2177-2576 versão online

Probabilidade de Recompra: Modelos de Machine Learning para Campanha de E-mail Marketing em E-Commerce

Autoria

NatÃ¡lia Cordeiro Zaniboni - nzaniboni@alumni.usp.br

Prog de Pós-Grad em Admin/Faculdade de Economia, Admin e Contab - PPGA/FEA/USP - Universidade de São Paulo

Pós Empresarial Inteligência de Mercado /ESPM - Esc Sup de Prop e MKT de São Paulo/Ass Esc Sup de Prop e MKT

Alessandra de Ã•vila Montini - amontini@usp.br

Prog de Pós-Grad em Admin/Faculdade de Economia, Admin e Contab - PPGA/FEA/USP - Universidade de São Paulo

Agradecimentos

A área de vendas das empresas deve acompanhar as transformações tecnológicas trazidas pela quarta revolução industrial. O crescimento da concorrência e a baixa taxa de respostas de campanhas de vendas, além do aumento de custos destas, aumentam a necessidade da identificação correta do público-alvo a ser alcançado. Modelos de Machine Learning podem indicar melhor este público e, neste trabalho, são propostos modelos de regressão logística e árvore de decisão para prever a probabilidade de recompra de clientes de um e-commerce brasileiro. Alguns fatores foram identificados como relevantes para melhorar a taxa de recompra, principalmente o valor dos produtos, o número de parcelas disponíveis e o local em que o cliente reside. Os modelos propostos, apesar de ter pequeno acréscimo na taxa de respostas, podem aumentar em até 64% o retorno financeiro de uma campanha.

Resumo

A área de vendas das empresas deve acompanhar as transformações tecnológicas trazidas pela quarta revolução industrial. O crescimento da concorrência e a baixa taxa de respostas de campanhas de vendas, além do aumento de custos destas, aumentam a necessidade da identificação correta do público-alvo a ser alcançado. Modelos de Machine Learning podem indicar melhor este público e, neste trabalho, são propostos modelos de regressão logística e árvore de decisão para prever a probabilidade de recompra de clientes de um e-commerce brasileiro. Alguns fatores foram identificados como relevantes para melhorar a taxa de recompra, principalmente o valor dos produtos, o número de parcelas disponíveis e o local em que o cliente reside. Os modelos propostos, apesar de ter pequeno acréscimo na taxa de respostas, podem aumentar em até 64% o retorno financeiro de uma campanha.

Probabilidade de Recompra: Modelos de *Machine Learning* para Campanha de E-mail Marketing em E-Commerce

Resumo

A área de vendas das empresas deve acompanhar as transformações tecnológicas trazidas pela quarta revolução industrial. O crescimento da concorrência e a baixa taxa de respostas de campanhas de vendas, além do aumento de custos destas, aumentam a necessidade da identificação correta do público-alvo a ser alcançado. Modelos de *Machine Learning* podem indicar melhor este público e, neste trabalho, são propostos modelos de regressão logística e árvore de decisão para prever a probabilidade de recompra de clientes de um e-commerce brasileiro. Alguns fatores foram identificados como relevantes para melhorar a taxa de recompra, principalmente o valor dos produtos, o número de parcelas disponíveis e o local em que o cliente reside. Os modelos propostos, apesar de ter pequeno acréscimo na taxa de respostas, podem aumentar em até 64% o retorno financeiro de uma campanha.

Palavras-chave: Recompra, *Machine Learning*, E-mail Marketing, Regressão Logística e Árvore de Decisão.

1. Introdução

A quarta revolução industrial traz transformações alimentadas pela digitalização, tecnologia da informação, robótica e inteligência artificial, e seu maior impacto se dá principalmente na tomada de decisão. A área de vendas sempre deve responder a alterações no ambiente, principalmente relacionadas à tecnologia. Com o advento Internet e bancos de dados, as informações se tornaram mais disponíveis e algumas das encomendas se afastaram de pedidos escritos para pedidos na internet, aumentando a incidência de e-commerces. Além disso, a inteligência artificial e o aprendizado de máquina representaram uma evolução nesta área ao automatizar a geração de listas de potenciais *prospects*, prever futuros cancelamentos e identificar clientes de maior lealdade (SYAM; SHARMA, 2018).

