UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Tayrone Duque Esteves

DESENVOLVIMENTO DE UM PROCESSO ESTRUTURADO PARA A MODELAGEM DE PREVISÃO DE *CHURN* EM MODELOS DE NEGÓCIO DE RECORRÊNCIA

Itajubá – MG

2025

UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Tayrone Duque Esteves

DESENVOLVIMENTO DE UM PROCESSO ESTRUTURADO PARA A MODELAGEM DE PREVISÃO DE *CHURN* EM MODELOS DE NEGÓCIO DE RECORRÊNCIA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Itajubá para a obtenção do título de Mestre em Engenharia de Produção.

Orientador: Prof. Dr. José Henrique de Freitas Gomes

Itajubá – MG

2025

AGRADECIMENTOS

Nada verdadeiramente importante na vida é conquistado individualmente. A conclusão de um mestrado não é exceção. Se cheguei até aqui, foi porque outras pessoas acreditaram em mim e me apoiaram durante uma jornada que, muitas vezes, foi mais difícil do que eu seria capaz de lidar sozinho. Por isso, com toda a sinceridade, agradeço a todos que compartilharam o caminho comigo.

Dentre essas pessoas, a mais importante foi a minha esposa, Thais. Em um dos momentos mais difíceis da minha vida até aqui, me deu a ideia de investir meu tempo na minha formação. Sem ela eu não teria nem coragem de começar. Seu amor pelo ensino e pela pesquisa são a minha maior inspiração e sem o seu amor e apoio eu não teria chegado a lugar algum. Eu te amo, princesa.

Aos meus pais, Vanda e Gilmar, e aos meus irmãos, Tales e Tulio, pelo amor, carinho e compreensão nos momentos de ausência. Obrigado por me ensinarem que o conhecimento liberta e que a jornada de aprender sempre vale a pena.

Aos amigos e colegas de profissão, William e Gabriel, que desde o início compreenderam o quanto tudo isso significava para mim e me apoiaram de todas as formas para chegar até aqui. Vocês foram fundamentais para que eu conseguisse o recurso mais importante para me tornar mestre: tempo para me dedicar. Com vocês eu aprendi, cresci e construí coisas das quais eu me orgulho muito.

Aos professores e colegas do Programa de Pós-Graduação em Engenharia de Produção da UNIFEI, por terem compartilhado comigo não só o conhecimento, mas a experiência que me abriu novos caminhos na vida. Em especial ao meu orientador, Prof. Dr. José Henrique de Freitas Gomes, por ter me ajudado na condução do trabalho sem nunca tirar minha autonomia.

Ao Charles, que com seu conhecimento e paciência foi uma luz guia na reta final desse projeto. Obrigado por compartilhar tanto comigo em tão pouco tempo.

Por fim, mas não menos importante, agradeço a Deus. Por ter me feito teimoso o suficiente para não desistir e ter colocado todas essas pessoas incríveis no meu caminho.

“A melhor forma de prever o futuro é criá-lo.”

Peter Drucker

RESUMO

O modelo de negócio de compras recorrentes, como os clubes de assinatura, tem crescido muito nos últimos anos. Como adquirir novos clientes tem um custo maior do que reter os atuais, a perda de clientes (*churn*) tem um impacto negativo na competitividade do negócio, e a capacidade de prevê-lo torna-se um diferencial estratégico. A diversidade de métodos disponíveis e a falta de um processo padronizado de avaliação de modelos são desafios tanto acadêmicos quanto práticos para o setor. Nesse contexto, esse trabalho estruturou um procedimento voltado para a construção de métodos de previsão de *churn* composto de três etapas: preparação dos dados, aplicação dos modelos e avaliação dos modelos. Para isso, primeiramente os dados são padronizados e balanceados, métricas são definidas para comparação entre os modelos e uma análise de estabilidade é realizada. Para validação deste método, sua aplicação foi realizada no caso real de um clube de assinaturas de livros, para o qual sete variáveis preditoras foram selecionadas para compor 3 modelos de previsão: regressão logística, uma combinação entre o algoritmo média-*k* e a regressão logística, e uma rede neural *mutilayer perceptron* (MLP). A regressão logística ficou selecionada como o melhor dentre os métodos testados, apresentando acuracidade de 71,1% e 44,6% de precisão. Os resultados encontrados são superiores aos do modelo atual aplicado na empresa estudada, que possui 71,8% de acuracidade e 12,1% de precisão. O procedimento estruturado mostrou-se efetivo na seleção de métodos e direcionamento de tomadas de decisão para o desenvolvimento de modelos de previsão de forma padronizada e robusta.

Palavras-chave: Previsão de *churn*; *Framework*, Regressão logística; Redes neurais; Clubes de assinatura.

ABSTRACT

The recurring purchase business model, such as subscription clubs, has grown significantly in recent years. Since acquiring new customers is more costly than retaining existing ones, customer loss (churn) negatively impacts business competitiveness, making the ability to predict it a strategic advantage. The variety of available methods and the lack of a standardized model evaluation process pose both academic and practical challenges for the sector. In this context, this study structured a procedure for developing churn prediction models, consisting of three stages: data preparation, model application, and model evaluation. First, the data is standardized and balanced, comparison metrics are defined, and lastly a stability analysis is conducted. To validate the method, its application was performed in a real case of a book subscription club. Seven predictive variables were selected to train three models: logistic regression, a combination of the *k*-means algorithm with logistic regression, and a multilayer perceptron (MLP) neural network. Logistic regression was identified as the best-performing method, achieving 71,1% of accuracy and 44,6% of precision. These results are superior to the actual model used by the company, that has 71,8% of accuracy and 12,1% of precision. Overall, the structured procedure proved effective in model selection and decision-making for churn prediction development in a standardized and robust way.

Keywords: Churn forecast; Framework; Logistic regression; Neural network; subscription-based markets.

LISTA DE FIGURAS

Figura 1 – Motivos de adesão a clubes de assinatura para consumidores dos EUA 13
Figura 2 – Curva de características operacionais do receptor 23
Figura 3 – *Framework* para construção de modelos de previsão de *churn* 25
Figura 4 – *Framework* para comparação entre modelos de previsão de churn 26
Figura 5 – Evolução do número de publicações sobre previsão de *churn* na última década 29
Figura 6 – Critérios para refinamento da base de pesquisa 30
Figura 7 – Principais mercados estudados para previsão de *churn* 31
Figura 8 – Principais ferramentas utilizadas para previsão de *churn* 32
Figura 9 – Principais grupos de variáveis preditoras utilizadas em modelos de previsão de *churn* 33
Figura 10 – Representação gráfica da regressão logística 38
Figura 11 – Representação de uma rede MLP e equação de saída do neurônio 40
Figura 12 – Método de pesquisa de modelagem e simulação. 42
Figura 13 – *Framework* proposto para construção de modelos de previsão de *churn* em mercados de compra recorrente. 45
Figura 14 – Proporção de cancelamentos (A) e proporção de cancelamento voluntários e involuntários (B) durante o mês de setembro de 2023..... 50
Figura 15 – Análise por *product_detail* 51
Figura 16 – Análise por *renewal_period* 51
Figura 17 – Análise por *payment_type* 52
Figura 18 – Distribuição de valores pagos pelos assinantes que cancelaram ou não as assinaturas 53
Figura 19 – Análise geral dos parâmetros da amostra. 54
Figura 20 – Análise de sensibilidade para os termos do modelo sem interação. 58
Figura 21 – Análise de sensibilidade para os termos do modelo com interação de primeiro grau. 58

Figura 22 – Análise de estabilidade do modelo de previsão de *churn* construído a partir da regressão logística. 66

LISTA DE TABELAS

Tabela 1 – Matriz de confusão ou matriz de erro..... 21
Tabela 2 – Resumo de publicações de modelos de previsão utilizando regressão logística 35
Tabela 3 – Descrição das variáveis do banco de dados..... 55
Tabela 4 – Valores médios dos centroides obtidos a partir do algoritmo média-K. 59
Tabela 5 – Parâmetros utilizados na construção da rede MLP 7:7-5-1:1. 59
Tabela 6 – Agrupamentos gerados pelo algoritmo média K. 62
Tabela 7 – Comparativo de acuracidade e precisão para os modelos avaliados. 64
Tabela 8 – constantes utilizadas para cálculo do retorno financeiro esperado. 64

LISTA DE SIGLAS E ABREVIAÇÕES

AUC	Área sob a curva de características operacionais do receptor
EMPC	Lucro máximo estimado para *churn* de consumidor
LM	Lucro máximo
LME	Lucro máximo esperado
MLP	*Multilayer perceptron*
NF	Negativos falsos
NV	Negativos verdadeiros
PF	Positivos verdadeiros
PV	Positivos falsos
RNA	Redes neurais artificiais
ROC	Curva de características operacionais do receptor
R-EMP	Lucro máximo esperado robusto
SMOTE	Técnica sintética de *oversampling* da minoria
TDL	*Top decile lift*

SUMÁRIO

1. INTRODUÇÃO 11

1.1. Justificativa da pesquisa 15

1.2. Objetivos 16

1.3. Método de pesquisa 17

1.4. Delimitações da pesquisa 17

1.5. Estrutura do trabalho 18

2. FUNDAMENTAÇÃO TEÓRICA 19

2.1. *Churn* 19

2.2. Modelos de previsão de *churn* 20

2.3. *Frameworks* para previsão de *churn* 24

2.4. Clubes de assinaturas de livros infantis 27

2.5. Análise sistemática da literatura 29

2.6. Técnicas de amostragem 36

2.7. Métodos de previsão e classificação de dados 37

2.7.1. Regressão logística 37

2.7.2. Redes Neurais 39

2.7.3. Média-K 40

3. MÉTODO DE PESQUISA 42

3.1. Definição do método de pesquisa 42

3.2. Objeto de estudo e coleta de dados 43

3.3. Modelo conceitual: procedimento para modelagem de previsão de *churn* em mercados de compra recorrente 43

3.3.1. Fase 1: Preparação dos dados 44

3.3.2. Fase 2: Aplicação dos modelos de previsão 46

3.3.3. Fase 3: Avaliação dos modelos 47

4. RESULTADOS 49

4.1. Fase 1: Preparação dos dados 49

4.1.1. Coleta e pré-tratamento da base de dados 49

4.1.2. Análise exploratória dos dados 49

4.1.3. Aplicação de técnicas de reamostragem e definição da amostra 53

4.1.4. Seleção de variáveis preditoras 55

4.2. Fase 2: Aplicação dos modelos 56

4.2.1. Seleção e aplicação dos métodos de previsão 56

4.2.2. Seleção dos indicadores de avaliação 60

4.3. Fase 3: Avaliação dos modelos 61

4.3.1. Análise dos indicadores de performance 61

4.3.2. Definição do melhor modelo 63

4.3.3. Análise de estabilidade do modelo 65

5. CONCLUSÕES 68

5.1. Contribuições do trabalho 69

5.2. Sugestão de trabalhos futuros 69

6. REFERÊNCIAS BIBLIOGRÁFICAS 71

ANEXOS - PUBLICAÇÕES 80

1. INTRODUÇÃO

O crescimento do mercado de compra recorrente traz um interesse cada vez maior das empresas de bens e serviços para o setor. Neste cenário de alta concorrência, o entendimento do perfil dos consumidores e o que impacta na sua decisão de compra se torna fundamental para a estratégia de crescimento das companhias. Este entendimento é aplicado, principalmente, na construção de estratégias eficiente de retenção de clientes e tem recebido muita atenção do meio prático e acadêmico nos últimos anos (Rahman, Janssens e Bogaert, 2025; Kumar, Lahiri e Dogan, 2018).

O fim do relacionamento de um cliente com uma empresa é chamado de *churn* e a construção de modelos capazes de prever este comportamento com antecedência pode ser um diferencial estratégico para empresas que operam no mercado de compra recorrente (Zhang, Moro e Ramos, 2022). Modelos de previsão de *churn* usam dados disponíveis dos clientes, aplicados através de diferentes técnicas e algoritmos, para prever a probabilidade que cada um tem de cancelar sua próxima compra (Adiputra e Wanchai, 2024).

Existe uma quantidade considerável de metodologias disponíveis para construção de modelos de previsão de *churn*, não existindo na literatura um consenso que abranja as especificidades de cada mercado e perfil de cliente (Ouf, Mahmoud e Abdel-Fattah, 2024). Por isso, muitas vezes, identificar o método mais eficaz para cada situação ou mercado pode levar muito tempo e consumir recursos, o que pode representar uma perda importante de competitividade para as empresas.

A acuracidade dos modelos de previsão sofre influência de uma série de fatores como a força da relação entre as variáveis preditoras e de resposta, a qualidade dos dados utilizados para alimentá-los e a natureza do método de previsão utilizado. Nesse contexto, o pré-tratamento de dados e a seleção das variáveis preditoras de maior impacto na resposta de saída do modelo (*feature engineering*) tem mostrado impacto positivo da acuracidade dos modelos (Ouf, Mahmoud e Abdel-Fattah, 2024).

As concatenações de métodos também têm criado impacto positivo na acuracidade dos modelos, tendo sido bastante explorada nos últimos anos. Nos

métodos compostos por mais de uma etapa (*ensemble*), os dados são inicialmente divididos em grupos que possuem características semelhantes para que posteriormente cada grupo seja submetido individualmente a outros métodos de previsão que fazem a classificação dos casos entre *churn* e não-*churn*.

Dentro do mercado de compras recorrentes, os clubes de assinatura têm se destacado, principalmente pelo seu crescimento de mercado nos últimos anos. Um clube de assinaturas pode ser descrito como um serviço recorrente de fornecimento de produtos específicos, seguindo intervalos fixos a partir de pagamentos recorrentes que podem ser mensais, trimestrais ou anuais (Kerschbaumer *et al.*, 2023). Esse tipo de serviço surge da mudança dos hábitos de consumo e da popularização do comércio *online*, e pode ser dividido em duas categorias: aqueles em que o assinante seleciona o produto que vai receber todos os meses e aqueles nos quais o produto é selecionado por especialistas, baseado nas características individuais de cada assinante (Li, Zhao e Ou, 2023).

Em uma pesquisa de mercado realizada por Jabil (2021), na qual foram entrevistados 1.124 consumidores americanos que realizaram pelo menos 5 compras *online* no ano anterior, foi mostrado que 68% dos entrevistados utilizam serviços de assinatura para compra de materiais domésticos como produtos de higiene pessoal, roupas e alimentos. A pesquisa mostra que esse tipo de compra é especialmente popular entre *millenials* e a geração Z.

A conveniência e praticidade são os principais motivos que levam os consumidores a tornarem-se clientes de empresas de fornecimento recorrente, como serviços por assinatura focados em reposição de produtos (Figura 1). Outros fatores como descontos concedidos à assinantes e a necessidade de armazenar menos produtos em casa também cumprem um papel importante na conquista de novos clientes (Jabil, 2021).

Apesar de todas as características de praticidade e do crescimento rápido de faturamento, a maior fatia do mercado de clubes de assinatura não está focada no reabastecimento de materiais de uso diário, mas sim no fornecimento de produtos curados por especialistas (Kumar, 2022). De acordo com Gupta, Eilert e Gentry (2020), para alguns consumidores, o principal motivo de adesão a clubes de assinatura está relacionado à surpresa de receber algo que foi curado por especialistas, ao fator surpresa associado ao fornecimento

recorrentemente que eleva e diferencia a experiência de compra, e ao crescente hábito de se auto presentear observado nos consumidores que buscam esse tipo de mercado. Este hábito fez com que o mercado de clubes de assinatura apresentasse um crescimento de faturamento significativo nos últimos anos e chamasse a atenção de diversos varejistas, incluindo grandes redes como Wallmart, Procter & Gamble e Sephora, que criaram seus próprios serviços de fornecimento por assinatura, configurando um novo e lucrativo modelo de negócio (Li, Zhao e Ou, 2023).

Figura 1 – Motivos de adesão a clubes de assinatura para consumidores dos EUA
Fonte: Adaptado de Jabil (2021)

Assim como no mercado varejista, existe uma diversidade grande de produtos que podem ser adquiridos de forma recorrente através dos clubes de assinatura. Andonova, Anaza e Bennett (2021) classificam as empresas que compõem esse mercado de acordo com a natureza e uso dos seus produtos, sendo as principais: produtos de beleza e higiene pessoal; refeições pré-preparadas; alimentos e bebidas; vestuário; itens infantis; artigos para animais de estimação; e produtos de uso doméstico diário. Cada um desses mercados busca atingir um público diferente atendendo necessidades específicas do dia a dia ou trazendo uma experiência de presentear.

O mercado de clubes de assinatura possui uma forte característica de recorrência e fidelização, sendo que o período de permanência dos clientes na base de assinantes e o alto custo de aquisição de novos clientes são fatores importantes que precisam ser acompanhados de perto pelos gestores desse tipo de negócio (Perišić e Pahor, 2022). Mesmo sob essas condições, esse mercado

apresentou um crescimento significativo na última década, faturando mais de 2,6 bilhões de dólares em 2016 somente no mercado americano (Chen *et al.*, 2018) e tendo apresentado um crescimento ainda maior em 2020, motivado principalmente pela pandemia do COVID-19 (Seetharaman, 2020). O faturamento global do segmento foi de U$ 28,1 bilhões em 2022 e tem previsão de crescer 16,5% ao ano atingindo U$ 73,6 bilhões em 2028 (IMARC, 2023).

Com esse crescimento expressivo, os clubes de assinatura se destacaram como um negócio rentável e isso trouxe diversas novas empresas para o mercado, tornando-o ainda mais competitivo (Cheng, 2019). A grande variedade de opções disponíveis para os assinantes torna fácil a troca de um fornecedor por outro. Isso fez com que a taxa de cancelamento de assinaturas (*churn*) crescesse de forma significativa nos últimos anos (Perišić, Jung e Pahor, 2022). Para lidar com essa realidade de mercado, gestores e acadêmicos tem se dedicado ao desenvolvimento de modelos de previsão capazes de identificar clientes com alta probabilidade de cancelamento (Kiguchi e Medi, 2022).

Os clubes de assinatura de livros são pioneiros neste modelo de negócio, sendo os primeiros que se tem registro, datando do início do século XIX, e vêm se adaptando desde então às mudanças sociais e tecnológicas para manterem-se relevantes. Atualmente, apesar de existirem centenas de empresas focadas neste nicho, esse ainda é um mercado muito pequeno, principalmente nos mercados emergentes (Lehmann, 2020). Norrick-Rühl (2020) afirma que o apelo dos clubes de assinatura de livros se sustenta atualmente com base nos “4 C’s”:

- **Curadoria:** em meio a uma disponibilidade cada vez maior de conteúdo, principalmente na era da internet, a seleção baseada nas preferências individuais de cada assinante passa a ser um serviço muito valorizado.
- **Conveniência:** receber os livros em casa periodicamente para manter um fluxo contínuo de leitura economiza horas de pesquisa para o assinante e traz praticidade a sua rotina.
- **Concessão:** o acesso a livros e conteúdo adicional a custo reduzido faz com que os leitores mais frequentes busquem clubes de assinatura como forma de economia.

- **Comunidade:** o senso de pertencimento proveniente do compartilhamento de preferências e experiências é também um atrativo para os clientes dos clubes de assinatura de livros.

Noorda (2019) realizou uma pesquisa com fundadores de clubes de assinatura de livros infantis no mercado americano e buscou entender as características desse mercado. Segundo a autora, o desenvolvimento desse mercado começa após o ano de 2010, puxado, principalmente, por mulheres (mães e avós) que buscam complementação educacional e desenvolvimento do hábito de leitura para as crianças de 0 a 12 anos. O sentimento de surpresa ao receber produtos curados e o ato de presentear também são características importantes que levam clientes a buscar serviços de recorrência como clubes de assinatura (Kumar, 2022). Para manter-se relevante e atrair novos consumidores que buscam uma experiência mais digital, muitas empresas têm lançado produtos como *e-books* e audiolivros (Noorda, 2019).

1.1. Justificativa da pesquisa

A quantidade de métodos que podem ser combinados para criação de modelos de previsão de *churn* faz com que a definição do modelo mais eficiente possa ser muito extensa e trabalhosa, principalmente em situações nas quais a aplicação de pré-tratamento de dados e técnicas para reamostragem precisam ser levados em consideração. Em geral, o processo envolve o teste de diversas hipóteses e pode levar a um consumo excessivo de tempo e recursos computacionais para se determinar o melhor, entre diversos métodos, que otimiza o modelo.

Somando-se a isso, as pesquisas relacionadas às características de mercado de compra recorrente são escassas (Li, Zhao e Ou, 2023), sendo essa lacuna de conhecimento ainda maior no mercado de clubes de assinatura (Noorda, 2019).

Quando analisado dentro do contexto brasileiro, nota-se uma ausência de publicações relacionadas ao tema. Para exemplificar, ao se fazer uma busca na base SCOPUS utilizando os termos “*children* AND *book* AND *subscription*”, apenas um dos artigos listados aborda temas relativos ao mercado brasileiro. O

trabalho em questão é um estudo de caso sobre uma *startup* e não aborda o tema do *churn*.

Além disso, o mercado de clubes de assinatura tornou-se altamente competitivo nos últimos anos e a taxa de cancelamentos de assinatura (*churn*) tem aumentado (Cheng, 2019; Ahmad, Jafar e Aljoumaa, 2019). Neste cenário, modelos de previsão capazes de classificar assinantes com base no seu risco de cancelamento tornam-se um diferencial estratégico, pois permitem a construção de campanhas eficientes para retenção de clientes (Amin et al., 2016; Sundarkumar e Ravi, 2015). Uma maior taxa de retenção reverte-se em receita adicional para as empresas, visto que o custo de aquisição de novos assinantes é consideravelmente superior ao custo de retenção dos clientes atuais (Bahnsen, Aouada e Ottersten, 2015).

1.2. Objetivos

Tendo em vista o que foi contextualizado anteriormente, o presente trabalho tem como objetivo geral desenvolver um procedimento estruturado, na forma de um *framework*, para a construção de modelos de previsão de *churn*, e aplicá-lo de forma prática a dados reais de uma base de assinantes de um clube de assinatura de livros, indicando os assinantes com maior probabilidade de cancelamento do serviço.

