Universidade Federal de Minas Gerais

Departamento de Estatística

Curso de Especialização em Estatística

Análise Estatística Aplicada ao Marketing: Estudo de Caso em *Marketing Mix Modeling*

Victor Gonçalves Campos Telles

Belo Horizonte

2024

Victor Gonçalves Campos Telles

Análise Estatística Aplicada ao Marketing: Estudo de Caso em *Marketing Mix Modeling*

Monografia apresentada ao Curso de especialização em Estatística da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Especialista em Estatística.

Área de concentração: Estatística

Orientador: Prof. Dr. Guilherme Lopes de Oliveira

Belo Horizonte

2025

Telles, Victor Gonçalves Campos.

T274a Análise estatística aplicada ao marketing: [recurso eletrônico] estudo de caso em Marketing Mix Modeling / Victor Gonçalves Campos Telles – 2025.
1 recurso online (49 f. il., color.) : pdf.

Orientador: Guilherme Lopes de Oliveira.

Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 43-45.

1. Estatística. 2. Análise de regressão. 3. Marketing – Aspectos financeiros. 4. Comércio eletrônico (Computação) (Computação) I. Oliveira, Guilherme Lopes de. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 337ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE VICTOR GONÇALVES CAMPOS TELLES.

Aos três dias do mês de fevereiro de 2025, às 13:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso do aluno **Victor Gonçalves Campos Telles**, intitulado: "Análise Estatística Aplicada ao Marketing: Estudo de Caso em *Marketing Mix Modeling"*, como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, o Presidente da Comissão, Professor Guilherme Lopes de Oliveira – Orientador, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado Aprovado condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 03 de fevereiro de 2025.

gov.br Documento assinado digitalmente
GUILHERME LOPES DE OLIVEIRA
Data: 04/02/2025 08:29:22-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Guilherme Lopes de Oliveira (orientador)
DECOM/CEFET-MG

MSc. Gabriel Oliveira Assunção
UFMG

gov.br Documento assinado digitalmente
GUILHERME AUGUSTO VELOSO
Data: 03/02/2025 22:13:04-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Guilherme Augusto Veloso
GET/UFF

Resumo

O presente trabalho explora a aplicação do *Marketing Mix Modeling* (*MMM*) como ferramenta estatística para avaliar o impacto de alavancas de marketing nos resultados de negócio, com foco em um e-commerce especializado. Foram utilizados métodos de regressão linear múltipla, regressão com regularização *lasso* e método *CART* (*classification and regression tree*), comparando a capacidade preditiva e interpretabilidade de cada técnica. Após a preparação e transformação dos dados, os modelos ajustados revelaram que variáveis relacionadas a preço, categorias de produto e investimento em anúncios são determinantes no desempenho da receita bruta. O estudo destaca a importância da parcimônia em modelos explicativos e da capacidade de simulação de cenários, evidenciando o potencial do *MMM* para orientar decisões estratégicas baseadas em dados.

Palavras-chave: marketing mix modeling; regressão linear múltipla; *lasso*; *cart*; análise de dados.

Abstract

This study explores the application of Marketing Mix Modeling (MMM) as a statistical tool to evaluate the impact of marketing levers on business outcomes, focusing on a specialized e-commerce platform. Multiple linear regression, Lasso regression, and CART methods were employed to compare predictive accuracy and interpretability. After data preparation and transformation, the adjusted models revealed that variables related to pricing, product categories and ad spending are significantly influence gross revenue performance. The study emphasizes the importance of parsimony in explanatory models and the ability to simulate scenarios, highlighting the potential of MMM to guide data-driven strategic decisions.

Keywords: marketing mix modeling; multiple linear regression; lasso; cart; data analysis.

Lista de Figuras

Lista de Abreviações

MMM – Marketing Mix Model
CMO – Chief Marketing Officer
CART – Classification and Regression Tree
ML – Machine Learning
ROI – Return Over Investiments
AIC – Akaike Information Criterio
LASSO - Least Absolute Shrinkage and Selection Operator
SKUs - Stock Keeping Unit
SLA – Service Level Aggrement
CRM – Customer Relationship Management
MSE – Mean Square Error
NPS – Net Promoter Score
GMV – Gross Merchandise Volume
SHAP – Shapley Additive Explanations
LIME – Local Interpretable Model-agnostic Explanations

Sumário

1. Introdução

Marketing é um ramo do conhecimento abrangente voltado para o estudo e desenvolvimento de metodologias para, por exemplo, atrair clientes, aumentar as vendas e maximizar as receitas de uma empresa. Em termos gerais, ele envolve uma série de ações e práticas que vão desde entender quem são os clientes até convencê-los de que um produto ou serviço atende suas necessidades. Segundo Kotler e Keller, dentre os ramos do marketing destacam-se: Pesquisa de Mercado e Segmentação de Público; Criação da Proposta de Valor; Estratégias de *Mixing* (4Ps do *Marketing*); Construção e Manutenção de Relacionamento com os Clientes e Análise de Resultados (Kotler & Keller, 2012).

Dando foco na parte do marketing que diz respeito às estratégias de *mixing*, também conhecido como 4Ps do marketing, é possível constatar o quão importante é quantificar o efeito de cada uma das quatro alavancas, ou cada um dos Ps (Produto, Preço, Praça e Promoção), no contexto produtivo de uma empresa. Por exemplo, alterar as características da embalagem, ou da comunicação sobre os benefícios do produto, pode alterar a quantidade de compradores. Ou, alterar preços dando descontos fortes em um período curto, tende a atrair mais clientes ou fazer com que os clientes usuais comprem em maior quantidade. Da mesma forma, promover o produto (Promoção nos 4Ps de marketing tem esse sentido) usando os canais mais adequados ao público-alvo e, integrados aos locais de compra (Praça), em tese, deveria aumentar a procura, e consequentemente, a receita da empresa que está aplicando as técnicas.

O professor emérito de marketing e publicidade, Neil H. Borden, foi a primeira pessoa a cunhar a expressão *Marketing Mix*, pois ele gostava de pensar na ideia de que um executivo de marketing teria a função de misturar ingredientes, do inglês "*mix ingredients*", de processos e políticas de marketing, com o objetivo de produzir lucro para as empresas (Borden, 1964). O referido artigo do Prof. Borden é tido como o texto base para o desenvolvimento do *marketing* moderno, e vai além de apenas introduzir a expressão, mas sim consolidar a ideia de que este ramo do conhecimento é composto por vários elementos estratégicos que podem ser ajustados para otimizar os resultados.

Borden identificou inicialmente 12 elementos que compunham o mix de *marketing*, incluindo fatores como produto, preço, marca, canais de distribuição, promoção, publicidade, pesquisa e planejamento, entre outros. A partir daí, ele argumentava que os profissionais de marketing podiam "misturar" esses elementos em proporções variáveis para melhor atender as necessidades do mercado e atingir os objetivos da empresa (Borden, 1964). Na concepção do Prof. Borden, estes ajustes de mix deveriam ser realizados de acordo com o mercado, considerando fatores internos e externos, visando encontrar a "combinação ótima" que atendesse aos interesses dos consumidores e atingisse os objetivos da companhia.

A estrutura proposta em Borden, N. H. (1964) serviu como fundação do marketing analítico e o planejamento de *marketing*. Com o passar do tempo, os 12 elementos se transformaram concisamente no que hoje se referencia como os 4Ps do *marketing*. A partir

da ideia de Borden, em conjunto com técnicas estatísticas e a capacidade de processar grandes massas de dados, décadas depois surgiu o conceito de *Marketing Mix Modeling* (*MMM*), que parte do uso de dados históricos, para quantificar o impacto de cada alavanca de negócio (4Ps do marketing), e por fim, ser capaz de trazer um cenário ótimo, ou capacidade de simular cenários.

De fato, o *MMM* tem sido aplicado em diversos contextos do ramo produtivo, para entender melhor as relações entre os canais de marketing e a métrica de negócio (receita bruta, por exemplo); para distinguir canais de marketing com alto *ROI* (*return over investment*, em português, retorno sobre o investimento) dos com baixo *ROI* e, finalmente, otimizar melhor seu orçamento de *marketing*; e por fim, fazer previsões de conversões futuras com base em entradas fornecidas (Shin, 2021). Portando, em ramos da indústria que sejam altamente dependentes de investimentos significativos em publicidade e campanhas promocionais, e frequentemente lidam com dados históricos complexos e necessitam entender o impacto de fatores como sazonalidade e variáveis externas no comportamento de consume, o *MMM* se torna ferramenta indispensável para uma gestão moderna e baseada em dados, proporcionando *insights* quantitativos poderosos na otimização de alocação de recursos (Hanssens, Parsons & Schultz, 2003).

Neste contexto, mostra-se interessante o aprofundamento e estudo de técnicas quantitativas que visam estimar o quanto cada uma das alavancas de negócio impactam na receita de uma empresa. Com esse objetivo, neste trabalho as técnicas envolvidas no *Marketing Mix Modeling* serão aplicadas em um conjunto de dados de uma empresa do ramo de varejo, bens de consumo, *e-commerce*, telecomunicações e serviços financeiros (Hanssens, Parsons & Schultz, 2003). Mais especificamente, tendo em vista o planejamento estratégico de determinado um período, conhecendo-se o orçamento disponível, e conhecendo o cenário atual de uma empresa, o objetivo prático de interesse na rotina de um analista de marketing da empresa é responder à pergunta: vale mais a pena fazer uma forte baixa nos preços ou investir em um anúncio na televisão ou no patrocínio de um grande evento? Ao longo das análises, discutiremos como os dados históricos da empresa podem ser usados para auxiliar a tomada de decisão dos gestores da empresa.

A sequência do trabalho está organizada da seguinte maneira. O Capítulo 2 traz um panorama dos métodos aplicados em *MMM*. No Capítulo 3 são apresentados os dados selecionados para o estudo, bem como os resultados da seleção e manipulação de variáveis (*Feature Engeneering*). O Capítulo 4 traz os resultados da aplicação do *MMM*. Finalmente, algumas discussões e considerações finais são apresentadas no Capítulo 5.

2. *Marketing Mix Modeling* (*MMM*)

O *Marketing Mix Modeling* (MMM) é uma análise estatística, usando modelos de regressão múltipla em dados de séries temporais de vendas, com a finalidade de estimar o impacto das várias táticas de *marketing* (*marketing* mix) nas vendas e, então, prever o impacto de conjuntos futuros de táticas. O propósito de usar *MMM* é entender o quanto cada alavanca de marketing contribui para as vendas e quanto gastar em cada entrada de marketing (Kumar, 2017). Outros autores como Kübler, R. (2023) e Jiang, A. (2022) sugerem uma abordagem beysiana ao tema, enquanto Wigren, R. & Cornell, F. (2020) fizeram uma análise aplicando diversos métodos ao problema do *Marketing Mix Modeling*.

As técnicas de *MMM* são amplamente utilizadas para explorar as relações entre variáveis que descrevem os 4P's do *marketing* e uma variável resposta relacionada a vendas (quantidade, volume, compradores etc.), como exemplificado na série temporal (linha sólida) de quantidade de vendas diárias (*Daily Sales*) de um produto apresentada na Figura 1. Uma segunda aplicação, após a exploração e entendimento das relações existentes entre cada componente e a variável resposta, o *MMM* passa a ser usado como fonte de simulação de cenários, porque periodicamente as empresas passam por revisões de orçamento ou necessitam reagir a algum movimento de mercado. Desta forma, a precisão e a interpretabilidade dos resultados são requisitos fundamentais de um modelo de mix de *marketing* ajustado.

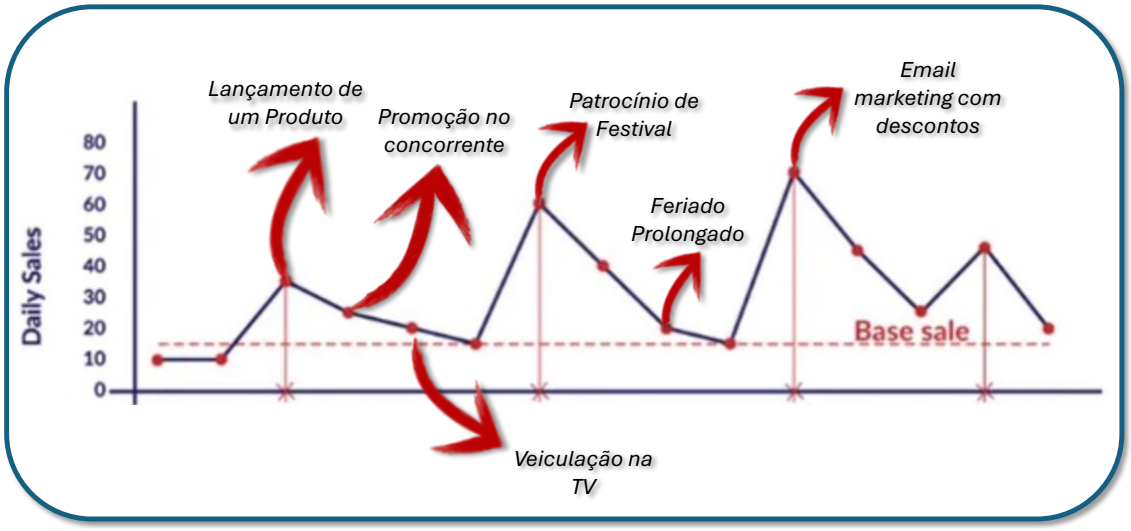

Figura 1 - Exemplo visual de eventos que impactam as vendas de uma empresa. Adaptado de Rahul (2019).

Tendo em vista os pontos apresentados acima, a técnica de regressão é amplamente reconhecida como a base estatística do Marketing Mix Modeling. Ela é utilizada para modelar a relação entre variáveis de *marketing*, como gastos com mídia, e vendas, permitindo às empresas avaliar o impacto de diferentes canais e estratégias. Segundo Hanssens, Parsons e Schultz (2003), a regressão, em suas formas linear e não-linear, é essencial para estimar a resposta de mercado a esforços publicitários e de promoção. A regressão linear múltipla é a técnica mais simples e frequentemente usada no *MMM*. Ela permite calcular os efeitos

médios das variáveis explicativas (por exemplo: TV, mídia digital e sazonalidade) sobre a variável resposta. No entanto, suas suposições, como linearidade e independência dos erros, podem ser desafiadas quando lidamos com séries temporais, como em campanhas de *marketing* contínuas (Hanssens, 2003).

