UNIVERSIDADE FEDERAL DE VIÇOSA

JOÃO MARCOS ALVES MODESTO RAMOS

PREVISÃO DO VALOR VITALÍCIO DO CLIENTE COM ABORDAGENS DE APRENDIZADO DE MÁQUINA

FLORESTAL - MINAS GERAIS
2024

JOÃO MARCOS ALVES MODESTO RAMOS

PREVISÃO DO VALOR VITALÍCIO DO CLIENTE COM ABORDAGENS DE APRENDIZADO DE MÁQUINA

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

Orientador: Fabrício Aguiar Silva

FLORESTAL - MINAS GERAIS
2024

Ficha catalográfica elaborada pela Biblioteca da Universidade Federal de Viçosa - Campus Florestal

T
R175p
2024

Ramos, João Marcos Alves Modesto, 1999-
Previsão do valor vitalício do cliente com abordagens de aprendizado de máquina / João Marcos Alves Modesto Ramos. – Florestal, MG, 2024.
1 dissertação eletrônica (52 f.): il. (algumas color.).

Orientador: Fabrício Aguiar Silva.
Dissertação (mestrado) - Universidade Federal de Viçosa, Instituto de Ciências Exatas e Tecnológicas, 2024.
Referências bibliográficas: f. 49-52.
DOI: https://doi.org/10.47328/ufvcaf.2024.016
Modo de acesso: World Wide Web.

1. Clientes - Fidelização. 2. Aprendizado do computador. I. Silva, Fabrício Aguiar, 1981-. II. Universidade Federal de Viçosa. Instituto de Ciências Exatas e Tecnológicas. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 23. ed. 006.31

JOÃO MARCOS ALVES MODESTO RAMOS

PREVISÃO DO VALOR VITALÍCIO DO CLIENTE COM ABORDAGENS DE APRENDIZADO DE MÁQUINA

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

APROVADA: 19 de julho de 2024.

Assentimento:

João Marcos Alves Modesto Ramos
Autor

Fabrício Aguiar Silva
Orientador

Aos meus mestres, que me guiaram nesta jornada, e a todos aqueles que me inspiraram a buscar o conhecimento e a excelência.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

À Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), pela concessão da bolsa de estudos.

À Universidade Federal de Viçosa (UFV), pela oportunidade de participar do programa de pós-graduação em Ciência da Computação.

Agradeço imensamente à minha família, pilar fundamental em toda a minha jornada. À minha mãe, por sempre acreditar em mim e me proporcionarem todas as oportunidades para que eu pudesse chegar até aqui. Aos meus tios, tia e avó, por serem um exemplo de caminho a a seguir. A vocês, dedico esta conquista, fruto de tanto amor e incentivo.

À minha companheira, agradeço por toda a paciência, compreensão e amor durante este período. Seu apoio foi essencial para que eu pudesse me dedicar integralmente a este trabalho. Sua presença em minha vida torna tudo mais leve e feliz.

Agradeço profundamente aos meus orientadores, pela dedicação, paciência e conhecimento transmitidos. Suas orientações foram cruciais para o desenvolvimento deste trabalho e para minha formação acadêmica. Sou imensamente grato por toda a confiança depositada em mim.

Agradeço as minhas colegas de trabalho, que sempre estiveram dispostos a ajudar e compartilhar seus conhecimentos.

“Ensinar não é transferir conhecimento, mas criar as possibilidades para a sua própria produção ou a sua construção."(Paulo Freire)

RESUMO

A mudança de paradigma para estratégias de negócio centradas no cliente demonstrou que focar na retenção de clientes e em relacionamentos de longo prazo resulta em modelos de negócios mais lucrativos e sustentáveis em comparação com abordagens centradas no produto. Essa mudança, facilitada pelos avanços na tecnologia, permite que as empresas personalizem suas ofertas com base nas preferências e comportamentos individuais dos clientes, levando a um aumento nos lucros e na satisfação dos clientes. O Valor Vitalício do Cliente (do inglês *Customer Lifetime Value* ou apenas CLV) é uma métrica importante para identificar os relacionamentos com clientes mais lucrativos. O CLV é definido como o valor total que um cliente contribui para uma empresa ao longo de todo o seu relacionamento. Estimar o CLV é desafiador devido aos diversos contextos dos relacionamentos com clientes, como interações contratuais versus não contratuais e discretas versus contínuas. Além disso, a disponibilidade de dados e as preocupações com a privacidade dificultam os modelos de previsão. Modelos probabilísticos tradicionais, como Cadeias de Markov, Pareto/NBD e BG/NBD, têm sido amplamente utilizados, ao contrário dos modelos baseados em aprendizado de máquina, que muitas vezes são limitados pela dependência de dados detalhados dos clientes, que nem sempre estão disponíveis ou não condizem com a ética. O objetivo desta dissertação é apresentar um modelo baseado em aprendizado de máquina para prever o CLV em diversos contextos sem utilizar dados sensíveis dos clientes, alcançando resultados competitivos com os métodos mais avançados. Os objetivos específicos incluem propor um modelo de aprendizado de máquina baseado em RFM e outro modelo que utiliza atributos mais complexos baseados em transações. Esta dissertação está organizada em dois artigos principais. O primeiro artigo, apresentado no Capítulo 2, desenvolve uma solução usando modelos probabilísticos e técnicas de aprendizado de máquina, mostrando que os modelos propostos superam os métodos probabilísticos tradicionais na previsão do número de transações. O segundo artigo, detalhado no Capítulo 3, aprimora o modelo incorporando dados relacionados às transações e testando-o em diferentes contextos, como o setor bancário, demonstrando um desempenho melhorado em todos os aspectos, particularmente na previsão do CLV. Como resultados, os modelos de aprendizado de máquina propostos fornecem uma solução generalizável para a estimativa do CLV que prioriza a privacidade dos clientes e se adapta a diversos contextos de negócios, aprimorando a capacidade das empresas de prever o valor dos clientes e personalizar seus serviços.

RAMOS, João Marcos Alves Modesto, M.Sc., Universidade Federal de Viçosa, Julho

de 2024. **Previsão do Valor Vitalício do Cliente com abordagens de aprendizado de máquina**. Orientador: Fabrício Aguiar Silva.

Palavras-chave: Customer Lifetime Value. CLV. LTV. Aprendizado de Máquina.

ABSTRACT

The paradigm shift towards customer-centric business strategies has demonstrated that focusing on customer retention and long-term relationships results in more profitable and sustainable business models compared to product-centric approaches. This shift, facilitated by advances in technology, allows companies to personalize their offerings based on individual customer preferences and behaviors, leading to increased profits and customer satisfaction.

Customer Lifetime Value (CLV) is a crucial metric for identifying the most profitable customer relationships. CLV is defined as the total value a customer contributes to a company over their entire relationship. Estimating CLV is challenging due to the various contexts of customer relationships, such as contractual versus non-contractual and discrete versus continuous interactions. Additionally, data availability and privacy concerns complicate predictive modeling. Traditional probabilistic models, such as Markov Chains, Pareto/NBD, and BG/NBD, have been widely used. However, these models are often limited by their reliance on detailed customer data, which is not always available or ethical to use.

The objective of this dissertation is to present a machine learning-based model to predict CLV in various contexts without using sensitive customer data, achieving competitive results with the most advanced methods. The specific objectives include proposing an RFM-based machine learning model and another model that utilizes more complex transaction-based attributes.

This dissertation is organized into two main articles. The first article, presented in Chapter 2, develops a solution using probabilistic models and machine learning techniques, showing that the proposed models outperform traditional probabilistic methods in predicting the number of transactions. The second article, detailed in Chapter 3, enhances the model by incorporating transaction-related data and testing it in different contexts, such as the banking sector, demonstrating improved performance in all aspects, particularly in CLV prediction.

As results, the proposed machine learning models provide a generalizable solution for CLV estimation that prioritizes customer privacy and adapts to various business contexts, enhancing companies' ability to predict customer value and personalize their services.

RAMOS, João Marcos Alves Modesto, M.Sc., Universidade Federal de Viçosa, July, 2024. **Customer Lifetime Value prediction with machine learning approaches**. Advisor: Fabrício Aguiar Silva.

Keywords: Customer Lifetime Value. CLV. LTV. Machine Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 Introdução

Com o passar do tempo, observou-se que, no contexto dos negócios, concentrar esforços no desenvolvimento de estratégias de marketing voltadas para a retenção de clientes e na manutenção de relacionamentos de longo prazo com clientes valiosos gera um modelo de negócio mais próspero e lucrativo do que focar apenas no desenvolvimento do produto. As organizações que adotaram essa mudança são conhecidas como organizações centradas no cliente e oferecem soluções completas de produtos projetadas com um profundo entendimento das necessidades dos clientes. Isso resulta em maiores lucros, maior envolvimento dos funcionários e clientes mais satisfeitos no setor privado. Esse fenômeno é decorrente do recente desenvolvimento de tecnologias digitais, que permitiu às empresas identificarem as preferências específicas e os padrões de comportamento de clientes individuais, atendendo perfeitamente a essas preferências e fornecendo um portfólio personalizado de produtos e serviços para cada indivíduo (Shin and Kim, 2022).

Com isso, surge a demanda de identificar quais relacionamentos entre empresas e clientes são os mais lucrativos, visto que, de acordo com o Princípio de Pareto aplicado em estratégias de marketing, 20% dos clientes resultam em 80% da receita, potencialmente quadruplicando o lucro de uma empresa ao atingir e reter clientes altamente lucrativos (Kruger, 2011).

Uma das maneiras de identificar os clientes mais lucrativos é através do cálculo do *Customer Lifetime Value* (*CLV*), que consiste no valor total de um cliente para uma companhia durante todo o seu período de relacionamento.

A definição e cálculo do *Customer Lifetime Value* (*CLV*) variam entre os autores. Segundo Dwyer (1997), o *CLV* representa o valor presente dos benefícios esperados subtraídos pelos encargos dos clientes. No trabalho de Aeron et al. (2012), um cliente lucrativo é definido como uma pessoa, empresa ou residência cuja receita excede os gastos da empresa para atraí-lo, vender e servir, e este excedente é considerado como o *CLV*. Em estudos como o de Kahreh et al. (2014), o *CLV* é visto como uma métrica para medir o valor real de um cliente em um segmento específico do mercado.

Contudo, um conceito recorrente nos estudos é a definição apresentada por Dwyer (1997), onde o *CLV* consiste na receita de um cliente subtraída dos custos relacionados àquele cliente, como de produção, marketing, dentre outros. Assim, dado um cliente i em um determinado período de tempo p, temos que o *CLV* é dado pelo valor $V_i(p)$ pago pelo cliente, subtraído pelos custos envolvidos $C_i(p)$ e pelos custos de aquisição AQ_i, tem-se a equação 1.1, onde d é a taxa de desconto, representando o valor do

dinheiro ao longo do tempo.

$$CLV_i(P) = \sum_{p=1}^{\mathrm{P}} \frac{(V_i(p) - C_i(p))}{(1+d)^p} - AQ_i \tag{1.1}$$

Como o objetivo é calcular esse valor para todo o período de vida do cliente, considera-se que $p = 1$ corresponde ao primeiro período, e P corresponde ao último período de tempo considerado na análise. No caso de o cliente não ter realizado alguma transação em um determinado período, temos que $CLV_i(p) = 0$. Em determinados contextos, pode não haver informações detalhadas sobre os custos. Neste caso, considera-se somente a receita, que é o valor pago pelo cliente $V_i(p)$, resultando na fórmula 1.2.

$$CLV_i(P) = \sum_{p=1}^{\mathrm{P}} \frac{V_i(p)}{(1+d)^p} \tag{1.2}$$

Ainda em relação à receita, podemos definir que ela pode ser considerada como o valor monetário médio por transação M que o cliente i realizou no período de tempo p, multiplicado pelo número de transações realizadas NT no período de tempo p. Dadas estas alterações, tem se a fórmula 1.3.

$$CLV_i(P) = \sum_{p=1}^{\mathrm{P}} \frac{NTi(p) \times M_i(p)}{(1+d)^p} \tag{1.3}$$

Baseado neste cálculo, busca-se estimar este valor em um horizonte de tempo h, com o intuito de estimar como o relacionamento entre cliente e companhia se comportará no futuro.

$$CLV_i(P+h) = \sum_{p=1}^{\mathrm{P+h}} \frac{NTi(p) \times M_i(p)}{(1+d)^p} \tag{1.4}$$

Baseado na formula 1.4, surgem diversos modelos que tentam realizar a predição de número de transações esperadas e o valor monetário médio por transação, com o intuito de prever o CLV. Os primeiros modelos que surgiram foram os modelos probabilísticos, onde assume-se que as observações (ou seja, as transações) são geradas por um processo físico que pode ser modelado utilizando distribuições de probabilidade. Entre eles, temos, como por exemplo Cadeias de Markov (Pfeifer and Carraway, 2000), Modelos de Pareto/NBD (Schmittlein et al., 1987) o método RFM (Hughes, 1996) e suas variações (Ullah et al., 2023; Cheng and Chen, 2009).

