Explainable AI For the Brazilian Stock Market Index: A Post-Hoc Approach to Deep Learning Models in Time-Series Forecasting

Lucas Rabelo de Araujo Morais[1], **Gabriel Arnaud de Melo Fragoso**[1],
Teresa Bernarda Ludermir[1], **Claudio Luis Alves Monteiro**[1]

[1]Informatics Center – Federal University Of Pernambuco (UFPE)
Recife – PE – Brazil

{lram2,gamf,tbl,clam}@cin.ufpe.br

Abstract. *Time-series forecasting is challenging when data lacks clear trends or seasonality, making traditional statistical models less effective. Deep Learning models, like Neural Networks, excel at capturing non-linear patterns and offer a promising alternative. The Bovespa Index (Ibovespa), a key indicator of Brazil's stock market, is volatile, leading to potential investor losses due to inaccurate forecasts and limited market insight. Neural Networks can enhance forecast accuracy, but reduce model explainability. This study aims to use Deep Learning to forecast the Ibovespa, striving to balance high forecasting accuracy with model interpretability, to improve decision-making in time-series forecasting and provide valuable insights into the economic landscape of Brazil*

1. Introduction

In the era of data, a vast amount of publicly available information is collected over time and represented as time-series data. Analyzing this type of data typically aims to capture trends, seasonality, visualize moving averages, detect outliers, or perform forecasting. Statistical models such as ARIMA or Holt-Winters have been widely used for time-series forecasting. Continuous advancements in this field have brought progress to various sectors, including finance, healthcare planning, marketing, public policies, and logistics. These advancements benefit not only developed countries but also help societies in developing countries like Brazil to manage their resources more efficiently.

However, forecasting becomes increasingly complex, requiring more sophisticated models. Deep Learning (DL) models are particularly useful in handling such complexities, though there is often a tradeoff between forecasting performance and explainability. In this challenging scenario, this work aims to use DL models to forecast the Brazilian stock market index, known as the Bovespa Index (Ibovespa). These neural networks can handle the variability and complex temporal dependencies present in the data. Additionally, this study aims to ensure the interpretability of the chosen model's results, discussing the behaviors captured by the model on both global and local scales.

Using Explainable AI (XAI) to interpret predictions of the Brazilian stock market index enhances forecast interpretability and decision-making. This approach to improve interpretability in predictions related to the Brazilian stock market was previously attempted by [Possatto 2022], who used SHAP to implement global explanations for predictions of black-box classification models of stock returns. However, this work focuses on the closing value of the Ibovespa, employing a regression approach. It aims to implement both local and global explanation methods to capture how economic features related to the Brazilian economy, such as the USD to BRL exchange rate, might affect the Bovespa Index on global and local scales. This approach enables both novice and expert investors to plan their actions according to the movement of economic features of the market and the projected values of the index.

2. Theoretical Framework

2.1. IBOVESPA

The Bovespa Index, known as Ibovespa, is a key performance indicator of stocks that aggregates the most important companies in the Brazilian capital market. Updated every four months, the index is derived from a theoretical portfolio of assets, composed of shares and units of companies listed on B3, representing approximately 80% of the number of trades and financial volume of the Brazilian capital market [A Bolsa do Brasil — B3]. In this context, the use of advanced machine learning techniques and XAI (Explainability of Artificial Intelligence) is crucial for understanding the behavior of the time series.

Currently, there are articles attempting to predict the Ibovespa using classical machine learning techniques. For instance, [Barbosa et al. 2021] compare the performance of two text classification techniques in predicting the movement of the Ibovespa, using the Naive Bayes SVM classifier and the BERT (Bidirectional Encoder Representations from Transformers) classifier, which is a neural network-based language model. Additionally, [Choinhet et al. 2021] evaluated the performance of the classification tree in predicting the future movement of the Ibovespa index, based on accumulated yield.

