Machine Learning para previsão de preços de ações em *High-Frequency Trading*.

Acácio Bonifácio de Marco

Faculdade de Computação e Informática – Universidade Presbiteriana Mackenzie
01302-907 – São Paulo – SP – Brazil

41803914@mackenzie.br

Abstract. *The main objective of this work is to employ traditional machine learning and deep learning models to forecast stock prices of companies in the Brazilian market. The goal is to identify which models and variables are more relevant for forecasting in high frequency trading.*

Resumo. *O objetivo principal deste trabalho é empregar modelos de aprendizado de máquina tradicionais e aprendizado profundo para previsão de preços de ações de empresas no mercado brasileiro. Busca-se identificar com quais modelos e variáveis faz-se uma previsão mais relevante em negociações de alta frequência.*

1. INTRODUÇÃO

A previsão de preços de ações é a maior dificuldade em finanças e é um assunto alvo de muita pesquisa, pois a capacidade de antecipar preços futuros de ativos geraria enormes ganhos aos investidores, o que trouxe muita atenção de diversos campos de estudo para o tópico. Kyong e Kyoung (2001) citam que os primeiros estudos que utilizaram redes neurais artificiais para previsão de ação de mercado foram feitos de 1990 até 1995.

Kamalov et al. (2021) disserta sobre as duas categorias principais de previsão de ação: Previsão de preço e previsão direcional do movimento. Aborda-se também o estudo que aplicou um conjunto de modelos da área de *Deep Learning* para previsão direcional do movimento de ações de empresas dos EUA utilizando dados históricos de alta frequência.

Brogaard (2010) define *High-Frequency Trading* (HFT), ou Negociação de Alta Frequência, como uma das formas de *trading* algorítmico onde um grande número de ordens são enviadas ao mercado em alta velocidade com tempos de execução de ida e volta medidas em microssegundos. De acordo com Hendershott e Riordan (2009), *Trading* Algorítmico é o "uso de algoritmos de computador para automaticamente fazer decisões de negociação, submeter ordens e gerenciar aquelas ordens após a submissão".

Larry Tabb et al. (2009) afirma que HFT surge como um produto do advento da internet e da análise quantitativa na década de 90, onde velocidade e acurácia eram necessárias para melhor execução de ordens de compra e venda.

Falkenberry (2002) nos mostra os desafios de manipular dados de alta frequência, ou *High-Frequency Data* (HFD), pois, quanto mais rápida é a negociação, maior é a probabilidade de erros ocorrerem na transferência de informações de trading.

Janiesch et al. (2021) fala sobre como *Machine Learning* trouxe diversos avanços em algoritmos sofisticados e eficientes técnicas de processamento. Como parte desses avanços podemos citar *Deep Learning* (DL). Pacheco e Pereira (2018) abordam as utilizações da técnica em diversas áreas de estudo como: Reconhecimento de imagens, de áudio, de caracteres e para reconhecimento facial.

Li e Bastos (2020) apresentam que estudos sobre modelos de previsão com dados de HFT tiveram LSTM (*Long-Short Term Memory*), uma técnica de rede neural recorrente usada em DL como a mais amplamente aplicada nesse cenário.

1.2 Problema da Pesquisa

Um dos desafios do trabalho é manipular os dados de alta frequência para poder gerar previsões de maior precisão, pois são aproximadamente 2.100 ticks por dia, ou seja, 530.000 por ano. Cada tick é a medida que representa o movimento mínimo para cima ou para baixo no preço de um título.

Podemos inferir que, uma extensa quantidade de dados como esta, torna necessário realizar uma limpeza de dados para que o conjunto de dados seja o menos errático possível, tornando a previsão capaz de alcançar maiores níveis de acurácia e precisão.

1.3 Objetivos da Pesquisa

Este estudo busca explorar o uso de modelos de aprendizado de máquina para a previsão de preços de ações de grandes empresas do mercado brasileiro, como o PETR4, VALE3 e BBAS3, com base em dados de HFT e dados diários de negociação desses papéis na Bolsa de São Paulo (B3). Para isso, serão empregados diferentes modelos de redes neurais profundas, como redes recorrentes e *transformers*. Os resultados serão comparados com os resultados de previsões tradicionais que incluem apenas dados de preço, sem uso de dados das operações de alta frequência.

