Previsão do Índice Bovespa Utilizando Variáveis Exógenas

Orlando da Silva Junior[1], Osvaldo Ribeiro dos Santos[2]

[1] Escola Politécnica – Universidade de São Paulo (USP) – São Paulo – SP

[2] Centro Universitário das Faculdades Metropolitanas Unidas (FMU) – São Paulo – SP

orlando.silvajr@usp.br, osvaldoribeiro70@gmail.com

Abstract. *Econometric models have adopted the use of exogenous variables to improve forecast performance. However, most models still focus on short-term applications. In this work, we study the use of exogenous variables for long-term forecasts of the future performance of the Bovespa index. We adopt a methodology based on artificial neural networks to consolidate, select, and model variables that encapsulate 15 years of information on the Brazilian financial market. The methodology also allows for the selection of the most relevant variables for the index prediction. The results show that an artificial neural network can predict the future performance of the Ibovespa with over 85% explanatory power of the selected variables, even during periods of high market volatility.*

Resumo. *Modelos econométricos têm adotado o uso de variáveis exógenas para melhorar a o desempenho de modelos de previsão. No entanto, a maior parte dos modelos ainda focam as aplicações de curto e curtíssimo prazos. Neste trabalho, estudamos o uso de variáveis exógenas para a previsão do rendimento futuro do índice Bovespa em prazos longos. Adotamos uma metodologia baseada em redes neurais artificiais para consolidar, selecionar e modelar variáveis que consolidam 15 anos de informações sobre o mercado financeiro brasileiro. A metodologia também permite a seleção das variáveis mais relevantes para a previsão do índice. Os resultados mostram que uma rede neural artificial é capaz de prever o desempenho futuro do Ibovespa com mais de 85% de explicabilidade das variáveis selecionadas, mesmo em períodos de alta volatilidade no mercado.*

1. Introdução

Há mais de duas décadas, pesquisadores tem utilizado técnicas Aprendizado de Máquina (AM) para prever o desempenho do mercado de ações. No entanto, a maior parte dos trabalhos tem observado apenas investimentos em prazos curtíssimos, de poucos minutos a um dia [Bhandari et al. 2002]. Essa ênfase em horizontes temporais extremamente curtos pode ser atribuída a várias razões, incluindo a alta volatilidade e a rápida mudança das condições de mercado que oferecem oportunidades para lucros rápidos, bem como a disponibilidade imediata dos dados.

Mais recentemente, modelos econométricos têm incorporado o uso de variáveis exógenas

para explicar as variações nos preços de ações e índices, além do próprio uso dos preços ao longo do tempo. Indicadores econômicos como inflação, taxa de juros e métricas de políticas fiscais de outros países têm permitido aperfeiçoar as abordagens estatísticas tradicionais [Banas and Utnik-Banas 2021] e também ampliar a capacidade de interpretabilidade de modelos neurais [Olivares et. al 2023].

Neste trabalho, estudamos o uso de variáveis exógenas para a previsão do rendimento do índice Bovespa (Ibovespa) nos 12 meses seguintes. No estudo, Redes Neurais Artificiais (RNA) são empregadas para a previsão do índice, considerando o uso de informações a partir da ótica do mercado financeiro brasileiro, incluindo preços de produtos vendidos por empresas negociadas na Bolsa de Valores B3, indicadores de desempenho da economia e do comércio exterior, indicadores de desequilíbrios nas contas públicas, níveis de inflação e de taxas de juros, além dos próprios índices de ações listadas na bolsa.

O conjunto de variáveis selecionadas para este estudo mapeia os principais fatores que podem influenciar os resultados das empresas, permitindo a identificação de comportamentos sazonais dentro do prazo de previsão. Por exemplo, o que aconteceria com o preço do minério de ferro em 12 meses após ele atingir um valor mínimo histórico?

