PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO

Previsão de Vendas do E-commerce usando Modelos Estatísticos e Métodos de Machine Learning

João Pedro Jesus de Abreu Martinez

Relatório de Projeto Final

Centro Técnico Científico - CTC
Departamento de Informática
Curso de Graduação em Engenharia da Computação

Rio de Janeiro
Janeiro de 2024

João Pedro Jesus de Abreu Martinez

Previsão de Vendas do E-commerce usando Modelos Estatísticos e Métodos de Machine Learning

Relatório de Projeto Final

Relatório de Projeto Final, apresentado ao programa de Engenharia de Computação da PUC-Rio como requisito parcial para a obtenção do titulo de Engenheiro de Computação.

Orientador: Prof. Cristiano Augusto Coelho Fernandes

Rio de Janeiro
Janeiro de 2024

Agradecimentos

Gostaria de expressar meus sinceros agradecimentos a todos que contribuíram para a realização deste trabalho final de conclusão de curso. Em primeiro lugar, quero expressar minha gratidão aos meus pais, Artur e Luciane, pelo apoio incondicional em todos os momentos desafiadores da minha vida. Seu incentivo e amor foram fundamentais para que pudesse concluir minha graduação.

Não poderia deixar de mencionar minha profunda gratidão ao prof. Cristiano Augusto Coelho Fernandes, meu orientador, pela dedicação, paciência e suporte ao longo do desenvolvimento deste trabalho. Seu comprometimento foi crucial para meu crescimento acadêmico e profissional. Agradeço também ao prof. Alvaro de Lima Veiga Filho por proporcionar oportunidades valiosas de experiência profissional ao longo da minha formação. Cada um de vocês desempenhou um papel significativo na minha jornada e sou imensamente grato por isso.

Agradeço também aos amigos que fiz ao longo da graduação e que tornaram essa jornada mais significativa, em especial a Matheus Nogueira e João Pedro de Paiva. Agradeço ainda aos meus amigos Jônatas Pessanha e Taylor Oliveira pelas experiências profissionais e projetos em que trabalhamos juntos.

Resumo

Martinez, João Pedro. Fernandes, Cristiano. Previsão de Vendas do E-commerce usando Modelos Estatísticos e Métodos de Machine Learning. Rio de Janeiro, 2024. 43p. Projeto de Graduação - Departamento de Engenharia Elétrica e Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Neste projeto investigou-se a acurácia preditiva de modelos estatísticos e métodos de *machine learning* aplicados à cinco séries temporais horárias de quantidade vendida, provenientes de um *e-commerce* varejista. Os modelos selecionados foram regressão dinâmica estimada por mínimos quadrados ordinários (MQO), Lasso e AdaLasso, além do método *random forest*. A acurácia preditiva foi investigada nos horizontes de previsão de 1 à 12 horas à frente, utilizando as métricas MAE e RMSE. Os resultados apontaram os modelos da família Lasso como aqueles de melhor desempenho conforme a métrica MAE. No caso do RMSE, verificou-se os melhores resultados associados ao modelo de regressão dinâmica que incorpora termos auto regressivos da quantidade vendida e variáveis *dummy* (RegrDin(3)). A implementação computacional dos modelos foi realizada utilizando as linguagens de programação Python e R.

Palavras-chave

SKU; *E-commerce*; Previsão de vendas; Modelos estatísticos; *Machine learning*.

Abstract

Martinez, João Pedro. Fernandes, Cristiano. E-commerce Sales Forecasting using Statistical Models and Machine Learning Methods. Rio de Janeiro, 2024. 43p. Undergraduate Thesis - Department of Electrical Engineering and Department of Informatics, Pontifical Catholic University of Rio de Janeiro.

In this project, the predictive accuracy of statistical models and machine learning methods applied to five hourly time series of sales quantity from a retail e-commerce was investigated. The selected models included dynamic regression estimated by ordinary least squares (OLS), Lasso, and AdaLasso, in addition to the random forest method. Predictive accuracy was assessed for forecast horizons ranging from 1 to 12 hours ahead, using the MAE and RMSE metrics. The results indicated that models from the Lasso family exhibited superior performance according to the MAE metric. Regarding RMSE, the best results were associated with the dynamic regression model that incorporates autoregressive terms of the sales quantity and dummy variables (RegrDin(3)). The computational implementation of the models was carried out using the programming languages Python and R.

Keywords

SKU; *E-commerce*; Sales forecasting; Statistical models; *Machine learning*.

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Motivação

Ao analisar o cenário vivido pelo mercado varejista nos últimos anos, percebe-se que este passou por condições muito adversas, em grande parte associadas à crise pandêmica do COVID-19. Uma amostra disso é a queda histórica mensal de 16,8% nas vendas do varejo, registrada em abril de 2020, sendo este o pior resultado registrado em 20 anos (IBGE, 2020). Assim, com o aumento do número de contaminados e tendo sido impostas as restrições de isolamento social, o *e-commerce* se tornou essencial para atender os consumidores e mitigar os impactos econômicos negativos das lojas físicas.

Atualmente, com a pandemia superada por meio da vacinação em massa, graças aos esforços heroicos de cientistas e profissionais da área da saúde, verifica-se que mesmo com o retorno às lojas físicas, o número de pessoas consumindo via *e-commerce* continua a crescer. Uma constatação disto se dá em estudos como o realizado pelas empresas NielsenIQ-Ebit e Bexs Pay e que resultaram no relatório Webshoppers - Edição 46. Neste documento é informado um faturamento de R$ 118,6 bilhões em vendas do comércio eletrônico no país, apenas no primeiro semestre de 2022, equivalente a uma alta de 6% quando comparado ao mesmo período em 2021 (FERNANDES, 2022). Em concordância com isto, outro levantamento feito pela Associação Brasileira de Comércio Eletrônico (ABComm) indica que o segmento deve movimentar R$ 185,7 bilhões até o fim deste ano, crescendo 9,5% ante 2022 (ABCOMM, 2023).

No entanto, com a aceleração da transformação digital do setor, foram proporcionadas condições favoráveis à mudanças no comportamento do consumidor, que não só passou a consumir mais nas lojas *online* do varejo, como também aumentou suas expectativas quanto à jornada de compras. Dentre essas expectativas, está elencada a praticidade que se espera ter na compra e rapidez na entrega dos produtos adquiridos. Com isso, é de extrema importância que as cadeias de suprimentos das varejistas otimizem sua gestão de estoque, evitando desperdícios e garantindo que haja produtos suficientes para atender à demanda.

Nesse contexto, beneficiados pelos avanços na tecnologia da informação e impulsionados pelo forte crescimento da geração e coleta de dados, os modelos estatísticos e métodos de *machine learning* têm se mostrado ferramentas valiosas. Na prática, sua aplicação torna-se uma vantagem competitiva na identificação dos padrões de consumo dos clientes, e por consequência, na tomada de decisões estratégicas por parte dos *players* do setor varejista. Dessa forma, pretende-se empregar modelagens distintas na previsão das vendas *e-commerce* de alguns produtos, sendo estes comercializados por uma empresa varejista. O intuito disto é viabilizar a construção de um ferramental que dê suporte à cadeia de suprimentos do setor como um todo.

1.2 Objetivos do trabalho

Este projeto final tem como objetivo estudar o emprego de diferentes modelos estatísticos e métodos de *machine learning* na previsão das vendas horárias de cinco *stock keeping units* (SKU), comercializados no *e-commerce* de uma empresa nacional pertencente ao setor varejista.

Para tanto, pretende-se utilizar a metodologia conhecida como *direct multi-step forecasting*, incorporando-a na implementação dos modelos selecionados para estudo. O intuito disto é possibilitar o emprego de diferentes horizontes de previsão, estimando um conjunto de H modelos preditivos - derivados de um dos modelos selecionados. Assim, cada estimação do conjunto deverá prever a i-ésima quantidade de vendas do SKU em questão, com $i \in \{1, ..., H\}$ (TAIEB et al., 2012).

Em sequência, espera-se estabelecer um comparativo entre as diferentes abordagens preditivas empregadas, a fim de identificar que modelo estatístico ou método de *machine learning* apresenta melhor adequação na previsão da quantidade vendida de um determinado SKU. Vale lembrar que nesse comparativo também serão considerados os H horizontes de previsão, de modo que mais de uma abordagem preditiva possa ter melhor adequação dado um dos SKU's.

Dentre os modelos estatísticos e métodos de *machine learning* encontrados na revisão da literatura, voltada à previsão de vendas no varejo *e-commerce*, estão incluídas desde abordagens mais tradicionais, até aquelas mais recentes. A partir disso, definiu-se a seguinte lista de modelos que serão empregados ao longo deste projeto:

- Regressão dinâmica, estimada via mínimos quadrados ordinários (MQO);
- Regressão Lasso (*Least absolute shrinkage and selection operator*);
- Regressão AdaLasso, com pesos estimados via regressão Ridge ou Lasso;
- *Random forest*, usando *block bootstrapping*.

Por último, espera-se entregar como produto final deste projeto códigos nas linguagens de programação Python e R, contendo funções capazes de realizar o tratamento das bases de dados analisadas; efetuar o treinamento e teste dos modelos estatísticos e métodos de *machine learning* estudados; comparar a performance preditiva destes por meio de métricas e testes de aderência adequados; mostrar os resultados obtidos por meio de visualizações como gráficos e tabelas.

1.3 Organização

Este texto está organizado em diversas seções que delineiam o escopo e a abordagem adotada no projeto. A Seção 2 realiza uma revisão concisa da literatura, abrangendo modelos estatísticos e métodos de *machine learning* empregados na previsão de séries temporais, desde os tradicionais até abordagens mais recentes.

Na Seção 3 são apresentados os fundamentos teóricos dos modelos e métodos escolhidos para realizar a previsão das séries horárias de quantidade vendida dos SKU's, proporcionando uma base teórica para compreender as estratégias adotadas. A Seção 4 discute os dados utilizados no estudo, descrevendo os tratamentos aplicados, as análises descritivas feitas e ajustes finais efetuados para otimizar o treinamento das abordagens preditivas.

A metodologia empregada é então descrita na Seção 5, esclarecendo os passos e procedimentos seguidos durante o estudo. Já a Seção 6 abrange a análise dos resultados de previsão para k-passos à frente, destacando os modelos com melhor desempenho conforme as métricas de aderência escolhidas. Finalmente, as Seções 7 e 8 concluem o texto, apresentando as impressões finais sobre os resultados obtidos, considerações finais e sugestões para trabalhos futuros.

2 Revisão da literatura

A previsão de demanda no varejo está intimamente ligada ao planejamento futuro dos varejistas e seu desempenho organizacional, com foco na melhoria dos processos associados à cadeia de suprimentos dos setor (FILDES; MA; KOLASSA, 2022). Alinhado com isto, outros estudos apontam que varejistas com um operacional de alto volume e margens baixas - o que inclui o cenário *e-commerce* - são fortemente beneficiados em termos de lucratividade, dada a aplicação desses métodos preditivos (FISHER; RAMAN, 2018) apud (FILDES; MA; KOLASSA, 2022).

Ademais, observa-se que grandes esforços tem sido direcionados ao melhoramento dos modelos de previsão nas últimas décadas. Vê-se que isto está intrinsecamente correlacionado com a transformação que o setor varejista tem passado no modo de estabelecer suas estratégias de tomada de decisão. De fato, tem-se transitado de uma perspectiva mais ligada à intuição para uma outra fundamentada em modelos estatísticos *data driven* (FISHER; RAMAN, 2018) apud (FILDES; MA; KOLASSA, 2022).

Afim de explorar o valor preditivo desses dados e gerar previsões de acurácia aceitável, é essencial que os modelos incorporem em sua formulação os fatores que mais afetam as vendas do varejista. Dentre esses fatores é possível listar alguns, como: ações promocionais de competidores, promoções feitas pelo próprio varejista, feriados com alto potencial de consumo (Natal, Páscoa, Dia das Mães etc.) e até eventos sazonais como a volta às aulas, vinculado à compra de materiais escolares (GEURTS; KELLY, 1986).

Entretanto, ressalta-se que identificar tais fatores ou encontrar formas de mensurá-los pode muitas vezes se tornar uma tarefa árdua no processo de modelagem da previsão de vendas. Exemplos de fatores são aqueles relacionados às preferências subjetivas dos consumidores por um determinado *stock keeping unit* (SKU) em detrimento de outro, sendo estes apenas variações do mesmo produto - como é o caso de toalhas de banho do mesmo modelo, porém de cores diferentes.

Assim, tratando-se de forma mais aprofundada dos modelos estatísticos adotados na previsão de vendas do varejo, enuncia-se inicialmente os modelos de regressão dinâmica. Estes modelos apresentam forte relevância, pois indo além dos métodos extrapolativos univariados, como o *exponential smoothing*, incluem múltiplas variáveis explicativas e apresentam a vantagem prática de serem facilmente interpretados e implementados (FILDES; MA; KOLASSA, 2022).

Num estudo feito a partir das vendas diárias de uma loja varejista de Portugal - contendo registros de janeiro de 2012 até abril de 2015, de 15 categorias de produtos e um total de 100 SKUs -, buscou-se comparar a capacidade preditiva de tais modelos frente ao modelo *autoregressive integrated moving average* (ARIMA), comumente usado na previsão de séries temporais (PINHO; OLIVEIRA; RAMOS, 2016). Comparando diversas variações de ambos os modelos, incluindo versões logarítmicas e outras com a adição de termos seno e cosseno - a fim de modelar a sazonalidade das séries de vendas -, concluiu-se que os modelos de regressão

dinâmica apresentavam melhor acurácia sob as métricas de erro utilizadas. Além disso, destaca-se o melhor desempenho obtido sobre períodos promocionais de venda, normalmente mais difíceis de serem previstos com boa precisão, considerando as mudanças abruptas ocasionadas no perfil da série histórica.

Já no caso da previsão de vendas de SKUs cujos perfis de demanda apresentam alta intermitência - períodos de contagem de vendas nula, intercalados com outros de contagem positiva -, se faz necessário modelagens capazes de prever tais séries temporais irregulares. Nesses casos, tem-se uma aplicação interessante dos modelos estatísticos denominados *score-driven models*, capazes de avaliar diferentes funções de distribuição probabilística para dados de contagem na modelagem de históricos de vendas intermitentes.

Verifica-se sua adequação sob esta ótica em estudos como o artigo publicado por (SARLO; FERNANDES; BORENSTEIN, 2023), que evidencia a boa performance preditiva de tais modelos frente a de outros métodos tradicionais de previsão intermitente, ao realizar análises sobre dados reais, obtidos de uma rede varejista brasileira.

Em outro estudo, publicado por (SINGH; BOOMA; EAGANATHAN, 2020), fez-se primeiramente uma extensa revisão da literatura associada à previsão de vendas no *e-commerce*, focada em diferentes métodos de *machine learning*. Dentre os métodos presentes nos trabalhos mencionados, cabe citar os de *convolutional neural network* (CNN), *recurrent neural network* (RNN), *random forest* e *gradient boosting*.

Ao longo dessa seção de revisão descreve-se a boa capacidade preditiva desses métodos, frente à outros métodos mais tradicionais na previsão de séries temporais. Ademais, também é descrito o comparativo feito em cada estudo citado entre os próprios métodos de *machine learning*.

Numa fase seguinte, os autores selecionaram os métodos *random forest* e *gradient boosting* para competir com os modelos estatísticos *autoregressive integrated moving average* (ARIMA) e *seasonal autoregressive integrated moving average* (SARIMA) na previsão de vendas *e-commerce*. Para tanto, foi utilizado o *Brazilian E-Commerce Public Dataset*, disponibilizado na plataforma Kaggle pela empresa Olist Store, contendo um histórico de cem mil pedidos de compra (OLIST; SIONEK, 2018).

Por fim, chegou-se à conclusão que ambos os métodos de *machine learning* escolhidos superavam os modelos ARIMA e SARIMA e que, dentre as quatro abordagens empregadas, observou-se o melhor desempenho no *random forest*, apresentando uma acurácia de 87,39% na fase de testes e os menores valores nas métricas de erro adotadas (SINGH; BOOMA; EAGANATHAN, 2020).

Já no artigo publicado por (BANDARA et al., 2019) utiliza-se um modelo do tipo *long short-term memory network* (LSTM) para prever as vendas no *e-commerce* da grande varejista mundial, Walmart. Sabendo-se que normalmente as políticas de estoque incluem produtos cujos padrões de vendas podem ser correlacionados,

implementou-se o modelo de modo a incorporar informações dessas múltiplas séries de vendas.

Tal processo ocorre explorando possíveis relações não-lineares entre as mesmas, a fim de realizar a previsão individual de um determinado produto. Como resultado final, o estudo apresenta uma acurácia preditiva superior por parte do modelo LSTM, quando comparada com a acurácia de outras técnicas consideradas estado da arte.

