

Evaluating user segments for predicting Customer Lifetime Value

Victória C. S. Rodrigues¹, João M. Ramos^{1,3}, Fabrício A. Silva^{1,3}, Linnyer B. R. Aylon^{2,3}

¹Universidade Federal de Viçosa (UFV)
Laboratório de Inteligência em Sistemas Pervasivos e Distribuídos (NESPED-Lab)
Florestal - MG - Brasil

²Universidade Estadual de Maringá (UEM) Maringá - PR - Brasil

³Manna Team

{victoria.rodrigues, joao.m.ramos, fabricio.asilva}@ufv.br

lbruiz@uem.br

Abstract. *The Customer Lifetime Value (CLV) is fundamental to business operations, offering a comprehensive understanding of customer value over time. This study aims to evaluate the impact of customer segmentation on CLV prediction, using machine learning models established in the literature to estimate the number of transactions and the average monetary value across three distinct datasets. A comparative analysis was conducted on the values related to the number of transactions and the average monetary value in two different scenarios: one where the model is trained exclusively with the characteristics of each segment and another where it is trained with all the data. The results reveal that customer segmentation can improve prediction accuracy, and the choice between segmenting customers and training with all data should be based on the specific characteristics and nature of the dataset.*

Resumo. *O Customer Lifetime Value (CLV) é fundamental para as operações comerciais, oferecendo uma compreensão abrangente do valor do cliente ao longo do tempo. Este estudo busca avaliar o impacto da segmentação dos clientes na previsão do CLV, utilizando modelos de aprendizado de máquina estabelecidos na literatura para estimar o número de transações e o valor monetário médio em três conjuntos de dados distintos. Foi feita uma análise comparativa dos valores relativos ao número de transações e ao valor médio monetário em dois cenários distintos: um em que o modelo é treinado exclusivamente com as características de cada segmento e outro em que é treinado com todos os dados. Os resultados obtidos revelam que a segmentação de clientes pode melhorar a precisão das previsões, e que a escolha entre segmentar os clientes e treinamento com todos os dados deve ser baseada nas características específicas do conjunto de dados e na natureza dos mesmos.*

1. Introdução

Compreender o comportamento dos clientes em relação ao negócio torna-se uma etapa crucial para o êxito de um empreendimento na atualidade. Nesse sentido, é importante direcionar a atenção para os clientes, e não apenas para os produtos e/ou serviços ofertados.

Conforme discutido por Rust et al. [2001], deve-se ter uma nova percepção sobre as atividades desenvolvidas para os clientes com base na sua importância para a empresa. Isso implica que não será levado em consideração apenas a lucratividade da oferta, mas sim a rentabilidade do cliente para o negócio. Desse modo, as empresas precisam ser capazes de identificar o valor potencial do relacionamento de cada cliente, ou seja, o valor que o cliente gera e poderá gerar para a companhia [Sharma, 2007].

Os clientes que apresentam uma contínua receita no negócio são considerados como retidos ou leais. Tal continuidade de receita pode ser denominada como *Customer lifetime value* (CLV) [Fader et al., 2005], que representa o valor individual do cliente no decorrer de um período de tempo. Pode-se considerar que este valor consiste na margem de lucro que uma empresa espera obter ao longo de todo relacionamento comercial com seu cliente durante as transações efetuadas. Dessa maneira, o foco principal está sobre como investir em determinados clientes pode tornar a empresa mais lucrativa [Malthouse and Mulhern, 2008].

A literatura indica que trabalhos envolvendo aprendizado de máquina têm demonstrado grande potencial na melhoria das previsões do CLV [Dai, 2022], superando significativamente os métodos tradicionais [Kumar et al., 2023]. No entanto, os trabalhos existentes frequentemente se concentram em técnicas avançadas sem examinar como variações na estratégia de treinamento, como a segmentação de clientes, influenciam os resultados, negligenciando a investigação aprofundada sobre como diferentes estratégias de modelagem afetam o desempenho dos modelos.

Os hábitos de compra variam de cliente para cliente. Por exemplo, alguns fazem compras de valores menores regularmente, enquanto outros realizam compras esporádicas, porém com valores maiores. Porém, todos os perfis de clientes têm sua importância na parcela de lucros da companhia, apesar dessas diferenças. No entanto, essa variação pode afetar os resultados do modelo, visto que um único modelo pode ter dificuldades em identificar esses padrões distintos.

A hipótese do presente estudo é que a criação de modelos específicos para cada segmento de clientes pode resultar em ganhos na precisão das previsões do CLV. Modelos segmentados podem capturar de forma mais eficaz os padrões de comportamento de compra de diferentes grupos de clientes, resultando em previsões mais precisas e estratégias de marketing mais eficientes.