Cadeias de Markov são modelos estocásticos que descrevem uma sequência de possíveis eventos, onde a probabilidade de um evento acontecer está relacionada ao estado obtido no último evento. Existem certos trabalhos que fazem uso destes

conceitos para modelar o relacionamento de clientes e calcular o *CLV*, como os trabalhos de Pfeifer and Carraway (2000) e de Khajvand et al. (2011).

Os modelos de Pareto/NBD são propostos para analisar uma base de clientes em modelos não-contratuais. Isto é feito através de derivações de expressões para entre outras coisas, (i) a probabilidade de que um cliente com um determinado histórico de transações ainda esteja ativo, e (ii) o número esperado de transações futuras para um cliente, condicionado ao seu histórico de transações (Schmittlein et al., 1987).

Os modelos RFM são aplicados em diversas áreas de Marketing, sendo utilizados para analisar o comportamento de um cliente, e fazer predições baseado nisto. São muito adotados em técnicas de segmentação de clientes, e baseia-se na combinação de três métricas:

- Recência: O quão recente foi feita a última transação.
- Frequência: O quão frequentes são feitas as transações.
- Valor monetário: O quanto, em valor monetário, a transação corresponde.

Baseando-se nestas métricas, estes modelos, que se baseiam no comportamento de compra dos clientes, ajustam uma distribuição de probabilidade ao valor RFM observado dos clientes. Os métodos baseados no RFM não são usados para estimar o CLV. Visto que identificam o comportamento passado do cliente, a proposta de utiliza-los para prever o CLV visa identificar os comportamentos dos clientes, possibilitando antecipar suas futuras atividades (Aslekar et al., 2019). Existem diversos trabalhos que buscam fazer variações deste método, como por exemplo, o método WRFM (Khajvand et al., 2011) que adiciona pesos nos parâmetros RFM dependendo da característica da indústria, ou o modelo de Cheng and Chen (2009), que o complementa utilizando o modelo RFMTC, que adiciona as métricas do tempo da primeira compra e a probabilidade de *churn*.

Os modelos probabilísticos são considerados importantes até o momento atual, visto que, de acordo com o trabalho de Sun et al. (2023), os resultados indicam que grande parte dos trabalhos fazem uso de modelo de cadeia de Markov, RFM e Pareto NBD.

Porém, tem surgido nos últimos anos a utilização de técnicas de aprendizado de máquina para este cálculo, como no trabalho de Desirena et al. (2019), que utiliza redes neurais para aumentar o CLV na indústria de seguros por meio de recomendações mais assertivas. Já o trabalho de Sun et al. (2023) faz uso de aprendizado de máquina para medir o CLV e a segmentação com base no valor do ciclo de vida do cliente.

Ainda assim, as soluções que surgem fazem uso de atributos pessoais dos clientes, e informações que em determinados contextos se tornam ausentes, como no trabalho

de Kailash et al. (2023) que utiliza dados demográficos, grau de escolaridade, renda e posses do cliente. A utilização desses tipos de dados dificulta a reprodutibilidade dos modelos em outros contextos, além de ferir a privacidade do cliente, preferindo-se utilizar os modelos probabilísticos. Portanto, ainda há a necessidade de uma solução que seja genérica para ser adotada em diferentes segmentos de mercado, e que não ameace a privacidade dos clientes.

1.1 Objetivos

O objetivo geral desta dissertação consiste em estimar o CLV, constituindo-se da estimativa do número médio de transações e o valor monetário médio esperado por transação de cada cliente, utilizando modelos de aprendizado de máquina, que funcione para diversos contextos sem utilizar dados sensíveis e que apresente resultados competitivos com o estado da arte. Os objetivos específicos estão descritos a seguir:

- Propor um modelo de aprendizado de máquina baseado em RFM.
- Propor um modelo de aprendizado de máquina baseado em atributos mais elaborados com base nas transações.

1.2 Justificativa

Uma estimativa efetiva do CLV pode trazer diversos benefícios tanto para as companhias quanto para seus clientes. Para as companhias, os benefícios incluem a identificação dos clientes mais valiosos e a capacidade de entender quais clientes estão em risco de *churn*[1]. Isso permite que a empresa tome medidas para recuperar o interesse desses clientes ou identifique os motivos do abandono. Para os clientes, os serviços podem ser ofertados de forma personalizada, de acordo com suas preferências, além de proporcionar um atendimento mais alinhado às suas necessidades.

Porém, este cálculo não é algo trivial, visto que apresenta uma série de dificuldades. A primeira consiste no fato de que o cálculo do CLV varia de acordo com o contexto. Por exemplo, em empresas que vendem serviços de assinaturas, onde os clientes possuem um contrato com a companhia, ou em um contexto de vendas em atacado, os métodos de cálculo podem diferir.

Pode-se separar o contexto do negócio de acordo com o tipo de relacionamento. Se o relacionamento é baseado em um contrato, onde se sabe quando o cliente ainda

[1]Churn é um indicador de quando um cliente encerra o seu relacionamento com a empresa.

é ativo na base de dados, chama-se de **contratual**. Caso contrário, é denominado **não-contratual**. Essa distinção ajuda a entender se o cliente está ou não em processo de *churn*.

Outra separação é de acordo com o momento que o cliente realiza as transações. Caso as transações sejam realizadas sempre em um momento específico, seja isto por uma assinatura ou por um período especificado pela companhia para realizar o pagamento, são chamados de **discretos**. Caso o cliente possa realizar transações em qualquer momento, são considerados **contínuas**. Esta divisão permite entender quando será feita a próxima transação do cliente.

Com estas duas separações, é possível definir uma gama de companhias, como podemos ver na figura 1.1 alguns exemplos de companhias utilizando estas duas divisões.

Tabela de Exemplos do Contexto de Negócios	**Não-Contratual** O relacionamento com o cliente não é regido por um contrato ou adesão.	**Contratual** As companhias geralmente tem uma expectativa de quando um cliente irá se tornar inativo.
Contínuas As transações ocorrem a qualquer momento.	Estadias de Hotel, Consultas Médicas, Lojas online, Drogarias, Bancos	Cartões de Crédito. Chips SIM pós-pago.
Discretas As transações ocorrem em um determinado momento.	Caridade, Presença em Eventos, Shows.	Assinatura de Revistas, Academias, Grande Parte dos Serviços de Seguros.

Figura 1.1: Separação de acordo com o contexto dos negócios

Os modelos probabilísticos, que são considerados de propósito geral, trabalham com o contexto não-contratual e contínuo, que são os mais desafiadores. Ainda assim, eles podem ser aplicados nos outros contextos, porém, são descartados as vantagens e atributos adicionais fornecidos, como a data esperada da próxima compra no contexto discreto e a atividade do cliente no contexto contratual.

Uma outra dificuldade encontrada consiste na utilização de diferentes tipos de dados. Pode-se observar em trabalhos como o de Vanderveld et al. (2016), que utiliza dados de um comércio eletrônico, contendo transações, informações sobre o engajamento, dados do cliente, tanto características e demográficos, enquanto Qi et al. (2015) utilizam dados de uma companhia telefônica, contendo informações sobre a satisfação, lealdade e algumas características de cada cliente. Portanto, replicar soluções que funcionam em um contexto pode não ser possível por falta de dados.

Além da diferença dos dados entre um contexto e outro, surge-se também a questão da privacidade do cliente, visto que há dados sensíveis sendo utilizados, podendo conter informações básicas, como características de um

cliente (Calabourdin and Aksenov, 2023), como informações mais pessoais, endereço (Kailash et al., 2023), horários que o mesmo realiza compras (Vanderveld et al., 2016) ou até mesmo sobre o estilo de vida (Haenlein et al., 2007).

Apesar dos avanços, grande parte dos trabalhos fazem uso de aprendizado de máquina somente na etapa de segmentação dos clientes, abordando a estimativa do CLV como um problema de classificação (Hiziroglu and Sengul, 2012; Kahreh et al., 2014; ABIDAR et al., 2023).

Este trabalho apresenta uma solução para os problemas descritos anteriormente, por meio da criação de um modelo de aprendizado de máquina de propósito geral, com enfoque no contexto não-contratual e contínuo. O modelo proposto visa priorizar a privacidade dos clientes utilizando somente informações das transações dos clientes, além de proporcionar uma estimativa mais precisa e adaptável do CLV, independentemente das variações de contexto e tipos de dados. Esta abordagem não só aprimora a capacidade das empresas de prever o valor dos clientes ao longo do tempo, mas também contribui para a melhoria do relacionamento com os clientes, ao possibilitar um serviço mais personalizado e eficiente.

1.3 Organização da dissertação

A organização desta dissertação está em conformidade com o padrão estabelecido pela Comissão do Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Viçosa (UFV). São apresentados dois artigos em formato de coletânea, sendo o primeiro publicado em uma conferência e o segundo submetido para o XXXVIII Simpósio Brasileiro de Banco de Dados (SBBD 2024).

O Capítulo 2 apresenta o artigo (Ramos and Silva, 2023), onde foi elaborada uma solução com base em modelos probabilísticos, fundamentada nos conceitos dos modelos RFM, juntamente com técnicas de aprendizado de máquina. Foi realizada uma comparação entre a proposta e os modelos probabilísticos, permitindo observar que, em relação ao número de transações previstas em situações reais, os modelos de aprendizado de máquina propostos se destacam em relação aos modelos da literatura. No que tange à previsão do valor monetário médio, observou-se que ambos os modelos comparados apresentam resultados competitivos, embora os modelos de aprendizado de máquina se sobressaiam.

O Capítulo 3 é composto de melhorias visando a inclusão de dados relacionados às transações, e a testagem de diversos modelos em diferentes contextos, como, por exemplo, no setor bancário. Observou-se melhorias em todos os aspectos analisados, especialmente no cálculo do CLV (*Customer Lifetime Value*).

No Capítulo 4, as conclusões e trabalhos futuros são discutidos com base nos dois artigos produzidos.

2 Customer Lifetime Value Prediction: A Machine Learning Approach[1]

Abstract: The *Customer Lifetime Value* (CLV) prediction is of paramount importance in various business models. However, this calculation varies according to the context and scope of the business. This work aims to predict the CLV using machine learning algorithms and compare it with existing models used in the literature in 3 different datasets. The choices and decisions taken for constructing the models are described, and it is shown that the machine learning models presented better results in the calculation of the expected number of transactions, and obtained very similar results in the model that calculates the average value per transaction.
Resumo: A previsão do *Customer Lifetime Value* (CLV) é de suma importância em diversos modelos de negócios. No entanto, esse cálculo varia de acordo com o contexto e o escopo do negócio. Este trabalho tem como objetivo realizar a previsão do CLV utilizando algoritmos de aprendizado de máquina e compará-lo com os principais modelos utilizados na literatura em 3 bases de dados diferentes. Durante o trabalho são descritas as escolhas e decisões tomadas para a construção dos modelos, sendo mostrado que os modelos de aprendizado de máquina apresentaram melhores resultados no cálculo do número esperado de transações, e obtiveram resultados muito semelhantes no modelo de cálculo do valor médio por transação.

2.1 Introdução

Nas últimas décadas, foi percebido que é mais vantajoso tentar manter as relações com os clientes atuais, do que prospectar novos clientes. De acordo com Kotler e Keller (Sargeant, 2003), os custos associados à obtenção de novos clientes podem ser cinco vezes maiores do que os custos associados com a manutenção de um bom relacionamento com clientes. Isto está fazendo com que as empresas mudem suas abordagens centradas no produto, onde o foco é vender mais produtos, para abordagens centradas no cliente (Aeron et al., 2012). Nas abordagens centradas no cliente, foca-se em fornecer serviços que fidelizam tal cliente, com o intuito de manter uma boa relação e aumentar a retenção de clientes (Chan, 2008).

Para isso, entender qual é a previsão de valor que cada cliente irá gerar para a

[1]Artigo publicado no XX Encontro Nacional de Inteligência Artificial e Computacional (ENIAC); ISSN: 2763-9061; QUALIS B4; Citação: Ramos and Silva (2023)

empresa é fundamental. Baseando-se nisso, foi identificado que cada cliente possui um determinado valor financeiro atribuído a seu relacionamento futuro com a companhia. Este valor é definido como *Lifetime Value* (LTV) ou *Customer Lifetime Value* (CLV)[2] (Dwyer, 1997). Ele também pode ser associado com a "vida útil" do cliente, definida como quanto tempo é esperado para que aquele cliente continue ativo e gerando receita para a empresa (Khajvand et al., 2011).

Um cálculo efetivo do CLV pode trazer diversos benefícios tanto para as companhias quanto para os seus clientes. Para as companhias, trazem benefícios como a identificação de clientes mais valiosos, uma vez que os 20% dos clientes mais valiosos são responsáveis por 80% do lucro da companhia (Ekinci et al., 2014). Além disso, permite entender quais clientes estão em risco de *churn*[3], para assim tentar trazer o interesse de volta ou identificar o motivo da evasão. Já para os clientes, os serviços podem ser ofertados de forma personalizada pela companhia de acordo com suas preferências, além de um atendimento de acordo com as suas necessidades.