As baixas taxas de resposta e o aumento dos custos das campanhas de marketing direto levaram profissionais de marketing a usar técnicas de *Machine Learning* para prever o comportamento das respostas à estas campanhas. O termo *Machine Learning* representa a ciência de fazer os computadores agirem sem ser explicitamente programados (LEE, 1995) e é uma subárea da Inteligência Artificial. Os modelos de *Machine Learning*, em marketing, possuem o objetivo de identificar os clientes mais propensos a aceitar uma campanha de vendas, seja ao clicar no e-mail ou comprar o produto oferecido. O objetivo do uso do modelo é focar a campanha nestes potenciais clientes, reduzindo custos e melhorando a taxa de resposta das campanhas. Em geral, os modelos apresentam pequena melhora à estas taxas de respostas, porém os resultados financeiros podem apresentar aumentos significativos (DEICHMANN et al., 2002).

Este trabalho teve como objetivo propor modelos de *Machine Learning* para cálculo da probabilidade de recompra de um cliente de uma empresa de e-commerce. Foi possível identificar os fatores que afetam a escolha de recompra do cliente, e os principais foram número de parcelas, valor do produto e estado em que o cliente reside. Com a probabilidade obtida, foi possível estimar que o retorno financeiro de uma campanha feita com base em um modelo de *Machine Learning* pode ser até 64% maior que uma campanha que não utiliza este modelo.

2. Revisão Bibliográfica

A revisão bibliográfica deste trabalho foca em aplicações de *Machine Learning* na área de marketing, especialmente em *Customer Relationship Management* (CRM), e na análise da lealdade do cliente. A Tabela 1 apresenta um resumo dos principais trabalhos de aplicação de técnicas de *Machine Learning* aplicadas à CRM. Notou-se um maior uso de técnicas estatísticas mais tradicionais, como Regressão Logística e Árvore de Decisão. Com o avanço da capacidade computacional, outras técnicas começaram a ser utilizadas, como Redes Neurais, *Random Forest* ou *Support Vector Machines*.

Tabela 1. Aplicações de *Machine Learning* em CRM

Objetivo	Técnicas Utilizadas	Autores
Segmentação	*K-Means*	(HRUSCHKA; NATTER, 1999) (QADADEH; ABDALLAH, 2018) (PANI; SAHU; MAJUMDAR, 2020) (ANITHA; PATIL, 2020) (ZHOU; ZHAI; PANTELOUS, 2020)
	Redes Neurais	(HRUSCHKA; NATTER, 1999)
	Support vector clustering	(HUANG; TZENG; ONG, 2007)
Campanhas	Regressão Logística	(KNOTT; HAYES; NESLIN, 2002) (TEZINDE; SMITH; MURPHY, 2002) (DEICHMANN et al., 2002) (MCCARTY; HASTAK, 2007) (AHN et al., 2011) (COUSSEMENT; HARRIGAN; BENOIT, 2015) (WANG, 2020)
	Árvore de Decisão	(ANAND et al., 1998) (AHN et al., 2011) (COUSSEMENT; HARRIGAN; BENOIT, 2015)
	Redes Neurais	(KNOTT; HAYES; NESLIN, 2002) (AHN et al., 2011) (COUSSEMENT; HARRIGAN; BENOIT, 2015)
	Análise Discriminante	(KNOTT; HAYES; NESLIN, 2002) (COUSSEMENT; HARRIGAN; BENOIT, 2015)
Retenção (*Churn*)	Regressão Logística	(BUREZ; VAN DEN POEL, 2009) (BALLINGS; VAN DEN POEL, 2012) (COUSSEMENT; LESSMANN; VERSTRAETEN, 2017) (DE CAIGNY; COUSSEMENT; DE BOCK, 2018) (JAIN; KHUNTETA; SRIVASTAVA, 2020)
	Árvore de Decisão	(BALLINGS; VAN DEN POEL, 2012) (DE CAIGNY; COUSSEMENT; DE BOCK, 2018)
	Random Forest	(BUREZ; VAN DEN POEL, 2009) (IDRIS; RIZWAN; KHAN, 2012)

	Support Vector Machine	(CHEN; FAN; SUN, 2012)
	Ensemble (Bagging/Boosting)	(BALLINGS; VAN DEN POEL, 2012) (COUSSEMENT; DE BOCK, 2013) (JAIN; KHUNTETA; SRIVASTAVA, 2020)