Este trabalho aborda essa lacuna ao analisar detalhadamente o impacto da segmentação de clientes na previsão do CLV usando modelos de regressão. Ao comparar o desempenho dos modelos treinados com dados segmentados versus dados agregados, busca-se compreender como a divisão dos clientes em grupos distintos pode melhorar a precisão das previsões e identificar estratégias mais eficazes para aprimorar o CLV. Os resultados gerais do trabalho mostram que a abordagem de segmentação resulta em uma melhora na precisão das previsões tanto para transações quanto para valores monetários. Essa abordagem permite capturar melhor as características específicas de cada segmento de clientes, apresentando uma melhoria média de até aproximadamente 80% nas métricas MSE, RMSE e MASE.

Este texto está organizado da seguinte forma: na seção 2 são apresentados os trabalhos relacionados; na seção 3 é feita a descrição detalhada dos materiais e métodos

empregados durante a pesquisa; na seção 4 são apresentados os resultados e a discussão; e, por último, as considerações finais estão contempladas na seção 5.

2. Trabalhos relacionados

A ampla aceitação da importância em estabelecer e fortalecer vínculos sólidos com a base de clientes existente é reconhecida, sobretudo considerando que o custo para conquistar novos clientes pode ser até cinco vezes mais elevado[Sargeant, 2001]. Diante desse cenário, diversos métodos foram desenvolvidos para aprofundar a compreensão do comportamento de compra dos clientes, permitindo às empresas identificar e cultivar relacionamentos com seus clientes mais valiosos.

Dentre estes métodos, a previsão do CLV e a segmentação de clientes são essenciais no ambiente de negócios atual. Com a utilização de uma grande quantidade de dados e modelos de aprendizado de máquina, os profissionais de marketing conseguem obter uma visão detalhada dos comportamentos dos clientes. Adicionalmente, a coleta de históricos de navegação e compras possibilita a oferta de promoções mais personalizadas e eficazes para cada cliente [Dai, 2022].

O emprego de técnicas avançadas de aprendizado de máquina permite aprimorar significativamente a previsão do CLV, resultando em uma alocação de recursos mais eficiente e em estratégias de marketing mais precisas na gestão do relacionamento com o cliente [Kumar et al., 2023]. Ilustrando este cenário, destaca-se, especialmente o trabalho de [Sun et al., 2023], cujo objetivo é segmentar os clientes utilizando modelos de aprendizado de máquina e análise de gerenciamento de relacionamento com o cliente (*Customer Relationship Management* ou CRM). Esta abordagem de segmentação é utilizada também neste trabalho, para identificar os segmentos do CLV em situações não contratuais, que consistem em situações onde não há um contrato envolvendo a transação entre as partes envolvidas.

Outro estudo relevante é o de [Maryani and Riana, 2017], que realiza a segmentação dos clientes usando o modelo de Recência, Frequência e Monetária (RFM) com o algoritmo *K-Means*, além da classificação de clientes com *Decision Tree* com o objetivo de fornecer recomendações no CRM para empresas de médio porte. Um terceiro método, explorado por [Fader et al., 2005], incorpora o modelo RFM em conjunto com os modelos probabilísticos Pareto/NBD para capturar o fluxo de transações ao longo do tempo dos clientes. Adicionalmente, é utilizado o modelo Gamma-Gamma para estimar o valor médio gasto em cada transação.

Como alternativa, trabalhos como o de [AutoresOmitidosparaRevisão] buscam investigar o uso de algoritmos de aprendizado de máquina para prever o CLV e compará-los com os principais modelos da literatura, tais como Pareto/NBD, BG/NDB e o modelo Gamma-Gamma. Baseado nos resultados apresentados, notou-se a superioridade dos modelos de aprendizado de máquina em relação aos modelos tradicionais mencionados.

Diante disso, o presente trabalho propõe a utilização de diferentes modelos de aprendizado de máquina para prever o número de transações e o valor monetário médio gasto, realizando uma comparação entre os diferentes modelos em dois cenários distintos: (A) utilizando apenas as características do segmento para treinar o modelo e (B) utilizando todos os dados disponíveis para o treinamento do modelo. O objetivo é entender o impacto

da segmentação dos clientes na etapa de treinamento dos modelos de aprendizado de máquina.

