*Proceedings of the 5th International Conference on Statistics: Theory and Applications (ICSTA'23)*
*Brunel University, London, United Kingdom – August 03-05, 2023*
*Paper No. 158*
*DOI: 10.11159/icsta23.158*

# Short and Long-term Forecasting of Emerging Market Data using ARIMA-based and Boosting Machine Learning Algorithms

**Md Al Masum Bhuiyan[1], Constanza Zurita-Valdebenito[1], Md Shamsul Alam[2], Nusrat Sarmin[1]**
[1]Austin Peay State University
601 College St, Clarksville, United States
bhuiyanm@apsu.edu; czuritavaldebenito@my.apsu.edu; nsarmin@my.apsu.edu
[2]The University of Texas at El Paso
500 W University Avenue, El Paso, United States
malam@my.utep.edu

***Abstract*** **-** In light of financial democratization brought about by economic growth, understanding emerging market dynamics benefits retail investors, industries, and governments in monitoring their economies. To better understand the complexity of the market movements, this work focuses on the autoregressive integrated moving average (ARIMA) method and Boosting algorithms in forecasting emerging stock market data. It is observed that the financial time series in emerging countries present a complex forecasting route due to volatility conditions, seasonality, trends, high-frequency, and non-stationary series conditions [1]. Thus, conventional statistical approaches show considerable uncertainty in both short and long-term predictions; that is the case with applying ARIMA-based models. So, the objective of this study is to predict a specific highly volatile emerging market stock using algorithms that boost uncertain conditions with the support of appropriate covariates, such as inflation and the consumer price index (CPI). In short- and long-term predictions, a performance comparison will be made among techniques such as ARIMA-based, Extreme Boosting-based (XGBoost), and Light Gradient Boosting Machine-based (LightGBM) algorithms. Since the data volatility could be captured by tuning appropriate Boosting hyperparameters and by selecting appropriate financial indicators, these algorithms can be categorized as practical techniques for determining economic growth dynamics in emerging markets.

***Keywords*****:** Emerging Market Data, High-Frequency, ARIMA, XGBoost, LightGBM, Forecasting, Covariates, Volatility.

## 1. Introduction

During the last decade, modeling emerging market data has become an intriguing phenomenon for the financial community. Emergent economies have experienced significant expansion because of several phenomena associated with globalization, rapid economic growth, and open economy policies [2]. Indeed, stock market liberalization in emerging economies worldwide has contributed to investment growth by reducing capital costs [3] [4]. This economic scenario has motivated many decision-makers worldwide about the importance of understanding stock market behaviors. Certainly, forecasting stock markets with high levels of accuracy is very important for these investors to minimize associated risks as much as possible. Several techniques have been applied when forecasting financial time series for short-term and long-term conditions. However, when dealing with high-frequency datasets that present conditions of uncertainty, the complexity of forecasting them increases.

The GBDT and LightGBM algorithms were used by optimizing feature selection methods to predict stock prices of S&P500 index [5]. A comparison was analyzed between ARIMA-based and LSTM-based models' performance when forecasting stock prices. The authors explained the advantages and disadvantages of these techniques, emphasizing that more complex neural networks can help to make better forecasting results [6]. A research study used the Istanbul Index to forecast its stock prices by applying simple ARIMA-based models. The results showed that the classic time series approach did not capture fluctuations and unique trends associated with each forecasting period [7]. Three algorithms, ARIMA, XGBoost, and LSTM, were applied in a study to capture the volatility aspects present in Amazon's stock market prices [8]. In 2022, a study concluded that the XGBoost is an efficient technique for forecasting short-term stock prices since it can capture the nonlinear dynamics and associated trends [9]. In 2023, authors argued that an ARIMA-based model is also an efficient tool to forecast short-term stock prices [10].

This study aims to develop two Boosting-based type methods: XGBoost and LightGBM for predicting the short-long-term stock prices of the Brazil Agro's market. In addition, ARIMA-based models will be built for forecasting the short-term stock prices of high frequency markets belonging to specific sectors in Argentina, Brazil, Mexico, and Poland Indices. The XGBoost technique generally offers a parallel boosting that accurately resolves various time series issues, such as trends and non-linearity conditions. However, computing the entire network depends on heavy memory space. On the contrary, the LightGBM approach uses less memory when computing the algorithm. Compared with traditional XGBoost, the LightGBM technique demonstrates faster operability and performance [11] [12] [15].