Contudo, o cálculo do CLV não é trivial. Em primeiro lugar, para cada tipo de companhia, podem ser necessários parâmetros e métricas específicos. Por exemplo, no contexto não-contratual, não há uma confirmação se aquele cliente ainda é um cliente ativo, enquanto nos meios contratuais, espera-se que o cliente ficará ativo até o fim do contrato (Singh and Jain, 2013). Em segundo lugar, o cálculo do CLV é poderoso para a análise de clientes, mas sua aplicação pode ser desafiadora (Fader et al., 2005b), visto que certas implementações são complexas e/ou computacionalmente custosas. Outra dificuldade encontrada consiste na volatilidade dos relacionamentos entre os clientes, sendo que clientes que apresentam não ser lucrativos, podem se tornar clientes lucrativos (Méndez-Suárez and Crespo-Tejero, 2021).

Existem diversos trabalhos na literatura que fazem da utilização de modelos probabilísticos para a previsão do CLV, como o trabalho de Mammadzada et al. (2021), que apresenta uma aplicação de modelos probabilísticos para um banco. Outro exemplo é o trabalho de Khajvand et al. (2011), que usa de duas soluções para estimar o CLV de uma companhia de beleza e saúde, utilizado o modelo RFM, que consiste em uma separação dos dados de acordo com a recência, frequência, e o valor monetário das compras para categorizar os clientes, e uma variação proposta pelo autor. Tem surgido nos últimos anos a utilização de técnicas de aprendizado de máquina para este cálculo, como no trabalho de Desirena et al. (2019), que utiliza redes neurais para aumentar o CLV na indústria de seguros por meio de recomendações mais assertivas. Já o trabalho de Sun et al. (2023) faz uso de

[2]Esses dois termos, LTV e CLV, serão usados para indicar o mesmo conceito neste texto.
[3]Churn é um indicador de quando um cliente encerra o seu relacionamento com a empresa.

aprendizado de máquina para medir o CLV e a segmentação com base no valor do ciclo de vida do cliente.

Apesar dos avanços, grande parte dos trabalhos fazem uso de aprendizado de máquina na etapa de segmentação dos clientes, e realizam o cálculo do CLV utilizando os modelos probabilísticos. O objetivo deste trabalho é usar o aprendizado de máquina para criar uma solução para calcular o CLV e comparar com os modelos probabilísticos, visto que, segundo o trabalho de Temor Qismat (2020), existem poucos estudos que fazem a comparação entre os métodos probabilísticos e os modelos de aprendizado de máquina.

O principal diferencial deste trabalho é que foi feita, em três bases de dados, a utilização de modelos de aprendizado de máquina para prever o valor do CLV em conjunto com modelos probabilísticos já consolidados na literatura. Foi verificada a eficácia dos modelos utilizando aprendizado de máquina, podendo complementar ou substituir qualquer um dos modelos probabilísticos no cálculo do CLV.

O restante deste texto está organizado contando com a apresentação de trabalhos relacionados na seção 2.2; os materiais e métodos utilizados durante a pesquisa na seção 2.3; a solução abordada se encontra na seção 2.4, e os resultados obtidos e métodos de avaliação na seção 2.5; por fim, as considerações finais na estão na seção 2.6.

2.2 Trabalhos Relacionados

Foi feita a divisão entre trabalhos que são relacionados ao cálculo do CLV e em abordagens de aprendizado de máquina.

2.2.1 Cálculo do *Customer Lifetime Value*

Ao prever o CLV, há uma necessidade de diferenciar de acordo com o tipo do relacionamento entre cliente e empresa, que pode ser contratual ou não-contratual (Reinartz and Kumar, 2000). Um relacionamento contratual implica que legalmente há uma relação entre o cliente e companhia, onde a companhia geralmente tem uma expectativa de quando um cliente irá se tornar inativo. No caso não-contratual, este relacionamento não é regido por um contrato ou adesão (Reinartz and Kumar, 2000). Além deste tipo de classificação, existe outra de acordo com o período de tempo em que ocorrem as compras, sendo que se as compras forem discretas, elas só podem ocorrer em um determinado momento, e caso sejam contínuas, podem ocorrer a qualquer momento (Temor Qismat, 2020).

Existe também a classificação em relação aos modelos, divididos entre modelos de comportamento do passado, e modelos de comportamento de passado-futuro.

Duas características diferenciam essas categorias: a primeira diferença baseia-se no pressuposto de que os clientes sujeitos a avaliações estarão ativos no futuro. A segunda diferença é se os custos dos clientes são incluídos nos modelos ou não. O primeiro grupo de modelos faz os cálculos incluindo a taxa de ativação futura de clientes e também os custos associados ao cliente, enquanto o último grupo não os leva em consideração. Os modelos de comportamento do cliente no futuro também podem ser separados em duas categorias com base no atributo de incluir ou não o custo de aquisição do cliente (Hiziroglu and Sengul, 2012).

Para ilustrar diversos contextos, existem soluções implementadas para o meio bancário, como o trabalho de Kahreh et al. (2014), onde se investiga o papel do CLV nos benefícios da segmentação. Existem também trabalhos voltados para o varejo, como o estudo de Khajvand et al. (2011), em que o objetivo é criar relacionamentos mais próximos e profundos com os clientes e maximizar o valor vitalício de um cliente para uma organização. Existem trabalhos para campanhas de doações, como o trabalho de Sargeant (2003), que analisa a contribuição que o CLV pode fazer para a arrecadação de fundos.

Por fim, certos métodos do cálculo de CLV acabam se destacando dos demais. Temos, por exemplo, o Modelo de Pareto/NBD (Schmittlein et al., 1987) e o método RFM (Hughes, 1996) e suas variações, como discutido a seguir.

O modelo de Pareto/NBD (Schmittlein et al., 1987) é feito para analisar uma base de clientes em cenários não-contratuais. Isto é feito através de derivações de uma série de expressões e, entre outras coisas, (i) a probabilidade de que um cliente com um determinado histórico de transações ainda esteja ativo, e (ii) o número esperado de transações futuras para um cliente, condicionado ao seu histórico de transações. É extremamente complexo e computacionalmente intenso. Este modelo assume que o cliente está ativo por um período observado de tempo até ficar inativo. Enquanto está ativo, o cliente é retratado por uma função de distribuição *Gamma*, conhecida como distribuição de Pareto. Nela, usa-se a recência, a frequência e a duração do período de observação para prever as compras futuras de um cliente.

Baseando-se nos resultados de Schmittlein et al. (1987), surgiu o modelo Gamma-Gamma (Fader et al., 2005a), expandindo o modelo de Pareto para permitir que seja possível estimar o valor gasto por transação no futuro. Ele assume que o valor médio segue uma distribuição *Gamma* com parâmetro de forma $px + q$ e parâmetro de escala $v + x\bar{z}$, sendo que $\bar{z}$ é a média observada do valor das transações e x o número de observações.

Surge também uma alternativa ao modelo de *Pareto*, que consiste no *BG/NBD* (Fader et al., 2005b), que tenta descrever a taxa com que clientes fazem compras e a taxa que eles deixam de consumir. Este modelo é mais simples em termos computacionais que o modelo de Pareto, e executa de maneira mais rápida, e

é apresentado que os resultados são bem semelhantes na literatura, sendo uma boa alternativa.

Por outro lado, os modelos RFM (Hughes, 1996) são aplicados em diversas áreas de Marketing, sendo utilizados para analisar o comportamento de um cliente, e fazer predições baseado nisto. São muito adotados em técnicas de segmentação de clientes, e baseia-se na combinação de três métricas:

- Recência: o quão recente foi feita a ultima transação.
- Frequência: o quão frequentes são feitas as transações.
- Valor monetário: o quanto, em valor monetário, a transação corresponde.

2.2.2 Abordagens de Aprendizado de Máquina

Em relação à utilização de abordagens de aprendizado de máquina na previsão de CLV, há uma carência de trabalhos na área. Porém, tem-se trabalhos como o de Sun et al. (2023), cujo objetivo é fazer a segmentação de clientes com base em algoritmos de aprendizado de máquina e modelos de análise de gerenciamento de relacionamento com o cliente e a criação de um modelo de identificação de segmentação de valor do cliente em uma relação não-contratual.

O trabalho de Vanderveld et al. (2016) desenvolve um sistema que prevê o CLV futuro, utilizando uma implementação do *Random Forest* e o engajamento do cliente pelo e-mail e aplicativos móveis, desenvolvendo um sistema que atualiza o CLV diariamente baseando nas interações. Este trabalho também desenvolve modelos para diferentes tipos de clientes, permitindo pesos diferentes para atributos relevantes para cada tipo. Este tipo de modelo, apesar de eficaz, é ideal apenas quando as interações com o cliente são comuns, e que estas interações sejam capturadas em dados abundantes, o que não é o contexto de vários negócios.

Alternativas mais modernas, como o trabalho de Temor Qismat (2020), utiliza modelos de aprendizado de máquina, e compara os resultando entre a mistura do Pareto/NBD em conjunto com o modelo Gamma-Gamma para categorizar seus clientes em 8 classes, de acordo com o CLV esperado usando a segmentação do RFM. Foi observado que modelos de aprendizado apresentaram resultados superiores. Porém, este trabalho cita que o modelo Pareto/NBD teria um desempenho muito melhor se comparado com um modelo de aprendizado de máquina que prevê o CLV diretamente, em vez dos segmentos.

Outra solução utilizando técnicas mais modernas, consiste no trabalho de Desirena et al. (2019), que faz uso de redes neurais para fazer recomendações de produtos para aumentar os lucros em uma indústria de seguros, um modelo contratual. Porém,

este trabalho não calcula o CLV futuro, mas apenas o CLV atual do cliente, e procura alternativas de produtos que possam aumentar o CLV atual.

Neste trabalho, é proposto um modelo de previsão do CLV do cliente baseado no passado-futuro, sem incluir os custos de aquisição. O objetivo deste trabalho é realizar uma comparação justa, e mostrar que os métodos de aprendizado de máquina possuem desempenho similar ou melhor do que os modelos probabilísticos, e que se combinados, podem produzir resultados relevantes. Diferente dos trabalhos anteriormente apresentados, busca-se criar um modelo que prediz o valor do CLV, ao invés da categoria do cliente, além de permitir que este modelo seja aplicável a contextos gerais e com dados esparsos.

2.3 Materiais e Métodos

Nesta seção são apresentados os dados utilizados neste trabalho, além da definição matemática do que é considerado o CLV, e os procedimentos para o seu cálculo.

2.3.1 Os Dados

Para este estudo, foram utilizados três conjuntos de dados. O objetivo de testar com diversas bases de dados consiste em validar os modelos em diferentes situações com diferentes comportamentos de compras.

O primeiro conjunto de dados, chamado CDNOW, contém o histórico de compras até o final de junho de 1998 de 23.570 indivíduos que fizeram sua primeira compra na CDNOW, uma empresa que operava um site de compras on-line que vendia CDs e produtos relacionados à música. Este conjunto foi usado no artigo original do modelo Gamma-Gamma (Fader et al., 2005a).

No segundo conjunto, chamado OpenCDP, foram utilizados dados de interações dos clientes por 5 meses (outubro de 2019 a fevereiro de 2020) de uma loja online de cosméticos de médio porte, sendo esta disponibilizada na plataforma Kaggle[4] pelo projeto Open CDP. Cada linha nos dados representa um entre quatro eventos possíveis de interação entre o cliente e a loja. Foram utilizados somente as interações de compra, contendo 110 mil usuários únicos, realizando 159 mil compras.

O terceiro conjunto de dados, chamado Olist, contém informações de 100 mil compras de 2016 a 2018 feitos em vários locais no Brasil na plataforma Olist[5], por 65 mil clientes, também disponibilizado na plataforma Kaggle[6].

[4]https://www.kaggle.com/datasets/mkechinov/ecommerce-events-history-in-cosmetics-shop
[5]https://olist.com/pt-br/
[6]https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

2.3.2 Cálculo do *Customer Lifetime Value*

A definição e cálculo do *Customer Lifetime Value* (CLV) varia entre os autores, como apresentado por Dwyer (1997), em que o CLV representa o valor presente dos benefícios esperados subtraídos pelos encargos dos clientes. Já no trabalho de Aeron et al. (2012), tem-se que um cliente lucrativo é definido como uma pessoa, empresa ou residência cujo a receita excede os gastos da empresa em atrai-lo, vender e servir aquele cliente ao longo do tempo. O excesso pode ser considerado como o CLV. Já em trabalhos como o de Kahreh et al. (2014), o CLV é visto como uma métrica para medir o real valor de um determinado cliente de uma parte especial do mercado. Este conceito é antigo, sendo que uma das primeiras definições de CLV foi feita por Kotler (1974) como o "valor presente do fluxo de lucro futuro esperado em um determinado horizonte de tempo de transação com o cliente". Porém, um conceito que permeia entre os trabalhos consiste na definição apresentada por Dwyer (1997), em que CLV consiste nos lucros de um determinado cliente subtraídos os custos para atraí-lo.

