UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Análise de Modelos Para Previsão de Séries Temporais Aplicado à Vendas no Segmento de Varejo

Arthur Ghiberti Polskih

Conclusion Course Paper to the Masters of Business Administration Program in Data Sciences

ICMC USP
SÃO CARLOS

Arthur Ghiberti Polskih

Análise de Modelos Para Previsão de Séries Temporais Aplicado à Vendas no Segmento de Varejo

Monografia apresentada ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados. *VERSÃO FINAL*

Área de Concentração: Ciências de Dados

Orientador: Prof. Dr. Francisco Aparecido Neves

USP – São Carlos
Janeiro de 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

GHIBERTI POLSKIH, ARTHUR
G788a Análise de Modelos Para Previsão de Séries Temporais Aplicado à Vendas no Segmento de Varejo / ARTHUR GHIBERTI POLSKIH; orientador Francisco Aparecido Neves. -- São Carlos, 2024.
62 p.

Trabalho de conclusão de curso (MBA em Ciência de Dados) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. . I. Aparecido Neves, Francisco, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Arthur Ghiberti Polskih

Time Series Models Analysis and Prediction Applied to Sales in Retail

Final Paper submitted to the Center for Mathematical Sciences Applied to Industry of the Institute of Mathematics and Computer Sciences – USP, in partial fulfillment of the requirements for the MBA in Data Science. *FINAL VERSION*

Concentration Area: Data Science

Advisor: Prof. Dr. Francisco Aparecido Neves

USP – São Carlos
January 2024

AGRADECIMENTOS

Agradeço e dedico esse trabalho de conclusão de curso aos meus pais, minha família, ao professor Francisco Aparecido Rodrigues e a todos os professores deste curso.

"The future influences the present just as much as the past."

Friedrich Nietzsche

RESUMO

POLSKIH, ARTHUR GHIBERTI. **Análise de Modelos Para Previsão de Séries Temporais Aplicado à Vendas no Segmento de Varejo** . 2024. 62 p. Monografia (MBA em Ciências de Dados) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Esta monografia explora técnicas de preparação, desenvolvimento e análise de dados em modelos de previsão de séries temporais com foco em vendas semanais. Utilizando o banco de dados *Walmart - Store Sales Forecasting*, este trabalho visou identificar os fatores que influenciam as vendas, especialmente durante feriados. Modelos como *ARIMA, SARIMA* e Suavização Exponencial foram estudados e comparados através de métricas de acurácia para avaliar sua eficácia na previsão. A abordagem metódica permitiu a integração e limpeza dos dados, garantindo confiabilidade nas análises. Alguns desafios encontrados, como a interpretação de dados anonimizados e limitações temporais dos dados, foram identificados e as técnicas adotadas são extensíveis para outros contextos de séries temporais no ambiente de negócios.

Palavras-chave: Séries Temporais, Previsão de Vendas, Aprendizado de máquina.

ABSTRACT

POLSKIH, ARTHUR GHIBERTI. **Time Series Models Analysis and Prediction Applied to Sales in Retail**. 2024. 62 p. Monografia (MBA em Ciências de Dados) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

This work seeks to investigate data preparation, development, and analysis techniques in time series forecasting models, focusing on weekly sales. Utilizing the *Walmart - Store Sales Forecasting* database from the *Kaggle* platform, the study aimed to pinpoint the factors affecting sales, particularly during holidays. Models such as *ARIMA, SARIMA*, and Exponential Smoothing were examined and compared using accuracy metrics to assess their predictive capabilities. The methodological approach enabled the integration and cleaning of data, ensuring reliable analyses. Challenges like interpreting anonymized data and temporal limitations of data were addressed, and the adopted techniques are scalable to other time series contexts business problems.

Keywords: Time Series, Sales Forecasting, Machine learning.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Visando aplicar os conhecimentos ensinados durante o curso de MBA em Ciências de Dados, esta monografia irá explorar técnicas de preparação, desenvolvimento e análise de dados e modelos de previsão de séries temporais.

1.1 Objetivos

O objetivo principal deste estudo é explorar e avaliar uma variedade de modelos de previsão de séries temporais. O banco de dados utilizado para os propósitos deste trabalho é o *Walmart - Store Sales Forecasting* disponibilizado pela empresa na plataforma online *Kaggle* [1]. A empresa disponibilizou este banco de dados, com o objetivo de analisar os fatores que impactam as vendas semanais de sua rede de lojas, principalmente em feriados e eventos esportivos.

Especificamente, modelos como ARIMA, SARIMA, e suavização exponencial serão examinados em detalhes. A comparação entre esses modelos será realizada com base em métricas de acurácia, permitindo uma análise da precisão e confiabilidade de cada modelo.

O presente estudo não apenas busca identificar modelos adequados para previsão

[1] A base original pode ser acessada através do link: <https://www.kaggle.com/competitions/walmart-recruiting-store-sales-forecasting/>.

deste banco de dados, mas também visa contribuir para a compreensão de como diferentes abordagens modelam e interpretam as complexidades associadas às séries temporais no contexto comercial.

CAPÍTULO

2

REVISÃO BIBLIOGRÁFICA

Este capítulo oferece uma revisão concisa, abordando estudos de caso, monografias e perspectivas históricas relacionadas a séries temporais e aprendizado de máquina. O objetivo principal é fornecer uma base teórica e prática que sustente e valide a metodologia adotada neste estudo.

2.1 Contextualização

2.1.1 Introdução à Previsão

A habilidade de previsão tem fascinado pessoas há milhares de anos. Por vezes sendo considerada como uma inspiração divina, e outras como uma atividade criminosa. A previsibilidade de um evento depende de certos fatores que incluem: a compreensão dos fatores que irão contribuir para seu resultado, quantos dados estão disponíveis e se as previsões podem afetar o próprio resultado que estamos tentando prever (HYNDMAN; ATHANASOPOULOS, 2018).

2.1.2 Séries Temporais

A série temporal pode ser vista como uma lista de números, juntamente com algumas informações temporais que indicam quando esses números foram registrados.

Para uma série temporal, observações vizinhas estão correlacionadas. Se em modelos de regressão a ordem das observações não importa, em modelos de séries temporais a ordem dos dados é crucial (HYNDMAN; ATHANASOPOULOS, 2018).

Os modelos utilizados para descrever séries temporais são processos estocásticos, isto é, processos controlados por leis probabilísticas. Qualquer que seja a classificação que façamos para os modelos de séries temporais, podemos considerar um número muito grande de modelos diferentes para descrever o comportamento de uma série particular (MORETTIN; TOLOI, 2018).

Os dois modelos mais populares para a previsão de séries temporais são a Suavização Exponencial, que se baseia em uma descrição da tendência e sazonalidade dos dados, e o *ARIMA (autoregressive integrated moving average),* que tenta descrever as autocorrelações dos dados (HYNDMAN; ATHANASOPOULOS, 2018).

Vale ressaltar que os métodos de Suavização Exponencial e *ARIMA* são técnicas "caixa-preta", pois são construídos sob a suposição de que os padrões históricos na série temporal continuarão se repetindo no futuro. No entanto, abordagens baseadas em regressão assumem que o comportamento da série temporal de interesse (variável resposta ou dependente) é influenciado por outras variáveis (variáveis independentes), sendo explorado por meio de regressão linear para possivelmente criar previsões mais precisas (ZEMKOHO, 2022).

Um estudo que pode ser destacado da aplicação de modelos *ARMA (Autoregressive Moving Average)* no varejo De Livera e seu grupo (LIVERA; HYNDMAN; SNYDER, 2011), exploram o uso do modelo em 3 diferentes contextos. Neste estudo, os modelos demonstraram sua capacidade de capturar a sazonalidade e tendências em aplicações de negócios como vendas semanais de gasolina nos Estados Unidos da América, demanda diária de um *call center*, assim como a demanda de eletricidade em feriados na Túrquia.

As técnicas de limpeza de dados são amplamente reconhecidas na literatura de ciências de dados. Em seu livro (BROWNLEE, 2020) destacou a imputação de médias móveis para tratar dados faltantes, uma técnica que não apenas permite a imputação dos dados faltantes, mas preserva a estrutura temporal dos dados. A remoção de duplicatas, crucial para a integridade dos dados, foi enfatizada por Tsay (TSAY, 2005) ao analisar

tendências financeiras, onde observações duplicadas poderiam levar a conclusões erradas. Além disso, ele também destacou a eliminação de valores improváveis, como preços negativos em dados financeiros, ressaltando a necessidade de dados realistas para modelagens confiáveis. Esses exemplos ilustram a relevância dessas técnicas de limpeza em diferentes contextos, sublinhando a importância de uma abordagem metódica e cuidadosa na preparação de dados para análises de séries temporais.

CAPÍTULO

3

METODOLOGIA

3.1 Metodologia

Nesta seção serão apresentadas as metodologias utilizadas para cumprir com os objetivos descritos. As análises foram desenvolvidas na linguagem de programação *Python*, e utilizando os pacotes necessários para efetuar a limpeza dos dados, análise exploratória e criação dos modelos. As etapas são orientadas com base no modelo *CRIPS-DM (ou Cross Industry Standard Process for Data Mining)* (SHEARER, 2000), uma metodologia comumente utilizada em problemas de ciência de dados, com as devidas alterações para compreender o objetivo deste estudo e o problema de séries temporais em questão.

