Data 07/11/2024

Melhores Compras | Predição de Vendas

Introdução e objetivos

Tem-se como objetivo, predizer o volume de vendas semanais a partir da base de dados, utilizando o algoritmo mais eficiente, ou seja, aquele que apresenta a maior precisão na previsão desse volume. A partir dessa abordagem, busca-se identificar o modelo capaz de gerar previsões mais acuradas, contribuindo para geração de insights financeiros importantes para o E-Commerce Melhores Compras.

Limpeza dos dados

Descrição das variáveis envolvidas:

- **Date**: indica a data em que a observação foi feita;
- **Weekly_Sales**: as vendas da empresa na semana, indicada pelo campo Date;
- **Holiday_Flag**: indica se houve um feriado na semana medida;
- **Temperature**: a temperatura média registrada naquela semana;
- **Fuel_Price**: preço do combustível. Deve ser dividido por 1.000 para se obter o preço em reais e centavos;
- **CPI**: índice que indica o nível de atividade econômica da região;
- **Unemployment**: nível de desemprego medido nacionalmente. Deve ser dividido por 1.000 para se obter a informação em percentual.

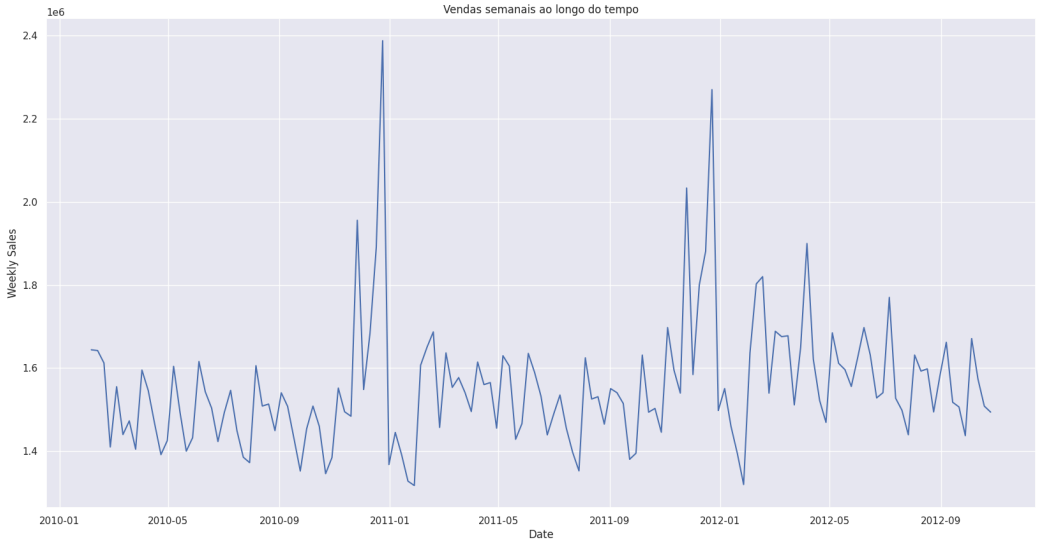
Pré-Processamento:

- Não houve dados duplicados na base
- Não houve valores negativos ou nulos
- **Fuel_Price:** Dados divididos por 1000 para adequação
- **Unemployment**: Dados divididos por 1000 para adequação
- **Date:** Transformação para o tipo datetime do pandas, garantindo que o formato DD/MM/YYYY