A segmentação é uma atividade essencial para os profissionais de marketing. Na academia, a ideia de segmentação surgiu na década de 1950 como objetivo de visualizar um mercado heterogêneo como um número menor de mercados homogêneos em resposta a diferentes preferências de produtos (SMITH, 1956). Desde então, diversas técnicas de *Machine Learning* vêm sendo utilizadas para segmentação, sendo a principal delas o *K-Means Clustering*, para vários diversos fins, como segmentação de produtos (HRUSCHKA; NATTER, 1999; PANI; SAHU; MAJUMDAR, 2020) ou de clientes (ANITHA; PATIL, 2020; QADADEH; ABDALLAH, 2018; ZHOU; ZHAI; PANTELOUS, 2020). Com o avanço da capacidade computacional, outras técnicas começaram a ser utilizadas, como Redes Neurais (HRUSCHKA; NATTER, 1999) ou *Support vector clustering* (HUANG; TZENG; ONG, 2007). Todas com o objetivo de prestar um atendimento mais personalizado ao cliente para fidelização.

Os custos crescentes das campanhas de marketing direto, juntamente com as taxas de resposta baixas, levaram muitos profissionais de marketing a usar técnicas de *Machine Learning* para modelar o comportamento das respostas à estas campanhas. Mesmo com uma pequena melhora à estas taxas de respostas, os resultados financeiros podem ter aumentos significativos (DEICHMANN et al., 2002). Estes modelos, em geral, identificam os clientes mais prováveis de aceitar uma oferta, para focar as campanhas nestes clientes, aumentando a conversão e diminuindo custos.

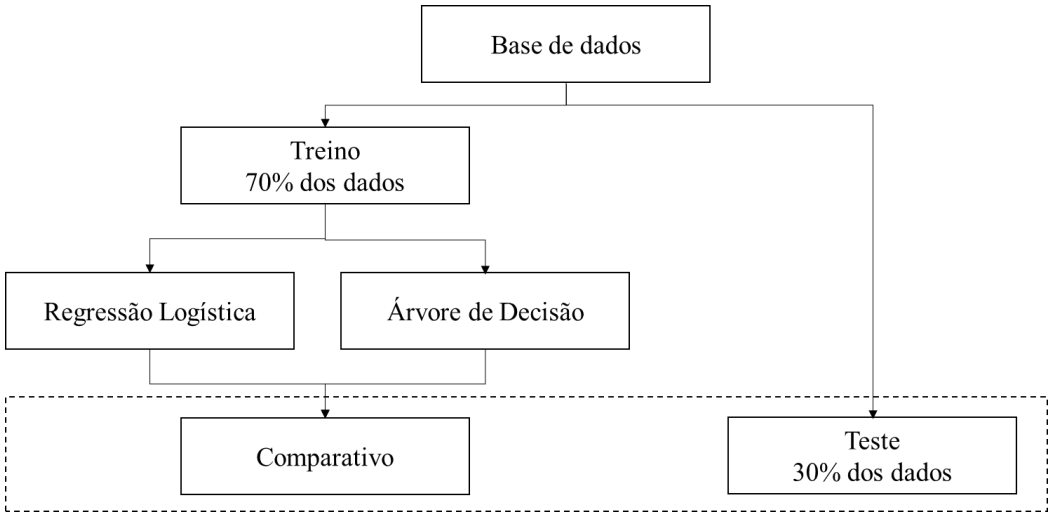
Foram identificadas aplicações de técnicas de *Machine Learning* para diferentes tipos de campanhas, como campanhas de *cross-selling*, em que se incentivam os clientes que já compraram o produto A para também comprar o Produto B, e são fundamentais para aumento da receita das empresas (AHN et al., 2011; ANAND et al., 1998; KNOTT; HAYES; NESLIN, 2002; YEH et al., 2010). Há também modelos para identificar clientes mais propensos a responder, clicar ou aceitar uma campanha de e-mail marketing (COUSSEMENT; HARRIGAN; BENOIT, 2015; DEICHMANN et al., 2002; MCCARTY; HASTAK, 2007; TEZINDE; SMITH; MURPHY, 2002) ou um anúncio em aplicativo (WANG, 2020).