A metodologia será desenvolvida com as seguintes fases:

1. Conhecimento do Negócio
2. Leitura e Entendimento dos Dados
3. Limpeza dos Dados
4. Avaliação de estacionariedade
5. Análise Exploratória dos Dados (EDA)

6. Feature Selection
7. Feature Engineering
8. Definição e Treino dos Modelos
9. Avaliação dos Modelos

3.1.1 Conhecimento do Negócio

Para o desenvolvimento deste estudo, a compreensão do contexto de negócios envolve a análise de dados de vendas do *Walmart*[1], uma das maiores cadeias de varejo nos Estados Unidos da América. Este estudo utilizou dados disponíveis para 45 lojas, explorando seus padrões de vendas. O desafio enfrentado pela rede varejista, caracterizado por variações inesperadas na demanda em datas de feriado, o que pode trazer problemas ocasionais de falta de estoque, destaca a necessidade de um algoritmo de aprendizado de máquina eficaz e capaz de lidar com essa variabilidade. (ADMIN, 2014)

Eventos promocionais, especialmente em períodos que antecedem feriados significativos, também podem apresentar impacto na previsão de vendas. Estes incluem o *Super Bowl*, Dia do Trabalho, Ação de Graças e Natal. Essa análise apresenta desafios à falta de dados históricos completos.

3.1.2 Leitura e Entendimento dos Dados

A fase de leitura dos dados constitui a base para a análise deste estudo. Envolve a coleta e o entendimento detalhado da origem e das variáveis dos dados fornecidos, assim como entender a necessidade de limpeza, criação de novas variáveis e avaliar a quantidade de informações disponíveis. A capacidade de identificar e utilizar adequadamente esses dados é o primeiro passo para desenvolver modelos que podem capturar com precisão as complexidades das vendas semanais.

O banco de dados fornecido pela rede de varejo compreende quatro arquivos principais, que serão explorados na seção de resultados.

[1] A base original pode ser acessada através do link: <https://www.kaggle.com/competitions/walmart-recruiting-store-sales-forecasting/>.

3.1.3 Limpeza dos Dados

No processo de limpeza dos dados, o objetivo é assegurar que o conjunto de dados seja preciso, completo e confiável, livre de erros e inconsistências. No contexto deste estudo, a limpeza de dados envolve várias etapas para garantir a qualidade e a integridade dos dados antes da modelagem.

A primeira etapa da limpeza dos dados é lidar com valores ausentes. Técnicas como imputação média, mediana ou médias móveis são frequentemente aplicadas, dependendo do tipo e da distribuição dos dados. Em contextos onde a imputação pode distorcer a análise, pode-se considerar a exclusão de linhas ou colunas com muitos valores ausentes.

Outliers, ou valores discrepantes, podem distorcer significativamente os resultados da análise. Métodos estatísticos como a análise de *boxplot* são comumente empregados. No caso deste estudo, a identificação de vendas, ou outras colunas como temperatura ou combustível, anormalmente altas ou baixas pode ajudar a refinar os modelos preditivos. A normalização (escalamento de dados em um intervalo entre dois valores) ou a padronização (transformação de dados para ter uma média e um desvio padrão com valores pré-determinados) é essencial para modelos que são sensíveis à escala das variáveis. Essas técnicas ajudam a garantir que as variáveis tenham peso equivalente nos modelos analíticos.

Inconsistências nos dados, como formatos de data variados ou categorias mal rotuladas, são corrigidas nesta fase. Esta fase também inclui avaliar o tipo dos dados em cada coluna, como por exemplo números inteiros, texto ou datas, e garantir que todos os valores estão tipificados corretamente.

Documentar cada passo do processo de limpeza de dados é considerado uma boa prática. Isso não apenas fornece transparência e reprodutibilidade, mas também ajuda a rastrear as modificações feitas no conjunto de dados original.

3.1.4 Avaliação de Estacionariedade

No contexto da modelagem de séries temporais, a avaliação da estacionariedade é uma etapa que precede a aplicação de técnicas avançadas de análise e predição. O teste

de *Dickey-Fuller Aumentado* (DICKEY; FULLER, 1979) é uma ferramenta estatística que pode ser aplicada neste processo. Este teste verifica a presença de uma raiz unitária na série temporal, que é um indicativo de não estacionariedade. Este teste opera sob a hipótese nula de que a série é não estacionária. Matematicamente, o teste analisa a significância do coeficiente na regressão:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \cdots + \delta_p \Delta y_{t-p} + \varepsilon_t \tag{3.1}$$

Onde:

- Δ é o operador de diferença,
- y_t é a série temporal no tempo t,
- α é o termo constante,
- βt é o coeficiente de uma tendência de tempo linear,
- γ é o coeficiente em y_{t-1},
- $\delta_1, \delta_2, \ldots, \delta_p$ são os coeficientes de atraso,
- ε_t é o termo de erro independente.

Se o valor p associado ao teste for inferior ao nível de significância de interesse, rejeita-se a hipótese nula, concluindo que a série é estacionária. Caso contrário, a série é considerada não estacionária. Quando uma série temporal é identificada como não estacionária, técnicas como a dessazonalização e a diferenciação tornam-se essenciais para transformá-la em uma forma estacionária. A dessazonalização envolve a remoção de padrões periódicos, como flutuações decorrentes de eventos sazonais ou cíclicos, enquanto a diferenciação — a subtração de uma observação pela sua anterior — visa eliminar tendências ou padrões de longo prazo. Essas técnicas podem ser adotadas porque diversos modelos de séries temporais, incluindo o modelo *ARIMA*, pressupõem que a série seja estacionária.

A diferenciação, em particular, pode ser aplicada através da função *diff()* da biblioteca *pandas*[2], onde cada valor da série é substituído pela diferença entre ele e seu valor anterior. Isso ajuda a estabilizar a média da série, permitindo que o modelo se concentre em flutuações significativas em torno de uma média constante.

3.1.5 Análise Exploratória dos Dados (EDA)

A Análise Exploratória dos Dados (EDA) proporciona ideias iniciais sobre qualquer conjunto de dados. A visualização é um componente integral da EDA. Gráficos como histogramas, *boxplots* e gráficos de dispersão serão utilizados para explorar a distribuição, variabilidade e relações entre as variáveis. Essas técnicas ajudam a identificar variáveis como a distribuição de vendas ao longo do tempo, a presença de *outliers* e a correlação entre variáveis como vendas e fatores econômicos. A análise de correlação é fundamental para entender as relações entre diferentes variáveis, e o uso de matrizes de correlação ajudará a identificar variáveis que têm forte correlação com as vendas semanais, um passo importante na seleção de variáveis para modelagem.

Dada a natureza temporal dos dados de vendas, técnicas de decomposição de séries temporais, como a análise *STL (decomposição de Tendência Sazonal usando LOESS)*, serão empregadas para desagregar as vendas em componentes de tendência, estacionariedade e resíduo. Isso pode proporcionar uma compreensão mais clara dos padrões subjacentes nas vendas, assim como ajudar a entender melhor como as vendas variam durante o ano e em diferentes períodos (CLEVELAND *et al.*, 1990).

As descobertas da EDA terão implicações diretas sobre as etapas subsequentes de modelagem preditiva. Compreender as variáveis dos dados guiará a seleção de modelos e técnicas a serem aplicadas, assegurando que os modelos sejam baseados em premissas válidas e dados confiáveis.

3.1.6 Feature Selection

Na etapa de seleção de variáveis (ou *Feature Selection*), o objetivo é identificar e selecionar as variáveis mais relevantes que influenciam a variável de resposta. Essa sele-

[2] A documentação original da biblioteca *pandas* pode ser acessada através do seguinte link: <https://pandas.pydata.org/docs/>.

ção tenta melhorar a eficiência dos modelos e garantir que eles não sejam sobrecarregados com dados irrelevantes ou redundantes.

Técnicas como análise de correlação e importância de variáveis podem ser aplicadas inicialmente, a análise de correlação pode auxiliar a identificar relações lineares entre variáveis, enquanto a importância de variáveis, obtida por meio de modelos como árvores de decisão, pode revelar quais variáveis têm maior impacto nas vendas.

Técnicas de redução de dimensionalidade, como Análise de Componentes Principais (PCA), podem ser exploradas para condensar informações em um número menor de variáveis explicativas, mantendo a maior parte da informação original. Esta abordagem pode ser particularmente útil se o conjunto de dados apresentar multicolinearidade.

Complementando a *Feature Selection*, a próxima etapa de *Feature Engineering* envolverá a criação de novas variáveis que podem melhorar a capacidade preditiva dos modelos. Este processo será discutido na próxima fase da metodologia. A seleção final de variáveis será validada através de técnicas de validação cruzada e sua contribuição para a melhoria do desempenho do modelo.

3.1.7 Feature Engineering

Feature Engineering, ou Engenharia de Variáveis, é uma etapa que visa transformar ou criar novas variáveis a partir dos dados brutos para melhorar a performance dos modelos. Uma abordagem comum, especialmente para dados de séries temporais, é a criação de variáveis baseadas em tempo. Isso inclui a decomposição da data em dia, mês, ano e outras unidades temporais relevantes.

Para a *Feature Engineering*, técnicas como codificação *One-Hot* podem ser aplicadas para transformar variáveis categóricas em um formato que possa ser efetivamente utilizado por modelos preditivos. Além disso, a criação de interações entre variáveis, onde novas variáveis são formadas a partir da combinação de duas ou mais variáveis, pode revelar relações complexas e não lineares que são importantes para prever as vendas.

As novas variáveis geradas serão validadas por meio de técnicas como validação cruzada de *Feature Selection*. Este processo assegura que as variáveis adicionadas ao modelo contribuem positivamente para o modelo.

3.1.8 Definição e Treino dos Modelos

A modelagem é a etapa onde selecionamos os modelos para treinar e gerar as previsões. No contexto deste estudo, a modelagem foca em identificar os modelos mais adequados para prever as vendas semanais. Dada a natureza temporal dos dados de vendas, modelos de séries temporais como *ARIMA* (*Autoregressive Integrated Moving Average*), *ARIMA* (*Seasonal Autoregressive Integrated Moving Average*) e Suavização Exponencial serão explorados.