AED - Análise Exploratória de Dados

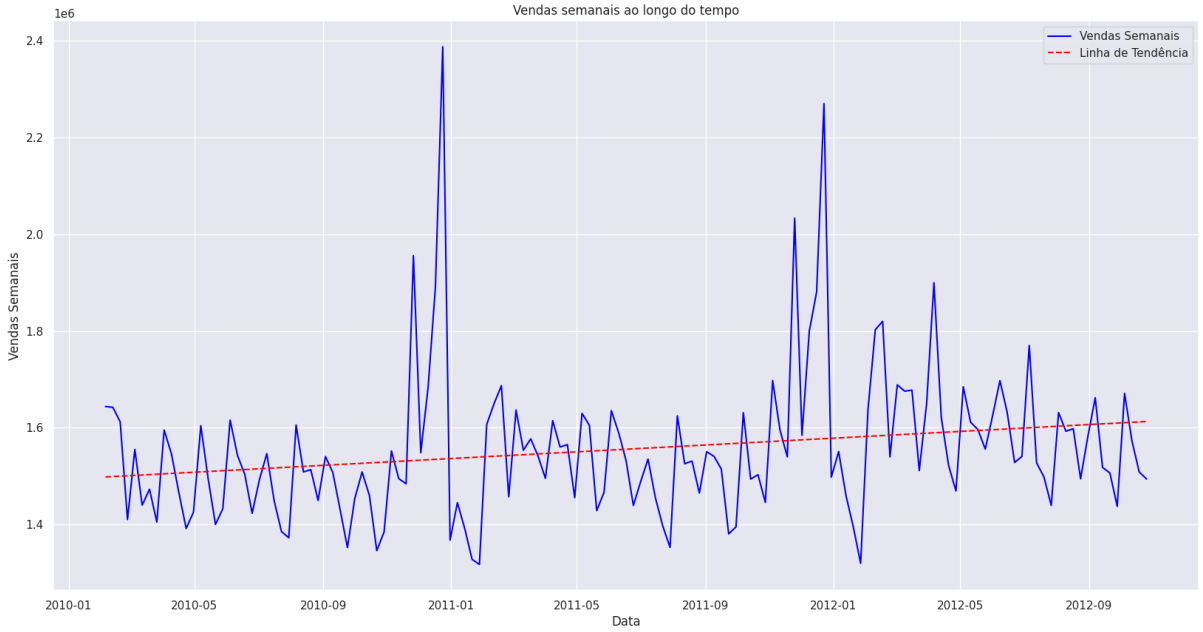
Visualização dos dados

1 - Gráfico de Linha: Comparação das Vendas ao Longo do Tempo



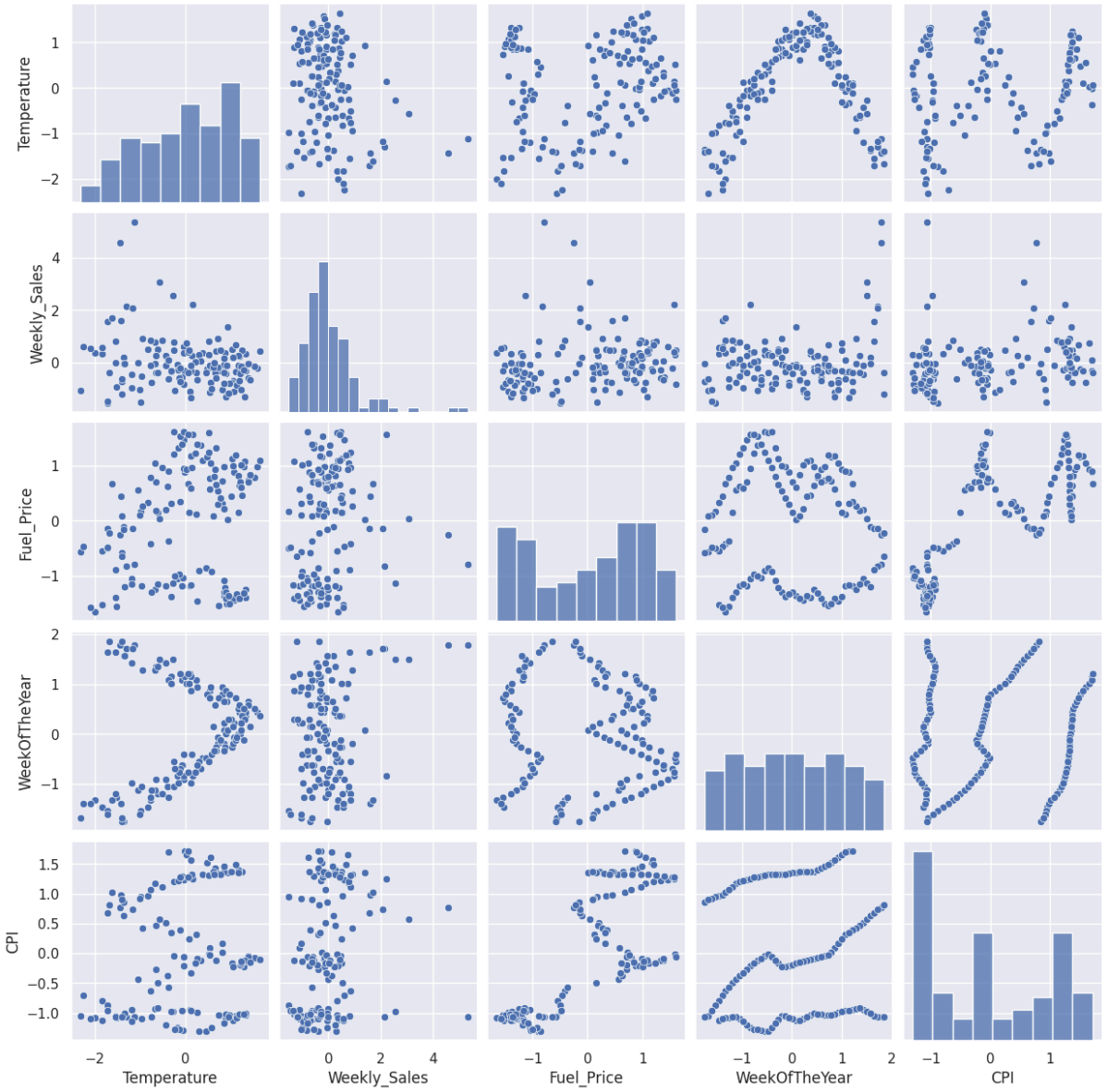
Observa-se um aumento significativo no volume de vendas durante o final de ano, além de flutuações constantes em outros períodos. Esse padrão é característico do mercado de e-commerce global, com picos de vendas impulsionados por eventos sazonais e promoções, e variações ao longo do ano refletindo o comportamento do consumidor. Esses padrões não são fenômenos isolados, mas sim parte de um comportamento amplamente reconhecido no setor, sendo essenciais para análise de vendas e estratégias de marketing.

2 - Gráfico de Linha: Linhas de Tendência Vendas ao Longo do Tempo



Com base nos dados, podemos entender que, à medida que a data avança, as vendas semanais apresentam uma tendência de crescimento. Isso sugere que existe uma correlação positiva entre as semanas e as vendas. A correlação positiva entre as duas variáveis indica que, quanto mais longe no tempo, mais altas são as vendas.

3 - Análise de Correlação com Gráficos de Dispersão: Identificação de Padrões e Relações entre Variáveis