Se um modelo de RNA for capaz de conectar uma grande quantidade de relacionamentos que respondam a perguntas como essa, o modelo poderá antecipar o comportamento do Ibovespa nos meses subsequentes. Além disso, essa informação poderá auxiliar investidores, permitindo que eles se comprometam com uma escolha de investimentos por um prazo maior estabelecido, em vez de manter a tradicional estratégia de comprar e vender ações com frequência, conhecida como *day trading*.

O enfoque deste trabalho em investimentos de médio e longo prazos visa atender investidores individuais que buscam acumular patrimônio para a compra de imóveis ou para a aposentadoria. Essa mudança de foco abre espaço para a utilização de um conjunto abrangente de variáveis relacionadas ao desempenho da economia e que impactam nos resultados das empresas.

As próximas seções detalham como os objetivos deste trabalho são alcançados. Na seção 2, revisamos os trabalhos relacionados e discutimos com a nossa proposta. A seção apresenta a metodologia que empregamos para a avaliação da previsão do Ibovespa, bem como os experimentos realizados. Na seção 4, apresentamos e discutimos os resultados obtidos pelos experimentos. Finalmente, na seção 5, concluímos o estudo.

2. Trabalhos Relacionados

Desenvolvimentos recentes indicam que as RNAs são um dos métodos mais eficazes em AM para a classificação e regressão de valores de ações ou índices [Hu, Zhao and Khushi 2021]. Apesar de serem consideradas pouco explicativas, RNAs são capazes de modelar as relações complexas e não lineares presentes nos dados financeiros. Modelos neurais também podem aprender e adaptar-se a padrões ocultos nos dados históricos de preços, volumes de negociação, indicadores econômicos e outras variáveis relevantes.

Entre as variáveis predominantes dos modelos econométricos, os indicadores técnicos e valores de ações e índices têm sido mais empregados [De Campos and De Figueiredo 2021]. Variáveis fundamentalistas são raras, em função da defasagem em sua divulgação. Alguns trabalhos ainda exploram a combinação de variáveis macroeconômicas com indicadores técnicos e valores de ações e índices [Bhandari et al.

2022, Lakshminarayanan and McCrae 2019]. Este trabalho adota essa estratégia, combinando diferentes indicadores de saúde da economia brasileira e das contas públicas com preços de *commodities* e índices de ações.

Para a avaliação experimental das previsões, a fixação de períodos temporais na validação cruzada com normalização dos dados é a abordagem mais usual [Patel et al. 2015]. Nesse tipo de avaliação, os dados são divididos em blocos temporais de treinamento e teste, respeitando a ordem cronológica dos eventos. Outras abordagens não convencionais enfatizam a reconstrução de conjuntos de treinamento e validação para o uso em uma nova rede neural ao final de cada dia de negociação [Nelson, Pereira and De Oliveira 2017]. Neste trabalho, entretanto, a aplicação desses métodos geraria conjuntos de avaliação muito pequenos e possivelmente concentradas em períodos atípicos, inviabilizando o procedimento experimental para longos prazos.

Em relação ao tipo de tarefa, a classificação tem sido adotada para cenários de curto prazo, uma vez que conhecer a direção do movimento é suficiente para a obtenção de bons resultados [Yuan et al. 2020, Long, Lu and Cui 2019]. No entanto, em cenários de longo prazo, a tarefa de regressão é a mais adequada para entender o nível de risco envolvido do investidor e a dimensão da oportunidade de negociação.

Por fim, o coeficiente de determinação (R^2) foi a principal métrica dos trabalhos que aplicaram a tarefa de regressão na previsão de índices e outros indicadores financeiros [Bhandari et al. 2022, Lakshminarayanan and McCrae 2019].

3. Metodologia

Neste trabalho, queremos estudar o uso de variáveis exógenas para a predição do rendimento do Ibovespa. Formulamos uma metodologia em etapas para obter a previsão do índice por meio de RNAs. A Figura 1 ilustra os passos dessa metodologia para a modelagem estatística e a previsão do índice.