Em mercados extremamente competitivos, a alta rotatividade dos clientes diminui a receita das empresas e gera elevados custos para adquirir novos clientes. Ao criar modelos de *Machine Learning* para prever os clientes que são mais prováveis a cancelar um produto (modelos de *Churn*), a empresa pode criar campanhas e promoções para retê-los. Notou-se diversas aplicações deste tipo de modelo na literatura, em setores de telecomunicações (COUSSEMENT; LESSMANN; VERSTRAETEN, 2017; IDRIS; RIZWAN; KHAN, 2012; JAIN; KHUNTETA; SRIVASTAVA, 2020), editorais (BALLINGS; VAN DEN POEL, 2012), varejo (CHEN; FAN; SUN, 2012) e jogos de azar (COUSSEMENT; DE BOCK, 2013). Alguns estudos testaram algoritmos em bases de dados de diversos segmentos, como De Caigny et al. (2018), que criaram modelos de regressão logística e árvore de decisão em 14 diferentes bases dos segmentos financeiro, varejo, editorial e telecomunicações. Burez e Van Den Poel (2009) estudaram o desbalanceamento de classes em modelos de *churn* aplicados aos segmentos financeiro, telecomunicações e supermercados.

3. Metodologia

O esquema metodológico do trabalho está apresentado na Figura 1. Primeiro selecionou-se uma base de dados do setor de e-commerce, que foi dividida em base de treino e teste. Esta divisão é necessária para a adequada avaliação do modelo e evitar possível *overffiting*. A escolha da proporção de 70% para treino e 30% para teste é a mais utilizada por equilibrar de forma adequada uma quantidade suficiente para testar o modelo sem perder capacidade preditiva (TREVIZAN et al., 2020). As técnicas de *Machine Learning* de Regressão Logística e Árvore de Decisão foram escolhidas por serem as mais utilizadas, conforme revisão bibliográfica, e apresentar interpretação mais fácil dos resultados. A qualidade destes modelos é mensurada e comparada ao final do trabalho.

Figura 1. Esquema metodológico



3.1. Base de dados

Foi utilizado um conjunto de dados público de comércio eletrônico brasileiro de pedidos feitos na Olist Store. O conjunto de dados possui informações de pedidos de 2016 a 2018 feitos em vários mercados no Brasil e contém informações do status do pedido, preço, pagamento, frete, localização do cliente, atributos do produto e, finalmente, revisões escritas pelos clientes. Os dados estão anonimizados.

3.2. Modelos de *Machine Learning*

O termo *Machine Learning* representa a ciência de fazer os computadores agirem sem ser explicitamente programados (LEE, 1995) e é uma subárea da Inteligência Artificial. Existem muitas técnicas que podem ser utilizadas para cálculo da probabilidade de ocorrência de um evento e, conforme análise da literatura, as técnicas de Regressão Logística e Árvore de Decisão são as mais utilizadas. Elas possuem boa acurácia, e a interpretabilidade destes modelos deve ser considerada na escolha (COUSSEMENT; HARRIGAN; BENOIT, 2015).

3.2.1. Regressão Logística

O primeiro modelo utilizado para explicar a probabilidade de $Y_i = 1$ é o modelo de regressão logística binária. A regressão logística binária consiste em relacionar uma variável resposta binária com variáveis explicativas, que podem ser categóricas, contínuas ou discretas.

Segundo Hosmer e Lemeshow (1989), a função ideal para modelar casos binários é a função logito. O modelo estima a probabilidade de recompra, que é apresentada na expressão (1)

$$\pi(x) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}, \quad (1)$$

em que β_0, β_1, ..., β_k são os parâmetros do modelo relacionados às k variáveis explicativas, $\pi(x)$ é a probabilidade estimada de recompra e X_k é a k-ésima variável explicativa.

Como os erros do modelo não possuem distribuição Normal e a variância não é constante, o método de estimação dos parâmetros deve ser o método da Máxima Verossimilhança. Os estimadores dos parâmetros obtidos pelo método de maximização da função de verossimilhança (Máxima Verossimilhança) são eficientes e produzem estimativas com distribuição Normal assintótica.

3.2.2. Árvore de Decisão

A árvore de decisão é uma metodologia que tem como objetivo classificar as observações em grupos baseado no seu perfil. A técnica estatística testa as possíveis decisões em cada fase (ou nó) e elege a melhor decisão baseada na capacidade de discriminação de acordo com o teste Qui-Quadrado (BREIMAN et al., 1984).