The present study is outlined below; section 2 describes the methodology for fitting the best models. The ARIMA and Boosting-based algorithms are discussed here. Section 3 deals with the information-related background of the data. During this stage, some essential emergent market data properties are described. Section 4 provides the results after applying the models for the datasets. A discussion is made to compare the models' performance. Finally, section 5 contains the conclusion and suitability of the study.

## 2. Methodology

### 2.1. ARIMA-based Models

The ARMA was created to measure the correlation of variables assuming a linear association of lags. These models cannot be applied to non-stationary data since they show mean and variance correlation over time. For that reason, ARIMA model can be used by applying classic regression tools and differencing methods. The ARIMA model utilizes previous information of a particular variable $x$ to forecast the observation at time *t*. The filtering of dynamic characteristics corresponding to trends, seasonality, cycles, and fluctuations that belong to the non-stationary time series condition is obtained by applying the Box-Jenkins technique. Therefore, the method reduces the mean and variance correlation, impacting in a model that behaves stationary. The ARIMA (p,d,q) model considers orders related to the autoregressive (AR) ($p$) and the moving average (MA) ($q$) terms. This model also considers a $d$-order, corresponding to the number of times the differencing was applied to obtain a stationary time series. Generally, the AR term corresponds to past steps used to predict the current value of $x$, the MA term is the obtained errors when regressing the series, and the differencing term is the order of non-seasonal differences applied to the model when reducing the dynamic characteristics [13]. The model can be expressed as follow:

$$\phi(B)\,(1 - B)^d x_t = \theta(B) w_t \tag{1}$$

From the above equation, the $B$-term corresponds to the backshift operator, and the $\phi(B)$, $x_t$, $\theta(B)$, and $w_t$ terms represent the autoregressive operator, current value, the moving average operator, and the residual parameter, respectively. The differencing parameter (*d)* reduces the high correlation of mean and variance metrics when obtaining a stationary time series condition. Based on the stationary of the model, the $q$ and $p$ order are selected from the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). Fitted models can be chosen based on a minimum Akaike Information Criterion (AIC) value [14].

### 2.2. Boosting Algorithms

Boosting algorithms perform based on decision trees. The *base learners* are the initial decision trees, thus, the weakest models of the entire network. The *training weak models* are the subsequent decision trees in the network; their residuals are lower than the previous trees; here, the model improves the performance if compared with the past. During a new training process, this model takes the highest error rates from previous steps. Finally, the *sequential training with respect to errors* is the stage where the tree increases in size enough to reduce all previous residuals to the most [12].

### 2.2.1. XGBoost-based Model

The XGBoost is a flexible and scalable algorithm that includes a penalization into the loss function that prevents overfitting conditions. The additive technique used in the XGBoost learning process enhances the model's performance. Furthermore, current trees use past errors to optimize the objective function. Compared to the LightGBM algorithm, the XGBoost behaves slower and presents a high-speed consumption during training [15]. The XGBoost structure contemplates a tree-based disposition with a level-wise arrangement [11]. This algorithm utilizes Tyler's expansion in the cost function among a second derivate [8]. Eq. (2) and (3) mathematically describes the objective function and regularization term. These expressions include the terms I, constant, $\gamma$, and $\lambda$ as the loss, constant, and customization parameters, respectively.

$$obj^{(t)} = \sum_{i=1}^{n} l[y_i, \hat{y}_i^{t-1} + f_t(x_i)] + \Omega(f_t) + constant \tag{2}$$

$$\Omega(f_t) = \Upsilon * T_t + \lambda \frac{1}{2}\sum_{j=1}^{T} w_j^2 \tag{3}$$

The second-order Taylor expansion is applied to Eq. (2), producing an updated objective function.

