PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO

Bernardo Ruiz Fernandes

Predição de Churn: Uma Abordagem Comparativa utilizando XGBoost e Random Forest

PROJETO FINAL DE GRADUAÇÃO

CENTRO TÉCNICO CIENTÍFICO - CTC

DEPARTAMENTO DE INFORMÁTICA

Curso de Graduação em Engenharia da Computação

Rio de Janeiro, julho de 2025

Bernardo Ruiz Fernandes

Predição de Churn: Uma Abordagem Comparativa utilizando XGBoost e Random Forest

Projeto Final II ENG1133

Relatório de Projeto Final, apresentado ao programa de Graduação do Departamento de Informática da PUC-Rio como requisito parcial para a obtenção do título de Engenheiro de Computação.

Orientador: Augusto Cesar Espindola Baffa

Rio de Janeiro

julho de 2025

Resumo

Bernardo Ruiz Fernandes; Baffa, Augusto. **Predição de Churn: Uma Abordagem Comparativa utilizando XGBoost e Random Forest**. Rio de Janeiro, 2025. 45p. Relatório de Projeto Final – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro

A expansão dos veículos elétricos no Brasil demanda infraestrutura de recarga robusta e aplicativos eficientes para suporte aos usuários. Neste contexto, a retenção de clientes representa desafio estratégico, considerando que aplicativos móveis apresentam altas taxas de abandono e que o custo de aquisição de novos clientes supera o de retenção. A predição de churn utilizando técnicas de *Machine Learning* emerge como abordagem para identificar usuários com propensão ao abandono, permitindo intervenções de retenção. Este trabalho compara o desempenho dos algoritmos XGBoost e Random Forest na predição de churn em aplicativos de recarga para veículos elétricos, avaliando simultaneamente o impacto de diferentes técnicas de balanceamento de dados. Os resultados demonstraram melhor resultado com XGBoost e revelaram que técnicas de balanceamento não proporcionaram melhorias significativas em algoritmos robustos.

Palavras-chave

Predição de churn; XGBoost; Random Forest; Veículos elétricos; Aplicativo de Recarga; Balanceamento de dados.

Abstract

Bernardo Ruiz Fernandes; Baffa, Augusto (Advisor). **Churn Prediction: A Comparative Approach using XGBoost and Random Forest**. Rio de Janeiro, 2025. 45p. Final Project Report - Department of Informatics, Pontifical Catholic University of Rio de Janeiro.

The expansion of electric vehicles in Brazil demands robust charging infrastructure and efficient applications to support users. In this context, customer retention represents a strategic challenge, considering that mobile applications present high abandonment rates and that the cost of acquiring new customers exceeds that of retention. Churn prediction using Machine Learning techniques emerges as an approach to identify users with propensity to abandon, enabling retention interventions. This work compares the performance of XGBoost and Random Forest algorithms in churn prediction for electric vehicle charging applications, simultaneously evaluating the impact of different data balancing techniques. The results demonstrated better performance with XGBoost and revealed that balancing techniques did not provide significant improvements in robust algorithms.

Keywords

Churn prediction; XGBoost; Random Forest; Electric vehicles; Charging application; Data balancing.

Sumário

Lista de figuras

Lista de tabelas

1 Introdução

Segundo o *Global EV Outlook* 2024 da Agência Internacional de Energia (IEA), as vendas globais de carros elétricos atingiram 14 milhões de unidades em 2023, um aumento de 35% em comparação com 2022, representando aproximadamente 18% do mercado automobilístico mundial (IEA, 2024). Já no Brasil, esse crescimento é ainda mais acentuado, com os veículos elétricos registrando um aumento de 91% no primeiro semestre de 2023 em comparação ao mesmo período do ano anterior (ANFAVEA, 2023). Essa expansão da frota de veículos elétricos (VEs) demanda uma infraestrutura de recarga ampla, garantindo a autonomia dos usuários.

Somado ao desafio da infraestrutura de recarga, existe o da retenção de clientes - *churn*, já que 71% dos usuários desistem dentro do primeiro trimestre após o *download* (LOCALYTICS, 2019). A importância dessa retenção se deve ao fato de que um aumento de 5% desta pode elevar os lucros entre 25% e 95% (REICHHELD; SCHEFTER, 2000). Além disso, o custo de aquisição de um novo cliente pode ser de cinco a 25 vezes maior do que o custo de reter um existente (GALLO, 2014).

Neste contexto, a predição de churn utilizando técnicas de *Machine Learning* é uma abordagem estratégica para identificar antecipadamente usuários com maior probabilidade de abandono, possibilitando intervenções proativas de retenção. De acordo com (AMIN et al., 2019) há estudos sobre a eficácia destas técnicas em setores como telecomunicações e serviços bancários (KERAMATI et al., 2016), no entanto, sua análise em aplicativos de recarga para VEs permanece inexplorada.

Assim, este estudo compara dois modelos de *Machine Learning*, XGBoost e Random Forest, na predição de churn em um aplicativo de recarga para veículos elétricos. A pesquisa analisa características dos usuários que abandonam a plataforma, implementa técnicas de balanceamento de dados e avalia o desempenho dos modelos. Os resultados determinarão a eficácia das técnicas de *Machine Learning* neste contexto e identificarão qual algoritmo oferece melhor relação entre precisão preditiva e eficiência computacional para implementações práticas.

2 Situação Atual

A predição de churn usando técnicas de *Machine Learning* (ML) demonstrou resultados significativos em setores específicos, conforme evidenciado por Sabbeh (2018), que documentou taxas de acurácia de até 96% em telecomunicações, e por Keramati et al. (2016), que relataram métricas de AUC superiores a 0,85 em serviços bancários. No entanto, sua aplicação específica no contexto de aplicativos de recarga para veículos elétricos ainda é incipiente, embora os fundamentos metodológicos sejam transferíveis.

Na análise de técnicas preditivas para churn, Sabbeh (2018) realizou uma comparação sistemática entre dez algoritmos de *Machine Learning*, incluindo Random Forest, XGBoost, SVM e Regressão Logística. Os resultados demonstraram superioridade dos métodos ensemble, com Random Forest e AdaBoost atingindo 96% de acurácia. Esta avaliação comparativa estabeleceu parâmetros de desempenho que norteiam a seleção de algoritmos em contextos de retenção de clientes.

Complementando a eficácia dos métodos ensemble, Al-Shatnawi e Faris (2020) usaram diferentes técnicas de balanceamento junto ao XGBoost para minimizar os efeitos negativos do desbalanceamento de classes. Seus experimentos demonstraram que SMOTE e Random Oversampling elevaram o desempenho do classificador, atingindo valores próximos a 84% com SMOTE.

Somado às técnicas de balanceamento, a engenharia de *features* apresenta uma correlação direta entre a seleção adequada de características e o aumento da precisão preditiva dos modelos (AMIN et al., 2016). No setor de serviços, a seleção precisa de variáveis relacionadas aos padrões de compras e localização geográfica mostrou-se determinante para identificar usuários com propensão a abandono (AL-SHATNWAI; FARIS, 2020).

No contexto específico dos veículos elétricos, a literatura sobre predição de churn em aplicativos de recarga é escassa. Entretanto, estudos sobre comportamento de usuários de VEs, como os de Franke e Krems (2013), identificam padrões de ansiedade relacionados à autonomia do veículo que podem impactar significativamente a experiência com aplicativos de recarga, afetando consequentemente as taxas de churn.

Assim, embora existam lacunas na literatura específica sobre predição de churn em aplicativos de recarga para veículos elétricos, as metodologias e

técnicas desenvolvidas em outros setores fornecem bases sólidas para esta investigação. A combinação de algoritmos robustos como XGBoost e Random Forest com técnicas adequadas de balanceamento de dados e engenharia de *features* representa uma abordagem promissora para enfrentar este desafio emergente no setor de mobilidade elétrica.

3 Objetivos do Trabalho

Este trabalho tem como objetivo geral desenvolver e comparar modelos preditivos baseados em XGBoost e Random Forest para identificar antecipadamente o comportamento de churn em aplicativos de recarga para veículos elétricos, permitindo intervenções proativas que aumentem a retenção dos usuários.

A pesquisa busca identificar os principais fatores comportamentais e padrões de uso que influenciam o abandono do aplicativo. Conforme demonstrado por Amin et al. (2019), a identificação precisa desses fatores é fundamental para estratégias eficazes de retenção.

Um objetivo complementar consiste em implementar e avaliar diferentes técnicas de balanceamento de dados como Random Oversampling, SMOTE, ADASYN e Borderline SMOTE. Al-Shatnawi e Faris (2020) demonstraram que essas técnicas podem melhorar significativamente o desempenho de modelos de predição de churn, especialmente em conjuntos de dados desbalanceados, situação comum neste tipo de problema.

