UNIVERSIDADE DE SÃO PAULO
ESCOLA POLITÉCNICA
PROGRAMA DE EDUCAÇÃO CONTINUADA

Fernando Gomes Papi

Um modelo de previsão de volatilidade para o Bitcoin utilizando dados on-chain e seleção de variáveis

São Paulo

2025

Fernando Gomes Papi

Um modelo de previsão de volatilidade para o Bitcoin utilizando dados on-chain e seleção de variáveis

Monografia apresentada ao Programa de Educação Continuada da Escola Politécnica da Universidade de São Paulo para obtenção do título de Especialista em Engenharia Financeira

Universidade de São Paulo

Escola Politécnica

Programa de Educação Continuada - Engenharia Financeira

Orientador: André Cury Maialy, PhD

São Paulo

2025

Agradecimentos

Agradeço a todo o corpo docente do MBA em Engenharia Financeira da Escola Politécnica da Universidade de São Paulo, com os quais tive o prazer de aprender durante este curso.

Em especial ao professor André Cury Maialy, por transmitir grande conhecimento e pela dedicação no curso de Apreçamento de Derivativos, e pelo suporte e paciência na orientação deste trabalho.

A todos os colegas de turma, pelas conversas e contribuições que tornaram esta experiência ainda mais rica.

Agradeço à minha família, pai, mãe e irmãos pelo incentivo de sempre, sem o qual eu nunca teria chegado a lugar algum.

E por fim, agradeço e dedico este trabalho à minha esposa Juliana, pelo incentivo e apoio na conclusão deste curso, e principalmente, no desenvolvimento deste trabalho, e por ser o motivo de que tudo isto vale a pena.

Resumo

Este trabalho apresenta um modelo de previsão de volatilidade para o Bitcoin utilizando dados on-chain e técnicas de seleção de variáveis, com foco na aplicação de métodos de aprendizado de máquina, especificamente o XGBoost, para prever a variância realizada (RV) deste ativo. O estudo investiga a viabilidade de utilizar dados de mercado, dados on-chain do Bitcoin (BTC) e do Dólar Tether (USDT) como regressores externos para prever a volatilidade futura. O objetivo principal é comparar o desempenho do XGBoost com o modelo HAR (Heterogeneous Autoregressive Model), amplamente utilizado na literatura de previsão de volatilidade, e avaliar se a incorporação de dados on-chain traz ganhos preditivos significativos.

A metodologia adotada envolve a seleção rigorosa de variáveis utilizando quatro abordagens complementares: Mean Decrease Impurity (MDI), Mean Decrease Accuracy (MDA), SHapley Additive exPlanations (SHAP) e Mutual Information (MI). Essas técnicas permitiram identificar as variáveis mais relevantes para a previsão da volatilidade, descartando ruídos e redundâncias. O conjunto de dados inclui informações de negociação do par BTC/USDT na corretora Binance, além de dados on-chain do Bitcoin e do USDT, coletados entre 2020 e 2024.

Os resultados mostram que o XGBoost superou o HAR em todas as métricas avaliadas, com destaque para o Skill Score de 0.968 na previsão de volatilidade para 60 dias, indicando uma redução de 93,8% no erro quadrático médio (MSE) em relação ao HAR. Além disso, o modelo com retreino diário demonstrou maior adaptabilidade a mudanças recentes na dinâmica de volatilidade, reduzindo o erro absoluto médio (MAE) em 16,7% e o erro percentual absoluto médio (MAPE) em 31 pontos percentuais em comparação ao HAR.

A análise SHAP revelou que variáveis da rede USDT responderam por 73% da importância global do modelo, destacando o papel do USDT como um canal de transmissão de volatilidade entre ativos. Isso valida a hipótese de que stablecoins, como o USDT, desempenham um papel crucial na dinâmica de volatilidade do Bitcoin.

O estudo conclui que a integração de aprendizado de máquina adaptativo, seleção rigorosa de variáveis e dados on-chain de stablecoins oferece um paradigma promissor para a modelagem de risco em ativos descentralizados. Além disso, sugere direções futuras de pesquisa, como a extensão do modelo para outras stablecoins algorítmicas e redes DeFi, a incorporação de variáveis macroeconômicas e o uso de novas arquiteturas de redes neurais para melhorar a previsão de volatilidade em mercados cripto.

Palavras-chave: Bitcoin, Volatilidade, Previsão, Aprendizado de Máquina, Dados On-Chain, XGBoost, HAR, USDT.

Abstract

This work presents a volatility forecasting model for Bitcoin using on-chain data and variable selection techniques, focusing on the application of machine learning methods, specifically XGBoost, to predict Bitcoin's realized variance (RV). The study investigates the feasibility of using market data, Bitcoin (BTC) on-chain data, and Tether (USDT) on-chain data as external regressors to forecast future volatility. The main objective is to compare the performance of XGBoost with the HAR (Heterogeneous Autoregressive Model), widely used in volatility forecasting literature, and to assess whether the inclusion of on-chain data provides significant predictive gains.

The methodology involves rigorous variable selection using four complementary approaches: Mean Decrease Impurity (MDI), Mean Decrease Accuracy (MDA), SHapley Additive exPlanations (SHAP), and Mutual Information (MI). These techniques allowed for the identification of the most relevant variables for volatility prediction, eliminating noise and redundancies. The dataset includes trading data from the BTC/USDT pair on the Binance exchange, as well as on-chain data from Bitcoin and USDT, collected between 2020 and 2024.

The results show that XGBoost outperformed HAR across all evaluated metrics, with a notable Skill Score of 0.968 for 60-day volatility forecasting, indicating a 93.8% reduction in mean squared error (MSE) compared to HAR. Additionally, the model with daily retraining demonstrated greater adaptability to recent changes in volatility dynamics, reducing the mean absolute error (MAE) by 16.7% and the mean absolute percentage error (MAPE) by 31 percentage points compared to HAR.

SHAP analysis revealed that variables from the USDT network accounted for 73% of the model's global importance, highlighting USDT's role as a channel for volatility transmission between assets. This validates the hypothesis that stablecoins, such as USDT, play a crucial role in Bitcoin's volatility dynamics.

The study concludes that the integration of adaptive machine learning, rigorous variable selection, and on-chain stablecoin data offers a promising paradigm for risk modeling in decentralized assets. Furthermore, it suggests future research directions, such as extending the model to other algorithmic stablecoins and DeFi networks, incorporating macroeconomic variables, and using new neural network architectures to improve volatility forecasting in crypto markets.

Keywords: Bitcoin, Volatility, Forecasting, Machine Learning, On-Chain Data, XGBoost, HAR, USDT.

Lista de Figuras

Conteúdo

1 Introdução

A volatilidade desempenha um papel central no gerenciamento de riscos e na precificação de ativos financeiros, pois flutuações de preços afetam diretamente a exposição a perdas e a avaliação de retornos. Em particular, a volatilidade impacta a precificação de derivativos, uma vez que a incerteza em relação à trajetória dos preços do ativo subjacente influencia tanto o custo de proteção quanto as oportunidades de arbitragem. Nos últimos anos, a maior disponibilidade de dados de alta frequência viabilizou a modelagem e previsão da Variância Realizada (RV), reconhecida como uma medida confiável de volatilidade (Andersen and Bollerslev (1998)). Dessa maneira, análises e técnicas eficazes de gestão da volatilidade possibilitam que investidores e gestores de portfólio tomem decisões embasadas, reduzindo riscos e potencializando ganhos no longo prazo.

O Bitcoin e outras criptomoedas, embora sejam ativos relativamente recentes no cenário financeiro, têm despertado enorme interesse devido aos expressivos retornos observados em seus estágios iniciais de negociação, acompanhados de riscos igualmente elevados resultantes de sua alta volatilidade. A expansão dos contratos derivativos em criptomoedas — com o Bitcoin figurando entre os maiores em volume de negociações — sinaliza o amadurecimento do mercado e a busca crescente por estratégias de proteção e alavancagem. Uma característica marcante do mercado cripto é a negociação contínua, 24 horas por dia, 7 dias por semana, gerando um volume maciço de informações sobre preço, volume e liquidez. Ao mesmo tempo, a tecnologia *blockchain* introduziu o conceito de "dados on-chain", que trazem transparência quanto às transações e à atividade de diversos participantes da rede. Tal abundância de dados, combinada à forte oscilação de preços, configura um ambiente desafiador, mas fértil, para pesquisas e investimentos orientados pela análise da volatilidade.

Em paralelo, surgiram ferramentas especializadas em analisar a atividade dos agentes nas redes *blockchain*, fazendo com que esses "dados on-chain" ganhassem relevância na investigação de padrões comportamentais e na antecipação de tendências de mercado. Nesse contexto, o advento das *stablecoins* — ativos cujo preço é lastreado em moedas fiduciárias, como o dólar — agregou ao ecossistema cripto novas possibilidades de arbitragem e gestão de risco. Dentre as stablecoins disponíveis, destaca-se o Dólar Tether (USDT), que por meio de alto volume de negociação e ampla adoção, tornou-se um elo central na dinâmica de entradas e saídas do mercado cripto. Tais movimentações de USDT, quando observadas em redes *blockchain*, podem preceder movimentos significativos no preço e na volatilidade de ativos como o Bitcoin, visto que influenciam a liquidez e a percepção de risco sistêmico dentro do universo das criptomoedas.

1.1 Objetivos de Pesquisa

Este trabalho tem como objetivo principal verificar a viabilidade de modelos de Aprendizado de Máquina — em especial o *XGBoost* — para a previsão da variância realizada do Bitcoin, negociado na maior corretora do mercado, a Binance, entre 2020 e 2024. O estudo investiga até que ponto dados de negociação, dados on-chain do Bitcoin (BTC) e dados on-chain do Dólar Tether (USDT) podem servir como regressores externos, carregando informações preditivas acerca da volatilidade futura. Pretende-se, ainda, comparar o desempenho do *XGBoost* com o modelo HAR, bastante utilizado na literatura de previsão de volatilidade, a fim de avaliar se a incorporação de *features* on-chain realmente fornece ganhos substantivos de previsão. Em última análise, busca-se elucidar o papel do USDT como potencial vetor de transmissão de volatilidade na rede Bitcoin, bem como fornecer evidências sobre a importância de um conjunto amplo de variáveis de mercado e on-chain na modelagem de riscos e oportunidades em mercados cripto.

1.2 Estrutura do Documento

Este documento organiza-se em cinco capítulos, além desta Introdução. No Capítulo 2, apresenta-se a Revisão Teórica e Bibliográfica sobre medidas de volatilidade, estudos prévios envolvendo variância realizada e fundamentos sobre criptomoedas, dados on-chain e stablecoins. As bases conceituais do modelo HAR e as discussões que o aproximam de modelos de aprendizado de máquina também são apresentadas nesse capítulo, estabelecendo o referencial teórico do trabalho.

O Capítulo 3 descreve a Metodologia utilizada, detalhando os dados de mercado (extraídos da corretora Binance) e os dados on-chain (do Bitcoin e do USDT). São explicados os critérios de limpeza, transformações das variáveis e o processo de seleção de *features*, que combina distintas abordagens de importância de variáveis (MDI, MDA, SHAP e MI) para mitigar ruídos e colinearidades em um universo amplo de informações. Apresenta-se, ainda, a formulação do modelo HAR e do *XGBoost*, as métricas de avaliação e a forma de particionamento *out-of-sample* (*OOS*).

No Capítulo 4, procede-se à Aplicação da Metodologia e apresentação dos Resultados, examinando tanto as análises *in-sample* quanto as previsões *out-of-sample*, em janelas de 7, 30 e 60 dias de volatilidade. São discutidas comparações quantitativas entre o *XGBoost* (com e sem retreino) e o modelo HAR, evidenciando ganhos estatisticamente significativos em diversas métricas. Também são explorados aspectos relacionados ao teste de Diebold-Mariano para aferir a significância das diferenças de desempenho.

Por fim, o Capítulo 5 e 6 trazem Resultados e a Conclusão, na qual se sintetizam as contribuições do estudo, enfatizando a relevância das variáveis on-chain do BTC e do USDT no incremento das previsões de volatilidade do Bitcoin. Avaliam-se as limitações da abordagem adotada, bem como se indicam direções para pesquisas futuras, tais como o uso de metodologias mais avançadas e a inclusão de outras fontes de dados.

2 Revisão Teórica e Bibliográfica

O objetivo deste capítulo é fornecer um embasamento teórico sobre os conceitos que fundamentam esta pesquisa.

2.1 Revisão de literatura

A volatilidade é um assunto amplamente discutido na literatura. Previsões de volatilidade encontram utilidade em gerenciamento de riscos, apreçamento e *hedging* de derivativos, *market making* e composição de portfólios, entre outras aplicações, segundo Engle and Patton (2007). Wilmott (2009) afirma que a volatilidade é difícil de medir e mais ainda de se prever, mas é uma das principais variáveis nos modelos de precificação de derivativos. A volatilidade é difícil de se medir porque geralmente sua formulação matemática requer dados históricos de retorno do ativo para ser calculada. Mas a volatilidade em si é uma medida instantânea, não histórica. Neste trabalho, será utilizada a medida de volatilidade apresentada na seção 2.2.1, sendo a medida desenvolvida por Andersen et al. (2001), adaptada ao ciclo ininterrupto de negociação dos cripto-ativos.

Com relação a modelos de previsão de volatilidade, principalmente utilizando-se de *machine learning*, o trabalho apresentado por Li and Tang (2024) propõe uma abordagem semelhante de comparação do uso de modelos não lineares com modelos estabelecidos, ainda que sua aplicação tenha sido focada em um universo grande de ativos do mercado convencional (S&P500). O trabalho desenvolvido por Branco et al. (2024) também possui contribuições significativas quanto à comparação de modelos lineares e não lineares para a previsão de volatilidade. Este trabalho também utiliza o modelo HAR Corsi (2009) como base de comparação.

O problema de predição de volatilidade para o mercado de cripto-ativos também dispõe de abundante literatura. O trabalho desenvolvido por Khan et al. (2023) apresenta uma comparação de diferentes modelos de redes neurais para a previsão de volatilidade de variadas criptomoedas, atingindo um resultado na faixa de 0.014 para o RMSE, similar ao demonstrado no Capítulo 5. O trabalho apresentado em Brauneis and Sahiner (2024) adiciona informação de sentimento , resultando em um poder preditivo ligeiramente superior (RMSE de 0.007). No entanto, a utilização de dados on-chain ainda é relativamente limitada. Um trabalho nesta direção foi apresentado por Chi et al. (2024), desenvolvendo uma estratégia de negociação baseado em um modelo de previsão de volatilidade usando dados on-chain.

O presente trabalho assume familiaridade com regressões a partir de modelos baseados em árvores de decisão, como o *Random Forest*, e especificamente, o XGBoost, desenvolvido por Chen and Guestrin (2016). Além do trabalho seminal sobre este modelo de *machine learning*, o trabalho apresentado em Nielsen (2016) traz uma revisão e análise aprofundada deste algoritmo

de aprendizado computacional.

Há literatura relevante também na área de seleção eficaz de variáveis. O presente trabalho tira inspiração do trabalho desenvolvido em Guyon and Elisseeff (2003), bem como das ideias expostas por de Prado (2018). Em Guyon and Elisseeff (2003), os autores concluem que *"Para esse fim, recomendamos o uso de um preditor linear de sua escolha (por exemplo, um SVM linear) e a seleção de variáveis de duas maneiras alternativas: (1) com um método de classificação de variáveis usando um coeficiente de correlação ou informação mútua; (2) com um método de seleção de subconjuntos iterativo que realiza seleção recursivamente, para frente (adicionando variáveis uma a uma) ou para trás (removendo variáveis uma a uma) ou com atualizações multiplicativas."*. O método presente neste trabalho, desenvolvido no Capítulo 3 tem embasamento semelhante, como será demonstrado.

2.2 Conceitos básicos

O intuito desta sub-seção é introduzir os conceitos base que formam este estudo, desde o modelo de retorno dos preços utilizado, o modelo de comparação de base (HAR), e as diversas formas de medição da importância de variáveis, que formam a base para a metodologia de seleção de variáveis apresentada no Capítulo 3.

2.2.1 Variância Realizada

Neste estudo, o objetivo é prever a variância realizada (Realized Variance, RV), descrita em Andersen et al. (2001), a qual constitui um estimador consistente da variação quadrática do processo de log-preço ao longo de um determinado período. Formalmente, seja p_t o logaritmo natural do preço do ativo no dia t, assumindo que o log-preço segue um processo genérico de difusão com saltos:

$$p_t = \int_0^t \mu_s \, ds + \int_0^t \sigma_s \, dW_s + J_t, \tag{2.1}$$

onde μ_t e σ_t denotam, respectivamente, os processos de drift e de volatilidade difusiva, W é um movimento Browniano padrão, J é um processo de saltos puros, e o intervalo de tempo unitário corresponde a um dia de negociação. É natural estender essa notação para preços intradiários usando p_t, $p_{t+1/n}$, $\ldots$, p_{t+1}, assumindo que os preços são observados em $n+1$ intervalos de tempo igualmente espaçados do dia t até o dia $t+1$. A variância realizada diária anualizada, calculada a partir da soma dos retornos quadráticos amostrados com frequência de 5 minutos dentro de um dia de negociação, é dada por:

$$\mathrm{RV}_t^d = 365 \times \sum_{i=1}^{n} r_{t-1+\frac{i}{n}}^2. \tag{2.2}$$

onde $r_{t-1+\frac{i}{n}} = p_{t-1+\frac{i}{n}} - p_{t-1+\frac{i-1}{n}}$ representa o retorno logarítmico ao longo do i-ésimo intervalo de tempo do dia t. Como o mercado de criptomoedas funciona continuamente 24 horas por dia, 365 dias por ano, estão incluídos os retornos quadráticos do período noturno na estimativa diária de RV. Conforme demonstrado em Andersen et al. (2001) , a RV é um estimador consistente da variação quadrática quando o número de intervalos n tende ao infinito. Para horizontes mais longos (por exemplo, semanal, mensal e trimestral), pode-se estimar RV pela média das RV diárias nos intervalos correspondentes. Formalmente, a RV de h dias à frente é definida como:

$$\mathrm{RV}_{t+1}^{t+h} = \frac{1}{h}\sum_{i=1}^{h}\mathrm{RV}_{t+i}^{d}, \tag{2.3}$$

em que $h = 7, 30, 60$ corresponde, respectivamente, à RV semanal, mensal e bimestral. O objetivo de pesquisa é desenvolver um modelo preditivo mais eficaz para as RV semanais, mensais e bimestrais. Para o cálculo empírico de RV, utilizamos a frequência de amostragem de cinco minutos, comumente adotada na literatura de volatilidade realizada, como exposto em Liu et al. (2015). A frequência de 5 minutos é comumente adotada para equilibrar a redução do erro de microestrutura (amostragem em alta frequência dos retornos) e a captura precisa da variação quadrática.

2.2.2 Modelo XGBoost (Extreme Gradient Boosting)

O XGBoost (Extreme Gradient Boosting), proposto por Chen and Guestrin (2016), é um algoritmo de aprendizado de máquina que combina múltiplas árvores de decisão de forma sequencial para corrigir progressivamente erros residuais. Projetado para eficiência e precisão em grandes volumes de dados, destaca-se em problemas de regressão complexos como previsão de volatilidade com centenas de variáveis preditoras.

Sua arquitetura baseia-se em três pilares principais:

- **Aprendizado Adaptativo**: Cada nova árvore foca nas instâncias onde as previsões anteriores falharam, refinando iterativamente o modelo,
- **Controle de Complexidade**: Mecanismos internos de regularização previnem sobreajuste (*overfitting*), mesmo com dezenas de *features* correlacionadas,
- **Otimização Computacional**: Processamento paralelo e técnicas de armazenamento eficiente permitem treinar modelos em escala sem hardware especializado.