Por fim, vê-se que o emprego de modelagens estatísticas e métodos de *machine learning* podem fornecer *insights* valiosos para o setor varejista, ajudando os agentes que o compõem a ter uma melhor visão do comportamento dos consumidores e a tomarem decisões mais assertivas nos seus processos internos. Inclui-se nisso um melhor planejamento de estoques, ações de *marketing*, encartes promocionais, dentre outros elementos favorecidos pela maior previsibilidade de vendas futuras.

3 Modelos e métodos preditivos

3.1 Regressão dinâmica

Considere um modelo de regressão dinâmica geral representado pela equação:

$$y_t = \beta' x_t + \varepsilon_t \tag{1}$$

onde:

y_t é a variável dependente no período t;

x_t é um vetor de variáveis independentes no período t, de dimensão $p \times 1$;

β' é um vetor de coeficientes, de dimensão $1 \times p$;

ε_t é o termo de erro no período t, tal que $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.

A equação de regressão dinâmica, indicada em (1), oferece três interpretações distintas:

1. **Regressão**: a primeira interpretação é de uma equação de regressão, onde y_t é modelado como a expectativa condicional de y_t dado x_t, expressa como $y_t = E[y_t|x_t] + \varepsilon_t$. Nesta abordagem, a função de regressão $E[y_t|x_t] = g(x_t)$ é presumida, e a equação resultante é linear apenas se a distribuição conjunta de y_t e x_t pertencer a uma classe específica de distribuições.

2. **Relação estrutural**: a segunda interpretação é baseada em uma relação estrutural subjacente. Vê-se a equação (1) como uma representação da verdadeira relação entre y_t e x_t. Esta interpretação implica que os parâmetros da equação têm significados subjetivos de acordo com o contexto da aplicação, indo além das relações puramente estatísticas.

3. **Equação de projeção dinâmica**: a terceira interpretação é que (1) representa uma projeção dinâmica. Isto significa que y_t é projetado em termos de x_t, proporcionando uma perspectiva de previsão dinâmica com base nas informações disponíveis até o instante de tempo t.

Em diferentes conjuntos de circunstâncias, essas interpretações podem coincidir ou diferir, de modo que destaca-se a importância de escolher uma interpretação adequada para testar hipóteses e só então realizar previsões. A análise começa com a definição desses tipos de modelo, a compreensão de suas propriedades e a identificação dos cenários em que eles diferem, lembrando que a distribuição conjunta de y_t e x_t desempenha um papel crucial, especialmente na consideração da exogeneidade de x_t para β. Para maior aprofundamento em regressão dinâmica é sugerida a leitura do livro Dynamic Econometrics de David Hendry, capítulos 4,5 e 6 (HENDRY, 1995).

3.2 Regressão Lasso

Como alternativa aos modelo de regressão dinâmica, cujos parâmetros são estimados utilizando o método de mínimos quadrados ordinários (MQO), encontra-se na literatura os chamados *shrinkage methods*. Como o nome em inglês sugere, esses métodos buscam reduzir a complexidade do modelo reduzindo o número de variáveis explicativas ou reduzindo o efeito daquelas com baixo poder explicativo sobre a variável dependente (WITTEN; JAMES, 2013).

Dentre os métodos de contração de modelo de regressão tem-se o *least absolute shrinkage and selection operator* (Lasso), que por construção é capaz não só de encolher alguns coeficientes estimados, mas também torná-los exatamente iguais a zero (TIBSHIRANI, 1996). Isso se mostra uma vantagem frente a outros métodos como a regressão Ridge, cuja formulação propicia apenas uma nulificação assintótica desses coeficientes.

Dessa forma, sob um incremento aceitável de viés, reduz-se a variância das previsões do modelo e obtém-se uma melhor interpretação empregando menos variáveis explicativas. As estimativas dos coeficientes do Lasso são encontradas resolvendo o problema de programação quadrática com restrições de desigualdades lineares indicadas na equação 2, como indicando no trabalho inicial de (TIBSHIRANI, 1996):

$$(\hat{\beta_0}, \hat{\beta_j}) = \underset{\beta}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \lambda \geq 0 \qquad (2)$$

Na prática, faz-se a minimização da soma dos quadrados dos resíduos (SQR) considerando um fator de penalização, λ, que atua sobre os coeficientes de cada variável explicativa do modelo. Se λ assume valor nulo, então retornamos às estimativas por MQO. Caso contrário, se o fator de penalização cresce indeterminadamente, o emprego de qualquer variável explicativa resulta num valor infinito da função objetivo, inviabilizando sua minimização.

Portanto, é preciso encontrar o valor ótimo de λ entre esses dois extremos, de modo que o modelo ajustado mantenha um subconjunto das variáveis explicativas originais que minimize SQR, mas cujo alto poder explicativo sobre a variável dependente compense a penalização dos seus coeficientes (i.e., terem estimativas não nulas).

Verifica-se na literatura algumas práticas comuns para realizar a busca pela otimilidade do fator de penalização, como o *cross-validation* (CV) (TIBSHIRANI, 1996) ou o uso de um critério de informação (ZOU; HASTIE; TIBSHIRANI, 2007), comparando os modelos ajustados para cada λ sob análise. Tratando-se da previsão de séries temporais cujos dados infringem certas condições do CV - como a premissa de que as observações são independentes e identicamente distribuídas (i.i.d), considera-se neste trabalho a definição do λ ótimo via critério de informação.

Em seções posteriores vinculadas à metodologia descreve-se o emprego dos critérios *Akaike Information Criterion* (AIC) e *Bayesian Information Criterion* (BIC).

3.3 Regressão AdaLasso

Apesar da regressão Lasso evidenciar em muitos casos *oracle properties* (FAN; LI, 2001), estudos posteriores mostram que isto é válido apenas sob uma condição necessária não trivial e portanto em alguns cenários não é possível obter conjuntamente estimadores ótimos e uma seleção de variáveis consistente (ZOU, 2006).

A partir disso, propõe-se uma uma versão modificada do Lasso, que pondera as penalizações impostas aos coeficientes que devem ser estimados. Isso se dá em um algoritmo de dois estágios, tal que o primeiro estima pesos w_j para cada coeficiente β_j, e o segundo estima um Lasso que multiplica λ ao penalizar os coeficientes. Por conta desta última característica atribui-se a esta versão a nomenclatura *adaptive Lasso* (AdaLasso), capaz de avaliar de forma particular o custo de manter cada variável explicativa no modelo (ZOU, 2006).

É possível encontrar diversas metodologias na literatura para o cálculo dos pesos w_j que na prática são estimados como $\hat{w}_j = 1/|\hat{\beta}_j^*|^\gamma$, sendo β_j^* um estimador $\sqrt{n}$-consistente obtido a partir do primeiro estágio deste modelo. No artigo que apresenta o AdaLasso pela primeira vez, sugere-se o uso dos coeficientes estimados mínimos quadrados ordinários, $\beta(\hat{ols})$.

Contudo, o próprio autor considera a possibilidade de outros estimadores consistentes, como $\beta(\hat{Ridge})$ e estudos mais recentes consideram inclusive uma regressão Lasso no primeiro estágio. Neste caso faz-se necessário adicionar um *offset* não nulo aos $\beta(\hat{Lasso})$ para evitar divisões por zero (GARCIA; MEDEIROS; VASCONCELOS, 2017).

Por fim, obtém-se os coeficientes estimados do modelo de regressão AdaLasso resolvendo o problema de otimização indicada na equação 3, cujo custo computacional é muito semelhante ao requerido no Lasso.

$$(\hat{\beta}_0, \hat{\beta}_j) = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\}, \lambda \geq 0 \qquad (3)$$

3.4 *Random forest*

Este método de *machine learning* foi proposto inicialmente por (BREIMANN, 2001) e se fundamenta na técnica de *bootstrap aggregation (bagging)* de múltiplas árvores de decisão, sendo estas destinadas a solucionar problemas classificação ou de regressão. Nestes últimos casos comumente adota-se como previsão a média dos valores previstos por cada uma dessas árvores internas, visando uma política de consenso entre estas e promover a redução da variância dos valores previstos.

Tendo descrito brevemente o funcionamento deste método, busca-se a seguir elucidar em mais detalhes os conceitos teóricos que o fundamentam.

Árvores de regressão

As árvores de regressão são um subtipo das árvores de decisão destinadas à previsão de variáveis dependentes de escala numérica. O intuito deste método é particionar o espaço dimensional criado pelas variáveis explicativas, X_p com $p \in \{1, 2, ..., P\}$, em regiões distintas e sem sobreposição, R_j com $j \in \{1, 2, ..., J\}$. Com isso, ajusta-se um mapeamento entre tais preditores e a variável dependente, Y, incluindo eventuais relações não lineares.

A construção dessa estrutura ocorre por meio de uma bipartição recursiva do espaço preditivo, gerando os nós e respectivas ramificações da árvore de decisão. Essa estratégia visa minimizar localmente uma métrica associada à impureza dos nós. No caso das árvores de regressão este valor é vinculado à estimativa da variância da variável dependente, utilizando comumente a soma dos quadrados dos resíduos (RSS).

Portanto, para cada região já criada, busca-se sucessivamente o par ótimo (X_p^*, s^*), sendo s um ponto de corte, capaz de minimizar o RSS da árvore ao delimitar duas novas regiões: uma onde $X_p^* \geq s^*$, outra em que $X_p^* < s^*$ (WITTEN; JAMES, 2013). Essa recursão persiste até que um dos critérios de parada sejam atendidos, por exemplo, que todas as regiões R_j ou nós terminais da árvore contenham uma quantidade de amostras de treino menor ou igual a um limiar a ser definido.

No entanto, a partição recursiva pode implicar numa estrutura muito complexa que performa bem no treino, mas apresenta uma variância elevada nas previsões *out-of-sample* com perda de generalização. Para evitar isto é sugerido na literatura a técnica *cost-complexity pruning* (BREIMAN et al., 1984), responsável por podar a árvore numa estrutura que balanceia a adequação às amostras de treino com o custo de particionar regiões em excesso.

Concluído o processo de construção da árvore, novas amostras que se enquadrem numa dada região R^* terão como previsão a média dos valores observados de Y, correspondentes às amostras na fase de treino e responsáveis pela delimitação de R^*.

Ilustra-se na Figura 1 a partição do espaço preditivo composto pelas variáveis explicativas X_1 e X_2 nas regiões R_j com $j \in \{1, 2, ..., 5\}$ (centro), obtidas a partir da construção da árvore de regressão (esquerda). Além disso, tem-se a visualização das previsões associadas a cada região nas alturas dos semiplanos correspondentes (direita).

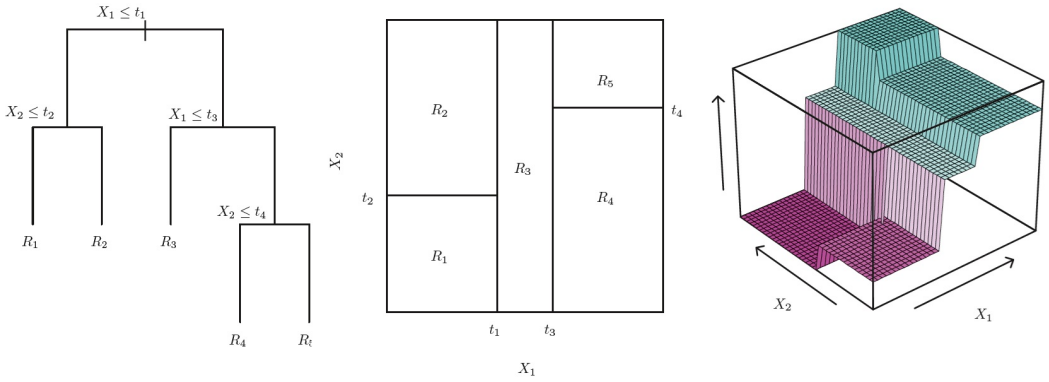

Figura 1: Visualização das regiões no espaço preditivo e previsões de uma árvore de regressão. Reprodução de parte da Figura 8.3 em (WITTEN; JAMES, 2013), p. 308.

Bootstrap aggregation (bagging)

Proposta no trabalho (BREIMAN, 1996), esta técnica de *ensemble* visa melhorar a estabilidade e a robustez dos modelos de aprendizado de máquina. No contexto de *tree bagging* a ideia é treinar múltiplas árvores de decisão independentes em subconjuntos aleatórios dos dados e em seguida combinar as previsões destas.

- **Amostragem Bootstrap:** a base do *bagging* é a amostragem *bootstrap*, que envolve a geração de múltiplas amostras de dados com reposição. Cada árvore é treinada em uma dessas amostras, o que introduz variação nos conjuntos de treinamento.
- **Independência:** as árvores são treinadas de forma independente umas das outras, o que reduz a correlação entre os modelos e aumenta a estabilidade do ensemble.
- **Combinação:** as previsões individuais das árvores são combinadas, muitas vezes por média (no caso de regressão) ou votação (no caso de classificação).

Pressupostos de IID (Independentes e Identicamente Distribuídos)

A técnica denominada *bagging* assume que os dados seguem o princípio de IID, ou seja, são independentes e identicamente distribuídos. Isso significa que cada ponto amostral é uma observação independente de uma distribuição de probabilidade com a mesma função de densidade. A amostragem com reposição durante o processo de *bootstrap* ajuda a aproximar essa condição.

Por conta disto, tal técnica tem sido extensivamente adotada em vários métodos de aprendizado de máquina, resultando em melhorias notáveis na redução de variância e no reforço da robustez dos modelos. Alguns exemplos de trabalhos que implementaram essa abordagem incluem (SIQUEIRA, 2023), (BRANDÃO et al., 2021) e (SANTOS; ROSSI, 2020).

Block bootstrap

Apesar das suas diversas aplicações a técnica de *bootstrapping* padrão não é compatível com a dependência temporal de uma série e portanto, ao fazer múltiplas amostragens desconstruiria esse padrão.

Portanto, trabalhos como (LIU; SINGH et al., 1992) e (POLITIS; ROMANO, 1991) buscam novas abordagens de empregar a técnica de amostragem em blocos. Dessa forma, é possível criar novas séries temporais por amostragem a partir da série temporal original, mantendo a estrutura de dependência temporal intacta intra-blocos.

O emprego de tais técnicas tem se mostrado benéfico em trabalhos recentes, como (MEDEIROS et al., 2021), cujo emprego do *block bootstrapping* em conjunto ao algoritmo de *random forest* demonstra superioridade de performance preditivà frente a outros modelos.

Cita-se ainda o trabalho publicado por (GOEHRY et al., 2021) que implementa diferentes variações desta técnica no pacote *rangerts* da linguagem de programação R. Nesta publicação demonstra-se resultados interessantes e um ganho de acurácia na predição de séries temporais se comparado ao *bootstraping* padrão do *random forest*.

Ademais, neste estudo explora-se uma base de dados de frequência horária e com uma sazonalidade diária expressiva, servindo de inspiração para o emprego destas técnicas na previsão das séries horárias de quantidade vendida dos SKU's.

4 Dados utilizados

4.1 Tratamento das bases de dados

Neste trabalho explorou-se três bases de dados disponibilizadas, denominadas *visits*, *sales*, *competitors*. Todas contém informações relevantes a cinco produtos distintos, identificados por seus *stock keeping units* (SKU).

A partir destas bases foram feitas as seguintes análises e tratamentos dos dados, a fim de unificá-las em registros horários para futuros treinamentos e testes dos modelos e métodos preditivos já enunciados.

4.1.1 Mascaramento de dados sensíveis

Desde o início do trabalho houve uma preocupação em proteger dados sensíveis presentes nas bases de dados que eventualmente pudessem identificar a empresa varejista detentora dos mesmos, assim como seus concorrentes e os produtos comercializados.

Para tanto, aplicou-se um mascaramento nas colunas que referenciavam as empresas do setor. No caso das séries de preço referentes aos concorrentes, fez-se a renomeação das respectivas colunas na base de dados para *price_compt_j*, sendo j um inteiro indicando o índice do concorrente. Para a série horária do preço próprio utilizou-se apenas a nomenclatura *price*.

De forma semelhante, todos os códigos que identificavam os SKU's foram mascarados e passaram a se chamar *prod_k*, sendo $k \in \{1, 2, ..., 5\}$, com objetivo de representar os cinco produtos distintos presentes nas bases de dados.

Além disso, fez-se o mascaramento dos valores das séries de preço de todos os SKU's para cada um dos *players* do setor considerados. Isso se deu por meio de uma transformação linear sobre as séries, de maneira que produtos com escalas de preços originalmente diferentes mantivessem essa relação mesmo estando mascarados.