Adicionalmente, também buscaremos explorar técnicas de tratamento e transformações dos dados e de explicabilidade dos modelos, o que poderá ajudar a indicar as características mais importantes na formação dos preços dos ativos.

1.4 Contribuições da pesquisa

Este estudo, assim, deve trazer as seguintes importantes contribuições:

a. Propõe e implementa diferentes transformações dos dados e modelos de *deep learning* para a previsão de preços de importantes ações do mercado brasileiro a partir de dados de preço diários e de operações HFT.

b. Busca identificar fatores das operações de HFT mais determinantes do preço dos ativos.

Ele está organizado do seguinte modo: após esta seção de introdução, a seção 2 apresenta a fundamentação teórica necessária para este estudo; a seção 3 apresenta a metodologia a ser empregada; e a seção 4 o cronograma previsto de desenvolvimento deste trabalho. Em todas essas seções é apresentado um resultado apenas preliminar, estando a pesquisa ainda em desenvolvimento.

2. REFERENCIAL TEÓRICO

Esta seção fornece uma revisão dos principais tópicos relacionados ao trabalho.

2.1 Negociação de Alta Frequência (HFT)

O aumento da popularidade de Negociações de Alta Frequência em conjunto com os avanços das tecnologias de IA para realização de previsões, e o acesso facilitado a dados de negociação, pavimentaram um caminho fértil para que HFT fosse apoiada por algoritmos de aprendizado de máquina nas decisões de compra e venda.

Aldridge (2013) caracterizou HFT pelos seguintes itens:

1. Toda operação que utiliza algoritmos de rápida execução.
2. Tecnologia de agilidade: geração de sinais, validação de modelos e execução em tempo supersônico.
3. Frequência de negociação em microssegundos.
4. Alto volume negociado com lotes de tamanho pequeno.
5. Atividade de negociação impraticável por humanos.

O tópico ganhou notoriedade, principalmente após o *Flash Crash,* que ocorreu em 6 de Maio de 2010, gerando preocupações sobre o impacto do uso de HFT e como ela poderia ser danosa a investidores menores. Brogaard et al. (2018) mostrou que existem poucas evidências de que HFT causasse *Extreme Price Movement* (EPM). Na verdade, foi observado que HFT provê liquidez durante esses movimentos por absorver desequilíbrios criados por negociadores de baixa frequência.

O site oficial da Nasdaq aponta que cerca de 50% do volume de *trading* de ações nos EUA é composto por HFT.

2.2 Previsão de preços

Suresh (2013) descreve que as duas principais abordagens para previsão de preços por analistas do mercado financeiro são: Análise Fundamentalista e Análise Técnica.

Uma Análise Fundamentalista é a examinação das forças subjacentes que afetam o bem-estar da economia, de grupos industriais e empresas. Combinam-se análises econômicas, industriais e empresariais para derivar um valor justo para uma ação. Esse valor é chamado de valor intrínseco e é usado para nortear os analistas.

A Análise Técnica é frequentemente utilizada como um suplemento para Análise Fundamentalista ao invés de um substituto dela. De acordo com a Análise Técnica, o preço de um ativo depende da lei da oferta e da procura. Ou seja, existe pouca correlação com o valor intrínseco, pois todo dado financeiro e informação de mercado já está refletido no seu valor.

2.3 *Machine Learning* e *Deep Learning*

Mahesh (2020) explica que o Aprendizado de Máquina, ou *Machine Learning* (ML), é usado para ensinar máquinas a lidar com dados mais eficientemente através de diferentes algoritmos para resolver problemas.

Shinde e Shah (2018) definem *Deep Learning* (DL), um subconjunto de Aprendizado de Máquina, como uma rede neural profunda, com um alto número de camadas e parâmetros, capaz de analisar e extrair conhecimento útil, tanto de grandes quantidades de dados quanto de dados coletados de diferentes fontes.