3. Materiais e métodos

Com o intuito de examinar como a segmentação de clientes, influenciam os resultados, foram propostos dois cenários:

- Cenário A: Neste cenário, os clientes são segmentados em grupos com características distintas em relação aos padrões de consumo. Para cada segmento, é desenvolvido um modelo específico que estima o comportamento dos clientes dentro daquele grupo. A segmentação visa capturar as particularidades de cada grupo, o que pode resultar em previsões mais precisas e relevantes.
- Cenário B: Neste cenário, é utilizado um único modelo para toda a base de clientes, sem realizar uma segmentação prévia. O modelo é treinado com um conjunto de dados que representa a população geral, buscando identificar padrões e comportamentos que se aplicam a todos os clientes de forma abrangente.

Ambos os cenários utilizam das mesmas etapas descritas posteriormente, com a exceção da segmentação.

3.1. Dados

Para validar as técnicas aplicadas, foram utilizados três conjuntos de dados em diferentes cenários comerciais e com diferentes comportamentos de aquisição dos clientes.

A primeira base de dados, denominada *CDNOW*¹, tem o histórico de compras de 23.570 clientes que fizeram sua primeira compra no primeiro trimestre de 1997 até junho de 1998. Os dados são de uma empresa que vendia CDs e produtos relacionados à música on-line. Este conjunto de dados está presente no artigo original do modelo Gamma-Gamma [Fader et al., 2005].

A segunda base de dados, denominada *Online Retail I* [mis, 2015], contém dados das transações ocorridas entre 12 de janeiro de 2010 a 12 de setembro de 2011 para um varejo on-line com sede no Reino Unido. O conjunto de dados está disponível na plataforma *UCI Machine Learning Repository*².

A terceira base de dados, também utilizando a plataforma *UCI Machine Learning Repository* [Chen, 2019], chamado *Online Retail II*³, contém todas as transações ocorridas em lojas online de varejo com sede no Reino Unido, sendo que as transações foram registradas entre 12 de janeiro de 2009 e 12 de setembro de 2011.

3.2. Segmentação

Para a segmentação, utiliza-se o modelo RFM (Recência, Frequência e Valor Monetário) [Fader et al., 2005] para extrair métricas importantes sobre o comportamento dos clientes. Essas variáveis possibilitam uma categorização de acordo com os padrões de consumo dos mesmos. Esta técnica de segmentação compreende três medidas, sendo elas: recência, frequência e o valor monetário.

¹<http://brucehardie.com/notes/026/>

²<https://archive.ics.uci.edu/dataset/352/online+retail>

³<https://archive.ics.uci.edu/dataset/502/online+retail+ii>

A recência pode ser definida pelo tempo entre a última compra do cliente e a data de referência da análise, sendo que esse intervalo pode ser medido em dias, semanas ou meses. Isto significa que clientes que têm um menor valor de recência efetuaram uma compra mais recentemente. Um exemplo dessa situação é mostrado na Figura 1, em que o cliente A tem uma recência maior que o cliente B quando observamos a última compra realizada por eles em relação à data de referência escolhida.

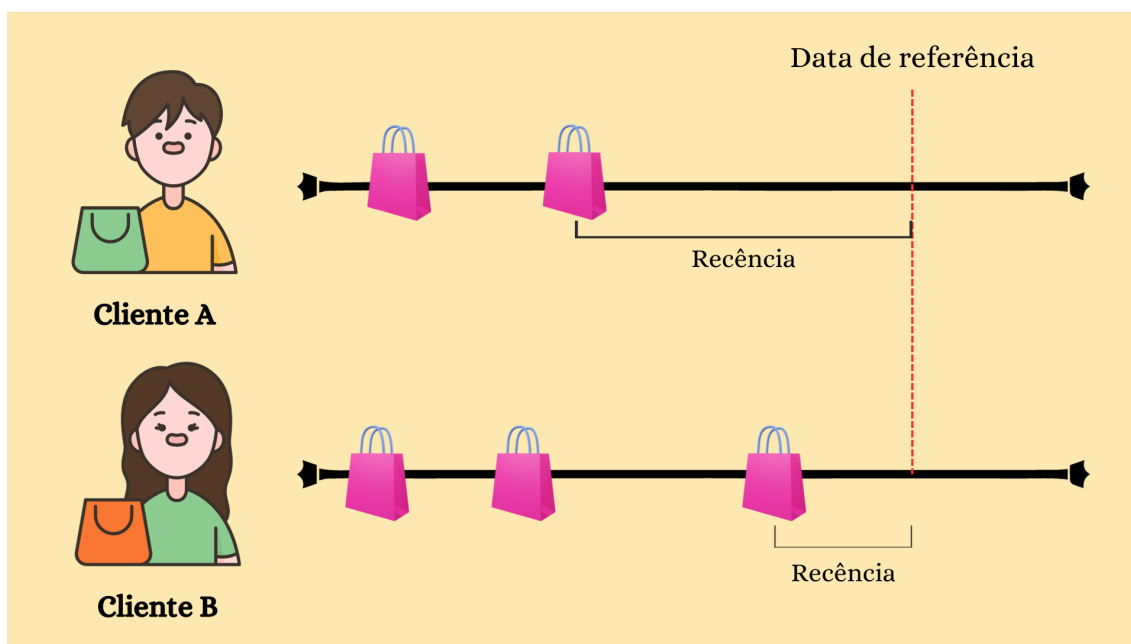


Figura 1. Ilustração de exemplo sobre o valor da recência.

