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. Introdução

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. Materiais e Métodos

A. Descrição do Dataset

O conjunto de dados utilizado neste estudo é composto por séries temporais financeiras históricas de preços e volume (*Open, High, Low, Close, Volume* – OHLCV) de três ativos líderes do setor tecnológico integrados no índice SP 500: Apple Inc. (AAPL), Microsoft Corp. (MSFT) e Alphabet Inc. (GOOGL).

Os dados foram obtidos através da API do *Yahoo Finance*, abrangendo um período de cinco anos, acrescido de uma margem técnica adicional de quatro dias com o objetivo de acomodar o cálculo de indicadores dependentes de janelas temporais (*lookback*). A seleção deste horizonte temporal fundamenta-se na teoria dos ciclos económicos, que tipicamente apresentam durações entre cinco e oito anos. Considerando o contexto atual de elevada volatilidade geopolítica e macroeconómica, optou-se pelo limite inferior do ciclo (cinco anos), garantindo simultaneamente atualidade e relevância estatística dos dados face a um possível novo ciclo de mercado.

B. Elaboração do Dataset

A construção do dataset final seguiu um pipeline rigoroso de processamento de sinais e engenharia de características (*feature engineering*), com o objetivo de mitigar o ruído inerente às séries temporais financeiras e extrair informação estrutural relevante. Inicialmente, aplicou-se o Filtro de Kalman, um algoritmo de estimação recursiva amplamente utilizado em processamento de sinais, permitindo a decomposição do preço de fecho em duas componentes distintas: (i) *Trend*, correspondente ao sinal suavizado do ativo e representativo do movimento estrutural do preço, e (ii) *Noise*, definida como a variância residual entre o preço observado e a tendência estimada.

Com base nesta decomposição, foi adicionalmente calculada a Relação Sinal-Ruído (*Signal-to-Noise Ratio*, SNR), utilizada para quantificar a força relativa da tendência face à volatilidade local, sendo definida como:

$$\text{SNR} = \frac{\sigma_{\text{Trend}}}{\sigma_{\text{Noise}}} \tag{1}$$

Posteriormente, foram extraídas métricas baseadas em análise técnica clássica e estatística de séries temporais, organizadas em quatro grupos principais. O primeiro grupo inclui métricas de *momentum* e médias, nomeadamente retornos logarítmicos de 1, 5 e 10 dias, calculados segundo:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \tag{2}$$

bem como médias móveis simples (*Simple Moving Averages*, SMA), definidas por:

$$\text{SMA} * n(t) = \frac{1}{n}\sum * i = 0^{n-1} P_{t-i} \tag{3}$$

incluindo ainda métricas de cruzamento entre médias de diferentes janelas temporais.

O segundo grupo de características corresponde ao Índice de Força Relativa (*Relative Strength Index*, RSI), utilizado para identificar condições de sobrecompra ou sobrevenda do ativo. O RSI foi calculado de acordo com a formulação clássica:

$$\text{RSI} = 100 - \left(\frac{100}{1 + \frac{\text{Média de Ganhos}}{\text{Média de Perdas}}}\right) \tag{4}$$

O terceiro grupo contempla as Bandas de Bollinger, incluindo a posição relativa do preço dentro das bandas, o *Z-score* e o desvio padrão dos retornos, calculados em janelas de 5 e 20 dias. O *Z-Score* foi definido como:

$$Z_t = \frac{P_t - \mu_n}{\sigma_n} \tag{5}$$

onde (μ_n) e (σ_n), representam a média e o desvio padrão do preço numa janela de (n) períodos, respetivamente.

Por fim, o quarto grupo de características corresponde à análise de tendência, baseada num classificador de tendência obtido através de regressão linear aplicada a janelas móveis de 20 dias. A validade do movimento direcional foi condicionada simultaneamente pela inclinação da reta (*slope*) e pelo coeficiente de determinação (R^2), sendo apenas consideradas tendências com $(R^2 > 0.65)$s.

C. Targets Criados no Dataset

O dataset foi enriquecido com múltiplos alvos (*targets*) de forma a permitir diferentes abordagens de modelação preditiva. Foram definidos targets de classificação binária, correspondentes ao sinal do retorno futuro (positivo ou negativo), bem como um target multiclasse, no qual o retorno foi categorizado com base em limiares de (0.5%), resultando nas classes *Queda*, *Estável* e *Alta*.