Pretende-se, ainda, desenvolver e comparar modelos utilizando XGBoost e Random Forest, avaliando seu desempenho quanto à acurácia, precisão, *recall*, F1-score e área sob a curva ROC. Sabbeh (2018) observou que esses algoritmos apresentam desempenho superior em problemas de predição de churn em telecomunicações, mas sua eficácia em aplicativos de recarga para VEs permanece inexplorada.

O sistema proposto visa apoiar desenvolvedores, gerentes de produto e equipes de marketing de aplicativos de recarga para VEs, permitindo identificar usuários com alto risco de abandono. Além disso, os resultados irão permitir melhorias na experiência do usuário, aspecto que Franke e Krems (2013) apontam como determinante na satisfação de proprietários de veículos elétricos.

Este estudo contribui para o avanço do estado da arte ao propor a primeira análise comparativa documentada de algoritmos de *Machine Learning* para predição de churn especificamente no contexto de aplicativos de recarga para VEs. Adicionalmente, a avaliação sistemática do impacto de diferentes técnicas de balanceamento fornece diretrizes metodológicas valiosas para problemas similares no setor de mobilidade elétrica, preenchendo uma lacuna significativa identificada na revisão de literatura.

4 Fundamentos Teóricos

Neste capítulo são apresentados os principais conceitos e técnicas utilizados para o desenvolvimento desta tese. Abordando tópicos desde temas técnicos de modelos de classificação e predição de churn, até assuntos mais voltados para a área de veículos elétricos.

4.1 Ecossistema de Veículos Elétricos

Este tópico apresenta o contexto dos veículos elétricos e sua infraestrutura de recarga, abordando desde o panorama global da mobilidade elétrica até as especificidades do cenário brasileiro.

4.1.1 Panorama Global da Mobilidade Elétrica

Os primeiros veículos elétricos (VEs) foram desenvolvidos por volta de 1830, com contribuições de inventores como Robert Anderson, Ányos Jedlik e Sibrandus Stratingh (EVBOX, 2023). Entre 1900 e 1912, a eletricidade alimentava um terço de todos os veículos nas estradas americanas, quando os carros elétricos eram preferidos por serem limpos, silenciosos e fáceis de manusear comparados aos ruidosos veículos a combustão (ENERGYSAGE, 2023).

Entretanto, o declínio dos veículos elétricos no início do século XX foi resultado em parte pela disparidade de preço, já que em 1912, um carro elétrico poderia custar até três vezes mais do que um veículo a gasolina (U.S. DEPARTMENT OF ENERGY, 2022). Simultaneamente, a descoberta de petróleo no Texas tornou a gasolina economicamente acessível. Essa combinação de fatores resultou na redução da produção de veículos elétricos em 1935 (U.S. DEPARTMENT OF ENERGY, 2022).

Contudo, três décadas depois, iniciou-se o ressurgimento dos veículos elétricos motivado por crises energéticas e preocupações ambientais (ENERGYSAGE, 2023). Os veículos elétricos estabeleceram-se como alternativa viável devido à sua integração eficiente com fontes renováveis de energia, reduzindo as emissões de ciclo de vida em até 69% quando comparados aos veículos convencionais (BIEKER, 2021).

Neste contexto, o marco decisivo ocorreu em 2008 com o lançamento do Tesla Roadster, o primeiro carro totalmente elétrico produzido em massa usando baterias de íon-lítio, capaz de percorrer mais de 320 quilômetros com uma única carga (VINFAST GLOBAL COMMUNITY, 2023). Desde então, as vendas globais de veículos elétricos apresentaram crescimento expressivo, atingindo quase 14 milhões de unidades em 2023, representando um aumento de 35% em relação a 2022 (IEA, 2024).

Um grande incentivador para esse crescimento foram as políticas públicas e incentivos governamentais, com mais de 90% das vendas globais de veículos leves cobertas por políticas que incentivam a adoção de VEs (IEA, 2023). A União Europeia tem apoiado a adoção comercial de VEs através de regulamentações e incentivos diversos, incluindo normativas de emissão de CO_2 para veículos pesados estabelecidas em 2019 (IEA, 2021).

As projeções de crescimento indicam que mais de um em quatro carros vendidos mundialmente em 2025 deve ser elétrico, já que as vendas globais de carros elétricos estão a caminho de ultrapassar 20 milhões (IEA, 2025). Este panorama internacional contrasta com a realidade brasileira, que enfrenta desafios próprios na adoção de veículos elétricos.

4.1.2 Cenário Brasileiro de Veículos Elétricos

O Brasil vem apresentando um crescimento no setor de mobilidade elétrica, com vendas que atingiram quase 94 mil veículos eletrificados em 2023, representando um aumento de 91% em relação a 2022 (CNN BRASIL, 2024). Este desempenho indica uma transformação significativa no mercado automotivo brasileiro, impulsionada principalmente pelos veículos plug-in (PHEV), que representaram 56% das vendas no período (ABVE, 2024), como podemos ver pela figura 4.1.

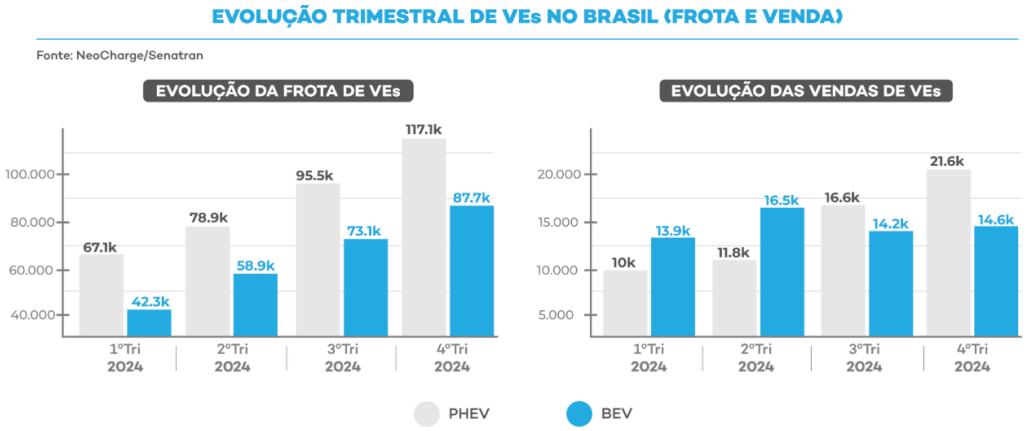

Figura 4.1: Evolução dos VEs no Brasil em 2024 (NeoCharge, 2024)

Apesar dos avanços, permanecem desafios estruturais significativos, principalmente relacionados à infraestrutura de recarga. O Brasil contava com aproximadamente 4.300 estações de recarga em dezembro de 2023, número considerado ainda insuficiente para suportar o crescimento acelerado da frota (PORTAL SUSTENTABILIDADE, 2024). A distribuição desigual desses pontos, concentrados majoritariamente nas regiões Sul e Sudeste, representa um obstáculo para a expansão nacional da mobilidade elétrica (FEI, 2023).

Os investimentos em infraestrutura têm aumentado de forma acelerada, com a rede de recarga superando a marca simbólica de 10 mil eletropostos em agosto de 2024, representando um crescimento de 179% em relação ao período anterior. Esta expansão tem permitido que os veículos elétricos (*BEV* e *PHEV*) ampliem sua participação no mercado automotivo brasileiro, representando 70% das vendas de veículos eletrificados entre janeiro e setembro de 2024 (ABVE, 2024).

Outros desafios para o mercado brasileiro incluem o alto custo dos veículos elétricos, a limitada oferta de modelos nacionais e a necessidade de maior capacitação técnica para manutenção. O cenário tende a melhorar com o anúncio de investimentos de montadoras como BYD, GWM e Stellantis para fabricação de veículos elétricos em território nacional, o que deve contribuir para redução de preços e ampliação do mercado (EXAME, 2023).

4.1.3 Pontos de Recarga

Os pontos de recarga para veículos elétricos no Brasil distribuem-se em dois segmentos: privados e públicos. Os carregadores privados respondem por mais de 83% do carregamento dos VEs nacionais (VOOLTA, 2024). Entretanto, a expansão dos pontos de recarga públicos — localizados em shoppings, rodovias, estacionamentos e postos de combustível — é fundamental para ampliar o alcance dos veículos elétricos e viabilizar viagens de longa distância.

Os pontos de recarga públicos são estações equipadas para fornecer energia aos veículos elétricos de forma segura e eficiente, funcionando por meio de diferentes tipos de conectores e tecnologias (TUPI MOBILIDADE, 2024). Assim, a funcionalidade dessas estações transcende a simples transferência de energia, incorporando sistemas integrados que monitoram o carregamento, exibem informações em tempo real e oferecem opções de pagamento digital.

A operação dos pontos de recarga públicos demanda sistemas integrados de autenticação, monitoramento e pagamento digital que estabelecem a interface tecnológica entre usuários e infraestrutura. Nesse âmbito, aplicativos como *PlugShare*, Tupi e NeoCharge são essenciais para viabilizar o acesso e operação desses pontos de recarga públicos.

