fenômeno que ocorre quando um cliente encerra seu contrato com uma empresa. Esse fenômeno pode ter um impacto significativo nos negócios das empresas, ao podar levar a perda de receita e até a inviabilidade da empresa continuar operando (Barsotti et al., 2024).

O Churn pode ser causado por diversos fatores, como: descontentamento com o produto ou serviço, os clientes podem estar insatisfeitos com a qualidade do produto ou serviço oferecido pela empresa, com o atendimento ao cliente ou com as condições do contrato. Ofertas mais competitivas, os clientes podem encontrar ofertas mais competitivas de outras empresas (AlShourbaji et al., 2023). Mudanças de hábitos: os clientes podem mudar seus hábitos de consumo e precisar mais dos serviços oferecidos pela empresa (Sulikowski; Zdziebko, 2021).

O gerenciamento do Churn é uma atividade importante para essas empresas, elas podem adotar diversas estratégias para reduzir o Churn, como: melhorar a qualidade do produto ou serviços: as empresas podem investir em melhorias na qualidade dos produtos ou serviços oferecidos, para os clientes estarem satisfeitos. Oferecer melhores condições de contrato, oferecendo condições de contrato mais vantajosas para os clientes, para que eles sejam menos propensos a trocar de empresa. Implementar programas de fidelização, implementando programas de fidelização para recompensar os clientes que permanecem fiéis à empresa (Sulikowski; Zdziebko, 2021).

A dinâmica vertiginosa do mercado, com suas demandas crescentes e a incessante evolução da tecnologia, exige uma reavaliação constante das estratégias de retenção tradicionais (Kim; Lee, 2021). Neste contexto, a aplicação da inteligência artificial (IA) surge como um farol orientador, prometendo não apenas mitigar as ameaças do Churn, mas revolucionar como a fidelização de clientes é concebida e implementada (Mustafa et al., 2021).

Nos últimos anos, o uso de IA para prever o churn de clientes em empresas de telecomunicações aumentou significativamente. Algoritmos de Machine Learning, como redes neurais, árvores de decisão e métodos ensemble, são aplicados com sucesso para analisar grandes conjuntos de dados e identificar padrões que indicam a probabilidade de um cliente cancelar seu contrato. A previsão de churn usando IA tornou-se uma área de pesquisa muito importante em telecomunicações (Kim; Lee, 2021). Esses modelos são amplamente explorados para determinar a possibilidade de os clientes encerrarem seus contratos com as empresas de telecomunicações. Eles são treinados em grandes conjuntos de dados históricos, incorporando uma variedade de variáveis, como comportamento do cliente, padrões de uso e histórico de interações (Ahmad et al., 2019).

A relevância incontestável desta investigação se delineia a partir da sua contribuição prospectiva ao panorama global de telecomunicações. Enquanto estudos pregressos delimitaram caminhos, este empreendimento planeja transcender, proporcionando não apenas uma avaliação comparativa tangível, mas também delineando os elementos latentes que conferem singularidade a cada abordagem (Saha et al., 2023). Aprofundar o entendimento sobre como algoritmos específicos se relacionam com a complexidade dos dados de clientes permitirá uma adaptação mais refinada das estratégias de retenção, culminando em um ecossistema resiliente e dinâmico (Ribeiro et al., 2023).

Uma abordagem cada vez mais comum é a personalização de estratégias de retenção com base no comportamento e nas preferências individuais dos clientes. A IA possibilita uma segmentação precisa do mercado e a criação de campanhas de retenção altamente direcionadas, que aumentam a eficácia e reduzem os custos. A personalização baseada no perfil de comportamento do cliente é cada vez mais valorizada. A IA é utilizada para segmentar os clientes em grupos mais específicos, permitindo ações de retenção singulares e direcionadas. Isso inclui a análise de padrões de uso, preferências de serviço e histórico de interações para adaptar as ofertas e comunicações às necessidades individuais de cada grupo de clientes (Sikri et al., 2024).

As empresas de telecomunicações estão aproveitando a IA para analisar grandes volumes de feedback dos clientes, incluindo mídias sociais, e-mails e interações de call center. Essa análise de sentimentos permite identificar rapidamente problemas e preocupações dos clientes, respondendo proativamente para evitar o churn. A IA é empregada para analisar o retorno dos clientes em várias plataformas, avaliações online e interações de atendimento ao cliente. Essa análise é usada para identificar problemas dos clientes, permitindo uma resposta proativa para resolver questões e melhorar a experiência dos clientes (Chang et al., 2024).

A automação de processos por meio de IA é implementada para fornecer respostas rápidas e personalizadas aos clientes em momentos críticos. *Chatbots* e assistentes virtuais são usados para resolver consultas e problemas de maneira eficiente, aumentando a satisfação do cliente e reduzindo a probabilidade de churn. A automação desempenha um papel crucial na retenção de clientes, com assistentes virtuais configurados para fornecer respostas instantâneas e personalizadas, agilizando pedidos e resolvendo as dificuldades. Isso eleva o bem-estar do cliente e diminui o tempo de resposta (Peng et al., 2023).

A IA permite a integração de dados de várias fontes, incluindo dados transacionais internos, dados comportamentais dos clientes e dados externos, como informações demográficas e econômicas. Essa abordagem holística melhora a precisão das previsões de churn e permite uma compreensão mais profunda do cliente. A IA integra dados complexos, internos e externos, e a análise combinada desses dados fornece previsões mais precisas de churn e um entendimento mais refinado do comportamento dos clientes propensos a deixar a base (Kim; Lee, 2021).

Com o aumento do uso de IA na retenção de clientes, surgem preocupações éticas e de privacidade. As empresas precisam garantir a transparência e o consentimento dos clientes no uso de seus dados, além de assegurar a segurança e proteção dessas informações contra acessos não autorizados. Embora a IA ofereça oportunidades significativas na retenção de clientes, também apresenta desafios éticos e de privacidade. As empresas devem garantir a conformidade com a LGPD para proteger os clientes no uso de seus dados, bem como manter esses dados seguros contra acesso não autorizado e uso indevido (Mustafa, Ling e Razak, 2021).

A retenção de clientes na indústria de telecomunicações, com foco na aplicação de IA é uma área em constante evolução. Desta forma, este estudo propôs uma incursão analítica meticulosa na interseção crucial entre a IA e retenção de clientes em telecomunicações, concentrando-se na avaliação comparativa de proeminentes abordagens: sendo 5 Modelos de Regressão e 2 de Classificação. O escopo abrangente desta pesquisa transcende a mera aplicação de algoritmos; busca, em sua essência, discernir nuances algorítmicas para informar estratégias empresariais substanciais. Nesse contexto, a IA, com sua capacidade de analisar grandes volumes de dados, surge como uma aliada promissora para antecipar e mitigar o *Churn* prevendo e segmentando e com isso aumentando a estabilidade e a lucratividade das empresas do ramo.

2 REFERENCIAL TEÓRICO

2.1 Churn em Telecomunicações

O *churn* é caracterizado pela sua capacidade de rotatividade de clientes, tendo se tornado um fenômeno crítico no setor de telecomunicações, refletindo a decisão de um cliente de encerrar seu relacionamento com a empresa. Segundo Barsotti et al. (2024), este comportamento acaba trazendo problemas financeiros devido, justamente, a perda de clientes que reduz a receita e pode, a longo prazo, ameaçar a sustentabilidade das empresas, especialmente em mercados competitivos onde as margens de lucro são estreitas.

O termo *Customer Churn* refere-se à saída de clientes de uma empresa de serviços. Isso também é conhecido como atrito, deserção ou rotatividade do cliente. Esse fenômeno, fisiológico para qualquer negócio de serviços, é de grande relevância para as empresas, e sua quantificação entra em um dos principais indicadores de desempenho típicos, a taxa de churn. A rotatividade de clientes tem um impacto substancial em muitos aspectos do negócio de uma empresa e implica custos de oportunidade: não só diminui as vendas da empresa, mas também não é fácil compensar atraindo novos clientes. Na verdade, a aquisição de um novo cliente é geralmente de 5 a 6 vezes, ou mesmo até 20 vezes, mais cara do que reter uma existente. Por esse motivo, as empresas adotam estratégias de retenção de clientes e a maioria delas depende da capacidade de prever quais clientes estão em risco de partida (Barsotti et al., 2024, p.403).

Para Sulikowski e Zdziebko (2021), se deve compreender e mitigar o *churn* justamente por, nos últimos anos, ter sido um fenômeno comum nos setores que pagam altos custos na aquisição de clientes. Porém, as razões para o *churn* do cliente podem ser divididas em dois tipos principais: obrigatório e voluntário. O primeiro é normalmente causada por circunstâncias súbitas, devido aos problemas financeiros do cliente, realocação, ou devido a uma oferta mais atraente fornecida pela concorrência; a voluntária é causada por fatores, que podem ser abordados diretamente pelo provedor, por exemplo, má qualidade do cliente serviço, “decepção com a qualidade do serviço, reclamações rejeitadas, maus-tratos ao cliente com temporário Pagamentos em atraso, etc” (Sulikowski; Zdziebko, 2021, p.4801).

