Aprendizado de máquina no aprimoramento da negociação de pares

Joaquim Rafael Mariano Prieto Pereira [1] - 10408805

Henrique Arabe Neres de Farias [1] - 10410152

Gustavo Matta [1] - 10410154

Lucas Trebacchetti Eiras [1] - 10401973

[1]Faculdade de Computação e Informática (FCI)

Universidade Presbiteriana Mackenzie São Paulo, SP – Brasil

<10408805@mackenzista.com.br,10410152@mackenzie.br>

<10410154@mackenzie.br,10401973@mackenzie.br>

2025

Resumo

Nas últimas décadas, a crescente demanda por estratégias automatizadas no mercado financeiro, aliada ao aumento da volatilidade e do volume de dados, tem impulsionado a aplicação de técnicas avançadas de análise e processamento de dados. Nesse contexto, este trabalho propõe uma abordagem híbrida que combina a tradicional estratégia de negociação de pares com algoritmos modernos de aprendizado de máquina. O objetivo é aprimorar a precisão das previsões do spread *entre ativos correlacionados e otimizar a execução das operações de* trade *de alta frequência com o menor custo possível. Para tanto, será realizada uma revisão teórica das técnicas clássicas de reversão à média e dos modelos preditivos atuais, seguida pela coleta de dados históricos de alta frequência da B3. Dois modelos (florestas aleatórias, redes neurais convolucionais) serão treinados para prever divergências de preços entre pares de ações. Espera-se como resultado uma estratégia com maior robustez e eficiência, capaz de gerar retornos superiores ao aproveitar ineficiências temporárias do mercado, contribuindo tanto para o âmbito acadêmico quanto para a prática da negociação de alta frequência.*

Palavras-chave: Negociação de pares, aprendizado de máquina, previsão, reversão, dados de alta frequência, redes neurais profundas.

1 Introdução

1.1 Contextualização

No contexto do mercado financeiro, a utilização de algoritmos e técnicas de inteligência artificial tem se intensificado, impulsionada pela busca de vantagem competitiva em panoramas de alta volatilidade e volume de transações. Estratégias de negociação automatizadas, como a negociação de alta frequência, operam em escalas de tempo diminutas para explorar ineficiências momentâneas do mercado (Zaharudin *et al.*, 2022).

1.2 Justificativa

Nesse sentido, prever retornos é um desafio devido à significativa variação dos preços dos ativos, os quais muitas vezes se comportam de maneira semelhante a ruído branco (Hadad *et al.*, 2024). A estratégia de negociação de pares, que explora a reversão à média entre ações intimamente correlacionadas, surge como uma alternativa para alavancar lucros em diferentes cenários macroeconômicos. Estudos como o de Gatev *et al.* (2006) demonstram historicamente a capacidade dessa técnica de gerar retornos superiores à média do mercado. Entretanto, tratando-se de dados com granularidade de minuto a minuto, a complexidade aumenta, exigindo o desenvolvimento de algoritmos computacionais robustos capazes de tratar a alta volatilidade dos dados, minimizar os custos operacionais e, consequentemente, garantir os lucros desejados.

1.3 Objetivo

Diante desse cenário, o trabalho tem como objetivo desenvolver uma abordagem híbrida que combine a tradicional negociação de pares com algoritmos inovadores de aprendizado de máquina, visando aprimorar a previsão e a execução dos trades com o menor custo possível. Para atingir esse objetivo, propõe-se: (1) analisar os métodos clássicos de reversão de pares; (2) aplicar modelos preditivos (florestas aleatórias, redes neurais convolucionais) para antecipar o espalhamento dos pares; (3) integrar a reversão com a previsão para desenvolver um algoritmo de execução automática; e (4) testar e avaliar os resultados utilizando dados reais obtidos diretamente da B3.

1.4 Opção do projeto

A escolha deste projeto concentra-se no aprendizado de máquina para o aprimoramento da negociação de pares, motivada pela crescente demanda por estratégias automatizadas no mercado financeiro. A intensa volatilidade e o volume de dados exigem técnicas avançadas para obtenção de vantagem competitiva.