Como objetivos específicos, ficam estabelecidos os seguintes tópicos:

- Estabelecer as lacunas de conhecimento referente à previsão de *churn* em modelos de negócio de compra recorrente, por meio de uma revisão sistemática dos artigos da literatura, e com destaque para os clubes de assinaturas de livros;
- Estruturar um procedimento que leve em consideração não somente a aplicação de diferentes modelos para a previsão de *churn*, mas também uma devida preparação dos dados, e análises comparativas que sustentem a definição do melhor modelo.
- Testar a aplicabilidade do *framework* criado para a previsão do *churn* mensal no contexto gerencial de um clube de assinaturas de livros;

- Avaliar a estabilidade em períodos subsequentes para o melhor modelo definido;
- Analisar as implicações técnicas e científicas das previsões obtidas com o processo estruturado frente ao objeto de estudo da pesquisa.

1.3. Método de pesquisa

Neste trabalho, foi utilizado o método de pesquisa de Modelagem e Simulação, adaptado de Bertrand e Fransoo (2002).

A empresa objeto de estudo trata-se de um clube de assinatura de livros focada no público infantil. O produto fornecido é composto por uma combinação de livros e itens lúdicos curados e selecionados para sete faixas etárias diferentes, abrangendo crianças de 0 até 12 anos de idade. Atendendo todo o território nacional, a empresa entrega atualmente, em mais de 5 mil municípios, mais de 180 mil kits todos os meses. Sendo assim, devido à importância estratégica que o controle do *churn* tem para o negócio, essa empresa desenvolveu uma ferramenta de previsão que classifica os seus assinantes de acordo com o risco de cancelamento. Porém, a acuracidade real desse modelo é muito baixa.

Quanto à coleta dos dados, primeiramente, foram mapeados os critérios utilizados para previsão e a performance do modelo atual aplicado pela empresa. Em seguida, os dados de todos os assinantes que compõem a base da empresa foram levantados, considerando o mês de setembro de 2023 como referência. Na aplicação do framework desenvolvido, esses dados foram utilizados como variáveis preditoras em modelos diferentes que foram comparados entre si. Por fim, para fins de validação dos resultados, a qualidade das previsões do modelo definido pelo *framework* foi comparada com a do atualmente utilizado pela empresa.

1.4. Delimitações da pesquisa

O escopo desse trabalho abrange a elaboração de um método estruturado, na forma de um *framework*, que possa ser aplicado na construção de modelos

de previsão de *churn* em mercados de compra recorrente, com foco, principalmente, em clubes de assinatura.

Além disso, será realizada a sua aplicação através do desenvolvimento de um modelo de previsão de *churn* utilizando dados reais de um clube de assinaturas de livros infantis. A qualidade do modelo será medida através de indicadores frequentemente utilizados na literatura, que são: acuracidade, precisão, *recall* e medida-F1.

Não estão incluídos nesse escopo compras esporádicas de produtos avulsos ou novos produtos que venham a ser lançados. Por fim, as especificidades do mercado editorial infantil e dos clubes de assinatura no Brasil, fazem com que o modelo construído seja dedicado especificamente a simular esse contexto de negócio, não sendo necessariamente aplicável a outros mercados.

1.5. Estrutura do trabalho

A presente dissertação foi estruturada na forma como se segue. Neste capítulo são apresentados os conceitos básicos que definem um clube de assinatura, um contexto geral do seu mercado, o desafio que a previsão de *churn* representa, e os objetivos, método e delimitações dessa pesquisa.

No Capítulo 2 é apresentada a fundamentação teórica, com a definição de *churn*, quais as principais ferramentas para construção e avaliação de modelos de previsão, a fundamentação básica dos métodos de previsão utilizados e a análise sistemática da literatura construída a partir da base SCOPUS (Elsevier).

O Capítulo 3 é dedicado ao método de pesquisa que foi utilizado para atingir-se os objetivos. No Capítulo 4, são apresentados os resultados obtidos após a aplicação do *framework* elaborado, a partir dados reais de um clube de assinatura de livros infantis.

Por fim, no Capítulo 5 constam as conclusões, contribuições da pesquisa e sugestões para investigações futuras.

2. FUNDAMENTAÇÃO TEÓRICA

2.1. *Churn*

Churn pode ser definido como o fim do relacionamento de uma empresa com o seu cliente, que muitas vezes se traduz na perda desse cliente para algum concorrente (Zhang, Moro e Ramos, 2022). É uma expressão muito utilizada para representar o encerramento do contrato entre um cliente e uma empresa no modelo de compra recorrente, incluindo os clubes de assinatura.

O *churn* tem um alto impacto no crescimento e na rentabilidade de um negócio, visto que a variação do tamanho da base de clientes de um clube de assinatura pode ser definida como a diferença entre o número de novas assinaturas vendidas e o de clientes que deixaram a base dentro de um determinado período (Bahnsen, Aouada e Ottersten, 2015). Primariamente, o *churn* pode ser classificado de acordo com o comportamento do assinante. No *churn* voluntário, a decisão de cancelamento da assinatura parte do cliente, geralmente movido por alguma insatisfação. Já no *churn* involuntário, a decisão parte da empresa, majoritariamente motivado por inadimplência (Lu *et al.*, 2014).

Os motivos que levam ao cancelamento de um serviço mudaram conforme as preferências do consumidor evoluíram. Inicialmente, o principal motivo de troca era o custo associado ao produto ou serviço, já atualmente os clientes valorizam mais a diferenciação e o valor agregado à experiência de consumo daquilo que estão comprando (Zhang, Moro e Ramos, 2022).

Uma maior acuracidade na previsão do crescimento da base de assinantes permite a operação de uma cadeia produtiva mais eficiente e com custos mais baixos em um mercado cada vez mais competitivo, principalmente considerando-se que o custo de aquisição de novos clientes é superior ao custo de retenção dos assinantes atuais (Xiahou e Harada, 2022).

Fader e Hardie (2009) afirmaram que o *churn* se comporta de maneiras diferentes dependendo se há ou não um contrato firmado entre o consumidor e a empresa. Segundo os autores, o tempo de inatividade, ou o tempo entre compras, é o conceito mais utilizado para classificar um cliente como *churn* em mercados não-contratuais. Porém, isso não reflete a realidade da relação do

consumidor com o negócio, visto que esse intervalo é arbitrário e pode variar muito para cada modelo de negócio. Essas características fazem com que o *churn* não possa ser definido como algo permanente, especialmente em modelos de negócio de compra recorrente, visto que nada impede um consumidor "inativo" de voltar a fazer negócio com a empresa (Tamaddoni, Stakhovych e Ewing, 2016).

2.2. Modelos de previsão de *churn*

Modelos de previsão de *churn* de alta acuracidade e que produzem dados de fácil interpretação são um diferencial estratégico para as empresas no mercado de compras recorrentes (Boukrouh e Azmani, 2025). De Caigny e De Bock (2018) listaram os principais impactos de negócio que têm as ações de retenção traçadas a partir da interpretação das respostas geradas por esses modelos:

- A construção de uma relação de longo prazo com o cliente;
- Aumento das oportunidades de venda;
- Oportunidade de conhecer melhor o perfil da base de assinantes;
- Menor necessidade de investimento em marketing para atração de novos assinantes;
- Fortalecimento do crescimento orgânico através da divulgação que seus próprios assinantes fazem dentro do seu círculo social.

Muitas abordagens diferentes são empregadas para construção de modelos de previsão de *churn*. As mais comuns têm como saída uma pontuação que indica o quão provável é que um assinante realize o cancelamento da assinatura, sendo que o limiar que separa essa escala de pontuação entre as duas classes fica, muitas vezes, a critério do gestor do negócio ou do pesquisador (Garrido, Verbeke e Bravo, 2018).

Além da probabilidade de cancelamento, outra informação importante que os modelos de previsão trazem aos tomadores de decisão são os motivos que

levam ao cancelamento, trazendo mais conhecimento sobre os perfis de clientes para dentro do negócio e tornando possíveis que as estratégias de retenção sejam mais focadas e eficientes (Li e Li, 2019).

Diante da variedade de métodos que vem sendo construídos e divulgados, é preciso definir critérios que permitam as comparações entre eles. A maioria dos gestores buscam modelos que performam melhor nas condições específicas do seu negócio. Um melhor desempenho traz mais confiabilidade ao método e ajuda na tomada de decisão em situações reais (Garrido, Verbeke e Bravo, 2018).

Existem diversas métricas utilizadas para medir a qualidade dos modelos de classificação. Dentre elas, uma das mais utilizadas é a matriz de confusão ou matriz de erro, pois permite uma visualização rápida do desempenho de classificação para modelos binários (Long *et al.*, 2019). A matriz de confusão pode ser montada conforme mostra a Tabela 1. Nela é possível observar os Positivos Verdadeiros (PV), que são os casos previstos pelo modelo como verdadeiros e que realmente são positivos; Negativos Verdadeiros (NV), que são os casos previstos como negativos e que realmente são negativos; Falsos Positivos (PF), que são os casos previstos como positivos, mas que na realidade são negativos (esses dados caracterizam o erro do tipo I ou α); e os Falsos Negativos (NF), que são os casos previstos como negativos, mas que na realidade são positivos (esses dados caracterizam o erro do tipo II ou β) (Long *et al.*, 2019).

Tabela 1 – Matriz de confusão ou matriz de erro

		Categoria prevista	
		Churn	**Não-*churn***
Categoria verdadeira	***Churn***	PV	NF
	Não-*churn*	PF	NV

PV: Positivos verdadeiros; NV: Negativos verdadeiros; FP: Falsos positivos; FN: Falsos negativos

Com os dados fornecidos pela matriz de confusão é possível calcular algumas métricas que permitem comparar modelos. De acordo com Lalwani et al. (2022), as principais são:

- ***Recall*:** taxa de positivos verdadeiros encontrada dentro de todos os positivos reais (Equação 1):

$$Recall = \frac{P_v}{P_v + N_f} \quad (1)$$

- **Precisão:** taxa de positivos verdadeiros dentro de todos os positivos indicados pelo modelo (Equação 2):

$$Precisão = \frac{P_v}{P_v + P_f} \quad (2)$$

- **Acuracidade:** taxa de previsões positivas corretas geradas pelo modelo (Equação 3):

$$Acuracidade = \frac{(P_v + N_v)}{(P_v + N_v + P_f + N_f)} \quad (3)$$

- **Medida F1:** valor que relaciona precisão e *recall*, representando a capacidade geral de previsão do modelo (Equação 4):

$$Medida\ F1 = \frac{(2 \times Precisão \times Recall)}{(Precisão + Recall)} \quad (4)$$

Os valores de *recall* e precisão, isoladamente, podem levar a análises equivocadas dos resultados de um modelo de previsão (Imani e Arabnia, 2023), principalmente quando a proporção de uma das classes dentro da população é significativamente menor que a outra. Por isso, a Medida F1 (*F1-score*) é mais utilizada, pois mostra o balanceamento entre o *recall* e a precisão (Kiguchi, Saeed e Medi, 2022).

Para modelos de previsão de *churn*, outro método de avaliação utilizado é a área sob a curva de características operacionais do receptor (AUC, sigla em inglês para *Area Under the Curve*) (Imani e Arabnia, 2023). Esta curva de características operacionais do receptor (curva ROC, sigla em inglês para *Receiver Operator Characteristics*) é construída plotando no eixo *x* o índice de

falsos positivos e no eixo *y* o índice de positivos verdadeiros, para valores limiares (*thresholds*) diferentes (Kim, Jun e Lee, 2014).

A área sob a curva ROC é um valor entre 0 e 1, que representa de forma compreensível a qualidade do modelo de previsão ou de classificação. Considerando que um classificador binário aleatório teria AUC igual 0,5 e um classificador binário perfeito teria AUC igual a 1, como ilustra a Figura 2, é possível afirmar que quando se comparam 2 ou mais modelos de classificação binária, aquele que apresentar o maior valor de AUC pode ser considerado como o melhor (Lu *et al.*, 2014).

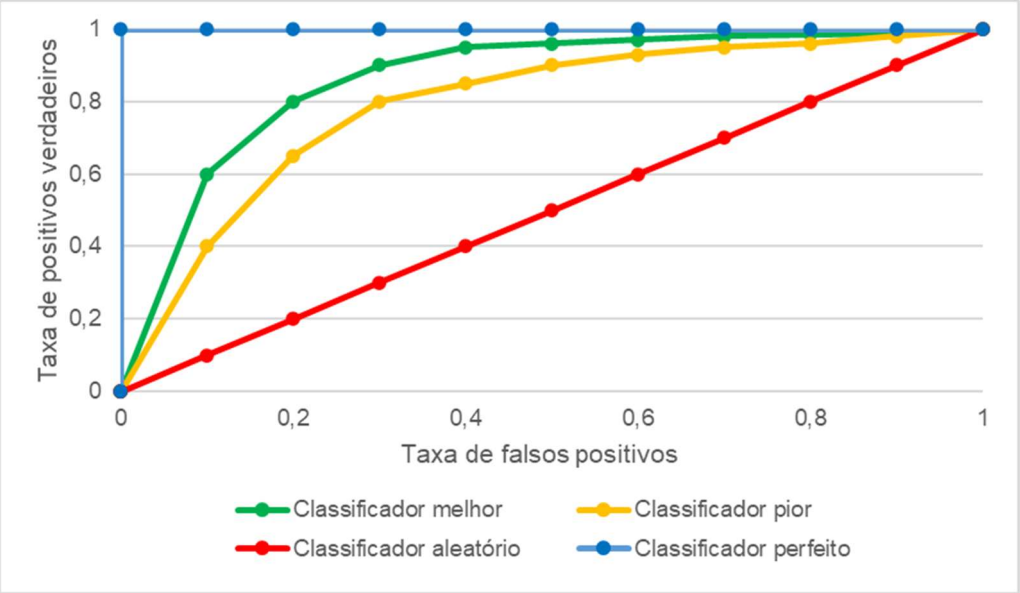

Figura 2 – Curva de características operacionais do receptor
Fonte: Adaptado de Lu et al., 2014

Um problema relacionado ao uso da AUC como a principal métrica de comparação entre modelos é que este método não considera o custo para o negócio relacionado às classificações corretas e incorretas (Garrido, Verbeke e Bravo, 2018). Para incorporar o impacto financeiro das classificações incorretas na avaliação dos modelos, Verbraken, Verbeke e Baesens (2013) propuseram duas abordagens diferentes, o Lucro Máximo (LM) e o Lucro Máximo Esperado (LME).

O cálculo do LM é feito considerando a taxa de retenção de assinantes, o custo da campanha (somando descontos oferecidos, custo de produtos adicionais e até mesmo do time de retenção necessário para a sua

operacionalização) e a proporção de clientes que receberão o incentivo para manter a assinatura, mas que não tinham intenção inicial de cancelá-la (a taxa de falsos positivos do modelo). Com esses dados, é possível estimar a lucratividade da campanha de retenção desenvolvida com base no modelo de previsão. Modelos mais acurados terão um maior retorno sobre o investimento.

Já o cálculo do LME é feito de maneira similar. Porém, ao invés de valores determinísticos para as variáveis citadas, são utilizadas distribuições de probabilidade para cada umas delas. Dessa forma, obtém-se uma distribuição de probabilidade para a lucratividade real da campanha de retenção.

2.3. *Frameworks* para previsão de *churn*

Existe uma considerável variedade de métodos e ferramentas disponíveis na literatura para previsão de *churn*. A aplicação individual ou a combinação entre métodos tem sido estudada em diferentes mercados, variando desde os mais tradicionais, como a regressão logística, até métodos altamente customizados para performar nas condições específicas de um modelo de negócio.

A construção de modelos através de dados brutos, ou seja, sem nenhum tipo de tratamento, pode prejudicar a qualidade das previsões ao inserir dados que não representam a realidade e que geram vieses falsos no modelo. Rao et al. (2020) estudaram os impactos da normalização, padronização e limpeza dos dados de entrada como uma etapa de pré-processamento na construção de um modelo de previsão de *churn.* Ao submeterem os dados de entrada por um pré-tratamento, os autores obtiveram acuracidades de 95% nas previsões utilizando regressão logística, redes neurais e algoritmos de aprendizado de máquina como *random forest*, *support vector machine* e *K-nearst neighbors*. A Figura 3 apresenta uma adaptação do *framework* utilizado pelos autores.

Além disso, métodos compostos por duas ou mais etapas têm sido estudados como uma forma de melhorar a qualidade das previsões. Em casos de classificação binária, como a identificação de *chun* e não-*churn*, realizar uma etapa de separação dos dados entre clientes com maior e menor probabilidade de cancelamento e posteriormente criar modelos específicos para cada grupo

mostrou ser capaz de melhorar a qualidade das previsões (Xiahou e Harada, 2022)

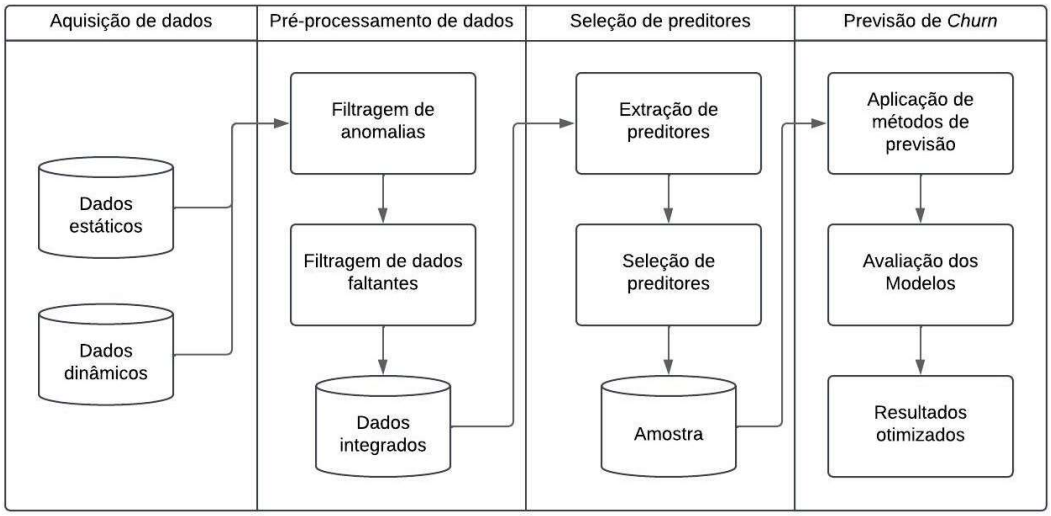

Figura 3 – *Framework* para construção de modelos de previsão de *churn*
Fonte: Adaptado de Rao et al. (2020)

Ouf, Mahmoud e Abdel-Fattah (2024) propuseram um *framework* para construção de modelos de previsão de churn em três etapas (Figura 4). Na primeira delas os dados são tratados e é realizada a seleção das variáveis preditoras com maior impacto na variável de resposta (*feature engineering*). Na segunda etapa, são aplicadas ferramentas de amostragem para minimizar o impacto do desbalanceamento das classes. As amostras construídas são então submetidas aos métodos selecionados de previsão. Por fim, na terceira etapa, as previsões realizadas por cada método são usadas para cálculo dos indicadores de qualidade dos modelos: acuracidade, precisão, *recall* e medida-F1. O melhor modelo é então selecionado baseado na análise dos indicadores.

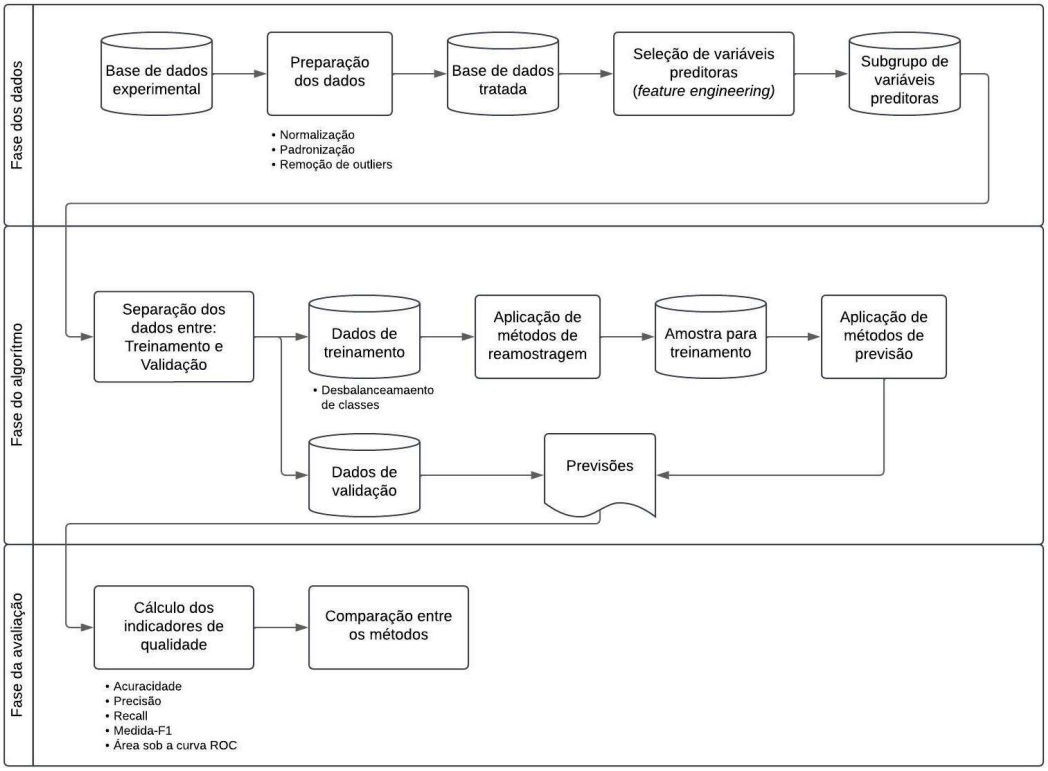

Figura 4 – *Framework* para comparação entre modelos de previsão de churn
Fonte: Adaptado de Ouf, Mahmoud e Abdel-Fattah (2024).

Guo et al. (2025) propuseram um método de previsão de churn para jogos eletrônicos capaz de processar grandes quantidades de dados (260 gigabites) de forma rápida e eficiente. Os autores destacam a importância do pré-tratamento dos dados e utilizam a regra de limpeza de vizinhança (*neighborhood cleaning rule*) para lidar com o problema do desbalanceamento das classes. As amostras construídas foram submetidas ao método *Spark* que obteve 96% de acuracidade na previsão, superando métodos tradicionais como regressão logística (82%) e redes neurais *multilayer perceptron* (77%).