Outra técnica, de modelos Bayesianos, que oferecem uma abordagem robusta para incorporar incertezas e informações prévias (*priors*), permitindo maior flexibilidade, melhor capacidade de modelar efeitos de saturação, e especialmente em contextos de dados limitados, também são comuns de serem utilizadas como ferramenta para modelagem de *MMM*. Especificamente para modelar efeitos decrescentes de gastos publicitários, transformações ou modelos não lineares são frequentemente utilizados. Esses modelos ajudam a representar melhor esta redução do retorno no longo prazo (*Diminishing Return*) (Jiang, 2022).

O método *CART* (*Classification and Regression Trees*) também tem aplicação no contexto de *Marketing Mix Modeling* (*MMM*), embora seja menos comum em comparação com abordagens estatísticas tradicionais, como regressões lineares ou modelos Bayesianos. Esta técnica tem tido aplicações especificas como: identificação de interações complexas, em outras palavras, pode ajudar a revelar que campanhas digitais e televisivas têm impacto sinérgico quando combinadas. Outra aplicação específica: modelagem de efeitos não lineares como retorno decrescente e saturação; e por fim segmentação de cliente, sendo possível primeiro segmentar os tipos de clientes e para cada tipo ter um *MMM* diferente (Hastie et al, 2009).

Por outro lado, técnicas modernas de aprendizado de máquina (em inglês, *machine learning* (*ML*)) ainda enfrentam alguns desafios que limitam sua adoção no campo do *Marketing Mix Modeling*. O objetivo principal do *MMM*, segundo Shmueli, G. (2010), é fornecer *insights* acionáveis, como *ROI* de canais específicos e modelos de *ML* são frequentemente considerados "caixas-pretas" e dificultam a explicação de como variáveis específicas afetam os resultados (Hanssens et al, 2003). Entretato, atualmente a interpretabilidade de modelos como redes neurais e *boosting* estão evoluindo significativamente com o desenvolvimento de métodos *pos-hoc* baseados em teoria dos jogos e análise de contribuição local. Técnicas como *SHAP* (*Shapley Additive Explanations*) (Lundberg & Lee, 2017) e *LIME* (*Local Interpretable Model-agnostic Explanations*) (Ribeiro et al., 2016) permitem quantificar a influência de cada variável nas predições, mesmo em modelos complexos. Essas abordagens traduzem a lógica não linear dos algoritmos em valores aditivos, tornando possível identificar padrões como interações entre variáveis ou efeitos de saturação, por exemplo. Com o avanço da popularização destas técnicas, é esperado que em breve, tenha-se mais influência de ML em modelos de mix de marketing.

Muitas técnicas de *ML* não foram projetadas para lidar com a complexidade de séries temporais com dependências de longo prazo, um elemento essencial no *MMM*. Técnicas estatísticas tradicionais, como regressões lineares ou Bayesianas, se adaptam melhor a essas exigências de acordo com Hastie et al., (2009). Além disto, o autor evidencia outro ponto de preocupação: modelos de aprendizados de máquina mais complexos correm maior risco de se ajustarem demais aos dados disponíveis, especialmente em conjuntos de dados pequenos, prejudicando sua generalização para campanhas futuras (Hastie et al, 2009).

2.1 Modelagem via Regressão Linear

Focando especificamente na análise de regressão linear múltipla, pode-se afirmar que este tipo de técnica serve como uma ponte entre parâmetros como gastos com marketing, preço e descontos, produtos em estoque, e vendas alcançadas, eventualmente maximizando o retorno sobre o investimento em todas as atividades de *marketing*. A modelagem de mix de *marketing* funciona segregando métricas de negócios em partes para analisar melhor os 4Ps do marketing, assim ajuda a distinguir atividades de marketing e promocionais, ou seja, drivers incrementais de drivers básicos e outros. Com uma solução *MMM* ideal, é possível medir o impacto de atividades de *marketing* individuais em receitas, volumes e preços de produtos vendidos no mercado (Jiang 2022).

De modo genérico, a ideia básica é desenvolver um modelo para predizer uma variável resposta Y_t em função de um conjunto de p covariáveis relacionadas às estratégias de marketing e investimentos da empresa anúncios (X_{1t}, X_{2t}, …, X_{pt}) como em um modelo de regressão linear múltipla da forma (Hegewald, 2019):

$$Y_t = \beta_0 + \sum_{i=1}^{p} (\beta_i * X_{it}) + \varepsilon_t$$

onde t=1, …, T indexa o tempo, sendo T o último período disponível para a análise; $\beta 0, \beta 1, \ldots, \beta p$ são os parâmetros do modelo e fornecem uma quantificação dos efeitos de interesse para a respectiva variável; e representa o termo de erro aleatório para o qual se assume independência para todo t e que, além disso, seguem uma distribuição Normal com média 0 e variância σ^2 fixa para todo t (homocedasticidade).

As suposições paramétricas feitas sobre o termo de erro do modelo permitem a definição de testes de hipóteses para verificar a significância do modelo, bem como do efeito individual de cada variável. A estimação dos parâmetros é feita, em geral, usando o método de mínimos quadrados que também gera uma estimativa para a variabilidade σ^2 do erro aleatório. Além disso, através da partição da variabilidade total da resposta Y entre a parte explicada pelo modelo e a parte explicada pelo erro surge uma medida típica de avaliação da qualidade do ajuste: o coeficiente de determinação R^2 que apresenta valores entre 0 e 1; sendo que quanto mais próximo de 1, melhor o ajuste. Mais detalhes sobre a metodologia podem ser encontrados, por exemplo, em Montegomery e Runger (2009).

A regressão *Lasso* (*Least Absolute Shrinkage and Selection Operator*) é uma técnica de aprendizado de máquina utilizada para realizar regressão linear, incorporando regularização para melhorar a interpretabilidade e o desempenho do modelo. Sua principal característica é a adição de um termo de penalização baseado na norma L_1 dos coeficientes à função de custo, o que resulta em coeficientes mais próximos de zero e, frequentemente, em alguns sendo exatamente zero. Essa penalização induz a seleção automática de variáveis, tornando o Lasso particularmente eficaz em cenários de alta dimensionalidade, onde há muitas variáveis preditoras (Hastie, Tibshirani & Friedman, 2009).

Neste trabalho, os ajustes dos modelos de regressão linear foram realizados com função base do *software* R (R Core Team, 2024) e a regressão *Lasso* foi ajustada com auxílio do pacote glmnet (Friedman et al., 2010) do R.

2.2 Modelagem via CART

O modelo *CART* (*classification and regression tree*) é uma abordagem não-paramétrica que segmenta o espaço de soluções em regiões homogêneas, com base em regras de decisão. Ao contrário da regressão linear, que assume uma relação linear entre as variáveis explicativas e a variável resposta, o *CART* identifica interações complexas e efeitos não lineares de maneira automática (Wigren, R. & Cornell, F., 2020).

O modelo resultante pode ser interpretado como uma árvore de decisão, onde cada nó representa uma variável explicativa e cada ramo indica uma condição que separa os dados em diferentes grupos. Os nós terminais (folhas) fornecem a previsão da receita bruta para cada segmento identificado. Para a árvore elaborada com a base de dados disponível, definiu-se os seguintes parâmetros de controle:

- *Minsplit* - Número mínimo de observações por Nós: 30
- *XVAL* – Quantidade de validações cruzadas: 1000
- *CP* – Parâmetro de complexidade: 0.00000354624, primeiro foi realizado a validação cruzada para entender qual o CP ótimo dado o xval utilizado
- *Maxdepth* – Ajusta o tamanho máximo de profundidade da árvore final: 30
- *Method* – Método de ajuste da árvore: *ANOVA*

Neste trabalho, o método CART dou aplicado usando o pacote rpart (Therneau & Atkinson, 2023) e rpart.plot (Milborrow, 2024) do *software* R (R Core Team, 2024).

2.3 Critérios de Avaliação dos Modelos

No estudo, a avaliação dos modelos foi baseada em três critérios quantitativos principais: Menor Erro Quadrático Médio (*MSE*), Maior R^2 Preditivo e Significância Estatística (p-valor). O *MSE* foi utilizado para medir a precisão das previsões, penalizando erros maiores de forma quadrática, enquanto o R^2 preditivo, calculado no conjunto de teste, avaliou a proporção da variabilidade da variável resposta explicada pelo modelo. Além disso, o p-valor foi empregado para verificar a significância estatística das variáveis, garantindo que apenas preditores relevantes fossem incluídos. Complementarmente, uma análise qualitativa foi realizada para avaliar a interpretabilidade e a representatividade das variáveis selecionadas, considerando sua aplicabilidade prática em cenários de negócio. Essa abordagem combinada permitiu equilibrar a capacidade preditiva dos modelos com a clareza interpretativa, essencial para a tomada de decisões estratégicas em *Marketing Mix Modeling* (Shmueli, 2010).

2.4 Utilização Prática dos Resultados da Modelagem

Elaborando de maneira adequada a modelagem, pode-se obter uma ferramenta com quatro boas funções analíticas, segundo Jiang, A. (2022):

1. Simular análise de hipóteses: O *MMM* permite que os *CMOs* (*chief marketing officer*, em português, diretor de marketing) prevejam os potenciais efeitos de receita das ações de marketing especificadas e fornece *insights* sobre quais decisões são baseadas. A análise de hipóteses tem como objetivo fornecer às empresas *insights* sobre potenciais consequências da implementação de mudanças ou estratégias específicas antes de realmente executá-las. Ao conduzir essas simulações, os profissionais de *marketing* podem tomar decisões mais informadas e mitigar os riscos associados a iniciativas não testadas.

2. Identifica as alavancas de desempenho: A modelagem de mix de *marketing* capacita as empresas a identificar quais foram os fatores que tornaram a campanha ou canal específico bem-sucedido. Também pode ser usada para desvendar as discrepâncias e obstáculos que podem estar dificultando uma campanha ou canal específico. Ao analisar dados históricos, os profissionais de *marketing* podem identificar os elementos que têm o maior impacto nas vendas e nos resultados comerciais.

3. Descobrindo sinergias e compensações: Através do *MMM*, nota-se a interação e as compensações entre diferentes variáveis de *marketing*. Ela ajuda as empresas a entender como as mudanças em um aspecto do mix de *marketing* podem impactar outros. Por exemplo, aumentar os gastos com publicidade pode levar a maiores vendas, mas às custas das margens de lucro. Essa compreensão de sinergias e compensações permite que as empresas façam escolhas informadas e encontrem um equilíbrio entre ganhos de curto prazo e sustentabilidade de longo prazo.

4. Quantifica o impacto das variáveis: A modelagem de mix de marketing decompõe a receita total em receita base e receita incremental. Em seguida, identifica os fatores que afetam a receita incremental e quantifica seu impacto individual. Também compara o *ROI* para diferentes canais no orçamento de *marketing*. Por meio de técnicas de análise estatística, a modelagem de mix de marketing mede a relação entre variáveis de *marketing* e receita. O modelo permite que os profissionais de *marketing* determinem até que ponto cada elemento de *marketing* contribui para o crescimento da receita.

E toda essa modelagem necessita de uma base de dados robusta o suficiente. Na maioria das vezes, devido à elevada autocorrelação das observações de uma variável ao longo do tempo e à correlação entre variáveis que medem efeitos similares, é preciso realizar uma seleção prévia minuciosa de quais efeitos devem entrar ou não no modelo. Isto tem o proposito de adequar às restrições e suposições típicas dos modelos de regressão aplicados, bem como evitar problemas de adequação dos modelos.

O processo de seleção e manipulação de variáveis (*feature engeneering*) envolve, muitas das vezes, a aplicação de técnicas de interpolação para preencher possíveis *missings* (dados faltantes) em longo da série temporal de observações de uma variável ou para se obter

medições em uma unidade temporal diferente da original, por exemplo, gerar dados semanais para dados originalmente medidos em escala mensal. Além disso, técnicas de redução de dimensionalidade, como análise fatorial, podem ser aplicadas para agregar variáveis fortemente correlacionadas e que poderiam levar a efeitos de multicolinearidade na modelagem. Mais detalhes sobre o processo de *feature engeneering* são apresentados no próximo capítulo com foco nos dados a serem abordados no estudo.

3. Definição dos Dados e *Feature Engeneering*

Na estatística aplicada, a definição precisa dos dados é crucial para a validade e a interpretabilidade dos resultados. Dados são informações coletadas sobre fenômenos, eventos ou objetos, e a primeira etapa em qualquer análise estatística é a identificação do tipo de dado que será utilizado, pois isso determina as técnicas e métodos a serem aplicados. Segundo Creswell e Creswell (2017), a clareza na definição dos dados é fundamental para garantir a rigorosidade dos métodos de pesquisa.

Além da natureza dos dados, é importante considerar sua fonte e método de coleta, pois isso pode impactar a qualidade e a confiabilidade das informações. Dados podem ser coletados por meio de pesquisas, experimentos, observações ou registros administrativos, e cada método possui suas vantagens e desvantagens. De acordo com Babbie (2016), a escolha adequada do método de coleta de dados é essencial para a obtenção de resultados significativos e representativos.

Em resumo, a definição dos dados é um passo fundamental que influencia toda a análise estatística, desde a escolha das ferramentas até a interpretação dos resultados. A compreensão adequada do tipo de dado em questão permite que os pesquisadores realizem análises mais precisas e tomadas de decisão informadas.

3.1 Base de dados bruta

Como uma breve contextualização, o desafio proposto pelo autor era aprofundar o estudo de modelos de mix de marketing devido seu histórico e desenvolvimento profissional. Entretanto, utilizar informações estratégicas de um ambiente altamente regulado de uma empresa em particular é algo fora de cogitação. Com estas restrições, o autor realizou pesquisas em plataformas especializadas em base de dados, como *Kaggle* e *GitHub*, e encontrou alternativas para enveredar no aprofundamento desejado.

Na plataforma *Kaggle* (https://www.kaggle.com/) um conjunto de bases de dados promissoras foi encontrado, e pode ser acessado pelo link: https://www.kaggle.com/datasets/datatattle/dt-mart-market-mix-modeling. Este conjunto de bases de dados diz respeito a um *e-commerce*, e é composto diferentes planilhas que contêm: (i) registros de investimentos realizados por tal *e-commerce* ao longo de 12 meses e com granularidade mensal, contendo o histórico do *NPS* (*net promoter score,* que em português significa pesquisa de satisfação do cliente) e do valor das ações no fechamento do mês; (ii) a hierarquia dos produtos vendidos neste *e-commerce*; (iii) as datas comemorativas do lugar onde este *e-commerce* opera; e (iv) dados das transações diárias ao longo de mais de 12 meses, onde constam dados do cupom fiscal (por exemplo, identificador do cliente, os itens comprados, o valor do item, o valor de desconto aplicado, o valor efetivamente recebido pela loja, se o produto havia no estoque, o método de pagamento, dentre outras informações transacionais).