Diferentemente de abordagens lineares como o HAR, o XGBoost não assume relações pré-definidas entre variáveis. Isso permite capturar:

- Padrões não lineares (ex.: impacto assimétrico de choques positivos vs. negativos na volatilidade),
- Interações complexas entre *lags* de diferentes horizontes temporais,
- Efeitos de limiar (ex.: volumes de negociação só impactam a volatilidade acima de um nível crítico).

Na prática financeira, três características são particularmente relevantes:

- **Integração de Dados Multifonte**: Combina naturalmente séries temporais (ex.: RV diária), variáveis macroeconômicas (ex.: taxa de juros) e dados não estruturados (ex.: sentimento de notícias),
- **Seleção Automática de Variáveis**: Identifica os *lags* e *features* exógenas mais preditivos, descartando redundâncias,
- **Atualização Dinâmica**: Recalibração eficiente diante de novos dados, crucial em mercados voláteis.

Sua principal limitação é a interpretabilidade: enquanto modelos como o HAR oferecem coeficientes linearmente quantificáveis, o XGBoost opera como uma *caixa cinza*, onde a importância das variáveis é inferida indiretamente por métricas como frequência de uso nas divisões das árvores. Ferramentas complementares (ex.: SHAP values) mitigam esta questão ao quantificar contribuições marginais das *features*.

A escolha pelo XGBoost justifica-se quando:

- Relações não lineares ou interações entre variáveis são suspeitas, mas difíceis de especificar a priori,
- O conjunto de dados inclui preditores heterogêneos (ex.: dados fundamentais, técnicos e comportamentais),
- A capacidade de processamento permite trade-offs entre complexidade e velocidade computacional.

Em comparação direta com o HAR, seu valor agregado reside na flexibilidade para modelar ambientes de mercado não estacionários, onde pressupostos lineares e de média móvel se mostram inadequados. Contudo, exige validação rigorosa via métricas out-of-sample e técnicas como *early stopping* para garantir que ganhos de performance não decorram de artefatos estatísticos.

2.2.3 Modelo HAR (Heterogeneous Autoregressive Model) - Benchmark

O Modelo HAR (*Heterogeneous Autoregressive Model*), proposto por Corsi (2009), é uma abordagem amplamente utilizada para prever a variância realizada (RV) em séries temporais financeiras de alta frequência. Sua estrutura captura a heterogeneidade de horizontes temporais na formação de expectativas de risco, refletindo o comportamento diferenciado de agentes que operam em escalas diárias, semanais e mensais.

A formulação canônica do HAR é dada por:

$$RV_{t+1} = \beta_0 + \beta_1 RV_t^d + \beta_2 RV_t^w + \beta_3 RV_t^m + \epsilon_{t+1}, \tag{2.4}$$

onde:

- RV_{t+1}: Variância realizada no dia $t+1$ (dependente).
- $RV_t^d = RV_t$: Variância realizada no dia t (horizonte diário).
- $RV_t^w = \frac{1}{5}\sum_{i=0}^{4} RV_{t-i}$: Média móvel da variância realizada nos últimos 5 dias (horizonte semanal).
- $RV_t^m = \frac{1}{21}\sum_{i=0}^{20} RV_{t-i}$: Média móvel da variância realizada nos últimos 21 dias (horizonte mensal).
- $\beta_0, \beta_1, \beta_2, \beta_3$: Coeficientes estimados por mínimos quadrados ordinários (MQO).
- ϵ_{t+1}: Termo de erro, assumido estacionário e não correlacionado serialmente.

A principal contribuição do HAR reside em sua capacidade de aproximar a dinâmica de memória longa da variância realizada por meio de uma estrutura autoregressiva hierárquica Corsi (2009). Essa característica o torna particularmente adequado para séries financeiras, onde a dependência temporal multiescala é ubíqua.

A linearidade do modelo permite analisar diretamente a contribuição relativa de cada horizonte temporal (β_1, β_2, β_3) para a previsão. Por exemplo, um β_2 estatisticamente significativo indica que o componente semanal da RV possui poder preditivo incremental sobre o componente diário.

E também por ser um modelo linear, o HAR demanda recursos computacionais mínimos, facilitando sua implementação em grandes conjuntos de dados ou para validação cruzada iterativa.

A comparação com o HAR será crítica para validar avanços metodológicos. Essa abordagem evita a complexidade excessiva e garante que ganhos aparentes de desempenho não decorram de *overfitting*.

Em resumo, o HAR não apenas fornece previsões economicamente interpretáveis, mas também estabelece um benchmark rigoroso para avaliar modelos preditivos alternativos. Sua simplicidade é, paradoxalmente, sua maior virtude em um campo onde a sofisticação matemática nem sempre se traduz em utilidade prática.

2.2.4 *Feature Importance* - MDI

A *feature importance* - ou importância da variável - constitui um conceito fundamental na análise de modelos preditivos, permitindo identificar a contribuição relativa de cada variável para a construção do modelo e a precisão de suas previsões. No caso mais simples, como em uma regressão linear, a importância de uma variável manifesta-se através da influência total que ela exerce na magnitude do valor predito, sendo esta proporcional ao coeficiente β_n associado à variável explicativa x_n na determinação do valor final de y.

Em algoritmos baseados em árvores, como *Random Forest* e *XGBoost*, o método *Mean Decrease Impurity* (MDI) estabelece-se como uma importante metodologia na mensuração da importância das variáveis. A fundamentação do MDI baseia-se na análise das divisões sucessivas dos nós em uma árvore de decisão, onde cada nó representa um subconjunto dos dados que será particionado, visando à maximização da homogeneidade dos grupos resultantes.

A equação 2.5 representa a função de ganho utilizada para avaliar a qualidade de uma divisão em árvores de decisão no algoritmo XGBoost. Este cálculo é fundamental para a construção eficiente de modelos de *gradient boosting*.

2.2.5 Formulação Matemática

$$\text{Ganho}(L,R) = \underbrace{\frac{\left(\sum_{i\in L} g_i\right)^2}{\sum_{i\in L} h_i + \lambda}}_{\text{Nó Esquerdo}} + \underbrace{\frac{\left(\sum_{i\in R} g_i\right)^2}{\sum_{i\in R} h_i + \lambda}}_{\text{Nó Direito}} - \underbrace{\frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}}_{\text{Nó Original}} - \gamma \tag{2.5}$$

onde:

- g_i: Gradiente da função de perda (primeira derivada), correspondente ao quadrado dos residuais da predição da iteração atual da árvore
- h_i: Hessiano da função de perda (segunda derivada), correspondente ao número de residuais do nó
- L, R, I: Subconjuntos de dados (Left, Right, e nó Original)
- λ: Termo de regularização L2 (previne sobreajuste)
- γ: Penalidade por complexidade (controla crescimento da árvore)

O ganho é calculado comparando a redução total da perda nos nós filhos (L e R) com a perda no nó pai (I). A divisão só é realizada se:

$$\text{Ganho}(L, R) > 0$$

Valores positivos indicam que a divisão melhora a capacidade preditiva do modelo, enquanto valores negativos sugerem que a divisão não é benéfica.

A importância MDI de uma variável é a soma dos ganhos de todas as divisões onde ela é utilizada, normalizada pelo número de árvores.

Em um problema de regressão de preços imobiliários, por exemplo, uma variável como área construída pode gerar divisões com elevado ganho ao separar imóveis em grupos distintos por dimensão. Tal comportamento resulta da capacidade dessa variável em gerar subgrupos com maior homogeneidade de preços, manifestada pela redução significativa da impureza - neste caso, a variância dos preços - em cada subgrupo resultante.

Exemplo Ilustrativo com Dados Sintéticos

Considere um sinal composto por três variáveis: X_1 (senoidal), X_2 (senoidal com frequência e amplitude diferentes) e X_3 (ruído Gaussiano). A Figura 1 mostra o sinal $Y = X_1 + X_2 + \epsilon$, onde $\epsilon \sim \mathcal{N}(0, 1)$ e ϵ é um ruído gaussiano não correlacionado a X_3.

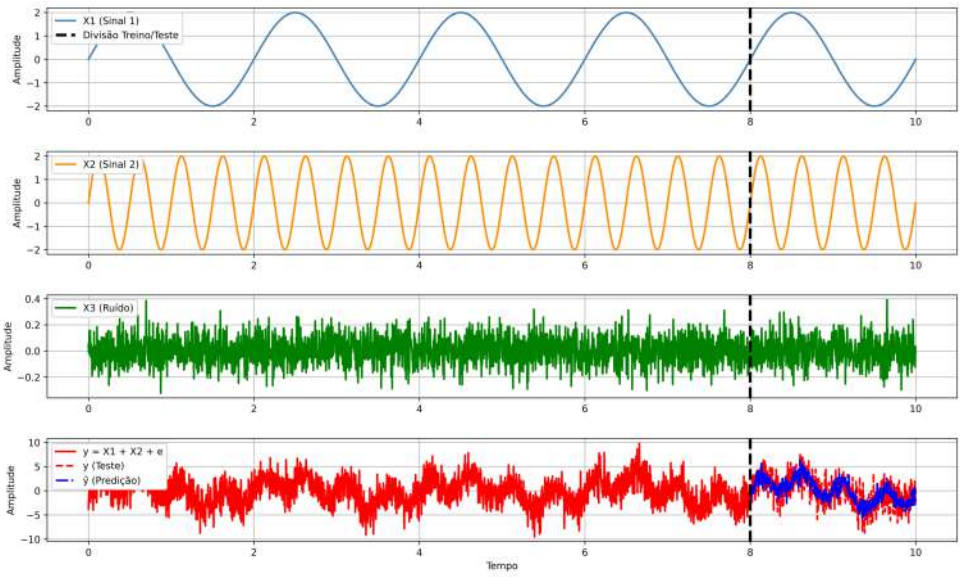

Figura 1 – Variáveis explanatórias e sinal. X_1 e X_2 são relevantes; X_3 é ruído.

A Figura 2 demonstra que o MDI atribui maior importância a X_1 e X_2, identificando corretamente sua relação com Y, enquanto X_3 (ruído) recebe importância próxima de zero.

Embora o *Mean Decrease Impurity* (MDI) seja amplamente utilizado como métrica de importância de variáveis, o método apresenta limitações significativas que devem ser consideradas em sua aplicação. A principal limitação refere-se à instabilidade do MDI na presença de multicolinearidade. Quando duas ou mais variáveis apresentam forte correlação, o algoritmo pode

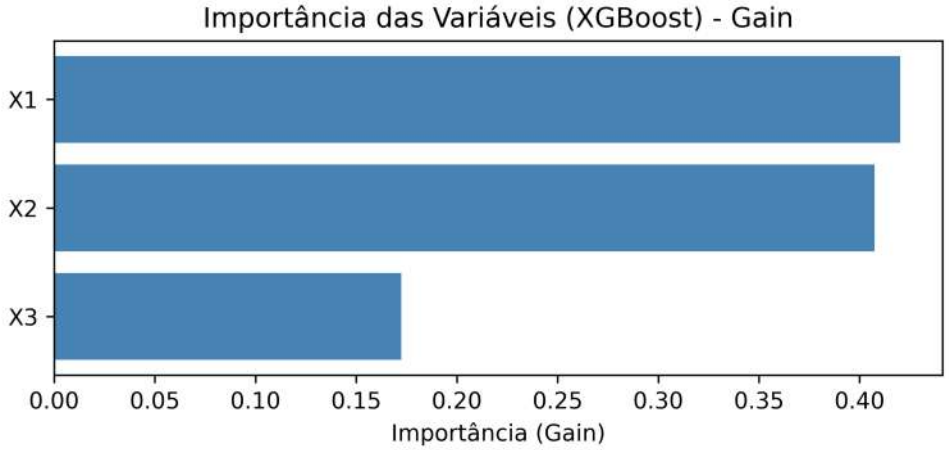

Figura 2 – Importâncias MDI no problema-exemplo. Valores elevados para X_1 e X_2 refletem sua relevância causal.

arbitrariamente atribuir maior importância a uma delas, distribuindo a importância total entre as variáveis correlacionadas de maneira inconsistente entre diferentes execuções do modelo. Este comportamento dificulta a interpretação precisa da relevância individual de cada variável.

Como demonstrado em Scornet (2021), o MDI também apresenta sensibilidade à estrutura das árvores e aos hiperparâmetros do modelo. A profundidade máxima das árvores, o número mínimo de observações por nó e outros parâmetros de regularização influenciam diretamente o cálculo das importâncias. Árvores mais profundas podem resultar em um viés em favor de variáveis que produzem múltiplas divisões sequenciais, mesmo quando divisões posteriores oferecem ganhos marginais reduzidos.

Adicionalmente, o método demonstra limitações na detecção de interações complexas entre variáveis. Como a importância é calculada considerando apenas divisões individuais, o MDI pode subestimar a relevância de variáveis que se tornam importantes apenas em conjunto com outras ou em interações não-lineares específicas.

Estas limitações sugerem a necessidade de complementar a análise do MDI com métodos alternativos de avaliação de importância de variáveis, como permutation importance (MDA) ou valores SHAP, discutidos a seguir.

2.2.6 *Feature Importance* - MDA

O *Mean Decrease Accuracy* (MDA), também conhecido como *permutation importance*, é um método universal para avaliação de importância de variáveis, aplicável a qualquer modelo preditivo. Sua lógica é intuitiva: se uma variável é preditiva, a aleatorização de seus valores deve degradar a performance do modelo. Formalmente, o MDA para a variável X_j é calculado como:

$$\text{MDA}(X_j) = \frac{1}{K} \sum_{k=1}^{K} \left(\mathcal{M}_{\text{original}} - \mathcal{M}^{(k)}_{\text{permutado}} \right), \tag{2.6}$$

onde:

- $\mathcal{M}_{\text{original}}$: Métrica de avaliação (e.g., R^2, RMSE) no conjunto de teste original.
- $\mathcal{M}^{(k)}_{\text{permutado}}$: Métrica após permutar X_j na k-ésima iteração.
- K: Número de permutações (tipicamente $K \geq 30$ para estabilidade).

Vantagens sobre o MDI

Ao contrário do MDI, que é intrínseco a modelos baseados em árvores, o MDA:

- Funciona para qualquer modelo (redes neurais, SVM, etc.).
- Não é influenciado pela escala das variáveis.
- Captura indiretamente interações entre variáveis, pois a permutação destrói relações não lineares.

Exemplo Ilustrativo com Dados Sintéticos

Utilizando o mesmo conjunto de dados da Seção 2.2.4 (X_1, X_2: sinais senoidais; X_3: ruído Gaussiano), a Figura 3 mostra o MDA calculado após 100 permutações. Como esperado, X_1 e X_2 apresentam MDA positivo, enquanto X_3 tem MDA próximo de zero, confirmando sua irrelevância.

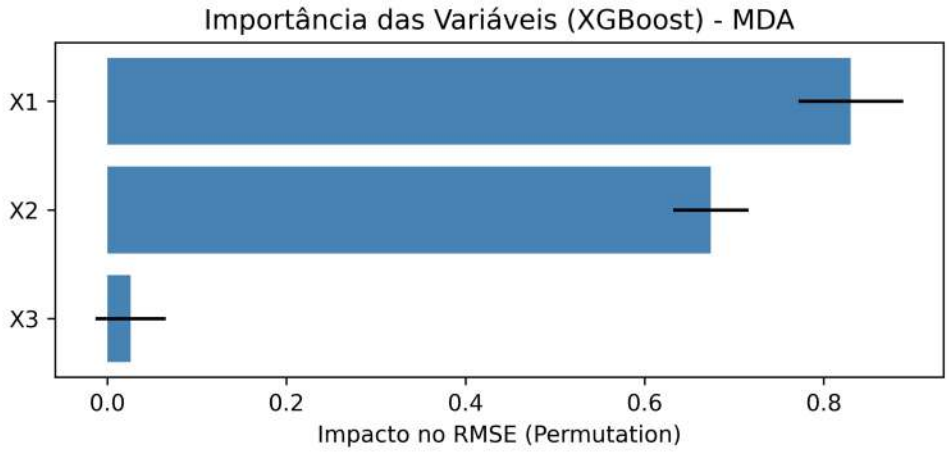

Figura 3 – Valores de MDA para o problema-exemplo. Barras de erro representam o desvio padrão sobre 100 permutações.

Limitações e Mitigações

Apesar de sua versatilidade, o MDA apresenta desafios práticos que exigem atenção. Em primeiro lugar, a multicolinearidade pode levar à subestimação da importância de variáveis correlacionadas, uma vez que a permutação de uma delas não elimina totalmente sua informação preditiva quando outras correlatas permanecem intactas Strobl et al. (2008). Para contornar esse problema, uma abordagem eficaz é permutar grupos de variáveis correlacionadas em conjunto. Por fim, o custo computacional do MDA - que requer K reavaliações do modelo por variável - pode ser mitigado mediante amostragem estratificada das permutações ou paralelização de tarefas. Na Seção 3, integramos o MDA a um protocolo híbrido que combina suas vantagens com o MDI e o SHAP, detalhado a seguir.

2.2.7 *SHAP Values - SHapley Additive exPlanations*

Os *SHAP (SHapley Additive exPlanations) values* são uma abordagem fundamentada na teoria dos jogos cooperativos para quantificar a contribuição de cada variável nas predições de modelos de aprendizado de máquina. Desenvolvido por Lundberg and Lee (2017), este método unifica técnicas de interpretação de modelos sob um framework matematicamente rigoroso, baseado nos valores de Shapley (Shapley (1953)).

Na teoria dos jogos de Shapley, as variáveis são tratadas como "jogadores"que colaboram para gerar a previsão do modelo. O SHAP value ϕ_i de uma variável i representa sua contribuição marginal média para a diferença entre a previsão individual $f(x)$ e o valor esperado do modelo $E[f(X)]$:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_x(S \cup \{i\}) - f_x(S) \right], \tag{2.7}$$

onde:

- F: Conjunto completo de M variáveis ($|F| = M$).
- S: Subconjunto de variáveis excluindo i ($S \subseteq F \setminus \{i\}$).
- $f_x(S)$: Previsão condicional do modelo usando apenas as variáveis em S.

Os SHAP values possuem três axiomas que garantem sua robustez interpretativa:

1. **Eficiência**: $\sum_{i=1}^{M} \phi_i = f(x) - E[f(X)]$, assegurando que a contribuição total das variáveis explica a diferença entre a previsão e a média.
2. **Consistência**: Se o impacto marginal de uma variável aumenta em um novo modelo, seu SHAP value não diminui.

3. **Aditividade**: Para modelos lineares $f(x) = \beta_0 + \sum_{i=1}^{M} \beta_i x_i$, $\phi_i = \beta_i(x_i - E[X_i])$. Sendo assim, para modelos lineares, os SHAP values se assemelham aos coeficientes da regressão linear.

Exemplo Ilustrativo

Assim como nos sub-capítulos anteriores, podemos observar a utilização dos valores SHAP como uma ferramenta de quantificação da importância das variáveis ilustradas figura 1.

Espera-se que os valores SHAP ajudem a distinguir o conjunto de variáveis explanatórias $\{X_1, X_2\}$, como demonstrado na figura 4.