O *Auto ARIMA*, uma extensão do *ARIMA*, será utilizado para automatizar o processo de seleção dos melhores parâmetros para o modelo *ARIMA*. Esta extensão é particularmente útil para conjuntos de dados complexos, onde a seleção manual de parâmetros pode ser desafiadora.

O modelo de Suavização Exponencial, também será explorado, e é reconhecido por sua simplicidade e eficácia, especialmente em dados com padrões sazonais claros.

Todos os modelos serão validados contra uma porção dos dados de teste, utilizando métricas que serão discutidas na próxima seção de Avaliação dos Modelos.

3.1.9 Avaliação dos Modelos

A Avaliação dos Modelos é a fase onde analisamos as métricas para obter uma visão quantitativa sobre a eficácia dos modelos preditivos. A seleção de métricas apropriadas é essencial para avaliar a precisão e a confiabilidade das previsões.

3.1.9.1 Erro Quadrático Médio (RMSE)

O RMSE (*Root Mean Square Error*), ou Erro Quadrático Médio, é uma das métricas mais comuns para avaliar modelos de regressão. Matematicamente, é definido como a raiz quadrada da média dos quadrados dos erros. A fórmula do RMSE é:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{3.2}$$

Onde:

- N é o número de observações,

- y_i é o valor observado no tempo i,
- $\hat{y}_i$ é o valor predito para o tempo i,

O RMSE é particularmente útil para quantificar o erro em unidades iguais às da variável dependente, tornando a sua interpretação simples e aplicada sobre o resultado.

3.1.9.2 Erro Médio Absoluto (MAE)

O MAE (*Mean Absolute Error*), ou Erro Médio Absoluto, mede a média das magnitudes dos erros entre as previsões e as observações. É expresso matematicamente como:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3.3}$$

Onde:

- N é o número de observações,
- y_i é o valor observado no tempo i,
- $\hat{y}_i$ é o valor predito para o tempo i,

A vantagem do MAE sobre o RMSE, é que ele é menos sensível a *outliers*, mas possui uma interpretaçào menos simples para públicos não técnicos.

3.1.9.3 Erro Percentual Médio Absoluto (MAPE)

O MAPE (*Mean Absolute Percentage Error*), ou Erro Percentual Médio Absoluto é uma medida do Erro Médio Absoluto como uma porcentagem. É calculado como:

$$\text{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3.4}$$

Onde:

- N é o número de observações,

- y_i é o valor observado no tempo i,
- $\hat{y}_i$ é o valor predito para o tempo i,

O MAPE é útil para compreender o erro em termos relativos, tornando-o intuitivo, especialmente para públicos não técnicos. Esta foi a métrica de avaliação oficial utilizada pelo *Walmart* na competição onde disponibilizou estes dados.

3.1.9.4 Comparação das Métricas para Avaliação de Modelos

Cada uma dessas métricas oferece informações sobre o desempenho do modelo. Enquanto o RMSE é mais sensível a *outliers*, o MAE fornece uma medida mais conservadora do erro. O MAPE, por outro lado, oferece uma visão percentual que é fácil de interpretar e comunicar. Para a avaliação de modelos de séries temporais, essas métricas fornecerão uma compreensão clara de como os modelos performam contra os valores de teste.

CAPÍTULO

4

RESULTADOS

4.1 Considerações Iniciais

Nesta seção serão exploradas uma série de modelos preditivos aplicados aos dados de vendas do *Walmart*. A análise exploratória e a modelagem estatística são empregadas para identificar padrões e fatores que influenciam as vendas semanais nas lojas. Esta investigação segue o roteiro metodológico estabelecido na fase de metodologia do estudo.

4.2 Leitura e Compreensão dos Dados

A primeira etapa dos resultados compreende a leitura e compreensão dos dados. O banco de dados fornecido pela rede varejista[1] possui quatro arquivos principais.

4.2.1 Leitura dos Dados

O arquivo *stores.csv* contém informações anonimizadas sobre 45 lojas, indicando o seu tipo e tamanho. As colunas incluem:

[1] A base original pode ser acessada através do link: <https://www.kaggle.com/competitions/walmart-recruiting-store-sales-forecasting/>.

- Store: Número identificador da loja, com valores entre 1 e 45;
- Type: Tipo da loja, separado nas categorias A, B ou C;
- Size: Tamanho da loja, com valores entre 34,9 mil a 220 mil.

O arquivo *features.csv* contém informações adicionais sobre as 45 lojas. As colunas incluem:

- Store: Número identificador da loja, com valores entre 1 e 45;
- Date: Valores de data, com intervalo de uma semana;
- Temperature: Temperatura média da região;
- Fuel Price: Custo do combustível na região;
- MarkDown1-5: Dados sobre eventos promocionais anonimizados. Mais informações sobre essas colunas não estão disponíveis;
- CPI: Índice de preços ao consumidor, uma medida que representa a média ponderada de preços de uma cesta de bens de consumo e serviços.

O arquivo *train.csv* contém dados sobre vendas semanais e indicação de feriados, com o objetivo de treinar os modelos de previsão. O arquivo *test.csv* é designado para a validação dos modelos, permitindo testar a precisão das previsões geradas contra os dados reais. Estes dados não vieram acompanhados da variável resposta para as datas de interesse, pois seriam analisados pela própria empresa. Desta forma, faremos a validação do modelo separando o arquivo *train.csv* em porções de treino e teste.

4.2.2 Compreensão dos Dados

O banco de dados apresenta uma variedade de colunas com informações identificáveis, como preço do combustível e taxas de desemprego, além de indicadores de feriados. No entanto, alguns dados estratégicos da empresa, como as variáveis *Markdown* de 1 a 5, são anonimizados, representando um desafio adicional na análise e interpretação.

A primeira manipulação dos dados envolve a consolidação dos arquivos *train.csv, features.csv e stores.csv.* Essa integração é feita com base nas colunas *Store* e *Date*, garantindo que todas as informações relevantes para uma semana de vendas em cada loja estejam presentes em uma única tabela consolidada.

4.3 Limpeza dos Dados

A limpeza dos dados é a etapa de manipulação e preparação dos dados para posterior análise estatística e modelagem. Essa fase envolve garantir a qualidade e a consistência dos dados, além de evitar o fenômeno conhecido como vazamento de dados. O vazamento de dados ocorre quando informações do conjunto de teste são inadvertidamente utilizadas durante o treinamento do modelo, comprometendo a validade das previsões. Para prevenir isso, é fundamental separar os dados em conjuntos de treino e teste, tipicamente entre 70% a 90% dos dados são usados para treinamento, e o restante para validação. Neste estudo, adotou-se uma divisão de 80-20%, mantendo a ordem cronológica dos dados para preservar a integridade temporal, garantindo que os dados de teste sejam a porção mais recente da série.

O conjunto de treino será utilizado não apenas para treinar os modelos, mas também para estabelecer critérios de normalização e outras manipulações estatísticas necessárias. Isso assegura que as análises e ajustes sejam baseados apenas nas informações disponíveis no conjunto de treino, simulando um cenário realista em que os dados de teste, representando dados desconhecidos, não são utilizados na fase de preparação.

O primeiro passo deste processo será identificar valores ausentes, representados por *NaN (Not a Number)*. Diversas técnicas podem ser empregadas para o tratamento de valores ausentes, variando conforme a natureza e a coluna dos dados. No conjunto de dados deste estudo, as colunas *Markdown1* até *Markdown5* apresentaram um número considerável de valores ausentes. Após análise, optou-se por preencher esses valores ausentes com zero. Essa decisão baseia-se na suposição de que a ausência de valores indica semanas sem promoções. Esta abordagem está alinhada com práticas comuns no setor de varejo, onde a não realização de uma promoção é uma ocorrência distinta de valores negativos ou arbitrariamente altos.

O segundo passo de limpeza dos dados é a avaliação de duplicatas nos dados.

Variável	Valores Ausentes
Markdown1	4155
Markdown2	4798
Markdown3	4389
Markdown4	4470
Markdown5	4140

Tabela 1 – Contagem de valores ausentes nas colunas Markdown.

Duplicatas são valores onde uma observação em duas ou mais linhas, possuem valores idênticos em todas as colunas. A análise dos dados revelou que não existem linhas duplicadas no conjunto de dados. Este resultado, é um indicativo positivo da qualidade dos dados fornecidos, e sua ausência simplifica o processo de limpeza e assegura que cada linha de dados representa uma observação única e válida.

A verificação dos tipos de dados neste conjunto de dados, mostra que a maioria das colunas possui tipos de dados apropriados para suas respectivas informações, como inteiros para *Store* e *Size*, e valores de ponto flutuante para *Weekly Sales*, *Temperature*, *Fuel Price*, *MarkDowns*, *CPI* (índice de preços ao consumidor) e *Unemployment*. Foi identificado que as colunas *Date* e *Type* não estavam no formato mais adequado para análise. A coluna *Date* apresentava-se como objeto e foi convertida para o formato de data (*datetime*), enquanto *Type*, também originalmente um objeto, foi transformada em uma categoria numérica. Colunas que possuem mais de um tipo de dados, como por exemplo números inteiros (1) e flutuantes (1.0), recebem essa denominação de objetos automaticamente pelo pacote *Pandas*.

Ao converter *Date* para o formato de data, aproveitou-se a oportunidade para criar novas variáveis temporais. As variáveis *Week* e *Month* foram obtidas a partir da informação da coluna que contém a informação de data, e representam em números inteiros ordinais a semana e o mês do ano vigente. Estas variáveis podem ser capazes de auxiliar o modelo a entender como as vendas semanais em anos diferentes podem ter a influência de sua época do ano.

Após a conversão, foi realizada uma verificação para garantir que as transformações foram aplicadas corretamente.

