UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
CURSO DE ENGENHARIA DE PRODUÇÃO MECÂNICA

João Vitor Goedert

Prevenção do cancelamento de licença de uso em uma empresa de *Software as a Service* a partir de modelos analíticos

Florianópolis
2024

João Vitor Goedert

Prevenção do cancelamento de licença de uso em uma empresa de *Software as a Service* a partir de modelos analíticos

Trabalho de Conclusão de Curso submetida ao curso de Engenharia de Produção Mecânica da Universidade Federal de Santa Catarina para a obtenção do título de Engenheiro Mecânico com habilitação em Engenharia de Produção.
Orientador: Prof. Mauricio Uriona Maldonado, Dr.

Florianópolis
2024

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Goedert, João Vitor
Prevenção do cancelamento de licença de uso em uma empresa de Software as a Service a partir de modelos baseados em dados. / João Vitor Goedert ; orientador, Mauricio Uriona Maldonado, 2024.
92 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia de Produção Mecânica, Florianópolis, 2024.

Inclui referências.

1. Engenharia de Produção Mecânica. 2. Software. 3. SaaS. 4. Cancelamento de Licenças de Uso. 5. Modelos Preditivos. I. Maldonado, Mauricio Uriona . II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Produção Mecânica. III. Título.

João Vitor Goedert

Prevenção do cancelamento de licença de uso em uma empresa de *Software as a Service* a partir de modelos analíticos

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de Engenheiro Mecânico com habilitação em Engenharia de Produção e aprovado em sua forma final pelo curso de Engenharia de Produção Mecânica.

Documento assinado digitalmente
Monica Maria Mendes Luna
Data: 21/02/2024 09:55:01-0300
CPF: ***.104.993-**
Verifique as assinaturas em https://v.ufsc.br

Profa. Mônica Maria Mendes Luna, Dra.
Coordenação do Curso

Banca Examinadora:

Documento assinado digitalmente
Mauricio Uriona Maldonado
Data: 21/02/2024 08:17:01-0300
CPF: ***.972.939-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Mauricio Uriona Maldonado, Dr.
Orientador

Prof. Guilherme Ernani Vieira, Dr.
Universidade Federal de Santa Catarina

Prof. Ricardo Villarroel Dávalos, Dr.
Universidade Federal de Santa Catarina

Florianópolis, 2024.

Este trabalho é dedicado aos meus pais e às minhas irmãs, que possibilitaram o meu sucesso nessa jornada.

AGRADECIMENTOS

Gostaria de expressar minha mais profunda gratidão à minha família. Aos meus pais, Guido e Luciane, que ao longo de toda a minha vida fizeram questão de prover todo o recurso necessário para a minha formação profissional, e às minhas irmãs, Gabriela, Thais e Ana, mulheres incríveis responsáveis por me tornar um homem melhor. Se nós somos a média das cinco pessoas a nossa volta, me considero sortudo de ter convivido com vocês nos últimos vinte e cinco anos, obrigado por todo o apoio, e por serem o porto seguro que tanto precisei em meus momentos de dificuldade.

Agradeço a Universidade Federal de Santa Catarina pela oportunidade de uma educação gratuita e de qualidade, que me proporcionou desde oportunidades maravilhosas de desenvolvimento pessoal e profissional, até amizades que vou levar para o resto da minha vida. Gostaria de agradecer o meu orientador, Mauricio, pela oportunidade de finalizar este trabalho, por todo o apoio no processo de desenvolvimento do mesmo e por confiar que eu seria capaz, mesmo que o desafio aparentasse ser impossível. Também gostaria de agradecer os professores Antonio Cezar Bornia e Mirna de Borba, pelas orientações e conselhos que guardarei no fundo do coração.

Agradeço ao grupo do "SC", por todos os momentos incríveis vividos, pelos conselhos valiosos, pelas aventuras, pelas risadas, por aquele abraço que fez toda a diferença. Também gostaria de agradecer ao "Conselho", por todo o carinho, direcionamentos profissionais e noites de fofocas e jogos dos últimos anos. Um agradecimento também ao "Rebanho do cetruco"e "Fut Acojar Seg 21h", por todo o entretenimento.

Sou grato a todas as amizades que fiz dentro da universidade. Gostaria de agradecer a todas aquelas construídas com os colegas de curso, durante minha passagem pelo Centro Acadêmico Livre de Engenharia de Produção e ao longo dos meus três anos como participante do PET Engenharia de Produção. Também agradeço os meus colegas de outros cursos, com os quais compartilhei vivências mais que especiais. Gostaria de deixar um agradecimento especial ao Mateus Lamin pela inspiração, à Ana Luiza Garcia pela parceria, ao Mateus Gato Cunha pelos momentos e ao Jhonatan Fiuza pelo carinho.

Queria deixar o agradecimento à empresa em que trabalho, que forneceu todo o suporte para o desenvolvimento deste projeto, em especial à Natalia Tomazi, por estar ali para mim em meus momentos de dificuldade. Por fim, gostaria de deixar meu agradecimento ao Artur Ferrari e ao Vinicius Haddad pela amizade e todo o suporte fornecido nos últimos tempos e aos queridos Felipe Daminelli e Johnatan Jankoski, eu não seria quem sou hoje se não fosse por vocês.

RESUMO

Altos índices de cancelamento de empresas de Software as a Service (SaaS) implicam em redução da receita total, dada a diminuição da receita recorrente mensal. Dessa forma, o presente trabalho tem como objetivo propor ações que ajudem a prevenção do cancelamento de licenças de uso de software em uma empresa de tecnologia com o auxílio de modelos baseados em dados. Para tal, foi realizado um estudo na literatura dos fatores relacionados ao cancelamento de licenças de uso de software e de estratégias para prevenção deste fenômeno. Foram utilizados dados de três diferentes bases, que foram trabalhadas para a seleção das melhores variáveis de entrada. Foram aplicados quatro diferentes modelos: Regressão Logística, Randon Forest, Support Vector Machine e XGBoost que foram comparados quanto à acurácia, precisão, sensibilidade, F-score e área sob a curva ROC para definição de qual método obtém um melhor resultado. Como resultado, baseando-se na revisão de literatura existente, e na resposta do modelo XGBoost apresentou melhores resultados com uma acurácia de 83,08% e uma área sob a curva ROC de 81,45%, foi possível propor seis diferentes ações: i. Implantação de uma equipe de sucesso do cliente personalizada para clientes propensos ao cancelamento, ii. Criação de um programa de valorização de clientes fiéis à marca, iii. Plano de melhoria de usabilidade dos clientes menos satisfeitos, iv. Plano de aumento de base de dados, v. Plano de correção da base de dados na fonte e vi. Melhoria do modelo com criação de novas variáveis.

Palavras-chave: Software, SaaS; Cancelamento de Licenças de Uso; Regressão Logística, Suport Vector Machine, Random Forest, XGBoost.

ABSTRACT

High cancellation rates for Software as a Service (SaaS) companies imply a reduction in total revenue, due to a decrease in monthly recurring revenue. Therefore, this work seeks to propose actions that help prevent the cancellation of software use licenses in a technology company with the help of data-based models. A literature review was carried out in order to find factors related to the cancellation of software licenses and strategies to prevent this phenomenon. Data from three different databases were used, in which were selected the best input variables. Four different models were applied: Logistic Regression, Randon Forest, Support Vector Machine and XGBoost, which were compared regarding accuracy, precision, sensitivity, F-score and area under the ROC curve to define which method obtains a better result. As a result, based on the existing literature review, and the response of the XGBoost model, it presented better results with an accuracy of 83.08% and an area under the ROC curve of 81.45%, it was possible to propose six different actions :i. Deploying a customized customer success team for churn-prone customers, ii. Creation of a program to value customers loyal to the brand, iii. Usability improvement plan for less satisfied customers, iv. Database Augmentation Plan, v. Database correction plan at source and vi. Improvement of the model with the creation of new variables.

Keywords: Software, SaaS; Client Churn, Logistic Regression, Suport Vector Machine, Random Forest, XGBoost.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANN	*Artificial Neural Networks* (Redes Neurais Artificiais)
AUC	*Area Under Curve* (Área sob a curva)
B2B	*Business To Business* (De Empresa Para Empresa)
B2C	*Business to Customer* (De Empresa Para Cliente)
CAC	Custo de Aquisição de Clientess
CS	*Customerr Success* (Sucesso do Cliente)
DT	*Decision Trees* (Árvores de Decisão
EULA	Acordos de licença com o usuário final
FN	Falso Negativo
FP	Falso Positivo
LGPD	Lei Geral de Proteção de Dados
LR	*Logistic Regression* (Regressão Logística)
MRR	Receita Reccorente Mensal
RFA	*Random Forest Algoritm*
ROC	Receiver Operating Characteristic
SaaS	*Software as a Service* (Software como Serviço)
SVM	*Support Vector Machine*
UFSC	Universidade Federal de Santa Catarina
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
XGB	*XGBoost*

SUMÁRIO

1 INTRODUÇÃO

Neste capítulo será feita uma introdução ao tema, passando pela descrição do problema, partindo para o levantamento dos objetivos geral e específicos e finalizando com uma justificativa que também apresentará a estrutura do trabalho.

1.1 CONTEXTUALIZAÇÃO

A receita mundial associado à indústria de *software* está crescendo exponencialmente, já se tornando um mercado multibilionário (GARTNER, 2021). As evoluções recentes do século da tecnologia trouxeram novas estratégias para garantir sucesso empresarial. Dentre essas estratégias, a migração da computação para a nuvem apresentou o início de uma nova era, a dos softwares de assinatura.

Software como serviço (SaaS) refere-se a *software* remotamente hospedado, desenvolvido, gerenciado e entregue via internet por um provedor de serviços (CHO; CHAN, 2015). Com a migração da computação para a nuvem dada a expansão da infraestrutura que facilita a entrega de *software* de forma virtual, além da popularização do uso de aplicativos, o SaaS ganhou força ao comercializar contratos de aluguel no formato de licença de uso. Dessa forma, o cliente pode utilizar de todos os recursos de *softwares* de qualidade pelo tempo que achar necessário. Nesse modelo de negócio, o lucro se dá pelo balanço entre a receita recorrente paga pelo cliente e o custo de manutenção da licença. Com a presença de um novo modelo de receita na negociação de *softwares* surgiram conceitos que vão além da do custo de venda, como Custo de Aquisição de Clientes (CAC) e retenção (VAIDYANATHAN; RABAGO, 2020).

O CAC se trata dos custos relacionados às campanhas e publicidades realizadas pelo setor de marketing e ao processo de vendas realizado pelo setor comercial. Dessa forma, a receita proveniente dos primeiros meses do cliente são utilizados para pagamento de despesas já desembolsadas. Nesse cenário, as principais opções para que a empresa alcance um desempenho financeiro favorável quando relacionado a custo incluem a redução dos gastos associados à aquisição de clientes ou cultivar a fidelidade, garantindo a satisfação e a retenção com relações de longo prazo.

A retenção se trata da manutenção da relação com o cliente, evitando o seu cancelamento e garantindo a receita recorrente mensal. Em diversos ramos empresariais que abordam a assinatura, o CAC pode ser até cinco vezes superior custo de retenção de um cliente existente(KURTZ; CLOW, 1997). Segundo Lalwani *et al.* (2022), a retenção de clientes é a mais barata das estratégias na redução de custos em empresas SaaS dado que a retenção representa aproximadamente 20% do custo de aquisição de um novo cliente. Além disso, no ponto de vista do cliente, trocar de fornecedor de produtos e serviços também é mais custoso se comparado a manter um fornecedor existente, dessa forma, a retenção é de interesse de todas as partes

envolvidas (CORREA BAHNSEN; AOUADA; OTTERSTEN, 2015).

Churn é o inverso da retenção, e nada mais é que o cliente que encerra seu relacionamento com uma empresa que fornece produtos ou serviços. No contexto de empresas SaaS, o *churn* significa o cancelamento de licenças de uso de *software*. A palavra "churn"deriva do conceito de churn rate, que é uma das métrica para medir a lealdade dos clientes de empresas SaaS e trata-se do levantamento do número de cancelamentos realizados em um determinado período quando comparado com toda a base de clientes ativos (GOLD, 2020).

Ou seja, o churn é uma expressão que determina a saída de clientes de um produto ou serviço. Dessa forma, reduzí-lo é, essencialmente, equivalente a aumentar a retenção de clientes Segundo Cister (2005), existem três diferentes categorias de *churn*. O *churn* involuntário se dá quando o desligamento parte do fornecedor de *software*, e costuma ser reflexo de inadimplência contratual. O *churn* voluntário deliberado que se dá quando o cliente decide trocar de fornecedor ou considera que a solução não é mais necessária para a realização de suas operações. Por fim, o *churn* voluntário acidental, ou inevitável, se dá quando o cliente encerra suas operações nas unidades que utilizavam da solução.

Ainda, o churn não é uma explessão que pode ser simplificada como apenas o cancelamento. Existem diversas interpretações para as diferentes formas de abordar o churn. O churn rate como volume de clientes perdidos é a apresentada anteriormente, onde é contada a quantidade de clientes perdidos dentro de toda a base. Já o churn de receita, ou MRR churn, calcula o quanto da receita foi perdida com os cancelamentos efetuados. A associação destas duas óticas possibilita análises mais aprofundadas do comportamento dos clientes, uma vez que uma empresa pode ter um grande volume de churn que representa uma baixa proporção de receita mas outra pode possuir uma grande perda de faturamento com um baixo índice de cancelamentos.

Por fim, Lalwani *et al.* (2022) afirma que a gestão do cancelamento pode ser feita de duas formas: (1) Reativa e (2) Proativa. Na abordagem reativa, a empresa aguarda o pedido de cancelamento recebido do cliente, em seguida, oferece para o mesmo planos atrativos para fomentar a retenção. Na abordagem proativa, prevê-se a possibilidade de cancelamento, conforme os planos são oferecidos aos clientes.

1.2 DESCRIÇÃO DO PROBLEMA

Apesar do sucesso do modelo SaaS o estudo *churn* ainda é uma pauta de grande importância. O relatório desenvolvido pela KeyBanc (2018) revelou uma taxa anual de *churn* em torno de 13,2% para empresas de *software* por assinatura. Ainda, um desafio no modelo de negócio SaaS, que amplia suas elevadas taxas de cancelamento de contrato, é sua natureza de receita recorrente, baseada em assinaturas. Se o provedor de SaaS não atende às necessidades e expectativas do cliente, o mesmo

possui total liberdade de cancelar rapidamente o serviço, resultando na perda total do investimento feito pelo provedor para conquistar aquela empresa como cliente (NORONHA, 2020). Dessa forma, como apresenta Ge *et al.* (2017) dado que apenas após de doze meses em média o investimento inicial para adquirir o cliente é reconquistado, a definição de estratégias bem estruturadas de retenção é essencial.

Para realizar uma estratégia de prevenção do cancelamento do cliente, é necessário antes levantar de quais as causas dessa atitude. Idris e Khan (2012) afirmam que muitas são as razões que resultam no cancelamento de um cliente, indo desde problemas relacionados à baixa qualidade, questões financeiras, altos custos, dentre outros que afetam diretamente o valor percebido pelo consumidor final. Ainda, essas causas são variáveis, haja vista que cada cliente possui necessidades e expectativas que variam conforme o contexto social e cultural. Dessa forma, mesmo que seja possível encontrar necessidades generalizadas para qualquer contexto, vê-se essencial o estudo caso a caso para ter maior assertividade para suprir as demandas do consumidor.

Ainda, expectativa é um termo subjetivo, que trata de um sentimento criado a partir de experiências anteriores com concorrência, discursos de venda e propagandas. O não atingimento das expectativas gera insatisfação, que pode ser decisiva para o cancelamento de clientes. Satisfação se trata de uma avaliação realizada pelo cliente quanto ao cumprimento das necessidades e expectativas criadas anteriormente a aquisição de um produto ou serviço (MINIARD; ENGEL; BLACKWELL, 2000).

Para manutenção de necessidades e expectativas e obter lealdade do consumidor, visto que estão cada vez mais exigentes e conscientes do seu poder de compra, as empresas de software como serviço contam com equipes de sucesso do cliente, especializadas em realizar ações para aumentar o engajamento, garantir a satisfação e previnir o abandono do cliente (CRISTINE, 2019). Estas equipes possuem grande responsabilidade por serem os principais pontos de contato entre a empresa e seus clientes, sendo encarregadas de resolver quaisquer problemas que possam surgir durante o ciclo de vida do cliente. Isso evidenciando a necessidade de manutenção e melhoria contínua dos processos desse setor a fim de proporcionar uma experiência mais eficiente e personalizada aos clientes.

Se tratando de eficiência de recursos, Cristine (2019) afirma que o conhecimento de quais são os clientes propensos ao cancelamento reduz os desperdícios com os departamentos de Sucesso do Cliente (CS). Para identificá-los é possível aplicar algoritmos de aprendizado de máquina, que são modelos baseados em informações financeiras, demográficas, geoespaciais, comportamentais, dentre outras. Ainda, estas técnicas permitem entender como esses fatores influenciam no cancelamento, o que pode ocasionar em tomadas de decisão mais assertivas no contexto de prevenção do cancelamento de licenças de uso. Dessa forma, levando em consideração a im-

portância da previsão do *churn* como uma estratégia para desenvolvimento do setor de tecnologia, mais especificamente dos comercializadores de *software* como serviço, o presente estudo levanta o seguinte questionamento: É possível definir ações que ajudem a prevenção de cancelamento de assinaturas em uma empresa SaaS com a utilização de modelos analíticos com a aplicação de *machine learning*?

1.3 OBJETIVOS

Nesta seção são descritos o objetivo geral e os objetivos específicos deste Trabalho de Conclusão de Curso.

1.3.1 Objetivo Geral

Propor ações que ajudem a prevenção do cancelamento de licenças de uso *software* em uma empresa de tecnologia com o auxílio de modelos baseados em dados.

1.3.2 Objetivos Específicos

- Levantar na literatura os fatores que levam ao cancelamento de licenças de uso de *software*.
- Levantar dados na empresa relacionadas ao cancelamento de licenças.
- Treinar modelos preditivos para prever o cancelamento de licenças.
- Propor ações para a prevenção de cancelamento de licenças.

1.4 JUSTIFICATIVA

Altos índices de cancelamento de empresas de software implicam em redução da receita total, dada a diminuição da Receita Recorrente Mensal (MRR). Segundo Amin *et al.* (2016), reduzir o cancelamento é extremamente importante em mercados altamente competitivos dado que a conquista de novos clientes nestes mercados é muito difícil, e atrair não assinantes pode custar até seis vezes mais do que manter um cliente atual. Apesar de ser um assunto que recentemente ganhou grande relevância, o aprendizado de máquina já foi muito aplicado a fim de prever os clientes propensos ao *churn*, com um grande aumento de pesquisas e publicações na última década (LUIZON, 2019).

Ahmad, Jafar e Aljoumaa (2019) concluiram que o aprendizado de máquina provou ser uma técnica altamente eficiente para prever o cancelamento dos clientes com base em dados previamente capturados. No contexto da Engenharia de Produção, algumas técnicas já são conhecidas como a clusterização, a regressão linear,

a regressão logística, dentre outros algoritmos que possibilitam converter dados em conhecimento a fim de realizar a tomada de decisão, aumentar a eficiência operacional, e consequentemente aumentar a lucratividade da empresa.

Um setor onde esta tecnologia já é muito presente desde cedo é no de telecomincações. Idris e Khan (2012) desenvolveu um modelos de classificação com a aplicação de *Random Forest Algoritm (RFA), Rotation Forest* e *RotBoost* para predição do cancelamento. Posteriormente Ahmad, Jafar e Aljoumaa (2019) aplicaram técnicas de *machine learning* em plataformas de *big data* afim de prever quais os clientes mais propensos ao cancelamento.

Entretanto, qualquer empresa que forneça serviços de assinatura possui a necessidade de controle da taxa de *churn*. Suh (2023) realizou um trabalho que analisa as informações sobre o comportamento do cliente de uma locadora de purificadores de água para residências, utilizando de modelos de *Random Forest Algoritm* e Gradient Boost para prever rotatividade.

Ainda, fora do contexto de assinaturas, Luizon (2019) formulou um conceito de *churn* para um *e-commerce*, se tratando possíveis clientes que não concretizaram a venda, e buscou prever quais seriam estes clientes com a aplicação de Árvores de Decisão, Redes Neurais Artificiais (ANN) e *Support Vector Machine* (SVM). Kumar e Leema (2023) abordaram a perda de clientes de bancos como uma forma de *churn*, e desenvolveu modelos predidivos baseados em ANN, SVM e RFA a fim de maximizar os lucros.

O presente trabalho busca fornecer insumos para a tomada de decisão empresarial a partir da análise de dados e aplicação de aprendizado de máquina em uma empresa SaaS, haja vista que existe falta de conhecimento e estudos aplicados em empresas desse tipo. Além disso, apesar de metodologias serem semelhantes em diferentes aplicações, empresas de *software* possuem dados diferentes que relacionam acesso à ferramenta, e a assinatura destes softwares implica numa necessidade de utilização de variáveis temporais além das categóricas, tornando o modelo único e dependente da estrutura da ferramenta em questão.

Dessa forma, apesar de os resultados não serem aplicáveis para diferentes empresas de tecnologia, o estudo se torna pertinente ao apresentar diferentes formas de estruturação dos dados para um segmento específico que apresenta resultados positivos e que contribuem para o aumento do faturamento.

1.5 ESTRUTURA DO TRABALHO

O presente trabalho é composto por cinco capítulos, sendo o primeiro a introdução aqui apresentada, com uma contextualização do tema seguida da definição do problema de pesquisa. Em seguida é definido o objetivo geral e os objetivos específicos, finalizando com a apresentação da justificativa do trabalho em relação ao objetivo

geral, enfatizando a necessidade da realização de novos estudos a respeito do tema.

O segundo capítulo é a fundamentação teórica, e possui como principal finalidade o entendimento do problema de pesquisa a fim de fundamentar a execução do trabalho. Neste, é apresentado o conceito de cancelamento de serviços e de licenças de uso de software, com o levantamento de fatores influenciadores nesse fenômeno e de estratégias para previnir sua ocorrência.

O terceiro capítulo apresenta os procedimentos metodológicos, com a apresentação dos materiais disponíveis para o estudo e os métodos de aprendizado de máquina, avaliação e otimização de aplicados. O capítulo finaliza com a apresentação do passo a passo da realização da pesquisa com cada etapa metodológica explicada.

O quarto capítulo aborda a aplicação da metodologia com a realização da análise exploratória, seleção e preparação dos dados, a apresentação dos resultados dos modelos de previsão e a proposição de ações para prevenção do cancelamento de licenças de uso de *software*. Por fim, o quinto capítulo apresenta as considerações finais e discorre quanto ao atingimento dos objetivos do trabalho, apresenta as limitações e propõe sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem como principal finalidade o entendimento do problema de pesquisa a fim de fundamentar a execução do trabalho. O capítulo inicia apresentando os conceito de serviços e cancelamento, mapeando os fatores que influenciam neste contexto. Posteriormente, são introduzidos conceitos de *software*, de licenças de uso de software e de rotatividade de clientes, de forma a embasar os tópicos seguintes. Dessa forma, foi possível levantar os fatores influenciadores no cancelamento de licenças e estratégias pertinentes para evitar o cancelamento no contexto de empresas de *software.*

2.1 CANCELAMENTO DE SERVIÇOS

Existe uma clara distinção entre organizações que oferecem serviços e aquelas que comercializam produtos físicos. Conforme delineado por Wirtz e Lovelock (2022), bens são objetos ou dispositivos físicos, enquanto serviços são ações prestadas por uma parte a outra, predominantemente intangível e, geralmente, não resulta na aquisição de propriedade nos fatores de produção por parte de nenhum dos envolvidos. Dessa forma, conforme descrito por Philip T Kotler *et al.* (2019), serviços englobam todas as tarefas, ações, conhecimentos e experiências que uma pessoa pode oferecer a outra. Dada a distinção marcante em relação aos bens tangíveis, é natural que os serviços demandem uma análise específica e diferenciada. Os serviços podem ser comercializados de duas diferentes maneiras: serviços fornecidos de forma pontual ou serviços de receita recorrente, e o presente estudo tratará como serviço apenas aquele que garante um contrato com faturamento recorrente

Nas empresas que prestam serviços, cada cliente representa um novo projeto, e apenas após atingir o ponto de equilíbrio (*payback)* desse projeto é que os investimentos feitos para adiquirir esse cliente são recuperados, começando então a gerar lucro para a organização. Além dos custos associados à aquisição de clientes, a organização também precisa levar em consideração os custos de manutenção de sua estrutura interna. Dessa forma fica evidente que a gestão do cancelamento desses serviços é importante não apenas para a geração de lucro, mas também para a própria sobrevivência da organização (MENEZES, 2014).

O cancelamento de serviços pode ser classificado, dadas suas características, em voluntário e involuntário (CISTER, 2005), podendo ele ser temporário, quando existe a intenção do cliente de retomar o serviço posteriormente, ou definitivo, quando não há a intenção de retomada. Segundo Hadden *et al.* (2007), o cancelamento involuntário quando o serviço deixa de ser prestado por uma escolha da empresa prestadora de serviços. É comum que isto ocorra em situações de inadimplência, onde o não pagamento da mensalidade por parte de um cliente resulta na interrompção da oferta

do serviço em questão.

O cancelamento voluntário ocorre quando o cliente toma a decisão de encerrar o relacionamento com a empresa, e pode ser dividido em duas subcategorias, o cancelamento voluntário acidental ou inevitável e o o cancelamento voluntário deliberado. É chamado de cancelamento voluntário acidental, quando o cliente é impedido por algum motivo de continuar com a empresa seja, ou seja, não existia intenção de cancelamento por parte do cliente porém mesmo assim o mesmo é forçado a encerrar o contrato, e comumente está relacionado à problemas financeiros. Já o cancelamento voluntário deliberado, ocorre a partir de uma tomada de decisão do cliente, e leva em consideração fatores econômicos, de qualidade, sociais, psicológicas, dentre outros. Para este caso, o cliente possui a intenção de realizar o cancelamento do serviço, seja por não ver mais benefícios suficientes nesse custo recorrente, ou por estar em busca de outro fornecedor que melhor aborde os fatores de tomada de decisão (MATTISON, 2006).