$$obj^{(t)} = R + \Omega(f_t) + constant \tag{4}$$

The R term from Eq. (4) is mathematically described as detailed in Eq. (5):

$$\sum_{i=1}^{n}[l(y_i, \hat{y}_i^{t-1}) + g_i f_i(x_i) + \frac{1}{2} h_i {f_t}^2(x_i)] \tag{5}$$

The above equation corresponds to the first and second derivatives of g and h terms are described as follows:

$$g_i = \partial_{\hat{y}_i^{t-1}} * l(y_i, \hat{y}_i^{t-1}), \qquad h_i = \partial^2_{\hat{y}_i^{t-1}} * l(y_i, \hat{y}_i^{t-1}) \tag{6}$$

The loss and weight functions are acquired by deriving the final equation. The final equation is obtained by replacing Eq. (3) and Eq. (6) with Eq. (4). Differencing the loss and weight expressions are displayed as follows:

$$w_j^* = -\frac{\sum g_i}{\sum h_i + \lambda}, \qquad obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum g_i)^2}{\sum h_i + \lambda} + \gamma * T \tag{7}$$

### 2.2.2. LightGBM-based Model

The LightGBM is an optimized-boosting, flexible, and scalable algorithm. Compared to the XGBoost algorithm, the LightGBM is faster and presents low-speed consumption during training [15]. Similarly, to the XGBoost, the LightGBM network uses iterative past errors in the training process to enhance the model's performance. The LightGBM algorithm has a leaf-wise disposition, where the trees grow vertically and horizontally [5]. This algorithm utilizes the optimized objective function's error function and the regular term. Eq. (10) mathematically expresses the iterative function [5]:

$$F_m(X) = \sum_{m=0}^{M} f_m(X) \tag{8}$$

Where the objective function is mathematically described as: $J = \sum_i L[y_i, F_m(X)] + \sum_k \Omega(f_k)$. The loss function and regular terms are represented by the parameters $L$ and $\Omega$. For details of the algorithm, we refer the reader to [15].

## 3. Data Background

Sampling stock market datasets from Argentina, Brazil, Mexico, and Poland indices were used from the Yahoo! Finance repository. The daily observations from Brazil and Brazil-Agro markets were obtained between 2018 - 2022 and 2015 - 2019, respectively. In general, the markets belong to the S&P Merval Index (MERVAL), the Brazil BOVESPA Index (IBOVESPA), Mexican Stock Exchange (BMV), and Warsaw Stock Exchange General Index (WIG). Samples include stock price attributes related to open, high, low, close, and volume trading prices in each period. For this study, the opening and closing prices are the historical indicators to be forecasted. Besides, the daily returns, daily, monthly, and annual volatility measurements are calculated during this research to evaluate the stock price changes and dispersion for all markets. Also, the financial covariates to be included in the Boosting-based models are inflation and the Consumer Price Index (CPI). The goal is to explain the dynamic behavior of certain emergent markets through these features. The volatility metrics (daily, monthly, and annual) are obtained using the measure of daily spread, adding the trading periods [16]. The CPI is the main change in consumer-based prices related to certain goods and services over time [17]. A data cleaning is computed due to the presence of missing values in the datasets.

Figure 1 illustrates the trend and volatility aspects for all stock markets. All markets present noticeable trends when analyzing the time series components. Argentina and Poland's stock markets show a combination of two tendencies; meaning that the stock price presents upward and downward combined trends. For Brazil, Brazil-Agro, and Mexico stock markets, an upward tendency is more pronounced for each corresponding stock price. In terms of volatility, all markets present fluctuations over time, with very pronounced peaks. Financial daily returns from all stock markets were examined in this stage. The financial crashes are evident in the large spikes and the evidence of dramatic volatility changes within short time intervals in the series.

Regarding the ACF results, the prices in all stock markets tend to be highly correlated with time and there are positive autocorrelation between a particular observation and its lagging over time. Indeed, these observations are directly related to the changes of mean and variance over time. In the case of the PACF results, all models present markable correlated spikes (significant non-zero autocorrelation) at lag one; however, when increasing the lag number, a combination of correlated and non-correlated (significant near to zero autocorrelation) spikes is noticeable. With these preliminary observations, models suggest non-stationary time series conditions.