A identificação e tratamento de valores improváveis também deve ser efetuada

na limpeza de dados. No contexto dos dados deste estudo, uma análise específica foi realizada para verificar a existência de valores de vendas semanais negativos ou outros *outliers* que pudessem indicar erros de registro ou entrada de dados. Para realizar essa avaliação, o método *describe()* do pacote *Pandas* foi utilizado no DataFrame. Este método oferece um resumo estatístico das colunas, fornecendo informações sobre a distribuição dos dados, incluindo média, mediana, mínimo, máximo e quartis.

A análise dos dados revelou que não existem vendas semanais negativas no conjunto de dados, um indicador de que os registros de vendas estão livres de erros óbvios de entrada. No entanto, foram identificados valores negativos nas colunas *MarkDown*, cuja interpretação não é clara devido à natureza anonimizada desses dados. Dada a incerteza sobre o significado exato destas colunas, optou-se por manter esses valores negativos. Esta decisão permite investigar mais profundamente essas variáveis antes de considerar qualquer exclusão ou transformação

4.4 Avaliação de Estacionariedade

Através do teste de *Dickey-Fuller* aumentado, avaliou-se caso a série *Weekly Sales* apresenta estacionariedade, uma propriedade que indica constância nas características estatísticas, como média e variância, ao longo do tempo. A estacionariedade é um indicativo que os modelos não serão influenciados por tendências ou padrões sazonais não removidos, que podem distorcer as previsões. Quando a série não é estacionária, métodos podem ser aplicados para transformá-la em uma forma estacionária, removendo influências de longo prazo e padrões sazonais.

O valor de estatística apresentou um valor de -11.5, sendo menor que os valores críticos em todos os níveis de significância críticos (1%, 5%, 10%), o que indica uma forte evidência contra a hipótese nula de não estacionariedade. O valor p associado ao teste foi inferior ao nível de significância 0,05, apresentando um valor de 4.7×10^{-21} Com estes valores, é possivel rejeitar a hipótese nula, concluindo que a série é estacionária.

Com a variável de interesse *Weekly Sales* já estacionária, prosseguiu-se com a modelagem sem a necessidade de diferenciação adicional.

4.5 Análise Exploratória dos Dados (EDA)

A análise exploratória deste estudo iniciou com a elaboração de uma matriz de correlação, uma ferramenta visual e quantitativa para explorar as relações entre diferentes variáveis. A matriz presente na Figuara 1 revelou correlações variadas. As células da matriz são coloridas em um espectro que varia de -1.0 a 1.0, onde células laranja, ou valores próximos a 1.0, indicam uma forte correlação diretamente proporcional, enquanto células azuis, ou valores próximos a -1.0, representam uma correlação inversamente proporcional forte.

A matriz revelou que muitas variáveis não possuem uma correlação forte entre si, com a maioria dos valores oscilando em torno de zero. Isso sugere que as relações entre as variáveis podem ser mais complexas do que correlações lineares simples e podem necessitar de variáveis adicionais ou de uma abordagem analítica mais sofisticada para serem compreendidas. Especificamente, em relação à variável de interesse, *Weekly Sales*, observamos algumas correlações que se destacam:

- *Weekly Sales x Store*: -0.34
- *Weekly Sales x Type*: -0.59
- *Weekly Sales x Size*: 0.80

Observa-se pouca correlação entre *Weekly Sales* e variáveis como feriados ou colunas *Markdown*. Apesar de a análise do impacto de feriados e promoções ser de grande interesse para a rede de varejo, essa correlação pode não ser tão linear quanto inicialmente presumido.

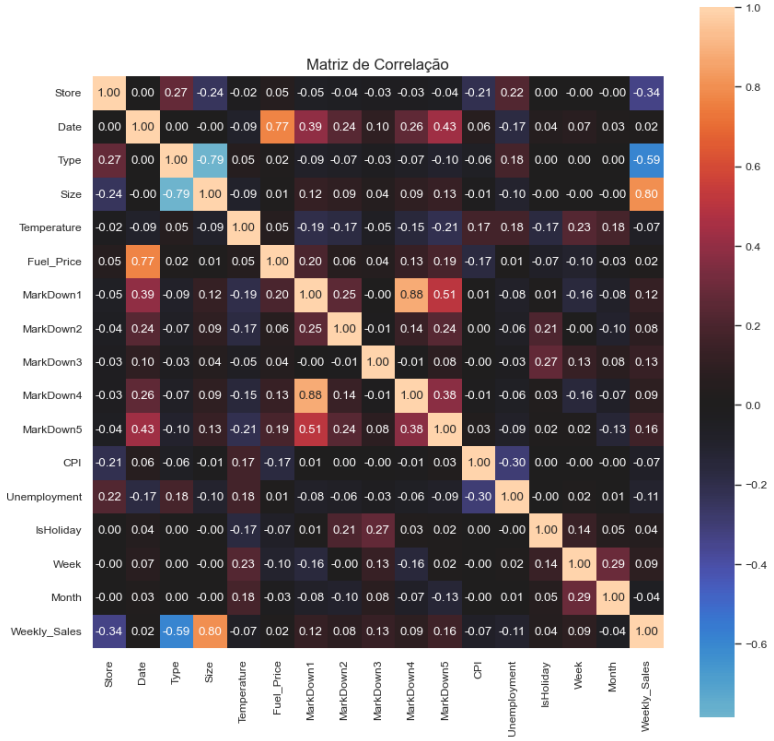


Figura 1 – Matriz de Correlação entre as Variáveis Iniciais.

O gráfico presente na Figura 2 foi elaborado empregando técnicas de agrupamento com base na variável *Week*, que indica o número da semana dentro do ano vigente. Para este gráfico, calculou-se a média das vendas semanais correspondentes ao longo de todo o período disponível no conjunto de dados. A análise deste gráfico revela um padrão relativamente constante e lateralizado de vendas entre as semanas 0 e 45, seguido por um notável pico nas últimas semanas do ano. Considerando que feriados significativos ocorrem neste período, essa tendência ascendente nas vendas pode ser interpretada como um indicativo de impacto desses feriados e uma evidência de sazonalidade durante esta época do ano.

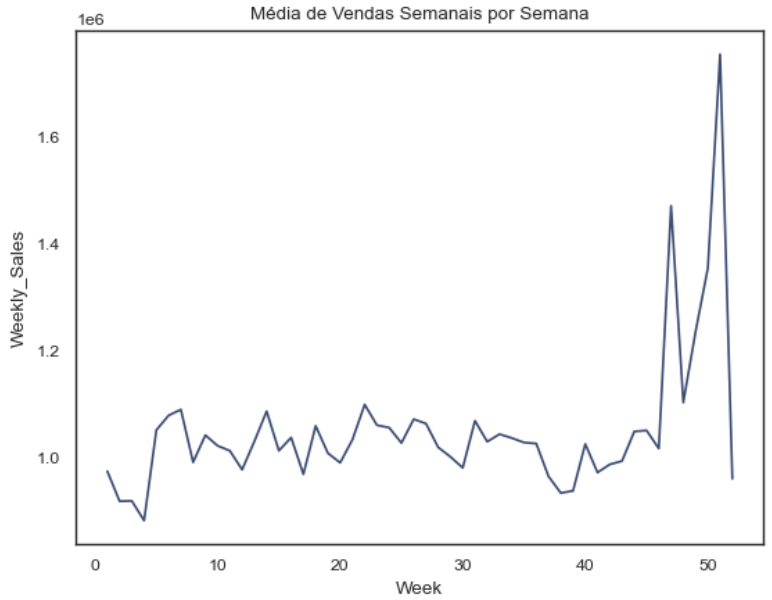


Figura 2 – Média de Vendas Semanais por Semana.

O gráfico presente na Figura 3 foi desenvolvido similarmente ao gráfico anterior, com a soma de vendas semanais agrupada por loja. Ele mostra variações substanciais nas vendas entre as lojas. Algumas lojas apresentam vendas significativamente maiores, o que pode ser atribuído a fatores como localização, tamanho ou tipo de loja. Essa disparidade aponta para a necessidade de modelar as vendas a nível de loja.

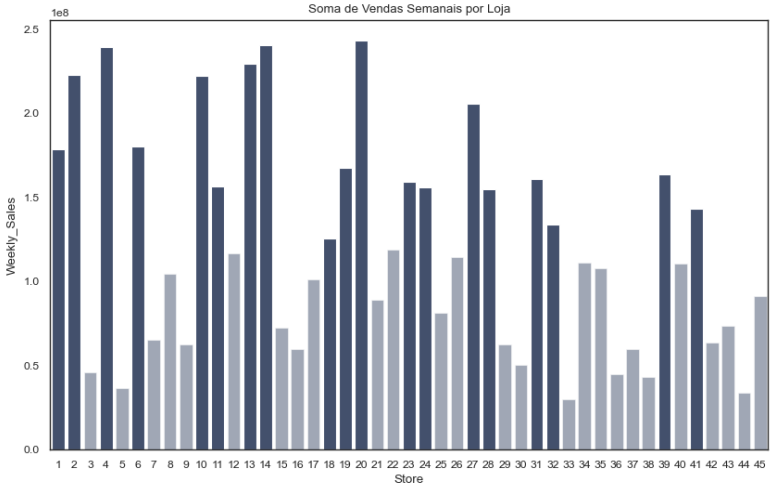


Figura 3 – Soma de Vendas Semanais por Loja.

O gráfico *box plot* presente na Figura 4 revela que as vendas tendem a ser ligeiramente maiores durante as semanas de feriado em comparação com as não feriados. No entanto, a variação nos dados de feriados não é particularmente significativa, o que pode sugerir que nem todos os feriados impactam as vendas igualmente.