Para mapear os fatores relevantes ao cancelamento de serviços, é necessário um entendimento acerca do comportamento do consumidor. Esse conceito refere-se à análise dos processos que ocorrem quando indivíduos ou grupos fazem escolhas que os levam a, adquir, utilizar ou descartar produtos, serviços, ideias ou experiências para atender a necessidades ou desejos (SOLOMON, 2016). Franzoso (2007) levanta que o comportamento do consumidor é influenciado por uma variedade de fatores que podem ser tanto individuais e psicológicos, quanto relacionados ao ambiente externo, composto por elementos econômicos, tecnológicos, políticos e culturais. Com base nesses fatores, o consumidor embarca em um processo decisório que orienta a seleção do produto ou serviço.

No que tange necessidades e desejos dos clientes, Sheth, Mittal e Newman (1999) definem necessidade como uma condição insatisfatória que leva o cliente a tomar uma atitude voltada a melhorar a sua situação, ou seja um elemento crucial no processo motivacional, impulsionando o cliente à ação. Já o desejo representa o anseio por mais satisfação o que é necessário para superar uma insatisfação inicial. Dessa forma, uma mudança de necessidades sem o estímulo de um desejo pode levar o cliente ao cancelamento de uma assinatura de um serviço.

Outro fator que pode levar um cliente a um cancelamento é a pressão da concorrência. O cancelamento por concorrência ocorre quando a decisão dos clientes de encerrar um serviço se dá devido à oferta de alternativas mais atrativas e competitivas no mercado (FRANZOSO, 2007). Nesse contexto, a concorrência apresenta propostas que parecem mais vantajosas, levando os usuários a optarem por cancelar o serviço atual em favor de uma oferta concorrente. Esse tipo de cancelamento destaca a importância da empresa em manter a competitividade, adaptar-se às necessidades dos usuários e oferecer um custo-benefício que supere as ofertas concorrentes, minimi-

zando assim a probabilidade de cancelamentos.

Se tratando de concorrência o valor de mercado é um elemento crucial na tomada de decisão do cliente, onde o principal objetivo é maximizar o valor, respeitando as limitações de recursos envolvidos. O valor de mercado refere-se à capacidade de um produto ou serviço em atender às necessidades dos clientes. Dado que os clientes variam em suas necessidades e desejos, um produto ou serviço pode ter um valor diferente, para diferentes clientes (FRANZOSO, 2007). Além disso, questões como contexto cultural e classe social podem alterar o valor percebido por um cliente, influenciando na sua tomada de decisão. O preço do serviço está incluso dentre os fatores que influenciam no valor percebido, porém a definição é mais subjetiva, levando em consideração outros elementos como a utilidade do serviço, a qualidade com que ele é realizado, o retorno percebido, dentre outros.

> "Valor entregue ao cliente é a diferença entre o valor total para o cliente e o custo total para o cliente. O valor total para o cliente é o conjunto de benefícios que os clientes esperam de um determinado produto ou serviço. O custo total para o cliente é o conjunto de custos em que os consumidores esperam incorrer para avaliar, obter, utilizar e descartar um produto ou serviço."(KOTLER, P., 2000)

O custo, ou o valor monetário que o cliente está disposto a dispender para adquirir um serviço pode variar com base em questões culturais, sociais, e econômicas. A cultura, de acordo com Solomon (2016), refere-se à acumulação de significados, rituais, normas e tradições compartilhados em uma organização ou sociedade. A cultura do consumidor desempenha um papel crucial na definição das prioridades globais, influenciando a associação a diferentes atividades e produtos, bem como determinando o sucesso ou fracasso de produtos específicos. A aceitação de um produto no mercado é favorecida quando seus benefícios estão alinhados com os desejos culturais. Vale notar que a cultura é dinâmica, evoluindo continuamente ao sintetizar ideias antigas com novas, e ela regula a sociedade, estabelecendo normas e padrões de comportamento que todos os membros conhecem.

Segundo Sheth, Mittal e Newman (1999) classe social refere-se à posição relativa dos indivíduos na sociedade, representando uma classificação global das pessoas com base em características semelhantes de posição social na comunidade. Aqueles agrupados na mesma classe social geralmente ocupam ocupações semelhantes e compartilham estilos de vida similares devido a faixas de renda e preferências comuns. No contexto econômico, à medida que uma economia se desenvolve, a distribuição do emprego entre agricultura, indústria e serviços sofre mudanças significativas, com um aumento expressivo na participação dos serviços (WIRTZ; LOVELOCK, 2022). Essa característica não se limita a países desenvolvidos; países em desenvolvimento, como México, Venezuela, Argentina e Brasil, também refletem esse padrão.

Por fim, a satisfação é um ponto crucial para a garantia de uma relação duradoura entre empresa e cliente. Segundo Miniard, Engel e Blackwell (2000), a satisfação é descrita como a avaliação que ocorre após o consumo, indicando que a opção escolhida atendeu ou superou, no mínimo, as expectativas adquiridas no processo de venda. Em outras palavras, a satisfação é afirmar que o desempenho foi tão bom quanto se esperava. As expectativas do cliente são criadas segundo diversos fatores como experiências anteriores, propagandas, discurso de venda e experiências com a concorrência (FRANZOSO, 2007), dessa forma, é crucial trabalhar essas expectativas a fim de evitar desapontamentos.

2.2 CANCELAMENTO DE LICENÇAS DE USO DE SOFTWARE

Software nada mais é que um programa, constituído por linguagem de programação e dados, que controlam as operações do sistema de um computador, abrangendo processamento de texto, programas de Internet, navegadores etc. SILVA *et al.* (2017) definem *software* como a parte lógica de um sistema, ou seja, o processamento e interpretação dos dados pela máquina para posterior armazenagem, a exemplo de aplicativos como o Word, Bloco de Notas, Corel Draw, Internet Explorer, entre outros.

Cerqueira (2000) promove três diferentes classificações para tipos *software*, sendo *software* sob medida, *software* de prateleira e *software* de mercado de nicho. O *software* sob medida é desenvolvido de acordo com as necessidades e especificações do cliente, como exemplos de sites exclusivos para empresas específicas, como aquelas que possuem marcas de carros ou escritórios de advocacia. O *software* de prateleira, também conhecido como *software* produto, é comercializado e distribuído em larga escala por revendedores, conhecidos por utilizar dos Acordos de licença com o usuário final (EULA), ou seja, as licenças de uso . Por fim, o *software* de mercado de nicho é direcionado para um grupo específico, geralmente pronto para uso e exigindo apenas instalação e utilização.

Ainda, é possível classificar o tipo de contrato que o *software* possui a fim de ententer as responsabilidades das diferentes partes envolvidas em sua comercialização. A classificação levantada por Mota (1999) classifica os contratos como:

- Contrato de distribuição ou revenda ou ainda representação de *software*: refere-se a um acordo legal estabelecido entre um desenvolvedor de *software* e um terceiro (distribuidor, revendedor ou representante) para a comercialização e distribuição dos produtos de *software*. Esse contrato define os termos, condições e responsabilidades de ambas as partes envolvidas na cadeia de distribuição do *software*.
- Contrato de desenvolvimento de sistema por encomenda: onde a empresa se compromete a criar um sistema personalizado de acordo com as especifi-

cações e requisitos fornecidos pelo cliente.

- O contrato de edição ou de *publishers*: semelhante ao que as editoras de livros realizam com seus autores, refere-se a um acordo entre um desenvolvedor ou detentor dos direitos autorais do *software* e uma empresa ou indivíduo responsável pela manutenção, distribuição, comercialização e promoção do *software*.
- Contrato de manutenção de computador: está relacionado ao contrato entre a empresa desenvolvedora de *software*, e uma pessoa ou empresa que possui autoridade para prestar serviços de manutenção, suporte e atualização do *software* nos computadores dos clientes
- Contrato de licença de uso de *software*: é um acordo entre o detentor dos direitos autorais do *software* (licenciante) e o usuário (licenciado), estipulando os termos e condições para o uso do *software*.

Um contrato de licença definido por Almeida (2021) é o tipo de contrato no qual o titular de um direito sobre algo incorpóreo (licenciante) concedendo a outra parte (licenciado) o uso temporário e remunerado desse direito. O contrato de licença trata de bens incorpóreos e é caracterizado pela diversidade de licenças contratuais, a exemplo do contrato de licença de uso de *software*. O contrato de licença de uso de *software* é um acordo entre o detentor dos direitos autorais (concedente) e o comprador, no qual o proprietário autoriza o uso do *software* de forma não exclusiva e por tempo indeterminado. O licenciante concede ao licenciado o direito de utilizar o *software* em seus servidores, sendo o detentor dos direitos autorais o desenvolvedor ou licenciante (MARQUES, 2014).

São diversos os tipos de licença de uso de *software*, de forma que cada tipo busca atender diferentes necessidades e preferências dos contratantes. A Licença gratuita de software aborda programas que podem ser utilizados sem a necessidade de pagamento, podendo não garantir necessariamente a liberdade total do *software*. Esse tipo de *software* é frequentemente conhecido como freeware, exigindo, mesmo sem custo, a aceitação de contratos que podem impor condições específicas, como restrições de ambiente de uso.

Tratando de softwares que buscam retorno financeiro, Navita (2023) define a licença de software proprietário. Neste tipo de licença é proibido copiar, redistribuir ou alterar o software. O descumprimento dos termos contratuais pode resultar em medidas judiciais pela desenvolvedora. Para superar essas limitações, é necessário entrar em contato com o desenvolvedor para obter permissões adicionais ou adquirir licenças que ofereçam mais funcionalidades.

Para confrontar as licenças proprietárias existem as licenças de *software* livre. Apesar do nome livre, isso não está associado à gratuidade do software, mas sim às autorizações referentes ao uso do software, que são mais abrangentes, incluindo

acesso ao código-fonte e modificações nas funcionalidades (GUIMARÃES, 2016). Isso implica que o usuário possui total liberdade para adaptar o programa conforme suas necessidades pessoais ou empresariais. No entanto, é crucial salientar que mesmo com essa permissão, o programa não é considerado de domínio público, ou seja, mantém algumas limitações associadas ao direito de propriedade intelectual.

Guimarães (2016) define a licença de aquisição perpétua como aquela em que a distribuidora aborda o *software* como um ativo comercial, garantindo aos clientes o direito vitalício de uso, embora as atualizações e serviços de manutenção não estejam necessariamente inclusos. Nesse caso, o comprador torna-se proprietário, obtendo acesso ao código-fonte para futuras atualizações e uma ampla liberdade de uso e distribuição.

Já o modelo de licenciamento para aluguel de *softwares*, conhecido como ASP (Provedor de Serviços de Aplicativos), em que a armazenagem do programa é feita na nuvem (EVEO, 2022). Esse tipo de licença elimina downloads e instalações e as empresas pagam uma mensalidade conforme o número de usuários ou recursos necessários, enquanto o provedor cuida de atualizações, manutenção, disponibilidade e segurança. Dessa forma, entende-se que este tipo de licença aborda o software como um serviço (SaaS) fornecido, trazendo à tona a necessidade de um controle acerca do seu cancelamento.

2.2.1 Software as a Service (Saas)

Empresas provedoras de SaaS são aquelas que gerenciam seus aplicativos em seus próprios datacenters para vários clientes. Nesse contexto, segundo Benlian (2009), o custo total para fornecimento de um software torna-se previsível, dado que pode ser obtido pela soma do custo de manutenção da infraestrutura necessária para disponibilizá-lo e do pessoal necessário para fazer acompanhamento, retenção e manutenção das contas. Dessa forma, o lucro mensal para este modelo de negócio pode ser calculado ao subtrair o custo da receita recorrente paga pelo cliente.

Bibi, Katsaros e Bozanis (2012) destacaram que, enquanto o desenvolvimento de software no local foca na customização do produto, o desenvolvimento baseado em nuvem restringe a customização, mas em contrapartida reduz o custo total do produto. Na ótica do cliente, o modelo SaaS elimina a necessidade de compra e manutenção de infraestrutura de TI, uma vez que todo o processamento e armazenamento de dados são realizados nos servidores do provedor de serviços. Além disso, o SaaS permite escalabilidade flexível, permitindo que os usuários aumentem ou diminuam os recursos do software conforme suas necessidades mudam ao longo do tempo (RITTINGHOUSE; RANSOME, 2009).

Entretanto, para análise de sucesso de empresas SaaS além dos custos de desenvolvimento e manutenção também deve ser considerado o Custo de Aquisição de

Clientes (CAC), que se trata dos custos relacionados a todas as atividades de venda do licenciamento do software. Dessa forma, a retenção de clientes é extremamente crítica para que as empresas de SaaS sobrevivam, especialmente em mercados competitivos, dado que os custos associados a aquisição de um novo cliente é muito maior do que o custo de manter um cliente existente (KURTZ; CLOW, 1997).

O cancelamento dos clientes, mais conhecida no contexto de SaaS como *churn*, é um termo usado para designar a perda de clientes ou consumidores. A rotatividade de clientes é um grande custo para uma empresa, dado que o custo é entre cinco a vinte e cinco vezes maior para adquirir um novo cliente do que para reter um existente (CHIRITA, 2021). Segundo Gold (2020), a palavra "churn"deriva do conceito de "*churn rate*", que representa a proporção de clientes que deixam uma empresa durante um período específico. Essa expressão descreve a saída de clientes de um produto ou serviço. Portanto, o *churn* propriamente dito pode ser definido como o cliente que realizou o cancelamento, se tratando do oposto da retenção de clientes. Dessa forma, reduzí-lo é, essencialmente, equivalente a aumentar a retenção de clientes.

O *churn rate* no contexto de SaaS é calculado como a porcentagem de clientes que realizaram cancelamentos em um determinado período, quando comparado com toda a base de clientes ativos. Strouse (1999) trouxe uma tradução complementar para o termo, pois afirma que o *churn* vai além da perda de clientes, mas sim se tratando da rotatividade. Isso se dá pelo fato que, para este modelo de negócio a perda de clientes implica na necessidade de reposição da evasão para a manutenção da base original. Mehta, Steinman e Murphy (2016) apontam que uma empresa de *software* por assinatura bem sucedida e com futuro promissor possui uma taxa de cancelamento abaixo de 8%.

Para Daré e Campomar (2007), a gestão do churn se concentra na criação de métodos que permitam avaliar e gerenciar a taxa de cancelamento dos clientes. Uma tática para ser aplicada na gestão do churn se baseia na antecipação de quais clientes possuem uma maior probabilidade de encerrar a relação com a empresa, com a tomada de ações ativas de um time de sucesso do cliente a fim de evitar essa situação (NESLIN *et al.*, 2006). Cristine (2019) afirma que o conhecimento de quais são os clientes propensos ao cancelamento reduz os desperdícios com os departamentos de gestão do relacionamento com o cliente, responsável por realizar ações de atendimento que previnem o abandono do cliente.

Entretanto, apesar da necessidade de identificação dos potenciais *churns* de organizações, o escopo deste reconhecimento deve limitar-se ao do *churn* voluntário deliberado, dado que apenas para este caso que é possível de alterar a decisão de evasão dos clientes.

2.2.2 Fatores Influenciadores no Cancelamento de Licenças

No universo de empresas SaaS é comum encontrar estruturas organizacionais que possuam uma área totalmente direcionada à manutenção do relacionamento com os clientes, chamada de Sucesso do Cliente(CS) (PIRES, 2023).

> "Customer Success é uma área focada em criar e definir estratégias para que o consumidor, seja empresa (B2B) ou consumidor final (B2C), alcance seus resultados esperados com o produto e/ou serviço ofertado, retendo os clientes pelo maior tempo possível, tornando-os leais emocionalmente, fazendo-os comprar mais produtos e/ou serviços e tornandoos promotores da marca, indicando-a para amigos e conhecidos e realizando propaganda orgânica"(SANTOS, 2022)

Segundo Santos (2022), sucesso do cliente "é uma área focado na análise de dados, provenientes de pesquisas de opiniões realizadas pela área, essas pesquisas ajudam a determinam as demandas e urgências que a equipe deverá focar como prioridade."Sendo assim, pode ser considerada uma equipe proativa, pois emprega as informações dos consumidores com o auxílio de levantamento de métricas que direcionam a tomada de decisão, para analisar e identificar ações, seja por envolver riscos e evitar o cancelamento ou oferecer oportunidades de aumento de planos contratuais (*upsell*) (STEINMAN; MURPHY; MEHTA, 2017). Dessa forma, é possível entender que o trabalho realizado pelas equipes de CS envolvem o engajamento, a promoção da satisfação, a retenção e a expansão da base de clientes.

Steinman, Murphy e Mehta (2017) apontam três benefícios significativos para as empresas com a implantação de times de *Customer Success* (CS). Primeiramente, esses times contribuem para a redução da perda de clientes, evitando não apenas prejuízos financeiros, mas também a possível repercussão negativa, que pode levar outros a cancelarem seus serviços em favor de concorrentes. Falha no trabalho das equipes de atendimento ao cliente implicam diretamente na satisfação e no valor percebido pelo cliente, o que pode levá-lo ao cancelamento de sua assinatura. Uma das principais falhas geradoras de insatisfação para o cliente é a de comunicação. A deficiência na comunicação no atendimento ao cliente gera frustração levando a uma percepção negativa e insatisfação da qualidade do serviço. Dessa forma problemas técnicos acabam sendo agravados por um suporte ineficaz, o que pode levar o cliente a cancelar suas licenças.

O segundo benefício com a implantação de times de CS em empresas trata do aumento do valor dos contratos com os clientes existentes, pois a satisfação leva a um desejo de expansão. Por último, esses times aprimoram a experiência e a satisfação do cliente, o que eleva a percepção de valor, a credibilidade e a confiança na marca. Clientes satisfeitos tornam-se defensores orgânicos, os chamados "clientes leais"ou promotores, recomendando a empresa a amigos e conhecidos, exercendo influência positiva em suas decisões de compra. Conforme Steinman, Murphy e Mehta (2017),

existem dois tipos de lealdade. O primeiro é a lealdade intelectual, na qual os consumidores permanecem fiéis por necessidade. O segundo é a lealdade emocional/atitudinal, em que o cliente é leal a uma marca ou produto porque tem um amor genuíno por eles. Para as empresas, o obter a lealdade do segundo tipo por parte de seus clientes é mais vantajoso, já que envolve redução de riscos de cancelamento por problemas pontuais, disposição para pagar preços mais altos e menor vulnerabilidade à competição

Ainda, em muitas organizações, é papel do time de sucesso do cliente realizar o procedimento de *onboarding* de novos clientes. Para melhorar a qualidade do serviço, a equipe de CS deve se concentrar em todos os pontos de contato antes, durante e após a implementação do software. O *onboarding* do cliente é o processo de ajudar um novo cliente a usar e obter valor de um produto ou serviço o mais rápido possível (ADAMS, 2019). Se trata da primeira fase da jornada do cliente ao contratar um serviço, que começa quando a jornada do comprador termina e o contrato é assinado. Nas primeiras semanas com o cliente, o objetivo é construir uma relação de confiança e mapear as necessidades e desejos do cliente para começar a usar o produto ou serviço. As implementações de produtos incluem consultas, sessões de treinamento, dentre outras atividades (WEBER, 2021). Van der Kooij e Pizarro (2015) descrevem o *onboarding* do cliente como o momento em que se realiza a integração, instalação e ativação de um serviço. Eles destacam que essa fase não apenas é crucial para reduzir o risco de *churn*, mas também para preservar o investimento da empresa no cliente.

Weber (2021) define dois modelos de *onboarding* para clientes: o *onboarding* genérico e o *onboarding* personalizado. O *onboarding* genérico é ideal para produtos ou serviços com opções de personalização limitadas, envolvendo o uso de modelos e processos simplificados para uma integração eficiente e padronizada. Por outro lado, o *onboarding* personalizado é empregado para proporcionar uma experiência estrategicamente adaptada às necessidades específicas do cliente, sendo mais apropriado para produtos ou serviços complexos, customizados e integrados a diversos softwares. O *onboarding* personalizado visa compreender as necessidades e desejos do cliente, bem como antecipar as mudanças e o valor que o produto ou serviço proporcionará. Geralmente, clientes com diversos grupos de usuários solicitam o *onboarding* personalizado para assegurar um uso relevante e direcionado do produto.

A má execução do *onboarding* de clientes é uma das principais causas de churn. Isso envolve processos demasiadamente longos, planos de engajamento engessados e interações negativas, levando os clientes a cancelar (SANDSTRÖM, 2022). Totango (2023) afirma que os clientes provavelmente cancelarão sua assinatura se tiverem interações negativas ou problemas ao usar o produto ou serviço no início do seu relacionamento com um fornecedor. Weber (2021) destaca que empresas de SaaS têm apenas 90 dias para transformar clientes em usuários engajados, sendo os primeiros

dias cruciais para o sucesso do cliente. Ter vários usuários ativos é uma maneira eficaz de reduzir o risco de churn. Quando vários membros da equipe concluem o processo de *onboarding* do cliente, a implementação é provavelmente mais abrangente e o uso do produto em um nível mais alto (WEBER, 2021). Dessa forma, um acompanhamento eficaz durante o onboarding leva à adoção definitiva do produto e posterior renovação e melhorias de pacotes de contratação, criando clientes promotores (TOTANGO, 2023).

Outro fator que pode levar ao cancelamento de clientes é a mudanças da necessidades do usuário. É importante entender quem são os clientes, quais são as suas necessidades e o que deve ser fornecido para satisfazer essas necessidades, com uma ampla análise dos concorrentes e o que eles realizam no mesmo contexto, a fim de não apenas cumprir, mas exceder as expectativas (WALLACE, 1992). As necessidades devem ser levantados já no início do processo de desenvolvimento, a fim de evitar impactos negativos na relação com o cliente e o custo da correção (DROMEY, 2003).

Existem dois tipos de necessidades que influenceiam no valor percebido de um cliente perante um produto ou serviço. As necessidades primárias referem-se ao que o indivíduo precisa em sua vida, mas não influenciam na escolha da marca, ou seja, os requisitos básicos que determinam a necessidade de um produto ou serviço. Já as necessidades secundárias dos consumidores estão associadas a fatores como identidade visual, preço e qualidade, ou seja são determinadas pela marca e podem conflitar com informações de concorrência. As necessidades secundárias são as que mais estão abertas a mudança por parte do cliente, dado que a ampla concorrência fomenta o desenvolvimento de produtos e serviços de maneira contínua. Dessa forma, as necessidades secundárias após mapeadas devem monitorados e fim de encontrar alterações nas expectativas do cliente alvo, influenciando no valor percebido. O marketing empresarial é um agente ativo ao desempenhar o papel de identificar e atender às mudanças nas necessidades dos consumidores, garantindo a satisfação do público-alvo (PAULILLO, 2023).

Uma necessidades extremamente relevante na relação de um cliente com um *software* é a usabilidade, que determina a experiência do usuário. Essa experiência implica na decisão de cancelamento de clientes, e nada mais é que uma medida de quão bem um usuário específico em um contexto específico pode usar um produto para atingir um objetivo definido de forma eficaz, eficiente e satisfatória (IDF, 2023). Mais de 90% dos clientes sentem que as empresas com as quais fazem negócios poderiam melhorar a usabilidade de seus produtos (CHIRITA, 2021). A usabilidade de um produto depende de quão bem seus recursos atendem as necessidades e contextos dos usuários. A usabilidade é definida com base nestes cinco elementos:

1. Eficácia do produto para apoiar os usuários na conclusão de ações com precisão.

2. Eficiência do produto para permitir que os usuários executem tarefas com rapidez e facilidade.
3. Engajamento dos usuários que acham agradável utilizar o produto de acordo com suas necessidades.
4. Tolerância a erros para apoiar as ações do usuário e só mostra um erro em situações genuínas de erro.
5. Facilidade de aprendizagem para novos usuários cumprirem metas sem a ajuda de alguém (IDF, 2023).

Um exemplo de déficit da experiência do usuário é em casos de criação de conta e autenticações (SHANMUGAM, 2022). Por exemplo, se tratando de cadastro de uma nova conta, 64,5% dos consumidores abandonarão um site se for solicitado a criar um nome de usuário e uma senha, e 66% dos consumidores abandonam um site se o processo para registrar sua conta for demasiado complexo (HNS, 2021). Se tratando de acesso, nos casos em que um usuário esquece a senha ou não consegue concluir a autenticação de dois fatores, 92% dos usuários preferem sair de um site do que recuperar ou redefinir suas credenciais de login. Estas situações normalmente implicam em frustração que pode levar o cliente optar pela concorrência (DAY, 2023). Isso pode ser extrapolado para o conceito de comercialização de softwares, caso a experiência do cliente com a usabilidade da plataforma não seja agradável, maiores são as chances de *churn*.

Outra necessidade em alta atualmente no contexto de *software* é a segurança que a provedora de tecnologia garante para os seus clientes. Segundo Waters (2005) para aplicações empresariais, a segurança dos dados é crucial e, no entanto, a gestão da segurança baseada na Internet requer conhecimentos especializados e atualizações quase constantes. No brasil, a Lei Geral de Proteção de Dados (LGPD), relacionada à segurança dos dados, estabelece diretrizes e regulamentos para o tratamento de informações pessoais. A lei visa proteger a privacidade e a segurança dos indivíduos, garantindo que as organizações processem e armazenem dados de maneira segura. Isso inclui a implementação de medidas técnicas e organizacionais para prevenir vazamentos, acessos não autorizados e garantir a integridade dos dados pessoais. Fornecedores de *software* como serviço tendem a ter uma infraestrutura de datacenter extremamente robusta e podem fornecer segurança de aplicativos que atenda ou exceda seus próprios padrões de segurança internos (WATERS, 2005). Entretanto, algumas organizações com elevadas necessidades de segurança têm questões culturais ou políticas que as fazem sentir-se mais seguras se tiverem a posse física de todos os seus dados. Nessas situações, o SaaS pode ser menos adequado para os negócios, independete do cumprimento das demais necessidades do cliente.