4.2 Aplicativos de Recarga para Veículos Elétricos

Os aplicativos de recarga para veículos elétricos constituem plataformas digitais que integram usuários e operadores por sistemas centralizados de gestão. Oferecem funcionalidades como localização de pontos de recarga, verificação de disponibilidade em tempo real, reserva de vagas e processamento de pagamentos (NANNICINI; LIBERTI, 2014).

A arquitetura utiliza protocolos como OCPP (*Open Charge Point Protocol*), que define a comunicação padronizada entre estações de carregamento e sistemas de gestão central, permitindo monitoramento remoto, controle de sessões e coleta de dados operacionais (OPEN CHARGE ALLIANCE, 2023).

Figura 4.2: Infográfico OCPP (GRIDX, 2024)

O mercado global projeta expansão de USD 16,69 bilhões em 2024 para USD 172,9 bilhões até 2033, com crescimento anual de 29,31% (IMARC GROUP, 2024). O crescimento das vendas de VEs — 14 milhões de unidades em 2023 — impulsiona demanda por soluções especializadas de software, como a Plataforma *PlugShare* com mais de 140.000 estações catalogadas (IEA, 2024).

No Brasil, aplicativos nacionais como Tupi e EZVolt competem desenvolvendo soluções adaptadas ao mercado local, enfrentando desafios específicos como integração com múltiplos sistemas de pagamento brasileiros e padronização de protocolos de comunicação entre diferentes fornecedores de infraestrutura (VOOLTA, 2024).

4.2.1 Tupi Mobilidade

A Tupi Mobilidade, fundada em 2019 como Tupinambá Energia, passou por rebranding em 2024 para focar exclusivamente no desenvolvimento de tecnologias de recarga para veículos elétricos (PRNEWSWIRE, 2024). A mudança estratégica ocorreu após a venda de sua rede física de carregadores para a Raízen em dezembro de 2023, permitindo concentração total em soluções de software.

Os dados de 2024 posicionam a Tupi Mobilidade como líder no segmento de aplicativos de recarga no Brasil, concentrando aproximadamente 50% dos proprietários de veículos elétricos nacionais e registrando crescimento mensal superior a 5.000 novos usuários (EXAME, 2024). Esta liderança fundamenta-se na conexão tecnológica de 1.015 estações de recarga através de sua plataforma

— representando o maior ecossistema digital integrado de mobilidade elétrica do Brasil, com 650 mil recargas processadas e 71 mil usuários ativos (TUPI MOBILIDADE, 2024).

O diferencial tecnológico da Tupi baseia-se na arquitetura dupla de produtos: o aplicativo Tupi Recarga para usuários finais e o software Tupi Conecta para operadores de infraestrutura, criando um ecossistema integrado de mobilidade elétrica. A empresa tornou-se a primeira B Corp certificada do setor de recarga elétrica no Brasil, validando altos padrões de desempenho social e ambiental (TUPI MOBILIDADE, 2024).

4.2.2 Desafios dos Aplicativos de Recarga

Um dos principais problemas enfrentados pelos aplicativos de recarga para veículos elétricos no Brasil é a retenção de usuários, uma vez que as empresas adotaram a prática de desenvolver aplicativos próprios que exibem exclusivamente carregadores vinculados aos seus sistemas. Essa decisão fragmenta a experiência do usuário e compromete a transparência do mercado (AMAZÔNIA ELETROVIAS, 2024).

Outro problema significativo reside na interoperabilidade de eletropostos e na ausência de padronização. Existem diferentes tipos de conectores, protocolos de comunicação e sistemas de pagamento, criando ambiente tecnológico heterogêneo. Na ausência de protocolo padronizado, os operadores enfrentam dificuldades significativas para integrar carregadores de diversos fabricantes em rede única (VOLTBRAS, 2024).

Nesse contexto, aplicativos como a Tupi buscam soluções para consolidação no mercado. Utilizam predição de churn para antecipar o abandono do aplicativo e desenvolvem múltiplas integrações específicas para conectar-se aos diversos sistemas operacionais disponíveis.

4.2 Churn

Este tópico aborda os conceitos fundamentais de churn e sua aplicação no contexto de aplicativos móveis. São apresentadas as definições conceituais, as

características específicas do abandono em ambientes digitais e os fatores comportamentais que antecedem a decisão de abandono.

4.2.1 Conceitos Fundamentais de Churn

O termo churn, derivado da expressão inglesa "customer churn", refere-se ao processo pelo qual clientes interrompem a utilização de determinado serviço (VERBEKE et al., 2012). No contexto empresarial, representa a taxa de abandono de clientes em período específico, sendo uma métrica fundamental para avaliação de modelos de negócio.

Neste âmbito, a definição temporal de churn varia conforme o setor e tipo de serviço, impactando diretamente a precisão dos modelos preditivos e a eficácia das estratégias de retenção. Empresas de telecomunicações frequentemente consideram churn após 30 dias de inatividade, enquanto serviços de streaming podem estabelecer períodos mais extensos (KUMAR; REINARTZ, 2016).

Somado a isso, adquirir novos clientes custa entre 5 a 25 vezes mais do que reter clientes existentes (REICHHELD; SASSER, 1990). Assim, a redução das taxas de churn por meio de estratégias proativas tornou-se prioridade para serviços baseados em recorrência.

4.2.2 Churn em Aplicativos

O churn em aplicativos é influenciado por fatores específicos do ambiente digital, como a facilidade de instalação e desinstalação dos produtos. Combinada com a abundância de alternativas disponíveis, resulta em taxas de abandono significativamente superiores aos serviços convencionais (LOCALYTICS, 2017).

Em razão desse cenário, aplicativos perdem aproximadamente 77% dos usuários nas primeiras 72 horas após a instalação inicial, com apenas 4% dos usuários permanecendo ativos após 90 dias (ADJUST, 2020). Situação que demanda estratégias de engajamento e retenção adaptadas ao digital.

4.2.3 Fatores Comportamentais de Abandono

O churn ocorre raramente de forma abrupta, mas sim apresentando sinais de intenção na interação do usuário com o serviço (BUCKINX; VAN DEN POEL, 2005). Dessa forma, a identificação de padrões comportamentais anteriores ao abandono é fundamental para desenvolvimento de modelos preditivos eficazes.

Um padrão observado que indica o churn é a redução na frequência de uso, diminuição do tempo de sessão e menor profundidade de navegação no aplicativo (BAECKE; VAN DEN POEL, 2011). Além disso, usuários que abandonam aplicativos apresentam períodos de inatividade crescentes, redução na utilização de funcionalidades e menor engajamento com notificações (LEMON; VERHOEF, 2016).

Particularmente nos aplicativos de recarga, existem fatores comportamentais como frustração com indisponibilidade de estações, experiências negativas de carregamento e dificuldades no processo de pagamento. A análise desses comportamentos permite identificar usuários em risco e implementar estratégias direcionadas de retenção antes da decisão final de abandono (SCHAAL et al., 2014).

4.3 *Machine Learning* para predição de Churn

O *Machine Learning* representa um subcampo da inteligência artificial que desenvolve algoritmos capazes de identificar padrões em dados históricos para realizar previsões automatizadas. Diferentemente da programação tradicional, os sistemas de ML extraem regras implícitas dos dados, eliminando a necessidade de codificação manual de todas as condições possíveis (RAMOS, 2024).

A taxonomia do aprendizado de máquina compreende três paradigmas principais. O aprendizado supervisionado utiliza dados com rótulos conhecidos para construir modelos preditivos. O não supervisionado explora estruturas em dados sem rótulos, identificando agrupamentos naturais. O aprendizado por reforço desenvolve estratégias por interações com ambientes dinâmicos, maximizando recompensas ao longo do tempo.

O desenvolvimento de modelos ML segue etapas sequenciais bem definidas. A coleta e preparação dos dados constitui a fase inicial, envolvendo limpeza e transformação das informações brutas. Posteriormente, ocorre a

seleção e treinamento de algoritmos apropriados ao problema específico. A avaliação da capacidade preditiva por métricas adequadas precede a implantação final em ambiente produtivo (GARCIA et al., 2020).

A construção de modelos robustos enfrenta o desafio do equilíbrio entre *underfitting* e *overfitting*. O *underfitting* surge quando algoritmos excessivamente simples não conseguem capturar padrões complexos nos dados, gerando baixo desempenho tanto no treino quanto no teste. O *overfitting* ocorre quando modelos muito complexos memorizam ruídos específicos do conjunto de treinamento, comprometendo sua capacidade de generalização para dados não observados (HANDELMA et al., 2018).

4.3.1 Aprendizado de Máquina Supervisionado

O principal objetivo dos algoritmos de *Machine Learning* Supervisionado, — foco principal desse projeto — é que o modelo aprenda com casos já rotulados para realizar a predição dos casos que ainda não tem classificação (MITCHELL, 1997). A predição de churn enquadra-se como classificação binária, categorizando usuários entre propensos ao abandono ou à retenção (JUNIOR, 2016).