Uma árvore de decisão é construída de forma recursiva. Inicia-se com o conjunto de dados, e este é dividido com base em uma das variáveis independentes, formando um subconjunto homogêneo em relação à esta variável e ao resultado do levantamento. Esse processo é repetido até que se formem vários subconjuntos homogêneos com as decisões de levantamento semelhantes entre si.

O resultado da árvore é formado por uma sequência de decisões obtidas de acordo com a decisão de levantamento. A situação de jogo de maior importância para a decisão do levantamento compõe o primeiro nó. Os demais nós são formados sucessivamente de acordo com a importância da variável. Cada combinação de decisões da árvore resulta em uma probabilidade do evento do levantamento.

4. Resultados

A base de dados foi manipulada de forma que apresentasse unicidade de registros por cliente, pois o objetivo é verificar se o cliente apresentou uma recompra ou não. Ela contém 95.389 clientes. Foram excluídos 707 clientes que apresentavam valores *missings* em mais da metade das variáveis selecionadas.

A parte inicial do modelo é a *feature engineering*, que é o processo de extrair e criar variáveis de uma base de dados. Foram criadas 11 variáveis, além das originais, totalizando 25 variáveis para a criação dos modelos.

A variável resposta (Y) é a recompra. A Tabela 2 apresenta a distribuição de frequências da variável resposta, e nota-se que apenas 3,1% dos clientes fizeram uma recompra no site. Isto mostra que a base é desbalanceada, assim como citado em Idris et. al. (2012), sendo necessário o balanceamento dos dados para modelagem. A base de modelagem, portanto, tem 5.916 clientes, sendo 2.958 que não apresentaram recompra (selecionados aleatoriamente) e 2.958 que apresentaram recompra.

Tabela 2. Distribuição de frequências da variável resposta

Recompra	Frequência Absoluta	Frequência Relativa
Não apresentou	92.431	96,9%
Apresentou	2.958	3,1%

4.1. Regressão Logística

O modelo de Regressão Logística é apresentado na Tabela 3. Foi utilizada uma confiança de 90% para seleção das variáveis, e os fatores que se apresentaram relevantes para identificar os clientes com maior chance de recompra são relacionadas ao cliente (estado) e à última compra realizada pelo cliente (parcelas, quantidade de produtos, produto de maior valor, valor do frete e nota da avaliação dada pelo cliente). O modelo não apresenta multicolinearidade, que é uma premissa do modelo de regressão logística.

Tabela 3. Modelo de Regressão Logística

Variável	Categoria	Coeficiente	P-Valor	Relação com a Recompra
(Intercepto)		-0,33246	0,000562	-
Estado do cliente	MG, PR, SC, BA, entre outros	-0,14740	0,036990	Estados de taxa chance de recompra
	PE, CE, MA, entre outros	-0,54543	0,000872	Estados de menor chance de recompra
Número de parcelas		0,08519	0,00000000000166	Quanto maior o número de parcelas, maior a chance de recompra
Quantidade de produtos		0,24088	0,000129	Quanto maior a quantidade de produtos comprados, maior a chance de recompra
Produto de maior valor	Entre R$100 e R$300	-0,31785	0,00002100028228	Quanto maior o valor dos produtos, menor a chance de recompra
	Maior que R$300	-0,73513	0,00000051643341	
Frete	Maior que R$18	-0,16642	0,026940	Quanto maior o valor do frete, menor a chance de recompra
Nota de avaliação	Nota 5 (maior nota)	0,12895	0,046162	Se o cliente deu uma avaliação de nota 5, maior a chance de recompra