2.2. Neural Networks

[Alkhatib et al. 2022] evaluated the performance of CNN, LSTM and GRU based neural networks in forecasting stock price data from Apple, Tesla, ExxonMobil and Snapchat. By applying feature-engineering to create two variables HiLo and OpSe and considering the high, low, open and volume values to predict the closing price, they demonstrated that feature engineering improved the models' performance. LSTM models particularly benefited from this approach. In time-series forecasting neural networks usually have one output layer. The inputs are linearly combined into equation 1. In the hidden and output layer each neuron has an activation function that receives z_j:

$$z_j = b_j + \sum_{i=1}^{12} w_{i,j} x_i \tag{1}$$

[Shiri et al. 2023] conducted comparative analysis on deep learning models, which included the variants of the recurrent structure of neural networks (RNNs). They summarized the architecture and mathematical formulations of the LSTM, GRU and Bidirectional LSTMs models. In Long Short Term Memory (LSTM) neural networks an LSTM unit is added to the RNN to address the issue of the vanishing gradient and and to capture long-term dependencies. The LSTM model consists of three main gates, the input gate, the forget gate and the output gate. Equation 2 refers to the hidden state (or internal memory of a simple RNN) which is employed in the LSTM unit.

$$h_t = g(Wx_t + Uh_{t-1} + b) \tag{2}$$

Bidirectional LSTMs have two LSTM layers: one to process input data in forward direction and another to process it in backward direction. By processing both directions the BiLSTM model can handle information about the past and the future, better capturing temporal dependencies in the data. The Gated Recurrent Unit (GRU) model is simpler than the LSTM and combines the input and forget gate into an update gate. Neural networks have shown a superior performance in several cases of stock market forecasting when compared to other ML models [Sonkavde et al. 2023, Alkhatib et al. 2022]. In certain conditions they also outperform statistical models [Hill et al. 1994] especially when data becomes too complex, and the forecasting performance of statistical models is reduced [Jat et al. 2018]. However there is a lack of transparency since neural networks are often perceived as black-box models [Alain and Bengio 2016]. This limitation which affects their usage in daily applications, is addressed by the field of Explainable AI (XAI).

2.3. Explainable AI

Research in Explainable Artificial Intelligence (XAI) primarily focuses on providing functional or technical explanations about the working of artificial intelligence algorithms [Bryan-Kinns 2024]. In the financial sector, investment companies are increasingly turning to AI to automate processes, reduce costs, and ultimately gain a competitive advantage. In this context, accordingly to [Longo et al. 2024] the quest for XAI is driven by the need to ensure the robustness and stability of AI systems, which may be subject to extreme market conditions and unexpected events. [Martins et al. 2024] state that SHAP and LIME are the preferred and most popular explainability methods used in the financial sector. Overall, the value in SHAP and LIME explanations lies in their ability to determine feature importance, where calculations are made to ascertain the weight each feature contributes to the prediction process. Despite their similarities, SHAP is primarily used for global explanations, while LIME tends to be used for local explanations.

In the field of time series forecasting, the exploration of XAI is still in its early stages, especially as a feature selection tool [van Zyl et al. 2024]. XAI for time series data is becoming increasingly crucial in fields such as finance, healthcare, and climate science. However, evaluating the quality of explanations, such as the attributions provided by XAI techniques, remains challenging [Schlegel and Keim 2023]. This is due to the inherent chronological order of time series data and the interaction of numerous variables, often resulting in high dimensionality and complexity. Therefore, incorporating XAI could not only improve the interpretability of the model but also address these challenging characteristics of time series data [van Zyl et al. 2024].

2.4. LIME

LIME is a post-hoc agnostic XAI method, developed by Ribeiro, Singh and Guestrin [Ribeiro et al. 2016] to produce locally faithful explanations for the predictions of classifiers and regressors. Considering a set of g explanations within a space G of interpretable models, for each explanation g, LIME minimizes the following function:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \left(\mathcal{L}(f, g, \pi_x) + \Omega(g) \right) \tag{3}$$

where x represents the instance being explained, $\Omega(g)$ is a measure of complexity (such as the depth of a decision tree), f is the black-box model, with $f(x)$ represents the prediction of x, $\pi_x(z)$ is a proximity measure between an instance z to x, in order to define the locality around x and $\mathcal{L}(f, g, \pi_x)$ is a measure of unfaithfulness of g. The function $\mathcal{L}$ needs to be minimized and $\Omega(g)$ needs to be low enough to ensure both interpretability and local fidelity for the explanation.