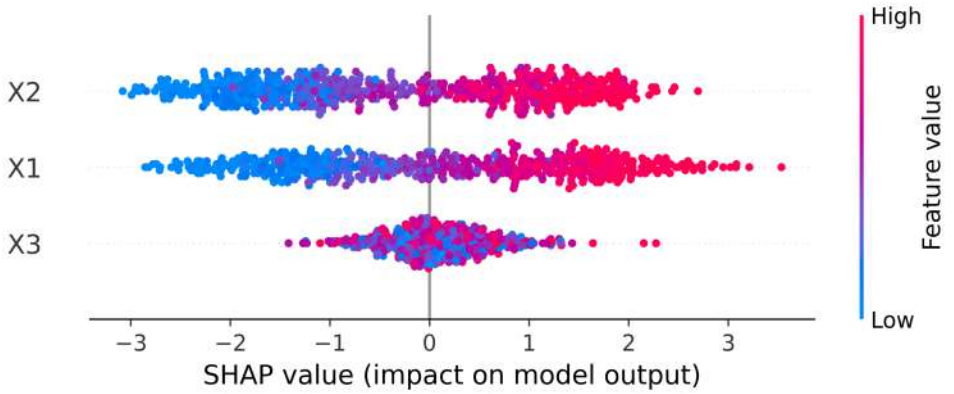

Figura 4 – Valores SHAP para todas as variáveis do problema-exemplo

Neste gráfico, podemos observar a relação entre o valor das variáveis, na escala de azul a vermelho, e o impacto na predição final, quantificado no eixo X. O valor esperado dos SHAP values para uma variável completamente aleatória é 0. Na prática, adotaremos o valor absoluto dos valores SHAP para rankear as variáveis, como explicado no Capítulo 3

Apesar de seu rigor teórico, os SHAP values apresentam desafios práticos:

- **Correlação entre Variáveis**: A premissa de independência entre variáveis na Equação (2.7) pode levar a contribuições distorcidas quando há multicolinearidade.

- **Interpretabilidade em Alta Dimensionalidade**: Para modelos com muitas variáveis, a visualização direta torna-se impraticável. Redução de dimensionalidade (e.g., PCA) ou agrupamento de variáveis correlacionadas são estratégias úteis.

Na Seção 3, aplicamos SHAP values para identificar preditores-chave da variância realizada, complementando as análises de MDI e MDA. Essa abordagem segue recomendações de Molnar (2022) para aumentar a confiabilidade das conclusões.

2.2.8 *Mutual Information*

A *Mutual Information* (MI) é uma medida fundamental da teoria da informação que quantifica a dependência estatística não linear entre variáveis aleatórias. Diferentemente de métricas baseadas em correlação linear, a MI captura relações arbitrárias entre variáveis preditoras e a resposta, sendo particularmente útil para análise exploratória em problemas complexos de previsão.

Fundamentação Matemática

Para variáveis contínuas, a MI entre X e Y é definida como:

$$I(X;Y) = \iint p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) dxdy, \tag{2.8}$$

onde $p(x,y)$ é a densidade conjunta, e $p(x)$, $p(y)$ são as densidades marginais. Na prática, a MI é estimada via métodos não paramétricos como o algoritmo k-vizinhos mais próximos (*k-Nearest Neighbors*) Kraskov et al. (2004), que aproxima as densidades através da geometria dos dados.

Propriedades Chave

- **Não-negatividade**: $I(X;Y) \geq 0$, com igualdade apenas para variáveis independentes.
- **Invariância a Monotonicidades**: Transformações invertíveis (e.g., logaritmo) não alteram $I(X;Y)$.
- **Universalidade**: Detecta qualquer relação funcional mensurável, incluindo não lineares e multimodais.

A MI é particularmente eficaz na fase exploratória de modelagem para identificar preditores não lineares ignorados por métodos lineares e priorizar variáveis para engenharia de *features* interativas.

Utilizando o conjunto de dados da Seção 2.2.4, a Figura 5 mostra a MI estimada entre cada variável explicativa (X_1, X_2, X_3) e o target Y. Como esperado, X_1 e X_2 apresentam MI significativa, enquanto X_3 (ruído) tem MI próxima de zero.

Análise Comparativa de Métricas

A Figura 6 compara o desempenho das quatro métricas discutidas (MDI, MDA, SHAP, MI) no problema-exemplo. Todas identificam corretamente X_1 e X_2 como relevantes, porém com magnitudes relativas levemente distintas:

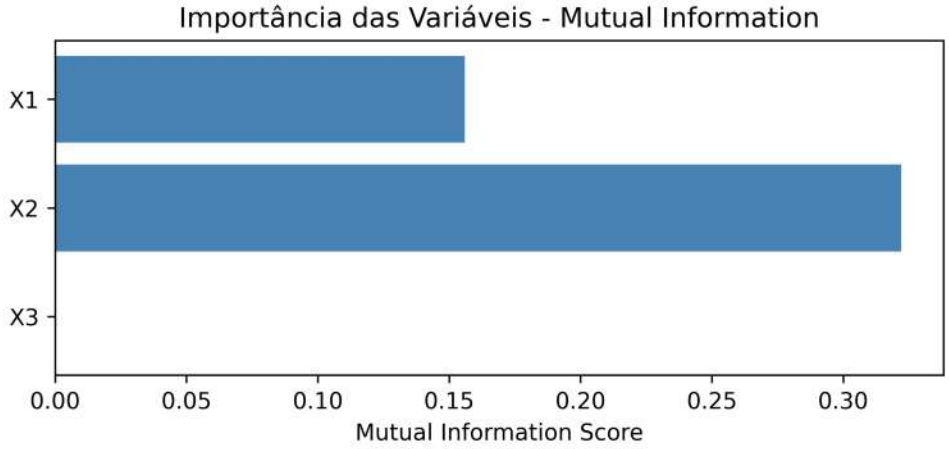

Figura 5 – Mutual Information estimada via k-NN ($k = 5$) para o problema-exemplo.

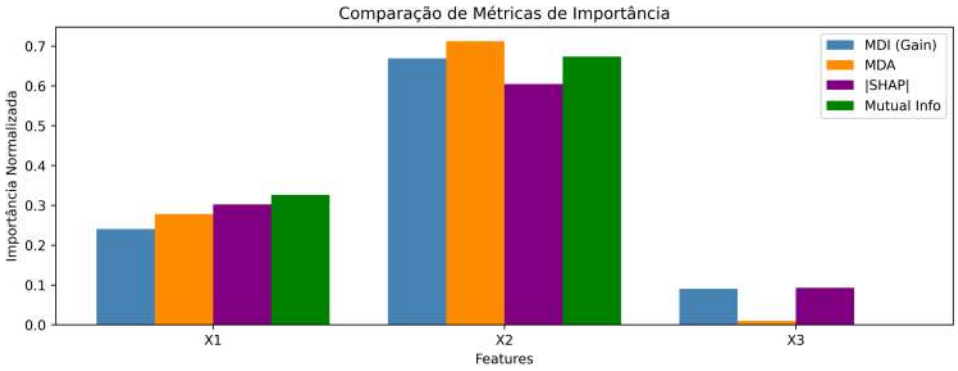

Figura 6 – Comparação padronizada das métricas de importância

A discrepância nas magnitudes reflete diferenças fundamentais:

- **MDI/MDA**: Sensíveis à capacidade preditiva condicional do modelo.
- **SHAP**: Mede impacto causal marginal na saída do modelo.
- **MI**: Quantifica associação estatística bruta, independente do modelo.

3 Metodologia

3.1 Introdução

Mercados de criptomoedas representam um ambiente desafiador para modelagem preditiva: combinam alta volatilidade, dados abundantes (como transações em *blockchain*) e ausência de mecanismos tradicionais de regulação. Essa combinação exige métodos robustos capazes de identificar padrões preditivos robustos em meio a ruído de alta frequência, especialmente em cenários de baixa relação sinal-ruído (*low signal-to-noise ratio*).

A identificação de padrões úteis em dados financeiros requer:

- **Seleção rigorosa**: Descarte de variáveis espúrias através de múltiplas métricas de importância;
- **Validação conservadora**: Testes *out-of-sample* em diferentes regimes de mercado;
- **Controle de robustez**: Garantia de que o desempenho não depende de *overfitting* a ruídos locais.

Este capítulo detalha uma metodologia em três estágios: (1) identificação de variáveis com poder preditivo consistente, (2) validação do desempenho em condições realistas (out-of-sample), e (3) comparação com benchmarks estabelecidos, como o modelo HAR

3.2 Dados e variáveis

Neste estudo, foi construído um conjunto de dados a partir de informações de negociação do par Bitcoin/USDT (btc/usdt, *de facto* a cotação do mercado para o valor do bitcoin), coletadas na corretora Binance, abrangendo o período de 1º de janeiro de 2020 a 31 de março de 2024, totalizando 1552 dias. As variáveis extraídas incluem, sobretudo, dados de operações (trades), volume negociado e ofertas (quotes), sendo inicialmente obtidos em formato bruto. Para o desenvolvimento dos modelos, o conjunto de dados foi re-amostrado para frequência diária para gerar os regressores.

Durante o processo de consolidação das informações, identificou-se um breve apagão de dados correspondente a dois dias (2020-02-09). Pelo fato de essa falha encontrar-se no início do período analisado e abranger apenas um intervalo muito pequeno, optou-se por preencher os valores ausentes com a média dos dias anteriores, minimizando eventuais distorções nos processos de treinamento e avaliação dos modelos subsequentes. Para mensurar a volatilidade, adotou-se a variância realizada diária, calculada como a soma dos retornos logarítmicos quadráticos

observados em janelas de cinco minutos ao longo de cada dia, conforme descrito em na seção 2.2.1.

Além disso, incorporaram-se dados *on-chain* de Bitcoin e USDT a partir da plataforma paga IntoTheBlock (https://www.intotheblock.com/). Esse acréscimo possibilitou enriquecer a base de dados com informações oriundas das tranasções na blockchain, com potencial para aprimorar a robustez das estimativas de volatilidade. No total, a base inicial reuniu 115 variáveis (features) e 1492 observações, após serem excluídos 120 dias destinados à avaliação fora da amostra (*out-of-sample*). Variáveis não estacionárias foram diferenciadas para garantir estacionariedade, condição necessária para modelagem estatística consistente.

Com base nas 115 variáveis iniciais, conduziu-se um processo de engenharia de features com o objetivo de identificar sinais adicionais capazes de aprimorar o poder preditivo em relação à variável de resposta. Esse procedimento envolveu a criação de novos indicadores e a aplicação de diferentes transformações, de modo a captar padrões mais sutis que não estariam explícitos nos dados originais.

Na sequência, o conjunto de dados (*in-sample*) foi segmentado em dois blocos: treino e validação. O bloco de treino serve para exploração dos dados, aprendizado e ajuste dos modelos, enquanto o de validação auxilia tanto na seleção das features quanto na aferição prévia dos resultados, funcionando como um indicativo de potencial *overfitting*. Se o desempenho obtido no conjunto de validação for muito superior ao verificado no teste *out-of-sample*, isso sinaliza que o modelo possivelmente "decorou" o conjunto de treino/validação, comprometendo sua capacidade de generalização ao contexto *out-of-sample*.

Para compor os conjuntos de treino e teste, optou-se pelas janelas fixas definidas na etapa inicial de divisão. Já para o teste em si, adota-se uma estratégia de expansão de janela, de forma que, a cada nova previsão (por exemplo, a cada três dias), os dados recentes são incorporados ao conjunto de treino para atualização do modelo. Essa abordagem garante que, mesmo em um cenário adversarial e em constante evolução, o modelo seja treinado com informações atualizadas, ampliando suas chances de capturar tendências e padrões emergentes.

É de vital importância ressaltar que, ao aferir resultados no conjunto de teste, nenhuma modificação na configuração do modelo deve ser avaliada, seja em adição/remoção de features ou configuração de hiper-parâmetros. Isso comprometeria a lisura da avaliação do conjunto de teste, aumentando em muito o risco de *overfitting* no conjunto de testes, o que levaria o modelo a ter uma performance baixa quando fosse de fato ser utilizado no mundo real, ao fazer previsões em dados não vistos em qualquer um dos sub-conjuntos.

3.2.1 Dados de Mercado

A fim de capturar os principais sinais do mercado a partir dos dados de negociação na corretora Binance, foram extraídas diversas variáveis relacionadas a volume, preço e desequilíbrios

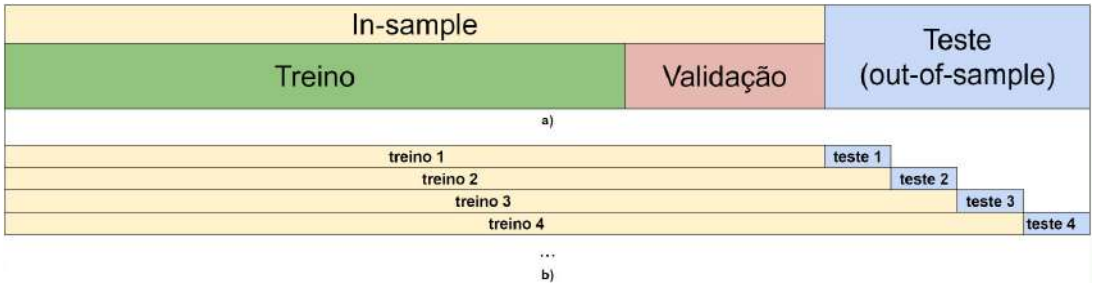

Figura 7 – a) Divisão inicial do conjunto de dados no momento da validação do modelo b) Abordagem de expansão da janela para cada período de testes (n=1)

Tabela 1 – Particionamento Temporal dos Dados

Conjunto	**Início**	**Fim**	**Dias**
Treino	2020-08-29	2023-07-27	1062
Validação	2023-07-28	2023-11-25	120
Teste (OOS)	2023-11-26	2024-03-25	120

na atividade de compra e venda. A seleção dessas variáveis foi guiada pela hipótese de que elas possuem capacidade de refletir dinâmicas de mercado relevantes para a previsão de volatilidade. Por exemplo:

- **Volume Imbalance**: Indica pressão de compra/venda, potencialmente antecipando movimentos de preço;

- **Spread Percentual**: Reflete condições de liquidez, que podem influenciar a volatilidade;

- **Retornos Logarítmicos**: Capturam variações percentuais de preço, essenciais para modelar volatilidade.

A Tabela 2 apresenta uma breve descrição de cada variável:

Tabela 2 – Variáveis de Mercado

Variável	**Descrição**
volume	Volume total negociado no intervalo.
buy_volume	Volume apenas de operações de compra.
sell_volume	Volume apenas de operações de venda.
buy_trades_count	Quantidade de transações de compra.
sell_trades_count	Quantidade de transações de venda.
price_avg	Preço médio no intervalo.
price_max	Maior preço registrado no intervalo.
price_min	Menor preço registrado no intervalo.
amount_avg	Média do tamanho das ordens executadas em moeda base (ex btc).
quote_amount_avg	Média do valor em moeda de cotação (ex.: USDT).
amount_max	Tamanho máximo de ordem executada.
amount_min	Tamanho mínimo de ordem executada.
log_return	Retorno logarítmico entre preços consecutivos.
log_return_pct	Retorno logarítmico expresso em porcentagem.
open	Preço de abertura do período.
close	Preço de fechamento do período.
volume_imbalance	Razão entre volume de compra e venda.
volume_imbalance_pct	Diferença relativa (em %) de volume de compra e venda.
activity_imbalance	Razão entre número de operações de compra e venda.
total_trades	Total de transações (compra + venda).
bid_amount	Soma das ordens de compra (bid) em aberto.
spread_pct	Diferença entre best bid e best ask em %.
amount_imbalance	Desequilíbrio no tamanho das ordens.
spread_pct_max	Maior spread percentual no intervalo.

Para garantir que as variáveis atendam às premissas de modelagem estatística, aplicou-se o teste Augmented Dickey-Fuller (ADF) com nível de significância de 1%. O teste ADF verifica a presença de raiz unitária, onde a hipótese nula H_0 assume que a série é não estacionária. Um p-valor abaixo do nível de significância (1%) rejeita H_0, indicando estacionariedade. O teste identificou que 14 das 24 variáveis necessitavam de diferenciação para atingir estacionariedade, incluindo *volume* e *price_avg*. A primeira diferença ($\Delta x_t = x_t - x_{t-1}$) foi utilizada para essas variáveis. A Figura 8 ilustra os resultados do teste.

Ao contrário de abordagens que removem ou tratam *outliers*, optou-se por mantê-los no conjunto de dados, uma vez que representam eventos reais do mercado (ex.: picos de volume durante notícias relevantes). Essa decisão preserva a integridade da distribuição dos dados e permite que os modelos capturem comportamentos extremos, que são particularmente relevantes em mercados de criptomoedas.

A completude dos dados foi de 99.8%, com valores faltantes preenchidos por interpolação linear. A corretora Binance foi escolhida por ser a líder em volume de negociação do par BTC/USDT.

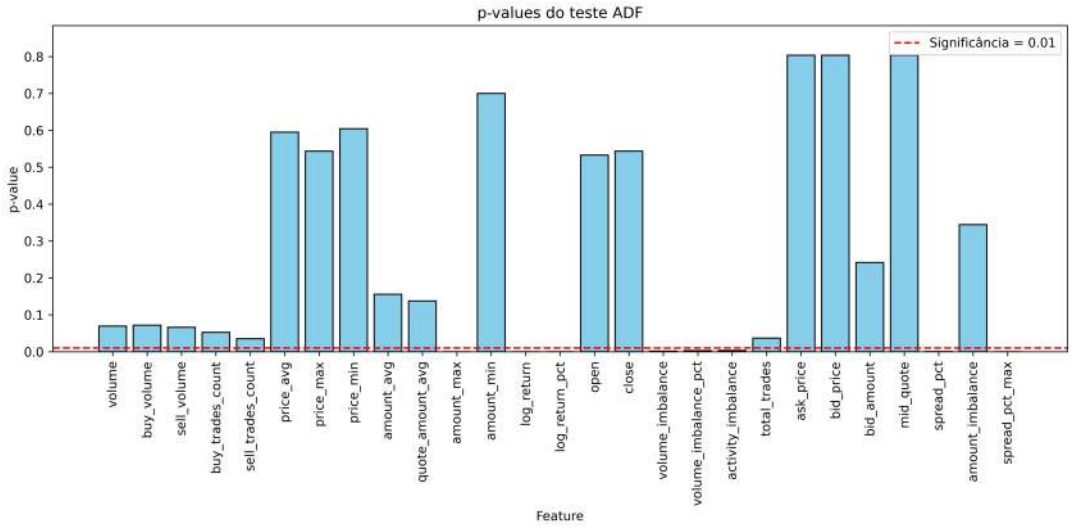

Figura 8 – Teste ADF para as variáveis de dados de negociação

3.2.2 Dados On-Chain Bitcoin

O Bitcoin é uma moeda digital descentralizada que opera em uma *blockchain* – um livro-razão público, imutável e distribuído, que registra todas as transações em blocos sequenciais. Embora a rede seja **pseudônima** (os participantes são identificados por endereços alfanuméricos, não por nomes reais), ela não é totalmente anônima. Cada transação fica permanentemente visível nesse registro aberto, permitindo rastrear padrões como:

- Transferências entre carteiras (ex.: "carteira X enviou 1 BTC para carteira Y"),
- Tempo de retenção de moedas (ex.: "quantos dias uma carteira guardou os BTC antes de movimentá-los"),
- Inatividade de endereços (ex.: "carteira Z não enviava BTC há 3 anos"),
- Fluxos de entrada/saída de *exchanges*.

A partir desses dados brutos, derivam-se as **métricas on-chain**: indicadores quantitativos que analisam o comportamento da rede e seus usuários. Essas métricas incluem, por exemplo:

1. **Tamanho médio das transações** (para detectar movimentos incomuns de grandes valores),
2. **Atividade de carteiras dormentes** (indicando possíveis vendas de "holders" antigos),
3. **Taxa de hash da rede** (medindo o poder computacional dedicado à segurança do sistema),
4. **Saldos acumulados em exchanges** (sinalizando tendências de compra/venda).

Essa análise – chamada *on-chain analytics* – transforma dados técnicos da *blockchain* em insights estratégicos, servindo como termômetro para:

- **Transparência**: Auditoria pública do suprimento de *BTC* em circulação,
- **Segurança**: Monitoramento de ataques ou concentração de poder na rede,
- **Comportamento de mercado**: Identificação de padrões de acumulação ou distribuição por grandes investidores.

