ANDRÉ LUIZ TIAGO SOARES

DESEVOLVIMENTO DE MODELO DE DEMANDA PARA VAREJISTA DO SETOR MOVELEIRO

São Paulo
2021

ANDRÉ LUIZ TIAGO SOARES

DESEVOLVIMENTO DE MODELO DE DEMANDA PARA VAREJISTA DO SETOR MOVELEIRO

Trabalho de Formatura apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do diploma de Engenheiro de Produção.

São Paulo
2021

ANDRÉ LUIZ TIAGO SOARES

DESEVOLVIMENTO DE MODELO DE DEMANDA PARA VAREJISTA DO SETOR MOVELEIRO

Trabalho de Formatura apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do diploma de Engenheiro de Produção.

Orientador: Profa. Dra. Linda Lee Ho

São Paulo

2021

Às pessoas com quem pude contar tanto quanto pude contar com o sol nascer todo dia: Mãe, Pai, Irmã, obrigado por sempre estarem lá por mim :-)

AGRADECIMENTOS

Eu sou uma pessoa que verdadeiramente tenho muito pelo que ser grato. Agora, nesse momento de transição acredito que seja a hora de abrir meu coração e confessar que os últimos sete anos, minha jornada de universitário, não foram fáceis. Tive que travar uma série de batalhas com demônios internos, e por mais de uma vez me vi em lugares sombrios.

Contudo, quando eu estava no que eu sem dúvida alguma posso considerar o mais difícil desses momentos, alguém me disse que "Tudo ia ficar bem". Essa frase, por mais simples que seja, teve um efeito poderoso sobre mim. De certa forma me trouxe para o presente. Quero dizer, tirou minha cabeça de preocupações abrangentes sobre o futuro, e me fez desapegar de inseguranças sobre passado.

Obviamente esse efeito não foi imediato. A frase foi tranquilizadora um primeiro momento, mas só depois de algum tempo de contemplação dela que eventualmente entendi que, mesmo se as coisas não se desenvolvessem da melhor maneira, ainda assim "Tudo ia ficar bem". O melhor que eu podia fazer era trazer meu foco para o momento presente e para as coisas sobre as quais eu tinha controle.

E alguns anos depois, no final das contas, parece de fato que as coisas de fato vão acabar bem.

Preciso agora mais que nunca ser gentil comigo mesmo, e agradecer ao André de 2017 por ter tomado as escolhas corretas no momento correto, por ter tido a coragem para superar adversidades e por, no frigir dos ovos, ter escolhido alimentar o seu lado mais nobre. Se eu pudesse ter uma conversa com aquela minha versão mais jovem, diria que seus esforços valeram a pena.

Mas por mais que eu esteja orgulhoso de tudo o que eu fiz nos últimos anos, não teria conseguido muito sem o suporte de certas pessoas.

Vou começar agradecendo ao ex-diretor Prof. Dr. José Roberto Castilho Piqueira e aos membros do Poli Recicla por ter acreditado em mim e me dado uma chance quando eu mais precisava. Gostaria de ter a honra de um dia fazer por alguém o que vocês fizeram por mim.

À Profa. Dra Linda Lee Ho, minha orientadora, obrigado por ter encontrado tempo para mim entre pesquisa, elaboração de provas, orientação de outros alunos e confecção belíssimos enfeites de crochê para o natal. Saiba que já nesse pouco de interação que tivemos durante a construção desse trabalho, a senhora já se tornou um exemplo para mim, pelo seu imenso nível de conhecimento, e também pela sua paciência, honestidade e principalmente pela paixão que você demonstra pelos seus alunos.

Agradeço também à Escola Politécnica como um todo. Na Escola, cresci tremendamente

não apenas em termos de conhecimento acadêmico, mas também em termos pessoais. Por causa dos desafios constantes propostos pela Escola, eu tive que conhecer a mim mesmo, e eu sou grato por isso.

Fora da escola, gostaria de agradecer à Mobly, primeiro pela oportunidade de estágio durante o último ano. Durante o último ano, tive a sorte de participar de um ambiente de trabalho engajante, e pude atuar com liberdade autonomia.

Vindo para minha vida pessoal, tenho que agredecer acima de tudo aos meus pais, Maria do Carmo e José. Desde criança, o lar construído por vocês sempre foi para mim um lugar de refúgio, um lugar em que eu sempre me senti seguro, em paz e em liberdade. O que eu tive durante toda minha vida é um privilégio por causa de vocês.

Mãe, admiro muito você por sempre encontrar satisfação pelas coisas simples, por manter uma boa disposição em qualquer situação e por ser absolutamente a pessoa mais cheia de amor que eu conheci. Você é uma mulher corajosa e que me inspira a ser corajoso quando eu preciso ser.

Pai, convenhamos que você é uma pessoa difícil de lidar, mas sempre esteve lá por mim quando eu precisava, inclusive para me levar para rua quando eu queria ir jogar bola na chuva. Sempre pude contar com você como pude contar que sol ia nascer de manhã. Nunca vou esquecer tudo o que você já fez por mim, e eu te amo.

Por fim, gostaria de agradecer também aos melhores amigos da minha vida, Ana Luiza, Matheus Duarte, Camillo Tiago, Brenno Enrico, Nathalia Reis, Isabella Tiago, Guilherme Rosa, Matheus Silva e Gabriel Segers. Vocês são todos grande parte da minha vida. Sem vocês a vida definitivamente seria mais difícil.

"It is said an Eastern monarch once charged his wise men to invent him a sentence, to be ever in view, and which should be true and appropriate in all times and situations. They presented him the words: "And this, too, shall pass away."How much it expresses! How chastening in the hour of pride! How consoling in the depths of affliction!"

(Lincoln, 1859)

"Tudo vai ficar bem"

(Dito por um amigo em um dia ruim)

RESUMO

Nesse trabalho é desenvolvido um modelo da demanda para uma empresa varejista do setor moveleiro focada no canal online. A variável de interesse é a receita bruta diária de uma categoria de produtos para dados observados entre 01/01/2018 e 31/03/2020. É importante para uma empresa ter esse tipo de modelo para ter um melhor entendimento de quais fatores que afetam a demanda. Esse conhecimento é útil para orientar ações que afetem a demanda. O trabalho começa com contextualização do problema dentro da empresa. Em seguida, é realizada uma análise exploratória de dados para identificar quais variáveis explicativas têm relação com variável resposta e quais variáveis são correlacionadas entre si. Das variáveis inicialmente apresentadas no conjunto de dados, é definido e especificado um modelo final que utiliza como variáveis explicativas, a quantidade de visitas diárias a páginas de produto, calculadas a partir da data e eventos de calendário representados por variáveis *dummy* que dizem respeito ao dia da semana e proximidade da *Black Friday*. Os parâmetros do modelo são estimados através do Método dos Mínimos Quadrados. Sobre diagnóstico do modelo utiliza-se um conjunto de métodos gráficos e estatísticas descritivas para demonstrar que o modelo apresentado satisfaz suposições do modelo de regressão linear múltiplas quanto aos resíduos e valores previstos. Com um modelo bem ajustado em mãos, uma interpretação do modelo para obtenção de *insights* sobre a demanda pode ser feita. Constata-se que a variável que melhor prevê a demanda é o número de visitas às paginas de produto, e seu coeficiente pode ser interpretado como uma taxa de conversão. As outras variáveis tem impacto sobre a demanda depois de já consideradas o número de visitas. É mostrado que essas variáveis estão relacionadas com diferentes taxas de conversão.

Palavras-chaves: Previsão de Demanda. Varejo. E-commerce. Setor moveleiro. Ciência de dados. Machine Learning.

ABSTRACT

In this work it is developed a model of demand for a retail company of the furniture sector focused on the online channel. The response variable is the daily gross revenue of a product category for data between 01/01/2018 and 31/03/2020. It is important for a business to have these kind of models to have a better understanding of the factors that affect demand. This knowledge is useful to guide action that may affect sales. The work starts at chapter 1 with the contextualization of the problem within the company. Next, in chapter 2 it is described the general multiple linear regression model that was utilized for the final model. After that, in chapter 4 it is done an exploratory data analisys to identify which explanatory variables are related to the response variable and which variables are correlated to each other. Of the variables initially presented in the dataset, it is defined and specified a model that utilizes as explanatory variables the quantity of visits to product pages, calculated variables based on date, and calendar events corresponding to weekdays and proximity to the *Black Friday*. The parameters of the model are estimated by the Least Squares Method. In the chapter about model diagnostics, chapter 5, it is utilized a combination of graphical methods and summary statistics do demonstrate that the presented model satisfies necessary assumptions of the multiple regression model about the residuals and predicted values. With a well fitted model in hands, in chaper 6, it is done the interpretation of the model for obtaining insights about the demand. It is verified that the single best predictor variable is the number of visits to product pages, and that it's coefficient can be interpreted as a conversion ratio. The other variables still impact the demand even after the number of visits is already considered. It is shown that these variables are related to variation of the conversion ratio.

Palavras-chaves: Demand forecasting. Retail. E-commerce. Furniture sector. Data science. Machine Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Definição do problema

O problema abordado nesse trabalho é a realização de uma análise de regressão a respeito das vendas de uma empresa varejista do setor de móveis, que é focada principalmente no canal online.

O objetivo da análise de regressão é investigar a relação funcional entre variáveis, mais especificamente, no relacionamento de um conjunto de variáveis explicativas sobre uma variável de interesse, que é chamada de variável resposta, ou variável dependente.

Um bom modelo de regressão explica sobre qual parte da variável de interesse pode ser explicada por efeitos genuínos que podem ser capturados a partir de dados disponíveis e qual componente é causada por variações aleatórias.

Dentro do contexto de varejo, existem diversos problemas que podem ser resolvidos através da análise de regressão. Um desses problemas é identificar as principais variáveis que têm efeitos sobre a venda.

Toda empresa tem o interesse de crescer, o que significa tomar ações com o intuito de fomentar o crescimento e o aumento de vendas. Surge naturalmente a pergunta: **Quais os principais fatores que afetam vendas?**. A resposta para essa pergunta é essencial para o direcionamento eficiente de ações a fim de gerar crescimento.

Através de modelos de regressão das vendas, é possivel fornecer inteligência de negócio crucial e, se dados estiverem disponíveis, é de interesse da empresa que essa análise seja realizada.

O modelo desenvolvido irá explicar a demanda diária de uma categoria de produtos a partir de dados internos disponíveis em bancos de dados da empresa.

Antes do detalhamento do escopo e objetivo do trabalho, será realizada uma descrição da empresa e contextualização de onde o trabalho foi desenvolvido

1.2 Contextualização e entendimento do negócio

A empresa em que o trabalho foi desenvolvido é na Mobly Comércio Varejista Ltda (MOBLY, 2021b) que será referida no trabalho como Mobly. Nessa seção serão descritas as principais características do negócio com o objetivo de elucidar o contexto em que o trabalho foi desenvolvido.

Será apresentado o modelo de negócios da Mobly, a descrição organizacional e o cenário do mercado de varejo de móveis em que a empresa está inserida. Primeiramente, será apresentado

uma visão geral da empresa.

Visão geral

A Mobly é uma empresa de varejo eletrônico, mais especificamente do nicho de móveis e focada principalmente no canal online. Foi fundada em 2011, pelos três fundadores Victor Noda, Marcelo Marques e Mário Fernandes, com investimento de fundos institucionais. O principal investidor, e grupo controlador, é o grupo alemão Rocket Internet, que posteriormente fez uma reestruturação corporativa e agregou todo portfólio de casa e decoração na *holding* Home24 (MOBLY, 2021b).

Em 2019, Mobly diversificou seus canais de venda com a abertura da primeira loja física em São Paulo. Hoje, a empresa continua aumentando a participação do canal físico, embora continue priorizando o canal digital. No ano seguinte, em março, a empresa abriu o capital, através da oferta pública inicial (IPO) na B3 (MOBLY, 2021a).

É uma empresa de grande porte e crescimento acelerado. Expressando o porte da empresa em termos de receita, entre 2019 e 2021, a receita bruta real (corrigida pelo IPCA) apresentou um CAGR (*Compound Annualized Growth Rate* - Crescimento anual médio) de aproximadamente 35%. Hoje, a receita anual e a capitalização de mercado estão em torno de R$ 700 MI (MOBLY, 2021a). Contudo ainda é um player pequeno se comparada com outras grandes empresas que operam no mercado de varejo eletrônico. A Magazine Luiza, como referência, tem receita anual na ordem de 4,5 bilhões de reais (MAGAZINE LUIZA, 2021).

Agora, será realizada uma análise mais detalhada da empresa, partindo de uma análise dos modelos negócios, que consiste na proposta de valor da empresa e o plano através do qual esse valor é alavancado para obtenção de resultado financeiro.

Os modelos de negócio

As atividades da Mobly podem ser separadas e descritas por dois modelos de negócio distintos: o varejo e o marketplace.

O negócio de varejo consiste na aquisição dos produtos do fornecedor e na revenda para os clientes finais através dos canais de venda da varejista, com uma margem sobre o custos de aquisição da mercadoria, logística e marketing. Já no modelo de marketplace, os fornecedores anunciam seus produtos no *site* da empresa e vendem diretamente para o cliente final, pagando para Mobly uma taxa sobre o valor de venda.

A proposta de valor para os dois modelos é similar. Tanto o varejo quanto o marketplace são formas de realizar a intermediação informacional e logística entre o fabricante de um produto e seu consumidor final. Para que a varejista justifique sua viabilidade como um negócio gerador de valor (que será posteriormente monetizado), é necessário gerar valor para as duas pontas do intermédio. As perguntas que tem que ser respondidas são: **"Por que o fornecedor vende através da varejista em vez de procurar o cliente diretamente?"** e **"Por que o cliente**

compra através da verejista em vez de procurar o fornecedor diretamente?".

Do lado do fornecedor, o que a varejista oferece é o acesso ao mercado consumidor, tanto do ponto de vista informacional ("Os consumidores irão estar cientes e conhecer o meu produto") quanto do ponto de vista logístico ("Serei capaz de distribuir o meu produto"). Para o cliente final, o motivo de procurar um intermediário para a realização de suas compras está relacionado com a acessibilidade, variedade e eficiência da compra. Tanto o *marketplace* quanto o varejo satisfazem esses critérios.

A principal similaridade dos dois modelos são os recursos e atividades chaves, que são essencialmente os mesmos. As atividades chave são: a logística, o marketing e o desenvolvimento de tecnologia, e os recursos chaves que possibilitam essas atividades são: a plataforma tecnológica (site e aplicativo), a marca e a rede de distribuição. Tanto esses recursos quanto as atividades são íntimamente relacionados à proposta de valor de realizar a ponte entre os fabricantes e o mercado consumidor. E fora o custo de aquisição da mercadoria do varejo, a estrutura de custos também é essencialmente a mesma: a logística, o marketing e os custos admnistrativos.

Quanto às diferenças, as mais marcantes são que no varejo os fornecedores são parceiros chave e a fonte de receita é proveniente da venda de produtos. No marketplace os fornecedores também são clientes, e a fonte de receita é originária de taxas sobre o preço de venda.

Um corolário disso é a receita da empresa em termos da escala de operação. No varejo, como a receita é proveniente do preço de venda dos produtos, a receita da empresa será da mesma ordem de grandeza de sua escala de operação (R$ 1 MI de mercadorias vendidas corresponderá a R$ 1 MI de receita bruta), enquanto para o *marketplace*, por obter sua receita através de taxa sobre o preço de venda, sua receita será apenas uma fração da escala de operação (R$ 1 MI em mercadorias vendidas corresponderão, talvez, a R$ 200 mil de receita bruta).

Inclusive por isso, no contexto de marketplace, é utilizado um outro indicador além da receita para medir a escala de operação: o *Gross Merchandise Value*, ou GMV, que é o total, em unidade monetária, de mercadorias vendidas atráves do *marketplace*. Esse indicador também poderia ser utilizado no contexto de varejo, mas normalmente não é, pois valores de receita costumam informar satisfatoriamente a escala de operação da empresa para esse modelo.

Também há algumas outras diferenças de menor importância em outras componentes dos modelos. Em linhas gerais, o modelo de varejo oferece maiores retornos sobre o GMV, já que a margem de contribuição do varejo é maior que as taxas de marketplace, mas em contrapartida inclui um escopo maior de atividades (e consequentemente custos) relacionados à negociações, marketing e atendimento ao cliente.

Figura 1 mostra uma representação gráfica do modelo de negócios da Mobly.

Entendendo os modelos de negócio, o próximo passo é compreender como foi construída uma estrutura oraganizacional que executa esse modelo. Essa estrutura organizacional será descrita em sequência.