Complementarmente, a competitividade cada vez mais acirrada do setor, faz com que o *churn* ascenda, onde empresas frequentemente engajam-se em guerra de preços e promoções agressivas para capturar a base de clientes dos concorrentes. Ofertas mais atrativas de outras operadoras, tais como planos de dados ilimitados, descontos em aparelhos, ou pacotes combinados com outros serviços, são poderosos incentivos para a mudança de fornecedor por parte do cliente (AlShourbaji et al., 2023). A dinâmica do mercado, portanto, exige que as

empresas de telecomunicações não apenas monitorem as ações dos concorrentes, mas também adaptem suas estratégias de retenção de maneira proativa para prevenir a perda de clientes.

Voltando as mudanças nos hábitos de consumo dos clientes, a crescente consciência do consumidor sobre questões de privacidade e o uso de dados pessoais impulsionou a demanda por maior transparência e controle sobre como as informações são coletadas e utilizadas pelas empresas, impactando diretamente as percepções de valor e lealdade à marca (Mustafa, Ling e Razak, 2021).

Para mitigar o *churn*, as empresas de telecomunicações têm adotado estratégias, que vão desde a melhoria da qualidade do serviço até a implementação de programas de fidelização e ofertas personalizadas. A personalização, facilitada pelo uso de tecnologias avançadas de análise de dados, permite que as empresas ajustem suas ofertas e serviços às preferências individuais dos clientes, aumentando a satisfação e, por consequência, reduzindo a probabilidade de *churn* (Chang et al., 2024). Essas estratégias são apoiadas pela crescente adoção de inteligência artificial (IA) e machine learning para prever o comportamento de *churn* e desenvolver intervenções proativas. Da mesma forma:

A fim de neutralizar a rotatividade, os operadores devem se preocupar com a lealdade do cliente e construir relacionamentos individuais, começando desde fornecer-lhes ofertas adaptadas às suas necessidades, proporcionando bônus de fidelidade, medindo o nível de satisfação etc. Essa abordagem orientada à lealdade é ditada pelo fato de que o custo de aquisição de clientes é geralmente muito mais alto do que manter um já existente. Graças à identificação bem-sucedida de clientes do grupo de risco, o operador pode procurar maneiras de diminuir o taxa de churn. No caso de clientes classificados como de alto risco de saída, o operador pode tomar medidas impedindo que aconteça (Sulikowski; Zdziebko, 2021, p.4801).

Contudo, apesar das inúmeras iniciativas para combater o *churn*, os desafios permanecem, especialmente em relação à integração de novas tecnologias e à gestão de dados de clientes de maneira ética e conforme as regulamentações de proteção de dados (Peng et al., 2023).

2.2 Aplicação da Inteligência Artificial na Previsão de Churn

A aplicação da Inteligência Artificial (IA) na previsão de *churn* vêm se expandindo devido à sua utilização dos setores que dependem da receita, como telecomunicações, serviços financeiros, e-commerce e assinaturas digitais. Prever o *churn* permite que as empresas identifiquem clientes em risco de abandonar o serviço, facilitando a adoção de estratégias proativas de retenção, como ofertas personalizadas e melhorias no atendimento ao cliente, como visto anteriormente. Nesse pretexto, a IA se destaca por sua capacidade de

analisar grandes volumes de dados e identificar padrões complexos que seriam difíceis de detectar por meio de métodos tradicionais, ao envolver “enorme variedade de subcampos, desde áreas de uso geral, como aprendizado e percepção, até tarefas específicas como jogos de xadrez, demonstração de teoremas matemáticos, criação de poesia e diagnóstico de doença” (Gomes, 2010, p.234). Ou seja, se consegue interpretar os fatores que influenciam o comportamento dos clientes e atuação de métodos de retenção de clientes e otimizar suas operações.

A aplicação de IA na previsão de *churn* envolve o uso de algoritmos que aprendem com dados históricos para prever comportamentos futuros, como explica Flores e Bess (2023). Segundo os autores, estes algoritmos são treinados para reconhecer padrões em dados como histórico de compras, interações com o serviço ao cliente, dados demográficos e outras métricas. Ao identificar os sinais precursores de *churn*, as empresas podem segmentar seus esforços de retenção, direcionando recursos para clientes com maior probabilidade de abandonar o serviço. Dessa forma, é essencial para a sustentabilidade de negócios, já que a retenção de clientes é geralmente mais econômica do que a aquisição de novos (Verbeke et al., 2011). Dessa forma, a aplicação de IA surge como um avanço para empresas que buscam prever e mitigar os riscos associados ao *churn*.

2.2.1 MODELOS DE MACHINE LEARNING

Os modelos de *Machine Learning* são fundamentais para a aplicação da IA na previsão de *churn*, segundo Fernandes (2024). Entre os modelos mais utilizados estão a regressão logística, as árvores de decisão, florestas aleatórias, máquinas de vetores de suporte (SVM) e redes neurais artificiais. A escolha do modelo depende de vários fatores, incluindo o tipo de dados disponíveis, o grau de complexidade do problema e a necessidade de interpretabilidade do modelo. A regressão logística, por exemplo, é um método linear amplamente utilizado em problemas de classificação binária e é particularmente útil quanto a relação entre as variáveis preditoras e o *churn* ser linear e/ou direto (Berson; Smith; Thearling, 2000).

Por outro lado, métodos não lineares, como as árvores de decisão e florestas aleatórias, conseguem capturar interações complexas entre variáveis que a regressão logística não consegue, tornando-os mais adequados para cenários com estruturas de dados mais complicadas (Breiman, 2001). Segundo Correia (2023), as florestas aleatórias, em particular, são reconhecidas por sua consistência e precisão, especialmente em conjuntos de dados de

alta dimensionalidade. As máquinas de vetores de suporte (SVM) também são utilizadas na previsão de *churn*, devido à sua capacidade de lidar com problemas de classificação e com conjuntos de dados onde as classes são altamente sobrepostas.

A agregação de previsões de um conjunto de previsores, como classificadores ou regressores, muitas vezes obterá melhores previsões do que com o melhor previsor individual. Um conjunto de previsores é chamado de ensemble, assim, esta técnica é chamada Ensemble Learning, e um algoritmo de Ensemble Learning é chamado de Ensemble method. Floresta aleatória é um ensemble, ou seja, um conjunto de árvores de decisão (aleatórias), que tem como o propósito minimizar o sobreajuste de cada modelo individual de árvore gerado para os dados de entrada. Apesar da sua simplicidade, é um dos mais poderosos algoritmos de aprendizado de máquina disponível atualmente (Correira, 2023, p. 24)

As redes neurais artificiais e, mais recentemente, as técnicas de *deep learning*, têm se mostrado extremamente promissoras na previsão de *churn*. Redes neurais, compostas por múltiplas camadas de processamento que aprendem representações hierárquicas dos dados, são particularmente úteis em cenários onde os padrões são complexos e não lineares (LeCun; Bengio; Hinton, 2015). Modelos de *deep learning*, como redes neurais convolucionais e redes neurais recorrentes, são aplicáveis a conjuntos de dados que envolvem sequências temporais ou dados espaciais, como registros de chamadas em telecomunicações ou transações bancárias ao longo do tempo (Goodfellow; Bengio; Courville, 2016). Esses modelos, embora muitas vezes menos interpretáveis, oferecem uma capacidade incomparável de modelar relações intrincadas nos dados, resultando em previsões mais precisas.

A avaliação da eficácia desses modelos é realizada via métricas como acurácia, precisão, recall e a área sob a curva ROC (*Receiver Operating Characteristic*). Tais métricas são cruciais para garantir que os modelos não apenas identificam corretamente os clientes em risco de *churn*, mas também que minimizam os erros de previsão, como falsos positivos e falsos negativos (Chawla, et al., 2002). A implementação de modelos de machine learning para a previsão de *churn*, portanto, não só proporciona uma vantagem competitiva para as empresas, mas também cria oportunidades para intervenções de retenção de clientes mais informadas e direcionadas.

2.3 Pré-processamento de Dados e Balanceamento de Classes

A previsão de *churn* é especialmente crítica para indústrias que dependem de modelos de receita recorrente. Através da capacidade de prever quais clientes têm maior probabilidade de cancelar um serviço permite que as empresas adotem medidas proativas, como intervenções personalizadas, descontos ou melhorias no atendimento ao cliente, para reduzir a taxa de cancelamento e maximizar a retenção de clientes (Verbeke et al., 2011).