Após algumas análises preliminares, foi constatado que a base de dados transacionais continha tratamentos não explicados na plataforma e não era uma base crua. Desta forma o autor realizou novas buscas, para encontrar o conjunto original, sem tratamentos. Tal fonte de dados foi encontrada no *GitHub*, https://github.com/palitr/Budget-Optimization-in-

Ecommerce-using-Market-Mix-Modelling/blob/master/Problem%20Statement.txt. A partir da melhor contextualização apresentada nesta fonte de dados, ficou definido que o *ecommerce* ao qual os dados se referem é o *ElecKart*, especializado em produtos eletrônicos e que fica localizado no Canadá. Para o alcance dos objetivos deste trabalho, tal fonte de dados foi validada, e usada para aprofundar o conhecimento em modelos aplicados ao mix de *marketing*.

Em suma, foi realizado o *download* de quatro bases de dados originais, a saber:

- ConsumerElectronics.csv: contém 20 colunas, tem a granularidade de item pedido, ou seja, detalha os *SKUs* (*stock keeping unit,* que em português significa Unidade de Manutenção de Estoque. É um código único que identifica produtos) contidos nos pedidos realizados ao longo do período analisado;
- MediaInvestment.csv: contém 12 colunas sobre os níveis de investimentos mensais em diferentes canais de comunicação (por exemplo, TV, rádio e mídia online) com o objetivo de atrair clientes, e obviamente vender mais;
- SpecialSale.csv: contém apenas 2 colunas e tem a granularidade de dia, com informações sobre datas comemorativas e sazonais para o comércio local;
- MonthlyNPSScore.csv: possui 3 colunas, e trás a posição de fechamento mensal das informações nos meses de referência.

De forma visual, o relacionamento entre os conjuntos de dados disponíveis e utilizados neste trabalho, pode ser entendido mais claramente observando a Figura 2.

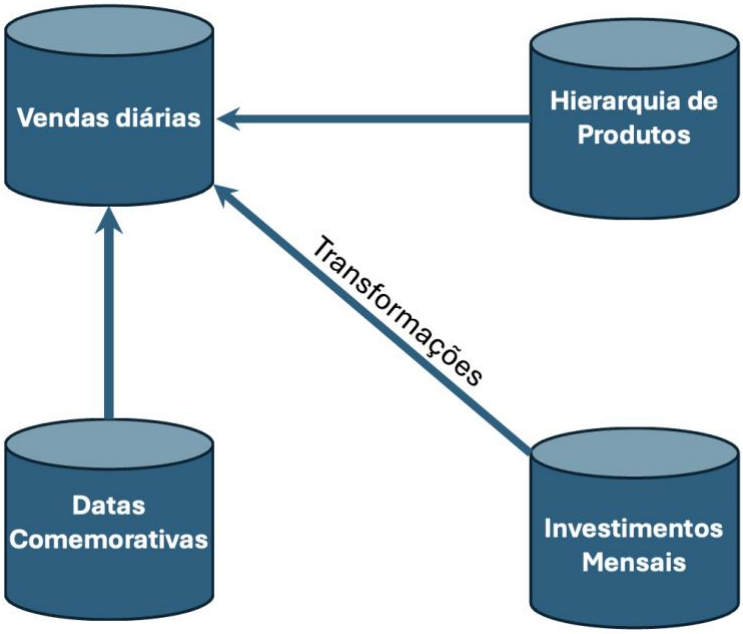

Figura 2 - Esquemático do relacionamento entre as bases de dados utilizadas no trabalho.

Por se tratar de muitas variáveis, o detalhamento e descrição de cada uma delas foi deixado no Apêndice A. Na sequência do texto, será dada ênfase às etapas de tratamento e manipulação da base de dados até a obtenção da base final usada na aplicação do *MMM*, ressaltando que algumas variáveis sem evidência de serem importantes para a análise (como, por exemplo, id do cliente, id da compra e datas sem relevância) foram descartadas.

3.2 Feature Engeneering

De acordo com Jiang (2022), o *MMM* é um modelo agregado e a granularidade de seus dados está tipicamente no nível de canal, país/região e/ou produto, em vez do nível do usuário. Os dados de gastos podem ser esparsos e voláteis em um nível diário, então dados agregados semanais ou mensais são muito comuns. Além disto, de acordo com Ni (2024), tipicamente são usados como variáveis preditoras em um modelo de mix de *marketing*: dados de investimento em comerciais, dados de performance como visitas, impressões, cliques; informações relativas a eventos como feriados, ou lançamentos de produtos; fatores de alteração de preços como promoções, descontos ou novos competidores.

Utilizando como ponto de partida os autores citados e com todos os dados listados a disposição, inicia-se a o processo de transformações e combinações a fim de criar ou adequar variáveis cruciais para atingir o objetivo proposto no trabalho. Necessitou-se fazer agregações e mudança na granularidade do dado das informações transacionais, para transformá-las em granularidade semanal e sair do nível de informação cliente/item/pedido. De forma similar, por outro lado, foi necessário trabalhar com interpolações com as informações de investimentos de marketing, *NPS* e ações. Tais procedimentos são descritos na sequência.

Para a aplicação adequada das técnicas estatísticas descritas neste trabalho, foi necessário lançar mão de uma técnica de manipulação de dados chamada *One-Hot Encoding* que nada mais é que uma técnica utilizada para transformar variáveis categóricas em um formato numérico adequado para modelos de *machine learning* e estatísticos. Essa abordagem cria *dummy variables* (variáveis binárias) para cada categoria única, onde cada coluna representa uma categoria e assume o valor 1 se a observação pertence àquela categoria, e 0 caso contrário. No caso prático do conjunto de dados utilizado no trabalho, havia uma coluna com a que descrevia a hierarquia de produtos (Game, Speaker, ...) que foi transformada em 7 colunas, uma para cada categoria e, posteriormente, uma destas colunas foi suprimida para ser usada como referência. Essa técnica evita a atribuição de ordem arbitrária às categorias, o que poderia distorcer a interpretação em algoritmos sensíveis à magnitude dos valores, como regressões lineares ou redes neurais (Géron, 2019)

3.2.1 Agregação, Interpolação e Criação de Variáveis Sazonais

As informações de investimento (MediaInvestment.csv) estão em granularidade mensal, enquanto a informação está disponível em granularidade diária para as demais bases. Para evitar a alta volatilidade de dados diários e a drástica redução no número de unidades

amostrais (tamanho da amostra) se considerássemos dados mensais, a análise será conduzida a nível semanal.

As variáveis diárias foram agregadas para o nível semanal por meio de somas ou médias, de acordo com o que fosse mais apropriado para cada indicador. Por sua vez, para a expansão das variáveis mensais para o nível semanal, uma boa estratégia de acordo com Bruce e Bruce (2019) é a utilização de *Splines*, técnica que permite a interpolação de pontos de forma suavizada. Foram aplicados três métodos de suavização usando a função *spline()* disponível no pacote stats (R Core Team, 2024) no *software R*. A Figura 3 mostra como é o investimento em TV mês a mês, em seu formato original, ainda antes da aplicação da técnica de suavização.

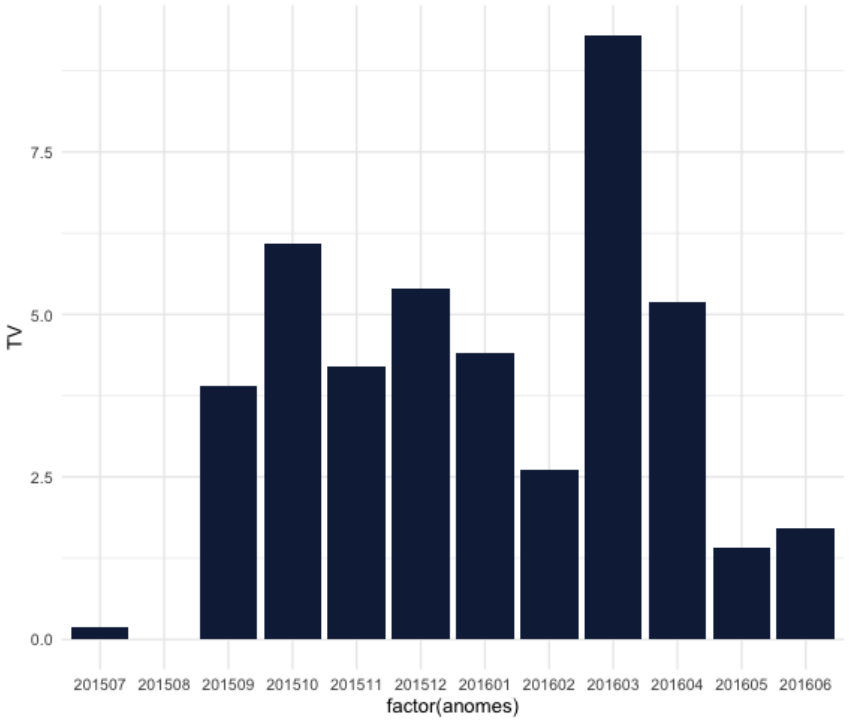

Figura 3 - Investimento em anúncios de TV, forma original do dado

Já a Figura 4 traz a comparação dos três métodos de interpolação aplicados para os 12 dados mensais de investimento em TV. Nota-se que a curva verde parece representar melhor as relações entre tempo e investimento em propaganda na TV, com isto o método *natural* foi o escolhido. Este método foi repetido para todas as informações mensais que necessitavam de interpolação, capacitando a obtenção de níveis de investimentos semanais, e mais do que isto, distinguir o nível de investimento de uma semana anterior e posterior.

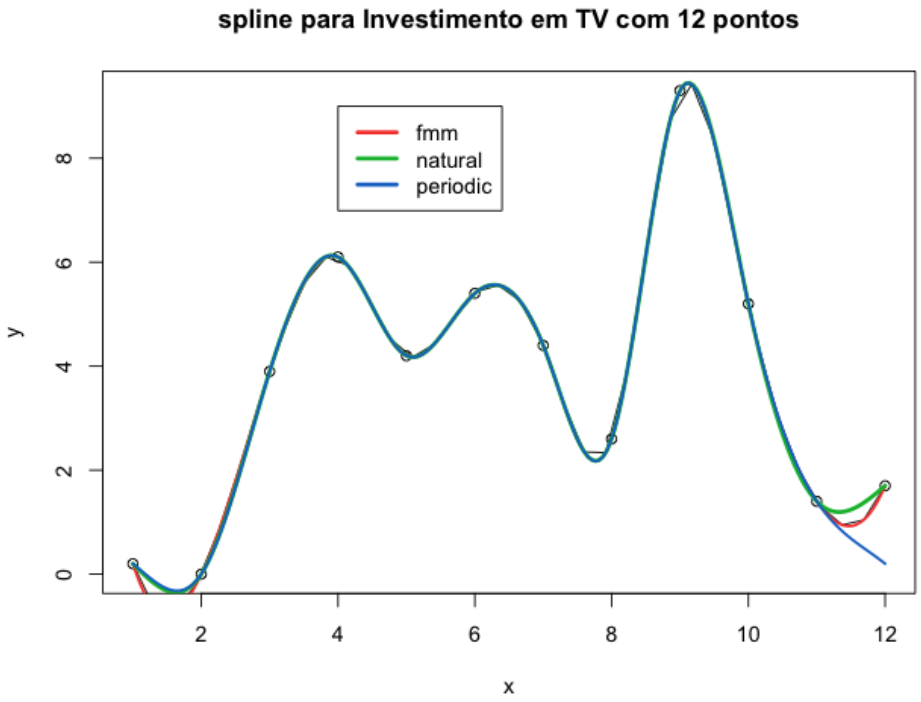

Figura 4 - Aplicação de métodos de interpolação via *splines* para os dados mensais de investimento em propaganda na TV.

Na Figura 5 observa-se no eixo y a esquerda a margem bruta semanal, representada pela linha azul, enquanto no eixo y a direita, o valor de investimento realizado na semana correspondente em veiculações na TV, cuja representação gráfica está na linha em amarelo.

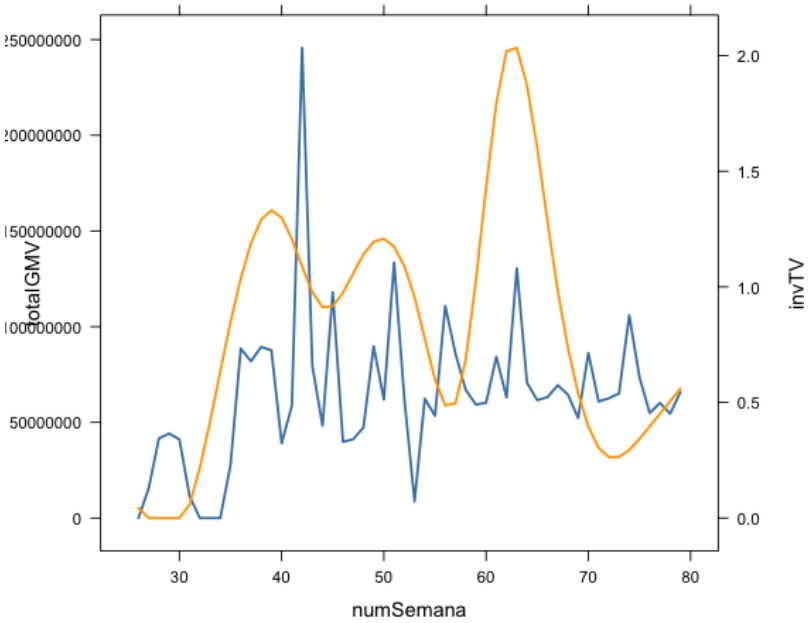

Figura 5 - Relação entre investimento em TV, após suavização, (curva em laranja) e receita bruta do *e-commerce* (curva em azul). No eixo x, tem-se a semana.