Analisando os *frameworks* propostos na literatura, é possível notar que, apesar da diversidade de métodos aplicados, algumas etapas parecem contribuir de forma positiva para determinação do melhor modelo de previsão de *churn*. A definição de métricas abrangentes para comparação entre modelos, o pré-tratamento dos dados e aplicação de técnicas de amostragem para casos de desbalanceamento de classes mostraram ter impacto positivo na qualidade das previsões e devem ser incluídos em processos que buscam identificar modelos mais otimizados.

A especificidade do mercado é outro fator que precisa ser levado em consideração. Quando se avalia o mercado de clubes de assinatura, por exemplo, não foram encontrados na literatura *frameworks* capazes de incorporar as características específicas deste modelo de negócio. O presente trabalho pretende contribuir com essa lacuna, propondo um método prático para avaliação de modelos de previsão de *churn* em mercados de compra recorrente.

2.4. Clubes de assinaturas de livros infantis

Dentro do cenário nacional, os clubes de assinatura de livros representam 25% dos clubes de assinatura e apresentaram um crescimento significativo entre 2019 e 2020, saindo de um faturamento R$ 13,17 milhões para R$ 36,15 milhões (Lima, 2022). Além disso, a representatividade dessas empresas para o mercado editorial no Brasil dobrou, saindo de 1,08% para 2,41%.

Porém, em 2021, o mercado sofreu com o cenário macroeconômico nacional, tendo uma queda de 40% no total de livros enviados (Lima, 2022). O cenário editorial como um todo continua desafiador. Comparando o faturamento do setor em 2022 com 2023, houve uma redução de 2,69%, mesmo com um aumento de 12,38% no preço médio dos livros (NIELSEN BOOKSCAN, 2024).

Em uma pesquisa encomendada pela Câmara Brasileira do Livro (CBL) em 2023, foram levantadas informações sobre o perfil de compradores de livros no Brasil, incluindo informações importantes sobre clubes de assinatura. Algumas características importantes são listadas a seguir (NIELSEN BOOKDATA, 2023):

- 49% dos entrevistados consomem livros exclusivamente através dos clubes de assinatura;
- Assim como no mercado americano, as mulheres são a maioria dos assinantes, representando 57% dos entrevistados;
- Livros impressos representam 49% dos tipos de assinatura, enquanto livros digitais correspondem a 39%, e audiolivros a 12%;
- Apesar de ser liderado por um clube de assinatura de livros infantis, o mercado é predominantemente focado no público adulto e juvenil;

- O hábito de leitura compete pela atenção dos entrevistados principalmente com serviços de *streaming* e uso de redes sociais.

Analisando o mercado de clubes de assinatura de livros infantis, Noorda (2019) realizou um estudo no mercado americano. A autora entrevistou proprietários e gestores que trabalhavam nesse ramo e traçou um panorama geral do produto oferecido, das principais características e dos principais desafios enfrentados pelas empresas. As conclusões gerais desse estudo são listadas a seguir:

- Existe uma preferência do público infantil por livros de capa dura;
- Itens extras adicionados ao *kit* são considerados um diferencial pelos clientes. Dentre esses itens, os principais são marca-páginas, cadernos e brinquedos relacionados à temática de leitura;
- Para a compra de livros, as empresas buscam parcerias com editoras grandes, pois precisam comprar volumes maiores a preços mais baixos que o varejo tradicional (livrarias);
- Já para os itens extras, a preferência é por parcerias locais com negócios menores;
- A maioria dos clubes de assinatura de livros infantis são empresas pequenas, com menos de 10 funcionários;
- As pessoas que mais buscam pelos clubes de assinatura são mulheres, geralmente mães ou avós. Além disso, o produto tem um apelo especial para pais que não podem passar muito tempo com os filhos;
- Plataformas de *e-commerce* são um dos principais canais de vendas para esse modelo de negócio. Segundo alguns entrevistados, essas plataformas auxiliam não só na divulgação, mas também na operacionalização das vendas de novas assinaturas.

A curadoria de livros por especialistas em desenvolvimento infantil, a busca por formas diferenciadas de presentear e o fator surpresa associado ao

recebimento dos produtos, são os principais motivos que levam à busca pelo serviço de clubes de assinatura de livros infantis (Noorda, 2019).

2.5. Análise sistemática da literatura

Para uma verificação na literatura sobre como o tema dessa pesquisa tem sido tratado por outros autores, foi realizado um levantamento na base de dados SCOPUS (Elsevier), utilizando os seguintes termos de busca*: churn AND (prediction OR forecast)*. Com isso, foi possível evidenciar um crescimento no número de publicações sobre modelos de previsão de *churn* na última década, conforme ilustrado graficamente na Figura 5. Esse comportamento indica uma preocupação crescente de pesquisadores e gestores de mercado com a eficiência das campanhas de retenção e de construção de relacionamento com o cliente, buscando modelos que permitam que as estratégias se tornem cada dia mais eficientes (Boozary *et al.*, 2025).

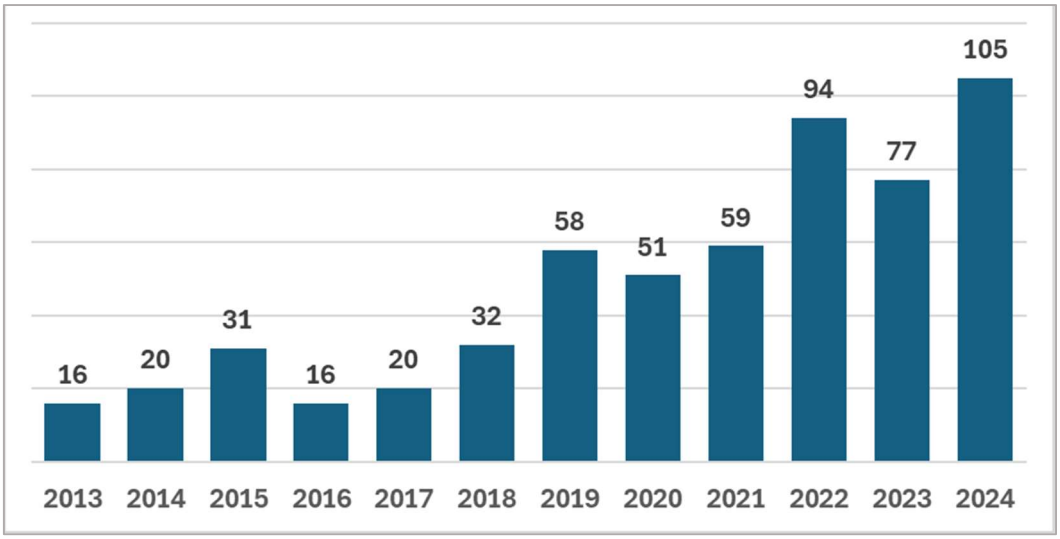

Figura 5 – Evolução do número de publicações sobre previsão de *churn* na última década
Fonte: autoria própria a partir da busca na base SCOPUS

Considerando o número inicial de artigos encontrados (1879 publicações), a Figura 6 descreve os critérios utilizados para filtrar e identificar as publicações mais relevantes. O mercado de telecomunicações desponta como aquele que mais recebeu atenção dos pesquisadores nos últimos dez anos, seguido do mercado bancário e *e-commerce* (Figura 7). Essa grande proporção do mercado

de telecomunicações, principalmente dos serviços de telefonia celular e canais por assinatura, é justificada pela maturidade do mercado, que apresenta taxas de crescimento mais baixas quando comparadas aos mercados mais novos, como o *e-commerce*, e a alta competitividade entre os principais agentes que torna fácil para o assinante mudar de um provedor para outro.

Nota-se que, apesar de ter crescido nos últimos anos, o modelo de negócio de clubes de assinatura permanece pouco explorado pela comunidade científica, o que evidencia a principal lacuna para as contribuições deste trabalho. A construção de um método estruturado para definição de um modelo de previsão de boa acuracidade e que gere dados de interpretação de maneira mais simplificada pode ajudar pesquisadores e gestores desse tipo de negócio a criar estratégias de retenção eficientes para redução das taxas de *churn*.

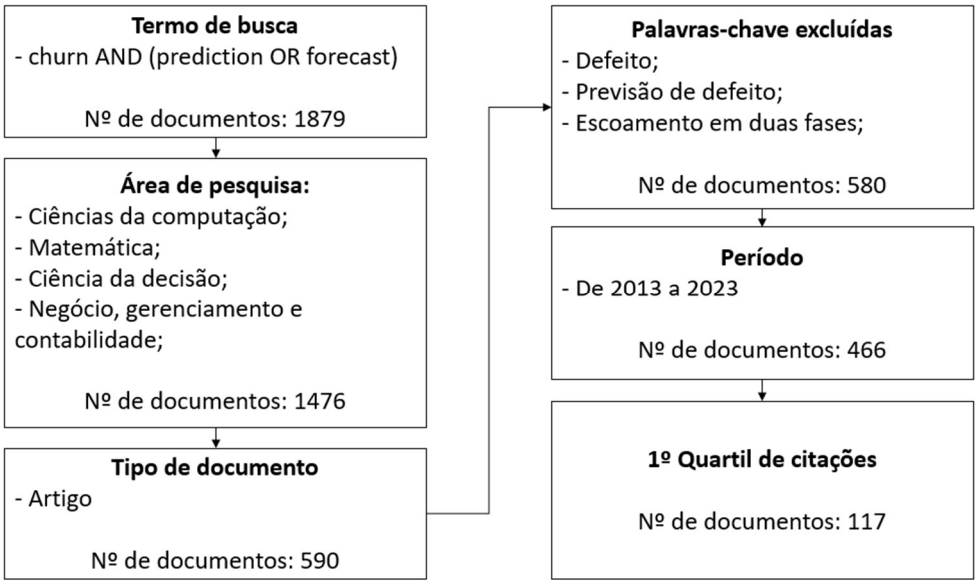

Figura 6 – Critérios para refinamento da base de pesquisa

Existem várias ferramentas para classificar os assinantes de acordo com o risco de cancelamento. De Caigny e De Bock (2018) atingiram uma AUC de 0,887 aplicando um modelo de previsão de *churn* com uma abordagem em dois estágios. No primeiro, os consumidores foram divididos em grupos utilizando o algoritmo de árvore de decisão e, no segundo, foi construído um modelo de previsão específico utilizando regressão logística para cada grupo. Lu et al. (2014) seguiram uma abordagem semelhante, porém dividindo os clientes em apenas dois grupos, de alto e baixo risco de cancelamento, e aplicando a

regressão logística em cada um desses grupos separadamente para criar um modelo de previsão específico, atingindo uma AUC de 0,723.

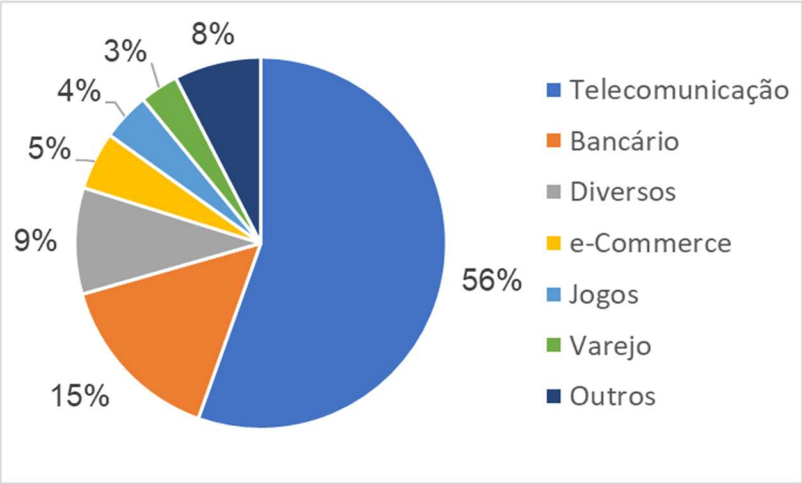

Figura 7 – Principais mercados estudados para previsão de *churn*
Fonte: autoria própria a partir da busca na base SCOPUS

Vafeiadis et al. (2015) compararam o desempenho de diversos tipos de ferramentas como regressão logística, redes neurais artificiais, *support vector machine* (SVM) e Naive Bayes para a previsão de *churn* no mercado de telecomunicações. Os autores buscaram entender o impacto do uso de algoritmos de *boosting* na capacidade preditiva do modelo e atingiram 97% de acuracidade e uma medida F1 de 84%, utilizando uma combinação de SVM e AdaBoost. Também no mercado de telecomunicações, Amin et al. (2016) abordaram o impacto do problema de desbalanceamento de classes, comparando a acuracidade de diferentes algoritmos de amostragem, mostrando que a função de difusão *mega-trend* (FDMT) teve performance superior ao *over* e *undersampling*, atingindo uma acuracidade de 94,6%.

A Figura 8 mostra as principais ferramentas utilizadas por pesquisadores dentre os trabalhos analisados. A maioria dos trabalhos propõem modelos com configurações personalizadas e compostos de diversas etapas focadas principalmente no tratamento dos dados brutos e na segmentação dos clientes em grupos com maior e menor probabilidade de cancelamento. Alguns dos artigos mais citados utilizaram ferramentas de agrupamento como *Random Forest*, *Social Network Analysis* e *Supporting Vector Machine*, associadas à regressão logística, para construir modelos que tem alta performance de

classificação binária (Caigny, De Coussement e Bock, 2018; Ullah et al., 2019; Amin et al., 2019).

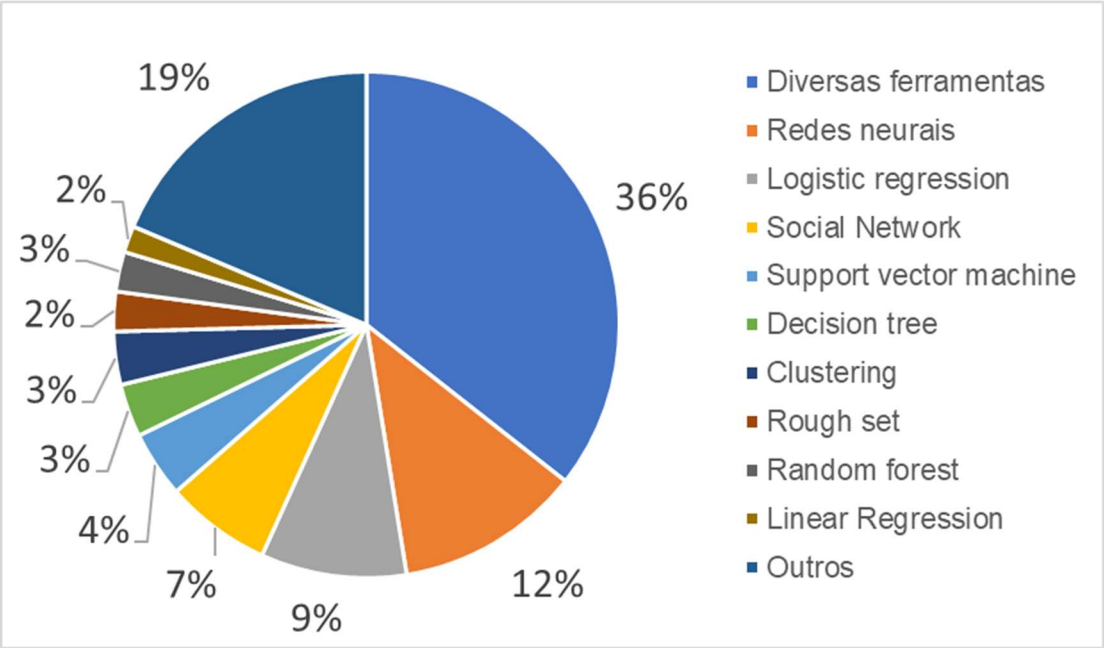

Figura 8 – Principais ferramentas utilizadas para previsão de *churn*
Fonte: autoria própria a partir da busca na base SCOPUS

Essa combinação de métodos permite criar modelos específicos para diferentes segmentos de dados, criando um sistema de previsão em duas etapas (Lalwani et al., 2022; Pustokhina et al., 2021). A Tabela 2 apresenta um resumo dos trabalhos publicados que utilizaram a regressão logística como ferramenta principal de previsão ou como referência para comparação com outros métodos. É importante destacar que entre os trabalhos mais citados listados na base SCOPUS, não há nenhum tratando do mercado de clubes de assinatura de livros, mostrando uma lacuna importante de conhecimento que este trabalho pretende abordar.

Técnicas de aprendizado de máquina tem sido cada vez mais utilizadas para previsão de *churn*, principalmente por serem capazes de aprender através do processamento de grandes bancos de dados (Gattermann-Itschert e Thonemann, 2021). Esses modelos são baseados tanto em características estáticas quanto dinâmicas dos assinantes (Gür Ali e Arıtürk, 2014).

Informações como valor da assinatura, tipo de plano e custo de frete, por exemplo, são consideradas características estáticas, enquanto o histórico de

compra e inadimplência, por outro lado, são consideradas características dinâmicas. Gür Ali e Arıtürk (2014) fizeram um levantamento dos principais grupos de parâmetros utilizados para compor modelos de previsão de *churn* e concluíram que critérios relacionados ao comportamento e a distribuição demográfica dos assinantes são os mais utilizados para o desenvolvimento de modelos de previsão (Figura 9).

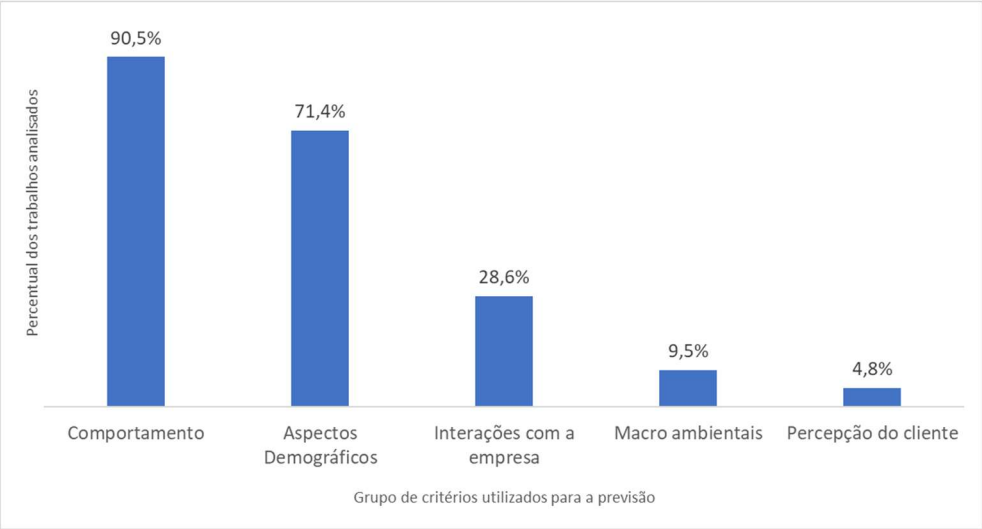

Figura 9 – Principais grupos de variáveis preditoras utilizadas em modelos de previsão de *churn*
Fonte: Adaptado de Gür Ali e Arıtürk (2014)

Esses são critérios estáticos e que se comportam bem para previsões com horizontes temporais mais curtos. Para melhorar a acuracidade a longo prazo, os autores afirmaram que é preciso incorporar fatores dinâmicos que representam a variabilidade da interação do assinante com o produto e a empresa ao longo do tempo.

De Caigny et al. (2020) utilizaram redes neurais artificiais para realizar análises textuais das interações dos clientes com a empresa e melhorar a performance de previsão de *churn* no setor bancário, atingindo um AUC de 0,899.

Para um contexto mais recente, nos últimos anos, outras ferramentas que têm recebido bastante destaque são as redes neurais e as técnicas de *deep learning*. A flexibilidade e a capacidade de simular relações entre as variáveis preditoras estão entre as características mais citadas para a escolha desse tipo de algoritmo. Entretando, mesmo nesses trabalhos, a regressão logística, muitas

vezes, é utilizada como modelo de referência em relação ao qual a capacidade preditora dos modelos é comparada, colocando-a como uma referência para pesquisadores da área. O que ainda a coloca como uma das técnicas mais relevantes para os pesquisadores da área.

De forma geral, a partir da análise sistemática da literatura observa-se que há uma variabilidade relativamente grande de métodos disponíveis que podem ser utilizados individualmente ou combinados entre si em modelos customizados. Porém, não há um processo estruturado estabelecido para construção de modelos de previsão de *churn*. Isso faz com que o processo de teste e identificação do melhor modelo para cada cenário torne-se oneroso.

Ao se analisar, dentre os casos disponíveis, os modelos que apresentaram os melhore resultados, é possível identificar algumas etapas que são comuns a esses métodos e que se destacam por consistentemente melhorarem o desempenho dos modelos. Dentro desse contexto, essa dissertação busca ajudar a sanar essa lacuna, propondo um método estruturado que contemple as etapas fundamentais para tornar mais eficiente a construção de modelos de previsão de churn.

Tabela 2 – Resumo de publicações de modelos de previsão utilizando regressão logística

Referências	Citações	Mercado	parâmetros preditores utilizados	Métricas de avaliação
De Caigny et al. (2018)	303	Diversos	13	AUC: 0,594 TDL: 1,444
Lu et al. (2014)	122	Telecomunicação	21	AUC: 0,724
Kim, Jun e Lee (2014)	62	Telecomunicação	36	AUC: 0,727 TDL: 31,5
Stripling et al. (2018)	55	Diversos	8	EMPC: 9,21
Günther et al. (2014)	38	Seguros	9	AUC: 15 vezes melhor que um classificador aleatório
Garrido, Verbeke e Bravo (2018)	19	Bancário	9	AUC: 0,422 Medida H: 0,538 EMP: 0,798 R-EMP: 0,798
Kiguchi, Saeed e Medi (2022)	15	Jogos	21	AUC: 0,9225 F1-score: 0,919
Gattermann-Itschert e Thonemann (2021)	14	Varejo	28	AUC: 0,65 TDL: 2,5
Zhang, Moro e Ramos (2022)	13	Telecomunicação	11	Acuracidade: 93,4%
Li e Li (2019)	13	e-Commerce	25	Acuracidade: 0,759 Precisão: 0,757 Recall: 0,836
Long et al. (2019)	13	Bancário	13	AUC: 0,82

AUC: Área sob a curva de características operacionais do receptor; TDL: *Top decile lift*; EMP: Lucro médio estimado; R-EMP: Lucro médio estimado robusto; EMPC: Lucro médio esperado para *churn* de consumidores

Fonte: autoria própria a partir da busca na base SCOPUS

2.6. Técnicas de amostragem

Um dos principais problemas que impactam a acuracidade de previsão dos modelos focados em *churn* é o desbalanceamento das classes. Este é um fenômeno observado em problemas de classificação onde uma classe é muito menos frequente que a outra, gerando um viés nos algoritmos de classificação (Wong, Seng e Wong, 2020). Um problema de desbalanceamento de classes acontece quando uma das classes representa menos de 10% de uma população (geralmente composta por duas classes) (Chawla *et al.*, 2003).