O uso de IA para a previsão de *churn* envolve a aplicação de algoritmos de aprendizado de máquina (*Machine Learning*) treinados com dados históricos de clientes para identificar padrões de comportamento que precedem o *churn*. Esses algoritmos conseguem processar grandes volumes de dados e detectar sinais sutis que indicam insatisfação do cliente, tais como mudanças na frequência de uso, interações negativas com o suporte ao cliente, ou alterações no padrão de compras. O uso de *Machine Learning*, como redes neurais artificiais e modelos de *deep learning*, permite que os sistemas identifiquem relações complexas e não lineares nos dados, oferecendo previsões mais precisas e úteis para os gestores (LeCun; Bengio; Hinton, 2015). Essas técnicas são úteis em contextos em que os padrões de *churn* são complexos e não facilmente discerníveis por métodos estatísticos tradicionais.

Como visto previamente, os modelos de *Machine Learning* comuns na previsão de *churn* incluem regressão logística, árvores de decisão, florestas aleatórias, máquinas de vetores de suporte (SVM) e redes neurais. A regressão logística é frequentemente utilizada devido à sua simplicidade e interpretabilidade, sendo uma escolha melhor quando a relação entre variáveis preditoras e *churn* é linear (Berson; Smith; Thearling, 2000). No entanto, para capturar interações mais complexas, são preferidos os métodos como árvores de decisão e florestas, por poderem lidar com não linearidades e interações entre variáveis, segundo Breiman (2001). Além disso, as redes neurais e modelos de *deep learning*, que envolvem múltiplas camadas de processamento de dados, podem capturar padrões mais sofisticados, especialmente quando os dados são de natureza temporal ou sequencial, como registros de chamadas em telecomunicações ou padrões de transações financeiras (Goodfellow; Bengio; Courville, 2016). Dessa forma, pode-se afirmar que:

As redes neurais são candidatas naturais para predição devido a sua capacidade de tolerar ruídos e sua não linearidade. O uso de redes neurais para a predição do tráfego começou no final da década de 80 com resultados encorajadores, e esse campo de pesquisa tem aumentado bastante desde então. O principal diferencial desse trabalho será o uso de uma nova técnica de aprendizado de máquina, chamada de Deep Learning, ou Aprendizagem Profunda, para o aprendizado de tráfegos de redes. Redes Neurais Artificiais (RNA), ou somente redes neurais, são modelos matemáticos inspirados na estrutura do cérebro humano. Uma RNA é composta por neurônios, sendo estruturas de processamento simples, separados em unidades fortemente conectadas. Os neurônios de uma RNA são organizados em camadas, uma camada pode ter vários neurônios e uma rede neural pode ter duas ou mais camadas (pelo menos uma camada de entrada e uma de saída) (Oliveira, 2014, p. 30–31).

Para a aplicação para previsão, pode-se usar a avaliação e validação dos modelos preditivos. As métricas comumente utilizadas para medir a eficácia dos modelos incluem acurácia, precisão, recall e a área sob a curva ROC (Receiver Operating Characteristic) que

ajudam a avaliar não apenas a capacidade do modelo de prever corretamente os casos de *churn*, mas também sua habilidade em evitar previsões incorretas, como falsos positivos e falsos negativos, que podem ter implicações em termos de custos e esforços de retenção (Chawla et al., 2002). Observa-se que os modelos que apresentam um bom equilíbrio entre essas métricas são melhores quanto a prática empresarial.

A curva Receiver Operating Characteristic (ROC) é uma técnica padrão para resumir o desempenho do classificador em uma série de compensações entre taxas de erro de verdadeiro positivo e falso positivo. A Área Sob a Curva (AUC) é uma métrica de desempenho tradicional aceita para uma curva ROC. O casco convexo ROC também pode ser usado como um método robusto para identificar classificadores potencialmente ótimos. Se uma linha passa por um ponto no casco convexo, então não há outra linha com a mesma inclinação passando por outro ponto com uma interceptação de verdadeiro positivo (TP) maior. Assim, o classificador naquele ponto é ótimo sob quaisquer suposições de distribuição em conjunto com essa inclinação (Chawla et al., 2002, p.322).

Dessa forma, a integração de modelos de *machine learning* na previsão de *churn* oferece uma vantagem competitiva significativa para as empresas, permitindo uma abordagem mais estratégica e baseada em dados para a gestão da retenção de clientes. Além disso, esses modelos podem ajudar a identificar os fatores subjacentes ao *churn*, proporcionando às empresas uma compreensão mais profunda das necessidades e expectativas dos clientes. Isso melhora a capacidade de prever o *churn* e orienta o desenvolvimento de estratégias de produto e serviço que podem melhorar a satisfação e lealdade do cliente (Verbeke et al., 2011). Portanto, isso torna a aplicação da IA na previsão de *churn* cada vez mais indispensável para empresas.

2.4 Implicações e Benefícios para as Empresas de Telecomunicações

As empresas de telecomunicações possuem problemas quanto à complexidade dos serviços que oferecem, à alta competitividade do mercado e às crescentes expectativas dos clientes por serviços mais personalizados e de alta qualidade. Nesse contexto, a implementação de IA pode transformar essas empresas, quanto ao entendimento do comportamento dos clientes.

As telecomunicações, é um setor caracterizado por intensa competição e margens de lucro pressionadas, encontram na IA uma ferramenta estratégica nova com múltiplas possibilidades de uso, como traçar a retenção de clientes, a gestão de recursos de rede e a prevenção de fraudes. A IA permite às empresas processar e analisar grandes volumes de

dados de maneira rápida e precisa, facilitando a tomada de decisões mais informadas e assertivas (Chandratreya et al., 2024).

O que antes era um sistema reativo, principalmente utilizando dados históricos, evoluiu usando IA para uma ferramenta preditiva e estratégica. Em poucas palavras, com a ajuda de grandes quantidades de dados e algoritmos avançados, a análise preditiva moderna baseada em inteligência artificial permite o gerenciamento antecipado do comportamento do cliente, detecção proativa de possíveis problemas e personalização abrangente da interação. Ela ainda melhora as experiências do cliente, bem como a utilização eficiente dos recursos e receitas disponíveis, sem mencionar a fidelidade do cliente amplificada (Chandratreya et al., 2024, p.306).

Um dos principais benefícios da IA para as telecomunicações é a melhoria na gestão da experiência do cliente. Algoritmos de aprendizado de máquina permitem a análise em tempo real de grandes volumes de dados de clientes, incluindo histórico de chamadas, padrões de uso e feedbacks, possibilitando a personalização dos serviços conforme as preferências individuais dos usuários. Isso não só aumenta a satisfação do cliente, mas também contribui para a retenção de clientes, um aspecto crítico para empresas que enfrentam altos índices de *churn*. A personalização de serviços com base em IA pode aumentar a retenção de clientes, porém, ao testar a regressão logística, árvores de decisão e outros algoritmos em um conjunto de dados de usuários de telecomunicações, a personalização de metodologias de modelagem e a abordagem de desafios relacionados a dados de alta dimensão na modelagem Naive Bayes (Afzal et al., 2024, p.6).

Além disso, as empresas de telecomunicações podem utilizar IA para otimizar suas operações e reduzir custos. A previsão de demanda, por exemplo, é uma área onde a IA pode oferecer significativos benefícios. Segundo Couluris et al. (2013), ao prever picos de uso de rede, as empresas podem alocar recursos, evitando gargalos de serviço e melhorando a qualidade da conexão para os usuários. Isso é particularmente relevante em um setor onde a infraestrutura de rede representa um custo fixo elevado, e a capacidade de otimizar o uso dessa infraestrutura pode resultar em economias substanciais. Modelos preditivos baseados em IA também podem ser utilizados para manutenção preditiva, identificando e corrigindo problemas de rede antes que afetem os clientes, o que pode reduzir significativamente o tempo de inatividade e os custos de manutenção (Chang, 2024).

Outro benefício crítico da IA nas telecomunicações é a capacidade de detectar e mitigar fraudes. A fraude é uma preocupação significativa para as empresas de telecomunicações, representando perdas anuais substanciais. Segundo Ritika (2022), os algoritmos de IA podem analisar padrões de comportamento e identificar atividades suspeitas em tempo real, permitindo que as empresas tomem medidas imediatas para prevenir fraudes antes que elas ocorram. Por exemplo, redes neurais artificiais e algoritmos de detecção de

anomalias têm se mostrado útil na identificação de padrões de fraude em chamadas e transações financeiras (Kshetri, 2021).

A implementação de IA também oferece vantagens na gestão de relacionamento com o cliente por *chatbots* e assistentes virtuais, que podem fornecer suporte 24/7, resolver problemas comuns rapidamente e encaminhar questões mais complexas para agentes humanos. Isso melhora a eficiência operacional e a experiência de cliente que valoriza a conveniência e a rapidez proporcionadas pelos assistentes virtuais, melhorando as percepções da marca e reduzindo o tempo de resolução de problemas (Huang; Rust, 2018).

