Relatório da A2 de Séries Temporais:

Modelagem de variável utilizando conceitos de Séries Temporais

Integrantes: Guilherme Carvalho, Guilherme Buss, Gustavo Bianchi, João Gabriel, Luís Felipe Marciano e Vinícius Nascimento.

01 de Dezembro de 2025

Introdução

Este trabalho aprofunda a modelagem da variável `volume` através de três etapas: (1) Análise de estacionariedade (KPSS/ADF e decomposição STL) confirmando a necessidade de transformação logarítmica; (2) Implementação de validação *Walk-Forward* com horizonte $H = 4$ para avaliação realista de *baselines* e Suavização Exponencial (ETS); e (3) Desenvolvimento de modelos de Regressão (Lasso) e SARIMAX com variáveis exógenas (`inv`, `users`) e diagnóstico de resíduos (Ljung-Box).

Resumo da Implementação Anterior

A etapa inicial utilizou Regressão Linear Múltipla com transformação logarítmica e validação fixa (52 semanas).

- **Baseline:** *Seasonal Naive* (RMSE 4.62), assumindo repetição anual.
- **Dados:** Apenas histórico da série e calendário (sem variáveis externas).
- **Modelo Final:** Lasso (RMSE 2.72), selecionado pelo equilíbrio entre desempenho e controle de *overfitting*.

1 Análise Exploratória e Pré-Processamento

A validação das premissas de modelagem confirmou a necessidade de tratamento para volatilidade e não-estacionariedade.

1.1 Análise Descritiva e Diagnóstico de Estacionariedade

A distribuição (Figura 1) evidencia forte assimetria positiva, volatilidade e *outliers*. A série é não-estacionária (Figura 2), apresentando tendência de alta e heteroscedasticidade. A Função de Autocorrelação (ACF) na Figura 3 corrobora o diagnóstico através do decaimento lento ("memória longa").

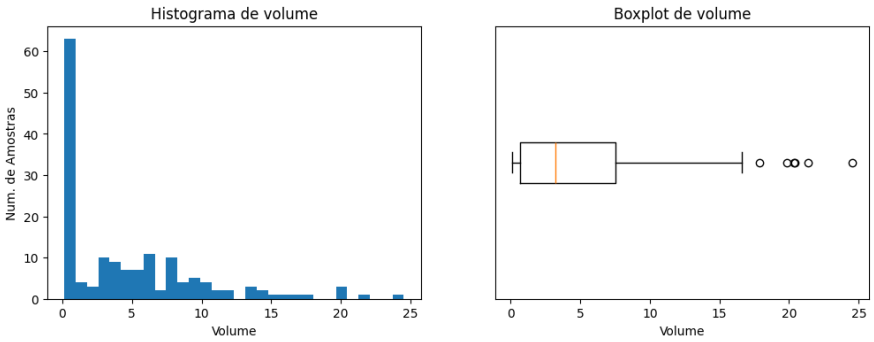

Figura 1: Histograma e Boxplot: assimetria e outliers evidentes.

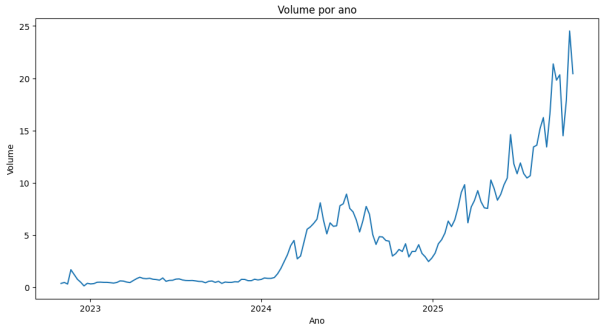

Figura 2: Série original: tendência e variância instável.

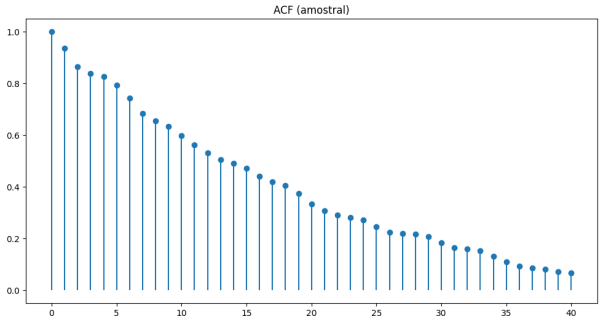

Figura 3: ACF original: decaimento lento indica não-estacionariedade.