O uso da informação sobre o passado recente de alguns indicadores pode ser importante por refletir num maior ou menor resultado nas vendas do *e-commerce* na semana em questão. Por exemplo, impactos de anúncios podem ter sua consequência temporal realizada mais do que sete semanas após o início do investimento ser feito pela empresa (Jiang, 2022) como pode ser visto na Figura 6, retirada do artigo de Jiang (2022). Este efeito retardado entre o impacto de um comercial e a efetiva conversão é chamado de *carryover effect*.

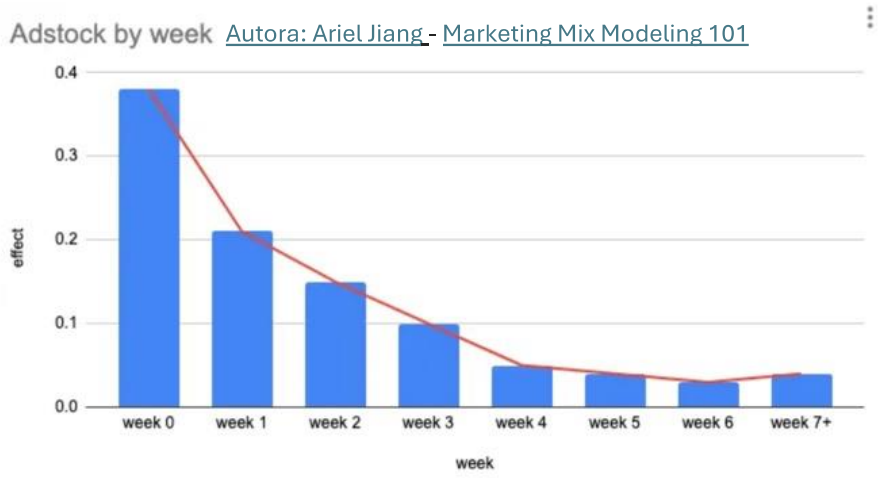

Figura 6 - Efeito do *carryover effect* por mais de sete semanas extraído Jiang (2022).

Tendo então a base de dados com todas as variáveis de interesse em nível semanal, a fim de capturar possíveis efeitos autorregressivos, foram criadas novas variáveis de investimentos e de fatores externos, contendo defasagem (*lag*) de ordem 1 (semana anterior) e ordem 2 (duas semanas atrás), além de uma variável representando a soma das 4 semanas anteriores. Isto foi feito para o nível de investimento em mídias, satisfação dos clientes e valor das ações. Por fim, chegou-se a uma base de dados contendo as informações em n=54 semanas de observações (1 ano) para 38 variáveis independentes e 1 variável resposta, conforme listado na Tabela 1.

Tabela 1 - Descrição da base de dados após agregações, interpolações, transformação e criação de variáveis defasadas

Variável	Tipo de dado	Conceito	Observações
totalGMV	Numérico	Receita Bruta Total	**Variável Resposta**
numSemana	Numérica Ordinal	indica o número da semana do período de análise	
flPrePago	Factor	0 indica que a venda foi feita *Cash on Delivery,* 1 indica que foi feito pré-pago	
Camera	Factor	Agrupamento médio de produtos, classificado como 1 caso seja câmera	Transformada pelo *One-hot-encocoding*, "AudioMP3Player" tomada como referência. As demais subcategorias viram uma coluna preenchidas com 0 e 1
CameraAssessory	Factor	Agrupamento médio de produtos, classificado como 1 caso seja acessório para câmera	Transformada pelo *One-hot-encocoding,* "AudioMP3Player" tomada como referência. As demais subcategorias viram uma coluna preenchidas com 0 e 1

GamingConsole	Factor	Agrupamento médio de produtos, classificado como 1 caso seja Console de Vídeo Game	Transformada pelo *One-hot-encocoding*, "AudioMP3Player" tomada como referência. As demais subcategorias viram uma coluna preenchidas com 0 e 1
HomeAudio	Factor	Agrupamento médio de produtos, classificado como 1 caso seja Áudio Residencial	Transformada pelo *One-hot-encocoding*, "AudioMP3Player" tomada como referência. As demais subcategorias viram uma coluna preenchidas com 0 e 1
Speaker	Factor	Agrupamento médio de produtos, classificado como 1 caso seja Auto Falante	Transformada pelo *One-hot-encocoding*, "AudioMP3Player" tomada como referência. As demais subcategorias viram uma coluna preenchidas com 0 e 1
TVVideoSmall	Factor	Agrupamento médio de produtos, classificado como 1 caso seja TVs portátil	Transformada pelo *One-hot-encocoding*, "AudioMP3Player" tomada como referência. As demais subcategorias viram uma coluna preenchidas com 0 e 1
Popular	Factor	classificação do produto de *tier* popular	Transformada pelo *One-hot-encocoding*, "Alto.Ticket" tomada como referência. As demais classes de produtos viram uma coluna cada preenchidas com 0 e 1
Média	Factor	classificação do produto de *tier* médio	Transformada pelo *One-hot-encocoding*, "Alto.Ticket" tomada como referência. As demais classes de produtos viram uma coluna cada preenchidas com 0 e 1
product_procurement_sla	Numérico	quantidade média de dias estimadas para loja adquirir o produto e tê-lo em seu estoque	
meanSLA	Numérico	tempo médio de entrega dos pedidos de produtos da semana	
qtySpecialDates	Numérico	quantidade de datas especiais na semana	
flPromocional	Factor	indica se o produto foi vendido com preços promocionais - descontos superiores a 10%	
totalProductMRP	Numérico	valor total dos produtos vendidos, sem desconto	
totalVlDescUnit	Numérico	valor total dos descontos aplicados nos produtos vendidos	
invTV	Numérico	investimento em TV na semana	
invSponsorship	Numérico	investimento em Patrocínios na semana	
invAffiliates	Numérico	investimento em Afiliados na semana	
invMidiaCRM	Numérico	investimento em Midia e *CRM* na semana	junção de Digital, Marketing Online e SEM
invTVLag1	Numérico	Investimento em TV semana anterior	
invTVLag2	Numérico	Investimento em TV de duas semanas antes	
invTVLag4Sum	Numérico	Soma do investimento em TV das 4 semanas anteriores	
invSponsorshipLag1	Numérico	Investimento em Patrocínios semana anterior	
invSponsorshipLag2	Numérico	Investimento em Patrocínios de duas semanas antes	

invSponsorshipLag4Sum	Numérico	Soma do investimento em Patrocínios das 4 semanas anteriores	
invAffiliatesLag1	Numérico	Investimento em Afiliados semana anterior	
invAffiliatesLag2	Numérico	Investimento em Afiliados de duas semanas antes	
invAffiliatesLag4Sum	Numérico	Soma do investimento em Afiliados das 4 semanas anteriores	
invMidiaCRMLag1	Numérico	Investimento em Midia e *CRM* semana anterior	junção de Digital, Marketing Online e SEM
invMidiaCRMLag2	Numérico	Investimento em Midia e *CRM* de duas semanas antes	junção de Digital, Marketing Online e SEM
invMidiaCRMLag4Sum	Numérico	Soma do investimento em Midia e *CRM* das 4 semanas anteriores	junção de Digital, Marketing Online e SEM
interpNPS	Numérico	*NPS* apurado da semana	
NPSLag1	Numérico	*NPS* apurado da semana anterior	
NPSLag2	Numérico	*NPS* apurado de duas semanas atras	
interpStock	Numérico	valor das ações na semana corrente	
StockLag1	Numérico	valor das ações na semana anterior	
StockLag2	Numérico	valor das ações há duas semanas antes	

3.2.2 Redução de Dimensionalidade

Tendo em vista que a base de dados descrita no Quadro 1 tem um número elevado de variáveis explicativas (p=38) e o tamanho amostral é relativamente pequeno (n=54), foi necessário lançar mão de técnicas para redução de dimensionalidade. Neste contexto, a análise fatorial ortogonal é uma técnica estatística amplamente utilizada para a redução de dimensionalidade, cujo objetivo principal é descrever a variabilidade do conjunto de dados original, um vetor aleatório X, em termo de um número menor de variáveis latentes, as quais são chamadas de fatores, e estão ligadas ao vetor de origem X através de um modelo linear (Mingoti, 2005).

A análise fatorial é dividida em duas principais vertentes: a análise fatorial exploratória (AFE) e a análise fatorial confirmatória (AFC). A análise fatorial exploratória é utilizada quando não se há hipóteses sobre redução de dimensões sobre os dados disponíveis e tem um caráter exploratório sobre as relações entre as características existentes. Por outro lado, a AFC, é utilizada para testar hipóteses pré-definidas (Kline, 2015). Especificamente neste estudo de caso, não havia uma hipótese clara de agrupamento dos dados, e foi necessário explorar as relações anteriormente as decisões, com isto foi utilizado o método AFE.

Segundo Mingoti, S. A. (2005), a análise fatorial pode ser aplicada às variáveis originais contidas em um vetor X e, para facilitar o entendimento e interpretação, pode ser preferível aplicar os conceitos principais utilizando as variáveis padronizadas pela média e desvio-

padrão. Esta sugestão foi seguida, visto que as variáveis podem ter ordem de grandezas ou unidades distintas.

Observando a Tabela 1 é possível notar um certo agrupamento natural das variáveis baseado em seu conceito ou temática. Deste modo, com o objetivo de facilitar a interpretabilidade dos fatores finais obtidos, a AFE foi aplicada partindo destas possíveis relações práticas. Cabe ressaltar que apenas as variáveis quantitativas foram consideradas na AFE, ou seja, as variáveis do tipo *factor* e booleanas foram mantidas em seu formato original para a aplicação do *MMM*.

As análises foram conduzias usando a função *principal()* do pacote *psych (Revelle, 2024)* do software R. Em caráter exploratório, sugere-se inicialmente obter os resultados com o numero de fatores m igual ao número de variáveis p do vetor original, sendo feita uma análise de quantos fatores deveriam ser retidos para explicar satisfatoriamente a variabilidade das p variáveis originais. Mingoti, S. A. (2005) sugere três critérios principais para determinar qual é a quantidade de fatores a serem utilizados na redução de dimensionalidade sem perda significante de informação. Neste estudo, optou-se por reter a quantidade de fatores associados a autovalores maiores que 1, conhecido como critério de Kaiser, o qual implica na retenção de fatores que explicam individualmente pelo menos uma variável (após padronização cada variável tem variância igual a 1). Por fim, a fim de maximizar a separabilidade das variáveis originais em fatores ortogonais, a rotação do tipo varimax foi considerada (Mingoti, 2005).

Análise Fatorial Exploratória dos Fatores Externos

O primeiro tema a ser tratado para a tentativa de redução de dimensionalidade foram os fatores externos, os quais são: a nota final de *NPS*; índice de satisfação dos clientes; e o valor de fechamento das ações da empresa (*Stock*). Em caráter exploratório, coloca-se o numero de fatores m igual ao número de variáveis p do vetor original, ou seja 6 fatores, uma vez que são 6 variáveis relacionadas ao agrupamento temático que foi chamado de fatores externos, são elas: interpNPS (nota do *NPS* interpolada na semana corrente), interpStock (valor da ação interpolada na semana corrente), NPSLag1 (nota do *NPS* interpolada com atraso de uma semana, que representa a nota de *NPS* da semana anterior a corrente), NPSLag2 (nota do *NPS* interpolada com atraso de duas semanas, que representa a nota de *NPS* duas semana anteriores a corrente), StockLag1 (valor da ação da empresa interpolada com atraso de uma semana, que representa o valor de fechamento da semana anterior a corrente), StockLag2 (nota do *NPS* interpolada com atraso de duas semanas, que representa a nota de *NPS* duas semana anteriores a corrente).

Neste caso, m=3 fatores apresentaram autovalor superior a 1 e foram retidos. Eles explicam 96% da variabilidade das p=6 variáveis originais (ver Quadro A1 do Apêndice B). Desta maneira, o grupo de fatores externos foi reduzido de 6 variáveis para 3 fatores. A interpretação dos 3 fatores obtidos foi feita da seguinte forma:

- RC3 - Fator de Satisfação dos Clientes de Curto Prazo, que é a combinação de maior peso das variáveis: interpNPS e NPSLag1.

- RC1 - Fator de Fechamento de valor de Ação de Curto Prazo, que é a combinação de maior peso das variáveis: interpStock e StockLag1.
- RC2 - Fatores Externos de longo prazo, que é a combinação dos fatores mais distantes da semana atual: NPSLag2 e StockLag2.

Análise Fatorial Exploratória do Investimento em TV

Para o agrupamento de variáveis relacionados ao nível de investimento realizado em veiculação de comerciais em televisão, as 4 variáveis utilizadas são: invTV, invTVLag1, invTVLag2, invTVLag4sum. Neste caso, m=2 fatores apresentaram autovalor superior a 1 e foram retidos. Eles explicam 99,9% da variabilidade das p=4 variáveis originais (ver Quadro A2 do Apêndice B). Desta maneira, o grupo de fatores de investimento em TV foi reduzido de 4 variáveis para 2 fatores. A interpretação dos 2 fatores obtidos foi feita da seguinte forma:

- RC1 - Fator Nível de Investimento em TV Realizado no Passado, que é a combinação de maior peso das variáveis: invTVLag2 e invTVLag4sum.
- RC2 - Fator Nível de Investimento em TV Recente, que é a combinação de maior peso das variáveis: invTV e invTVLag1.

Análise Fatorial Exploratória do Investimento em Patrocínios

Outra vez foi usado o mesmo processo dos outros dois agrupamentos de variáveis anteriores, primeiro a *EFA* com todas as variáveis disponíveis que dizem respeito ao investimento realizado em patrocínios, e depois analisado os critérios de corte e interpretação dos fatores resultantes. Tem-se 4 variáveis que fazem parte deste agrupamento. São elas: invSponsorship, invSponsorshipLag1, invSponsorshipLag2, invSponsorshipLag4sum. Neste caso, m=2 fatores apresentaram autovalor superior a 1 e foram retidos. Eles explicam 99% da variabilidade das p=4 variáveis originais (ver Quadro A3 do Apêndice B). Desta maneira, o grupo de fatores de investimento em patrocínios foi reduzido de 4 variáveis para 2 fatores. A interpretação dos 2 fatores obtidos foi feita da seguinte forma:

- RC1 - Fator Nível de Investimento em Patrocínios Realizado no Passado, que é a combinação de maior peso das variáveis: invSponsorshipLag2 e invSponsorshipLag4sum.
- RC2 - Fator Nível de Investimento em Patrocínios Recente, que é a combinação de maior peso das variáveis: invSponsorship e invSponsorshipLag1.