The Loss function $\mathcal{L}$ can be calculated as:

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z) \left[f(z) - g(z') \right]^2 \tag{4}$$

The proximity measure $\pi_x(z)$ is considered to be an exponential kernel on a distance function D (which may vary depending on the type of data), where $\pi_x(z) = exp(-D(x,z)^2/\sigma^2)$ and z' are perturbed samples associated with the predictions $f(z)$, if G is the space of linear models then $g(z')$ can be assumed as $g(z') = w_g z'$.

2.5. SHAP

The main concept underlying the definition of the Shapley value is the notion of marginal contribution (**Equation 5**) [Fryer et al. 2021], which signifies the increment in the evaluation of a particular submodel upon the inclusion of a specific feature. The marginal contribution of feature i to submodel S is precisely defined as the difference in evaluation when feature i is added to submodel S.

$$M_i(S) = C(S \cup \{i\}) - C(S) \tag{5}$$

The Shapley value of feature i is determined as a weighted average across all marginal contributions of feature i [Fryer et al. 2021], specifically over $M_i(S)$ for every subset S of F that does not include i, where F delineates the overall scenario of the feature set (**Equation 6**).

$$\phi_i = \sum_{S \in 2^{(}F \setminus \{i\})} \omega(S) M_i(S) \tag{6}$$

3. Data and Materials

Python 3.10 was used to run the `yahoo finances` api to collect time-series data from the Brazilian Stock Market Index, with daily data ranging from January 1, 2007, to October 17, 2023. Eighty percent of the data was set aside for training, and 20% for validation. MinMax normalization was fitted on the training set and also applied to the validation data. The models were adjusted to forecast the closing price (figure 1). To validate the best model the predicted values were also compared to testing windows of 3, 7 and 15 days and time-series cross validation was applied to check if the difference between the models were statistically significant.

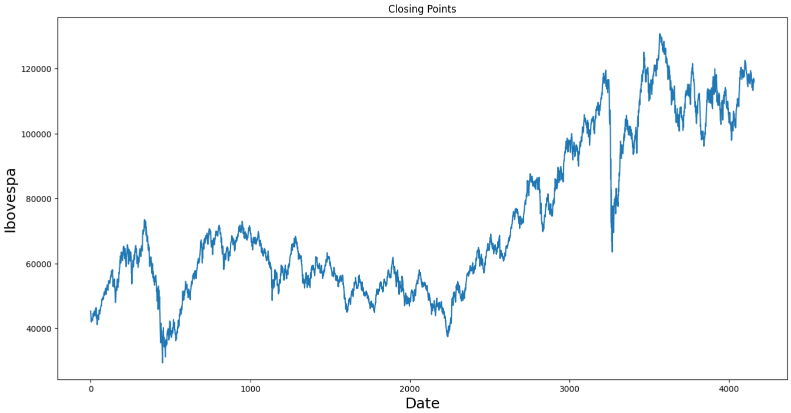

Figure 1. Ibovespa Closing

In the the autocorrelation plot (Figure 2) the last 60 days of the Ibovespa present a statistical significant correlation with the actual day. However the top 10 feature importances of the Regression Tree algorithm (Figure 3) show that up to lag 6 there is a sequential importance of lags 1, 2 ,3, 5 and 6, with the previous day (lag 1) being the most important to the forecast. This highlights the importance of temporal dependence and the recent history of the Bovespa index in predicting its future behavior. Therefore, using the lags of the last 6 days aims to capture complex temporal patterns without generating a lot of noise for the model.

