A solution for predicting the Customer Lifetime Value of different market segments

J. M. A. M . Ramos[1], F. A. Silva[1]

Universidade Federal de Viçosa, Florestal, MG, Brazil
Manna Team
joao.m.ramos,fabricio.asilva@ufv.br

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, demonstrating its applicability, in addition to being compared with five reference solutions from the literature. The results showed that it was possible to achieve a reduction of up to 34% in CLV calculation error in the scenarios.

CCS Concepts: • **Information systems** → *Business intelligence*; • **Computing methodologies** → **Feature selection**; **Machine learning algorithms**.

Keywords: LTV,CLV,Machine Learning

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, potencialmente quadruplicando o lucro de uma empresa ao atingir e reter clientes altamente lucrativos [Kruger 2011].

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 gerar um modelo genérico para estimar o CLV, surgiram abordagens probabilísticas, como a de *Pareto/Gamma-Gamma* [Fader and Hardie 2013] ou *Beta-Geometric/Negative Binomial Distribution* (BG/NBD) [Fader et al. 2005]. 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 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.

O restante deste texto está organizado com a apresentação de trabalhos relacionados na Seção 2; as especificações da solução na Seção 3; os métodos de avaliação e resultados obtidos na Seção 4; por fim, as considerações finais estão na Seção 5.

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 I.

Trabalho	Abordagem	Categoria	Contexto dos Dados	Dados Necessários
[Fader et al. 2005]	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
[Khajvand et al. 2011]	Modelo RFM	Não-Contratual, Contínuo	Comércio Varejista	Transações
[Hiziroglu and Sengul 2012]	Modelo RFM	Não-Contratual, Contínuo	Comércio Eletrônico	Clientes, transações e demográficos
[Fader and Hardie 2013]	Modelo Probabilístico	Não-Contratual, Contínuo	Comércio Eletrônico	Transações
[Ekinci et al. 2014]	Redes Neurais	Contratual, Contínuo	Setor Bancário	Clientes, transações e produtos
[Qi et al. 2015]	Modelo Conceitual	Contratual, Discreta	Telecomunicação.	Satisfação e lealdade do cliente, demográficos
[Vanderveld et al. 2016]	Aprendizado de Máquina	Não-Contratual, Contínuo	Comércio Eletrônico	Demográficos, transações, produtos, relacionamentos
[Bauer and Jannach 2021]	Redes Neurais, Stacking	Não-Contratual, Contínuo	Comércio Eletrônico	Demográficos, transações, clientes, produto
[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
[Sun et al. 2023]	Aprendizado de Máquina	Não-Contratual, Contínuo	Comércio Eletrônico	Demográficos, transações, clientes, produtos
[Comlan and Adiba 2024]	Cadeias de Markov	Contratual, Contínuo	Streaming	Clientes, compra de pacotes e atividades de visualização

Table I. Características dos Trabalhos Relacionados

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 a abordagem *Pareto* [Schmittlein et al. 1987], 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; a abordagem *Gamma-Gamma* [Fader and Hardie 2013], uma extensão do modelo de Pareto, que permite a estimativa do valor médio monetário das futuras transações; e o modelo *BG/NBD* [Fader et al. 2005], que é uma alternativa ao modelo de Pareto, apresentando resultados semelhantes, mas com maior eficiência computacional.

2.2 Abordagens com aprendizado de máquina

A utilização de algoritmos de aprendizado de máquina para a previsão do CLV é recente [Temor Qismat 2020]. Este tipo de abordagem permite a utilização de informações sobre o cliente, dados demográficos ou até dados sobre as interações do cliente com a empresa. 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. Apesar disto, as soluções apresentam bons resultados quando comparadas aos modelos clássicos [Ramos and Silva 2023].

Existem duas abordagens principais utilizando aprendizado de máquina para estimar o CLV: como um problema de classificação ou de regressão. Na abordagem de classificação, os clientes são segmentados e tratados de maneira diferente com base no valor que representam para a empresa. Por exemplo, o trabalho de [Abidar et al. 2023] utiliza dados de um site varejista para separar os clientes em três segmentos: baixo, médio e alto valor. De forma similar, [Haenlein et al. 2007] realiza a classificação de clientes em um contexto bancário, categorizando-os de acordo com os serviços utilizados e interesses. O estudo de [Sun et al. 2023], por sua vez, usa dados de uma loja online para medir o CLV atual e utilizá-lo na segmentação dos clientes.

A abordagem de regressão visa estimar um valor numérico que representa o CLV do cliente. O trabalho de [Kailash et al. 2023] usa o conjunto de dados IBM Watson para avaliar e comparar o desempenho de diversos algoritmos de regressão em uma indústria de seguros. O estudo de [Vanderveld et al. 2016] utiliza dados de uma plataforma de comércio online, atualizando diariamente o CLV dos clientes. Pesquisas como a de [Kumar et al. 2023] utilizam o modelo ARIMA e o comparam com outros modelos de aprendizado de máquina. Além disso, [Bauer and Jannach 2021] propõem um *framework* de previsão de CLV que combina o desempenho de duas abordagens: uma baseada em redes neurais sequenciais e outra em um modelo de regressão, demonstrando sua eficácia em dois cenários de aplicação no comércio eletrônico.