4.1.2 Obtenção da frequência horária

Dentre os tratamentos de dados realizados, ressalta-se que a fim de estudar a quantidade vendida de cada *stock keeping unit*, hora a hora, foi necessário agrupar os registros de venda da base de dados *sales* numa mesma hora fechada.

Define-se hora fechada como aquela que agrega informações de todos registros pertencentes ao intervalo [hh:00:00, hh:59:59] (hora-minuto-segundo), tendo fixado uma data no formato yyyy-mm-dd (ano-mês-dia). Atribuiu-se então os valores obtidos na coluna *date_hour* no formato yyyy-mm-dd hh:00:00, com hh indo de 0 até 23 horas.

Com isso, ao agregar os múltiplos registros originais da base de dados *sales*, somou-se as quantidades vendidas dos registros vinculados a uma mesma em hora fechada, tendo fixado um dos SKU's.

Quanto à base de dados *competitors*, também verificou-se diversos registros dos preços próprio e dos concorrentes para um dado intervalo horário. A fim de consolidar esses valores numa frequência horária e associá-los às séries horárias de vendas e visitas, fez-se um tratamento análogo, agregando os preços sob uma mesma hora fechada e obtendo sua média aritmética.

Com essa operação, foi criada a coluna *price* para designar a série horária do preço médio próprio, além das colunas *price_compt_j*, tal que $j \in \{1, 2, ..., 6\}$, indexando as séries horárias de preço médio dos concorrentes.

Por fim, destaca-se que os registros da base de dados *visits* já se encontravam no formato de hora fechada. Portanto, não foi necessário fazer nenhum tipo de agregação dos seus registros e posterior operação de soma do número de visitas.

4.1.3 Tratamento dos valores faltantes

Após realizar a operação de união das três bases de dados já tratadas individualmente, fez-se uma operação para tratar possíveis registros faltantes de cada SKU para uma data e hora qualquer.

Considerando os valores mais antigo e mais recente na coluna *date_hour*, completou-se a base de dados com os eventuais registros referentes às datas-horas faltantes, para cada um dos cinco SKU's. Destaca-se que cada registro foi adicionado apresentando valores faltantes, NaN, para as colunas diferentes de *date_hour* e *product_id* (guarda os diferentes valores dos SKU's já mascarados).

Ademais, notou-se que além dos registros (linhas da base de dados unificada) as colunas também apresentavam valores faltantes. Cita-se como exemplo registros com valores válidos para o número de visitas, mas com valores faltantes (NaN) para a quantidade vendida.

Infere-se que a nível de frequência horária, apesar de haver visitas registradas na base *visits* para uma certa data-hora, não havia vendas associadas a esta na base *sales*. Com a operação de união de ambas as bases de dados, pareadas pelos valores de data-hora e SKU, gerou-se indiretamente tais valores faltantes.

Desse modo, tratou-se os valores faltantes de quantidade vendida atribuindo-lhes valor nulo. Seguindo um raciocínio análogo, atribuiu-se valor nulo aos valores faltantes na série horária de visitas.

Já no caso de valores faltantes nas séries de preços assumiu-se uma falha na metodologia adotada na sua obtenção. Identifica-se na base de dados a ausência de valores válidos ao longo de intervalos horários completos de determinadas datas.

Portanto, com o objetivo de eliminar tais descontinuidades nas séries de preço, sob a lógica de que nenhum *player* do mercado trabalharia subitamente sem preços associados aos produtos comercializados, tratou-se os valores faltantes repetindo o último valor válido da série de preços em questão; no caso do primeiro valor da série já ser NaN, foi considerada a média da série anterior ao tratamento.

4.1.4 Remoção e agrupamento de concorrentes

Comparando todas as séries de preço percebeu-se que dentre os cinco SKU's presentes na base de dados, as séries denominadas por *price_compt_5* e *price_compt_6* apresentavam valores válidos apenas nos registros referentes ao SKU identificado por *prod_3*.

Ainda sob a análise deste SKU, observou-se apenas 313 registros diferentes de NaN na série *price_compt_5*. No caso da série *price_compt_6*, um número ainda menor de 31 registros. Dessa forma, decidiu-se pela exclusão de suas respectivas colunas da base de dados, considerando sua fraca consistência e também para tornar mais equilibrada a comparação entre os preços dos concorrentes remanescentes.

Dando continuidade às análises das séries horárias de preço dos concorrentes, percebeu-se ao longo da janela temporal observada que muitas vezes *price_compt_j*, com $j \in \{2, 3, 4\}$, se sobrepunham independentemente do SKU. Feito o cálculo posterior da matriz de correlação de todas as séries de preço, verificou-se então uma alta correlação entre essas três séries, evidenciando comportamentos muito semelhantes nas políticas de preços adotadas por estes concorrentes.

Devido a essas observações, decidiu-se adotar uma série de preço criada artificialmente a partir da média aritmética das séries horárias de preço desses três concorrentes, renomeando-a para *price_compt_2*.

Assim, ao final desses tratamentos sequenciais, permaneceram na base de dados unificada e tratada as séries de preço denominadas pelas colunas por *price*, *price_compt_1* e *price_compt_2*.

4.1.5 Base de dados tratada

Por fim, feitos os tratamentos descritos obteve-se uma base de dados de frequência horária, sem valores faltantes, contendo um total de 14640 registros. Posteriormente tomou-se a decisão de fragmentar esta base em cinco bases, uma para cada SKU, contendo efetivamente 2928 registros correspondentes às horas fechadas entre as datas 01/06/2022 até 30/09/2022.

Destaca-se que cada registro da base de dados tratada contém informações horárias da quantidade vendida de um certo SKU (*qty*), o número de visitas às plataformas *e-commerce* da empresa (*visit*), o preço médio próprio (*price*), além dos preços médios de cada concorrente remanescente (*price_compt_1*, *price_compt_2*).

4.2 Análises descritivas dos dados

Concluído os tratamento dos dados descritos anteriormente, decidiu-se explorar os dados contidos na base de dados resultante.

Dessa forma, são mostradas nessa subseção todas as análises feitas, incluindo a visualização das séries horárias de quantidade vendida e levantamento de hipóteses sobre a mesma; análise de *box plots* horários e de dias da semana;

estudo das funções de autocorrelação (FAC) e autocorrelação parcial (FACP) das séries horárias de quantidade vendida.

Ressalta-se que todas as análises foram feitas para cada um dos cinco *stock keeping units* disponíveis, porém para efeitos de exemplificação são mostrados apenas os resultados obtidos para o SKU identificado por *prod_4*.

4.2.1 Visualização da série horária de quantidade vendida

A partir da visualização da série horária de quantidade vendida mostrada na Figura 2 verifica-se um decaimento na quantidade de vendas do produto ao longo da janela temporal dos dados disponíveis, fenômeno que também ocorre para os demais produtos presentes na base, apesar de não mostrados aqui.

Nota-se também alguns picos na série temporal, mesmo nos períodos em que a média dos valores aparenta ser maior. Um exemplo disso são os picos percebidos próximos ao dia 12 de junho de 2022, Dia dos Namorados. Com isso, infere-se que outras datas comemorativas possam ter um efeito similar sobre as séries de vendas, gerando uma alteração pontual nos padrões de consumo, com a tendência por parte dos consumidores de presentear alguém querido.

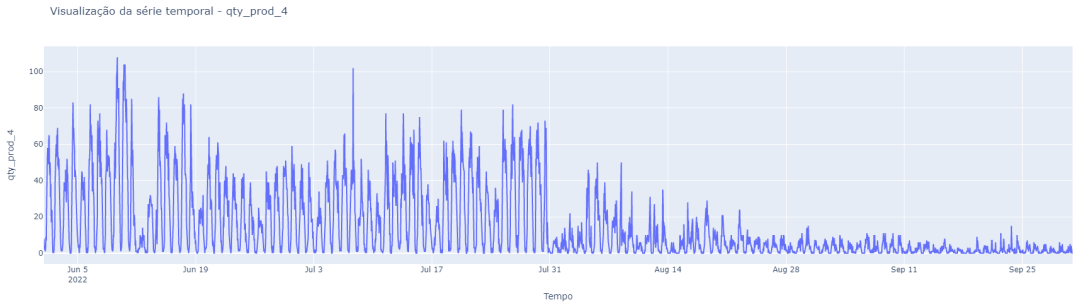

Figura 2: Série horária da quantidade vendida do SKU denominado prod_4.

4.2.2 *Box plots* da quantidade vendida

Analisando os *box plots* dos SKU's notou-se os padrões aqui exemplificados, tal que ao longo da semana observa-se que a mediana da quantidade vendida costuma ser ligeiramente ser maior nas quartas e quintas-feiras (ver Figura 3), e para alguns produtos verifica-se ainda um aumento médio nas segundas-feiras - diferentemente do *prod_4*. Isso foi interessante, pois a priori esperava-se que tais aumentos na mediana da quantidade vendida fosse percebido nos dias que compõem o final de semana.

Já na investigação dos resultados obtidos com os *box plots* da quantidade vendida sob as horas do dia, exemplificado na Figura 4, evidencia-se um claro efeito de sazonalidade diária. Vê-se que a mediana das vendas inicia mais baixa nas primeiras horas do dia que constituem a madrugada (coerente com a rotina de sono do consumidor comum), e que com o passar do dia, a mediana das horas fechadas

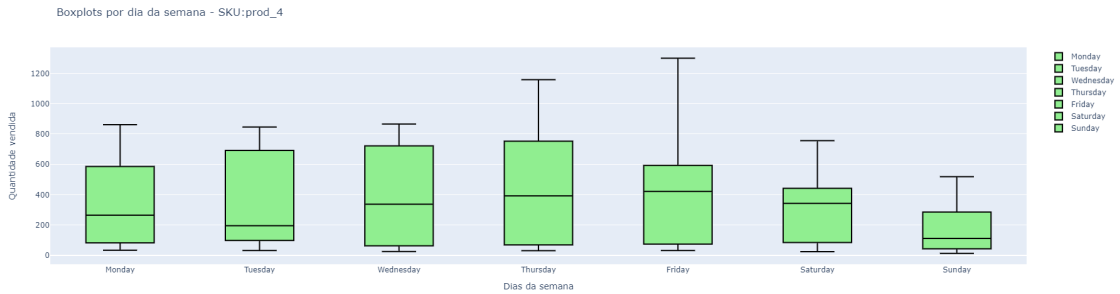

Figura 3: Box plot da quantidade vendida para os dias da semana.

seguintes aumenta. Isso se dá consistentemente até atingir uma estabilização entre as horas 12 e 15, tal que da hora 16 em diante a mediana começa a decrescer, indicando o recomeço do comportamento diário dos consumidores.

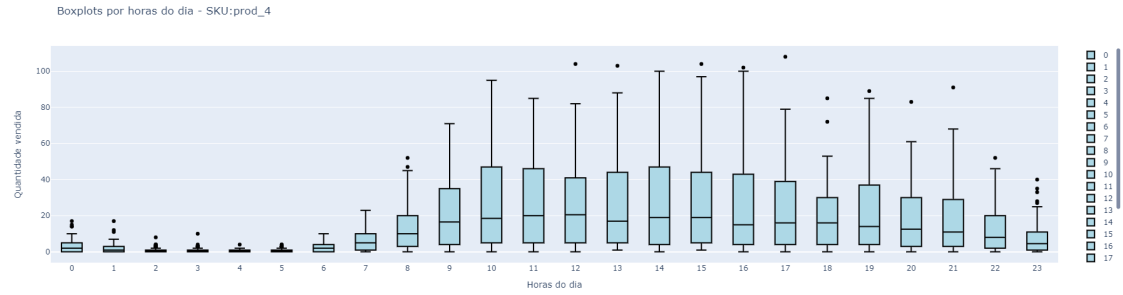

Figura 4: Box plot da quantidade vendida para as horas do dia.

4.2.3 Análise da FAC/FACP

No estudo das funções de autocorrelação (FAC) e de autocorrelação parcial (FACP) confirmou-se a característica de sazonalidade diária evidenciada pelo *box plot* horário, de modo que no gráfico da FAC, na Figura 5, nota-se uma alta correlação conjunta dos diversos *lags* (horas anteriores ao instante de tempo atual, t) sobre o instante de tempo t, mesmo os *lags* de 24 e até 48 horas passadas.

Já no caso da FACP identifica-se que a correlação isolada do instante de tempo atual com a primeira e da segunda horas passadas são as de maior magnitude, apresentando fortes indícios de que esses *lags* são relevantes na modelagem da quantidade quantidade vendida uma hora à frente.

4.3 Ajustes finais das bases de dados

A partir das bases de dados tratadas obtidas por SKU e das análises descritivas feitas decidiu-se fazer alguns ajustes para prosseguir com o uso destes dados no treinamento dos modelos e métodos preditivos.

Da seção de apêndice A é possível ver que as séries horárias de quantidade vendida dos SKU apresentam escalas distintas de variabilidade entre si,

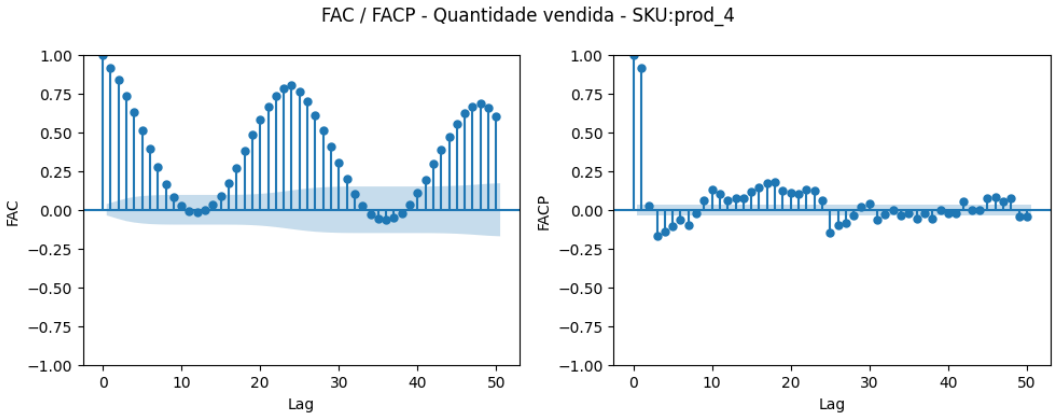

Figura 5: Visualização da FAC e FACP da quantidade vendida.

aplicou-se a transformação logarítmica sobre as mesmas com o objetivo de nivelar essas diferenças. Sob a mesma ótica, entendeu-se como necessário aplicar tal transformação nas séries de preços médios próprio e dos concorrentes, além das séries de visitas.

A partir desse ponto, considera-se para o treinamento dos modelos estatísticos e métodos de *machine learning* a variável dependente $log(q_t)$, denominada *log_qty* nas bases de dados ajustadas. O mesmo vale para as variáveis explicativas, que na base ajustada recebem o prefixo *log_* referindo-se a $log(v_t)$, $log(p_t)$, $log(pc_{1,t})$ e $log(pc_{2,t})$ - séries logarítmicas de visitas e preços médios próprio e dos concorrentes.

Um ponto importante a ser ressaltado é que empiricamente verifica-se contagens nulas nas séries horárias de quantidade vendida (*qty*) e do número de visitas às plataformas *e-commerce* (*visit*). Portanto, adiciona-se um *offset* de uma unidade a estas séries para que suas transformações logarítmicas não resultem em menos infinito nestes casos, mas sim em zero.

A fim de empregar nas abordagens preditivas a sazonalidade diária e semanal, exploradas anteriormente nas Figuras 4 e 3, acrescentou-se às bases de dados variáveis *dummy*. Aquelas correspondentes às horas fechadas do dia são denominadas pelas colunas *hour_h*, $h \in \{1, 2, ..., 23\}$. Já as colunas *weekday_w*, $w \in \{1, 2, ..., 6\}$, se referem aos dias da semana.

Observa-se que apesar da sazonalidade diária incluir 24 horas fechadas, considera-se apenas 23 destas no emprego das respectivas *dummies*. Se houvesse uma vigésima-quarta *dummy* para indicar a hora fechada haveria uma redundância de informação, uma vez que quando todas as demais *dummies* dessa categoria se anulam, implicitamente indica-se a ocorrência da vigésima-quarta hora fechada.

Além disso, a inclusão desta impossibilitaria o cálculo dos coeficientes estimados via método dos mínimos quadrados ordinários (MQO), durante o treinamento dos modelos de regressão dinâmica. A causa disto está na não invertibilidade de uma das matrizes do método, por conta da multicolinearidade vinculada à vigésima-quarta *dummy* de hora fechada em excesso.

O mesmo raciocínio se aplica às *dummies weekday_w* e por isso acrescenta-se às bases de dados tratadas dos SKU's apenas 6 colunas referentes a estas variáveis, apesar da sazonalidade semanal compreender um período de tamanho 7.