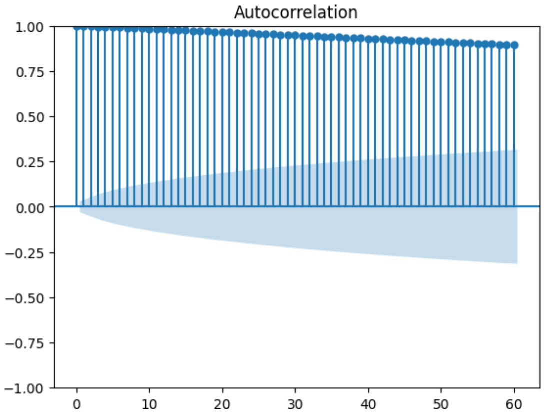

Figure 2. Autocorrelation of Previous Lags

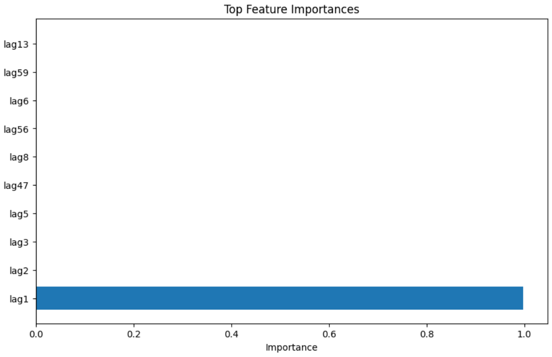

Figure 3. Feature Importance of Previous Lags

In univariate time-series, lag features of the target variable are used to forecast the target variable. However, this work encapsulates other features that might impact the prediction of the index and would be useful to know when providing an explanation for the forecast, such as the USD to BRL exchange rate, the return on the variation between the highest and lowest values of the Ibovespa in the day, and features that indicate whether the closing day is during or after a holiday or public observance day (e.g., Valentine's Day). Table 1 summarizes information about the features, including their main source and description.

Table 1. Description and meaning of variables, $i = 1, ..., 6$

Variable	Description	Source
lag$i$Close	i^{th} day lagged bovespa closing score	[Yahoo Finance API]
lag$i$USD	i^{th} day lagged USD to BRL exchange rate	[Yahoo Finance API]
lag$i$variation_low_high	i^{th} day lagged USD to BRL exchange rate	[Yahoo Finance API]
selic	Dummy variable informing the day that interest rates are defined	[Brazil's Central Bank]
4_observance	Dummy variable informing if it's an observance day (or the day after)	[`date-holidays` JavaScript Library]
4_optional	Dummy variable informing if it's an optional holiday (or the day after)	[`date-holidays` JavaScript Library]
4_public	Dummy variable informing if it's a public holiday (or the day after)	[`date-holidays` JavaScript Library]

Six-day Lagged values of the closing index (lag$i$Close), USD to BRL exchange rates (lag$i$USD) and the return of the variation between the highest and lowest index (lag$i$variation_low_high) were used as previous information to forecast the closing value for the next day. Information about holidays is not lagged since it is possible to know in advance if the target day is after an observance day or a holiday. The lag$i$variation_low_high is calculated as follows in equation 7, so that a value of 2 means that the highest value of the last ith day was twice as high as the lowest value:

$$lag(i)variation\_low\_high = \frac{high_i - low_i}{low_i} \tag{7}$$

The GRU model consists of a single GRU layer with 128 units followed by a dense layer for output. Similarly, the LSTM model comprises a single LSTM layer with 128 units and a dense output layer. The BiLSTM model utilizes a bidirectional LSTM layer with 128 units for simultaneous processing of temporal information in both directions, followed by a dense output layer. All models are compiled with the Adam optimizer and mean squared error loss function.

3.1. Error Metrics and Model Evaluation

The error metrics used to evaluate the models' forecasting performance are the *Root Mean Squared Error* (RMSE), the *Mean Absolute Scaled Error* (MASE) and the *Mean Absolute Percentage Error* (MAPE). A MASE higher than one indicates that the model performs worse than the baseline and should not be considered. More details about these error metrics and evaluation techniques, such as Time-Series Cross Validation can be found in [Hyndman and Athanasopoulos 2021].