Já a frequência diz respeito a quantidade de compras realizadas pelo cliente em um determinado intervalo de tempo, conforme ilustrado na Figura 2. Quanto maior a frequência, mais compras esse cliente realizou na empresa no período avaliado.

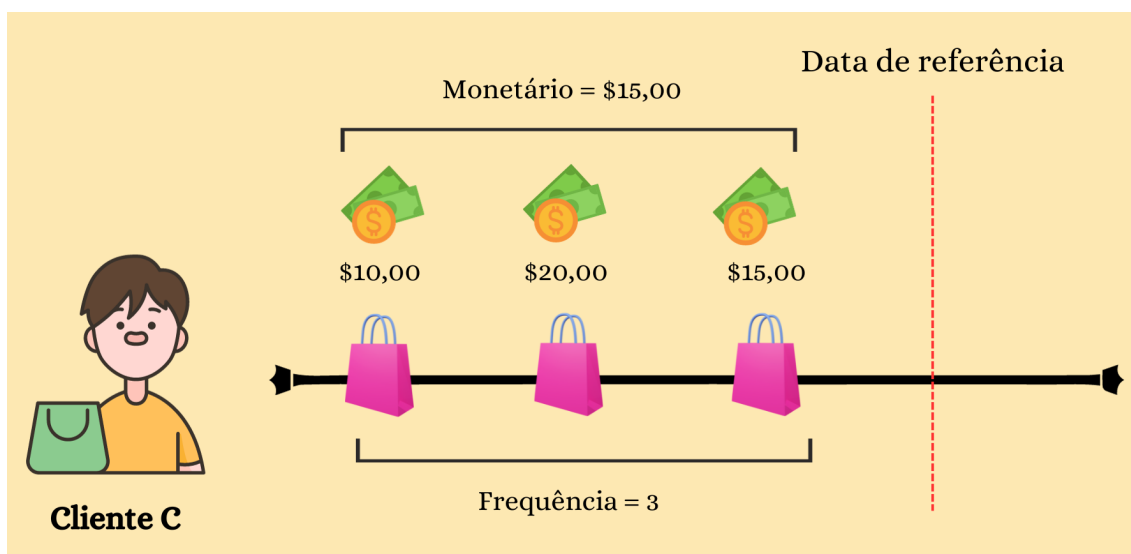


Figura 2. Ilustração de exemplo sobre a frequência e valor monetário.

O valor monetário representa o valor médio gasto pelo cliente em um determinado intervalo de tempo, conforme representado na Figura 2. Ou seja, é a soma do valor de todas as compras realizadas pelo cliente dividido pelo número de compras em um intervalo de tempo.

Neste estudo, a segmentação é realizada com base nas métricas de recência, frequência e valor monetário obtidas através do modelo RFM. Os clientes são classificados em oito grupos distintos com base nesses valores, conforme detalhado na Tabela 1. Para determinar se um valor é considerado alto ou baixo, utilizam-se as médias gerais de recência, frequência e valor monetário. Recência abaixo da média é classificada como “baixa”, enquanto recência acima da média é classificada como “alta”. Quanto à frequência e ao valor monetário, valores abaixo da média são classificados como “baixos”, e valores acima da média são classificados como “altos”. É importante destacar que a tabela foi elaborada com base no trabalho de [Sun et al., 2023]. No entanto, uma modificação foi feita em relação à representação do valor de recência, onde, ao contrário da abordagem original que considera “alto” como indicativo do melhor valor de recência, optamos por utilizar “baixo” para representar essa condição e vice-versa.

Tipo do Cliente	Recência	Frequência	Monetário	Segmento
Clientes importantes de alto valor	Baixo	Alto	Alto	1
Clientes importantes em desenvolvimento	Baixo	Baixo	Alto	2
Clientes importantes para proteger	Alto	Alto	Alto	3
Clientes importantes para retenção	Alto	Baixo	Alto	4
Clientes de valor geral	Baixo	Alto	Baixo	5
Clientes em desenvolvimento	Baixo	Baixo	Baixo	6
Clientes para retenção	Alto	Alto	Baixo	7
Clientes perdidos	Alto	Baixo	Baixo	8

Tabela 1. Segmentação dos clientes.