Análise Fatorial Exploratória do Investimento em Afiliados

O agrupamento de dados que diz respeito ao investimento realizado em afiliados também possui 4 variáveis: invAffiliates, invAffiliatesLag1, invAffiliatesLag2, invAffiliatesLag4sum. Ao se realizar a análise fatorial exploratória, assim como nos demais

agrupamentos de variáveis que tratam gastos da *ElecKart* em formas de atrair clientes, houve a indicação de se manter apenas m=2 fatores que explicam 99,9% da variabilidade das p=2 variáveis originais (ver Quadro A4 do Apêndice B). A interpretação dos m=2 fatores obtidos é:

- RC1 - Fator Nível de Investimento em Afiliados Realizado no Passado, que é a combinação de maior peso das variáveis: invAffiliatesLag2 e invAffiliatesLag4sum.
- RC2 - Fator Nível de Investimento em Afiliados Recente, que é a combinação de maior peso das variáveis: invAffiliates e invAffiliatesLag1.

Análise Fatorial Exploratória do Investimento em Mídia e *CRM*

Assim como os demais agrupamentos relacionados a investimentos para trazer clientes, o investimento em mídia e *CRM (Customer Relationship Manager, gerenciamento do relacionamento com o cliente)*, possui 4 variáveis: invMidiaCRM, invMidiaCRMLag1, invMidiaCRMLag2 e invMidiaCRMLag4sum. Também neste caso, m=2 fatores apresentaram autovalor superior a 1 e foram retidos. Eles explicam 99% da variabilidade das p=4 variáveis originais (ver Quadro A5 do Apêndice B). A interpretação dos 2 fatores obtidos foi feita da seguinte maneira:

- RC1 - Fator Nível de Investimento em Mídia e *CRM* Realizado no Passado, que é a combinação de maior peso das variáveis: invMidiaCRMLag2 e invMidiaCRMLag4sum.
- RC2 - Fator Nível de Investimento em Mídia e *CRM* Recente, que é a combinação de maior peso das variáveis: invMidiaCRM e invMidiaCRMLag1

3.3 Base de Dados Final Preparada

Após todos os processos de agregações, interpolações, criações de variáveis defasadas e redução de dimensão, a base de dados final que será utilizada na aplicação dos *MMM* é apresentada no Quadro 2. São 29 variáveis explicativas e a variável resposta.

Tabela 2 - Descrição da base de dados final após aplicação da análise fatorial

Variável	Tipo de dado	Conceito	Observações
totalGMV	Numérico	Receita Bruta Total	Variável Resposta
numSemana	Numérica Ordinal	indica o número da semana do período de análise	
flPrePago	Factor	0 indica que a venda foi feita *Cash on Delivery*, 1 indica que foi feito pré-pago	
Camera	Factor	Agrupamento médio de produtos, classificado como 1 caso seja câmera	Transformada pelo *One-hot-code*, "AudioMP3Player" tomada como referência. As demais subcategorias viram uma coluna preenchidas com 0 e 1

CameraAssessory	Factor	Agrupamento médio de produtos, classificado como 1 caso seja acessório para câmera	Transformada pelo *One-hot-code*, "AudioMP3Player" tomada como referência. As demais subcategorias viram uma coluna preenchidas com 0 e 1
Game	Factor	Agrupamento médio de produtos, classificado como 1 caso seja game	Transformada pelo *One-hot-code*, "AudioMP3Player" tomada como referência. As demais subcategorias viram uma coluna preenchidas com 0 e 1
GamingAccessory	Factor	Agrupamento médio de produtos, classificado como 1 caso seja acessório para games	Transformada pelo *One-hot-code*, "AudioMP3Player" tomada como referência. As demais subcategorias viram uma coluna preenchidas com 0 e 1
GamingConsole	Factor	Agrupamento médio de produtos, classificado como 1 caso seja Console de Vídeo Game	Transformada pelo *One-hot-code*, "AudioMP3Player" tomada como referência. As demais subcategorias viram uma coluna preenchidas com 0 e 1
HomeAudio	Factor	Agrupamento médio de produtos, classificado como 1 caso seja Áudio Residencial	Transformada pelo *One-hot-code*, "AudioMP3Player" tomada como referência. As demais subcategorias viram uma coluna preenchidas com 0 e 1
Speaker	Factor	Agrupamento médio de produtos, classificado como 1 caso seja Auto Falante	Transformada pelo *One-hot-code*, "AudioMP3Player" tomada como referência. As demais subcategorias viram uma coluna preenchidas com 0 e 1
TVVideoSmall	Factor	Agrupamento médio de produtos, classificado como 1 caso seja TVs portátil	Transformada pelo *One-hot-code*, "AudioMP3Player" tomada como referência. As demais subcategorias viram uma coluna preenchidas com 0 e 1
Popular	Factor	classificação do produto de *tier* popular	Transformada pelo *One-hot-code*, "Alto.Ticket" tomada como referência. As demais classes de produtos viram uma coluna cada preenchidas com 0 e 1
Média	Factor	classificação do produto de *tier* médio	Transformada pelo *One-hot-code*, "Alto.Ticket" tomada como referência. As demais classes de produtos viram uma coluna cada preenchidas com 0 e 1
product_procurement_sla	Numérico	quantidade média de dias estimadas para loja adquirir o produto e tê-lo em seu estoque	
meanSLA	Numérico	tempo médio de entrega dos pedidos de produtos da semana	
qtySpecialDates	Numérico	quantidade de datas especiais na semana	
flPromocional	Factor	indica se o produto foi vendido com preços promocionais - descontos superiores a 10%	Se o produto foi vendido com desconto maior que 10% esta variável recebe 1, senão recebe 0
totalProductMRP	Numérico	valor total dos produtos vendidos, sem desconto	
totalVlDescUnit	Numérico	valor total dos descontos aplicados nos produtos vendidos	
Fator Nível de Investimento em TV Realizado no Passado	Numérico	Fator Nível de Investimento em TV Realizado no Passado, que é a combinação de maior peso das variáveis: invTVLag2 e invTVLag4sum	
Fator Nível de Investimento em TV Recente	Numérico	Fator Nível de Investimento em TV Recente, que é a combinação de maior peso das variáveis: invTV e invTVLag1	
Fator Nível de Investimento em Afiliados Realizado no Passado	Numérico	Fator Nível de Investimento em Afiliados Realizado no Passado, que é a combinação	

		de maior peso das variáveis: invAffiliatesLag2 e invAffiliatesLag4sum.	
Fator Nível de Investimento em Afiliados Recente	Numérico	Fator Nível de Investimento em Afiliados Recente, que é a combinação de maior peso das variáveis: invAffiliates e invAffiliatesLag1.	
Fator Nível de Investimento em Patrocínios Realizado no Passado	Numérico	Fator Nível de Investimento em Patrocínios Realizado no Passado, que é a combinação de maior peso das variáveis: invSponsorshipLag2 e invSponsorshipLag4sum	
Fator Nível de Investimento em Patrocínios Recente	Numérico	Fator Nível de Investimento em Patrocínios Recente, que é a combinação de maior peso das variáveis: invSponsorship e invSponsorshipLag1	
Fator Nível de Investimento em Mídia e CRM Realizado no Passado	Numérico	Fator Nível de Investimento em Mídia e *CRM* Realizado no Passado, que é a combinação de maior peso das variáveis: invMidiaCRMLag2 e invMidiaCRMLag4sum.	
Fator Nível de Investimento em Mídia e CRM Recente	Numérico	Fator Nível de Investimento em Mídia e *CRM* Recente, que é a combinação de maior peso das variáveis: invMidiaCRM e invMidiaCRMLag1	
Fator de Satisfação dos Clientes de Curto Prazo	Numérico	*NPS* apurado da semanaFator de Satisfação dos Clientes de Curto Prazo, que é a combinação de maior peso das variáveis: interpNPS e NPSLag1	
Fator de Fechamento de valor de Ação de Curto Prazo	Numérico	Fator de Fechamento de valor de Ação de Curto Prazo, que é a combinação de maior peso das variáveis: interpStock e StockLag1	
Fatores Externos de longo prazo	Numérico	Fatores Externos de longo prazo, que é a combinação dos fatores mais distantes da semana atual: NPSLag2 e StockLag2	

3.4 Estratégia de modelagem

Uma análise preliminar foi realizada usando modelos de regressão linear simples com cada uma das variáveis explicativas. O intuito dessa abordagem é analisar o efeito individual de cada uma delas, além da significância estatística, ao nível de 5% de significância. Na sequência, foram ajustados seis modelos com múltiplas variáveis utilizando três diferentes técnicas: Regressão Linear Múltipla (3 modelos – um com 100% das variáveis, um com seleção automática de variáveis, e um com seleção manual de variáveis); Regressão *Lasso* (*Least Absolute Shrinkage and Selection Operator*) (Friedman, 2010) e *CART* (Therneau & Atkinson, 2023). Devido o foco do *MMM* está em sua interpretabilidade, as abordagens mais comuns têm sua base em regressões (Wigren & Cornell, 2020) contudo, este não é um pré-requisito único, por isso a última técnica citada, apesar de não ter uma interpretação clara das relações entre os preditores, foi executada para testar a capacidade preditiva, para captura de insights e projeção de cenários futuros.

Os modelos de regressão linear múltipla foram ajustados usando o método dos mínimos quadrados e a seleção automática de variáveis foi realizada usando o critério *AIC* usando método *stepwise*. Para todas as técnicas aplicados, os dados referentes as últimas quatro semanas foram extraídos da base de dados e chamados de Base de Teste, os demais dados foram chamados Base de Treino, caracterizando a utilização do método de avaliação de modelos chamado *DataSplit*.

Como critério de avaliação dos modelos, foram calculados o R^2 preditivo, que é calculado utilizando o conjunto de teste, e o erro quadrático médio (em inglês, *mean squared error* (*MSE*)) também para a base de teste. Quanto maior o R^2 preditivo e quanto menor o *MSE*, melhor o ajuste. Além disto, foram analisados o nível de complexidade do modelo, interpretabilidade e usabilidade para simular cenários de investimento de marketing, venda de determinadas categorias e o nível de descontos a serem aplicados aos consumidores.

Todas as análises foram feitas utilizando o *software R* (R Core Team, 2024) e a interface de desenvolvimento *R Studio* (Posit team, 2024), com auxilio dos pacotes: tidyverse (Wickham et al, 2019), dplyr (Wickham et al, 2023), ggplot2 (Wickham, 2016), car (Fox & Weisberg, 2019), esquisse (Mayer & Perrier, 2024), stats (R Core Team, 2024), dlm (Petris, 2010), caret (Kuhn, 2008), psych (Revelle, 2024), rpart (Therneau & Atkinson, 2023), rpart.plot (Milborrow, 2024), glmnet (Friedman et al, 2010).

4. Resultados

4.1 Análise Descritiva das Variáveis

Antes de iniciar a modelagem estatística, foi realizado uma análise exploratória final dos dados. Para os dados categóricos, foi plotado o *boxplot*, para tentar observar como se comportam com relação a variável resposta, também foi observado a dispersão das categorias, percentualmente (Figura 7).

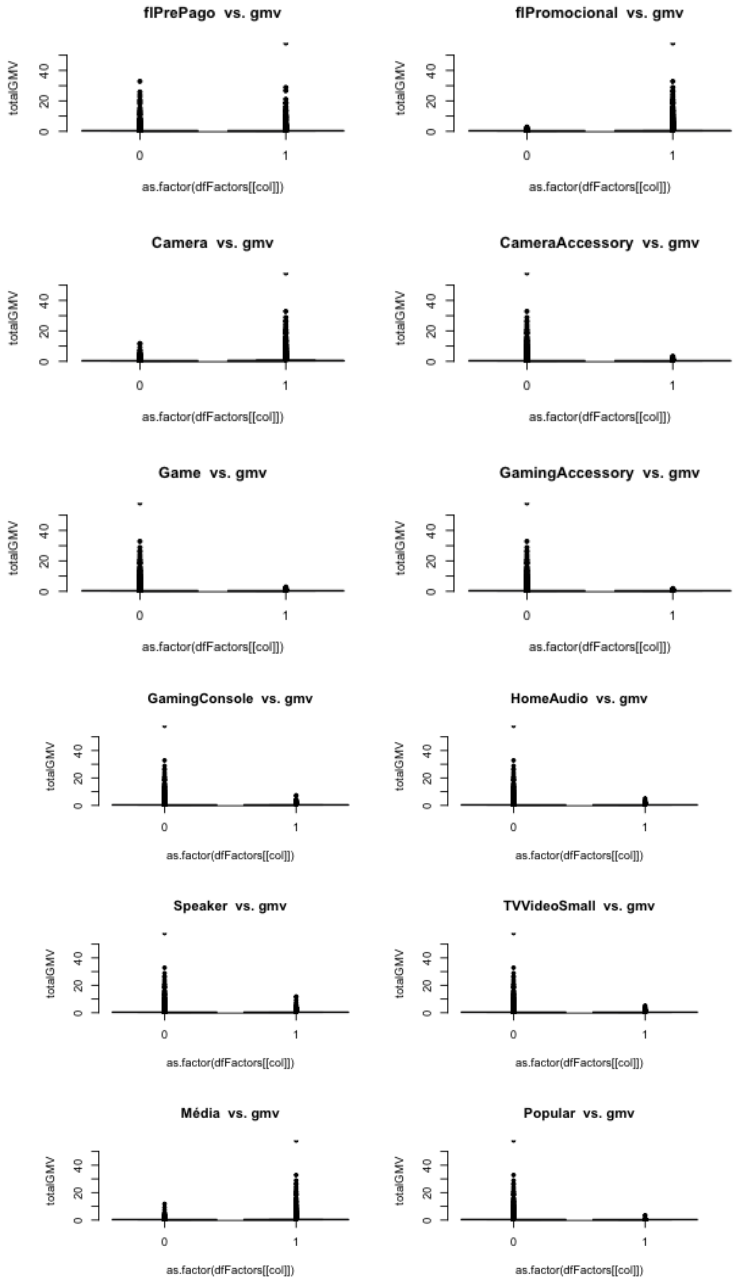

Figura 7 - Análise Exploratória das variáveis categóricas.