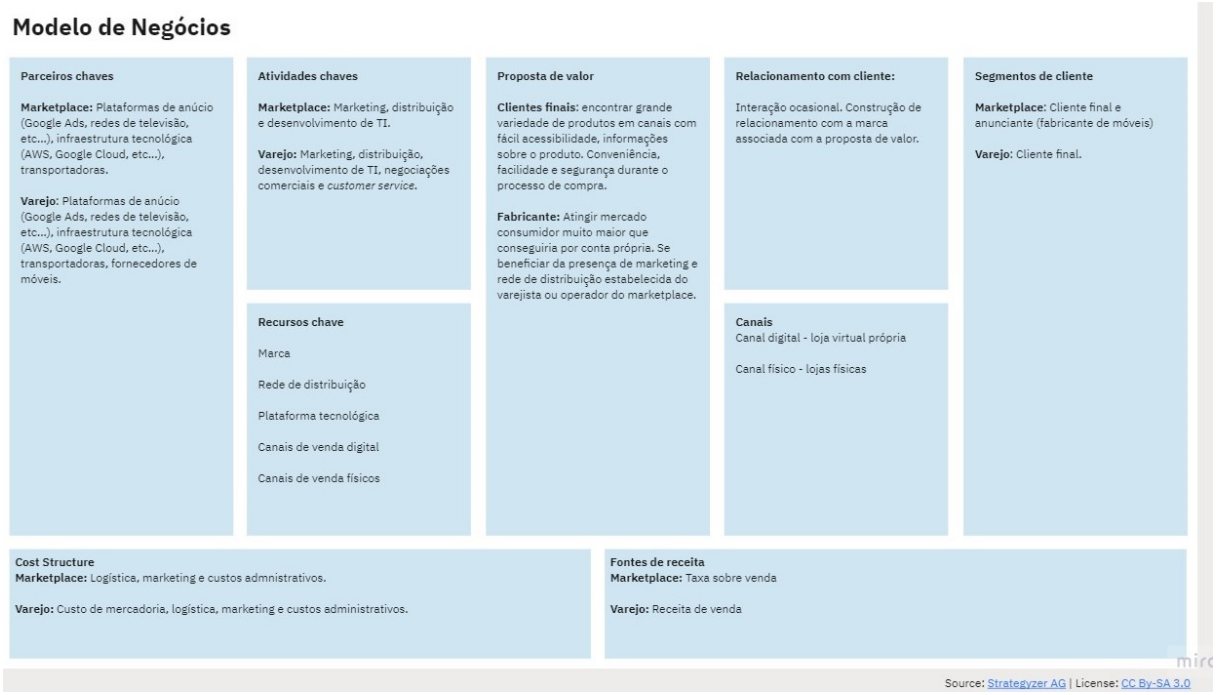

Figura 1 – *Business Model Canvas* da Mobly

Estrutura organizacional da empresa

Os departamentos da empresa podem ser separados em quatro grupos: áreas de tecnologia, áreas de negócios, áreas logísticas e áreas financeiras e de gestão de pessoas. Nos próximos parágrafos é apresentada uma descrição breve de todos departamentos da Mobly e suas atividades.

a- Áreas de Tecnologia

Tecnologia da Informação - TI: Responsável pelo desenvolvimento de aplicações de tecnologia para diversas áreas da empresa. Subdivida em *squads* (pequenos times) que tratam dos diversos produtos em desenvolvimento.

Infraestrutura: Responsável pelo desenvolvimento e manutenção da infraestrutura física de tecnologia e ambientes em nuvem; segurança da informação; sistemas internos e atendimento a colaboradores.

Business Intelligence - BI: Responsável pela construção de pipelines de dados (disponibilização de dados para empresa) e fomentação de inteligência de negócio através de análise de dados e ciência de dados.

b - Áreas de Negócios

Marketing: Responsável pela elaboração de conteúdo criativo e gerenciamento de canais de marketing, com o intuito de atrair potenciais clientes para o site e lojas físicas.

Marketplace: Gerenciamento das atividades de marketplace. Isso inclui o *marketplace in*, que é atrair produtos de anunciantes para o marketplace da Mobly e o *marketplace out*, que é a venda de produtos Mobly em outros marketplaces.

CVO (Comissão de Vendas Offline): Responsável pela maioria das atividades relacionadas a lojas físicas, desde a implantação até a operação das lojas.

Comercial: Responsável pela prospecção e negociação com parceiros para fornecimento de SKUs para as lojas e marketplaces; criação de produtos exclusivos e definição de sortimento de lojas físicas.

Supply Chain: A área de Supply Chain é responsável pela gestão operacional dos fornecedores, compreendendo os processos de abastecimento, gestão de pedidos e estoques de CDs (Centros de Distribuição) e lojas.

Jurídico: Acompanhamento de ações judiciais, elaboração e análise de contratos e documentos a serem celebrados pela empresa.

Planejamento - CM: Análise do desempenho de portfólio. Realiza a precificação e seleção para campanhas promocionais de produtos.

Business-to-business - B2B: Responsável pelas vendas corporativas para outras pessoas jurídicas (escritórios, hoteis, etc...). Lida com a propecção de clientes, negociação e acompanhamento dos pedidos.

Logística Armazém: As atividades do armazém incluem o recebimento de pedidos, a gestão de estoque, o faturamento expedição de produtos para o cliente final e o recebimento de itens de reversa.

Transportes: Cuida da movimentação de mercadorias tanto nas direções do fornecedor ao centro de distribuição quanto do centro de distribuição até o cliente final.

CDC (Centro do Cliente): Atendimento aos clientes, anunciantes do marketplace e acompanhamento da percepção da empresa em redes sociais.

c - Áreas Financeira e Gestão de Pessoas

Financeiro: Lida com a compra de alguns insumos e a contratação de alguns serviços; as questões fiscais; gestão do fluxo de caixa e meios de pagamento; planejamento financeiro da empresa.

Relação com Investidores: Relacionamento comunicação entre a empresa, acionistas e agentes reguladores.

Gente e Gestão: Responsável pela contratação de funcionários, definição de remuneração, bônus, benefícios e comunicação interna da empresa.

Beneficiamento: Departamento de industrialização própria da Mobly. Esse departamento está diretamente envolvido na gestão do processo de produção de produtos de desenvolvimento próprio em fábricas parceiras.

Uma organização, contudo, não existe isolada. A empresa existe dentro de um ecossistema de negócios relacionado a um ramo específico do mercado de varejo, que é o *ecommerce*

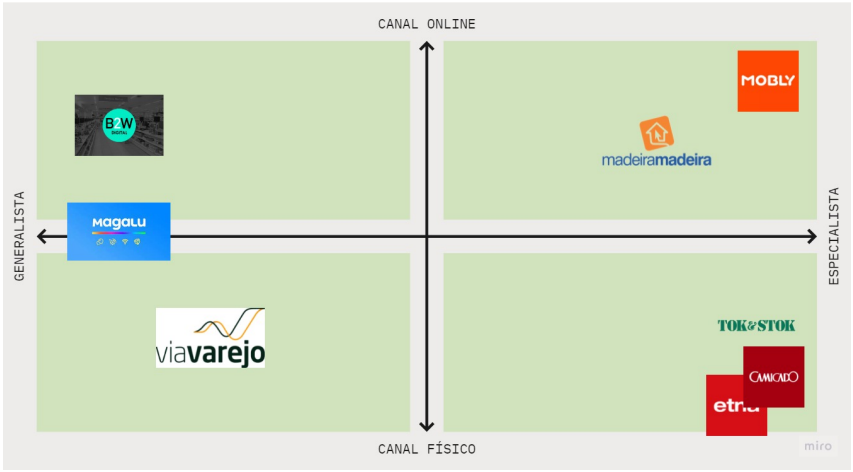

Figura 2 – Posicionamento estratégico dos principais *players* do varejo de móveis

de móveis e decoração. Em seguida, é realizada uma análise da posição da empresa dentro desse contexto de mercado.

O mercado e o contexto estratégico

O mercado de varejo eletrônico de móveis é considerado um mercado fragmentado. Segundo analista da Suno Research (RIVAS, 2021), os maiores 5 maiores *players* de mercado concentravam apenas 13% de marketshare.

Os players desse mercado podem ser divididos em duas categorias. Em primeiro lugar, temos varejistas de menor porte, especializadas na categoria de móveis e decoração. Os principais *players* que se enquadram nessa categoria, incluem MadeiraMadeira, Tok&Stok e WestWing além da própria Mobly.

Contudo, a competição no setor vai muito além disso, pois também atuam nesse mercado grandes grupos varejistas, mais generalistas, ou seja, que não são focados no nicho de móveis especificamente. Nessa categoria se enquadram Magazine Luiza, Via Varejo e B2W (MAGAZINE LUIZA, 2021; VIA VAREJO, 2021; AMERICANAS S.A, 2021).

Como eixo secundário para avaliação do posicionamento estrategico, pode ser considerado o canal focal (online ou físico) ou o público alvo. Das empresas especializadas no nicho de móveis e decoração, WestWing e Tok&Stok, ambas tem suas vendas concentradas pelo canal físico e tem como público alvo as classes A e B (FONSECA, 2021). Já a Mobly, juntamente com MadeiraMadeira, são focadas no canal online e tem como público alvo as classes média e baixa.

A representação gráfica do posicionamento estratégico dos principais players está na Figura 2.

Isso conclui a descrição da empresa. Agora, será discutido em mais detalhes o escopo e objetivo do modelo que será desenvolvido.

1.3 Objetivo e delimitação de escopo

O escopo do modelo é delimitado por quatro restrições: a unidade de negócio, o canal de vendas, a categoria e o período. Serão estudadas as vendas de varejo online, da categoria de Guarda-Roupas ocorridas no período entre 1 de Janeiro de 2018 até 31 de Março de 2020.

A decisão sobre a limitação às vendas de varejo justifica-se pelo fato da variável resposta do modelo ser a receita bruta, que tem fontes distintas no varejo e no *marketplace*. No varejo, a fonte de receita é a venda ao cliente final, enquanto no *marketplace* a receita é proveniente de taxas cobradas aos anunciantes do site. Isso significa que a escala de receita é diferente entre um modelo e outro. Por exemplo, a venda de um item cujo preço seja R$1000,00 será indicado como resposta em R$1000,00 de receita pelo modelo de varejo, mas somente uma fração disso pelo modelo de *marketplace*.

Dos canais de venda, limitou-se ao canal de venda online porque é razoável supor que os fatores que afetam a demanda sejam radicalmente diferentes entre um canal e outro. Digamos, hipoteticamente, que um dia chuvoso afete negativamente as vendas em loja física mas afete positivamente o canal online. Para incorporar as vendas dos dois canais em um único modelo, seria necessário mudar a granularidade da amostra de dades de treinamento, para separar, dia-a-dia, as vendas online das vendas em loja, criar uma variável categórica para indicar o canal de vendas e provavelmente modelar as interações mais importantes entre o canal de vendas e as outras variáveis exógenas. A essa altura, seria mais simples e justificável criar dois modelos diferentes para cada canal de vendas. O desenvolvimento de um modelo para lojas físicas, ou mesmo a extensão desse modelo para o caso geral pode ser um objeto de trabalho futuro, mas no momento optou-se pela abordagem de um caso reduzido simplificado para explorar o potencial da ferramenta.

Além disto, a restrição a uma única categoria justifica-se principalmente por questões computacionais. O conjunto de dados utilizado é uma agregação diária de um conjunto de dados original com granularidade a nível SKU. Embora o dataset final depois da agregação tenha o mesmo tamanho independentemente da quantidade de categorias incluídas, a extração e preparação dos dados a partir do conjunto original seria inviável. O *dataset* original tem granularidade de venda por SKU por dia, com dezenas de milhares de SKUs ativos em todas as categorias, a tabela completa para todas as categorias alcançaria facilmente milhões de linhas e teria tamanho da ordem de alguns gigabytes de memória, definitivamente mais do que pode habilmente ser armazenado na mémoria RAM de um único computador típico. O processamento desse volume de dados deveria idealmente ser realizado por um *cluster* máquinas trabalhando em paralelo.

Se adotada a suposição que, desde que a análise esteja limitada a uma única unidade de negócios e canal de vendas, e que os principais fatores que afetam a demanda não sejam radicalmente diferentes entre uma categoria e outra, os passos executados nesse trabalho poderiam

ser replicados para outras categorias.

O objetivo do trabalho é desenvolver um modelo que seja útil para identificar quais fatores afetam a demanda. A principal utilidade do modelo não será proveniente das previsões geradas, mas dos *insights* que podem ser extraídos dos parâmetros do modelo final e suas interpretações. Por isso, inclusive, que há a preferência pelo modelo de regressão linear, cujo um dos pontos fortes é a interpretabilidade dos coeficientes de regressão parcial.

O modelo final deve expandir o entendimento da dinâmica de vendas e ser utilizado como uma ferramenta para auxiliar na decisão e orientar ações que possam, direta ou indiretamente, impactar vendas.

1.4 Roteiro do trabalho

Este trabalho será estruturado da seguinte maneira. No Capítulo 2 é feita uma breve revisão bibliográfica da literatura sobre modelos de regressão linear múltipla, que é a técnica utilizada para o desenvolvimento do modelo. Nesse Capítulo, é descrita a forma do modelo, quais as suposições subjacentes do modelo e métodos estatísticos e gráficos para realização do diagnóstico de regressão a fim de avaliar quando o modelo está propriamente ajustado.

No Capítulo 3, são discutidos o ferramental e recursos tecnológicos em geral que podem ser utilizados para realizar os passos de extração dos dados, prepraração dos dados, especificação e estimação dos parâmetros do modelo e também diagnóstico da regressão.

No Capitulo 4, é realizada a análise exploratória do conjunto de dados, calculado a partir do conjunto de dados original, que contém a variável resposta e as variáveis explicativas. Na fase de análise exploratória, são identificadas quais das potenciais variáveis exógenas tem maior poder explicativo sobre a variação da variável resposta.

Tendo anotado quais variáveis explicativas tem maior poder preditivo, a especificação do modelo final é realizada através de um processo iterativo de especificação do modelo, ajuste aos dados e diagnóstico dos resultados, até a obtenção de um modelo satisfatório. Após a obtenção do modelo final é feito o diagnóstico deste modelo de regressão. Estes passos são realizados no Capítulo 5.

No Capítulo 6, com um modelo bem ajustado, estamos em posição de realizar a interpretação dos coeficientes do modelo obtido. Essa interpretação consiste na tradução dos coeficientes para ou grandezas da vida real ou noções intuitivas. É a partir dessa interpretação que iremos ver o que um modelo de previsão pode nos ensinar sobre como funciona a dinâmica de vendas do varejo online de móveis.

Por fim, no Capítulo 7, é apresentada a conclusão do trabalho, onde é feita uma discussão a respeito os benefícios dos benefícios derivados da análise de regressão sobre as vendas, como essa técnica pode aumentar competência de tomada de decisão estratégica da empresa, o que os

resultados do modelo sugerem a respeito de como ações devem ser direcionadas para o aumento de vendas e quais os próximo passos podem ser tomados para a extensão da utilidade do modelo.

2 UMA BREVE REVISÃO BIBLIOGRÁFICA SOBRE ANÁLISE DE REGRESSÃO MÚLTIPLA

Nesse Capítulo é feita uma breve revisão da literatura disponível sobre modelos de regressão linear múltipla, diagnóstico de regressão e seleção de modelos. O modelo de regressão linear múltipla é uma extensão do modelo de regressão linear simples (que observa apenas uma variável explicativa) para múltiplas variáveis.

O modelo de regressão linear múltipla é utilizado para prever uma variável resposta a partir de várias variáveis explicativas. Nesse modelo, a estimativa da variável resposta é calculada a partir da somatória dos produtos dos valores das variáveis explicativas pelos seus respectivos coeficientes de regressão parcial. Esses coeficientes, que podem ser chamados também de parâmetros do modelo, são valores constantes que são estimados a partir do ajuste do modelo aos dados disponíveis (CHATTERJEE; HADI, 2012).

Primeiramente, é expressado formalmente qual a forma do modelo de regressão linear múltipla. Ao ajustar um modelo de regressão linear múltipla, estamos também adotando algumas suposições sobre as variáveis explicativas e sua relação com a variável resposta, notoriamente, que a relação é linear. Estatísticas de teste calculados sobre o modelo também costumam assumir algumas suposições sobre os resíduos do modelo. Essas e outras suposições são descritas logo depois de declarada a forma geral do modelo.

Em seguida, é tratado como os parâmetros do modelo (coeficientes de regressão parcial) podem ser estimados a partir de uma amostra de dados usando o Método dos Mínimos Quadrados, e também como deve ser realizada a interpretação dos coeficientes dentro do contexto do problema.

Depois disso, são apresentados estatísticas descritivas, testes de hipótese e métodos gráficos que pode ser utilizados para avaliar se podemos aceitar o modelo ajustado com um certo nível de significância estatística. Esses métodos gráficos devem ser aliados a estatísticas descritivas para discernir sobre quando chegamos a um modelo satisfatório e, se não chegarmos, sugerir ações remediadoras, em termos de exclusão, inclusão ou transformação das variáveis repostas.

O desenvolvimento de modelos de regressão também necessita a escolha de variáveis que entram no modelo de regressão. Existem procedimentos iterativos de especificação, ajuste e avaliação do modelo que já estabelem critérios a respeito de variáveis para serem incluídas ou excluídas do modelo até a obtenção de um modelo. Esse Capítulo é encerrado com a apresentação de algumas dessas técnicas.

2.1 Regressão Linear Múltipla

Descrição do modelo

O modelo de regressão linear múltipla é uma extensão do modelo de regressão linear simples. Esse é um modelo descritivo da variável Y, utilizando as k variáveis explicativas $X_1, X_2, ..., X_k$. O modelo de regressão linear simples é o caso limite em que $k = 1$.

De acordo com Chatterjee e Hadi (CHATTERJEE; HADI, 2012), vemos que o modelo para uma observação Y_i pode ser expresso pela equação (1).

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i = \beta_0 + \sum_{j=1}^{k} \beta_i X_{ji} + \varepsilon_i \quad (1)$$

O modelo descrito em (1) assume algumas suposições que serão exploradas na próxima Seção.

2.2 Suposições do modelo de regressão

Ao adotar o modelo de regressão linear múltipla, também estamos assumindo suposições que não foram até aqui formalmente declaradas. Chatterjee e Hadi (CHATTERJEE; HADI, 2012) declara que as suposições podem ser separadas em suposições sobre a forma do modelo, sobre os resíduos e sobre as variáveis explicativas.

a - Suposições sobre a forma do modelo

Segundo (CHATTERJEE; HADI, 2012), supõe-se que a variável resposta está linearmente relacionada com as variáveis explicativas. Essa suposição é chamada de *suposição de linearidade*.

A confirmação dessa suposição é feita antes da determinação do modelo através de análise gráfica por meio de gráficos de dispersão. Para um modelo de regressão linear simples, um gráfico de dispersão entre a variável resposta Y e a variável explicativa X é suficiente para verificar dessa suposição. No entanto, essa verificação é mais difícil em modelos de regressão linar múltipla, especialmente se o número de variáveis respostas for muito alto. Nesses casos, é necessário técnicas gráficas mais complexas, como por exemplo o gráfico tipo matriz.