Apesar dos benefícios, a aplicação de IA nas telecomunicações também apresenta problemas, como questões relacionadas à privacidade de dados, a necessidade de infraestrutura tecnológica avançada e a demanda por profissionais qualificados para gerenciar e interpretar os resultados dos algoritmos de IA. A segurança dos dados é uma preocupação, e as empresas de telecomunicações devem garantir que as informações dos clientes sejam protegidas contra acessos não autorizados, especialmente quando algoritmos de IA são utilizados para analisar dados pessoais sensíveis. Porém, como aponta Mikalef et al. (2019), a implementação de soluções de IA requer investimentos em infraestrutura tecnológica e a contratação de especialistas em ciência de dados, o que pode representar um obstáculo para algumas empresas.

3 METODOLOGIA

3.1 Área de Estudo

A metodologia adotada neste estudo é voltada para a aplicação de técnicas de Ciência de Dados para analisar o churn de clientes no setor de Telecom. Foi utilizado o *dataset Telco Customer Churn* (“WA_Fn-UseC_-Telco-Customer-Churn.csv”), frequentemente empregado em práticas de aprendizado de máquina, análise preditiva e engenharia de características, visando compreender e prever o comportamento de cancelamento de clientes. A análise focou na identificação das variáveis que mais influenciam na decisão de cancelar o serviço e no desenvolvimento de modelos de classificação e regressão para prever a probabilidade de *churn*. A relevância do estudo está em aprimorar as estratégias de retenção de clientes e otimizar as operações da empresa.

3.2 Planejamento dos Experimentos

3.2.1 PREPARAÇÃO DOS DADOS

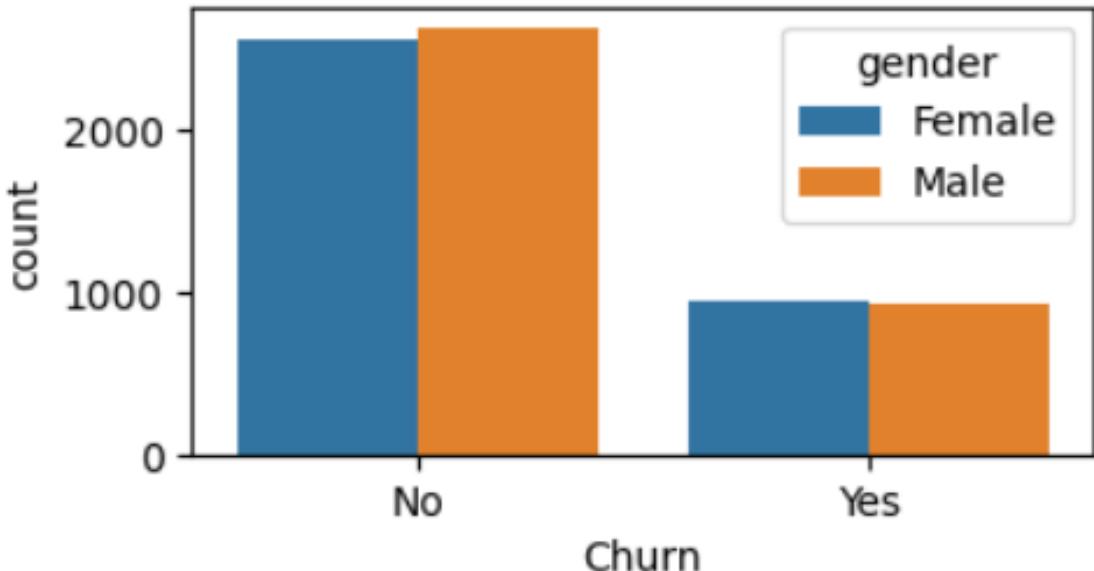
Na preparação dos dados, a primeira etapa consistiu na importação do conjunto de dados, carregado no formato CSV. Em seguida, procedeu-se à limpeza dos dados, que envolveu a identificação e tratamento de valores nulos, duplicatas e inconsistências. Variáveis categóricas como '*No internet service*' e '*No phone service*' foram transformadas para '*No*', e a coluna *TotalCharges* foi convertida para tipo numérico para garantir a consistência dos dados. Para a codificação das variáveis categóricas, foram aplicadas diferentes técnicas: o *LabelEncoder* foi utilizado para transformar rótulos de classes em valores numéricos, sendo adequado para variáveis target; o *StandardScaler* foi empregado para normalizar características numéricas, ajustando-as para média 0 e desvio padrão 1, o que é particularmente útil para atributos contínuos; e o *OneHotEncoder* foi utilizado para transformar variáveis categóricas em uma representação binária, evitando a interpretação de ordens nas categorias. A escolha dessas ferramentas permitiu o ajuste apropriado da base de dados, sendo bastante heterogênea.

3.2.2 ANÁLISE EXPLORATÓRIA DOS DADOS (EDA)

A Análise Exploratória dos Dados envolveu a geração de visualizações, como gráficos de barras, para explorar a distribuição das variáveis, incluindo a variável de *churn* por gênero,

conforme ilustrado na Figura 1. Além disso, foram obtidas e analisadas as estatísticas descritivas das variáveis numéricas e categóricas, proporcionando uma visão geral das características dos dados.

Figura 1. Exemplo de gráfico usado para explorar a distribuição de variáveis



Fonte: Elaborado pelo autor (2024)

3.2.3 DIVISÃO DOS DADOS

Na divisão dos dados, as variáveis foram separadas em features (variáveis independentes) e *target* (variável dependente, neste caso, o *Churn*). Para a criação dos conjuntos de treinamento e teste, utilizou-se o método *train_test_split*, que dividiu os dados em 70% para treinamento e 30% para teste, assegurando uma base robusta para a validação dos modelos.

3.2.4 PRÉ-PROCESSAMENTO DOS DADOS:

O pré-processamento dos dados envolveu o escalonamento e codificação das features, realizados através do *ColumnTransformer*. Esta etapa foi essencial para garantir que todas as variáveis fossem tratadas consistentemente, especialmente em uma base com múltiplas características categóricas e numéricas.

3.2.5 TREINAMENTO E AVALIAÇÃO DOS MODELOS:

Durante o treinamento e avaliação dos modelos, foram testados diversos algoritmos de classificação, incluindo *Decision Tree*, *Logistic Regression*, *Random Forest* e *XGBoost*. Para cada modelo, foram criados pipelines que integraram as etapas de pré-processamento e treinamento, facilitando um fluxo de trabalho estruturado e eficiente. A desempenho dos

modelos foi avaliada utilizando métricas como acurácia e relatórios de classificação, para identificar o melhor modelo com base na acurácia.

3.2.6 AJUSTE DE HIPERPARÂMETROS:

Por fim, o ajuste de hiperparâmetros foi realizado utilizando o RandomizedSearchCV, uma técnica que permite encontrar os melhores hiperparâmetros eficientemente para cada modelo. Após o ajuste, os modelos otimizados foram reavaliados, e seus desempenhos comparados aos resultados iniciais, destacando as melhorias alcançadas em termos de acurácia e robustez dos modelos ajustados.

4 RESULTADOS

4.1 Resultados Sem HiperParametros

Os resultados apresentados na Figura 2 mostram o desempenho dos modelos de classificação sem a utilização de otimização dos hiperparâmetros. Inicialmente, o modelo de Árvore de Decisão apresentou uma acurácia de 71,93%. Apesar de um desempenho aceitável na classe majoritária (classe 0), com precisão de 82% e recall de 81%, houve uma significativa disparidade em relação à classe minoritária (classe 1), cuja precisão foi de apenas 47% e recall de 51%, resultando em um f1-score de 0,49.

O modelo de Regressão Logística, por sua vez, apresentou uma acurácia de 78,54%. As métricas para a classe majoritária foram elevadas, com precisão de 83% e recall de 89%. No entanto, para a classe minoritária, os valores de precisão e recall caíram para 62% e 49%, respectivamente.

Já o modelo Random Forest alcançou uma acurácia de 79,03%, com um desempenho equilibrado na classe majoritária (precisão de 83% e recall de 86%). Contudo, a classe minoritária continuou a mostrar desempenho inferior, com precisão de 64% e recall de 48%, refletindo problemas semelhantes aos encontrados nos outros modelos, principalmente quanto a identificação da classe menos representada.

Figura 2. Resultados sem HiperParâmetros

Model: Decision Tree					
Accuracy: 0.7193					
		precision	recall	f1-score	support
	0	0.82	0.80	0.81	1033
	1	0.47	0.51	0.49	374
	accuracy			0.72	1407
	macro avg	0.65	0.65	0.65	1407
	weighted avg	0.73	0.72	0.72	1407

Model: Logistic Regression					
Accuracy: 0.7854					
		precision	recall	f1-score	support
	0	0.83	0.89	0.86	1033
	1	0.62	0.49	0.55	374
	accuracy			0.79	1407
	macro avg	0.73	0.69	0.70	1407
	weighted avg	0.77	0.79	0.78	1407

Model: Random Forest					
Accuracy: 0.7903					
		precision	recall	f1-score	support
	0	0.83	0.90	0.86	1033
	1	0.64	0.48	0.55	374
	accuracy			0.79	1407
	macro avg	0.73	0.69	0.71	1407
	weighted avg	0.78	0.79	0.78	1407

Model: XGBoost					
Accuracy: 0.7655					
		precision	recall	f1-score	support
	0	0.82	0.87	0.85	1033
	1	0.57	0.47	0.52	374
	accuracy			0.77	1407
	macro avg	0.70	0.67	0.68	1407
	weighted avg	0.75	0.77	0.76	1407

Fonte: Elaborado pelo autor (2024)

O modelo XGBoost apresentou uma acurácia de 76,55%. Na classe majoritária, o desempenho foi razoável, com precisão de 82% e recall de 87%. Entretanto, a classe minoritária mostrou um desempenho moderado, com precisão de 57% e recall de 52%.