Inicialmente, foram coletados dados entre 2007 e 2022 referentes a diferentes índices de ações brasileiros e americanos, bem como preços de commodities, volumes de comércio exterior brasileiro, consumo de energia, saúde das contas públicas e índices inflacionários. O período considerado inclui crises econômicas e momentos de grande otimismo, incluindo a crise imobiliária nos Estados Unidos, a recessão brasileira de 2015, a reforma da previdência de 2016, a pandemia de 2020 e a recuperação dos mercados no ano seguinte.

Esse conjunto de dados inicial é composto por 19 variáveis selecionadas que formam um banco de dados consolidado com 15 anos de informações sobre o mercado financeiro brasileiro. Os indicadores de preços de commodities foram escolhidos por estarem diretamente relacionados às receitas das principais empresas negociadas na Bolsa, como Petrobrás (PETR3) e Vale (VALE3). Indicadores de comércio exterior e consumo de energia também foram escolhidos para a composição do conjunto de dados, uma vez que consistem em indicadores de atividade econômica sem a defasagem comum na divulgação do Produto Interno Bruto (PIB). Outros indicadores estão relacionados sobre a saúde de contas públicas, índices inflacionários que podem influenciar as taxas de juros na economia e índices de ações. Como esses últimos indicadores podem gerar custos para as empresas e afetar o equilíbrio de investimentos entre ações e renda fixa, eles podem trazer informações que permitam à RNA identificar as sazonalidades.

Na etapa seguinte, as variáveis do conjunto de dados foram pré-processadas. O indicador do Ibovespa foi corrigido pelo indicador de inflação IPCA. As variáveis explicativas empregadas no modelo foram normalizadas pelo escore padrão (Z-score).

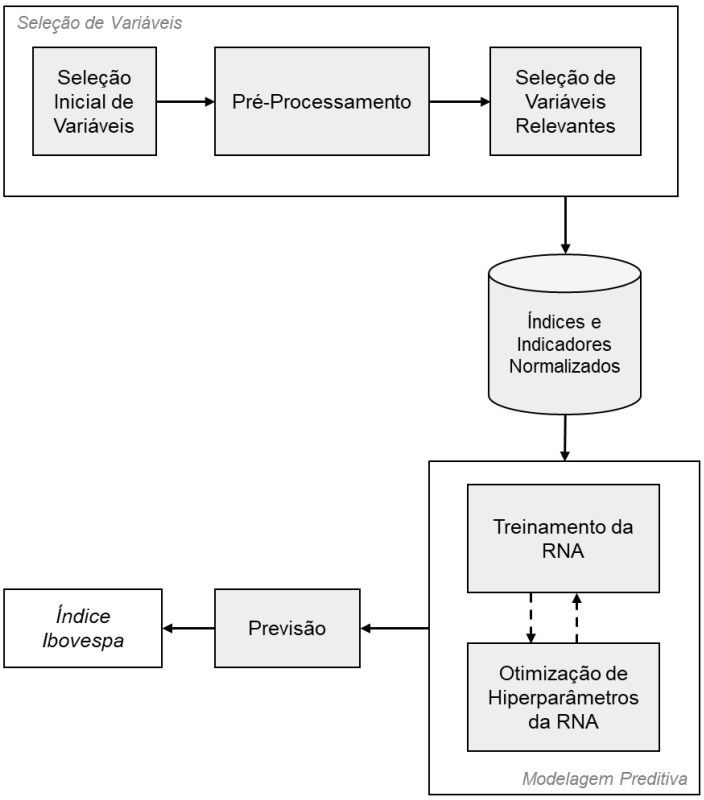

Figura 1. Metodologia para a previsão do índice Bovespa.