2.2.3 Estratégias de Retenção e Prevenção do Cancelamento

Retenção de clientes é essencial para alcançar o sucesso dada a grande competitividade dos mercados atuais, sendo o elemento crucial para consolidar a presença no mercado, com um foco direto na satisfação do cliente (CHRISTOPHER, 2022). Entretanto, segundo Vavra (1993), a retenção não exclui mas vai além da satisfação do cliente, com uma ótica que é uma nova forma de produzir lucro para organizações, assim implicando em maior competitividade de mercado. Não obstante, a retenção também pode ser considerada sinônimo de força de marca, e está diretamente associada à elevação da participação da empresa no mercado direcionado. Em um estudo realizado por Hennig-Thurau e Klee (1997) foi desenvolvido um modelo conceitual que determinam retenção de clientes como um resultado da satisfação dos clientes, além da qualidade e da confiança percebida por eles e do comprometimento realizado por parte do fornecedor de serviços.

Entretanto, nem todos os clientes devem ser alvo de ações ativas de retenção. Em seu trabalho, Ganesh, Arnold e Reynolds (2000) concluiram que, apesar dos benefícios das ações de aumento da retenção e da lealdade de clientes, o esforço não deve ser direcionado para a totalidade da base de clientes, sendo que existem clientes que estão totalmente satisfeitos com o atendimento oferecido. Além dos clientes já satisfeitos, existe uma parcela de clientes que não poderão ser retidos por fatores que vão além do controle, não importa qual seja o nível de esforço empregado pelo fornecedor. Dessa forma, é preciso aplicar investimentos para direcionar esforços nos clientes que possuem real possibilidade de retenção a fim de buscar maior eficiência nas estratégias de prevenção do *churn*.

A existência de um time de *customer success*, mencionada no tópico anterior, já abrange uma série de estratégias que buscam a redução das taxas de cancelamento. Para isso, as marcas devem entender a jornada inteira de seus clientes e como cada interação contribui para uma experiência conectada mais ampla (COTTON, 2022). Assim, as equipes de CS podem fornecer experiências de suporte e atendimento personalizadas, de forma a fornecer valor e captar a lealdade dos mesmos (SALLES, 2023). Para estes times, a comunicação é essencial, sendo que um atendimento lento pode minar a confiança no negócio. Dessa forma, Curvelo (2022) aponta o crescimento do conceito de *omnichannel* para suporte e atendimento, ou seja, o uso de diversos canais de comunicação, como números de telefone e e-mail evidentes nas ferramentas, chatbots de atendimento em tempo real em sites e atendimento via redes sociais, garantindo que o acesso ao suporte está facilitados para os diferentes tipos de clientes.

São estratégias comuns aplicadas por times de sucesso do clientes o contato proativo com clientes propensos a dificuldades e dúvidas, visitas *in loco* a clientes estratégicos, a adoção de programas de fidelidade, definição de um processo claro de *onboarding*, personalizado para cada cliente e a aplicação de pesquisas de satisfação

do cliente (CURVELO, 2022). A abordagem proativa por parte das equipes de sucesso do cliente é crucial para fortalecer as relações com os clientes em um ambiente digital. Para Cotton (2022), ao antecipar necessidades e identificar possíveis problemas antes que se tornem significativos, as equipes podem oferecer soluções personalizadas e eficazes, demonstrando um compromisso contínuo com o sucesso do cliente. Segundo Salles (2023), esse contato proativo não apenas contribui para a fidelização, construindo relacionamentos mais sólidos, mas também oferece oportunidades para apresentar novos recursos, coletar feedback contínuo e, fundamentalmente, reduzir a probabilidade de churn. A comunicação ativa com os clientes existentes pode ser realizada por meio de agendando chamadas ou reuniões para identificar as dificuldades ou necessidades que possam surgir durante o uso da solução, notificações como alertas sobre manutenções programadas ou atualizações do sistema e plataformas de mídia social e e-mails, normalmente com conteúdos informativos que agreguem valor ao negócio do cliente.

Embora as interações virtuais, seja por chamadas telefônicas ou reuniões remotas, sejam frequentes, realizar visitas presenciais ao cliente (*in loco*) é uma estratégia fundamental pois fortalece os vínculos e proporciona uma compreensão mais detalhada do seu cotidiano e necessidades específicas (CURVELO, 2022). Essas visitas não apenas fortalecem o relacionamento, mas também permitem uma análise direta das soluções implementadas, identificando possíveis melhorias e otimizações. Além disso, o contato face a face pode ser considerado mais um canal para esclarecimento de dúvidas, mostrando de forma clara um compromisso genuíno com a construção de uma parceria duradoura com o cliente.

No tópico anterior já foi descrito o processo de *onboarding*, e seu papel como um fator influenciador na lealdade do cliente com a marca, ao estabelecer uma introdução consistente e adaptada às necessidades específicas de cada usuário. Curvelo (2022) afirma que adotar um novo produto ou serviço pode gerar desconforto inicial para os clientes. Se eles não conseguirem navegar eficientemente no produto ou serviço logo no início, é provável que percam o interesse rapidamente. Dessa forma, é papel dos times de CS estabelecer um processo, ou roteiro, de *onboarding* que seja ao mesmo tempo padronizado para atender o mínimo das expectativas dos clientes, mas flexível para ser adaptato à realidade de cada um. Essa abordagem não apenas gerencia as expectativas dos clientes, mas também proporciona controle sobre o ritmo de apresentação de informações, e a personalização desse processo é fundamental, considerando que cada consumidor possui objetivos específicos, garantindo assim a resolução de todas as dúvidas.

O processo de *onboarding* trata de um treinamento inicial que comunica as necessidades básicas para a utilização do produto oferecido. Entretanto, os produtos estão em constante desenvolvimento, e novos métodos de aplicação frequentemente

são criados de forma a facilitar a usabilidade do cliente. Ainda, as necessidades do cliente estão em constante mudança, e um recurso da ferramenta que anteriormente não era utilizado pode tornar-se necessário com essa mudança. Dessa forma, vê-se necessária a oferta contínua de materiais educativos, como tutoriais, guias detalhados e vídeos explicativos de maneira constante para garantir o domínio efetivo do software, possibilitando uma exploração abrangente das características e funcionalidades proporcionadas (SALLES, 2023). Este comprometimento educacional não apenas fortalece os laços entre a empresa e o cliente, mas também contribui para a eficiência operacional, assegurando que os usuários alcancem o máximo potencial do software, resultando em experiências mais produtivas e gratificantes.

Depois que o consumidor efetuou a primeira compra, e está treinado para utilizar a ferramenta, vê-se necessario buscar formas de fazer com que ele queira voltar a adquirir produtos ou serviços da sua marca (ROUTH; ROY; MEYER, 2021). Uma estratégia comum para as empresas reduzirem a rotatividade é fornecer incentivos aos clientes com probabilidade de abandonar, para que mantenham o seu relacionamento com a empresa (NESLIN *et al.*, 2006). Estes incentivos dependem de qual o objetivo que busca ser alcançado, indo desde recompensas para clientes fiéis, até um desconto para um cliente propenso ao cancelamento. Outro exemplo são os incentivos do tipo compre um e leve outro, que pode ser projetado para promover *cross sell* em diferentes categorias de marca. Por outro lado, uma venda relâmpago pode ser projetada para reduzir o tempo entre as compras. Routh, Roy e Meyer (2021) postulam que, se diferentes aspectos do comportamento do cliente afetam diferentes vias de rotatividade, então o design de campanhas deve ser diferente com base na identificação destas vias.

Empresas que compreendem a conexão entre vários elementos mensuráveis da experiência do cliente e o comportamento individual do cliente ganham uma vantagem competitiva na identificação e tomada de ações baseando-se nos fatores que influenciam a fidelidade e a rotatividade (KON, 2004). Isso possibilita que a estratégia da diferenciação possa ser aplicada a fim de garantir que o *software* não apenas atenda às necessidades gerais, mas também àquelas inerentes as particularidades de cada cliente. Para Correia, Barreto e Alves (2024) essa identificação das necessidades possibilita a identificação de produtos ou serviços suplementares que atendam às demais demandas do cliente. Nesse sentido, a capacidade de formar alianças estratégicas com outras organizações, principalmente aquelas atuantes em setores complementares, pode gerar novas perspectivas de mercado e agilizar o processo de expansão (CARLOS, 2023).

A insatisfação do cliente, que pode levar ao cancelamento, surge quando há falta de clareza sobre um produto ou serviço, ou quando não há suporte efetivo para resolver problemas específicos (COTTON, 2022). Dessa forma, as pesquisas de satisfação do

cliente surgem para identificar os problemas específicos que impactam a empresa fornecedora de *software*. A coleta regular de feedback ajuda a identificar necessidades e pontos de melhoria do negócio, assim direcionando a criação de novas estratégias de negócio. Entretanto, apenas realizar pesquisas de *feedback* não significa o sucesso das estratégias realizadas. Kon (2004) afirma que as pesquisas de feedback quando não concebidas e interpretadas adequadamente podem ser altamente enganosas na identificação e priorização de áreas-chave nas quais concentrar os investimentos em retenção. A captura e avaliação da satisfação do cliente possui um alto custo envolvido, o que leva as empresas a dar grande credibilidade aos resultados, chegando ao ponto de vincular a remuneração de colaboradores aos níveis de satisfação do cliente.

No entanto, a satisfação declarada do cliente só é relevante se indicar como os clientes irão alterar o seu comportamento no futuro. A satisfação em termos absolutos significa pouco; por exemplo, quase todos os viajantes frequentes diriam que estão insatisfeitos com a pontualidade da sua companhia aérea regular, mas a maioria reconhece que o problema é endémico nas viagens aéreas modernas e é pouco provável que escolham uma companhia aérea diferente para a sua próxima viagem. Por outro lado, esses mesmos viajantes poderão afirmar no mesmo inquérito que estão muito satisfeitos com a qualidade da refeição a bordo. No entanto, a oferta de um concorrente de refeições antes do voo em voos noturnos ainda pode levar muitos a mudar de companhia aérea(KON, 2004).

Assim, Kon (2004) afirma que no contexto de serviços de assinatura, a baixa satisfação declarada do cliente com a confiabilidade e a qualidade deve ser vista no contexto das expectativas dos clientes e da experiência com os serviços dos concorrentes. É importante interpretar adequadamente a satisfação declarada com os vários elementos da experiência e elaborar um índice de satisfação que considere a importância relativa dos componentes individuais. Dessa forma, é possível concluir que as expectativas dos clientes desempenham um papel crucial na avaliação da satisfação a fim de cultivar lealdade e reduzir as taxas de rotatividade. Curvelo (2022) afirma que quando as experiências proporcionadas por um serviço ficam aquém das expectativas do cliente, a probabilidade de insatisfação aumenta. Portanto, entender e gerenciar as expectativas desde o início do relacionamento é fundamental para garantir a satisfação contínua do cliente. A experiência do cliente não se limita apenas ao desempenho do produto ou serviço, mas abrange todos os pontos de contato, desde a fase de pré-compra até o suporte pós-venda.

Para o entendimento das necessidades dos clientes, é possível utilizar de ferramentas analíticas de dados, como a análise de métricas relacionadas ao comportamento do usuário em ambientes digitais, visando identificar padrões e preferências (CANDIDO *et al.*, 2023). Adicionalmente, a exploração de plataformas de mídia social também oferece insights valiosos acerca das percepções e exigências dos clientes.

Dessa forma, ao adotar uma abordagem abrangente, as organizações podem ajustar suas estratégias de acordo com as necessidades emergentes, implementando ações corretivas antes que o cliente considere o cancelamento e assim solidificando assim os vínculos com os clientes de maneira duradoura e substancial. Kon (2004) afirma que uma melhor compreensão dos fatores de rotatividade só será valiosa se for seguida por um plano de ação que considere as informações levantadas. Muitas empresas são frequentemente vítimas da paralisia da análise, não incentivando a mudança no investimento e recurso da empresa para alteração do comportamento e design do produto.

Ainda, Curvelo (2022) afirma que um grande desafio é identificar as necessidades que os clientes ainda não enxergaram. Embora os clientes possam expressar com precisão suas necessidades imediatas, compreender o que desejam ou precisarão no futuro é um desafio considerável, e é papel das empresas antecipar as necessidades com o estudo das transformações na sociedade e inovando no meio digital. Para Correia, Barreto e Alves (2024), "a inovação constante é a espinha dorsal do sucesso a longo prazo em SaaS.". Existe uma diferença entre inovação incremental e disruptiva. A primeira caracteriza-se por aprimoramentos graduais nas soluções preexistentes, ao passo que a segunda abraça mudanças radicais que reconfiguram fundamentalmente a prestação de serviços. Encontrar um equilíbrio apropriado entre essas duas formas de inovação é essencial para assegurar a continuidade da relevância e a manutenção da liderança no cenário de mercado (BOSAK, 2021). A aplicação de inovações voltadas à retenção com o auxílio de análise de métricas de comportamento do usuário garantem que estão sendo tomadas ações para solucionas as causas mais profundas da rotatividade (KON, 2004).

Algumas pesquisas tentam encontrar as melhores estratégias para identificar as necessidades dos clientes e os principais indicadores que influenciam a rotatividade. Em seu trabalho, Routh, Roy e Meyer (2021) determinou uma estrutura de modelagem que se baseia em três áreas de pesquisa existentes: programas de gerenciamento de rotatividade de clientes para eventos múltiplos, estimativa do risco de rotatividade em eventos concorrentes e modelos de previsão que estimam a rotatividade de clientes. Esses estudos buscam a aplicação de modelos de aprendizado de máquina a fim de, com base em padrões de comportamento, identificar quem são os clientes propensos à rotatividade e suas causas. Segundo Correia, Barreto e Alves (2024), a introdução de técnicas de aprendizado de máquina e inteligência artificial possibilita uma personalização mais apurada das soluções, assegurando uma experiência do usuário excepcional e fortalecendo os vínculos com a base de clientes existente, possibilitando que as empresas a superem desafios inesperados e prosperem em um contextos empresariais dinâmicos.

2.3 MACHINE LEARNING

O *Machine Learning* é um conceito que está compreendido dentro da área de Inteligência Artificial, que por sua vez é uma área cujo surgimento se deu nos estudos de Ciência da Computação (BATISTA, 2022). De acordo com Mitchell (1997), o aprendizado de máquina é definido como o estudo de algoritmos computacionais que podem melhorar automaticamente por meio da experiência. As técnicas de *Machine Learning* se dedicam principalmente à previsão e à classificação, utilizando informações aprendidas a partir de um conjunto de dados de treinamento como base, assim identificando padrões e regularidades, para realizar tarefas específicas (SCHNEIDER, 2016).

São diversas as possibilidades de aplicação do *machine learning* e as técnicas aplicadas costumeiramente dependem da base de dados que está sendo utilizada, e de qual resultado busca ser encontrado (ROLLINS, 2015), assim se trata de uma tecnologia que pode ser aplicada em diversas áreas, como reconhecimento de fala, processamento de imagem, análise de dados etc. Sivakumar e Gunasundari (2017) reforçam a necessidade de atenção no pré-processamento de dados. Esta que é uma etapa em que é feita uma limpeza de dados irrelevantes ou mal coletados de forma que não atendem os padrões ou pré-requisitos para uma boa entrada no modelo.

Burkov (2019) levanta que existem três tipos diferentes de aplicação de machine learnig: o Sistema de Aprendizado Supervisionado, que tem como objetivo prever uma variável dependente a partir de uma lista de variáveis independentes; o Sistema de Aprendizado Não-Supervisionado, que normalmente está associado ao uso de uma grande quantidade de informações não rotuladas, que buscam ser separadas ou "*clusterizadas*; e por fim o Sistema de Aprendizado por Reforço, no qual a máquina toma decisões sequenciais em um ambiente com o objetivo de maximizar uma recompensa cumulativa.

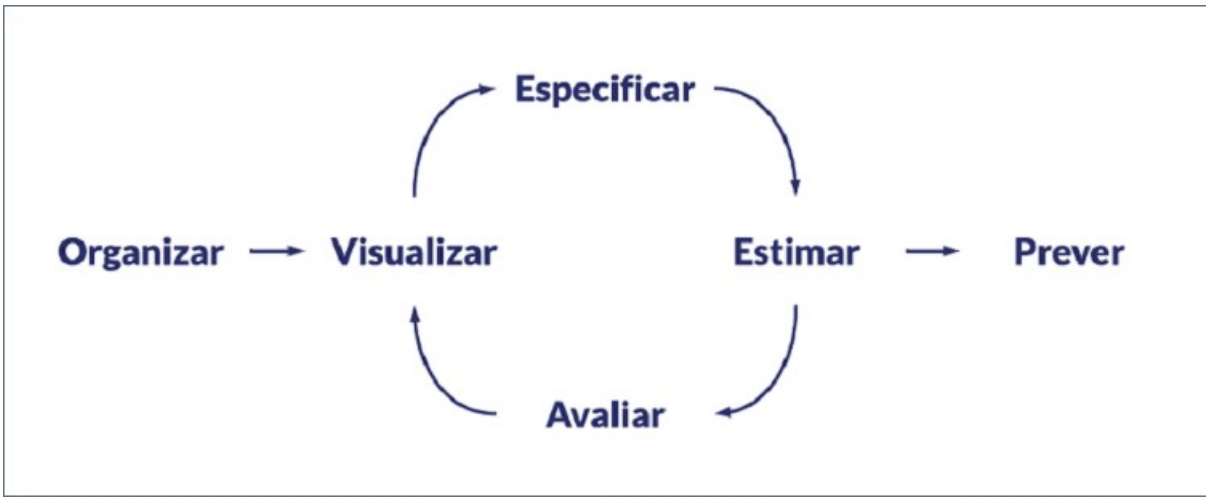
Makridakis, Spiliotis e Assimakopoulos (2018) afirmam que uma aplicação muito comum de técnicas de *machine learning* é na preparação de modelos de predição. De acordo com Hastie, Tibshirani e Friedman (2009), a modelagem preditiva é uma abordagem que utiliza algoritmos e técnicas estatísticas para construir modelos capazes de fazer previsões ou estimativas sobre eventos futuros com base em dados históricos. Essa técnica busca identificar padrões e relações nos dados existentes para realizar projeções e inferências.

A modelagem preditiva é um conceito amplo, e trata das técnicas utilizadas na construção de modelos que buscam prever comportamentos futuros (NYCE, 2007). Ao construir um modelo preditivo, é essencial considerar a seleção adequada de variáveis, as técnicas de pré-processamento de dados a serem utilizadas e a validação e avaliação dos modelos para garantir a qualidade das previsões.

Estes modelos costumam ser Sistemas de Aprendizado Supervisionado, onde o

resultado é encontrado a partir de iterações de tentativa e erro com alterações na base de dados, na base de entrada, até que se encontre métricas de avaliação aceitáveis para o treinamento. Ou seja, considera-se que neste tipo de aplicação existe um ciclo de tratamento, modelagem e análise dos resultados. Silva (2023) define o fluxo de trabalho para a realização de uma modelagem preditiva, composto por Organizar, Vizualizar, Especificar, Estimar, Avaliar e por fim Prever (Figura 1).

Figura 1 – Fluxo de Modelagem Preditiva



Fonte: Silva (2023)

O sistema de aprendizado supervisionado é uma abordagem de aprendizado de máquina em que um modelo é treinado utilizando exemplos de entrada e saída conhecidos. Nesse tipo de aprendizado, o objetivo é mapear corretamente as entradas para as saídas desejadas com base nos exemplos fornecidos durante o treinamento (MITCHELL, 1997).

Hastie, Tibshirani e Friedman (2009) explicam que o aprendizado supervisionado é um dos principais exemplos de aplicações do aprendizado de máquina, onde um modelo é treinado usando um conjunto de exemplos rotulados.

Nesta abordagem, são fornecidos dados históricos vindos de fontes externas ao modelo sendo tanto os dados de entrada quanto os de saída, ou hipóteses, conhecidos de forma a mapeá-los, encontrando o menor erro (SCHNEIDER, 2016). Entretanto, encontrar a menor taxa de erro não implica diretamente que o modelo será ótimo ao serem inseridos valores desconhecidos. É possível que após o treinamento, o modelo encontre uma solução muito genérica e simples para um problema complexo, ocorrendo o *underfitting*. Por outro lado é possível que o modelo acabe se adequando demais à hipótese ocorrendo o *overfitting*, que impede uma aplicação generalizada para novos dados de entrada (SILVEIRA *et al.*, 2019).

Dentre os sistemas de aprendisado supervisionado estão os métodos de classificação. Segundo Höppner *et al.* (2020), a previsão do cancelamento de clientes pode ser formulada como um problema de classificação binária. O objetivo é atribuir clientes a uma das duas classes: 0 para cliente não propensos ao cancelamento e 1 para os clientes propensos ao cancelamento. Este resultado se dá com base em suas características observadas, os parâmetros de entrada do modelo.

2.3.1 *Estudos Correlatos*

Existem diversos estudos já realizados no contexto da aplicação de técnicas de modelagem preditiva na identificação de probabilidades de *churn*. Em um levantamento bibliográfico realizado por Schneider (2016), foram analisados 80 diferentes estudos quanto à aplicação de diferentes modelos preditivos e ao ramo de atuação das empresas objeto de estudo. Como resultado, as técnicas de *machine learning* mais utilizadas na época englobavam principalmente Árvores de decisão (DT), Regressão Logística (LR), Redes Neurais Artificiais (ANN), *Suport Vector Machine* (SVM), o algoritmo de *Random Forest Algoritm* (RFA) e algoritmos que aplicam técnicas de *Boosting* como o XGBoost (XGB).

No contexto de serviços, o estudo de Keaveney (1995) investigou como incidentes críticos fez com que os clientes mudassem de um provedor de serviços para outro. Na pesquisa foram identificados oito categorias de incidentes críticos pelos quais os clientes trocam de prestadores de serviços: preços, inconveniência, falhas de serviços principais, resposta a falhas de serviço, concorrência, problemas éticos e troca involuntária. Apesar de a maioria destas causas ser controlável, uma troca involuntária não podem ser abordados diretamente pelos esforços da empresa, tornando impossível a sua detecção.

Quanto ao ramo de atuação, um setor relevante onde esta tecnologia foi muito aplicado é o de telecomunicações. Nessa linha, Idris e Khan (2012) desenvolvram um modelo de classificação com a aplicação de RFA, *Rotation Forest* e *RotBoost* para predição do cancelamento. Posteriormente Ahmad, Jafar e Aljoumaa (2019) aplicaram técnicas de machine learning em plataformas de big data afim de prever quais os clientes mais propensos ao cancelamento.

Outro setor em que muito é aplicado mecanismos de predição de cancelamento é o das instituições financeiras e bancos. Kumar e Leema (2023) abordaram a perda de clientes de bancos como uma forma de churn e desenvolveu modelos predidivos baseados em ANN, SVM e RFA a fim de maximizar os lucros. Charandabi (2023) realizou um estudo onde comparou o desempenho de seis técnicas de modelos de classificação supervisionada para propor um modelo eficiente para prever a rotatividade de clientes no setor bancário europeu. Baby *et al.* (2023) foi além e desenvolveu um modelo para o setor bancário usando ANN que para prever desligamentos voluntários de clientes,

e aplicou técnicas de validação cruzada que melhoram o desempenho em relação à exatidão e taxa de precisão.

Entretanto, são diversos os setores em que essas tecnologias podem ser aplicadas, como *marketing*, *e-commerce* e empresas que aplicam serviços de assinatura B2B e B2C. Suh (2023) realizou um trabalho que analisa as informações sobre o comportamento do cliente de uma locadora de purificadores de água para residências, utilizando de modelos de *Random Forest Algoritm* e *Gradient Boost* para prever rotatividade. Ainda, fora do contexto de assinaturas, Luizon (2019) formulou um conceito de churn para um *e-commerce* abordando possíveis clientes que não concretizaram a venda, e buscou prever quais seriam estes clientes com a aplicação de Árvores de Decisão, Redes Neurais Artificiais e *Suport Vector Macine*.

No contexto de softwares, existe uma ampla pesquisa de predição do fenômeno do *churn* na indústria de jogos (TAO *et al.*, 2022) (PERIŠIĆ; PAHOR, 2023). Ainda que exista pouco conhecimento desenvolvido no contexto de previsão de cancelamento em *softwares* B2B, alguns estudos como o de Marín Díaz, Galán e Carrasco (2022) foram realizados. Nele, foram aplicados algoritmos com o acréscimo de informação de transações de compra como recência, frequência, importância e duração (RFID) aos modelos tradicionais. Já o estudo de De Caigny *et al.* (2021) inovou ao utilizar um modelo chamado *uplift logit leaf* que combina desempenho preditivo com interpretabilidade, oferecendo novas aplicações na forma de visualização dos resultados mais direcionados ao *marketing* industrial.

3 METODOLOGIA

No presente capítulo serão apresentados os procedimentos metodológicos que serão aplicados na pesquisa. Primeiramente é definido o tipo de pesquisa desenvolvida, passando posteriormente para um levantamento dos materiais e métodos para aplicá-la. Por fim, o capítulo traz o levantamento das etapas metodológicas, indo desde a revisão de literatura realizada, passando pelo levantamento e tratamento de dados, partindo para o treinamento dos modelos e finalizando com a proposição de ações para a prevenção do cancelamento.

3.1 ENQUADRAMENTO DA PESQUISA

Segundo Gil (2008), é necessário classificar uma pesquisa cientifica de forma a organizá-la, ajudando na racionalidade das etapas para que o trabalho possa ser desenvolvido dentro das limitações de tempo e recursos, e mesmo assim alcançar os resultados propostos. Dessa forma, quanto a natureza o presente trabalho é classificado como uma pesquisa aplicada, haja vista que os resultados solucionam um problema real da empresa na qual está sendo realizada, com uma abordagem quantitativa.

3.2 MATERIAIS E MÉTODOS

Nesta seção são apresentados os dados disponíveis para serem utilizados na pesquisa. Em seguida os métodos de predição, de avaliação e de otimização são explicados.