Time-Series Cross Validation was applied to statistically validate the choice of the best model. A statistical significance level of 5% was considered when applying non-parametrical tests to identify differences among groups of explanations and DL models. Fifty folds of expanding windows were used to validate the models, and a Kruskal-Wallis test was conducted to detect statistical differences between models/explanations. A post-hoc Conover test was also applied to determine which group was statistically different [McKnight and Najab 2010, Pamplona and Jorge 2020].

3.2. Classification Models

A binary decision approach was also proposed, as explaining the prediction based on the ibovespa score may not be accessible to individuals who are not invested in the stock market. The binary answer approach switches the regression to a classification problem, where the target variable has two classes: one that indicates a increase in the BOVESPA Index compared to the previous day and another used when the opposite occurs. The Extreme Gradient Boosting algorithm (XGBoost) [Chen and Guestrin 2016] achieved the best performance among the models considered. Some hyperparameters were set in the model such as a learning rate of 0.001, 1000 gradient boosted trees and the regularization parameter gamma with a value of 0.2.

4. Results

In Figure 4, when comparing predictions it is clear that the DL models are better at capturing seasonal patterns and the trend of the time-series than the baseline model. Table 2 shows the error metrics for the validation set. MAPE reveals that the GRU, LSTM and BiLSTM models with respective percentual errors of 2.06%, 1.33% and 1.24%, have higher precision in their predictions than the baseline (9.90%), which highlights that the predictions of DL models are very close to the real values. When comparing the three metrics (RMSE, MAPE and MASE) the BiLSTM model has the lowest error metrics. The lowest MASE also reiterates that the BiLSTM is superior in predicting the validation set when compared to GRU and LSTM. Furthermore the time-series crossvalidation results in normalized $\overline{\text{RMSE}}$ for GRU (0.0046), LSTM (0.0061) and BiLSTM (0.0042), with a p-value of 0.0044 (Kruskal-Wallis test) gives statistical evidence at the level of 5% of significance to reject the hypothesis that the error metrics between the models are the same. The Conover test (results available in Table 2) using the BiLSTM model as reference, reaveals that this model is different from the others. When comparing the LSTM or GRU models, there is no statistical evidence to say that these models bring different error metrics.

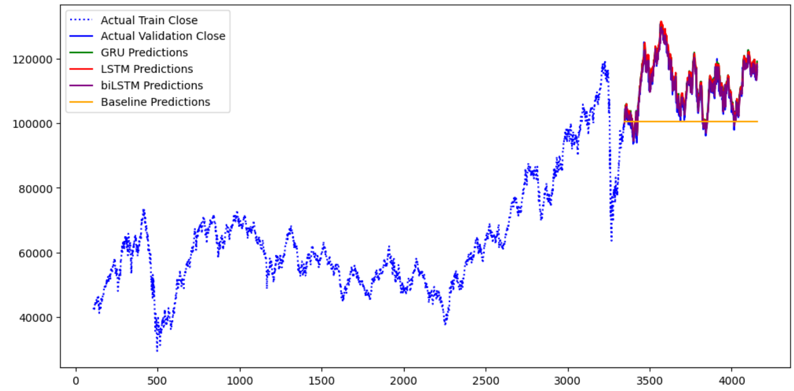

Figure 4. Predicitons in Validation Set

Model	**Validation Set**			**Cross Validation**	
	RMSE	**MAPE (%)**	**MASE**	$\overline{\text{RMSE}}$	**P-value**
GRU	2722.77	2.06	0.20	0.0046	0.0114
LSTM	1856.50	1.33	0.13	0.0061	0.0114
BiLSTM	1742.44	1.24	0.12	0.0042	1
Baseline	13599.37	9.90	1.00	-	-

Table 2. Error Metrics in Validation Set and Conover Test's P-value

For generalization to future timesteps, Table 3 for the shorth-term (3 days) and medium-term (7 days) time windows, the BiLSTM model has the best error metrics which is highlighted by the lower MASE of the model in comparison to GRU and LSTM. For the short-term window the model BiLSTM predictions are in average 1.4% distant from real values whereas GRU and LSTM are more than 2% far. The GRU model performs worse than LSTM and BiLSTM in generalization, with a MASE above 1 in medium and short-term windows, making it worse than the baseline. For long-term (15 days) predictions, the LSTM model slightly outperforms the BiLSTM. However, given the interest in forecasting only the next day (short-term forecast) and the BiLSTM's consistency across all time windows, the validation set and the time-series cross-validation statistically highlight it as the best model. Therefore, the BiLSTM was chosen to explain the predictions and how features associated with the stock market contributed to the forecast.