Além disso, no intuito de explorar o caráter auto regressivo das séries horárias de quantidade vendida dos SKU's, acrescentou-se colunas referentes aos *lags* da variável dependente, *qty*. Também foram adicionados *lags* da série horária de visitas à plataforma de vendas *e-commerce* da varejista e das séries de preço médio próprio e dos concorrentes.

Destaca-se que os *lags* empregados para o realizar o acréscimo das variáveis explicativas descritas apresentam graus contidos no conjunto $\{1, 2, 3, 4, 12, 24, 48\}$. O propósito disto foi avaliar indiretamente por meio das métricas de aderência a capacidade explicativa em curto e médio prazo dos valores passados sobre os valores futuros das séries de vendas, incluindo à dinâmica das variáveis exógenas delimitadas pelas séries de visita e preços.

Com o acréscimo das colunas de *lags*, cujo maior grau resulta em 48 horas fechadas passadas, tem-se por construção uma perda de mesma ordem de registros da base de dados, reduzindo de 2928 para 2880 registros em cada base por SKU. A partir disso, considerou-se as previsões para $k = 1$, reservando para o período de testes três semanas de dados horários, ou 504 pontos amostrais um passo a frente, restando 2376 registros para realizar o treinamento dos modelos e métodos preditivos.

Deste último ajuste, considerou-se a janela temporal compreendida pelo registros de treino e então fez-se a adição de *dummies* associadas às datas comemorativas compreendidas nesse intervalo. Para os feriados de corpus christi (16/06/2022) e dia da independencia do brasil (07/09/2022), empregou-se apenas uma *dummy*, gerando as colunas *corpus_christi* e *independence_day*.

Já no caso do dia dos namorados(12/06/2023) e dia dos pais (14/08/2022), acrescentou-se *dummies* para as datas em si e para 4 dias anteriores e 2 dias subsequentes. Com isso, gerou-se as colunas de radical *valentines_day* e *fathers_day*.

A razão disto foi mapear o perfil de alta anterior à data comemorativa, uma queda brusca ao longo da data e a normalização das vendas nos dias seguintes. Para exemplificar esse comportamento, mostra-se na Figura 6 um *zoom* dos dias mapeados por estas *dummies* para o dia dos namorados, na série horária de quantidade vendida do SKU 4.

Por fim, evidenciou-se picos de vendas na transição entre os meses de junho e julho de 2022, alguns ocorrendo até o sétimo dia de julho. Pesquisando a respeito, viu-se que certos grupos de profissionais vinculados ao serviço público brasileiro, como militares das forças armadas, costumam receber no início de julho a primeira metade do décimo terceiro salário, juntamente com o salário de junho.

Enxergando uma eventual correlação desses pagamentos com a variação do comportamento padrão de consumo dos SKU's neste intervalo, acrescentou-se *dummies* referentes ao primeiro dia do mês de julho, além dos quatro dias anteriores

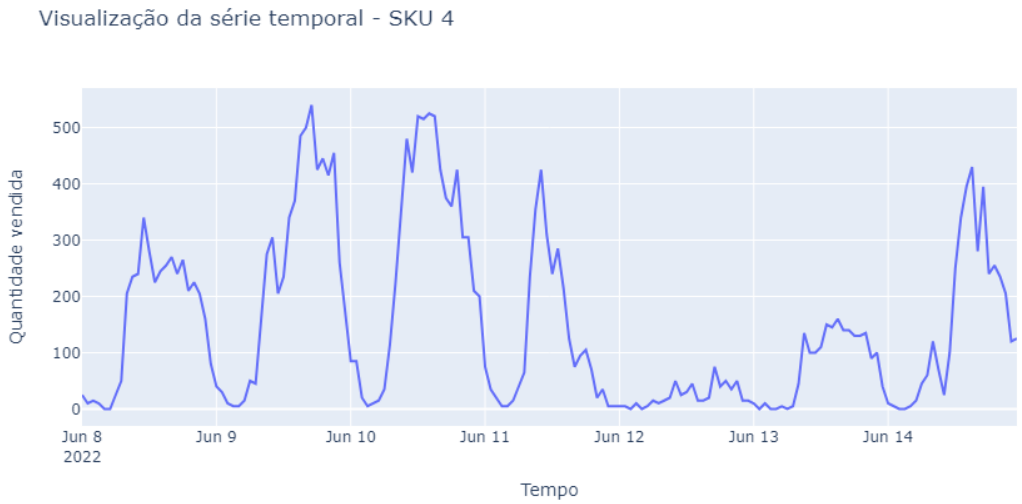

Figura 6: Zoom desde 4 dias anteriores até 2 dias seguintes ao dia dos namorados.

e sete dias seguintes. Tais variáveis são denominadas pelas colunas de radical *half_13_salary*.

Assim, considerando a frequência horária das bases de dados tratadas e posteriormente ajustadas, explicita-se que no caso das *dummies* referentes a feriados, datas comemorativas e ao pagamento da primeira metade do décimo terceiro, o valor dos registros para as respectivas colunas incidem em 1 para todas as horas fechadas destes dias.

4.4 Variáveis de interesse

Por fim, faz-se uma breve recapitulação das variáveis contidas nas versão final das bases de dados para cada SKU, após tratamentos e ajustes para treinamento dos modelos e métodos preditivos. Considerando os graus dos *lags* iguais a L, sendo $L = \{1, 2, 3, 4, 12, 24, 48\}$, e fixado um SKU, tem-se nas bases ajustadas colunas referentes a:

Variável dependente

- $log(q_t)$: transformação logarítmica da série horária de quantidade vendida, acrescida do *offset* de 1 unidade; corresponde à coluna *log_qty*.

Variáveis independentes

- $log(q_{t-j})$: *lags* da transformação logarítmica da série horária de quantidade vendida, acrescida do *offset* de 1 unidade; corresponde às colunas *log_qty_lag_j*, tal que $j \in L$.
- $log(v_{t-j})$: *lags* da transformação logarítmica da série do número de visitas à plataforma *e-commerce*, acrescida do *offset* de 1 unidade; correspondem às

colunas *log_visit_lag_j*, $j \in L$.

- $log(p_{t-j})$: *lags* da transformação logarítmica da série de preço médio próprio; corresponde às colunas *log_price_lag_j*, $j \in L$.
- $log(pc_{1,t-j})$: *lags* da transformação logarítmica da série de preço médio do concorrente 1; correspondem às colunas *log_price_compt_1_lag_j*, $j \in L$.
- $log(pc_{2,t-j})$: *lags* da transformação logarítmica da série de preço médio do concorrente 2; correspondem às colunas *log_price_compt_2_lag_j*, $j \in L$.
- $D_{h,t}$: *dummies* para expressar a sazonalidade diária; correspondem às colunas *hour_h*, $h \in \{1, 2, ..., 23\}$.
- $D_{w,t}$: *dummies* para expressar a sazonalidade semanal; correspondem às colunas *weekday_w*, $w \in \{1, 2, ..., 6\}$.
- $D_{v,t-4+m}$: *dummies* para expressar o dia dos namorados (valentine's day), $m \in \{0, 1, ..., 6\}$; correspondem às colunas *valentines_day*, *valentines_day_lag_l*, $l \in \{1, 2, 3, 4\}$, e *valentines_day_inercia_i*, $i \in \{1, 2\}$.
- $D_{f,t-4+m}$: *dummies* para expressar o dia dos namorados (father's day), $m \in \{0, 1, ..., 6\}$; correspondem às colunas *fathers_day*, *fathers_day_lag_l*, $l \in \{1, 2, 3, 4\}$, e *fathers_day_inercia_i*, $i \in \{1, 2\}$.
- $D_{cc,t}$: *dummy* para expressar o feriado de Corpus Christi; corresponde à coluna *corpus_christi*.
- $D_{idp,t}$: *dummy* para expressar o feriado do dia da independência do Brasil; corresponde à coluna *independence_day*.
- $D_{sal,t-4+i}$: *dummies* para expressar possíveis dias de pagamento da metade do décimo terceiro salário, $i \in \{0, 1, ..., 11\}$; correspondem às colunas *half_13_salary*, *half_13_salary_lag_l*, $l \in \{1, 2, 3, 4\}$, e *half_13_salary_inercia_i*, $i \in \{1, 2, ..., 7\}$.

5 Metodologia

5.1 Pós-processamento k-passos à frente

Para estimar os modelos propostos anteriormente considera-se uma partição da base de dados em observações destinadas ao treinamento e ao teste da performance preditiva. Porém, além disso é adotado neste trabalho a metodologia de previsão denominada *direct multi-step forecasting*.

Dessa forma, deve-se estimar para cada k-ésimo horizonte de previsão almejado uma versão derivada dos modelos, mantendo fixo o conjunto de variáveis explicativas disponíveis, mas ajustando a variável dependente a ser prevista. Cada

passo à frente mais adiante no futuro implica em valores faltantes da variável dependente nos registros da base de dados, das quais faziam parte anteriormente.

Assim, ressalta-se que diferentes horizontes de previsão requerem bases de dados ajustadas de tamanhos distintos. Numa tentativa de minimizar possíveis problemas quanto a ao treinamento e posterior generalização dos modelos ajustados, garantiu-se a mesma janela temporal de treino em todas as estimações feitas. Esta janela corresponde às primeiras 2376 data-horas de cada série temporal.

5.2 Padronização das variáveis

Dado que o projeto busca explorar a capacidade preditiva de alguns *shrikage methods*, como Lasso e AdaLasso, implementa-se a padronização da variável dependente e das variáveis explicativas, com exceção das *dummies* de marcação das sazonalidades diária e semanal, além de eventos especiais como feriados.

O intuito desse procedimento é nivelar eventuais discrepâncias nas escalas dos preditores e evitar que alguns sejam mais penalizados por conta da sua escala original, ao invés da sua contribuição real na explicação da variável dependente. Por outro lado, alguns métodos de *machine learning*, como o *random forest* apresentam menor sensibilidade à escalas variadas. Porém, verifica-se de modo geral a padronização aos preditores não impacta negativamente sua capacidade preditiva do método.

Dada a motivação da padronização, explica-se como foi feito tal procedimento. Havendo ambas as partições de treinamento e teste da base de dados, ajustada para o horizonte de previsão devido, aplica-se a padronização sobre as variáveis contínuas do conjunto de treinamento, guardando seus valores de média e desvio padrão amostrais. É com esses valores então que por fim faz-se a padronização sobre as variáveis explicativas correspondentes, presentes no conjunto de testes.

Procedeu-se desta forma ao invés de padronizar todas as séries contínuas antes de fragmentar as observações em treino e teste para evitar vazamento de informações do teste para a fase de treino. Além disso, a adoção deste tipo de padronização busca refletir de maneira mais próxima cenários realistas em que não têm-se acesso à observações futuras dos dados.

Por fim, destaca-se que devido a padronização da variável dependente, $log(q_t)$, após feitas as previsões de cada modelo aplica-se uma reversão da padronização sobre o valor previsto. Emprega-se neste caso os mesmos valores de média e desvio padrão amostrais obtidos da padronização desta variável no conjunto de treinamento.

5.3 Reversão da transformação logarítmica

Conforme descrito em seções anteriores, o treinamento dos modelos e métodos preditivos feitos neste trabalho consideram a transformação logarítmica da série horária da quantidade vendida, $log(q_t)$. No entanto, busca-se na prática

prever a quantidade vendida de cada SKU em sua escala original.

Para tanto, após reverter a padronização das previsões obtidas, é preciso realizar a reversão da transformação logarítmica de $\hat{log(q_t)}$, exponenciando os valores previstos. Além disso, recorda-se que antes de aplicar tal transformação foi preciso adicionar uma unidade às séries de quantidade vendida para não operar o logaritmo de contagens nulas.

No entanto, não basta exponenciar a série de valores previstos e subtraí-la uma unidade para que a reversão da transformação esteja completa. Para tanto, se faz necessário multiplicá-las ainda por um fator de variância. Assim sendo, sob normalidade dos resíduos do ajuste em fase de treinamento dos modelos e métodos preditivos, com σ_ε^2 sendo a variância destes, obtém-se que:

$$\hat{q_{t+k|t}} = \exp(\hat{log(q_{t+k})} - 1) \cdot \exp\left(\frac{\sigma_\varepsilon^2}{2}\right) \tag{4}$$

No caso de não normalidade do resíduos $\hat{\varepsilon_t}$, pode-se mostrar que (WOOLDRIDGE, 2015):

$$\hat{q_{t+k|t}} = \exp(\hat{log(q_{t+k})} - 1) \cdot \hat{\alpha_0} \tag{5}$$

$$\hat{\alpha_0} = \frac{1}{n}\sum_{i=1}^{n}\exp(\hat{\varepsilon_i}) \tag{6}$$

Portanto, para cada modelo estatístico e método de *machine learning* estimado, aplicou-se o teste de hipótese de Jarque-Bera sobre seus resíduos padronizados para avaliar sua normalidade, a um nível de significância de $\alpha = 5\%$. No caso de p-valores inferiores a 0.05, rejeita-se a hipótese nula de normalidade dos resíduos.

Destaca-se que esta análise foi acoplada às implementações do código de cada abordagem preditiva estudada neste trabalho. Assim, reverteu-se automaticamente a transformação logarítmica das previsões utilizando fatores de variância adequados.

5.4 Regressão dinâmica

Para estimar os modelos de regressão dinâmica foi utilizada a biblioteca *statsmodels*, pertencente à linguagem de programação Python. Nesta biblioteca tem-se disponível uma classe denominada OLS capaz de treinar tais modelos via método dos mínimos quadrados (MQO). Dessa forma, estima-se os coeficientes de todas variáveis explicativas que se deseje inserir na dinâmica do modelo.

Neste trabalho decidiu-se trabalhar com três modelos de regressão dinâmica. O primeiro modelo desta categoria, denominado RegrDin(1), faz uma regressão dos *lags* considerados neste trabalho do log da série de visitas à plataforma de vendas *e-commerce* da companhia varejista. Além disso, também regride-se a variável $log(q_t)$ sobre as variáveis *dummy* presentes nas bases de dados ajustadas.

O intuito deste modelo é de avaliar o quão bem a dinâmica das visitas consegue prever a quantidade vendida. Sua formulação é tal qual a Equação 7.

Já o segundo modelo, RegrDin(2), realiza uma regressão dos *lags* do log das séries de preço médio próprio e dos concorrentes sobre a variável dependente. Adicionalmente, a variável $log(q_t)$ é regredida sobre todas as variáveis *dummy* já explicitadas previamente. Busca-se com este modelo avaliar se a dinâmica das séries de preço é capaz fornecer previsões assertivas quanto a quantidade vendida. Vê-se sua formulação na Equação 8.

Por fim, o terceiro modelo busca avaliar a capacidade preditiva dos termos auto regressivos da variável dependente $log(q_t)$ juntamente com todas as variáveis *dummy* que foram acrescentadas à base de dados tratada após os ajustes finais. A este denomina-se RegrDin(3) e vê sua formulação na Equação 9.

$$\log(q_t) = \alpha_0 + \sum_{j \in L} \alpha_j \, log(v_{t-j}) + D_t + \varepsilon_t \tag{7}$$

$$\log(q_t) = \beta_0 + \sum_{j \in L} \left[\beta_j \log(p_{t-j}) + \beta_{1,\,j} \log(pc_{1,\,t-j}) + \beta_{2,\,j} \log(pc_{2,\,t-j}) \right] + D_t + \varepsilon_t \tag{8}$$

$$\log(q_t) = \gamma_0 + \sum_{j \in L} \gamma_j \, log(q_{t-j}) + D_t + \varepsilon_t \tag{9}$$

$$\begin{aligned} D_t = \sum_{h=1}^{23} \delta_h D_{h,t} + \sum_{w=1}^{6} \pi_w D_{w,t} + \phi D_{cc,t} + \tau D_{idp,t} \quad + \\ \sum_{m=0}^{6} \left[\omega_m D_{v,t-4+m} + \rho_m D_{f,t-4+m} \right] + \sum_{i=0}^{11} D_{sal,t-4+i} \end{aligned} \tag{10}$$

5.5 Regressão Lasso

Para estimar a regressão Lasso utilizou-se a classe *LassoLarsIC* da biblioteca *scikit-learn*, pertencente à linguagem de programação Python. Na implementação deste modelo foram aproveitados diversos recursos internos da biblioteca, que faz otimização do fator de penalização λ por meio da minimização dos critérios de informação AIC ou BIC.

Adotou-se como metodologia apenas a minimização do BIC, pois foi observado empiricamente melhores resultados na seleção das variáveis explicativas, ao zerar os coeficientes estimados de maneira mais expressiva, mas principalmente nas previsões feitas, se mostrando mais assertivas do que quando otimizado pela minimização do AIC.