Nota-se, a partir da visualização da Figura 7, que há muitos pontos fora do *boxplot*, em todas as variáveis do tipo fator utilizadas no projeto. Para as variáveis que representam hierarquia de produtos, o fato de pertencer àquele tipo de produto, ameniza a quantidade de pontos sobressalentes a caixa do gráfico, ou seja, produtos daquele tipo, parecem ter uma dispersão de receita bruta mais controlada quando comparada aos seus pares, pertencentes a mesma categoria. A Tabela 3, complementa a Figura 7, e traduz como as observações utilizadas para treinar e testar os modelos propostos neste trabalho, estão distribuídas quanto a pertencimento ou não às variáveis categóricas.

Tabela 3 - Distribuição percentual de pertencimento das observações às variáveis categóricas

Variável Categórica	NÃO (%)	SIM (%)
flPrePago	52%	48%
flPromocional	43%	57%
Camera	90%	10%
CameraAccessory	86%	14%
Game	87%	13%
GamingAccessory	88%	12%
GamingConsole	93%	7%
HomeAudio	89%	11%
Speaker	87%	13%
TVVideoSmall	91%	9%
Média	59%	41%
Popular	69%	31%

Também é importante entender como as variáveis numéricas ser relacionam com a variável resposta, e para isto pode-se lançar mão da análise de correlação, que está plotada na Figura 8. Nota-se a alta correlação da Receita Bruta com o valor dado de desconto nos produtos vendidos, também é curioso que o NPS Recente tenha uma correlação negativa e forte com os investimentos em anúncios. A correlação positiva entre o indicador da semana e os investimentos em afiliados também chama atenção.

Além da análise de correlação, na Tabela 4, pode-se entender cada variável de forma de forma independente. As variáveis numéricas foram padronizadas, usando a função *scale()* do *software* R. A variável que representa o tempo médio de entrega e a variável que representa o desconto concedido, são as que apresentam as maiores dispersões individuais. Depois de entendido como as variáveis se relacionam entre si e como se relacionam com a variável resposta, pode-se avançar sobre as estratégias de modelagem estatística.

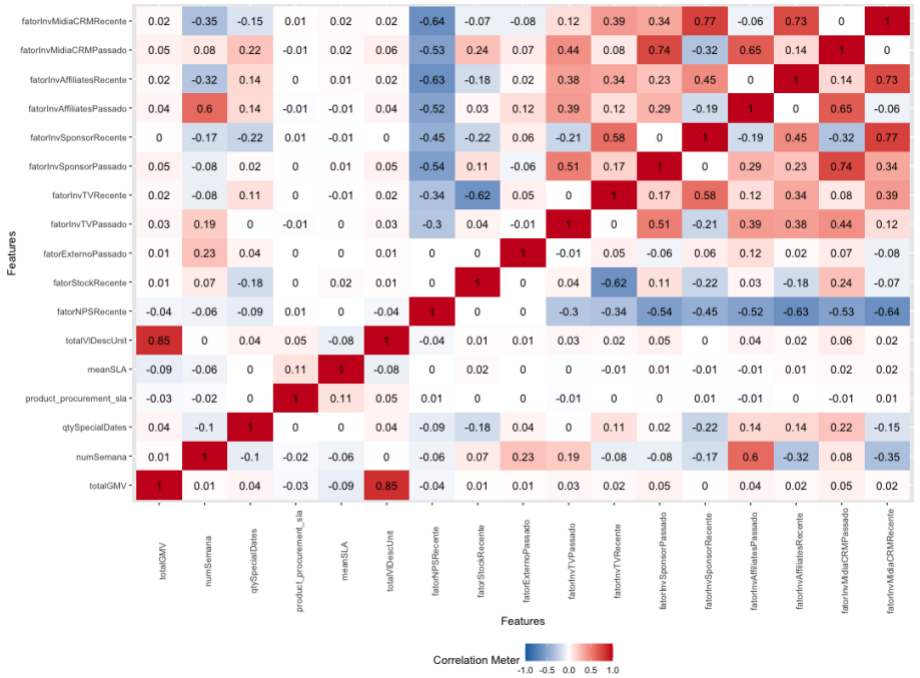

Figura 8 - Tabela de correlação entre as variáveis numéricas usadas no projeto

Tabela 4 - Análise descritiva das variáveis numéricas.

Variável Numérica	Média	Desvio Pad.	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo
numSemana	1,25E-14	1,00	-1,90	-0,83	0,05	0,85	1,66
qtySpecialDates	4,09E-15	1,00	-0,58	-0,58	-0,58	0,83	3,65
product_procurement_sla	-2,23E-16	1,00	-1,09	-0,65	-0,20	0,69	4,70
meanSLA	-1,48E-16	1,00	-2,91	-0,58	-0,02	0,50	10,50
totalVlDescUnit	1,15E-16	1,00	-0,20	-0,20	-0,19	-0,11	54,50
fatorNPSRecente	9,19E-15	1,00	-4,54	-0,68	-0,09	0,34	2,81
fatorStockRecente	2,94E-15	1,00	-5,98	-0,70	0,23	0,81	1,40
fatorExternoPassado	-1,77E-14	1,00	-7,04	-0,16	0,11	0,37	1,12
fatorInvTVPassado	-5,71E-14	1,00	-1,67	-0,95	-0,03	0,70	2,37
fatorInvTVRecente	1,61E-14	1,00	-1,29	-0,88	-0,20	0,61	2,58
fatorInvSponsorPassado	1,92E-14	1,00	-1,58	-0,83	-0,09	0,40	2,56
fatorInvSponsorRecente	-3,01E-14	1,00	-1,59	-0,65	-0,15	0,53	2,38
fatorInvAffiliatesPassado	-5,33E-14	1,00	-3,03	-0,17	0,35	0,56	1,54
fatorInvAffiliatesRecente	2,99E-14	1,00	-3,26	-0,01	0,23	0,62	1,36
fatorInvMidiaCRMPassado	3,28E-14	1,00	-1,84	-0,47	-0,05	0,29	2,86
fatorInvMidiaCRMRecente	-3,81E-14	1,00	-1,66	-0,54	-0,12	0,31	2,85

4.2 Ajuste de modelos de regressão linear simples

A base de dados descrita na Tabela 2 foi utilizada para prever a receita bruta (*gmv*) de um período (N semanas) do *e-commerce,* com base nos investimentos em anúncios (Promoção), na variabilidade de categorias e *tiers* vendidos (Produtos), nos descontos que estão sendo aplicados (Preço), no método de pagamento utilizado e no *SLA* (Service Level Agreement, que em português a tradução literal é acordo de nível de serviço, e neste contexto significa o tempo esperado para a entrega (Praça)), além de fatores exógenos a venda, como o andamento dos valores da ação da companhia e o nível de satisfação dos clientes. Simbolicamente, temos a seguinte equação, paramétrica, que sintetiza o problema, aplicando na equação proposta por Hegewald, M. (2019):

$$GMV_t = \beta_0 + \beta_1 * Promoção_t + \beta_2 * Produtos_t + \beta_3 * Preço_t + \beta_4 * Praça_t + \beta_5 * Externo_t + \varepsilon$$

Inicialmente, a fim de realizar a seleção de variável, foi realizada uma regressão linear individual entre cada uma das vinte e nove variáveis independentes e a variável resposta. Os resultados estão resumidos na Tabela 5.

Tabela 5 - Estimativas obtidas via ajuste de modelo de regressão linear entre cada uma das variáveis preditora individualmente e a variável resposta.

Variável	Coeficiente	p-valor	R^2 (individual)
totalProductMRP	0,965	0,000	0,932
totalVlDescUnit	0,851	0,000	0,724
Camera	0,228	0,000	0,052
Média	0,155	0,000	0,024
flPromocional	0,154	0,000	0,023
Popular	-0,132	0,000	0,017
meanSLA	-0,089	0,000	0,008
GamingAccessory	-0,053	0,000	0,002
Game	-0,051	0,000	0,002
fatorInvMidiaCRMPassado	0,051	0,000	0,002
fatorInvSponsorPassado	0,046	0,000	0,002
CameraAccessory	-0,044	0,000	0,002
fatorNPSRecente	-0,043	0,000	0,001
fatorInvAffiliatesPassado	0,037	0,000	0,001
qtySpecialDates	0,037	0,000	0,001
flPrePago	-0,034	0,000	0,001
Speaker	0,032	0,000	0,001
TVVideoSmall	-0,029	0,000	0,000
fatorInvTVPassado	0,029	0,000	0,000
HomeAudio	-0,028	0,000	0,000
product_procurement_sla	-0,027	0,000	0,000
fatorInvAffiliatesRecente	0,024	0,000	0,000
fatorInvTVRecente	0,020	0,001	0,000

fatorInvMidiaCRMRecente	0,018	0,005	0,000
numSemana	0,011	**0,073**	0,000
fatorExternoPassado	0,011	**0,078**	0,000
GamingConsole	0,011	**0,087**	0,000
fatorStockRecente	0,006	**0,296**	0,000
fatorInvSponsorRecente	0,004	**0,488**	0,000

Observando a Tabela 5 é possível notar que as duas variáveis da dimensão preço são as mais importantes, seguidas de variáveis da dimensão produto. As variáveis que remetem ao investimento em anúncios estão mais para o fim da tabela e, embora tenham p-valor inferior ao nível de significância de 0,05, não são as que mais explicam a Receita bruta individualmente. Da lista de 29 variáveis, apenas 5 não tem significância estatística (destacadas em negrito): o identificador da semana, o fator externo do passado, o fator que traduz o preço das ações recentemente, o fato de vender categoria de consoles de videogame e investimento em patrocínios recentes.

Com relação ao sinal do coeficiente, a grande maioria das variáveis contribuem positivamente. As que têm sinal negativo são: venda de produtos de *tier* populares, o que faz sentido, visto que são os produtos mais baratos, então vender muito não contribui tanto para o crescimento da receita bruta; as variáveis de *SLA*, o que era de se esperar, quanto maior o tempo de entrega, menos contribui negativamente a receita. Além disso, chama a atenção algumas categorias de produtos, terem relação negativa com a receita bruta, como por exemplo categoria HomeAudio, TVVideoSmall e Game, além do *NPS* Recente.

Com esta análise preliminar, tem-se uma base de expectativa gerada sobre o que se esperar da modelagem combinada das variáveis usando modelos de regressão linear múltipla clássica, regressão *Lasso* e *CART*.

4.3 Ajustes de modelos com múltiplas variáveis

A Tabela 6 apresenta um resumo do ajuste dos cinco modelos com múltiplas variáveis utilizando três diferentes técnicas: Regressão Linear Múltipla (3 modelos – um com 100% das variáveis, um com seleção automática de variáveis, e um com seleção manual de variáveis); Regressão *Lasso*; e *CART*. Foi utilizada a técnica de *Data Split*, ou particionamento de dados, para separar os dados em conjunto de treinamento, para ajustar o modelo e conjunto de teste, para avaliar a performance dos mesmos, e compará-los entre si. O conjunto de treinamento, recebeu dados da série temporal de julho de 2015 até aproximadamente fim de maio de 2016. Já o conjunto de teste recebeu as últimas 4 semanas de dados.

Analisando a Tabela 6, nota-se que os ajustes possuem R^2 preditivos razoáveis em torno ou superiores a 70% em todos os modelos, para os dados de treino e de teste. Importante destacar que a variável totalProductMRP que individualmente tinha maior relação com o GMV não foi incluída nos modelos. Isso se deu pelo fato de que ao realizar testes incluindo a variável totalProductMRP, que representa o valor nominal do preço dos produtos, ela causa

um *overfitting* nos ajustes, levando a R^2 preditivos maiores que 0,99, característica de Data Leakage, que é um problema crítico em modelagem estatística, ocorrendo quando informações do conjunto de teste ou dados futuros "vazam" para o conjunto de treinamento, resultando em avaliações otimistas e modelos que falham em generalizar para novos dados (Kaufman et al., 2012). Esse fenômeno pode surgir de diversas formas, e no caso deste trabalho, a fonte de vazamento seria o uso da candidata a preditora totalProductMRP, que é uma informação que só estará disponível após a predição. Além disto, em cenários de simulação, não se sabe o montante nominal total e, portanto, esta não seria uma alavanca a ser trabalhada pelo time de negócio. Por outro lado, pode-se simular uma forte rebaixa de preços em determinadas categorias, ou não ter descontos em outras, ou até mesmo parar de vender uma categoria ou um *tier* especifico de produtos. Também, tem-se como alavanca de negócio a variação do nível de investimentos, ou valores de *NPS*, ou preço das ações, e analisar o impacto que pode causar na receita bruta. Tais fatores foram incluídos nas análises.

Tabela 6 - Medidas de comparação dos modelos múltiplos

Tipo de Modelo	R^2 dados de treino	MSE dados de treino	R^2 dados de teste	MSE dados de teste	Quantidade Variáveis	Qualitativo das Varíaveis
RL com todas as variáveis	0,72272	0,29089	0,66997	0,14265	28	Há representatividade de todas as vertentes
RL com Seleção Automática	0,72291	0,29069	0,66909	0,14303	20	Há representatividade de todas as vertentes
RL com Seleção Manual	0,72299	0,29061	0,66810	0,14346	19	Manutenção de todas as variáveis de anuncios
Regressão Lasso	0,75233	0,25983	0,66934	0,14292	20	Exclusão de variáveis de investimento em anúncios
CART	0,82198	0,18676	0,83234	0,07246	26	Há representatividade de todas as vertentes

Nota: MSE= mean squared error

Para fins de elencar o modelo a ser utilizado pelo time de planejamento estratégico da *ElecKart*, com base na Tabela 6, duas estratégias se destacam. Para fins de exploração das sinergias entre as diferentes variáveis, e potencialização das alavancas de negócio conjuntas, além de previsão dado uma estratégia já definida, deveria ser usado o modelo *CART*, o qual apresenta os menores erros de predição e maior R^2. Para fins elaboração e análise de cenários, projeção de distribuição ótima de investimento para o próximo ciclo e projeção de vendas, seria a regressão linear com seleção automática de 20 variáveis. Ele apresentou um

R^2 preditivo próximo ao dos demais, além do erro quadrático médio também em linha, possui uma menor complexidade pela quantidade de variáveis e as variáveis selecionadas abarcam todas as vertentes de alavancas de negócio, portanto, parece ser a melhor escolha. Além disso, vale destacar que o método *CART* foi menos parcimonioso (contém mais variáveis) que o modelo de regressão linear destacado. O modelo de regressão permite a interpretação direta dos efeitos de cada fator.