As variáveis *on-chain*, extraídas diretamente da análise de movimentação de transações gravadas na rede Bitcoin, oferecem informações complementares ao tradicional conjunto de dados de mercado e têm o potencial de fornecer indicativos sobre a dinâmica de preços e a volatilidade do ativo. Em particular, a análise de fluxos de criptomoedas para corretoras (e vice-versa), bem como o monitoramento do comportamento de detentores de longo e curto prazo, permite identificar potenciais pontos de inflexão na oferta e na demanda. A Tabela 3 lista os principais indicadores *on-chain* de Bitcoin utilizados neste estudo, evidenciando suas características e a forma como cada um pode contribuir para a compreensão dos movimentos e da volatilidade do mercado.

Para garantir que as variáveis atendam às premissas de modelagem estatística, aplicou-se o teste Augmented Dickey-Fuller (ADF) com nível de significância de 1%. O teste identificou que 30 das 41 variáveis necessitaram de diferenciação para atingir estacionariedade, incluindo *btc_onchain_netflow_exchange_binance* e *btc_onchain_miners_volume_share*. A primeira diferença ($\Delta x_t = x_t - x_{t-1}$) foi utilizada para essas variáveis. A Figura 9 ilustra os resultados do teste.

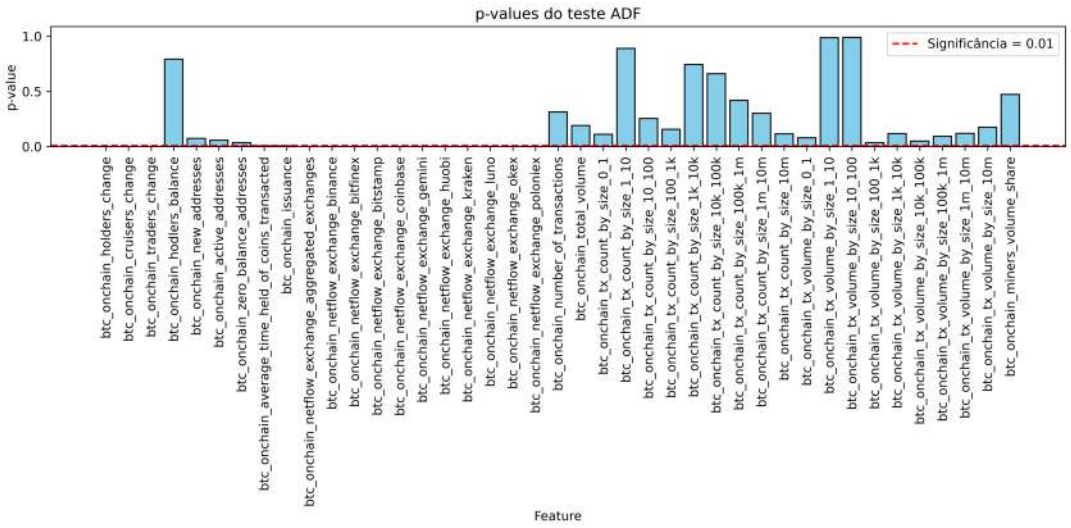

Figura 9 – Teste ADF para as variáveis on-chain de Bitcoin

Assim como nos dados de mercado, optou-se por manter outliers nos dados on-chain, uma vez que representam eventos reais da rede Bitcoin (ex.: grandes transferências entre carteiras).

Tabela 3 – Variáveis *on-chain* de Bitcoin

Variável	**Descrição**
btc_onchain_holders_change	Variação no número de holders de BTC.
btc_onchain_cruisers_change	Variação no número de usuários de médio prazo (cruisers).
btc_onchain_traders_change	Variação no número de traders (curto prazo).
btc_onchain_hodlers_balance	Saldo total mantido pelos hodlers (longo prazo).
btc_onchain_new_addresses	Quantidade de novos endereços criados.
btc_onchain_active_addresses	Número de endereços ativos na rede.
btc_onchain_zero_balance_addresses	Número de endereços com saldo zero.
btc_onchain_average_time_held_of_coins_transacted	Tempo médio de retenção das moedas transacionadas.
btc_onchain_issuance	Emissão total de BTC (novas moedas).
btc_onchain_netflow_exchange_aggregated_exchanges	Fluxo líquido agregado de/para corretoras.
btc_onchain_netflow_exchange_binance	Fluxo líquido de/para a Binance.
btc_onchain_netflow_exchange_bitfinex	Fluxo líquido de/para a Bitfinex.
btc_onchain_netflow_exchange_bitstamp	Fluxo líquido de/para a Bitstamp.
btc_onchain_netflow_exchange_coinbase	Fluxo líquido de/para a Coinbase.
btc_onchain_netflow_exchange_gemini	Fluxo líquido de/para a Gemini.
btc_onchain_netflow_exchange_huobi	Fluxo líquido de/para a Huobi.
btc_onchain_netflow_exchange_kraken	Fluxo líquido de/para a Kraken.
btc_onchain_netflow_exchange_luno	Fluxo líquido de/para a Luno.
btc_onchain_netflow_exchange_okex	Fluxo líquido de/para a OKEx.
btc_onchain_netflow_exchange_poloniex	Fluxo líquido de/para a Poloniex.
btc_onchain_number_of_transactions	Número total de transações na rede.
btc_onchain_total_volume	Volume total transacionado em BTC.
btc_onchain_tx_count_by_size_0_1	Contagem de transações entre 0 e 1 BTC.
btc_onchain_tx_count_by_size_1_10	Contagem de transações entre 1 e 10 BTC.
btc_onchain_tx_count_by_size_10_100	Contagem de transações entre 10 e 100 BTC.
btc_onchain_tx_count_by_size_100_1k	Contagem de transações entre 100 e 1.000 BTC.
btc_onchain_tx_count_by_size_1k_10k	Contagem de transações entre 1.000 e 10.000 BTC.
btc_onchain_tx_count_by_size_10k_100k	Contagem de transações entre 10.000 e 100.000 BTC.
btc_onchain_tx_count_by_size_100k_1m	Contagem de transações entre 100.000 e 1 milhão de BTC.
btc_onchain_tx_count_by_size_1m_10m	Contagem de transações entre 1 milhão e 10 milhões de BTC.
btc_onchain_tx_count_by_size_10m	Contagem de transações acima de 10 milhões de BTC.
btc_onchain_tx_volume_by_size_0_1	Volume total em transações de 0 a 1 BTC.
btc_onchain_tx_volume_by_size_1_10	Volume total em transações de 1 a 10 BTC.
btc_onchain_tx_volume_by_size_10_100	Volume total em transações de 10 a 100 BTC.
btc_onchain_tx_volume_by_size_100_1m	Volume total em transações de 100 a 1.000 BTC.
btc_onchain_tx_volume_by_size_1k_10k	Volume total em transações de 1.000 a 10.000 BTC.
btc_onchain_tx_volume_by_size_10k_100k	Volume total em transações de 10.000 a 100.000 BTC.
btc_onchain_tx_volume_by_size_100k_1m	Volume total em transações de 100.000 a 1 milhão de BTC.
btc_onchain_tx_volume_by_size_1m_10m	Volume total em transações de 1 milhão a 10 milhões de BTC.
btc_onchain_tx_volume_by_size_10m	Volume total em transações acima de 10 milhões de BTC.
btc_onchain_miners_volume_share	Fração do volume total em custódia de mineradores.

Essa decisão preserva a integridade da distribuição dos dados e permite que os modelos capturem comportamentos extremos, que são particularmente relevantes em mercados de criptomoedas.

A plataforma *IntoTheBlock* foi utilizada por sua abrangência e confiabilidade na coleta de dados on-chain. No entanto, é importante ressaltar que dados on-chain podem não capturar completamente a dinâmica do mercado, especialmente em períodos de alta volatilidade ou fragmentação de liquidez.

3.2.3 Dados On-Chain USDT

A USDT (Tether) é uma *stablecoin* centralizada, lastreada em reservas fiduciárias (como o dólar americano) e emitida pela empresa Tether. Suas transações são registradas em *blockchains* públicas (como Ethereum, Tron e Solana), funcionando como um "dólar digital"para facilitar negociações no ecossistema cripto. Apesar de sua natureza centralizada, todas as movimentações de USDT são públicas e auditáveis nas redes onde operam, gerando dados críticos para análise.

Assim como o Bitcoin, a USDT é **pseudônima**: endereços de carteira são identificados por códigos alfanuméricos, mas a Tether reserva-se o direito de congelar saldos ou bloquear endereços. Exemplos de padrões rastreáveis incluem:

- Grandes transferências para *exchanges* (ex.: "100 milhões de USDT enviados à Binance"),
- Emissões (*minting*) ou destruições (*burning*) de tokens por contratos controlados pela Tether,
- Atividade de "whales" (grandes carteiras institucionais).

As métricas on-chain da USDT focam em estabilidade, liquidez e confiança no lastro:

1. **Suprimento Circulante**: Total de USDT em circulação (emitidos não queimados),
2. **Reservas Auditáveis**: Comparação entre o saldo emitido e as reservas declaradas pela Tether,
3. **Fluxo entre Cadeias**: Migração de USDT entre blockchains (ex.: Ethereum → Tron),
4. **Blacklist de Endereços**: Número de carteiras bloqueadas por suspeita de fraude ou cumprimento regulatório.

As variáveis *on-chain* do USDT fornecem um panorama detalhado do comportamento de oferta, demanda e fluxos na rede Dólar Tether, permitindo identificar possíveis desequilíbrios ou mudanças na dinâmica do stablecoin mais utilizado em operações de arbitragem, transferência e proteção de valor dentro do ecossistema de criptomoedas. Além disso, os dados relacionados a grandes transações e distribuição de holdings podem revelar movimentos estratégicos de investidores de diferentes portes, enquanto o monitoramento de endereços e fluxo para corretoras potencialmente antecipa pressões de compra ou venda de outros criptoativos, em especial o Bitcoin. A Tabela 4 lista os principais indicadores *on-chain* de USDT utilizados neste estudo, evidenciando suas características e a forma como cada um pode contribuir para a compreensão dos movimentos e da volatilidade do mercado.

Assim como nas variáveis anteriores, para garantir que as variáveis atendam às premissas de modelagem estatística, aplicou-se o teste Augmented Dickey-Fuller (ADF) com nível de significância de 1%. O teste identificou que 31 das 49 variáveis necessitaram de diferenciação para atingir estacionariedade, incluindo por exemplo *usdt_onchain_number_of_large_transactions* e *usdt_onchain_whales*. A primeira diferença foi utilizada para essas variáveis. A Figura 10 ilustra os resultados do teste.

Assim como nos dados de mercado e de Bitcoin, optou-se por manter outliers nos dados on-chain de USDT, uma vez que representam eventos reais da rede (ex.: grandes transferências

Tabela 4 – Variáveis *on-chain* de USDT

Variável	Descrição
usdt_onchain_number_of_large_transactions	Contagem de transações de valor considerável em USDT.
usdt_onchain_circulating_supply	Quantidade total de USDT em circulação.
usdt_onchain_usdt_balance_by_holdings_0_1	Saldo de USDT em carteiras com holdings entre 0 e 1 USDT.
usdt_onchain_usdt_balance_by_holdings_1_10	Saldo de USDT em carteiras com holdings de 1 a 10 USDT.
usdt_onchain_usdt_balance_by_holdings_10_100	Saldo de USDT em carteiras com holdings de 10 a 100 USDT.
usdt_onchain_usdt_balance_by_holdings_100_1k	Saldo de USDT em carteiras com holdings de 100 a 1.000 USDT.
usdt_onchain_usdt_balance_by_holdings_1k_10k	Saldo de USDT em carteiras com holdings de 1.000 a 10.000 USDT.
usdt_onchain_usdt_balance_by_holdings_10k_100k	Saldo de USDT em carteiras com holdings de 10.000 a 100.000 USDT.
usdt_onchain_usdt_balance_by_holdings_100k_1m	Saldo de USDT em carteiras de 100.000 a 1 milhão de USDT.
usdt_onchain_usdt_balance_by_holdings_1m_10m	Saldo de USDT em carteiras de 1 milhão a 10 milhões de USDT.
usdt_onchain_usdt_balance_by_holdings_10m	Saldo de USDT em carteiras com mais de 10 milhões de USDT.
usdt_onchain_usdt_txs_volume_total_volume	Volume total em transações de USDT.
usdt_onchain_netflow_ratio	Relação entre fluxo de entrada e saída de USDT.
usdt_onchain_usdt_large_txs_volume_total_volume	Volume total de grandes transações de USDT.
usdt_onchain_beta_coefficient	Beta do USDT em relação a referência de mercado ou ativo específico.
usdt_onchain_inflow	Fluxo total de entrada de USDT em carteiras monitoradas.
usdt_onchain_market_cap	Capitalização de mercado do USDT.
usdt_onchain_usdt_txs_count_0_1	Contagem de transações de 0 a 1 USDT.
usdt_onchain_usdt_txs_count_1_10	Contagem de transações de 1 a 10 USDT.
usdt_onchain_usdt_txs_count_10_100	Contagem de transações de 10 a 100 USDT.
usdt_onchain_usdt_txs_count_100_1k	Contagem de transações de 100 a 1.000 USDT.
usdt_onchain_usdt_txs_count_1k_10k	Contagem de transações de 1.000 a 10.000 USDT.
usdt_onchain_usdt_txs_count_10k_100k	Contagem de transações de 10.000 a 100.000 USDT.
usdt_onchain_usdt_txs_count_100k_1m	Contagem de transações de 100.000 a 1 milhão de USDT.
usdt_onchain_usdt_txs_count_1m_10m	Contagem de transações de 1 milhão a 10 milhões de USDT.
usdt_onchain_usdt_txs_count_10m	Contagem de transações com mais de 10 milhões de USDT.
usdt_onchain_outflow	Fluxo total de saída de USDT de carteiras monitoradas.
usdt_onchain_total	Número total de carteiras (endereços) de USDT.
usdt_onchain_total_with_balance	Número de carteiras com saldo positivo em USDT.
usdt_onchain_total_zero_balance	Número de carteiras com saldo zero em USDT.
usdt_onchain_whales	Contagem de grandes detentores (whales) de USDT.
usdt_onchain_investors	Contagem de investidores médios de USDT.
usdt_onchain_retail	Contagem de detentores de pequeno porte (retail).
usdt_onchain_average_time_held_of_coins_transacted	Tempo médio de retenção das moedas transacionadas em USDT.
usdt_onchain_new_addresses	Número de novos endereços de USDT criados.
usdt_onchain_active_addresses	Número de endereços ativos de USDT.
usdt_onchain_zero_balance_addresses	Número de endereços de USDT com saldo zero.
usdt_onchain_number_of_transactions	Número total de transações com USDT.
usdt_onchain_usdt_inflow_volume_aggregated_exchanges	Volume de entrada total de USDT em corretoras agregadas.
usdt_onchain_usdt_inflow_volume_binance	Volume de entrada de USDT na Binance.
usdt_onchain_usdt_inflow_volume_crypto_com	Volume de entrada de USDT na Crypto.com.
usdt_onchain_usdt_outflow_volume_aggregated_exchanges	Volume de saída total de USDT em corretoras agregadas.
usdt_onchain_usdt_outflow_volume_binance	Volume de saída de USDT na Binance.
usdt_onchain_usdt_outflow_volume_crypto_com	Volume de saída de USDT na Crypto.com.
usdt_onchain_usdt_btc	Relação USDT-BTC na rede (por exemplo, pares de endereços).
usdt_onchain_usdt_eth	Relação USDT-ETH na rede (por exemplo, contratos ERC-20).
usdt_onchain_hodlers_1y	Número de endereços com USDT retido por mais de 1 ano.
usdt_onchain_cruisers_1_12m	Número de endereços com USDT retido de 1 a 12 meses.
usdt_onchain_traders_1m	Número de endereços com USDT retido por até 1 mês.

entre carteiras). Essa decisão preserva a integridade da distribuição dos dados e permite que os modelos capturem comportamentos extremos, que são particularmente relevantes em mercados de criptomoedas. A plataforma *IntoTheBlock* também foi utilizada para a coleta dos dados onchain de USDT.

3.2.4 Engenharia de Variáveis

Nesta seção, definimos formalmente as transformações aplicadas às variáveis originais, com o intuito de gerar novas *features* capazes de capturar padrões de curto e longo prazo na série temporal. As Equações 3.1 a 3.4 ilustram o cálculo das principais métricas empregadas: médias móveis, desvios-padrão móveis, diferenças e defasagens (*lags*).

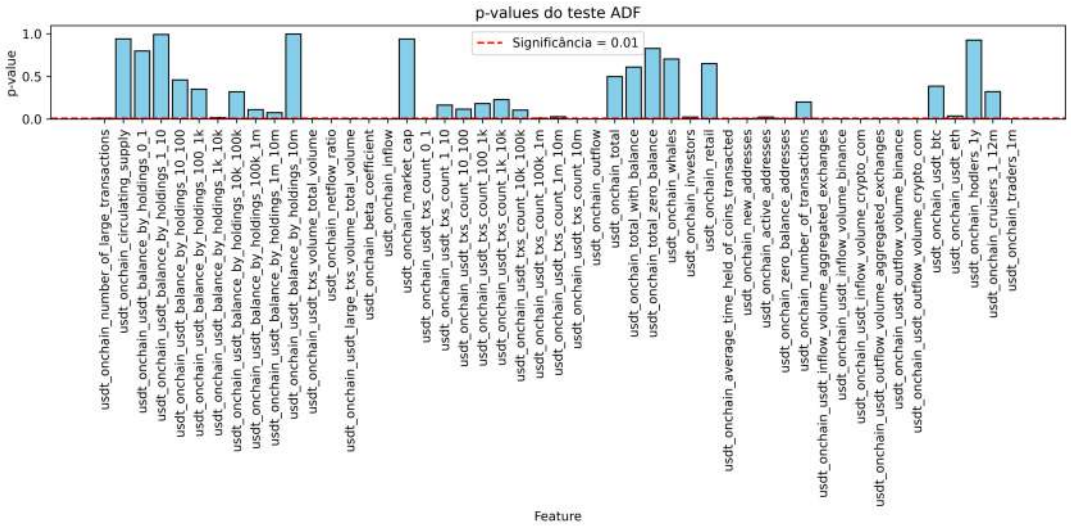

Figura 10 – Teste ADF para as variáveis on-chain de USDT

Rolling Average (Média Móvel)

$$\text{roll\_avg}_t^{(h)} = \frac{1}{h}\sum_{k=0}^{h-1} x_{t-k}, \tag{3.1}$$

onde $\text{roll\_avg}_t^{(h)}$ representa a média móvel de ordem h no instante t, e x_t é o valor da variável original no tempo t.

Rolling Standard Deviation (Desvio-Padrão Móvel)

$$\text{roll\_std}_t^{(h)} = \sqrt{\frac{1}{h}\sum_{k=0}^{h-1}\left(x_{t-k} - \text{roll\_avg}_t^{(h)}\right)^2}, \tag{3.2}$$

em que $\text{roll\_std}_t^{(h)}$ é o desvio-padrão móvel de ordem h no instante t, calculado em torno da média móvel $\text{roll\_avg}_t^{(h)}$.

Rolling Difference (Diferença em Janela)

$$\text{roll\_diff}_t^{(h)} = x_t - x_{t-h}, \tag{3.3}$$

onde $\text{roll\_diff}_t^{(h)}$ denota a diferença entre o valor atual da variável e o valor de h períodos atrás.

Lag (Defasagem)

$$\text{L}_t^{(l)} = x_{t-l}, \tag{3.4}$$

em que $\text{L}_t^{(l)}$ corresponde ao valor de x_t deslocado em l períodos no passado. Neste trabalho, empregamos defasagens de 1, 3, 5, 7, 9, 15, 30, 60, 120 e 240 dias.

As janelas foram escolhidas para capturar padrões em diferentes horizontes temporais:

- **Curto Prazo (1-9 dias)**: Variações diárias e semanais;
- **Médio Prazo (15-60 dias)**: Tendências mensais;

- **Longo Prazo (120-240 dias)**: Comportamentos sazonais e de longo prazo.