3.3. Algoritmos de aprendizado de máquina

Com os valores retornados após o processamento do RFM, é possível estimar o número de transações e o valor monetário dos clientes de acordo com um intervalo de tempo previamente estabelecido. Para isso foram selecionados sete algoritmos de aprendizado de máquina, sendo eles: *LassoCV*, *ElasticNet*, *Random Forest Regressor*, *Gradient Boosting Regressor*, *Histogram-based Gradient Boosting*, *Regression Tree* e *LightGBM*.

Em relação à otimização dos hiperparâmetros, foi utilizado o algoritmo *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 número 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 $N_{est} \in \{200, 800, 1000\}$. 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 $L_{Max} \in \{-1, 1, 2, 4\}$. Para o

Regression Tree, usa-se a profundidade máxima $P_{max} \in \{-1, 1, 2, 4\}$ e o número mínimo de folhas $L_{Min} \in \{1, 2, 4\}$. Já o para o *LightGBM* 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 $P_{max} \in \{3, 6, 10\}$.

No cenário A, para cada segmento é realizado este processo de escolha de algoritmo, visto que diferentes segmentos podem apresentar diferentes comportamentos, e no cenário B, este processo é feito por conjunto de dados.

3.4. Separação dos dados

Após realizar a análise exploratória, limpeza e processamento dos dados utilizando o modelo RFM, obtém-se um conjunto de dados com as seguintes características: identificador do cliente, data da compra e valor gasto na transação. Dado que se trata de uma série temporal, seguindo o trabalho apresentado por Vishwas and Patel [2020], opta-se por uma estratégia específica para a separação dos dados de treino e validação.

Para ambos os conjuntos de dados, são utilizados 80% para as datas mais antigas para o treinamento dos modelos e reservamos 20% para fins de teste. Um ponto relevante considerado foi o período de tempo; neste estudo, optamos por agrupar as transações semanalmente como unidade temporal.

3.5. Métricas de avaliação

Existem diversas métricas fundamentais para avaliar o desempenho de modelos de aprendizado de máquina. Neste trabalho iremos utilizar três: o Erro Quadrático Médio (*Mean Squared Error* ou *MSE*) e sua contraparte com raiz quadrada, a Raiz do Erro Quadrático Médio (*Root Mean Squared Error* ou *RMSE*). O MSE calcula a diferença média quadrática entre os valores preditos e reais, enquanto o RMSE aplica uma raiz quadrada a essa métrica para manter a mesma escala dos dados originais.

Outra métrica empregada é o Erro Médio em Escala Absoluta (*Mean Absolute Scaled Error* ou *MASE*). Sua fórmula inclui o Erro Médio Absoluto (MAE), que representa a média das diferenças absolutas entre valores observados e previstos. O denominador normaliza essa métrica, levando em consideração a média das diferenças absolutas entre observações consecutivas na série temporal. Esta métrica é particularmente útil ao lidar com dados de séries temporais visto que é independente de escala, o que significa que pode ser usado para comparar modelos em diferentes escalas ou unidades de medida [Franses, 2016].

3.6. Seleção do modelo para os segmentos

Para cada segmento analisado, foram aplicados os algoritmos de aprendizado de máquina previamente descritos, com o objetivo de identificar os resultados mais eficazes, conforme indicado pelas métricas de avaliação. Neste processo, utilizou-se a métrica MASE (*Mean Absolute Scaled Error*) como critério principal. Esta escolha baseia-se na robustez da MASE, que é especialmente adequada para avaliar a qualidade das previsões em modelos de séries temporais. A MASE apresenta vantagens significativas em contextos com variabilidade e sazonalidade distintas, proporcionando uma avaliação mais precisa e confiável do desempenho dos modelos. [SILVA, 2023].

4. Resultados e discussões

É importante destacar que, com o objetivo de simplificar e aprimorar a visualização dos resultados, os segmentos foram designados numericamente de 1 a 8, conforme mostrado na Tabela 1. A seguir serão apresentados os resultados obtidos em cada conjunto de dados visando estimar o número de transações e valor médio monetário.

4.1. Número de transações

É feita uma comparação entre os valores relativos ao número de transações para as métricas MSE, RMSE e MASE nos dois cenários distintos. O cenário A, que consiste na segmentação, tende a ter desempenho melhor ou semelhante ao treinamento com todos os dados em datasets variados, como no caso do *CDNOW* e *Retail I*. Porém, a eficácia de uma abordagem sobre a outra pode depender do dataset. O treinamento com todos os dados mostra-se mais eficaz em alguns casos, como no *Retail II*.