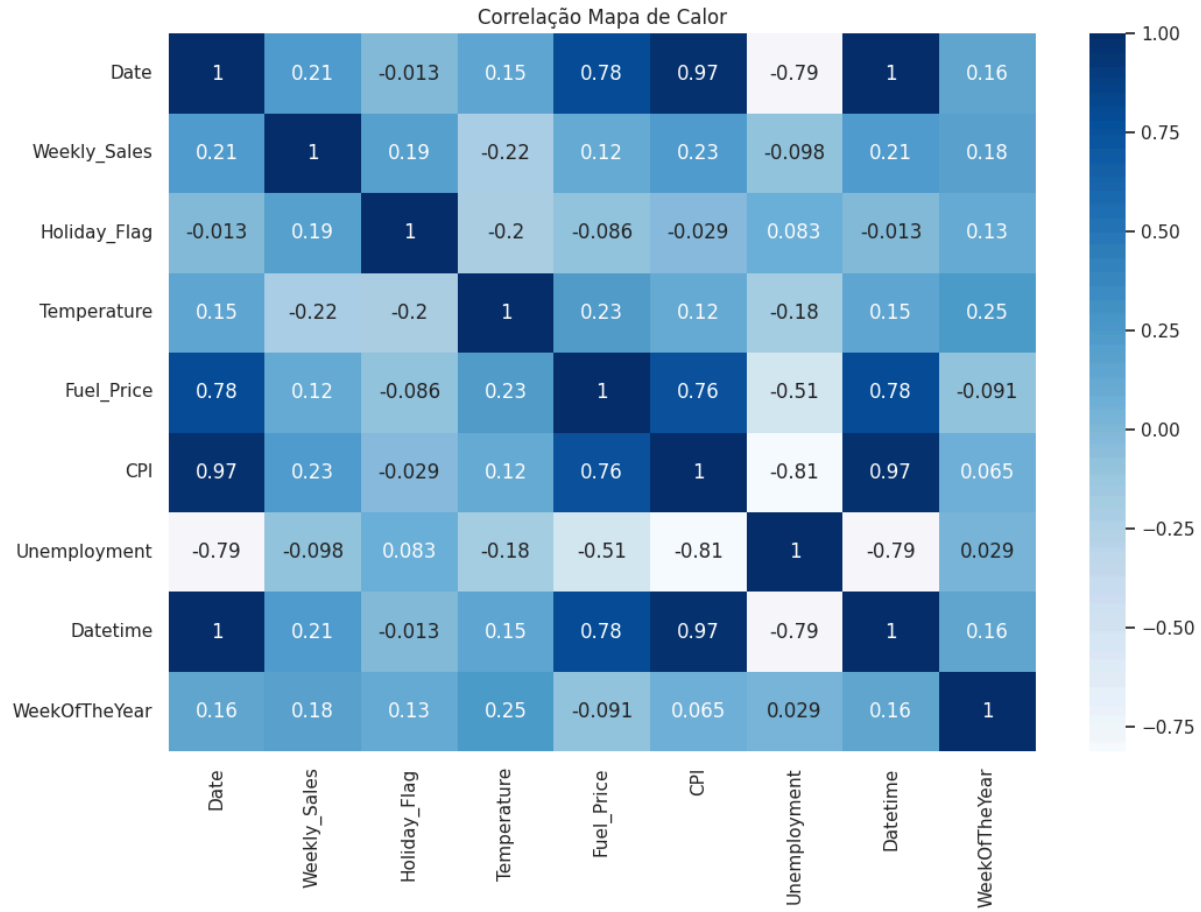


Com base na imagem acima não foram encontradas correlações fortes entre a variável Weekly_Sales e as variáveis independentes. Isso sugere que outros métodos de análise.

4 - Visualização da Correlação entre Variáveis na Tabela Sales

Antes de selecionar o algoritmo de regressão, é fundamental analisar a correlação, ou seja, a relação entre a variável **'weekly_sales'** e as demais variáveis que possam influenciar suas variações. Com base nessa análise, faz sentido avaliar as

variáveis ***Date***, ***Holiday_Flag***, ***Fuel_Price*** e ***CPI***, pois são aquelas que apresentam as maiores correlações com a variável de interesse.



Modelagem - Aprendizado de Máquina

Técnicas/Algoritmos de Regressão Linear

Gradient Boosting: é usado para melhorar a precisão das previsões. Ela funciona juntando várias "árvores de decisão" simples, onde cada nova árvore tenta corrigir os erros das árvores anteriores. Com base nessa técnica, vamos verificar quatro tipos de regressores lineares.

Processamento

Foi-se dividido os dados da seguinte forma: 80% dos dados para treinamento e 20% para testes, permitindo que o modelo aprenda com a maior parte das informações e seja validado em dados inéditos, o que favorece uma boa capacidade de generalização.