O trabalho de Temor Qismat (2020) definiu que, dado que P_{it} representa o valor pago pelo cliente i em um tempo t, C_{it} representa o custo de servir o cliente i em um tempo t, AC_i o custo de aquisição do cliente i, r_{it} é a probabilidade do cliente estar ativo no tempo t, T é o horizonte de tempo estimado para o cálculo do CLV e por último, d sendo a taxa de desconto, temos a seguinte fórmula:

$$CLV_i = \sum_{t=1}^{\mathrm{T}} \frac{(P_{it} - C_{it}) \times r_{it}}{(1+d)^t} - ACi$$

Porém, em alguns cenários os custos envolvidos não estão disponíveis, e o CLV se resume à receita esperada vinda de cada cliente. Essa alternativa é adotada neste trabalho, e o CLV é calculado como a receita deste cliente ajustada a uma taxa de desconto ao longo do tempo. Com isto, tem-se a seguinte fórmula:

$$CLV_i = \sum_{t=1}^{\mathrm{T}} \frac{N_{it} \times V_{it}}{(1+d)^t} \tag{2.1}$$

Onde N_{it} consiste no número de transações esperadas que aquele determinado cliente i irá realizar no horizonte de tempo t, e que V_{it} corresponde ao valor esperado por transação daquele mesmo cliente. Além disso, para todos os dados, foi considerado que a taxa de desconto será de 6%, visando simular a inflação anual, e em relação ao horizonte estimado, foi escolhido 3 meses. Com isto, podemos dividir as etapas do cálculo de acordo com o diagrama na Figura 2.1, que serão detalhados a seguir.

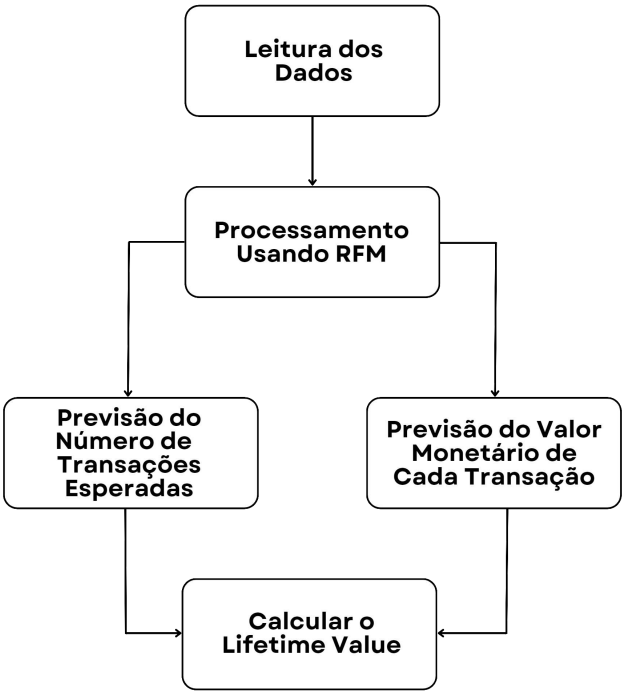


Figura 2.1: Diagrama das etapas para o cálculo do CLV

2.3.3 Processamento Usando RFM

Primeiramente, é preciso extrair métricas dos clientes usando uma das alternativas do método RFM. Neste trabalho, foi escolhido o RFMT (Ullah et al., 2023) (*Recency, Frequency, Monetary, Time*), visto que ele adiciona a quanto tempo o cliente existe na base de dados ao modelo RFM, com o objetivo de melhorar o entendimento sobre os comportamentos e a fidelidade dos clientes. Dado que um cliente i possui x transações no período observado, e que $z_1, z_2, ...z_x$ denote o valor de cada transação e que $d_1, d_2, ..d_x$ denote a data de cada transação, os cálculos são feitos conforme descrito abaixo.

A recência R consiste no tempo de atividade do cliente quando ele fez a última compra. Isso é igual ao intervalo entre a primeira compra de um cliente e sua última compra na unidade de tempo escolhida.

$$R_i = d_x - d_1$$

A frequência F representa o número de compras que o cliente fez, porém, subtraí-se um do número total de compras, para que assim, clientes que são considerados *One-time-buyers*[7], sejam facilmente detectados, portanto:

$$F_i = x - 1$$

[7]Um consumidor que compra apenas um produto ou serviço de uma empresa.

O valor monetário M representa o valor médio das compras de um determinado cliente. Isso é igual à soma de todas as compras de um cliente dividida pelo número total de compras.

$$M_i = \overline{z} = \sum_{i=1}^{x} z_i / x$$

T, oriundo de *Time*, representa quanto tempo o cliente existe até o final do tempo de observação d_y . Isso equivale à duração entre a primeira compra de um cliente e o final do período em estudo.

$$T_i = d_y - d_1$$

2.3.4 Estimativa do número de transações

O processo de estimar o número de transações esperadas consiste em, baseado nos valores retornados do Modelo RFMT, estimar o número de transações que um determinado cliente irá realizar no horizonte de tempo esperado. Para isto, foi proposto neste trabalho um modelo usando aprendizado de máquina para realizar esta estimativa, e para comparar os resultados da solução elaborada, assim como o trabalho de Temor Qismat (2020), foram escolhidos os modelos Pareto/NBD (Schmittlein et al., 1987) e *BG/NBD* (Fader et al., 2005b).

2.3.5 Estimativa do valor monetário

Para obter o lucro esperado por um determinado cliente, é necessário estimar o valor de cada transação que aquele determinado cliente irá realizar, baseando-se nos valores das transações passadas e os retornos do modelo RFMT.

Pensando nisto, foi proposta neste trabalho uma solução usando aprendizado de máquina, e esta foi comparada com um dos métodos mais utilizados na literatura, que consiste no modelo Gamma-Gamma (Fader et al., 2005a), que estima o valor médio monetário por transação de cada cliente.

2.4 Solução Proposta

Nesta seção é apresentada a solução proposta neste trabalho, além dos detalhamentos sobre como foi feita a separação dos dados, métricas utilizadas e escolhas feitas no processo de desenvolvimento do modelo.

2.4.1 Modelos de aprendizado de máquina

Os modelos baseados em aprendizado de máquina incorporam muitos parâmetros e atributos, opondo os modelos probabilísticos. Isto pode produzir uma melhor acurácia e melhores resultados, sendo essa a hipótese deste trabalho.

Para realizar a predição do número de transações e o valor monetário médio por transação, foram escolhidos diferentes algoritmos preditivos conhecidos na literatura, utilizando diversas metodologias para tentar encontrar determinados comportamentos entre os dados. Foram escolhidos os seguintes algoritmos: *LassoCV*, *Elastic Net*, *Kernel Ridge*, *Random Forest Regressor*, *Gradient Boosting Regressor*, *Histogram-based Gradient Boosting Regression Tree*, *LightGBM* e *Extreme Gradient Boosting Regressor*.

Em relação à otimização dos hiperparâmetros, foi utilizado o algoritmo do *Grid Search* para buscar os melhores hiperparâmetros de cada modelo. Para o algoritmo *LassoCV*, utiliza-se o número de $\alpha \in \{100, 200, 500, 100\}$ e o numero máximo de iterações $N \in \{1000, 1500, 2000\}$. Para o algoritmo *Elastic Net*, utiliza-se $\alpha \in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$, o número máximo de iterações $N \in \{1000, 1500, 2000\}$ e o parâmetro de mistura $l_1 \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. Para o algoritmo *Random Forest Regressor*, utiliza-se o número mínimo de folhas $L_{Min} \in \{1, 2, 4\}$, o número mínimo de divisões $D_{Min} \in \{2, 5, 10\}$ e o número de estimadores $N_{est} \in \{200, 800, 1000\}$. Já para o *Kernel Ridge*, foi utilizado $\alpha \in \{0.001, 0.01, 0.1, 1\}$. Para o *Gradient Boosting Regressor*, utiliza-se o número de estimadores $N_{est} \in \{500, 1000, 2000\}$, a taxa de aprendizado $L_r \in \{0.001, 0.01, 0, 1\}$, a profundidade máxima $P_{max} \in \{1, 2, 4\}$ e a fração de amostras $F_{amostra} \in \{0.5, 0.75, 1\}$. No *Histogram-based Gradient Boosting Regression Tree* utiliza-se a taxa de aprendizado $L_r \in \{0.001, 0.01, 0, 1\}$, a profundidade máxima $P_{max} \in \{-1, 1, 2, 4\}$, e o número máximo de folhas $L_{Max} \in \{-1, 1, 2, 4\}$. Tanto para o *LightGBM* quanto para o *Extreme Gradient Boosting Regressor* utiliza-se o número de estimadores $N_{est} \in \{100, 500, 1000\}$, a taxa de aprendizado $L_r \in \{0.01, 0.05, 0, 1\}$ e a profundidade máxima $P_{max} \in \{3, 6, 10\}$.

2.4.2 Separação dos dados

Para realizar o treinamento dos dados é necessário fazer uma separação em treino e teste. Baseando-se nisso, foi escolhido realizar esta separação de acordo com a data das transações, sendo que as 80% transações mais antigas serão utilizadas para o treinamento e validação dos modelos, e o restante para testes, tanto para os modelos de aprendizado de máquina quanto para os probabilísticos da literatura. Estes períodos podem ser chamados também de período de calibração (treino e validação) e observação (teste). Para criar os modelos, cada semana é considerada um período

de tempo, e as transações são agrupadas semana a semana.

Portanto, após a etapa do processamento usando RFM, cada linha nos dados representa um cliente para um período (i.e., semana) de tempo específico, sendo que cada uma destas representações possui seus atributos extraídos do RFMT levando em consideração os períodos decorrentes até o período especificado.

2.4.3 Métricas de avaliação

Para todos os modelos, foram utilizadas as mesmas métricas de validação, sendo que a principal consiste no Erro Quadrático Médio (*Mean Squared Error* ou MSE) - que consiste na diferença média quadrática entre os valores estimados e o valor real. Outra métrica utilizada consiste no Erro Médio Absoluto (*Mean Absolute Error* ou MAE). O erro absoluto também é utilizado, que consiste na diferença entre o valor estimado e o valor real.

2.4.4 Estimativa do número transações

Para estimar o número de transações, foram utilizados os atributos fornecidos pelo RFMT, além de qual o horizonte de tempo esperado para realizar a previsão. O alvo deste modelo consiste no número de transações que este cliente irá realizar no futuro. Para isto, utiliza-se 70% dos dados do período de calibração para treino, e os 30% restante para validação com o objetivo de definir qual dos algoritmos e respectivos hiperparâmetros obtêm um melhor desempenho. Com isto, obtêm-se o modelo que mais represente o comportamento dos dados, e os resultados estão na Tabela 2.1.

Modelo	**MSE na Base CDNOW**	**MSE na na Base Olist**	**MSE na na Base OpenCDP**
LassoCV	0,9227	0,0212	0,2431
ElasticNet	0,9229	0,0213	0,2428
RandomForestRegressor	0,6820	0,0206	0,2195
GradientBoostingRegressor	0,5912	**0,0201**	0,2034
HistGradientBoostingRegressor	0,6436	0,0213	0,2308
XGBRegressor	**0,5600**	0,0202	**0,1937**
LGBMRegressor	0,6207	0,0208	0,2240

Tabela 2.1: Comparação de Modelos em Relação ao Erro Quadrático Médio na Predição de Transações

Todos os modelos apresentaram bons resultados, porém, o que apresentou um melhor resultado em todas as bases foi o *XGBRegressor*. Por esta razão, ele foi escolhido para realizar a predição do número de transações para as 3 bases de dados.

2.4.5 Estimativa do valor monetário

Para o cálculo do valor monetário por transação, foram utilizados os mesmos atributos do modelo de transações, sendo que a diferença consiste no alvo. Neste caso, a variável alvo é a média dos valores de transação total do cliente, considerando os dados do futuro. Da mesma forma que na estimativa de transações, é feita a separação dos 70% dos dados de calibração para obter o modelo que melhor representa o comportamento dos dados, sendo testados diversos modelos, e os resultados encontrados estão na Tabela 2.2.

Modelo	**MSE na Base CDNOW**	**MSE na na Base Olist**	**MSE na na Base OpenCDP**
LassoCV	120,1387	465,0415	238,6791
ElasticNet	120,1009	464,9450	238,4658
RandomForestRegressor	102,1443	456,3697	221,2709
GradientBoostingRegressor	**99,3803**	457,5863	222,2082
HistGradientBoostingRegressor	142,3978	875,0773	323,0098
XGBRegressor	99,7793	**452,3832**	**221,0846**
LGBMRegressor	109,5694	460,6623	292,1478

Tabela 2.2: Comparação de Modelos em Relação ao Erro Quadrático Médio na Predição do Valor Monetário

O *XGBRegressor* obteve os melhores resultados para as bases Olist e OpenCDP, tendo tido o segundo melhor resultado para a base CDNOW. Como a diferença nesta última base foi muito pequena, o *XGBRegressor* foi selecionado para as três bases.

2.5 Avaliação

Nesta seção são apresentados os resultados obtidos neste trabalho, nas situações de cálculo de avaliação do número de transações e valor monetário esperado por cada transação, além do cálculo do CLV.