Essa diferença entre a quantidade de membros de cada classe pode fazer com que algoritmos de classificação como regressão logística, *random forest* e *support vector machine* apresentem um alto nível de acuracidade apenas classificando todos os membros de uma população como a classe majoritária, o que não faz sentido no contexto prático da previsão de *churn* (Li *et al.*, 2019). Por exemplo, em uma população de 100.000 clientes na qual 3.000 são possíveis cancelamentos, um algoritmo que classifique todos eles com baixo risco de cancelamento teria uma precisão de 97%, porém esse resultado não faz sentido do ponto de vista prático.

As ferramentas desenvolvidas para contornar esse problema focam em aplicar métodos de amostragem diferentes entre as classes. Outra forma de equilibrar a amostragem é atribuir pesos diferentes a cada classe. Dentre os métodos aplicados, o *oversampling* e o *undersampling* são os mais comuns (Chawla *et al.*, 2002). O *oversampling* gera amostras aleatórias da classe minoritária, o que gera o risco de prejudicar a capacidade de generalização do modelo devido ao *overfitting*. Já o *undersampling* remove aleatoriamente amostras da classe majoritária, gerando o risco de remover casos com características importantes para o desenvolvimento do modelo (Li *et al.*, 2019).

Diversas abordagens já foram propostas para contornar esses riscos. Dentre elas, o SMOTE (técnica sintética de *oversampling* da minoria, na sigla em inglês), desenvolvido por Li et al. (2019), está entre os mais citados na literatura. O SMOTE é um método de *oversampling* que busca equilibrar as amostras criando dados sinteticamente através de métodos de interpolação. O SMOTE pode ser combinado ao *undersampling* da classe majoritária para melhorar ainda mais a qualidade das previsões de um modelo (Chawla *et al.*, 2003).

O principal risco associado ao uso do SMOTE está na adição de ruído aos dados, que pode prejudicar a capacidade de previsão do modelo, além de apresentar dificuldade para lidar com situações nas quais a sobreposição das classes é muito alta (Li *et al.*, 2019). Algumas variações do SMOTE foram propostas para melhorar a sua robustez, como o DeepSMOTE (Dablain, Krawczyk e Chawla, 2023), o SMOTEboost (Chawla *et al.*, 2003), o SMOTE IPF (Sáez *et al.*, 2015), o Borderline-SMOTE (Han, Wang e Mao, 2005) e o Safe-Level-SMOTE (Bunkhumpornpat, Sinapiromsaran e Lursinsap, 2009). Essas ferramentas se diferenciam entre si pelo método utilizado para gerar dados aleatórios para a classe minoritária, buscando gerar o mínimo de ruído possível, e tornar mais clara para o modelo de previsão a fronteira entre as classes da população.

2.7. Métodos de previsão e classificação de dados

2.7.1. Regressão logística

A regressão logística é uma técnica efetiva e relativamente de implementação, para testar a relação entre uma variável dependente categórica e variáveis independentes, podendo estas serem contínuas ou categóricas (Lu *et al.*, 2014). Ela pode ser definida como uma técnica utilizada para construir um modelo linear que diferencia uma amostra em duas categorias (muitas vezes identificadas como 0 ou 1), baseado em variáveis preditoras (Gattermann-Itschert e Thonemann, 2021).

A regressão logística pode ser representada pelas Equações 5 e 6, onde y ϵ **Y** = {0,1}, para um conjunto de variáveis x_i, sendo x ϵ **X** = {x_1, x_2, ..., x_n}.

$$y = \frac{1}{1+ e^{-x}} \tag{5}$$

$$x = \beta_0 + \sum_1^n \beta_i x_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n = \ln\left(\frac{y}{1-y}\right) \tag{6}$$

A Figura 10 mostra uma representação gráfica da regressão logística.

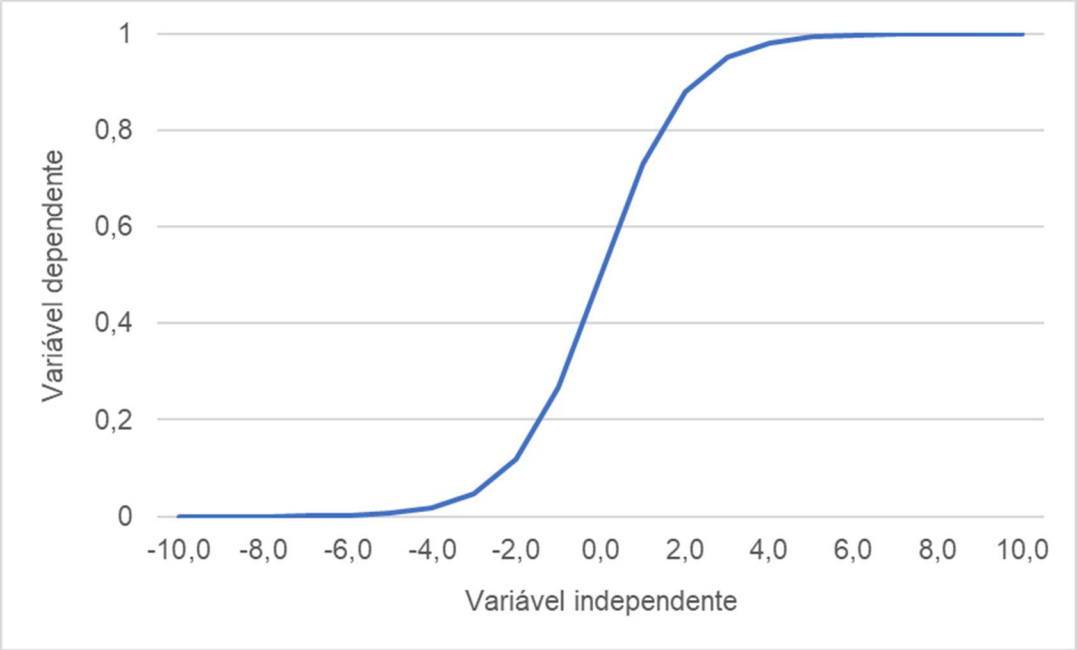

Figura 10 – Representação gráfica da regressão logística

Fonte: Adaptado de (Montgomery e Runger, 2011)

A regressão logística apresenta melhor desempenho para conjuntos de dados pequenos (Perlich, Provost e Simonoff, 2003) e representa muito bem relações lineares entre as variáveis de entrada. Porém, tem dificuldade para representar a relação de interação entre elas (De Caigny e De Bock, 2018; Zhang, Moro e Ramos, 2022). Devido a essas características, essa ferramenta é muitas vezes tratada como a base de comparação para novos métodos em desenvolvimento e, também, pode ser combinada a outras ferramentas para formar modelos de previsão compostos (*ensemble*).

Verbeke et al. (2012) utilizaram um banco de dados públicos, compostos por dados de diversos tipos de empresas, e avaliaram a qualidade das previsões de diferentes ferramentas. Nesse estudo, os autores mostraram através da comparação da área sob a curva de características operacionais (ROC), que a regressão logística apresentou uma acuracidade superior quando comparada aos demais métodos.

A absorção de variáveis categóricas e as lacunas da base de dados incompletas são limitações para o uso da regressão logística como ferramenta de previsão, porém essas limitações podem ser contornadas de maneira simples através do tratamento da base de dados (Coussement, Lessmann e Verstraeten,

2017). Outro aspecto positivo dos modelos baseados em regressão logística é a facilidade de interpretar os dados de saída, já que o método não é uma “caixa preta” como outros mais complexos, o que facilita o entendimento e auxilia na tomada de decisão de gestores de negócio em situações reais (De Caigny e De Bock, 2018).

A seleção das variáveis independentes que serão utilizadas como preditoras do modelo tem um impacto significativo na sua acuracidade. Segundo Alboukaey, Joukhadar e Ghneim (2020), análises de *churn* baseadas em parâmetros que são ajustados mensalmente é menos eficaz do que análises baseadas em parâmetros ajustados diariamente, pois tem uma performance de previsão menor.

2.7.2. Redes Neurais

Saha et al. (2023) definem redes neurais artificiais (RNA) como uma representação matemática do sistema nervoso humano e suas conexões, a qual é composta por três camadas: a camada de entrada, camada oculta e a camada de saída. As RNA são um dos métodos mais utilizados para modelar sistemas de previsão, pois são capazes de representar relações lineares e não lineares entre as variáveis preditoras com alta acuracidade (Bianchesi *et al.*, 2022).

Dentre as redes neurais as redes *multilayer perceptron* (MLP) e função de base radial (RBF) são as mais aplicadas para resolução de problemas complexos como a previsão de *churn* (Khodabandehlou e Zivari Rahman, 2017). Keramati et al. (2014) compararam a performance de diferentes métodos de aprendizado de máquina na previsão de *churn* de uma empresa de telecomunicações e mostraram que redes neurais atingiram 85% de acuracidade, o que representa uma performance superior a outros algoritmos como árvore de decisão e *support vector machine*.

A Figura 11 apresenta uma representação esquemática de uma rede neural multicamadas e a equação geral que descreve a relação entre as entradas e saídas de cada neurônio, sendo: v a saída; H o número total de neurônios na camada anterior; u_h a saída do neurônio de número h da camada anterior; w_h o peso atribuído à conexão; w_0 o erro; e f a função de ativação não linear.

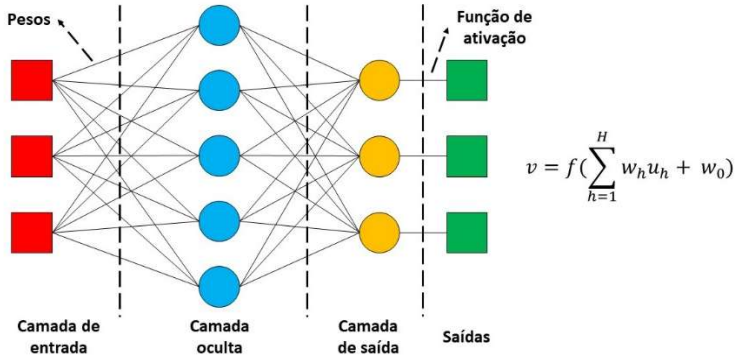

Figura 11 – Representação de uma rede MLP e equação de saída do neurônio
Fonte: Adaptado de Bianchesi et al. (2022)

A capacidade de processar grandes quantidades de dados de forma rápida e de representar de forma eficaz relações não lineares entre as variáveis preditoras são uma das principais vantagens da aplicação das redes neurais quando comparadas a outros métodos de previsão de *churn*.

2.7.3. Média-K

Métodos de *clustering* estão entre os métodos de aprendizado de máquina mais aplicados na construção de modelos de previsão de churn (Khine e Myo, 2023). Através deles, os dados de uma determinada variável de resposta são agrupados de acordo com a similaridade entre as suas características (variáveis preditoras).

O algoritmo média-K (*K-means*) pode ser definido como um processo de aprendizado não supervisionado onde os dados são agrupados em "K" subgrupos através de uma função que busca minimizar a média das distâncias entre cada ponto com o centroide do seu respectivo subgrupo (da Silva e Salcedo, 2024).

Liu et al. (2023) analisaram o impacto da aplicação do algoritmo média-K em conjunto com outros métodos de previsão para identificar os clientes com maior probabilidade de *churn* em uma empresa de telecomunicações. Os autores atingiram 79,5% de acuracidade combinando o algoritmo média-K à regressão logística e 90,8% ao combiná-lo com uma rede neural MLP.

Em um trabalho semelhante Khine e Myo (2023) propuseram um modelo de previsão de *churn* para o mercado bancário através da combinação do algoritmo

média-K com redes neurais MLP. O método proposto apresentou uma acuracidade média entre os *clusters* de 85% sendo que dentre os subgrupos a acuracidade máxima foi de 94%.

Da Silva e Salcedo (2024) dividiram os dados da base de compras de uma empresa do mercado de *e-commerce* em 5 subgrupos através da aplicação do algoritmo média-K. Os autores avaliaram a performance na previsão de *churn* de quatro algoritmos atingindo acuracidade de 68,7% com a regressão logística, 69,6% com *random forest*, 69,4% com *gradient boosting* e 71,5% com *light gradiente boosting*.

A segregação dos dados com perfis semelhantes facilita a diferenciação entre as categorias da variável de resposta dentro de cada subgrupo, por isso o *clustering* tem sido utilizado com sucesso como uma etapa prévia à aplicação de outros métodos de previsão de *churn* (De Caigny, Coussement and Bock, 2018).

3. MÉTODO DE PESQUISA

3.1. Definição do método de pesquisa

O presente trabalho seguiu o método de pesquisa de modelagem e simulação, fundamentado na estrutura de Bertrand e Fransoo (2002). Segundo os autores, a construção de um modelo de simulação é baseada na premissa de que é possível simular, totalmente ou em partes, o comportamento de sistemas reais de forma a auxiliar no processo de tomada de decisão no gerenciamento de operações. A relação causal e quantitativa entre as variáveis de um processo, quando mapeadas corretamente, permitem prever com relativa qualidade o estado futuro de uma operação e não somente descrevê-la em seu estado atual.

Os métodos de pesquisa para modelagem e simulação na gestão de operações podem ser classificados como sendo axiomáticos ou empíricos, de acordo com a forma de construção do modelo, e cada uma dessas categorias pode ser subdividida como sendo normativas ou descritivas, de acordo com o seu objetivo de aplicação (Bertrand e Fransoo, 2002). Na Figura 12, é mostrado de forma esquemática as etapas para a condução de um estudo por modelagem e simulação, no qual um modelo de previsão baseado em dados empíricos pode ser elaborado para que seja possível medir o impacto de possíveis mudanças/melhorias futuras em um processo.

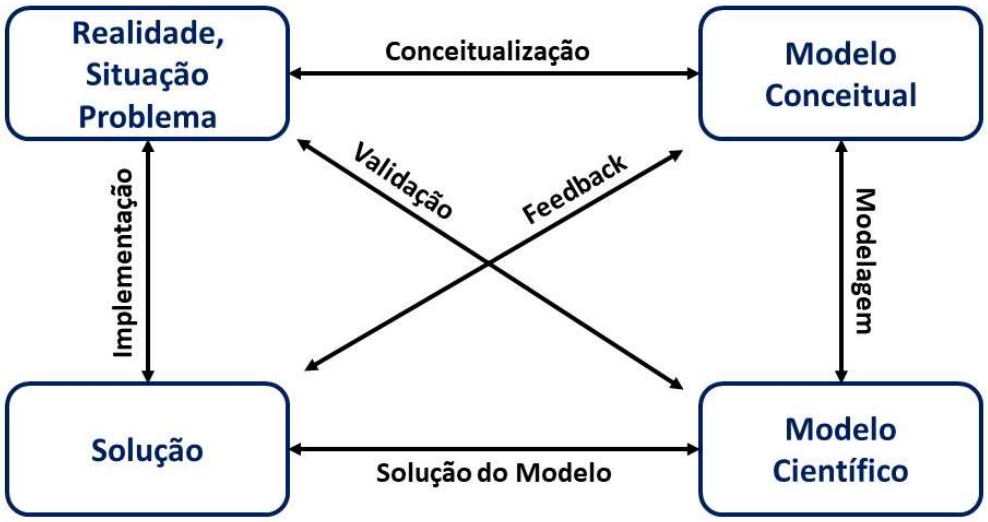

Figura 12 – Método de pesquisa de modelagem e simulação.
Fonte: adaptado de Bertrand e Fransoo (2002)

3.2. Objeto de estudo e coleta de dados

O processo proposto nessa dissertação incorpora etapas importantes para melhoria da qualidade das previsões dos modelos analisados. Para validá-lo, foi realizada a sua aplicação prática utilizando o cenário real de um clube de assinaturas de livros que atua no mercado brasileiro e possui mais de 180 mil assinantes.

A empresa objeto de estudo utiliza atualmente um modelo de previsão de churn que classifica seus assinantes quanto ao risco de cancelamento. Com uma acuracidade de 71,8%, precisão de 12,1%, *recall* de 11,7% e medida F1 de 11,9%, o modelo atual foi construído com base na regressão logística e não leva em consideração etapas de preparação como balanceamento de classes e, posteriormente, testes de estabilidade. Dessa forma, a falta dessas etapas faz com que a capacidade do modelo de identificar os *churn* verdadeiros seja baixa, o que, consequentemente, o custo da sua aplicação prática seja alto.

Para a coleta dos dados, a base de assinantes no período de setembro de 2023 foi armazenada e tratados com o objetivo de prever os cancelamentos que ocorreram durante o mês de outubro do mesmo ano através da aplicação do *framework* proposto neste trabalho. Para se determinar o melhor modelo, diferentes métodos de previsão foram testados e comparados entre si, e os indicadores de acuracidade, precisão, *recall* e medida F1 foram os selecionados para esta comparação eles.

A validação da estabilidade do modelo selecionado levou em consideração dados extraídos de períodos diferentes e posteriores para a geração de novas previsões. Com as novas informações geradas, foram calculados os indicadores de cada período e comparados para verificar se houve variação na performance do modelo entre os períodos considerados.

3.3. Modelo conceitual: procedimento para modelagem de previsão de *churn* em mercados de compra recorrente

O *framework* estruturado neste trabalho para construção de modelos de previsão de *churn* para mercados de compra recorrente encontra-se representado na Figura 13. O processo foi elaborado em três fases macro e

sequenciais: preparação dos dados, aplicação dos modelos e avaliação dos modelos.

3.3.1. Fase 1: Preparação dos dados

O processo tem início com a coleta dos dados brutos de clientes durante um período pré-determinado. As informações que compõem esse banco de dados muitas vezes vão depender da natureza do negócio analisado, mas em geral são priorizadas variáveis de caráter demográfico, comercial ou ligadas à interação entre cliente e empresa, como informações de atendimento, dúvidas e reclamações (Gür Ali e Arıtürk, 2014).

Dessa forma, esses dados brutos precisam ser inicialmente tratados para um melhor processamento posterior pelos modelos de previsão. O pré-tratamento dos dados busca padronizá-los em um formato que remova possíveis fontes de erro, por exemplo, pela eliminação de dados faltantes e/ou valores discrepantes (outliers). Em casos específicos, dependendo do método de previsão aplicado, pode ser necessário fazer uma transformação nos dados, buscando aproximá-los de uma distribuição normal de probabilidades.

Com os dados refinados, prossegue-se com uma análise exploratória inicial, buscando identificar padrões de distribuição indiquem, por exemplo, as variáveis preditoras mais promissoras e uma distribuição geral da base de clientes entre essas variáveis. Outro ponto importante é entender, dentro da população, a proporção entre as classes de *churn* e não-*churn*. Uma taxa de *churn* menor que 10% dentro da população indica um problema de desbalanceamento das classes que, se não tratado, pode gerar um viés de previsão e impactar negativamente a qualidade dos modelos avaliados (Li *et al.*, 2019).

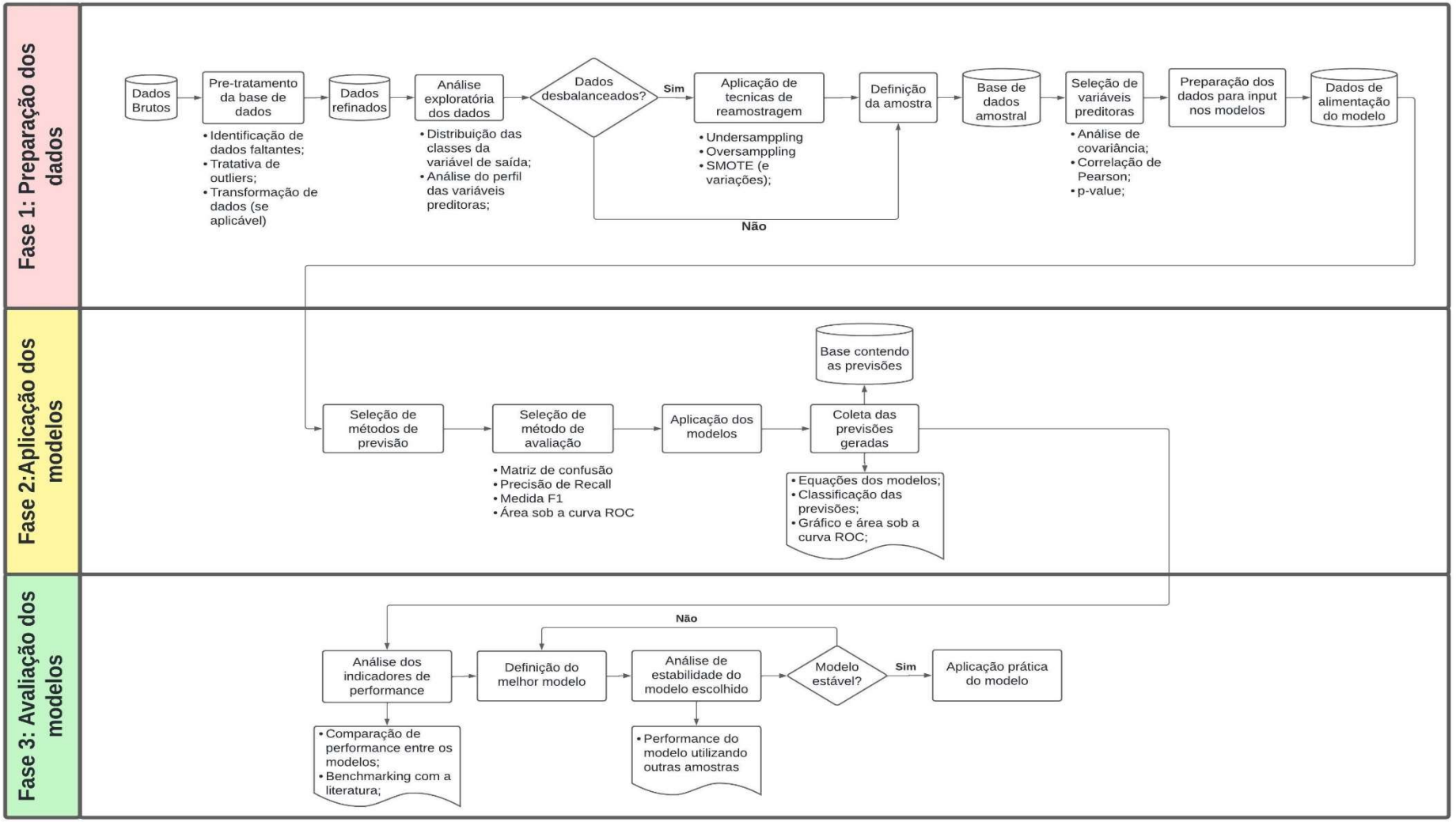

Figura 13 – *Framework* proposto para construção de modelos de previsão de *churn* em mercados de compra recorrente.
Fonte: Autoria própria.