A divisão dos dados em conjuntos de treino, validação e teste estabelece a metodologia para desenvolvimento robusto. O conjunto de treinamento alimenta o aprendizado inicial, enquanto a validação orienta ajustes de hiperparâmetros. O conjunto de teste fornece avaliação imparcial do desempenho final do modelo (HANDELMA et al., 2018).

4.3.2 Desafios em Predição de Churn

O desbalanceamento de classes representa o principal desafio na predição de churn. Naturalmente, a proporção de usuários que abandonam serviços é inferior àqueles que permanecem ativos, criando assimetria nos dados de treinamento. Esta característica impacta significativamente o desempenho dos algoritmos tradicionais (CHAWLA et al., 2002).

Algoritmos convencionais tendem a favorecer a classe majoritária durante o processo de aprendizado, resultando em modelos com alta acurácia geral, mas

baixa capacidade de identificar casos de churn. Assim, a acurácia tradicional torna-se métrica inadequada para avaliação de desempenho neste contexto.

Neste cenário, métricas específicas como precisão, *recall* e F1-score fornecem avaliação mais apropriada. A área sob a curva ROC (AUC-ROC) oferece medida robusta da capacidade discriminativa do modelo, sendo menos sensível ao desbalanceamento de classes. Estas métricas permitem avaliação mais precisa do desempenho real dos algoritmos em cenários de churn (BRADLEY, 1997).

4.3.3 Random Forest

O algoritmo Random Forest, desenvolvido por Breiman (2001), representa uma extensão do método de árvores de decisão baseada no conceito de *ensemble learning*. Este algoritmo constrói múltiplas árvores de decisão durante o treinamento e produz previsões através da agregação dos resultados individuais.

A robustez do Random Forest deriva de duas fontes principais de aleatoriedade. Primeiramente, cada árvore é treinada em uma amostra bootstrap diferente do conjunto original, técnica conhecida como *bagging*. Simultaneamente, em cada nó de divisão, apenas um subconjunto aleatório das variáveis é considerado para determinar a melhor divisão, reduzindo a correlação entre as árvores individuais (BREIMAN, 2001). O esquema visual do processo de treinamento do Random Forest é apresentado na Figura 4.3 a seguir.

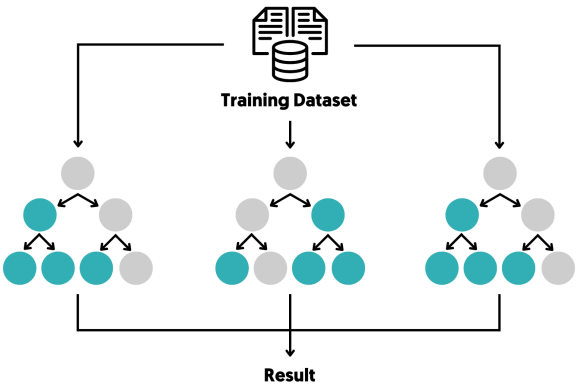

Figura 4.3: Infográfico da Random Forest (DIDA, 2024)

Para predição de churn, o Random Forest apresenta características particularmente adequadas. O algoritmo demonstra robustez natural a dados desbalanceados através da agregação de múltiplas árvores, reduzindo a tendência de favorecer a classe majoritária. Somado a isso, o método oferece interpretabilidade através de medidas de importância das variáveis, permitindo identificar quais fatores comportamentais mais influenciam a decisão de abandono dos usuários (LIAW; WIENER, 2002).

4.3.4 Extreme Gradient Boosting (XGBoost)

O XGBoost, desenvolvido por Chen e Guestrin (2016), implementa uma versão otimizada do algoritmo *gradient boosting* que constrói modelos sequencialmente. Cada novo modelo corrige os erros dos anteriores, criando um *ensemble* forte a partir de aprendizes fracos.

O algoritmo utiliza funções de perda diferenciáveis e incorpora termos de regularização para controlar o *overfitting*. A implementação inclui otimizações como tratamento automático de valores ausentes, paralelização de operações e técnicas de poda para reduzir complexidade computacional, resultando em melhor desempenho em competições de *Machine Learning* (CHEN; GUESTRIN, 2016).

Para predição de churn, o XGBoost oferece vantagens específicas em cenários de dados desbalanceados. O algoritmo permite otimização direta de métricas como AUC-ROC através de funções objetivo customizadas, superando limitações da acurácia tradicional. Adicionalmente, a capacidade de ajustar pesos das classes durante o treinamento e a flexibilidade na definição de *thresholds* de classificação proporcionam controle refinado sobre a sensibilidade na detecção de casos de abandono (NIELSEN, 2016).

5 Projeto e Especificação do Sistema

Este projeto desenvolve um sistema de predição de churn de usuários no contexto de aplicativos para recarga de veículos elétricos utilizando algoritmos de *Machine Learning*. O sistema utiliza dados históricos de comportamento dos usuários do aplicativo Tupi Recarga para identificar padrões que antecedem o abandono da plataforma.

Este capítulo apresenta os recursos tecnológicos utilizados na implementação e as características dos dados empregados nos modelos.

5.1 Arquitetura do Sistema

O sistema estrutura-se em quatro módulos separados que operam sequencialmente para transformar dados brutos em um modelo que pode realizar a predição de churn, conforme ilustrado na Figura 5.1. Esta arquitetura modular facilita a manutenção e permite separação clara de responsabilidades.

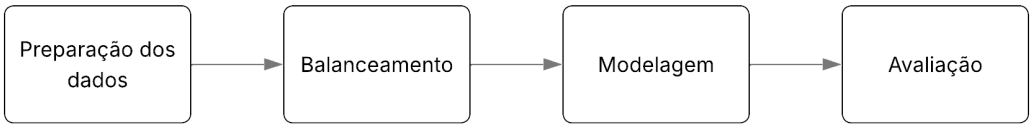

Figura 5.1: Fluxo de execução dos módulos

- **Módulo de Preparação de Dados:** esse primeiro módulo integra as duas bases de dados da Tupi Mobilidade que estão disponíveis em formato csv e implementa rotinas de limpeza para tratamento de valores ausentes. Sua saída é uma base de dados tratada e agregada nos moldes do que será apresentado em 5.2.2.
- **Módulo de Balanceamento:** o segundo módulo implementa cinco estratégias para tratamento de dados desbalanceados, aplicadas exclusivamente no conjunto de treinamento para preservar a integridade

do conjunto de teste. O objetivo desse módulo é permitir a comparação futura de forma justa entre os métodos de balanceamento.

- **Módulo de Modelagem:** no terceiro módulo temos a implementação dos algoritmos XGBoost e Random Forest com hiperparâmetros otimizados. Gerencia o treinamento dos modelos, validação cruzada e geração de predições probabilísticas. Aqui os algoritmos são aplicados nos conjuntos de dados gerados pelo módulo de balanceamento. Os Hiperparâmetros dos modelos serão apresentados nas próximas seções.

- **Módulo de Avaliação:** por fim, no quarto módulo calculam-se métricas de performance para classificação binária e extrai a importância das *features* através de métodos nativos dos algoritmos, fornecendo informações sobre fatores comportamentais determinantes para o churn. Esse módulo permite a avaliação de resultados para definição dos melhores modelos.

5.2 Especificação dos Dados

A construção do *dataset* final envolveu a integração de duas bases principais da Tupi Mobilidade, a fim de otimizar a qualidade e disponibilidade dos dados para o projeto. A Figura 5.2 ilustra de forma simplificada a estrutura de dados da Tupi Mobilidade e as relações entre as entidades do sistema para as bases utilizadas nesse projeto.

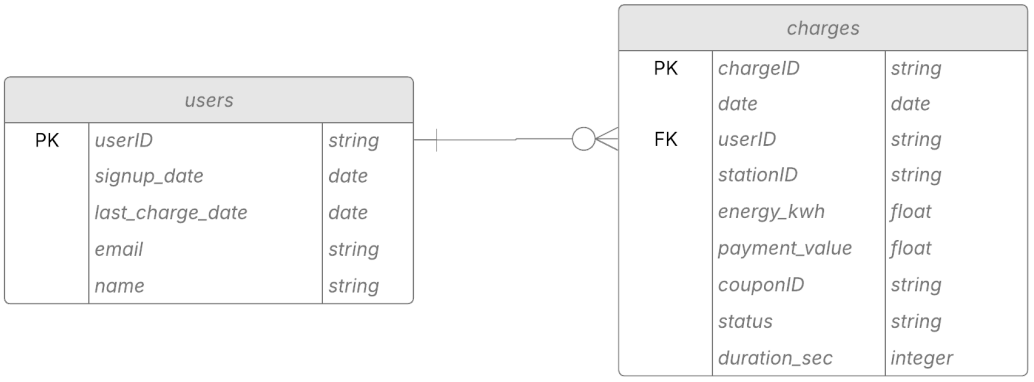

Figura 5.2: Modelo dos dados disponíveis

A respeito das bases, na denominada **`users`** contém informações de cadastro do usuário, incluindo identificadores únicos, datas de registro e características do perfil. A fim de manter o caráter anônimo dos dados, os dados de **`email`** e de **`name`** não serão fornecidos no material apresentado, esses dados não são necessários para a construção do modelo.