Quando a suposição de linearidade não for verdadeira, uma possibilidade é realizar transformações das variáveis explicativas de forma que a relação com as variáveis transformadas seja linear com a variável resposta. Essa operação chama-se *linearização*.

b - Suposições sobre os resíduos

Dos resíduos, Chatterjee e Hadi (CHATTERJEE; HADI, 2012) diz que a suposição é de que são *independentemente e identicamente distribuídos (i.i.d)*. Isto é, os resíduos devem ser compatíveis com uma distribuição normal com média zero e variância σ^2, constante (suposição de homoscedasticidade).

Quando essa suposição não for satisfeita, o modelo deve ser reespecificado, ou talvez pela mudança das variáveis resposta, ou até o próprio descarte do modelo de regressão linear.

c - Suposições sobre as variáveis explicativas

São feitas três suposições, de acordo com Chatterjee e Hadi (CHATTERJEE; HADI, 2012).

- As variáveis explicativas são não aleatórias, isto é, são definidas a priori antes da "geração" da resposta. Essa hipótese é satisfeita pelas fases de formulação do problema e coleta de dados. Se pela formulação no problema deseja-se, por exemplo, realizar uma análise com uma base de dados coletados automaticamente de forma não-experimental (que é caso desse trabalho), essa suposição é claramente invalidada. Nessa situação, o modelo gerado ainda se sustenta, mas sua interpretação muda, pois as conclusões geradas pelo modelo são condicionais, a depender das condições em que os dados foram coletados.

- As variáveis explicativas são medidas sem erro. Essa suposição depende da coleta de dados. Variáveis contínuas coletadas por instrumentos de medição terão o erro de medida intrínsico ao instrumento utilizado. Erros de medição afetam a variância dos resíduos e estimativas dos coeficientes de regressão.

- A variáveis explicativas são linearmente independentes umas das outras. Se isso não for satisfeito, a solução dos mínimos quadrados não será unica. Esse problema se chama de problema de *colinearidade* e deve ser resolvido na fase de especificação do modelo, pois não é possível realizar o ajuste do modelo aos dados nesse caso. Essa suposição pode ser verificada por meio de uma matriz de correlações que facilita a identificação de variáveis altamente correlacionadas.

A seguir falaremos sobre a estimação dos parâmetros do modelo (1).

2.3 Estimação dos parâmetros

Segundo Chatterjee e Hadi (CHATTERJEE; HADI, 2012) o vetor de parâmetros $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ é estimado através do *Método dos Mínimos Quadrados*. É um método que encontra um conjunto $\boldsymbol{\beta}$ tal que minimize a soma quadrática dos erros $S(\beta_0, \beta_1, \ldots, \beta_k)$ descrita em (2).

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=i}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i}, ..., \beta_k x_{kn})^2 \tag{2}$$

A estimativa dos parâmetros $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ que minimizam (2) podem ser obtidas através da solução de um sistema de equações chamadas *equações normais*. O processo para a solução desse sistema de equações pode ser realizado através do uso de pacotes computacionais.

Podemos escrever o modelo de regressão múltipla em notação matricial como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$$

com

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix}; \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{pmatrix} \text{ e } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{pmatrix}$$

Os estimadores dos parâmetros (a solução dos sistemas de equações) escritos em forma matricial:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

e

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X'X})^{-1}\sigma^2$$

Obtidas as estimativas dos coeficientes, é sempre relevante saber interpretá-los.

Interpretação dos coeficientes

Segundo Chatterjee e Hadi (CHATTERJEE; HADI, 2012), Os coeficientes β_i são chamados de *coeficientes de regressão parcial* e a interpretação mais simples é que o coeficiente representa o efeito variável X_i quando as demais são constantes. Ou seja, coeficiente β_i representa o efeito da variável X_i na variável resposta Y, fixados os valores das outras X_j, $j \neq i$.

Para ilustrar a relevância da interpretação dos coeficientes num modelo de regressão múltipla. Imagine por exemplo que estivéssemos criando um modelo de regressão linear simples para estimar o custo de frete de entrega (Y) de um produto usando como única variável explicativa seu peso (X_1) ou seu volume (X_2).

Pode-se imaginar que itens maiores são em geral mais pesados que itens menores. Isso é equivalente a dizer que as duas variáveis explicativas seriam correlacionadas entre si. Desse forma, em um modelo de regressão linear simples, tanto a variável de peso quanto de volume serviriam como sinalizadores do quão volumoso e pesado o item é.

Em outras palavras, pelo fato das duas variáveis explicativas serem correlacionadas, o efeito de uma variável estará "embutida" uma na outra. Se o peso for escolhido como variável explicativa, seu coeficiente representará não somente o efeito do item ser pesado sobre o frete, mas também o efeito do volume, considerando que um item pesado é provavelmente também volumoso.

Se, em vez de utilizarmos um modelo de regressão linear simples, utilizarmos um modelo de regressão linear múltipla que use as duas variáveis, cada um dos coeficientes trará o efeito "puro" de cada variável. O coeficiente do volume representará o efeito do volume sobre o frete dado que já sabíamos seu peso.

É por isso que, em um modelo de regressão linear múltipla, os coeficientes de regressão parcial de cada variável são diferentes de seus respectivos coeficientes de regressão caso cada uma dessas variáveis fosse ajustada a um modelo de regressão linear simples.

No exemplo dado, de estimar o custo de frete com base no volume e peso dos itens, como é razoável esperar que as duas variáveis sejam correlacionadas, podemos esperar também que no modelo de regressão linear múltipla os coeficientes compartilhassem o efeito causado pelo frete do item ser "grande e pesado". Dessa forma itens volumosos mas leves (pouco densos), teriam um custo de frete estimado menor que um item volumoso com peso proporcional ao volume.

O modelo de regressão linear simples que utiliza-se somente o volume estaria estimando o frete com base no volume e no peso esperado para aquele volume. Esse modelo, diferente do modelo de regressão linear múltipla, seria insensível para os casos de itens muito densos (pouco volume e muito peso) e pouco densos (muito volume e pouco peso).

Um caso ainda mais estranho na regressão linear múltipla é quando duas variáveis explicativas são positivamente correlacionadas entre si, mas tem efeitos opostos sobre a variável resposta. Nesse caso, é possível que nenhuma das variáveis sozinha tenha consiga explicar a variável resposta, mas as duas em conjunto conseguem.

O mesmo pode ocorrer quando duas variáveis explicativas são negativamente correlacionadas, mas tem efeitos similares sobre a variável resposta. Para esses casos, somente é possível explicar a variável reposta através do modelo de regressão linear múltipla.

A conclusão mais relevante sobre a interpretação dos coeficientes de regressão é que devem ser evitadas leituras isoladas dos coeficientes de regressão. Seu significado deve ser avaliado considerando as outras variáveis já introduzidas no modelo.

Outra consideração importante sobre os coeficientes de correlação, é avaliar se os coeficientes do modelo ajustado estão capturando efeitos reais que as variáveis explicativas tem sobre a variável resposta ou flutuações aleatórias da amostra de dados utilizada para o ajuste. Isso pode se realizado através de testes de hipótese descritos na próxima seção.

2.4 Testes de hipótese

Considere que um modelo de regressão múltipla expresso em (1) foi ajustado a um conjunto de dados. Geralmente há interesse em testar as hipóteses nula e alternativa expressas (3)

$$\begin{aligned} &H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \\ &H_1 : \exists \text{ pelo menos um } \beta_i \neq 0 \ (i = 1, ..., k) \end{aligned} \tag{3}$$

Para testar a teste de hipótese expressa em (3), emprega-se a estatística F dada por:

$$F = \frac{SSReg/k}{SSRes/(n-1-k)} \tag{4}$$

onde $SSReg = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$ e $SSRes = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ são respectivamente a soma de quadrados devido à regressão e a soma de quadrados dos resíduos do modelo expresso em (1), $\hat{y}_i$ o valor previsto da observação y_i segundo modelo (1) e $\overline{y}$, a média amostral dos valores da variável resposta. Rejeita a hipótese nula da equação (3) se F em (4) for maior que F_c da distribuição de Fisher com k graus de liberdade no numerador e $n-1-k$ graus de liberdade no denominador escolhido o erro do tipo I α. Além da realização destes testes de hipótese em modelo de regressão múltipla, há interesse em testar a equivalência de dois modelos de regressão múltipla construídos com as seguintes características principalmente quando a hipótese nula expressa em (4) não é verdadeira. Esta é uma estratégia que pode ser utilizada para selecionar vários bons modelos-candidato de regressão.

Sejam dois subconjuntos de variáveis explicativas A_j e A_w, tal que $A_j \subset A_w$, tendo j variáveis explicativas em A_j e $w = j + p$ variáveis em A_w. Sejam $\hat{y}_i^j$, $\hat{y}_i^w$, respectivamente, os valores previstos da observação y_i segundo os modelos construídos com os conjuntos A_j e A_w. As equações (5) e (6) denotam respectivamente as somas de quadrados dos resíduos dos modelos construídos empregando os conjuntos A_j e A_w .

$$SSRes(A_j) = \sum_{i=1}^{n}(y_i - \hat{y}_i^j)^2 \tag{5}$$

$$SSRes(A_w) = \sum_{i=1}^{n}(y_i - \hat{y}_i^w)^2 \tag{6}$$

De (KUTNER et al., 2005), para a realização do teste de hipótese (de equivalência de dois modelos de regressão construídos com dois subconjuntos aninhados), utilizamos a estatística F descrita em (7).

$$F = \frac{[SSRes(A^j) - SSres(A^w)]/p)}{SSres(A\hat{}w)/(n-w-1)} \tag{7}$$

Note que (7) sempre é positiva, pois o modelo reduzido (A_j) nunca terá uma soma quadrática de resíduos menor que o modelo menos reduzido (A_w). A noção intuitiva da estatística F é que, quanto melhor a melhoria de acurácia entre o modelo reduzido (A_j) e o modelo menos reduzido (A_w), maior será o valor da estatística F. Rejeitamos então a hipótese nula quando o valor dessa estatística ultrapassar um valor crítico, ou seja, quando a melhoria de acurácia entre o modelo reduzido ao modelo menos reduzido é grande o suficiente para ser estatísticamente significante. Especificamente, rejeita a hipótese de igualdade entre os dois modelos se F em

$(7) > F_c$ da distribuição de Fisher com p graus de liberdade no numerador e $n-w-1$ graus de liberdade no denominador escolhido o erro do tipo I igual a α.

Ainda segundo Draper e Smith (DRAPER; SMITH, 1998) pode-se demonstrar que $\hat{\beta}_i \sim N(\beta_i, D_{ii}\sigma^2) \rightarrow \dfrac{\hat{\beta}_i - \beta_i}{\sqrt{D_{ii}\sigma^2}} \sim N(0,1)$, $i = 0, \ldots, k$ com D_{ii} sendo o (i+1)-ésimo elemento da diagonal da matriz $\left(\mathbf{X'X}^{-1}\right)$. Desta forma, testes de hipótese dos coeficientes individualmente do tipo $H_0 : \beta_i = \beta_{i0}$ também podem ser feitos bem como a determinação de intervalos de confiança dos coeficientes individualmente.

Esse método poderia ser utilizado para testar muitas diferentes hipóteses, que normalmente dependerão da formulação do problema. Por exemplo, poderia ser testada a hipótese nula de que os coeficientes são iguais a valores pré-estabelecidos, ou se o efeito de uma variável é maior ou menor que o efeito de outra.

Esses testes de hipótese podem ser usados em processos de seleção de modelos, que são descritos da próxima Seção.

2.5 Seleção do modelo

A seleção do modelo consiste na escolha de um subconjunto (A_w) de variáveis explicativas para a determinação de um modelo de regressão linear múltipla.

Draper e Smith (DRAPER; SMITH, 1998) diz que, ao realizar a seleção de um modelo, por um lado, desejamos utilizar variáveis suficientes para que valores previstos confiáveis possam ser encontrados. Por outro lado, também queremos evitar colocar variáveis demais para manter a variância de $\hat{Y}$ baixa e também para redução de custos envolvidos com a geração de previsões a partir de modelos muito complexos.

Draper e Smith (DRAPER; SMITH, 1998) reconhece que não existe um único procedimento que deva ser utilizado, e faz a sugestão de quatro: *All Possible Regressions*, *Best Subset Regression*, *Stepwise Regression Backward Elimination*, além de algumas variações.

a - *All Possible Regressions* e *Best Subset Regression*

O procedimento de *All Possible Regressions* consiste na comparação de todos os modelos que podem ser construídos com as variáveis disponíveis. A quantidade de modelos a serem definidos para comparação será igual a 2^k, onde k é o número de variáveis disponíveis.

Os modelos definidos dessa forma devem ser agrupados de acordo com o número de variáveis utilizados no modelo. Em seguida, os modelos são avaliados a partir de um critério que pode ser R^2, ou s^2 ou a estatística p.

É observado, para cada grupo, o melhor desempenho possível (de acordo com o critério utilizado) para cada grupo. Deve ser selecionado o número de variáveis a partir do qual não os aumentos de performance com o aumento do número de variáveis não seja significativa.

Será escolhido um dos modelos com melhor performance dentro grupo com aquele número de variáveis.

A **Best Subset Regression** é uma variação desse método para reduzir o número de modelos a serem testados. Nesse procedimento, deve ser usado um programa que cria uma lista com as melhores K equações para cada grupo de modelos. Ou seja, os melhores modelos com 1 variável, duas variáveis, 3 variáveis, etc.

b - ***Stepwise Regression***

O procedimento de *Stepwise Regression*, inicia com um modelo composto pela variável X_i que seja melhor correlacionada com a variável resposta.

Em cada interação desse procedimento, as variáveis que não estão no modelo são ordenadas pelos seus $F - valores$ parciais, a variável com o maior valor sera utilizada para a definição de um novo modelo. É então conferido se a a melhoria da estatística F do novo modelo é significante, e é registrado também a melhoria no R^2. Se a melhoria não for significante, o novo modelo é rejeitado e o modelo antecessor é adotado como modelo final.

Para cada estágio desse procedimento, depois de ser adicionada ao modelo uma variável nova, é conferido quais variáveis já presentes no modelo podem ser excluídas. Essa passo é realizado porque variáveis que podem ser boas preditoras no começo podem se tornar surpéfolas depois de adicionadas outras variáveis.

c - **Backwards Elimination**

Esse procedimento inicia com um modelo de regressão que utiliza todas as variáveis explicativas. Para cada variável é calculado o teste F parcial considerando como se aquela variável tivesse sido a última a entrar no modelo.

O menor valor do teste F parcial é comparado com o nível de significância pré estabelecido. Se esse valor F for menor que o valor pré-estabelecido, a variável em questão é removida do modelo, os testes F parciais são recalculado e o procedimento se repete.

Variáveis irão sendo eliminadas do modelo enquanto o teste o menor teste F parcial for menor que o valor crítico. Quando isso não for satisfeito, nenhuma outra variável é eliminada do modelo, que é selecionada como modelo final.

Esses quatro procedimentos descritos nessa seção são algumas alternativas que podem ser utilizadas para a seleção de modelos. Quando um modelo final é selecionado, é necessário realizar um diagnóstico com métodos gráficos e estatísticas descritivas, que são descritos na próxima Seção.

2.6 Diagnóstico de Regressão

Esse diagnóstico pode ser realizado apenas após a especificação do modelo e estimação dos parâmetros. Essa suposição é, inclusive, necessária para validar as estatísticas t dos coeficientes de regressão, pois o cálculo dessa estatística depende dessa suposição.

Novamente, utiliza-se de análise gráfica para confirmação dessa suposição. Recomenda-se traçar o histograma dos resíduos e o gráfico *quantil-quantil* dos resíduos *versus* a variável resposta.

O diagnóstico de regressão consiste em alguns passos que devem ser realizados, antes e depois do ajuste do modelo para avaliar se o modelo está bem ajustado ou não. Queremos avaliar se as suposições do modelo de regressão estão sendo satisfeitas e testes de hipóteses para aceitar ou rejeitar o modelo com base em um nível significância estabelecido.

É reforçado por Chatterjee e Hadi (2012) (CHATTERJEE; HADI, 2012) que nenhum dos métodos por si só é suficiente para gerar um diagnóstico adequado. Devem ser usados conjuntamente métodos gráficos e estatíticas descritivas para avaliar o modelo.

Os métodos gráficos incluem visualizações que podem ser criadas antes e depois do ajuste do modelo, para confirmar as hipóteses de linearidade, independência das variáveis explicativas e as diversas suposições a respeito dos resíduos.

Métodos estatísticos incluem estatísticas descritivas e testes de hipótese para confirmar que as variáveis utilizadas de fato são informativas quanto à variável resposta e o quão bem o modelo é capaz de prever a variável resposta.

Métodos gráficos

Nesta seção, serão explorados em mais detalhes métodos gráficos que podem ser utilizados para a confirmação de hipóteses do modelo de regressão. Como já mencionado anteriormente, estatísticas descritivas não são suficientes para realizar o diagnóstico adequado de um modelo de regressão. De fato, algumas das estatísticas descritivas só tem relevância quando algumas suposições forem verdadeiras.

Para ilustrar a relevância de fazer gráficos de dispersão antes de ajustar um modelo de regressão simples, na Figura 3, temos gráficos de dispersão de um famoso conjunto de dados didáticos chamado de Quarteto de Anscombe. Esse é um conjunto de dados elaborado por Anscombe (ANSCOMBE, 1973) para demonstrar a importância dos métodos gráficos.

As quatro variáveis explicativas, quando ajustadas em um modelo de regressão linear simples à variável resposta, tem estatísticas descritivas similares: mesma correlação com a variável resposta, mesmo coeficiente de regressão linear e mesmo coeficiente de determinação, mesmos $\overline{x}$ e $\overline{y}$ e etc. Contudo, os gráficos de dispersão dessas variáveis em relação à variável resposta revelam relações diferentes.