3.2.1 Materiais

Para a criação de modelos de aprendizado de máquina é necessário a aplicação de linguagens de programação, dessa forma este estudo utilizará a linguagem de programação Python como ferramenta de manipulação de dados e criação dos modelos. O Ambiente de Desenvolvimento Integrado (IDE) para a aplicação do python será o Google Colaboratory, uma plataforma gratuita baseada na nuvem oferecida pelo Google, que proporciona um ambiente interativo para desenvolvimento de código Python. Uma das principais vantagens do Colab é a sua integração com o Google Drive, permitindo o armazenamento e compartilhamento fácil dos dados. Essa escolha visa aproveitar a praticidade, acessibilidade e os recursos robustos fornecidos pelo Google Colaboratory para facilitar o desenvolvimento do estudo com uma coleta de dados rápida e eficiente. Os dados disponíveis para o presente estudo vêm de três diferentes fontes:

1. Dados Financeiros: mensalmente, os dados de faturamento dos clientes são atualizados. Desses dados são retiradas informações como tempo de vida, receita recorrente mensal, data do primeiro faturamento, dentre outras.
2. Dados do Hubspot: os dados presentes no CRM da empresa envolvem informações referentes às empresas que são utilizadas pelas equipes Comercial e de Sucesso do Cliente. Além disso, são fornecidas as transações comerciais chamadas de "Négócios". Sempre que uma transação comercial é iniciada é criado um novo negócio, e um novo pedido de cancelamento por parte de um cliente é considerado uma transação comercial. Por fim, o software possui informações de atendimento, chamados "Tíquetes", que são criados sempre que um cliente solicita auxílio da equipe de Sucesso do Cliente
3. Dados do Sistema: no contexto de empresas SaaS, todos os dados gerados pelos clientes ficam de domínio da fornecedora do software. Dessa forma, dados de uso do sistema como o volume de utilização de funcionalidades estão disponíveis para consulta.

O Quadro 1 apresenta o resumo dos dados disponíveis para a criação dos modelos de predição.

O modelo utilizará todos os dados históricos de clientes que em algum momento contrataram a solução comercializada pela empresa objeto de estudo. A base de Dados Financeiros é o local onde esta informação está disponível, e ela é filtrada anteriormente à aplicação de qualquer análise, de forma a excluir os observações com indicações positivas para safra e internacional. Essa exclusão se dá pelo fato que estas observações possuem formas de registros de dados que não são compatíveis com o montante geral da base, assim totalizando 4253 empresas.

Entretanto ao invés de focar na resposta *churn* ou não *churn*, a pesquisa terá como intenção a previsão da intenção de cancelamento, ou seja, a intenção de *churn*. Isso se dá pois, nem todos os clientes que apontam a intenção de cancelar efetivamente cancelam suas assinaturas. Além disso, existem casos de clientes que apontam intenção de cancelamento em mais de uma ocasião. Dessa forma, considerando estes aspectos a base de dados financeiros é mesclada com a base de Negócioo do Hubspot considerando apenas o cancelamento dos clientes e ex-clientes, que são os registros da intenção, totalizando 4435 observações. Estas obeservações estão organizados da forma: registros de intenção de cancelamento e registros sem intenção de cancelamento, sendo o segundo definido como qualquer cliente que nunca abriu um negócio de cancelamento.

Os dados das bases Empresa do Hubspot e Tíquetes do Hubspot complementam a base de dados, com informações categóricas para o caso das empresas e numéricas para o caso da contagem de tíquetes. Além disso, os dados do sistema são

Quadro 1 – Dados disponíveis para o estudo.

Dados Financeiros	
Dado	**Descrição**
Codigo Alternativo	Código único de cada cliente, que é utilizado como chave primária
Descricao	Indica qual solução é contratada pelo cliente.
Safra	Booleano que indica se o cliente comercializa produtos que possuem ou não uma safra bem definida.
Internacional	Booleano que indica se o cliente está localizado ou não em território brasileiro.
Segmento	Segmento de atuação de do cliente, que pode ser tanto Produtor quanto Distribuidor.
Origem	Qual fonte de prospecção trouxe o cliente para a base. Pode ser Inbound e Outbound.
Codigo Periodo	Qual a frequência de pagamento desse cliente.
Primeiro Faturamento	Data do primeiro registro de faturamento do cliente
Ultimo Faturamento	Data do último registro de faturamento do cliente
MRR	Valor da receita recorrente mensal do cliente.
Empresa do Hubspot	
Dado	**Descrição**
Codigo Alternativo	Código único de cada cliente, que é utilizado como chave primária
Estado	Estado em que a empresa cliente está localizada.
Segmentação SC	Segmentação de nível de serviço definida pela equipe de sucesso do cliente.
Negócio do Hubspot	
Codigo Alternativo	Código único de cada cliente, utilizado como chave primária
Data de Criacao Negocio	Data em que o negócio foi criado dentro do CRM.
Tipo de Negocio	Informação que categoriza a negociação quanto à sua natureza. Pode um negócio de venda comercial, ou uma solicitação de cancelamento.
Etapa do negocio	Informação que indica se a negociação com o cliente está em andamento ou se já foi encerrada.
Pipeline	Informação que categoriza o negócio quanto à equipe responsável.
Tíquetes do Hubspot	
Codigo Alternativo	Código único de cada cliente, que é utilizado como chave primária
Data de Criacao Tiquete	Data em que o tíquete foi criado dentro do CRM.
Dados do Sistema	
Id_sap	Código de cadastro no sistema. Pode ser associado ao Código Alternativo
Qtd_etiquetas_k (k = 1, 2, ..., 12)	Quantidade de etiquetas impressas pelo cliente no k-ésimo mês anterior à intenção de cancelamento.

Fonte: Elaborado pelo autor

.

relativos à data de abertura do negócio de cancelamento, realizando um levantamento da atividade realizada pelo cliente nos 12 últimos meses que antecedem a expressão da intenção. Para os clientes que não possuem registro de abertura de negócio de cancelamento, o mês de junho de 2023 foi escolhido como referência para coleta dos dados de atividade dos meses anteriores. Dessa forma, foi possível maximizar o número de clientes que não possuem intenção de cancelar disponíveis, ao passo

que também houve garantia de que estes realmente não cancelaram em um momento futuro. Assim, o Quadro 2 apresenta a base de dados organizados e disponível para modelagem.

Quadro 2 – Base de dados para modelagem

Dado	**Descrição**
Resposta	Binário que possui valor 1 indicando que é um registro de cancelamento e 0 para um registro sem cancelamento
Data de Criacao Negocio	Data em que o negócio foi criado dentro do CRM.
Segmento	Segmento de atuação de do cliente
Origem	Qual fonte de prospecção trouxe o cliente para a base.
Codigo periodo	Qual a frequência de pagamento do cliente.
Lifetime	Diferença da data de primeiro ao último pagamento realizado pelo registro.
MRR	Valor da receita recorrente mensal do cliente.
Estado	Estado em que a empresa cliente está localizada.
Segmentação SC	Segmentação de nível de serviço definida pela equipe de sucesso do cliente.
Contagem Tickets	Número de tickets de atendimento de suporte abertos pelo registro.
Qtd_etiquetas_k (k = 1, 2, ..., 12)	Quantidade de etiquetas impressas pelo cliente no k-ésimo mês anterior à intenção de cancelamento.

Fonte: Elaborado pelo autor

.

Dessa forma, os dados disponíveis para o estudo são divididos em variáveis categóricas e numéricas. São cinco variáveis categóricas, sendo a primeira o "Segmento", que é subdividido em sete diferentes categorias: Produtor, Distribuidor, Produtor e Distribuidor, Indústria, Cooperativas, Varejo, e Outros, que é referente à qualquer segmento não mencionado anteriormente. A variável "Origem"possui três diferentes classificações, os clientes com origem *Outbound*, com prospecção ativa do time comercial, aqueles com origem *Inbound*, que ingressaram na base por meio de campanhas de marketing e os provenientes de indicação de outros clientes. A terceira variável é o estado, onde há o registro de todas as vinte e seis unidades federativas, com o acréscimo do distrito federal. A quarta variável se trata do "Código do Período", onde a exclusão dos clientes do tipo "bimestral"resultou em uma variável com quatro diferentes classes: anual, semestral, trimestral e mensal. Por fim, com a exclusão das filiais dos clientes, tres diferentes classes de "Segmentação SC"foram obtidas, os clientes de alta prioridade (classe A), os de média prioridade (classe B) e os de baixa prioridade (classe C). Por fim, variáveis numéricas são "Data de Criação do Negócio", o "Lifetime", que trata do tempo acumulado desde a aquisição do software, representando a fidelidade do cliente, o MRR que é referente ao ticket médio que o cliente possui, a contagem histórica de tíquetes, e as quantidades de etiquetas nos k meses anteriores ao período de análise.

3.2.2 Modelos de Predição

Diversos são os métodos de classificação para a aplicação em modelos de previsão. Neste trabalho serão utilizados: Regressão Logística, Random Forest, Support Vector Machine e XGBoost.

3.2.2.1 Regressão Logística

A regressão é um dos processos estatísticos para estimar como as variáveis estão relacionadas umas com as outras. A análise de regressão logística é um modelo de classificação estatística probabilística que é usado para classificação binária ou previsão binária de um valor categórico que depende de um ou mais parâmetros (AGRESTI, 2015).

De acordo com Zumel e Mount (2016), a modelagem por Regressão Logística permite prever probabilidades ou taxas, e seus coeficientes indicam como se espera que o evento se comporte, ou seja, essa técnica permite compreender como cada variável afeta o cenário em análise. O modelo de regressão logística considera a variável resposta como a variável dependente, e os parâmetros de entrada como variáveis independentes.

Segundo Fernandes *et al.* (2020), dado que a variável dependente em um modelo logístico é binária, a probabilidade prevista pelo modelo também será algum valor dentro do intervalo de zero a um. Assim, à medida que a variável independente se aproxima de zero, a probabilidade tende a se aproximar de zero. Por outro lado, o aumento do valor da variável independente também aumenta a probabilidade. A mesma analogia pode ser feita para um modelo de Regressão Logística Multivariada, ou seja, com mais de uma variável independente.

A curva logística apresentada na Figura 2 é a base deste método, e é traduzida como a conexão entre a variável dependente e as variáveis independentes, e apresenta uma percepção visual probabilidade obtida variando de zero a um (BATTISTI; SMOLSKI, 2019). O ajuste da curva denota o quanto a variável independente Como apresentado por Schneider (2016), justamente a abordagem de uma probabilidade em um intervalo contínuo é o que traz uma vantagem na aplicação do método ao possibilitar uma análise do nível de confiança da previsão com base no valor probabilístico encontrado.

3.2.2.2 *Support Vector Machine (SVM)*

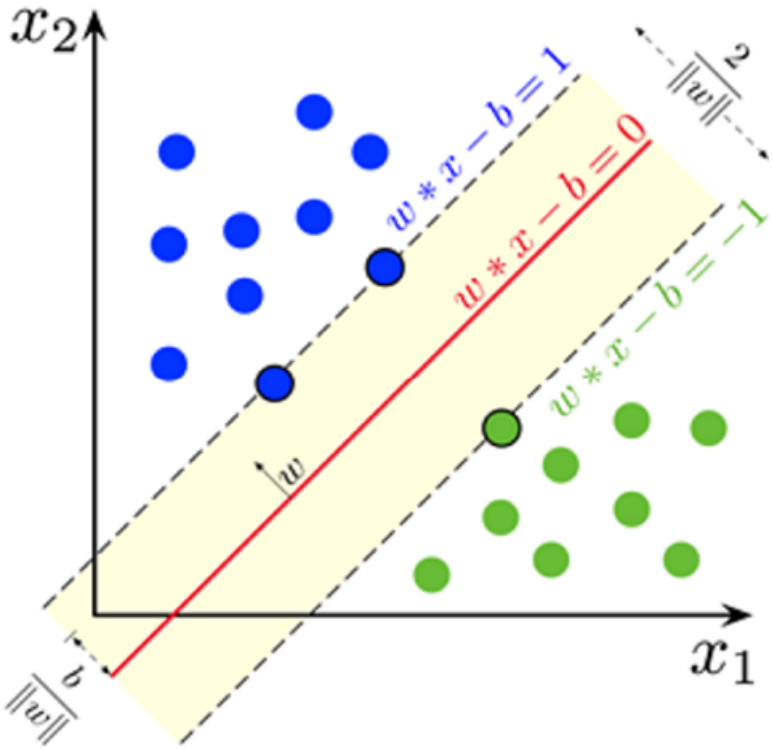
Máquinas de Vetores de Suporte (SVM) é uma técnica de sistema de aprendizado supervisionado que pode ser aplicado para resolver problemas de classificação Vapnik (1995) . Esta técnica se baseia na criação de uma dimensão a mais na base de dados, que é considerada a dimensão de classificação. Conforme apresentado por

Figura 2 – Curva Logística

Fonte: BATTISTI e Smolski (2019)

Schneider (2016), é criado um "hiperplano"de separação ideal, ou seja, uma fronteira de decisão que separa e classifica as observações com base nesta nova dimensão. Este hiperplano é criado analisando a distância entre as observações, e é criado ao encontrar os vetores de suporte e suas margens (ou módulos). (Figura 3).

Figura 3 – Maximização da margem entre os vetores de um modelo SVM



Fonte: Larhman (2018)

Apesar de ser um sistema desenvolvido mais recentemente, é um método consi-

derado muito promissor entre os estudiosos (SCHNEIDER, 2016). Luizon (2019) afirma que dentre as diversas vantagens da aplicação do SVM estão: uma boa capacidade de generalização, ou seja a facilidade de aplicar os resultados para novos dados de entrada; a robustez mesmo na utilização de dados de entrada que possuem grandes dimensões; convexividade da função objetivo, que possui apenas um mínimo global; e por fim uma teoria de criação embasada em aplicação de matemática e estatística.

3.2.2.3 Árvore de Decisão

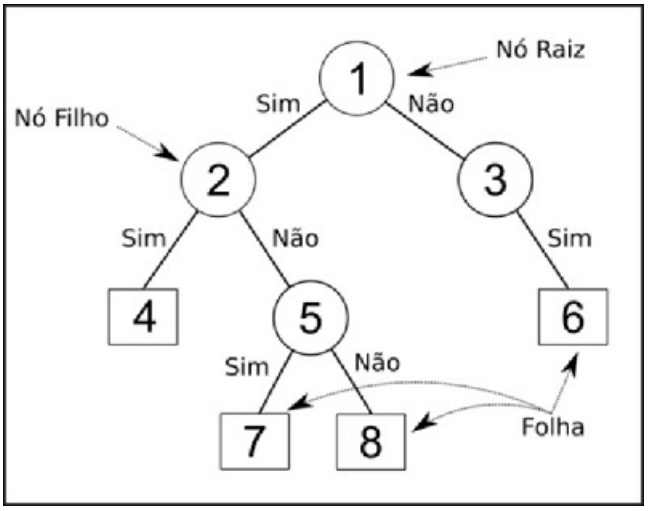
Um método popular para lidar com problemas de classificação binária é uma árvore de decisão. Breiman *et al.* (1984) definiu os métodos de Árvore de decisão como algoritmos de aprendizado de máquina supervisionado que podem ser utilizados tanto para classificação quanto para regressão. Höppner *et al.* (2020) afirmam que os modelos de árvores de decisão aplicados à classificação, além de serem fáceis de usar e fornecerem alta interpretabilidade, também possibilitam lidar com estruturas de dados complexas, como não linearidades, e possibilitam o uso de dados de entrada categóricos.

A Figura 4 apresenta uma representação diagramada do método. O funcionamento da ávore consiste em estabelecer “Nós de decisão” que se relacionam entre si por uma hierarquia. No início, o método procura pelo atributo que fornece a melhor informação para se tornar o nó raiz e divide a árvore em subárvores com base nesse nó. Da mesma forma, a subárvore é dividida com base em outras informações, ou seja, outras variáveis de entrada, e a divisão continua até que um nó folha (ou nó terminal) seja alcançado (HUANG; KECHADI; BUCKLEY, 2012). Assim, o principal papel do método é definir qual será a hierarquia entre os nós, e quais valores-chave serão utilizados nos atributos da base de dados para então atribuir um valor aos registros, ou encaixá-los em alguma classe específica. Apesar de ser um método muito popular, ele, assim como a maioria dos classificadores, pode enfrentar dificuldades ao lidar com conjuntos de dados de treinamento desequilibrados, e não raro acabam aprendendo padrões irregulares, tendenciando em direção à uma classe majoritária (*overfitting*), resultando em menor precisão para demais classes (SCHNEIDER, 2016).

3.2.2.3.1 Random Forest

Random Forest (Floresta Aleatória) é um método que combina várias árvores de decisão criadas a partir de amostras aleatórias dos dados de treinamento, processo chamado de *ensemble*. A previsão final é feita agregando as previsões de todas as árvores e retorna como resultado a média dentre as classes. Schneider (2016) discute que, ao utilizar diversas árvores com diferentes combinações de treinamento, é possível

Figura 4 – Representação da Árvore de Decisão



Fonte: **inproceedings**

reduzir a variância quando comparado com os modelos tradicionais de árvores de decisão, assim também diminuindo o *overfitting*.

A técnica de *ensemble* mais cumumente utilizado para a construção de árvores aleatórias é o *Bagging* (Figura 5). O *bagging* consiste na construção de subamostras do conjunto de dados inicial, de forma a criar árvores com diferentes entradas (SCHNEIDER, 2016).

Lauretto (2010) definiu a construção de um algoritmo de Random Forest nos seguintes passos:

i. Divisão aleatória do conjunto de dados em k conjuntos de teste e treinamento;

ii. Criação do modelo de árvore de classificação a partir do conjunto de treinamento e estimativa de erros a partir do conjunto de teste. Este procedimento é repetido uma quantidade k de vezes resultando nas árvores de classificação (de 1 até k);

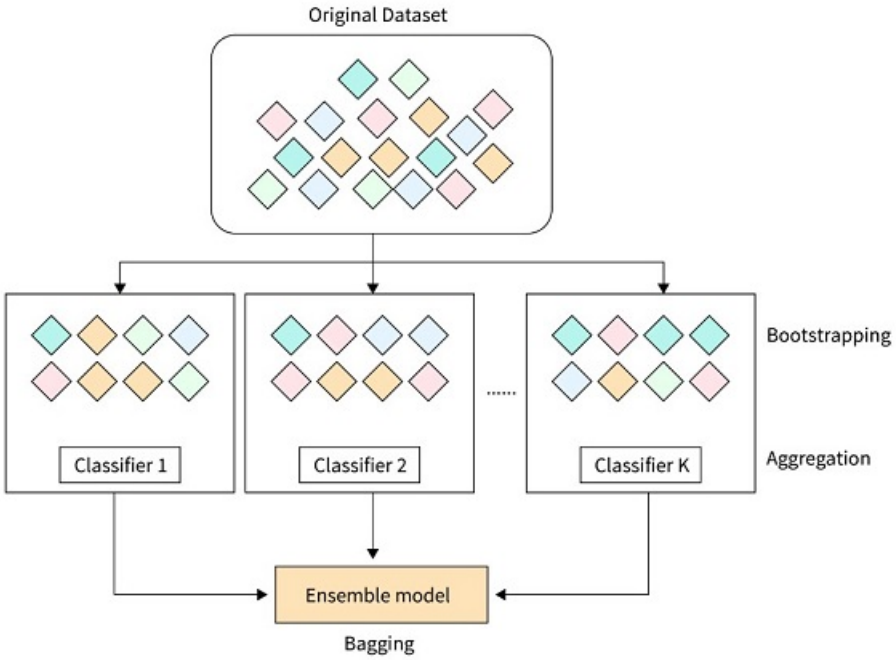
iii. Para cada elemento do conjunto de teste, compara-se a classe prevista com a classe verdadeira e a proporção de previsões erradas define o erro de classificação;

iv. Repete-se a divisão aleatória dos dados nos conjuntos teste e treinamento e o resultado final consiste nas taxas médias de erro das as árvores criadas.

3.2.2.3.2 XGBoost

Na seção 3.2.2.3.1 é apresentado o conceito de *ensemble* para caracterizar a técnica de *bagging*, necessária para a construção de modelos de *random forest*. Além do *bagging*, uma técnica muito utilizada para a construção de modelos preditivos é a técnica de *boosting* (Figura 6), que se trata da construção de um modelo que é treinado k vezes, sendo que a entrada do modelo k utiliza como base a saída do modelo k-1

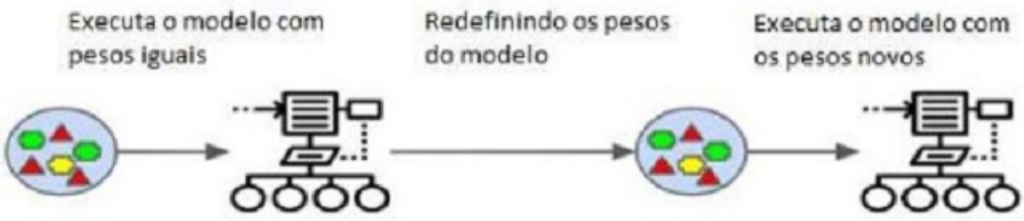
Figura 5 – Representação do método de *Bagging*



Fonte: GUPTA (2023)

(SCHNEIDER, 2016). Ou seja, enquanto o *bagging* realiza o treinamento de k modelos independentes, o *boosting* executa cada modelo identificando quais entradas são mais bem sucedidas e atribuindo um peso maior para elas na próxima iteração (SILVEIRA *et al.*, 2019).

Figura 6 – Funcionamento do *Boosting*



Fonte: Rogojan (2017)

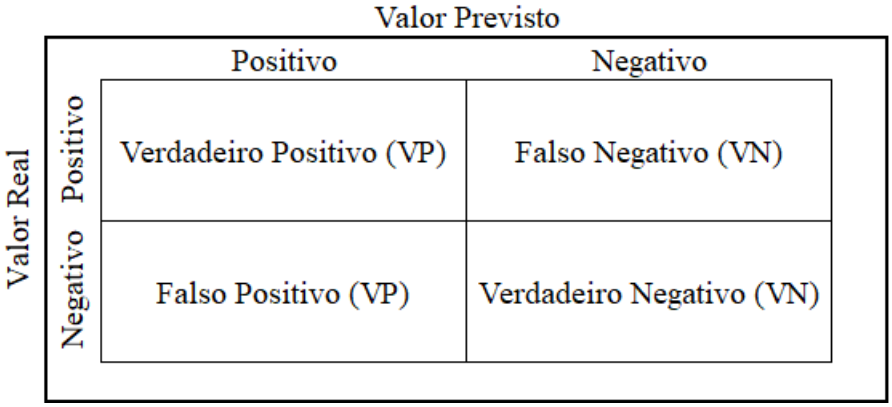
Dessa forma, o *XGBoost*, que pode ser entendido como *eXtreme Gradiente Boosting* se trata de um modelo que aplica a técnica de *boosting* para o treinamento. Segundo Lalwani *et al.* (2022), é um método que implementa o algoritmo de árvore de decisão somada a um aumento de gradiente. O aumento de gradiente segue uma abordagem em que novos modelos são usados para calcular o erro ou os resíduos do

modelo aplicado anteriormente e, em seguida, ambos são combinados para fazer a próxima iteração. Dessa foram, o *XGBoost* pode ser considerado um dos algoritmos mais eficazes para a construção de modelos de previsão, se destacando por, dentre outros fatores, evitar o overfitting de dados (DHALIWAL; NAHID; ABBAS, 2018). Ele tem a capacidade de elevar ao máximo o desempenho das máquinas usadas e tem produzido resultados que se destacaram quando comparados a outros modelos de aprendizado de máquina nos últimos anos (SILVEIRA *et al.*, 2019).

3.2.3 Avaliação dos Modelos

A maioria das métricas de desempenho para modelos de classificação com resposta binária são extraídas da matriz de confusão. Essa matriz tabula os números de classificações corretas e incorretas com base nas respostas produzidas pela previsão do modelo, quando comparadas com as informações reais armazenadas (HÖPPNER *et al.*, 2020). A Figura 7 esquematiza a matriz de confusão com a definição dos seus valores. Valores Verdadeiro Positivo (VP) indicam que quando aplicado nos dados de teste, o modelo preveu corretamente a intenção de cancelamento dos clientes, já valores Verdadeiro Negativo (VN) indicam que o modelo preveu corretamente os clientes que não tinham pretensão de cancelamento. Os valores Falso Negativo (FN) indicam que o cliente mostrou intenção de cancelamento, porém o modelo não identificou esta intenção, e por fim os valores Falso Positivo (FP) trata das previsões errôneas realizadas pelo modelo, que indicou uma intenção não existente.

Figura 7 – Matriz de Confusão.



Fonte: Elaborado pelo autor.

Com a matriz de confusão estruturada é possível calcular outras Métricas de avaliação como acurácia, precisão e a área sob a curva ROC (*Receiver Operating Characteristic*).

3.2.3.1 Acurácia

A acurácia é uma métrica para avaliar modelos de classificação, e nada mais é do que a fração das previsões que o modelo acertou. Esta métrica é encontrada dividindo os número total de previsões corretas (VP e VN) pelo número total de previsões realizadas, ou seja, o tamanho da base de teste.

Entretanto, em certas situações, o modelo pode alcançar uma alta acurácia, mas, mesmo assim, seu desempenho pode ser insatisfatório e ter uma baixa performance (CHEN *et al.*, 2020). Isso se dá pois nem sempre a base de dados está balanceada, ou seja, com a mesma proporção de valores verdadeiros positivos e negativos. Por exemplo, em uma base desbalanceada onde 95% dos casos são valores verdadeiros negativos, o modelo pode indicar todos as previsões como negativas e ainda acertaria 95%. Além disso, esta métrica também não diferencia falso positivo de falso negativo, assim limitando a análise caso essa informação seja relevante.

3.2.3.2 Score F1

Chen *et al.* (2020) apresentam o Score F1, também conhecido como *F-measure* ou *F-Score*, como uma métrica que relaciona a precisão com a revocação de um modelo. A precisão indica qual a proporção de positivos verdadeiros (VP) foi identificada corretamente se comparada com todos os valores indicados como positivos pelo modelo (VP + FP). Se trata de uma métrica que explica se o modelo está viciado a gerar um excesso de falsos positivos.

Já a revocação, também conhecida como sensibilidade, indica a proporção de positivos verdadeiros (VP) foi identificada corretamente se comparada com todos os valores realmente positivos (VP + FN). Portanto, se trata de uma métrica que busca entender o quanto foi acertado dentre os valores realmente positivos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Assim, é possível determinar o Score F1 a partir da média harmônica, ou seja, o dobro do produto dividido pela soma, entre precisão e sensibilidade. Uma característica importante dessa métrica é que, por se tratar de uma média harmônica, caso precisão ou sensibilidades sejam próximas de zero, o resultado do *F-score* também se aproximará de zero (CHEN *et al.*, 2020). Assim, é uma métrica que premite entender se o modelo é capaz tanto de acertar suas predições (precisão alta), quanto recuperar os exemplos da classe de interess (revocação alta), sendo um bom resumo da qualidade do modelo.