1.2 Decomposição e Transformação Logarítmica

A decomposição STL (Figura 4) isolou a tendência de crescimento e a sazonalidade (52 semanas). Aplicou-se a transformação logarítmica ($\log(y)$) seguida de diferenciação. A Figura 5 demonstra a eficácia do tratamento: variância estabilizada e decaimento rápido nas funções ACF/PACF, apta para modelagem.

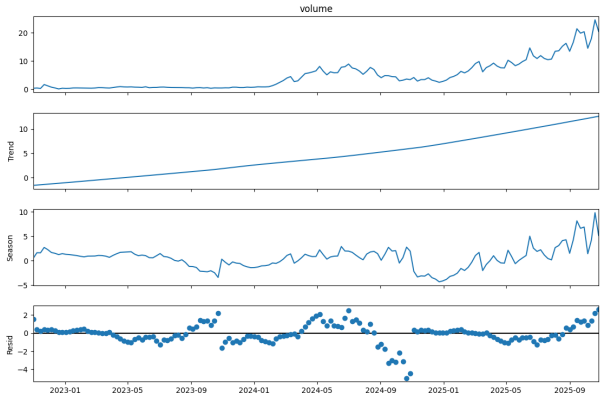

Figura 4: Decomposição STL da série volume.

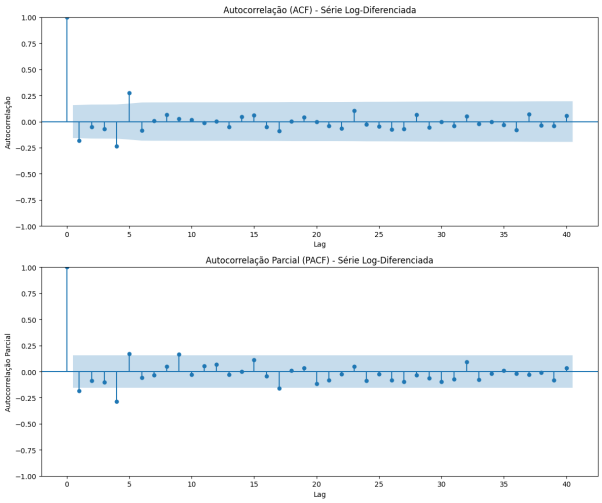

Figura 5: ACF/PACF após log e diferenciação.

2 Estratégia de Validação e Modelos de Referência

Substituiu-se a janela de teste fixa da fase anterior por uma estratégia dinâmica para maior robustez operacional.

2.1 Validação Walk-Forward (Janela Deslizante)

Implementou-se validação ***Walk-Forward*** com horizonte curto ($H = 4$). O processo simula a operação real: (1) treino com histórico disponível, (2) previsão de 4 semanas e (3) avanço mensal da janela para re-treinamento, avaliando a capacidade de adaptação do modelo a mudanças recentes.

2.2 Novos Baselines e Tendência Recente

A inclusão do método Drift alterou o foco da modelagem:

- **Seasonal Naive:** Desempenho ruim (RMSE 7.42), indicando inconsistência no padrão anual.
- **Drift:** Melhor baseline (RMSE **2.28**), sugerindo que a tendência de curto prazo supera a sazonalidade anual como preditor.

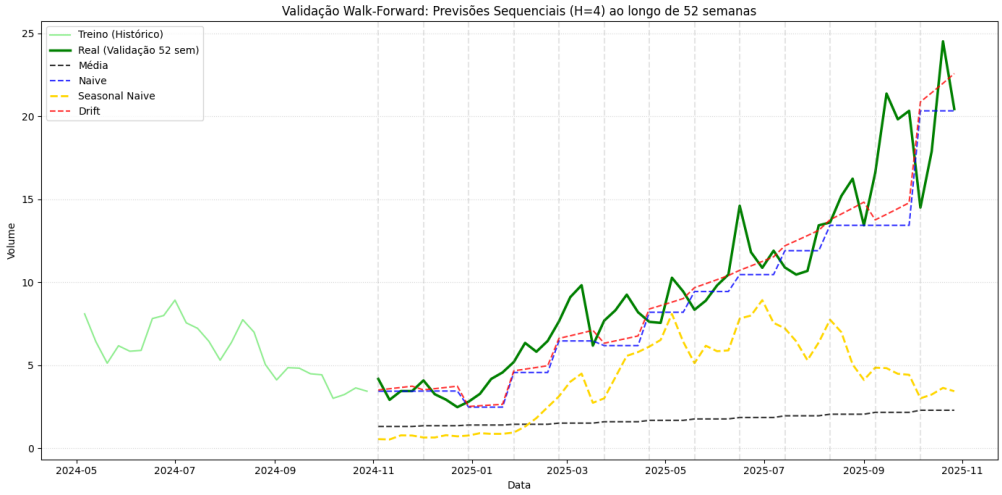

Figura 6: Baselines na validação Walk-Forward: destaque para o Drift.