Transações										
Dataset	Cenário	Métrica	Segmentos							
			1	2	3	4	5	6	7	8
CDNOW	Cenário A	MSE	0,20	1,21	1,34	0,49	0,57	0,08	1,19	0,24
		RMSE	0,45	1,10	1,16	0,70	0,75	0,28	1,09	0,49
		MASE	0,63	1,32	0,86	0,81	0,66	0,68	0,88	0,69
	Cenário B	MSE	0,22	1,24	1,33	0,53	0,51	0,08	1,53	0,20
		RMSE	0,46	1,12	1,15	0,73	0,71	0,29	1,24	0,45
		MASE	1,11	1,06	2,49	1,09	1,28	0,22	2,31	0,91
Retail I	Cenário A	MSE	6,13	1,28	2,63	1,24	2,47	0,82	2,87	1,24
		RMSE	2,48	1,13	1,62	1,11	1,57	0,91	1,69	1,11
		MASE	1,19	1,45	1,04	1,19	0,84	1,39	1,19	1,67
	Cenário B	MSE	6,53	1,23	2,70	1,17	2,95	0,81	2,79	1,05
		RMSE	2,56	1,11	1,64	1,08	1,72	0,90	1,67	1,03
		MASE	2,21	1,22	1,75	1,18	1,70	0,85	1,84	1,12
Retail II	Cenário A	MSE	3,33	2,06	2,94	1,10	10,14	0,90	3,07	1,17
		RMSE	1,83	1,43	1,72	1,05	3,18	0,95	1,75	1,08
		MASE	0,93	1,36	1,12	1,62	3,07	1,51	1,24	1,66
	Cenário B	MSE	2,04	2,12	2,98	1,05	9,06	0,89	2,88	0,95
		RMSE	1,43	1,46	1,73	1,03	3,01	0,95	1,70	0,97
		MASE	1,65	1,48	1,91	1,12	3,21	0,89	1,79	1,00

Tabela 2. Comparativo entre abordagens em relação ao número esperado de transações.

Com base nos resultados e nas características de cada segmento de clientes, pode-se concluir que a eficácia da modelagem específica (Cenário A) varia significativamente de acordo com o perfil dos clientes.

Clientes de alto valor e alta frequência, como os dos segmentos 1 (Clientes importantes de alto valor) e 3 (Clientes importantes para proteger), apresentaram um desempenho superior quando utilizado a segmentação. Esses clientes têm baixo índice de recência e alto valor monetário, o que indica que suas transações são regulares e de grande importância para a empresa. A modelagem específica permite capturar melhor os padrões de comportamento desses clientes, resultando em estimativas mais precisas.

Por outro lado, segmentos com baixa frequência de transações, como os segmentos 2 (Clientes importantes em desenvolvimento) e 6 (Clientes em desenvolvimento), demonstraram piora significativa com a modelagem específica. Esses segmentos, caracteri-

zados por baixa frequência, se beneficiam mais de uma abordagem genérica que considera um conjunto mais amplo de dados.

Segmentos de clientes com alta frequência mas menor valor monetário, como o segmento 5 (Clientes de valor geral) e o segmento 7 (Clientes para retenção), também mostraram melhorias ao utilizar a modelagem específica. Esses clientes, embora tenham transações frequentes, não contribuem tanto em termos de valor monetário, mas a precisão nas estimativas é crucial para estratégias de retenção e otimização do relacionamento com o cliente.

Os clientes do segmento 8 (Clientes perdidos), que têm alta recência e baixa frequência e valor monetário, apresentaram resultados menos favoráveis com a modelagem específica, sugerindo que a abordagem genérica pode capturar melhor as poucas transações que esses clientes ainda realizam.

Portanto, a conclusão é que a segmentação deve ser aplicada estrategicamente. Clientes com alta frequência de transações, independentemente do valor monetário, tendem a se beneficiar mais dessa abordagem, pois seus padrões de compra são mais estáveis e previsíveis. Já os clientes em desenvolvimento ou com transações esporádicas podem obter melhores resultados com a modelagem genérica, que permite uma visão mais ampla e menos sensível às variações individuais. A segmentação cuidadosa e a escolha da abordagem de modelagem apropriada são essenciais para otimizar a precisão das estimativas de transações e, conseqüentemente, melhorar as estratégias de retenção e desenvolvimento de clientes.

4.2. Valor monetário

A análise dos resultados das predições de valores monetários revela diferenças significativas na eficácia das abordagens de modelagem quando adaptadas aos diferentes perfis de clientes como apresentados na Tabela 3.