XGBoost

- Alto desempenho e precisão. Ele corrige erros de modelos anteriores, sendo rápido e eficaz em vários tipos de dados.

MSE:	0.27	**Erro Quadrático Médio:** Calcula a média dos erros elevado ao quadrado, no caso, quanto menor, melhor.
RMSE:	0.52	**RMSE (Raiz do Erro Quadrático Médio)**: É uma variação do MSE que indica, em média, o quão perto as previsões estão dos valores reais. Quanto menor, melhor a precisão.
MAE:	0.41	**Média do Erro Absoluto):** Mede o erro médio sem se preocupar se é positivo ou negativo. Quanto menor, melhor.
R2:	0.72	**Coeficiente de Determinação:** Mostra o percentual da variação dos dados. Quanto mais perto de 1, melhor.

SVM (Support Vector Machine):

- Ele funciona corrigindo os erros cometidos pelos modelos anteriores, o que o torna mais forte a cada etapa. É rápido e pode lidar bem com diferentes tipos de dados.

MSE:	1.10	**Erro Quadrático Médio:** Calcula a média dos erros elevado ao quadrado, no caso, quanto menor, melhor.
RMSE:	1.05	**RMSE (Raiz do Erro Quadrático Médio)**: É uma variação do MSE que indica, em média, o quão perto as previsões estão dos valores reais. Quanto menor, melhor a precisão.
MAE:	1.75	**Média do Erro Absoluto):** Mede o erro médio sem se preocupar se é positivo ou negativo. Quanto menor, melhor.
R2:	-0.10	**Coeficiente de Determinação:** Mostra o percentual da variação dos dados. Quanto mais perto de 1, melhor.

Random Forest

- Baseado em várias árvores de decisão, o Random Forest melhora a precisão combinando múltiplos modelos, tornando-o robusto e menos propenso ao overfitting.

MSE:	0.78	**Erro Quadrático Médio:** Quanto menor, melhor.

RMSE:	0.99	**RMSE (Raiz do Erro Quadrático Médio)**:
MAE:	0.70	**Média do Erro Absoluto):** Mede o erro médio sem se preocupar se é positivo ou negativo. Quanto menor, melhor.
R2:	0.02	**Coeficiente de Determinação:** Mostra quanto o modelo explica dos dados. Quanto mais perto de 1, melhor.

Linear Regression

- É um método que ajusta uma linha reta para representar a relação entre uma variável dependente e uma ou mais variáveis independentes, permitindo prever valores futuros com base nessa relação linear.

MSE:	0.98	**Erro Quadrático Médio:** Quanto menor, melhor.
RMSE:	0.99	**RMSE (Raiz do Erro Quadrático Médio)**:
MAE:	0.70	**Média do Erro Absoluto):** Mede o erro médio sem se preocupar se é positivo ou negativo. Quanto menor, melhor.
R2:	0.02	**Coeficiente de Determinação:** Mostra quanto o modelo explica dos dados. Quanto mais perto de 1, melhor.

Análise dos Resultados: Testando a Performance de 4 Modelos

XGBRegressor:
Este modelo foi o melhor de todos. Ele obteve os melhores resultados, com o menor MSE (erro quadrático) e RMSE (raiz do erro), o que significa que seus erros foram mais baixos em comparação com os outros modelos. Além disso, o R^2 (coeficiente de determinação) foi de 66%, o que indica que ele consegue explicar bem as variações nos dados — ou seja, ele fez boas previsões, capturando a maior parte do comportamento dos dados.

RandomForestRegressor:
Esse modelo também teve um bom desempenho, mas não tanto quanto o XGBRegressor. Ele obteve resultados razoáveis em termos de MSE e RMSE, mas seu R^2 foi de 41%. Isso significa que ele conseguiu explicar uma boa parte das variações nos dados, mas não tão bem quanto o XGBoost. Em termos simples, ele fez boas previsões, mas não foi perfeito.