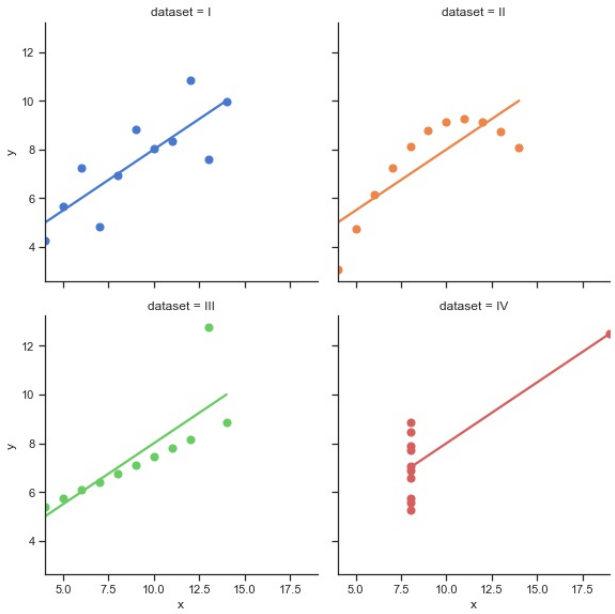

Figura 3 – Quarteto de Anscombe

O segundo conjunto, por exemplo, viola a suposição de linearidade, pois a relação com a variável resposta é claramente não linear, que deveria ser linearizada (ou utilizar um modelo não linear). O terceiro conjunto tem relação linear, mas é influenciada por um *outlier*.

Métodos gráficos devem ser utilizados para confirmar hipóteses antes e depois de estimar os coeficientes de regressão (ajuste do modelo). Serão apresentados alguns métodos gráficos importantes, começando pelos métodos antes do ajuste do modelo.

a - Antes do ajuste do modelo

Antes do ajuste do modelo, é relevante verificar o comportamento das variáveis explicativas. Queremos observar se há presença de outliers na variável auxiliar ou resposta, se algumas variáveis auxiliares são altamente correlacionadas entre si e se algumas delas tem uma relação de linearidade forte com a variável resposta.

Segundo Chatterjee e Hadi (CHATTERJEE; HADI, 2012), gráficos que envolvem uma variável só, que podem ser chamados de *univariados* ou *unidimensionais* e são úteis para verificar a distribuição de uma única variável e identificação de outliers. Os dois exemplos mais populares de gráficos univariados são o *boxplot* e o histograma.

Já para verificar a relação entre as variáveis, pode-se traçar, dois-a-dois, os gráficos de dispersão das variáveis. Para poucas variáveis, é conveniente traçar um gráfico tipo matriz, como exemplificado na Figura 4 que possibilita a melhor visualização, numa única figura, das distribuições das variáveis nas diagonais e os gráficos de dispersão duas a duas no resto fora da diagonal do gráfico.

O gráfico tipo matriz não é muito adequado para muitas variáveis. Como o tamanho total do gráfico é proporcional ao quadrado de variáveis incluídas, o gráfico de matriz tende a ficar

muito grande e de difícil leitura para um número grande de variáveis.

Para analisar um conjunto de dados com muitas variáveis, sugere-se traçar o gráfico de matriz para subconjuntos menores das variáveis.

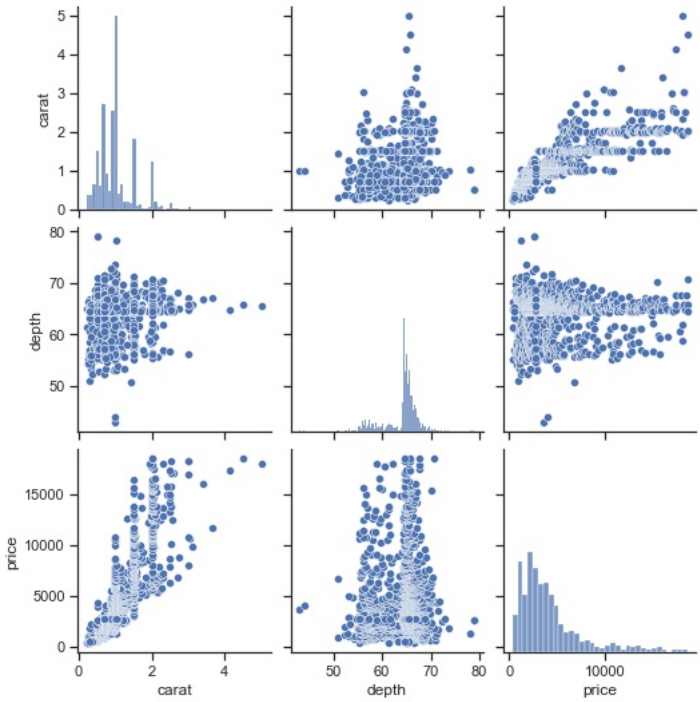

Figura 4 – Exemplo de gráfico de matriz

Quando se trata de identificar relações entre a as variáveis explicativas e a variável resposta, gráficos de dispersão ainda são limitados, pois não conseguem capturar relacionamentos multivariados. Isto é, quando uma variável só apresenta um relacionamento significativo quando é levada em consideração junto com outras.

Após o ajuste do modelo

Gráficos gerados após do modelo podem ser utilizados para confirmar as hipóteses de linearidade e normalidade, para identificação de outliers e observações muito influentes e para realizar o diagnóstico dos efeitos das variáveis explicativas (**??**).

Para conferir as hipóteses de linearidade e normalidade, podem ser traçado o histograma dos resíduos ε_i. Essa visualização irá revelar se o os resíduos obedecem uma distribuição normal. Outra alternativa é o gráfico de probabilidade normal, também chamado de gráfico quantil-quantil.

O gráfico quanti-quantil traça os pontos ordernados de um vetor, no caso, dos resíduos, *versus* um vetor de tamanho n extraídos de uma distribuição normal. Se os resíduos obedecerem uma distribuição normal, os pontos em um gráfico quantil-quantil cairão ao longo de uma linha de quarenta e cinco graus.

Um exemplo desse tipo de visualização está apresentado na figura 5. Nessa figura foi

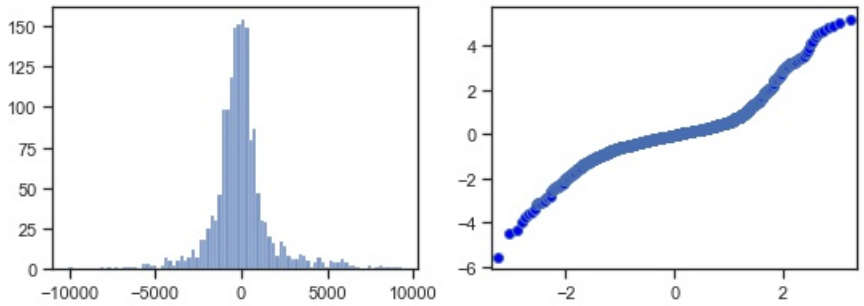

Figura 5 – Exemplo de gráfico de quantil-quantil (à direita)

traçada o gráfico quantil quantil do mesmo conjunto de valores do histograma à esquerda. Embora o histograma aparente ser uma distribuição normal, o gráfico quantil-quantil revela claremente que esse conjunto de valores não obedece uma distribuição normal, já que os pontos nesse gráfico não estão ocupando a linha de 45º.

Além disso, também é desejável que os resíduos não tenham relação com as variáveis explicativas nem com as variáveis resposta. Isso pode ser verificado através de gráficos de dispersão entre os resíduos ε_i e as variáveis explicativas x_i, ou os valores previstos $\hat{y}_i$.

Foi utilizado um *dataset* didático para determinar um modelo de regressão múltipla de uma variável resposta com três variáveis explicativas. Os gráficos de dispersão entre os resíduos do modelo e as variáveis utilizadas estão apresentadas na figura 6. Essa figura também mostra o exemplo de um modelo mal ajustado, porque claramente há relação entre os resíduos as as variáveis X_1 e Y.

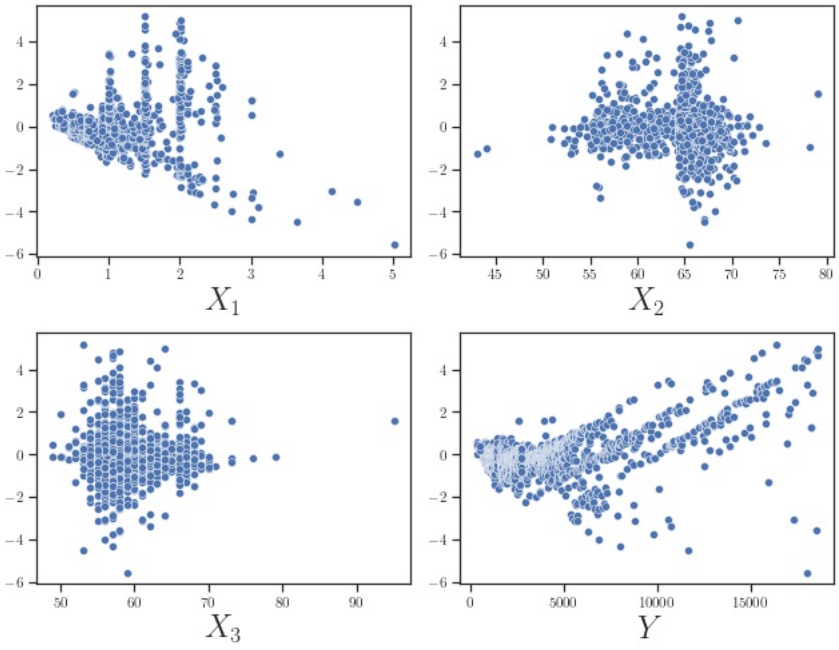

Figura 6 – Exemplo de gráfico de dispersão entre resíduos e demais variáveis

Isso conclui a breve revisão bibliográfica. No próximo capítulo tratará de como a teoria

levanta até aqui será aplicada para o trabalho em questão, e qual ferramental será utilizado.

3 METODOLOGIA

Depois da breve revisão de literatura sobre analise de regressão, é descrito nesse Capítulo como foram implementadas as técnicas presentes na literatura para o desenvolvimento do trabalho. É descrito, primeiramente, quais são as fases que devem ser executadas em um projeto de desenvolvimento de modelos, desde a formulação do problema até a entrega do modelo final. Também é apresentada a estratégia utilizada para navegar entre as fases.

É detalhado para cada fase quais os passos devem ser realizados, quais técnicas presentes na teoria estão incorporados e qual ferramental utilizado para a aplicação prática dessas técnicas.

O Capítulo inicia com a descrição do processo de desenvolvimento de modelos estatísticos.

3.1 Visão geral e natureza iterativa do processo de desenvolvimento de modelos

Diversas fontes de literatura sugerem estratégias de desenvolvimento de modelos preditivos. Essas fontes costumam sugerir fases de desenvolvimento muito similares ou análogas, e todas as fontes costumam sugerir que seja introduzido algum elemento iterativo ao processo.

Para o contexto mais amplo de ciência de dados, é muito popular a aplicação da metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*) (SCHRÖER; KRUSE; GÓMEZ, 2021). Essa metodologia sugere que o processo de desenvolvimento de modelos preditivos (notar que essa metodologia não se limita à analise de regressão) seja dividida em seis fases: entendimento do problema, entendimento dos dados, preparação dos dados, criação de modelos, avaliação de modelos e implantação em produção dos modelos.

Também há estratégias sugeridas para o contexto específico de análise de regressão. Para Kutner et al. (KUTNER et al., 2005) a estratégia inicia com uma fase de análise exploratória de dados, seguido de ciclo entre desenvolvimento, validação e revisão de modelos até que um ou mais modelos satisfatórios sejam encontrados. E Chatterjee e Hadi (CHATTERJEE; HADI, 2012) sugere uma estratégia bem parecida, que omite a fase análise de exploratória de dados, mas entra em mais detalhes das fases de desenvolvimento e validação dos modelos, separando-as em subpassos.

Todos esses métodos também apresentam alguma característica iterativa. A metodologia CRISP-DM é uma metodologia cíclica, em que os modelos desenvolvidos possibilitam melhor entendimento do problema, o que por sua vez possibilita o desenvolvimento de melhores modelos no futuro. CRISP-DM também é bem flexível com iterações entre as fases de entendimento do negócio, entendimento dos dados, preparação dos dados e criação de modelos.

Pode-se dizer que CRISP-DM é uma metodologia de alto nível que faz o delineamento

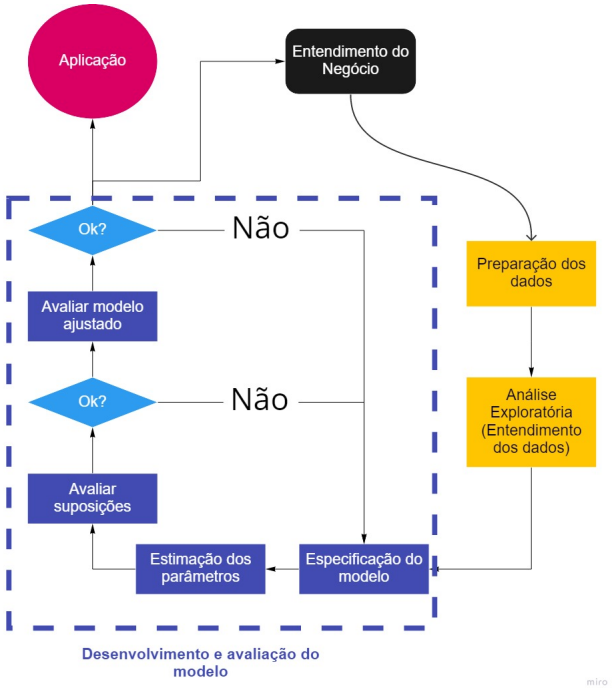

Figura 7 – Estratégia de desenvolvimento do modelo

de todas as fases do desenvolvimento do modelo desde a obtenção dos dados até a implantação de modelos em produção, sem entrar em detalhes sobre cada fase. As estratégias específicas de análise de regressão assumem que os dados já estão preparados, e entram em detalhes sobre a especificação e avaliação dos modelos, sem também se preocupar como a aplicação do modelo para o propósito planejado.

Notar ainda que a fase de análise exploratória de dados sugerida por Kutner et al. (KUTNER et al., 2005) é análoga a fase de entendimento dos dados da CRISP-DM. Uma possível interpretação disso é que as estratégias de análise de regressão podem ser "encaixadas"dentro da metodologia CRISP-DM no lugar das fases de modelagem e avaliação de modelos.

E é baseada nessa interpretação que foi definida a estrategia de desenvolvimento desse trabalho, que é ilustrada na Figura 7. A estratégia pode ser descrita como um ciclo CRISP-DM em que as as estratégias de análise de regressão inseridas no lugar das fases de desenvolvimento e avaliação de modelos. Pode-se dizer também que essa estratégia é uma versão mais detalhada da CRISP-DM, específica para análise de regressão.

3.2 Preparação dos dados

A fase de preparação de dados inclui as atividades de realizar a obtenção (no caso, extração de dados), manipulações e pré-processamento dos dados.

Como os dados utilizados para esse trabalho são dados internos da empresa disponibilizados em *Datalake*, a extração de dados significa a elaboração e execução de consultas

(*queries*) SQL (*Structured Query Language*) nesse *Datalake*. Atráves de SQL é possível extrair um conjunto de dados com a granularidade, escopo e colunas desejadas.

Depois de extraído o conjunto de dados, algumas transformações precisam ser realizadas para fim de análise.

A primeira alteração realizada foi o ajuste de todos os valores monetários pela inflação. Os valores de inflação utilizados foram o Índice de Preços ao Consumidor Amplo, extraído via a API (*Application Programming Interface*) do Sistema Gerenciador de Séries temporais do Banco Central. Os dados extraídos dessa forma são a variação mensal do indicador, que podem integrados para calcular um fator multiplicador para ajustar valores monetários para o valor de um mês base. O mês base utilizado foi Março de 2020, último mês do conjunto de dados.

Outra manipulação necessária foi a criação de variáveis auxiliares. Todas variáveis calculadas eram obtidas a partir do valor de data. Essas variáveis incluem o mês, o dia da semana, proximidade da *Black Friday*, etc. Essas manipulações foram feitas através dos pacotes *panda* e *numpy* de *python*. A seguir uma breve descrição destes pacotes.

pandas é uma ferramenta de análise de dados de código aberto que oferece APIs para representar conjunto de dados atráves dos objetos *pandas.Series* e *pandas.DataFrames*. Esses objetos possuem métodos que abstraem operações comumente realizadas sobre dados. Essas operações incluem seleções, agrupamentos, reordenação, indexação, operação de vetores, etc. *pandas* é construído com base em *numpy*, um pacote de computação científica que também oferece objetos para representar *arrays* multidimensionais e diversas rotinas de operações sobre esses *arrays* (THE PANDAS DEVELOPMENT TEAM, 2020) (HARRIS et al., 2020).

Listing 3.1 – Exemplo de utilização de pandas e numpy

```
# df e um objeto pandas.DataFrame, que representa uma "tabela"
# Neste exemplo, estamos usando os metodos numpy.sin e numpy.cos
# para criar colunas adicionais no DataFrame, que sao senoides
# calculadas com base na coluna 't' do mesmo DataFrame

df['cos30'] = numpy.cos(2*math.pi*df['t']/30)
df['sin30'] = numpy.sin(2*math.pi*df['t']/30)
df['cos60'] = numpy.cos(2*math.pi*df['t']/60)
df['sin60'] = numpy.sin(2*math.pi*df['t']/60)
```

Além dessas manipulações, também costuma ser necessário realizar um passo de pré-processamento, que consiste em outras alterações adicionais sobre *Array-likes* (*DataFrames* e *Arrays*) para torná-los compatíveis com pacotes de modelagem estatísticas (a serem discutidos nas próximas seções).

Normalmente pacotes de modelagem estatística em python não costumam suportar

cor	X_1	X_2
Vermelho	1	0
Azul	0	1
Verde	0	0

Tabela 1 – Exemplo de codifição em variáveis *Dummy*

valores nulos e variáveis categóricas. O *dataset* já foi extraído sem valores em branco, então foi necessário somente transformar as variáveis categóricas.