Já a segunda base apresentada na Figura 5.2, a charges, temos o histórico completo de carregamento com datas, durações, valores monetários, consumo energético e status de conclusão. Essa base está conectada com a base de **`users`** pelo campo **`userID`**, que é o identificador único do usuário.

Esses conjuntos de dados extraídos do banco de dados da Tupi Recarga representam parte de um sistema mais complexo de informações, a caráter de foco no projeto e de compreensão somente serão usados dados que estão sendo abrangidos por esses dois conjuntos.

5.2.1 Dataset Final

O primeiro módulo de código tem como objetivo adquirir, agrupar e tratar os dados das bases de dados apresentadas anteriormente. Este módulo implementa as agregações ao nível de usuário, uma vez que nosso objetivo é prever o churn do usuário. As *features*, campos do *dataset* final, foram idealizadas com base no aprendizado teórico de importância de dados como consumo, valor transacionado, duração de sessão, diversidade de compras, entre outros campos idealizados especificamente com o contexto de aplicativos de recarga para carros elétricos

O *dataset* resultante apresenta 51.306 registros distribuídos em 13 *features* preditoras, além da variável target churn. A estrutura inclui um identificador alfanumérico (**`userID`**) e 12 *features* numéricas derivadas do comportamento transacional. Em seguida serão apresentadas as 13 *features* do *dataset*:

- **userID:** identificador único do usuário, representado por string alfanumérica anonimizada para garantir conformidade com diretrizes de privacidade de dados.

- **transaction_frequency_monthly:** valor de transações médias por mês com recarga, expressa como valor numérico de ponto flutuante. Esta métrica quantifica o padrão de engajamento recorrente do usuário com a plataforma. Quanto mais recargas mensais, maior o valor da feature.

- **total_transactions:** contagem total de transações completadas pelo usuário ao longo de todo o período de análise. Valor numérico inteiro que representa o volume histórico de atividade.

- **avg_transaction_duration:** duração média das transações em segundos, calculada como média de todas as sessões de recarga do usuário. Valor numérico de ponto flutuante.

- **avg_days_between_transactions:** intervalo médio em dias entre transações consecutivas, representando a regularidade temporal dos padrões de uso. Valores negativos (-1) indicam usuários com transação única.

- **total_transaction_value:** valor monetário total transacionado pelo usuário em reais (R$), representado como ponto flutuante. Métrica que quantifica a contribuição financeira histórica do usuário.

- **avg_transaction_value:** valor médio por transação em reais (R$), calculado através da divisão do valor total pelo número de transações. Indicador do perfil de gasto por sessão.

- **monthly_spend_avg:** gasto médio mensal do usuário em reais (R$), obtido pela divisão do valor total transacionado pelo número de meses ativos. Métrica que normaliza o comportamento financeiro temporal.

- **station_diversity:** quantidade de estações de recarga distintas utilizadas pelo usuário. Valor numérico inteiro que representa diversidade e flexibilidade de uso.

- **weekend_transaction_ratio:** proporção de transações realizadas em finais de semana, expressa como valor decimal entre 0 e 1. Indicador de padrões de mobilidade e uso recreativo.

- **night_transaction_ratio:** proporção de transações realizadas no período noturno (18h às 6h), expressa como valor decimal entre 0 e 1. Métrica que caracteriza preferências temporais de uso.

- **failed_transaction_attempts:** contagem de tentativas de transações não concluídas com sucesso. Valor numérico inteiro que indica potenciais problemas de experiência do usuário.

- **coupon_usage_frequency:** frequência de utilização de cupons promocionais pelo usuário. Valor numérico que quantifica a resposta a estratégias de retenção e promoções.

- **avg_energy_consumed:** quantidade média de energia consumida por transação, medida em quilowatts-hora (kWh). Valor numérico de ponto flutuante que caracteriza o perfil de consumo energético.

- **tenure_days:** número de dias desde o cadastro do usuário até a data de coleta dos dados. Valor numérico inteiro que representa a antiguidade da relação comercial.

- **churn:** variável target binária que indica o status do usuário: 0 para usuários ativos e 1 para usuários que abandonaram a plataforma. Esta é a variável dependente que os modelos preditivos buscam classificar.

5.3 Bibliotecas e Dependências

As seguintes bibliotecas do Python foram necessárias para a construção do projeto:

- **Scikit-learn:** núcleo das operações de *Machine Learning*, incluindo divisão de dados, implementação do Random Forest e métricas de avaliação (acurácia, precisão, *recall*, F1-score, curvas ROC). Inclui *StandardScaler* para normalização de *features*.

- **Imblearn:** técnicas de *oversampling* para *datasets* desbalanceados - SMOTE, ADASYN, BorderlineSMOTE e RandomOverSampler.

- **XGBoost:** implementação do modelo de *gradient boosting* utilizado no projeto.

Essa combinação de bibliotecas cria o ambiente para desenvolvimento de soluções de classificação, desde o pré-processamento até a avaliação final dos modelos.

5.4 Configuração dos Algoritmos

A configuração adequada dos hiperparâmetros constitui etapa fundamental para obtenção de modelos robustos em predição de churn. Os algoritmos selecionados foram configurados baseando-se em práticas estabelecidas na literatura e características específicas do *dataset* de aplicativos de recarga.

5.4.1 Configuração do Random Forest

Serão apresentados nessa seção os hiperparâmetros utilizados no modelo do Random Forest apresentando uma justificativa para cada decisão.

- **n_estimators=100**: define o número de árvores de decisão. Valores inferiores a 50 tendem a produzir *underfitting*, enquanto valores superiores a 200 oferecem ganhos marginais com aumento significativo do tempo de treinamento. Para *datasets* com aproximadamente 50.000 registros, 100 estimadores fornecem estabilidade adequada nas predições (BREIMAN, 2001).

- **max_depth=10**: controla a profundidade máxima de cada árvore individual, limitando a complexidade do modelo para prevenir *overfitting*. Em problemas de churn com 13 *features*, profundidades entre 8 e 12 demonstram eficácia (LIAW; WIENER, 2002).

- **class_weight='balanced'**: ajusta automaticamente os pesos das classes inversamente proporcionais às suas frequências no *dataset*. Neste contexto, com 70% de churn, o parâmetro atribui maior peso à classe minoritária (usuários ativos), compensando o desbalanceamento natural dos dados.

Os demais parâmetros permaneceram com valores padrão do scikit-learn configurações que demonstraram robustez em múltiplos domínios de aplicação.

5.4.2 Configuração do XGBoost

O algoritmo XGBoost foi parametrizado considerando suas características específicas de *gradient boosting*:

- **n_estimators=100**: número de árvores construídas sequencialmente no processo de *boosting*. Diferentemente do Random Forest, cada árvore corrige erros das anteriores, tornando o modelo mais sensível a este parâmetro. Valores entre 50 e 150 demonstraram eficácia em problemas de churn (CHEN; GUESTRIN, 2016).

- **max_depth=6**: profundidade máxima das árvores individuais. O XGBoost utiliza árvores mais rasas que o Random Forest devido ao aprendizado sequencial — árvores entre 3 e 8 níveis capturam interações complexas sem comprometer a generalização.

- **learning_rate=0.1**: taxa de aprendizado que controla a contribuição de cada árvore para o modelo final. Valores menores (0.01-0.05) exigem mais estimadores para convergência, enquanto valores maiores (0.2-0.3) aceleram o treinamento, mas aumentam a instabilidade. O valor 0.1 estabelece convergência estável com eficiência computacional adequada.

- **objective='binary:logistic'**: define a função objetivo para classificação binária. Esta configuração produz probabilidades calibradas que facilitam a interpretação dos resultados.

- **eval_metric='logloss'**: métrica de avaliação durante o treinamento que monitora a perda logarítmica. Esta escolha é para maximizar a qualidade das probabilidades preditas, aspecto fundamental em sistemas de retenção.

- **scale_pos_weight**: calculado automaticamente como razão entre classes (negativa/positiva), compensando o desbalanceamento através de penalização diferenciada dos erros de classificação.

Parâmetros de regularização permaneceram em valores padrão configurações que oferecem regularização L2 moderada sem comprometer a capacidade preditiva em *datasets* de tamanho médio.

5.4.3 Configuração dos Algoritmos de Balanceamento

Para avaliar o impacto de diferentes abordagens no tratamento de dados desbalanceados, implementamos cinco estratégias distintas aplicadas exclusivamente no conjunto de treinamento. Esta metodologia preserva a integridade da avaliação ao evitar vazamento de informação para o conjunto de teste.