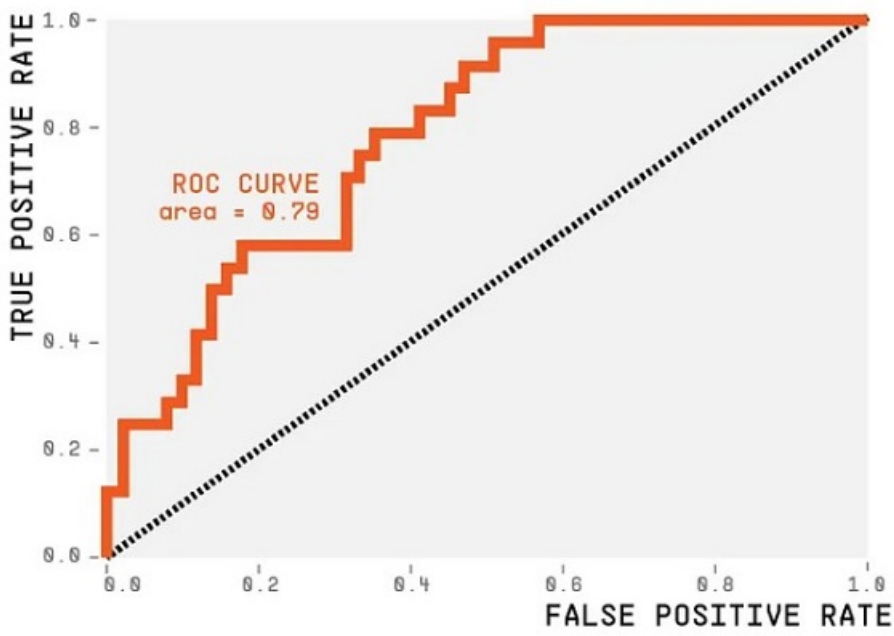
3.2.3.3 Curva ROC (*Receiver Operating Characteristic*)

Höppner *et al.* (2020) afirma que, para problemas de classificação binária, uma escolha popular para selecionar o modelo vencedor é a métrica ROC AUC (Figura 8),

devido à sua simplicidade e interpretação intuitiva. A partir dos valores VP e FP, é possível definir a Curva ROC, que faz uma relação entre sensibilidade e especificidade em diferentes pontos de corte de probabilidade. A especificidade é o oposto da revocação apresentada na seção 3.2.3.2. Se trata da divisão do número de verdadeiros negativos (VN) pela soma de todos os valores realmente negativos (VN + FP) e busca entender o quanto foi acertado dentre os valores realmente negativos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A relação da curva se dá com a taxa de verdadeiro positivo (sensibilidade) no eixo y e a taxa de falso positivo no eixo x (1 - especificidade). Esta métrica possibilita avaliar a qualidade de um modelo, uma vez que, quanto mais próxima a curva se encontrar do canto superior esquerdo, maior será a área sob a curva e, consequentemente, melhor será o desempenho do modelo (BRADLEY, 1996).

Figura 8 – Curva ROC



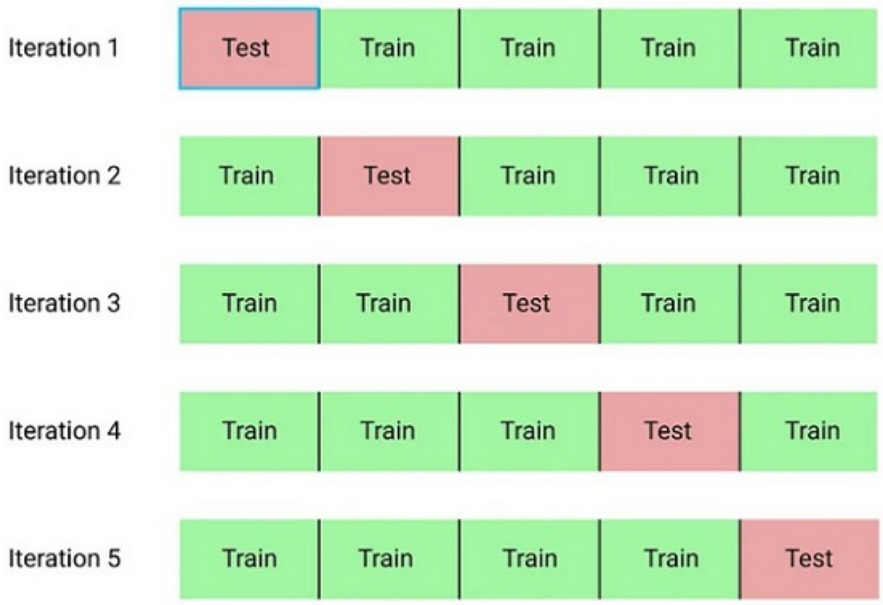
Fonte: Chen *et al.* (2020)

3.2.4 Validação Cruzada

Nem sempre encontrar boas métricas apresentadas anteriormente implica diretamente em bons resultados do modelo ao serem inseridos dados não vistos anteriormente. Segundo Hastie, Tibshirani e Friedman (2009), o *split* de dados é uma das formas que são aplicadas para validação de modelos de *machine learning*. Entretanto ela não é única, e alternativas como a validação cruzada com o uso de *k-folds* surgem para aumentar a confiança no modelo.

O *K-folds* é uma técnica de análise que substitui o *split* original dos dados. Aplicá-la consiste em dividir dos dados em K partes com apriximadamente a mesma quantidade de registros em cada uma delas. Assim o algoritmo realiza o treinamento dos dados K vezes, onde a cada iteração o subconjunto diferente é utilizado para teste e o restante para treino (Figura 9). Dessa forma o treinamento garante que utiliza todos os dados disponíveis em algum momento, e o resultado final será a média dos resultados encontrada pelas iterações (SHAIKH, 2018).

Figura 9 – Iterações da Validação Cruzada *K-Folds*



Fonte: Shaikh (2018)

3.2.5 Otimização de Hiperparâmetros

Os parâmetros e hiperparâmetros constituem elementos essenciais em modelos preditivos. Em algoritmos de aprendizado de máquina, os parâmetros são ajustados durante o processo de aprendizado, exercendo influência direta no desempenho do algoritmo. Já os hiperparâmetros são variáveis do algoritmo definidas antes do treinamento, representando características mais estruturais, como o número de neurônios em uma rede neural, o número de árvores de um random forest etc (MIURA, 2020). A escolha adequada dos hiperparâmetros é crucial para otimizar o desempenho do modelo preditivo, uma vez que decisões inadequadas podem comprometer sua utilidade ou distanciá-lo do desempenho ótimo.

Dessa forma, segundo Miura (2020) a tunagem de hiperparâmetros, refere-se ao processo de identificação da combinação mais eficaz de configurações para os hiperparâmetros de um modelo de aprendizado de máquina. A otimização de hiperparâmetros busca determinar a configuração ideal para melhorar métricas como precisão e generalização, entre outras. Métodos como busca em grade, busca aleatória e otimização bayesiana são comumente empregados para encontrar a combinação ótima que maximiza o desempenho do modelo.

3.3 ETAPAS METODOLÓGICAS

Neste tópico estão descritas as etapas metodológicas desenvolvidas no decorrer do projeto. As atividades desenvolvidas no trabalho seguem as etapas apresentado no quadro Quadro 3.

Quadro 3 – Etapas metodológicas

Etapa	Descrição
Levantamento de Razões do Cancelamento de Licença	Buscar na literatura as principais razões e estratégias para evitar o cancelamento de licenças
Levantamento dos Dados Relacionados ao Cancelamento	Analisar os dados para identificar de padrões e tendências, para então definir as variáveis de entrada e tratá-las.
Treinamento dos Modelos de Predição	Aplicar iterativamente diferentes conjuntos de dados de forma a identificar a melhor avaliação com os devidos hiperparâmetros.
Proposição de Ações para a Prevenção do Cancelamento	Com base na literatura, propor planos de ação para melhoria dos processos de retenção e aumentar a assertividade dos modelos.

Fonte: Elaborado pelo autor

.

3.3.1 Levantamento de Razões do Cancelamento de Licenças

O método de busca desempenha um papel crucial em revisões sistemáticas de literatura, pois determina quais estudos serão incluídos ou excluídos na análise. Uma estratégia eficaz deve ser sensível para identificar estudos pertinentes e específicos, evitando a inclusão de estudos irrelevantes ou duplicados. Além disso, é fundamental que a estratégia seja transparente, registrável e adaptável a diversas bases de dados, facilitando a replicação e verificação por outros pesquisadores. Recomenda-se a utilização de termos de busca padronizados e operadores booleanos para aprimorar a busca, sendo crucial a atualização da estratégia ao longo do processo de pesquisa.

A estratégia de pesquisa adotada para a realização do referencial teórico empregou as palavras-chave ("cancelamento"OR "churn"OR "data graphics") AND ("serviço"OR "service"OR "software"OR "SaaS"OR "licença de uso"OR "user licence") nas bases de dado do Google Acadêmico e Scopus, com o apoio da Central de Periódicos da Capes e do Repositório Institucional da UFSC.

A Scopus é amplamente reconhecida como uma das maiores bases de dados multidisciplinares, abrangendo diversas áreas de estudo. Para aprimorar a eficácia das buscas na Scopus, é aconselhável empregar operadores booleanos, truncamento e palavras-chave específicas relacionadas ao tema de pesquisa. A utilização da funcionalidade de pesquisa avançada na Scopus proporciona maior precisão, permitindo a aplicação de filtros para refinar os resultados.

O Google Acadêmico é uma plataforma de pesquisa desenvolvida pelo Google que fornece uma maneira fácil de pesquisar literatura acadêmica revisada por pares, incluindo artigos, teses, livros, conferências e patentes. Ele oferece uma ampla variedade de recursos para estudantes, pesquisadores e acadêmicos, permitindo o acesso a informações acadêmicas de diversas disciplinas.

A Central de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) é um serviço oferecido pelo governo brasileiro que disponibiliza acesso online a uma extensa coleção de periódicos científicos internacionais. Seu objetivo é fornecer à comunidade acadêmica brasileira recursos de alta qualidade para pesquisa e estudo. Já o Repositório Institucional da Universidade Federal de Santa Catarina (UFSC) é uma plataforma online que abriga, preserva e disponibiliza a produção acadêmica e científica da instituição. Ele reúne uma variedade de documentos, como teses, dissertações, artigos, relatórios técnicos, entre outros, produzidos por membros da comunidade acadêmica da UFSC. O principal objetivo do repositório é promover o acesso aberto ao conhecimento gerado na universidade, facilitando a disseminação e compartilhamento da produção intelectual. Estas duas bases serviram de suporte para a pesquisa realizada nas outras bases, dada a possibilidade de buscar referências nacionais de forma simplificada.

Dentre os resultados, foram identificados trabalhos recentes que possuem rela-

ção direta com o tema do estudo, e suas referências mais relevantes foram levantadas a fim utilizar de fontes mais primitivas com conceitos base. Ainda, foi utilizado de ferramentas de busca da web com palavras-chave específicas para coleta de informações pontuais que complementam a revisão da literatura.

3.3.2 Levantamento dos Dados Relacionados ao Cancelamento

Com os dados disponíveis definidos, é possível identificar padrões, tendências e relacionamentos entre as variáveis o que pode incluir a análise de estatísticas descritivas, visualização de dados e análise de correlação. Para a pesquisa em questão, é utilizada a biblioteca matplotlib para realizar a visualização dos dados e realizar as análises pertinentes. Dessa forma, é possível compreender os dados, identificar padrões e relações entre as informações e selecionar as variáveis relevantes para o modelo, além indicar quais são as transformações e tratamentos como normalização, codificação de variáveis categóricas e o tratamento de valores ausentes necessárias para aplicar o modelo.

Uma análise relevante a ser feita nesta etapa é referente às correlações presentes no modelo, com a verificação de multicolinearidade. O resultado deste procedimento apresenta uma Matriz de Correlação, que apresenta o quanto cada variável está associada com outra. Dessa forma, para os casos em que os coeficiêntes entre duas variáveis são muito altos, é provável que apenas uma dela possa ser inserida no modelo sem afetá-lo negativamente. Além desta análise, diversas outras podem ser aplicadas como o a avaliação das variáveis de entrada em comparação com a variável resposta, análise da curva normal de probabilidade e com a utilização de histogramas para variáveis contínuas e a análise de *outliers* com os *boxplot*.

A partir da análise exploratória de dados é possível selecionar as variáveis independentes que serão utilizadas como entrada do modelo. Com as variáveis selecionadas parte-se para o *split* dos dados de treino e teste. Essa técnica trata-se de separar a base de dados inicial em uma determinada proporção, para que uma parte deles sejam utilizadas para treinar o modelo, enquanto o restante é mantido isolado durante o treinamento para que, posteriormente, seja avaliado o desempenho do modelo com dados não vistos anteriormente. Para o presente estudo a base inicial será dividida de forma que 80% dos dados serão utilizados para treino e o restante para teste.

Com os dados selecionados, ainda se fazem necessárias algumas transformações para aplicação de modelos de classificação de resposta binária. Dentre as manipulações possíveis, está a criação de colunas dummies para as variáveis categóricas, possibilitando a aplicação dos modelos e a normalização de variáveis contínuas para balancear os resultados dos mesmos. A transformação de variáveis categóricas em variáveis dummy é uma prática que precisa ser realizada sempre que forem aplicados

modelos que aceitam apenas variáveis numéricas. Esse tipo de variável nada mais é que uma representação binária criada para expressar uma variável que possui duas ou mais categorias. A técnica consiste em separar uma coluna com duas características em duas colunas referentes a cada característica. Assim, é atribuído o valor 1 para os registros que possuem essas características nas referidas colunas e o valor -1 para o caso contrário. Quando se lida com uma variável que possui 3 ou mais categorias, é essencial criar n-1 dummies, uma vez que a última variável automaticamente implica a exclusão das demais.

Ainda, é necessário realizar normalização das variáveis numéricas, de forma a garantir que toda a base respeite o mesmo intervalo, assim garantindo que não haverá diferentes pesos entre elas. Por fim foi aplicada a técnica de *oversampling* nos dados base, de forma a reduzir o desequilíbrio no conjunto de dados que inicialmente estava de 3 para 1. Essa abordagem envolve aumentar a representação da classe minoritária, replicando instâncias ou gerando dados sintéticos dessa classe. O objetivo é equilibrar as proporções entre as classes, melhorando assim o desempenho do modelo.

3.3.3 Treinamento dos Modelos de Predição

Com as variáveis selecionadas e transformadas, e o *split* dos dados realizado, é possível construir modelos estatísticos ou de aprendizado de máquina adequados. Foram aplicados quatro diferentes algoritmos para treinamento do conjunto de dados: Regressão Logística, *Random Forest*, *Suport Vector Machine* e *XGBoost*, para encontrar aquele que melhor se adeque às variáveis de entrada mencionadas anteriormente. Cada um dos algoritmos foi treinado usando o mesmo conjunto de dados como base com o split tradicional e com a aplicação de validação cruzada, realizando um comparativo entre o previsto pelo modelo e o valor real existente.

Os modelos foram contruídos utilizando o Python dentro do Ambiente de Desenvolvimento Integrado (IDE) disponibilizado pela Google, o *Google Collaboratory*. A biblioteca sklearn foi utilizada para a construção dos modelos de Regressão Logística, Random Forest e SVM, além de ser a biblioteca que possibilita a coleta des métricas de avaliação dos modelos e a implantação dos mecanismos de validação cruzada. O XGBClassifier é o modelo presente dentro da biblioteca xgboost, direcionada especificamente para a construção destes tipos de modelo. Por fim, a otimização de hiperparâmetros foi feita utilizando a biblioteca Scikit Learn para cada um dos modelos, a fim de encontrar a combinação que gera os melhores resultados.

Para comparar os dados foram utilizadas cinco diferentes métricas: Acurácia, Precisão, Sensibilidade, F-Score e a Área sob a Curva ROC (AUC). Isso se dá pois uma métrica sozinha não é suficiente para avaliar a qualidade dos modelos, dessa forma a combinação das cinco minimiza a tomada de conclusões equivocadas.

3.3.4 Proposição de Ações para a Prevenção do Cancelamento

A proposição de ações direcionadas à prevenção do cancelamento de licenças é fundamentada nos resultados obtidos nos modelos preditivos, levando em consideração as variáveis críticas identificadas durante a análise, sustentada pela revisão de literatura apresentada no Capítulo 2. Dessa forma, foi possível levantar as possíveis ações que a empresa pode adotar, alinhadas à literatura e aos resultados com a proposição processos ainda não realizados pela empresa para prevenir o cancelamento de licenças.

Esta proposição utiliza de conceitos de Engenharia de Produção, com a ótica de gestão estratégica e com uma metodologia para traçar os planos de ação. Por fim, foram levantadas pontos de melhoria para processos na empresa que envolvem dados a fim de aumentar a variedade e direcionar o registro de informações com base em fatores da literatura que são importantes para entender e evitar o cancelamento de licenças, a fim de possibilitar futuros incrementos na modelagem.

4 PREVISÃO DO CANCELAMENTO DE LICENÇAS

Neste capítulo será apresentada a análise dos resultados do presente trabalho. A seção de levantamento de dados trata da Análise Exploratória de Dados que fundamenta o tópico seguinte de seleção e preparação dos dados. A segunda seção aborda os resultados encontrados com os modelos e a seção final apresenta propostas de ações para prevenção do cancelamento de licenças de uso de software.

4.1 LEVANTAMENTO DE DADOS

Nesta seção é apresentada a análise detalhada dados, com todas as análises realizadas para seleção das variáveis independentes do modelo. A seção finaliza apresentando as as técnicas necessárias para preparar os dados de entrada dos modelos preditivos.

4.1.1 Análise Exploratória de Dados

Para realizar a seleção e preparação dos dados é necessário primeiramente analisá-los a fim de identificação de informações que podem afetar o modelo tanto de forma positiva quanto negativa, e a biblioteca Pandas do Python permite a realização de análises estatísticas para este fim.

Como mencionado no capítulo anterior, existem 4435 obeservações na base inicial de dados. Entretanto, utilizando o comando ".info()"é possível ter um panorama geral dos dados, e ao analisar sua saída percebe-se existem colunas com dados faltantes. Dentre as informações faltantes, evidencia-se um grande número de dados faltantes nas colunas referentes às impressões de etiquetas (etiquetas_mes_k). Esta problemática se dá por incompatibilidade nas chaves primárias entre os dados disponíveis no sistema da empresa e em outras bases de dados. A Figura 10 apresenta a aplicação do comando, onde mostra que o numero de linhas não nulas para a informação de etiquetas varia de 3456 a 3714, ou seja, apenas 78% estão delas possuem estas informações.

Figura 10 – Comando “info()” aplicado no conjunto de dados inicial.

```
df_analise.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4435 entries, 0 to 4434
Data columns (total 23 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   resposta                     4435 non-null   object
 1   segmento                     4217 non-null   object
 2   origem                       4435 non-null   object
 3   cod_periodo_menor            4435 non-null   object
 4   lifetime                     4435 non-null   int64
 5   mrr                          4435 non-null   object
 6   vida_financeiro              4435 non-null   int64
 7   data_negocio                 4435 non-null   datetime64[ns]
 8   contagem_tickets             4435 non-null   float64
 9   Estado                       4275 non-null   object
 10  Segmentação SC - Rastreador  4275 non-null   object
 11  etiqueta_mes_1               3714 non-null   object
 12  etiqueta_mes_2               3714 non-null   object
 13  etiqueta_mes_3               3714 non-null   object
 14  etiqueta_mes_4               3714 non-null   object
 15  etiqueta_mes_5               3714 non-null   object
 16  etiqueta_mes_6               3714 non-null   object
 17  etiqueta_mes_7               3714 non-null   object
 18  etiqueta_mes_8               3672 non-null   object
 19  etiqueta_mes_9               3620 non-null   object
 20  etiqueta_mes_10              3555 non-null   object
 21  etiqueta_mes_11              3501 non-null   object
 22  etiqueta_mes_12              3456 non-null   object
dtypes: datetime64[ns](1), float64(1), int64(2), object(19)
memory usage: 797.0+ KB
```

Fonte: Elaborado pelo autor.

A partir desta informação, viu-se interessante o entendimento da distribuição da variável resposta na base de dados (Figura 11). Dentre todo o conjunto de dados, 71% representa variáveis do tipo "negativa"para o cancelamento de licenças de uso, com o complementar referente ao positivo. Dado que datasets com mais de 50% das entradas pertencendo a uma única classe são considerados desbalanceados, favorecendo a classe majoritária. Para melhorar o desempenho do algoritmo de machine learning será necessário realizar uma técnica de amostragem, e neste estudo foi utilizado o *oversampling*, dado que existe uma limitação no volume de dados disponível.

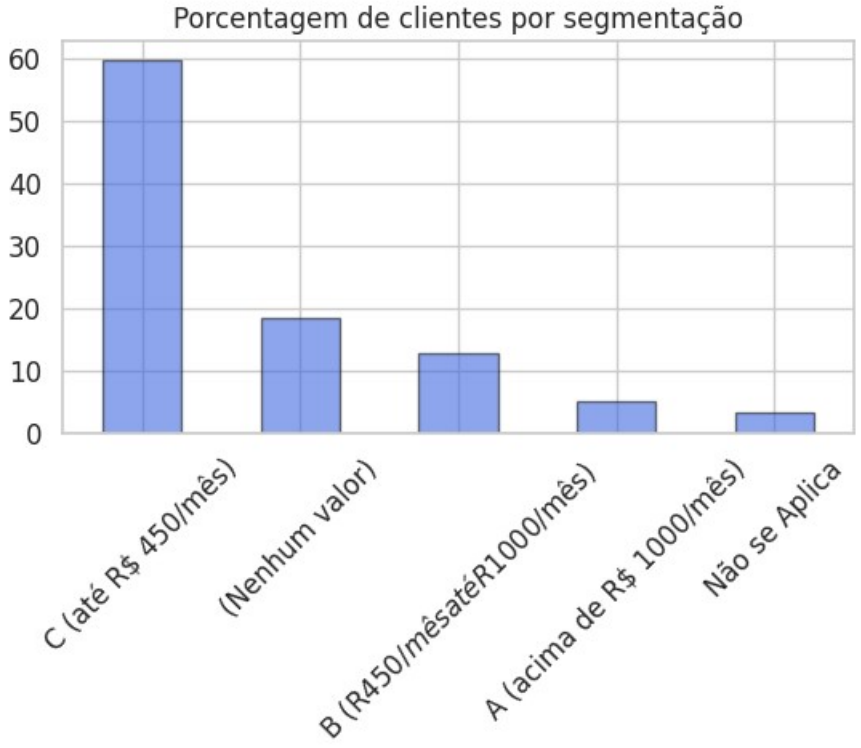
Figura 11 – Comportamento da Variável Resposta

Fonte: Elaborado pelo autor.

Uma análise importante a ser realizada é a análise das variáveis categóricas que servirão de entrada no modelo. Uma forma de realizar esta análise é com a aplicação de gráficos de barra a fim de entender o perfil de cada uma das variáveis independentes, além de encontrar falhas e oportunidades de manipulações dos dados a fim de aumentar a acertividade do modelo.

A Figura 12 mostra uma análise da variável "Segmentação SC". Como mencionado anteriormente, esta variável aborda a presença de um especialista com atendimento personalizado, e em diversas situações ela é definida com base no ticket médio do cliente. Entretanto, não apenas clientes com tickets altos possuem especialistas, clientes estratégicos, clientes que contratam outras ferramentas e clientes que possuem uma relação duradoura também são alvo de especialistas. Na análise, percebe-se uma grande quantidade de clientes com "Nenhum valor"atribuído. Isso se dá pela falta de atualização de dados na fonte de coleta. Dessa forma, estes dados podem ser corrigidos com a informação de "mrr", de forma a não retirá-los da base, dada a limitação no volume de dados.

Figura 12 – Presença de dados incompletos em Segmentação SC

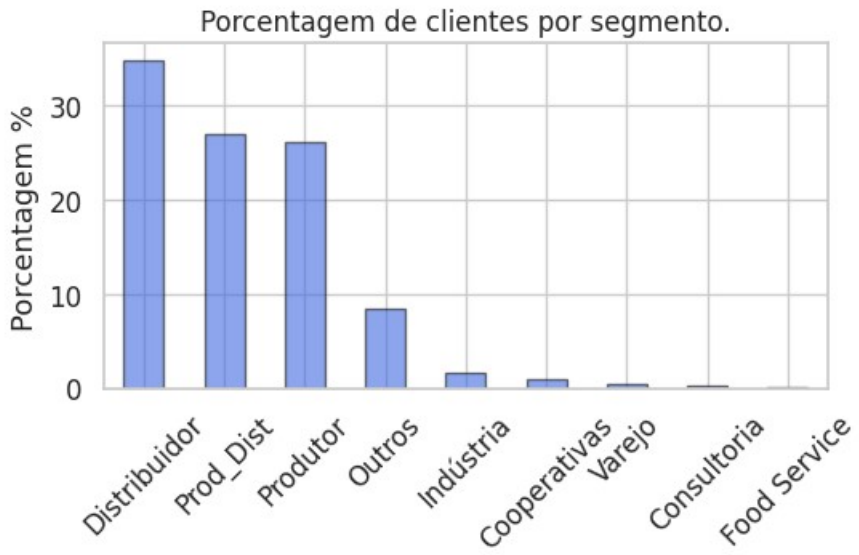


Fonte: Elaborado pelo autor.

Entretanto, ainda na Figura 12 observa-se a presença da categoria "Não Aplicável. Essa categoria refere-se a registros de filiais de outras empresas já presentes no modelo, dessa forma estes dados trazem informações redundantes e que deve ser filtrados para a execução do modelo.