2.3 Modelos de Suavização Exponencial (ETS)

Avaliamos a família Holt-Winters (Linear, Damped e Sazonal). O modelo **Sazonal** apresentou desempenho inferior, reforçando que impor uma sazonalidade rígida gera ruído. O **Holt Linear** obteve o melhor desempenho global (RMSE **2.16**), capturando a tendência local adaptativa melhor que os baselines.

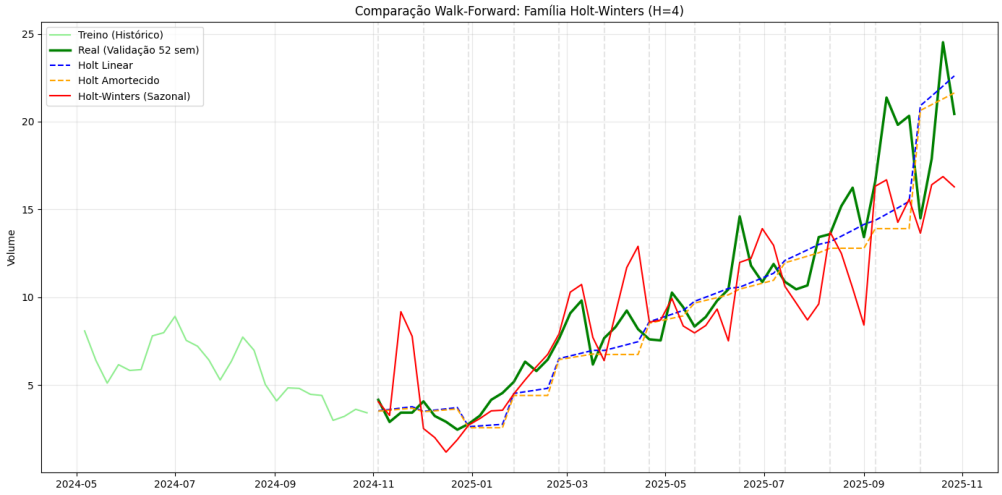

Figura 7: Performance dos modelos ETS: Holt Linear como benchmark.

3 Regressão Múltipla, Variáveis Exógenas e SARIMAX

Buscou-se superar o *benchmark* (Holt Linear: 2.16) incorporando variáveis externas (`inv`, `users`).

3.1 Engenharia de Features e Data Leakage

Devido ao horizonte $H = 4$, o uso de *lags* imediatos $(t - 1)$ de variáveis exógenas causaria *data leakage*, pois esses dados não estariam disponíveis para prever a 4ª semana. Utilizaram-se *lags* de 4 semanas (`ratiolag4`) para garantir a integridade temporal. O modelo linear inicial (MLR v1) falhou no teste Ljung-Box, indicando resíduos autocorrelicionados.

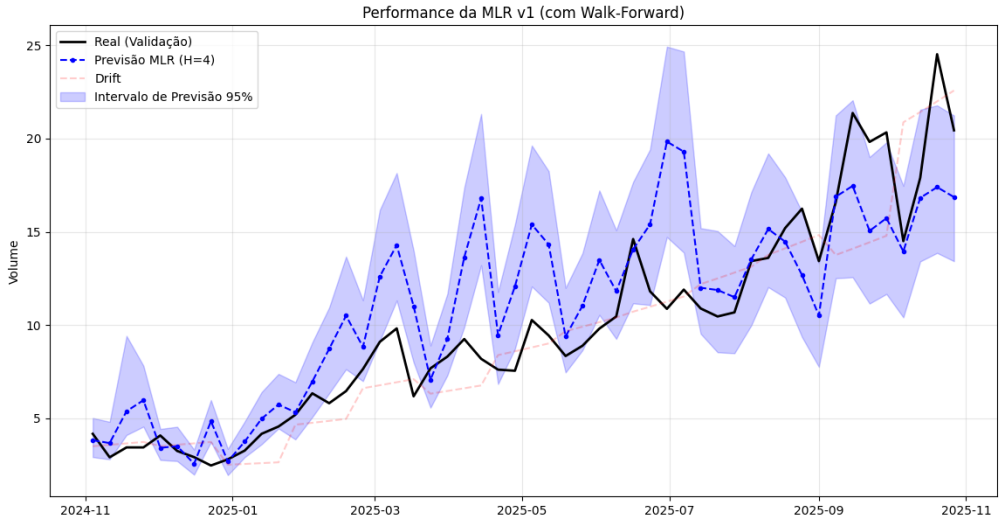

Figura 8: Previsão MLR v1 vs. Real.