		Monetário								
Dataset	Cenário	Métrica	Segmentos							
			1	2	3	4	5	6	7	8
CDNOW	Cenário A	MSE	451.00	3.557.39	148.335.81	66.275.48	176.51	1.935.31	2.866,58	7.609.54
		RMSE	21.24	59.64	385.14	257.44	13.29	43.99	53,54	87.23
		MASE	0.21	0.67	0.65		1,73	0.05	0,28	1,06
	Cenário B	MSE	5.313,53	4.687,79	175.128,14	66.768,76	652,35	1.956,17	2.688.38	7.680,40
		RMSE	72,89	68,47	418,48	258,40	25,54	44,23	51.85	87,64
		MASE	0,54	0,45	2,63	1.70	0,52	0.10	0.66	0.42
Retail I	Cenário A	MSE	2.497.89	629,65	6.379,71	16.805,34	54.72	16.428.600.00	790,60	399.43
		RMSE	49.98	25,09	79,87	129,64	7.40	4.053.22	28,12	19.99
		MASE	0.79	0.46	0.70	1,30	0,62	9,82	0,57	0,66
	Cenário B	MSE	2.852,42	602.75	6.325.89	16.751.63	104,29	16.428.650,00	307.89	448,50
		RMSE	53,41	24.55	79.54	129.43	10,21	4.053,23	17.55	21,18
		MASE	1,28	0,63	1,30	1.18	0.42	5.03	0.49	0.53
Retail II	Cenário A	MSE	1.823.062,00	5.155.03	14.985.53	3.969.19	128,71	1.584.11	1.231.74	740,01
		RMSE	1.350,21	71.80	122.42	63.00	11,34	39.80	35.10	27,20
		MASE	19,02	0.91	0.71	1,29	0,54	1,04	0,84	1,62
	Cenário B	MSE	1.796.301.00	5.167,05	15.747,17	4.005,53	110.63	1.587,42	1.232,76	733.37
		RMSE	1.340.26	71,88	125,49	63,29	10.52	39,84	35,11	27.08
		MASE	12.85	0,97	1,64	0.95	0.32	0.51	0.64	0.51

Tabela 3. Comparativo entre abordagens em relação ao valor monetário.

Para clientes do segmento 1 (Clientes importantes de alto valor) e do segmento 5 (Clientes de valor geral), que possuem alta frequência de transações e variam em termos de valor monetário, a modelagem específica mostrou um desempenho superior. Isso su-

gere que a abordagem do cenário A é capaz de capturar com maior precisão os padrões de comportamento financeiro, resultando em previsões mais acuradas.

Por outro lado, para o segmento 2 (clientes importantes em desenvolvimento) e 6 (clientes em desenvolvimento), que apresentam baixa frequência e valor monetário, a modelagem genérica do cenário B demonstrou ser mais eficaz. Esses segmentos, caracterizados por padrões de transação menos consistentes, se beneficiam de uma abordagem que considera um conjunto mais amplo de dados, ao invés de modelos específicos que podem não capturar adequadamente a variabilidade no comportamento financeiro desses clientes visto que possuem poucas quantidades de dados.

Segmentos como 3 (clientes importantes para proteger) e 4 (clientes importantes para retenção) apresentaram um desempenho mais variado. A eficácia da modelagem específica e genérica parece depender do contexto e do conjunto de dados utilizado, mostrando que a escolha da abordagem pode ser menos previsível para esses grupos, cujas características combinam alta recência com alta ou baixa frequência de transações.

Finalmente, para o segmento 8 (clientes perdidos), que têm alta recência, baixa frequência e baixo valor monetário, a modelagem genérica foi mais eficaz. Esse resultado sugere que uma abordagem mais abrangente pode ser mais adequada para prever o comportamento financeiro de clientes que estão menos ativos, capturando melhor as tendências gerais sem depender de características específicas que podem ser menos relevantes para esses clientes.

Em resumo, os resultados indicam que a eficácia da modelagem para previsões de valores monetários varia com o perfil dos clientes. A segmentação se torna mais adequada para clientes de alta frequência e valor monetário, enquanto modelos genéricos são preferíveis para clientes com baixa frequência e valor monetário.

5. Conclusão e Trabalhos Futuros

Em síntese, o presente estudo teve como objetivo segmentar os clientes e avaliar o impacto da criação de modelos distintos para cada segmento na previsão do número de transações e no valor monetário médio, comparativamente a um único modelo para toda a base de dados. Foram utilizados modelos de aprendizado de máquina consolidados na literatura e três conjuntos de dados que representam o comportamento de compra dos clientes.