Em seguida, o número de variáveis foi consolidado a fim de diminuir a multicolinearidade e reduzir eventuais correlações entre dados semelhantes. Para a determinação do número ideal de variáveis, utilizamos o Fator de Inflação da Variância (VIF). Quanto menor o valor de VIF, menor será a multicolinearidade entre os atributos. Na Equação 2, o VIF da j-ésima variável corresponde ao coeficiente de determinação R^2 dessa variável em relação às demais. Quando R_j^2 é igual a zero, o valor de VIF será igual a 1.

$$VIF_j = \frac{1}{(1-R_j^2)} \tag{2}$$

O valor de R^2 é computado a partir da Equação 3, onde Y é o rendimento do Ibovespa nos 12 meses seguintes, $\bar{Y}$ é a sua média e $\hat{Y}$ é a previsão desse rendimento feita pela RNA.

$$R^2 = 1 - \left[\frac{\sum(Y-\hat{Y})^2}{\sum(Y-\bar{Y})^2}\right] \tag{3}$$

Exaustivos experimentos com métodos automáticos de seleção de variáveis sobre o conjunto de dados inicial mostraram que o melhor equilíbrio o entre VIF e R^2 considera apenas sete variáveis. A seção seguinte detalha os resultados para a seleção de variáveis relevantes.

Para a etapa de modelagem preditiva, construímos uma rede neural Perceptron de múltiplas camadas (MLP) com duas camadas ocultas totalmente conectadas de 256 e 128 neurônios, em ordem. Após cada camada oculta, incluímos uma camada de *dropout* para anular, com 10% de probabilidade p, em cada passagem *forward* do treinamento, aleatoriamente alguns elementos do tensor de entrada. O uso dessa camada também permitiu escalar as demais saídas por um fator $1/(1-p)$. Os pesos e vieses associados aos neurônios nas camadas ocultas foram selecionados aleatoriamente para os experimentos. Para o treinamento, utilizamos 300 épocas e o erro médio quadrático como função de loss. A taxa de aprendizagem da RNA é sempre 0.0001.

Na avaliação da previsão, o conjunto de dados de índices e indicadores normalizados foi dividido em duas partições disjuntas, chamadas Treinamento (Tr) e Teste (Ts). A partição Tr recebeu 75% de todos os dados. Os dados de cada partição foram selecionados aleatoriamente para cada um dos 100 experimentos que realizamos para a previsão do Ibovespa. Os experimentos são avaliados por R^2, sendo também obtidos a média e desvio padrão de cada experimento.

4. Resultados e Discussão

Esta seção apresenta e discute os resultados experimentais da nossa metodologia aplicada para a previsão do índice Bovespa utilizando variáveis exógenas e RNAs.

Durante o processo de seleção de variáveis relevantes, reduzimos o conjunto de dados inicial de 19 para sete variáveis a partir de métodos exaustivos de seleção automática. O gráfico da Figura 2 mostra a distribuição de cada variável analisada a partir dos quartis de resultados (969 combinações de 19 variáveis, tomadas 3 a 3).

Essa análise inicial indica quais variáveis devem ser escolhidas para a composição do banco de dados prévio à modelagem preditiva. A combinação [2, 7, 9, 10, 13, 16, 17] é utilizada em todos os experimentos. Com a redução no número de variáveis significativas, o valor de R^2 permanece em 0.85, com $VIF = 4.24$. Dessa forma, as variáveis selecionadas continuam explicando o modelo de previsão com mais de 85% de precisão.

Entre as variáveis selecionadas, estão indicadores de desempenho da economia, como consumo de energia e volume de importações; preços de *commodities*, como minério de ferro, gás natural e gasolina; e índices de ações brasileiro e americano. Indicadores de desempenho de contas públicas não foram selecionados para o conjunto final, indicando que esse tipo de variável possui maior relação com outros prazos de variação do Ibovespa.

Além disso, a correlação entre as variáveis selecionadas e o rendimento futuro do Ibovespa variou de 0 (correlação fraca) a 0.64 (correlação moderada). Esses resultados mostram que o parâmetro da correlação não contribui na análise de dados para a escolha de variáveis.