Caso o problema de desbalanceamento de dados seja identificado, técnicas de reamostragem devem ser aplicadas para mitigá-lo e gerar uma amostra que tenha um perfil mais balanceado de proporção entre as classes. A aplicação de métodos estatísticos para geração de pontos artificiais da categoria minoritária (*oversampling*) ou retirada de pontos da classe majoritária (*undersampling*) estão entre as técnicas mais aplicadas no tratamento desse tipo de problema (Chawla *et al.*, 2003). Métodos mais recente, como SMOTE e suas variações, podem apresentar resultados ainda melhores na composição da amostra final, pois evitam a retirada da amostra de dados importantes para o entendimento do perfil do cliente (Dablain, Krawczyk e Chawla, 2023).

Nos casos em que a amostra final contém muitas variáveis preditoras, a seleção daquelas que tem maior relação com a decisão de cancelamento por parte do cliente, e a remoção de variáveis possivelmente redundantes, é importante para melhorar a eficiência e reduzir o tempo de processamento necessário para métodos de previsão mais complexos. Análise de covariância, correlação de Pearson e até mesmo o valor-*p* fornecido por alguns testes de hipóteses, como forma de avaliar a relação de cada variável de entrada com os dados de *churn*, ajudam na construção de modelos mais precisos.

Por fim, a Fase 1 termina com a obtenção de uma amostra contendo dados padronizados, com classes bem distribuídas e variáveis preditoras relevantes selecionadas para o abastecimento dos modelos de previsão.

3.3.2. Fase 2: Aplicação dos modelos de previsão

A aplicação individual ou uma combinação de métodos traz uma miríade de opções para a construção de modelos de previsão. O uso, por exemplo, de técnicas de *clustering* para segmentar os dados amostrais, combinado com a aplicação de métodos de regressão, tem sido bastante explorado recentemente como forma de aumentar a precisão nos modelos de previsão de *churn* (De Caigny, Coussement e De Bock, 2018).

Como visto no capítulo anterior, a revisão sistemática dos artigos que trabalharam em mercados semelhantes ao objeto de estudo dessa pesquisa é um bom ponto de partida para selecionar os métodos que se deseja testar. Além disso, a inclusão de métodos mais tradicionais de previsão, como a regressão

logística, tende a ajudar no estabelecimento de um referencial quando se deseja comparações com de modelos mais customizados.

Considerando que o objetivo do *framework* elaboado neste trabalho é a identificação do melhor modelo entre os considerados no contexto da previsão de *churn*, é importante definir as métricas que serão usadas para avaliar a qualidade de cada modelo. Nesse contexto, Os indicadores derivados da matriz de confusão estão entre os mais utilizados, principalmente porque permitem uma avaliação abrangente dos resultados (Imani e Arabnia, 2023).

Para o caso dos métodos regressivos, a variável de resposta é dada como a probabilidade de *churn* de cada cliente. Assim, a classificação final entre cada classe depende do valor limiar (*threshold*) que é definido para separar o *churn* do *não-churn*. E consequentemente variando-se esse valor limitar, obtém-se taxas diferentes para positivos verdadeiros (PV) e positivos falsos (PF).

A partir dos valores de PV e PF obtidos em cada valor limiar considerado, pode-se construir a curva de características operacionais do receptor (ROC) e calcular a área sob esta curva (Kim, Jun e Lee, 2014). A área sob a curva ROC varia entre 0 e 1 e representa a probabilidade que um modelo tem de classificar um cliente de forma correta (Lu *et al.*, 2014).

Para mercados onde o custo de retenção dos clientes é muito alto, os falsos positivos gerados pelo modelo podem representar um impacto financeiro significativo. Para esse tipo de situação, pode-se utilizar os indicadores LM e LME proposto por Garrido, Verbeke e Bravo (2018) que incorporam o custo de retenção na comparação entre modelos.

Ao fim da Fase 2, espera-se obter as equações ou representações matemáticas para cada modelo selecionado. Além disso, é possível realizar uma análise de sensibilidade, que pode trazer informações importantes sobre o impacto de cada variável preditora na decisão do cliente de cancelar a próxima compra. Essas informações podem ser valiosas para construção de campanhas de retenção mais eficazes no futuro.

3.3.3. Fase 3: Avaliação dos modelos

De posse das previsões geradas e dos indicadores de performance de cada modelo testado, parte-se então para a definição do melhor método de previsão.

Assim, é importante que essa decisão leve em consideração os aspectos de aplicação prática, como a estabilidade dos modelos e suas complexidades de implementação.

A análise de estabilidade é feita aplicando-se aos modelos dados de períodos diferentes daqueles que foram utilizados na sua construção. Esses dados precisam passar primeiro pelas mesmas etapas de padronização utilizadas para composição da amostra inicial. Risselada, Verhoef e Bijmolt, (2010a) afirmaram que a análise de estabilidade nos modelos de previsão de *churn* é importante para avaliar a deterioração da capacidade preditiva dos modelos com o tempo. Além disso, essa etapa ajuda a avaliar possíveis vieses de sazonalidade que possam ter sido incorporadas ao modelo, prejudicando sua utilização em situações reais. Idealmente, um modelo com alta capacidade de generalização deve ser capaz de sustentar a qualidade das previsões ao ser aplicado à conjuntos de dados de diferentes períodos.

Ao final da Fase 3, espera-se obter um modelo relativamente robusto e estável, capaz de gerar previsões com boa acuracidade e estabilidade em diferentes períodos, e que tenha implementação viável no cotidiano da empresa em questão. Com isso, busca-se não somente reduzir o churn por meio da previsão de possíveis cancelamentos, mas também uma identificação de perfis semelhantes que possam nortear a construção de campanhas proativas de retenção, que atuem de forma mais eficaz entre os clientes da empresa.

Um método construído dessa forma ajuda não só a reduzir o *churn* através da previsão de possíveis cancelamentos, como também a identificação de perfis semelhantes que podem nortear a construção de campanhas proativas de retenção que atuem de forma eficaz nos critérios mais importantes para a tomada de decisão dos clientes.

4. RESULTADOS

Neste capítulo será realizada a aplicação do *framework* proposto para construção de um modelo de previsão de *churn* utilizando os dados reais de um clube de assinaturas de livros infantis localizado no Brasil.

4.1. Fase 1: Preparação dos dados

4.1.1. Coleta e pré-tratamento da base de dados

A empresa objeto de estudo atende mais de 180 mil assinantes mensalmente em todo o Brasil. O período selecionado para construção do modelo contém os dados de setembro de 2023 com o objetivo de prever os cancelamentos ocorridos durante o mês de outubro do mesmo ano. Para isso, foram identificados os assinantes que realizaram o cancelamento da assinatura durante o mês de outubro de 2023 e adicionado uma coluna ao banco de dados com a informação de "*churn*" para identificar os que cancelaram e "não-*churn*" para os que permaneceram na base até último dia do mês.

Os dados brutos foram coletados, organizados e tratados para facilitar as análises posteriores. Nesse processo foram identificadas e removidas inconsistências, dados faltantes e *outliers* resultando em uma base de dados refinados composta por 180.440 pontos.

4.1.2. Análise exploratória dos dados

A partir dos dados refinados foram realizadas algumas análises para avaliar a distribuição geral dos dados de acordo com cada uma das variáveis disponíveis.

A proporção de cancelamentos dentro do mês de outubro, como mostra a Figura 14, é de 8,9%, representando aproximadamente 16 mil potenciais cancelamentos. É importante destacar que a maior parte desses cancelamentos é revertido através de ações de retenção que buscam contato ativo com os clientes oferecendo métodos alternativos de pagamentos e benefícios como

descontos e produtos adicionais. A importância desse tipo de ação pode ser constatada a partir da análise da Figura 14 que mostra que grande parte dos cancelamentos acontece de forma involuntária, principalmente por inadimplência.

Através da análise das informações coletadas, buscou-se encontrar critérios relevantes para a construção de um modelo capaz de prever quais assinantes têm maior probabilidade de cancelar o serviço no decorrer do mês seguinte. Para ter um entendimento inicial de quais critérios poderiam ser mais ou menos relevantes, foi feita uma análise gráfica comparando a distribuição da população analisada de acordo com os critérios disponíveis e comparado com a distribuição da população que realizou e a que não realizou o cancelamento da assinatura.

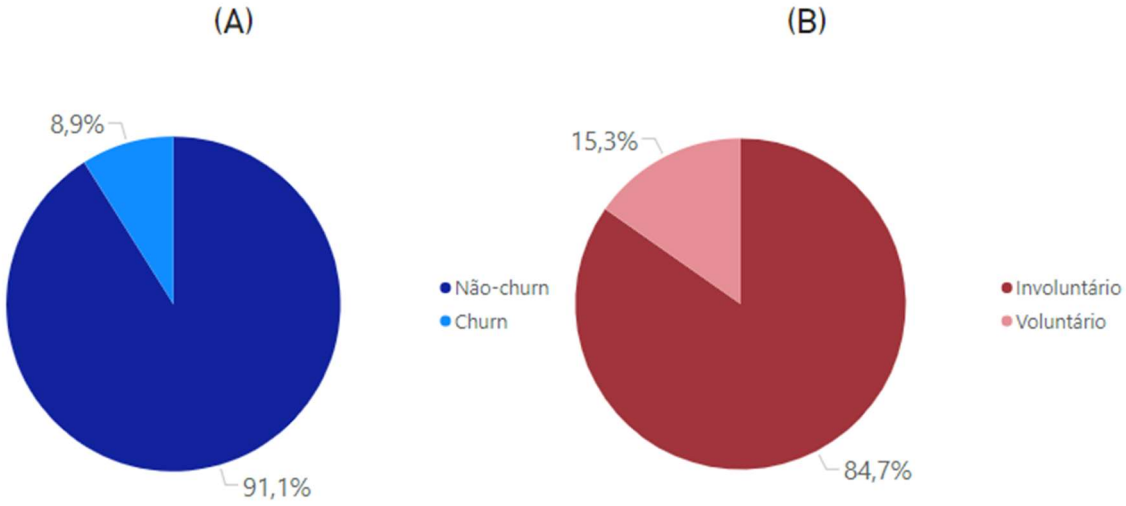

Figura 14 – Proporção de cancelamentos (A) e proporção de cancelamento voluntários e involuntários (B) durante o mês de setembro de 2023

É possível observar pela Figura 15 que há pouca diferença entre a composição da base total e das populações que cancelaram e não cancelaram com relação ao produto adquirido (*product_detail*). Isso pode indicar que o tipo de plano assinado pelo cliente não possui um alto impacto na previsibilidade do cancelamento da assinatura. Um comportamento similar pode ser observado na Figura 16 ao se analisar a variável *renewal_period*, com exceção de um aumento na categoria de pagamentos mensais, que apesar de representar uma pequena parcela da base, tem uma proporção maior entre a população que cancela quando comparada a população que não cancela a assinatura.

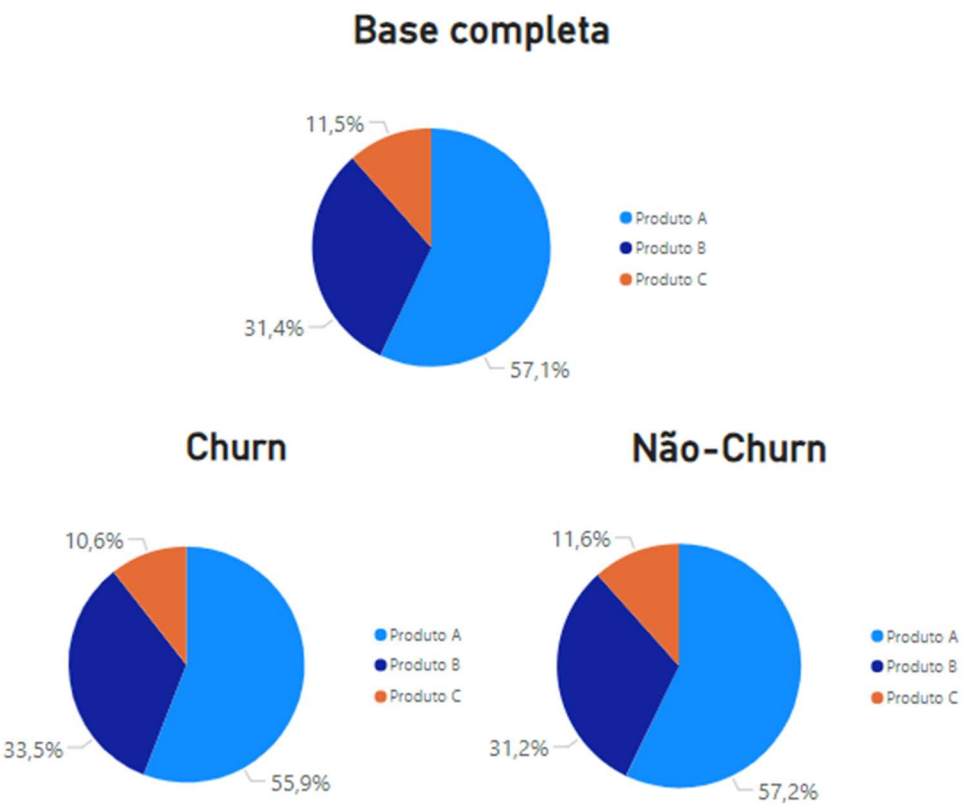

Figura 15 – Análise por *product_detail*

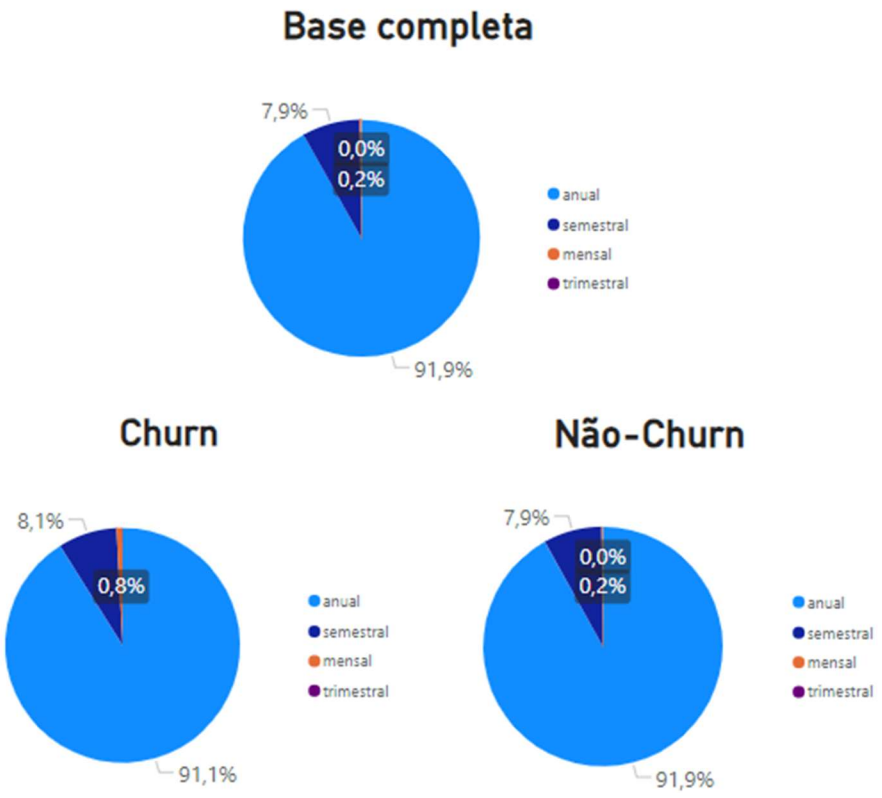

Figura 16 – Análise por *renewal_period*

Em contrapartida, quando se considera o critério *payment_type* (Figura 17), observa-se uma proporção significativamente maior de pagamentos recorrentes na população que realizou o cancelamento da assinatura quando comparada à que não cancelou, indicando que essa talvez seja uma variável preditora importante para a construção do modelo. Dzieciolowski (2015) afirma que a intenção de cancelamento do assinante pode mudar de acordo com a quantidade de tempo restante para o fim do contrato, pois a incorrência de multas para sua interrupção é uma prática de mercado e o valor pago pelo assinante é proporcional ao tempo remanescente do contrato. Dessa forma, pagamentos recorrentes, que possuem contrato de apenas 1 mês, não possuem multa de cancelamento e por isso permitem a saída sem custo do assinante. O mesmo racional se aplica a cancelamentos realizados próximos ou após a última parcela para pagamentos recorrentes.

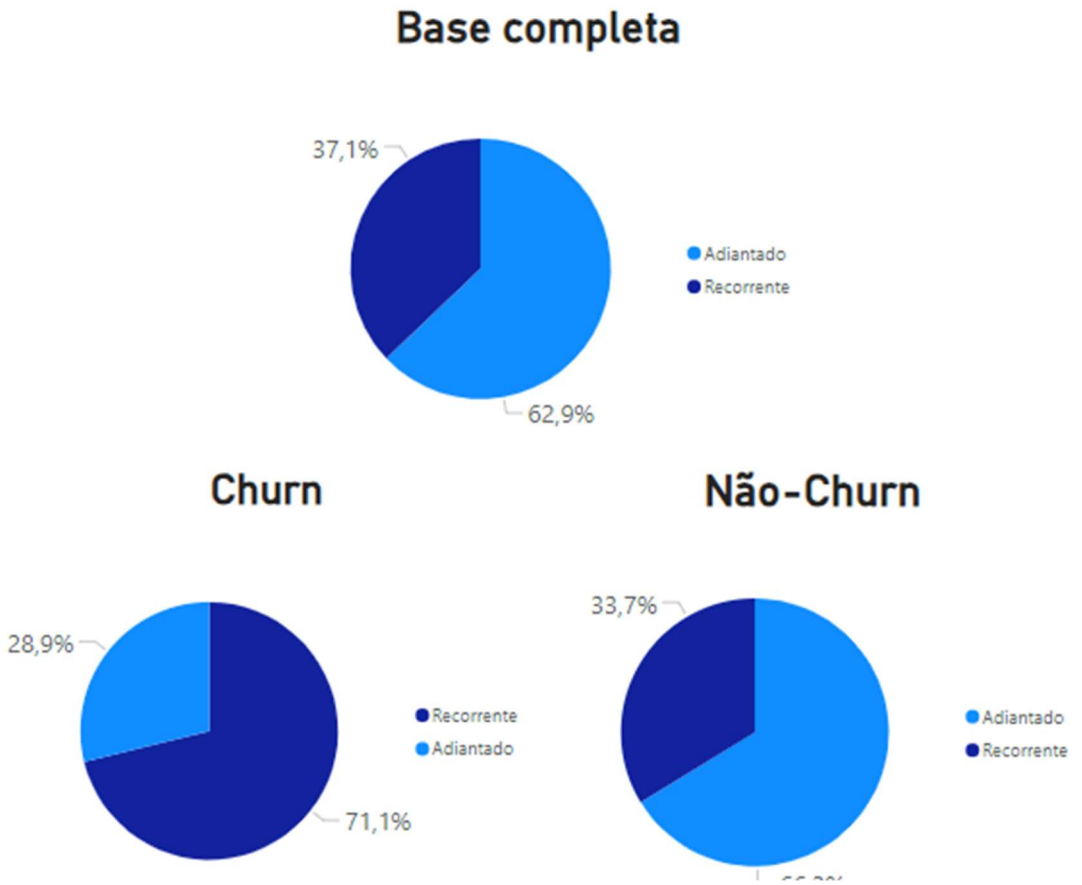

Figura 17 – Análise por *payment_type*

Outro fator importante que pode influenciar a decisão do cliente sobre manter ou não assinatura ativa é o valor pago pelo produto, pois isso tem um

impacto direto no custo-benefício percebido. a Figura 18 mostra um box-plot dos valores pagos pelas pessoas que tiveram suas assinaturas canceladas e as que não tiveram. Pode-se observar que não há uma diferença significativa na distribuição do valor pago pela assinatura entre os clientes que cancelaram ou não a sua assinatura.