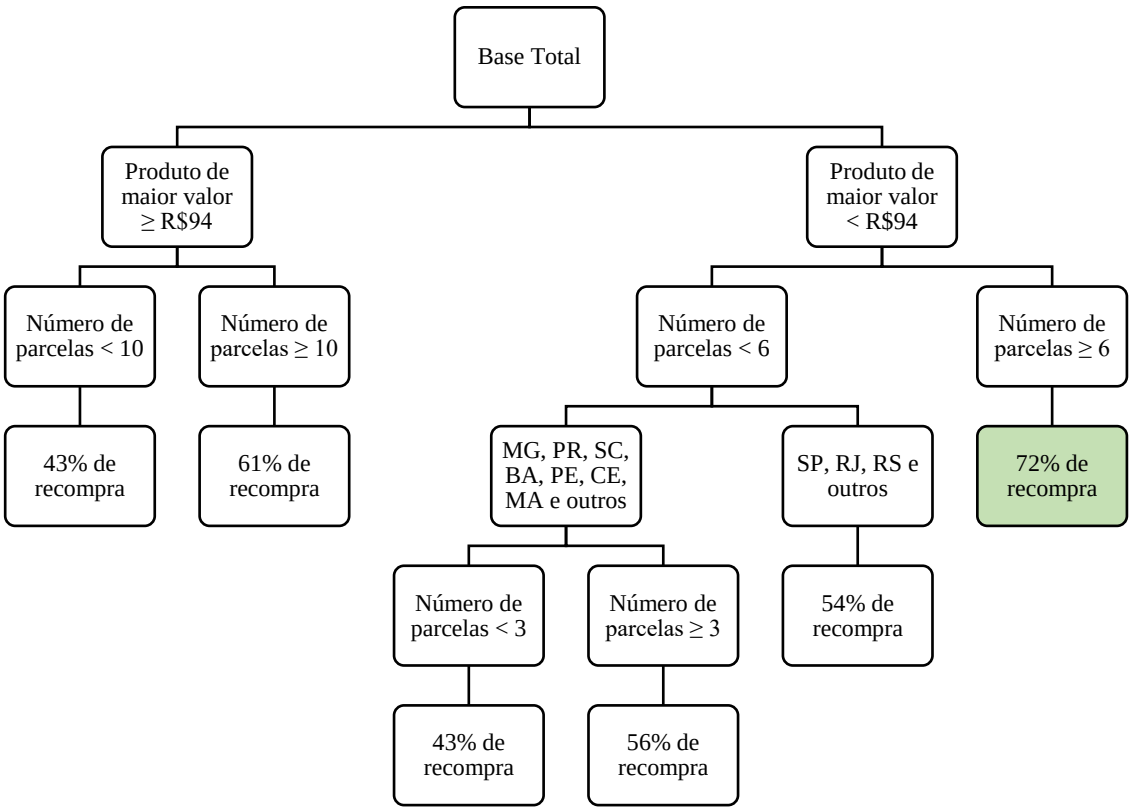
Nota-se que os estados de menor chance de compra são PE, CE, MA (entre outros – são apresentados os maiores estados de cada categoria na tabela). Os estados com maior chance de recompra são SP, RJ e RS.

Com relação às variáveis relacionadas à última compra do cliente, notou-se que, caso o e-commerce possibilite um parcelamento em um número maior de parcelas, a chance de recompra aumenta. Quanto maior a quantidade de produtos que o cliente comprou na sua última compra, maior a chance de recompra. Quanto maior o valor dos produtos e do frete, menor a chance de recompra. Se o cliente deu uma avaliação de nota 5, maior a chance de recompra.

4.2. Árvore de Decisão

O modelo de Árvore de Decisão é apresentado na Figura 2. Nota-se que o grupo que apresenta maior chance de recompra é dos clientes em que, em sua última compra, o produto comprado de maior valor é menor que R$94 e a compra foi parcelada em 6 ou mais parcelas. Outro ponto importante é analisar a recompra dos clientes que compraram produtos de valores maiores que R$94 reais. Caso este cliente tenha parcelado em 10 vezes ou mais, sua chance de recompra é maior.

Figura 2. Modelo de Árvore de Decisão



4.3. Comparativo

A Tabela 4 apresenta as medidas de qualidade do ajuste dos modelos de Regressão Logística e Árvore de Decisão. O percentual de classificação correta é calculado pela divisão do total de

clientes classificados corretamente pelo modelo e o total de clientes. O cálculo foi feito para a base de treino e base de teste. A base de teste apresenta qualidade menor que a base de treino, o que é esperado.

O retorno financeiro foi calculado da seguinte forma: Supôs-se que seria feita uma campanha de e-mail marketing para apenas 20% dos clientes da base de dados, que totalizam 19.078 dos 95.389 clientes, ao custo de R$0,50 por cliente. Uma campanha foi feita com a seleção aleatória de clientes, e a outra campanha foi realizada com os clientes de maior probabilidade de recompra, segundo os modelos de *Machine Learning*. Somou-se o valor dos produtos da recompra dos clientes atingidos por cada campanha. O retorno é o aumento, em percentual, do valor da recompra (menos os custos) comparado com a campanha aleatória.

