Rede Neural Multivariada Para Apreçamento de Ativos no Mercado Financeiro Brasileiro do Setor Varejista

Lucas Dejard[1], **Diego L. Cardoso**[1]

[1]Laboratório de Pesquisa Operacional – Universidade Federal do Pará (UFPA) – Belém, PA – Brazil

lucas.mendonca@castanhal.ufpa.br, diego@ufpa.br

Abstract. *Using public data from the Monthly Survey of Trade (PMC) provided by Brazilian Institute of Geography and Statistics (IBGE) and historical stock data from the retailer Grupo Mateus S.A (GMAT3), this paper proposes a multivariate Long Short-Term Memory (LSTM) model that achieves higher accuracy in predicting the closing price of the GMAT3 ticker on the Brazilian stock exchange compared to a univariate model, what shows the model's utility in the pricing of assets within the Brazilian retail sector and highlight the research potential of predictive modeling for time series analysis.*

Resumo. *Utilizando dados públicos da Pesquisa Mensal do Comércio(PMC) fornecida pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e dados públicos da ação da varejista Grupo Mateus S.A (GMAT3), neste artigo é proposto um modelo de Long Short-Term Memory (LSTM) multivariado que apresenta maior precisão na predição do valor de fechamento no ativo GMAT3 da bolsa de valores brasileira quando comparado com o modelo univariado, demonstrando sua utilidade no apreçamento de ativos do setor de varejo brasileiro e potencial de pesquisa da linha de modelos preditivos em séries temporais.*

1. Introdução

O mercado de capitais desempenha um papel fundamental na economia moderna, facilitando a captação de recursos para empresas e oferecendo oportunidades de rentabilidade para investidores. No entanto, a predição de preços de ativos financeiros é uma tarefa intrinsecamente complexa devido à alta volatilidade, não linearidade e influência de múltiplos fatores exógenos, conhecida na literatura como a hipótese do mercado eficiente (EMH) [Fama 1970].

Em contrapartida, a validade estrita da EMH é frequentemente contestada pela literatura de Finanças Comportamentais, argumentando que os investidores não atuam de forma puramente racional, estando sujeitos a vieses cognitivos e reações emocionais, distorcendo os preços. Essa perspectiva sugere que os mercados possuem ineficiências informacionais e anomalias temporais. É neste hiato de eficiência que modelos baseados em *Deep Learning* encontram aplicabilidade, pois possuem a capacidade de identificar padrões não lineares sutis e dependências de longo prazo que refutam a aleatoriedade total dos preços [Shiller 2003].

Com a grande produção e disponibilidade de dados no momento atual da internet, modelos multivariados, em que os dados utilizados advêm de diferentes fontes e formatos, têm sido desenvolvidos e têm atingido desempenho acima da média. Esse tipo de modelo

é essencial para aprender não apenas o padrão temporal, mas também a relação de causa e efeito (ou correlação forte) entre diferentes variáveis [Gopali et al. 2024].

No contexto brasileiro, o setor varejista é particularmente sensível às flutuações econômicas, como inflação, taxas de juros e poder de compra da população. Embora métodos estatísticos tradicionais como *Autoregressive Integrated Moving Average* (ARIMA) e *Generalized Autoregressive Conditional Heteroscedasticity* tenham sido amplamente utilizados, eles muitas vezes falham em capturar dependências de longo prazo e padrões não-lineares complexos. É neste cenário que as Redes Neurais Artificiais, especificamente as redes *Long Short-Term Memory* (LSTM), têm demonstrado superioridade [Churi et al. 2023].

A literatura recente em Sistemas de Informação aponta para a eficácia de modelos híbridos que incorporam análise fundamentalista, variáveis macroeconômicas e ou *analytics* de redes sociais. [Bantis et al. 2023] demonstraram que a integração de dados de busca do *Google Trends* a modelos de fatores dinâmicos aprimora significativamente a acurácia de *forecasting* do Produto Interno Bruto (PIB), apresentando ganhos de performance em economias emergentes como o Brasil. [Leippold et al. 2022] utilizando redes neurais identificaram que a forte presença de investidores de varejo no mercado chinês aumenta a previsibilidade dos retornos a curto prazo, especialmente em ações de pequena capitalização (*small caps*) enquanto [Wang 2022] utiliza indicadores econômicos e aprendizado profundo focados exclusivamente na previsão de receitas do setor varejista.