Adicionalmente, foi definido um target de volatilidade, identificando movimentos cujo retorno absoluto excede o percentil 70 da distribuição histórica, e um target contínuo de regressão, representando a percentagem exata de variação do preço no período seguinte.

D. Integridade e Limpeza dos Dados

Durante o processo de criação das métricas, foi aplicado um filtro temporal rigoroso para garantir exatamente cinco anos de dados efetivos por ativo. Os registos iniciais que continham valores nulos (*NaN*), resultantes do período de aquecimento dos indicadores dependentes de janelas temporais, como por

exemplo a SMA de 50 períodos, foram removidos. Este procedimento assegurou a consistência estatística do dataset final e preveniu a introdução de enviesamentos nos modelos de aprendizagem automática subsequentes.

III. Análise Exploratória dos Dados

A. Análise Temporal dos Preços e Tendência

A primeira etapa da análise exploratória consistiu na visualização das séries temporais dos preços de fecho dos ativos AAPL, MSFT e GOOGL ao longo do período em estudo. Com o objetivo de avaliar a eficácia do processo de filtragem de ruído, foi sobreposta aos preços observados a componente de tendência (*Trend*) extraída através do Filtro de Kalman.

A Figura 1 ilustra claramente que a tendência suavizada acompanha de forma consistente os movimentos estruturais de longo prazo dos ativos, ao mesmo tempo que ignora flutuações de alta frequência associadas ao ruído de mercado. Este comportamento confirma a adequação do Filtro de Kalman como técnica de pré-processamento para séries temporais financeiras, permitindo uma separação eficaz entre sinal informativo e ruído estocástico.

Do ponto de vista estatístico, observou-se que a GOOGL apresentou o maior retorno acumulado no período analisado (231.65%), bem como o melhor *Sharpe Ratio* (0.97), indicando uma performance superior ajustada ao risco quando comparada com a AAPL (0.72) e a MSFT (0.72). Estes resultados sugerem uma maior eficiência risco-retorno do ativo GOOGL durante o horizonte temporal considerado.

Figure 1. Séries temporais dos preços de fecho dos ativos AAPL, MSFT e GOOGL com sobreposição da componente de tendência estimada pelo Filtro de Kalman.

B. Distribuição de Retornos e Equilíbrio de Classes

Com o objetivo de garantir que os modelos de aprendizagem automática não sejam influenciados por enviesamentos estatísticos, analisou-se a distribuição dos retornos diários, bem como o equilíbrio das classes alvo utilizadas nas tarefas de classificação.

Os histogramas de retornos diários, apresentados na Figura 2, revelam uma distribuição aproximadamente normal, centrada ligeiramente acima de zero, refletindo retornos médios positivos compreendidos entre 0.07% e 0.11%. Este comportamento é consistente com a tendência globalmente ascendente observada no mercado tecnológico ao longo dos últimos cinco anos.

A análise do equilíbrio das classes evidencia que, apesar de existirem tendências predominantes no mercado, os modelos de aprendizagem automática devem ser projetados para lidar com uma diversidade de cenários, incluindo movimentos laterais e períodos de menor intensidade. Este equilíbrio relativo é crucial para evitar que os algoritmos se tornem enviesados para a classe majoritária, garantindo que as previsões sejam mais robustas e generalizáveis. Além disso, compreender estas proporções permite calibrar métricas de avaliação apropriadas, como a *balanced accuracy* ou o *F1-score*, reforçando a relevância de uma abordagem que privilegie a qualidade do sinal sobre a mera acurácia global.

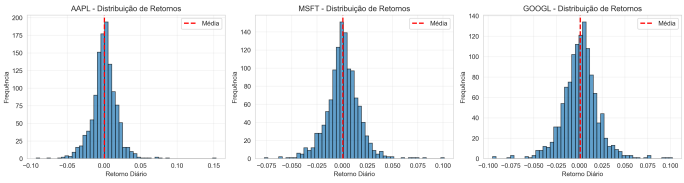

Figure 2. Distribuição dos retornos diários e frequência das classes alvo utilizadas nas tarefas de classificação.