O projeto se enquadra na Opção ML/DL/VC/PLN, empregando algoritmos de Machine Learning (ML) e Deep Learning (DL) para solucionar um problema de predição de negócio. A negociação de pares, que explora a reversão à média entre ações correlacionadas, tem um histórico de gerar retornos superiores (Gatev et al., 2006). Contudo, sua aplicação em alta frequência requer algoritmos robustos para lidar com a complexidade dos dados e otimizar custos.

Assim, propomos uma abordagem híbrida, combinando negociação de pares com algoritmos de aprendizado de máquina (como florestas aleatórias e redes neurais). O objetivo é aprimorar a precisão das previsões do spread entre ativos e otimizar a execução das operações de trade de alta frequência, visando o menor custo. A integração de machine

learning em estratégias de negociação de pares pode aumentar a precisão na previsão e a robustez do sistema (Shah et al., 2022).

Este projeto busca, portanto, desenvolver uma estratégia mais robusta e eficiente, capaz de gerar retornos superiores ao aproveitar ineficiências temporárias do mercado brasileiro, unindo fundamentos estatísticos e o poder preditivo da inteligência artificial.

2 Referencial Teórico

Para Zaharudin, Young e Hsu (2022), o conceito de negociação de alta frequência está relacionado ao uso de programas de computador ou estratégias sofisticadas e de alta velocidade para gerar, encaminhar e executar ordens sobre ativos. Nesse contexto, encontra-se a negociação de pares, que se baseia no princípio de que duas ações altamente correlacionadas tendem a ter seus preços caminhando juntos; assim, quando ocorre um desvio entre eles, são realizadas operações de compra sobre a ação em queda e venda sobre a ação em crescimento, assumindo que o preço retornará à média histórica (Gatev *et al.*, 2006). Essa técnica se fundamenta em princípios estatísticos, como a cointegração e a regressão linear, e possui um largo histórico de aplicação para ganhos no mercado financeiro, principalmente em *Wall Street*. Inicialmente aplicada em dados diários, de maior granularidade, a técnica tem sido adaptada para funcionar em ambientes de alta frequência, onde a complexidade é incrementada devido à quantidade massiva de dados e à necessidade de decisões em curtíssimo período (Hadad *et al.*, 2024).

Paralelamente, o aprendizado de máquina corresponde ao ramo da Inteligência Artificial responsável por criar modelos que aprendem por meio de dados para realizar tarefas de forma autônoma (Tatsat, Puri e Lookabaugh, 2024). Assim, os avanços nos algoritmos de aprendizado de máquina permitem modelar relações não lineares e identificar padrões complexos em séries temporais. Técnicas como Florestas Aleatórias (Donick e Lera, 2021) e modelos de *deep learning*, como as Redes Neurais Convolucionais (Chen et al., 2018), são particularmente eficazes na redução do erro preditivo e na captura de dependências temporais em mercados financeiros.

Diante disso, combinando os fundamentos estatísticos da negociação de pares com a capacidade preditiva dos algoritmos de aprendizado de máquina, configura-se uma estratégia promissora para enfrentar os desafios do mercado financeiro contemporâneo no contexto da bolsa de valores brasileira.

3 Descrição do Problema

O mercado financeiro contemporâneo é caracterizado por uma intensa volatilidade e um volume massivo de transações, cenários que têm impulsionado a busca por estratégias de negociação automatizadas e, em particular, as de alta frequência. Neste ambiente dinâmico, a capacidade de prever retornos apresenta um desafio significativo, dada a variação dos preços dos ativos, que frequentemente se assemelham a um comportamento de ruído branco (Hadad et al., 2024).

A estratégia de negociação de pares, que explora a reversão à média entre ativos correlacionados, oferece um caminho para alavancar lucros. No entanto, sua aplicação em um contexto de alta frequência, que envolve dados com granularidade de minuto a minuto, acarreta uma complexidade considerável. Tal complexidade exige o desenvolvimento de algoritmos computacionais extremamente robustos, capazes não apenas de processar a alta volatilidade inerente a esses dados, mas também de minimizar os custos operacionais associados às transações, garantindo a lucratividade esperada.