Analisando a variável segmento também observa-se um ponto de atenção. As categorias "Distribuidor", "Prod_Dist"e "Produtor"são bem distribuídas no conjunto de dados (Figura 13). Entretanto, outras categorias menos relevantes aumentam a granularidade da variável, e não trazem informações suficientes para modelagem. Dessa forma, aproveitando a existência da categoria "Outros", as categorias "Indústria", "Cooperativas", Varejo", "Consultoria"e "Food Service"foram agrupadas de forma a representarem um grupo em comum, à jusante da cadeia produtiva de alimentos.

Figura 13 – Distribuição dos dados por segmento

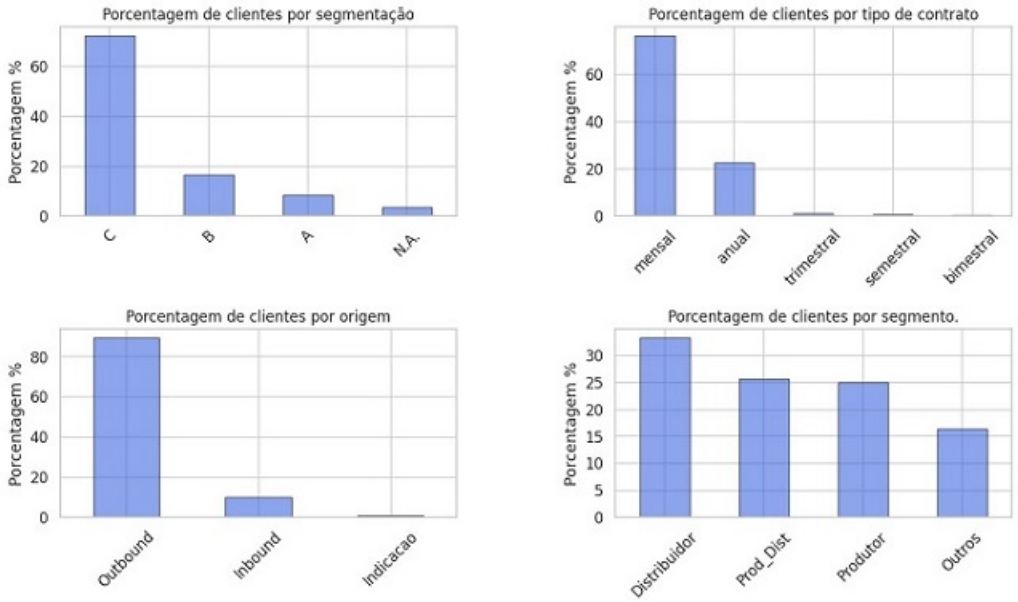


Fonte: Elaborado pelo autor.

Dessa forma, é possível analisar todas as variáveis categóricas em um bloco, como mostra a Figura 14. A variável referente ao tipo de período de pagamento registrado no contrato apresenta que a maior porção dos registros se tratam de empresas que realizam o pagamento mensal ou anual. Ao analisar os registros referentes aos demais períodos, identificou-se que atualmente não são mais realizados contratos do tipo "bimestral", assim, da mesma forma que o "Não Aplicável"para a Segmentação SC, estes dados serão retirados do modelo de forma a aumentar sua classificação.

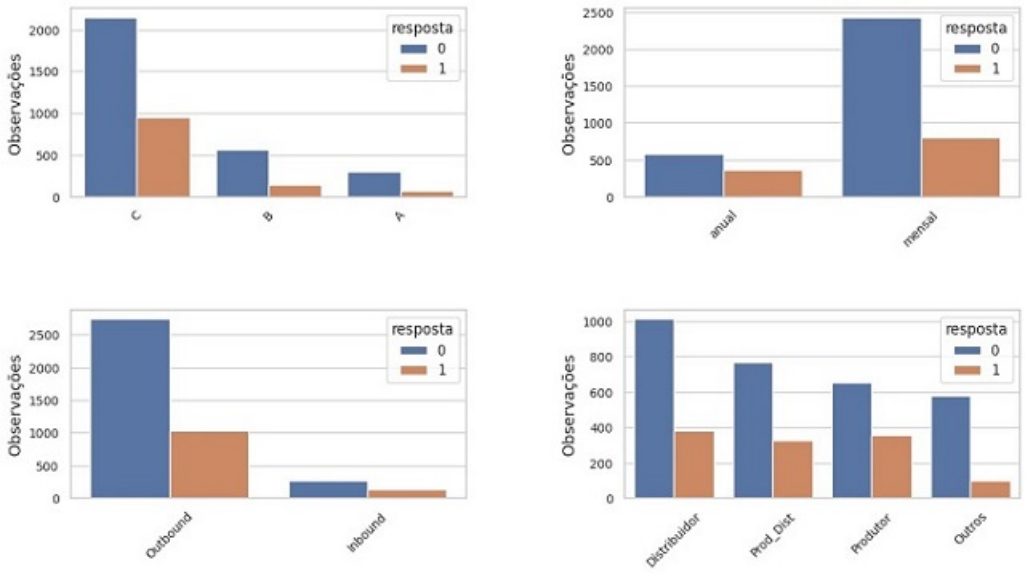
Ainda, precebeu-se que a maioria dos dados referente à origem são do tipo *outbound*, sendo que clientes que possuem a origem via indicação não possuem uma frequência suficiente, dessa forma, essa categoria de origem será retirada da modelagem. Por fim, é possível analisar a distribuição da resposta em relação às variáveis categóricas (Figura 15). A variável resposta possui uma distribuição uniforme na maioria das variáveis, ou seja, apesar do desbalanceamento dos dados entre churn e não churn, as proporções são constantes na maioria das categorias indicando a possibilidade de *oversampling*.

Figura 14 – Variáveis categóricas: gráficos de barra



Fonte: Elaborado pelo autor.

Figura 15 – Variáveis categóricas: gráficos de barra por resposta



Fonte: Elaborado pelo autor.

Outra análise feita na base de dados foi referente às variáveis numéricas. O comando ".describe()"traz informações estatísticas referentes às colunas quantitativas existentes na base, como média, desvio padrão, valores mínimo e máximo além do último valor dos quartis para as variáveis numéricas (Figura 16). Neste relatório é possível identificar uma anormalidade em "contagem_tickets", onde até o terceiro quartil seu valor máximo é de 5 tíquetes, mas o valor máximo para esta variável dentro da base de dados é de 848 tíquetes. Assim, preparando o boxplot para esta variável, entende-se como *outliers* os registros que possuem 12 ou mais tíquetes registrados em seu histórico (Figura 17).

Figura 16 – Comando “.describe()” para a análise estatística

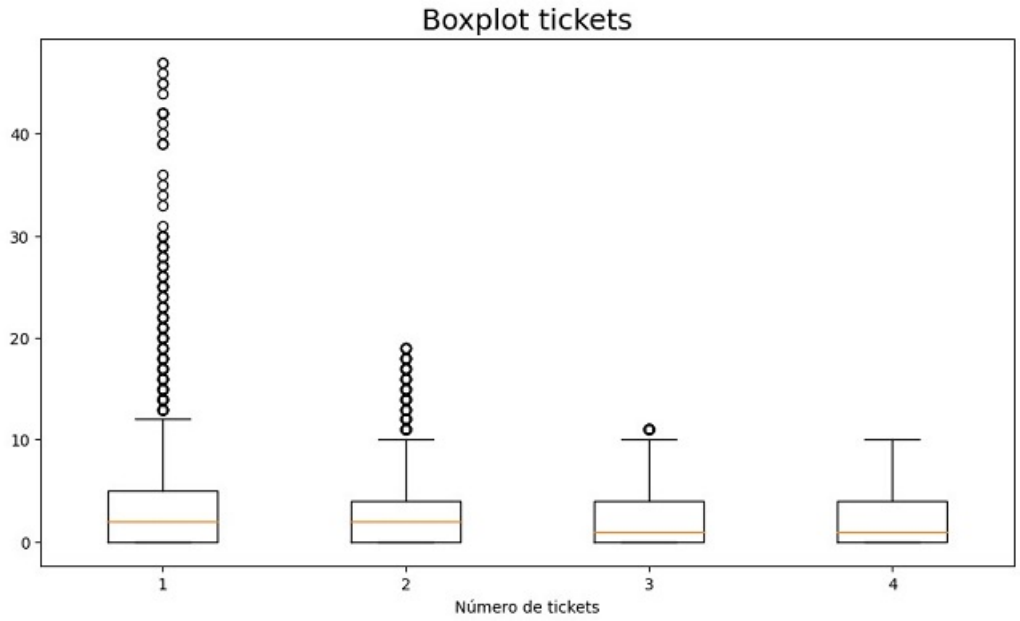
	lifetime	vida_financeiro	data_negocio	contagem_tickets
count	4435.000000	4435.000000	4435	4435.000000
mean	33.824803	31.136640	2023-02-14 00:45:27.395716096	4.079594
min	1.000000	0.000000	2021-07-01 00:00:00	0.000000
25%	12.000000	10.000000	2023-02-01 00:00:00	0.000000
50%	32.000000	27.000000	2023-06-01 00:00:00	2.000000
75%	50.000000	48.000000	2023-06-01 00:00:00	5.000000
max	87.000000	86.000000	2023-07-01 00:00:00	848.000000
std	24.842485	25.174794	NaN	15.551266

Fonte: Elaborado pelo autor.

Dado que a base já possui um tamanho limitado, viu-se a necessidade de entender qual o volume destes outliars, pois apenas retirá-los pode acarretar em prejuízo de treinamento dada a falta de volume. A Figura 18 apresenta essa análise, onde observou-se que estes dados totalizam 245 observações, sendo 65 do tipo positivo para a abertura de negócio de cancelamento e 180 do tipo negativo. De forma a não prejudicar a análise, optou-se por realizar uma suavização desta *feature*, de forma a não excluir estes dados da análise.

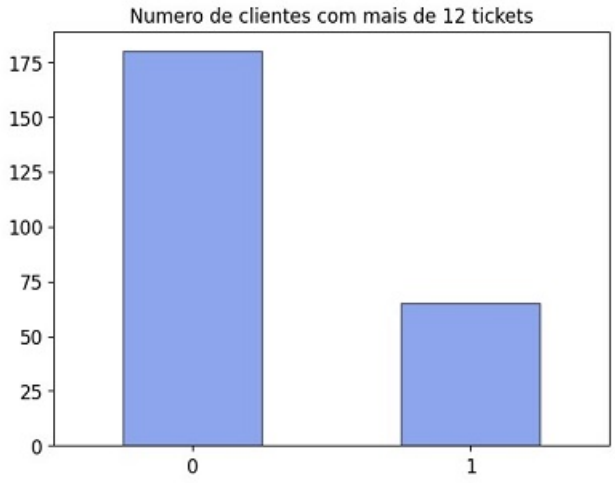
Ainda se tratando das variáveis numéricas, é possível analisar a relação entre elas. A Figura 19 apresenta um modelo que sumariza a média do tícket dos clientes com o período em que contrataram a solução. Percebe-se a existência de uma leve tendência de crescimento entre os dados, mas não o suficiente para afirmar uma relação direta entre elas.

Figura 17 – Boxplot com análise para 50, 20, 12 e 11 tíquetes, respectivamente



Fonte: Elaborado pelo autor.

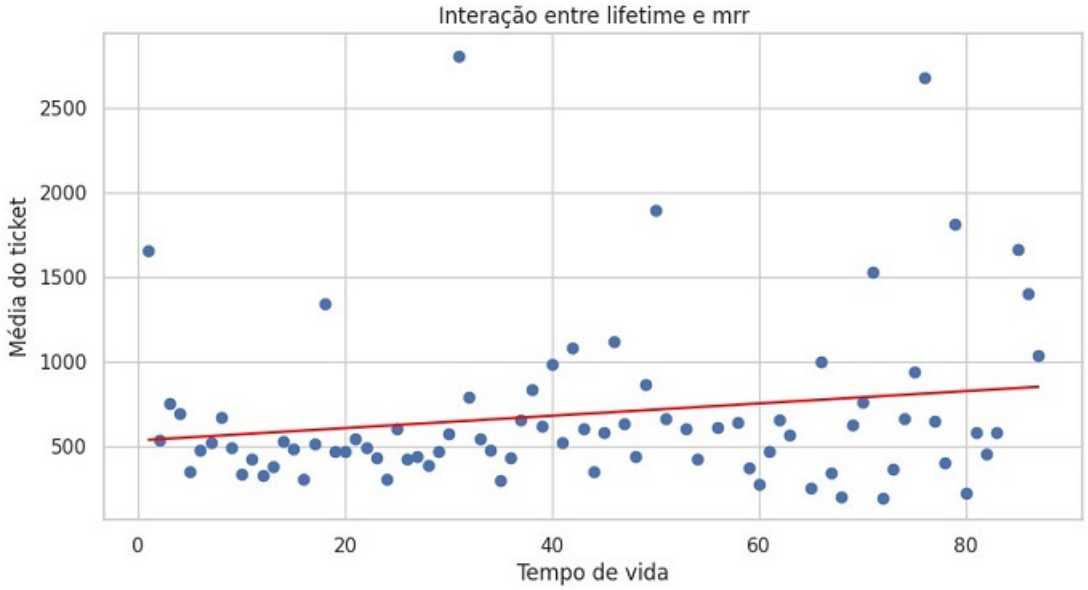
Figura 18 – Número de clientes com mais de 12 tíquetes



Fonte: Elaborado pelo autor.

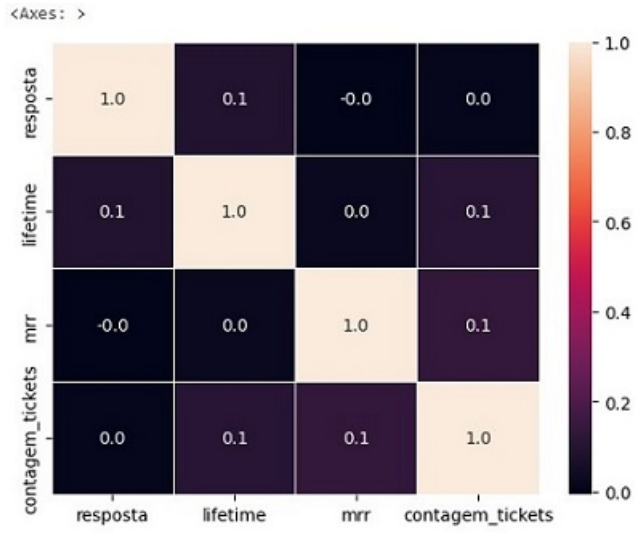
Dessa forma, foi realizada uma análise de correlação com todas as variáveis, incluindo o cliclo de vida e o ticket médio, para garantir que não seriam criados vícios nos resultados da modelagem que levam ao *overfitting*. A análise de correlaçao é apresentada na (Figura 20).

Figura 19 – Interação entre lifetime e MRR



Fonte: Elaborado pelo autor.

Figura 20 – Análise da correlação entre as variáveis independentes



Fonte: Elaborado pelo autor.

Por fim, após retirar e corrigir todos os dados, foi realizada uma última verificação dos dados faltantes com o comando ".isnull()", que reforçou a defasagem no levantamento dos dades de etiquetas que devem ser ignorados (Figura 21).

Figura 21 – Levantamento de valores nulos por *feature*

```
resposta                          0
segmento                          0
origem                            0
cod_periodo_menor                 0
lifetime                          0
mrr                               0
vida_financeiro                   0
data_negocio                      0
contagem_tickets                  0
Estado                          159
Segmentação SC - Rastreador       0
etiqueta_mes_1                  662
etiqueta_mes_2                  662
etiqueta_mes_3                  662
etiqueta_mes_4                  662
etiqueta_mes_5                  662
etiqueta_mes_6                  662
etiqueta_mes_7                  662
etiqueta_mes_8                  704
etiqueta_mes_9                  752
etiqueta_mes_10                 815
etiqueta_mes_11                 867
etiqueta_mes_12                 910
dtype: int64
```

Fonte: Elaborado pelo autor.

4.1.2 Seleção e Perparação dos Dados

As informações levantadas na seção anterior direcionaram filtragens necessárias para a implementação dos modelos. Após as filtragens e manipulações necessárias realizadas, foram escolhidas as variáveis para servirem de entrada do modelo. A base de dados final contém 3163 observações, e as variáveis selecionadas podem ser visualizadas na Figura 22.

Figura 22 – *Features* do modelo

```
Data columns (total 22 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   resposta                     3163 non-null   float64
 1   segmento                     3163 non-null   object
 2   origem                       3163 non-null   object
 3   cod_periodo_menor            3163 non-null   object
 4   lifetime                     3163 non-null   float64
 5   mrr                          3163 non-null   float64
 6   data_negocio                 3163 non-null   object
 7   contagem_tickets             3163 non-null   float64
 8   Estado                       3163 non-null   object
 9   Segmentação SC - Rastreador  3163 non-null   object
 10  etiqueta_mes_1               3163 non-null   float64
 11  etiqueta_mes_2               3163 non-null   float64
 12  etiqueta_mes_3               3163 non-null   float64
 13  etiqueta_mes_4               3163 non-null   float64
 14  etiqueta_mes_5               3163 non-null   float64
 15  etiqueta_mes_6               3163 non-null   float64
 16  etiqueta_mes_7               3163 non-null   float64
 17  etiqueta_mes_8               3163 non-null   float64
 18  etiqueta_mes_9               3163 non-null   float64
 19  etiqueta_mes_10              3163 non-null   float64
 20  etiqueta_mes_11              3163 non-null   float64
 21  etiqueta_mes_12              3163 non-null   float64
dtypes: float64(16), object(6)
memory usage: 568.4+ KB
```

Fonte: Elaborado pelo autor.

A Figura 23 apresenta o código utilizado para a criação das colunas *dummies*. Ele foi aplicado nas variáveis "segmento", "origem", "cod_periodo_menor"e "Estado". Ao final do processo, a base que anteriormente possuía 22 colunas, passa a ter 51, com o acréscimo das colunas *dummies.*

Figura 23 – Criação das variáveis *dummies*

```
cat_vars = ['segmento',
            'origem',
            'cod_periodo_menor',
            'Estado']

for i in cat_vars:
    if (df_train[i].dtype == np.str or df_train[i].dtype == np.object):
        for j in df_train[i].unique():
            df_train[i+'_'+j] = np.where(df_train[i] == j,1,-1)
        remove.append(i)
df_train = df_train.drop(remove, axis=1)
```

Fonte: Elaborado pelo autor.

A Figura 24 apresenta o código utilizado para realizar o *split* inicial dos dados, com 80% dos dados direcionados para treinamento do modelo e os 20% para a realização de testes e coleta das métricas. Ainda, na figura é apresentado o código necessário para realizar a normalização dos dados, que para o presente modelo estarão entre 0 e 1.

Figura 24 – *Split* e normalização dos dados

```
# Separando o dataframe (X) da variável resposta (y)
y = df_train['resposta'].values
X = df_train.drop(columns = ['resposta'])

# Normalização do dataframe
from sklearn.preprocessing import MinMaxScaler
features = X.columns.values
scaler = MinMaxScaler(feature_range = (0,1))
scaler.fit(X)
X = pd.DataFrame(scaler.transform(X))
X.columns = features
```

Fonte: Elaborado pelo autor.

Por fim, como mencionado anteriormente, para balanceamento dos dados é necessário realizar um *oversampling* nos dados de treino. O *script* para a realização destre procedimento é apresentado na Figura 25. Dessa forma, após a realizada de todas as transformações, a base final de dados de treino (X_train) possui um total 3944 linhas e 50 colunas com valores variando entre 0 e 1, sendo quinze com variáveis numéricas normalizadas, e trinta e cinco colunas *dummies*.

Figura 25 – Aplicação do *oversampling*.

```
# Criando o split tradicional
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=101)

# criando uma instância do SMOTE
smote = SMOTE()

# Oversampling
X_resampled, y_resampled = smote.fit_resample(X_train, y_train)

X_train = X_resampled
y_train = y_resampled
```

Fonte: Elaborado pelo autor.

4.2 RESULTADOS DOS MODELOS DE PREDIÇÃO

Com as *features* definidas e o *dataframe* de teste finalizado, é possível ajustar os hiperparâmetros para obter resultados mais precisos. A biblioteca sklearn oferece a função "GridsearchCV", que viabiliza a avaliação do modelo em um conjunto predefinido de hiperparâmetros. Para o modelo de regressão logística, apenas dois hiperparâmetros foram otimizados. O hiperparâmetro "C", responsável por controlar o quão tolerante a erros de classificação será o modelo treinado, obteve o melhor valor de 1000, e o parâmetro "*penalty*", que adiciona uma penalidade a fim de reduzir o *overfitting* obteve "l2"como melhor valor. Dessa forma, foi possível treinar o modelo, e os resultados estão apresentados na Tabela 1.

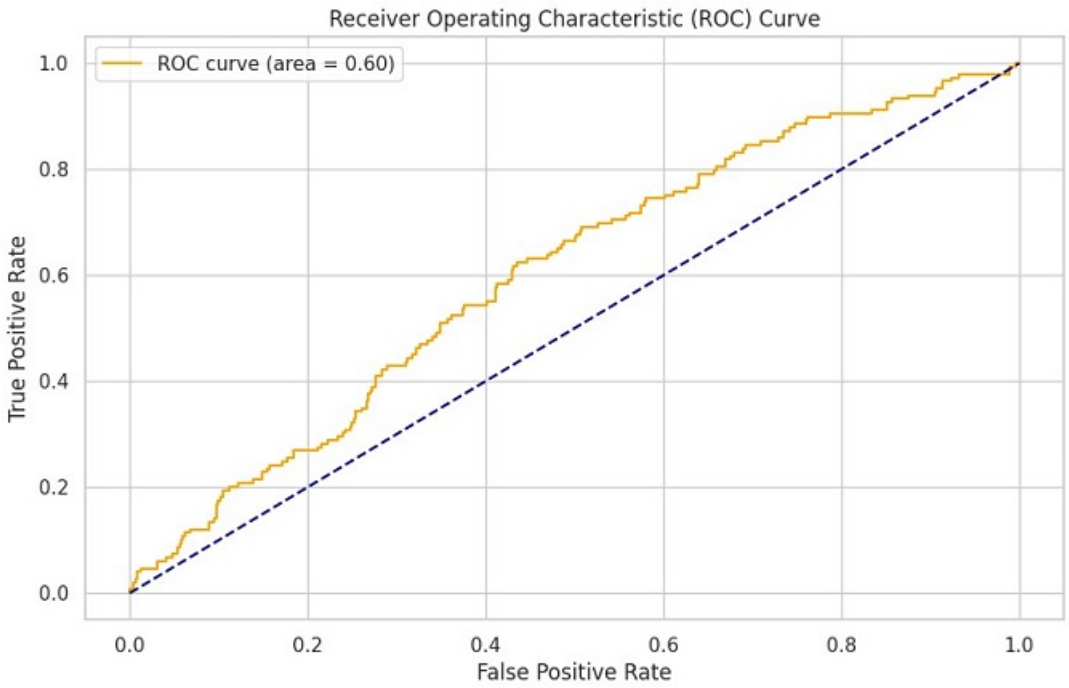
Tabela 1 – Resultado Modelo Regressão Logística

Métrica	**Split Tradicional**	**Validação Cruzada**
Acurácia	60.51%	77.71%
Precisão	27.36%	31.41%
Sensibilidade	53.69%	2.97%
F-score	39.02%	6.21%

Fonte: Elaborado pelo autor.

Analisando os resultados, apesar de alcançar uma acurácia relativamente alta quando analisada a validação cruzada, uma sensibilidade e um F-score baixo quando aplicada essa técnica. Isso significa que apesar de acertar uma boa proporção dos valores, o modelo está viciado em apontar as previsões como negativas, ou seja, há uma baixa assertividade para verdadeiros positivos, sendo adequado a prever o *churn*. Ainda o modelo alcançou uma área sob a curva ROC de 60% (Figura 26), com uma baixa capacidade de distinguir entre as classes.

Figura 26 – Curva ROC (AUC) Regressão Logística



Fonte: Elaborado pelo autor.

Analisando os resultados referentes ao algoritmo de *Random Forest*, a Tabela 2 exibe os hiperparâmetros testados na tunagem, com sua descrição, os valores de teste utilizados e o melhor valor encontrado.

Com os hiperparâmetros definidos, foi realizado o treinamento do modelo de *Random Forest*. Os resultado apresentados na Tabela 3 apontam que a modelagem foi razoavelmente adequada à predição do *churn*, possuindo uma acurácia de 82% com a validação cruzada. Apesar de possuir uma sensibilidade baixa (abaixo de 50%), o modelo acertou uma proporção relevante de casos de *churn* quando comparado ao modelo de regressão logística.

Tabela 2 – Tunagem de hiperparâmetros Random Forest

Hiperparâmetro	Descrição	Valores Testados	Melhor Valor
min_samples_split	Quantidade mínima de amostras que um nó deve conter para se dividir em outros nós.	[2, 5]	2
min_samples_leaf	Quantidade mínima de amostras que um nó deve conter após ser dividido.	[1, 3]	1
max_depth	Profundidade máxima da árvore. Árvores profundas correm o risco de apresentar overfitting.	[10, 50, 100]	50
n_estimatores	Número de árvores usadas no modelo.	[50, 100, 500]	500
max_features	Direciona número de features a serem considerados para fazer a melhor divisão das árvores	['auto', 'sqrt', 'log2']	'log2'
bootstrap	Indica a necessidade de realização de bagging adicional.	[True, False]	False

Fonte: Elaborado pelo autor.