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 para 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 abordando o CLV como uma abordagem de regressão.

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 II foram propostos. Esses atributos extraem características do relacionamento do cliente com a empresa, sem violar a privacidade do mesmo.

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.

Table II. Descrição das variáveis e fórmulas

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.

Neste trabalho, adota-se $h = 4$ semanas, visando realizar uma previsão de curto prazo. Vale destacar que, apesar do CLV considerar o valor vitalício, a ausência de dados impede uma análise de longo prazo, como observado na 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, Elastic-Net, Random Forest Regressor, Kernel Ridge, Gradient Boost Regressor, XGBoost Regressor e Ligth GBM*) com o intuito de estimar o valor do $NTx_u(p+h)$ e $M_u(p+h)$, sendo que aquele com o melhor resultado para cada base foi escolhido. 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.

4. AVALIAÇÃO E RESULTADOS

Com o intuito de avaliar o modelo desenvolvido, foram escolhidos trabalhos do estado da arte com diferentes características. A abordagem *Sequence-Based* [Bauer and Jannach 2021] utiliza de redes neurais recorrentes, porém, na proposta original, os autores utilizam dados dos clientes e produtos, que não foram utilizados por coerência com as outras soluções. A abordagem *ML-T-Based* segue a metodologia sugerida por Bauer e utiliza *stacking* dos valores do modelo *Sequence-Based*. A abordagem *ML-RFM-Based* [Ramos and Silva 2023] emprega atributos do modelo RFM (Recência, Frequência e Monetário) em um algoritmo de aprendizado de máquina. O modelo de *Pareto* [Schmittlein et al. 1987] é utilizado para estimar o número esperado de transações e é considerado um dos principais

modelos probabilísticos. Alternativamente, o modelo *BG/NBD* (*Beta-Geometric*) [Fader et al. 2005] é computacionalmente mais eficiente e apresenta resultados similares ao Pareto. Por fim, o modelo *Gamma-Gamma* [Fader and Hardie 2013], é um modelo probabilístico que permite calcular o valor monetário médio esperado por transação.

Em relação à divisão dos dados, neste trabalho, a abordagem escolhida para usar em todos os modelos foi a *Expanding Window*. Além disto, foi feito o ajuste de hiper-parâmetros com os valores de cada iteração dos dados de treino usando o *Exhaustive Grid Search*. Já 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 não refletem a precisão diferente das que trabalham com valores monetários absolutos.

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)[1] 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)[2] contém transações bancárias anonimizadas e reais de um banco da República Tcheca. O terceiro conjunto de dados (B3)[3] 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)[4] contém informações sobre compras realizadas de 2016 a 2018 em vários locais do Brasil na plataforma Olist. O quinto conjunto (B5)[5] inclui dados de interações dos clientes por 5 meses (de out. de 2019 a fev. de 2020) em uma loja online de cosméticos de médio porte sendo consideradas apenas as interações de compra.

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 Tabela III 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. Transações por Usuário	Max. Transações por Usuário	Moda 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

Table III. Características dos Trabalhos Relacionados

4.2 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 Tabela IV 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

[1]https://www.brucehardie.com/datasets/
[2]https://data.world/lpetrocelli/czech-financial-dataset-real-anonymized-transactions
[3]https://www.kaggle.com/datasets/shivamb/bank-customer-segmentation?resource=download
[4]https://olist.com/pt-br/
[5]https://www.kaggle.com/datasets/mkechinov/ecommerce-events-history-in-cosmetics-shop

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.

Base	Métrica	Número de Transações					
		Sequence-Based	BG/NBD	Pareto	ML-RFM-Based	ML-T-Based	Modelo Proposto
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**
	Tempo de Exec. (s)	473	**0,14**	0,47	0,43	1,60	0,94
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**
	Tempo de Exec. (s)	2677	**0,13**	3,67	1,01	4,94	3,02
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
	Tempo de Exec. (s)	3993	**0,16**	0,65	1,39	6,12	3,46
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**
	Tempo de Exec. (s)	1529	**0,10**	0,55	0,71	2,41	1,27
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**
	Tempo de Exec. (s)	480	**0,43**	347	1,42	20	11

Table IV. Resultados de cada um dos modelos de previsão de transações

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.