- **Baseline sem modificação**: utilizou os dados originais mantendo a proporção natural de 70% churn e 30% usuários ativos. Esta configuração serviu como referência para comparação quantitativa com as demais técnicas de balanceamento.

- **Random Oversampling**: duplicou aleatoriamente instâncias da classe minoritária até alcançar proporção equilibrada de 50% para cada classe. Neste contexto, a técnica oferece abordagem computacionalmente simples que estabelece parâmetros de comparação para métodos mais sofisticados.

- **SMOTE (Synthetic Minority Oversampling Technique)**: gerou exemplos sintéticos da classe minoritária por interpolação linear entre k-vizinhos mais próximos no espaço de características (CHAWLA et al., 2002). O algoritmo utiliza 5 vizinhos por padrão, criando amostras sintéticas que preservam a distribuição estatística original.

- **ADASYN (Adaptive Synthetic Sampling)**: estendeu a abordagem SMOTE adaptando a densidade de geração sintética baseada na complexidade local de cada instância (HE et al., 2008). Assim, regiões com maior dificuldade de aprendizado recebem proporcionalmente mais exemplos sintéticos.

- **Borderline SMOTE**: concentrou a geração de exemplos sintéticos exclusivamente em instâncias próximas à fronteira de decisão entre classes (HAN et al., 2005). Esta variação identifica automaticamente pontos de borda através da análise dos k-vizinhos, teoricamente oferecendo melhor capacidade discriminativa.

Somado à aplicação controlada no conjunto de treinamento, todas as técnicas mantiveram os hiperparâmetros padrão das respectivas implementações na biblioteca imblearn. Esta padronização elimina vieses de configuração específica e permite comparação direta entre metodologias.

6 Implementação e Avaliação

Este capítulo documenta o processo de implementação, os obstáculos encontrados e os resultados obtidos na comparação entre XGBoost e Random Forest e dos diferentes tipos de balanceamentos.

6.1 Metodologia da Implementação

Primeiramente foi realizado o tratamento dos dados e definidos parâmetros fundamentais como o período para categorização do churn e especificações dos campos que compõem o *dataset* final. Estes aspectos conceituais estabeleceram as bases para execução experimental consistente dos modelos preditivos.

Em seguida, a divisão dos dados utilizou estratificação 80/20 para treino e teste, preservando a proporção original de classes em ambos os conjuntos. Esta metodologia assegurou representatividade estatística adequada para o *dataset* com 51.306 registros e distribuição de 70% churn.

Para cada uma das cinco técnicas de balanceamento, treinamos ambos os algoritmos separadamente, resultando em dez combinações experimentais. Neste contexto, todas as modificações de balanceamento foram aplicadas exclusivamente no conjunto de treino para evitar vazamento de informação. O conjunto de teste permaneceu inalterado durante todo o processo experimental.

Após a aplicação das dez combinações, os modelos gerados foram validados com base nos mesmos dados de teste. Assim, cada combinação foi avaliada sob condições idênticas, viabilizando comparações diretas dos resultados obtidos. Esta padronização eliminou vieses metodológicos na análise comparativa.

Por fim, foram calculadas as métricas padrão de avaliação de modelos de classificação, priorizando F1-Score como indicador principal. Somado a isso, os resultados foram organizados em *dataset* estruturado para análise comparativa sistemática. Estes dados são apresentados detalhadamente na seção 6.4.1.

6.2 Definição de Parâmetros Críticos

A classificação de churn exigiu estabelecer um *threshold* temporal específico para o contexto de aplicativos de recarga. Definimos 20 dias de inatividade como critério de abandono, baseando-nos na literatura sobre retenção em aplicativos móveis que indica baixa probabilidade de retorno espontâneo após períodos de 14-21 dias (LOCALYTICS, 2019).

Junto a decisão de 20 dias de inatividade como categorização para o churn, foi decidido utilizar dados dos últimos 6 meses de atividades a fim de manter um volume de dados que não tivessem padrões de consumos antigos, mas que tivessem volume significativo para o modelo.

Esta definição resultou em uma distribuição desbalanceada, com 70% dos usuários classificados como churn e 30% como ativos. Embora esta proporção possa parecer elevada, reflete a realidade dos aplicativos de recarga, onde muitos usuários experimentam o serviço brevemente antes de descontinuar o uso. Este cenário de desbalanceamento tornou-se ideal para avaliar o impacto das técnicas de balanceamento, alinhando-se com os objetivos da pesquisa.

6.4 Resultados

As métricas de performance incluíram *Accuracy*, *Precision*, *Recall*, F1-Score e AUC-ROC, selecionadas especificamente para problemas de classificação desbalanceada. O F1-Score serviu como métrica principal para equilibrar adequadamente *Precision* e *Recall*, aspectos fundamentais na predição de churn.

6.4.1 Performance Comparativa dos Modelos

A Tabela 6.1 apresenta os resultados consolidados dos experimentos realizados, comparando os dois algoritmos com diferentes técnicas de balanceamento.

Balanceamento	Modelo	F1-Score	AUC-ROC
Sem Balanceamento	Random Forest	0,9057	0,8926
Sem Balanceamento	XGBoost	0,9230	0,9130
Random Oversampling	Random Forest	0,8820	0,8934
Random Oversampling	XGBoost	0,9013	0,9125
SMOTE	Random Forest	0,8856	0,8894
SMOTE	XGBoost	0,9120	0,9086
ADASYN	Random Forest	0,8779	0,8908
ADASYN	XGBoost	0,9114	0,9065
Borderline SMOTE	Random Forest	0,8710	0,8898
Borderline SMOTE	XGBoost	0,9132	0,9096

Tabela 6.1: Resultados obtidos com os modelos

Os resultados demonstram superioridade consistente do XGBoost em relação ao Random Forest em todas as configurações testadas. O modelo XGBoost sem balanceamento aparece como a melhor opção, com F1-Score de 0,9230 e AUC-ROC de 0,9130. Essa melhor pontuação do XGBoost se dá pelo modelo utiliza *gradient boosting* sequencial, onde cada árvore corrige especificamente os erros das anteriores, criando aprendizado adaptativo. Assim, essa abordagem sequencial permite capturar nuances comportamentais que o Random Forest não consegue detectar.

Neste contexto, a superioridade do F1-Score do XGBoost (0,9230 vs 0,9057) indica capacidade superior de equilibrar precisão e *recall* — importante em sistemas de retenção onde tanto falsos positivos quanto falsos negativos geram custos operacionais significativos. Somado ao F1-Score, a métrica AUC-ROC confirma a robustez discriminativa do XGBoost (0,9130 vs 0,8926). Valores acima de 0,90 indicam uma capacidade de distinguir entre usuários que permanecerão ativos e aqueles que abandonarão a plataforma. Esta característica mostra-se ser importante para implementações práticas, onde a confiabilidade das predições determina o sucesso das estratégias de retenção.

6.4.2 Análise das Técnicas de Balanceamento

Contrariamente às expectativas estabelecidas pela literatura, todas as técnicas de balanceamento degradam a performance do modelo superior. O XGBoost sem balanceamento superou todas as configurações alternativas, com diferenças variando entre 1,1% (SMOTE) e 2,3% (Random Oversampling) no F1-Score. Esta degradação indica que algoritmos modernos lidam com desbalanceamento moderado.

O fenômeno observado alinha-se com Chen e Guestrin (2016), que demonstram a capacidade do XGBoost de otimizar diretamente métricas como AUC através de funções objetivo customizadas. Neste contexto, técnicas de balanceamento podem introduzir ruído sintético que compromete a qualidade dos padrões aprendidos, especialmente quando o algoritmo já possui mecanismos internos para compensar o desbalanceamento através de pesos de classe adaptativos.

6.4.3 Importância das Features

A identificação das *features* mais relevantes auxilia na melhoria dos modelos preditivos de churn. Neste contexto, denomina-se feature cada variável preditora utilizada pelos algoritmos para estabelecer padrões discriminativos entre usuários.

Os dados de importância das *features* foram obtidos através das funções nativas dos modelos, que calculam a contribuição de cada variável para a redução da impureza nas divisões das árvores de decisão.

Os resultados foram apresentados na Tabela 6.2 referentes as *features* da Random Forest e na Tabela 6.3 referentes as *features* do XGBoost. Nas tabelas, lê-se que quanto maior o índice de importância, maior a relevância daquela feature para auxiliar o modelo a categorizar os usuários propensos ao churn.

A análise de importância das variáveis revelou que o XGBoost prioriza total_transactions (34,8%) como principal preditor, seguido por tenure_days (22,0%) e avg_days_between_transactions (13,6%). O Random Forest

apresentou distribuição mais equilibrada, com tenure_days (28,7%) como variável principal.