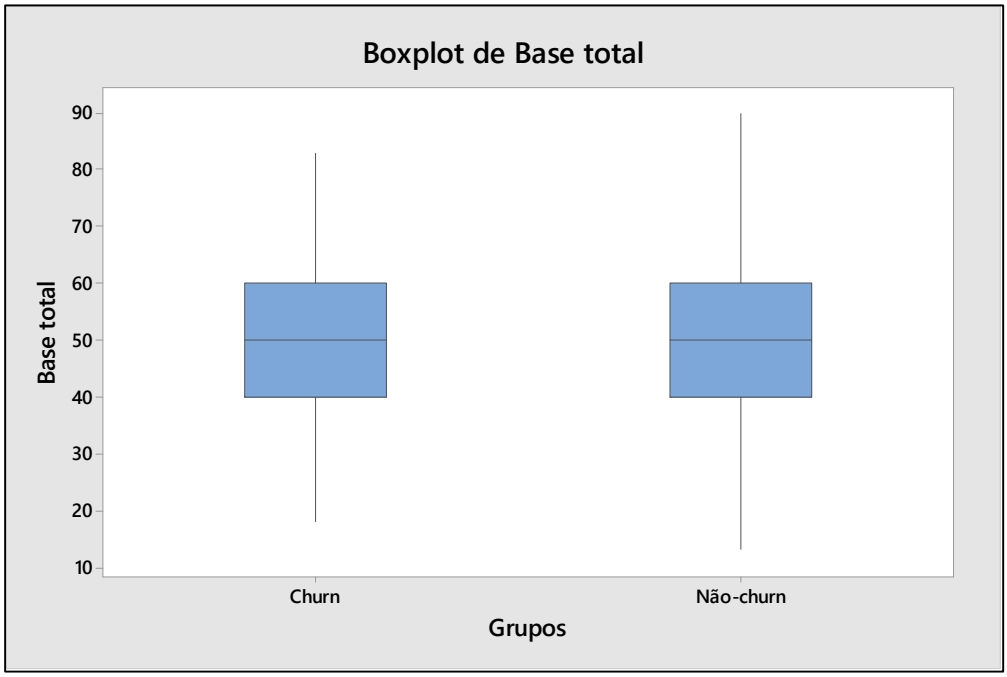

Figura 18 – Distribuição de valores pagos pelos assinantes que cancelaram ou não as assinaturas

O valor mensal pago e a parcela de desconto aplicada para cada assinante também podem ser considerados fatores de decisão para o cancelamento de assinaturas, principalmente porque, no momento da renovação do contrato, o valor pago pelo assinante é corrigido automaticamente e os descontos são retirados. Esse valor de desconto pode ser calculado subtraindo do parâmetro *paid_price* o valor do *plan_price* para cada assinante. Essa nova variável *desconto* foi calculada para todos os clientes da base e incluída como uma variável preditora nos modelos testados nos tópicos seguintes.

4.1.3. Aplicação de técnicas de reamostragem e definição da amostra

Como pode ser observado na Figura 14, a taxa de cancelamentos dentro da população analisada é de 8,9%. Uma diferença tão grande na proporção entre

os assinantes que cancelaram e não cancelaram suas assinaturas configura um problema de desbalanceamento de classes. Se não tratada no momento de se construir a amostra, essa distribuição dos dados pode gerar um viés nos modelos testados, fazendo com que identifiquem todos os casos como não-*churn* e ainda apresentando altos valores de acuracidade (Li *et al.*, 2019).

Neste estudo, optou-se por realizar uma subamostragem aleatória (*random undersampling*), coletando-se uma menor quantidade de pontos da classe majoritária (não-*churn*). Para isso, a população foi dividida entre as classes e um número aleatório variando entre 0 e 1 foi atribuído a cada caso, os 8.000 maiores números da classe "não-*churn"* e os 2.667 da classe "*churn*" foram selecionados para compor a amostra final, o que constituiu uma proporção de 1 para 4 (25% de *churn*) entre as classes. A Figura 19 mostra a configuração da amostra de acordo com os parâmetros analisados para a população.

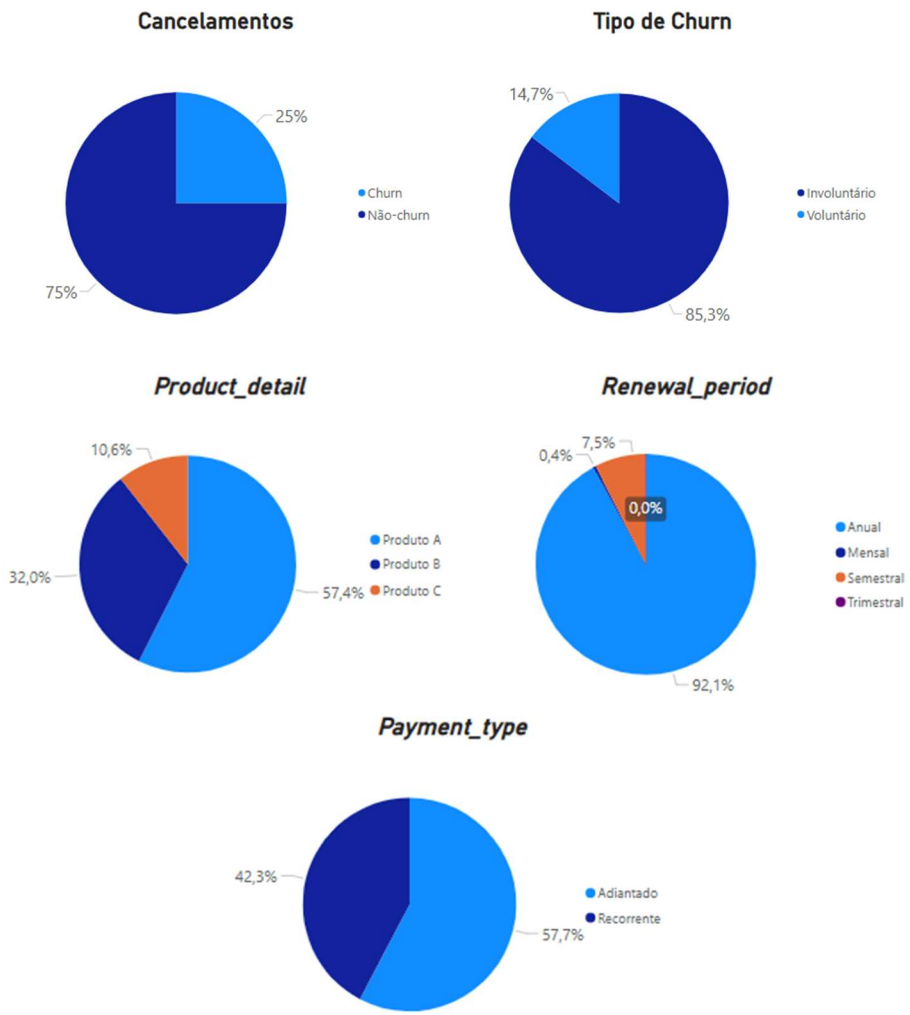

Figura 19 – Análise geral dos parâmetros da amostra.

4.1.4. Seleção de variáveis preditoras

As variáveis de entrada foram selecionadas de acordo com as informações individuais de cada assinante disponíveis no banco de dados da empresa e que são atualizadas mensalmente.

É importante destacar que o perfil sazonal do *churn* foi avaliado, pois alguns meses possuem histórico de cancelamentos maior do que os outros, e por isso podem representar um maior risco para o total de cancelamentos. A Tabela 3 descreve a identificação, a descrição dos dados coletados como variáveis preditoras do modelo e os possíveis valores assumidos:

Tabela 3 – Descrição das variáveis do banco de dados

Identificação	Notação	Descrição	Valores possíveis
fk_subscription	FK	Identificação numérica única de cada caso	
subscription_date	SD	Data de início da assinatura	De 28/07/2014 a 30/09/2023
product_detail	PD	Tipo de produto adquirido	0, 1 ou 2
payment_type	PT	Tipo de pagamento	Adiantado ou Recorrente
renewal_period	RP	Ciclo de renovação	Anual, Semestral, Trimestral ou Mensal
plan_price	PLP	Preço original do plano	Entre R$ 20,00 e R$ 90,00
paid_price	PDP	Preço pago atualmente pelo plano	Entre R$ 20,00 e R$ 90,00
subscription_cycle	SC	Parcela atual do plano	Entre 1 e 100
Desconto	D	Diferença entre preço pago e preço original	Entre R$ 0,00 e R$ 54,90

tipo de churn	TC	Tipo de *churn*	Voluntário ou Involuntário
churn	C	Cancelamento de assinatura	*Churn* ou Não-*churn*

Fonte: Autoria própria

4.2. Fase 2: Aplicação dos modelos

4.2.1. Seleção e aplicação dos métodos de previsão

A quantidade de trabalhos acadêmicos que buscam construir modelos de previsão de *churn* para clubes de assinatura de livros é pequena, ainda mais quando com foco no público infantil. Por isso, para selecionar os métodos a serem testados nesse trabalho, considerou-se as ferramentas mais utilizadas para mercados análogos como *e-commerce*, jogos eletrônicos e de telecomunicações.

A regressão logística está entre os métodos mais aplicados para avaliar a performance de novos modelos de previsão, sendo muitas vezes utilizada como referência para comparação entre métodos mais complexos (Verbeke *et al.*, 2012). Por isso sua aplicação foi realizada individualmente neste trabalho.

De posse dos dados da amostra, foi utilizado o software Minitab 18® para gerar a equação de regressão. Foram incluídos não só as variáveis independentes individualmente, como as possíveis interações de primeiro grau, buscando incorporar ao modelo possíveis relações entre elas.

Analisando os dados fornecidos pelo software, foram identificados termos da equação que apresentavam *p-value* superior 0,05. Com o intuito de melhorar a taxa de acertos do modelo, foram realizadas 2 iterações utilizando os mesmos dados, porém removendo o termo com maior *p-value*. As equações 7 e 8 mostram as equações finais utilizada neste trabalho como modelo onde "*P(churn)*" representa a probabilidade de cancelamento calculada para cada assinante representada como um valor entre 0 e 1.

$$P(churn) = \frac{e^{Y\prime}}{(1+e^{Y\prime})} \quad (7)$$

$$Y' = 15{,}2 + 4{,}5 * PD - 6{,}647 * PT - 0{,}76 * RP - 0{,}484 * PLP + 0{,}0706 * PDP + 0{,}0065 * SC - 1{,}495 * PD * PT - 0{,}1519 * PD * RP - 0{,}0712 * PD * PLP + 0{,}0259 * PD * PDP + 0{,}01494 * PD * SC + 0{,}1373 * PT * RP + 0{,}083 * PT * PLP - 0{,}00397 * PT * PDP + 0{,}03177 * PT * SC + 0{,}01638 * RP * PLP - 0{,}00357 * RP * PDP - 0{,}00038 * RP * SC + 0{,}002656 * PLP * PDP + 0{,}00362 * PLP * D - 0{,}000566 * PLP * SC + 0{,}000246 * PDP * D - 0{,}000218 * PDP * SC \quad (8)$$

Através dos coeficientes da equação do modelo e os valores possíveis para as variáveis preditoras, é possível avaliar o quanto cada termo impacta na probabilidade de um assinante ser classificado como *churn*. A Figura 20 mostra a análise de sensibilidade construída a partir da equação acima considerando somente as variáveis preditoras e a Figura 21 apresenta a mesma análise para os termos originados a partir das suas interações.

As variáveis *plan_price* e *paid_price* apresentaram efeitos opostos sobre a probabilidade de cancelamento quando comparadas individualmente aos demais preditores. O efeito do valor percebido pelo cliente com relação ao produto pode ser avaliado ao ser observar que a probabilidade de *churn* é menor entre assinantes do produto 2 (que possui o maior valor de venda) versus assinantes do produto 0 (que possui o menor valor de venda). Em contra partida, a variável *paid_price* indica que um aumento no valor pago pelo produto gera um aumento significativo na probabilidade de cancelamento.

O comportamento da variável *subscription_cycle* indica que a permanência do assinante dentro da base do clube de assinaturas tem um efeito negativo na probabilidade de cancelamento, ou seja, clientes mais longevos tendem a ser mais fiéis à marca e ao produto.

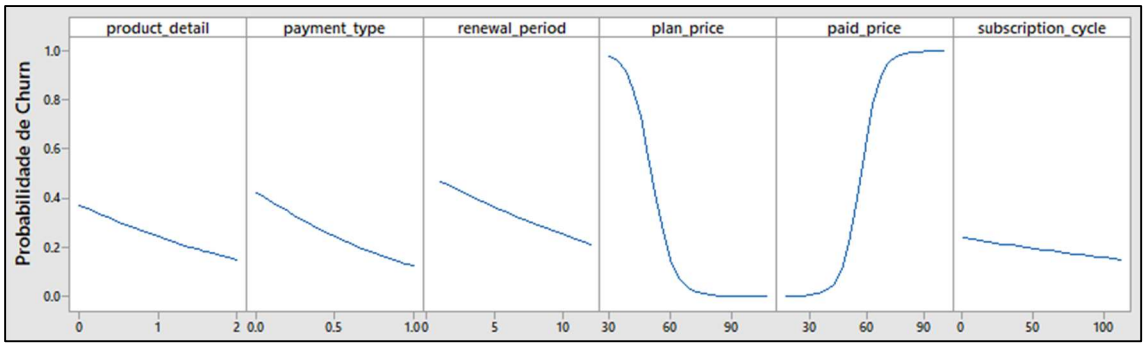

Figura 20 – Análise de sensibilidade para os termos do modelo sem interação.

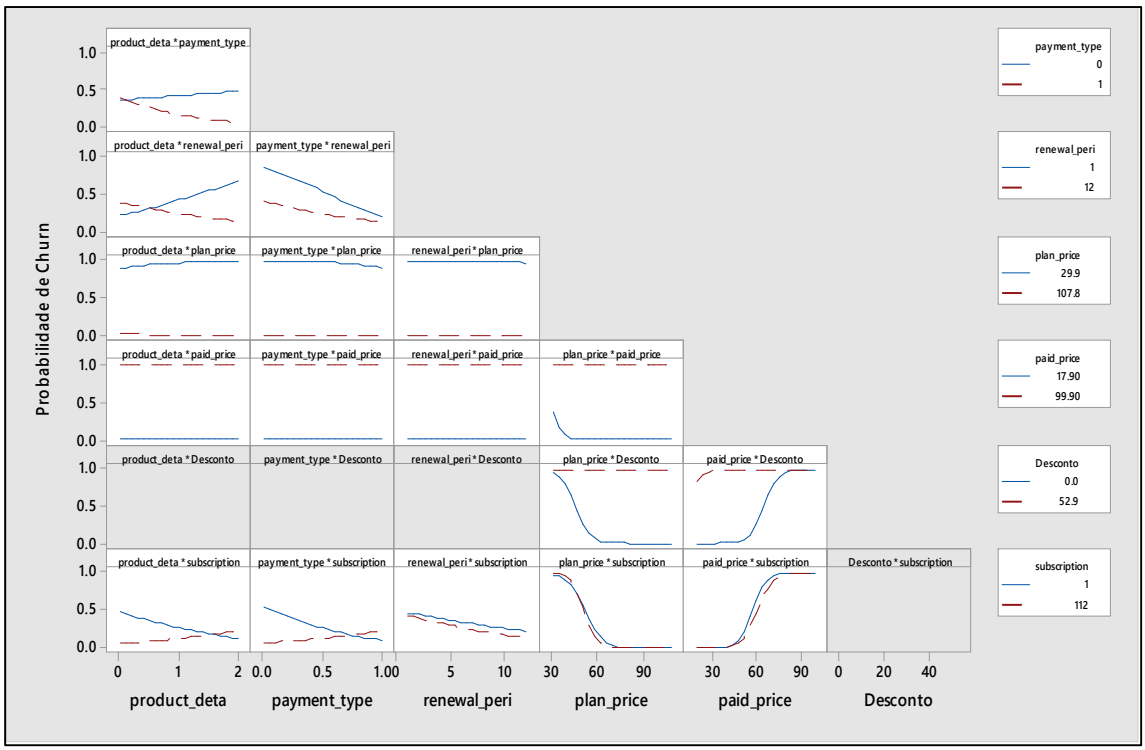

Figura 21 – Análise de sensibilidade para os termos do modelo com interação de primeiro grau.

A análise das interações entre as variáveis pode ajudar a indicar termos do modelo que tem pouca interferência no resultado da previsão gerada pelo modelo. Esses termos podem ser identificados por apresentarem pouca ou nenhuma inclinação no gráfico como as interações entre *paid_price* e *product_detail*, *payment_type* e *renewal_period*. A retirada desses termos do modelo contribui para a sua simplificação sem necessariamente impactar de forma significativa a qualidade das previsões realizadas.

Modelos compostos pela combinação de métodos de *clustering* e regressivos têm sido muito estudados em trabalhos recentes de previsão de *churn* (Liu *et al.*, 2023). Para avaliar sua performance no mercado de clubes de assinatura de livros infantis, foi testada a aplicação do algoritmo média-K combinado à regressão logística. Na primeira etapa, aplicou-se o algoritmo de *clustering* dividindo a amostra em 2 subgrupos. A cada subgrupo foi então aplicada separadamente a regressão logística, gerando-se um conjunto de previsões para cada um deles. É importante ressaltar que nas equações geradas para cada *cluster* também foram incluídas as interações de primeiro grau entre as variáveis preditoras.

A performance do modelo foi avaliada considerando tanto as classificações individuais de cada cluster quanto o resultado global. A Tabela 4 apresenta os dados dos centroides obtidos para cada variável após aplicação do algoritmo média-K.

Tabela 4 – Valores médios dos centroides obtidos a partir do algoritmo média-K.

Cluster	PD	PT	RP	PNP	PDP	D	SC	Casos	%
1	1,19	1,00	11,56	57,29	50,95	6,35	18,58	6156	57,71
2	1,24	0,00	11,43	58,35	50,56	7,79	21,72	4511	42,29

Devido à capacidade de simular relações não lineares entre as variáveis e a eficiência no processamento de grandes quantidades de dados, as redes neurais têm sido cada vez mais utilizadas na previsão de *churn* para mercados de compras recorrentes, como os de telecomunicações, bancários e de jogos eletrônicos (Khodabandehlou e Zivari Rahman, 2017). Para comparar a performance desse tipo de método com os mais tradicionais, foi construída uma rede neural MLP utilizando os mesmos 16 mil casos que compõem a amostra o seu treinamento e validação, sendo 50% dos dados utilizados no treinamento, 25% na validação e 25% como teste. A Tabela 5 mostra os parâmetros utilizados na rede MLP 7:7-5-1:1.

Tabela 5 – Parâmetros utilizados na construção da rede MLP 7:7-5-1:1.

Quantidade de neurônios	Camada de entrada	7
	Camada oculta	5
	Camada de saída	1

Parâmetros de treinamento	Algoritmo fase 1	Backpropagation
	Número de épocas	100
	Taxa de aprendizado	0,01
	Momentum	0,3
	Algoritmo fase 2	Gradiente descendente conjugado
	Número de épocas	500
Funções de ativação	Camada oculta	Tangente hiperbólica
	Camada de saída	Função logito

4.2.2. Seleção dos indicadores de avaliação

Pela sua natureza complementar, a acuracidade, a precisão, o recall e a medida F1 foram os indicadores selecionados para comparação entre os modelos. O intuito principal é eliminar um provável viés ao se considerar somente a acuracidade. Devido à baixa proporção de *churn* presente na população, uma alta taxa de negativos verdadeiros pode induzir a seleção de modelos que possuem algum viés de classificação.

Em um cenário ideal, uma alta precisão e um alto recall resultam em uma medida F1 próxima de 1. Isso indicaria um modelo robusto e capaz de identificar corretamente clientes com alta probabilidade de cancelamento. Em contrapartida, um modelo com um viés que favorece a classificação de assinantes como não-*churn* tende a apresentar uma acuracidade alta, porém uma precisão e recall baixos, resultando em uma medida-F1 mais próxima de 0.

Considerando o custo de retenção atrelado à possíveis campanhas construídas utilizando as previsões geradas, uma alta taxa de positivos falsos pode ser gerar um investimento desnecessário para a empresa que aplica o modelo selecionado. Por isso, uma seleção correta de indicadores de análise de performance pode ser um diferencial competitivo para situações de aplicação real.

Além disso, ao se realizar a análise de estabilidade, avaliar o comportamento combinado dos indicadores pode trazer ainda mais informações sobre a capacidade de generalização do modelo. Modelos robustos e capazes de manter esses indicadores em valores próximos ao da amostra inicial demonstram uma melhor capacidade de generalização.

4.3. Fase 3: Avaliação dos modelos

4.3.1. Análise dos indicadores de performance

A amostra foi submetida primeiramente à regressão logística utilizando como termos independentes tanto as 7 varáveis independentes (Tabela 3) como as possíveis interações de primeiro grau entre elas. A função logito gerou como saída a probabilidade de cancelamento expressa como um número entre 0 e 1. O valor limiar (*threshold*) utilizado para separar as classes foi 0,3, ou seja, assinantes com um risco estimado maior do que 30% de cancelamentos foram classificados como *churn*.

Com as previsões geradas pela regressão logística, foi construída a matriz de confusão e calculados os indicadores de performance do modelo. Com acuracidade de 71,1% a regressão apresentou um resultado ligeiramente abaixo modelo atual. Em contrapartida, com uma precisão calculada de 44,6%, um recall de 64,6% e uma medida F1 de 52,8% a regressão logística mostrou que é capaz de prever uma quantidade maior de *churns* verdadeiros do que o modelo atualmente utilizado pela empresa.

A performance geral do modelo pode estar relacionada a falta de variáveis com forte relação com o processo de decisão do consumidor. A possível adição de novos preditores e uma análise de sensibilidade para selecionar aqueles com maior relação com o risco de cancelamento podem melhorar ainda mais os resultados da regressão logística.

Hoang e Cam, (2025) atingiram 58% de acuracidade e 50% de precisão utilizando a regressão logística para prever o *churn* do mercado de jogos eletrônicos utilizando 13 preditores. Buckinx e Poel, Van den (2005) também utilizaram regressão logística para prever o *churn* no mercado varejista e atingiram 80,0% de acuracidade utilizando 10 preditores. Sahar (2018) atingiu uma acuracidade máxima de 87,4% utilizando 13 dos 17 preditores disponíveis em um modelo de regressão logística utilizado na previsão de *churn* do mercado de telecomunicações. De maneira geral, o resultado do modelo obtido por regressão logística neste trabalho está dentro da faixa obtida pelos demais

trabalhos descritos na literatura, o que mostra a aplicabilidade do método dentro do cenário analisado.

O segundo método avaliado foi a combinação entre o método de *clustering* e a regressão logística. A Tabela 6 mostra as características de cada grupo gerado pelo algoritmo média K. A regressão logística foi aplicada separadamente em cada um dos conjuntos de dados, foram mantidas as variáveis preditoras e o valor limiar utilizado anteriormente para identificar as classes.

Tabela 6 – Agrupamentos gerados pelo algoritmo média K.

	Não-churn	**Churn**	**Total**
Grupo 1	775	5381	6156
Grupo 2	1892	2619	4511

Quando analisado o resultado combinado dos dois grupos, a acuracidade global do método foi de 76,3% e a precisão foi de 55,9%, sendo melhor quando comparada a aplicação direta da regressão logística. Porém com um recall de 24,7% e uma medida F1 de 30,2% a combinação de média-K e regressão logística mostrou uma menor capacidade de identificação dos *churns* verdadeiros quando comparada ao método anterior.