O problema central, portanto, reside na necessidade de aprimorar a precisão das previsões do spread (diferença de preço) entre pares de ativos correlacionados e de otimizar a execução de operações de trade de alta frequência para que ocorram com o menor custo possível. As abordagens tradicionais, embora eficazes em contextos de menor granularidade, demonstram limitações quando confrontadas com o volume e a velocidade dos dados atuais, necessitando de inovações que integrem técnicas avançadas de aprendizado de máquina para superar essas barreiras e explorar ineficiências temporárias do mercado de forma mais eficiente e robusta.

4 Aspectos Éticos e Responsabilidade no Uso da IA para o Desenvolvimento da Solução

O uso de inteligência artificial na negociação de alta frequência no mercado financeiro traz importantes questões éticas e de responsabilidade. A capacidade da IA de processar dados e executar operações rapidamente exige uma análise cuidadosa para evitar consequências negativas.

4.1 Aspectos Éticos:

1. Equidade e Acesso Justo: A negociação de alta frequência impulsionada por IA pode criar uma vantagem desproporcional para alguns, marginalizando investidores menores e gerando dúvidas sobre a justiça do mercado.

2. Transparência e Explicabilidade (XAI): Modelos complexos de aprendizado de máquina, como as Redes Neurais Convolucionais, podem ser vistos como "caixas-pretas". Mesmo com Florestas Aleatórias, que geralmente são mais interpretáveis, a complexidade inerente à análise de dados de alta frequência ainda pode dificultar a compreensão exata das decisões. A falta de clareza sobre suas decisões dificulta a atribuição de responsabilidade em caso de falhas e impede auditorias eficazes.

3. Estabilidade do Mercado e Riscos Sistêmicos: Algoritmos de IA podem amplificar a volatilidade do mercado e causar *flash crashes*, criando riscos sistêmicos onde falhas em um sistema podem desestabilizar todo o ambiente financeiro.

4. Viés e Discriminação: Mesmo com dados de mercado, existe o risco de que os modelos captem ou amplifiquem vieses históricos, levando a resultados desfavoráveis para certos ativos ou participantes.

4.2 Responsabilidade no Desenvolvimento da Solução:

1. Robustez e Validação Rigorosa: É fundamental desenvolver algoritmos robustos e validá-los rigorosamente com dados reais (B3), utilizando métricas como retorno acumulado, índice de Sharpe e *drawdown* para garantir confiabilidade.

2. Testes de Estresse: A solução deve ser testada em cenários extremos e de "*cisnes negros*", além dos dados históricos típicos, para avaliar seu comportamento sob pressão e evitar contribuições para a desestabilização do mercado.

3. Desenvolvimento Ético por Design: A ética deve ser integrada desde o início do design do algoritmo, incorporando limites de risco claros e mecanismos para mitigar vieses, buscando modelos mais interpretáveis. Embora Redes Neurais Convolucionais sejam mais complexas, a busca por interpretabilidade, mesmo que parcial, é essencial.

4. Monitoramento Contínuo e Intervenção Humana: Após a implementação, é crucial haver monitoramento humano constante e a capacidade de intervenção manual rápida para corrigir anomalias ou adaptar-se a eventos imprevistos.

5. Conformidade Regulatória: A solução deve estar em total conformidade com as regulamentações financeiras, incluindo normas sobre manipulação de mercado, transparência e proteção ao investidor.

Em resumo, a integração de aprendizado de máquina na negociação de pares, utilizando modelos como Florestas Aleatórias e Redes Neurais Convolucionais, deve ser guiada por uma responsabilidade abrangente, assegurando que a inovação tecnológica seja utilizada de forma ética e segura para todos os participantes do mercado, além da otimização de lucros.