Por fim, ressalta-se que todas as estimações feitas para este modelo, k-passos à frente e para todas as séries horárias de quantidade vendida, incluíram a estimação de um intercepto, juntamente com os coeficientes das variáveis explicativas elencadas na Subseção 4.4. Faz-se uma última ressalva quanto aos valores de

testados para λ, os quais foram definidos automaticamente pelo algoritmo interno da classe *LassoLarsIC*, usada na implementação do modelo.

5.6 AdaLasso

Para este modelo foram consideradas duas variações da estimação dos pesos dos coeficientes das variável independente: a primeira estima uma regressão Ridge, já a segunda estima uma regressão Lasso. Ambas as variações foram implementadas utilizando o pacote *glmnet*, da linguagem de programação R.

Um detalhe a ser mencionado é que no primeiro passo as duas variações são estimadas sem intercepto. Já no segundo passo, estima-se o modelo Lasso com as penalizações $\hat{w_j}$ e λ, mas também um intercepto. A razão disto se fundamenta empiricamente nos melhores resultados obtidos dessa maneira, se comparados a estimação com intercepto desde o primeiro passo.

Outro ponto relevante é que este pacote não fornece uma forma nativa de obter o valor do BIC. No entanto, assumindo como premissa que os termos de erro são independentes e identicamente distribuídos, de acordo com uma distribuição normal, e que a derivada da log-verossimilhança em relação à variância é zero, mostra-se que (KOOPMANS, 1995):

$$BIC = n \cdot log(RSS/n) + k \cdot log(n) \tag{11}$$

Sendo assim, a partir da equação (11) calculou-se o respectivo critério de informação para cada λ avaliado automaticamente pelo pacote *glmnet*. Estima-se então no primeiro passo os pesos $\hat{w_j} = 1/|\hat{\beta_j^*}|^\gamma$ e $\gamma = 1$, a partir dos β_j^* associados ao BIC mínimo. No segundo passo, emprega-se $\hat{w_j}$ na penalização dos coeficientes a serem estimados para cada variável explicativa, otimizando a regressão por meio do λ que minimize o BIC.

Ainda no primeiro passo, implementou-se como explicitado na fundamentação teórica a adição de um *offset* não nulo aos coeficientes estimados $\hat{\beta_j^*}$ estimados via regressão Lasso, tomando como referência a metodologia adotada em (GARCIA; MEDEIROS; VASCONCELOS, 2017). Novamente, o objetivo desta operação é evitar que certos pesos $\hat{w_j}$ resultem numa divisão por zero no caso de $\hat{\beta_j^*} = 0$ (variável explicativa descartada pelo Lasso), ocasionando em erros no treinamento desta variação do AdaLasso.

5.7 Random forest

A implementação deste método de *machine learning* se fundamentou no pacote *rangerts* da liguagem R. Com este pacote foi possível explorar variações do *bootstrapping* padrão do algoritmo de *random forest*, cujos próprios desenvolvedores do pacote apontam resultados promissores no emprego de *block bootstrapping* para a previsão de séries temporais de frequência horária (GOEHRY et al., 2021).

Ao todo foram considerados três hiperparâmetros ao longo do processo de

fine-tuning do método: o tamanho dos blocos para realizar o *bootstrapping* do treinamento das árvores de regressão internas (*block_sizes*), o tipo de *bootstrapping* em blocos utilizado (*bootstraps*) e por fim, as m-variáveis explicativas disponíveis para realizar a bipartição sucessiva em cada nó das árvores (*m_preds*).

Um tópico de destaque que diz respeito à metodologia adotada para o *fine-tuning* é o conjunto de dados de validação, criado internamente para realizar este procedimento. Por padrão, considerou-se um período equivalente a 25% dos registros pertencentes ao conjunto de treinamento passado para a instância de *random forest*.

Além disso, foram estabelecidas cinco *seeds* distintas para a geração de números aleatórios. Com isso, para cada configuração dos hiperparâmetros, garante-se escolhas mais variadas dos índices de blocos que irão compor as séries temporais de treinamento do RF, além de tornar os resultados aqui apresentados reprodutíveis. Desse modo, avalia-se o RMSE médio de cada instância de RF para definir qual combinação dos valores considerados para os hiperparâmetros apresenta melhor capacidade preditiva.

Determinada a melhor configuração dos hiperparâmetros, refez-se o treinamento do *random forest* com tais valores, realizando por fim sua previsão sobre o conjunto de registros destinados ao período de testes, respectivos ao horizonte de previsão k-passos à frente em questão.

No caso de *block_sizes* foram explorados valores $\{7, 12, 24, 36, 48, 60, 72\}$. Destaca-se desse conjunto os valores múltiplos de 24, pois sugere-se como heurística no artigo de referência do pacote testar os múltiplos da sazonalidade mais proeminente da série temporal em questão. No caso das séries horárias de quantidade vendida dos SKU's identifica-se como a mais proeminente a sazonalidade diária. Os demais valores foram avaliados para testar uma gradação mais granular do tamanho dos blocos.

Quanto as m-variáveis explicativas disponíveis ao longo do treinamento das árvores de regressão do algoritmo, tomou-se como base o valor citado na literatura de $\sqrt{p}$, sendo p neste caso o número total de variáveis explicativas presentes nas bases de dados ajustadas (ver Subseção 4.4). A partir deste valor de referência, considerou-se ao todo para *m_preds* o conjunto de valores $\{9, 7, 5, 3\}$.

Por último, dentre os tipos de *block bootstrapping* disponíveis no pacote, testou-se os tipos *moving block*, *circular*, *non-overlapping*, cujas referências são citadas brevemente na fundamentação teórica. Com relação a quantidade de árvores de regressão internas, treinou-se 1000 destas para cada instância de *random forest*.

5.8 Métricas de aderência

Mean Absolute Error (MAE)

Esta métrica de aderência calcula a média aritmética dos erros absolutos cometidos pelo modelo sobre o conjunto de dados de avaliação. Por considerar o somatório os desvios absolutos entre o valor observado da quantidade vendida, y_t e o valor previsto, $\hat{y}_t$, o MAE dá o mesmo peso a desvios de magnitude baixa ou elevada, como mostrado na sua formulação em (12).

Portanto, decidir por qual melhor modelo dentre todos os estimados através do MAE implica um cenário operacional que preveja maior robustez à ocorrência de *outliers*. Na prática, dá-se maior importância ao agregado dos erros, sejam eles negativos ou positivos, do que a eventuais valores extremos destes. No fim, tais valores são diluídos no valor médio, conforme a amostra de teste se torna mais volumosa.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_{t+k} - y_{t\hat{+}k|t}| \tag{12}$$

Root Mean Squared Error (RMSE)

Por outro lado, quando se deseja dar mais ênfase a valores extremos dos erros de previsão é possível empregar a métrica RMSE. A partir de sua formulação em (13) verifica-se que os erros considerados são elevados ao quadrado; sendo assim, quanto o maior o desvio, positivo ou negativo, mais elevado será a contabilização deste ponto amostral para a avaliação da capacidade preditiva dos modelos.

Novamente, entende-se a escolha desta métrica na tomada de decisão pela otimalidade preditiva dentre as estimações cabe a um profissional do setor operacional do varejo, assumindo a premissa de que modelo escolhido deve produzir predições com uma menor incidência de valores extremos.

Na prática, subavaliações ou superavaliações extremas da quantidade vendida, fixado um horizonte de predição, representariam maiores impactos na cadeia operacional de vendas da empresa varejista.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_{t+k} - y_{t\hat{+}k|t})^2} \tag{13}$$

5.9 Implementação dos códigos do projeto

Para estimar todos os modelos propostos implementou-se códigos na linguagens de programação Python e R, das quais cita-se abaixo alguma informações relevantes. Dentre essas, estão elencadas a versão da linguagem instalada, as dependências do projeto quanto às bibliotecas utilizadas e quais foram as principais responsáveis pela implementação dos modelos preditivos, abordados na seção de

fundamentação teórica.

Python

- **Versão**: 3.10.7
- **Principais dependências**: statsmodels.api.OLS, scikit-learn.LassoLarsIC, pandas, numpy, plotly, pickle.
- **Abordagens preditivas implementadas**: modelos de regressão dinâmica, estimados via método dos mínimos quadrados ordinários (MQO) e regressão Lasso, com otimização do fator de penalização λ via BIC.

R

- **Versão**: 4.3.1 (2023-06-16 ucrt)
- **Principais dependências**: rangerts, glmnet, ggplot2, tibble, tseries, Metrics.
- **Abordagens preditivas implementadas**: modelo AdaLasso, incluindo ambas as variações de estimação dos pesos w_j via regressão Ridge ou Lasso. Implementação do método de *machine learning random forest* com uso de *block bootstraping*.

Para acessar os códigos desenvolvidos ao longo deste projeto, desde o tratamento das bases de dados, até o módulo que estima os modelos e métodos preditivos com os resultados em gráficos e tabelas, segue o link de acesso público:

https://drive.google.com/drive/folders/1smX6T8xarmo9UuqbndXib4s9Uqzk4EMt?usp=sharing

6 Resultados

Nesta seção, aborda-se as estimações de cada um dos modelos e métodos preditivos estudados para os k-passos à frente, $k \in \{1, 2, ...12\}$. Sendo assim, busca-se definir o melhor modelo para prever a quantidade vendida de um dado SKU, fixado o horizonte de previsão, por meio das métricas de aderência adotadas. Ademais, faz-se a inspeção visual do passo $k = 1$ sob o conjunto de testes a nível de exemplificação.

Optou-se por não adentrar em todos os detalhes de interpretação. Cita-se como exemplo a importância das variáveis selecionadas pelo *random forest*, a análise de cada hiperparâmetro avaliado no *fine-tuning* deste método e o estudo individualizado dos coeficientes estimados de cada variável explicativa selecionada pelos modelos da família Lasso. Entende-se que, dado o volume conjunto das abordagens preditivas, k-passos à frente e séries temporais analisadas, tais estudos levariam a um desvio do principal objetivo deste trabalho.

Além disso, é crucial observar que a maioria dos modelos estimados realiza uma regressão que pode extrapolar os valores observados durante a fase de treinamento, apresentando coeficientes de diferentes magnitudes e sinais. Ao examinar os gráficos das séries horárias observadas e previstas para $k = 1$, fica evidente que essas regressões frequentemente resultam em valores previstos negativos.

Por construção, é sabido que valores negativos não condizem com o domínio da variável dependente de interesse (quantidade de vendas no varejo *e-commerce*). No entanto, optou-se por manter esses valores como uma simplificação do trabalho, em vez de aplicar uma regularização dos mesmos. Essa decisão foi tomada com o receio de que esses ajustes prejudicassem a consistência dos estimadores e as premissas originais assumidas pelos modelos estatísticos.

No que diz respeito ao modelo *random forest* (RF), é importante destacar que suas previsões não foram consideradas ótimas para nenhum SKU. Observa-se que na publicação (GOEHRY et al., 2021), são ajustados modelos RF com *block bootstrapping* também sobre séries de frequência horária.

Entretanto, essas séries referem-se à carga elétrica de edificações, as quais normalmente apresentam características mais perenes. A maior inconstância das séries de varejo, decorrente de padrões de consumo diferentes daqueles vinculados ao setor elétrico, pode ser a causa implícita dessa divergência nos resultados.

Adicionalmente, prossegue-se com as análises dos resultados, as quais estão separadas por SKU.

SKU 1

A partir das métricas na Tabela 1, identifica-se ao longo de todos os horizontes de previsão que os melhores resultados sob a métrica MAE estão atrelados ao modelo AdaLasso, com pesos estimados por regressão Ridge. As exceções são os horizontes de previsão com $k \in \{4, 7, 12\}$, cujos melhores MAE estão ligados aos modelos AdaLasso (com pesos estimados via Lasso primário) e Lasso (para $k = 7$).

Já no caso da métrica RMSE observa-se superioridade unânime sobre todos os horizontes de previsão por parte do modelo RegrDin(3), que emprega as diversas variáveis *dummy* em conjunto com os termos auto regressivos da variável dependente.

Também vê-se na Figura 7 que no caso do horizonte de previsão de uma hora à frente, de fato os modelos aparentam reproduzir previsões mais fiéis aos valores observados

Model	Metric	T + 1	T + 2	T + 3	T + 4	T + 5	T + 6	T + 7	T + 8	T + 9	T + 10	T + 11	T + 12
RegrDin(1)	RMSE	3.80	3.80	3.79	3.77	3.78	3.79	3.80	3.80	3.81	3.81	3.81	3.81
	MAE	1.60	1.59	1.59	1.59	1.61	1.62	1.64	1.64	1.64	1.64	1.63	1.63
RegrDin(2)	RMSE	3.83	3.83	3.84	3.85	3.85	3.84	3.84	3.84	3.85	3.86	3.86	3.87
	MAE	1.66	1.66	1.65	1.65	1.64	1.64	1.65	1.66	1.65	1.65	1.66	1.66
RegrDin(3)	RMSE	**3.42**	**3.45**	**3.47**	**3.48**	**3.51**	**3.51**	**3.49**	**3.42**	**3.53**	**3.49**	**3.38**	**3.51**
	MAE	1.50	1.58	1.69	1.79	1.87	1.90	1.91	1.91	1.94	1.85	1.81	1.81
Lasso	RMSE	3.54	3.57	3.63	3.70	3.78	3.81	3.82	3.80	3.82	3.76	3.75	3.72
	MAE	1.41	1.46	1.47	1.51	1.57	1.59	**1.59**	1.59	1.60	1.57	1.57	1.57
AdaLasso (w = ridge)	RMSE	3.56	3.62	3.69	3.75	3.77	3.81	3.87	3.84	3.84	3.83	3.78	3.78
	MAE	**1.38**	**1.42**	**1.44**	1.50	**1.51**	**1.55**	1.60	**1.57**	**1.56**	**1.56**	**1.53**	1.54
AdaLasso (w = lasso)	RMSE	3.54	3.61	3.70	3.76	3.80	3.81	3.86	3.85	3.86	3.85	3.80	3.74
	MAE	1.39	1.42	1.44	**1.49**	1.53	1.55	1.59	1.59	1.59	1.57	1.54	**1.50**
Random Forest	RMSE	3.56	3.54	3.55	3.52	3.67	3.76	3.77	3.72	3.78	3.76	3.75	3.76
	MAE	1.65	1.88	2.15	2.13	2.56	2.74	2.76	2.68	2.75	2.69	2.69	2.65

Tabela 1: Métricas de aderência k-passos à frente; série temporal SKU 1.

da série de vendas são os modelos AdaLasso (w = Ridge) e RegrDin(3). Isto fica mais nítido nos picos de vendas repentinos que ocorrem na série; apesar de não serem precisos sempre, tais modelos conseguem acompanhar esses picos nas suas previsões, diferente dos demais modelos que não parecem encapsular tal comportamento e mantém previsões muito próximas a zero durante toda a série.

Destaca-se que esta série temporal tem como particularidade baixas contagens se comparada as séries dos outros SKU's explorados. Observa-se ainda múltiplos intervalos de intermitência entre contagens nulas de vendas e valores positivos, marcados pelos picos evidenciados no período de testes. Portanto, entende-se que apesar de suas limitações, os melhores modelos obtiveram uma acurácia razoável, destacada na Figura 8.

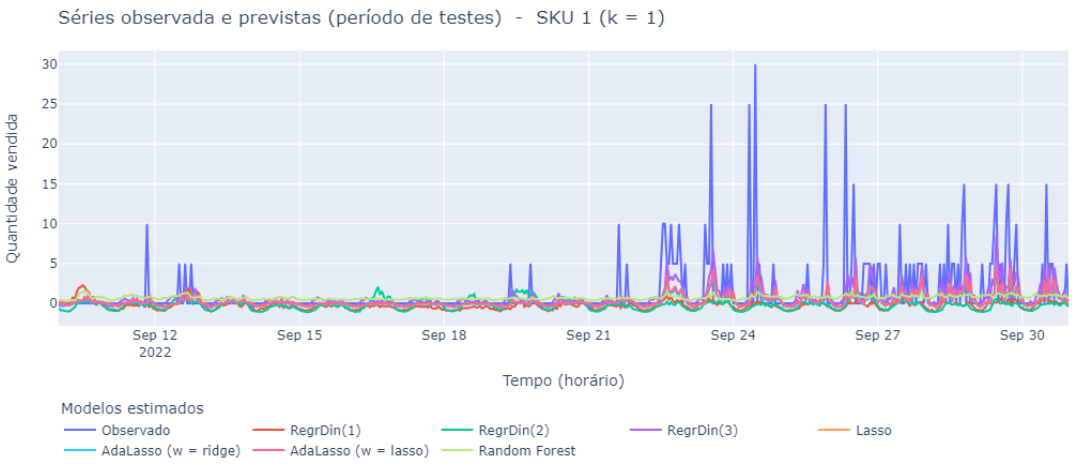

Figura 7: Séries observada e previstas, um passo à frente, SKU 1.