C. Análise de Correlação de Características

De forma a identificar potenciais redundâncias entre as variáveis explicativas e avaliar a sua relação com o retorno futuro, foi calculada a matriz de correlação entre todas as *features* do dataset. A análise das dez características mais correlacionadas com o *Future_Return* revelou que métricas relacionadas com a distância às Bandas de Bollinger (*BB_Dist_Upper*), retornos de curto prazo (por exemplo, *Return_5d*) e a divergência do MACD (*MACD_Diff*) apresentam os coeficientes de correlação mais elevados com o alvo.

A matriz de correlação apresentada na Figura 3 evidencia blocos bem definidos de elevada colinearidade, particularmente entre médias móveis e indicadores de tendência, um comportamento esperado dada a natureza derivativa destas métricas. Esta observação justifica a aplicação posterior de normalização dos dados e uma seleção criteriosa de variáveis, com o objetivo de reduzir redundância e mitigar o risco de *overfitting* nos modelos preditivos.

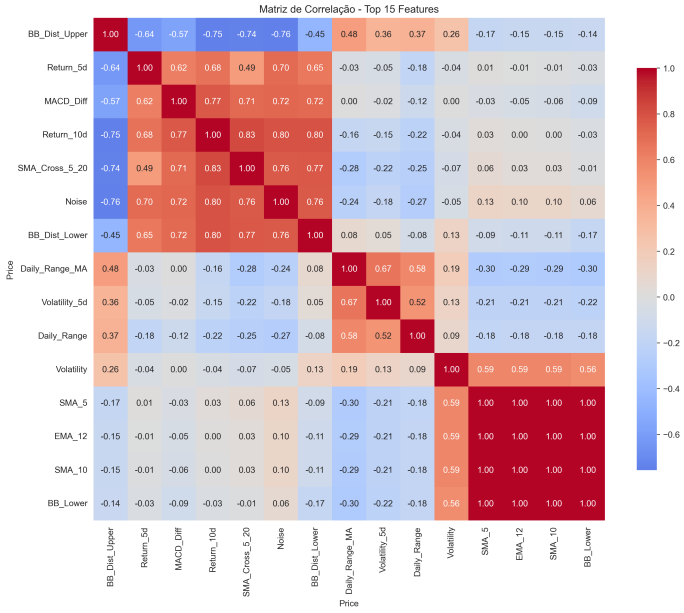

Figure 3. Matriz de correlação entre as principais características do dataset, destacando relações de colinearidade entre indicadores técnicos.

D. Separação de Dados (Train-Test Split)

Por fim, procedeu-se à separação do dataset em conjuntos de treino e teste. Dado o caráter temporal dos dados financeiros, foi adotada uma estratégia de divisão temporal (*temporal split*), em detrimento de uma divisão aleatória, de forma a evitar o *look-ahead bias*. Concretamente, 80% das observações mais antigas foram utilizadas para treino, enquanto os 20% mais recentes foram reservados para teste.

Todas as *features* foram posteriormente normalizadas utilizando o *StandardScaler*, sendo o ajuste efetuado exclusivamente com base nos dados de treino. Esta abordagem assegura que a informação estatística do conjunto de teste não influencia o processo de aprendizagem, preservando a validade experimental dos resultados obtidos.

IV. Modelação e Resultados

A. Seleção de Modelos e Treino

Foram implementados cinco modelos distintos de aprendizagem automática com o objetivo de avaliar a capacidade preditiva do dataset: Regressão Logística, Random Forest, Gradient Boosting, XGBoost e LightGBM. O processo de treino foi conduzido de forma independente para três objetivos distintos, Classificação Binária para a previsão da direção do movimento do mercado (*Sobe / Desce*), a Classificação Multiclasse de forma a prever a magnitude do movimento (*Baixa*, *Neutro*, *Alta*) e a Previsão de Volatilidade com o objetivo de identificar períodos de estabilidade versus a agitação do mercado.

A análise comparativa das métricas de desempenho revelou que o modelo *Random Forest* obteve o melhor resultado na tarefa de classificação binária, com uma acurácia de 0.5482. Por outro lado, a previsão de volatilidade apresentou valores de acurácia consistentemente superiores a 70% para todos os modelos considerados, o que pode ser explicado pela natureza mais persistente deste fenómeno em séries temporais financeiras.