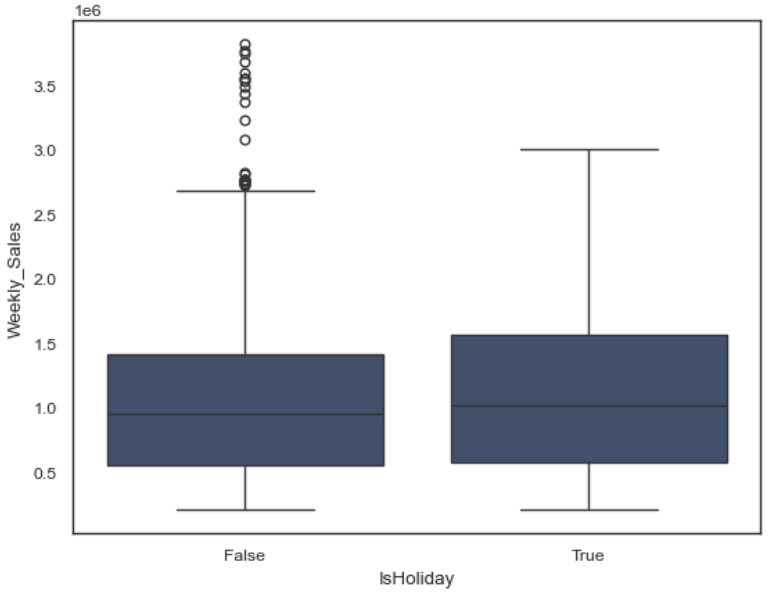


Figura 4 – Relação entre Vendas Semanais e Feriados.

O gráfico de dispersão que relaciona *Weekly Sales* e *Fuel Price*, presente na Figura 5, mostra uma distribuição ampla de pontos sem uma tendência clara. A linha de tendência horizontal sugere que não há uma relação linear significativa entre o preço do combustível e as vendas semanais. Essa falta de correlação pode indicar que o preço do combustível, inicialmente, não é um determinante direto das vendas.

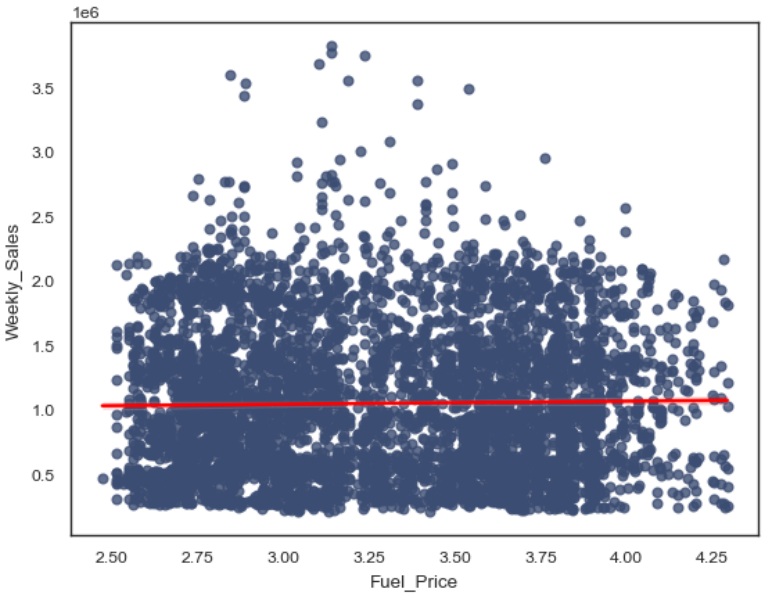


Figura 5 – Dispersão entre Vendas Semanais e Preço do Combustível.

Similarmente, o gráfico na Figura 6 relacionando *Weekly Sales* e *CPI* não exibe uma correlação linear evidente, como representado pela linha de tendência horizontal. Isso implica que variações no CPI não influenciam diretamente as vendas semanais de maneira uniforme. No entanto, é possível notar 3 áreas clusterizadas de CPI, e a inclusão de uma variável categórica, pode auxiliar um modelo a enxergar esta relação de forma mais evidente. Foi criada uma nova coluna categórica do CPI, chamada *CPI Categorical*. Separamos os valores em 3 faixas do mínimo ao máximo, onde cada data receberá um número de 0 a 2 baseado em seu CPI naquela ocasião.

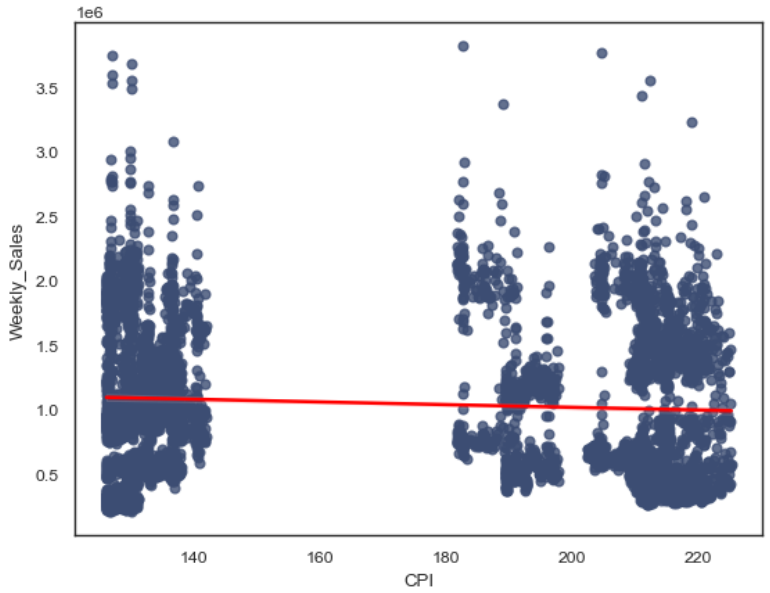


Figura 6 – Dispersão entre Vendas Semanais e CPI.

O gráfico de dispersão entre *Weekly Sales* e *Size*, presente na Figura 7 mostra uma tendência positiva, indicando que lojas maiores tendem a ter vendas maiores. Isso demonstra a capacidade da rede de varejo de avaliar o potencial de vendas em uma certa região, antes de investir em uma loja de tamanho grande, assim como a capacidade de lojas maiores atenderem maiores volumes de vendas. A coluna *Size* apresenta números flutuantes como resultado, por isso, uma nova coluna categórica do tamanho de loja, chamada *Size Categorical* foi criada separando os valores em 4 quartis do mínimo ao máximo, onde cada loja receberá um número de 0 a 3 baseado em seu tamanho. As menores lojas ficarão contidas na categoria 0 de tamanho, enquanto as maiores ficarão contidas na categoria 3.

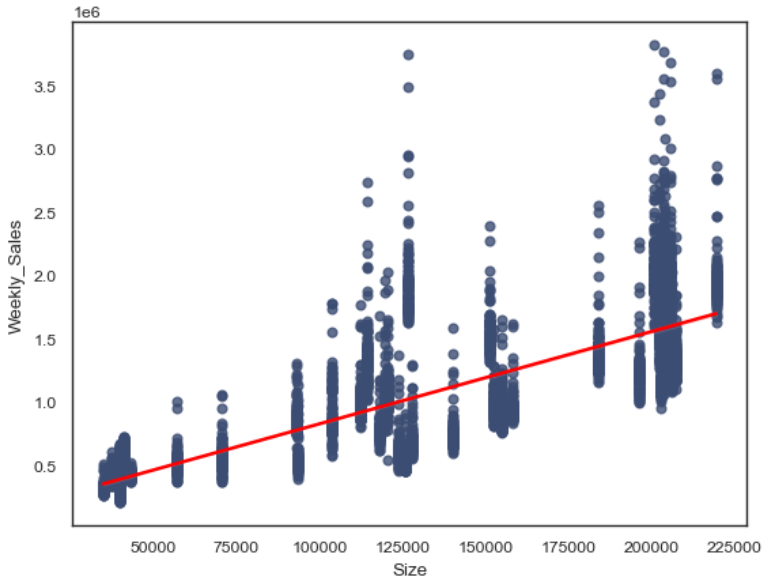
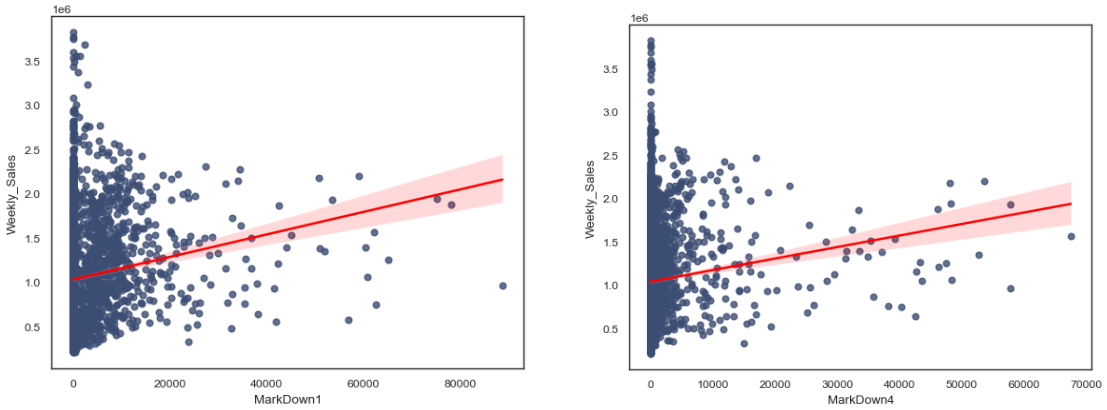


Figura 7 – Dispersão entre Vendas Semanais e Tamanho da Loja.

Analisando os gráficos de dispersão entre *Weekly Sales* e *Markdown* na Figura 8 mostram tendências positivas, sugerindo que maiores valores de Markdown podem estar associados a um aumento nas vendas. No entanto, a baixa disponibilidade de dados nas maiores faixas de Markdown nos traz uma margem de erro que aumenta progressivamente.



(a) Markdown 1 (b) Markdown 4

Figura 8 – Dispersão entre Vendas Semanais e Markdown.

