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.

05 de Outubro de 2025

1 Introdução

O objetivo deste trabalho é desenvolver um modelo para prever a série temporal da variável `volume`. A metodologia seguiu três etapas principais: (1) uma análise exploratória para identificar tendências, sazonalidade e a necessidade de transformações; (2) a definição de métricas de avaliação e a criação de modelos de *baseline* para referência ; e (3) a construção iterativa de um modelo de regressão linear múltipla com variáveis derivadas do tempo, buscando um equilíbrio entre performance preditiva e robustez estatística.

2 Análise Exploratória e Pré-Processamento

2.1 Análise Descritiva e Estacionariedade

A análise descritiva inicial da variável `volume` (Tabela 1) revelou uma forte assimetria positiva (1.03) e alta volatilidade, sugerindo a presença de picos e uma distribuição não normal.

Tabela 1: Estatísticas descritivas da variável `volume`.

Estatística	**Valor**	**Estatística**	**Valor**
Média	4.12	Assimetria	1.0339
Mediana	3.00	Valor Mínimo	0.14
Desvio Padrão	4.04	Valor Máximo	16.59

A análise visual da série (Figura 1, esquerda) confirma que ela não é estacionária, exibindo uma clara tendência de crescimento e variância crescente ao longo do tempo. O gráfico de autocorrelação (ACF) associado (Figura 1, direita) mostra um decaimento muito lento, um sinal clássico de não-estacionariedade que viola premissas importantes para a modelagem.

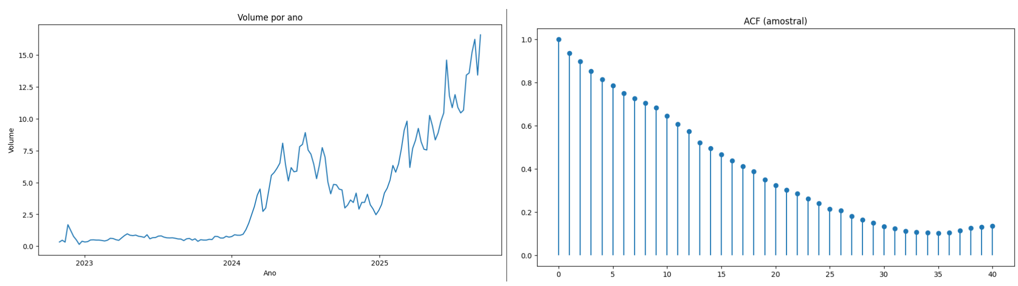

Figura 1: Série temporal original de `volume` (esquerda) e sua função de autocorrelação (ACF) (direita).

2.2 Transformação e Decomposição da Série

Dado o crescimento aparentemente exponencial e a variância instável, aplicamos uma transformação logarítmica para estabilizar a série.A Figura 2 demonstra a eficácia da transformação, resultando em uma série com comportamento mais regular e autocorrelações que decaem rapidamente.

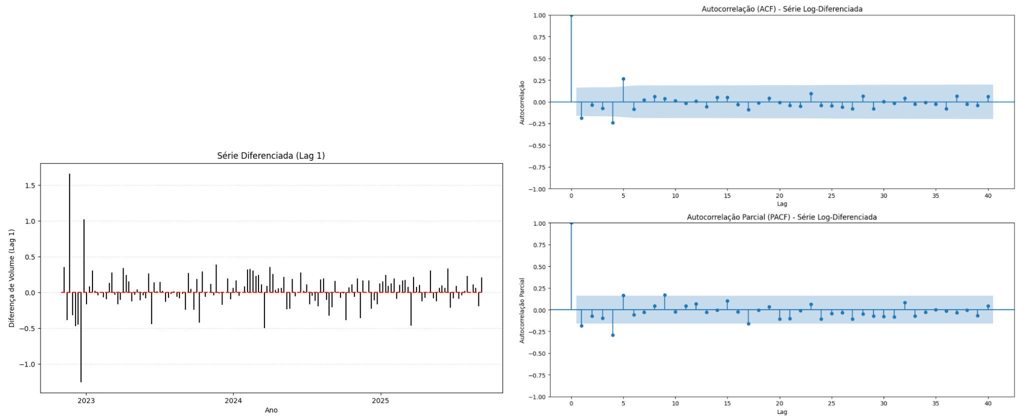

Figura 2: Série após transformação logarítmica (esquerda) e seu respectivo ACF (direita).