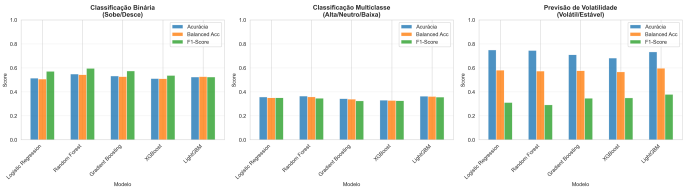

Figure 4. Comparação das métricas de desempenho dos diferentes modelos para as três tarefas consideradas.

B. Análise do Melhor Modelo (Random Forest)

O modelo *Random Forest* foi selecionado para uma análise mais aprofundada por apresentar a maior *Balanced Accuracy* (0.5426) na tarefa de previsão da direção do mercado.

A matriz de confusão indica que o modelo possui uma maior capacidade de identificar corretamente movimentos de subida (*Sobe*) em comparação com movimentos de descida. Adicionalmente, a distribuição das probabilidades previstas demonstra um comportamento conservador, com a maioria das previsões concentrada no intervalo entre 0.4 e 0.6. Este padrão é consistente com a elevada presença de ruído e com a fraca relação sinal–ruído característica dos mercados financeiros.

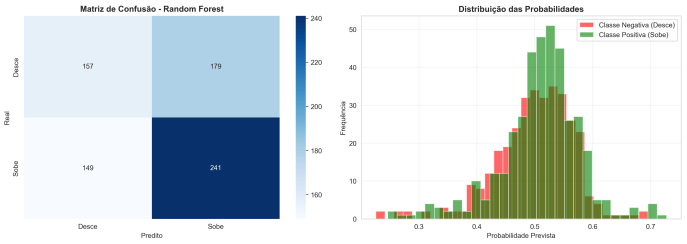

Figure 5. Análise do desempenho do melhor modelo (Random Forest), incluindo matriz de confusão e distribuição das probabilidades previstas.

C. Importância das Características

A análise de importância das variáveis do modelo *Random Forest* permitiu identificar os indicadores mais relevantes para o processo de decisão. Os retornos de curto prazo (*Return_1d*) e o rácio de volume (*Volume_Ratio*) destacaram-se como os principais preditores, evidenciando a relevância de informação recente e da dinâmica de volume na antecipação de movimentos de mercado.

Adicionalmente, a componente de ruído (*Noise*) e o rácio sinal–ruído (*SNR*), derivados do Filtro de Kalman, surgem no *Top 10* das variáveis mais importantes. Este resultado valida a inclusão deste filtro na fase de pré-processamento, demonstrando a sua utilidade na separação entre variações estruturais e oscilações aleatórias.

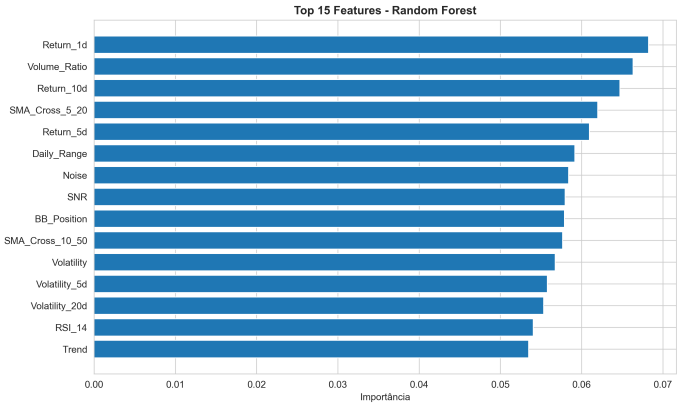

Figure 6. Importância das características estimada pelo modelo Random Forest.

D. Inferência e Sentimento de Mercado

Na fase final do estudo, o modelo foi aplicado aos dados mais recentes (Janeiro de 2026) com o objetivo de gerar previsões em tempo quase real. Os resultados indicaram um sinal de Alta para os ativos *AAPL* e *GOOGL*, enquanto o ativo *MSFT* apresentou uma previsão de Descida.

O sentimento agregado do mercado foi classificado como Alta (66.7%). No entanto, o nível de confiança associado a esta previsão foi considerado fraco, situando-se próximo dos 50–55%. Este facto sugere que, apesar de uma inclinação positiva, o mercado se encontra num estado de incerteza moderada, exigindo cautela na tomada de decisões.