A decomposição STL para as vendas semanais na Figura 9, sugere uma clara presença de sazonalidade e tendência no comportamento das vendas. A análise indica que as vendas têm pico em determinados períodos do ano. A componente de tendência mostra como as vendas evoluem ao longo do tempo, com a presença de uma ligeira tendência de alta. A componente sazonal demonstra flutuações periódicas que se repetem em intervalos regulares, como semanas específicas que consistentemente mostram aumento ou redução nas vendas. Estas variações parecerem ocorrer principalmente entre os meses de Novembro e Janeiro de cada ano. Os resíduos, que representam a variação nas vendas semanais não explicada pela tendência ou sazonalidade, parecem ser relativamente estáveis com alguns desvios, sugerindo que o modelo de decomposição está capturando bem os componentes principais da série temporal.

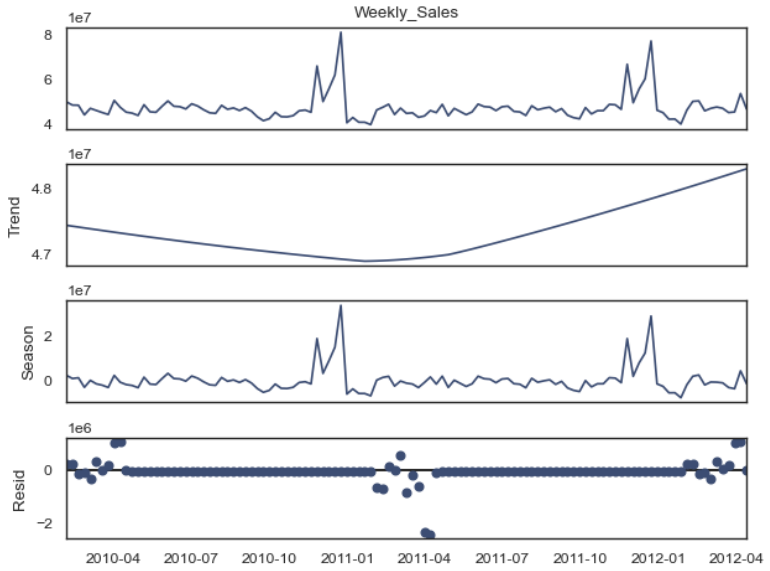


Figura 9 – Decomposição STL de Vendas Semanais.

A função de autocorrelação (ACF) e autocorrelação parcial (PACF) mostra como as vendas semanais estão correlacionadas com seus valores passados. Essas análises podem ser utilizadas na fase de seleção de modelos *ARIMA*, indicando o número de atrasos (*lags*) a serem usados para os termos autoregressivos e médias móveis.

Os gráficos na Figura 10 apresentam significância para os 2 primeiros *lags*, enquanto o ACF ainda apresenta uma sutil significância para o terceiro *lag*. Após um

longo período dentro dos intervalos, existe um novo pico de significância no *lag* 52. Os dados de venda estão divididos em semanas, e a presença de um nível de significância no *lag* 52 fornece evidências para a hipótese observada de que existe um pico anual em alguns períodos do ano.

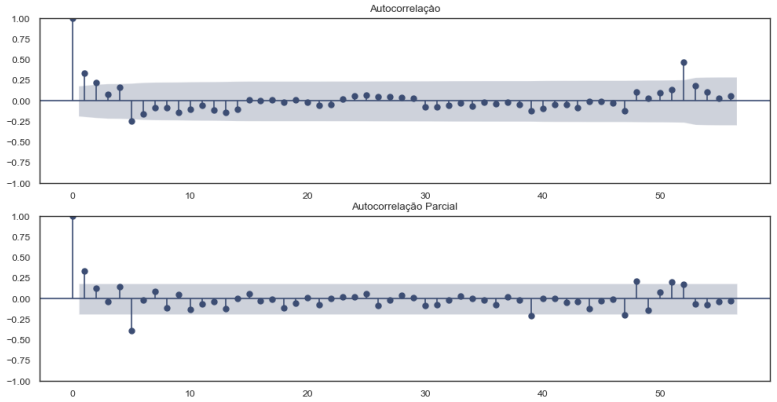


Figura 10 – Autocorrelação e Autocorrelação Parcial.

4.6 Feature Selection

Na etapa de *Feature Selection*, ou seleção de variáveis, aplicamos um conjunto diversificado de métricas para avaliar a relevância de cada uma. Essas métricas incluem:

- Importância de Recursos do *Random Forest* - que mede o quanto cada recurso diminui a impureza dentro dos nós das árvores de decisão;
- Análise de Componentes Principais (PCA) - que destaca a variância explicada por cada recurso, servindo como uma técnica de redução de dimensionalidade;
- Eliminação Recursiva de Características (RFE) - que classifica recursos pela ordem de importância ao progressivamente remover os menos significativos em um modelo de aprendizado supervisionado;
- Informação Mútua - que quantifica a quantidade de informação que uma variável contém sobre outra.

A lógica por trás dessas métricas está fundamentada na identificação de padrões e relações dentro dos dados, permitindo interpretar a força e a natureza das correlações entre as variáveis independentes e a variável de interesse. Ao interpretar essas métricas, devemos garantir que as variáveis selecionadas para o modelo sejam não apenas estatisticamente válidas, mas também relevantes para o contexto de negócios.

	RF Importance	PCA	RFE Ranking	Mutual Info
Store	0.139	0.151	3	1.673
Type	0.002	0.128	14	0.397
Size	0.440	0.105	1	1.446
Temperature	0.008	0.074	8	0.059
Fuel_Price	0.005	0.070	9	0.069
MarkDown1	0.001	0.064	15	0.033
MarkDown2	0.000	0.058	20	0.020
MarkDown3	0.003	0.053	11	0.033
MarkDown4	0.002	0.047	10	0.022
MarkDown5	0.001	0.043	17	0.055
CPI	0.059	0.039	4	0.515
Unemployment	0.017	0.036	6	0.693
IsHoliday	0.001	0.030	16	0.004
Week	0.057	0.029	5	0.015
Month	0.002	0.024	13	0.058
Size_Categorical_1	0.002	0.021	12	0.097
Size_Categorical_2	0.020	0.011	7	0.139
Size_Categorical_3	0.231	0.005	2	0.395
CPI_Categorical_1	0.000	0.000	18	0.049
CPI_Categorical_2	0.001	0.000	19	0.087

Tabela 2 – Seleção de Variáveis Iniciais.

Combinando as análises apresentadas na Tabela 2, *Size* e *Store* aparecem como variáveis significativas em todos os métodos de seleção, enquanto as colunas *MarkDowns* e *IsHoliday* parecem ter menos correlação com a variável de interesse. Poderia-se considerar a possibilidade de excluir *MarkDowns* devido à sua baixa importância e informação mútua. No entanto, devido ao contexto do trabalho, e o objetivo de medir o impacto de promoções e a previsibilidade em épocas de feriado, optou-se por manter essas colunas. A coluna *CPI Categorical* não demonstrou relevância em nenhuma das

métricas avaliadas, e desta forma, foi retirada do conjunto de treino e testes.

4.7 Feature Engineering

A *Feature Engineering*, ou engenharia de variáveis, consiste em transformar e criar dados derivados para fornecer ao modelo. Ao introduzir novas variáveis, como características temporais e totais de promoções, e ao normalizar os dados para garantir consistência na escala, é possivel melhorar significativamente a capacidade dos modelos de capturar padrões e tendências ocultas nos dados brutos. O conhecimento do negócio impulsiona essa fase.

O principal objetivo da rede de varejo *Walmart* ao divulgar esta base de dados na plataforma *Kaggle* é a previsão de como os feriados impactam nas vendas semanais. Isso nos orienta na criação de variáveis de impacto, como a contagem regressiva para feriados, que encapsulam o comportamento de compra e antecipação dos consumidores frente a estas datas. A agregação de *MarkDowns* em uma única métrica de *TotalMarkDown* fornece uma visão simplificada do impacto cumulativo das promoções. Isso nos permite avaliar a eficácia das estratégias promocionais de forma agregada e ajustar nossas abordagens de modelagem para refletir o efeito sinérgico dessas variáveis. O *StandardScaler* é uma técnica de pré-processamento que transforma cada variável do conjunto de dados para que tenha uma média de 0 e um desvio padrão de 1. O processo envolve subtrair a média de cada observação e então dividir pelo desvio padrão da característica. A aplicação do *StandardScaler* a variáveis como *Temperature*, *Fuel Price*, *CPI*, e *Unemployment* padroniza esses recursos, trazendo-os para uma escala comum e reduzindo o viés de escala.

Na modelagem de séries temporais, a engenharia de variáveis de atraso (ou lag) e médias móveis é uma prática adotada para entender as dinâmicas temporais nos dados. Neste estudo, foram introduzidas variáveis de atraso para as 1, 2 e 3 semanas anteriores, bem como médias móveis para as 2, 3 e 4 semanas antecedentes. A inclusão dessas variáveis de atraso e médias móveis permite que o modelo considere as relações entre os valores passados e o valor presente, oferecendo informações sobre o momento atual em que as vendas estão ocorrendo e as tendências de curto prazo.