Como ilustram as Equações 3.1 a 3.4, a aplicação das médias móveis e dos desvios-padrão móveis visa captar tendências e variações em diferentes horizontes (curto, médio e longo prazo), enquanto as diferenças em janela permitem identificar variações pontuais de maior magnitude. Já as defasagens (*lags*) visam capturar efeitos de causa e consequência entre as variáveis e o objetivo.

3.3 Métricas de Avaliação

A avaliação quantitativa do desempenho de modelos preditivos requer métricas objetivas que mensuram diferentes aspectos da qualidade das predições. Este trabalho utiliza cinco métricas complementares: Skill Score, Erro Médio Absoluto (MAE), Raiz do Erro Quadrático Médio (RMSE), Coeficiente de Determinação (R^2) e Erro Percentual Absoluto Médio (MAPE).

O Skill Score constitui uma métrica que avalia a melhoria relativa do modelo em comparação com um modelo de referência. Neste trabalho, o modelo de referência é o modelo HAR (Heterogeneous Autoregressive), conforme detalhado na Seção 2.2.3. Sua formulação matemática é expressa por:

$$\text{Skill} = 1 - \frac{\text{MSE}_{\text{modelo}}}{\text{MSE}_{\text{referência}}} \tag{3.5}$$

onde MSE representa o erro quadrático médio. O Skill Score varia no intervalo $(-\infty, 1]$, onde valores positivos indicam que o modelo supera o desempenho da referência, o valor 1 representa um modelo perfeito, e valores negativos indicam desempenho inferior ao modelo de referência.

O Erro Médio Absoluto (MAE - Mean Absolute Error) quantifica a magnitude média dos erros de previsão em termos absolutos, preservando a unidade original da variável predita:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.6}$$

onde y_i representa o valor observado, $\hat{y}_i$ o valor predito, e n o número total de observações. O MAE apresenta a vantagem de ser diretamente interpretável na escala da variável de interesse.

A Raiz do Erro Quadrático Médio (RMSE - Root Mean Square Error) penaliza erros maiores mais severamente devido à sua natureza quadrática:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.7}$$

O RMSE também mantém a unidade original da variável predita, porém atribui peso maior a desvios extremos devido ao termo quadrático. A comparação entre MAE e RMSE fornece insights sobre a distribuição dos erros: quando RMSE é substancialmente maior que MAE, indica a presença de erros de grande magnitude.

O Coeficiente de Determinação (R^2) mensura a proporção da variância na variável dependente que é explicada pelo modelo:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.8}$$

onde $\bar{y}$ representa a média dos valores observados. O R^2 varia no intervalo $(-\infty, 1]$, onde 1 indica um ajuste perfeito, 0 indica que o modelo não apresenta poder preditivo superior à média simples, e valores negativos indicam desempenho inferior à média.

O Erro Percentual Absoluto Médio (MAPE - Mean Absolute Percentage Error) é uma métrica auxiliar que expressa o erro em termos percentuais, facilitando a interpretação em contextos onde a escala da variável predita varia significativamente:

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3.9}$$

No entanto, o MAPE pode ser enganoso quando os valores observados (y_i) estão próximos de zero, pois divisões por valores muito pequenos inflam o erro percentual. Por essa razão, o MAPE é utilizado apenas como uma métrica auxiliar para facilitar o entendimento, mas nunca como referência para melhorias nos modelos.

Limitações das Métricas

É importante ressaltar que cada métrica possui limitações:

- **Skill Score**: Depende da escolha do modelo de referência, que pode não ser ótimo em todos os cenários;
- **R^2**: Pode ser inflado em séries não estacionárias ou com tendências fortes;
- **MAE e RMSE**: Não são invariantes à escala, dificultando comparações entre diferentes variáveis;

- **MAPE**: Pode ser enganoso para valores observados próximos de zero e não deve ser usado como métrica principal para otimização de modelos.

A utilização conjunta destas métricas proporciona uma avaliação abrangente do desempenho do modelo. O Skill Score quantifica a melhoria relativa sobre o modelo HAR, MAE e RMSE fornecem medidas absolutas de erro em escalas complementares; o R^2 avalia a capacidade explicativa do modelo em termos da variância dos dados, e o MAPE oferece uma interpretação percentual do erro. Esta abordagem multi-métrica permite identificar diferentes aspectos da qualidade das predições e fundamentar comparações objetivas entre diferentes modelos.

3.4 Seleção de variáveis

O conjunto de dados resultante da aplicação da engenharia de features ao conjunto de dados inicial faz com que o total de variáveis a serem testadas passe de 1000, atingindo quase o mesmo número de observações disponíveis, causando um efeito chamado "maldição da dimensionalidade". É improvável que cada uma dessas variáveis apresente poder preditivo em relação à volatilidade; na realidade, muitas exibem comportamento essencialmente aleatório e desassociado da variável de resposta.

Além disso, é possível que a variável de resposta seja parcialmente explicada por diferentes componentes, provenientes tanto das variáveis explicativas quanto de sua interação. Nesse sentido, métodos lineares que tentam captar apenas relações diretas (como o *F-test* ou a correlação de Pearson) tendem a ser insuficientes. Isso ocorre porque tais técnicas podem descartar, prematuramente, atributos que revelariam poder preditivo ao serem considerados em conjunto com outros. Por esse motivo, recorremos a ferramentas mais avançadas, já apresentadas no Capítulo 2, capazes de avaliar a importância das *features* de forma mais robusta.

Conforme afirma de Prado (2018) em seu Capítulo 8, uma das principais lições da aplicação de Aprendizado de Máquina a dados financeiros é que:

> "O *backtest* não é uma ferramenta de pesquisa. A *feature importance* sim."

Partindo dessa premissa, torna-se imprescindível estabelecer um processo consistente de seleção das variáveis, baseado nas suas importâncias, a serem utilizadas no modelo.

Assim, para um conjunto de dados $\{X, y\}$, adiciona-se propositalmente um número reduzido de variáveis totalmente aleatórias, a fim de servir como referência base para as demais. A escolha de um número de até 10% do número total de variáveis, limitado a 20, busca equilibrar rigor estatístico e eficiência computacional. Um número excessivo de variáveis aleatórias eleva o custo computacional (especialmente para MDA, que requer o treinamento de um modelo para cada feature testada) e aumenta a probabilidade de falsos positivos (que eliminariam possíveis

boas variáveis apenas por chance). Limitar $n_{\text{rand}} \leq 20$ mitiga esses riscos sem comprometer a sensibilidade do método.

Em seguida, treina-se um modelo baseado em árvores de decisão, como *Random Forest* ou *XGBoost*, para se obter a *Mean Decrease Impurity* (MDI) e a *Mean Decrease Accuracy* (MDA). Paralelamente, realiza-se a predição em um conjunto de validação, obtendo-se os valores de SHAP, que mensuram a contribuição isolada de cada atributo para a predição final, atribuindo maior relevância às variáveis decisivas no resultado. Por fim, calcula-se a *Mutual Information* (MI) entre cada atributo e a variável-alvo.

Espera-se que MDI, MDA e SHAP exibam alta concordância no ranqueamento de variáveis importantes, uma vez que todas dependem do modelo preditivo adotado e capturam interações não lineares entre *features*. Em contraste, a Mutual Information (MI) – que mede dependência estatística independente do modelo – pode divergir significativamente, pois não considera sinergias entre variáveis explicativas.

A concordância entre os rankings é quantificada pela correlação de Kendall's τ ponderada (*weighted*), que atribui maior importância à concordância nas primeiras posições dos rankings, onde discordâncias têm maior impacto na performance preditiva.

Idealmente, as variáveis puramente aleatórias devem exibir valores nulos ou próximos de zero nas métricas empregadas. Desse modo, para MDI, MDA e SHAP, seleciona-se o conjunto de variáveis cujo valor supere o maior valor observado entre as variáveis aleatórias, retendo, para cada uma delas, o subconjunto de variáveis que compõe pelo menos 50% do valor total do somatório das importâncias. Em relação à Mutual Information (MI), por não depender de um modelo preditivo subjacente, a estratégia adotada é escolher um número limitado de variáveis (*k best*), escolhendo-se $k = \sqrt{N}$, onde N é o número total de variáveis do dataset a ser filtrado (convém notar que esta é uma regra empírica para a seleção). Em seguida, obtém-se a união entre todos os subconjuntos, resultando em um grupo de atributos que consistentemente demonstram maior relevância preditiva do que o ruído aleatório, com uma margem de segurança. Esse procedimento contribui para mitigar o risco de *overfitting* e reforçar a robustez do modelo frente às complexidades do mercado financeiro.

O seguinte pseudo-algoritmo descreve a metodologia para seleção de variáveis baseada em importância, utilizando MDI, MDA, SHAP e Mutual Information (MI):

Algorithm 1 Seleção de Variáveis Baseada em Importância

Passo 1: Adicionar Variáveis Aleatórias
$N \leftarrow$ número total de variáveis originais
$n_{\text{rand}} = \min\left(20, 1 + \left\lceil \frac{N}{50} \right\rceil\right)$ ▷ Aprox. 2% de N, máximo 20
$X_{\text{rand}} = \{r_1, r_2, \ldots, r_{n_{\text{rand}}}\}$, onde $r_i \sim \mathcal{U}(0,1)$
$X_{\text{aug}} = X \cup X_{\text{rand}}$ ▷ Conjunto de dados aumentado
Passo 2: Treinar Modelo e Calcular Importâncias
model $\leftarrow$ RandomForest/XGBoost(X_{aug}, y) ▷ Treinar modelo
$\text{MDI} = \{\text{MDI}_1, \ldots, \text{MDI}_{N+n_{\text{rand}}}\}$ ▷ Importância por impureza
$\text{MDA} = \{\text{MDA}_1, \ldots, \text{MDA}_{N+n_{\text{rand}}}\}$ ▷ Impacto na acurácia
$\text{SHAP} = \{\text{SHAP}_1, \ldots, \text{SHAP}_{N+n_{\text{rand}}}\}$ ▷ Contribuição marginal
Passo 3: Calcular Mutual Information (MI)
for cada $x_i \in X_{\text{aug}}$ **do**
    $\text{MI}_i = I(x_i; y)$ ▷ I: informação mútua
end for
Passo 4: Definir Limiares de Significância
$\tau_{\text{MDI}} = \max(\{\text{MDI}_i \mid x_i \in X_{\text{rand}}\})$
$\tau_{\text{MDA}} = \max(\{\text{MDA}_i \mid x_i \in X_{\text{rand}}\})$
$\tau_{\text{SHAP}} = \max(\{\text{SHAP}_i \mid x_i \in X_{\text{rand}}\})$
Passo 5: Selecionar Variáveis por Método
$S_{\text{MDI}} = \{x_i \in X \mid \text{MDI}_i > \tau_{\text{MDI}}\}$
$S_{\text{MDA}} = \{x_i \in X \mid \text{MDA}_i > \tau_{\text{MDA}}\}$
$S_{\text{SHAP}} = \{x_i \in X \mid \text{SHAP}_i > \tau_{\text{SHAP}}\}$
$k = \left\lfloor \sqrt{N} \right\rfloor$
$S_{\text{MI}} = \{x_i \in X \mid \text{MI}_i \in \text{top-}k(\text{MI})\}$
Passo 6: Unir Subconjuntos e Retornar
$S = S_{\text{MDI}} \cup S_{\text{MDA}} \cup S_{\text{SHAP}} \cup S_{\text{MI}}$
return S

3.5 Avaliação Out-of-Sample e Comparação com o Modelo HAR

A avaliação do desempenho do modelo proposto em dados não vistos (*out-of-sample*, OOS) é uma etapa crucial para garantir a generalização e a robustez da metodologia. Além disso, a comparação com um modelo baseline, como o HAR (*Heterogeneous Autoregressive Model*), fornece uma referência clara para avaliar se a complexidade adicional do modelo proposto é justificada.

A avaliação Out-of-sample é essencial para o desenvolvimento de um modelo de *machine learning*:

- **Evitar Overfitting**: Testar o modelo em dados não utilizados durante o treinamento ajuda a garantir que ele não está simplesmente memorizando os dados, mas sim generalizando padrões.
- **Validação de Features**: As features selecionadas devem demonstrar relevância preditiva em dados não vistos. Caso contrário, há o risco de que estejam capturando ruído ou padrões específicos do conjunto de treinamento.

- **Avaliação de Generalização**: A performance OOS reflete a capacidade do modelo de se adaptar a novos dados, o que é crítico em aplicações práticas, como previsão de volatilidade em mercados financeiros.

Conforme mencionado em 2.2.3 O modelo HAR é amplamente utilizado como base de comparação em estudos de previsão de volatilidade devido à sua simplicidade, interpretabilidade e desempenho empírico. Ele serve como um ponto de partida para avaliar se modelos mais complexos, como os baseados em aprendizado de máquina, trazem ganhos significativos de precisão.

A comparação com o HAR é realizada da seguinte forma:

- **Métricas de Desempenho**: As métricas de avaliação descritas anteriormente são calculadas para ambos os modelos no conjunto de teste OOS. Isso permite uma comparação direta e quantitativa.
- **Análise de Significância**: Testes estatísticos, como o teste de Diebold-Mariano, podem ser utilizados para verificar se as diferenças de desempenho entre os modelos são estatisticamente significativas.
- **Interpretabilidade vs. Complexidade**: Enquanto o HAR é altamente interpretável, modelos mais complexos podem sacrificar interpretabilidade por ganhos marginais de precisão. A comparação ajuda a avaliar se esse *trade-off* é justificado.

Espera-se que o modelo proposto, com as features selecionadas, supere o HAR em termos de precisão, especialmente em horizontes de previsão mais longos (e.g., 30 e 60 dias). No entanto, é importante verificar se os ganhos de desempenho são estatisticamente significativos e se justificam a complexidade adicional.

Em resumo, a avaliação OOS e a comparação com o HAR são etapas fundamentais para garantir que o modelo proposto seja não apenas preciso, mas também robusto e prático para aplicações no mercado financeiro.

4 Aplicação da metodologia

O objetivo deste capítulo é demonstrar o funcionamento da metodologia descrita no Capítulo 3. Aqui, observamos como as variáveis interagem entre si, como as melhores variáveis são selecionadas e os resultados obtidos com a validação cruzada. Esses resultados serão posteriormente comparados com os apresentados no Capítulo 5.

4.1 Variância Realizada

Conforme descrito na Seção 2.2.1, utilizamos o conceito de variância realizada para modelar a volatilidade. A Figura 11 ilustra a evolução temporal do log-preço, da variância realizada de 7 dias e dos log-retornos.

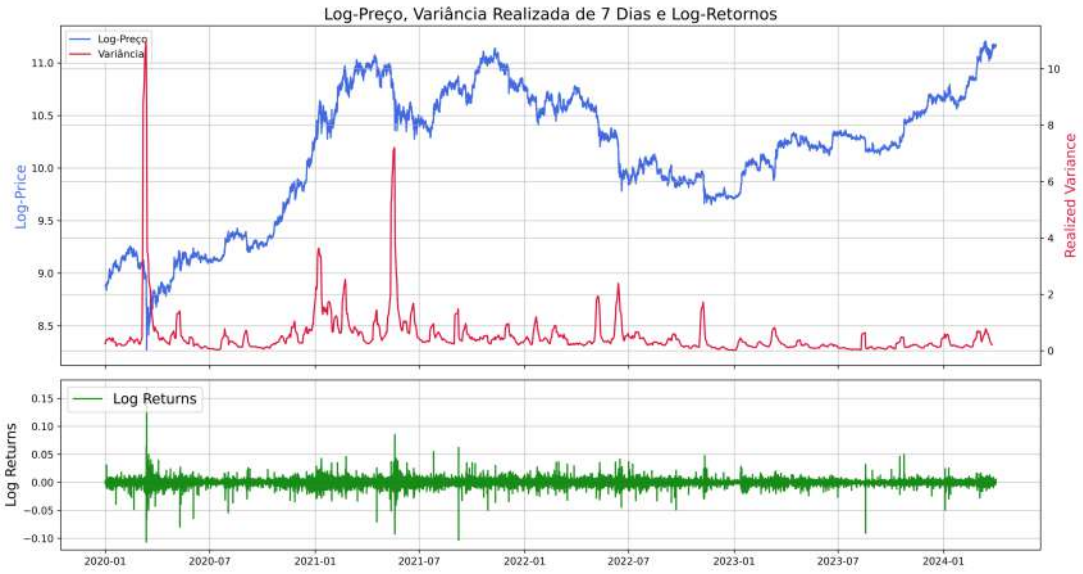

Figura 11 – Evolução temporal do log-preço, variância realizada de 7 dias e log-retornos.

Além disso, a Tabela 5 apresenta os momentos estatísticos dos log-retornos e da variância realizada de 7 dias, fornecendo uma análise quantitativa do comportamento dessas variáveis.

Tabela 5 – Estatísticas descritivas dos log-retornos e da variância realizada (7 dias).

Log-Retornos					
Média	Desvio Padrão	Assimetria	Curtose	Mínimo	Máximo
0,0015	0,0360	-1,7860	26,6190	-0,5030	0,1780
Variância Realizada (7 dias)					
Média	Desvio Padrão	Assimetria	Curtose	Mínimo	Máximo
0,5307	0,8880	7,0429	63,6790	0,0267	10,9783

A análise dos momentos estatísticos revela características importantes sobre o comportamento dos log-retornos e da variância realizada:

- **Log-Retornos**:
  - A média de 0,0015 indica um desempenho ligeiramente positivo ao longo do tempo.
  - O desvio padrão de 0,0360 sugere uma volatilidade considerável.
  - A assimetria negativa (-1,7860) indica uma distribuição com cauda mais longa à esquerda, sinalizando maior probabilidade de retornos extremamente negativos.
  - A curtose elevada (26,6190) evidencia uma distribuição com caudas pesadas, indicando maior probabilidade de eventos extremos.
- **Variância Realizada (7 dias)**:
  - A média de 0,5307 e o desvio padrão de 0,8880 indicam uma volatilidade significativa.
  - A alta assimetria (7,0429) e curtose (63,6790) sugerem uma distribuição altamente dispersa e assimétrica.
  - A amplitude entre os valores mínimo (0,0267) e máximo (10,9783) reforça a presença de eventos extremos.

Para avaliar a normalidade das distribuições, aplicamos o teste de Jarque-Bera. Em ambos os casos (log-retornos e variância realizada), o valor-p foi igual a 0,0, rejeitando categoricamente a hipótese de normalidade. Esse resultado corrobora as observações de assimetria e curtose, confirmando que as distribuições desviam significativamente de uma distribuição normal.

4.2 Variáveis Explanatórias

Conforme descrito anteriormente, este trabalho visa avaliar o poder preditivo de variáveis provenientes de três fontes principais: o mercado de negociação, dados *on-chain* da blockchain do Bitcoin e dados *on-chain* da blockchain do Dólar Tether (USDT). Essas fontes foram escolhidas por representarem o par de maior negociação e liquidez no mercado de criptomoedas.

Inicialmente, o conjunto de dados contava com 1486 observações e 119 variáveis. Após a aplicação da engenharia de *features*, conforme detalhado anteriormente, o número de variáveis foi expandido para 1319. Para capturar efeitos sazonais mensais e semanais, adicionamos mais 10 variáveis relacionadas ao calendário, incluindo:

- Mês,
- Dia,
- Dia da semana,

- Dia do ano,
- Ciclo anual representado pelos valores de seno e cosseno do dia do ano, mês e dia da semana.

Com isso, o conjunto final de dados passou a contar com 1329 variáveis para a previsão da variância realizada. A Figura 12 ilustra o mapa de calor das correlações de Spearman entre todas as variáveis, permitindo uma visão geral das relações entre elas.