Traduzindo, os resultados demonstram que, sem ajustes nos hiperparâmetros, todos os modelos enfrentam dificuldades na detecção da classe 1.

4.2 Resultados Com HiperParametros

Os resultados ilustrados na Figura 3 refletem no modelo de Árvore de Decisão, após os ajustes, um aumento da sua acurácia para 78,18%. Houve uma melhora na precisão da classe minoritária, que subiu para 66%, e no f1-score, que alcançou 0,48.

Figura 3. Resultados com HiperParametros

Best Model: Decision Tree				
Accuracy: 0.7818				
	precision	recall	f1-score	support
0	0.80	0.93	0.86	1033
1	0.66	0.38	0.48	374
accuracy			0.78	1407
macro avg	0.73	0.65	0.67	1407
weighted avg	0.76	0.78	0.76	1407

Best Model: Logistic Regression				
Accuracy: 0.7846				
	precision	recall	f1-score	support
0	0.83	0.89	0.86	1033
1	0.62	0.49	0.55	374
accuracy			0.78	1407
macro avg	0.72	0.69	0.70	1407
weighted avg	0.77	0.78	0.78	1407

Best Model: Random Forest					
Accuracy: 0.8010					
	precision	recall	f1-score	support	
0	0.83	0.91	0.87	1033	
1	0.67	0.49	0.57	374	
accuracy			0.80	1407	
macro avg	0.75	0.70	0.72	1407	
weighted avg	0.79	0.80	0.79	1407	

Best Model: XGBoost					
Accuracy: 0.7967					
	precision	recall	f1-score	support	
0	0.84	0.90	0.87	1033	
1	0.65	0.52	0.57	374	
accuracy			0.80	1407	
macro avg	0.74	0.71	0.72	1407	
weighted avg	0.79	0.80	0.79	1407	

Fonte: Elaborado pelo autor (2024)

Para o modelo de Regressão Logística, a otimização resultou em um pequeno aumento na acurácia, que chegou a 78,64%. Embora a precisão da classe minoritária tenha se mantido em 62%, houve uma leve queda no recall, impactando o f1-score.

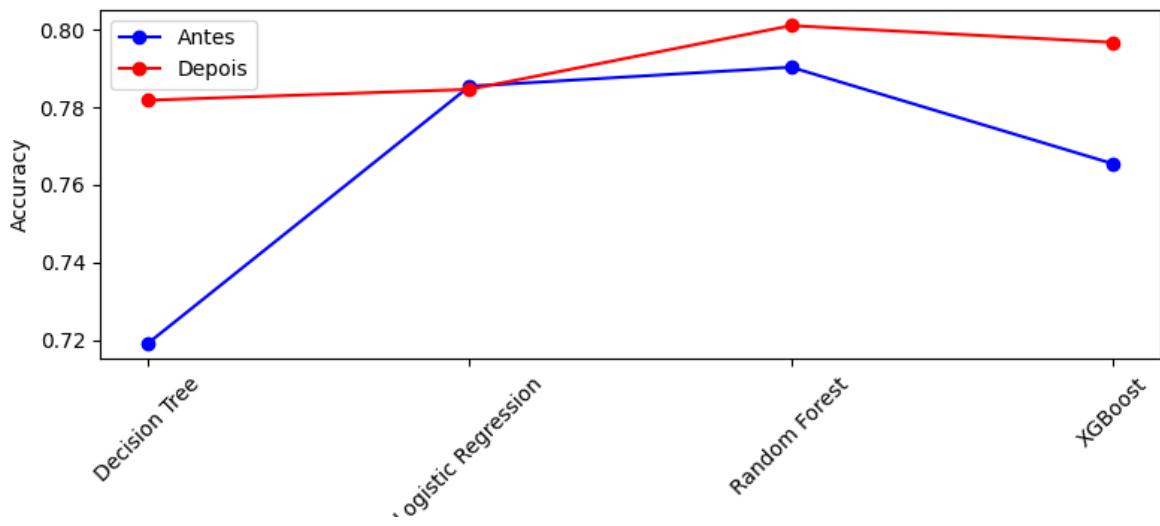
O modelo Random Forest se destacou com a melhor desempenho após a otimização, alcançando uma acurácia de 80,10%. As métricas para a classe minoritária melhoraram consideravelmente, com precisão de 67% e recall de 49%. Esse ajuste nos hiperparâmetros permitiu ao modelo capturar mais a classe minoritária.

Já no XGBoost, após os ajustes, apresentou uma acurácia de 79,67%. Houve uma melhora na classe minoritária, com a precisão subindo para 65% e o recall para 52%. Ou seja, otimizou-se, por equilíbrio entre as classes, os hiperparâmetros, embora o modelo ainda mostre um desempenho superior na classe majoritária.

4.3 Comparação

No Gráfico 1, observa-se a comparação das acurárias dos quatro modelos antes e depois da otimização dos hiperparâmetros, sob um aumento para todos os modelos após a otimização, com destaque para o modelo Random Forest, que apresentou o maior ganho, superando os 80% de acurácia após o ajuste. O modelo de Regressão Logística também apresentou um bom aumento na acurácia, posicionando-se como o segundo melhor desempenho. Por outro lado, o modelo XGBoost, embora tenha mostrado melhorias após a otimização, ainda ficou atrás dos demais.

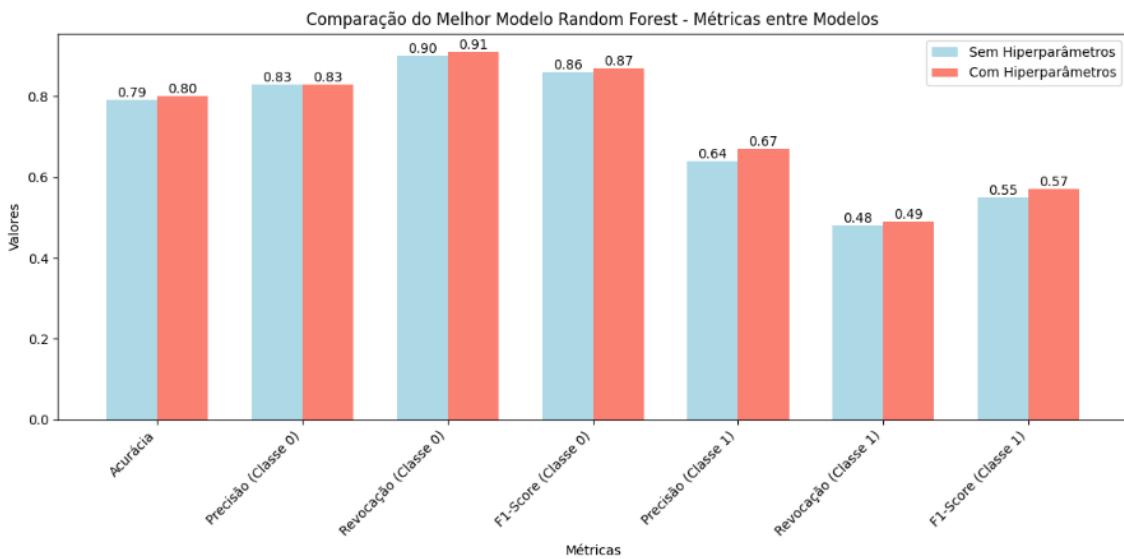
Gráfico 1. Gráfico Acurácia dos 4 modelos de Comparação antes e depois dos hiperparâmetros



Fonte: Elaborado pelo autor (2024)

O Gráfico 2 fornece uma visão mais detalhada do desempenho do modelo Random Forest, comparando os resultados com e sem otimização dos hiperparâmetros em termos de acurácia, precisão, recall e f1-score. É evidente que, com a otimização, todas as métricas de desempenho melhoraram. A acurácia aumentou de aproximadamente 0,79 para 0,80, enquanto a precisão passou de 0,83 para 0,85, e o recall de 0,67 para 0,69. O f1-score também apresentou um incremento, destacando o impacto positivo da otimização dos hiperparâmetros no equilíbrio entre precisão e recall. Esse comportamento confirma que a otimização não apenas melhora a detecção das classes majoritárias, mas também aperfeiçoa a identificação de casos menos frequentes, o que é especialmente importante em contextos em que a classe minoritária possui alta relevância.