Model	**3 Days**			**7 Days**			**15 Days**		
	RMSE	**MAPE (%)**	**MASE**	**RMSE**	**MAPE (%)**	**MASE**	**RMSE**	**MAPE (%)**	**MASE**
GRU Model	4224.67	3.70	1.50	3455.65	2.86	1.11	2663.43	1.95	0.81
LSTM Model	2373.89	2.03	0.83	1799.14	1.41	0.55	1650.93	1.27	0.53
BiLSTM Model	1665.41	1.40	0.57	1357.29	1.07	0.42	1796.39	1.30	0.55

Table 3. Error Metrics in Different Test Windows

When looking for a global explanation of the BiLSTM model, the Shapley Values (Figure 5) reveals that the closing index for the previous day (lag 1) is the most important feature. Lower values of the index tend to be associated with lower values in the future while higher values tend to be associated with higher forecasts. Lags 2, 4 and 3 also hold significant importance and bring similar interpretations to the forecasts. The next important features are the lags of the USD to BRL exchange rate, where a lower exchange rate during the previous day (lag1USD) results in higher values for the Bovespa index. In contrast, the other lags often lead to mixed outcomes since they are not as significant as the previous day.

For the local explanation, using LIME values (Figure 7), a random day was selected showing that previous values of the closing Bovespa index had the most positive contribution to the forecast. In contrast, the USD to BRL exchange rate being higher than \$ 3.36 in lag 1 (which is in the top 10 features in the plot) contributed negatively to the forecast. However most of the previous exchange rate lags had a positive contribution. In fact exchange rate was \$5.15 in lag 6 and \$5.04 in lag 1 (Figure 6), which means a reduction of 2.14% in almost one week, indicating that the variation in the previous day may have had a negative impact in the forecasts. Three other local explanations were generated with random seeds, but the Kruskal test didn't give statistical evidence to state that the values provided by the explanations were different (p-value=0.9597).

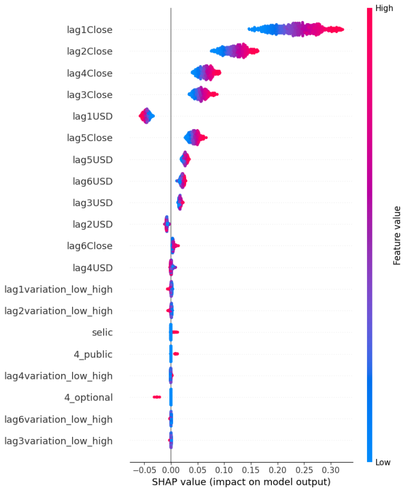

Figure 5. SHAP Beeswarm Plot Of BiLSTM Model

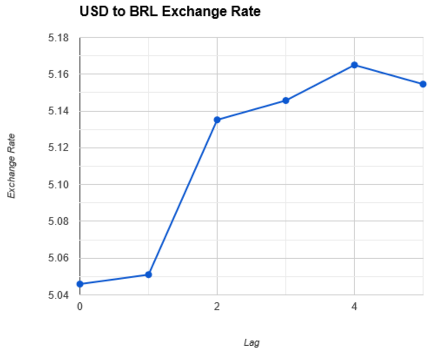

Figure 6. USD To BRL Lag Values

5. Discussion

The XGBOOST classifier achieved 53% of accuracy; however, some studies have reported achieving at least 88% Directional Accuracy (DA) in predicting specific stocks [Alzaman 2023]. Nonetheless, since the focus of this work is not on building a classifier model and the stock market as a whole may encapsulate more noise than specific stocks that are tied to specific fields of the market sectors, Shapley values were applied (Figure 8). This approach for the stock market may be more feasible for users who are not well-acquainted with it, as it is easier to understand whether the market went up or down and to explain the features that may have contributed for the prediction.