As variáveis categóricas precisam passar um passo de codificação (*encoding*). Esse passo é a tradução dessas variáveis para variáveis quantitativas. Existem dois principais tipos de codificação. Variáveis categóricas ordinais (ex: "pequeno", "grande"e "médio") podem ser codificadas como uma única variável quantitativa discreta, em que o menor valor equivale a zero, o segundo menor a 1 e assim por diante. No pacote *scikit-learn*, essa técnica se chama *Label Encoding* (PEDREGOSA et al., 2011). No entanto, em caso de usar variáveis ordinais em modelos de regressão recomenda-se fortemente que ela seja codificada como variável nominal, vide parágrafo a seguir.

Variáveis categóricas nominais (ex: "vermelho", "verde"ou "azul") podem ser codificadas em $k-1$ colunas binárias onde k correspondem ao número de níveis da variável categórica, cada uma correspondente a um nível. Um exemplo desse tipo de codificação é apresentado na Tabela 1. Esse tipo de codifição pode ser chamado de *One-Hot Encoding* ou codificação em variáveis *dummy*. Note que não existe uma única codificação e no exemplo a cor verde foi escolhida como a categoria de referência.

Para o caso desse trabalho, toda codificação de variáveis categóricas necessárias pode ser feita através de métodos do pacote *pandas*. Para casos mais complexos, seria necessário utilizar pacotes específicos de *Machine Learning*, que contém métodos mais robustos, como *scikit-learn* (PEDREGOSA et al., 2011).

Um passo opcional do pré-processamento dos dados é o ajuste de escala (*scaling*). Esse passo consiste em trazer todas as variáveis para a mesma escala. Uma forma popular de fazer isso é a padronização das variáveis de forma que tenham média igual a zero e desvio padrão igual a um. Esse passo normalmente deve ser feito caso estejam sendo utilizados modelos sensíveis a escala.

Segundo Chatterjee e Hadi (CHATTERJEE; HADI, 2012), modelos de regressão linear não são sensíveis a escala, mas um método de ajuste de escala, como a padronização de variáveis ainda pode ser feita. O maior impacto dessa decisão é na forma como devem ser interpretados os coeficientes de regressão do modelo depois do ajuste.

Nesse trabalho, optou-se por não realizar ajuste de escala.

3.3 Análise exploratória

Na analise preliminar, ou análise exploratória de dados, utilizam-se métodos gráficos para explorar a distribuição das variáveis, o relacionamento entre elas e validar algumas suposições do modelo de regressão linear.

O objetivo desse passo é escolher quais variáveis entram no modelo final. É desejável que as variáveis apresentem relação com a variável resposta. Contudo, há casos, como discutido no Capítulo 2, em que uma variável sozinha não apresenta correlação com a variável reposta, mas um conjunto duas ou mais variáveis apresenta correlação. Portanto, pode ser difícil de reconhecer o poder preditivo de uma variável resposta na fase de analise exploratória.

Para a visualização da distribuição das variáveis, pode-se utilizar-se gráficos de das variáveis em relação ao tempo, pois essa visualização mostra simultaneamente a variabilidade da variável e seu relacionamento com o tempo. Já para identificar o relacionamento das variáveis entre si e com a variável resposta, são utilizados gráficos tipo matriz e matrizes de correlação. Para criação os gráficos deste capítulo, são utilizados os pacotes de python *matplotlib* e *seaborn* (HUNTER, 2007) (WASKOM, 2021).

3.4 Ciclo de desenvolvimento, avaliação e revisão dos modelos

Esse ciclo começa com a especificação de um modelo a partir da seleção de um subconjunto de variáveis disponíveis no conjunto de dados. Os parâmetros desse modelo são estimados e os resultados do modelo são avaliados primeiramente através da análise dos resíduos e, se confirmadas as suposições sobre resíduos do modelo de regressão linear, então o modelo é avaliado pelas estatísticas t, F e seu coeficiente de determinaçào R^2.

Esse ciclo se repete até que obtenha-se um modelo que satisfaça as suposições do modelo de regressão e tenha uma acurácia satisfatória.

Na análise de regressão, é comum inclusive processos, muitas vezes automatizados, com criterios estabelecidos para a seleção das variáveis que, a cada interação, entram ou saem da especificação do modelo. Draper e Smith (DRAPER; SMITH, 1998), por exemplo, descreve três desses métodos: *Best subset*, *Stepwise Regression* e *Backwards Elimination*. No desenvolvimento desse trabalho, contudo, foi usado um procedimento manual, determinado principalmente por observações da análise exploratória.

O modelo da primeira iteração usava, como variáveis, a quantidade de dias desde o ínicio do período englobado pelo dataset e eventos de calendário. A seleção desse subconjunto de variáveis para o modelo inicial se deu por duas razões.

A primeira delas é que o gráfico da variável resposta em relação ao tempo revela que, no período análisado, a variável resposta seguiu uma aparente tendência linear, com picos na *Black Friday*, o que sugere que essas duas variáveis poderiam resultar em uma boa previsão inicial.

O segundo motivo é que ambas as variáveis são calculadas a partir da data, o que significa que esse modelo inicial ilustraria o que seria capaz de ser previsto somente com nenhuma informação além da data.

Essas observações são retomadas em mais detalhes no Capítulo 4.

A seleção de variáveis nas iterações seguintes foram realizadas de forma manual, sem critério pré-estabelecido, com base principalmente na análise residual do modelo e estatísticas de teste. A decisão sobre adicionar informação sobre dias da semana, por exemplo, aconteceu por ter-se percebido uma autocorrelação dos resíduos em um período de sete dias, que foi removida pela adição da informação sobre dia da semana.

A estimação dos parâmetros do modelo foi feita através da implementação do Método dos Mínimos quadrados de pacotes estatísticos de *python*.

No momento em que este trabalho está sendo desenvolvido (2021) as duas principais opções de pacotes de modelagem estatística em python são *scikit-learn* (PEDREGOSA et al., 2011) e *statsmodels* (SEABOLD; PERKTOLD, 2010).

Scikit-learn é um pacote de *Machine Learning* que oferece a implementação de diversos modelos de classificação, regressão e clusterização de dados; *ensemble* de modelos; seleção de variáveis (redução de dimensionalidade) e seleção de modelos. É um pacote com muitos recursos focada para aplicações de *Machine-Learning* e Ciência de Dados em geral. Em comparação, statsmodels oferece uma gama menor de modelos, focada em modelos lineares generalizados, análise de séries temporais e econometria.Em contra partida, oferece mais opções de estatísticas descritivas, testes de hipótese e análise dos resíduos. Os dois pacotes também trabalham muito bem com o pacote *pandas*, de forma que objetos *pandas.DataFrame* e *pandas.Series* podem ser passados como argumentos nas chamadas de definição dos modelos e ajuste dos parâmetros tanto em um pacote como outro.

Ambos pacotes foram utilizados em diferentes momentos de desenvolvimento do trabalho. Deu-se preferência, contudo, para o pacote *statsmodels*, por oferecer ferramentas que habilitaram melhor análise residual.

Listing 3.2 – Demonstração statsmodels

```
1
2 # A variavel X e um DataFrame pandas com as variaveis explicativas
3 # Esse exemplo demonstra como a realizaodo a definicao e ajuste de um modelo
4 # de regressao linear com o pacote statsmodels
5
6 import statsmodels.api as sm
7
8 X = sm.add_constant(df[lin_reg_features])
9
```

```
10 mod3 = sm.OLS(df['y']/1000, X.astype(float)).fit()
11
12 df['y_pred'] = mod3.predict(X.astype(float))*1000
```

A validação do modelo é realizada pela análise residual através de métodos gráficos e avaliação de estatísticas descritivas.

A análise residual consiste na visualização do histograma dos resíduos do modelo para verificar que obedecem a uma distribuição normal com média zero; na visualização dos resíduos ao longo do tempo para verificar sua estacionaridade; e no gráfico quantil-quantil para verificar se não há anormalidade para valores extremos.

De estatísticas descritivas, as principais utilizadas são os p-valores das estatísticas t dos coeficientes que devem todas respeitar o valor crítico de 5% e o coeficiente de determinação R^2.

4 ANÁLISE EXPLORATÓRIA DE DADOS

Para a criação de bons modelos de regressão é necessário entender os dados em mãos. Este Capítulo é dedicado a essa investigação, utilizanco métodos gráficos estabelecidos no Capítulo 2.

Queremos entender o formato dos dados. Quantos registros (linhas) nosso conjunto de dados tem, o que cada registro representa e quais variáveis (colunas) temos para cada registro.

Uma boa análise exploratória de dados também inclui a verificação da distribuição de cada variável e as relações que existem entre elas.

Esse capítulo será organizado em seções, a primeira delas tratando da avaliação do formato dos dados, seguida de uma seção para análise da variável resposta. As variáveis explicativas foram separadas seis subconjuntos, que são avaliadas cada uma em uma seção.

Em cada uma das seções sobre os subconjuntos de variáveis explicativas é feita a verificação da distribuição das variáveis contidas no subconjunto e das relações que existem entre elas.

A noção do que está presente no conjunto de dados dá o suporte necessário para entender de quais variáveis explicativas podemos partir para iniciar os ciclo de especificação, avaliação e revisão dos modelos.

4.1 Formato dos dados

O *dataset* utilizado descreve medidas agrupadas da categoria de Guarda-Roupas, dia-a-dia, entre 01/01/2018 até 31/03/2021. O total de registros é 821, correspondendo aos 821 dias de operação do período analisado.

Além do filtro de categoria e período de análise, os dados incluídos no dataset limitam-se as vendas pelo canal *webshop*, isto é, vendas pelo site da Mobly. Portanto, o *dataset* não inclui vendas de lojas físicas, *outlets* nem de outros *marketplaces* digitais fora da marca Mobly.

Então, como visto, a granularidade do *dataset* é diário e, para cada dia descrito, há 25 medidas, que são ou medidas agrupadas de características dos SKUs da categoria de Guarda-Roupas, ou variáveis de eventos de calendário.

A variável resposta é o volume de vendas, dia-a-dia, medido em unidade monetária e corrigido pelo IPCA amplo, da categoria de Guarda-Roupas no período descrito pelo *dataset*. Assim, essa variável resposta também tem 821 registros, correspondentes aos dias abordados pelo dataset.

Das 25 variáveis, 22 delas são quantitativas (números inteiros ou reais). As últimas 3

variáveis são categóricas e codificadas como 33 variáveis *dummy* . A tabela com as variáveis e seus tipos de dados é apresentada no Apêndice.

Em resumo, as 26 variáveis presentes no *dataset* podem ser agrupadas nos seguintes subconjuntos:

- Variável resposta
- Variáveis calculadas a partir da data - 5 variáveis quantitativas e 3 categóricas
- Variáveis relativas a precificação de itens - 3 variáveis quantitativas
- Variáveis relativas a condições de frete - 2 variáveis quantitativas
- Variáveis relativas a marketing - 3 variáveis quantitativas
- Variáveis relativas a estoque - 2 variáveis quantitativas
- Variáveis relativas a características de SKUs - 6 variáveis quantitativas

A analise detalhada de cada variável é inciada com a variável resposta.

4.2 Variável resposta

A variável resposta é a venda diária da categoria de Guarda-roupas, medido em unidade monetária. A evolução ao longo do tempo dessa variável é apresentada na Figura 8. Nessa figura, nota-se que a venda diária parece seguir uma tendência de crescimento linear ao longo do período analisado. Destaca-se também os picos próximos do período de *Black Friday*.

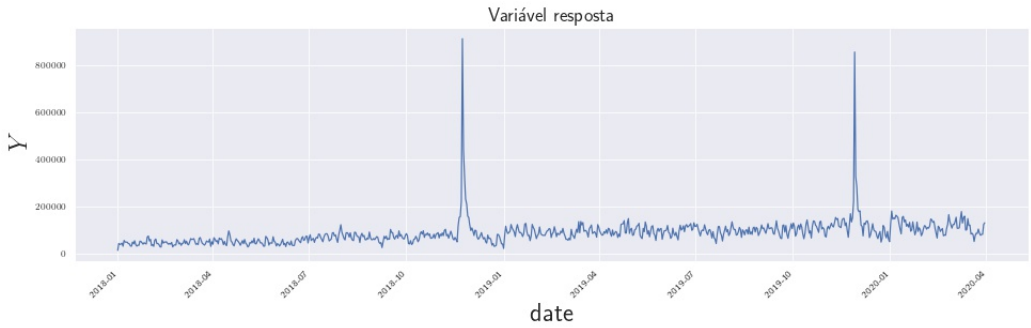

Figura 8 – Evolução temporal da variável resposta

Em seguida, vamos analisar as variáveis calculadas a partir da data.

4.3 Variáveis criadas a partir da data

Essas variáveis são aquelas que são calculadas a partir da informação de data. A data em si é utilizado como indexador do conjunto de dados e não é em nenhum momento utilizada como variável explicativa. Em vez disso, utilizamos a variável T, que mede a quantidade de dias desde o início do dataset. Tem valor zero para o dia 01/01/2018 e valor de 820 para o dia 31/03/2020, que é o último dia englobado pelo *dataset*.

A partir do valor T são calculadas outras quatro variáveis quantitativas: $S_1 = \sin(30T)$, $C_1 = \cos(30T)$, $S_2 = \sin(60T)$ e $C_2 = \cos(60T)$. São o seno e cosseno ajustados para ter frequência de 30 ou 60 dias. A intenção da utilização dessas variáveis como variáveis explicativas é capturar possíveis efeitos cíclicos mensais ou bimestrais.

As demais outras variáveis calculadas a partir da data são eventos de calendário. Todas, variáveis *dummy*.

As variáveis $W_i, i \in (2,...,7)$ dizem respeito aos dias da semana, sendo W_2 se o dia da semana corresponde a terça-feira e ,... e W_7 correspondente ao domingo. Os meses são representados pelas variáveis $M_i, i \in (2,...12)$, sendo M_2 correspondente ao mês de Fevereiro, ... e M_{12} correspondente ao mês de Dezembro. Notem que o dia semana segunda-feira e o mês de Janeiro foram escolhidos como categorias de referência.

Além disso temos as variáveis relativas à *Black Friday* que sinalizam o próprio dia da *Black Friday* quanto os dias imediatamente antes ou depois. Essas variáveis são $B_i, i \in (-7,...6)$, em que B_{-7} denota o dia sete dias antes da *Black Friday*, B_{-1} denota o dia imediatametne anterior, B_0 é o próprio dia da *Black Friday*, B_1 é o dia imediatamente depois e B_6 é o dia seis dias depois.

A próximo subconjunto de variáveis a serem analisadas são as variáveis relativas à precificação de produtos.

4.4 Variáveis de preço

O conjunto de dados contém três variáveis descritivas de preço: P_1, P_2 e P_3.

Na base de dados os SKUs são cadastrados com dois preços, o preço comum e o preço especial. Enquanto o SKU estiver em período promocional, será visível no site o preço especial e, caso contrário, será visível apenas o preço ordinário. Notar que, mesmo quando um produto está em promoção, ele ainda tem um preço ordinário cadastrado na base. Outra forma de compreender esses dados é: o preço comum é o preço "de", e promocional é o o preço "por"de um produto.

P_1 é a média dos preços comuns de todos os SKUs, P_2 é a média dos preços especiais considerando somente SKUs em promoção. P_3 é a média do preço de venda no site de todos SKUs, isto é, para SKUs em promoção considera-se o preço promocional, e para os demais

SKUs considera-se o preço comum.

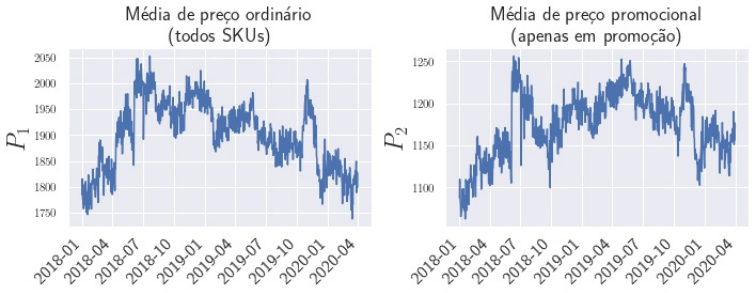

Figura 9 – Evolução temporal do preço ordinário e preço promocional

A evolução temporal das variáveis relativas ao preço são apresentadas na Figura 9. A média de preço ordinários aumenta até o meio de 2018, e segue em queda desde então. O preço promocional tem uma dinâmica similar, com pico aproximadamente em Julho de 2019.

Figura 10 – Evolução temporal do preço do site

O preço do site, apresentado na Figura 10 também apresenta variações, mas é mais estável. Tem um leve aumento até abril de 2019, e depois se mantém em estabilidade.

O gráfico tipo matriz entre essas três variáveis, apresentado na Figura 11 mostra que as três variáveis são correlacionadas entre si. A variável P_3 apresenta uma correlação muito forte com tanto com P_1 quanto com P_2, sendo a correlação entre P_3 e P_2 visivelmente forte.

4.5 Variáveis relativas ao frete

São incluídas duas medidas descritivas do frete: F_1 e F_2 que são, respectivamente, a média do tempo de entrega e média do valor de frete para São Paulo. A evolução temporal das duas variáveis está apresentada na Figura 12.

O valor médio de frete mantém a mesma escala de variância ao longo de todo período. A escala de variância do tempo médio de frete muda drasticamente a partir do começo de 2020.