SKU 2

Observando a Tabela 2, nota-se uma maior alternância entre os melhores modelos, se fixada a métrica de aderência em análise. Analisando o RMSE, percebe-se que o melhor modelo na maioria dos horizontes de previsão foi RegrDin(3), seguido pelos modelos Lasso e RegrDin(1).

Fazendo uma análise similar do MAE, vê-se que na maioria dos horizontes de

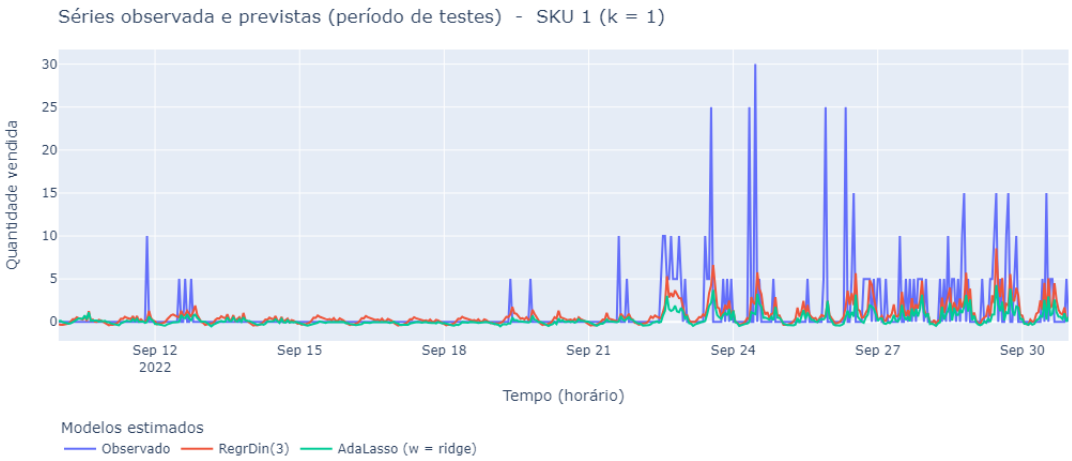

Figura 8: Séries observada e previstas pelos melhores modelos, um passo à frente, SKU 1.

previsão que os modelos foram estimados, o modelo Lasso alcançou os melhores resultados, com exceção dos instantes de previsão $T + k$, tal que $k \in \{4, 5, 10\}$, cujo melhor modelo foi o RegrDin(1); já no instante $T + 9$, teríamos como ótimo o ajuste do modelo RegrDin(3).

Model	Metric	T + 1	T + 2	T + 3	T + 4	T + 5	T + 6	T + 7	T + 8	T + 9	T + 10	T + 11	T + 12
RegrDin(1)	RMSE	19.84	20.10	20.37	**20.36**	**20.31**	20.23	20.17	20.17	20.14	**20.24**	20.06	19.86
	MAE	10.71	10.88	11.00	**10.94**	**10.93**	10.97	10.93	11.01	10.99	**11.08**	11.11	10.84
RegrDin(2)	RMSE	20.44	20.53	20.67	20.84	20.87	20.75	20.51	20.55	20.76	20.74	20.72	20.83
	MAE	10.96	11.09	11.22	11.23	11.18	11.13	10.96	11.03	11.20	11.18	11.21	11.33
RegrDin(3)	RMSE	19.46	20.28	**20.12**	20.38	20.59	**20.09**	**20.10**	**19.97**	**19.83**	20.27	20.32	20.00
	MAE	10.70	11.15	11.03	11.30	11.70	10.88	11.08	11.10	**10.81**	11.34	11.23	11.03
Lasso	RMSE	**18.88**	**19.73**	20.46	20.73	20.46	20.29	20.20	20.17	20.67	20.48	**20.05**	**19.77**
	MAE	**10.27**	**10.68**	**11.00**	11.39	11.04	**10.87**	**10.78**	**10.83**	11.13	11.10	**10.96**	**10.59**
AdaLasso (w = ridge)	RMSE	19.84	20.98	21.44	21.47	21.43	21.15	21.50	21.64	21.51	22.09	21.96	21.75
	MAE	10.89	11.38	11.67	11.72	11.64	11.43	11.58	11.74	11.67	12.44	12.53	12.18
AdaLasso (w = lasso)	RMSE	19.83	21.00	21.85	21.53	21.47	21.24	21.18	21.56	21.70	21.66	21.81	21.61
	MAE	10.82	11.43	12.16	11.77	11.67	11.45	11.31	11.65	11.87	11.87	12.33	11.96
Random Forest	RMSE	20.10	20.86	21.38	21.78	22.20	22.59	22.84	22.93	22.86	23.02	22.87	22.26
	MAE	11.19	11.69	12.05	12.44	12.58	12.83	13.11	13.13	12.98	13.18	13.05	12.49

Tabela 2: Métricas de aderência k-passos à frente; série temporal SKU 2.

Por inspeção visual das previsões sobre os pontos amostrais de teste um passo à frente, nas Figuras 9 e 10, percebe-se que as curvas previstas pelos modelos Lasso e RegrDin(3) realmente são as que melhor acompanham a série horária da quantidade vendida para este SKU. Além de acompanhar a sazonalidade diária, apresentam previsões mais elevadas que a de outros modelos em picos de vendas observadas, como nos dias 25 e 26 de setembro de 2022.

É interessante destacar que dentre as variáveis explicativas selecionadas pelo modelo Lasso encontra-se lags das variável dependente, de curto e longo prazo (ver Subseção 4.4), assim como o modelo de regressão dinâmica RegrDin(3), cujo conjunto de variáveis explicativas consiste em todos os lags trabalhados da variável dependente, além de todas as dummies presentes na base de dados tratada.

Desse modo, entende-se que a série horária de quantidade vendida do SKU 2 apresenta aspectos de um processo auto regressivo com sazonalidade, mas também associado à dinâmica de outras variáveis exógenas. Retomando o exemplo das estimativas um passo à frente (k = 1), verifica-se no modelo Lasso coeficientes estimados não nulos

para lags de visita à plataforma de vendas e de algumas variáveis de preço.

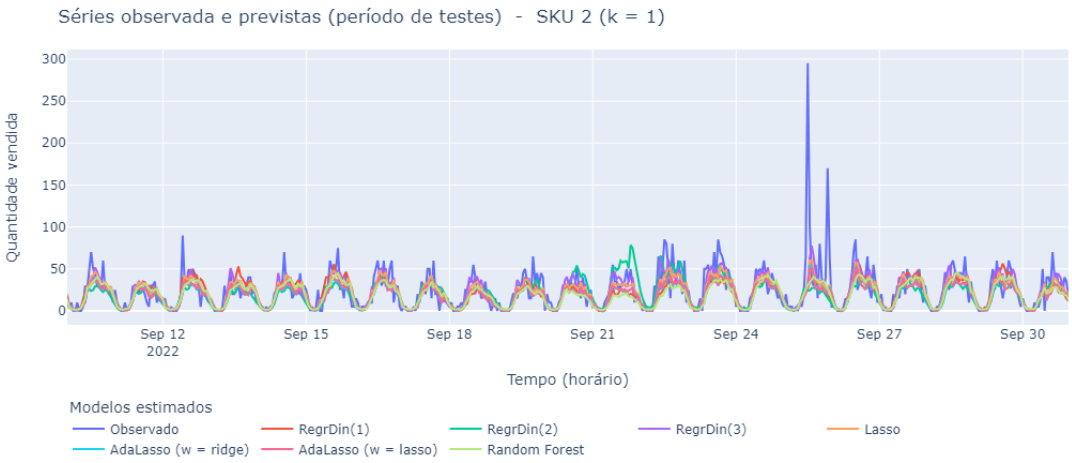

Figura 9: Séries observada e previstas, um passo à frente, SKU 2.

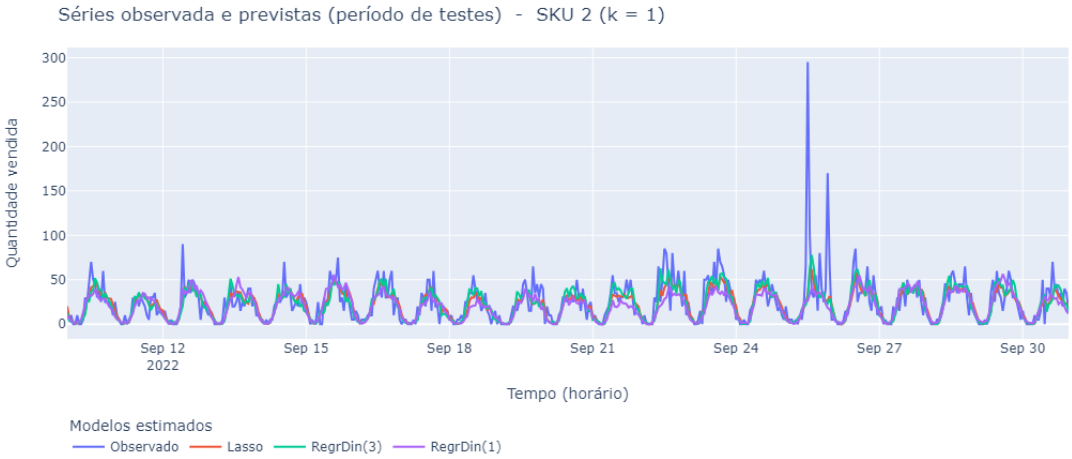

Figura 10: Séries observada e previstas pelos melhores modelos, um passo à frente, SKU 2.

SKU 3

Da análise das métricas de aderência dos modelos ajustados sobre a série horária deste SKU, na Tabela 3, identifica-se como os melhores modelos RegrDin(3) e AdaLasso, com pesos dos coeficientes estimados via regressão Ridge. Em termos de RMSE, nota-se que para horizontes de previsão de mais curto prazo, com $k \in \{1, 2, 3, 4\}$, o modelo AdaLasso performa melhor. A partir de $k = 5$, o melhor modelo passa a ser RegDin(3).

Tratando da métrica MAE, verifica-se um maior destaque para RegrDin(3) que apresenta um MAE ótimo para todos os horizontes de previsão considerados, a menos do primeiro. Para este horizonte tem-se o MAE ótimo vinculado ao modelo estimado pelo AdaLasso.

Sabe-se que o modelo RegDin(3) é uma regressão dinâmica com intercepto, que regride os termos auto regressivos da variável dependente, $log(q_t)$, além das dummies presentes na base de dados. Com isso, entende-se que o perfil da série horária da quantidade vendida do SKU 3 também apresenta um perfil auto regressivo e com sazonalidade diária.

Ressalta-se ainda uma menor sensibilidade às variações do preço próprio e dos concorrentes, evidenciada pelos piores valores obtidos pelas métricas de aderência de RegrDin(2) em todos os horizontes de previsão. Dentre estes, toma-se como exemplo novamente as previsões para $k = 1$ da Figura 11, no qual o maior descolamento da série horária observada se dá pela série prevista por RegrDin(2).

Model	Metric	T + 1	T + 2	T + 3	T + 4	T + 5	T + 6	T + 7	T + 8	T + 9	T + 10	T + 11	T + 12
RegrDin(1)	RMSE	26.99	27.16	29.48	28.81	29.67	29.09	29.37	28.85	29.49	29.49	31.17	30.43
	MAE	11.74	11.86	12.95	12.76	13.42	13.28	13.53	13.30	13.17	13.23	13.88	13.59
RegrDin(2)	RMSE	60.28	58.43	56.11	55.29	53.47	52.20	51.64	51.81	53.10	51.89	52.31	52.01
	MAE	40.90	39.99	38.55	37.92	37.07	36.62	36.10	36.21	36.85	36.39	36.41	35.89
RegrDin(3)	RMSE	17.49	20.03	21.16	22.02	**22.15**	**22.87**	**23.10**	**23.75**	**24.79**	**25.50**	**26.28**	**26.80**
	MAE	7.47	**8.63**	**9.36**	**10.12**	**10.40**	**10.94**	**10.92**	**11.35**	**11.90**	**12.12**	**12.54**	**12.83**
Lasso	RMSE	14.65	19.38	22.14	26.65	30.70	34.29	37.42	38.82	42.61	41.49	45.33	47.56
	MAE	7.51	10.27	11.43	14.04	16.38	18.54	21.45	21.44	24.60	22.23	22.26	21.74
AdaLasso (w = ridge)	RMSE	**13.52**	**18.08**	**20.12**	**21.07**	25.82	30.54	30.71	30.92	34.83	37.43	42.48	42.19
	MAE	**6.50**	9.22	11.03	11.12	14.00	16.86	17.30	17.37	19.50	19.78	20.31	19.04
AdaLasso (w = lasso)	RMSE	14.28	18.22	20.27	21.69	25.89	26.05	28.60	28.58	29.13	33.16	38.54	39.40
	MAE	6.81	8.92	10.04	11.54	13.71	14.19	15.52	15.24	15.21	17.68	18.30	18.22
Random Forest	RMSE	19.88	23.05	25.46	26.68	27.70	28.01	28.59	28.65	28.38	27.41	27.43	27.24
	MAE	9.26	11.23	13.29	15.63	18.15	19.00	19.62	18.27	18.14	16.95	16.27	15.49

Tabela 3: Métricas de aderência k-passos à frente; série temporal SKU 3.

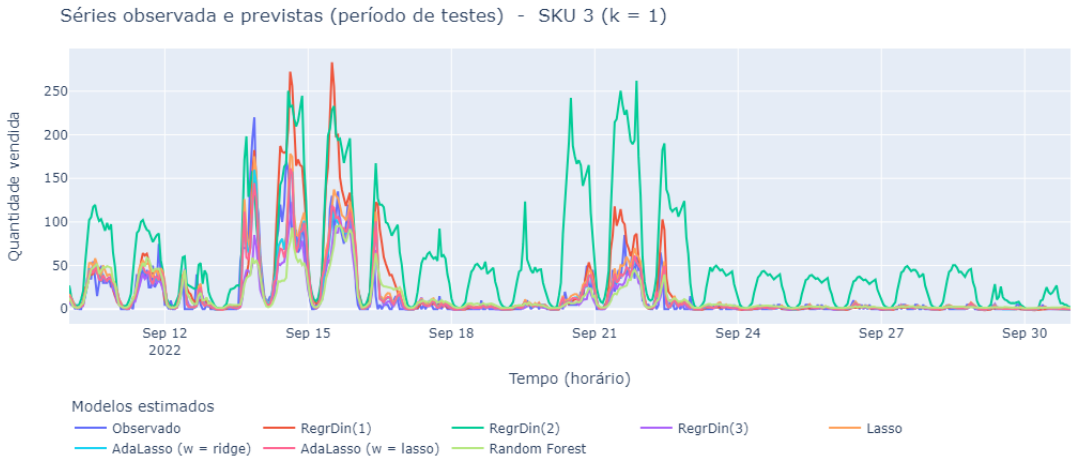

Figura 11: Séries observada e previstas, um passo à frente, SKU 3.

Destaca-se por fim os dias com picos de vendas, em que também evidencia-se o descolamento das previsões obtidas pelo modelo RegrDin(1) - emprega lags do número de visitas à plataforma de vendas da varejista. Portanto, infere-se que em situações de número de vendas atípicas, a dinâmica da quantidade vendida horas antes demonstra melhor acurácia preditiva sobre a dinâmica do número de visitas.

Por fim, vê-se na Figura 12 a série observada do SKU 3 e das séries previstas pelos modelos de melhor ajuste no período de testes: RegrDin(3) e AdaLasso, com pesos estimados via regressão Ridge. Destaca-se a adequação de ambas as séries previstas mesmo em dias atípicos, com fortes picos de vendas, seguidos de pontos amostrais de contagem nula ou próximo disto.