Para entender melhor suas componentes, a série transformada foi decomposta pelo método STL, com período sazonal de 52 semanas. A decomposição (Figura 3) confirmou a presença de uma tendência e de um padrão sazonal anual, que serão fundamentais para a criação de variáveis no modelo preditivo.

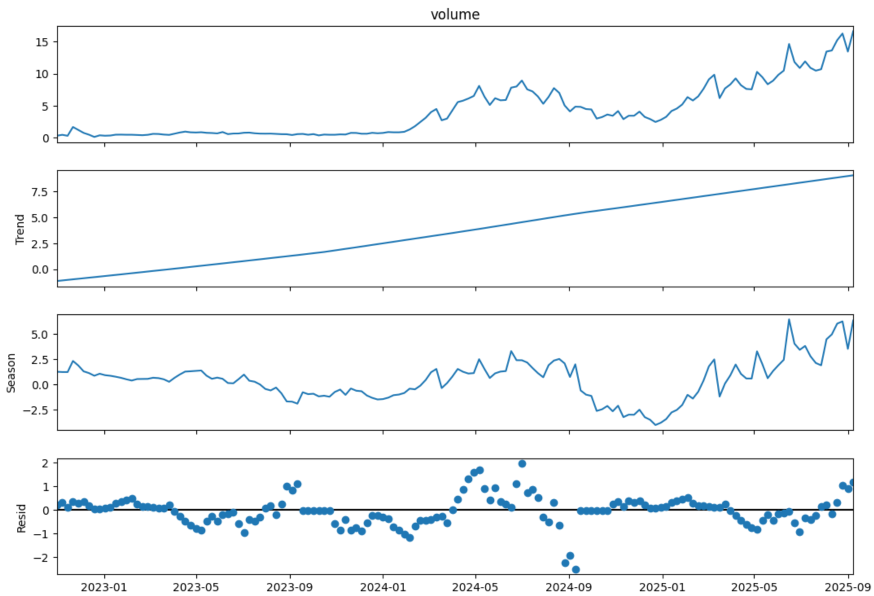

Figura 3: Decomposição da série `log(volume)` em tendência, sazonalidade e resíduos.

3 Metodologia de Modelagem

3.1 Métricas de Avaliação e Estratégia de Validação

O desempenho dos modelos foi avaliado em um conjunto de teste composto pelas últimas 52 semanas dos dados.As métricas utilizadas foram o Erro Absoluto Médio (MAE), a Raiz do Erro Quadrático

Médio (RMSE) e o Erro Percentual Absoluto Médio (MAPE). Como a modelagem foi feita na escala logarítmica, as previsões foram revertidas para a escala original (via exponenciação) antes do cálculo dos erros.

3.2 Modelos Baseline

Para estabelecer um ponto de referência, foram implementados três modelos simples:

- **Modelo da Média:** Prevê a média histórica do conjunto de treino.
- **Modelo Naive:** Prevê que o próximo valor será igual ao último observado.
- **Modelo Seasonal Naive:** Prevê que o valor será igual ao observado no mesmo período do ciclo anterior (neste caso, 52 semanas antes).

Os resultados (Tabela 2) indicam que o modelo **Seasonal Naive** apresentou o melhor RMSE (4.62), sendo mais eficaz em evitar grandes erros. Por essa razão, ele foi adotado como o principal benchmark a ser superado pelos modelos de regressão.

Tabela 2: Resultados dos modelos de Baseline no conjunto de teste.

Modelo	**MAE**	**RMSE**	**MAPE**
Mean	6.4976	7.5939	0.7906
Naive	3.7627	4.8566	0.4440
Seasonal Naive	4.0827	4.6211	0.5958

3.3 Construção dos Modelos de Regressão Múltipla (MLR)

Foram desenvolvidas três versões de um modelo MLR, adicionando complexidade de forma iterativa:

- **MLR v1:** Modelo inicial com features de tendência (`time_index`), sazonalidade (`week_of_year`, `month`) e lags da variável alvo para capturar dependência temporal.
- **MLR v2:** Adição de features cíclicas (seno/cosseno para mês e semana), mais lags para capturar dependências de curto e longo prazo, e uma média móvel de 4 semanas.
- **MLR v3 (Lasso):** Simplificação do v2 com regularização L1 (Lasso) para penalizar coeficientes grandes, reduzir o overfitting e selecionar as features mais relevantes.Lags intermediários e a média móvel foram removidos.