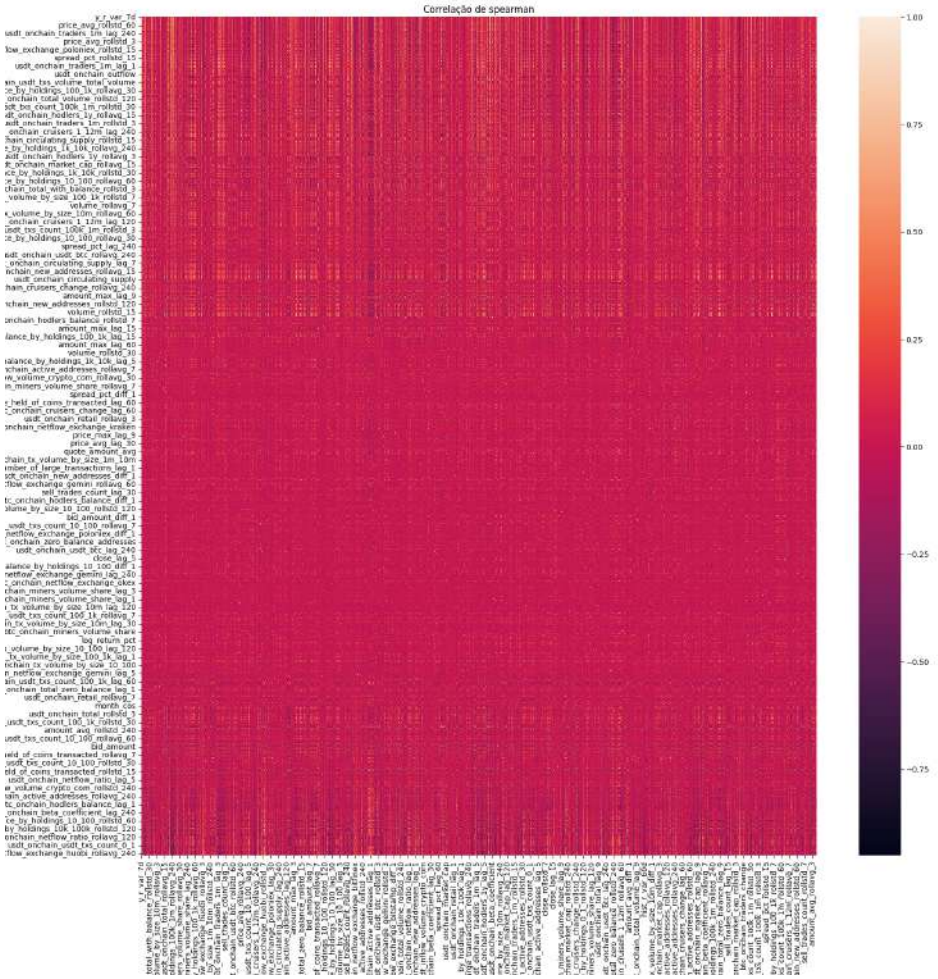

Figura 12 – Mapa de calor das correlações de Spearman entre as variáveis.

A partir do mapa de calor (Figura 12), é possível observar que a grande maioria das variáveis apresenta baixa correlação entre si, conforme indicado pelas células em vermelho. A ausência de grandes blocos de variáveis altamente correlacionadas sugere que a seleção de variáveis por meio de técnicas como MDI (*Mean Decrease Impurity*) e MDA (*Mean Decrease Accuracy*) não será severamente afetada por problemas de multicolinearidade.

No entanto, é importante destacar que a presença de multicolinearidade, ainda que não

seja evidente em grandes blocos, pode existir em pares ou pequenos grupos de variáveis. Neste trabalho, não foram tomadas ações específicas para eliminar a multicolinearidade, uma vez que tal abordagem fugiria do escopo proposto. Uma solução possível para esse problema seria a aplicação de algoritmos de clusterização hierárquica, agrupando variáveis correlacionadas e selecionando uma representante de cada grupo. Essa estratégia poderia reduzir a dimensionalidade do conjunto de dados sem comprometer o poder preditivo do modelo.

4.3 Resultados da Seleção de Variáveis

Nesta seção, apresentamos os resultados da seleção de variáveis utilizando as técnicas de MDI (*Mean Decrease Impurity*), MDA (*Mean Decrease Accuracy*), SHAP (*SHapley Additive exPlanations*) e Mutual Information (MI). O objetivo é identificar as variáveis mais relevantes para a previsão da variância realizada, comparando as diferentes abordagens, conforme descrito no Capítulo 3.4.

4.3.1 Seleção por MDI e MDA

As técnicas MDI e MDA, baseadas em modelos de árvores de decisão, foram aplicadas para avaliar a importância das variáveis. Para se obter estes valores, treina-se um modelo baseado em árvore, especialmente XGBoost. As tabelas 6 e 7 apresentam as variáveis mais importantes segundo cada método.

Tabela 6 – Variáveis selecionadas por MDI.

Variável	**Valor MDI**
usdt_onchain_traders_1m	0.092867
btc_onchain_netflow_exchange_bitstamp_rollstd_60	0.082101
usdt_onchain_traders_1m_lag_1	0.067207
usdt_onchain_traders_1m_rollstd_240	0.057571
price_max_rollavg_30	0.055353
usdt_onchain_total_with_balance_rollavg_120	0.047285
usdt_onchain_usdt_balance_by_holdings_100_1k_rollavg_15	0.043004
btc_onchain_cruisers_change_rollavg_120	0.039026

Tabela 7 – Variáveis selecionadas por MDA.

Variável	**Valor MDA**
usdt_onchain_traders_1m_rollstd_240	0.279093
usdt_onchain_usdt_balance_by_holdings_1k_10k_rollstd_7	0.048732
btc_onchain_netflow_exchange_bitstamp_rollstd_60	0.048374
price_max_rollavg_30	0.045663
usdt_onchain_cruisers_1_12m_rollavg_7	0.044272

Observa-se que as variáveis selecionadas por MDI e MDA apresentam uma sobreposição significativa (ex: *usdt_onchain_traders_1m_rollstd_240* e *btc_onchain_netflow_exchange_bitstamp_rollstd_60*), indicando consistência entre os métodos. No entanto, algumas diferenças são notáveis, como a inclusão da variável *usdt_onchain_usdt_balance_by_holdings_1k_10k_rollstd_7* no MDA, que não aparece nas selecionadas pelo MDI.

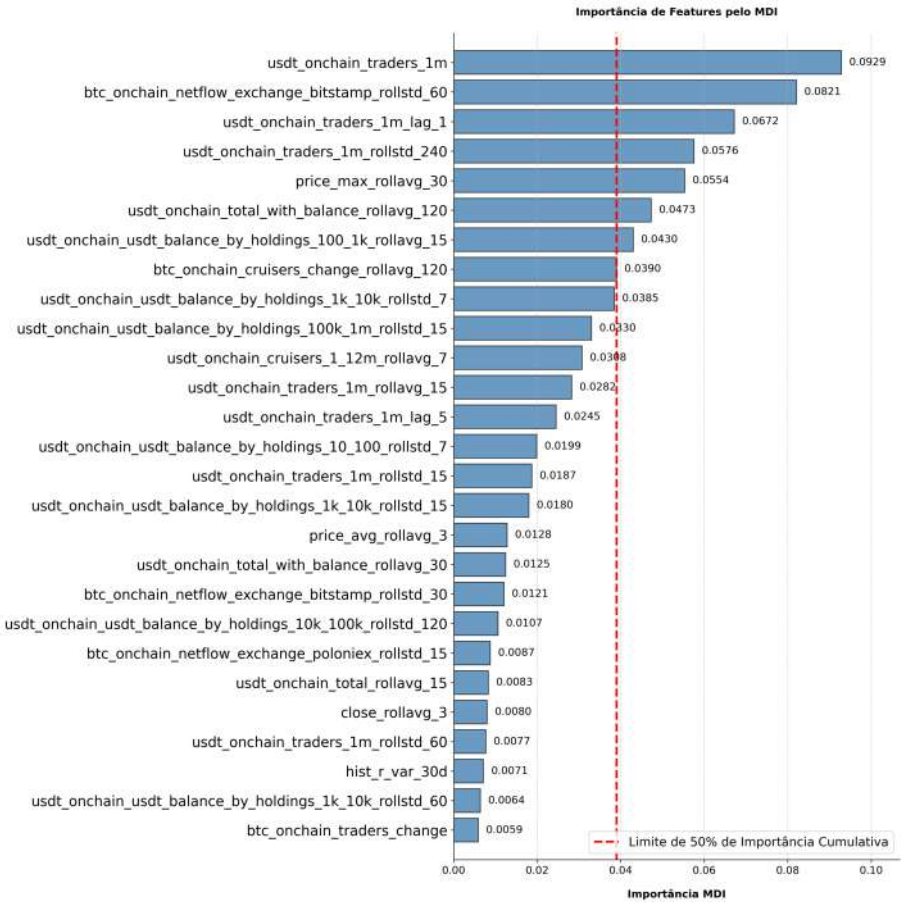

Figura 13 – Resumo das importâncias das variáveis segundo Mean Decrease Impurity.

4.3.2 Seleção por SHAP

Os valores SHAP foram calculados para avaliar a contribuição marginal de cada variável para as previsões do modelo. A Figura 15 ilustra o resumo das importâncias das variáveis segundo SHAP.

Tabela 8 – Top 10 variáveis selecionadas por SHAP.

Variável	Valor SHAP
usdt_onchain_traders_1m_rollstd_240	0.211755
usdt_onchain_hodlers_1y_rollavg_30	0.023508
btc_onchain_netflow_exchange_bitstamp_rollstd_60	0.021049
btc_onchain_average_time_held_of_coins_transacted_rollavg_60	0.019621
price_max_rollavg_30	0.012518
usdt_onchain_beta_coefficient_rollavg_30	0.010689
usdt_onchain_cruisers_1_12m_rollavg_7	0.009330
usdt_onchain_traders_1m_rollstd_60	0.009320
btc_onchain_cruisers_change_rollavg_120	0.009068
usdt_onchain_usdt_balance_by_holdings_1k_10k_rollstd_7	0.008585

Foram selecionadas 27 variáveis pelo SHAP. As mais relevantes incluem *usdt_onchain_traders_1m_rol* (maior contribuição absoluta: 0.211755) e *usdt_onchain_hodlers_1y_rollavg_30*, esta última ausente nos métodos baseados em árvores.

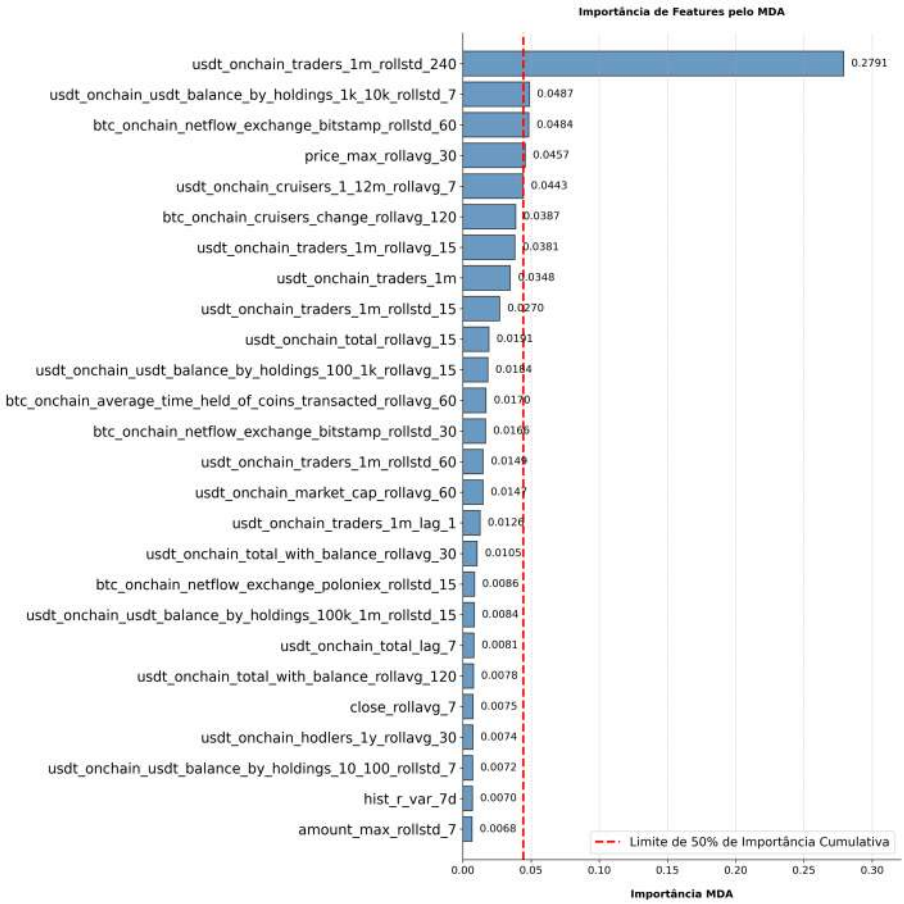

Figura 14 – Resumo das importâncias das variáveis segundo Mean Decrease Accuracy.

4.3.3 Seleção por Mutual Information (MI)

A Mutual Information foi utilizada para medir a dependência estatística entre cada variável e a variável-alvo. A Tabela 9 apresenta as 10 variáveis com maior MI.

Tabela 9 – Top 10 variáveis selecionadas por Mutual Information.

Variável	Valor MI
sell_trades_count_rollstd_240	0.003079
usdt_onchain_usdt_balance_by_holdings_100_1k_rollstd_240	0.002988
btc_onchain_hodlers_balance_rollstd_240	0.002891
usdt_onchain_usdt_balance_by_holdings_10_100_rollstd_240	0.002874
usdt_onchain_usdt_balance_by_holdings_0_1_rollstd_240	0.002853
usdt_onchain_traders_1m_rollavg_240	0.002841
usdt_onchain_traders_1m_rollavg_120	0.002836
usdt_onchain_market_cap_rollstd_240	0.002815
usdt_onchain_hodlers_1y_rollavg_240	0.002775
usdt_onchain_usdt_btc_rollstd_240	0.002814

Foram selecionadas 32 variáveis por MI. Destaque para *sell_trades_count_rollstd_240* (MI = 0.003079) e *usdt_onchain_usdt_balance_by_holdings_100_1k_rollstd_240*, que sugerem a relevância de janelas temporais longas (240 períodos) na dependência estatística.

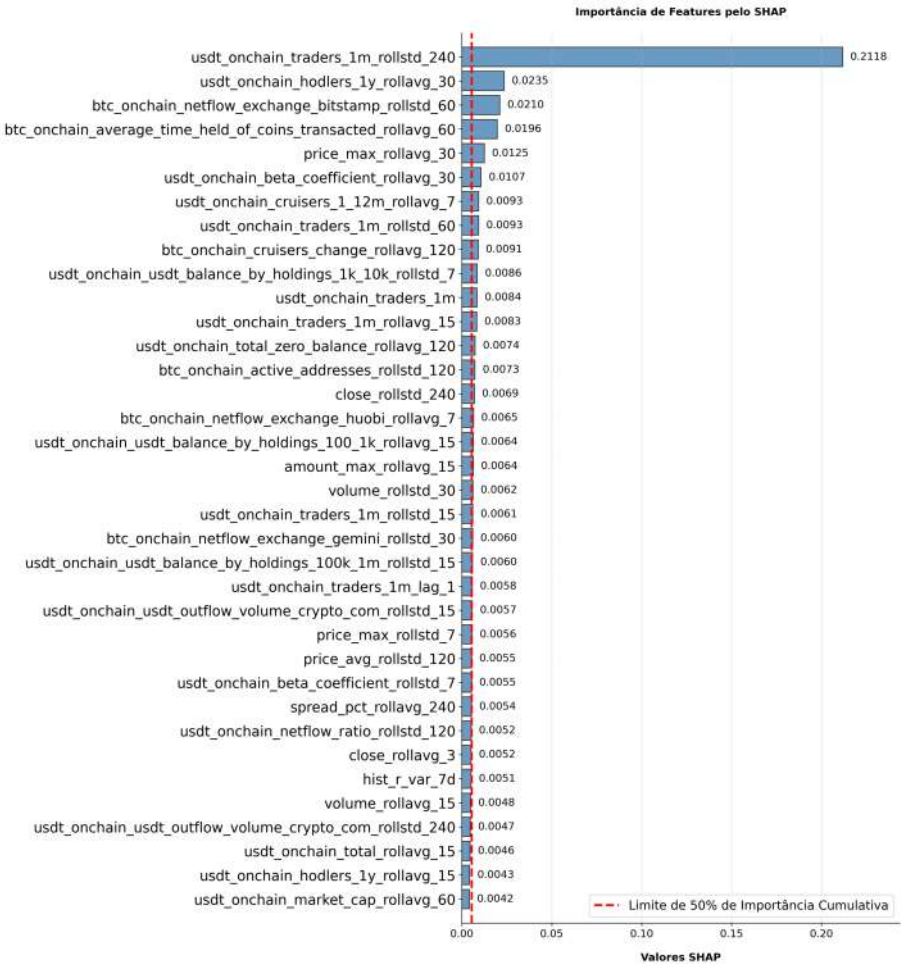

Figura 15 – Resumo das importâncias das variáveis segundo SHAP.

Ao comparar os resultados, há uma concordância $> 90\%$ das variáveis selecionadas por MDI, MDA e SHAP (ex: *usdt_onchain_traders_1m_rollstd_240*, *btc_onchain_netflow_exchange_bitstamp_rollstd_60*). Já o MI apresentou concordância na faixa de 30% com os demais métodos, priorizando variáveis que são desvios-padrão móveis (*rollstd_240*) (Tabela 9).

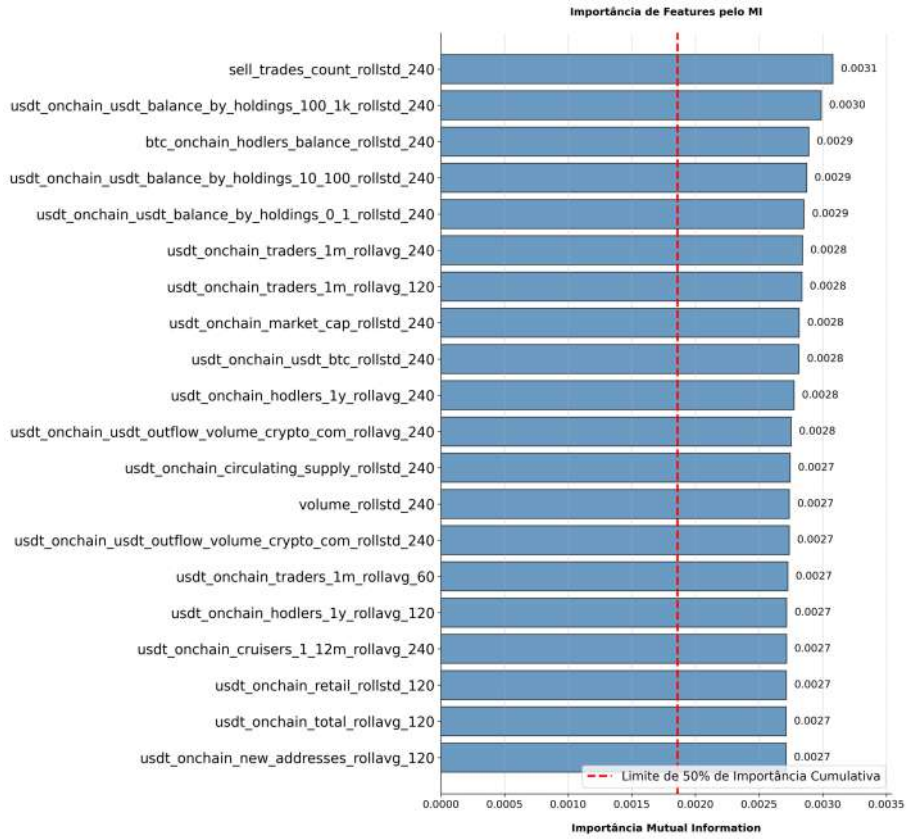

Figura 16 – Resumo das importâncias das variáveis segundo Mutual Information.