As principais técnicas de DL são: Autoencoder; *Deep Belief Network*; *Convolutional Neural Network*; *Recurrent Neural Network*; *Recursive Neural Network*; e *Direct Deep Reinforcement Learning*, podendo serem vistas nos seguintes domínios de aplicação:

- Visão computacional;
- Previsão;
- Análise semântica;
- Processamento de linguagem natural;
- Recuperação de informação; e
- Gestão de relacionamento com cliente.

Alguns exemplos das principais áreas de aplicação de técnicas de *Deep Learning* são: Reconhecimento de imagem; Reconhecimento de fala automático; Processamento de áudio e fala; Classificação de textos e documentos; Análise de imagem; Análise de diagnóstico médico; e Bioinformática.

O estudo cita também um trabalho de previsão para Representação e Negociação de Sinal Financeiro, que foi bem implementado utilizando *Direct Deep Reinforcement Learning* por Deng et al. em 2017.

2.4 Dados de Alta Frequência (HFD)

Os dados de alta frequência, também referido como dados *tick-by-tick*, registram cada movimentação mínima no preço da ação. Portanto, o volume de dados é dependente da volatilidade do ativo. Quanto maior a volatilidade, maior é a quantidade de ticks gerados.

Dacorogna (2001) aponta que observações de alta frequência em apenas um dia de mercado líquido podem ser equivalentes à quantidade de dados diários coletados em 30 anos. Essa vastidão de dados faz com que quase toda fonte de HFD contenha erros, por isso a limpeza dos dados é uma necessidade. Para lidar com esses erros, um dos algoritmos utilizados é o "data filter" que tem por objetivo eliminar outliers anormais.

Falkenberry (2002) cita alguns dos desafios deixados pelas características estatísticas de HFD, os quais são:

- Natureza assíncrona dos dados.
- A miríade dos possíveis tipos de erros, como: ticks ruins isolados, múltiplos ticks ruins em sucessão, erros decimais, erros de transposição, e a perda da porção decimal de um número.
- O tratamento do tempo.

- Diferenças na frequência de ticks entre títulos.
- Padrões sazonais na frequência de ticks intradiários.
- *Bid-ask bounce*.
- A inabilidade de explicar um dado errado.

O grau de filtragem deve ser calibrado adequadamente, pois, caso não seja, pode gerar problemas, como dados inúteis serem mantidos ou gerar dados excessivamente limpos que mudariam a realidade dos dados e suas propriedades estatísticas.

Hautsch (2011) aponta que existem 5 níveis principais de detalhamento de informações em HFD: i) *Trade data*; ii) *Trade and quote data*; iii) *Fixed Level Order Book data*; iv) *Messages on all limit order activities*; e v) *Data on order book snapshots*.

Abaixo apresentamos um tipo de conjunto de HFD:

SYMBOL	DATE	TIME	EX	PRICE	SIZE	COND	CORR	G127
MSFT	2009-06-01	36601	Z	21.1900	200		0	0
MSFT	2009-06-01	36601	Z	21.1900	1000		0	0
MSFT	2009-06-01	36601	Z	21.1900	100		0	0
MSFT	2009-06-01	36601	B	21.1900	400	@F	0	0
MSFT	2009-06-01	36601	B	21.1900	400	@F	0	0
MSFT	2009-06-01	36602	D	21.1912	470		0	0
MSFT	2009-06-01	36602	Z	21.1900	200		0	0
MSFT	2009-06-01	36602	Q	21.1900	900		0	0
MSFT	2009-06-01	36602	Q	21.1900	100	@F	0	0
MSFT	2009-06-01	36602	Q	21.1900	100	@F	0	0
MSFT	2009-06-01	36602	Q	21.1900	300		0	0
MSFT	2009-06-01	36602	Q	21.1900	100		0	0
MSFT	2009-06-01	36602	D	21.1900	100	@F	0	0
MSFT	2009-06-01	36602	D	21.1900	100	@F	0	0