Feature	Índice de importância - Random Forest
tenure_days	28,70%
avg_days_between_transactions	11,19%
total_transactions	10,62%
total_transaction_value	8,55%
avg_transaction_duration	6,18%
avg_energy_consumed	6,08%
monthly_spend_avg	5,97%
avg_transaction_value	5,57%
transaction_frequency_monthly	4,74%
failed_transaction_attempts	3,07%

Tabela 6.2: Top 10 campos importantes para o Random Forest

Feature	Índice de importância - XGBoost
total_transactions	34,78%
tenure_days	21,95%
avg_days_between_transactions	13,56%
transaction_frequency_monthly	4,29%
coupon_usage_frequency	3,76%
weekend_transaction_ratio	3,27%
avg_transaction_value	2,66%
monthly_spend_avg	2,48%
total_transaction_value	2,46%
avg_energy_consumed	2,42%

Tabela 6.3: Top 10 campos importantes para o XGBoost

Ambos os algoritmos identificaram o tempo de relacionamento com o serviço (tenure_days) entre as três variáveis mais importantes, confirmando que usuários com maior histórico apresentam menor propensão ao churn. Esta descoberta alinha-se com a literatura sobre retenção de clientes, onde Reichheld e Schefter (2000) demonstram que o tempo de relacionamento correlaciona-se positivamente com a lealdade.

A presença de failed_transaction_attempts entre as *features* relevantes no Random Forest sinaliza que experiências negativas impactam significativamente a retenção. Assim, melhorias na confiabilidade da infraestrutura de recarga e na estabilidade do aplicativo constituem investimentos estratégicos para redução do churn.

6.5 Avaliação dos Objetivos

O sistema desenvolvido atendeu aos objetivos propostos na pesquisa, fornecendo percepções valiosas sobre predição de churn em aplicativos de recarga para veículos elétricos.

A comparação entre XGBoost e Random Forest demonstrou superioridade clara do XGBoost, com F1-Score de 0,9230 contra 0,9057 do Random Forest. Esta diferença de performance estabelece parâmetros de referência para implementações futuras no setor.

A avaliação das técnicas de balanceamento revelou que estas degradam a performance dos modelos, com reduções no F1-Score entre 1,1% e 2,3%. Este achado contraria expectativas convencionais e indica que algoritmos robustos conseguem aprender efetivamente em dados desbalanceados.

A identificação de *features* determinantes apontou total_transactions, tenure_days e avg_days_between_transactions como principais preditores de churn, fornecendo diretrizes práticas para estratégias de retenção baseadas em evidências. O *pipeline* reproduzível foi estabelecido através da modularização do sistema, permitindo a replicação dos experimentos e adaptação para outros aplicativos do domínio.

6.6 Contribuições da Solução

Esta pesquisa oferece três contribuições principais para o estado da arte em predição de churn aplicada ao setor de mobilidade elétrica. A aplicação específica preenche uma lacuna identificada na literatura sobre predição de churn em aplicativos de recarga para veículos elétricos. A comparação sistemática entre XGBoost e Random Forest estabelece parâmetros de referência para futuras implementações no setor.

A avaliação do impacto de técnicas de balanceamento em algoritmos robustos contraria expectativas estabelecidas pela literatura tradicional. Os resultados demonstram que algoritmos modernos conseguem extrair padrões discriminativos eficazes mesmo em cenários com 70% de desbalanceamento.

A identificação de que volume total de transações, tempo de relacionamento e intervalos entre carregamentos constituem os principais preditores fornecem percepções acionáveis para gestores de produto e equipes de marketing no setor de recarga elétrica.

6.7 Aprendizados do Projeto Atual

A implementação revelou desafios técnicos e limitações metodológicas que oferecem diretrizes valiosas para futuras implementações no domínio de aplicativos de recarga para veículos elétricos.

O desbalanceamento natural dos dados mostrou-se menos problemático que o esperado para algoritmos *ensemble robustos*. A distribuição 70/30 (*churn*/ativo) foi tratada efetivamente pelos modelos sem necessidade de técnicas de balanceamento, questionando práticas convencionais em *datasets* com desbalanceamento moderado. Esta descoberta indica que a complexidade adicional das técnicas de balanceamento pode ser desnecessária quando algoritmos modernos já incorporam mecanismos internos de compensação.

As limitações metodológicas incluem a especificidade dos dados para o contexto da Tupi Mobilidade e aplicativos de recarga no Brasil. Os resultados podem não ser diretamente generalizáveis para outros operadores ou mercados

com características distintas de infraestrutura e comportamento de usuários. Neste contexto, a validação cruzada com outros datasets do setor torna-se fundamental para confirmar a robustez dos achados.

A definição do *threshold* temporal de 20 dias para classificação de *churn*, embora fundamentada na literatura de aplicativos móveis, pode variar conforme características específicas do negócio. Implementações futuras devem considerar análise de sensibilidade para diferentes períodos de inatividade baseados no contexto operacional específico de cada aplicativo.

Somado às limitações temporais, a ausência de variáveis relacionadas à experiência do usuário — como tempo de resposta do aplicativo, falhas de conectividade e satisfação com a interface — representa uma lacuna que pode impactar a precisão preditiva. A incorporação dessas métricas qualitativas em trabalhos futuros pode elevar significativamente o desempenho dos modelos.

6.8 Sugestões para Trabalhos Futuros

Os resultados obtidos nesta pesquisa identificaram várias direções promissoras para investigações futuras que podem aprofundar o conhecimento sobre predição de churn em aplicativos de mobilidade elétrica.

6.8.1 Investigação Aprofundada do Balanceamento de Dados

A degradação observada no desempenho dos modelos quando aplicadas técnicas de balanceamento pode estar relacionada ao fato de que tanto *XGBoost* quanto *Random Forest* já incorporam mecanismos internos de compensação para dados desbalanceados. O *XGBoost* utiliza o parâmetro *scale_pos_weight* e o *Random Forest* emprega *class_weight='balanced'*, que ajustam automaticamente os pesos das classes durante o treinamento.

Trabalhos futuros devem investigar se a aplicação simultânea de balanceamento externo (SMOTE, ADASYN) e interno dos algoritmos cria redundância que compromete a qualidade do aprendizado. Neste contexto, recomenda-se comparar o desempenho dos modelos em três configurações distintas: sem balanceamento interno nem externo, apenas com balanceamento interno, e com ambas as abordagens combinadas.

6.8.2 Aplicação de *Recursive Feature Elimination* (RFE)

O uso de *Recursive Feature Elimination* representa uma abordagem promissora para compreender por que as técnicas de balanceamento não melhoraram o desempenho dos modelos. O RFE permite identificar sistematicamente quais *features* são mais impactadas pela introdução de dados sintéticos, revelando possíveis conflitos entre a informação original e a gerada artificialmente.

A implementação do RFE possibilitará avaliar se as *features* mais importantes para predição de churn — *total_transactions*, *tenure_days* e *avg_days_between_transactions* — mantêm sua relevância após a aplicação de técnicas de balanceamento. Esta análise pode explicar se a degradação observada resulta da diluição da importância dessas variáveis ou da introdução de ruído em dimensões específicas do espaço de características.

7 Considerações Finais

Este trabalho teve como objetivo comparar modelos de *Machine Learning* — XGBoost e Random Forest — para predição de churn em aplicativos de recarga para veículos elétricos. A pesquisa buscou identificar qual algoritmo oferece melhor relação entre precisão preditiva, além de avaliar o impacto de diferentes técnicas de balanceamento de dados.

A análise dos resultados demonstrou que o XGBoost obteve desempenho superior, alcançando F1-Score de 0,9230 e AUC-ROC de 0,9130, confirmando sua eficácia para problemas de classificação em *datasets* desbalanceados. Contrariamente às expectativas, as técnicas de balanceamento de dados não melhoraram a performance dos modelos, indicando que algoritmos robustos extraem informações relevantes mesmo em cenários com 70% de desbalanceamento.

A análise de importância das variáveis revelou que volume total de transações, tempo de relacionamento com o serviço e intervalos entre transações constituem os principais preditores de abandono. Estes achados fornecem percepções valiosas para estratégias de retenção no setor de mobilidade elétrica.

Para trabalhos futuros, recomenda-se explorar períodos alternativos de inatividade, validar os resultados em *datasets* de outros aplicativos de recarga e incorporar variáveis relacionadas à ansiedade de autonomia dos usuários de veículos elétricos.

Assim, este trabalho confirma a eficácia de técnicas de *Machine Learning* para predição de churn em aplicativos de recarga para veículos elétricos. O XGBoost demonstrou superioridade consistente, contribuindo para o avanço do conhecimento sobre aplicação de algoritmos de ML em dados naturalmente desbalanceados do setor de mobilidade elétrica brasileira.