Base	Métrica	Valor Monetário				
		Sequence-Based	Gamma-Gamma	ML-RFM-Based	ML-T-Based	Modelo Proposto
B1	MAE	6,20	10,30	1,52	**0,45**	1,67
	RMSE	7,10	13,60	2,87	**1,47**	2,92
	Tempo de Exec. (s)	348	**0,05**	0,58	1,60	0,86
B2	MAE	164,79	643,82	109,53	**59,03**	106,40
	RMSE	210,94	1046,13	180,09	**96,67**	174,58
	Tempo de Exec. (s)	2677	**0,04**	1,75	4,94	3,16
B3	MAE	639,76	95,00	25,38	**14,52**	28,32
	RMSE	784,73	115,37	39,32	**23,87**	40,32
	Tempo de Exec. (s)	3993	**0,06**	1,67	6,12	4,33
B4	MAE	68,12	31,88	4,80	5,91	**3,08**
	RMSE	79,45	66,23	16,15	19,18	**15,52**
	Tempo de Exec. (s)	1529	**0,04**	0,84	2,41	1,30
B5	MAE	15,03	3,95	5,41	6,79	**3,24**
	RMSE	19,26	7,60	7,92	10,84	**5,43**
	Tempo de Exec. (s)	480	**0,04**	1,65	20	11

Table V. Resultados de cada um dos modelos de previsão do valor monetário médio

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

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 d de 1% ao ano, dado pelo seguinte cálculo:

$$CLV_i = \sum_{t=1}^{\mathrm{T}} \frac{N(p) \times M(p)}{(1+d)^t} \tag{1}$$

Na Tabela VI 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 (Tabela V), 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 aproximadamente 34%, como observado na base B5. É importante frisar que a melhora, até mesmo em pequenos valores, pode ter um impacto significante na predição, visto que os erros impactam diretamente as estratégias de investimento em marketing e vendas. Um *CLV* inflacionado pode justificar gastos maiores que não se sustentam ao longo do tempo, enquanto um *LTV* subestimado pode levar a uma abordagem conservadora, resultando em menor crescimento.

Base	Métrica	CLV			
		Sequence-Based	ML-RFM-Based	ML-T-Based	Modelo Proposto
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**

Table VI. Resultados da previsão do *CLV*

5. CONCLUSÃO E TRABALHOS FUTUROS

Foi apresentada uma solução genérica para cálculo do CLV em diversos contextos, sendo avaliada em 5 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. Pode-se observar que a solução proposta obteve melhoras no número esperado de transações e no valor monetário médio de transação, resultando em um cálculo do *CLV* mais preciso e próximo da realidade, permitindo estimar com maior acurácia o lucro. Esta melhora traz um impacto significante, visto que os erros do modelo podem se propagar, a precisão na estimativa tanto do número de transações quanto do valor médio por transação é crucial para uma avaliação adequada do *CLV*, impactando diretamente a tomada de decisão estratégica e a alocação de recursos em uma empresa.

Como trabalhos futuros, 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.

Agradecimentos

Os autores agradecem ao apoio do Manna Team, da Fundação Araucária, da Softex, CNPq (Número 421548/2022-3) e a Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

REFERENCES

Abidar, L., ZAIDOUNI, D., ASRI, I. E., and ENNOUAARY, A. 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), 2023.

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

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

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

Ekinci, Y., Uray, N., and Ülengin, F. A customer lifetime value model for the banking industry: a guide to marketing actions. *European Journal of Marketing* 48 (3/4): 761–784, 2014.

Fader, P. S. and Hardie, B. G. The gamma-gamma model of monetary value. *February* vol. 2, pp. 1–9, 2013.

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

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

Hiziroglu, A. and Sengul, S. Investigating two customer lifetime value models from segmentation perspective. *Procedia-Social and Behavioral Sciences* vol. 62, pp. 766–774, 2012.

Kailash, H., Kanwar, K., Sonia, S., and Kant, R. 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)*. pp. 538–541, 2023.

Khajvand, M., Zolfaghar, K., Ashoori, S., and Alizadeh, S. Estimating customer lifetime value based on rfm analysis of customer purchase behavior: Case study. *Procedia computer science* vol. 3, pp. 57–63, 2011.

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

Kumar, A., Singh, K. U., Kumar, G., Choudhury, T., and Kotecha, K. 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)*. pp. 1–7, 2023.

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

Qi, J., Qu, Q.-X., Zhou, Y.-P., and li, L. The impact of users' characteristics on customer lifetime value raising: evidence from mobile data service in china. *Information Technology and Management* vol. 16, pp. 273–290, 12, 2015.

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

Reinartz and Kumar. On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of Marketing* vol. 64, pp. 17–35, 10, 2000.

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

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

Temor Qismat, Y. F. *Comparison of classical RFM models and Machine learning models in CLV prediction.* M.S. thesis, Norwegian Business School BI Open, Oslo, 2020.

Vanderveld, A., Pandey, A., Han, A., and Parekh, R. 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. Association for Computing Machinery, New York, NY, USA, pp. 293–302, 2016.