Figura 1. Registro de trades para Microsoft da base de dados TAQ (Trades and Quotes)

O registro acima contém as seguinte informações: Símbolo da ação (SYMBOL); Data de negociação (DATE); Momento da cotação (TIME), em segundos; Corretora em que a negociação ocorreu (EX), Preço da transação (PRICE), Tamanho do trade (SIZE); Condição da venda (COND); Indicador de correção da corretividade de um trade(CORR); G trades (G127), que representam membros da NYSE (Bolsa de Valores de Nova Iorque) fazendo negociação por conta própria e *block trades* (ordem em bloco).

SYMBOL	DATE	EX	TIME	BID	BID SZ	OFFER	OFF SZ	MODE
MSFT	2009-06-01	Z	36001	21.1100	43	21.1300	38	12
MSFT	2009-06-01	T	36001	21.1100	97	21.1200	6	12
MSFT	2009-06-01	T	36001	21.1100	92	21.1200	6	12
MSFT	2009-06-01	T	36001	21.1100	82	21.1200	6	12
MSFT	2009-06-01	I	36001	21.1100	9	21.1200	5	12
MSFT	2009-06-01	T	36001	21.1100	72	21.1200	6	12
MSFT	2009-06-01	B	36001	21.1100	30	21.1300	22	12
MSFT	2009-06-01	D	36001	21.1000	8	21.2100	2	12
MSFT	2009-06-01	B	36001	21.1100	31	21.1300	22	12
MSFT	2009-06-01	B	36002	21.1100	30	21.1300	22	12
MSFT	2009-06-01	B	36002	21.1100	21	21.1300	22	12
MSFT	2009-06-01	T	36002	21.1100	72	21.1200	5	12
MSFT	2009-06-01	T	36002	21.1100	78	21.1200	5	12
MSFT	2009-06-01	I	36002	21.1100	9	21.1300	33	12

Figura 2. Registro de cotações para Microsoft da base de dados TAQ (Trades and Quotes)

Acima pode-se observar um registro contendo SYMBOL, DATE, TIME e EX como anteriormente, mais o preço da oferta de compra (BID), o tamanho da oferta de compra em lotes de 100 *shares* (BID SZ), o preço da oferta de venda (ASK), o tamanho da oferta de compra em lotes de 100 *shares* (OFF SZ) e a condição da cotação (MODE).

Para fins de comparação, temos abaixo um conjunto tradicional de dados no período diário das ações da Microsoft retirado da base Yahoo! Finance:

Date	Open	High	Low	Close*	Adj Close**	Volume
Jun 04, 2009	21.77	21.90	21.58	21.83	16.54	42,330,000
Jun 03, 2009	21.31	21.76	21.29	21.73	16.46	56,039,600
Jun 02, 2009	21.36	21.98	21.20	21.40	16.21	48,935,700
Jun 01, 2009	21.00	21.50	20.86	21.40	16.21	57,317,100

Figura 3. Registro de preços no período diário para Microsoft

Verifica-se que os dados de alta frequência fornecem muito mais informações do que os tradicionais, que se apoiam apenas em preços OHLC (*Open*, *High*, *Low* e *Close*) e volume negociado.

2.6 Trabalhos correlatos

Rundo (2019) aplica a técnica de LSTM com *Reinforcement Learning Layer* para previsão de tendência financeira para FX (*Foreign Exchange*), também conhecido como mercado de câmbio, em sistemas de negociação de alta frequência. Com uma acurácia média de aproximadamente 85%, o algoritmo mostrou ser capaz de prever tendências de médio-curto prazo de uma moeda através de sua tendência histórica e por meios de correlação de dados com outras moedas utilizando técnicas conhecidas no campo financeiro. A parte final incluiu um *grid trading engine* que performará operações de alta frequência com o objetivo de maximizar o lucro e reduzir *drawdown.*

O sistema de *trading* foi validado em EUR/USD e confirmou alta performance em termos de ROI, ou Retorno de Investimento (98.23%), e *drawdown* reduzido (15.97%).