Após a integração de novas variáveis derivadas, e a transformação de colunas

existentes, uma nova análise foi realizada para avaliar sua importância. A utilização de métodos com os mesmos critérios de seleção da seção de *Feature Selection*, pode ser conferida na Tabela 3:

	RF Importance	PCA	RFE Ranking	Mutual Info
Temperature	0.001	0.340	3	0.060
Fuel_Price	0.002	0.090	5	0.069
CPI	0.002	0.083	4	0.516
Unemployment	0.001	0.076	8	0.693
TotalMarkDown	0.000	0.061	14	0.078
Weekly_Sales_Lag_1	0.018	0.058	2	1.743
Weekly_Sales_Lag_2	0.001	0.057	10	1.561
Weekly_Sales_Lag_3	0.000	0.052	13	1.459
Weekly_Sales_MA_2	0.970	0.049	1	2.413
Weekly_Sales_MA_3	0.002	0.041	6	2.103
Weekly_Sales_MA_4	0.001	0.033	11	1.980
Size_Categorical_1	0.001	0.022	12	0.097
Size_Categorical_2	0.001	0.018	9	0.139
Size_Categorical_3	0.002	0.008	7	0.396
Proximity_Holiday_1	0.000	0.005	15	0.003
Proximity_Holiday_2	0.000	0.003	19	0.002
Proximity_Holiday_3	0.000	0.001	18	0.002
Proximity_Holiday_4	0.000	0.000	16	0.000
Proximity_Holiday_5	0.000	0.000	17	0.000

Tabela 3 – Seleção de Novas Variáveis e Variáveis com Alterações.

Os resultados apresentados na Tabela 3 mostram que, após a aplicação de *Standard Scaler*, as variáveis *Temperature*, *Fuel Price*, *CPI* e *Unemployment* mostraram uma melhora nas métricas avaliadas. Isso pode indicar que, embora não sejam as principais impulsionadoras das vendas, podem oferecer alguma informação sobre as condições externas que afetam o comportamento do consumidor. As variáveis de atraso e médias móveis mostraram resultados promissores, o que reforça a adoção desse tipo de variáveis para problemas de séries temporais. A variável *TotalMarkDown*, embora não seja uma das variáveis de maior significância, apresentou relevância especialmente na perspectiva da Informação Mútua, sugerindo que a agregação das promoções pode capturar efeitos cumulativos sobre as vendas. A variável *Proximity Holiday* apresentou resultados ruins,

mostrando baixa importância e Informação Mútua próxima de zero. Isso pode indicar que a contagem regressiva para feriados não tem tanto impacto nas vendas quanto inicialmente se supunha ou que o efeito é muito específico e não generalizado em todas as semanas.

Após a validação cruzada, optou-se por descartar as seguintes variáveis, a fim de não sobrecarregar os modelos com informações de baixo valor:

- *Proximity Holiday*
- *CPI Categorical*

4.8 Definição e Treino dos Modelos

A fase de modelagem começa com a preparação dos dados para treinamento. Alguns modelos podem apresentar dificuldades em processar tipos de dados como booleanos (Verdadeiro ou Falso) ou dados do tipo *category*. Para contornar isso, transformamos esses tipos de dados em números inteiros, onde cada número representa uma categoria distinta.

Foram selecionados três modelos para avaliação, como discutido anteriormente: *ARIMA*, *SARIMA* e Suavização Exponencial. A principal distinção entre *ARIMA* e *SARIMA* consiste na presença de ordens sazonais no modelo *SARIMA*, que buscam capturar a periodicidade e a magnitude dos efeitos sazonais sobre os dados. Considerando a segmentação dos dados por loja e um intervalo de tempo relativamente curto para treinamento, optou-se por uma abordagem de modelo conjunto, ou "ensemble", para cada tipo de modelo. Os dados foram divididos por loja, e cada modelo foi treinado individualmente para cada uma delas, com os resultados sendo combinados em um modelo abrangente. Assim, é possível realizar previsões específicas para cada loja em um determinado período e, em seguida, calcular a média geral para estimar o desempenho total das vendas da rede de varejo.

4.8.1 ARIMA

Para o modelo *ARIMA*, a definição das ordens pode ser baseada nos gráficos de autocorrelação e autocorrelação parcial. Uma abordagem mais moderna é utilizar a ferramenta de *Auto ARIMA*, que testa automaticamente várias combinações de ordens para determinar o modelo com o menor Critério de Informação de Akaike (AIC). O AIC é uma medida de qualidade de modelo que equilibra a complexidade do modelo com o seu desempenho, sendo útil para selecionar modelos com eficiência. Para a seleção dos modelos *ARIMA*, adotou-se o método *Auto ARIMA*. Considerando a vasta quantidade de lojas envolvidas na análise, torna-se impraticável discutir individualmente as ordens selecionadas para cada loja. No entanto, essa ferramenta garante que, apesar do volume significativo de dados, cada loja receba um modelo *ARIMA* otimizado para suas características e padrões de vendas específicos, sem a necessidade de uma análise detalhada e manual de cada configuração de modelo individual.

4.8.2 SARIMA

Para o modelo *SARIMA*, a intenção inicial era empregar o método *Auto ARIMA* para determinar as ordens e ordens sazonais apropriadas. No entanto, a análise dos gráficos de autocorrelação revelou um ciclo sazonal de aproximadamente 52 semanas, tornando o *Auto ARIMA* computacionalmente impraticável para este estudo. Optou-se por um modelo *SARIMA* mais genérico, com as ordens e ordens sazonais baseada nos gráficos de autocorrelação e autocorrelação parcial. Embora uma abordagem mais detalhada e específica para cada loja pudesse potencialmente levar a um modelo ensemble com maior precisão, dentro do escopo deste estudo, um modelo genérico oferece uma avaliação adequada das tendências e padrões sazonais gerais.

É importante reconhecer que, em séries temporais com longos períodos sazonais, a modelagem SARIMA pode apresentar desafios incluindo complexidade nas relações temporais e considerações sobre o uso computacional, que são aspectos importantes em aplicações práticas de negócios.

4.8.3 Suavização Exponencial

O método de Suavização Exponencial foi também selecionado para análise neste estudo. Este modelo é particularmente eficaz para séries temporais que apresentam padrões claros de nível, tendência e sazonalidade. Diferentemente dos modelos *ARIMA* e *SARIMA*, que se concentram em ajustar os dados baseando-se em suas relações autoregressivas e médias móveis, a Suavização Exponencial aplica um método de ponderação que atribui pesos decrescentes às observações mais antigas. Esta técnica é útil para previsões de curto prazo e em cenários onde a série temporal exibe um comportamento relativamente estável e previsível.

Um dos principais benefícios da Suavização Exponencial é sua capacidade de adaptar-se rapidamente a mudanças nos dados. No entanto, é importante notar que este modelo pode apresentar limitações ao lidar com padrões sazonais muito complexos ou mudanças abruptas na tendência.

4.9 Avaliação dos Modelos

Os resultados da modelagem no conjunto de validação foram agregados a uma única tabela, agrupada por Loja e Data. Na Figura 11 encontra-se ilustrados os resultados dos modelos *ARIMA*, *SARIMA* e Suavização Exponencial aplicados ao conjunto de validação. As vendas reais durante o período de treinamento e teste são representadas, respectivamente, pelas linhas contínuas azul e laranja. As predições de cada modelo são representadas pelas linhas tracejadas, diferenciadas por cor. Ao observar o gráfico, nota-se que todos os modelos conseguem acompanhar as tendências e flutuações semanais das vendas, com variações em sua acurácia.

A linha de predições do modelo *ARIMA* acompanha de perto a trajetória das vendas reais, mas com desvios notáveis em certos pontos. O modelo *SARIMA*, projetado para capturar a sazonalidade, mostra um alinhamento similar. Já as predições do modelo de Suavização Exponencial tendem a suavizar os picos e vales, indicando uma reatividade menor às mudanças rápidas no volume de vendas. Enquanto as médias das predições fornecem uma boa indicação geral do desempenho do modelo, uma investigação mais detalhada a nível individual de cada loja pode desvendar áreas específicas de melhoria, revelando modelos que apresentam performance abaixo do esperado e ajustá-los para

melhor refletir as dinâmicas de vendas locais. Portanto, embora os resultados atuais sejam promissores como ponto de partida, um exame mais aprofundado seria necessário para otimizar a precisão das previsões e garantir sua aplicabilidade.

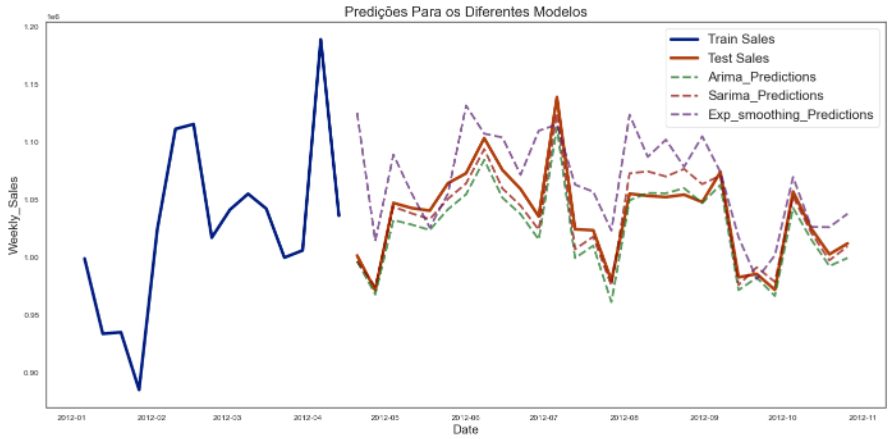


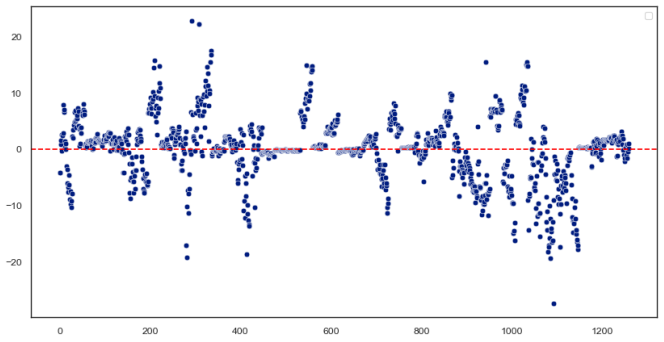
Figura 11 – Resultados das Previsões para os Modelos ARIMA (verde tracejado), SARIMA (laranja tracejado) e Suavização Exponencial (roxo tracejado)

