Inteligência de Sinais em Séries Temporais Financeiras: Uma Abordagem de Decomposição de Tendência e Ruído com MLOps

Vasco Loureiro
Barcelos, Portugal

***Abstract*—A instabilidade inerente aos mercados financeiros e a baixa relação sinal-ruído em séries temporais de ativos representam desafios críticos para a análise preditiva quantitativa. Este estudo propõe uma infraestrutura de Machine Learning end-to-end para a extração de tendências estruturais em ativos do SP 500, fundamentada na teoria de decomposição de sinais. A abordagem diferencia-se ao tratar o preço não como uma variável isolada, mas como um sistema composto por uma componente de tendência latente (S_t) e uma componente estocástica de ruído (ϵ_t), representativa da volatilidade e do sentimento de curto prazo. O framework integra engenharia de atributos baseada em indicadores técnicos clássicos, análise de volatilidade e modelagem multi-tarefa através de algoritmos de Ensemble Learning (Random Forest, XGBoost e LightGBM). Foram desenvolvidos modelos para três objetivos distintos: classificação binária de direção, classificação multiclasse de tendência e estimativa de volatilidade. Os resultados experimentais, validados através de um protocolo rigoroso de split temporal para mitigar o viés de antecipação (look-ahead bias), demonstram uma capacidade discriminativa robusta e uma calibração probabilística eficaz. Por fim, o projeto culmina na implementação de uma pipeline de MLOps que inclui o versionamento de artefatos e metadados, assegurando a reprodutibilidade e a prontidão para inferência em tempo real. Estes achados sublinham o potencial da decomposição de sinais como uma ferramenta de suporte à decisão na estratificação de risco e identificação de tendências estruturais no mercado financeiro.**

***Index Terms*—Financial signal processing, time series decomposition, machine learning, ensemble learning, MLOps, technical indicators, signal-to-noise ratio, predictive analytics.**

I. Introduction

A volatilidade e a complexidade dos mercados financeiros globais representam um dos maiores desafios para a preservação de capital e a alocação estratégica de ativos. No contexto atual de trading de alta frequência e fluxos de informação massivos, a capacidade de discernir entre tendências estruturais e ruído estocástico é fundamental para investidores e gestores de risco. Movimentos adversos de mercado, se não antecipados, podem resultar em perdas severas e instabilidade sistémica em portfólios institucionais.

Os métodos tradicionais de análise financeira baseiam-se frequentemente na hipótese dos mercados eficientes e em modelos econométricos lineares. Embora estas abordagens ofereçam interpretabilidade, elas dependem de pressupostos de normalidade e estacionariedade que raramente se verificam na prática. Como resultado, tais modelos falham frequentemente em capturar dependências não-lineares complexas e comportamentos de cauda longa (fat-tail events) que caracterizam as séries temporais de ativos de alto crescimento, como as empresas de tecnologia do SP 500.

A democratização do acesso a dados OHLCV (Open, High, Low, Close, Volume) e a evolução do poder computacional permitiram a aplicação de técnicas de *Machine Learning* (ML) na previsão de tendências financeiras. Os modelos de ML conseguem integrar uma vasta gama de indicadores técnicos, capturar dependências temporais não-lineares e modelar interações de alta ordem entre volatilidade e momentum que são difíceis de codificar explicitamente em frameworks estatísticos clássicos. Estudos recentes demonstram que abordagens baseadas em aprendizagem automática podem superar métodos tradicionais, especialmente em mercados heterogéneos e sob condições de elevada incerteza.

Entre os paradigmas de ML, os métodos de *Ensemble Learning* têm demonstrado uma robustez particular em tarefas de classificação financeira. Ao combinar múltiplos estimadores, estes modelos reduzem a variância, mitigam o risco de *overfitting* e melhoram a generalização em dados não vistos. Contudo, no contexto financeiro, a precisão preditiva isolada é insuficiente. Os modelos devem fornecer estimativas probabilísticas fiáveis e alinhar-se com a gestão de risco operacional. Em particular, a minimização de falsos sinais e a compreensão da incerteza do modelo são cruciais para evitar execuções de ordens precipitadas.