2.5.1 Baselines

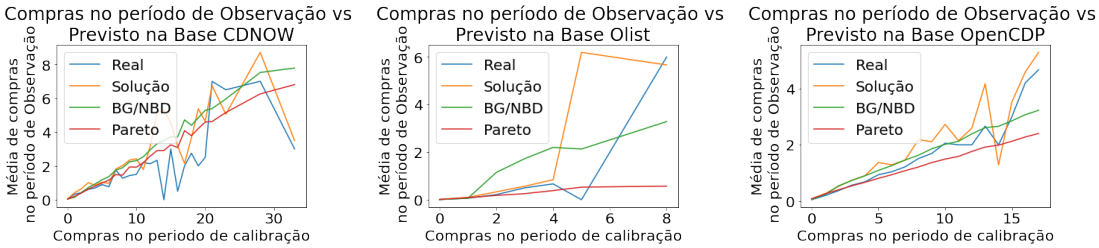
Para comparar os resultados, as soluções base de Pareto e *BG/NBD* foram usadas para estimar o número de transações, e a solução *Gamma-Gamma* para estimar o valor monetário. Na Tabela 2.3 estão descritos os parâmetros das distribuições probabilísticas obtidas após o treinamento para cada uma das bases de dados. Nos modelos de Pareto e *BG/NBD*, necessitam-se de três variáveis como entrada, a recência, a frequência e o T. Já no modelo *Gamma-Gamma*, é necessário o valor monetário e a frequência de compras do usuário como entrada.

Modelo	Pareto				BG/NBD				Gamma-Gamma		
	Alpha	Beta	R	S	A	Alpha	B	R	P	Q	S
Base CDNOW	16,60	9,80	0,62	0,43	0,10	6,43	0,25	0,27	4,07	0,93	3,79
Base Olist	150,75	5,67	0,23	0,54	0,01	5,55	0,00	0,01	3,92	0,82	3,62
Base OpenCDP	11,49	48,95	0,38	0,00	0,01	4,57	0,02	0,19	3,92	0,82	3,62

Tabela 2.3: Parâmetros dos Modelos

2.5.2 Resultados

Pode-se observar na Figura 2.2 o número de transações que foram observadas no período de treino (calibração) no eixo X, e a média de transações previstas no eixo Y. É possível observar que, o modelo de aprendizado de máquina (i.e., Solução) teve um desempenho melhor, no sentido de que ele identifica o comportamento dos clientes de maneira mais semelhante aos valores reais. O mesmo é válido para o Erro Quadrático Médio (MSE) como mostra a Tabela 2.4.



(a) Base CDNOW (b) Base Olist (c) Base OpenCDP

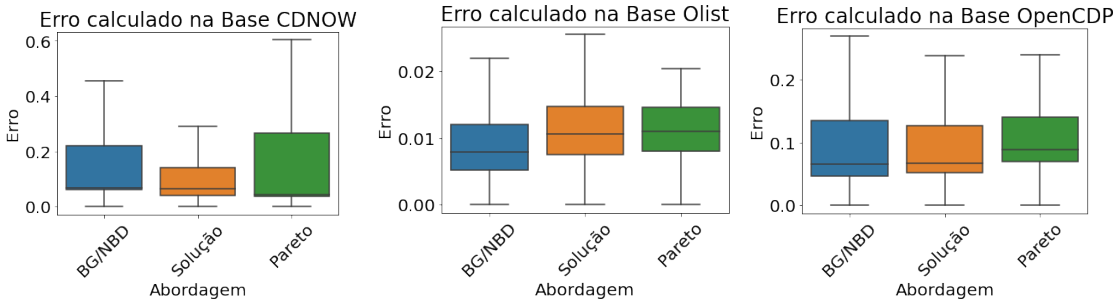
Figura 2.2: Transações no período de treino por compras previstas

Modelo	MSE na Base CDNOW	MSE na na Base Olist	MSE na na Base OpenCDP
Pareto	0,3918	0,0108	0,105
BG/NBD	0,4172	0,0113	0,106
Solução Proposta	**0,3730**	**0,0105**	**0,095**

Tabela 2.4: Comparação de Modelos de Predição do Número de Transações

Em relação ao erro absoluto (MAE), pode-se observar na Figura 2.3 que todas as abordagens conseguiram um bom resultado, mas ainda assim a solução proposta, utilizando aprendizado de máquina, se sobressaiu em relação às outras em duas das três bases.

Já em relação ao valor monetário médio por transação, pode-se observar na Tabela 2.5, que o modelo proposto de aprendizado de máquina obteve resultados significativamente mais próximos do correto, com erros médios absolutos bem menores. Isto pode ser observado mais precisamente na Figura 2.4, onde o erro



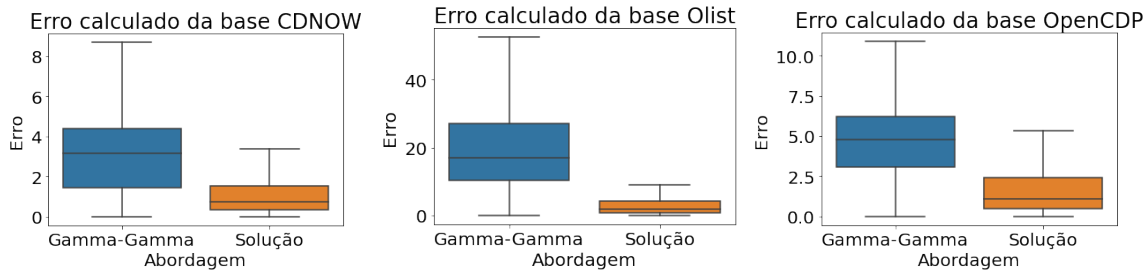
(a) Base CDNOW (b) Base Olist (c) Base OpenCDP

Figura 2.3: Erro absoluto entre os modelos de predição de transações

absoluto é mostrado com o modelo de aprendizado de máquina com resultados melhores.

Modelo	**MAE na Base CDNOW**	**MAE na Base Olist**	**MAE na Base OpenCDP**
Gamma Gamma	4,10299	26,88304	6,27010
Solução Proposta	**1,4830**	**6,6965**	**2,7155**

Tabela 2.5: Comparação de Modelos de Predição de valor médio por transação



(a) Base CDNOW (b) Base Olist (c) Base OpenCDP

Figura 2.4: Erro absoluto entre os modelos de predição de valor monetário médio por transação

2.5.3 Cálculo do CLV

Para se obter o real valor do CLV do cliente, foi utilizado o número de transações que ocorreram no período de testes, e multiplicado pelo valor médio por transação calculado utilizando todas as transações do cliente, descontando a inflação, conforme equação 2.1. Foram testadas todas as seis combinações com modelos probabilísticos e de aprendizado de máquina.

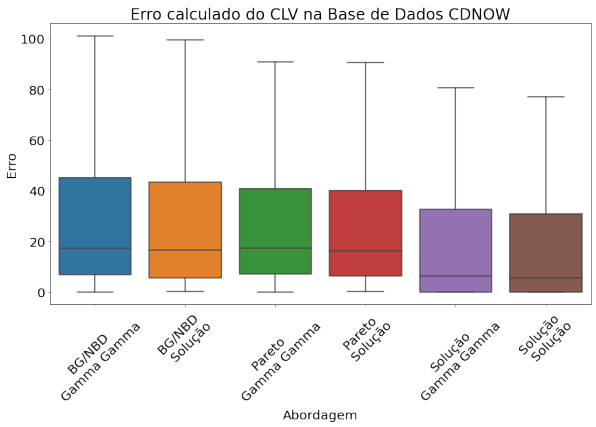
Na Tabela 2.6 pode-se observar os resultados da utilização de cada um dos modelos de predição de transações em combinação com os de predição monetária, e

os erros obtidos. Pode-se observar que, quando as duas soluções propostas utilizando aprendizado de máquina são usadas, todos os resultados foram melhores.

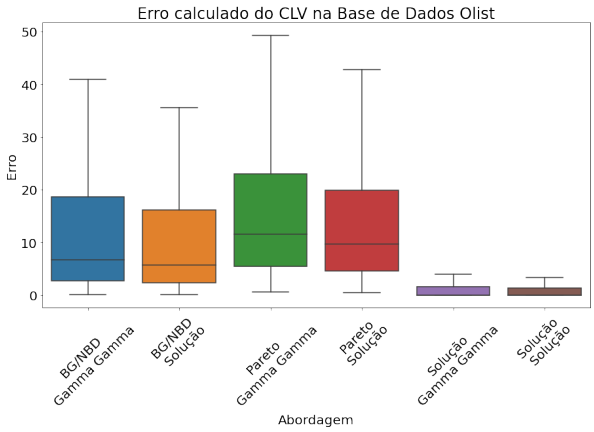
Modelo de Predição de Transações	Modelo de Predição Monetária	Base CDNOW		Base Olist		Base OpenCDP	
		MSE	MAE	MSE	MAE	MSE	MAE
BG/NBD	Gamma-Gamma	5025,7020	38,1459	15552,7530	34,0561	46915,0220	131,7316
Pareto/NBD	Gamma-Gamma	3571,3690	33,8526	3873,5680	24,5138	3025,2570	114,4087
Solução	Gamma-Gamma	2454,3811	24,3039	3313,0941	9,4711	1274,3249	18,5551
BG/NBD	Solução	4828,3215	36,6086	13705,5407	31,1360	40317,2660	120,5055
Pareto/NBD	Solução	3460,7106	32,6195	3611,1815	22,1368	25454,5611	103,7789
Solução	Solução	**2360,0799**	**23,7094**	**3256,7795**	**9,1423**	**1245,7343**	**18,0845**

Tabela 2.6: Comparação de Modelos

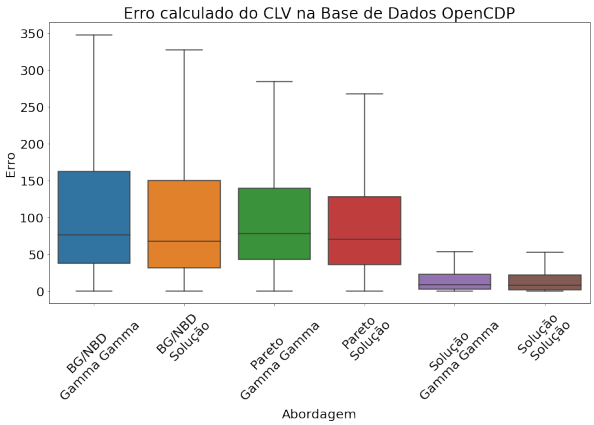
Pode-se observar que o erro calculado na junção da utilização de cada um dos modelos na Figura 2.5. Vemos que a utilização dos modelos de aprendizado de máquina proposto neste trabalho no cálculo do número de transações esperadas alcançaram um resultado significativamente melhor. Já no cálculo do valor monetário, todos os modelos apresentam resultados bem semelhantes, porém, o erro é um pouco maior quando utilizado o Gamma-Gamma.



(a) Erro absoluto do CLV na base da CDNow



(b) Erro absoluto do CLV na base da Olist



(c) Erro absoluto na do CLV base do OpenCDP

Figura 2.5: Erro absoluto entre o cálculo do CLV

2.6 Considerações Finais

Neste trabalho, foi proposta uma solução para o cálculo do *Customer Lifetime Value* (CLV) usando aprendizado de máquina. Foi feita uma comparação da proposta com modelos consolidados na literatura. Foi possível observar que em relação ao número de transações previstas em uma situação real, os modelos de aprendizado de máquina propostos se destacam em relação aos modelos da literatura. Em relação aos modelos monetários, foi visto que ambos os modelos apresentam resultados bem competitivos, apesar dos modelos de aprendizado de máquina se sobressaíram. Já em relação ao cálculo do CLV, foi apresentado a maneira como este trabalho considera este cálculo, levando em conta o número de transações esperados, e o valor esperado por transação. Foi feito o cálculo e a solução proposta apresentou resultados mais precisos em bases reais.

As principais limitações consistem nos dados encontrados, visto que para análises mais profundas são necessário uma quantidade maior de dados, além de possuir mais atributos. Em relação a trabalhos futuros, sugere-se melhorar os modelos de aprendizado de máquina, além de testar com outras bases de dados e contextos e além de utilizar modelos diferentes. Neste trabalho foram utilizados somente os atributos fornecidos pelo modelo RFMT, por motivos de realizar uma comparação justa com os modelos probabilísticos. Porém, é importante adicionar mais atributos com o objetivo de tornar o cálculo do CLV mais preciso. Também sugere-se adicionar mais etapas, como a predição de *Churn* e/ou as interações do cliente na etapa do cálculo do CLV, visando obter uma métrica mais realista e mais próxima da realidade.

3 Um solução para previsão do *Customer Lifetime Value* para diferentes segmentos de mercado[1]

Abstract: The estimation of Customer Lifetime Value (CLV) has gained importance in understanding the relationship between companies and their customers. However, its implementation often encounters difficulties, ranging from solutions being highly specific to the context for which they are developed to the attributes used violating customer privacy. This work aims to build a generic solution for CLV prediction, using only attributes derived from the date and value of customer transactions with the company. The developed model was evaluated on five different datasets, in addition to being compared with five reference solutions from the literature. Once validated, the proposed model was applied in a real-world scenario involving a large pharmacy chain, demonstrating its applicability and importance to the market. The results showed that it was possible to achieve an improvement of up to 17.93% in CLV calculation error in the validation scenarios, and 2.9% in the real-world scenario.
Resumo: A estimativa do *Customer Lifetime Value* (CLV) tem ganhado relevância na compreensão do vínculo entre empresas e seus clientes. No entanto, sua implementação frequentemente se depara com dificuldades, desde o fato das soluções serem altamente específicas ao contexto para o qual são desenvolvidas, como os atributos utilizados violarem a privacidade dos clientes. Este trabalho tem como objetivo a construção de uma solução genérica para previsão do CLV, utilizando apenas atributos extraídos da data e do valor das transações do cliente com a empresa. O modelo desenvolvido foi avaliado em cinco conjuntos de dados distintos, além de ser comparado com outras cinco soluções de referência da literatura. Depois de validado, o modelo proposto foi aplicado em um cenário real de uma grande drogaria, demonstrando a sua aplicabilidade e importância para o mercado. Os resultados mostraram que foi possível obter uma melhora de até 17,93% no erro do cálculo do CLV nos cenários de validação, e de 2,9% no cenário real.