O gráfico tipo matriz apresentado na Figura 13 mostra que as duas variáveis não estão correlacionadas. Os histogramas indicam que a distribuição do tempo médio do frete, com

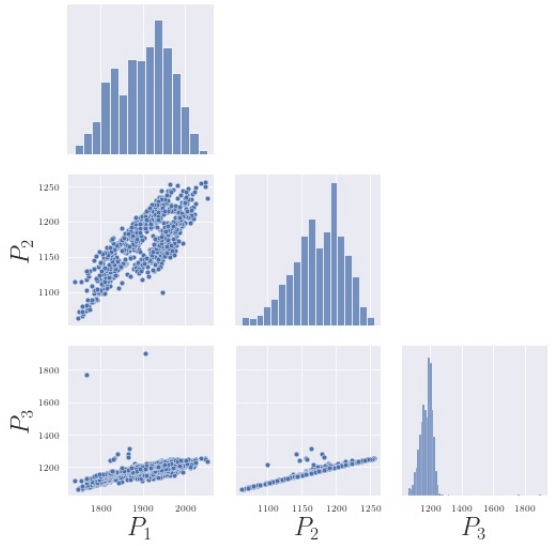

Figura 11 – Gráfico tipo matriz das variáveis de preço

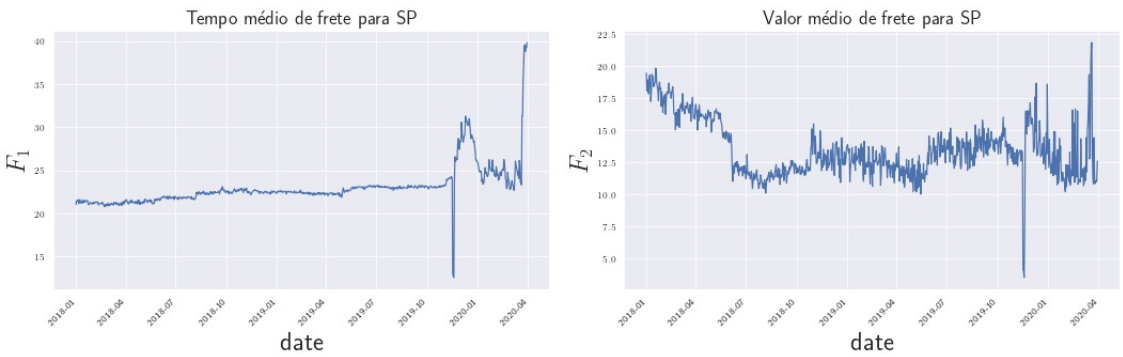

Figura 12 – Evolução temporal das variáveis relativas ao frete

muitos valores concentrados em uma pequena faixa, mas com uma cauda muito grande de valores extremos, provavelmente correspondentes ao periodo a partir de 2020.

4.6 Variáveis relativas à marketing digital

Há três variáveis com essa temática: M_1, M_2 e M_3.

M_1 é a soma total de impressões de catálogo de todos produtos da categoria, e M_2 é a soma total de todas as visitas à páginas de produto da categoria. M_3 é o total em reais dos custos de marketing com a categoria no dia.

Vemos na Figura 14 a evolução das impressões de catálogo e custo de marketing dos produtos. As impressões de catálogo tem uma tendência de crescimento, com variabilidade relativamente constante. Contudo essa variável tem muitos registros de impressão zero, incluindo um grande intervalo sem registro de impressões de catálogo no final de 2019. O uso dessa variável como variável explicativa exigiria, possivelmente, a remoção dos registros em que as

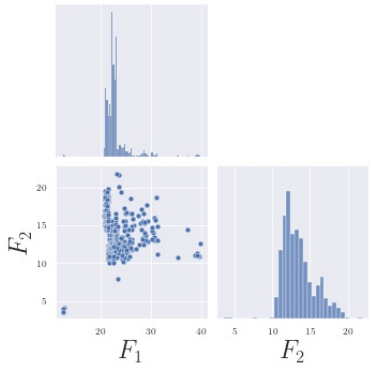

Figura 13 – Gráfico tipo matriz das variáveis relativas ao frete

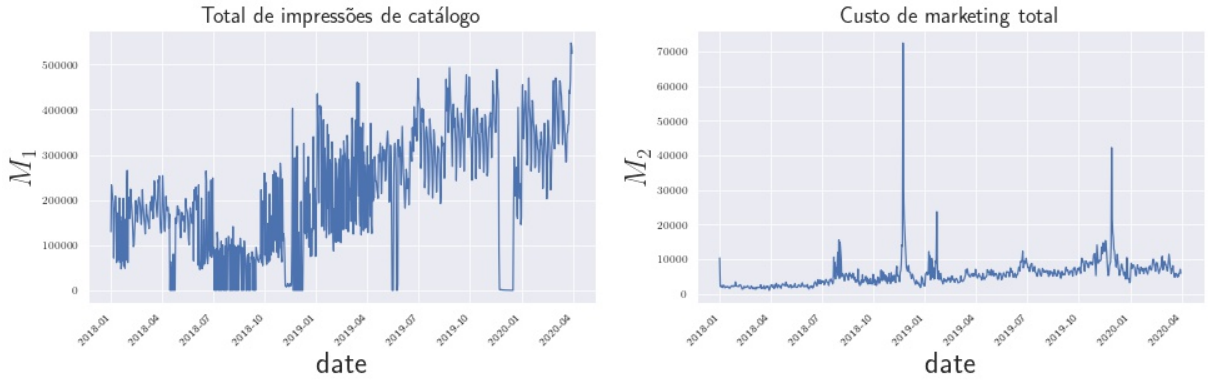

Figura 14 – Evolução temporal das impressões de catálogo e custo de marketing

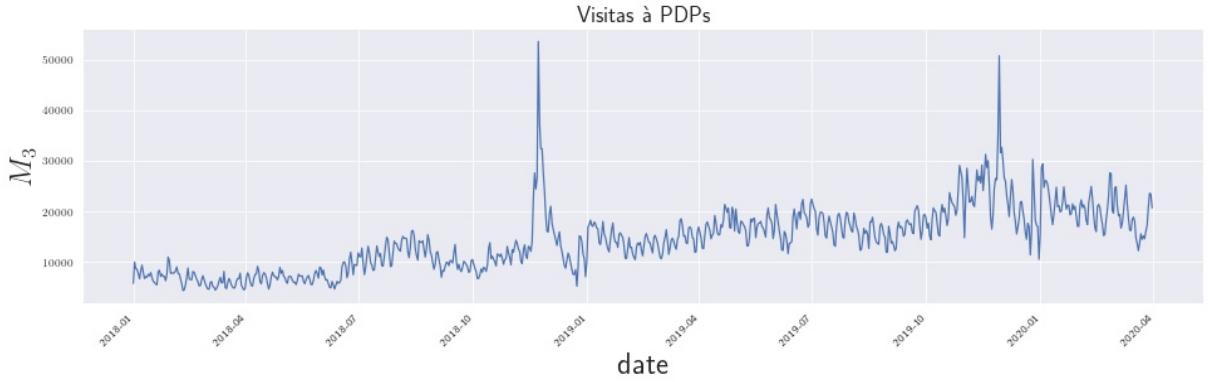

Figura 15 – Evolução temporal das visitas à PDPs

impressões de catálogo é igual a zero.

O custo de marketing também tem uma tendência de crescimento ao longo do tempo, com picos bem evidentes no final de 2018 e 2019, correspondentes à *Black Friday* dos dois anos.

Na Figura 15, temos a evolução temporal das visitas à Páginas de Produto (PDPs). Assim como o custo de marketing, também tem uma tendência de crescimento, sem grandes alterações variabilidade em função do tempo.

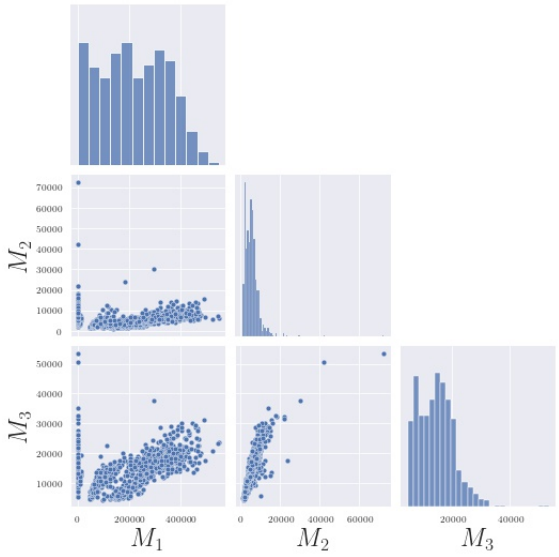

Figura 16 – Gráfico tipo matriz das variáveis relativas à marketing

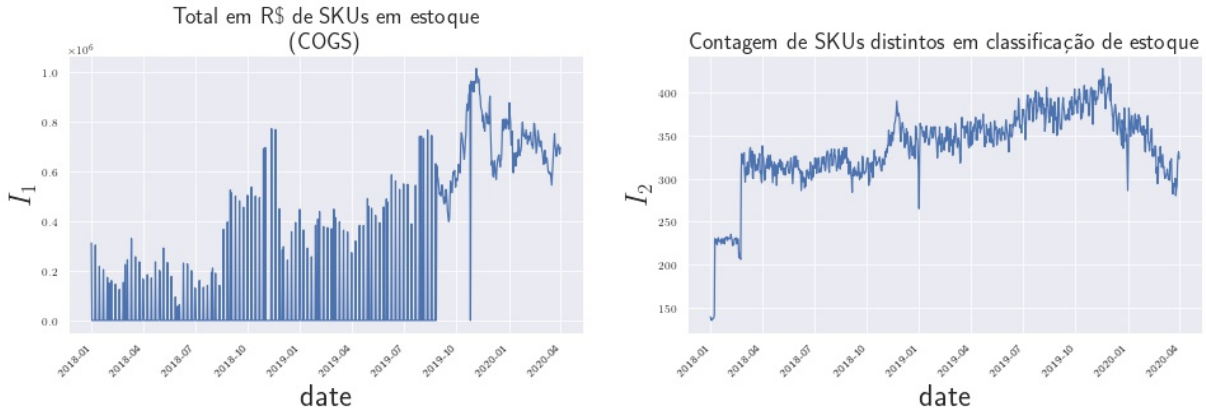

Figura 17 – Evolução temporal das variáveis relativas à estoque

Vemos na Figura 16 que as três variáveis são correlacionadas entre si. Essa é uma observação previsível, já que as três variáveis estão positivamente correlacionadas com o tempo.

4.7 Variáveis de estoque

Há duas variáveis relativas a estoque. I_1 é total de itens em estoque medidos em reais de custo do produto (custo da aquisição dos itens do fornecedor - *sell in*) e I_2 é contagem de SKUs distintos em classificação de estoque, isto é, SKUs cuja manutenção de estoque é considerada desejável.

Figura 17 mostra a evolução temporal dessas duas variáveis. Vemos que o custo em estoque, assim como impressões de catálogo, tem muitos dias antes antes de 2020 registrados com valor zero.

Mas como no gráfico tipo matriz apresentado na Figura 18 nota-se que as duas variáveis

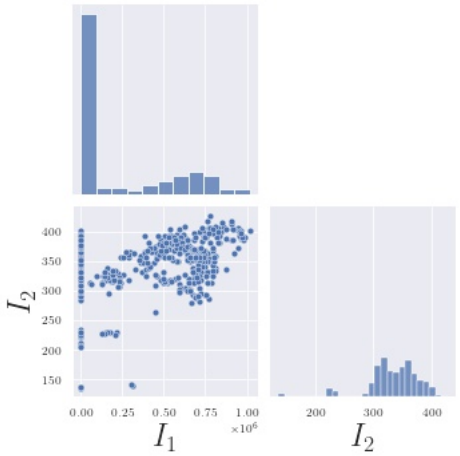

Figura 18 – Gráfico tipo matriz das variáveis relativas à estoque

estão correlacionadas entre si, pode ser necessário utilizar somente uma das variáveis.

4.8 Características de SKUs

Essas variáveis são contagem de SKUs distintos que possuem uma determinada característica. U_1 é a contagem de SKUs em promoção; U_2 é a contagem total de SKUs distintos; U_3 é a contagem total de SKUs fora de linha; U_4 é a contagem de SKUs visíveis;U_5 é a contagem de SKUs exclusivos e U_6 é a contagem de SKUs no programa de beneficiamento (a Mobly acompanha parte do processo produtivo do produto).

O número de SKUs em beneficiamento (U_6) é constante até 2020, o que provavelmente impacta negativamente o poder explanatório dessa variável. Todas demais variáveis são positivamente correlacionadas com o tempo de forma que é esperado algum grau de correlação entre elas.

Na Figura 20, vemos que as variáveis de contagem de SKUs total, em promoção e visíveis é tão similar que a inclusão de duas delas poderia causar problemas de colinearidade em modelo.

Na Figura 21, vemos o gráfico tipo matriz entre contagem de SKUs total, skus exclusivos e fora de linha. Observa-se que há uma certo nível de correlação entre essas variáveis que também não é forte o suficiente para justificar a escolha de necessariamente apenas uma delas.

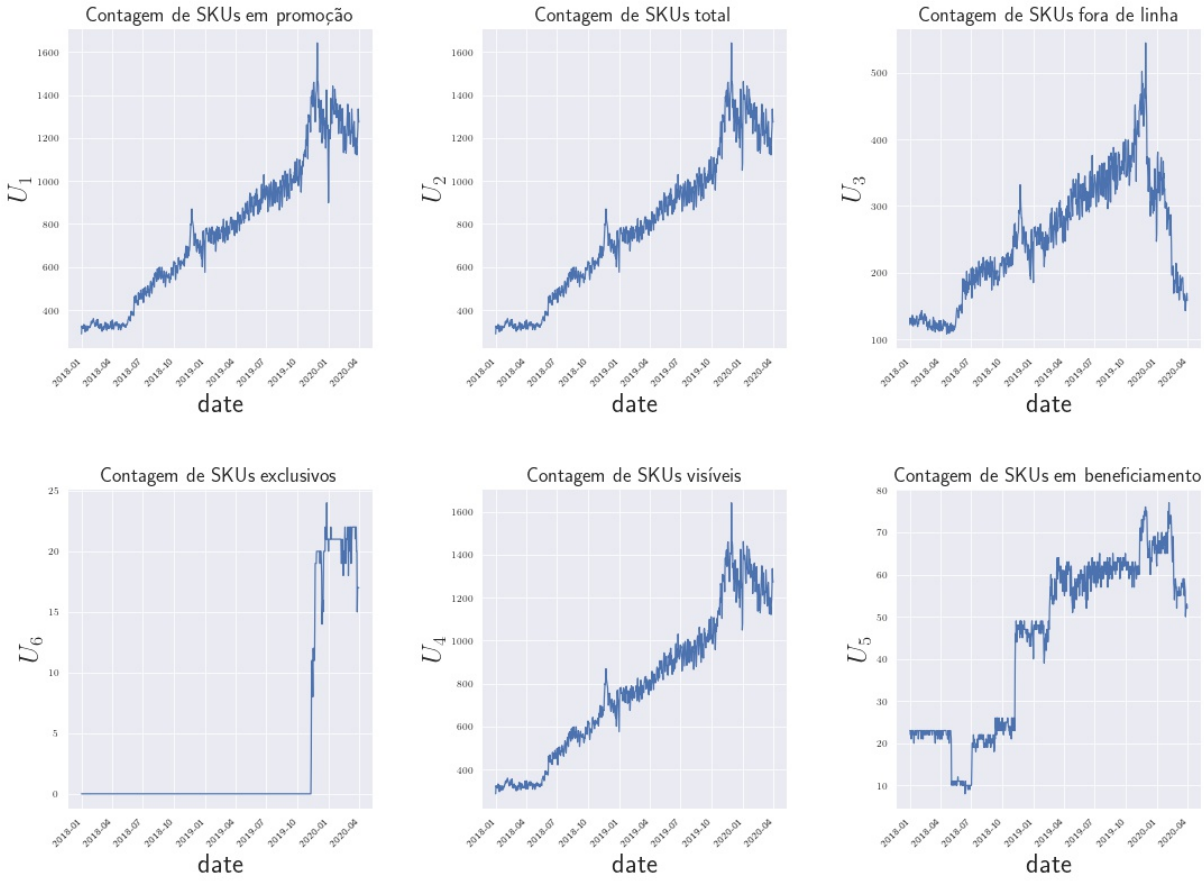

Figura 19 – Evolução temporal das variáveis relativas à SKUs

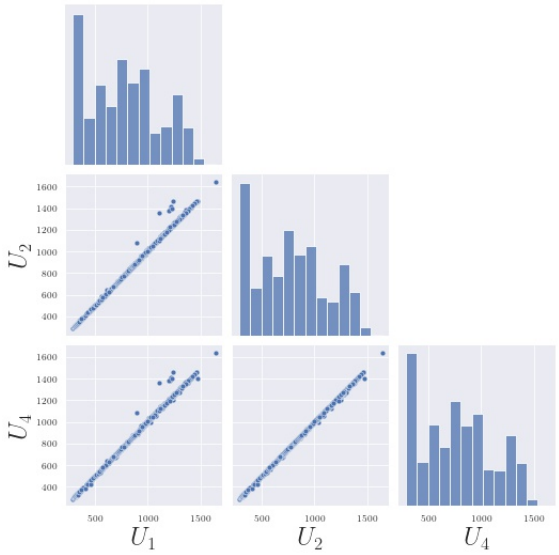

Figura 20 – Gráfico tipo matriz entre contagem de SKUs total, em promoção e visíveis

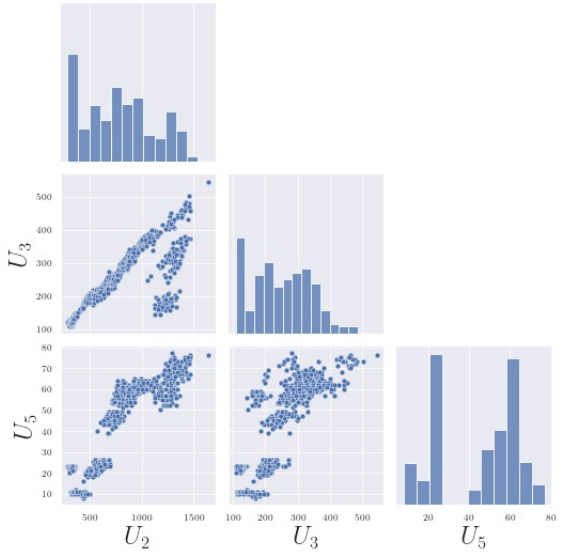

Figura 21 – Gráfico de matriz entre contagem de SKUs total, exclusivos e fora de linha

5 ESPECIFICAÇÃO E DIAGNÓSTICO DO MODELO

Nesse capítulo é tratado sobre a forma final do modelo escolhido, os valores dos coeficientes estimados e da aplicação dos métodos gráficos e estatísticos de diagnóstico de modelo para confirmar que o modelo satizaz as suposicões do modelo de regressão linear múltipla.