Este trabalho de pesquisa propõe um modelo preditivo focado no setor varejista da B3. A principal contribuição e diferencial deste desenho de pesquisa reside na integração de séries temporais financeiras tradicionais: *Open*, *High*, *Low*, *Close* (OHLC) com indicadores macroeconômicos oficiais do governo brasileiro, especificamente a *Pesquisa Mensal do Comércio* (PMC) do Instituto Brasileiro de Geografia e Estatística (IBGE), criando um modelo multivariado. O objetivo é investigar se a inclusão de dados do comércio real aumenta a acurácia preditiva dos modelos de Deep Learning para ações desse setor.

2. Metodologia e Desenho da Pesquisa

A metodologia deste trabalho segue uma abordagem quantitativa experimental, estruturada nas etapas de Coleta de Dados, Pré-processamento, Modelagem e Avaliação.

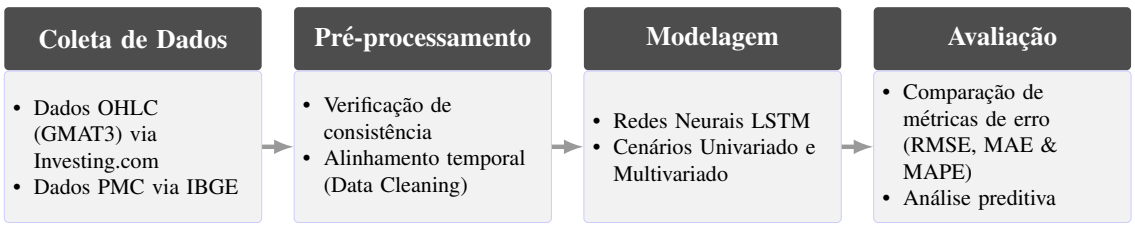

Figura 1. Fluxo metodológico da pesquisa: etapas de processamento e análise.

2.1. Coleta e Fontes de Dados

Para o estudo de caso inicial e validação do software, foi selecionada a ação GMAT3, uma das maiores empresas do varejo alimentar do país, com forte presença regional e sensibilidade direta ao consumo das famílias. Os dados são coletados de duas fontes primárias:

1. Dados de Mercado (B3): Coletados via gerador de relatório do site Investing.com[1], que disponibiliza dados em tempo real do mercado financeiro (OHLC).
2. Dados Macroeconômicos (IBGE): Coletados via API do Sistema IBGE de Recuperação Automática (SIDRA)[2], que fornece estatísticas oficiais de desempenho do setor, especificamente o *"Índice de base fixa da receita nominal de vendas no comércio varejista (1) e comércio varejista ampliado (2), por atividades"*

2.2. Pré-processamento e Engenharia de Atributos

Os dados do PMC são constantemente atualizados e passam por revisão, o que altera seus valores e ocasiona a mudança de ano-base. A série temporal mais atual do PMC tem como ano-base 2022, o que limitou a utilização dos dados para a análise do modelo ao período de janeiro de 2022 a setembro de 2025 (03/01/2022 a 30/09/2025). Visto que os dados possuem periodicidades diferentes (ações são diárias, PMC é mensal), realizou-se um processo de alinhamento temporal. Os dados da PMC foram repetidos para preencher a granularidade diária, permitindo que o modelo LSTM processe as entradas simultaneamente. Os dados foram normalizados utilizando a técnica *Standard Scaler* transformando cada *feature* (característica) para ter média zero e desvio padrão um (normalização Z-score). Ele subtrai a média do valor e divide pelo desvio padrão, essencial para a convergência eficiente de redes neurais baseadas em gradiente.

A seleção das variáveis exógenas baseou-se na análise de correlação entre os indicadores da PMC e o preço de fechamento da ação. Conforme ilustrado na Figura 2, observa-se uma correlação positiva moderada entre os setores de hipermercados, supermercados, farmacêuticos e atacado com os dados de mercado do ativo financeiro utilizado para validação, justificando sua inclusão no vetor de características do modelo.