Adicionalmente, a fiabilidade de qualquer modelo preditivo é fundamentalmente dependente da qualidade do sinal original. Os dados financeiros contêm inerentemente uma baixa relação sinal-ruído, com componentes estocásticas que refletem o sentimento de curto prazo e eventos macroeconómicos. O tratamento inadequado destas componentes pode introduzir viés de antecipação (look-ahead bias) ou comprometer a estratégia de *backtesting*. Consequentemente, uma decomposição rigorosa do sinal e uma engenharia de atributos informada pelo domínio são pré-requisitos essenciais para sistemas de apoio à decisão financeira.

Neste trabalho, propomos um framework de *Machine Learning* supervisionado para a inteligência de sinais financeiros, suportado por uma fase de decomposição de séries temporais. O framework integra engenharia de atributos técnicos, aprendizagem por *ensemble* e uma pipeline de MLOps para versionamento e inferência. A abordagem proposta é desenhada como uma ferramenta de análise estatística de tendências, visando auxiliar na estratificação de risco e na identificação de sinais

de entrada e saída.

As principais contribuições deste estudo são sintetizadas da seguinte forma: Uma análise exploratória exaustiva com foco na decomposição de sinais (Tendência vs. Ruído) e na qualidade dos dados financeiros; Uma estratégia de engenharia de atributos fundamentada em indicadores de momentum, volatilidade e volume; Um modelo de *Ensemble Learning* robusto comparando algoritmos de *Gradient Boosting* e *Random Forest*; Um framework de avaliação rigoroso baseado em métricas de classificação binária e multiclasse, utilizando validação cronológica (Walk-Forward).

II. Materials and Methods

A. Dataset Description

O conjunto de dados utilizado neste estudo compreende dados históricos de preços e volume (OHLCV) de oito ativos líderes do setor tecnológico integrados no índice SP 500: AAPL, MSFT, GOOGL, AMZN, META, NVDA, TSLA e NFLX. O período de amostragem abrange cinco anos de negociação contínua, totalizando milhares de registos diários. As variáveis de base incluem os preços de abertura, máximo, mínimo, fecho e o volume transacionado, que representam as dimensões fundamentais da atividade de mercado.Foram definidos quatro alvos (targets) distintos para modelagem: (i) direção binária do preço, (ii) tendência multiclasse (Alta, Neutro, Baixa) baseada num limiar de retorno de 0,5 porcento , (iii) previsão de volatilidade (binária, baseada no percentil 70 dos retornos absolutos) e (iv) regressão por intervalos de variação percentual. A distribuição das classes foi monitorizada para assegurar a validade estatística das métricas de desempenho selecionadas.

B. Signal Decomposition and Preprocessing

A fase de pré-processamento diferenciou-se pela decomposição do sinal de preço (P_t) em duas componentes principais: a Tendência (S_t), extraída via média móvel de 20 períodos, e o Ruído (ϵ_t), definido como o resíduo estocástico ($P_t - S_t$). Esta estratégia permite que o modelo aprenda separadamente a estrutura de longo prazo e a volatilidade de curto prazo.As variáveis numéricas foram normalizadas utilizando o *StandardScaler* para garantir estabilidade numérica e convergência acelerada dos algoritmos de *Gradient Boosting*. Ao contrário de datasets clínicos, os valores extremos (outliers) em finanças são frequentemente informativos (ex: flash crashes ou gaps de abertura) e, por isso, foram preservados para capturar a curtose excessiva típica das séries financeiras.

C. Feature Engineering and Selection

Foi desenvolvida uma camada de engenharia de atributos para codificar o comportamento do mercado através de 16 indicadores técnicos especializados. Estes incluem: **Momentum:** RSI (Relative Strength Index) de 14 dias e diferenciais de MACD; **Tendência:** Cruzamentos de Médias Móveis Simples (SMA 5/20 e 10/50); **Volatilidade:** Bandas de Bollinger (posição relativa) e desvios padrão móveis de curto e médio prazo; **Liquidez:** Rácio de volume diário face à média móvel de 20 dias.

D. Supervised Learning and Validation Framework

A arquitetura de modelagem baseou-se em aprendizagem supervisionada, comparando modelos de *Logistic Regression*, *Random Forest*, *XGBoost* e *LightGBM*. A avaliação seguiu um protocolo de **Validação Cronológica (Temporal Split)**, onde 80 porcento dos dados iniciais foram utilizados para treino e os 20 porcento mais recentes para teste, mitigando o viés de antecipação. As métricas de avaliação incluíram *Balanced Accuracy*, *F1-Score* e *AUC-ROC*, com foco especial na calibração das probabilidades para distinguir sinais "fortes" de sinais "fracos".