LinearRegression (Regressão Linear):
Este modelo teve um desempenho mais fraco. O R^2 foi de apenas 2%, o que significa que ele não conseguiu explicar quase nada das variações nos dados. Em

outras palavras, o modelo não acertou muito bem as previsões, porque a relação entre as variáveis não é bem representada de forma linear, ou seja, ele não capturou os padrões complexos presentes nos dados.

SVR (Support Vector Regression):
O SVR foi o modelo que teve o pior desempenho. O R^2 foi negativo, o que é um sinal de que o modelo não conseguiu fazer previsões boas. Na verdade, um R^2 negativo indica que ele errou mais do que acertou, ou seja, ele não conseguiu se ajustar bem aos dados e suas previsões foram bem imprecisas.

Testes e Validações

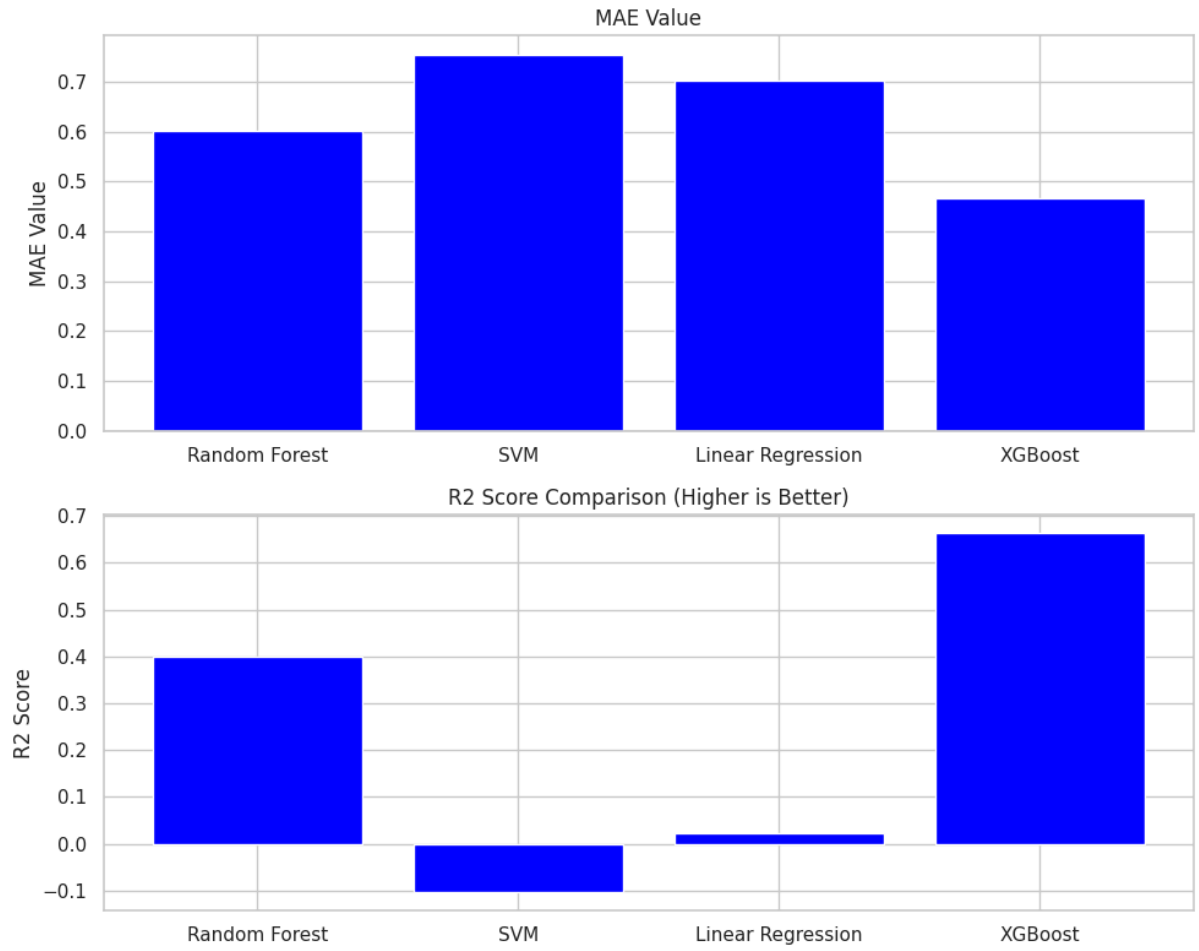
Validação cruzada

Nosso modelo passou por um processo de validação cruzada, dividido em 5 partes (chamadas de 'folds'). Em cada parte, o modelo foi treinado e testado, e o RMSE (que mede o erro médio das previsões) foi calculado. Os resultados para cada uma das 5 partes foram os seguintes:

Fold 1: RMSE	0.58
Fold 2: RMSE	0.48
Fold 3: RMSE	0.74
Fold 4: RMSE	0.61
Fold 5: RMSE	1.26

Esses valores indicam o erro do modelo em cada divisão dos dados. O RMSE médio de todos os testes foi de 0.74, o que nos dá uma boa ideia do erro geral do modelo. Ou seja, em média, o modelo teve um erro de 0.74 unidades nas suas previsões, o que é um indicativo de como ele se saiu ao longo de diferentes partes dos dados.