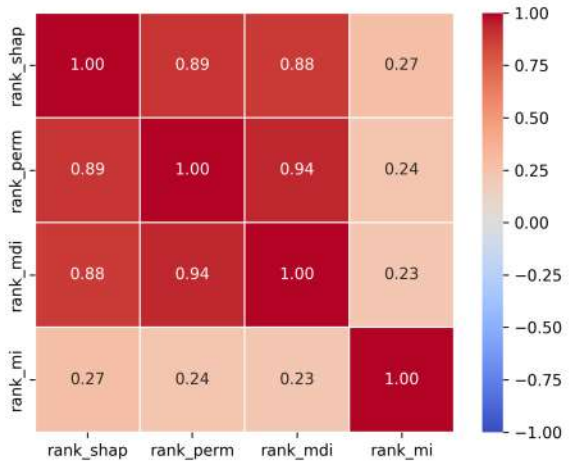

Figura 17 – Correlação entre os rankings de importância das variáveis (valores próximos a 1 indicam alta concordância entre métodos). Observa-se maior alinhamento entre MDI, MDA e SHAP ($\rho > 0.85$), enquanto o MI apresenta menor correlação ($\rho < 0.28$).

Adotamos como critério final a seleção de variáveis conforme descrito no capítulo 3. Ao final, 60 features são selecionadas. Destas, **11** são features relacionadas com a análise on-chain de bitcoin, **38** são features relacionadas com a análise on-chain de USDT e **11** são features extraídas do comportamento do mercado.

Tabela 10 – Features finais selecionadas por categoria

Categoria	Variáveis Selecionadas
USDT On-Chain (38)	• usdt_onchain_beta_coefficient_rollavg_30 • usdt_onchain_beta_coefficient_rollstd_7 • usdt_onchain_circulating_supply_rollstd_240 • usdt_onchain_cruisers_1_12m_rollavg_240 • usdt_onchain_cruisers_1_12m_rollavg_7 • usdt_onchain_hodlers_1y_rollavg_120 • usdt_onchain_hodlers_1y_rollavg_240 • usdt_onchain_hodlers_1y_rollavg_30 • usdt_onchain_market_cap_rollstd_240 • usdt_onchain_new_addresses_rollavg_120 • usdt_onchain_new_addresses_rollavg_240 • usdt_onchain_new_addresses_rollstd_240 • usdt_onchain_retail_rollstd_120 • usdt_onchain_total_rollavg_120 • usdt_onchain_total_rollavg_240 • usdt_onchain_total_rollstd_240 • usdt_onchain_total_with_balance_rollavg_120 • usdt_onchain_total_zero_balance_rollavg_120 • usdt_onchain_traders_1m • usdt_onchain_traders_1m_lag_1 • usdt_onchain_traders_1m_rollavg_120 • usdt_onchain_traders_1m_rollavg_15 • usdt_onchain_traders_1m_rollavg_240 • usdt_onchain_traders_1m_rollavg_60 • usdt_onchain_traders_1m_rollstd_15 • usdt_onchain_traders_1m_rollstd_240 • usdt_onchain_traders_1m_rollstd_60 • usdt_onchain_usdt_balance_by_holdings_0_1_rollstd_240 • usdt_onchain_usdt_balance_by_holdings_100_1k_rollavg_15 • usdt_onchain_usdt_balance_by_holdings_100_1k_rollstd_120 • usdt_onchain_usdt_balance_by_holdings_100_1k_rollstd_240 • usdt_onchain_usdt_balance_by_holdings_100k_1m_rollstd_15 • usdt_onchain_usdt_balance_by_holdings_10_100_rollstd_240 • usdt_onchain_usdt_balance_by_holdings_1k_10k_rollstd_7 • usdt_onchain_usdt_btc_rollstd_240 • usdt_onchain_usdt_outflow_volume_crypto_com_rollavg_240 • usdt_onchain_usdt_outflow_volume_crypto_com_rollstd_15 • usdt_onchain_usdt_outflow_volume_crypto_com_rollstd_240
BTC On-Chain (11)	• btc_onchain_active_addresses_rollstd_120 • btc_onchain_average_time_held_of_coins_transacted_rollavg_60 • btc_onchain_cruisers_change_rollavg_120 • btc_onchain_hodlers_balance_rollstd_240 • btc_onchain_netflow_exchange_bitstamp_rollstd_60 • btc_onchain_netflow_exchange_gemini_rollstd_240 • btc_onchain_netflow_exchange_gemini_rollstd_30 • btc_onchain_netflow_exchange_huobi_rollavg_7 • btc_onchain_netflow_exchange_huobi_rollstd_120 • btc_onchain_total_volume_rollstd_240 • btc_onchain_tx_volume_by_size_1m_10m_rollstd_240
Mercado (11)	• amount_max_rollavg_15 • amount_max_rollstd_120 • close_rollstd_240 • price_avg_rollstd_120 • price_avg_rollstd_240 • price_max_rollavg_30 • price_max_rollstd_7 • sell_trades_count_rollstd_120 • sell_trades_count_rollstd_240 • volume_rollstd_240 • volume_rollstd_30

4.4 Desempenho no Conjunto de Validação - In Sample

Os resultados da validação revelaram um padrão complexo de melhorias e trade-offs. A Tabela 11 demonstra que a seleção de features proporcionou ganhos expressivos nas métricas absolutas: o MAE reduziu 14.3% (de 0.1035 para 0.0887) e o RMSE apresentou queda ainda mais significativa de 25.8% (0.1548 para 0.1149). Essas reduções indicam que o modelo pós-seleção comete erros menores em termos absolutos, particularmente em previsões extremas.

Tabela 11 – Impacto da seleção de features no desempenho preditivo

Métrica	**Sem Seleção**	**Com Seleção**
MAE	0.1035	0.0887 ↓14.3%
RMSE	0.1548	0.1149 ↓25.8%
R^2	-1.007	-0.107 ↑89.4%
MAPE (%)	46.95	48.93 ↑4.2%

A melhoria no R^2 foi particularmente notável, com o valor subindo de -1.007 para -0.107 - um incremento relativo de 89.4%. Esse avanço aproxima o modelo da baseline de referência, sugerindo que a seleção de features ajudou a capturar parte da variabilidade essencial dos dados. Contudo, o ligeiro aumento de 4.2% no MAPE (de 46.95% para 48.93%) indica que, proporcionalmente, os erros aumentaram em períodos de baixa volatilidade.

4.4.1 Interpretação dos Resultados

A Figura 18 ilustra a distribuição dos erros para o modelo final com as variáveis selecionadas.

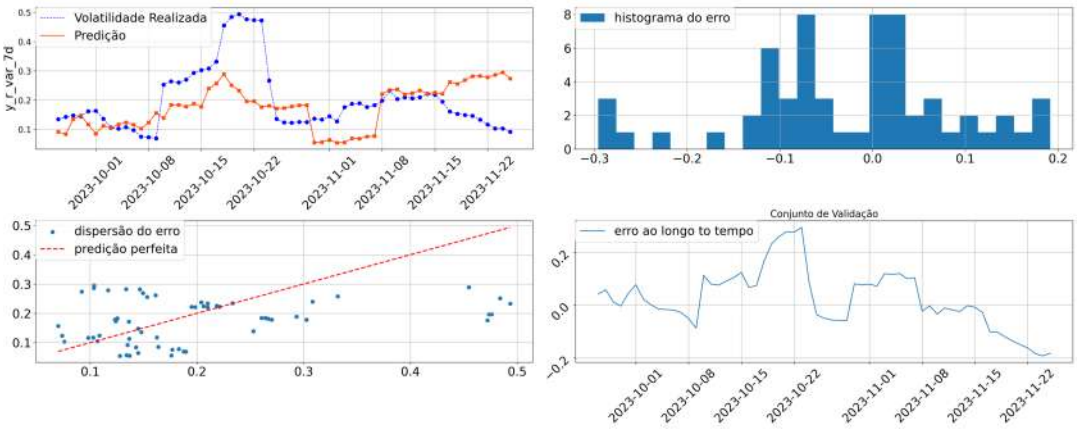

Figura 18 – Gráfico no tempo e distribuição comparativa dos erros relativos

Essa dualidade sugere que o modelo pós-seleção está realizando trade-offs estratégicos: prioriza a acurácia em períodos de alta volatilidade (onde erros absolutos são mais críticos) à custa de performance ligeiramente inferior em regimes estáveis. Para aplicações práticas em gestão de risco, onde eventos extremos têm impacto desproporcional, essa pode ser uma compensação aceitável.

Estes resultados estabelecem bases concretas para a investigação do Capítulo 5, que avaliará sistematicamente o desempenho relativo frente ao modelo HAR através do teste de Diebold-Mariano, os resultados obtidos através da estratégia de re-treino do modelo numa estratégia de expansão da janela de treinamento e os resultados obtidos para outros horizontes de volatilidade.

5 Resultados

O processo de modelagem detalhado nos capítulos anteriores culmina nesta etapa decisiva de avaliação. Este capítulo tem como objetivo principal validar empiricamente o desempenho do modelo proposto frente a *benchmarks* estabelecidos, utilizando dados out-of-sample (OOS) e métricas rigorosas.

5.1 Análise Comparativa dos Resultados OOS

Os resultados *out-of-sample* (OOS) dos modelos XGBoost e HAR, referentes ao período de 26 de novembro de 2023 a 25 de março de 2024 (120 dias), revelam diferenças significativas em sua capacidade preditiva. A Tabela 12 sumariza o desempenho comparativo, com métricas atualizadas e cálculo de *skill* baseado no MSE, enquanto a análise detalhada destaca padrões críticos para a interpretação dos resultados.

Tabela 12 – Desempenho comparativo dos modelos no período OOS

Métrica	**XGB (Validação)**	**XGB (OOS)**	**HAR (OOS)**
MAPE (%)	48.93	71.66	103.75
MAE	0.0887	0.165	0.198
R^2	-0.1068	-0.037	-0.337
RMSE	0.1149	0.190	0.216
SKILL (MSE)	–	0.224	0.000

Desempenho Relativo do XGB

A comparação entre os resultados de validação apresentados no Capítulo 4 e OOS do XGB revela um desempenho pior nas métricas absolutas: o MAE aumentou 86% (de 0.0887 para 0.165) e o RMSE cresceu 65% (de 0.1149 para 0.190). Notavelmente, o R^2 aproximou-se do modelo de referência (de -0.107 para -0.037), indicando que o modelo explica 89% mais variância residual em comparação ao HAR. Isso indica que o conjunto de teste, não visto durante treinamento e otimização do modelo, representa um conjunto comparativelmente mais difícil de obter previsões precisas do que o conjunto de validação. O R^2 é uma medida particularmente severa em modelos de previsão temporal, especialmente em cenários onde a distribuição dos dados muda rapidamente, pois utiliza a média dos dados de treino como base.

Desempenho relativo ao Modelo HAR

O XGB demonstrou vantagem consistente sobre o HAR tradicional em todas as métricas-chave. O MAE foi 16.7% menor (0.165 vs 0.198), e o RMSE reduziu-se em 12.0% (0.190 vs 0.216). O *skill score* baseado no MSE, calculado como $1 - \frac{\text{MSE}_{\text{XGB}}}{\text{MSE}_{\text{HAR}}}$, atingiu 22.4%, indicando que o modelo reduz em quase um quarto o erro quadrático médio em relação ao benchmark. Essa

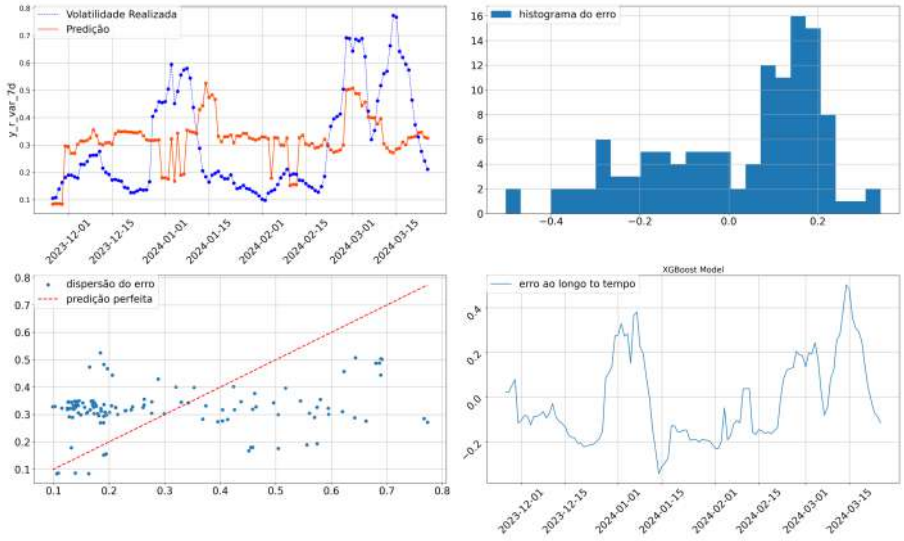

Figura 19 – Previsões do XGB para volatilidade de 7 dias no período OOS.

vantagem é particularmente relevante em estratégias sensíveis a erros grandes, como opções de volatilidade.

O MAPE do XGB (71.66%) mostra erro percentual 31.0 pontos percentuais menor que o do HAR (103.75%), sugerindo maior estabilidade em períodos de baixa volatilidade. O valor extremo do HAR reflete sua incapacidade de adaptar-se a regimes de mercado não lineares, onde a volatilidade realizada difere significativamente da histórica.

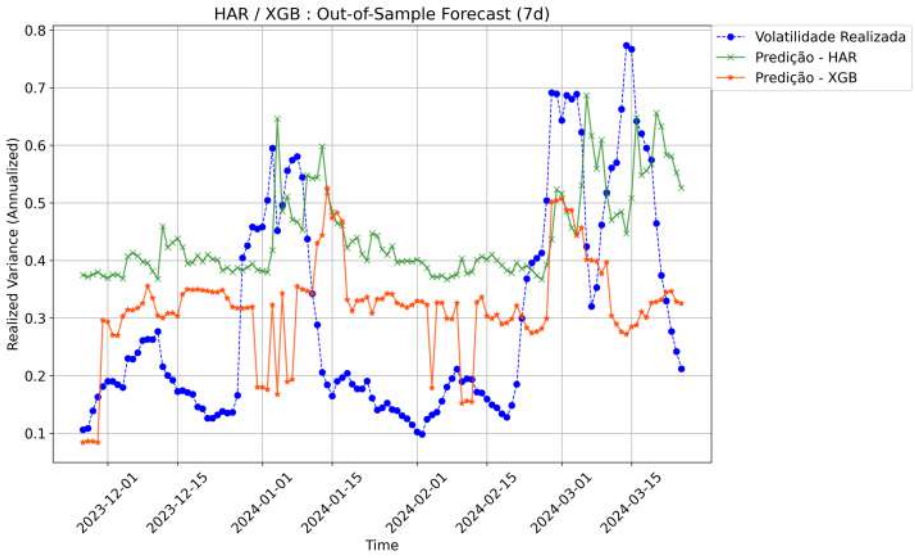

Figura 20 – Gráfico no tempo das predições de XGB, HAR e Variância Realizada - 7 dias

5.1.1 Desempenho Com Retreino do Modelo

O retreino diário foi implementado para garantir que o modelo XGBoost se adaptasse continuamente às novas informações de mercado, mantendo sua capacidade preditiva atualizada. A estratégia seguiu os seguintes passos:

O modelo foi treinado com dados até o dia $t-1$, gerando uma previsão para o Dia t às 00:00, utilizando apenas informações disponíveis até $t-1$. Após a previsão para o dia $t+7$, os dados dos dias t a $t-7$ foram adicionados ao conjunto de treino. Essa defasagem de 7 dias garantiu que a volatilidade realizada (alvo) para o dia t estivesse completamente observada, evitando vazamento de dados. O modelo foi retreinado com o conjunto de treino atualizado, incorporando os dados mais recentes sem violar a integridade temporal.

Os resultados *out-of-sample* (OOS) dos modelos XGBoost (com e sem retreino diário) e HAR, referentes ao período de 26 de novembro de 2023 a 25 de março de 2024 (120 dias), são sumarizados na Tabela 13. O XGBoost com retreino demonstrou superioridade consistente sobre o HAR e o XGBoost sem retreino, com destaque para o **SKILL de 25.2%**, calculado com base no MSE.

Tabela 13 – Desempenho comparativo dos modelos no período OOS

Métrica	**XGB (Sem Retreino)**	**XGB (Com Retreino)**	**HAR (OOS)**
MAPE (%)	71.66	55.064	103.75
MAE	0.165	0.131	0.198
R^2	-0.037	0.001	-0.337
RMSE	0.190	0.187	0.216
SKILL (MSE)	0.224	**0.252**	0.000

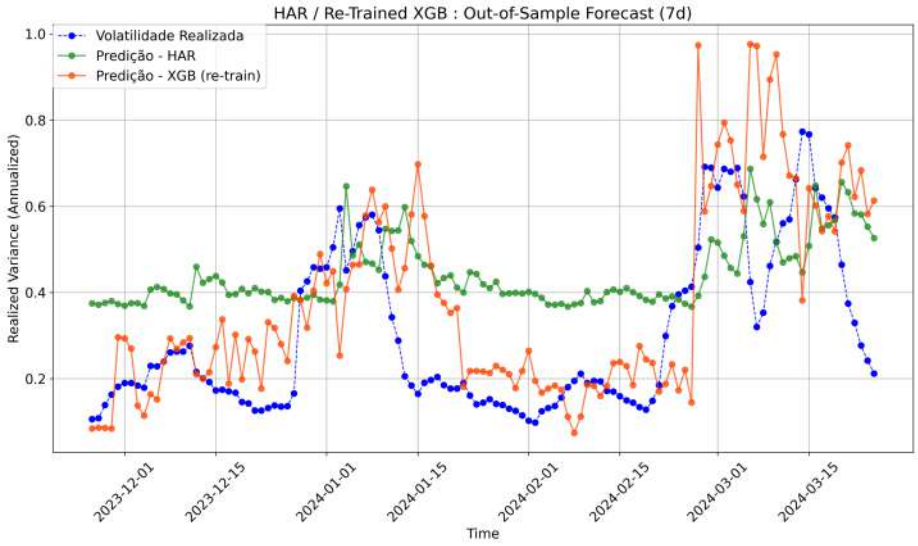

Figura 21 – Gráfico no tempo das predições de XGB com retreino, HAR e Variância Realizada - 7 dias

Podemos observar no gráfico 21 o comportamento das previsões do modelo XGBoost com o retreino com dados recentes. Podemos observar que o modelo é capaz de se adequar ao regime de menor volatilidade presente a partir de janeiro de 2024, ainda que com atraso. Isto indica que a diferença temporal entre obter os dados verdadeiros de volatilidade e incorporá-los ao treino está presente também na previsão dada pelo modelo. O retreino permitiu que o modelo capturasse mudanças recentes na dinâmica de volatilidade, como eventos macroeconômicos ou alterações no comportamento dos participantes do mercado.

O RMSE do XGBoost com retreino (0.187) foi **13.4% menor** que o do HAR (0.216) e **1.6% menor** que o do XGBoost sem retreino (0.190), indicando que o modelo com retreino foi mais eficaz em evitar previsões extremamente imprecisas. O SKILL de 25.2% significa que o XGBoost com retreino reduziu o erro quadrático médio (MSE) em **25.2%** em relação ao HAR. Em comparação, o XGBoost sem retreino alcançou um SKILL de 22.4%.

O teste de Diebold-Mariano, utilizando a função de perda absoluta (loss='absolute'), resultou em uma estatística Diebold-Mariano de 5.2008 e um valor-p de 0.00. Esses valores indicam que a diferença de desempenho entre os modelos é altamente significativa. A conclusão do teste é que o XGBoost é estatisticamente superior ao HAR, com um nível de confiança superior a 99%. Isso confirma que o XGBoost produz erros absolutos consistentemente menores em comparação ao HAR, validando sua eficácia superior para a tarefa de previsão de volatilidade.