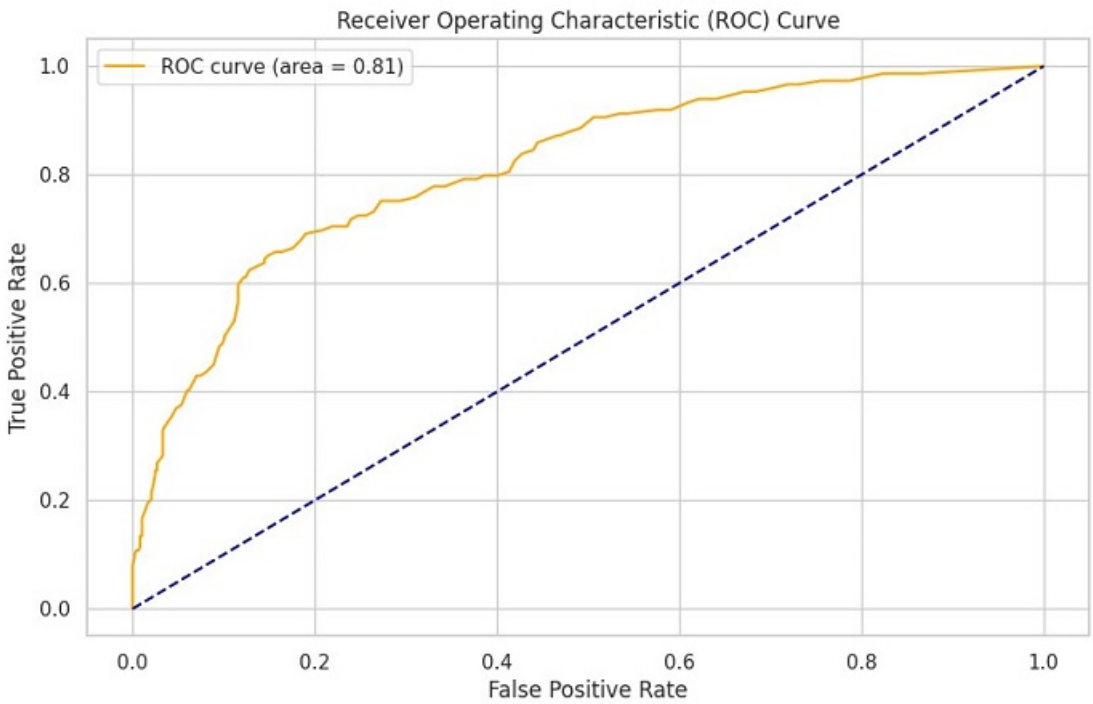
Tabela 3 – Resultado Modelo Random Forest

Métrica	Split Tradicional	Validação Cruzada
Acurácia	81.36%	82.42%
Precisão	46.30%	60.91%
Sensibilidade	62.42%	46.16%
F-score	61.18%	54.00%

Fonte: Elaborado pelo autor.

Ainda, o modelo teve uma ótima eficiência em distinguir as classes positivas (*churn*) e negativas (não *churn*), obtendo uma área sob a curva ROC de 81.38% (Figura 27).

Figura 27 – Curva ROC (AUC) Random Forest



Fonte: Elaborado pelo autor.

Se tratando do modelo de máquinas de vetores de suporte, são 3 parâmetros relevantes: "C", já apresentado anteriormente, o "*gamma*", que se trata de pesos que são introduzidos para a distância entre amostras, dando maior ou menor importância à amostras distantes ou próximas da fronteira de decisão e o "kernel", que determina se o modelo será linear ou não linear com a utilização das funções de Kernel. A Tabela 4 apresenta os valores testados e o melhor valor encontrado para este modelo preditivo. Dessa forma, foi possível treinar o modelo com as variáveis de entrada e os hiperparâmetros definidos, encontrando os resultados apresentados na Tabela 5.

Tabela 4 – Tunagem de hiperparâmetros *Support Vector Machine*

Hiperparâmetro	**Valores Testados**	**Melhor Valor**
C	[0.1, 1, 10, 100, 1000]	1000
gamma	[1, 0.1, 0.01, 0.001, 0.0001]	1
kernel	['rbf', 'linear']	'rbf'

Fonte: Elaborado pelo autor.

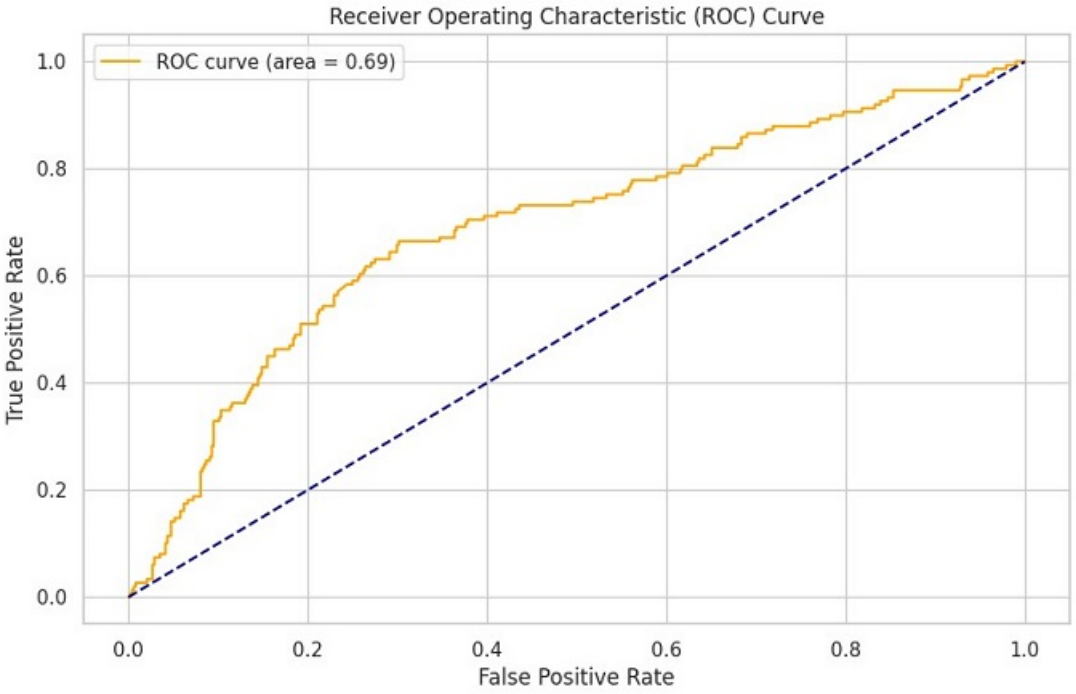
Tabela 5 – Resultado Modelo *Support Vector Machine*

Métrica	**Split Tradicional**	**Validação Cruzada**
Acurácia	70.93%	75.59%
Precisão	34.27%	34.60%
Sensibilidade	59.06%	27.27%
F-score	48.88%	32.12%

Fonte: Elaborado pelo autor.

Para o SVM o resultado foi mediano, superior ao do modelo de regressão logística, porém inferior ao modelo de *Random Forest*. Este modelo também apresentou dificuldade em prever os casos positivos de churn, apresentando uma sensibilidade baixa, e não conseguiu distinguir de forma satisfatória as classes de predição obtendo uma área sob a curva ROC de 69.14% (Figura 28) .

Figura 28 – Curva ROC (AUC) Suport Vector Machine



Fonte: Elaborado pelo autor.

Por fim a Tabela 6 apresenta a tunagem do último algoritmo, o Extreme Gradiente Boost. O algoritmo possui hiperparâmetros semelhantes ao do *Random Forest*, com a adição de parâmetros direcionados à técnica de *boosting*.

Tabela 6 – Tunagem de hiperparâmetros XGBoost

Hiperparâmetro	**Descrição**	**Valores Testados**	**Melhor Valor**
colsample_bytree	Fração das colunas que serão utilizadas em cada árvore.	[0.5, 0.6, 0.7, 0.8, 0.9]	0.9
learning_rate	Diminui o peso em cada etapa para evitar overfitting	[0.01, 0.1, 0.3]	0.1
max_depth	Profundidade máxima da árvore. Árvores profundas correm o risco de apresentar overfitting.	[3, 5, 7, 9, 11]	9
min_child_weight	Indica a soma mínima dos pesos de todas as observações de uma mesma folha.	[2, 3, 5, 7, 9]	2
n_estimatores	Número de árvores usadas no modelo.	[100, 200]	200
subsample	Indica a fração de observações usadas como amostra para cada árvore.	[0.6, 0.8, 1]	0.8

Fonte: Elaborado pelo autor.

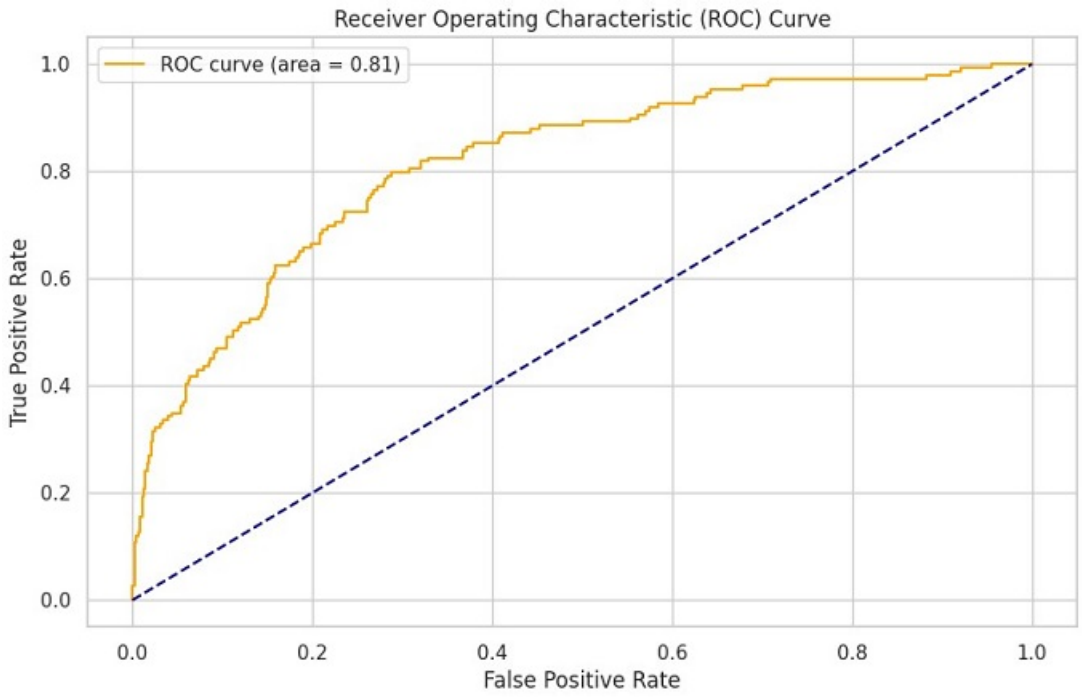
Aplicando os melhores hiperparâmetros encontrados, foi possível treinar o modelo de XGBoost obtendo os resultados apresentados na Tabela 7. Da mesma forma que o *Random Forest*, o modelo encontrou resultados razoáveis, não alcançando 50% de acerto dos casos de *churn* porém, proporcionalmente, acertando mais as predições se comparado com o outro modelo. Ainda, o modelo distinguiu com de forma eficiênte as classes positivas e negativas, alcançando uma área sob a curva ROC de 81.45% (Figura 29) .

Tabela 7 – Resultado Modelo XGBoost

Métrica	**Split Tradicional**	**Validação Cruzada**
Acurácia	78.20%	83.08%
Precisão	39.73%	63.29%
Sensibilidade	53.69%	45.7%
F-score	53.69%	55.24%

Fonte: Elaborado pelo autor.

Figura 29 – Curva ROC (AUC) XGBoost



Fonte: Elaborado pelo autor.

A Tabela 8 apresenta o resumo dos resultados obtidos aplicando a validação cruzada. O XGBoost foi o modelo que apresentou melhor resultado, independente da métrica avaliada. A acurácia encontrada pelo modelo foi de 81,79%, ou seja, essa é a porcentagem de verdadeiros positivos e verdadeiro negativos obtidos dentro de todo o universo de teste. A precisão de 66,36% indica que o modelo não está emitindo um número exagerado de falsos positivos, e a sensibilidade de 55,25% indica que mais da metade dos dados verdadeiros positivos (clientes que realmente deram *churn*) foram previstos, combinação que resultou em um F-Score balanceado de 59,25%. Os demais modelos apresentaram uma acurácia relativamente alta, todos acima de 75%, entretanto todos pecaram na obtenção de valores satisfatórios para sensibilidade, ou seja, estava viciado em prever os casos como "não *churn*".

Tabela 8 – Resultados dos Modelos de Classificação

Modelo	**Regressão Logística**	**Random Forest**	**SVM**	**XGBoost**
Acurácia	77.71%	82.42%	75.59%	83.08%
Precisão	31.41%	60.91%	34.6%	63.29%
Sensibilidade	2.97%	46.16%	27.27%	45.70%
F-score	6.21%	54.00%	32.12%	55.24%
AUC	59.82%	81.38%	69.14%	81.45%

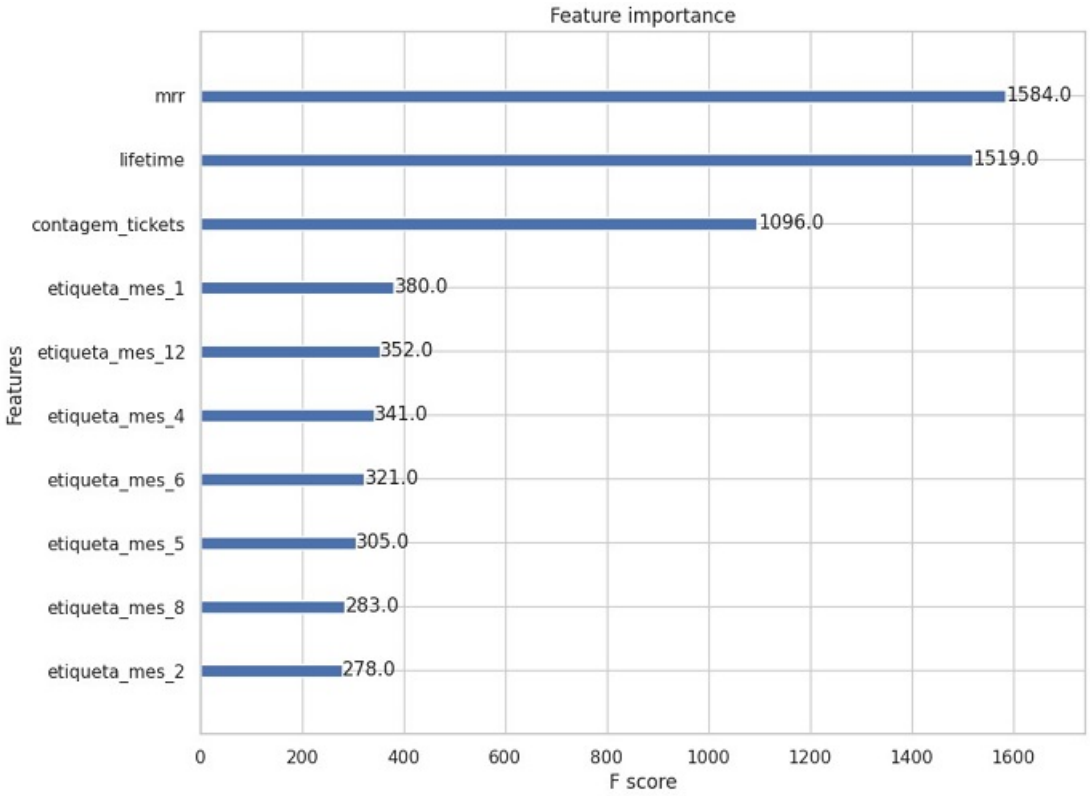
Fonte: Elaborado pelo autor.

Analisando o compilado dos resultados com o desempenho dos modelos de *Random Forest* e XGBoost, é notório que a utilização de metodos de *ensemble* implica em um aumento no desempenho de modelos de *machine learning*, principalmente quando aplicadas bases de dados com limitações quanto ao volume de dados disponíveis. XGB foi superior a todos os outros modelos na maioria das métricas, com exceção da sensibilidade que foi superado pelo *Random Forest*. A curva ROC, o XGboost obteve um resultado similar ao modelo de Random Forest, acima de 80% , confirmando que estes modelos conseguiu distinguir de forma suficiente as duas classes de "*churn*"e "não *churn*". Dessa forma, é possível concluir que o XGBoost é o modelo que melhor se encaixa ao conjunto de dados fornecidos, e deve ser utilizado para a futura aplicação prática.

A partir da escolha do melhor modelo, é possível analisar as variáveis de entrada a fim de obter informações úteis para a aplicação futura de planos de ação. Dessa forma, buscou-se identificar quais variáveis tiveram um impacto mais significativo na previsão para compreensão de que tipo de informação realmente está sendo útil para realizar a predição, e revelando quais variáveis tiveram pouca ou nenhuma importância para o resultado encontrado. Outra importância da identificação do impacto das variáveis é no entendimento da variável resposta, pois possibilita compreender quais fatores possuem um peso maior na identificação do perfil dos clientes propensos ao *churn*. O comando "*plot_importance()*"presente na biblioteca do XGBoost premite a criação de gráficos com a importância de cada uma das variáveis do modelo. Existem algumas formas de analisar a importância e a primeira é a padrão do comando, chamada de "*weight*". Este tipo de análise de importância se baseia na contagem de aparições de cada uma das variáveis nos nós de decisão das árvores criadas, implicando num peso relativo à ela (Figura 30).

As variáveis que apresentaram um peso maior nos *splits* do modelo foram, respectivamente o MRR, o *lifetime* do cliente e a contagem de tíquetes que o cliente possui no seu histórico. Essa análise exalta a importância destas variáveis, e evidencia a necessidade de garantir que estes dados estejam corretos quando inseridos no modelo para futuras previsões. Entretanto, existem situações em que mesmo com

Figura 30 – Importância do tipo "*weight*"para as variáveis de entrada

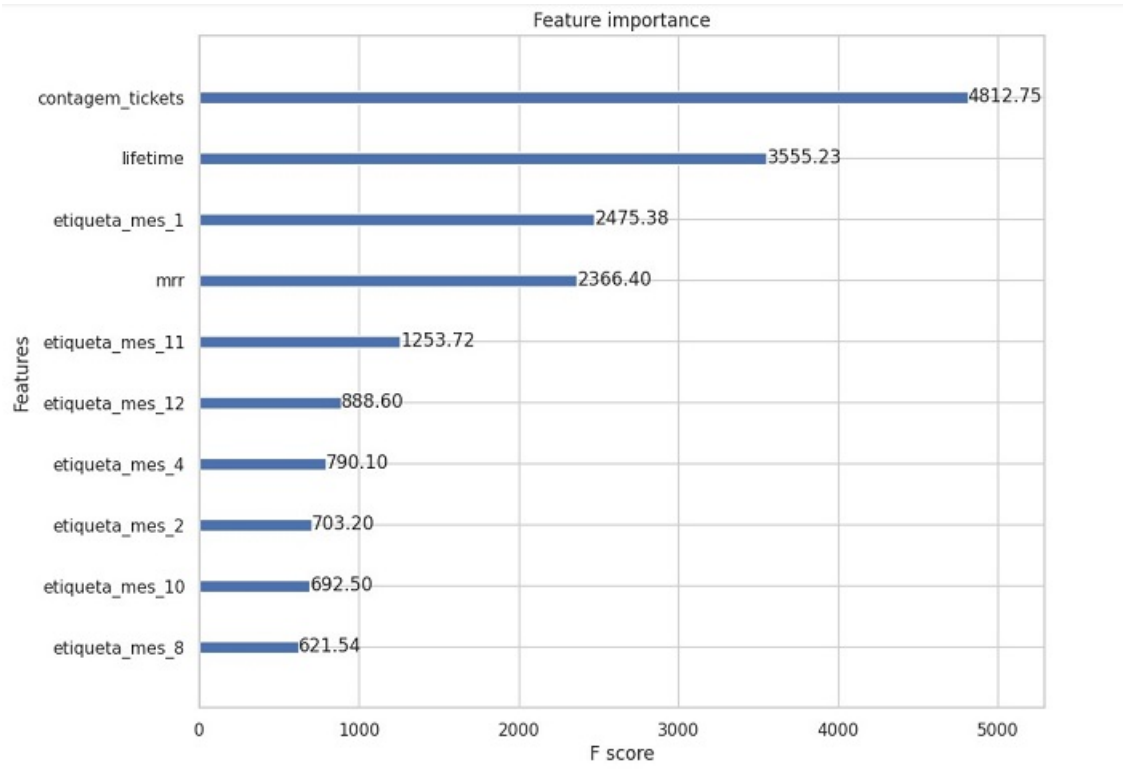


Fonte: Elaborado pelo autor.

poucas aparições em nós de decisão, uma variável implica em um grande ganho para a obtenção de uma resposta do modelo. Dessa forma, foi inserido o parâmetro *importance_type* como '*total_gain*', que calcula a importância da variável relativa ao quanto ela contribui para o modelo em termos de *F-Score* (Figura 31).

Com esta análise, a contagem de tíquetes se mostrou superior às demais variáveis, o *lifetime* apresentou maior importância que o MRR, e a variávei referente ao número de etiquetas impressas no mês anterior ao cancelamento apresentou um impacto relevante para a previsão do cancelamento. Dessa forma, é entender quais dados são mais sensíveis para a saúde do modelo, e que necessitam de um maior controle e governança, além de tornar possível o planejamento e execução de ações de prevenção do *churn* que serão apresentadas no próximo tópico.

Figura 31 – Importância do tipo "*total_gain*"para as variáveis de entrada



Fonte: Elaborado pelo autor.

4.3 PROPOSTA DE AÇÕES PARA PREVENÇÃO DO CANCELAMENTO

A literatura apontou algumas estratégias comuns em empresas de software que buscam evitar o cancelamento de licenças de uso. A modelagem preditiva possibilita a identificação de quais são os clientes propensos ao *churn*, que em um ambiente empresarial altamente competitivo colabora para o direcionamento dos recursos empresariais a fim de maximizar a eficiência mediante às limitação de custo. Ainda, foi possível selecionar as estratégias ainda não aplicadas pela empresa e cruzá-las com as informações de referentes às variáveis críticas identificadas, de forma a realizar a proposição de planos de ação para reter clientes propensos ao cancelamento além de aumentar a lealdade de clientes fiéis.

Sucesso do Cliente Personalizado Pela Identificação de Clientes Propensos ao Cancelamento: Se faz necessário adicionar uma nova segmentação para os clientes da empresa que possuem essa propensão ao cancelamento. Dentro desse segmento de clientes propensos ao cancelamento, devem ser preparadas subdivisões com base nos resultados da análise preditiva, levando em consideração o ticket médio, tempo de vida do cliente, usabilidade e problemas frequentes na aplicação da ferramenta que geram abertura de tíquetes. Dessa forma, essas subdivisões devem ser

mapeadas e caracterizadas levando em consideração suas necessidades, preferências e desafios. Os aspectos mencionados anteriormente ainda devem ser utilizados para a criação de um modelo de "saúde do cliente", dado que a um primeiro momento este pode ser um cliente de risco, mas após as ações de engajamento termine se tornando um cliente estável.

Com a nova segmentação de clientes e o modelo de saúde do cliente bem definidos, deve ser adotado um atendimento personalizado, com a criação de uma equipe especializada em atender e engajar os clientes propensos ao cancelamento com responsáveis por cada uma das classes definidas anteriormente. Dessa forma, é possível direcionar esforços na identificação de sinais de insatisfação possibilitando uma intervenção proativa por parte do especialista, oferecendo suporte adicional, treinamento personalizado, atualizações de produto e ofertas e incentivos direcionados.

O treinamento personalizado para estes clientes deve ser um programa contínuo, com a criação de conteúdo rico como webinars, tutoriais personalizados e materiais educativos que destacam as funcionalidades mais relevantes da ferramenta para este segmento específico de cliente. A educação fortalece a percepção de valor do cliente e reduz a probabilidade de cancelamento devido a falta de compreensão ou subutilização do produto. Por fim, devem ser criadas ofertas e incentivos para clientes propensos ao cancelamento, como descontos em renovações, upgrades gratuitos ou acesso antecipado a novos recursos demonstrando o comprometimento da empresa em manter o cliente satisfeito ao atender as necessidades específicas de cada grupo.

Programa de Valorização para Clientes de Longa Data: no modelo preditivo de cancelamento de licenças de uso, a variável *lifetime* indica o tempo pelo qual o cliente é cliente da organização. Essa variável quando filtrada em valores mais elevados na identificação de clientes propensos ao *churn* indica que mesmo que haja a intenção de cancelamento, existe uma relação duradoura de confiança entre ambas as partes do negócio. É possível aproveitar desta relação para engajar o cliente com estratégias diferenciadas como a publicidade com a participação de clientes de longa data e programas de indicação e eventos exclusivos para estes cliente.

Publicidades como destaques em *newsletters* e midias sociais, somadas à campanhas de marketing com a participação de clientes, reconhecendo as contribuições dos mesmos no sucesso do negócio aumentam o valor percebido pelo participante. Ainda, benefícios e recompensas exclusivos denotam uma valorização do cliente e fortalece laços. Dessa forma deve ser desenvolvido um programa de indicação exclusivo para clientes de longa data, com recompensas significativas por cada indicação bem-sucedidas. Por fim, sugere-se a organização de eventos exclusivos para clientes de longa data, sejam eles educativos, voltados à *networking* ou até comemorativos.

Por fim, não apenas estratégias para clientes já propensos ao cancelamento devem ser adotadas quando se tratando de valorização de clientes de longa data. É

importante o desenvolvimento de um programa de recompensas para clientes que mantenham uma relação de longo prazo, com descontos progressivos ou outros benefícios à medida que o tempo de vida do cliente aumenta. Dessa forma, além de reter um cliente propenso ao *churn*, a medida ajuda a aumentar a lealdade de clientes estáveis, possibilitando que os mesmos tornem-se promotores da marca.

Coleta de Feedbacks Quanto a Usabilidade e Funcionalidades com Clientes Propensos ao Cancelamento: na fundamentação teórica foi levantada a importância das pesquisas de satisfação para coleta de informações específicas de necessidades e expectativas dos clientes. No modelo de predição do *churn*, a quantidade de impressões de etiquetas dos clientes nos meses anteriores foi um fator que apareceu com um peso considerável nos quesitos de peso e ganho do modelo. Dessa forma, é recomendável realizar pesquisas de satisfação exclusivas com clientes que estão com baixa utilização do sistema, e que fazem parte da segmentação de clientes propensos ao cancelamento, de forma a identificar pontos de melhoria. Estas pesquisas devem contar com perguntas abertas e fechadas que abordem aspectos específicos do *software* , eque podem estar impactando a satisfação do cliente. Este processo destaca a importância da opinião dos clientes na melhoria contínua do produto, aumentando o valor percebido e a satisfação.