4 Resultados e Análise dos Modelos

4.1 MLR v1: Baseline de Regressão

O primeiro modelo de regressão superou significativamente os baselines, com um **RMSE de 2.60**. No entanto, a análise de resíduos (Figura 4) revelou um problema: o teste de Ljung-Box apresentou p-valores baixos (e.g., 0.0017 para lag 10), indicando que os resíduos não eram independentes e que o modelo não capturou toda a estrutura de autocorrelação.

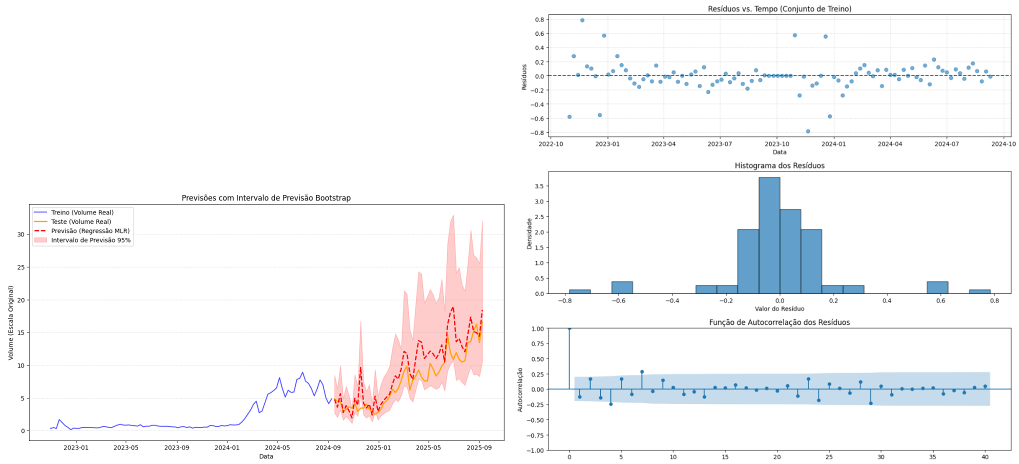

Figura 4: Previsão do MLR v1 (esquerda) e diagnóstico dos resíduos (direita).

4.2 MLR v2: Complexidade e Overfitting

O MLR v2 foi construído para corrigir as deficiências do v1. Embora tenha resolvido o problema da autocorrelação nos resíduos (teste de Ljung-Box com p-valores altos, >0.2), suas métricas de erro aumentaram drasticamente (**RMSE de 10.97**). Isso indicou que o excesso de complexidade levou a um overfitting, onde o modelo se ajustou demais ao ruído dos dados de treino e perdeu capacidade de generalização.

