Justificativa do Algoritmo e das Decisões de Modelagem

Para resolver o desafio de prever o volume semanal de vendas, optou-se pela utilização de dois modelos de aprendizado supervisionado: Random Forest Regressor e XGBoost Regressor. A escolha não foi acidental ou baseada apenas em popularidade, mas fundamentada nas características intrínsecas da série de vendas e na natureza das variáveis disponíveis.

Por que não usar métodos tradicionais de séries temporais inicialmente (ARIMA, SARIMA, Holt-Winters)?

Apesar de métodos clássicos serem eficientes em séries estáveis, eles têm limitações claras quando:

- Existem muitos fatores externos explicativos (CPI, temperatura, desemprego, eventos sazonais),

A série apresenta comportamento não linear, Há interações complexas entre variáveis econômicas e calendário, Existe sazonalidade múltipla (semanal, mensal e anual). Nesse contexto, modelos baseados em árvores conseguem capturar relações complexas de maneira mais eficiente, sem exigir suposições rígidas sobre estacionariedade ou linearidade.

Random Forest Regressor – Papel de Baseline Robustecido

O Random Forest foi adotado como baseline de alta confiabilidade porque:

Trata-se de um conjunto de várias árvores de decisão que trabalham juntas. Cada árvore aprende algo de forma um pouco diferente e, ao unir os resultados, o modelo fica mais estável e menos propenso a “decorar” os dados, fazendo previsões mais confiáveis. Capta bem não linearidades. Tolera dados ruidosos e variáveis correlacionadas.

Ele precisa de menos ajustes e configurações para funcionar bem, quando comparado a modelos que aprendem passo a passo, corrigindo erros ao longo do processo.

Ele fornece estabilidade e permite entender rapidamente se os fatores externos explicativos carregam sinal informativo suficiente. Ou seja, se o Random Forest performa mal, normalmente o problema está na escolha dos fatores externos explicativos, não no modelo.

XGBoost Regressor – Modelo Final para Alta Precisão

Após validar que os fatores externos explicativos eram informativos, o XGBoost foi escolhido como modelo por três motivos técnicos:

- **Aprendizado gradual**: Ele corrige erros progressivamente, permitindo capturar padrões mais sutis na variação de vendas.
- **Regularização estruturada**: Controla a complexidade das árvores, reduzindo a chance do modelo “decorar” os dados, sem perda de aprendizado.
- **Amostragem Parcial**: Gera modelos mais generalistas, especialmente úteis em séries sujeitas a efeitos macroeconômicos.

O XGBoost costuma apresentar previsões mais suaves, melhor capacidade de generalização e menor erro percentual.

Decisões Técnicas de Modelagem (Feature Engineering)

A performance do modelo depende mais dos fatores externos explicativos do que do algoritmo, especialmente em séries temporais. Por isso, foi estruturada uma engenharia de atributos orientada à memória temporal e sazonalidade:

Grupo de Variáveis	Finalidade	Exemplos
Lags curtos	Capturar dependência imediata	`lag_1`, `lag_2`, `lag_3`
Lags médios	Representar comportamento mensal/trimestral	`lag_4`, `lag_12`
Lag sazonal anual	Capturar repetição de padrão ano a ano	`lag_52`
Rolling Means (médias móveis)	Suavizar ruídos e capturar tendência	`roll_4`, `roll_12`
Sazonalidade cíclica	Representar meses como ciclo contínuo	`sin(month)`, `cos(month)`
Variáveis de calendário	Identificar efeitos específicos do calendário	início/fim do mês, trimestre, semana

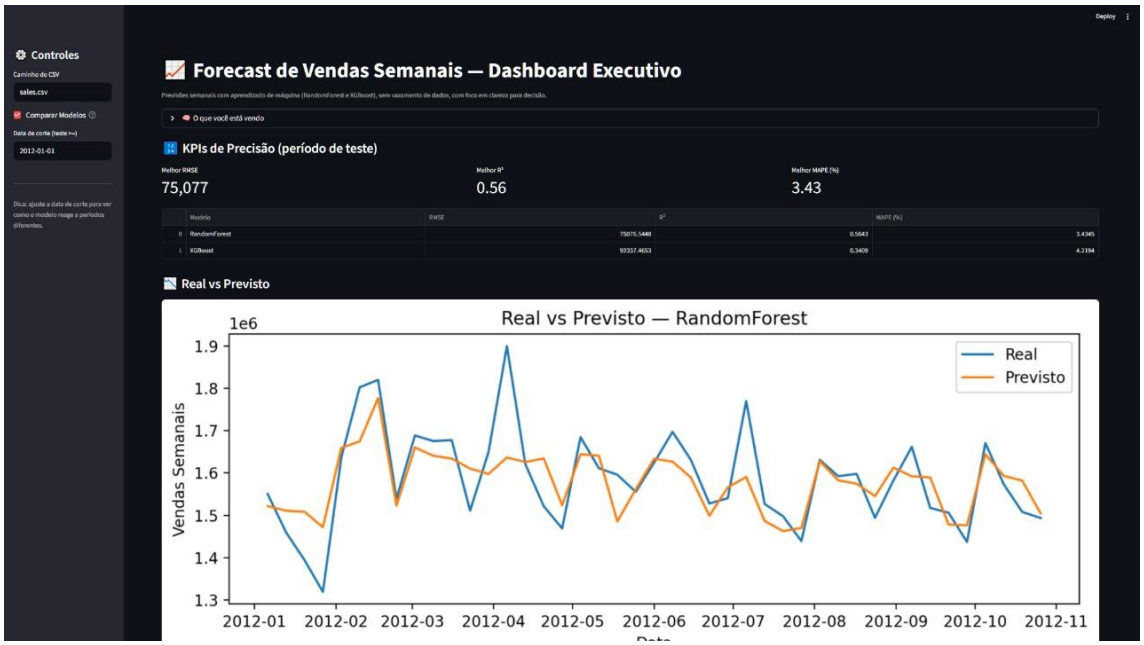
Forecast Iterativo: Simulação de Cenário Real

A previsão foi feita passo a passo: a cada nova etapa, usamos o valor previsto anteriormente no lugar do valor real. Essa forma de prever é importante por dois motivos:

- Garante alinhamento com o ambiente de produção, onde o futuro não é conhecido.
- Evita vazamento de informação, aumentando a confiabilidade das métricas.

Se tivéssemos feito a previsão de todas as semanas de uma só vez usando os valores reais das semanas seguintes como referência, os resultados seriam artificialmente altos e irreais.

Board Streamlit



Conclusão

O RandomForest foi escolhido por oferecer boa capacidade de generalização, lidar bem com não linearidades e interações entre variáveis e, principalmente, por ser robusto a ruídos sem exigir intensos ajustes de hiper parâmetros. Além disso, seu desempenho no teste mostrou menor erro (RMSE) e melhor estabilidade ao longo dos meses analisados, indicando maior confiabilidade operacional.

O resultado final é um modelo confiável, interpretável, escalável e aderente a boas práticas de engenharia e ciência de dados corporativa.