[1]Artigo submetido para publicação no 39º Simpósio Brasileiro de Banco de Dados (SBBD); ISSN: 2763-8979; QUALIS A4

3.1 Introdução

Nos últimos anos, houve uma significativa transição no mercado, passando de uma abordagem centrada no produto, que foca na melhoria dos processos de produção para aumentar os lucros, para uma abordagem centrada no cliente (ABIDAR et al., 2023). Essa mudança visa aprimorar o relacionamento entre a empresa e o cliente, com o objetivo de aumentar a lucratividade. Destaca-se, portanto, a importância de se conhecer os clientes, aliado a um aumento nos investimentos em estratégias de marketing e retenção.

Entretanto, uma das dificuldades enfrentadas pelas empresas é compreender em que estágio se encontra o relacionamento com cada cliente e estimar o valor que ele pode gerar para a empresa. Isso se deve à diversidade de hábitos e comportamentos dos clientes. Para abordar essa questão, surge o conceito de Valor Vitalício do Cliente, ou *Customer Lifetime Value* (CLV)(Kumar et al., 2023). O CLV é uma métrica fundamental que permite às empresas avaliarem o valor que cada cliente agrega ao longo do tempo, levando em consideração não apenas as compras imediatas, mas também a possibilidade de futuras interações e a fidelidade do cliente. Ao entender e utilizar o CLV, as empresas podem direcionar estratégias mais eficazes para maximizar o valor do relacionamento com o cliente a longo prazo.

Porém, estimar o CLV não é trivial visto que existem diversos contextos de negócios e tipos de relacionamentos. Um exemplo consiste no padrão de compra do produto ofertado, visto que determinadas empresas oferecem serviços que são pagos em intervalos especificados de tempo (i.e., discretos), enquanto em outras os clientes compram livremente (i.e., contínuos). Outra dificuldade envolve saber se o cliente ainda é ativo na empresa, quando o relacionamento não exige um contrato explícito.

Com o intuito de facilitar as estimativas de CLV, surgiram abordagens probabilísticas como o modelo de *Pareto/Gamma-Gamma* (Schmittlein et al., 1987) ou *Beta-Geometric/Negative Binomial Distribution* (BG/NBD) (Fader et al., 2005b), que assumem diversas premissas. Visando adicionar mais informações contextuais, surgiram abordagens de aprendizado de máquina, que permitem a utilização de atributos específicos do comportamento de cada cliente para enriquecer os resultados. Porém, essas abordagens geralmente demandam atributos específicos de cada contexto e muitas vezes sensíveis (e.g., dados pessoais e de movimentações financeiras), dificultando assim a sua generalização e aplicação em outros contextos.

O objetivo deste trabalho é criar uma solução genérica para estimar o CLV, utilizando apenas atributos extraídos de transações básicas que são disponibilizadas em diversos contextos sem invadir a privacidade dos clientes. A solução proposta foi comparada com cinco outras do estado da arte, em cinco bases de dados públicas de diferentes contextos, com resultados que chegam a até 17,93% de melhoria.

Este trabalho foi desenvolvido como uma demanda do mercado, por meio de uma empresa parceira, e se encaixa na trilha industrial e de aplicações do SBBD. A solução proposta foi aplicada em um cenário real de uma drogaria, com dados recentes, e os resultados foram melhorados em aproximadamente 2,9% em relação às soluções da literatura.

O restante deste texto está organizado com a apresentação de trabalhos relacionados na Seção 3.2; as especificações da solução na Seção 3.3; os métodos de avaliação e resultados obtidos na Seção 3.4; os resultados obtidos no teste com a base de dados real na Seção 3.5; por fim, as considerações finais estão na Seção 3.6.

3.2 Trabalhos Relacionados

No que diz respeito ao cálculo do CLV, Reinartz and Kumar (2000) afirmam que o mesmo irá depender do relacionamento do cliente com a empresa. Se as transações são efetuadas em momentos específicos, é considerado um contexto **discreto**, e quando as compras podem ocorrer a qualquer momento, é considerado um contexto **contínuo**. Este tipo de separação fornece detalhes importantes em relação à frequência das compras, facilitando ou dificultando as predições. Ainda nos relacionamentos, no contexto **contratual** o cliente tem alguma forma de "contrato" com a empresa, informando quando desejar encerrar o relacionamento; já no caso **não-contratual** essa obrigação não existe. Vale ressaltar que neste trabalho, o foco está em relacionamentos contínuos, em que as transações podem ocorrer a qualquer momento, sendo contratual (e.g., bancos) ou não-contratual (e.g., varejo).

Os trabalhos encontrados na literatura podem ser organizados de acordo com as técnicas e dados utilizados, como discutido a seguir e apresentado na Tabela 3.1.

3.2.1 Abordagens clássicas

O trabalho de POPA et al. (2021) destaca a importância da previsão do *Customer Lifetime Value* (CLV) como parte integrante da estratégia de marketing contemporânea. Ele identifica os principais estudos na área e os temas abordados. Nota-se que muitos desses estudos empregam abordagens clássicas, entre as quais se destacam:

- *Pareto* (Schmittlein et al., 1987): Abordagem bem reconhecida que utiliza os atributos do modelo RFM (Recência, Frequência, Monetário) para estimar o número esperado de transações de um cliente em um determinado período de tempo.
- Gamma-Gamma (Fader et al., 2005a): Extensão do modelo de Pareto, permitindo a estimativa do valor médio monetário das futuras transações.

Trabalho	Abordagem	Categoria	Contexto dos Dados	Dados Necessários
Bauer and Jannach (2021)	Redes Neurais, Stacking	Não-Contratual, Contínuo	Comércio Eletrônico	Demográficos, transações, clientes, produto
Qi et al. (2015)	Modelo Conceitual	Contratual, Discreta	Telecomunicação.	Satisfação e lealdade do cliente, demográficos
Calabourdin and Aksenov (2023)	Cadeias de Markov	Não-Contratual, Contínuo	Comércio Eletrônico	Cliente, produtos e o contexto de aquisição do cliente
Kailash et al. (2023)	Aprendizado de Máquina	Contratual, Discreta	Seguradora	Clientes, demográficos, educacionais, renda, posses do cliente e dados do relacionamento
ABIDAR et al. (2023)	Aprendizado de Máquina	Não-Contratual, Contínuo	Comércio Eletrônico	Demográficos, transações, clientes, produtos
Kumar et al. (2023)	Séries Temporais	Não-Contratual, Contínuo	Comercio Eletrônico	Transações e comportamentos dos clientes
Comlan and Adiba (2024)	Cadeias de Markov	Contratual, Contínuo	Streaming	Clientes, compra de pacotes e atividades de visualização
Fader et al. (2005b)	Modelo Probabilístico	Não-Contratual, Contínuo	Comércio Eletrônico	Clientes e transações
Haenlein et al. (2007)	Cadeias de Markov	Contratual, Contínuo	Setor Bancário	Clientes, produtos, e atividade do usuário
Ekinci et al. (2014)	Redes Neurais	Contratual, Contínuo	Setor Bancário	Clientes, transações e produtos
Hiziroglu and Sengul (2012)	Modelo RFM	Não-Contratual, Contínuo	Comércio Eletrônico	Clientes, transações e demográficos
Khajvand et al. (2011)	Modelo RFM	Não-Contratual, Contínuo	Comércio Varejista	Transações
Sun et al. (2023)	Aprendizado de Máquina	Não-Contratual, Contínuo	Comércio Eletrônico	Demográficos, transações, clientes, produtos
Vanderveld et al. (2016)	Aprendizado de Máquina	Não-Contratual, Contínuo	Comércio Eletrônico	Demográficos, transações, produtos, relacionamentos
Fader et al. (2005a)	Modelo Probabilístico	Não-Contratual, Contínuo	Comércio Eletrônico	Transações

Tabela 3.1: Características dos Trabalhos Relacionados

- BG/NBD (Fader et al., 2005b): Esta é uma alternativa ao modelo de Pareto, apresentando resultados semelhantes, mas com maior eficiência computacional.

Além dessas, surgem outras alternativas para prever o CLV, como os modelos baseados em cadeias de Markov, que descrevem uma sequência de eventos possíveis, onde a probabilidade de um evento ocorrer está relacionada ao estado obtido no evento anterior. Existem estudos que exploram esse conceito para modelar o relacionamento com os clientes e estimar o CLV, como Calabourdin and Aksenov (2023) e Comlan and Adiba (2024). Outra abordagem clássica consiste em utilizar os modelos RFM e comparar sua eficácia com outros modelos, como demonstrado no trabalho de Hiziroglu and Sengul (2012). Além disso, há alternativas para os modelos baseados no RFM que incorporam outros atributos para estimar o CLV. Entre elas, destacam-se Cheng and Chen (2009), que utiliza a probabilidade de *churn*, Khajvand et al. (2011) que tenta incorporar o número de itens consumidos, e Ullah et al. (2023) que, além do *churn*, leva em consideração o tempo de permanência do cliente na base de dados.

3.2.2 Abordagens com aprendizado de máquina

A utilização de algoritmos de aprendizado de máquina para a previsão do CLV é recente. Apesar disto, as soluções apresentam bons resultados quando comparadas aos modelos clássicos. Este tipo de abordagem permite a utilização de informações sobre o cliente, dados demográficos ou até mesmo dados sobre as interações do cliente com a empresa, com o intuito de melhorar os resultados. Porém, à medida que tais dados podem trazer melhorias, eles aumentam a complexidade dos modelos e afetam a privacidade dos clientes, junto com a dificuldade de generalização. Em suma, existem duas abordagens para o uso de aprendizado de máquina para CLV: classificação e regressão.

Classificação

Esta abordagem consiste em segmentar os clientes, que devem ser tratados de maneira diferente. Alguns trabalhos lidam com o CLV como um problema de classificação, objetivando identificar os clientes mais valiosos. O trabalho de ABIDAR et al. (2023) utiliza dados de um site varejista com o objetivo de separar clientes em três segmentos: baixo, médio e alto valor. O modelo proposto consegue separar em média 90% dos clientes nas categorias corretas, com uma revocação de 0,7. Outro exemplo consiste no trabalho de Haenlein et al. (2007), que realiza a classificação em um contexto bancário, com o intuito de classificar clientes de acordo com os serviços utilizados e interesses. Já o trabalho de Sun et al. (2023) usa dados de uma loja online, e concentra-se em medir o CLV atual e usá-lo na segmentação do cliente.

Regressão

A abordagem por regressão tem como objetivo estimar um valor numérico que representa o CLV do cliente. O Trabalho de Kailash et al. (2023) utiliza o conjunto de dados IBM Watson para avaliar e comparar o desempenho de diversos regressores, em um contexto contratual e discreto de uma indústria de seguros. O trabalho de Vanderveld et al. (2016) utiliza dados de uma plataforma de comércio online, contendo informações do usuário, dados demográficos e interações do cliente para estimar o CLV. Trabalhos como de Kumar et al. (2023) utilizam o ARIMA em comparação com outros modelos de aprendizado de máquina. Já Bauer and Jannach (2021) propõem um *framework* de previsão de CLV e demonstra sua eficácia em dois cenários de aplicação no domínio do comércio eletrônico, combinando o desempenho de duas abordagens, uma baseada em redes neurais sequenciais e outra baseada em um modelo de regressão. A utilização da combinação de ambas as abordagens traz ganhos significativos, visto que adiciona o comportamento temporal

do cliente, levando em conta a sazonalidade nos modelos de regressão.

As soluções descritas acima demonstram as vantagens de utilizar as abordagens de aprendizado de máquina em relação às clássicas, visto que permitem a utilização de diversos atributos, contendo informações das transações, dos produtos, dos clientes e até dados demográficos com o intuito de melhorar a predição. Porém, isso dificulta a replicação da solução em contextos em que esses dados não estejam disponíveis, além de impactarem na privacidade dos clientes. Nestes trabalhos, os modelos são feitos para solucionar o problema contextualizado, e não objetiva-se estender para um propósito mais geral. Com o objetivo de preencher essa lacuna, neste presente trabalho foi desenvolvida uma solução genérica baseada somente em transações básicas.

3.3 Solução genérica baseada em transações

O objetivo da proposta deste artigo é propor uma solução para previsão do CLV utilizando apenas atributos extraídos das transações, sem a necessidade de informações sensíveis ou detalhadas dos clientes, produtos e suas interações com a empresa. Pensando nisto, seja $TX = \{tx_1, ..., tx_n\}$ o conjunto de transações de um cliente, em que $tx_i = \langle d_i, m_i \rangle$, sendo d_i a data quando foi feita a operação e m_i o valor monetário daquela transação.