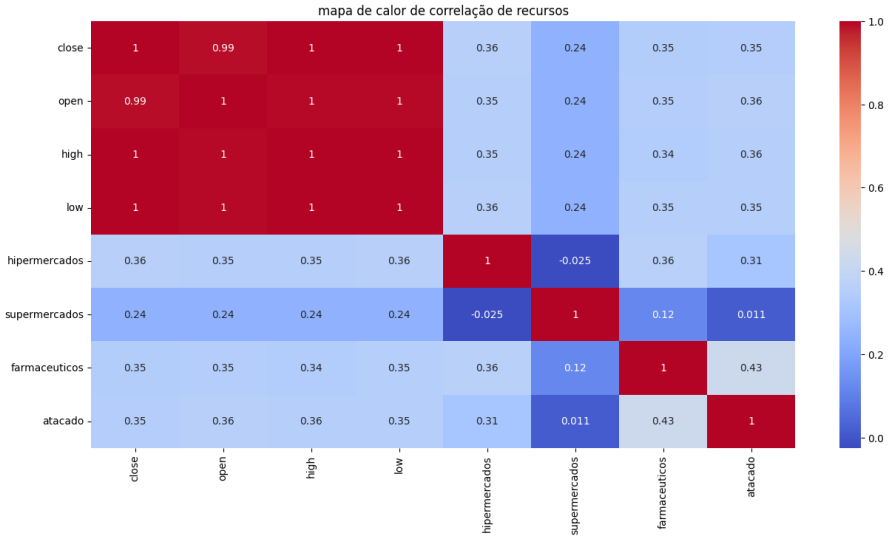

Figura 2. Matriz de correlação (Pearson) entre dados de mercado (OHLC) e setores da PMC.

[1] https://www.investing.com
[2] https://apisidra.ibge.gov.br/

2.3. Arquitetura do Modelo Proposto

O artefato desenvolvido em Python utiliza a biblioteca Keras/TensorFlow. A arquitetura implementada consiste em uma rede sequencial composta por:

- **Camada de Entrada:** Janela deslizante (*look-back*) configurada no formato $(N, \textit{Features})$, processada por uma camada LSTM de 64 unidades com retorno de sequências para capturar padrões temporais complexos.
- **Camadas Ocultas:** Uma segunda camada LSTM de 64 unidades (sem retorno de sequência) para condensar a informação temporal em um vetor de características, seguida por uma camada *Dense* de 64 neurônios com função de ativação ReLU para interpretação não-linear das características.
- **Regularização:** Uma camada de *Dropout* (20%) posicionada antes da saída para mitigar o *overfitting*, desativando aleatoriamente neurônios durante o treinamento.
- **Camada de Saída:** Uma camada *Dense* com um único neurônio e ativação linear, responsável por projetar o valor final previsto para a ação.

A função de perda (*loss function*) utilizada é o erro quadrático médio, adequada para problemas de regressão, otimizada pelo algoritmo Adam.

3. Resultados Preliminares e Discussão

Atualmente, o modelo é capaz de ingerir os dados históricos da GMAT3 e os dados da PMC. Foi realizada a comparação de desempenho do modelo em dois cenários:

1. Univariado: Utilizando apenas o histórico de preços da ação.
2. Multivariado: Adicionando a série histórica da PMC como variável exógena.

A Tabela 1 apresenta a evolução das métricas de erro em diferentes cenários de treinamento. Observa-se que, com 20 épocas, a diferença entre o modelo univariado e o multivariado é marginal, indicando que a rede ainda não havia convergido o suficiente para extrair padrões complexos das variáveis exógenas. No entanto, ao estender o treinamento para 200 épocas, o modelo enriquecido com os dados da PMC (Multivariado) superou o modelo base em todas as métricas, atingindo um *Mean Absolute Percentage Error* (MAPE) de 1.31% contra 1.58% do modelo univariado. A redução da raiz do erro quadrático médio (RMSE) para R$0.13 demonstra que a inclusão de indicadores do comércio varejista contribui efetivamente para minimizar os erros de predição, validando a hipótese de que dados da economia real refinam a acurácia de modelos financeiros.

Tabela 1. Comparativo de Métricas de Erro: Univariado vs. Multivariado (+PMC)

Modelo	Épocas	RMSE (R$)	MAE (R$)	MAPE (%)
Univariado	20	R$0.17	R$0.13	1.76%
Multivariado (+PMC)	20	R$0.17	R$0.12	1.69%
Univariado	200	R$0.15	R$0.11	1.58%
Multivariado (+PMC)	**200**	**R$0.13**	**R$0.09**	**1.31%**

Diante dos resultados validados em análise estatística, as observações empíricas sugerem que a inclusão da PMC ajuda o modelo a reagir melhor a tendências de médio prazo, enquanto o preço diário captura a volatilidade de curto prazo.