Quando analisados separadamente, o agrupamento 1 apresentou 87,5% de acuracidade, 72,2% de precisão, 1,7% de recall e 3,3% de medida-F1 com uma densidade de cancelamentos de 12,6%. Em comparação, o agrupamento 2 apresentou valores menores de precisão e acuracidade, mas um recall e uma medida F1 maiores com uma densidade de cancelamentos de 41,9%. Isso mostra que a densidade de cancelamentos em cada *cluster* influencia diretamente na performance do modelo.

Xiahou, Xiancheng e Harada (2022) aplicaram a mesma combinação de métodos para prever o *churn* no mercado de e-commerce atingindo 90,7% de acuracidade e 85,3% de precisão. Os autores utilizaram o algoritmo média-K para dividir a amostra em 3 grupos e aplicaram a regressão logística individualmente em cada um deles. Isso mostra que o número de centroides definidos na etapa de pré-processamento pode impactar positivamente a performance do modelo.

Por fim, uma rede neural MLP foi construída utilizando o software Statistica 7®. Após a realização de alguns testes iniciais, a estrutura que apresentou os melhores resultados foi configurada com 7 neurônios na camada de entrada (um para cada variável preditora), 5 neurônios na camada oculta e 1 neurônio na camada de saída. A função de ativação utilizada entre a primeira e a segunda camada foi a tangente hiperbólica e entre a segunda e a terceira camada foi a função logito. O valor limiar utilizado para classificar as saídas da rede MLP foi otimizado pelo próprio software no valor de 0,3611.

Assim como foi feito com os outros modelos, foi gerada a matriz de confusão a partir das previsões geradas pela rede MLP e os indicadores foram calculados. O método apresentou um acuracidade de 75,6% e uma precisão de 51,7%, superando a regressão logística, mas um recall de 38,3% e uma medida F1 de 44,0%.

Hoang e Cam (2025) aplicaram redes neurais para prever o *churn* no mercado de jogos eletrônicos e alcançaram uma acuracidade de 68,5% com uma precisão de 66,7% enquanto Gordini e Veglio (2017) atingiram 88,5% de acuracidade prevendo *churn* no mercado de *e-commerce*.

Assim como a regressão logística, a performance da rede neural também pode se beneficiar da adição de novos preditores. A adição de mais neurônios na camada oculta e testes envolvendo novos parâmetros de aprendizado também podem ajudar a aumentar a qualidade das previsões.

4.3.2. Definição do melhor modelo

A Tabela 7 traz a compilação dos resultados obtidos para cada um dos métodos avaliados em comparação com o modelo utilizado pela empresa. Nota-se que a regressão logística foi o método com o maior valor de recall, o que resultou em uma medida F1 quase 5 vezes maior do que a do modelo atualmente utilizado pela empresa.

A lucratividade foi calculada utilizando uma adaptação do processo proposto por Neslin et al. (2006) no qual os modelos de classificação são avaliados com base no retorno financeiro estimado para campanhas de retenção construídas a partir de suas previsões.

Tabela 7 – Comparativo de acuracidade e precisão para os modelos avaliados.

	Modelo Atual	RL	MK+RL (Cluster 1)	MK+RL (Cluster 2)	MK+RL (Total)	Rede MLP
Recall	11,7%	64,6%	1,7%	34,1%	24,7%	38,3%
Precisão	12,1%	44,6%	72,2%	55,7%	55,9%	51,7%
Acuracidade	71,8%	71,1%	87,5%	61,0%	76,3%	75,6%
Medida F1	11,9%	52,8%	3,3%	42,3%	34,2%	44,0%
Retorno da campanha (R$)	-275.988	80.514	942	38.760	39.702	57.258

RL: Regressão logística; MK – Algoritmo média-K; MLP: Multilayer perceptron.

A partir dos dados das matrizes de confusão construídas a partir das previsões geradas por cada método, foram estimados os parâmetros e aplicados à Equação 9 onde N é o número total de assinantes, ŋ é a fração de assinantes impactados pela campanha, γ é a fração de cancelamentos reais que foram revertidos pela campanha, LTV é o valor adicional pago por cada cliente retido (*lifetime value*), d é o custo do incentivo oferecido para cada assinante, π_0 é o percentual de *churn* da amostra e λ é o coeficiente "*lift*" e mede o quão mais provável é o cancelamento de um assinante que foi contactado em relação ao total da amostra.

$$Retorno = N * ŋ * \Big(\big(\gamma * LTV + d * (1-\gamma)\big) * \pi_0 * \lambda - d\Big) \quad (9)$$

Para que fosse possível comparar os cenários, foram feitas algumas estimativas sobre o escopo e possíveis impactos das campanhas de retenção. Essas estimativas dessas constantes foram aplicadas igualmente a cada um dos modelos e são apresentadas na Tabela 8.

Tabela 8 – Constantes utilizadas para cálculo do retorno financeiro esperado.

Constante	Notação	Valor
Custo do incentivo	d	R$ 30,00
Percentual de cancelamentos reais revertidos	γ	70%
Valor adicional pago por cada cliente retido	CLV	R$ 150,00
Percentual de *churn* na amostra	$\boldsymbol{\pi_0}$	25,0%

Através da análise conjunta da medida F1 e do retorno esperado da campanha de retenção, pode-se afirmar que a regressão apresentou uma capacidade superior as dos demais modelos para a identificação de cancelamentos reais dentro das condições selecionadas. Uma maior taxa de positivos verdadeiros resulta em uma maior eficácia de retenção e, consequentemente, um maior retorno sobre o investimento realizado nas campanhas (Burez e Poel, 2007).

Bahnsen, Aouada e Ottersten (2015) afirmam que o custo de aquisição de um novo cliente é até 5 vezes maior do que o custo de retenção de clientes atuais. Sendo assim, uma taxa de retenção real superior a 20% dos clientes impactados já é o suficiente para que o retorno sobre o investimento seja superior quando comparado ao investimento em aquisição. Apresentando uma capacidade de identificar 44,6% de *churn* real dentre todos os casos classificados como *churn*, o modelo pode ser considerado economicamente viável.

A inclusão de mais variáveis que expressem a relação direta do assinante com a empresa pode fortalecer ainda mais o modelo. Histórico de inadimplência, nível de satisfação com o produto ou mesmo histórico de compras adicionais podem ajudar a aumentar a taxa de positivos verdadeiros do modelo (Risselada, Verhoef e Bijmolt, 2010).

Variáveis como essas não estão disponíveis no banco de dados da empresa atualmente, mas a sua inclusão é fortemente recomendada como forma de gerar modelos com capacidade de previsão ainda melhores e, consequentemente, maior eficácia de retenção de clientes.

4.3.3. Análise de estabilidade do modelo

Por ter apresentado uma melhor capacidade de identificação dos clientes que efetivamente cancelaram a sua assinatura no mês de outubro de 2023 através da análise dos dados de setembro de 2023, a regressão logística foi selecionada para seguir para a próxima etapa do fluxo e passar pela análise de estabilidade.

A Equação 8 gerada pela regressão logística construída na etapa anterior foi aplicada aos dados da base completa de assinantes em 3 períodos diferentes

com espaço de 6 meses, ou seja, os meses de outubro de 2023, abril de 2024 e outubro de 2024. Os mesmos indicadores de acuracidade e precisão foram avaliados e são apresentados na Figura 22. É importante destacar que os dados brutos dos períodos selecionados passaram pelo mesmo processo de tratamento para remição de inconsistências. Além disso, para testar a capacidade de generalização do modelo, a equação foi aplicada diretamente a população, sem que fossem construídas amostras balanceadas.

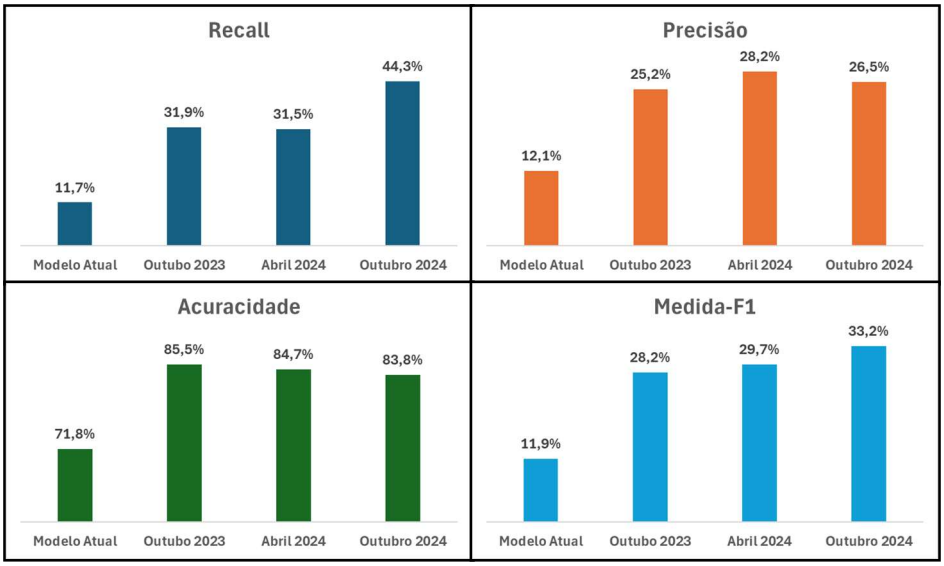

Figura 22 – Análise de estabilidade do modelo de previsão de *churn* construído a partir da regressão logística.

Como pode-se observar, o modelo baseado em regressão logística foi capaz de manter a acuracidade acima de 80% mesmo quando alimentado com dados referentes a um período 1 ano a frente daqueles usados na sua construção. A precisão e o recall do modelo, apesar de terem mantido valores superiores ao do modelo atual utilizado pela empresa, apresentaram uma redução quando comparados aos valores obtidos anteriormente. O mesmo pode ser observado no indicador medida F1, indicando uma perda da capacidade do modelo de identificar *churns* verdadeiros.

Esse comportamento pode ser consequência da baixa densidade de cancelamentos nos períodos avaliados, que teve média de 9,4%, quando comparada aos 25% da amostra utilizada na construção do modelo. Risselada, Verhoef e Bijmolt (2010) observaram um comportamento semelhante ao avaliarem o decréscimo do TDL de um modelo construído a partir da regressão

logística quando aplicado a períodos diferentes. Segundo os autores, a performance dos modelos pode ser melhorada estabelecendo-se uma rotina periódica de reavaliação ou retreinamento que podem ou não incluir novas variáveis e pesos atribuídos.

Li et al. (2023) avaliaram a estabilidade e a capacidade de generalização de algoritmos de aprendizado de máquina na previsão de *churn*. No respectivo trabalho, foram aplicados métodos como *random forest*, *support vector machine* e *extreme gradiente boosting* em 30 experimentos consecutivos e calculados os valores máximos, mínimos e o desvios padrão da área sob a curva ROC e da medida F1. Os autores também concluem que a perda de capacidade de previsão dos modelos gera a necessidade de reavaliação periódica sob o risco de gerar investimentos desnecessários em retenção caso a taxa de acertos seja muito reduzida.

Quanto aos métodos avaliados, existem possibilidades a serem exploradas buscando melhorar ainda mais os resultados obtidos. A análise de sensibilidades das variáveis preditoras utilizadas na regressão logística pode trazer informações importantes sobre fatores com baixa relação causal com a variável de resposta. A adição de mais centroides à aplicação do algoritmo média-K pode segmentar melhor os dados, tornando mais fácil a identificação da classe de cada caso. Por fim, a utilização de diferentes funções de ativação e de aprendizado, assim como o teste de outros tipos de redes neurais como as de função de base radial (RBF) ou linear, podem contribuir para a construção de modelos mais precisos e com melhor capacidade de generalização.

5. CONCLUSÕES

O *framework* proposto para construção de modelos de previsão de *churn* mostrou-se capaz de tornar mais eficiente o processo de identificação do melhor método previsão dentro do contexto dos mercados de compras recorrente. A inclusão de etapas dedicadas ao tratamento de dados e balanceamento da amostra mostrou ter um impacto positivo na qualidade final das previsões geradas por diferentes métodos.

A definição de métricas padronizadas permitiu a comparação dos métodos avaliados de forma ampla, trazendo informações além da acuracidade para o processo de tomada de decisão. A inclusão da precisão, do recall e da medida-F1 como critérios de avalição tornou possível analisar a performance dos modelos sob o prisma do custo de retenção de clientes, o que pode contribuir para decisões práticas em situações reais de negócio.

Além disso, a inclusão da avaliação da estabilidade como critério de seleção mostrou-se importante para a avaliação da capacidade de generalização e do estabelecimento de períodos para retreinamento do modelo. Esse processo traz mais dinamismo para os métodos e contribui para manter a qualidade das previsões a longo prazo, tendo um potencial impacto importante no retorno financeiro das campanhas de retenção.

A aplicação prática do *framework* proposto neste trabalho permitiu a avaliação da sua performance em uma situação real de mercado. Ao submeter os dados de um clube de assinaturas de livros infantis ao processo, foi possível demonstrar com sucesso a importância de cada uma das etapas.

A regressão logística se mostrou o mais eficaz para previsão de *churn* dentre os métodos avaliados, apresentando acuracidade de 76,4% e 52,8% de medida-F1, além de um retorno financeiro esperado superior ao dos outros métodos como *clustering* e até redes neurais MLP.

Avaliando a estabilidade do método, a regressão logística manteve níveis satisfatórios de acuracidade ao receber dados provenientes de outros meses. Porém a baixa densidade de *churn* na população e a falta de disponibilidade de variáveis com relação mais forte com o processo de tomada de decisão do cliente prejudicou a precisão e o recall do modelo.

5.1. Contribuições do trabalho

Essa dissertação contribuiu com lacunas importantes de conhecimento da área de previsão de *churn* em mercados de compra recorrente. A estruturação do processo de construção de modelos de previsão de *churn* com a inclusão das etapas de tratamento de dados, construção de amostras e análise de estabilidade fornece uma forma padronizada para se testar diversos métodos diferentes e comparar os resultados dentro do mesmo contexto.

Além disso, ao endereçar o mercado de clubes de assinatura que, até o momento, apresenta uma falta de publicações que enderecem as condições específicas desse mercado, também ajuda a compreender um pouco melhor as especificidades do modelo de negócio.

Por fim, através da aplicação prática do *framework* realizada através dos dados de um clube de assinaturas de livros infantis, foi possível construir um modelo de previsão com resultados superiores ao modelo atualmente utilizado pela empresa e que tem potencial para impactar positivamente os resultados de negócio.

5.2. Sugestão de trabalhos futuros

Apesar da melhora na capacidade de previsão do modelo proposto baseado em regressão logística versus o modelo atual utilizado pela empresa objeto de estudo, existem ainda possibilidades que podem ser avaliadas para atingir previsões ainda mais acuradas.

A principal delas é a inclusão de mais variáveis preditoras ao modelo, para que possa refletir ainda mais o processo de tomada de decisão do assinante. Variáveis relacionadas a histórico de inadimplência, compras adicionais realizadas além do *kit* mensal e o índice de satisfação com o produto e o atendimento podem trazer mais acuracidade ao modelo ao incorporar a ele mais parâmetros dinâmicos.

Em relação aos métodos de previsão avaliados, as redes neurais apresentam um potencial ainda a ser explorado. Otimizações relacionadas à

estrutura da rede, número de neurônios por camada e ao algoritmo de aprendizado podem, além de fornecer previsões mais acuradas, trazer mais estabilidade ao modelo.

Por fim, existem diversos métodos disponíveis na literatura que podem ser testados e comparados com os resultados apresentados até aqui. Algoritmos de aprendizado de máquina como *random forest*, *suport vector machine*, *gradiente boosting* e árvores de decisão podem ser avaliados individualmente ou de forma combinada a outros métodos, como a regressão logística e redes neurais, buscando previsões ainda melhores.

A grande diversidade de métodos atualmente disponíveis e validados em outros mercados abre muitas possibilidades de continuidade da aplicação do *framework* proposto no contexto dos clubes de assinatura. Futuras aplicações de modelos diferentes vão ajudar não só a validar o processo estruturado proposto, mas também a gerar mais conhecimento sobre o mercado de clubes de assinatura.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ADIPUTRA, I. N. M.; WANCHAI, P. CTGAN-ENN: a tabular GAN-based hybrid sampling method for imbalanced and overlapped data in customer churn prediction. **Journal of Big Data**, v. 11, n. 1, p. 121, 2 set. 2024.

AHMAD, A. K.; JAFAR, A.; ALJOUMAA, K. Customer churn prediction in telecom using machine learning in big data platform. **Journal of Big Data**, v. 6, n. 1, 2019.

ALBOUKAEY, N.; JOUKHADAR, A.; GHNEIM, N. Dynamic behavior based churn prediction in mobile telecom. **Expert Systems with Applications**, v. 162, p. 113779, dez. 2020.

AMIN, A.; AL-OBEIDAT, F.; SHAH, B.; ADNAN, A.; LOO, J.; ANWAR, S. Customer churn prediction in telecommunication industry using data certainty. **Journal of Business Research**, v. 94, p. 290–301, 2019.

AMIN, ADNAN; ANWAR, S.; ADNAN, A.; NAWAZ, M.; HOWARD, N.; QADIR, J.; HAWALAH, A.; HUSSAIN, A. Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study. **IEEE Access**, v. 4, p. 7940–7957, 2016.

AMIN, A; ANWAR, S.; ADNAN, A.; NAWAZ, M.; HOWARD, N.; QADIR, J.; HAWALAH, A.; HUSSAIN, A. Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study. **IEEE Access**, v. 4, p. 7940–7957, 2016.

ANDONOVA Y.; ANAZA A. N.; BENNETT D. H. S. Riding the subscription box wave: Understanding the landscape, challenges, and critical success factors of the subscription box industry. **Business Horizins**, v. 64, p. 631–646, 2021.

BAHNSEN, A. C.; AOUADA, D.; OTTERSTEN, B. A novel cost-sensitive framework for customer churn predictive modeling. **Decision Analytics**, v. 2, n. 1, dez. 2015.

BIANCHESI, N. M. P.; MATTA, C. E. DA; STREITENBERGER, S. C.; ROMÃO, E. L.; BALESTRASSI, P. P.; COSTA, A. F. B. A nonlinear time-series prediction methodology based on neural networks and tracking signals. **Production**, v. 32, 2022.

BOOZARY, P.; SHEYKHAN, S.; GHORBANTANHAEI, H.; MAGAZZINO, C. Enhancing customer retention with machine learning: A comparative analysis of

ensemble models for accurate churn prediction. **International Journal of Information Management Data Insights**, v. 5, n. 1, p. 100331, jun. 2025.

BOUKROUH, I.; AZMANI, A. Explainable machine learning models applied to predicting customer churn for e-commerce. **IAES International Journal of Artificial Intelligence (IJ-AI)**, v. 14, n. 1, p. 286, 1 fev. 2025.

BUCKINX, W.; POEL, D. VAN DEN. Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. **European Journal of Operational Research**, v. 164, n. 1, p. 252–268, jul. 2005.

BUNKHUMPORNPAT, C.; SINAPIROMSARAN, K.; LURSINSAP, C. Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem. *Em*: [s.l: s.n.]. p. 475–482.

CAIGNY, A. DE; COUSSEMENT, K.; BOCK, K. W. DE. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. **European Journal of Operational Research**, v. 269, n. 2, p. 760–772, set. 2018.

CAIGNY, A. DE; COUSSEMENT, K.; BOCK, K. W. DE; LESSMANN, S. Incorporating textual information in customer churn prediction models based on a convolutional neural network. **International Journal of Forecasting**, v. 36, n. 4, p. 1563–1578, 2020.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 1 jun. 2002.

CHAWLA, N. V.; LAZAREVIC, A.; HALL, L. O.; BOWYER, K. W. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. *Em*: **Knowledge discovery in databases> PKDD 2003**. [s.l.] Springer, 2003. v. 2838p. 107–119.

CHEN, T.; FENYO, K.; YANG, S.; ZHANG, J. **Thinking inside the subscription box: new research on e-commerce consumers**. Disponível em: <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/thinking-inside-the-subscription-box-new-research-on-ecommerce-consumers>. Acesso em: 23 maio. 2022.

CHENG, A. **Subscription Boxes May Have Lost The Hype, But Don't Count Them Out Just Yet**. Disponível em: <https://www.forbes.com/sites/andriacheng/2019/12/20/the-crowded-

subscription-box-industry-still-holds-growth-promise/?sh=5b1c526f21ae>. Acesso em: 29 out. 2023.

COUSSEMENT, K.; LESSMANN, S.; VERSTRAETEN, G. A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. **Decision Support Systems**, v. 95, p. 27–36, mar. 2017.

DABLAIN, D.; KRAWCZYK, B.; CHAWLA, N. V. DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. **IEEE Transactions on Neural Networks and Learning Systems**, v. 34, n. 9, p. 6390–6404, set. 2023.

DZIECIOLOWSKI, K. **Predicting customer attrition with Markov chain**European Conference on Data Mining 2015, ECDM 2015 and International Conferences on Intelligent Systems and Agents 2015, ISA 2015 and Theory and Practice in Modern Computing 2015, TPMC 2015. **Anais**...Las Palmas de Gran Canaria: jul. 2015

F., S. Machine-Learning Techniques for Customer Retention: A Comparative Study. **International Journal of Advanced Computer Science and Applications**, v. 9, n. 2, 2018.

FADER, P. S.; HARDIE, B. G. S. Probability Models for Customer-Base Analysis. **Journal of Interactive Marketing**, v. 23, n. 1, p. 61–69, 1 fev. 2009.

GARRIDO, F.; VERBEKE, W.; BRAVO, C. A Robust profit measure for binary classification model evaluation. **Expert Systems with Applications**, v. 92, p. 154–160, 2018.

GATTERMANN-ITSCHERT, T.; THONEMANN, U. W. How training on multiple time slices improves performance in churn prediction. **European Journal of Operational Research**, v. 295, n. 2, p. 664–674, 2021.

GORDINI, N.; VEGLIO, V. Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. **Industrial Marketing Management**, v. 62, p. 100–107, abr. 2017.

GÜNTHER, C.-C.; TVETE, I. F.; AAS, K.; SANDNES, G. I.; BORGAN, Ø. Modelling and predicting customer churn from an insurance company. **Scandinavian Actuarial Journal**, n. 1, p. 58–71, 2014.

GUO, R.; XIONG, W.; ZHANG, Y.; HU, Y. Enhancing game customer churn prediction with a stacked ensemble learning model. **The Journal of Supercomputing**, v. 81, n. 1, p. 178, 14 jan. 2025.