5.1 Especificação do modelo

O modelo final escolhido é um modelo de regressão linear múltipla com quinze variáveis, sendo dez delas variáveis dummy. A forma do modelo está descrita na equação 8.

$$\hat{y} = \beta_0 + \beta_1 M_3 + \sum_{i=2}^{8} \beta_i B_{i-3} + + \sum_{i=9}^{11} \beta_i W_{i-5} + \beta_{12} S_1 + \beta_{13} C_1 + \beta_{14} S_2 + \beta_{15} C_2 \tag{8}$$

A descrição de todas as variáveis e os valores de seus respectivos coeficientes estão representados na Tabela 2.

Já na Tabela 3, temos o sumário das estatísticas do modelo completo.

Destaca-se também nessa tabela o coeficiente de determinação R^2 de 0.925. Esse valor indica que o modelo final consegue explicar 92,5% da variação da variável resposta.

	Descrição	**coef**	**Erro padrão**	**t**	**P> |t|**
β_0	Constante	21.7461	1.572	13.833	0.000
β_1	M_3 - Número de visitas à PDPS	0.0046	9.55e-05	47.807	0.000
β_2	B_{-1} - Dia anterior à *Black Friday*	58.5972	10.833	5.409	0.000
β_3	B_0 - Dia da *Black Friday*	631.0528	11.432	55.201	0.000
β_4	B_1 - 1 dia depois da *Black Friday*	216.8261	10.998	19.716	0.000
β_5	B_2 - 2 dias depois da *Black Friday*	143.5442	10.927	13.137	0.000
β_6	B_3 - 3 dias depois da *Black Friday*	45.7867	10.841	4.223	0.000
β_7	B_4 - 4 dias depois da *Black Friday*	49.2270	10.793	4.561	0.000
β_8	B_5 - 5 dias depois da *Black Friday*	35.0426	10.773	3.253	0.001
β_9	W_4 - Quarta-feira	-6.7806	1.588	-4.269	0.000
β_{10}	W_5 - Quinta-feira	-10.7142	1.589	-6.742	0.000
β_{11}	W_6 - Sexta-feira	-7.6949	1.574	-4.889	0.000
β_{12}	S_1 - $\sin 30t$	-2.3029	0.746	-3.087	0.002
β_{13}	C_1 - $\cos 30t$	-2.7117	0.748	-3.626	0.000
β_{14}	S_2 - $\sin 60t$	2.1390	0.751	2.850	0.004
β_{15}	C_2 - $\cos 60t$	-1.8335	0.752	-2.439	0.015

Tabela 2 – Estimativas dos coeficientes de regressão

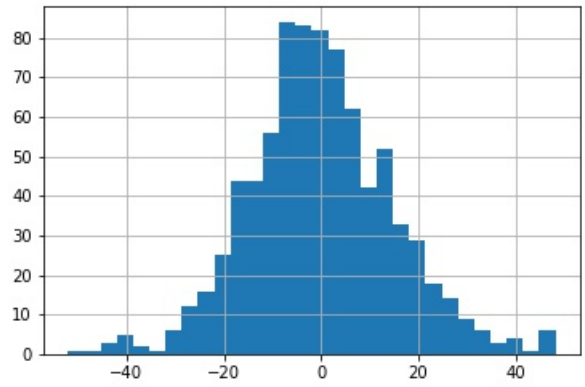

Figura 22 – Histograma dos resíduos

Antes de fazer a interpretação dos coeficientes do modelo final que será objeto do próximo capítulo, iremos fazer o diagnóstico do modelo ajustado.

5.2 Diagnóstico do modelo

Segundo o Capítulo 2, o diagnóstico do modelo é realizado primeiramente pela utilização de métodos gráficos para a confirmação de suposições sobre os resíduos. Mais especificamente, deseja-se confirmar que os resíduos são *i.i.d.*

Na Figura 22, vemos o histograma dos resíduos, que indica que os resíduos do modelo obedecem a uma distribuição normal com média zero. Ao fazer um teste de aderência Kolmogorov-Smirnov sobre os resíduos do modelo testando a hipótese nula de que os resíduos são originados de uma distribuição normal, obtemos o p-valor de 0.2226, o que significa que não podemos rejeitar, a um nível de 5% de significância a hipótese de os resíduos pertencem a uma distribuição normal.

Na Figura 23, temos a o gráfico de série temporal dos resíduos ao longo do tempo. O primeiro fato a ser observado é que os resíduos tem média zero ao longo do tempo, sem apresentar nenhum comportamento de tendência, sazonalidade ou ciclicidade.

Figura 24 apresenta o gráfico de dispersão entre os resíduos e o número de visitas. Esse gráfico também não indica nenhuma relação entre os resíduos e essa variável. Os resíduos aparentam ter média igual a zero, mesma variância e amplitude independentemente da quantidade de visitas. Vemos também que existem alguns (poucos) pontos que são outliers em números de visitas. Todos esses pontos correspondem a dias próximos da *Black Friday*, que estão destacados em laranja.

Nas Figuras 25 e 26 temos os gráficos de dispersão com as variáveis cíclicas de trinta e sessenta dias, respectivamente. As observações a respeito dessa relação são similares às variáveis

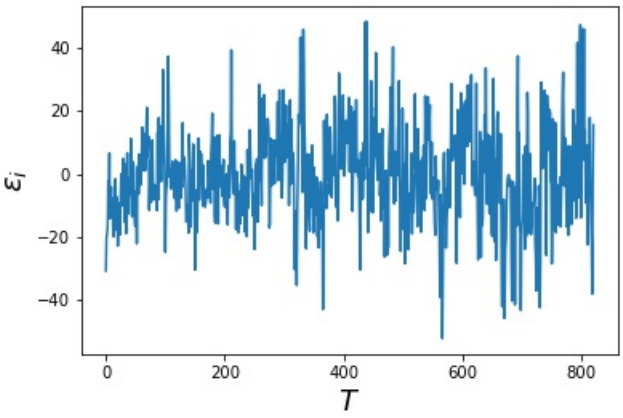

Figura 23 – Série temporal dos resíduos

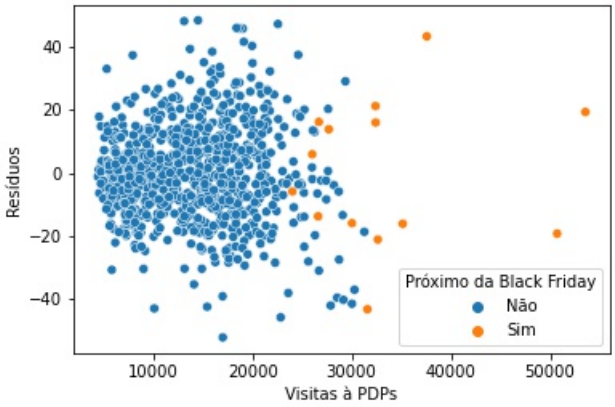

Figura 24 – Resíduos em relação à visitas

já observadas. Os resíduos não parecem apresentar nenhuma relação com essas variáveis pois, independentemente de seu valor, os resíduos continuam tendo média zero e variância e amplitude similar.

Na Figura 27, temos o gráfico Quantil-Quantil dos resíduos. Observa-se que os pontos no gráfico estão distribuidos ao longo da linha de 45º. Isto indica que os resíduos estão aderentes a uma distribuição distribuição normal.fazer teste de aderencia

Os métodos gráficos utilizados mostram que os resíduos satisfazem ãs suposições sobre o modelo de regressão linear. Os resíduos obedecem a uma distribuição normal com média zero, e não apresentam nenhuma relação significante com nenhuma das variáveis explicativas observadas nesse modelo.

Com essas suposições confirmadas, estamos em posição de completar o diagnóstico do modelo através da utilização de estatísticas descritivas. Tabela 2 mostra um sumário do modelos a nível das variáveis explicativas. Esta tabela contém os valores das estatísticas t e seus p-valores

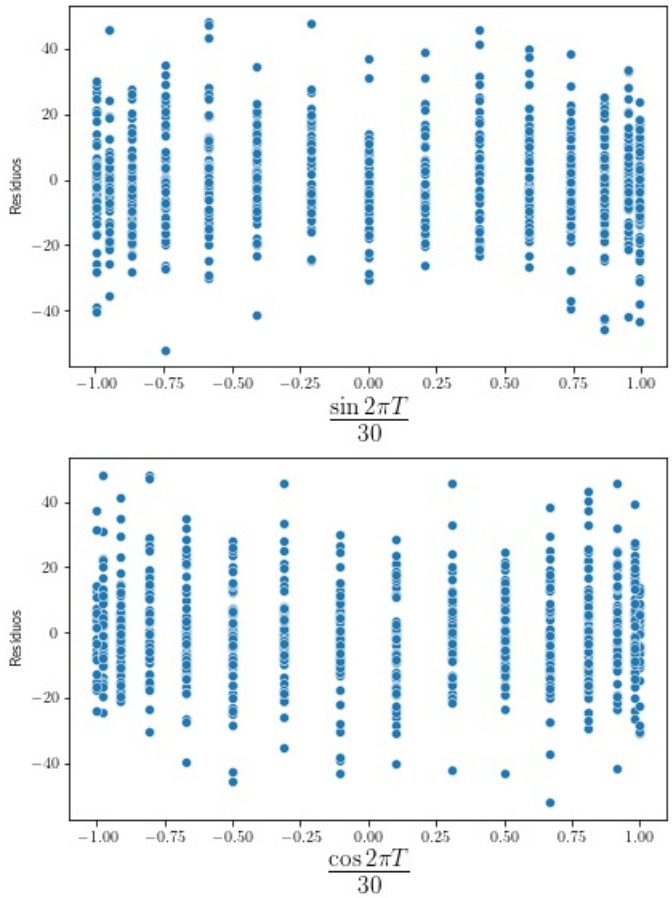

Figura 25 – Resíduos em relação às variáveis cíclicas de 30 dias

Dep. Variable:	y	**R-squared:**	0.925
Model:	OLS	**Adj. R-squared:**	0.923
Method:	Least Squares	**F-statistic:**	659.6
Date:	Tue, 09 Nov 2021	**Prob (F-statistic):**	0.00
Time:	10:38:56	**Log-Likelihood:**	-3385.9
No. Observations:	821	**AIC:**	6804.
Df Residuals:	805	**BIC:**	6879.
Df Model:	15		

Tabela 3 – Estatísticas descritivas do modelo

para todas as variáveis.

Na coluna de p-valores, observervamos que todos os p-valores respeitam o nível de significância de 2,5%. Em outras palavras, pode-se afirmar, a um nível de 2,5% de significância, que todas variáveis utilizadas pelo modelo são variáveis infomativas que contribuem para o poder de predição do modelo.

Pode-se prosseguir com a interpretação dos coeficentes do modelo.

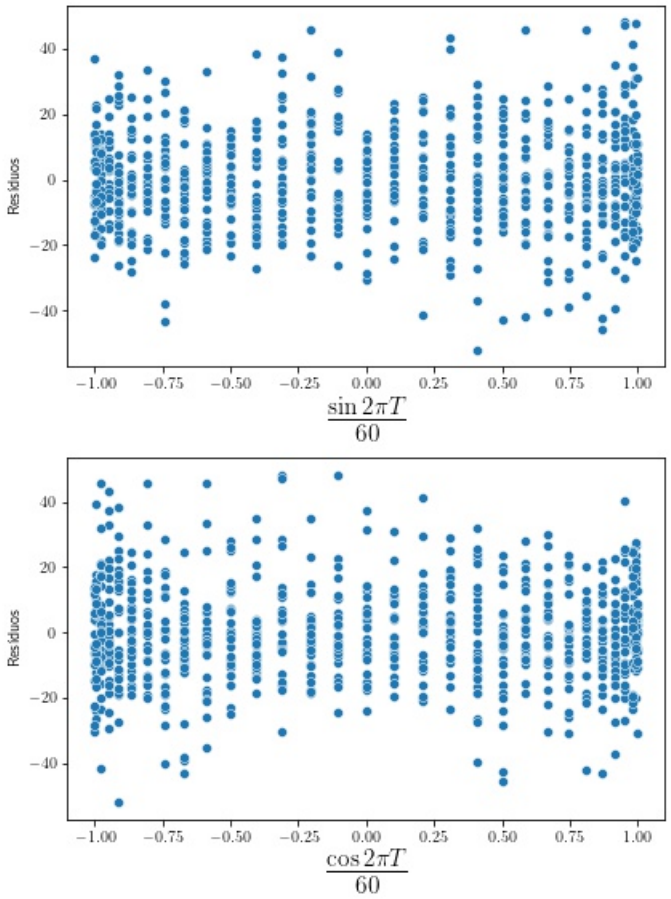

Figura 26 – Resíduos em relação às variáveis cíclicas de 60 dias

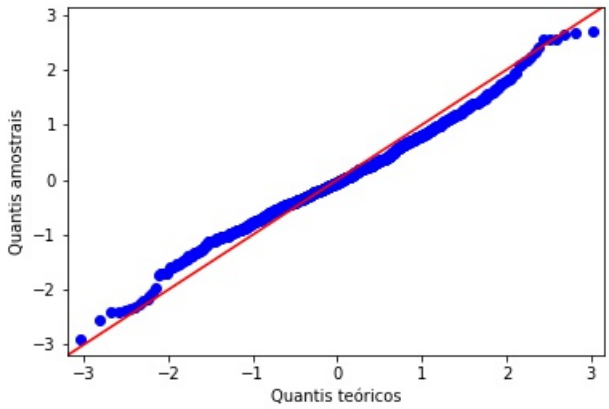

Figura 27 – Gráfico quantil-quantil dos resíduos

Com todas as suposições sobre os resíduos necessárias confirmadas através da análise gráfica, com todas variáveis explicativas sendo significantes e com um coeficiênte de determinação alto, concluí-se que o modelo apresentado está bem ajustado, e satisfaz os objetivos desse trabalho.

6 INTERPRETAÇÃO DOS RESULTADOS

Das variáveis que entraram no modelo final, a que tem o maior coeficiente de correlação com a variável resposta, ou ainda, aquela que melhor explica por conta própria a variável resposta é o número de visitas às páginas de produtos.

É por essa variável, portanto, que começaremos a realizar a interpretação dos insights fornecidos pelo modelo. Essa escolha é adequada também porque o coeficiente de regressão dessa variável pode ser facilmente compreendida como uma grandeza física: taxa de conversão entre visitas a página de produto e realização da compra, medida em reais por 1000 visitas.

O coeficiente gerado pelo ajuste do modelo é de 0.0046 milhares de reais por visitas, ou seja, R$ 4,6 por visita. Esse valor sugere inicialmente que, para a categoria de questão, entre 2018 e 2020, a Mobly poderia esperar faturar uma média de R$ 4,6 por visita.

Esse é um resultado muito interessante, pois ao mesmo tempo que demonstra a importância financeira para atrair visitas à paginas de produto, também estabelece um valor por visita. Esse resultado poderia ser usado, por exemplo, para calcular a que custo a aquisição de visitas é vantajosa para a empresa. Contudo, será discutido como não podemos interpretar esse valor por conta própria sem antes entender como a taxa de conversão se relaciona com as outras variáveis do modelo.

Vemos na Figura 28 a evolução da taxa de conversão da categoria durante o período estudado. De fato, pode-se constatar que a taxa de conversão se manteve em torno de um valor estável. Contudo, destaca-se também que esse valor é visivelmente diferente do que aparece no modelo. Enquanto o modelo sugere um valor próximo a R$ 4,5 por visita, na Figura 28 a verdadeira taxa de conversão parece estar mais próxima de R$ 6 por visita.

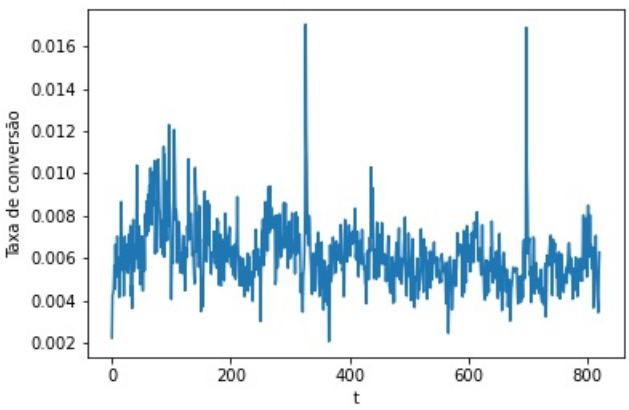

Figura 28 – Taxa de conversão

Isso pode ser entendido da seguinte maneira. O valor do modelo próximo de R$ 4,50 por visita não é a taxa de conversão final da empresa, mas a taxa de conversão antes de ser considerados os efeitos que outras variáveis explicativas têm sobre a taxa de conversão.

No Capítulo 2, foi discutido que a interpretação dos coeficientes de regressão parciais não são fáceis de interpretar, e que a maneira mais correta de interpretá-los, é que o coeficiente de regressão parcial de uma variável representa o efeito dessa variável após o efeito das demais variáveis já terem sido consideradas. Este conceito é pouco intuitivo, e deve ser abordado com cuidado.

A conclusão que devemos chegar é que, enquanto o coeficiente de regressão da variável M_3 (visitas) representa uma taxa de conversão "base", os coeficientes de regressão das demais variáveis sinalizam efeitos que afetam essa taxa de conversão.

Essa interpretação já aponta na direção da resposta sobre o porquê do coeficiente de regressão de visitas ser diferente da taxa de conversão real da empresa, e essa resposta é porque essa diferença e explicada pelo efeito das demais variáveis sobre a taxa de conversão.