A capacidade de generalização do modelo proposto pode ser visualizada na Figura 3. O gráfico apresenta o comparativo entre os dados reais (conjunto de teste) e as predições realizadas pelo modelo Multivariado após 200 épocas de treinamento. Nota-se que a rede neural consegue acompanhar a tendência de recuperação do ativo observada no último semestre, capturando a volatilidade com um atraso (*lag*) reduzido.

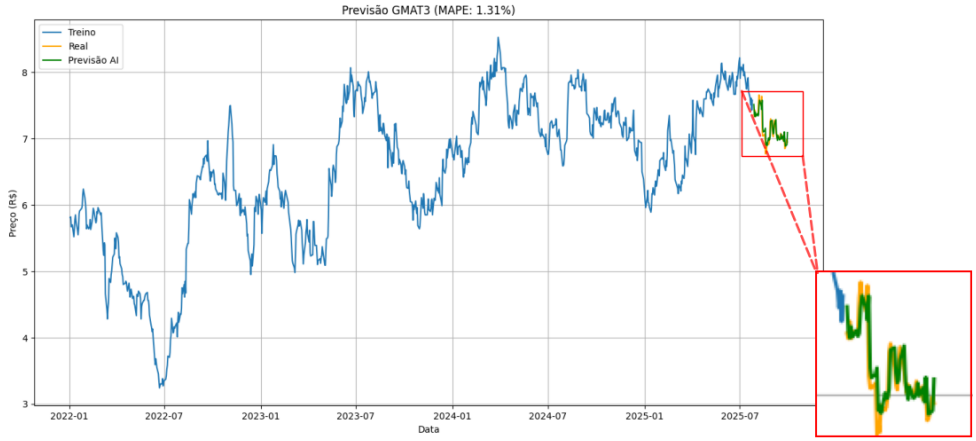

Figura 3. Comparativo entre valores reais e preditos para GMAT3 (Modelo Multivariado com PMC).

Um desafio encontrado refere-se ao atraso na divulgação de dados do IBGE (publicados com semanas de atraso em relação ao tempo real). O desenho da pesquisa está sendo ajustado para considerar esse *delay* na janela de predição, garantindo que o modelo não utilize dados futuros que não estariam disponíveis em um cenário de *trading* real.

4. Contribuições Esperadas e Próximos Passos

Este trabalho visa contribuir para a área de Sistemas de Informação ao demonstrar como a integração de dados governamentais abertos (*open data*) pode enriquecer sistemas de suporte à decisão financeira. Diferente de abordagens puramente especulativas, propõe-se uma ferramenta que une a ”economia real (dados do IBGE)”com o mercado financeiro.

Como próximos passos para a finalização da pesquisa de Mestrado, define-se:

1. Expansão do *Dataset*: Incluir outras varejistas de grande porte (ex: Magazine Luiza, Lojas Renner) para validar a generalização do modelo.
2. Aprimoramento da Arquitetura: Testar mecanismos de atenção (*attention mechanisms*) para identificar quais dias na janela temporal histórica têm maior peso na decisão da rede.
3. Comparativo de *Benchmarking*: Comparar formalmente os resultados com modelos clássicos (ARIMA) e outros modelos de ML (XGBoost).
4. Análise de Atraso: Implementar uma estratégia de *forecasting* da própria PMC para mitigar o impacto do atraso na divulgação dos dados oficiais.

Referências

Bantis, E., Clements, M. P., and Urquhart, A. (2023). Forecasting gdp growth rates in the united states and brazil using google trends. *International Journal of Forecasting*, 39(4):1909–1924.

Churi, A., Chakraborty, D., Khatwani, R., Pinto, G., Shah, P., and Sekhar, R. (2023). Stock price prediction using deep learning and sentiment analysis. In *2023 2nd International Conference on Futuristic Technologies (INCOFT)*, pages 1–6. IEEE.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417.

Gopali, S., Siami-Namini, S., Abri, F., and Namin, A. S. (2024). A comparative multivariate analysis of var and deep learning-based models for forecasting volatile time series data. *IEEE Access*, 12:155423–155436.

Leippold, M., Wang, Q., and Zhou, W. (2022). Machine learning in the chinese stock market. *Journal of Financial Economics*, 145(2):64–82.

Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives*, 17(1):83–104.

Wang, C.-H. (2022). Considering economic indicators and dynamic channel interactions to conduct sales forecasting for retail sectors. *Computers & Industrial Engineering*, 165:107965.