3.2 Refinamento com Lasso (MLR v3)

Com regularização Lasso e inclusão de *lags* estratégicos, o MLR v3 atingiu RMSE **2.18**. Os resíduos foram aprovados no teste Ljung-Box (até *lag* 30), mas a análise por horizonte (Figura 9) revelou aumento da incerteza nas semanas mais distantes (2 e 4), validando a dificuldade de previsão no fim

do ciclo mensal.

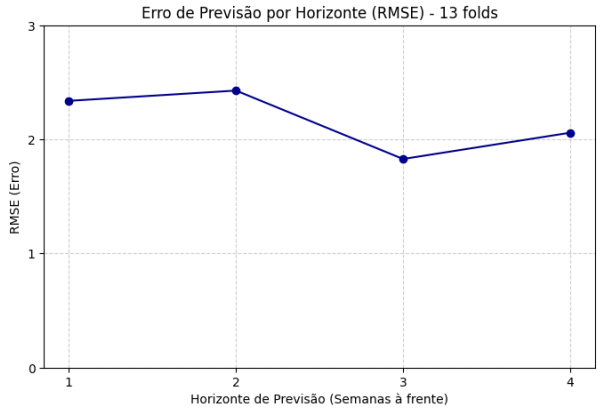

Figura 9: RMSE por Horizonte: erro cresce com a distância.

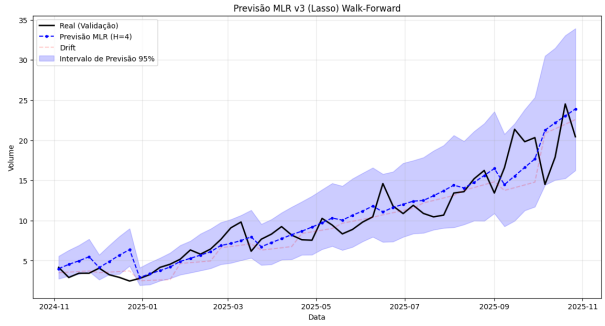

Figura 10: Previsão Lasso (MLR v3) com intervalos de confiança.

3.3 SARIMAX e Hipótese do 'Ano Base'

O SARIMAX (Auto-ARIMA com exógenas) não superou o RMSE do Holt Linear. Testou-se ainda a hipótese de que o "Ano 2"seria uma anomalia, usando o padrão estável do "Ano 1"como regressor exógeno. Apesar do apelo teórico, não houve ganho de performance, reforçando a superioridade da adaptação dinâmica sobre padrões históricos rígidos.

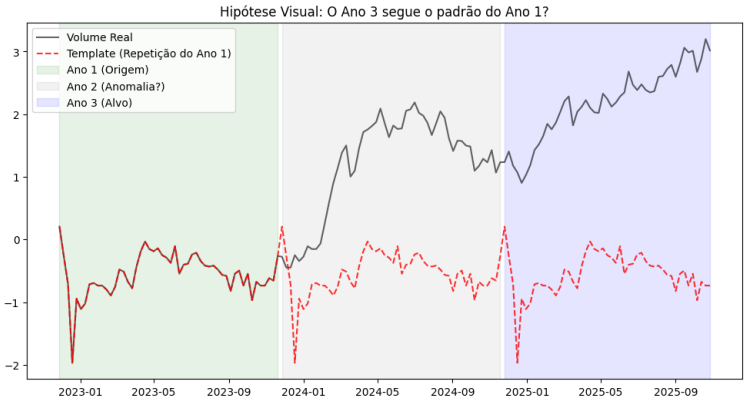

Figura 11: Comparação visual: Volume real vs. Template Ano 1.

Conclusão Final

Para o horizonte $H = 4$, a complexidade não resultou em maior acurácia. Apesar da validação rigorosa e uso de exógenas, o modelo de **Suavização Exponencial Holt Linear** (RMSE 2.16) prevaleceu. Conclui-se que a série é dominada pela inércia de tendência recente, onde o ajuste local supera a capacidade explicativa de padrões sazonais anuais (que se provaram inconsistentes) ou variáveis externas.