A metodologia também é empregada para a obtenção do rendimento futuro do Ibovespa nos 12 meses seguintes. A Figura 3 mostra os resultados para a previsão do

índice Bovespa durante o período de 15 anos analisados. O erro médio quadrático de treinamento é igual a 0.148 ± 0.014 e a mesma métrica para teste é igual a 0.208 ± 0.046. A pouca variação mostra a eficácia da metodologia na previsão do indicador para longos prazos.

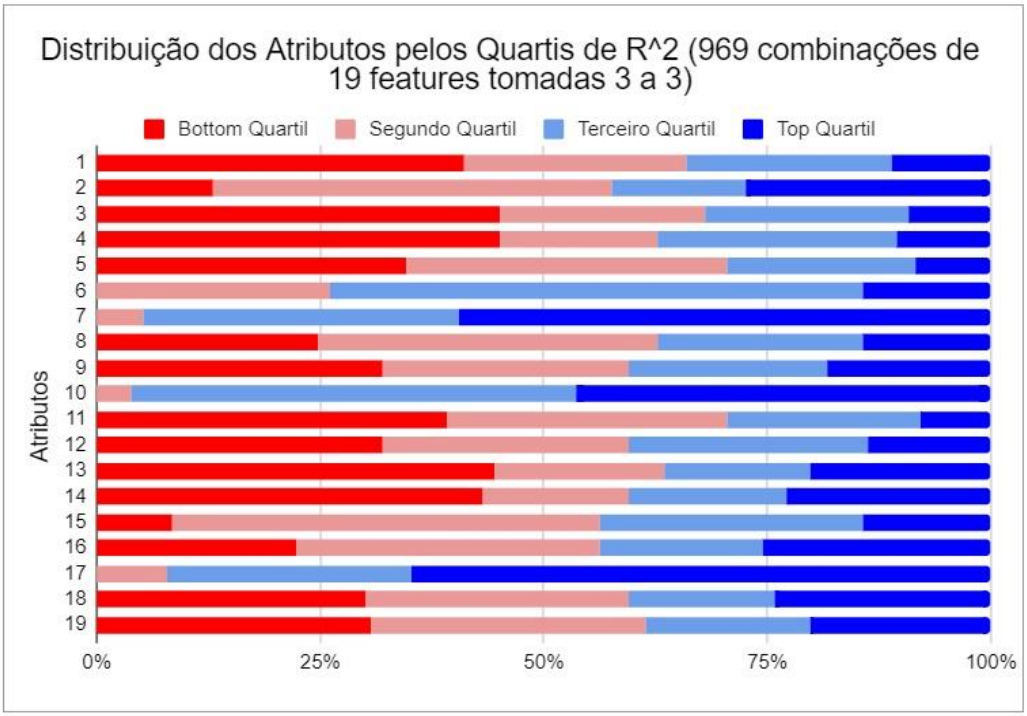

Figura 2. Distribuição R^2 das variáveis iniciais para a seleção de variáveis relevantes.