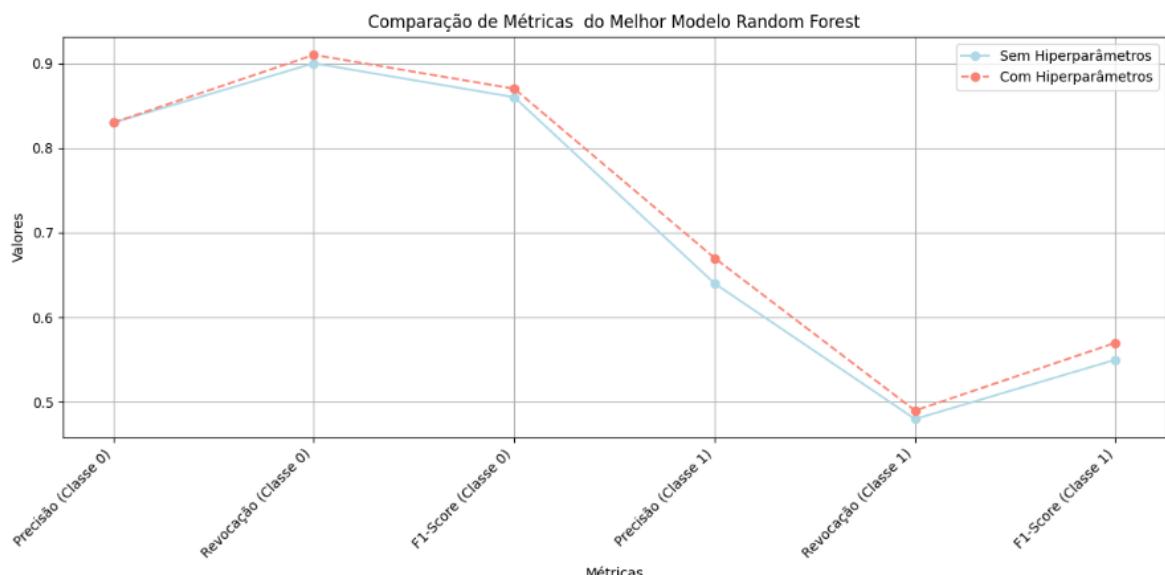
Gráfico 2. Desempenho do melhor modelo, Random Forest Com e Sem otimização de hiperparâmetros, para comparar a Acurácia, Precisão, Revocação e F1-score entre os dois modelos.



Fonte: Elaborado pelo autor (2024)

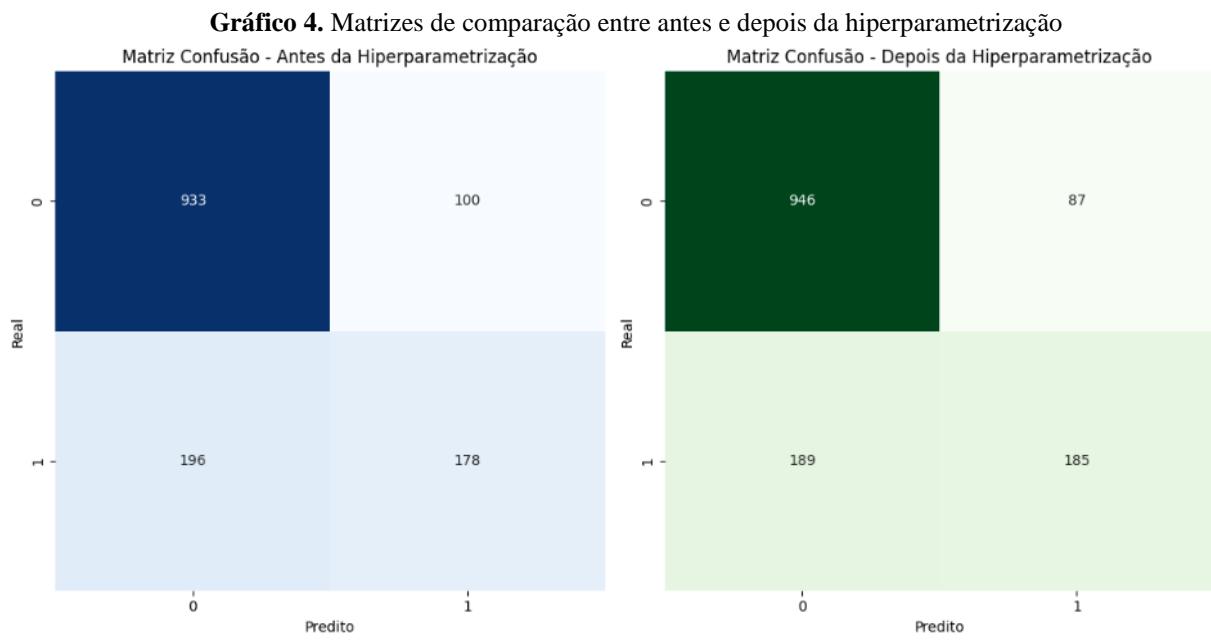
O Gráfico 3 reforça essa análise ao comparar visualmente as métricas de desempenho do Random Forest antes e depois da otimização. O gráfico de linhas mostra uma tendência clara de melhoria após a otimização, com as métricas de precisão e f1-score exibindo os maiores ganhos. Isso indica que, além de aumentar a acurácia geral do modelo, a otimização dos hiperparâmetros contribui para o balanceamento das previsões entre as classes, reduzindo a ocorrência de falsos negativos e falsos positivos, críticos para a aplicação prática do modelo.

Gráfico 3. Desempenho do melhor modelo, Random Forest Com e Sem otimização de hiperparâmetros, para comparar a Acurácia, Precisão, Revocação e F1-score entre os dois modelos.



Fonte: Elaborado pelo autor (2024)

Por fim, o Gráfico 4 apresenta as matrizes de confusão do modelo Random Forest antes e depois da otimização dos hiperparâmetros. Antes da otimização, o modelo teve 185 verdadeiros positivos e 946 verdadeiros negativos, com 57 falsos positivos e 189 falsos negativos. Após a otimização, notou-se um aumento nos verdadeiros positivos para 194 e uma redução dos falsos negativos para 179, enquanto os verdadeiros negativos aumentaram ligeiramente para 954. Esse resultado indica que a otimização dos hiperparâmetros não apenas melhorou a capacidade do modelo de prever corretamente ambas as classes, mas também contribuiu para a redução de erros críticos, especialmente os falsos negativos, que podem ter um impacto significativo dependendo da aplicação específica do modelo.



TP = 185 Verdadeiros Positivos
 TN = 946 Verdadeiros Negativos
 FP = 57 Falsos Positivos
 FN = 189 Falsos Negativos

Fonte: Elaborado pelo autor (2024)

Estes gráficos mostram o desempenho do nosso melhor modelo de classificação. Abaixo detalhamos cada valor e seus respectivos cálculos, para podermos calcular as métricas para avaliar o desempenho dos nossos modelos:

Interpreta-se e se calcula as principais métricas a partir dos dados fornecidos. Aqui estão as definições e os cálculos para as principais métricas de desempenho de um modelo de classificação:

Accuracy (Acurácia): 0.8038 A acurácia é a proporção de previsões corretas entre o total de previsões feitas.

$$Accuracy = \frac{VP+VN}{Total} = \frac{185+946}{1407} \approx 0,8038 \quad (1)$$

TP = 185 Verdadeiros Positivos

TN = 946 Verdadeiros Negativos

FP = 87 Falsos Positivos

FN = 189 Falsos Negativos

Inicialmente, observa-se que a acurácia geral do modelo é de aproximadamente 0,8038, indicando que cerca de 80,38% das previsões feitas pelo modelo estão corretas. Esta medida é calculada pela razão entre a soma dos verdadeiros positivos e verdadeiros negativos (185 e 946, respectivamente) sobre o total de previsões (1407). Apesar de ser uma métrica robusta, a acurácia, por si só, pode não refletir adequadamente o desempenho do modelo, especialmente em conjuntos de dados desbalanceados, o que nos leva a considerar outras métricas como a precisão, o recall e o F1-score.

A precisão, que mede a proporção de verdadeiros positivos em relação a todas as previsões positivas, foi calculada como 0,68. Isso sugere que 68% das previsões positivas feitas pelo modelo são corretas, mas indica uma margem para erros onde falsos positivos ainda são significativos. Já o recall, que avalia a capacidade do modelo em identificar corretamente todos os casos positivos, é de apenas 0,49, o que demonstra que o modelo está deixando de identificar uma proporção considerável de verdadeiros positivos, com um número de 189 falsos negativos.