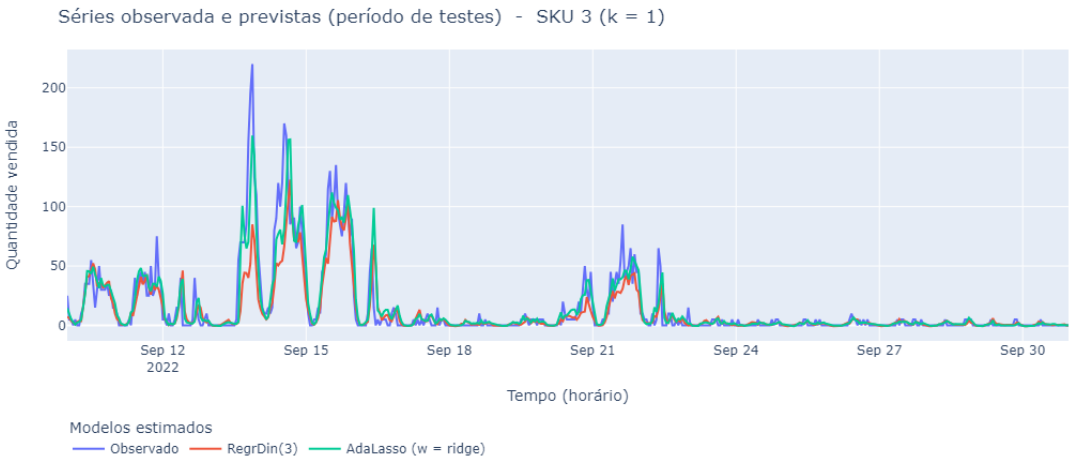

Figura 12: Séries observada e previstas pelos melhores modelos, um passo à frente, SKU 3.

SKU 4

No caso deste SKU, obteve-se nas métricas de aderência da Tabela 4 a determinação do modelo AdaLasso (com pesos estimados via Lasso) como o melhor modelo sob a ótica da métrica MAE, para a maior parte dos horizontes de previsão. Por outro lado, nota-se uma predominância de otimalidade do modelo RegrDin(1) sob a avaliação da métrica RMSE, k-passos à frente. Poucas são as exceções, que ocorrem nos horizontes de previsão $T+1$ e $T+8$, nos quais dependendo da métrica analisada é tem-se o ótimo nos modelos AdaLasso (w = Ridge) ou Lasso.

Model	Metric	T + 1	T + 2	T + 3	T + 4	T + 5	T + 6	T + 7	T + 8	T + 9	T + 10	T + 11	T + 12
RegrDin(1)	RMSE	9.16	9.34	9.58	9.79	10.13	10.10	10.29	10.28	10.13	10.19	9.66	9.43
	MAE	6.18	6.31	6.45	6.66	6.84	6.83	6.92	6.92	6.86	6.95	6.71	6.52
RegrDin(2)	RMSE	41.66	41.70	41.90	41.71	40.52	40.23	40.16	39.30	39.23	39.49	39.11	38.02
	MAE	25.63	26.03	26.39	26.67	26.07	25.94	26.02	25.27	25.09	25.33	25.20	24.51
RegrDin(3)	RMSE	9.18	10.05	11.08	12.11	13.07	14.47	15.32	15.81	15.51	15.36	14.98	13.88
	MAE	5.98	6.70	7.35	8.12	8.60	9.44	10.06	10.57	10.28	10.39	10.16	9.49
Lasso	RMSE	8.94	10.11	11.14	10.95	10.46	11.17	11.45	11.02	10.38	10.10	10.22	9.66
	MAE	5.70	6.48	6.84	7.10	6.93	7.39	7.68	7.45	7.07	6.94	7.09	6.49
AdaLasso (w = ridge)	RMSE	8.96	9.98	11.32	10.17	10.54	10.65	10.76	10.24	10.14	9.83	9.98	9.49
	MAE	5.65	6.26	6.77	6.66	6.84	6.93	7.14	6.81	6.77	6.63	6.80	6.31
AdaLasso (w = lasso)	RMSE	8.98	9.86	11.13	11.68	10.37	10.51	11.14	10.28	9.77	9.42	9.11	9.16
	MAE	5.66	6.12	6.67	7.12	6.67	6.81	7.22	6.76	6.53	6.35	6.08	5.96
Random Forest	RMSE	8.98	9.39	10.21	13.66	18.66	23.75	24.96	24.50	23.12	20.90	18.15	14.74
	MAE	5.76	6.41	7.33	10.07	13.72	17.83	19.02	18.89	18.24	17.09	15.28	12.33

Tabela 4: Métricas de aderência k-passos à frente; série temporal SKU 4.

Vê-se novamente na Figura 13 um forte descolamento das séries observada e prevista pelo modelo RegrDin(2), que envolve apenas a dinâmica *lags* das variáveis de preço em conjunto as *dummies* abordadas. Desse modo, entende-se que este modelo têm baixa capacidade preditiva da quantidade vendida do SKU 4, similar ao SKU 3. Esse descolamento é refletido pelos valores mais elevados de RMSE deste modelo, em todos os k-passos à frente, uma vez que tal métrica penaliza mais fortemente erros de maior magnitude.

Apesar dos modelos AdaLasso (w = Lasso) e RegrDin(1) apresentarem os melhores valores de MAE e RMSE, respectivamente, para a maioria dos horizontes de previsão, considera-se na Figura 14 a adequação dos modelos Lasso e AdaLasso (w = Ridge) para

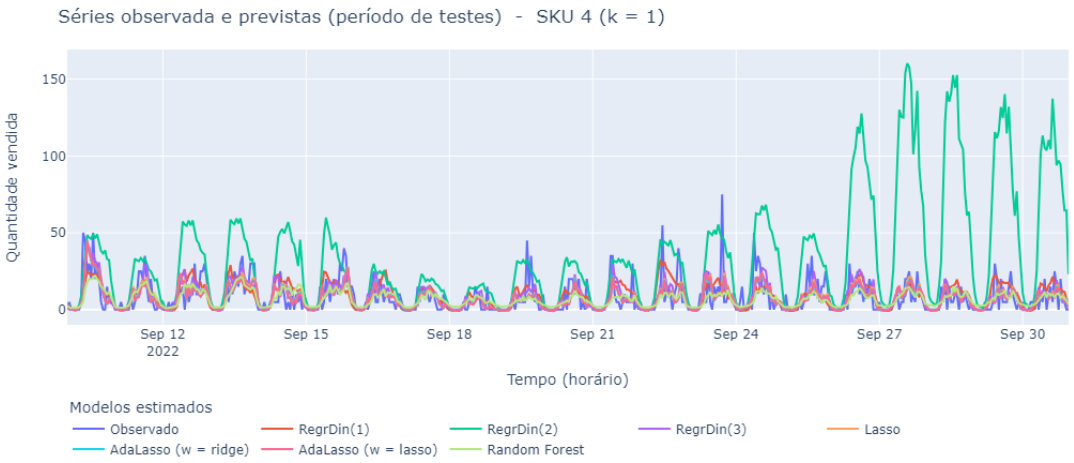

Figura 13: Séries observada e previstas, um passo à frente, SKU 4.

$k = 1$ por se mostrarem os melhores modelos nesse instante. Observa-se que ambos os modelos fornecem previsões um passo à frente capazes de acompanhar a sazonalidade da série de quantidade vendida do SKU 4 e a tendência de alta seguida de queda nas vendas evidenciado pelos picos no gráfico.

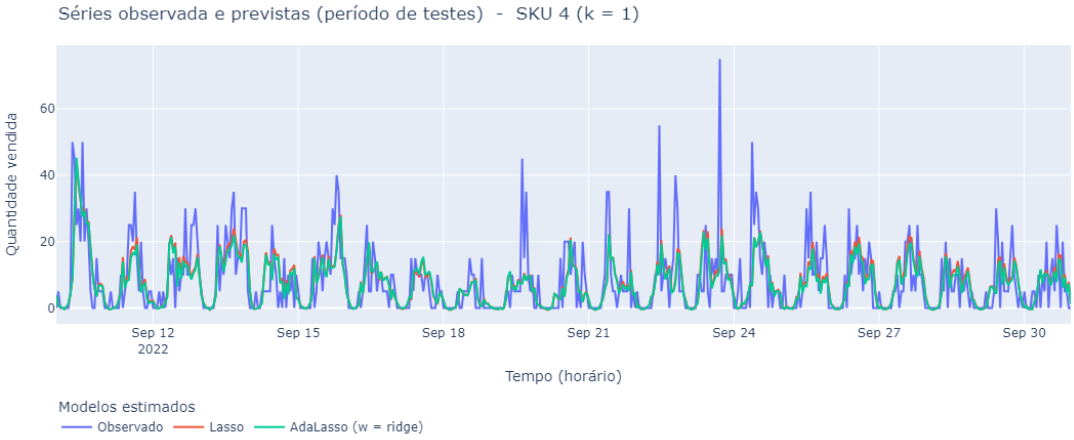

Figura 14: Séries observada e previstas pelos melhores modelos, um passo à frente, SKU 4.

SKU 5

Por meio da Tabela 5 percebe-se unanimidade na identificação do melhor modelo quando faz-se a análise pela métrica MAE. Neste caso, atribui-se esse título ao modelo Lasso. Sob a análise da métrica de aderência RMSE, vê-se que em dois terços dos horizontes de previsão considerados tem-se como melhor modelo RegrDin(3). Nesta regressão dinâmica incorpora-se termos auto regressivos da quantidade vendida, além das *dummies* já apresentadas. No outro terço dos horizontes de previsão, correspondentes a $T + k$ com $k \in \{5, 6, 7, 8\}$, verifica-se por meio do RMSE que o modelo Lasso obtém os melhores resultados.

Da Figura 15 nota-se a terceira ocorrência de um descolamento das previsões feitas pelo modelo de regressão dinâmica envolvendo as séries preços, RegrDin(2), o que mais uma vez reflete nos valores de RMSE mais elevados ao longo dos k-passos à frente

Model	Metric	T + 1	T + 2	T + 3	T + 4	T + 5	T + 6	T + 7	T + 8	T + 9	T + 10	T + 11	T + 12
RegrDin(1)	RMSE	36.18	35.98	35.52	34.72	34.32	34.02	34.16	34.63	34.70	35.10	35.22	35.06
	MAE	20.88	20.53	20.10	19.69	19.44	19.16	19.23	19.48	19.44	19.54	19.53	19.52
RegrDin(2)	RMSE	50.47	50.07	49.38	48.17	47.85	47.65	47.89	47.93	46.95	48.83	49.30	48.45
	MAE	35.99	35.59	35.05	34.32	34.18	34.12	34.51	34.25	34.00	35.05	35.01	34.77
RegrDin(3)	RMSE	24.22	25.81	26.09	26.94	28.25	29.15	29.29	29.92	29.25	29.00	29.03	28.71
	MAE	13.79	14.77	15.26	15.63	16.54	16.55	16.55	17.06	16.63	17.01	16.64	16.39
Lasso	RMSE	25.29	27.82	28.52	28.02	28.05	28.33	28.54	29.70	30.26	30.87	31.00	30.29
	MAE	13.64	14.44	14.69	14.58	14.53	14.53	14.80	15.72	16.06	16.58	16.62	16.12
AdaLasso (w = ridge)	RMSE	25.82	28.46	29.32	29.10	29.10	29.41	29.80	30.85	31.50	31.62	32.02	31.37
	MAE	13.92	15.00	15.26	15.27	15.30	15.31	15.44	16.41	16.86	17.22	17.35	17.18
AdaLasso (w = lasso)	RMSE	26.92	30.09	30.24	30.60	29.44	29.82	31.89	30.79	31.41	32.56	32.33	32.04
	MAE	14.41	15.85	15.81	16.32	15.65	15.59	17.04	16.34	16.83	17.84	17.59	17.69
Random Forest	RMSE	31.02	31.50	32.77	32.91	31.90	31.54	32.11	32.52	32.29	31.09	30.68	30.26
	MAE	16.80	17.11	18.47	18.99	19.24	20.22	21.35	22.25	21.72	20.49	19.91	19.36

Tabela 5: Métricas de aderência k-passos à frente; série temporal SKU 5.

avaliados.

Por outro lado, enquanto RegrDin(2) superavalia a quantidade vendida ao longo do período de testes, verifica-se que o modelo RegrDin(1) realiza subavaliações da mesma. Percebe-se isto também da análise da Figura 15, em que a série prevista de quantidade vendida do modelo RegrDin(1) apresenta valores inferiores aos das séries previstas pelos demais modelos e inferior aos da própria série observada.

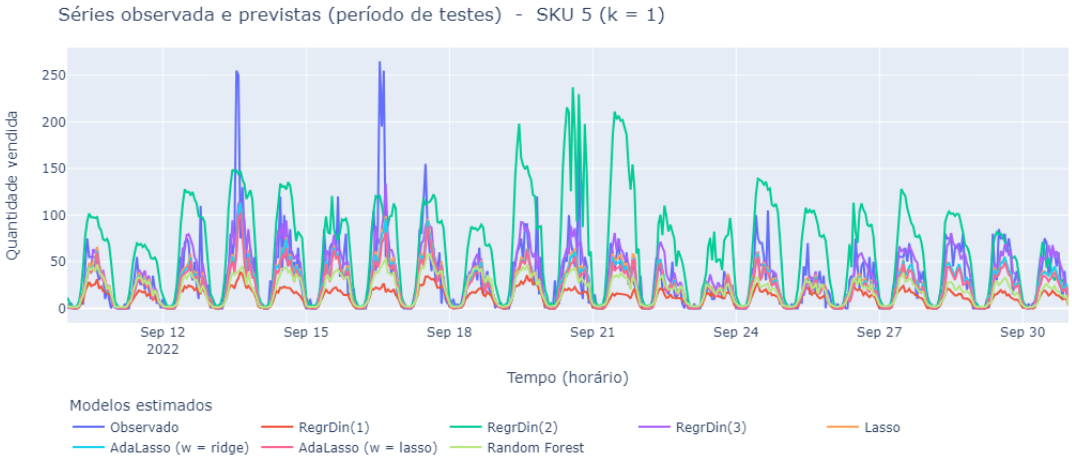

Figura 15: Séries observada e previstas, um passo à frente, SKU 5.

Observando ainda os picos de vendas mais destoantes ocorridos nos dias 13 e 16 de setembro e dos restante da série horária observada do SKU 5, no período de testes para $k = 1$, observa-se que os modelos que melhor conseguem acompanhar essa disrupção da alta nas vendas são de fato os modelos RegrDin(3) e Lasso. Isto fica ainda mais evidente ao isolar as séries previstas desses modelos frente a série observada no mesmo período, na Figura 16.

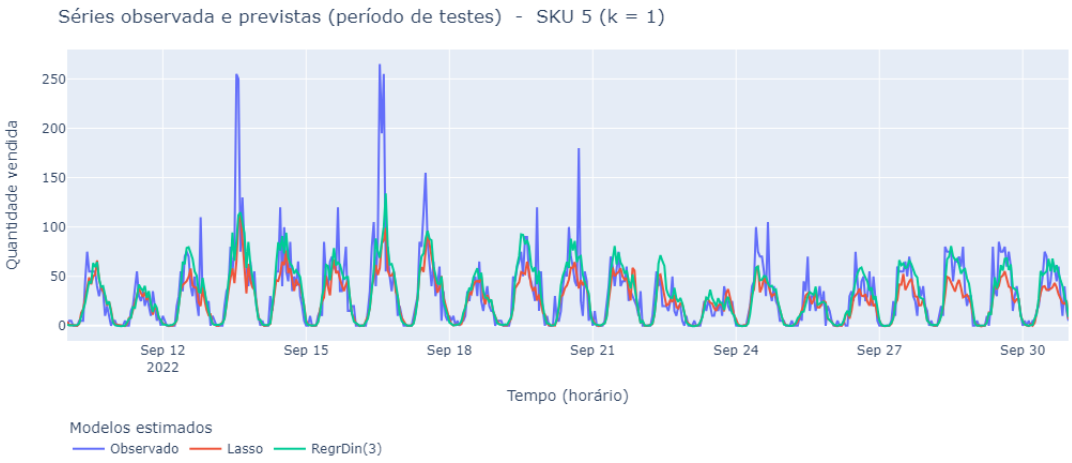

Figura 16: Séries observada e previstas pelos melhores modelos, um passo à frente, SKU 5.

7 Conclusão

Neste trabalho explorou-se diferentes modelos estatísticos e métodos de *machine learning* na predição k-passos à frente, $k \in \{1, 2, ..., 12\}$, das séries horárias de quantidade vendida de cinco SKU's, comercializados na plataforma *e-commerce* de uma companhia varejista.

Por fim, tendo feito a análise dos resultados segmentada por SKU, mostra-se na Tabela 6 sua relação com os modelos estimados e a indicação pelas métricas de aderência MAE e RMSE de quais dentre esses modelos apresentou um maior número de vezes o melhor valor para a respectiva métrica.

	RegrDin(1)	RegrDin(2)	RegrDin(3)	Lasso	AdaLasso (w = ridge)	AdaLasso (w = lasso)	Random forest
SKU 1			RMSE		MAE		
SKU 2			RMSE	MAE			
SKU 3			RMSE / MAE				
SKU 4	RMSE					MAE	
SKU 5			RMSE	MAE			

Tabela 6: Melhores modelos ajustados por SKU, via contagem k-passos à frente das métricas de aderência ótimas.