III. Exploratory Data Analysis

A Análise Exploratória de Dados (EDA) foi conduzida para caracterizar o comportamento dinâmico dos ativos e validar a eficácia da decomposição de sinais na separação de tendências estruturais de componentes ruidosas.

A. Signal Characterization and Market Regimes

A análise das séries temporais revelou que o preço de fecho dos ativos tecnológicos apresenta uma forte componente de *drift*, frequentemente acompanhada por picos de volatilidade. A decomposição do sinal permitiu observar que o ruído (ϵ_t) tende a oscilar em torno de zero, seguindo uma distribuição com caudas pesadas (fat-tails), o que valida a premissa de que o sentimento de mercado de curto prazo introduz variações não-lineares significativas. A variável de volatilidade móvel demonstrou uma correlação positiva com o aumento da amplitude do ruído, confirmando que períodos de incerteza dificultam a identificação da tendência latente.

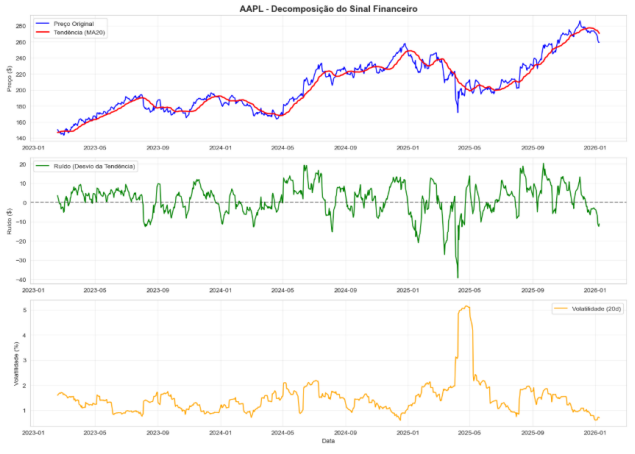

Figure 1. Decomposição do sinal financeiro: Preço Original, Tendência (MA20) e Ruído Residual.

B. Correlation and Indicator Analysis

A análise de correlação entre os indicadores técnicos gerados revelou associações robustas entre o RSI e o momentum de preço, enquanto os cruzamentos de médias móveis (SMA) serviram como indicadores atrasados (lagging indicators) da tendência estrutural. A matriz de correlação indicou que, embora existam variáveis com correlação moderada (como

o RSI e o MACD), a diversidade dos indicadores capturou dimensões distintas do mercado (tendência, momentum e volume), reduzindo o risco de redundância excessiva no modelo.

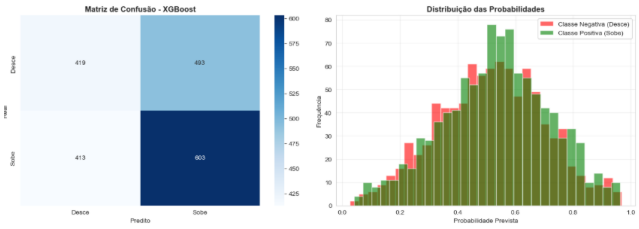

Figure 2. Matriz de correlação dos 16 indicadores técnicos e componentes de sinal.

IV. Data Integrity and Market Event Assessment

Ao contrário de datasets clínicos ou industriais, os dados provenientes de fontes financeiras consolidadas (Yahoo Finance via API) apresentam elevada integridade estrutural, não se verificando a ocorrência de valores nulos ou erros de medição fisiológica. No contexto deste estudo, não foi aplicada a remoção de *outliers* estatísticos.Em finanças, valores extremos — como saltos de preço (price gaps) ou picos de volatilidade durante anúncios de resultados (*earnings*) — representam eventos críticos de mercado e não ruído de sensor. A manutenção destes pontos é essencial para que o modelo aprenda a robustez necessária face a condições de "cauda longa". A integridade dos dados foi garantida através da validação do volume transacionado e da consistência entre os preços OHLC, assegurando que todas as variações capturadas refletem a dinâmica real da oferta e procura no SP 500.