Lahmiri e Bekiros (2021) aplicaram *Deep Forward Neural Network* (DFFNN) em um conjunto de dados composto por 65.535 amostras para análise e previsão de dados de preço de alta frequência de Bitcoin. Utilizou-se de 3 tipos de algoritmos diferentes em conjunto com DFFNN e o que se sobressaiu foi o algoritmo Levenberg-Marquadt, mostrando ser o mais efetivo e fácil de implementar.

Arévalo et al. (2016) descreve o uso de Redes Neurais Profundas, ou *Deep Neural Networks* (DNN), como base para estratégias de alta frequência. Os dados de treino e teste utilizados foram de transações *tick-by-tick* da AAPL (Apple) de setembro a novembro de 2008. A melhor DNN encontrada teve 66% de acurácia direcional, atingindo 81% de sucesso no período de testes.

3. METODOLOGIA DA PESQUISA

Para verificar a hipótese colocada neste trabalho e atingir os objetivos, são propostas as seguintes atividades de pesquisa:

1. Coleta e caracterização de dados HFT
2. Estudo de técnicas para agregação de dados HFT e para previsão de
3. Aplicação e avaliação dos resultados obtidos
4. Análise dos dados
5. Preparação de artigo para submissão

Nosso trabalho tem como natureza a pesquisa aplicada, abordando o problema de forma quantitativa com a finalidade metodológica. A pesquisa, usará como meio o Laboratório.

- Quanto à Natureza, esta é uma pesquisa Aplicada, pois será feita em cima de dados concretos;
- Quanto a forma de Abordagem, a pesquisa pode ser considerada Quantitativa, uma vez que desenvolve técnicas quantitativas, ou seja, com cálculos e medidas objetivas;
- Quanto aos Fins, a pesquisa é Metodológica, já que apresenta uma metodologia de trabalho, inclusive com o desenvolvimento de um modelo preditivo;
- Quanto aos Meios, serão utilizados os seguintes recursos: Bibliografia, Documental, Estudo de Caso e Laboratório.

4. CRONOGRAMA

As atividades desta pesquisa deverão se desenvolver de acordo com o cronograma apresentado a seguir:

ATIVIDADE	**MÊS**			
	1	**2**	**3**	**4**
1. Preparação dos dados	■			

ATIVIDADE	MÊS			
	1	2	3	4
2. Construção e treinamento de modelos				
3. Refinamento dos modelos e avaliação				
4. Preparação de artigo para submissão				

Referências

Kyong Joo Oh, Kyoung-jae Kim, Analyzing stock market tick data using piecewise nonlinear model, Expert Systems with Applications, Volume 22, Issue 3, 2002.

Kamalov, Firuz and Gurrib, Ikhlaas and Rajab, Khairan, Financial Forecasting with Machine Learning: Price Vs Return (March 18, 2021). Kamalov, F., Gurrib, I. & Rajab, K. (2021). Financial Forecasting with Machine Learning: Price Vs Return. Journal of Computer Science, 17(3), 251-264. https://doi.org/10.3844/jcssp.2021.251.264, Available at SSRN: https://ssrn.com/abstract=3808539

Brogaard, Jonathan. "High frequency trading and its impact on market quality." Northwestern University Kellogg School of Management Working Paper 66 (2010).

BROGAARD, J. *et al.* High frequency trading and extreme price movements, J. Financ. Econ., 2018

ALDRIDGE, I. High-Frequency trading: a practical guide to algorithmic strategies and trading systems. 2nd. New Jersey: John Wiley & Sons, 2013.

MALCENIECE, L.; MALCENIEKS, K.; PUTNINS, T. J. High frequency trading and comovement in financial markets, Journal of Financial Economics,Volume 134, Issue 2, 2019.

HENDERSHOTT, T.; JONES, C.M.; MENKVELD, A.J. Does Algorithmic Trading Improve Liquidity?, THE JOURNAL OF FINANCE, VOL. LXVI, NO. 1, 2011

VIRGILIO, G.P.M. High frequency trading: a literature review. Financial Markets and Portfolio Management, n. 33, p.183-208, 2019. Disponível em: <https://link.springer.com/article/10.1007/s11408-019-00331-6>. Acesso em 15/05/2022.