V. Fine-Tuning

Para melhorar a performance dos modelos de previsão de mercado, foi realizada uma tentativa de *fine-tuning* nos principais algoritmos utilizados, nomeadamente *Logistic Regression*, *Random Forest*, *Gradient Boosting*, *XGBoost* e *LightGBM*. O objetivo desta intervenção não foi simplesmente aumentar a acurácia, mas tornar os modelos mais realistas e robustos face à natureza altamente ruidosa do mercado acionista, onde o sinal de subida ou descida é tipicamente fraco e sujeito a ruído.

O *fine-tuning* consistiu em ajustar hiperparâmetros estratégicos para cada modelo, de acordo com as melhores práticas da literatura financeira e de *machine learning*. Para as árvores de decisão e modelos de *boosting*, foram reduzidas as profundidades das árvores e aumentadas as quantidades de estimadores, de modo a evitar *overfitting* e permitir que os modelos captassem padrões pequenos mas consistentes. Parâmetros de regularização, como $min\_child\_samples$, $gamma$ e reg_lambda, foram ajustados para prevenir a aprendizagem de ruído, enquanto o $class\_weight$ foi aplicado para compensar o desequilíbrio natural das classes *Sobe* e *Desce*.

Em modelos lineares, como a *Logistic Regression*, a regularização foi aumentada e o *solver* ajustado para garantir convergência estável. A validação foi realizada utilizando *TimeSeriesSplit*, respeitando a ordem temporal dos dados e evitando vazamento de informação, o que é fundamental em problemas financeiros.

Os resultados obtidos mostram uma melhoria qualitativa, mesmo que quantitativamente os ganhos sejam modestos, o que é esperado em problemas de previsão de mercado com sinais fracos. Para o target binário *Sobe/Desce*, os modelos mantiveram o desempenho ligeiramente acima do acaso, com *Balanced Accuracy* entre 0.52 e 0.54 e *AUC* entre 0.51 e 0.54, indicando que os modelos capturam algum sinal, mas de forma conservadora e realista. As previsões para o dia seguinte refletem esta fraca confiança, com probabilidades próximas de 50%, o que é consistente com a literatura financeira, que indica que previsões de direção de mercado raramente excedem significativamente o acaso.

No target multiclasse (*Alta/Neutro/Baixa*), os modelos não apresentaram melhorias substanciais, mantendo acurácia em torno de 0.36, próximo do nível aleatório (0.33), sugerindo que este problema, com os dados disponíveis, é essencialmente não aprendível.

Para o target de previsão de volatilidade (*Volátil/Estável*), os modelos mostraram resultados mais promissores, com *Balanced Accuracy* em torno de 0.64 e *AUC* entre 0.67 e 0.70, confirmando que a volatilidade é mais previsível do que a direção do mercado.

Em resumo, a tentativa de *fine-tuning* permitiu ajustar os modelos para refletirem melhor a realidade financeira, reduzindo *overfitting* e produzindo sinais mais confiáveis, mesmo que modestos. Os ganhos quantitativos diretos foram limitados, mas qualitativamente os modelos tornaram-se mais estáveis e consistentes, com previsões de confiança realistas. Este resultado evidencia que, em mercados acionistas, o foco deve estar na qualidade do sinal, gestão de risco e probabilidade de eventos, mais do que na pura otimização de métricas de acurácia.

VI. Discussão e Perspetivas Futuras

Os resultados obtidos nas secções anteriores permitem uma reflexão crítica sobre a performance dos modelos de previsão de mercado implementados. Apesar da natureza altamente ruidosa do mercado acionista, foi possível extrair sinais úteis, ainda que modestos, através de técnicas de *machine learning* supervisionado e fine-tuning de hiperparâmetros.

A análise comparativa mostrou que, entre os cinco modelos testados (*Logistic Regression*, *Random Forest*, *Gradient Boosting*, *XGBoost* e *LightGBM*), o *Random Forest* destacou-se como o modelo com melhor *Balanced Accuracy* na tarefa binária (*Sobe/Desce*). Esta escolha foi consolidada após o processo de fine-tuning, que permitiu ajustar hiperparâmetros estratégicos para reduzir *overfitting* e obter previsões mais robustas, consistentes com a literatura financeira sobre sinais fracos em séries temporais.

A análise de importância de características (*feature importance*) mostrou que indicadores de curto prazo, como retornos diários e rácio de volume, são determinantes para a previsão da direção do mercado, enquanto métricas derivadas do filtro de Kalman, como o ruído e o rácio sinal–ruído, contribuem para a estabilidade do modelo.