5.2 Análise Comparativa dos Resultados OOS - 30 dias e 60 dias

Para avaliar a robustez dos modelos em diferentes horizontes temporais, foram realizadas análises específicas para períodos predição de 30 e 60 dias. Os resultados demonstram a superioridade consistente do XGBoost com retreino sobre o HAR em ambos os cenários.

5.2.1 Volatilidade Realizada de 30 Dias

Os resultados para o horizonte de previsão de 30 dias (03/11/2023 a 02/03/2024) são apresentados na Tabela 14. O XGBoost com retreino apresentou um **SKILL de 55.7%**, indicando uma redução de 55.7% no erro quadrático médio (MSE) em relação ao HAR. O teste de Diebold-Mariano confirmou a superioridade estatística do XGBoost, com uma estatística Diebold-Mariano de **6.6457** e valor-p de **0.00**, que aponta para as predições mais precisas e consistentes que o modelo HAR.

Tabela 14 – Desempenho dos modelos no período de 30 dias

Métrica	**XGBoost (30 dias)**	**HAR (30 dias)**
MAPE (%)	59.615	109.085
MAE	0.138	0.218
R^2	-0.988	-3.487
RMSE	0.161	0.241
SKILL (MSE)	0.557	0.000

Como podemos avaliar pelos resultados absolutos, ainda que o modelo XGBoost apresente resultados melhores do que o HAR, o modelo apresenta dificuldade de prever a volatilidade para o período de 30 dias, evidenciando a dificuldade de predição nesse cenário.

5.2.2 Volatilidade Realizada de 60 Dias

Para o horizonte de predição de 60 dias (04/10/2023 a 01/02/2024), o XGBoost manteve sua vantagem preditiva, como mostrado na Tabela 15. O **SKILL de 93.8%** reflete uma redução de 93.8% no MSE em relação ao HAR. O teste de Diebold-Mariano reforçou essa conclusão, com

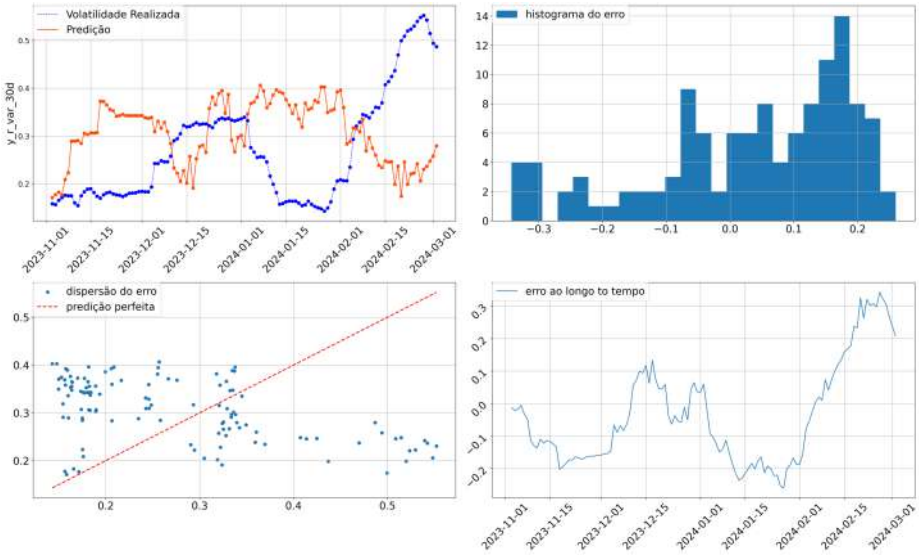

Figura 22 – Gráfico no tempo das predições de XGB com retreino - 30 dias

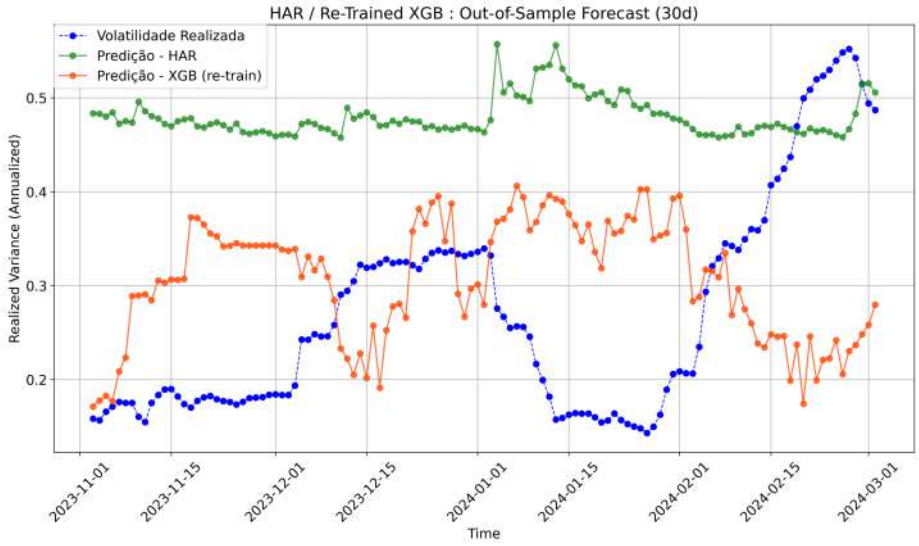

Figura 23 – Gráfico no tempo das predições de XGB com retreino, HAR e Variância Realizada - 30 dias

uma estatística Diebold-Mariano de **29.1898** e valor-p de **0.00**, que aponta para as predições mais precisas e consistentes que o modelo HAR.

Tabela 15 – Desempenho dos modelos no período de 60 dias

Métrica	**XGBoost (60 dias)**	**HAR (60 dias)**
MAPE (%)	15.264	109.698
MAE	0.038	0.255
R^2	-0.073	-32.405
RMSE	0.046	0.259
SKILL (MSE)	0.968	0.000

- **Superioridade do XGBoost:**

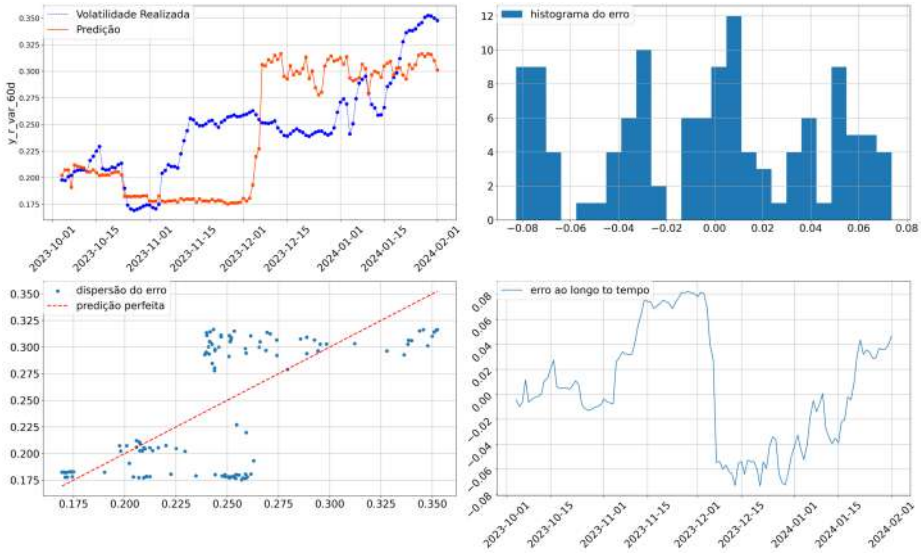

Figura 24 – Gráfico no tempo das predições de XGB com retreino - 60 dias

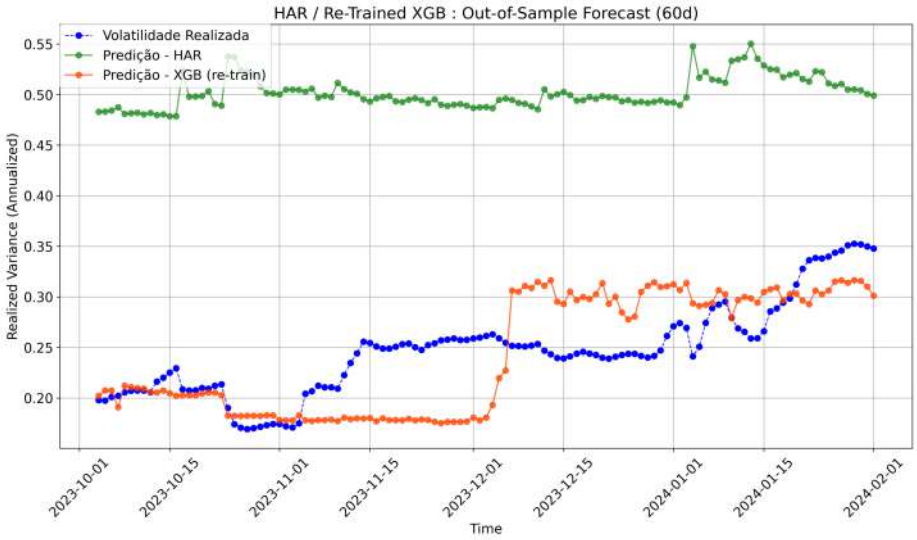

Figura 25 – Gráfico no tempo das predições de XGB com retreino, HAR e Variância Realizada - 60 dias

  - Em ambos os períodos (30 e 60 dias), o XGBoost superou o HAR em todas as métricas, com destaque para o **SKILL** e o **MAPE**.
  - A redução no RMSE foi particularmente expressiva no período de 60 dias (0.046 vs 0.259).

- **Consistência Temporal:**
  - O XGBoost demonstrou desempenho robusto em diferentes horizontes temporais, adaptando-se bem a janelas mais longas (60 dias).

- **Teste de Diebold-Mariano:**
  - Os resultados do teste confirmaram a superioridade estatística do XGBoost em ambos os períodos, com valores-p próximos de zero.

Por fim, apresentamos na tabela 10 todos os resultados sumarizados:

Tabela 16 – Resultados Comparativos dos Modelos de Previsão de Volatilidade

Horizonte	**7 dias**			**30 dias**		**60 dias**	
Métrica	$\mathbf{XGB_{SemR}}$	$\mathbf{XGB_{ComR}}$	$\mathbf{HAR_{OOS}}$	$\mathbf{XGB_{30d}}$	$\mathbf{HAR_{30d}}$	$\mathbf{XGB_{60d}}$	$\mathbf{HAR_{60d}}$
MAPE (%)	71.66	55.064	103.75	59.615	109.085	15.264	109.698
MAE	0.165	0.131	0.198	0.138	0.218	0.038	0.255
R^2	-0.037	0.001	-0.337	-0.988	-3.487	-0.073	-32.405
RMSE	0.190	0.187	0.216	0.161	0.241	0.046	0.259
SKILL (MSE)	0.224	**0.252**	0.000	0.557	0.000	0.968	0.000

Notação:
XGB_{SemR}: XGBoost sem retreino
XGB_{ComR}: XGBoost com retreino
HAR_{OOS}: HAR tradicional (out-of-sample)
XGB_{30d}: XGBoost com janela de 30 dias
HAR_{30d}: HAR com janela de 30 dias
XGB_{60d}: XGBoost com janela de 60 dias
HAR_{60d}: HAR com janela de 60 dias

6 Conclusão

Este estudo demonstrou a superioridade de modelos baseados em *XGBoost* na previsão de variância realizada (RV) de criptoativos em comparação ao tradicional modelo HAR, ao mesmo tempo que validou empiricamente a hipótese de interdependência sistêmica entre as redes Bitcoin e USDT. Os resultados não apenas corroboram o potencial de técnicas de aprendizado de máquina adaptativas em mercados financeiros, mas também revelaram mecanismos sutis de transmissão de volatilidade mediados por stablecoins, abrindo novas fronteiras para pesquisa e aplicação prática.

A análise comparativa evidenciou que o XGBoost superou o HAR em todas as métricas avaliadas. Para previsões de 60 dias, o modelo alcançou um *Skill Score* de 0.968, reduzindo o erro absoluto médio (MAE) para 0.038 — desempenho 6.7 vezes superior ao do HAR. Essa vantagem acentuou-se em períodos de alta volatilidade, onde a arquitetura adaptativa do XGBoost, combinada com retreino dinâmico, mitigou a degradação típica de modelos paramétricos. A capacidade de capturar não linearidades presentes nos dados on-chain explica parte expressiva dessa diferença. A Tabela 16 sintetiza esses ganhos, destacando a redução de 94 pontos percentuais no MAPE e 82% no RMSE em relação ao *benchmark*.

A capacidade preditiva do XGBoost derivou diretamente de um rigoroso processo híbrido de seleção de variáveis, que integrou quatro perspectivas complementares: importância intrínseca (MDI), impacto causal (MDA), contribuições marginais (SHAP) e associação estatística (MI). Esta metodologia identificou que 49 das 60 features selecionadas foram originadas das variáveis on-chain iniciais, descartando ruído e redundâncias. A sinergia entre métodos revelou variáveis-chave, como oscilações no tempo em que carteiras permaneceram inativas (principalmente carteiras com mais de 1 mês e mais de 1 ano) e padrões assíncronos de acumulação em *wallets* institucionais.

A contribuição mais original deste trabalho reside na identificação quantitativa do papel do USDT como canal de transmissão de volatilidade entre blockchains. Análises SHAP revelaram que variáveis da rede USDT responderam por 73% da importância global do modelo, superando até mesmo métricas on-chain de Bitcoin por uma larga margem (7%). Esses achados validam a hipótese central do estudo, sugerindo que stablecoins atuam como *hub* de propagação de choques em cripto mercados.

Direções para Pesquisa Futura

- **Extensão para stablecoins algorítmicas e redes DeFi:** Propõe-se estender a análise para stablecoins algorítmicas (e.g., DAI) e redes DeFi (e.g., Ethereum), onde mecanismos de governança podem introduzir fontes adicionais de volatilidade.

- **Assimetrias informacionais:** Investigar a assimetria informacional entre *whales* e investidores *retail*, sobretudo em eventos de grande transferência de USDT, poderia revelar padrões de *front-running* ou picos de volatilidade abrupta.
- **Modelos HAR híbridos e não lineares:** O desenvolvimento de modelos HAR híbridos, enriquecidos com componentes não lineares (como aqueles derivados de SHAP values), pode unir a interpretabilidade dos métodos tradicionais à flexibilidade dos modelos de *machine learning*.
- **Incorporação de variáveis macroeconômicas:** Uma vez que todo USDT é convertido do USD, faz-se pertinente incluir indicadores relativos à base monetária $M2$ e a condições macroeconômicas (inflação, juros etc.) de grandes economias, pois tais variáveis podem oferecer poder preditivo adicional.
- **Volatilidade implícita e mercado de opções:** Além das variáveis históricas, sugere-se incluir volatilidade implícita e curvas de volatilidade retiradas do mercado de opções, *open interest* e volume negociado de derivativos. Esses fatores podem aprimorar a previsão da Volatilidade Realizada Futura.
- **Indicadores de Sentimento:** O uso de indicadores de medo/ganância baseados em menções em redes sociais (ex.: X/Twitter) pode capturar sinais de euforia ou pânico, antecipando mudanças bruscas de volatilidade.
- **Novas arquiteturas de redes neurais:** Modelos como LSTM e *Transformers* — ou *ensembles* que combinem diversos preditores — podem melhorar a capacidade de adaptação a diferentes regimes de volatilidade.
- **Métricas de impacto econômico:** Por fim, medir o desempenho do modelo em termos de métricas econômicas diretas (como *Variance Risk Premium*) pode elucidar eventuais limitações do modelo atual em cenários de gestão de risco.

Considerações Finais

Ao integrar aprendizado de máquina adaptativo, seleção rigorosa de variáveis e dados on-chain de stablecoins, este trabalho estabelece um paradigma metodológico para modelagem de risco em ativos descentralizados. Mais do que validar técnicas computacionais, os resultados iluminam a complexa rede de interdependências em cripto mercados, nos quais stablecoins transcendem seu papel monetário para se tornarem sensores de risco em tempo real. Assim, abre-se caminho para soluções cada vez mais robustas de gerenciamento de risco, impulsionando a evolução do ecossistema cripto e contribuindo para o amadurecimento dos mercados descentralizados.

Bibliografia

Torben G. Andersen and Tim Bollerslev. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4):885–905, 1998. ISSN 00206598, 14682354. URL http://www.jstor.org/stable/2527343.

Torben G. Andersen, Tim Bollerslev, Francis X. Diebold, and Paul Labys. The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, 96(453): 42–55, 2001. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2670339.

Rafael R. Branco, Alexandre Rubesam, and Mauricio Zevallos. Forecasting realized volatility: Does anything beat linear models? *Journal of Empirical Finance*, 78:101524, 2024. ISSN 0927-5398. doi: https://doi.org/10.1016/j.jempfin.2024.101524. URL https://www.sciencedirect.com/science/article/pii/S0927539824000598.

Alexander Brauneis and Mehmet Sahiner. Crypto volatility forecasting: Mounting a HAR, sentiment, and machine learning horserace. *Asia-Pac. Financ. Mark.*, December 2024.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

Yeguang Chi, Qionghua, Chu, and Wenyan Hao. Return-forecasting and volatility-forecasting power of on-chain activities in the cryptocurrency market, 2024. URL https://arxiv.org/abs/2411.06327.

Fulvio Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009.

Marcos Lopez de Prado. *Advances in Financial Machine Learning*. Wiley Publishing, 1st edition, 2018. ISBN 1119482089.

Robert F. Engle and Andrew J. Patton. 2 - what good is a volatility model?*. In John Knight and Stephen Satchell, editors, *Forecasting Volatility in the Financial Markets (Third Edition)*, Quantitative Finance, pages 47–63. Butterworth-Heinemann, Oxford, third edition edition, 2007. ISBN 978-0-7506-6942-9. doi: https://doi.org/10.1016/B978-075066942-9.50004-2. URL https://www.sciencedirect.com/science/article/pii/B9780750669429500042.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3(null):1157–1182, March 2003. ISSN 1532-4435.

Farman Ullah Khan, Faridoon Khan, and Parvez Ahmed Shaikh. Forecasting returns volatility of cryptocurrency by applying various deep learning algorithms. *Futur. Bus. J.*, 9(1), June 2023.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6), June 2004. ISSN 1550-2376. doi: 10.1103/physreve.69.066138. URL http://dx.doi.org/10.1103/PhysRevE.69.066138.

Sophia Zhengzi Li and Yushan Tang. Automated volatility forecasting. *Management Science*, 2024. URL https://api.semanticscholar.org/CorpusID:273709933.

Lily Y. Liu, Andrew J. Patton, and Kevin Sheppard. Does anything beat 5-minute rv? a comparison of realized measures across multiple asset classes. *Journal of Econometrics*, 187 (1):293–311, 2015. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2015.02.008. URL https://www.sciencedirect.com/science/article/pii/S0304407615000329.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022. URL https://christophm.github.io/interpretable-ml-book.

Didrik Nielsen. Tree boosting with xgboost - why does xgboost win "every"machine learning competition? 2016. URL https://api.semanticscholar.org/CorpusID:114191144.

Erwan Scornet. Trees, forests, and impurity-based variable importance, 2021. URL https://arxiv.org/abs/2001.04295.

Lloyd S Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton, 1953.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests, 2008. URL http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-2821-0.

Paul Wilmott. *Frequently asked questions in quantitative finance : including key models, important formulæ, popular contracts, essays and opinions, a history of quantitative finance, sundry lists, the commonest mistakes in quant finance, brainteasers, plenty of straight-talking, the Modellers' Manifesto and lots more*. John Wiley & Sons, Chichester, England, 2 edition, September 2009.