Tabela 4. Medidas de qualidade do ajuste dos modelos

Medida	Regressão Logística	Árvore de Decisão
Percentual de classificação correta – base de treino	56,29%	57,23%
Percentual de classificação correta – base de teste	54,87%	53,24%
Retorno	64,27%	10,95%

Nota-se que o percentual de classificação correta não é muito alto, pois muitas vezes é esperado um percentual acima de 70% para um modelo de *Machine Learning*. Porém, mesmo com o pequeno ganho em termos de acerto, o ganho financeiro chega até a ser 64% maior em uma campanha utilizando um modelo de *Machine Learning*. O resultado está condizendo com o encontrado na literatura, em que há uma pequena melhora à taxas de respostas, porém os resultados financeiros possuem aumentos significativos (DEICHMANN et al., 2002).

5. Considerações Finais

Por meio dos modelos de *Machine Learning*, foi possível identificar os fatores que afetam a probabilidade de recompra do cliente da empresa de e-commerce. O valor do produto, o número de parcelas e o estado em que o cliente reside foram as principais variáveis dos modelos de Regressão Logística e Árvore de Decisão.

No geral, quanto maior o valor do produto, menor a chance de recompra. O valor do produto, ou da transação, foi identificado diversas vezes na literatura como um fator relevante para os modelos (ANAND et al., 1998; ANITHA; PATIL, 2020; PANI; SAHU; MAJUMDAR, 2020) No segmento do varejo e e-commerce, não houve um consenso na literatura da relação entre valor da compra e recompra. Em alguns estudos, compras de valores maiores apresentaram maior frequência de compras (ANITHA; PATIL, 2020). Em outros, compradores de maior frequência apresentaram menor ticket médio, assim como encontrado no presente estudo (ZHOU; ZHAI; PANTELOUS, 2020). Em muitas técnicas de *Machine Learning*, como Redes Neurais ou Ensemble, não é possível identificar esta relação, pois os modelos não são interpretáveis, e este fator foi considerado na escolha dos modelos deste trabalho.

Com relação a quantidade de parcelas, é identificado que o parcelamento é visto como um evento positivo, e o valor das parcelas não é percebido como uma dívida pelos consumidores, aumentando a chance de novas compras destes (MINIBAS-POUSSARD; BINGOL; ROLAND-LEVY, 2018).

O estado em que o cliente reside reflete, indiretamente, o rendimento ou potencial de compra do cliente. Clientes com potencial maior possuem maior frequência de compras, logo, maior chance de recompra (KNOTT; HAYES; NESLIN, 2002). Os estados como São Paulo e Rio de Janeiro possuem maiores probabilidades de recompra nos modelos.

Os modelos apresentados possuem pequeno incremento na qualidade de previsão, porém possuem um grande potencial de aumento na receita de campanhas de e-mail marketing, incrementando em até 64% a receita ao utilizar um modelo de *Machine Learning* na definição de um público-alvo.

Como sugestões de trabalhos futuros, pode-se utilizar outros métodos de balanceamento da base de dados com objetivo de não se perder tantos registros para o modelo e aumentar, possivelmente, o percentual de classificação correta deste. Também foram utilizadas apenas as variáveis do estado em que o cliente reside, porém pode-se utilizar variáveis com maior granularidade, como cidade ou até mesmo geolocalização.

Referências Bibliográficas

AHN, H. et al. Facilitating cross-selling in a mobile telecom market to develop customer classification model based on hybrid data mining techniques. **Expert Systems with Applications**, v. 38, n. 5, p. 5005–5012, 2011.

ANAND, S. S. et al. A data mining methodology for cross-sales. **Knowledge-Based Systems**, v. 10, n. 7, p. 449–461, 1998.

ANITHA, P.; PATIL, M. M. RFM model for customer purchase behavior using K-Means algorithm. **Journal of King Saud University - Computer and Information Sciences**, Article In press, 2020.

BALLINGS, M.; VAN DEN POEL, D. Customer event history for churn prediction: How long is long enough? **Expert Systems with Applications**, v. 39, n. 18, p. 13517–13522, 2012.

BREIMAN, L. et al. **Classification and regression trees**. 1. ed. Monterey: Chapman and Hall, 1984.