MARKWICK, Dean. Order Flow Imbalance - A High Frequency Trading Signal

. 2 de fevereiro de 2022. Disponível em: <https://dm13450.github.io/2022/02/02/Order-Flow-Imbalance.html>. Acesso em: 8 de nov. de 2022.

LARRY TABB et al., US Equity High Frequency Trading: Strategies,

Sizing and Market Structure, TABB GROUP, Sept. 2009.

KEARNS, M.; KULESZA, A.; NEVMYVAKA, Y., 2010, Empirical Limitations on High Frequency Trading Profitability. arXiv:1007.2593v2

FALKENBERRY, T. N. (2002), "High Frequency Data Filtering", Technical Report, Tick Data.

MYKLAND, P.A.; & Zhang, L., 2012, The Econometrics of High Frequency Data.

PACHECO, C. A. R., & PEREIRA, N. S. (2018). Deep Learning Conceitos e Utilização nas Diversas Áreas do Conhecimento. *Revista Ada Lovelace*, *2*, 34–49. Recuperado de http://anais.unievangelica.edu.br/index.php/adalovelace/article/view/4132

OLIVEIRA, W., Software para reconhecimento de espécies florestais a partir de imagens digitais de madeiras utilizando Deep Learning. Universidade Tecnológica Federal do Paraná. Programa de Pós - Graduação em Tecnologias Computacionais para o Agronegócio. Medianeira, 2018.

SILVA, L. C. P.; OSÓRIO, F. S., Sistema autônomo e inteligente de reconhecimento facial para autorização de entrada de pessoal em ambientes restritos, USP – Universidade de São Paulo – SP, 2016.

SILVA, J. M. S. C. F, Detecção de convulsões epiléticas em eletroencefalogramas usando Deep Learning. ISEP – Instituto Superior de Engenharia do Porto. 2017.

RUNDO F. Deep LSTM with Reinforcement Learning Layer for Financial Trend Prediction in FX High Frequency Trading Systems. *Applied Sciences*. 2019; 9(20):4460. https://doi.org/10.3390/app9204460

LAHMIRI, S., BEKIROS, S. Deep Learning Forecasting in Cryptocurrency High-Frequency Trading. *Cogn Comput* 13, 485–487 (2021). https://doi.org/10.1007/s12559-021-09841-w

A S, SURESH. (2013). A STUDY ON FUNDAMENTAL AND TECHNICAL ANALYSIS. 2. 44-59.

MAHESH, BATTA. "Machine learning algorithms-a review." *International Journal of Science and Research (IJSR).[Internet]* 9 (2020): 381-386.

HAUTSCH, NIKOLAUS. *Econometrics of financial high-frequency data*. Springer Science & Business Media, 2011.

Janiesch, C., Zschech, P. & Heinrich, K. Machine learning and deep learning. *Electron Markets* 31, 685–695 (2021). https://doi.org/10.1007/s12525-021-00475-2

Borovkova, S, Tsiamas, I. An ensemble of LSTM neural networks for high-frequency stock market classification. *Journal of Forecasting*. 2019; 38: 600– 619. https://doi.org/10.1002/for.2585

A. W. Li and G. S. Bastos, "Stock Market Forecasting Using Deep Learning and Technical Analysis: A Systematic Review," in *IEEE Access*, vol. 8, pp. 185232-185242, 2020, doi: 10.1109/ACCESS.2020.3030226.

Dacorogna, M. M. (2001). An introduction to high-frequency finance. San Diego: Academic Press.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F..
Inteligência Artificial: Uma
Abordagem de Aprendizagem de Máquina
. Rio de Janeiro: LTC, 2011.

Arévalo, Andrés, et al. "High-frequency trading strategy based on deep neural networks." *International conference on intelligent computing*. Springer, Cham, 2016.