4.9.1 Análise de Resíduos

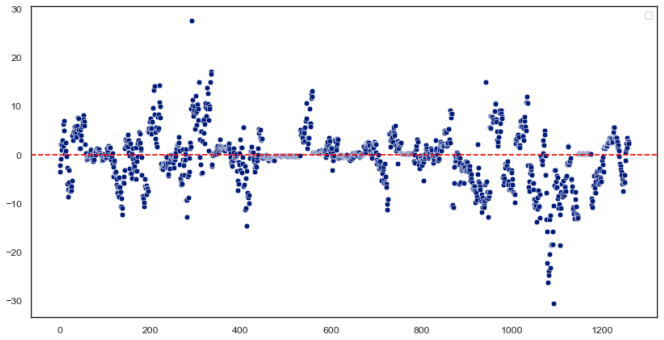
Resíduos são as discrepâncias entre os valores observados e as previsões dos modelos, com o cenário ideal sendo uma distribuição aleatória dessas diferenças, indicando a ausência de autocorrelação. Isto é, espera-se que os resíduos se distribuam aleatoriamente em torno do valor zero, sem exibir padrões discerníveis ou correlações, sugerindo que o modelo capturou toda a informação disponível. Os resíduos que mostram padrões ou autocorrelações significativas podem sinalizar inadequações na modelagem, tais como a omissão de variáveis importantes ou uma escolha ruim de parâmetros no modelo. No contexto deste estudo, os resíduos foram analisados em termos percentuais para proporcionar uma escala ajustada em relação aos altos volumes de vendas semanais da rede de varejo.

Conforme ilustrado na Figura 12, os resíduos dos três modelos - ARIMA, SARIMA e Suavização Exponencial - apresentam comportamentos similares. Embora a maioria dos resíduos esteja alinhada em torno do zero percentual, a variabilidade observada sugere que há espaço para melhorias. Dentro do escopo deste estudo, não será realizada uma análise detalhada por loja ou um ajuste fino (*fine-tuning*) dos modelos.

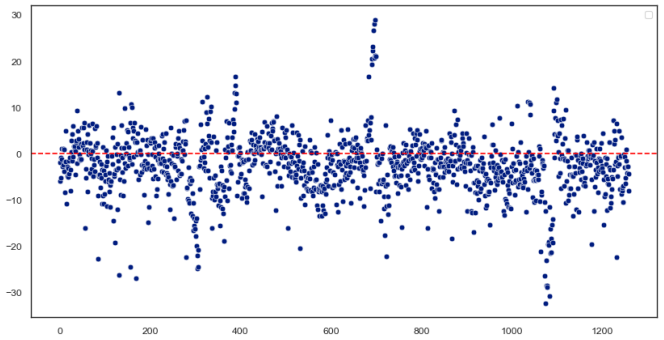
No entanto, a modelagem é um processo repetitivo e dinâmico; análises subsequentes e ajustes baseados em novos ciclos de treinamento podem revelar oportunidades para otimizar ainda mais o desempenho preditivo.



(a) ARIMA



(b) SARIMA



(c) Suavização Exponencial

Figura 12 – Análise de Resíduos Percentual.

4.9.2 Métricas de Avaliação

De acordo com a metodologia estabelecida, foi avaliado o desempenho dos modelos implementados utilizando as métricas MAE (Erro Absoluto Médio), RMSE (Erro Quadrático Médio) e MAPE (Erro Percentual Médio Absoluto). Os resultados dessas métricas são apresentados na Tabela 4. Os modelos ARIMA e SARIMA exibiram um desempenho similar, com MAPE de 3.74% e 3.66%, respectivamente. O modelo de Suavização Exponencial demonstrou erros ligeiramente maiores, atingindo um MAPE de 5.05%.

Importante ressaltar que esta análise foi conduzida sobre as médias consolidadas de vendas semanais por data. Uma inspeção mais detalhada poderia identificar variações significativas de desempenho entre os modelos a nível de cada loja. Tal análise seria essencial para direcionar o processo de ajuste fino (*fine-tuning*) dos modelos, buscando melhorias nos que apresentam desempenho abaixo do esperado. Além disso, é importante considerar que o conjunto de dados de validação não englobou os períodos de pico de vendas observados nas semanas entre os mese de Novembro e Janeiro dos anos anteriores. Tais picos representariam um teste mais rigoroso para os modelos, desafiando-os a capturar os aumentos sazonais de vendas e potencialmente expondo uma incapacidade de previsão não reveladas pela análise atual.

Modelo	MAE	RMSE	MAPE (%)
ARIMA	37 747.14	41 974.24	3.74
SARIMA	36 021.48	41 648.59	3.66
Suavização Exponencial	51 769.72	64 414.72	5.05

Tabela 4 – Métricas de Avaliação para Cada Modelo.

CAPÍTULO

5

CONCLUSÃO

5.1 Conclusão

Este estudo adotou uma metodologia baseada na metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*) para a análise de séries temporais, com o intuito de aplicar modelos de séries temporais aplicados aos dados de vendas semanais do *Walmart*. O objetivo principal foi explorar os padrões e fatores que impactam essas vendas, utilizando a limpeza e análise exploratória de dados como instrumentos para extrair ideias e informações, e estabelecer modelos preditivos iniciais.

A compreensão dos dados foi alcançada por meio da integração e limpeza do conjunto de dados separados em treino e validação, a fim de evitar o vazamento de dados para o conjunto de validação, garantindo a qualidade e consistência dos dados analisados. A aplicação de técnicas de *Feature Selection* e *Feature Engineering* permitiu aprimorar o conjunto de dados, ressaltando variáveis significativas e transformando outras para melhor representar a dinâmica das vendas no varejo. Os resultados indicaram que as vendas são influenciadas por uma série de fatores internos e externos. No entanto, o impacto de variáveis como feriados, preço do combustível e CPI (índice de preços ao consumidor) revelou-se mais complexo do que inicialmente previsto.

As métricas de avaliação mostraram um bom potencial preditivo dos modelos, com o Erro Médio Quadrático a 4% para os modelos *ARIMA* e *SARIMA*, e ligeiramente

superiores a 5% para o de Suavização Exponencial. A inclusão de períodos de pico de vendas nas análises poderia constituir um teste mais rigoroso para os modelos, potencialmente expondo limitações não observadas na avaliação do presente estudo. A análise residual aponta para oportunidades de otimização, particularmente na personalização dos modelos a nível de cada loja, e a oportunidade de efetuar-se o *fine-tuning*, ou ajuste fino.

Os desafios encontrados durante o estudo incluíram a interpretação das variáveis Markdown, estrategicamente anonimizadas pela empresa fornecedora, e a baixa quantidade de dados disponíveis para treinamento e teste. A complexidade computacional na configuração de um modelo *SARIMA* com períodos sazonais extensos e grande quantidade de lojas também exigiu a definição manual das ordens do modelo, no lugar de uma personalização para cada loja como inicialmente previsto.

Os resultados deste trabalho oferecem uma perspectiva empresarial sobre a aplicação de análise de séries temporais, e reforça as principais dificuldades e considerações necessárias para essas modelagens. As técnicas utilizadas são transferíveis para outros contextos de séries temporais, desde que devidamente adaptadas às variáveis específicas e ao problema comercial em questão, visando aprimorar a capacidade preditiva e tomada de decisões estratégicas.

REFERÊNCIAS

ADMIN, W. C. W. C. **Walmart Recruiting - Store Sales Forecasting**. Kaggle, 2014. Disponível em: <https://kaggle.com/competitions/walmart-recruiting-store-sales-forecasting>. Citado na página 26.

BROWNLEE, J. **Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python**. Machine Learning Mastery, 2020. Disponível em: <https://books.google.com.br/books?id=uAPuDwAAQBAJ>. Citado na página 22.

CLEVELAND, R. B.; CLEVELAND, W. S.; MCRAE, J. E.; TERPENNING, I. Stl: A seasonal-trend decomposition procedure based on loess (with discussion). **Journal of Official Statistics**, v. 6, p. 3–73, 1990. Citado na página 29.

DICKEY, D. A.; FULLER, W. A. Distribution of the estimators for autoregressive time series with a unit root. **Journal of the American Statistical Association**, Taylor Francis, v. 74, n. 366a, p. 427–431, 1979. Citado na página 28.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: Principles and Practice**. 2. ed. [S.l.]: OTexts, 2018. Citado nas páginas 21 e 22.

LIVERA, A. M. D.; HYNDMAN, R. J.; SNYDER, R. D. Forecasting time series with complex seasonal patterns using exponential smoothing. **Journal of the American Statistical Association**, [American Statistical Association, Taylor Francis, Ltd.], v. 106, n. 496, p. 1513–1527, 2011. ISSN 01621459. Disponível em: <http://www.jstor.org/stable/23239555>. Citado na página 22.

MORETTIN, P. A.; TOLOI, C. M. **Análise de séries temporais: modelos lineares univariados**. [S.l.]: Editora Blucher, 2018. Citado na página 22.

SHEARER, C. The crisp-dm model: the new blueprint for data mining. **Journal of data warehousing**, THE DATA WAREHOUSE INSTITUTE, v. 5, n. 4, p. 13–22, 2000. Citado na página 25.

TSAY, R. **Analysis of financial time series**. 2. ed.. ed. Hoboken, NJ: Wiley-Interscience, 2005. (Wiley series in probability and statistics). ISBN 978-0-471-69074-0. Disponível em: <http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+483463442&sourceid=fbw_bibsonomy>. Citado na página 22.

ZEMKOHO, A. A basic time series forecasting course with python. **SN Operations Research Forum**, v. 4, 12 2022. Citado na página 22.