5 Dataset

Nossa base dados consiste em 3 conjuntos de ações para negociações em pares, cada conjunto tem o valor negociado das duas ações repeditamente a cada minuto. Os ativos em cada um dos dois conjuntos são correlacionados, propícios para o uso da estratégia de pair trading, além de serem ativos extremamente consolidados e com comportamentos de natureza menos volátil, o que ajuda na obtenção de resultados mais fiéis a realidade. As conjutos de ações são BBDC3 e BBDC4, ITAU4 e ITAU3 e PETR4 e PETR3. O dataset foi obtido através do professor Eli Haddad, que teve acesso a essas informações na B3, portanto trataremos de dados reais para este trabalho. Os dados já estão devidamente tratados/estruturados em csv.

Dataset e seu detalhamento disponiveis no repositorio: <https://github.com/joaquimrafael/AI_Project>

6 Metodologia

Trata-se de uma pesquisa de natureza aplicada, com abordagem quantitativa e caráter experimental (Gil, 2022; Marconi & Lakatos, 2022). A metodologia proposta estrutura-se em etapas sequenciais para viabilizar o desenvolvimento e a validação da solução.

Na primeira etapa, realiza-se uma revisão bibliográfica aprofundada sobre os principais tópicos relacionados ao estudo: negociação de pares e estratégias de reversão à média (Gatev *et al.*, 2006; Hadad *et al.*, 2024), negociação algorítmica de alta frequência (Zaharudin *et al.*, 2022) e modelos de aprendizado de máquina para previsão de séries temporais financeiras (Fischer & Krauss, 2018; Shah *et al.*, 2022; Oreshkin *et al.*, 2021). Essa revisão teórica fornece o embasamento necessário para a compreensão do problema e das soluções já propostas na literatura.

A segunda etapa consiste no levantamento e preparação dos dados históricos de alta frequência que serão utilizados nos experimentos. Serão selecionados pares de ações listadas na B3, obtendo-se suas séries de preços intradiários, e em seguida será realizada a limpeza e organização desses dados. Essa preparação inclui a verificação da qualidade e integridade das informações (por exemplo, detecção de *outliers* e tratamento de valores ausentes), o que é essencial dado o alto volume de ruído típico em ambientes de negociação de alta frequência (Zaharudin *et al.*, 2022). Ao final dessa etapa, espera-se ter uma base de dados consistente e pronta para alimentar os modelos preditivos.

A terceira etapa envolve o desenvolvimento e treinamento de modelos preditivos de aprendizado de máquina para estimar o espalhamento (spread) entre os membros de cada par de ações. Serão exploradas técnicas de aprendizado de máquina, como Florestas Aleatórias, e modelos de *deep learning*, como Redes Neurais Convolucionais. A seleção dessas técnicas apoia-se em sua utilização bem-sucedida em trabalhos recentes de predição financeira (Donick & Lera, 2021; Chen et al., 2018; Shah et al., 2022). Os modelos serão treinados com dados históricos separados em conjuntos de treinamento e teste, seguindo práticas recomendadas de validação (por exemplo, usando parte dos dados para avaliar a capacidade preditiva em relação aos dados não vistos).

Em seguida, na quarta etapa, procede-se à integração da estratégia de negociação de pares com as saídas fornecidas pelos modelos preditivos, desenvolvendo-se um algoritmo de execução automática de ordens. Nesta fase, a lógica tradicional de identificação de divergências e convergências de preços entre pares (Gatev *et al.*, 2006) será combinada com os sinais gerados pelos modelos de previsão para decidir dinamicamente as entradas e saídas das operações. A implementação desse algoritmo busca minimizar a latência e maximizar a eficiência na execução das ordens, conforme sugerido por Chen *et al.* (2018) em contextos de alta frequência. O resultado dessa etapa será um protótipo funcional do sistema de *trading* automatizado.