8 Referências Bibliográficas

ABVE - ASSOCIAÇÃO BRASILEIRA DO VEÍCULO ELÉTRICO. **Infraestrutura de recarga acelera no país e apresenta crescimento de 179%**. 2024. Disponível em: https://abve.org.br/infraestrutura-de-recarga-acelera-no-pais-e-apresenta-crescimento-de-179/. Acesso em: 03 maio 2025.

AL-SHATNAWI, A. M.; FARIS, M. **Predicting Customer Retention using XGBoost and Balancing Methods**. International Journal of Advanced Computer Science and Applications, v. 11, n. 7, p. 704-712, 2020.

AMAZÔNIA ELETROVIAS. **A importância do PlugShare para a comunidade de motoristas de veículos elétricos. 2024**. Disponível em: https://amazoniaeletrovias.com/a-importancia-do-plugshare-para-a-comunidade-de-motoristas-de-veiculos-eletricos/. Acesso em: 17 abr. 2025.

AMIN, A. et al. **Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study**. IEEE Access, v. 7, p. 134964-134980, 2019.

ANFAVEA. **Anuário da Indústria Automobilística Brasileira 2023**. Associação Nacional dos Fabricantes de Veículos Automotores. Disponível em: https://anfavea.com.br/anuario-2023. Acesso em: 12 abr. 2025.

BIEKER, Georg. **A global comparison of the life-cycle greenhouse gas emissions of combustion engine and electric passenger cars. Berlin: International Council on Clean Transportation (ICCT)**, 2021. Disponível em: https://theicct.org/publication/a-global-comparison-of-the-life-cycle-greenhouse-gas-emissions-of-combustion-engine-and-electric-passenger-cars/. Acesso em: 07 maio 2025.

BRADLEY, A. P. **The use of the area under the ROC curve in the evaluation of machine learning algorithms**. Pattern Recognition, v. 30, n. 7, p. 1145-1159, 1997.

CHAWLA, N. V. et al. **SMOTE: Synthetic Minority Over-sampling Technique**. Journal of Artificial Intelligence Research, v. 16, p. 321-357, 2002.

CNN BRASIL. **Venda de carros elétricos no Brasil bate recorde em 2023, diz associação**. 4 jan. 2024. Disponível em: https://www.cnnbrasil.com.br/economia/

macroeconomia/venda-de-carros-eletricos-no-brasil-bate-recorde-em-2023-diz-associacao/. Acesso em: 06 maio 2025.

ENERGYSAGE. **The History of Electric Vehicles: From Then to Now**. Boston, 2023. Disponível em: https://www.energysage.com/electric-vehicles/the-history-of-electric-vehicles-from-then-to-now/. Acesso em: 22 abr. 2025.

EVBOX. **History of the electric car**. Amsterdam, 2023. Disponível em: https://blog.evbox.com/electric-cars-history. Acesso em: 18 abr. 2025.

EXAME. **A onda dos híbridos e elétricos: R$ 117 bilhões de investimentos foram anunciados**. 2023. Disponível em: https://exame.com/negocios/a-onda-dos-hibridos-e-eletricos-r-117-bilhoes-de-investimentos-foram-anunciados/. Acesso em: 23 abr. 2025.

EXAME. **Eletrificados: rebranding da 'Tupinambá', que virou 'Tupi', foi ideia do marketing ou dos clientes?** 28 maio 2024. Disponível em: https://exame.com/bussola/eletrificados-rebranding-da-tupinamba-que-virou-tupi-foi-ideia-do-marketing-ou-dos-clientes/. Acesso em: 21 maio 2025.

FEI - CENTRO UNIVERSITÁRIO. **EletroPOSTO FEI**. São Bernardo do Campo, 2023. Disponível em: https://fei.edu.br/eletroposto/. Acesso em: 10 maio 2025.

FRANKE, T.; KREMS, J. F. **Understanding charging behaviour of electric vehicle users. Transportation Research Part F: Traffic Psychology and Behaviour**, v. 21, p. 75-89, 2013.

GALLO, A. **The Value of Keeping the Right Customers. Harvard Business Review**, 2014. Disponível em: https://hbr.org/2014/10/the-value-of-keeping-the-right-customers. Acesso em: 14 maio 2025.

GARCIA, A. L. et al. **A cloud-based framework for machine learning workloads and applications**. IEEE Access, v. 8, p. 18681–18692, 2020.

HANDELMA et al. **Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods**. 2018. Disponível em: https://sci-hub.se/10.2214/ajr.18.20224. Acesso em: 28 abr. 2025.

IEA – INTERNATIONAL ENERGY AGENCY. **Global EV Outlook 2023**. Paris: IEA, 2023. Disponível em: https://www.iea.org/reports/global-ev-outlook-2023. Acesso em: 16 maio 2025.

IEA – INTERNATIONAL ENERGY AGENCY. **Global EV Outlook 2024**. Paris: IEA, 2024. Disponível em: https://www.iea.org/reports/global-ev-outlook-2024. Acesso em: 05 maio 2025.

IEA – INTERNATIONAL ENERGY AGENCY. **Global EV Outlook 2025**. Paris: IEA, 2025. Disponível em: https://www.iea.org/reports/global-ev-outlook-2025. Acesso em: 09 abr. 2025.

IMARC GROUP. **Electric Vehicle Charging Station Market Size, Forecast 2033**. 2024. Disponível em: https://www.imarcgroup.com/electric-vehicle-charging-station-market. Acesso em: 25 maio 2025.

JUNIOR, E. E. R. **Estratégias para Classificação Binária Um estudo de caso com classificação de e-mails**. 2016. Disponível em: https://jreduardo.github.io/ce064-ml/work-master.pdf. Acesso em: 11 maio 2025.

KERAMATI, A. et al. **Improved churn prediction in telecommunication industry using data mining techniques**. Applied Soft Computing, v. 42, p. 24-38, 2016.

LOCALYTICS. **Mobile Apps: What's A Good Retention Rate?** 2019. Disponível em: http://info.localytics.com/blog/mobile-apps-whats-a-good-retention-rate. Acesso em: 20 abr. 2025.

MITCHELL, T. M. **Machine Learning**. New York: McGraw-Hill, 1997.

NANNICINI, G.; LIBERTI, L. **A Mobile Application to Assist Electric Vehicles' Drivers with Charging Services**. IEEE Transactions on Intelligent Transportation Systems, v. 15, n. 4, p. 1-12, 2014. DOI: 10.1109/TITS.2014.2345678.

NEOCHARGE. **Evolução dos VEs no Brasil em 2024**. 2024. Disponível em: https://www.neocharge.com.br. Acesso em: 04 maio 2025.

OPEN CHARGE ALLIANCE. **OCPP (Open Charge Point Protocol)**. 2023. Disponível em: https://openchargealliance.org/protocols/open-charge-point-protocol/. Acesso em: 02 maio 2025.

PORTAL SUSTENTABILIDADE. **Infraestrutura de recarga no Brasil**. 2024. Disponível em: https://www.portalsustentabilidade.com.br. Acesso em: 18 abr. 2025.

PRNEWSWIRE. **Tupinambá anuncia rebranding da marca e lança seu novo modelo de negócio**. 23 mai. 2024. Disponível em: https://www.prnewswire.com/br/comunicados-para-a-imprensa/tupinamba-anuncia-rebranding-da-marca-e-lanca-seu-novo-modelo-de-negocio-302154720.html. Acesso em: 27 maio 2025.

RAMOS, B. A. **Machine Learning para Previsão de Churn**. 2024. 48 f. Projeto Final – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2024.

REICHHELD, F. F.; SCHEFTER, P. **E-loyalty: your secret weapon on the web**. Harvard Business Review, v. 78, n. 4, p. 105-113, 2000.

SABBEH, S. F. **Machine-learning techniques for customer retention: a comparative study**. International Journal of Advanced Computer Science and Applications, v. 9, n. 2, p. 273-281, 2018.

TUPI MOBILIDADE. **Tupimob - tecnologia para simplificar recarga de carros elétricos**. 2024. Disponível em: https://tupimob.com/. Acesso em: 15 abr. 2025.

U.S. DEPARTMENT OF ENERGY. **The History of the Electric Car**. Washington, D.C., 2022. Disponível em: https://www.energy.gov/articles/history-electric-car. Acesso em: 08 maio 2025.

VINFAST GLOBAL COMMUNITY. **A Brief History of Electric Vehicles**. 2023. Disponível em: https://community.vinfastauto.us/forums/discussion/a-brief-history-of-electric-vehicles/. Acesso em: 13 abr. 2025.

VOLTBRAS. **Veículos elétricos em frotas no Brasil: tendências e desafios**. 13 jun. 2024. Disponível em: https://voltbras.com/frotas/veiculos-eletricos-em-frotas-no-brasil-tendencias-e-desafios/. Acesso em: 26 maio 2025.

VOOLTA. **Pontos de recarga no Brasil: um guia completo**. 2024. Disponível em: https://voolta.com.br/blog/pontos-de-recarga-no-brasil/. Acesso em: 19 maio 2025.