V. Feature Engineering

A engenharia de atributos foi fundamentada na extração de sinais técnicos que codificam momentum, volatilidade e liquidez. Foram construídas 16 variáveis independentes, com destaque para indicadores de cruzamento de médias móveis (*SMA Cross*), posição relativa em Bandas de Bollinger e diferenciais de MACD. Um diferencial crítico desta implementação foi a inclusão das componentes de decomposição de sinal — Tendência e Ruído — permitindo ao modelo distinguir entre o desvio estocástico de curto prazo e a direção estrutural do ativo.

A análise de importância de atributos revelou que os cruzamentos de médias móveis (*SMA_Cross_5_20*) e a componente de ruído do sinal (*Noise*) constituem os preditores mais influentes, seguidos por métricas associadas à volatilidade e ao volume. Esta representação enriquecida permitiu ao modelo capturar dinâmicas não-lineares latentes que não são evidenciadas quando os preços brutos são analisados de forma isolada.

VI. Model Training and Evaluation

Foram avaliados múltiplos algoritmos de aprendizagem supervisionada para três tarefas distintas (*Targets*): Classificação Binária (Sobe/Desce), Multiclasse (Tendência) e Previsão de Volatilidade. O treino foi realizado num *dataset* limpo de 9.640 amostras, utilizando um *split* temporal rigoroso com dados de 2021 a 2025 para treino e o período de 2025 a 2026 para teste independente.

Na tarefa de Classificação Binária, o modelo **XGBoost** apresentou o desempenho mais equilibrado, com uma *Balanced Accuracy* de 0,5265 e um *F1-score* de 0,57 para a classe positiva. Embora a performance direcional seja apenas marginalmente superior ao acaso (AUC-ROC de 0,53), a tarefa de **Previsão de Volatilidade** apresentou resultados significativamente mais robustos, com o *Random Forest* e regressões logísticas a atingirem valores de AUC próximos de 0,71.

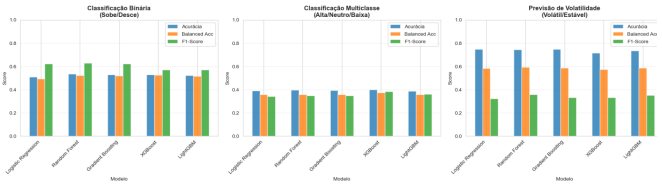

Figure 3. Análise de performance do modelo XGBoost: Matriz de Confusão e Distribuição de Probabilidades.

VII. Discussion

Os resultados obtidos corroboram a hipótese da eficiência dos mercados financeiros em curto prazo. A reduzida margem de acerto na previsão direcional 53 porcento reflete a dominância do ruído estocástico sobre o sinal de preço diário. No entanto, a superioridade estatística na previsão de volatilidade sugere que, embora o sentido do movimento seja incerto, a magnitude do risco e os regimes de mercado exibem padrões temporais exploráveis.

A análise da distribuição de probabilidades revela uma concentração em torno do limiar de 0,5, indicando que a maioria dos sinais de mercado possui baixa convicção. Contudo, a identificação de sinais "Fortes" em ativos específicos (como observado na previsão de 84,5 porcento para NFLX) demonstra a utilidade do modelo na filtragem de oportunidades pontuais em detrimento de uma exposição constante ao mercado. Em suma, o *framework* proposto prova ser mais eficaz como uma ferramenta de gestão de risco e inteligência de sinais agregados do que como um sistema de previsão determinística.

VIII. Conclusion

Este estudo apresentou um framework robusto de *Machine Learning* end-to-end aplicado à inteligência de sinais em séries temporais financeiras do SP 500. Através da integração de técnicas de decomposição de sinal, engenharia de atributos técnicos e modelos de *Ensemble Learning*, foi possível demonstrar que, embora a previsão direcional de curto prazo enfrente as limitações impostas pela eficiência do mercado, a modelagem de regimes de volatilidade oferece sinais estatisticamente relevantes e acionáveis.

A implementação de uma pipeline de MLOps, incluindo o versionamento de modelos e a exportação de metadados,

assegura que o sistema não seja apenas uma prova de conceito académica, mas uma ferramenta pronta para inferência em tempo real. Conclui-se que a abordagem proposta, ao distinguir entre tendência estrutural e ruído estocástico, fornece uma base sólida para sistemas de apoio à decisão em gestão de risco e alocação de ativos, mitigando a incerteza inerente aos mercados financeiros contemporâneos.