4.4 Interpretação dos Modelos de melhor performance

Nesta subseção, será realizada a interpretação das equações dos dois modelos de melhor performance ajustados: a Regressão Linear com Seleção Automática de Variáveis e o modelo *CART* (*Classification and Regression Trees*). Em resumo, enquanto a regressão linear com seleção automática oferece uma solução mais interpretável e ajustada às necessidades de simulação de cenários, o modelo *CART* se destaca por sua flexibilidade em identificar relações não lineares e interações importantes entre variáveis, além de ser um bom preditor para um cenário estruturado. Ambos os modelos, quando combinados, podem fornecer uma visão abrangente para a otimização das alavancas de *marketing*. Porém, a complexidade do modelo *CART* em contexto de vários preditores pode dificultar a sua utilização prática. Comentários mais específicos sobre cada uma das abordagens são apresentados na sequância.

Regressão Linear Múltipla com Seleção Automática de Variáveis

O modelo de regressão linear múltipla com base em mínimos quadrados para ajustar a reta de regressão, foi ajustado com seleção automática de variáveis utilizando o critério *AIC* (*Akaike Information Criterion*) para incluir apenas as variáveis mais relevantes, eliminando aquelas que não contribuem significativamente para a explicação da receita bruta (*GMV*). As variáveis selecionadas no modelo final representam os principais drivers de negócio, como as dimensões investimentos em mídias e fatores externos (valor das ações), descontos aplicados, categorias de produtos vendidos.

Também foi ajustado um modelo utilizando a técnica da Regressão *Lasso*, uma técnica de aprendizado de máquina utilizada para realizar regressão linear, apesar de a penalização da *Lasso* induzir uma seleção automática de variáveis, obtém-se resultados distintos da técnica de regressão linear múltipla clássica, apoiada na seleção automática de variáveis pelo critério *AIC*. A lista de variáveis final de cada uma delas contém variáveis preditoras distintas, e obviamente, resultados ligeiramente diferentes. Neste trabalho, ao ajustar a regressão *lasso*, não se obteve como atributos do modelo variáveis relacionadas aos investimentos em anúncios, e através da observação da Tabela 6, nota-se que os resultados são muito próximos em termos de predição e ajuste. Por isso, optou-se por eleger a Regressão Linear Múltipla Clássica como a mais adequada ao estudo.

A equação gerada pelo modelo pode ser representada genericamente como:

$$\widehat{totalGMV}_t = -0.0016 + (-0.0775 * product\_procurement\_sla) + (0.0189 * flPromocional) + (0.1267 * Camera) + (0.0207 * CameraAccessory) + (0.0199 * Game) + (-0.0086 * GamingAccessory) + (0.0472 * GamingConsole) + (0.0128 * HomeAudio) + (0.0071 * Speaker) + (0.0100 * TVVideoSmall) + (0.0532 * Média) + (-0.0180 * Popular) + (0.8258 * totalVlDescUnit) + (0.0076 * fatorStockRecente) + (-0.0091 * fatorInvTVPassado) + (0.0244 * fatorInvSponsorPassado) + (0.0237 * fatorInvAffiliatesPassado) + (0.0226 * fatorInvAffiliatesRecente) + (-0.0319 * fatorInvMidiaCRMPassado) + (-0.0138 * fatorInvMidiaCRMRecente)$$

onde:

- $\widehat{totalGMV}_t$ representa o valor estimado (média) da variável resposta, a receita bruta padronizada no tempo t.
- -0.0016 é o intercepto do modelo, que contém a informação da categoria AudioMP3Player e *tier* de produtos de alto *ticket*.
- Os valores numéricos multiplicados são os coeficientes das respectivas variáveis explicativas no tempo t (o subscrito t foi suprimido para amenizar notação).

A interpretação dos coeficientes indica o impacto marginal de cada variável explicativa na receita bruta, mantendo todas as outras constantes. Por exemplo, um coeficiente positivo para a variável de investimento em *Afiliados Recente* significa que, ao aumentar os gastos nesse canal em uma unidade, espera-se um incremento proporcional ao coeficiente na média da receita bruta padronizada. Similarmente, o valor negativo, por exemplo, em *investimentos Midia e CRM Recente*, não quer dizer destruição de valor, mas sim que o aumento de uma unidade em tais variáveis tem impacto negativo na média receita bruta padronizada. A seleção automática de variáveis permitiu reduzir a multicolinearidade e simplificar o modelo, garantindo maior robustez preditiva.

Modelo *CART* (*Classification and Regression Trees*)

Com base nos dados disponíveis e usando as parametrizações especificadas na Seção 2.2, ajustou-se o modelo *CART*. A Figura 9 mostra a estimativa do grau de importância de cada variável utilizada. Percebe-se que é visualmente discrepante o nível de importância que a variável *totalVlDescUnit* (que descreve o valor total de desconto aplicado na semana) tem frente as outras preditoras. Portanto, para o contexto da base de dados utilizada, a utilização da informação sobre aplicação de descontos parece ser crucial para previsibilidade de Receita Bruta, destacando-se como uma estratégia de *marketing* muito importante.

A vantagem do *CART* é sua capacidade de capturar interações entre variáveis, como a sinergia entre investimentos em mídia digital e patrocínios. No entanto, esse modelo pode ser menos parcimonioso, pois tende a utilizar um número maior de variáveis e criar regras complexas. Como pode ser observado na Figura 10 o modelo ajustado é complexo, sendo quase inviável a visualização dos particionamentos por envolver muitas variáveis.

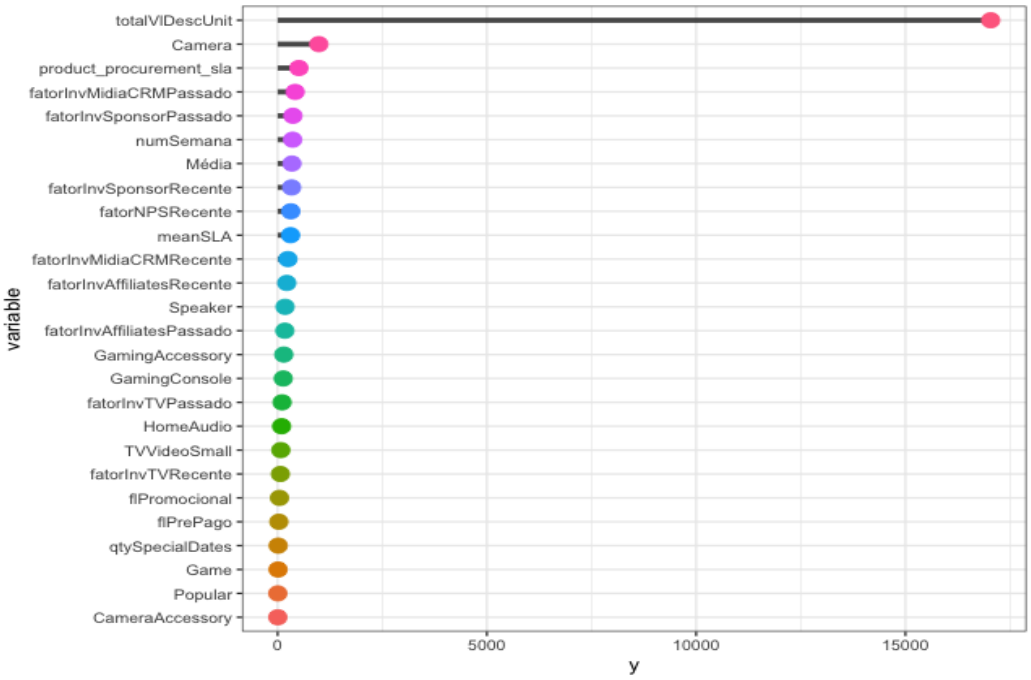

Figura 9 - Importância de cada variável no modelo CART.

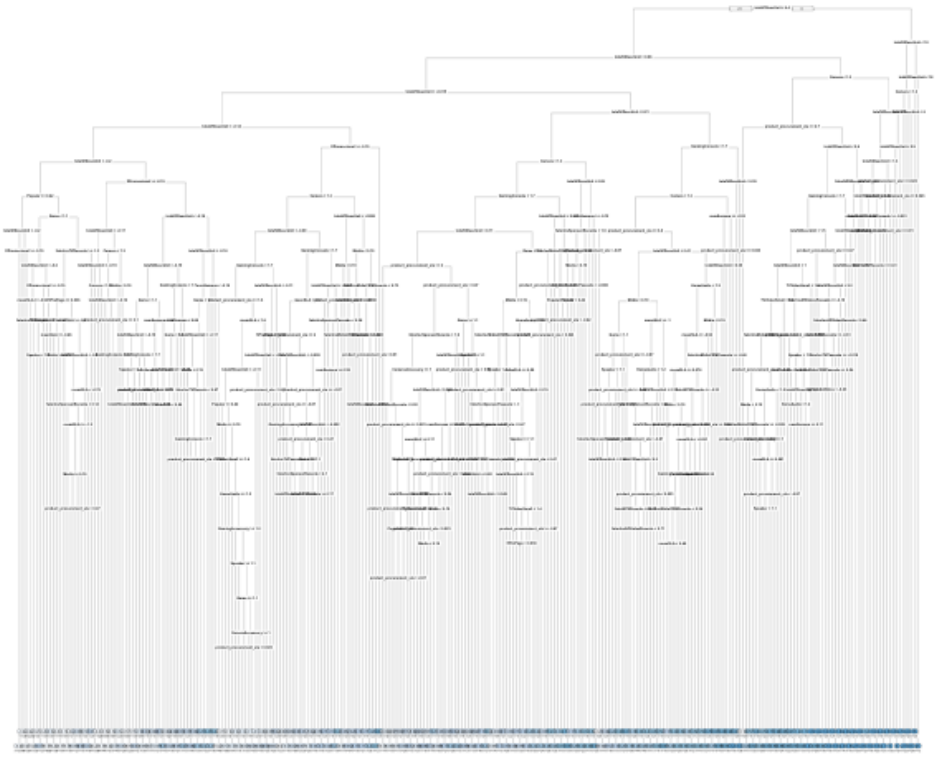

Figura 10 - Modelo *CART* ajustado

5. Considerações Finais

O *Marketing Mix Modeling* (*MMM*) demonstrou ser uma técnica estatística crucial para mensurar o impacto de diferentes alavancas de marketing nas métricas de negócio, como a receita bruta. Ao longo deste trabalho, foram aplicados métodos de regressão linear múltipla e técnicas de ajuste via árvores de decisão, destacando-se a capacidade do *MMM* em identificar relações entre variáveis explicativas e a variável resposta. Os modelos ajustados são capazes de fornecer estimativas de receita bruta (*GMV*) para cada estratégia de *marketing*, seja alterações em investimentos em determinado canal de anúncios, alterações da política de descontos para o período, ou retorno esperado para um período sazonal. Além disso, eles podem ser utilizados para simular cenários e fornecer ferramentas robustas para a equipe de marketing das empresas.

A análise realizada mostrou que a aplicação de modelos estatísticos baseados em regressões é fundamental para captar efeitos sazonais e variáveis exógenas que influenciam o comportamento de vendas. Foi evidenciado que o ajuste de modelos pode ser aprimorado por meio da seleção de variáveis relevantes e da redução de multicolinearidade, problemas comuns em bases de dados utilizadas para modelagem de mix de marketing.

Os principais desafios na construção de um modelo de *MMM* está em modelar relações não lineares entre as covariáveis e a variável alvo. Como exemplo, pode-se citar a necessidade de tratar dependências temporais (*carryover effect*) comuns em campanhas de marketing de longa duração. Neste trabalho, para mitigar tais dependências, lançou-se mão do uso de defasagens temporais (*lags*) em algumas variáveis, melhorando a precisão nas estimativas.

Outro ponto relevante abordado foi a dificuldade de utilizar métodos de aprendizado de máquina (*ML*) no contexto do *MMM*. Embora esses métodos ofereçam alto poder preditivo, ainda está se popularizando técnicas para garantir a interpretabilidade dos seus resultados, como *SHAP* e *LIME*. E este fator é essencial para tomada de decisão em marketing. Modelos como árvores de decisão e regressões penalizadas (*Lasso*) foram explorados para equilibrar precisão preditiva e clareza interpretativa. A regressão *Lasso* resultou em modelo com a mesma quantidade de preditoras do modelo de regressão linear clássico, porém concentrando os atributos em algumas dimensões de negócio, apesar de obter capacidade preditiva praticamente igual nos dados de validação, enquanto o modelo *CART* apresentou maior poder preditivo, mas foi menos parcimonioso.

Há oportunidade de incrementar o trabalho aplicando transformações logarítmicas a fim de tentar modelar outros efeitos não lineares, como a saturação de uma propaganda ou patrocínio (*Diminishing Return*). Outro ponto a incrementar é a exploração de relações sinérgicas entre os preditores, a título de exemplo: testar se as variáveis de investimento em diferentes canais juntas trazem maior poder de predição, ou promoção associado a alguma categoria específica. Além disso, há margem para avaliar a utilização de modelos baseados em técnicas Bayesianas, para a base de dados do trabalho, por ser uma alternativa robusta para cenários de alta incerteza, onde há necessidade de incorporar informações prévias e tratar dados limitados. Também é possível incrementar o trabalhando criando mais

variáveis e testando-as na contribuição para aumentar o poder preditivo do modelo, por exemplo, não foi testado o uso defasagem ou avanço temporal em feriados e datas comemorativas, que em geral podem movimentar receitas para os varejistas, seja previamente, como é o caso do Natal no Brasil, ou após, como o *Boxing Day* no Reino Unido.

Os resultados obtidos reforçam que o *MMM* é amplamente utilizado em indústrias que possuem altos investimentos em publicidade, como varejo, bens de consumo e *e-commerce*. Esses setores se beneficiam de insights quantitativos gerados pelo *MMM* para otimizar seus orçamentos de marketing, identificar canais de alto e baixo *ROI* (*return over Investments*) e prever cenários futuros.

Referências

Babbie, E. (2016). The Practice of Social Research (14th ed.). Cengage Learning.

Borden, N. H. (1964), The Concept of the Marketing Mix. Journal of Advertising Research. Harvard Business School.

Bruce, P. and Bruce, A. (2019) Estatística Prática para Cientista de Dados. Alta Books. ISBN: 978-85-508-0603-7.

Creswell, J. W., & Creswell, J. D. (2017). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches (5th ed.). SAGE Publications.