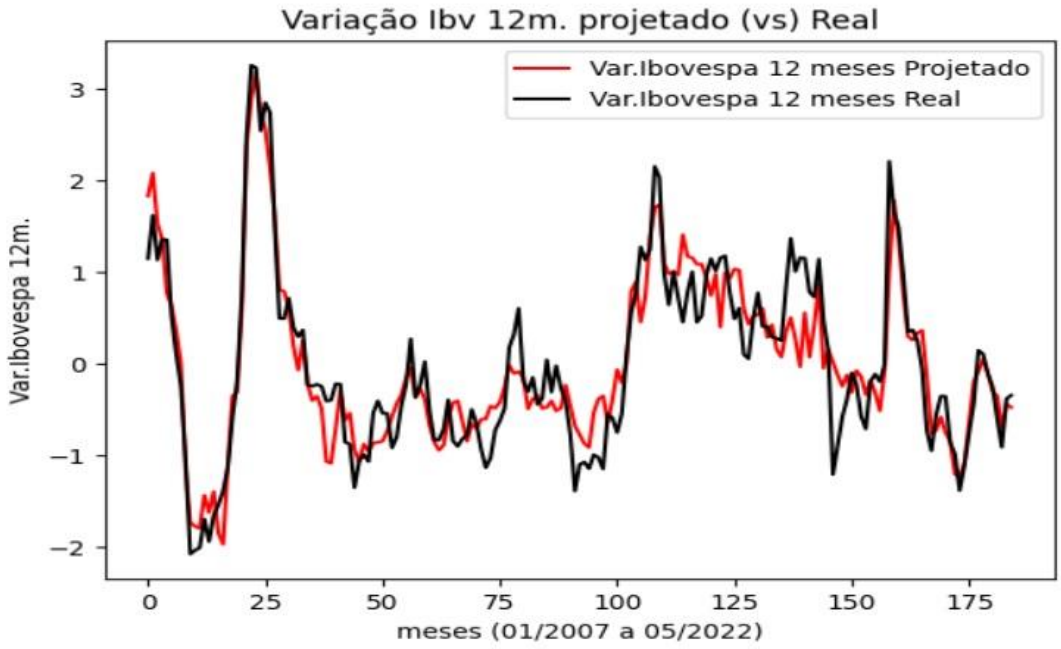

Figura 3. Previsão do índice Bovespa ao longo de 15 anos.

5. Considerações Finais

Os resultados mostram que a RNA treinada a partir da nossa metodologia é capaz de prever o desempenho futuro do Ibovespa em 12 meses, considerando ainda períodos de alta volatilidade do mercado de ações brasileiro. A metodologia confirma o potencial dos experimentos para a construção de estratégias de investimento de longo prazo.

Este estudo mostra ainda que a escolha de variáveis com base em correlação pode tornar o modelo de decisão sensível às variáveis. Variáveis explicativas sem correlação com a variável alvo também se mostram significativas para a RNA. Estudos futuros devem ampliar o número de variáveis, incluindo mais índices de ações de outros países, em especial do Sul Global.

Referências

Banas, J. and Utnik-Banas, K. (2021). Evaluating a seasonal autoregressive moving average model with an exogenous variable for short-term timber price forecasting. Forest Policy and Economics, v. 131.

Bhandari, H. N., Rimal, B., Pokhrel N. R., Rimal, R., Dahal K. R. and Khatri, R. K. C. (2022). Predicting stock market index using LSTM. *Machine Learning with Applications*.

De Campos, L. M. L. and De Figueiredo, Y. F. C. (2021). Avaliação de redes neurais profundas para a previsão de preço das ações da Petrobrás. *Revista Gestão & Tecnologia*, [S. l.], v. 21, n. 3.

Hu, Z., Zhao, Y. and Khushi, M. (2021). A survey of forex and stock price prediction using deep learning. *Applied System Innovation*.

Lakshminarayanan, S. K. and McCrae, J. P. (2019). A Comparative Study of SVM and LSTM Deep Learning Algorithms for Stock Market Prediction. *Artificial Intelligence and Cognitive Science (AICS)*.

Long, W., Lu Z. and Cui, L. (2019). Deep learning-based feature engineering for stock price movement prediction. *Knowledge-Based Systems*, v. 164.

Nelson, D. M.Q., Pereira, A. C. M. and De Oliveira, R. A. (2017). Stock market's price movement prediction with LSTM neural networks. *International Joint Conference on Neural Networks (IJCNN)*.

Olivares, K. G. et al. (2023). Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx. International Journal of Forecasting.

Patel, J., Shah, S., Thakkar, P. and Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*.

Yuan, X., Yuan J. and Ain, Q. UI (2020) Integrated long-term stock selection models based on feature selection and machine learning algorithms for China stock market. *IEEE Access*, v. 8.