O caso das variáveis relativas a *Black Friday* é bem interessante para ilustrar esse efeito. Todos os dias próximos ao evento têm efeitos positivos sobre a venda. Uma interpretação menos cuidadosa poderia chegar a conclusão que esses coeficientes ilustram nada mais que o aumento de vendas típico da *Black Friday*. Entretanto, deve-se recordar que o modelo já utiliza a variável de visitas, que já é uma variável sinalizadora da escala de vendas. Em outras palavras, é perfeitamente plausível que a variável visitas já poderia perfeitamente ter "embutida"em si o efeito da *Black Friday*, pois já existe aumento de vendas durante a *Black Friday*.

De fato, no Capítulo 4, observa-se na Figura 15 que o número de visitas de fato tem picos durante os dias de *Black Friday*. Se fosse adotado que a *Black Friday* tem efeito sobre o número de visitas, mas não sobre a taxa de conversão, então informações sobre *Black Friday* seriam desnecessárias se já tivéssemos o número de visitas.

Em resumo, se o número de visitas já está incorporado no modelo, variáveis relativas à *Black Friday* não necessariamente teriam um efeito significativo. O fato que isso ocorre indica que a *Black Friday* não somente tem um efeito sobre o número total de visitas, mas também sobre o a taxa de conversão. Isso é ilustrado na Figura 29. Essa visualização confirma que a taxa de conversão é visivelmente maior nos dias próximos da *Black Friday*.

As possíveis explicações para essa mudança na taxa de conversão levanta hipóteses muito interessantes sobre o comportamento do consumidor perto da *Black Friday*. É possível, por exemplo, que muitos clientes façam suas pesquisas de produtos e decisões de compra em antecedência a *Black Friday*, e no dia da *Black Friday* acessem as páginas de produtos com a intenção de comprar.

Essa hipótese, sendo verdadeira, sugere que campanhas de comunicação sobre campanhas de *Black Friday* alguns dias ou semanas antes do evento podem ser tão importante quanto

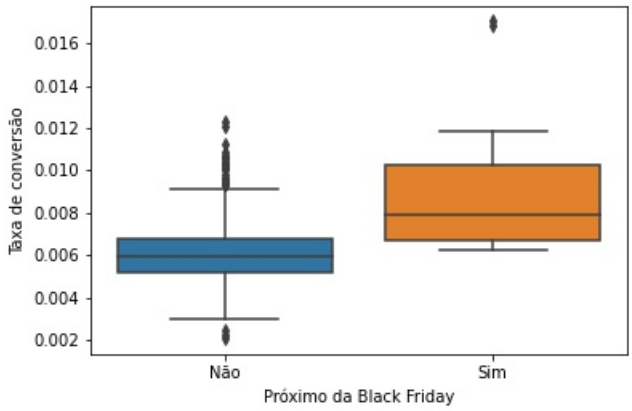

Figura 29 – Taxa de conversão próximo da Black Friday

comunicação durante o evento.

Alternativamente, é possível também que o aumento da taxa de conversão seja devido a um senso de urgência do cliente criado pelas campanhas, que leva a maior decisividade por parte do cliente sobre a compra.

Essas são duas hipóteses, entre várias outras que podem ser consideradas. O importante é que, de qualquer forma, o modelo reconheceu um efeito real e significativo, cujo entendimento certamente evidenciará oportunidades de ações que podem ser tomadas para gerar aumento de vendas.

Outra categoria de variáveis com comportamento similar é de dias da semana. Não são todos dias da semana que aparecem no modelo. Os únicos são quinta-feira, sexta-feira e sábado, todos com efeitos negativos. Uma consequência disso é que os demais dias tem um efeito "positivo". Isto é, a previsão do modelo para esses dias será, em geral, mais positivos que para quintas, sextas e sábados, porque todos esses dias tem efeitos negativos.

O significado no mundo real desses coeficientes também pode ser relacionado com taxas de conversão, na mesma forma que as variáveis de *Black Friday*. Os dias representados no modelo tem efeitos negativos sobre a taxa de conversão. Isso é ilustrado na Figura 30.

E de forma similar ao efeito da *Black Friday*, esse efeito também pode ser bem esclarecedor sobre o comportamento do consumidor. É possível que os clientes estejam fazendo pesquisas quinta, sexta e sábado, para tomar a decisão e realizar a compra entre domingo e quarta-feira.

Uma possível ideia que poderia ser utilizada para capitalizar sobre esse efeito seria utilizar uma estratégia de comunicação no site em que priorize mostrar uma maior variedade de produtos que possam despertar o interesse dos clientes entre quinta e sábado, e gerar maior senso de urgência para compra nos dias de maior conversão.

As próximas variáveis a serem interpretadas serão as variáveis cíclicas. São utilizadas

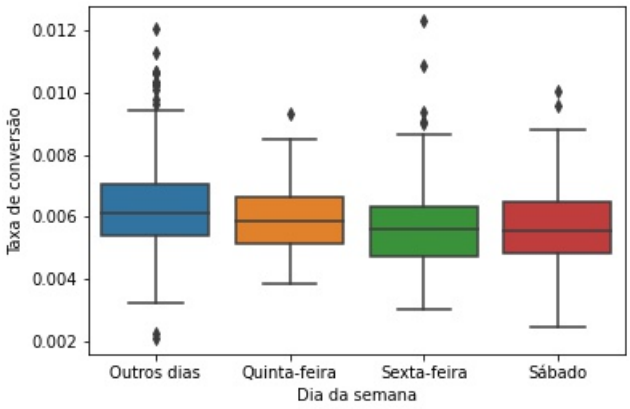

Figura 30 – Taxa de conversão por dia da semana

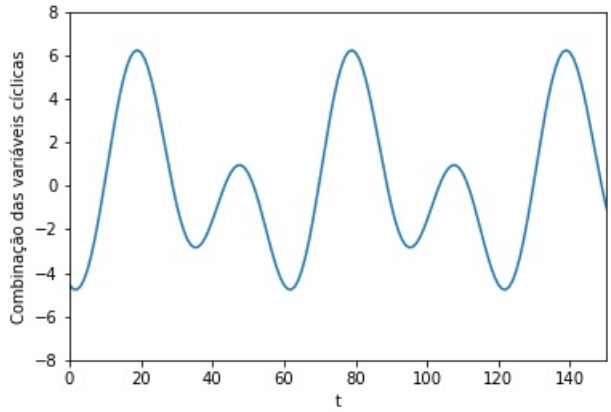

Figura 31 – Resultado da soma das variáveis cíclicas

quatro variáveis cíclicas, cujo efeito total, isto é, a soma de seus valores multiplicados pelos seus coeficientes de correlação, está apresentado na Figura 31. Observa-se que as quatro variáveis cícliclas, juntas, representam um efeito cíclico de que pode ser descrito como um efeito positivo no meio do mês (próximo do dia quinze) e negativo perto das viradas dos meses. Pode ser entendido que esse efeito é, em geral, forte em um mês e fraco em outro.

Além da interpretação prática dos coeficientes de regressão, outra característica do modelo que queremos avaliar é seu desempenho em termos de geração de previsões.

O coeficiente de determinação do modelo, como mencionado no Capítulo 5 é de 0.925, o que significa que o modelo é capaz de explicar 92,5% da variância da variável resposta. Para entender como esse nível de acurácia "se parece", pode-se referenciar a Figura 32.

Nessa figura, estão representados em laranja os valores Y reais da variável resposta e em azul os valores $\hat{Y}$ previstos pelo modelo. Na visualização, para fim de legibilidade, estão somente os últimos 220 valores do *dataset*.

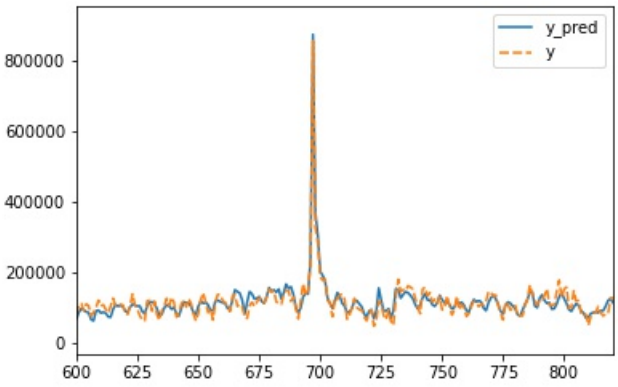

Figura 32 – Valores previstos vs valores reais

Observa-se que com esse nível de acurácia, o modelo consegue explicar quase perfeitamente a venda diária da categoria de Guarda-Roupas no período estudado incluindo pequenas flutuações semanais dos dias próximos à *Black Friday*.

Sabendo o número de visitas à PDPs, a data, e informações que podem ser derivadas da data, é possível explicar muito bem as vendas de uma categoria de produtos.

7 CONCLUSÃO

Foi realizada uma análise de regressão das vendas de *webshop* da cateogoria de Guarda-Roupas para dados observando entre 01/01/2021 e 31/03/2020. Foi possível desenvolver um modelo de regressão linear múltipla que, utilizando apenas informações sobre o número de visitas a páginas de produtos e informações derivadas da data, é capaz de explicar 92,5% da variação das vendas.

O diagnóstico do modelo foi realizado com diversas características descritivas e métodos gráficos que verificam que o modelo desenvolvido satisfaz as suposições do modelo de regressão linear múltipla, e portanto descreve os principais efeitos genuínos que afetam as vendas que podem ser capturados a partir dos dados utilizados.

O principal benefício oferecido pelo modelo à empresa é que, entendendo os principais fatores que afetam as vendas, é possível planejar ações eficazes para o aumento de vendas.

É sugerido pelo modelo que a principal variável que deve ser afetada para o aumento de vendas é o número de visitas. Isso significa que conseguir que clientes acessem as páginas de produtos é um marco especialmente importante no fluxo de vendas.

Os clientes do canal online, também chamado de *webshop*, tem que passar por uma série de estágios até a decisão pela compra. Esse processo se inicia com o acesso ao site da Mobly, observar a listagem de produtos, clicar em algum produto, decidir por adicioná-lo ao carrinho de compras e executar o pagamento.

Alternativamente, clientes também podem ser redirecionados ao site da Mobly a partir de sites de comparação de preços ou por serviços de anúncios. Nesse caso, é possível que já sejam direcionados diretamente para as páginas de produto.

O que a análise de regressão realizada nesse trabalho está dizendo é que um marco extremamente importante durante esse fluxo é quando o cliente acessa uma página de produto. Uma boa ilustração disso é que, ainda segundo o modelo, se a empresa soubesse o número de visitas à páginas de produtos futuro, conseguiria prever as vendas com R^2 de 0.925.

Esse trabalho começou com a pergunta: **"Quais as principais variáveis que impactam as vendas?"** e foi constatado que, para as vendas de *webshop*, visitas é uma variável particularmente importante.

Esse trabalho, em consequência disso, sugere três linhas linhas de trabalho que podem ser seguidas pela empresa para aumentar o resultado financeiro.

A primeira dessas linhas vêm de que é possível atribuir um determinado valor que uma visita (R$ 4,5 por visita), trás à Mobly, e margens financeiras podem ser utilizadas para calcular o valor retornado por visitas em termos de lucro operacional.

Assim pode existir uma iniciativa de otimização de investimentos em canais de marketing digital. O custo por visita proveniente de um canal pode ser calculado a partir do valor gasto naquele canal dividido pela quantidade de visitas geradas pelo canal.

É possível, através dessas medidas, calcular para cada canal, um "investimento ótimo"em marketing que maximize o lucro gerado.

Outra ideia é considerar que, ainda sobre a ideia que pode ser atribuído um valor de receita bruta para cada visita, e que a margem de venda dos produtos é conhecida, então é possível otimizar uma seleção de produtos cuja exposição no site seja altamente rentável.

Notar que tanto produtos com margem alta mas que geram poucas visitas; quanto produtos com alta geração de visitas mas pouca margem não seriam os mais rentáveis.

Os melhores produtos seriam aqueles que tenham tanto alta margem e alta geração de visitas. E se existir um *tradeoff* entre essas duas características, os melhores produtos seriam os que se posicionarem em um ponto ótimo nessa curva.

A terceira linha de ação vêm de que se foi constatada a importância da variável visitas, então naturalmente pode-se perguntar: **Mas o que gera visitas?**. E essa questão pode ser expandida em uma gama de outras questões relacionadas.

Qual a maior fonte de visitas da Mobly? Vêm de dentro do próprio site ou redirecionados de sites externos?

O posicionamento dos produtos nas páginas de cátalogo influencia visitas? E a imagem, afeta? Que tal o preço? Talvez avaliações?

Se a empresa conseguir responder essas questões e tomar ações que aumentem o número de visitas à páginas de produto, definitivamente os resultados serão excelentes.

Por uma perspectiva mais ampla esse trabalho também é uma demonstração de como a dados podem ser utilizados para gerar valor dentro do contexto de negócio do varejo. Junto com modelagem de demanda, nesse setor há diversos processos e decisões de negócios que podem ser melhorados através do suporte de dados, modelos estatísticos e análise criteriosa de dados.

Através de uma análise estatística criteriosa, foi possível alavancar os dados disponíveis para desenvolver um modelo cuja a interpretação irá dar suporte importante para tomada de decisões táticas e estratégicas futuras, e certamente também mudar como certas ações são avaliadas.

REFERÊNCIAS

AMERICANAS S.A. *Earnings Release 3Q 2021*. 2021. Citado na página 26.

ANSCOMBE, F. J. Graphs in statistical analysis. *The American Statistician*, v. 27, 2 1973. ISSN 00031305. Citado na página 39.

CHATTERJEE, S.; HADI, A. S. *Regression Analisys by Example*. Cairo: John Wiley & Sons, Inc., 2012. ISBN 978-0-470-90584-5. Citado 8 vezes nas páginas 31, 32, 33, 34, 39, 40, 45 e 48.

DRAPER, N. R.; SMITH, H. *Applier regression analysis*. Canada: John Wiley & Sons, Inc., 1998. ISBN 0-471-17082-8. Citado 2 vezes nas páginas 37 e 49.

FONSECA, M. E-commerce de descoberta: como a westwing navega o oceano vermelho das compras online. *Infomoney*, 21 abr. 2021 2021. Citado na página 26.

HARRIS, C. R. et al. Array programming with NumPy. *Nature*, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: <https://doi.org/10.1038/s41586-020-2649-2>. Citado na página 47.

HUNTER, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007. Citado na página 49.

KUTNER, M. H. et al. *Applied Linear Statistical Models*. Nova Yorque: McGraw-Hill/Irwin, 2005. ISBN 0-07-238688-6. Citado 3 vezes nas páginas 36, 45 e 46.

LINCOLN, A. *Address before the Wisconsin State Agricultural Society*. 1859. Disponível em: <http://www.abrahamlincolnonline.org/lincoln/speeches/fair.htm>. Acesso em: 30 nov. 2021. Citado na página 9.

MAGAZINE LUIZA. *Divulgação de Resultados 3T21*. [S.l.], 2021. Citado 2 vezes nas páginas 22 e 26.

MOBLY. *Management Report of the Third Quarter of 2021 results*. [S.l.], 2021. Citado na página 22.

MOBLY. *Relação com investidores*. 2021. Disponível em: <https://investors.mobly.com.br/en/>. Acesso em: 25 nov. 2021. Citado 2 vezes nas páginas 21 e 22.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado 2 vezes nas páginas 48 e 50.

RIVAS, K. Ação da mobly 'derreteu' mais de 50que esperar? *Invest News*, 10 set. 2021 2021. Disponível em: <https://investnews.com.br/financas/mobly-o-que-esperar-da-acao/>. Acesso em: 25 nov. 2021. Citado na página 26.

SCHRÖER, C.; KRUSE, F.; GÓMEZ, J. M. A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, v. 181, 2021. ISSN 18770509. Citado na página 45.

SEABOLD, S.; PERKTOLD, J. statsmodels: Econometric and statistical modeling with python. In: *9th Python in Science Conference*. [S.l.: s.n.], 2010. Citado na página 50.

THE PANDAS DEVELOPMENT TEAM. *pandas-dev/pandas: Pandas*. Zenodo, 2020. Disponível em: <https://doi.org/10.5281/zenodo.3509134>. Citado na página 47.

VIA VAREJO. *Earnings Release 3Q2021*. 2021. Citado na página 26.

WASKOM, M. L. seaborn: statistical data visualization. *Journal of Open Source Software*, The Open Journal, v. 6, n. 60, p. 3021, 2021. Disponível em: <https://doi.org/10.21105/joss.03021>. Citado na página 49.

Apêndices

APÊNDICE A – TABELA COMPLETA DAS VARIÁVEIS DO CONJUNTO DE DADOS

Variaveis numéricas	tipo
date	object
t	float64
avg_price	float64
special_skus	float64
avg_special_price	float64
avg_actual_price	float64
avg_delivery_time_sp	float64
avg_shipping_amount_sp	float64
catalog_impressions	float64
visits	float64
mkt_cost	float64
available_cogs_net_stock	float64
wanted_stock_classification	float64
skus_count	int64
out_of_order_count	float64
active_count	float64
visible_count	float64
private_label_count	float64
imported_count	float64
process_costing_count	float64
cos30	float64
sin30	float64
cos60	float64
sin60	float64

Variáveis categóricas	tipo
black_friday_m1	bool
black_friday_m2	bool
black_friday_m3	bool
black_friday_m4	bool
black_friday_m5	bool
black_friday_m6	bool
black_friday_m7	bool
black_friday_p0	bool
black_friday_p1	bool
black_friday_p2	bool
black_friday_p3	bool
black_friday_p4	bool
black_friday_p5	bool
black_friday_p6	bool
month_1	int64
month_2	int64
month_3	int64
month_4	int64
month_5	int64
month_6	int64
month_7	int64
month_8	int64
month_9	int64
month_10	int64
month_11	int64
month_12	int64
weekday_0	int64
weekday_1	int64
weekday_2	int64
weekday_3	int64
weekday_4	int64
weekday_5	int64
weekday_6	int64
y	float64

Tabela 4 – Variáveis do conjunto de dados