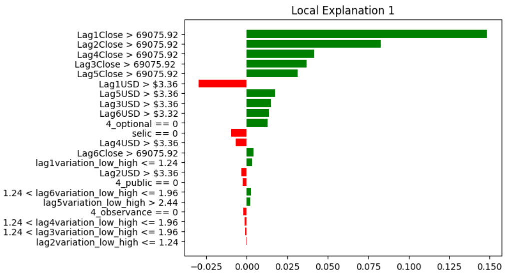

Figure 7. Local Explanation with LIME Values

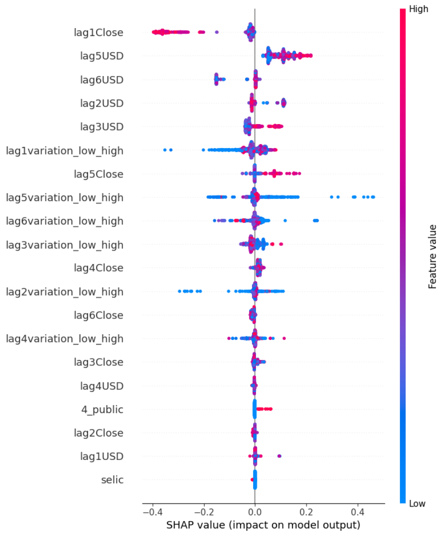

Figure 8. SHAP Beeswarm Plot Of XGBoost Model

The Shap and LIME explanations (Figures 5 and 7) of the BiLSTM model (regression approach) help to shed light on the model's black-box nature. Furthermore the behaviour associated with the importance of the USD to BRL exchange rate in the global explanation highlights a pattern that can also be verified in the work of [Tabak 2006], where the author suggests a causal pattern between exchange rates and stock prices in the Brazilian market. The BiLSTM model achieved the best forecasting performance when compared to the other considered models. [Siami-Namini et al. 2019] recommend using BiLSTM instead of LSTM for forecasting problems in time series, arguing that some additional features not captured by LSTM might be captured by BiLSTM.

6. Conclusion

The emphasis on interpretability of the XGBoost and BiLSTM models for financial time series is a significant contribution. Traditionally, these models are considered "black boxes" due to their complexity. This study progresses by applying agnostic model methods to elucidate the internal decision-making of these models on time series data. Such work is rarely found in the literature and in public code repositories. With this contribution, the study becomes one of the few to publicly provide an approach for SHAP and LIME in time series, accessible in the GitHub repository at this link [de Araujo Morais 2024].

Overall, the findings highlights the efficacy of DL models, particularly BiLSTM, in accurately forecasting the Ibovespa index. The study advocates for the usage of post-hoc explainability, by changing the perspective of the problem from a time-series forecasting to a binary classification. It emphasizes that explanations should be accessible to stakeholders from all fields, in order to offer valuable insights when navigating the complexities of the stock market.

It is important to keep in mind that the interpretation of LIME and SHAP for time series is an approximation and may not fully capture temporal complexity. This work does not apply Temporal Fusion Transformers, a state-of-art solution that uses attention to provide interpretability of deep learning models for time-series forecasting [Lim et al. 2021]. The main reason for this is that the model is not simple to apply, as it requires a different approach to pre-processing the time-series data. However in future research and extension, Temporal Fusion Transformers and other state-of-art solutions shall be explored.

References

A Bolsa do Brasil — B3. B3. `https://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/ibovespa.htm`. Last accessed on 2024-03-01.

Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv:1610.01644 [stat.ML].*

Alkhatib, K., Khazaleh, H., Alkhazaleh, H. A., Alsoud, A. R., and Abualigah, L. (2022). A new stock price forecasting method using active deep learning approach. *Journal of Open Innovation: Technology, Market, and Complexity*, 8(2):96.

Alzaman, C. (2023). Forecasting and optimization stock predictions: Varying asset profile, time window, and hyperparameter factors. *Syst Soft Comput*, 5:200052.