Precision (Precisão): A precisão é a proporção de verdadeiros positivos em relação ao total de positivos previstos.

$$Precision = \frac{VP}{VP+FP} = \frac{185}{185+87} \approx 0,68 \quad (2)$$

Recall (Revocação): A revocação é a proporção de verdadeiros positivos em relação ao total de positivos reais.

$$Recall = \frac{VP}{VP+FN} = \frac{152}{185+189} \approx 0,49 \quad (3)$$

F1-Score: O F1-Score é a média harmônica da precisão e revocação, sendo útil para equilibrar a precisão e a revocação.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{0,68 \times 0,49}{0,68 + 0,49} \approx 0,57 \quad (4)$$

O F1-score, uma métrica que harmoniza precisão e recall, resultou em 0,57, evidenciando um equilíbrio modesto entre essas duas métricas. O valor relativamente baixo do F1-score aponta para a necessidade de melhorias adicionais no modelo para alcançar um balanceamento entre minimizar falsos positivos e falsos negativos.

Analizando as médias ponderadas, que consideram o suporte de cada classe, as métricas mostram valores levemente superior, com a precisão ponderada em 0,79, recall ponderado em 0,80 e F1-score ponderado também em 0,79. Estas médias ponderadas indicam um desempenho melhorado ao considerar o tamanho de cada classe, sugerindo que o modelo é melhor em classes com maior representação no conjunto de dados.

Macro Average: A média macro calcula a métrica para cada classe e depois tira a média, sem considerar o número de exemplos de cada classe.

$$Macro Avg = \frac{1}{N} \sum_{i=1}^N \text{Métrica}_i \quad (5)$$

$$Macro Avg Precision = \frac{0,83 + 0,68}{2} = 0,755 \quad (6)$$

$$Macro Avg Recall = \frac{0,92 + 0,49}{2} = 0,705 \quad (7)$$

$$Macro Avg F1 - Score = \frac{0,87 + 0,57}{2} = 0,72 \quad (8)$$

Adicionalmente, ao examinar as médias macro — que calculam a média simples das métricas para cada classe sem considerar o suporte —, onde a média macro da precisão é 0,755, do recall é 0,755, e do F1-score é 0,72. Essas médias são úteis para avaliar o desempenho geral do modelo em todas as classes de forma igualitária, sem a influência do tamanho de cada classe.

Weighted Average: A média ponderada leva em consideração o suporte (número de amostras) de cada classe ao calcular a média.

$$Weighted Avg = \frac{\sum_{i=1}^N (\text{Métrica}_i \times \text{Suporte}_i)}{Total \ de \ Amostras} \quad (9)$$

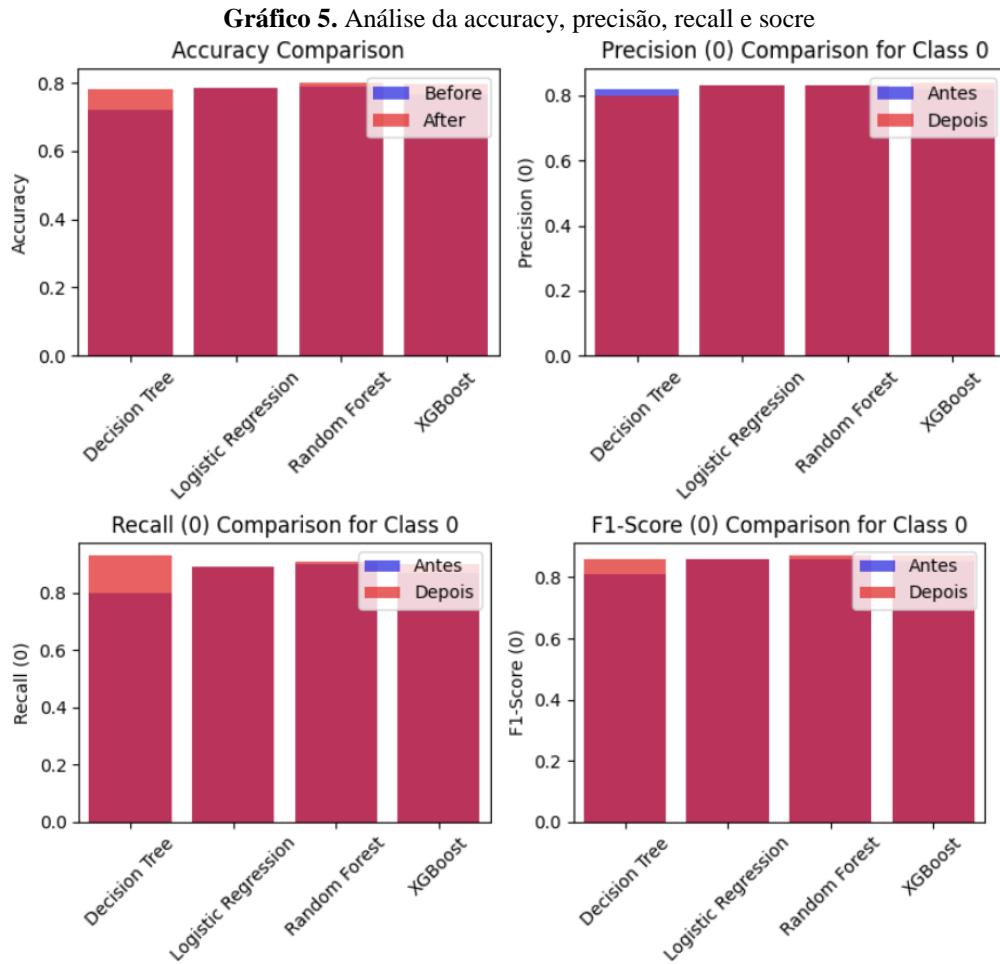
$$Weighted Avg Precision = \frac{0,83 \times 1033 + 0,68 \times 374}{1033 + 374} \approx 0,79 \quad (10)$$

$$Weighted Avg Recall = \frac{0,92 \times 1033 + 0,49 \times 374}{1033 + 374} \approx 0,80 \quad (11)$$

$$Weighted Avg F1 - Score = \frac{0,87 \times 1033 + 0,57 \times 374}{1033 + 374} \approx 0,79 \quad (12)$$

Identificação de Erros: analisam-se os falsos positivos e falsos negativos para entender os dos modelos antes e depois dos ajustes dos hiperparâmetros.

Os gráficos 5 e subsequentes mostram uma comparação das métricas antes e depois da otimização dos hiperparâmetros, confirmando melhorias em todas as métricas avaliadas. Por exemplo, no caso da classe 0 (gráficos superiores), observam-se aumentos claros na acurácia, precisão, e recall após os ajustes de hiperparâmetros, reforçando a necessidade de se acompanhar esta etapa no processo de modelagem. A otimização levou a uma redução dos falsos negativos e, consequentemente, melhorou o F1-score da classe.



4.4 Discussão

A análise comparativa entre os modelos revelou que, sem a otimização de hiperparâmetros, todos os modelos tiveram um desempenho significativamente melhor na previsão da classe majoritária (não *churn*) em comparação com a classe minoritária (*churn*). Esse resultado é consistente com estudos anteriores que apontam a dificuldade inerente na previsão de classes minoritárias em problemas de *churn*, onde a classe positiva representa uma pequena parcela dos dados e traz desafios adicionais na identificação correta de padrões de comportamento associados ao cancelamento do serviço (Saha et al., 2023).

No caso específico da Regressão Logística, o modelo apresentou uma acurácia geral de 78,54%, com desempenho razoável na classe majoritária, mas falhou em capturar adequadamente os clientes em risco de *churn*, evidenciado pela baixa precisão e recall na classe minoritária (62% e 49%, respectivamente), denibstrando que existe uma limitação do modelo em lidar com dados não lineares e de estrutura complexa, o que é uma característica comum em conjuntos de dados de *churn*, onde os fatores determinantes para o cancelamento

de serviços são diversos. Modelos baseados em árvores, como a Árvore de Decisão e o Random Forest, apresentaram maior desempenho em termos de acurácia geral, com o Random Forest alcançando 79,03%. No entanto, esses modelos também enfrentaram dificuldades na previsão da classe minoritária, sugerindo que, embora possam capturar interações intrincadas entre variáveis, ainda há um desafio significativo na generalização dos resultados para os casos de *churn*.

Após a otimização dos hiperparâmetros, todos os modelos mostraram melhorias nas métricas de desempenho, evidenciando a importância dessa etapa no pipeline de machine learning. O Random Forest, em particular, destacou-se com uma acurácia otimizada de 80,10% e um equilíbrio melhorado entre precisão e recall na classe minoritária (67% de precisão e 49% de recall). Ou seja, pode-se dizer que os resultados bateram com as premissas de Breiman (2001) sobre o desempenho do Random Forest em contextos de alta dimensionalidade e desafios relacionados à variabilidade dos dados, e reforçam sua aplicabilidade em problemas de previsão de *churn* onde a diversidade e a quantidade de dados são desafiadoras. A otimização dos hiperparâmetros, utilizando técnicas como o RandomizedSearchCV, permitiu ajustes finos que potencializaram o desempenho do modelo, melhorando sua capacidade de generalizar e capturar corretamente os padrões associados à classe minoritária.