Fox J, Weisberg S (2019). _An R Companion to Applied Regression_, Third edition. Sage, Thousand Oaks CA. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>.

Friedman J, Tibshirani R, Hastie T (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." Journal of Statistical Software, 33(1), 1–22. doi:10,18637/jss.v033.i01.

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media. ISBN: 978-1-492-03264-9

Giovanni Petris (2010). An R Package for Dynamic Linear Models. Journal of Statistical Software, 36(12), 1-16. URL: https://www.jstatsoft.org/v36/i12/.

Gopinath, R. (2024) (https://medium.com/@mail2rajivgopinath/comprehensive-guide-to-bayesian-marketing-mix-modeling-techniques-parameters-and-practical-b761b366bc31)

Hanssens, D. M., Parsons, L. J., & Schultz, R. L. (2003). Market Response Models: Econometric and Time Series Analysis. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Hegewald, M. (2019) Marketing Mix Modelling Step-by-Step — Part 1: Building the Model (https://medium.com/swlh/marketing-mix-modelling-step-by-step-part-1-702c793d91fd)

Jiang, A. (2022) (https://towardsdatascience.com/marketing-mix-models-102-the-good-the-bad-and-the-ugly-f5895c86b7c3)

Jiang, A. (2022) Marketing Mix Modeling 101 (https://towardsdatascience.com/marketing-mix-modeling-101-d0e24306277d)

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in Data Mining: Formulation, Detection, and Avoidance. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(4), 1-21. DOI: 10.1145/2382577.2382579

Kisilevich, S. (2012) (https://towardsdatascience.com/improving-marketing-mix-modeling-using-machine-learning-approaches-25ea4cd6994b)

Kline, P. (2015). Psychometrics and Psychological Assessment. British Journal of Clinical Psychology, 54(1), 1-10,

Kotler, P., & Keller, K. L. (2012). Administração de Marketing. Pearson Prentice Hall.

Kübler, R. (2023) How to Optimize Your Marketing Budget (https://towardsdatascience.com/how-to-optimize-your-marketing-budget-63707c18ba36)

Kuhn, Max (2008). "Building Predictive Models in R Using the caret Package." Journal of Statistical Software, 28(5), 1–26. doi:10,18637/jss.v028.i05, https://www.jstatsoft.org/index.php/jss/article/view/v028i05.

Kumar, R. (2017) Market Mix Modeling (MMM) — 101 (https://towardsdatascience.com/market-mix-modeling-mmm-101-3d094df976f9)

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30. DOI: 10.48550/arXiv.1705.07874

Meyer F, Perrier V (2024). _esquisse: Explore and Visualize Your Data Interactively_. R package version 1.2.0, <https://CRAN.R-project.org/package=esquisse>.

Milborrow S (2024). _rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'_. R package version 3.1.2, <https://CRAN.R-project.org/package=rpart.plot>.

Mingoti, S. A. (2005) Análise de dados através de métodos de estatística multivariada. Editora UFMG. ISBN 978-85-7041-451-9.

Montgomery, D. C., e Runger, G. C. (2009). Estatística aplicada e probabilidade para engenheiros (4th ed.).

Ni, A. (2024) Market Mix Model (MMM) - 02 Data (https://medium.com/@amy2023/market-mix-model-mmm-data-263cd816c617).

Posit team (2024). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL http://www.posit.co/.

R Core Team (2024). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rahul, P. (2019) Marketing Mix Modelling (MMM) — Factors impacting MMM (Part 3) (https://medium.com/@patnalarahul/marketing-mix-modelling-mmm-factors-impacting-mmm-part-3-a777d348b0c3)

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. DOI: 10.1145/2939672.2939778

Shmueli, G. (2010). "To Explain or To Predict?" Statistical Science, 25(3), 289-310

Therneau T, Atkinson B (2023). _rpart: Recursive Partitioning and Regression Trees_. R package version 4.1.23, https://CRAN.R-project.org/package=rpart

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." _Journal of Open-Source Software_, *4*(43), 1686. doi:10,21105/joss.01686 <https://doi.org/10,21105/joss.01686>.

Wickham H, François R, Henry L, Müller K, Vaughan D (2023). _dplyr: A Grammar of Data Manipulation_. R package version 1.1.4, <https://CRAN.R-project.org/package=dplyr>.

Wickham H, ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Wigren, R & Cornell, F (2020). Markeng Mix Modelling: A comparave study of statistical models. Linköpings universitet. Linköping

William Revelle (2024). _psych: Procedures for Psychological, Psychometric, and Personality Research_. Northwestern University, Evanston, Illinois. R package version 2.4.6, <https://CRAN.R-project.org/package=psych>.

Zeileis A, Hothorn T (2002). "Diagnostic Checking in Regression Relationships." R News, 2(3), 7–10, https://CRAN.R-project.org/doc/Rnews/.

APÊNDICE A: Descrição das variáveis na base de dados bruta

Detalhamento da base de dados ConsumerElectronics.csv: Esta base contém 20 colunas, tem a granularidade de item pedido, ou seja, detalha os SKUs contidos nos pedidos realizados ao longo do período analisado. As variáveis presentes na base são:

- fsn_id contém o código do SKU;
- order_date contém a data e hora da realização do pedido;
- Year contém o ano do pedido;
- Month contém o mês do pedido;
- order_id contém o código do pedido;
- order_item_id contém código do item dentro do pedido, foi descartada de início;
- gmv contém a receita obtida com aquela venda em unidades monetárias, esta é a variável de interesse;
- units contém a quantidade daquele SKU contida no pedido;
- deliverybdays contém o tempo em dias uteis para entrega do pedido, porém esta coluna possui muita sujeira;
- deliverycdays contém o tempo em dias corridos para entrega do pedido, e também possui muita sujeira;
- s1_fact.order_payment_type contem o tipo de pagamento utilizado: COD (abreviação para cash on delivery) quando o pagamento é realizado no momento da entrega, ou Prepaid quando o pagamento é realizado no momento da finalização do pedido online
- sla contém o tempo médio de para entrega do produto;
- cust_id contém a identificação do cliente;
- pincode, foi descartada no início;
- product_analytics_super_category contém o agrupamento macro de produtos, todos os registros possuem o mesmo valor, foi descartada no início;
- product_analytic_category contém o agrupamento de produtos intermediário;
- product_analytic_sub_category contém a subcategoria, um detalhamento maior;
- product_analytic_vertical contém um agrupamento mais granular que a subcategoria, último nível antes do SKU;
- product_mrp contém o preço máximo de venda do produto em unidade monetária, o preço nominal;
- product_procurement_sla contém o tempo médio para a loja adquirir o produto;

Detalhamento da base de dados MediaInvestment.csv: Esta base de dados trás os níveis de investimentos mensais em diferentes canais com o objetivo de atrair clientes, e obviamente vender mais. As variáveis presentes na base são:

- Year contém o ano do investimento realizado;

- Month contém o mês do investimento realizado;
- Total.Investment contém o valor total investido pela ElecKart;
- TV contém o valor investido em inserções na TV;
- Digital contém o valor investido em estratégias de ativação digital (CRM);
- Sponsorship contém o valor investido em patrocinios;
- Content.Marketing contém o valor investido em marketing de conteúdo;
- Online.marketing contém o valor investido em estratégias de mídia online;
- Affiliates contém o valor investido em programas de afiliados;
- SEM contém o valor investido em estratégias de search;
- Radio contém o valor investido em inserções no rádio;
- Other contém o valor investido em outros canais;

Detalhamento da base de dados SpecialSale.csv: Essa tabela contém apenas duas colunas e tem a granularidade de dia, que são:

- Date contém a data que é considerada data sazonal;
- Sales.Name contém a descrição da época sazonal;

Detalhamento da Base de dados Monthly NPS Score.csv: Este conjunto de dados possui três coluans, e trás a posição de fechamento mensal das informações nos meses de referência por meio das seguintes variáveis:

- Date contém a data de referência para as informações;
- NPS contém a informação de Net Promoter Score, índice comum no mercado para indicar o nível de satisfação dos clientes;
- Stock.Index contém a informação do valor das ações da ElecKart, em unidade monetária.

APÊNDICE B: Resultados da Análise Fatorial Exploratória

Quadro A1: Análise Fatorial Exploratória dos Fatores Externos

```
## Principal Components Analysis
## Call: principal(r = X, nfactors = length(colNomesNum))
## Standardized loadings (pattern matrix) based upon correlation matrix
##             RC4   RC1   RC2   RC3   RC5   RC6 h2                 u2 com
## interpNPS    0,98  0,13 -0,03 -0,10 -0,13  0,00  1  0,00000000000000078 1.1
## interpStock  0,14  0,96  0,04 -0,22 -0,03  0,00  1  0,00000000000000144 1.2
## NPSLag1      0,97  0,15  0,09  0,10  0,16  0,00  1  0,00000000000000167 1.1
## NPSLag2      0,21 -0,04  0,97 -0,08  0,10 -0,02  1 -0,00000000000000022 1.1
## StockLag1    0,15  0,94  0,16  0,28  0,02  0,01  1  0,00000000000000200 1.3
## StockLag2   -0,16  0,27  0,94  0,12 -0,10  0,03  1  0,00000000000000022 1.3
##
##                  RC4  RC1  RC2  RC3  RC5 RC6
## SS loadings         2.00 1.92 1.85 0,16 0,06   0
## Proportion Var       0,33 0,32 0,31 0,03 0,01   0
## Cumulative Var       0,33 0,65 0,96 0,99 1.00   1
## Proportion Explained  0,33 0,32 0,31 0,03 0,01   0
## Cumulative Proportion 0,33 0,65 0,96 0,99 1.00   1
##
## Mean item complexity =  1.2
## Test of the hypothesis that 6 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0
##  with the empirical chi square  0  with prob <  NA
##
## Fit based upon off diagonal values = 1
```

Quadro A2: Análise Fatorial Exploratória dos Fatores de Investimento em TV

```
## Principal Components Analysis
## Call: principal(r = X, nfactors = length(colNomesNum), rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##              RC1  RC2   RC3   RC4 h2                u2 com
## invTV       0,42 0,91 -0,01  0,00  1 0,00000000000000011 1.4
## invTVLag1    0,65 0,75  0,09 -0,01  1 0,00000000000000133 2.0
## invTVLag2    0,85 0,52  0,08  0,02  1 0,00000000000000122 1.7
## invTVLag4sum 0,91 0,41 -0,02 -0,01  1 0,00000000000000144 1.4
##
##                    RC1  RC2  RC3 RC4
## SS loadings          2.16 1.83 0,02   0
## Proportion Var       0,54 0,46 0,00   0
## Cumulative Var       0,54 1.00 1.00   1
## Proportion Explained  0,54 0,46 0,00   0
## Cumulative Proportion 0,54 1.00 1.00   1
##
## Mean item complexity =  1.6
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0
##  with the empirical chi square  0  with prob <  NA
##
## Fit based upon off diagonal values = 1
```

Quadro A3: Análise Fatorial Exploratória dos Investimentos em Patrocínio

```
## Principal Components Analysis
## Call: principal(r = X, nfactors = length(colNomesNum), rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                    RC1  RC2   RC3   RC4 h2                u2 com
## invSponsorship       0,36 0,93 -0,01  0,01  1  0,00000000000000033 1.3
## invSponsorshipLag1    0,63 0,77  0,12 -0,01  1 -0,00000000000000155 2.0
## invSponsorshipLag2    0,86 0,50  0,10  0,04  1  0,00000000000000044 1.6
## invSponsorshipLag4sum 0,93 0,37 -0,03 -0,02  1  0,00000000000000078 1.3
##
##                    RC1  RC2  RC3 RC4
## SS loadings          2.13 1.84 0,03   0
## Proportion Var       0,53 0,46 0,01   0
## Cumulative Var       0,53 0,99 1.00   1
## Proportion Explained  0,53 0,46 0,01   0
## Cumulative Proportion 0,53 0,99 1.00   1
##
## Mean item complexity =  1.6
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0
##  with the empirical chi square  0  with prob <  NA
##
## Fit based upon off diagonal values = 1
```

Quadro A4: Análise Fatorial Exploratória dos Investimentos em Afiliados

```
## Principal Components Analysis
## Call: principal(r = X, nfactors = length(colNomesNum), rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                    RC1  RC2   RC4   RC3 h2                 u2 com
## invAffiliates       0,50 0,87 -0,01  0,00  1  0,00000000000000011 1.6
## invAffiliatesLag1    0,65 0,76  0,09  0,02  1 -0,00000000000000178 2.0
## invAffiliatesLag2    0,80 0,59  0,05  0,06  1 -0,00000000000000022 1.9
## invAffiliatesLag4sum 0,86 0,51 -0,01 -0,03  1  0,00000000000000011 1.6
##
##                     RC1  RC2  RC4 RC3
## SS loadings          2.05 1.94 0,01   0
## Proportion Var       0,51 0,48 0,00   0
## Cumulative Var       0,51 1.00 1.00   1
## Proportion Explained  0,51 0,48 0,00   0
## Cumulative Proportion 0,51 1.00 1.00   1
##
## Mean item complexity =  1.8
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0
##  with the empirical chi square  0  with prob <  NA
##
## Fit based upon off diagonal values = 1
```

Quadro A5: Análise Fatorial Exploratória dos Investimentos em Mídia e CRM

```
## Principal Components Analysis
## Call: principal(r = X, nfactors = length(colNomesNum), rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                   RC1  RC2   RC4   RC3 h2                 u2 com
## invMidiaCRM        0,37 0,93 -0,01  0,01  1  0,00000000000000022 1.3
## invMidiaCRMLag1    0,64 0,76  0,12 -0,01  1 -0,00000000000000244 2.0
## invMidiaCRMLag2    0,87 0,49  0,10  0,05  1  0,00000000000000000 1.6
## invMidiaCRMLag4sum 0,93 0,37 -0,03 -0,03  1  0,00000000000000067 1.3
##
##                     RC1  RC2  RC4 RC3
## SS loadings          2.16 1.81 0,03   0
## Proportion Var       0,54 0,45 0,01   0
## Cumulative Var       0,54 0,99 1.00   1
## Proportion Explained  0,54 0,45 0,01   0
## Cumulative Proportion 0,54 0,99 1.00   1
##
## Mean item complexity =  1.6
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0
##  with the empirical chi square  0  with prob <  NA
##
## Fit based upon off diagonal values = 1
```