BUREZ, J.; VAN DEN POEL, D. Handling class imbalance in customer churn prediction. **Expert Systems with Applications**, v. 36, n. 3 PART 1, p. 4626–4636, 2009.

CHEN, Z. Y.; FAN, Z. P.; SUN, M. A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. **European Journal of Operational Research**, v. 223, n. 2, p. 461–472, 2012.

COUSSEMENT, K.; DE BOCK, K. W. Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. **Journal of Business Research**, v. 66, n. 9, p. 1629–1636, 2013.

COUSSEMENT, K.; HARRIGAN, P.; BENOIT, D. F. Improving direct mail targeting through customer response modeling. **Expert Systems with Applications**, v. 42, n. 22, p. 8403–8412, 2015.

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

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

DEICHMANN, J. et al. Application of multiple adaptive regression splines (mars) in direct response modeling. **Journal of Interactive Marketing**, v. 16, n. 4, p. 15–27, 2002.

HOSMER, D.; LEMESHOW, S. **Applied Logistic Regression**. 2. ed. Nova York: John Wiley & Sons, 1989. v. 6

HRUSCHKA, H.; NATTER, M. Comparing performance of feedforward neural nets and K-means for cluster-based market segmentation. **European Journal of Operational Research**, v. 114, p. 346–353, 1999.

HUANG, J. J.; TZENG, G. H.; ONG, C. S. Marketing segmentation using support vector clustering. **Expert Systems with Applications**, v. 32, n. 2, p. 313–317, 2007.

IDRIS, A.; RIZWAN, M.; KHAN, A. Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies. **Computers and Electrical Engineering**, v. 38, n. 6, p. 1808–1819, 2012.

JAIN, H.; KHUNTETA, A.; SRIVASTAVA, S. Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. **Procedia Computer Science**, v. 167, n. 2019, p. 101–112, 2020.

KNOTT, A.; HAYES, A.; NESLIN, S. A. Next-product-to-buy models for cross-selling applications. **Journal of Interactive Marketing**, v. 16, n. 3, p. 59–75, 2002.

LEE, J. A. N. Computer Pioneers. **The Institute of Electrical and Electronics Engineers, Inc., Los Alamitos, CA**, 1995.

MCCARTY, J. A.; HASTAK, M. Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. **Journal of Business Research**, v. 60, n. 6, p. 656–662, 2007.

MINIBAS-POUSSARD, J.; BINGOL, H. B.; ROLAND-LEVY, C. Behavioral control or income? An analysis of saving attitudes and behavior, credit card use and buying on installment. **Revue Europeenne de Psychologie Appliquee**, v. 68, n. 6, p. 205–214, 2018.

PANI, A.; SAHU, P. K.; MAJUMDAR, B. B. Expenditure-based segmentation of freight travel markets: Identifying the determinants of freight transport expenditure for developing marketing strategies. **Research in Transportation Business and Management**, n. October 2019, p. 100437, 2020.

QADADEH, W.; ABDALLAH, S. Customers Segmentation in the Insurance Company (TIC) Dataset. **Procedia Computer Science**, v. 144, p. 277–290, 2018.

SMITH, W. R. Product differentiation and market segmentation as alternative marketing strategies. **Journal of Marketing**, v. 21, n. 1, p. 3–8, 1956.

SYAM, N.; SHARMA, A. Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. **Industrial Marketing Management**, v. 69, n. December 2017, p. 135–146, 2018.

TEZINDE, T.; SMITH, B.; MURPHY, J. Getting permission: Exploring factors affecting permission marketing. **Journal of Interactive Marketing**, v. 16, n. 4, p. 28–36, 2002.

TREVIZAN, B. et al. A comparative evaluation of aggregation methods for machine learning over vertically partitioned data. **Expert Systems with Applications**, v. 152, p. 113406, 2020.

WANG, R. J. H. Branded mobile application adoption and customer engagement behavior. **Computers in Human Behavior**, v. 106, n. September 2019, p. 106245, 2020.

YEH, I. C. et al. Cosmetics purchasing behavior - An analysis using association reasoning neural networks. **Expert Systems with Applications**, v. 37, n. 10, p. 7219–7226, 2010.

ZHOU, J.; ZHAI, L.; PANTELOUS, A. A. Market segmentation using high-dimensional sparse consumers data. **Expert Systems with Applications**, v. 145, 2020.