Barbosa, B. A., Marcacini, R. M., and Rezende, S. O. (2021). Predição do movimento do índice ibovespa a partir de notícias. In *Workshop de Matemática, Estatística e Computação Aplicadas à Indústria - WMECAI*. ICMC-USP.

Brazil's Central Bank. Brazil's Central Bank Interest Rates. *Brazilian Central Bank*. `https://www.bcb.gov.br/controleinflacao/historicotaxasjuros`, last accessed on 2024-02-25.

Bryan-Kinns, N. (2024). Reflections on explainable ai for the arts (xaixarts). *Interactions*, 31(1):43–47.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794. ACM.

Choinhet, R., Schmidt, C. E., and Chies, L. (2021). Aplicação da árvore de classificação na predição do movimento do índice ibovespa. *Sociedade Brasileira de Automática*, 1(1):512–517.

`date-holidays` JavaScript Library. `date-holidays`. *GitHub*, Commenthol. `https://github.com/commenthol/date-holidays`, last accessed on 2024-02-25.

de Araujo Morais, L. R. (2024). Forecasting paper xai. Available at: `https://github.com/marreapato/Forecasting_Paper_XAI`, last accessed on 2024-10-01.

Fryer, D., Strümke, I., and Nguyen, H. (2021). Shapley values for feature selection: The good, the bad, and the axioms. *IEEE Access*, 9:144352–144360.

Hill, T., Marquez, L., O'Connor, M., and Remus, W. (1994). Artificial neural network models for forecasting and decision making. *International Journal of Forecasting*, 10(1):5–15.

Hyndman, R. J. and Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 3rd edition.

Jat, D. S., Dhaka, P., and Limbo, A. (2018). Applications of statistical techniques and artificial neural networks: A review. *Journal of Statistics and Management Systems*, 21(4):639–645.

Lim, B., Arık, S. , Loeff, N., and Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764.

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., and Stumpf, S. (2024). Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301.

Martins, T., Almeida, A. M. d., Cardoso, E., and Nunes, L. (2024). Explainable artificial intelligence (xai): A systematic literature review on taxonomies and applications in finance. *IEEE Access*, 12:618–629.

McKnight, P. E. and Najab, J. (2010). Kruskal-wallis test. In *The Corsini Encyclopedia of Psychology*, volume 1, pages 1–10. Wiley.

Pamplona, D. and Jorge, C. (2020). An overview of air delay: A case study of the brazilian scenario. *Transportation Research Interdisciplinary Perspectives*, 7:1–13.

Possatto, A. B. (2022). Painting the black box white: Interpreting an algorithm-based trading strategy. *Revista Brasileira de Finanças*, 20(3):105–138.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. *arXiv:1602.04938 [cs.LG]*.

Schlegel, U. and Keim, D. A. (2023). A deep dive into perturbations as evaluation technique for time series xai. In Longo, L., editor, *Explainable Artificial Intelligence*, pages 165–180, Cham. Springer Nature Switzerland.

Shiri, F. M., Perumal, T., Mustapha, N., and Mohamed, R. (2023). A comprehensive overview and comparative analysis on deep learning models: Cnn, rnn, lstm, gru. *arXiv:2305.17473*.

Siami-Namini, S., Tavakoli, N., and Siami-Namini, A. (2019). A comparative analysis of forecasting financial time series using arima, lstm, and bilstm. *arXiv:1911.09512*.

Sonkavde, G., Dharrao, D. S., Bongale, A. M., Deokate, S. T., Doreswamy, D., and Bhat, S. K. (2023). Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications. *International Journal of Financial Studies*, 11(3):94.

Tabak, B. M. (2006). The dynamic relationship between stock prices and exchange rates: Evidence for brazil. *International Journal of Theoretical and Applied Finance*, 9(8):1377–1396.

van Zyl, C., Ye, X., and Naidoo, R. (2024). Harnessing explainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of grad-cam and shap. *Applied Energy*, 353:122079.

Yahoo Finance API. Yahoo Finance. `yfinance`. *PyPI*, `https://pypi.org/project/yfinance/`, last accessed on 2024-02-25.