Seja $I = 7$ dias o período de agrupamento das transações para o cálculo dos atributos, e d_{min} e d_{max}, a primeira e a última data de transação considerando todos os clientes. Seja p_i o período da transação tx_i, dado por $p_i = \lfloor \frac{d_i - d_{min}}{I} \rfloor + 1$ e tem-se então $P = \lceil \frac{d_{max} - d_{\min}}{I} \rceil$ períodos de tempo no total.

Seja $TX_p \subset TX$ o subconjunto de transações até o período $p, (1 \leq p \leq P)$:

$$TX_p = \{tx_i | tx_i \in TX, p_i \leq p\}$$

Além disso, seja tx_f a última transação do conjunto TX_p. Com base nisso, os atributos descritos na Tabela 3.2 foram propostos. Esses atributos extraem características do relacionamento do cliente com a empresa, sem violar a privacidade do mesmo.

Com esses atributos, define-se uma matriz com uma linha para cada cliente e os atributos propostos calculados para o cliente no período de tempo p analisado:

- $NTx_u(p+h)$ = Número de transações esperadas do cliente u em h períodos de tempo no futuro.
- $M_u(p+h)$ = Valor monetário médio esperado por transação do cliente u em h períodos de tempo no futuro.

Atributo	Fórmula	Descrição
$N(p)$	$N(p) = \lvert\{p_i \lvert tx_i \in TX_p\}\rvert$	Número de períodos distintos que ocorreram uma transação, até o período p.
$SM(p)$	SM(p) =$\sum_{tx_i \in TX_p} m_i$	Valor monetário acumulado até o período p
$R(p)$	$R(p) = \frac{d_f - d_1}{I}$	Tempo de atividade do cliente com a empresa quando realizou a transação mais recente tx_f, em relação ao período p.
$F(p)$	$F(p) = N(p) - 1$	Quantidade de períodos distintos que o cliente realizou uma transação até o período p sem incluir a primeira compra.
$M(p)$	$M(p) = \begin{cases} \frac{1}{F(p)}(SM(p) - m_1), & \text{se } F(p) = 0 \\ 0, & \text{caso contrário} \end{cases}$	Valor monetário médio por período, até o período p, além de subtrair o valor da primeira transação.
$T(p)$	$T(p) = p - p_1$	Quantidade de períodos de tempo o cliente está ativo desde a sua primeira compra até o período p
$SL(p)$	$SL(p) = p - p_f$	Número de períodos de ociosidade do cliente, desde a última transação tx_f até o período p.

Tabela 3.2: Descrição das variáveis e fórmulas

Neste trabalho, adota-se h = 4 semanas, visando realizar uma previsão de curto prazo. Vale destacar que, apesar do CLV considerar o valor vitalício (i.e., a longo prazo), a ausência de dados impede que uma análise com períodos longos seja feita, como observado em todos os trabalhos da literatura. Além disso, uma estimativa de curto prazo sendo realizada periodicamente traz benefícios significativos para as empresas.

Com os valores informados, foram treinados modelos com diferentes algoritmos (i.e., *Lasso, ElasticNet, Random Forest Regressor, Kernel Ridge, Gradient Boost Regressor, XGBoost Regressor e Ligth GBM*). O ajuste dos hiper-parâmetros foi feito com a técnica de *Exhaustive Grid Search*, com os dados do período de treino objetivando estimar os dados do período de validação. A métrica RMSE foi utilizada para a escolha do melhor modelo para a coleta dos resultados de erros da próxima seção.

3.4 Avaliação e Resultados

Com o intuito de avaliar o modelo desenvolvido, foram escolhidos trabalhos do estado da arte com diferentes características:

- *Sequence-Based* [Bauer and Jannach (2021)]: Uma abordagem utilizando redes neurais recorrentes. Na proposta original, os autores utilizam dados dos clientes e produtos, que não foram utilizados por coerência com as outras soluções.
- *ML-T-Based*: Modelo de aprendizado de máquina proposto como parte do trabalho Bauer and Jannach (2021), junto com uma abordagem de *stacking*. Diferente da abordagem *Sequence-based*, não utiliza-se redes neurais.
- *ML-RFM-Based* [Ramos and Silva (2023)]: Utiliza atributos do modelo RFM (Recência, Frequência e Monetário) em um algoritmo de aprendizado de máquina.
- *Pareto*[Schmittlein et al. (1987)]: Modelo utilizado para estimar o número esperado de transações, sendo um dos principais modelos probabilísticos.
- *BG/NBD*[Fader et al. (2005b)](*Beta-Geometric*): Modelo alternativo ao Pareto, computacionalmente mais eficiente e apresenta resultados similares.
- *Gamma-Gamma*[Fader et al. (2005a)]: Modelo probabilístico que permite calcular o valor monetário médio esperado por transação.

As métricas utilizadas foram o Erro Médio Absoluto (MAE) e a Raiz do Erro Quadrático Médio (RMSE), visto que quando lidam-se com valores monetários, outras medidas baseadas em porcentagens são menos informativas do que aquelas que trabalham com valores monetários absolutos. Em todas os modelos foi feita o ajuste dos hiper-parâmetros, utilizando a técnica *Exhaustive Grid Search*.

3.4.1 Dados utilizados

Para este estudo, foram utilizados cinco conjuntos de dados de diferentes contextos com o objetivo de validar os modelos em diversas situações.

O primeiro conjunto de dados (B1)[2] contém o histórico de compras até o final de junho de 1998 da empresa CDNOW, que operava um site de compras online especializado na venda de CDs e outros produtos musicais. O segundo conjunto (B2)[3] contém transações bancárias anonimizadas e reais de um banco da República

[2]https://www.brucehardie.com/datasets/
[3]https://data.world/lpetrocelli/czech-financial-dataset-real-anonymized-transactions

Tcheca. O terceiro conjunto de dados (B3)[4] consiste em outro conjunto bancário, com mais de 1 milhão de transações realizadas por mais de 800 mil clientes em um banco na Índia. O quarto conjunto de dados (B4)[5] contém informações sobre compras realizadas de 2016 a 2018 em vários locais do Brasil na plataforma Olist. O quinto conjunto (B5)[6] inclui dados de interações dos clientes ao longo de 5 meses (de outubro de 2019 a fevereiro de 2020) em uma loja online de cosméticos de médio porte. Foram consideradas apenas as interações de compra, envolvendo 110 mil usuários únicos e 159 mil compras.

Em todas as bases de dados, foram removidos os valores nulos e *outliers* em que o valor monetário total esteja fora dos percentis de 1% e 80%. Na Figura 3.1 são apresentados os dados após esse processo de limpeza. Pode-se observar diferentes números de usuários, transações e períodos em cada contexto, com o objetivo de simular uma variedade de ambientes e avaliar o desempenho de cada abordagem em diferentes configurações.

Base	Período (Semanas)	Usuários	Total de Transações	Média Transações por usuário	Min. de Transações por usuário	Max. de Transações por usuário	Moda de Transações por usuário
B1	78	312	1015	3,253	1	9	2
B2	49	4595	14067	3,061	1	5	3
B3	313	2550	552233	216,562	26	612	181
B4	104	1273	4489	3,526	3	16	3
B5	22	52479	62882	1,198	1	6	1

Figura 3.1: Tabela com a descrição de cada base de dados utilizada

3.4.2 Separação dos dados

Para analisar dados temporais, é comum segmentá-los em janelas temporais, como os métodos *Sliding Window* e *Expanding Window*. No primeiro, a janela de treinamento e teste é movida ao longo do tempo, descartando dados antigos e adicionando novos. No segundo, a cada iteração, os períodos mais recentes são incluídos na janela de treinamento, mantendo a de teste avançando. Ambas as abordagens têm vantagens: o *Sliding Window* alcança um equilíbrio favorável entre a precisão do modelo e o tempo de treinamento, especialmente quando se trata de testar dados de alta frequência, enquanto o *Expanding Window* é usado com mais frequência em séries temporais semanais, mensais ou trimestrais, onde o número de pontos históricos é limitado[7]. A escolha da abordagem depende do objetivo da

[4]https://www.kaggle.com/datasets/shivamb/bank-customer-segmentation?resource=download
[5]https://olist.com/pt-br/
[6]https://www.kaggle.com/datasets/mkechinov/ecommerce-events-history-in-cosmetics-shop
[7]https://www.uber.com/blog/omphalos/

análise e das características dos dados. Neste trabalho, a abordagem escolhida para usar em todos os modelos foi a *Expanding Window*, com exceção do *Sequence-Based*, que com o intuito de seguir a implementação apresentada no trabalho de Bauer and Jannach (2021), foi treinado com a janela anterior à janela usada para testes.

3.4.3 Resultados

O modelo proposto foi comparado com as soluções da literatura na previsão de três fatores: o número esperado de transações, o valor médio por período e o valor esperado do CLV para as próximas 4 semanas.

Na Figura 3.2 são apresentados os resultados comparativos para o número de transações. É possível observar que a solução desenvolvida apresentou resultados superiores em comparação com os outros modelos na maioria dos casos. Especificamente, em relação ao RMSE, pode-se observar melhorias significativas em todos os dados, com destaque para uma redução de até 12% na base (B5). Essa melhoria é crucial, pois indica a capacidade da solução em fazer previsões mais precisas e confiáveis.

Além disso, mesmo nos casos em que houve uma leve piora no desempenho, como na base (B3) onde o RMSE aumentou em aproximadamente 1%, ainda manteve resultados competitivos em comparação com os outros modelos. Essas pequenas variações podem ser atribuídas a nuances nos conjuntos de dados, como o número variado de transações por cliente, e destacam a importância da adaptação do modelo a diferentes contextos.

Um aspecto interessante a se notar é a superioridade dos modelos probabilísticos em relação ao MAE na base (B5). Isso sugere que, quando lida-se com um baixo número de transações por usuário e um período de treinamento curto, abordagens probabilísticas podem ser mais eficazes para fazer previsões precisas. No entanto, mesmo nesse cenário, a solução proposta baseada em aprendizado de máquina demonstrou resultados competitivos, destacando sua versatilidade e robustez.

Ao analisar os resultados apresentados na Figura 3.3, pode-se observar uma tendência consistente em relação ao desempenho do modelo em prever o valor monetário médio por transação em diferentes contextos. A proposta deste trabalho mostrou-se mais eficaz do que a abordagem baseada no RFM (Recência, Frequência, Monetário), evidenciando a relevância da inclusão de informações específicas sobre o comportamento das transações para melhorar a precisão das previsões.

Observa-se também que a solução desenvolvida trouxe melhorias significativas em relação aos modelos base. Em particular, destaca-se uma melhoria de até 35% no cálculo do MAE na base B4, indicando uma capacidade robusta do nosso modelo em prever com precisão o valor médio por período em um ambiente de comércio

Base	Métrica	Número de Transações					
		Sequence-Based	BG/NBD	Pareto	ML-RFM-Based	ML-T-Based	Solução
B1	MAE	0.79	1.20	1.19	0.17	0.15	0.17
	RMSE	1.05	1.53	1.52	0.32	0.33	0.32
B2	MAE	1.05	1.58	1.58	0.43	0.44	0.43
	RMSE	1.23	1.71	1.71	0.51	0.50	0.49
B3	MAE	70.51	161.38	161.38	0.64	0.89	0.63
	RMSE	99.02	180.63	180.64	0.79	1.13	0.80
B4	MAE	0.97	0.43	0.43	0.07	0.09	0.05
	RMSE	1.23	0.89	0.89	0.20	0.24	0.18
B5	MAE	0.72	0.18	0.18	0.29	0.31	0.22
	RMSE	0.90	0.52	0.52	0.40	0.50	0.35

Figura 3.2: Resultados relacionados no número de transações esperadas

eletrônico.

No entanto, ao analisar o desempenho em um contexto geral, nota-se variações nos resultados. Nos casos em que houve uma piora, como na base B3, onde o MAE aumentou em 11%, é importante investigar as possíveis razões por trás dessa variação. Questões como valores monetários divergentes pela base de dados e em um intervalo grande de análise devem ser investigados. Por outro lado, nos casos em que observa-se melhorias, como na base B4, onde houve uma melhoria de 35%, podendo-se atribuir esse sucesso à capacidade do modelo em capturar padrões específicos do comportamento do cliente e adaptar-se de forma eficaz às características únicas de cada conjunto de dados.

Base	Métrica	Valor Monetário				
		Sequence-Based	Gamma-Gamma	ML-RFM-Based	ML-T-Based	Solução
B1	MAE	6.20	10.30	1.52	0.45	1.67
	RMSE	7.10	13.60	2.87	1.47	2.92
B2	MAE	164.79	643.82	109.53	59.03	106.40
	RMSE	210.94	1046.13	180.09	96.67	174.58
B3	MAE	639.76	95.00	25.38	14.52	28.32
	RMSE	784.73	115.37	39.32	23.87	40.32
B4	MAE	68.12	31.88	4.80	5.91	3.08
	RMSE	79.45	66.23	16.15	19.18	15.52
B5	MAE	15.03	3.95	5.41	6.79	3.24
	RMSE	19.26	7.60	7.92	10.84	5.43

Figura 3.3: Resultados em relação ao valor monetário

Já em relação ao cálculo do CLV, como os custos envolvidos não foram informados nas bases de dados, é considerada a receita gerada pelo cliente até o período p, dada pela multiplicação do número de transações esperadas $NTx_u(p)$ e o valor monetário médio por transação $M_u(p)$, ajustada pela taxa de desconto de 1% ao ano.