Com os *feedbacks* coletados, deve ser realizada uma análise abrangente das respostas a fim de identificar padrões comuns relacionados à usabilidade e funcionalidades, ajudando a priorizar áreas de melhoria dentro da ferramenta. Dessa forma, com base nos insights obtidos, deve ser planejada a implementação das melhorias no *software*, que quando finalizadas devem ser comunicadas aos clientes que às propuseram, destacando como essas mudanças agregam valor aos serviços oferecidos e oficializando a personalização direcionada.

Aumentar a Disponibilidade de Observações: Ao longo da preparação do modelo de predição do *churn*, ficou evidente a falta de dados para o treinamento do modelo, levando ao ponto de ser necessária a aplicação de técnicas de balanceamento com a replicação de dados da variável positiva (*oversampling*). A única forma de aumentar o volume de dados associados à essa variável resposta é com a ocorrência de novos cancelamentos por parte de clientes, que é justamente o que busca-se evitar com a realização do presente estudo. Entretanto, é impossível reduzir a taxa de *churn* de uma empresa de *software* a zero, e dessa forma, anualmente o modelo deve ser incrementado a fim de adicionar os novos casos de cancelamento do ano anterior, a fim de aumentar gradativamente a confiança do modelo.

Plano de Correção da Base de Dados: observou-se diversos problemas na compatibilidade entre as chaves presentes nas diferentes fontes fonte, que afetou negativamente a manipulação dos mesmos ocasionando em cortes de observações que poderiam aumentar a confiabilidade do modelo. Dessa forma, deve-se criar um projeto

voltado unificação dos dados, unificando as chaves primárias e secundárias entre as diferentes fontes, e definindo um modelo de consulta padrão para evitar a perda de informações na coleta e manipulação de dados. Ainda, a base do Hubspot já possui um sistema de governança implementado, porém dado o vulume de informações faltantes nesta base deve ser incluído no plano de preenchimento das mesmas para aumentar a disponibilidade.

Inclusão e Criação de Novas Variáveis: o modelo utilizou apenas de uma informação de usabilidade, que é o número de etiquetas impressas nos últimos doze meses de relacionamento do cliente com o *software*. Entretanto, a ferramenta possui uma série de funcionalidades além da impressão de etiquetas que podem ser mapeadas e inseridas no modelo, a fim de incrementar o treinamento no levando em consideração a usabilidade. Após o mapeamento das informações complementares, novas métricas podem ser criadas para quantificar a utilização do cliente de uma forma mais completa, melhorar o modelo de predição e perceber necessidades que os clientes ainda não enxergaram.

Por fim, já foi sugerido anteriormente a realização de pesquisas de *feedback* personalizadas para os clientes propensos ao cancelamento. Porém o modelo desenvolvido peca na utilização de variáveis que medem a satisfação do cliente e qualidade percebida. Dessa forma, deve ser definido um processo de pesquisa de satisfação do cliente com o serviço prestado que envolva todos os clientes da base, e não apenas aqueles que solicitam atendimento, a fim de coletar dados quantitativos para definição de métricas de satisfação.

5 CONCLUSÃO

Dado a necessidade de desenvolvimento dos processos em empresas de *Software as a Service*, o objetivo geral deste trabalho de conclusão de curso foi propor ações que ajudem a prevenção do cancelamento de licenças de uso *software* em uma empresa de tecnologia com o auxílio de modelos baseados em dados, e o atingimento dos objetivos específicos corroborou para a conclusão do objetivo geral.

Inicialmente a literatura serviu como base para direcionar a pesquisa, com a identificação de fatores que influenciam no cancelamento de licenças como a necessidade, expectativa e satisfação dos clientes, contextos culturais e sociais e usabilidade da ferramenta, que serviu de apoio para a escolha das variáveis de entrada dos modelos de previsão. Ainda, foram levantadas estratégias que objetivam a prevenção do cancelamento de licenças comumente utilizadas por empresas do mesmo ramo. Com o embasamento literário, os dados disponíveis para a modelagem foram analisados e trabalhados, possibilitando a escolha de vinte e duas variáveis de entrada.

Foram treinados quatro diferentes modelos preditivos: Regressão Logistica, Máquinas de Vetores de Suporte, *Random Forest* e XGBoost, com as devidas otimizações de hiperparâmetros, e os resultados foram avaliados utilizando com validação cruzada com cinco métricas: acurácia, precisão, sensibilidade, F-Score e Área sob a curva ROC. O XGBoost alcançou o melhor resultado com uma acurácia de 83,08% e AUC de 81,45%. Dadas limitações de volume, incompatibilidade e desbalanceamento entre nas bases de dados, torna-se relevante atualizações periódicas no modelo, tratamento dos dados base e a criação e inclusão de novas variáveis de entrada para a melhoria das demais métricas, dado o baixo desempenho apresentado por elas.

Ainda, além da identificação dos clientes propensos ao cancelamento, foi possível levantar as variáveis de entrada com maior relevância na predição do cancelamento de licenças de uso. A partir do cruzamento dos resultados do modelo com as estratégias encontradas na literatura, foi possível traçar planos de ação a fim de melhorar a eficiência operacional da empresa quando se tratando da produtividade dos times de sucesso do cliente.

Além das limitações relacionados à base de dados, é importante ressaltar que esta pesquisa aborda o caso de uma empresa de *software* específica, com suas particularidades quanto ao tipo de negócio e perfil de cliente. Portanto, os resultados não podem ser generalizados para outros contextos que estão aquém do ambiente em que a empresa objeto de estudo está inserida.

Como a aplicação de modelos de predição se mostrou um agente direcionador da tomada de decisão no contexto de *softwares* como serviço, e estudo pode ser adaptado para diferentes realidades em trabalhos futuros. Ainda, pontos de melhoria na base de dados fazem parte das ações futuras planejadas por este estudo, como

mencionado anteriormente. Por fim, sugere-se trabalhar na identidicação da influência positiva e negativa de variáveis, possibilitando, com o entendimento do porquê aquele cliente está com em risco de cancelamento, desenvolver métricas de saúde do cliente para reter o cliente de forma personalizada.

REFERÊNCIAS

ADAMS, Rick. **Practical customer success management: a best practice framework for rapid generation of customer success**. [*S.l.*], 2019.

AGRESTI, Alan. **Foundations of Linear and Generalized Linear Models**. [*S.l.*], 2015.

AHMAD, Abdelrahim; JAFAR, Asef; ALJOUMAA, Kadan. **Customer churn prediction in telecom using machine learning in big data platform**. [*S.l.*], mar. 2019.

ALMEIDA, Carlos Ferreira de. **Contratos II-5a Edição**. [*S.l.*], 2021.

AMIN, Adnan; ANWAR, Sajid; ADNAN, Awais; NAWAZ, Muhammad; HOWARD, Newton; QADIR, Junaid; HAWALAH, Ahmad; HUSSAIN, Amir. **Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study**. [*S.l.*], out. 2016.

BABY, Bestin; DAWOD, Zainb; SHARIF, Saeed; ELMEDANY, Wael. **Customer Churn Prediction Model Using Artificial Neural Networks (ANN): A Case Study in Banking**. [*S.l.*], 2023.

BATISTA, Igor Carvalho de Brito. **Análise de influência de características no modelo de predição de Churn**. 2022. B.S. thesis – Universidade Federal do Rio Grande do Norte.

BATTISTI, DE; SMOLSKI, FM da S. **Software R: curso avançado**. [*S.l.*], 2019.

BENLIAN, Alexander. **A transaction cost theoretical analysis of software-as-a-service (SAAS)-based sourcing in SMBs and enterprises.** [*S.l.*], jan. 2009. P. 25–36.

BIBI, Stamatia; KATSAROS, Dimitrios; BOZANIS, Panayiotis. **Business Application Acquisition: On-Premise or SaaS-Based Solutions?** [*S.l.*], mai. 2012. P. 86–93.

BOSAK, Nathália Prestes. **O impacto do parque tecnológico no perfil comercial das empresas de software como serviço: um estudo de caso sobre o Tecnosinos**. [*S.l.*], 2021.

BRADLEY, Andrew. **The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms**. [*S.l.*], nov. 1996. P. 1145–1159.

BREIMAN, L.; FRIEDMAN, J.; STONE, C.J.; OLSHEN, R.A. **Classification and Regression Trees**. [*S.l.*], 1984. ISBN 9780412048418. Disponível em: https://manuals.google.com.br/manuals?id=JwQx-WOmSyQC.

BURKOV, Andriy. **Deep Learning**. [*S.l.*], jan. 2019. ISBN 199957950X.

CANDIDO, Alexandre Gomes *et al.* **EXPLORANDO A INFLUÊNCIA DA VISUALIZAÇÃO DE DADOS NO CONTEXTO DO MARKETING DIGITAL: Uma revisão sistemática da literatura**. [*S.l.*], 2023.

CARLOS, Sarah Fonseca. **Estudo sobre as competências dos profissionais de customer success em uma empresa de software como serviço.** [*S.l.*], 2023.

CERQUEIRA, Tarcisio Queiroz. **Software: lei, comércio, contratos e serviços de informática**. [*S.l.*], 2000.

CHARANDABI, Sina Esmaeilpour. **Prediction of customer churn in banking industry**. [*S.l.*], 2023.

CHEN, Dehua; NIGRI, Eduardo; OLIVEIRA, Gibram; SEPULVENE, Luis; ALVES, Tiago. **Métricas de Avaliação em Machine Learning: Classificação**. [*S.l.*], 2020. Disponível em: https://medium.com/kunumi/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-em-machine-learning-classifica%C3%A7%C3%A3o-49340dcdb198. Acesso em: 11 fev. 2023.

CHIRITA, Silvana. **How to reduce customer churn with user-centric authentication**. [*S.l.*], 2021. Disponível em: https://blog.typingdna.com/reduce-customer-churn-with-user-centric-authentication/. Acesso em: 1 out. 2023.

CHO, Vincent; CHAN, Aman. **An integrative framework of comparing SaaS adoption for core and non-core business operations: An empirical study on Hong Kong industries**. [*S.l.*], jun. 2015. P. 629–644.

CHRISTOPHER, Martin. **Logística e gerenciamento da cadeia de suprimentos**. [*S.l.*], 2022.

CISTER, Angelo Maia. **Mineração de dados para a análise de atrito em telefonia móvel. Tese (Doutorado em Engenharia).** 2005. Tese (Doutorado) – Universidade Federal do Rio de Janeiro.

CORREA BAHNSEN, Alejandro; AOUADA, Djamila; OTTERSTEN, Björn. **A novel cost-sensitive framework for customer churn predictive modeling**. [*S.l.*], jun. 2015. P. 1–15.

CORREIA, Geovanni dos Santos; BARRETO, Gabriel Santos de Lima; ALVES, Nathalia de Meneses. **Crescimento e expansão no uso de software como serviço (SaaS): estratégias e obstáculos para empresas de tecnologia**. [*S.l.*], 2024. e14902–e14902.

COTTON, Pam. **Sete dicas para te ajudar a reduzir a taxa de cancelamento de clientes**. [*S.l.*], 2022. Disponível em: https://br.hubspot.com/blog/service/reduzir-taxa-cancelamento. Acesso em: 1 out. 2023.

CRISTINE, Sara. **5 estratégias de fidelização para diminuir o Churn**. [*S.l.*], 2019. Disponível em: https://blog.octadesk.com/estrategias-de-fidelizacao-reduzir-churn/. Acesso em: 10 jul. 2023.

CURVELO, Rakky. **Como se tornar uma empresa com uma cultura de customer centric?** [*S.l.*], 2022. Disponível em: https://br.hubspot.com/blog/service/customer-centric?hubs_content=br.hubspot.com%2Fblog%2Fservice%2Freduzir-taxa-cancelamento&hubs_content-cta=necessidades%20do%20cliente. Acesso em: 1 out. 2023.

DARÉ, PR; CAMPOMAR, Marcos Cortez. **Retenção de clientes à luz do gerenciamento de Churn: um estudo no setor de telecomunicações**. 2007. Tese (Doutorado) – Dissertaçãoo de Mestrado em Administraçãoo. Universidade de São Paulo.

DAY, Jenn. **A Definition of Customer Churn**. [*S.l.*], 2023. Disponível em: https://www.ngdata.com/what-is-customer-churn/. Acesso em: 1 out. 2023.

DE CAIGNY, Arno; COUSSEMENT, Kristof; VERBEKE, Wouter; IDBENJRA, Khaoula; PHAN, Minh. **Uplift modeling and its implications for B2B customer churn prediction: A segmentation-based modeling approach**. [*S.l.*], 2021. P. 28–39.

DHALIWAL, Sukhpreet Singh; NAHID, Abdullah-Al; ABBAS, Robert. **Effective intrusion detection system using XGBoost**. [*S.l.*], 2018. P. 149.

DROMEY, R Geoff. **Software quality—prevention versus cure?** [*S.l.*], 2003. P. 197–210.

EVEO. **Licença de software: como funciona e quais são os tipos existentes**. [*S.l.*], 2022. Disponível em: `https://blog.eveo.com.br/licenca-de-software`. Acesso em: 1 out. 2023.

FERNANDES, Antônio Alves Tôrres; FILHO, Dalson Britto Figueiredo; ROCHA, Enivaldo Carvalho da; SILVA NASCIMENTO, Willber da. **Leia este artigo se você quiser aprender regressão logística**. Curitiba, 2020.

FRANZOSO, Júlia Medina. **O cenário de abandono da linha pré-paga pelos clientes da Claro**. [*S.l.*], 2007.

GANESH, Jaishankar; ARNOLD, Mark J; REYNOLDS, Kristy E. **Understanding the customer base of service providers: an examination of the differences between switchers and stayers**. [*S.l.*], 2000. P. 65–87.

GARTNER. **Contracting by 2026 — and So Do Risks**. [*S.l.*], 2021. Disponível em: `https://www.gartner.com/en/documents/4009001`.

GE, Yizhe; HE, Shan; XIONG, Jingyue; BROWN, Donald E. **Customer churn analysis for a software-as-a-service company**. [*S.l.*], 2017. P. 106–111.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. [*S.l.*], jan. 2008. P. 220. ISBN 9788522451425.

GOLD, Carl. **Fighting Churn with Data: The science and strategy of customer retention**. [*S.l.*], 2020.

GUIMARÃES, Leandro. **Licença de software: o que é e quais são seus tipos?** [*S.l.*], 2016. Disponível em:

https://www.knowsolution.com.br/licenca-software-e-seus-tipos/. Acesso em: 1 out. 2024.

GUPTA, BINAY KUMAR. **Bagging in Machine Learning**. [*S.l.*], 2023. Disponível em: https://www.scaler.com/topics/machine-learning/bagging-in-machine-learning/. Acesso em: 11 fev. 2023.

HADDEN, John; TIWARI, Ashutosh; ROY, Rajkumar; RUTA, Dymitr. **Computer assisted customer churn management: State-of-the-art and future trends**. [*S.l.*], 2007. P. 2902–2917.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)**. [*S.l.*], fev. 2009. ISBN 0387848576.

HENNIG-THURAU, Thorsten; KLEE, Alexander. **The impact of customer satisfaction and relationship quality on customer retention: A critical reassessment and model development**. [*S.l.*], 1997. P. 737–764.

HNS. **Why passwords are to blame for loss of revenue, identity attrition and poor customer experiences**. [*S.l.*], 2021. Disponível em: https://www.helpnetsecurity.com/2021/04/01/dependency-on-traditional-password-systems/. Acesso em: 1 out. 2023.

HÖPPNER, Sebastiaan; STRIPLING, Eugen; BAESENS, Bart; BROUCKE, Seppe vanden; VERDONCK, Tim. **Profit driven decision trees for churn prediction**. eng. [*S.l.*], 2020. P. 920–933.

HUANG, Bingquan; KECHADI, Mohand Tahar; BUCKLEY, Brian. **Customer churn prediction in telecommunications**. eng. OXFORD, 2012. P. 1414–1425.

IDF. **Usability**. [*S.l.*], 2023. Disponível em: https://www.interaction-design.org/literature/topics/usability. Acesso em: 1 out. 2023.

IDRIS, Adnan; KHAN, Asifullah. **Customer churn prediction for telecommunication: Employing various various features selection techniques and tree based ensemble classifiers**. [*S.l.*], dez. 2012. P. 23–27. ISBN 978-1-4673-2249-2.

KEAVENEY, Susan M. **Customer switching behavior in service industries: An exploratory study**. [*S.l.*], 1995. P. 71–82.

KEYBANC. SaaS Survey Results. *In*: KEYBANC CAPITAL MARKETS.

KON, Martin. **Customer churn**. [*S.l.*], 2004. P. 54–60.

KOTLER, Philip. **Administração de Marketing–A edição do Novo Milênio. 10ª Edição**. [*S.l.*], 2000.

KOTLER, Philip T; PEREIRA, Elaine; NASCIMENTO, Herbert do; LEE, NANCY R. **Marketing social**. [*S.l.*], 2019.

KUMAR, Soni; LEEMA, Nelson. **PCP: Profit-Driven Churn Prediction using Machine Learning Techniques in Banking Sector**. [*S.l.*], jan. 2023. P. 303.

KURTZ, D.L.; CLOW, K.E. **Services Marketing**. [*S.l.*], 1997. ISBN 9780471180340. Disponível em: `https://manuals.google.com.br/manuals?id=SO4JAQAAMAAJ`.

LALWANI, Praveen; MISHRA, Manas; CHADHA, Jasroop; SETHI, Pratyush. **Customer churn prediction system: a machine learning approach**. [*S.l.*], fev. 2022. P. 1–24.

LARHMAN. **File:SVM margin.png**. [*S.l.*], 2018. Disponível em: `https://commons.wikimedia.org/wiki/File:SVM_margin.png`. Acesso em: 1 out. 2023.

LAURETTO, Marcelo. **Árvores de decisão**. [*S.l.*], 2010.

LUIZON, Bruna Franciany Girata. **Análise preditiva de churn em um e-comerce (Monografia)**. 2019. Tese (Doutorado) – Universidade Federal do Paraná.

MAKRIDAKIS, Spyros; SPILIOTIS, Evangelos; ASSIMAKOPOULOS, Vassilios. **Statistical and Machine Learning forecasting methods: Concerns and ways forward**. [*S.l.*], 2018. e0194889.

MARÍN DÍAZ, Gabriel; GALÁN, José Javier; CARRASCO, Ramón Alberto. **XAI for Churn Prediction in B2B Models: A Use Case in an Enterprise Software Company**. [*S.l.*], 2022. P. 3896.

MARQUES, Tatiana Freire. **O não cumprimento do contrato de licença de uso de software perante o direito brasileiro e o direito português**. [*S.l.*], 2014.

MATTISON, Rob. **The telco churn management handmanual**. [*S.l.*], 2006.

MEHTA, Nick; STEINMAN, Dan; MURPHY, Lincoln. **Customer success: How innovative companies are reducing churn and growing recurring revenue**. [*S.l.*], 2016.

MENEZES, Diego Kilpp de. **Fatores que levam os clientes a cancelar seus serviços de telefonia móvel**. [*S.l.*], 2014.

MINIARD, Pauli W; ENGEL, James; BLACKWELL, Roger. **Comportamento do consumidor**. [*S.l.*], 2000.

MITCHELL, Tom M. **Machine Learning**. [*S.l.*], mar. 1997. P. 432. ISBN 0070428077.

MIURA, Lucas. **Modelos de Predição | Otimização de Hiperparâmetros em Python**. [*S.l.*], 2020. Disponível em: `https://medium.com/turing-talks/modelos-de-predi%C3%A7%C3%A3o-otimiza%C3%A7%C3%A3o-de-hiperpar%C3%A2metros-em-python-3436fc55016e`. Acesso em: 1 out. 2023.

MOTA, Mauricio. **A boa-fé nos contratos de licença de uso de software**. [*S.l.*], 1999. P. 12.

NAVITA. **Quais os tipos de licença de softwares? Conheça elas aqui!** [*S.l.*], 2023. Disponível em: `https://navita.com.br/blog/quais-os-tipos-de-licenca-de-softwares-conheca-elas-aqui/`. Acesso em: 1 out. 2023.

NESLIN, Scott A; GUPTA, Sunil; KAMAKURA, Wagner; LU, Junxiang; MASON, Charlotte H. **Defection detection: Measuring and understanding the predictive accuracy of customer churn models**. [*S.l.*], 2006. P. 204–211.

NORONHA, Maria Gabriela Ferreira de. **A aplicação do marketing de relacionamento em provedores de software as a service (SaaS) B2B para minimizar taxas de churn**. 2020. Tese (Doutorado) – Universidade de São Paulo.

NYCE, Charles. **Predictive Analytics White Papper**. [*S.l.*], 2007.

PAULILLO, JÚLIO. **Dicas de como atender às necessidades e expectativas dos clientes**. [*S.l.*], 2023. Disponível em: https://www.agendor.com.br/blog/necessidades-e-expectativas-dos-clientes/. Acesso em: 1 out. 2023.

PERIŠIĆ, Ana; PAHOR, Marko. **Clustering mixed-type player behavior data for churn prediction in mobile games**. [*S.l.*], 2023. P. 165–190.

PIRES, André de Assis. **Responsabilidades e desafios do Customer Success Manager na indústria de software**. 2023. Tese (Doutorado).

RITTINGHOUSE, John W.; RANSOME, James F. **Cloud Computing: Implementation, Management, and Security**. [*S.l.*], 2009.

ROGOJAN, Ben. **Boosting and Bagging: How To Develop A Robust Machine Learning Algorithm**. [*S.l.*], 2017. Disponível em: https://medium.com/@SeattleDataGuy/how-to-develop-a-robust-algorithm-c38e08f32201. Acesso em: 11 fev. 2023.

ROLLINS, JB. **Foundational Methodology for Data Science, IBM Anal**. [*S.l.*]: document, 2015.

ROUTH, Pallav; ROY, Arkajyoti; MEYER, Jeff. **Estimating customer churn under competing risks**. [*S.l.*], 2021. P. 1138–1155.

SALLES, Ingrid. **O papel do Customer Success na redução de churn e aumento da retenção de clientes**. [*S.l.*], 2023. Disponível em: https://www.linkedin.com/pulse/o-papel-do-customer-success-na-redu%C3%A7%C3%A3o-de-churn-e-aumento-salles/?originalSubdomain=pt. Acesso em: 1 out. 2023.

SANDSTRÖM, Lisa. **Improving customer onboarding in a B2B SaaS company**. [*S.l.*], 2022.

SANTOS, Luana Cruz. **Análise da área de customer success de empresa SAAS: estudo de caso de uma startup SaaS de Brasília**. [*S.l.*], 2022.

SCHNEIDER, Pedro Henrique. **Análise preditiva de Churn com ênfase em técnicas de Machine Learning: uma revisão**. 2016. Tese (Doutorado).

SHAIKH, Rahil. **Cross Validation Explained: Evaluating estimator performance.** [*S.l.*], 2018. Disponível em: https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85. Acesso em: 11 fev. 2023.

SHANMUGAM, Annadurai. **Improving Customer Experience and Reducing Customer Churn by Implementing Frictionless, Adaptive, and Risk-Based Authentication Controls for Digital Assets**. 2022. Tese (Doutorado) – Utica University.

SHETH, Jagdish N; MITTAL, Banwari; NEWMAN, Bruce I. **Comportamento do cliente: indo além do comportamento do consumidor**. [*S.l.*], 1999.

SILVA, Andressa Alves da; REGO BARROS FILHO, Fernando do; MULLER, Jessica Cardoso; TAVARES, Michel Alexandre Mesquita. **CONTRATO DE SOFTWARE**. [*S.l.*], 2017.

SILVA, Fernando da. **FLUXO DE TRABALHO PARA MODELAGEM PREDITIVA**. [*S.l.*], 2023. Disponível em: https://analisemacro.com.br/econometria-e-machine-learning/fluxo-de-trabalho-para-modelagem-preditiva/.

SILVEIRA, Ian Vieira *et al.* **Modelo de previsão de demanda com o uso de aprendizado supervisionado de máquina: um estudo de caso em uma empresa de varejo**. [*S.l.*], 2019.

SIVAKUMAR, A; GUNASUNDARI, R. **A survey on data preprocessing techniques for bioinformatics and web usage mining**. [*S.l.*], 2017. P. 785–794.

SOLOMON, Michael R. **O Comportamento do consumidor-: comprando, possuindo e sendo**. [*S.l.*], 2016.

STEINMAN, Dan; MURPHY, Lincoln; MEHTA, Nick. **Customer Success: como as empresas inovadoras descobriram que a melhor forma de aumentar a receita é garantir o sucesso dos clientes**. [*S.l.*], 2017.

STROUSE, Karen G. **Marketing telecommunications services: new approaches for a changing environment**. [*S.l.*], 1999.

SUH, Youngjung. **Machine learning based customer churn prediction in home appliance rental business**. [*S.l.*], abr. 2023.

TAO, Jianrong *et al.* **Explainable AI for cheating detection and churn prediction in online games**. [*S.l.*], 2022.

TOTANGO. **What is Customer Churn?** [*S.l.*], 2023. Disponível em: `https://www.totango.com/customer-churn`. Acesso em: 1 out. 2023.

VAIDYANATHAN, Ashvin; RABAGO, Ruben. **The Customer Success Professional's Handmanual: How to Thrive in One of the World's Fastest Growing Careers–While Driving Growth For Your Company**. [*S.l.*], 2020.

VAN DER KOOIJ, Jacco; PIZARRO, Fernando. **Blueprints for a SaaS sales organization**. [*S.l.*], 2015.

VAVRA, Terry G. **Marketing de relacionamento: aftermarketing**. [*S.l.*], 1993.

WALLACE, Thomas F. **Customer-driven strategy: winning through operational excellence**. [*S.l.*], 1992.

WATERS, Bret. **Software as a service: A look at the customer benefits**. [*S.l.*], 2005. P. 32–39.

WEBER, Donna. **Onboarding Matters: How Successful Companies Transform New Customers Into Loyal Champions**. [*S.l.*], 2021.

WIRTZ, Jochen; LOVELOCK, Christopher. **Services Marketing: People, Technology, Strategy, 9th edition**. [*S.l.*], jan. 2022. ISBN 978-194-4659-00-4.

ZUMEL, Nina; MOUNT, John. **vtreat: a data. frame Processor for Predictive Modeling**. [*S.l.*], 2016.