A análise comparativa dos resultados ilustra claramente que, embora todos os modelos tenham se beneficiado da otimização, o Random Forest mostrou-se superior, destacando-se como a abordagem mais robusta e eficiente entre as testadas. Isso se alinha com os achados de Fernandes (2024), que destaca a consistência do Random Forest para aplicações práticas em churn prediction, particularmente ao buscar um equilíbrio entre a precisão e a capacidade de recuperação de casos de *churn*. Além disso, os gráficos de desempenho apresentados no estudo ilustram que as métricas de precisão e F1-score melhoraram significativamente após a otimização, indicando que os ajustes nos hiperparâmetros não apenas aumentaram a acurácia geral, mas também contribuíram para um balanceamento mais adequado entre as previsões corretas e incorretas.

Contudo, mesmo com as melhorias, todos os modelos mantiveram uma disparidade no desempenho entre as classes majoritária e minoritária, refletindo um desafio contínuo na previsão de *churn*. Essa limitação sugere que, para avançar na precisão dessas previsões, pode ser necessário explorar modelos mais sofisticados, como redes neurais profundas, que têm mostrado potencial para capturar padrões mais intrincados e não lineares nos dados (LeCun et al., 2015). A aplicação de técnicas de *deep learning*, que utilizam múltiplas camadas de

processamento para aprender representações hierárquicas dos dados, poderia oferecer melhorias substanciais, especialmente em cenários onde os sinais de *churn* são sutis e multifacetados. Além disso, a integração de estratégias como *oversampling* ou técnicas de ensemble mais avançadas, que combinam múltiplos modelos para melhorar a precisão das previsões, que ajudariam a superar as limitações com os modelos testados.

6 CONCLUSÃO

A análise dos dados dos clientes envolveu um processo detalhado que incluiu a limpeza e transformação dos dados, análise exploratória, e a aplicação de algoritmos de aprendizado de máquina, com foco na otimização dos hiperparâmetros para comparar os resultados antes e depois das otimizações. Técnicas de balanceamento de classes, como o RandomUnderSampler e o SMOTE, foram utilizadas para lidar com o desbalanceamento dos conjuntos de dados, e atuaram de maneira complementar: enquanto o RandomUnderSampler reduz o número de exemplos da classe majoritária, o SMOTE gera exemplos sintéticos da classe minoritária para alcançar um equilíbrio entre as classes.

Após a otimização dos hiperparâmetros, a maioria dos modelos apresentou melhorias nas métricas de desempenho, especialmente em termos de acurácia e nas métricas específicas de classe. O Random Forest e o XGBoost se destacaram com as maiores melhorias, sendo que o Random Forest emergiu como o modelo mais consistente e eficaz para a tarefa, tornando-se a escolha ideal para a previsão de *churn*. O desempenho para a classe majoritária (não *churn*) melhorou consideravelmente após a otimização, enquanto a previsão da classe minoritária (*churn*) ainda apresentou variabilidade, o que sugere que a identificação da classe dominante foi favorecida pela otimização, mas a identificação precisa da classe minoritária continua a requerer ajustes adicionais. A avaliação comparativa entre os modelos destacou o Random Forest como a melhor escolha, com o XGBoost como uma alternativa forte, especialmente em cenários onde um equilíbrio entre precisão e recall é crucial.

Dessa forma, os resultados deste estudo reforçam a importância da aplicação de técnicas de machine learning na análise de *churn* e sublinham a necessidade contínua de ajustar e refinar os modelos para lidar com as complexidades dos dados de clientes. A compreensão de como cada modelo se comporta em diferentes condições e como os ajustes nos hiperparâmetros e nas técnicas de balanceamento impactam nas estratégias de retenção de clientes mais direcionadas e personalizadas que deverão ser adotadas.

Ademais, recomenda-se a continuidade da otimização dos modelos, juntamente com a inclusão de novos dados e variáveis que possam capturar aspectos adicionais do comportamento do cliente. Isso inclui explorar modelos como redes neurais profundas para capturar padrões mais intrincados e interações não lineares nos dados.

REFERÊNCIAS

- AFZAL, Mohammad et al. Cross-Sector Application of Machine Learning in Telecommunications: Enhancing Customer Retention through Comparative Analysis of Ensemble Methods. **IEEE Access**, 2024.
- AHMAD, A. K.; JAFAR, A.; ALJOUMAA, K. Customer churn prediction in telecom using machine learning in big data platform. **Journal of Big Data**, v. 6, n. 1, 20 mar. 2019.
- ALSHOURBAJI, I. et al. An efficient churn prediction model using gradient boosting machine and metaheuristic optimization. **Scientific Reports**, v. 13, n. 1, p. 14441, 2 set. 2023.
- BARSOTTI, A. et al. A Decade of Churn Prediction Techniques in the TelCo Domain: A Survey. **SN Computer Science/SN computer science**, v. 5, n. 4, 6 abr. 2024.
- BERSON, A.; SMITH, S.; THEARLING, K. **Building Data Mining Applications for CRM**. [s.l.] McGraw-Hill Companies, 2000.
- BREIMAN, Leo. Random forests. **Machine Learning**, v. 45, p. 5-32, 2001.
- CHANDRATREYA, Abhijit et al. AI-driven Predictive Analytics in Customer Relationship Management. **Nanotechnology Perceptions**, p. 305-321, 2024.
- CHANG, E. **AI-Powered Fraud Detection for Telecom Security**. Disponível em: <<https://yotelecom.co.uk/ai-fraud-detection-telecom-security/>>. Acesso em: 6 set. 2024.
- CHANG, V. et al. Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models. **Algorithms**, v. 17, n. 6, p. 231–231, 27 maio 2024.
- CHAWLA, Nitesh V. et al. SMOTE: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321-357, 2002.
- CORREIA, J.V.A. **Utilizando Aprendizado de Máquina para Predição de Diagnósticos da Doença de Alzheimer**. Dissertação, Universidade Federal do Ceará, Quixadá, 2023.
- COULOURIS, George et al. **Sistemas Distribuídos: Conceitos e Projeto**. Bookman Editora, 2013.
- FERNANDES, P. M.S. **Aplicação de Machine learningem contextos de previsão de Churn–Uma revisão de literatura**. Seven, publicações acadêmicas, 2024.
- FLORES, Márcio José; BESS, Alexandre Leal. **Inteligência Artificial aplicada a negócios**. Editora Intersaberes, 2023.
- GOMES, D.S. Inteligência Artificial: conceitos e aplicações. **Revista Olhar Científico**, v. 1, n. 2, p. 234-246, 2010.
- GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. Deep feedforward networks.

Deep learning, n. 1, 2016.

HUANG, Ming-Hui; RUST, Roland T. Artificial intelligence in service. **Journal of service research**, v. 21, n. 2, p. 155-172, 2018.

KIM, S.; LEE, H. Customer Churn Prediction in Influencer Commerce: An Application of Decision Trees. **Procedia Computer Science**, v. 199, p. 1332–1339, 2022.

KRISHNAN, R.; KRISHNAVENI, C. V.; PRASAD, AV Krishna. Telecom Churn Prediction using Machine Learning. **World Journal of Advanced Engineering Technology and Sciences**, v. 7, n. 2, p. 087-096, 2022.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **nature**, v. 521, n. 7553, p. 436-444, 2015.

MIKALEF, Patrick et al. Big data analytics capabilities and innovation: the mediating role of dynamic capabilities and moderating effect of the environment. **British journal of management**, v. 30, n. 2, p. 272-298, 2019.

MUSTAFA, N.; SOOK LING, L.; ABDUL RAZAK, S. F. Customer churn prediction for telecommunication industry: A Malaysian Case Study. **F1000Research**, v. 10, p. 1274, 13 dez. 2021.

OLIVEIRA, T.P. **Predição de tráfego, usando redes neurais artificiais, para gerenciamento adaptativo de largura de banda em roteadores**. Dissertação [mestrado] Universidade Federal de Uberlândia, Minas Gerais, 2014.

PENG, K.; PENG, Y.; LI, W. Research on customer churn prediction and model interpretability analysis. **PLOS ONE**, v. 18, n. 12, p. e0289724–e0289724, 8 dez. 2023.

RIBEIRO, H. et al. Determinants of churn in telecommunication services: a systematic literature review. **Management Review Quarterly**, 28 fev. 2023.

RITIKA, H. J. Fraud detection and management for telecommunication systems using artificial intelligence (AI). In: **2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)**. IEEE, 2022. p. 1016-1022.

SAHA, L. et al. Deep Churn Prediction Method for Telecommunication Industry. **Sustainability**, v. 15, n. 5, p. 4543, 3 mar. 2023.

SIKRI, A. et al. Enhancing customer retention in telecom industry with machine learning driven churn prediction. **Scientific reports**, v. 14, n. 1, 7 jun. 2024.

SULIKOWSKI, P.; ZDZIEBKO, T. Churn factors identification from real-world data in the telecommunications industry: case study. **Procedia Computer Science**, v. 192, p. 4800–4809, 2021.

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