Na Figura 3.4 são apresentados os resultados comparativos para o CLV. Pode-se observar que a solução proposta apresenta resultados superiores em todos os contextos, mesmo naqueles em que os resultados foram piores para o valor monetário (Figura 3.3), já que o CLV também é afetado pelo número de transações. Em um contexto geral, em relação ao RMSE, pode se observar uma melhoria significativa nos resultados. No pior caso, identifica-se uma melhora de aproximadamente 1%, enquanto no melhor caso, uma melhoria de até 14%, como observado na base B4.

Base	Métrica	CLV			
		Sequence-Based	ML-RFM-Based	ML-T-Based	Solução
B1	MAE	13.42	4.07	2.92	3.80
	RMSE	15.83	6.89	6.68	6.65
B2	MAE	503.07	186.29	192.59	181.52
	RMSE	638.78	273.02	273.04	269.03
B3	MAE	199145.30	2905.36	2994.83	2537.90
	RMSE	294464.30	3903.57	3861.02	3411.03
B4	MAE	107.86	3.95	8.39	3.24
	RMSE	128.22	18.22	27.21	15.59
B5	MAE	20.50	3.69	9.77	2.43
	RMSE	25.65	6.48	14.85	6.91

Figura 3.4: Resultados em relação ao CLV

Em resumo, os resultados indicam que o modelo genérico proposto tem o potencial de ser uma ferramenta valiosa para previsão do CLV para empresas de diferentes segmentos, sem a necessidade de dados sensíveis dos clientes.

3.5 Aplicação em um Caso Real do Mercado

No intuito de verificar o desempenho de cada modelo em um contexto real e atual, foram utilizados dados de uma drogaria, durante um período de 16 semanas, com uma média de 2.755 transações por semana. Com um total de 54.991 usuários durante esse período, a drogaria registrou um volume significativo de 151.525 transações no total.

Os dados foram disponibilizados por uma empresa parceira, seguindo as normas da LGPD, sendo anonimizados e sem nenhuma informação que possa permitir a rastreabilidade do cliente.

Os resultados revelam a eficácia da solução proposta, que obteve o melhor desempenho, com MAE de 0,57 e um RMSE de 0,62 para a previsão de número de transações, um ganho de no mínimo 3% em relação às soluções base. Já em relação ao valor monetário médio, a solução proposta apresentou uma melhora no erro de aproximadamente 2% em relação ao MAE. E em relação ao cálculo do CLV, a solução proposta obteve o melhor desempenho, com MAE de 20,50 e um RMSE de 28,93, um ganho de 2,9% em relação às soluções bases.

Esses valores indicam uma variação média relativamente baixa entre as transações previstas e os valores reais, sugerindo uma precisão sólida do modelo. Embora os outros métodos também tenham produzido resultados próximos, a proposta destaca-se como a mais promissora para prever o número de transações no cenário real da drogaria, junto com o CLV previsto de cada cliente. Vale ressaltar que um ganho de 2% a 3% pode parecer baixo, mas é muito significativo no longo prazo para uma drogaria com volume grande de clientes.

3.6 Conclusão e Trabalhos Futuros

Foi apresentada uma solução genérica para cálculo do CLV em diversos contextos, sendo avaliada em 6 bases de dados diferentes, com diferentes configurações e contextos. Além disso, a solução genérica proposta foi comparada com outras 5 soluções da literatura, que também se consideram genéricas, e foi visto que a extração dos atributos propostos foi benéfica para os resultados.

Como trabalho futuro, sugere-se que a solução genérica seja testada em outros cenários, como contratuais e discretos, visto que há uma escassez de dados com essas características. Além disto, englobar novas métricas no cálculo que podem ser retiradas dos poucos atributos fornecidos, como o cálculo do *churn* ou a lealdade do cliente.

4 Conclusões e Trabalhos Futuros

Neste trabalho, foi destacada a importância do cálculo do *Customer Lifetime Value* (CLV) e as principais dificuldades encontradas em seu desenvolvimento.

Na primeira etapa, foi proposta uma solução para o cálculo do CLV utilizando apenas atributos fornecidos pelo modelo RFM. Esta solução foi comparada com modelos consolidados na literatura, revelando que os modelos de aprendizado de máquina se destacam na previsão do número de transações em situações reais. Em termos de previsão monetária, ambos os modelos apresentaram resultados competitivos, com vantagem para os modelos de aprendizado de máquina. O cálculo do CLV, conforme apresentado neste trabalho, considera o número esperado de transações e o valor esperado por transação. Em testes com bases de dados reais, a solução proposta demonstrou resultados mais precisos.

Já na segunda etapa, a solução foi desenvolvida e apresentada como uma abordagem genérica para o cálculo do CLV em diversos contextos, sendo avaliada em seis bases de dados diferentes com variadas configurações. Além disso, a solução genérica proposta foi comparada com outras cinco soluções da literatura, também consideradas genéricas, evidenciando que a extração dos atributos propostos contribuiu positivamente para os resultados, obtendo o melhor desempenho, com uma melhoria mínima de aproximadamente 1% e máxima de até 14%.

Foi possível analisar o desempenho dos modelos probabilísticos em comparação aos modelos de aprendizado de máquina em diferentes contextos, e percebe-se que os modelos de aprendizado de máquina possuem resultados superiores e conseguem acompanhar melhor as tendências dos clientes. Porém, é importante frisar que, os resultados dos modelos probabilísticos podem ser úteis, visto que os atributos fornecidos pelo RFM se mostraram importantes pelos resultados obtidos. Além disso, a solução elaborada apresentou resultados competitivos, utilizando somente dados de transações, permitindo que mantenha a privacidade sobre informações pessoais dos clientes. Usar estes tipos de dados permite que a solução seja considerada genérica, permitindo ser usada em diversos negócios, desde o contexto bancário à comércios eletrônicos. Por fim, foi feita a validação utilizando dados reais e recentes de uma drogaria, e os resultados obtidos validaram a eficácia da solução.

Em relação a trabalhos futuros, sugere-se a adição de parâmetros relacionados a padrões de compra dos clientes, bem como métricas que identifiquem a satisfação e a probabilidade de *churn*. Recomenda-se também que a solução genérica seja testada em outros cenários, como contextos contratuais e discretos. Pode-se também testar

em diferentes períodos temporais, visto que o cálculo do CLV à longo prazo pode ser interessante para estimar o desempenho da determinada companhia. Outra abordagem, ainda com enfoque na privacidade, consiste em utilizar informações sobre os produtos consumidos, ou dados contextuais da companhia que está sendo aplicado o cálculo.

Referências Bibliográficas

ABIDAR, L., ZAIDOUNI, D., ASRI, I. E., and ENNOUAARY, A. (2023). Predicting customer segment changes to enhance customer retention: A case study for online retail using machine learning. *International Journal of Advanced Computer Science and Applications*, 14(7).

Aeron, H., Kumar, A., and Moorthy, J. (2012). Data mining framework for customer lifetime value-based segmentation. *Journal of Database Marketing &amp Customer Strategy Management*, 19(1):17–30.

Aslekar, A., Sahu, P., and Pahari, A. (2019). Big data analytics for customer lifetime value prediction. *Telecom Business Review*.

Bauer, J. and Jannach, D. (2021). Improved customer lifetime value prediction with sequence-to-sequence learning and feature-based models. *ACM Trans. Knowl. Discov. Data*, 15(5).

Calabourdin, A. V. and Aksenov, K. A. (2023). Streaming bayesian modeling for predicting fat-tailed customer lifetime value.

Chan, C. C. H. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. *Expert Systems with Applications*, 34(4):2754–2762.

Cheng, C.-H. and Chen, Y.-S. (2009). Classifying the segmentation of customer value via rfm model and rs theory. *Expert Systems with Applications*, 36(3, Part 1):4176–4184.

Comlan, M. and Adiba, E. (2024). Customer lifetime value in streaming: a markov chain approach. In *2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, pages 1–6. IEEE.

Desirena, G., Diaz, A., Desirena, J., Moreno, I., and Garcia, D. (2019). Maximizing customer lifetime value using stacked neural networks: An insurance industry application. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 541–544.

Dwyer, F. R. (1997). Customer lifetime valuation to support marketing decision making. *Journal of Direct Marketing*, 11(4):6–13.

Ekinci, Y., Uray, N., and Ulengin, F. (2014). A customer lifetime value model for the banking industry: a guide to marketing actions. *European Journal of Marketing*, 48.

Fader, P. S., Hardie, B. G., and Lee, K. L. (2005a). RFM and CLV: Using Iso-Value Curves for Customer Base Analysis. *Journal of Marketing Research*, 42(4):415–430.

Fader, P. S., Hardie, B. G. S., and Lee, K. L. (2005b). "counting your customers" the easy way: An alternative to the pareto/NBD model. *Marketing Science*, 24(2):275–284.

Haenlein, M., Kaplan, A. M., and Beeser, A. J. (2007). A model to determine customer lifetime value in a retail banking context. *European Management Journal*, 25(3):221–234.

Hiziroglu, A. and Sengul, S. (2012). Investigating two customer lifetime value models from segmentation perspective. *Procedia - Social and Behavioral Sciences*, 62:766–774. World Conference on Business, Economics and Management (BEM-2012), May 4–6 2012, Antalya, Turkey.

Hughes, A. (1996). Boosting response with rfm. *Marketing Tools*, 5:4–7.

Kahreh, M. S., Tive, M., Babania, A., and Hesan, M. (2014). Analyzing the applications of customer lifetime value (CLV) based on benefit segmentation for the banking sector. *Procedia - Social and Behavioral Sciences*, 109:590–594.

Kailash, H., Kanwar, K., Sonia, S., and Kant, R. (2023). Machine learning algorithms for predicting customers' lifetime value: A systematic evaluation. In *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 538–541.

Khajvand, M., Zolfaghar, K., Ashoori, S., and Alizadeh, S. (2011). Estimating customer lifetime value based on rfm analysis of customer purchase behavior: Case study. *Procedia Computer Science*, 3:57–63. World Conference on Information Technology.

Kotler, P. (1974). Marketing during periods of shortage. *Journal of Marketing*, 38(3):20–29.

Kruger, E. (2011). Top market strategy: Applying the 80/20 rule.

Kumar, A., Singh, K. U., Kumar, G., Choudhury, T., and Kotecha, K. (2023). Customer lifetime value prediction: Using machine learning to forecast clv and enhance customer relationship management. In *2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–7.

Mammadzada, A., Alasgarov, E., and Mammadov, A. (2021). Application of bg / nbd and gamma-gamma models to predict customer lifetime value for financial institution. In *2021 IEEE 15th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–6.

Méndez-Suárez, M. and Crespo-Tejero, N. (2021). Why do banks retain unprofitable customers? a customer lifetime value real options approach. *Journal of Business Research*, 122:621–626.

Pfeifer, P. E. and Carraway, R. L. (2000). Modeling customer relationships as markov chains. *Journal of Interactive Marketing*, 14(2):43–55.

POPA, A.-L., SASU, D. V., and TARCZA, T. M. (2021). Investigating The Importance Of Customer Lifetime Value In Modern Marketing - A Literature Review. *Annals of Faculty of Economics*, 30(2):410–416.

Qi, J., Qu, Q.-X., Zhou, Y.-P., and li, L. (2015). The impact of users' characteristics on customer lifetime value raising: evidence from mobile data service in china. *Information Technology and Management*, 16:273–290.

Ramos, J. and Silva, F. (2023). Customer lifetime value prediction: A machine learning approach. In *Anais do XX Encontro Nacional de Inteligência Artificial e Computacional*, pages 486–500, Porto Alegre, RS, Brasil. SBC.

Reinartz, W. J. and Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of Marketing*, 64(4):17–35.

Sargeant, A. (2003). Using donor lifetime value to inform fundraising strategy. *Nonprofit Management and Leadership*, 12:25 – 38.

Schmittlein, D. C., Morrison, D. G., and Colombo, R. (1987). Counting your customers: Who are they and what will they do next? *Management Science*, 33(1):1–24.

Shin, D. and Kim, D. (2022). A historical paradigm shift from producer centricity to customer centricity: An organization theory approach to the enabling effects of digital technology. *Korean Academy of Management*.

Singh, S. and Jain, D. (2013). Measuring customer lifetime value: Models and analysis. *SSRN Electronic Journal*.

Sun, Y., Liu, H., and Gao, Y. (2023). Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. *Heliyon*, 9(2).

Temor Qismat, Y. F. (2020). Comparison of classical rfm models and machine learning models in clv prediction. Master's thesis, Norwegian Business School BI Open, Oslo.

Ullah, A., Mohmand, M. I., Hussain, H., Johar, S., Khan, I., Ahmad, S., Mahmoud, H. A., and Huda, S. (2023). Customer analysis using machine learning-based classification algorithms for effective segmentation using recency, frequency, monetary, and time. *Sensors*, 23(6):3180.

Vanderveld, A., Pandey, A., Han, A., and Parekh, R. (2016). An engagement-based customer lifetime value system for e-commerce. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 293–302, New York, NY, USA. Association for Computing Machinery.