Com isso, conclui-se que sob a ótica da métrica RMSE, o modelo RegrDin(3) se mostra o mais adequado na previsão da maioria dos instantes futuros considerados, para quatro dos cinco SKU's em análise. Portanto, entende-se que o modelo de regressão dinâmica que incorpora os *lags* da variável dependente, $log(q_t)$, e as *dummies* presentes na base de dados tratada é capaz de prover uma maior resistência a erros de alta ou baixíssima magnitude.

A exceção se evidencia para o SKU 4, o qual uma análise similar levaria à escolha pelo modelo RegrDin(1) - apresentando maior explicabilidade dos valores futuros da quantidade vendida deste SKU, por meio da dinâmica de visitas registradas às plataformas de vendas *online*.

A respeito da métrica de aderência MAE, conclui-se empiricamente da Tabela 6 que os SKU's estudados têm como melhor modelo para a maioria dos horizonte de previsão

analisados, uma variação do modelo AdaLasso (pesos estimados via regressão Ridge, para o SKU 1, e via regressão Lasso, no caso do SKU 4) ou seu modelo precursor em termos de fundamentação téorica: Lasso, nos SKU's 2 e 3.

Nota-se em dentre as cinco séries temporais horárias dos SKU's, 80% delas aparentam ser melhor previstas por modelos que realizam seleção de variáveis, empregando modelos estatísticos. A exceção a esta conclusão se dá no SKU 1, cujo melhor modelo segundo a métrica MAE, a maioria dos horizontes de previsão, se dá no emprego do modelo RegreDin(1).

Ressalta-se esta predominância dos modelos Lassso e variações do AdaLasso como melhores modelos via MAE, pois apesar do método de *machine learning random forest* (RF) também apresentar a propriedade de seleção de variáveis, esta não se mostrou capaz de superar a dos modelos previamente enunciados.

Além disso, acrescenta-se o fato de que por meio desse modelos estatísticos obtém-se não só uma boa acurácia, como também um ganho de interpretabilidade frente ao RF. Um exemplo disto, seria avaliar a variação percentual da quantidade vendida dada uma variação percentual das variáveis explicativas selecionadas (por exemplo, número de visitas ou oscilações de preço próprio e dos concorrentes).

Isto pois a menos das *dummies*, todos os regressores estão numa relação log-log com a variável dependente, implicando em coeficientes estimados que indiretamente estimam a elasticidade entre tais variáveis.

8 Considerações finais

Feita a conclusão deste estudo faz-se em sequência a abordagem das considerações finais a respeito deste trabalho.

No que tange a estimação do modelos estatísticos e métodos de *machine learning*, considera-se para trabalhos futuros retomar estas séries temporais de quantidade vendida e analisá-las sob as estimações de outros modelos. Dentre estes, tem-se:

- ***Score-driven models***: a partir das pesquisas feitas a respeito desses modelos, vê-se um potencial preditivo competitivo para estas séries temporais, evidenciado por outros trabalhos vinculados ao setor de varejo, como em (SARLO; FERNANDES; BORENSTEIN, 2023).
- **Métodos de *deep learning***: *Recurrent Neural Networks* (RNN) e *Echo State Networks* (ESN). Publicações recentes apontam resultados promissores da aplicação destes métodos em problemas de previsão de séries temporais, como em (SHAHI; FENTON; CHERRY, 2022) e (SALAMAI; AGEELI; EL-KENAWY, 2022).
- **Outras variações da regressão Lasso**: retomando os resultados da Tabela 6, verifica-se que quatro dentre as cinco séries temporais analisadas foram melhor previstas pelos modelos Lasso ou AdaLasso. Portanto, considera-se outros modelos da mesma família, como *sparse-group Lasso* em (BABII; GHYSELS; STRIAUKAS, 2022).

Referências bibliográficas

ABCOMM. **Previsão de vendas no e-Commerce para os Próximos 5 anos**. 2023. Acessado em 19/04/2023. Disponível em: ⟨https://dados.abcomm.org/previsao-de-vendas-online⟩.

BABII, A.; GHYSELS, E.; STRIAUKAS, J. Machine learning time series regressions with an application to nowcasting. **Journal of Business & Economic Statistics**, Taylor & Francis, v. 40, n. 3, p. 1094–1106, 2022.

BANDARA, K.; SHI, P.; BERGMEIR, C.; HEWAMALAGE, H.; TRAN, Q.; SEAMAN, B. Sales demand forecast in e-commerce using a long short-term memory neural network methodology. In: SPRINGER. **Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part III 26**. [S.l.], 2019. p. 462–474.

BRANDÃO, M. S.; GODINHO-FILHO, M.; JUNIOR, W. A.; BATTISSACCO, B. C.; MARÇOLA, J. A. Melhoria da categorização de produtos a partir do uso de algoritmos de aprendizado de máquina e medidas de similaridade. **Revista Produção Online**, v. 21, n. 4, p. 2093–2124, 2021.

BREIMAN, L. Bagging predictors. **Machine learning**, Springer, v. 24, p. 123–140, 1996.

BREIMAN, L.; FRIEDMAN, J.; STONE, C.; OLSHEN, R. **Classification and Regression Trees**. Taylor & Francis, 1984. ISBN 9780412048418. Disponível em: ⟨https://books.google.com.br/books?id=JwQx-WOmSyQC⟩.

BREIMANN, L. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001.

FAN, J.; LI, R. Variable selection via nonconcave penalized likelihood and its oracle properties. **Journal of the American statistical Association**, Taylor & Francis, v. 96, n. 456, p. 1348–1360, 2001.

FERNANDES, D. **Vendas no e-commerce batem marca de R$ 118,6 bilhões no 1º semestre no Brasil**. 2022. Acessado em 24/04/2023. Disponível em: ⟨https://www.ecommercebrasil.com.br/noticias/vendas-e-commerce-bilhoes-semestre-brasil⟩.

FILDES, R.; MA, S.; KOLASSA, S. Retail forecasting: Research and practice. **International Journal of Forecasting**, Elsevier, v. 38, n. 4, p. 1283–1318, 2022.

FISHER, M.; RAMAN, A. Using data and big data in retailing. **Production and Operations Management**, Wiley Online Library, v. 27, n. 9, p. 1665–1669, 2018.

GARCIA, M. G.; MEDEIROS, M. C.; VASCONCELOS, G. F. Real-time inflation forecasting with high-dimensional models: The case of brazil. **International Journal of Forecasting**, Elsevier, v. 33, n. 3, p. 679–693, 2017.

GEURTS, M. D.; KELLY, J. P. Forecasting retail sales using alternative models. **International Journal of Forecasting**, Elsevier, v. 2, n. 3, p. 261–272, 1986.

GOEHRY, B.; YAN, H.; GOUDE, Y.; MASSART, P.; POGGI, J.-M. Random forests for time series. 2021.

HENDRY, D. F. **Dynamic econometrics**. [S.l.]: Oxford university press, 1995.

IBGE. **Vendas no varejo caem 16,8% em abril, pior resultado em 20 anos: Agência de Notícias**. 2020. Disponível em: ⟨https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/27963-vendas-no-varejo-caem-16-8-em-abril-pior-resultado-em-20-anos⟩.

KOOPMANS, L. H. **The spectral analysis of time series**. [S.l.]: Elsevier, 1995.

LIU, R. Y.; SINGH, K. et al. Moving blocks jackknife and bootstrap capture weak dependence. **Exploring the limits of bootstrap**, v. 225, p. 248, 1992.

MEDEIROS, M. C.; VASCONCELOS, G. F.; VEIGA, Á.; ZILBERMAN, E. Forecasting inflation in a data-rich environment: the benefits of machine learning methods. **Journal of Business & Economic Statistics**, Taylor & Francis, v. 39, n. 1, p. 98–119, 2021.

OLIST; SIONEK, A. **Brazilian E-Commerce Public Dataset by Olist**. Kaggle, 2018. Disponível em: ⟨https://www.kaggle.com/dsv/195341⟩.

PINHO, J. M.; OLIVEIRA, J. M.; RAMOS, P. Sales forecasting in retail industry based on dynamic regression models. In: **Advances in Manufacturing Technology XXX**. [S.l.]: IOS Press, 2016. p. 483–488.

POLITIS, D. N.; ROMANO, J. P. **A circular block-resampling procedure for stationary data**. [S.l.]: Purdue University. Department of Statistics, 1991.

SALAMAI, A. A.; AGEELI, A. A.; EL-KENAWY, E.-S. M. Forecasting e-commerce adoption based on bidirectional recurrent neural networks. **Computers, Materials & Continua**, v. 70, n. 3, p. 5091–5106, 2022.

SANTOS, J.; ROSSI, R. Aprendizado de máquina não supervisionado baseado em redes heterogêneas para agrupamento de textos. In: SBC. **Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.], 2020. p. 35–46.

SARLO, R.; FERNANDES, C.; BORENSTEIN, D. Lumpy and intermittent retail demand forecasts with score-driven models. **European Journal of Operational Research**, Elsevier, v. 307, n. 3, p. 1146–1160, 2023.

SHAHI, S.; FENTON, F. H.; CHERRY, E. M. Prediction of chaotic time series using recurrent neural networks and reservoir computing techniques: A comparative study. **Machine learning with applications**, Elsevier, v. 8, p. 100300, 2022.

SINGH, K.; BOOMA, P.; EAGANATHAN, U. E-commerce system for sale prediction using machine learning technique. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [S.l.], 2020. v. 1712, n. 1, p. 012042.

SIQUEIRA, N. R. M. **Utilização de aprendizado de máquina para classificação de perfis de consumo de energia elétrica nas diferentes regiões do Brasil**. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Norte, 2023.

TAIEB, S. B.; BONTEMPI, G.; ATIYA, A. F.; SORJAMAA, A. A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition. **Expert systems with applications**, Elsevier, v. 39, n. 8, p. 7067–7083, 2012.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, Oxford University Press, v. 58, n. 1, p. 267–288, 1996.

WITTEN, D.; JAMES, G. **An introduction to statistical learning with applications in R**. [S.l.]: springer publication, 2013.

WOOLDRIDGE, J. **Introductory Econometrics: A Modern Approach**. Cengage Learning, 2015. ISBN 9781473754393. Disponível em: ⟨https://books.google.com.br/books?id=HveHAQAACAAJ⟩.

ZOU, H. The adaptive lasso and its oracle properties. **Journal of the American statistical association**, Taylor & Francis, v. 101, n. 476, p. 1418–1429, 2006.

ZOU, H.; HASTIE, T.; TIBSHIRANI, R. On the “degrees of freedom” of the lasso. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 35, n. 5, p. 2173 – 2192, 2007. Disponível em: ⟨https://doi.org/10.1214/009053607000000127⟩.

Apêndice A - Séries horárias dos SKU's

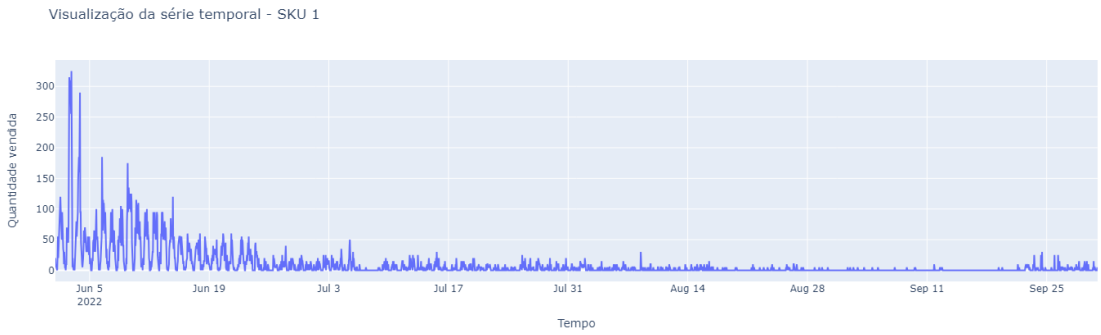

Figura 17: Série horária da quantidade vendida do SKU 1.

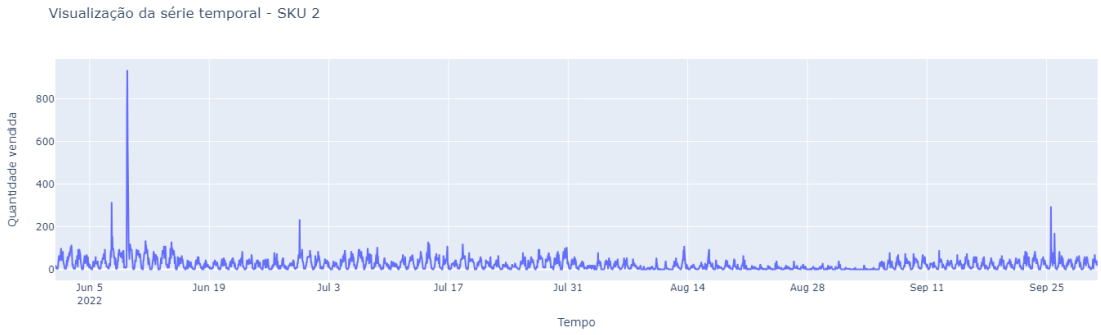

Figura 18: Série horária da quantidade vendida do SKU 2.

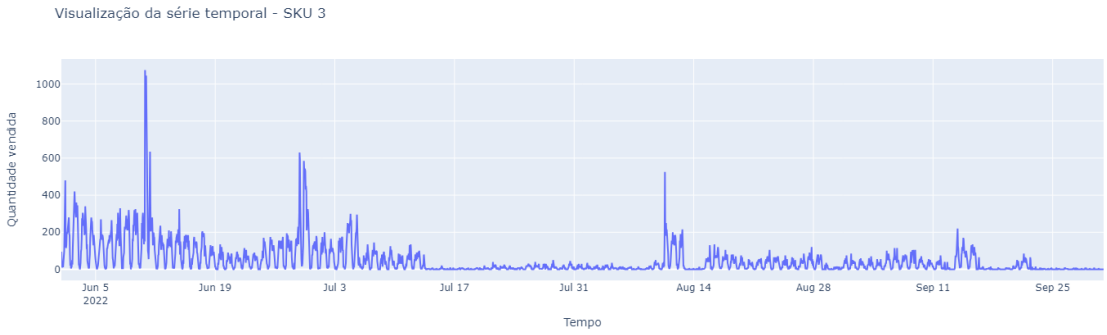

Figura 19: Série horária da quantidade vendida do SKU 3.

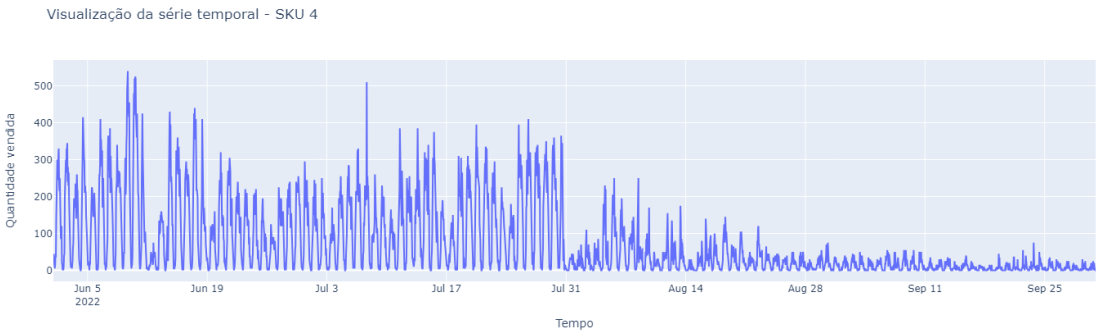

Figura 20: Série horária da quantidade vendida do SKU 4.

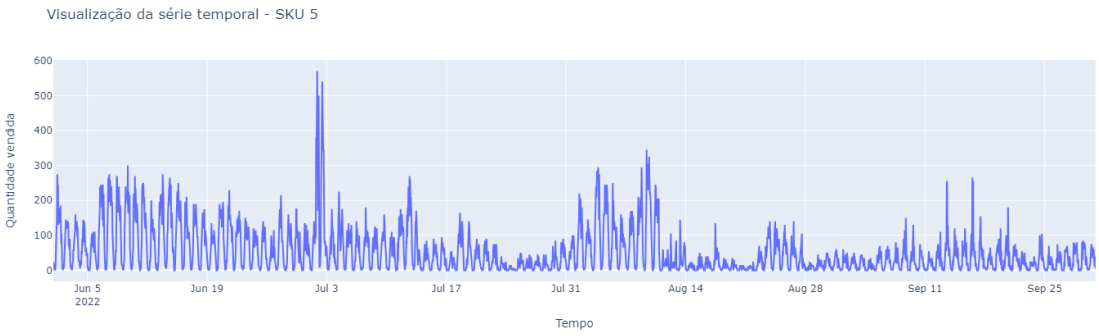

Figura 21: Série horária da quantidade vendida do SKU 5.