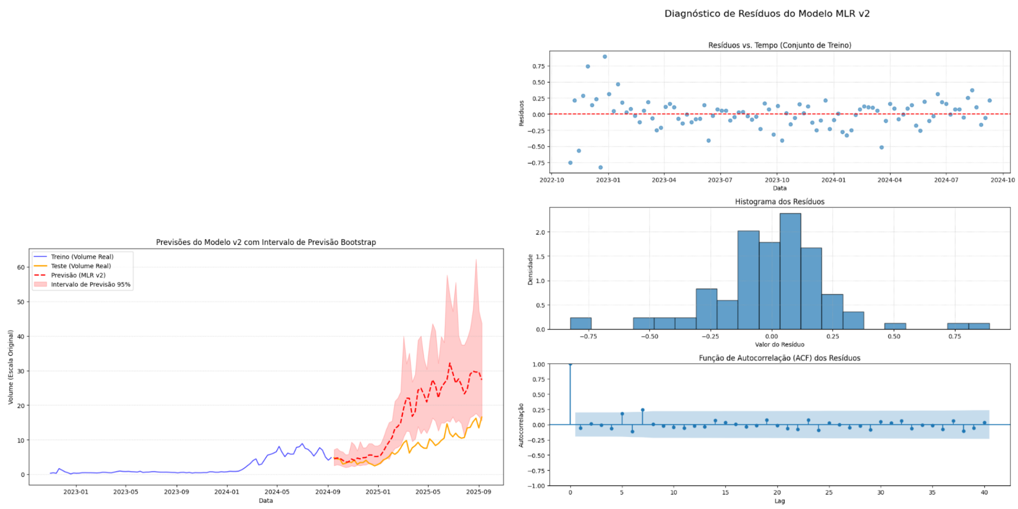

Figura 5: Previsão do MLR v2 (esquerda) e diagnóstico dos resíduos (direita).

4.3 MLR v3 (Lasso): O Modelo Equilibrado

O MLR v3 buscou um meio-termo, simplificando as features do v2 e aplicando regularização Lasso. O resultado foi um modelo com **RMSE de 2.72**, ligeiramente superior ao v1, mas com resíduos estatisticamente válidos (Figura 6).O teste de Ljung-Box confirmou a independência dos resíduos (p-valores >0.2), e os gráficos mostraram distribuição normal e homocedasticidade.

Tabela 3: Comparativo de performance dos modelos de regressão.

Modelo	MAE	RMSE	MAPE (%)	Diagnóstico Ljung-Box
MLR v1	**1.94**	**2.61**	**28.32**	Falhou (resíduos correlacionados)
MLR v2	8.86	10.98	101.80	Aprovado (mas com overfitting)
MLR v3 (Lasso)	2.12	2.72	29.18	**Aprovado (modelo robusto)**

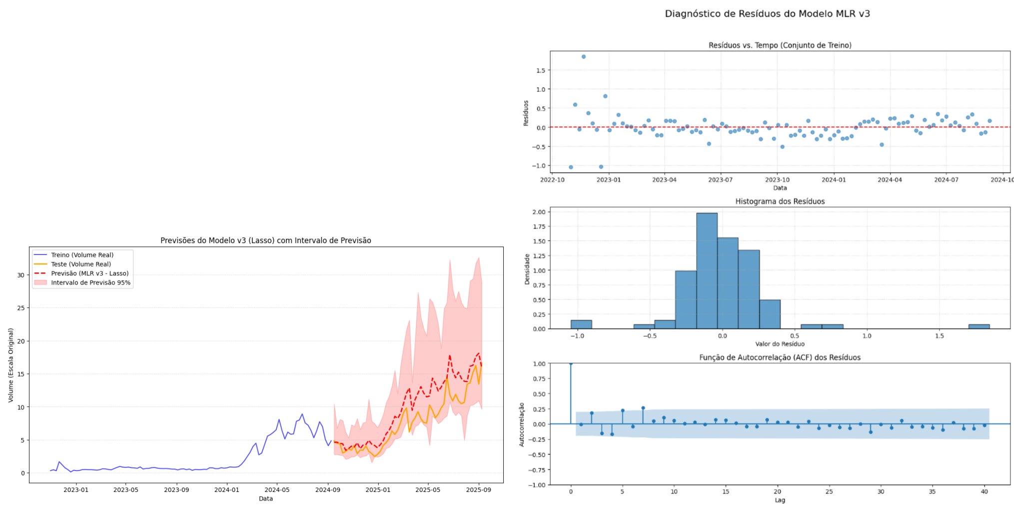

Figura 6: Previsão do MLR v3 (Lasso) (esquerda) e diagnóstico dos resíduos (direita).

5 Conclusão

Este trabalho demonstrou a importância de um processo iterativo na modelagem de séries temporais. Embora o modelo MLR v1 tenha apresentado as melhores métricas de erro, ele era estatisticamente falho. O modelo MLR v3 (Lasso), por outro lado, representa um **trade-off clássico**: sacrificamos uma pequena fração da performance preditiva em troca de um modelo muito mais robusto, confiável e generalizável.Ele evita o overfitting do v2 e corrige as deficiências estatísticas do v1, representando um excelente ponto de partida a partir do qual futuras melhorias, como o uso de modelos nãolineares, podem ser exploradas com segurança.