Os resultados apontam para as seguintes conclusões: em geral, a abordagem de segmentação tem mostrado um desempenho superior em muitos casos, especialmente para dados de transações e em algumas situações nos dados monetários. A segmentação parece capturar melhor as particularidades dos diferentes grupos de clientes, resultando em melhores métricas de erro. A abordagem de segmentação se torna mais adequada para clientes de alta frequência e valor monetário, enquanto modelos genéricos são preferíveis para clientes com baixa frequência e valor monetário. Como o intuito do CLV foca nos clientes mais lucrativos, a abordagem de segmentação se torna mais efetiva neste contexto.

Estes resultados permitem um melhor entendimento sobre a relação entre a companhia e os clientes, auxiliando a determinar quanto investir na aquisição de novos clientes e na retenção dos mesmos. Além disso, a segmentação possibilita que a empresa direcione esforços personalizados para os clientes mais valiosos, aumentando a retenção e a satisfação desses clientes. Por fim, analisar estes clientes pode ajudar a identificar

quais produtos e serviços são mais valorizados por eles, auxiliando no desenvolvimento da companhia.

Em relação aos trabalhos futuros, sugere-se implementar novos modelos de aprendizado de máquina, implementar outros métodos de segmentação dos clientes e utilizar outras bases de dados com diferentes contextos. Também sugere-se a investigação aprofundada sobre outras diferentes estratégias de modelagem que podem consequentemente afetar o desempenho dos modelos.

Agradecimentos

As pessoas autoras gostariam de agradecer ao apoio do Manna Team⁴, da Fundação Araucária, da Softex, CNPq(Número 421548/2022-3) e a Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

Referências

- Online Retail. UCI Machine Learning Repository, 2015. DOI: <https://doi.org/10.24432/C5BW33>.
- Autores Omitidos para Revisão. Título Omitido para Revisão. In *Local de Publicação Omitido para Revisão*.
- Daqing Chen. Online Retail II. UCI Machine Learning Repository, 2019. DOI: <https://doi.org/10.24432/C5CG6D>.
- Xinqian Dai. Customer lifetime value analysis based on machine learning. In *Proceedings of the 6th International Conference on Information System and Data Mining, ICISDM '22*, page 13–17, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450396257. doi: 10.1145/3546157.3546160. URL <https://doi.org/10.1145/3546157.3546160>.
- Peter S Fader, Bruce GS Hardie, and Ka Lok Lee. Rfm and clv: Using iso-value curves for customer base analysis. *Journal of marketing research*, 42(4):415–430, 2005.
- Philip Hans Franses. A note on the mean absolute scaled error. *International Journal of Forecasting*, 32(1):20–22, 2016. ISSN 0169-2070. doi: <https://doi.org/10.1016/j.ijforecast.2015.03.008>. URL <https://www.sciencedirect.com/science/article/pii/S0169207015000448>.
- Ankit Kumar, K. Singh, Gaurav Kumar, Tanupriya Choudhury, and K. Kotecha. Customer lifetime value prediction: Using machine learning to forecast clv and enhance customer relationship management. *2023 7th International Symposium on Multi-disciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–7, 2023. doi: 10.1109/ISMSIT58785.2023.10304958.
- Edward Malthouse and Frank Mulhern. Understanding and using customer loyalty and customer value. *Journal of Relationship Marketing*, 6(3-4):59–86, 2008.
- Ina Maryani and Dwiza Riana. Clustering and profiling of customers using rfm for customer relationship management recommendations. In *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, pages 1–6. IEEE, 2017.
- Roland T Rust, Valarie Zeithaml, and Katherine N Lemon. *O valor do cliente: o modelo que está reformulando a estratégia corporativa*. Bookman, 2001.
- Adrian Sargeant. Using donor lifetime value to inform fundraising strategy. *Nonprofit Management and Leadership*, 12(1):25–38, 2001.

⁴<https://manna.team/>

- Arun Sharma. The metrics of relationships: measuring satisfaction, loyalty and profitability of relational customers. *Journal of Relationship Marketing*, 6(2):33–50, 2007.
- Raphael Albuquerque Xavier SILVA. Estudo de caso sobre aplicações de modelos clássicos de previsão de demanda para uma empresa de delivery de alimentos. B.S. thesis, 2023.
- Yuechi Sun, Haiyan Liu, and Yu Gao. Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. *Heliyon*, 9(2), 2023.
- B V Vishwas and Ashish Patel. *Hands-on Time Series Analysis with Python: From Basics to Bleeding Edge Techniques*. Apress, 2020.