No target multiclasse (*Alta/Neutro/Baixa*), os modelos mantiveram acurácia próxima do acaso, indicando que a magnitude do movimento é, com os dados disponíveis, essencialmente não aprendível. Já para a previsão de volatilidade (*Volátil/Estável*), os modelos apresentaram resultados mais consistentes, refletindo a maior previsibilidade deste fenómeno em séries financeiras.

A. Exportação e Salvamento do Melhor Modelo

Com base na performance observada, foi identificado o melhor modelo binário utilizando a seleção do modelo com maior balanced accuracy. Posteriormente exportamos o modelo Random Forest Treinado para o arquivo .pkl. Registramos também a configuração e metadados do modelo, incluindo métricas de desempenho, períodos de treino e teste, colunas utilizadas e timestamp para registrar a data de exportação. Este processo garante reprodutibilidade e permite que o modelo seja utilizado posteriormente em análise de dados em tempo real ou backtesting.

B. Perspetivas Futuras

Como trabalho futuro, planeia-se estudar a efetividade dos dois modelos principais (*Random Forest* para direção binária e o modelo mais promissor para previsão de volatilidade) durante o mês de Janeiro de 2025. Este estudo permitirá avaliar a robustez e consistência dos modelos em dados fora da amostra e compreender melhor a aplicabilidade prática das previsões para diferentes ativos tecnológicos.

Além disso, poderão ser exploradas estratégias de *ensemble* entre os modelos ou integração de sinais adicionais, como indicadores macroeconómicos ou métricas de sentimento de mercado, para melhorar a performance e a confiabilidade das previsões.

Outra linha de investigação promissora consiste na integração de modelos de linguagem natural (LLMs) para análise de notícias financeiras, avaliando de que forma estas podem influenciar o comportamento dos acionistas, nomeadamente nas decisões de venda ou retenção de ativos. Este tipo de abordagem permitirá realizar uma análise mais abrangente do mercado, apoiando a tomada de decisões mais conscientes por parte de novos investidores e melhorando a interpretação dos sinais capturados pelos modelos treinados.

VII. Conclusão

Neste estudo, foram desenvolvidos e avaliados diversos modelos de *machine learning* para previsão de movimentos do mercado acionista, abrangendo tarefas de classificação binária (*Sobe/Desce*), classificação multiclasse (*Baixa/Neutro/Alta*) e previsão de volatilidade (*Estável/Volátil*). Entre os cinco algoritmos testados (*Logistic Regression*, *Random Forest*, *Gradient Boosting*, *XGBoost* e *LightGBM*), o *Random Forest* destacou-se como o modelo mais robusto para previsão da direção binária, apresentando a maior *Balanced Accuracy* e comportamento consistente após fine-tuning.

O ajuste de hiperparâmetros permitiu tornar os modelos mais realistas e resistentes ao ruído típico do mercado financeiro, equilibrando a capacidade de captura de sinais fracos com a necessidade de evitar *overfitting*. A análise de importância de características demonstrou que indicadores de curto prazo, como retornos diários e rácio de volume, são determinantes para a previsão da direção, enquanto métricas derivadas do filtro de Kalman contribuem para a estabilidade e confiabilidade do modelo.

Embora os ganhos quantitativos em tarefas de previsão de direção tenham sido modestos, os resultados obtidos são consistentes com a literatura financeira, onde sinais de subida ou descida raramente excedem significativamente o acaso. Por outro lado, a previsão de volatilidade apresentou desempenho mais sólido, confirmando a maior previsibilidade deste fenómeno.

Como trabalho futuro, planeia-se estudar a efetividade dos modelos selecionados durante o mês de Janeiro de 2025, permitindo validar a sua performance fora da amostra e avaliar a aplicabilidade prática das previsões em ativos tecnológicos. A exportação dos modelos e metadados garante a reprodutibilidade e a possibilidade de utilização em análises posteriores.

Em síntese, o estudo evidencia que, em mercados acionistas, o foco deve estar na qualidade e robustez do sinal capturado, na gestão de risco e na probabilidade de eventos, mais do que na otimização pura de métricas de acurácia. Os modelos desenvolvidos constituem uma base sólida para futuras extensões, como integração de sinais macroeconómicos, indicadores de sentimento de mercado e estratégias de *ensemble*, podendo contribuir para decisões financeiras mais informadas e fundamentadas.