GUPTA, A.; EILERT, M.; GENTRY, J. W. Can I surprise myself? A conceptual framework of surprise self-gifting among consumers. **Journal of Retailing and Consumer Services**, v. 54, p. 101712, 1 maio 2020.

GÜR ALI, Ö.; ARITÜRK, U. Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. **Expert Systems with Applications**, v. 41, n. 17, p. 7889–7903, dez. 2014.

HAN, H.; WANG, W.-Y.; MAO, B.-H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *Em*: [s.l: s.n.]. p. 878–887.

HOANG, H. D.; CAM, N. T. Do they like your game? Early-stage churn prediction using a two-phase neural network system. **Engineering Applications of Artificial Intelligence**, v. 144, p. 110102, mar. 2025.

IMANI, M.; ARABNIA, H. R. Hyperparameter Optimization and Combined Data Sampling Techniques in Machine Learning for Customer Churn Prediction: A Comparative Analysis. **Technologies**, v. 11, n. 6, p. 167, 26 nov. 2023.

IMARC. **Subscription Box Market Report by Type (Replenishment Subscription, Curation Subscription, Access Subscription), Gender (Male, Female), Application (Clothing and Fashion, Beauty, Food and Beverages, Pet Food, Baby Products, Health and Fitness, and Others), and Region 2024-2032**. [s.l: s.n.]. Disponível em: <https://www.imarcgroup.com/subscription-box-market>. Acesso em: 31 out. 2023.

JABIL. **Connected packaging perceptions and attitudes: A customer insight surveywww.jabil.com/dam/jcr:ecdb74e6-c34f-4c30-aa34-c10269617db6/2021-connected-packaging-survey.pdf**. [s.l: s.n.]. Disponível em: <www.jabil.com/dam/jcr:ecdb74e6-c34f-4c30-aa34-c10269617db6/2021-connected-packaging-survey.pdf>. Acesso em: 9 set. 2023.

KERAMATI, A.; JAFARI-MARANDI, R.; ALIANNEJADI, M.; AHMADIAN, I.; MOZAFFARI, M.; ABBASI, U. Improved churn prediction in telecommunication industry using data mining techniques. **Applied Soft Computing Journal**, v. 24, p. 994–1012, 2014.

KERSCHBAUMER, R. H.; KREIMER, D.; FOSCHT, T.; EISINGERICH, A. B. Subscription commerce: an attachment theory perspective. **International**

Review of Retail, Distribution and Consumer Research, v. 33, n. 1, p. 92–115, 2023.

KHINE, S. T.; MYO, W. W. **Mining Customer Churns for Banking Industry using K-means and Multi-layer Perceptron**2023 IEEE Conference on Computer Applications (ICCA). **Anais**...IEEE, 27 fev. 2023

KHODABANDEHLOU, S.; ZIVARI RAHMAN, M. Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. **Journal of Systems and Information Technology**, v. 19, n. 1/2, p. 65–93, 13 mar. 2017.

KIGUCHI, M.; SAEED, W.; MEDI, I. Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest. **Applied Soft Computing**, v. 118, 2022.

KIM, K.; JUN, C.-H.; LEE, J. Improved churn prediction in telecommunication industry by analyzing a large network. **Expert Systems with Applications**, v. 41, n. 15, p. 6575–6584, 2014.

KUMAR, H. **Subscription Box Industry Trends and Opportunities for Businesses**.

LALWANI, P.; MISHRA, M. K.; CHADHA, J. S.; SETHI, P. Customer churn prediction system: a machine learning approach. **Computing**, v. 104, n. 2, p. 271–294, 2022.

LEHMANN, M. A Box for 'Bad Readers'? Bookish Gifts in Subscription Book Boxes. **Publishing Research Quarterly**, v. 36, n. 3, p. 365–377, 23 set. 2020.

LI, J.; BAI, X.; XU, Q.; YANG, D. Identification of Customer Churn Considering Difficult Case Mining. **Systems**, v. 11, n. 7, p. 325, 25 jun. 2023.

LI, X.; LI, Z. A hybrid prediction model for e-commerce customer churn based on logistic regression and extreme gradient boosting algorithm. **Ingenierie des Systemes d'Information**, v. 24, n. 5, p. 525–530, 2019.

LI, Y.; WEI, J.; KANG, K.; WU, Z. An efficient noise-filtered ensemble model for customer churn analysis in aviation industry. **Journal of Intelligent & Fuzzy Systems**, v. 37, n. 2, p. 2575–2585, 9 set. 2019.

LI, Z.; ZHAO, X.; OU, G. Understanding customer satisfaction in curated subscription services: Moderating roles of subscription time and information disclosure. **Information & Management**, v. 60, n. 6, p. 103834, set. 2023.

LIMA, M. **Clubes de livros podem salvar mercado editorial? Leia mais em: https://forbes.com.br/forbes-money/2022/07/clubes-de-livros-por-assinatura-crescem-e-sao-aposta-do-mercado-editorial/**. Disponível em: <https://forbes.com.br/forbes-money/2022/07/clubes-de-livros-por-assinatura-crescem-e-sao-aposta-do-mercado-editorial/>. Acesso em: 14 dez. 2023.

LIU, Y.; FAN, J.; ZHANG, J.; YIN, X.; SONG, Z. Research on telecom customer churn prediction based on ensemble learning. **Journal of Intelligent Information Systems**, v. 60, n. 3, p. 759–775, 14 jun. 2023.

LONG, H. V.; SON, L. H.; KHARI, M.; ARORA, K.; CHOPRA, S.; KUMAR, R.; LE, T.; BAIK, S. W. A New Approach for Construction of Geodemographic Segmentation Model and Prediction Analysis. **Computational Intelligence and Neuroscience**, v. 2019, p. 1–10, 13 maio 2019.

LU, N.; LIN, H.; LU, J.; ZHANG, G. A Customer Churn Prediction Model in Telecom Industry Using Boosting. **IEEE Transactions on Industrial Informatics**, v. 10, n. 2, p. 1659–1665, maio 2014.

MONTGOMERY, D. C.; RUNGER, G. C. **Applied Statistics and Probability for Engineers**. 5ª ed. [s.l.] John Wiley & Sons, 2011.

NESLIN, S. A.; GUPTA, S.; KAMAKURA, W.; LU, J.; MASON, C. H. Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. **Journal of Marketing Research**, v. 43, n. 2, p. 204–211, 1 maio 2006.

NIELSEN BOOKSCAN. **Painel do varejo de livros no Brasil. Resultados 2023 x 2022**. Disponível em: <https://snel.org.br/wp/wp-content/uploads/2024/01/SNEL_13_2023_-_13T_2023.pdf>. Acesso em: 31 jan. 2024.

NILSEN BOOKDATA. **Panorama do consumo de livros. Um estudo sobre o perfil e hábitos de compradores de livros no Brasil**. Disponível em: <https://cbl.org.br/wp-content/uploads/2023/12/Pesquisa-Panorama-do-Consumo-de-Livros_para-publicacao_V1.pdf>. Acesso em: 31 jan. 2024.

NOORDA, R. The Element of Surprise: A Study of Children's Book Subscription Boxes in the USA. **Publishing Research Quarterly**, v. 35, n. 2, p. 223–235, 15 jun. 2019.

NORRICK-RÜHL, C. **Book Clubs and Book Commerce**. [s.l.] Cambridge University Press, 2020.

OUF, S.; MAHMOUD, K. T.; ABDEL-FATTAH, M. A. A proposed hybrid framework to improve the accuracy of customer churn prediction in telecom industry. **Journal of Big Data**, v. 11, n. 1, p. 70, 9 maio 2024.

PERIŠIĆ, A.; JUNG, D. Š.; PAHOR, M. Churn in the mobile gaming field: Establishing churn definitions and measuring classification similarities. **Expert Systems with Applications**, v. 191, p. 116277, abr. 2022.

PERLICH, C.; PROVOST, F.; SIMONOFF, J. S. **Tree Induction vs. Logistic Regression: A Learning-Curve AnalysisJournal of Machine Learning Research**. [s.l: s.n.].

PUSTOKHINA, I. V; PUSTOKHIN, D. A.; RH, A.; JAYASANKAR, T.; JEYALAKSHMI, C.; DÍAZ, V. G.; SHANKAR, K. Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms. **Information Processing and Management**, v. 58, n. 6, 2021.

RAHMAN, S.; JANSSENS, B.; BOGAERT, M. Profit-driven pre-processing in B2B customer churn modeling using fairness techniques. **Journal of Business Research**, v. 189, p. 115159, fev. 2025.

RAO, S.; POOJARY, P.; SOMAIYA, J.; MAHAJAN, P. A COMPARATIVE STUDY BETWEEN VARIOUS PREPROCESSING TECHNIQUES FOR MACHINE LEARNING. **International Journal of Engineering Applied Sciences and Technology**, v. 5, n. 3, p. 431–438, 31 jul. 2020.

RISSELADA, H.; VERHOEF, P. C.; BIJMOLT, T. H. A. Staying Power of Churn Prediction Models. **Journal of Interactive Marketing**, v. 24, n. 3, p. 198–208, 1 ago. 2010a.

___. Staying Power of Churn Prediction Models. **Journal of Interactive Marketing**, v. 24, n. 3, p. 198–208, 1 ago. 2010b.

SÁEZ, J. A.; LUENGO, J.; STEFANOWSKI, J.; HERRERA, F. SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. **Information Sciences**, v. 291, p. 184–203, jan. 2015.

SAHA, L.; TRIPATHY, H. K.; GABER, T.; EL-GOHARY, H.; EL-KENAWY, E.-S. M. Deep Churn Prediction Method for Telecommunication Industry. **Sustainability**, v. 15, n. 5, p. 4543, 3 mar. 2023.

SEETHARAMAN, P. Business models shifts: Impact of Covid-19. **International Journal of Information Management**, v. 54, 1 out. 2020.

SILVA, E. N. DA; MAGALHAES, F. B.; SALCEDO, W. J. **Customer Churn Prediction in E-Commerce base Using Machine Learning and LIME Algorithm**2024 IEEE ANDESCON. **Anais**...IEEE, 11 set. 2024

STRIPLING, E.; BROUCKE, S. VANDEN; ANTONIO, K.; BAESENS, B.; SNOECK, M. Profit maximizing logistic model for customer churn prediction using genetic algorithms. **Swarm and Evolutionary Computation**, v. 40, p. 116–130, 2018.

SUNDARKUMAR, G. G.; RAVI, V. A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. **Engineering Applications of Artificial Intelligence**, v. 37, p. 368–377, jan. 2015.

TAMADDONI, A.; STAKHOVYCH, S.; EWING, M. Comparing Churn Prediction Techniques and Assessing Their Performance. **Journal of Service Research**, v. 19, n. 2, p. 123–141, 25 maio 2016.

ULLAH, I.; RAZA, B.; MALIK, A. K.; IMRAN, M.; ISLAM, S. U.; KIM, S. W. A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. **IEEE Access**, v. 7, p. 60134–60149, 2019.

VAFEIADIS, T.; DIAMANTARAS, K. I.; SARIGIANNIDIS, G.; CHATZISAVVAS, K. C. A comparison of machine learning techniques for customer churn prediction. **Simulation Modelling Practice and Theory**, v. 55, p. 1–9, 2015.

VERBEKE, W.; DEJAEGER, K.; MARTENS, D.; HUR, J.; BAESENS, B. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. **European Journal of Operational Research**, v. 218, n. 1, p. 211–229, 1 abr. 2012.

VERBRAKEN, T.; VERBEKE, W.; BAESENS, B. A novel profit maximizing metric for measuring classification performance of customer churn prediction models. **IEEE Transactions on Knowledge and Data Engineering**, v. 25, n. 5, p. 961–973, 2013.

WILL M. BERTRAND, J.; FRANSOO, J. C. Operations management research methodologies using quantitative modeling. **International Journal of Operations & Production Management**, v. 22, n. 2, p. 241–264, fev. 2002.

WONG, M. L.; SENG, K.; WONG, P. K. Cost-sensitive ensemble of stacked denoising autoencoders for class imbalance problems in business domain. **Expert Systems with Applications**, v. 141, p. 112918, mar. 2020.

XIAHOU, X; HARADA, Y. B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. **Journal of Theoretical and Applied Electronic Commerce Research**, v. 17, n. 2, p. 458–475, 2022.

XIAHOU, XIANCHENG; HARADA, Y. B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. **Journal of Theoretical and Applied Electronic Commerce Research**, v. 17, n. 2, p. 458–475, 6 abr. 2022.

ZHANG, T.; MORO, S.; RAMOS, R. F. A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation. **Future Internet**, v. 14, n. 3, 2022.

ANEXOS - PUBLICAÇÕES

ESTEVES, T; GOMES, J. H. F. Uma Revisão Sistemática Da Literatura Sobre Modelos De Previsão De Churn Em Modelos De Negócio De Recorrência. **XLIV Encontro Nacional de Engenharia de Produção**. Porto Alegre, 2024.

XLIV ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO
"Reindustrialização no Brasil"
Porto Alegre, Rio Grande do Sul, Brasil, 22 a 25 de outubro de 2024.

UMA REVISÃO SISTEMÁTICA DA LITERATURA SOBRE MODELOS DE PREVISÃO DE *CHURN* EM MODELOS DE NEGÓCIOS DE RECORRÊNCIA

Tayrone Duque Esteves (Universidade Federal de Itajubá)

José Henrique de Freitas Gomes (Universidade Federal de Itajubá)

Neste trabalho, foi realizada uma análise sistemática da literatura mais recente disponível na base SCOPUS com intuito de identificar os principais aspectos relacionados à modelagem de churn. As principais ferramentas de previsão, variáveis preditoras mais comuns e os mercados de atuação foram identificados. Observa-se que as empresas de telecomunicação se destacam como o principal mercado para modelos de previsão, seguidos pelo mercado bancário e o de comércio digital. Quanto às ferramentas para construção de modelos, existe uma grande variabilidade com diferentes graus de customização, com destaque para a regressão logística, utilizada como modelo de referência, as redes neurais que vem recebendo um destaque maior nos últimos anos e as técnicas de clustering. A qualidade das previsões geradas pelos modelos é tida como um diferencial estratégico importante para os modelos de negócios com características de recorrência, alimentando estratégias de retenção e permitindo um melhor entendimento dos comportamentos associados ao cancelamento de assinaturas.

Palavras-chave: Churn, Previsão, Revisão sistemática de literatura, Empresas de assinatura.

CUNHA, F. C.; ESTEVES, T. D.; LIMA, R. S.; PINHO, A. F. Logística Reversa De Equipamentos Eletrônicos – Um Estudo De Caso No Sul De Minas Gerais. **XLII encontro nacional de engenharia de produção**. Fortaleza, 2023.

XLIII ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO
"A contribuição da engenharia de produção para desenvolvimento sustentável das organizações: Cadeias Circulares, sustentabilidade e tecnologias"
Fortaleza, Ceará, Brasil, 16 a 19 de outubro de 2023.

LOGÍSTICA REVERSA DE RESÍDUOS DE EQUIPAMENTOS ELETROELETRÔNICOS - UM ESTUDO DE CASO NO SUL DE MINAS GERAIS

Flávia Gontijo Cunha (Universidade Federal de Itajubá)

Tayrone Duque Esteves (Universidade Federal de Itajubá)

Renato da Silva Lima (Universidade Federal de Itajubá)

Alexandre Ferreira de Pinho (Universidade Federal de Itajubá)

O Brasil alcançou em 2019 a primeira posição no ranking de maior gerador de Resíduos de Equipamento Elétrico e Eletrônico (REEE) na américa do sul. Infelizmente a falta de conscientização da população faz com que a grande maioria desses resíduos sejam descartados juntamente com o resíduo doméstico, e destinados a aterros sanitários ou redes informais, condenando não somente o meio ambiente, como também a saúde da população. Portanto torna-se essencial conhecer as práticas de logística reversa (LR) nos municípios e propor melhorias necessárias. Dessa forma o objetivo do presente artigo foi conhecer as práticas de logística reversa de resíduos de equipamentos elétricos eletrônicos adotadas pela população do município de Itajubá, Minas Gerais. Para isso foi utilizado o método survey, empregando o estudo de amostra aleatória da população, e utilizando o aplicativo Google Forms® para a coleta dos dados. O estudo foi realizadocom 386 pessoas do município. Foi verificado que 63% dos entrevistados não conheciam a forma correta de descarte. Foi identificado também que grande parte das pessoas guarda os resíduos obsoletos em casa, algo que dificulta o processo de logística reversa, e o eletrônico mais descartado pela população são os celulares, seguidos pelos materiais de informática. Sugere-se que o município de Itajubá busque pela execução de campanhas de conscientização, separação e destinação correta dos resíduos eletrônicos, a fim de ocorrer o bom funcionamento do sistema de logística reversa. Outro fator que contribuiria para a adoção de práticas de LR eficientes e eficazes, seria a adoção da fiscalização de todo o processo.

Palavras-chave: Logística Reversa, Resíduos de equipamentos eletroeletrônicos, e-waste.

ESTEVES, T. D.; GOMES, J. H. F.; LIMA, R. S. Análise Do Impacto No Nível De Serviço Logístico Da Redução Da Participação Dos Correios Nas Entregas De Um Clube De Assinaturas De Livros Infantis. **37º congresso de ensino e pesquisa em transportes**. Santos, 2023.

37° ANPET CONGRESSO DE PESQUISA E ENSINO EM TRANSPORTES SANTOS-SP 06 A 10 NOVEMBRO DE 2023

ANÁLISE DO IMPACTO NO NÍVEL DE SERVIÇO LOGÍSTICO DA REDUÇÃO DA PARTICIPAÇÃO DOS CORREIOS NAS ENTREGAS DE UM CLUBE DE ASSINATURA DE LIVROS INFANTIS

Tayrone Duque Esteves
José Henrique de Freitas Gomes
Renato da Silva Lima
Instituto de Engenharia de Produção e Gestão (IEPG)
Universidade Federal de Itajubá (UNIFEI)

RESUMO

Considerando que a logística eficiente é um diferencial estratégico em clubes de assinatura, esse trabalho avaliou o impacto da redução do volume transportado pelos Correios para um clube de assinaturas de livros infantis, usando como critérios as entregas no prazo, reclamações recebidas e custo médio do transporte. No cenário 1, entregas com até 15% a mais de custo via transportadoras privadas foram redistribuídas e no cenário 2 foi realizada uma redistribuição ainda maior, de forma que nos cenários 1 e 2 o volume transportado pelos correios passou de 12 para 10,5 e 8%, respectivamente. No cenário 1 a entrega dentro do prazo aumentou em 0,36%, com redução de 3,9% no custo médio do frete, enquanto para o cenário 2 estes valores foram de 0,51% e 0,6%, o que mostra que o foco na melhoria do nível de serviço de pode trazer uma redução no custo total das entregas.

1. INTRODUÇÃO

O termo "clube de assinatura" é aplicado ao serviço de curadoria, processamento e entrega de diversos tipos de produto, periodicamente, no endereço de escolha dos clientes em um modelo de assinatura recorrente, em geral mensal, trimestral ou anual (Noorda, 2019). Atualmente no mercado, existem diversas empresas que vêm fornecendo esse tipo de serviço para uma variedade de produtos como livros, revistas, cosméticos, roupas e artigos de uso doméstico (Andonova *et al.*, 2021).

O mercado de clubes de assinatura possui uma forte característica de recorrência e fidelização. Mesmo sob essas condições, esse mercado vem apresentando um crescimento significativo na última década, faturando mais de 2,6 bilhões somente no mercado americano em 2016 (Chen et al., 2018) e tendo apresentado um crescimento ainda maior em 2020 motivado principalmente pela pandemia do COVID-19 (Seetharaman, 2020).

Pais, familiares e educadores têm buscado formas de complementar a formação e o desenvolvimento das crianças, principalmente através da leitura e de atividades educacionais interativas. Essas necessidades combinadas à produtos de qualidade, uma curadoria cuidadosa e direcionamento de produtos específicos para cada faixa etária e momento de aprendizado, tem tornado o mercado educacional muito promissor para o crescimento de clubes de assinatura (Noorda, 2019).

De acordo com Bahsen et al. (2015) a aquisição de novos assinantes e a retenção daqueles que já estão na base de clientes são os dois principais objetivos das companhias que fornecem serviços por assinatura. Por isso o nível de serviço oferecido ao cliente se torna um diferencial estratégico para qualquer clube de assinatura.

Swain e Garza (2022) definem nível de serviço como como um acordo entre um fornecedor e

ESTEVES, T. D.; GOMES, L. H. F. Modelagem De Um Processo De Previsão De *Churn* Para Um Clube De Assinaturas De Livros Infantis. **IEPG Summit 2023**. Itajubá, 2023

IEPG SUMMIT'23

PENSANDO O FUTURO COM INOVAÇÃO E SUSTENTABILIDADE

MODELAGEM DE UM PROCESSO DE PREVISÃO DE *CHURN* PARA UM CLUBE DE ASSINATURA DE LIVROS INFANTIS

Tayrone Duque Esteves – tayrone.esteves@gmail.com
José Henrique de Freitas Gomes – ze_henriquefg@unifei.edu.br

Palavras-chave: *Churn*, Previsão, Modelagem, Clube de assinatura.

1 INTRODUÇÃO

O termo "clube de assinatura" é aplicado ao serviço de curadoria e entrega de produto, periodicamente, em um modelo recorrente (Noorda, 2019). Dentre os clubes de assinatura, o fornecimento de material educacional tem se mostrado muito promissor, uma vez que pais e educadores estão buscando formas de complementar a formação e o desenvolvimento das crianças (Noorda, 2019).

Bahnsen, Aouada e Ottersten (2015) afirmaram que a aquisição de novos assinantes e a retenção daqueles que já estão na base de clientes são os dois principais objetivos das companhias que fornecem serviços por assinatura. A quantidade de clientes que deixam de assinar um serviço é definido como "*churn*" (Óskarsdóttir et al., 2018). Desta forma, a previsão do *churn* se apresenta como um dos principais desafios para os clubes de assinatura (He et al., 2021).

Neste contexto, este trabalho tem como objetivo construir um modelo que seja capaz de classificar os assinantes que compõem a base entre aqueles com alta e baixa probabilidade de cancelamento, para que sirva de insumo na tomada de decisão estratégica em um clube de assinatura de livros infantis.