Por fim, a quinta etapa abrange os testes experimentais e a avaliação de desempenho da estratégia proposta utilizando dados reais da B3. O algoritmo híbrido será submetido a simulações de negociação (*backtesting*) em um período histórico, comparando seus resultados com os obtidos por uma estratégia clássica de negociação de pares e por modelos puramente preditivos individualmente. A avaliação da estratégia de negociação seguirá métricas consagradas na área, incluindo retorno acumulado, índice de Sharpe e *drawdown*, de modo semelhante ao procedimento adotado por Gatev et al. (2006) na mensuração de retornos. Adicionalmente, para a avaliação da precisão dos modelos puramente preditivos

na estimativa do valor futuro do *spread*, serão empregadas métricas de regressão como MAE (Mean Absolute Error), MSE (Mean Squared Error), MAPE (Mean Absolute Percentage Error) e R2 (Coeficiente de Determinação). Dessa forma, será possível verificar em que medida a integração entre aprendizado de máquina e negociação de pares proporciona melhorias estatisticamente significativas. Espera-se observar um desempenho superior da estratégia híbrida, tanto em termos de lucro quanto de controle de risco, o que validaria a contribuição desta pesquisa.

Referências

1 Challu, C. et al. Nhits: Neural hierarchical interpolation for time series forecasting. In: *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI 2023)*. [S.l.: s.n.], 2023. v. 37, p. 6989–6997.

2 Chen, Y.-Y.; Chen, W.-L.; Huang, S.-H. Developing arbitrage strategy in high-frequency pairs trading with filterbank cnn algorithm. In: *2018 IEEE International Conference on Agents (ICA)*. [S.l.: s.n.], 2018. p. 113–116.

3 Donick, D.; Lera, S. C. Uncovering feature interdependencies in high-noise environments with stepwise lookahead decision forests. *Scientific Reports*, v. 11, n. 1, p. 9238, 2021.

4 Fischer, T.; Krauss, C. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, v. 270, n. 2, p. 654–669, 2018.

5 Gatev, E.; Goetzmann, W. N.; Rouwenhorst, K. G. Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, v. 19, n. 3, p. 797–827, 2006.

6 GIL, A. C. Como classificar as pesquisas? In: _____. *Como elaborar projetos de pesquisa*. 7. ed. Barueri, SP: Atlas, 2022. p. 40–57. ISBN 6559771636.

7 Hadad, E. et al. Machine learning-enhanced pairs trading. *Forecasting*, v. 6, n. 2, p. 434–455, 2024.

8 Li, J. et al. Stock prediction based on deep learning and its application in pairs trading. In: *2022 International Symposium on Networks, Computers and Communications (ISNCC)*. [S.l.: s.n.], 2022. p. 1–6.

9 Lim, B. et al. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, v. 37, p. 1748–1764, 2021.

10 MARCONI, M. d. A.; LAKATOS, E. M. Técnicas de pesquisa. In: _____. *Fundamentos de metodologia científica*. 9. ed. São Paulo: Atlas, 2022. p. 202–246. ISBN 8597026561. Reimpr.

11 Oreshkin, B. N. et al. N-beats neural network for mid-term electricity load forecasting. *Applied Energy*, v. 293, p. 116918, 2021.

12 Sarmento, S. M.; Horta, N. Enhancing a pairs trading strategy with the application of machine learning. *Expert Systems with Applications*, v. 158, p. 113490, 2020.

13 Shah, A. et al. A stock market trading framework based on deep learning architectures. *Multimedia Tools and Applications*, v. 81, n. 10, p. 14153–14171, 2022.

14 Siami-Namini, S.; Tavakoli, N.; Siami Namin, A. A comparison of arima and lstm in forecasting time series. In: *17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. [S.l.: s.n.], 2018. p. 1394–1401.

15 Siami-Namini, S.; Tavakoli, N.; Namin, A. S. The performance of lstm and bilstm in forecasting time series. In: *2019 IEEE International Conference on Big Data (Big Data)*. [S.l.: s.n.], 2019. p. 3285–3292.

16 Tatsat, H.; Puri, S.; Lookabaugh, B. *Blueprints de aprendizado de máquina e ciência de dados para finanças: desenvolvendo desde estratégias de trades até robôs Advisors com Python*. [S.l.]: Alta Books, 2024. EPub. ISBN 9788550821726.

17 Zaharudin, K. Z.; Young, M. R.; Hsu, W.-H. High-frequency trading: Definition, implications, and controversies. *Journal of Economic Surveys*, v. 36, n. 1, p. 75–107, 2022.