Análise Comparativa: Algoritmos Preditores



Neste contexto, dispõe-se das ferramentas necessárias para identificar o preditor mais eficiente e assertivo para os volumes de vendas semanais. A análise demonstra que o XGBoost é, de fato, o modelo com maior potencial de precisão, uma vez que apresenta o menor erro médio absoluto (MAE), o que resulta em um R^2 mais elevado e, consequentemente, maior capacidade de previsão. O **R^2** (coeficiente de determinação) indica a proporção da variabilidade dos dados que o modelo consegue explicar; quanto mais próximo de 1, melhor o modelo está ajustado aos dados.

Otimização de Modelo: XGBoost

Para aprimorar a precisão do modelo, realizamos a otimização dos hiperparâmetros. Esses hiperparâmetros são ajustes feitos no processo de aprendizado que ajudam o modelo a se adaptar melhor aos dados. Encontrar os valores ideais para esses parâmetros é crucial para obter um bom equilíbrio entre desempenho e generalização, evitando problemas como o overfitting (quando o modelo se ajusta aos dados de treino) ou o underfitting (quando o modelo não consegue capturar os padrões importantes dos dados).

Resultado do Modelo com Hiperparâmetros Otimizados:

Valor obtido: 39.602 (presumivelmente o RMSE)

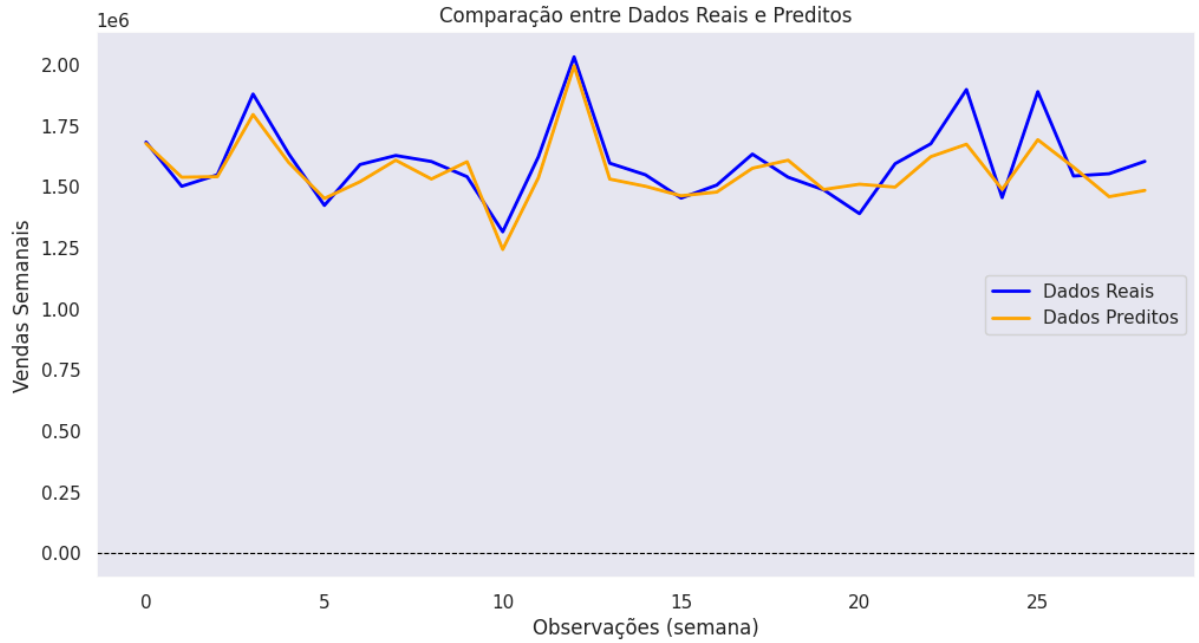
Interpretação:

RMSE de 39.602 significa que, após a otimização dos hiperparâmetros, o erro médio nas previsões do modelo foi de 39.602 unidades. Isso representa uma melhoria significativa em relação ao modelo anterior, com uma redução do erro nas previsões. Quanto menor o valor do RMSE, melhor o modelo, pois ele está fazendo previsões mais precisas e próximas dos valores reais.

Resultado do Modelo Sem Hiperparâmetros Otimizados:
RMSE: 50.546

Interpretação:
O RMSE de 50.546 significa que, em média, o modelo está errando 50.546 unidades nas suas previsões. Ou seja, toda vez que o modelo tenta prever algo (por exemplo, o valor de vendas semanais), ele erra em cerca de 50.546 unidades.

Análise Comparativa: Dados Reais vs. Preditos para as Semanas do Conjunto de Teste



Os dados preditos apresentam um padrão linear consistente, o que reflete a relação clara entre as variáveis nos dados utilizados. Esse comportamento é característico de modelos de regressão aplicados a dados com tendências estáveis e pouco voláteis. Apesar de sua simplicidade, o modelo consegue capturar com precisão os padrões principais dos dados, o que resulta em previsões confiáveis para cenários mais previsíveis. Isso mostra que o modelo é eficiente em representar relações estáveis, mesmo que com limitações para capturar variações muito complexas ou sazonais.

Conclusões

Percebe-se, portanto, a importância da análise exploratória de dados a posteriori do levantamento de hipótese; Fomos da série temporal de vendas semanais para o mapa de correlação entre as variáveis; Dessa forma, tendo uma análise be, fundamentada para performar o melhor modelo preditor de volumes de vendas para a melhores compras. Entre os algoritmos testados, o **XGBoost** se destacou por sua alta performance, mostrando-se o mais eficaz na previsão dos volumes de vendas, mas claro, para chegar nesse patamar, foi necessário realizar um ajuste cuidadoso dos hiperparâmetros, garantindo que o modelo estivesse bem configurado para se tornar um eficiente preditor de vendas, mesmo com relativo poucos dados.

Próximos passos

Em termos de volume de dados, vale ressaltar que uma forma eficaz de melhorar a performance do **XGBoost** é expô-lo a um maior volume de dados, especialmente dados históricos. Além disso, a partir dessa análise, ficou claro o impacto de fatores externos em um segmento tão relevante para o setor comercial. Na base de dados, incluímos variáveis macroeconômicas da região, como temperatura, índice de potencial econômico e taxa de desemprego, que exerceram influência significativa sobre o comportamento observado.
Nesse sentido, seria relevante acrescentar a esses fatores a variável inflação - o aumento generalizado e contínuo dos preços dos bens e serviços ao longo do tempo, com certeza possui uma relação - quiçá - uma correlação com o volume de vendas no setor comercial.