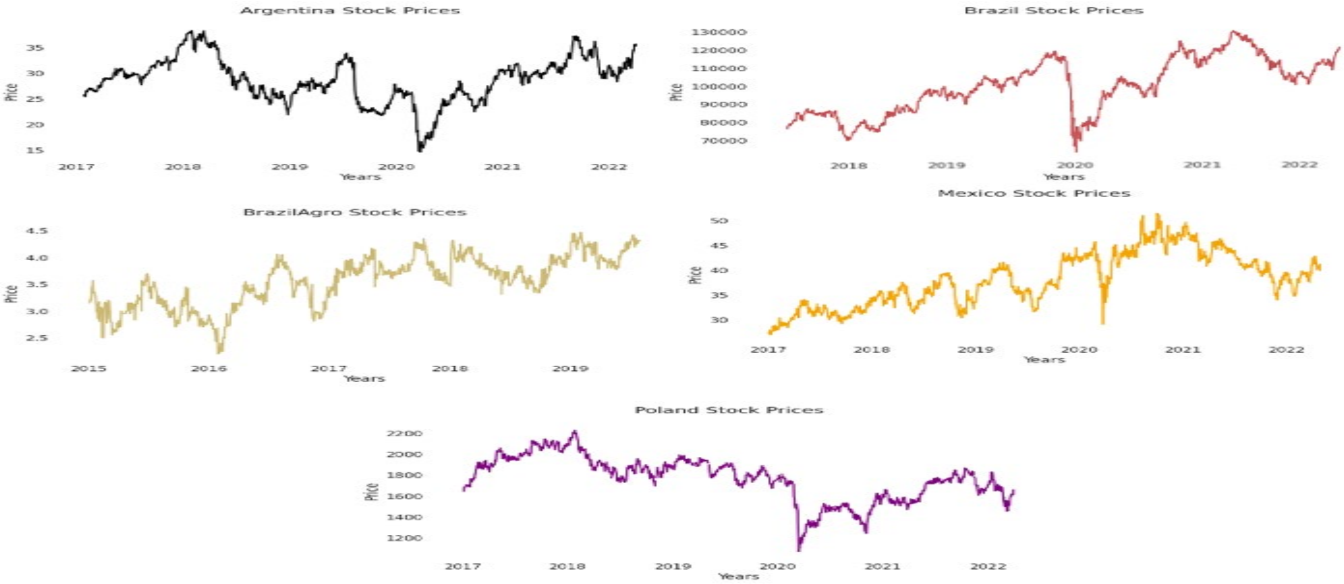

Fig. 1: 2017 – 2022 Stock Prices for Argentina, Brazil, Brazil-Agro, Mexico, and Poland Markets.

# 4. Results

## 4.1. ARIMA-based Model

The Argentina, Brazil, Mexico, and Poland stock markets are analyzed using ARIMA-based models to forecast their opening stock prices. The trend and seasonality characteristics for each time series were reduced by applying a logarithm transformation and a differencing method to make the data stationary. Table 1 presents the stationary test results, where the the null hypothesis states that the model is non-stationary, while the alternative rejects this condition. The rejection is based on a confidence level of 95 %. The test results for each market are detailed in Table 1.

Table 1: Dickey-Fuller (ADF) Test Results for the Stock Markets.

| Stock Market | p-value (before transformation) | p-value (after transformation) |
|---|---|---|
| Argentina | 0.28 | 0.00 |
| Brazil | 0.31 | $9.58x10^{-17}$ |
| Mexico | 0.10 | $2.51x10^{-18}$ |
| Poland | 0.24 | $5.10x10^{-13}$ |

To avoid the over-difference condition, the $d$ selected term for all stock markets is 1. This selection is determined by the p-values lower than the significance level of 0.05 shown in Table 1. The $q$ and $p$ were selected based on the ACF and PACF results. The orders correspond to the lags located over the significance level. The best-fitted models are selected based on the minimum AIC values shown in Table 2. The Ljung-Box test is used to verify the residuals white noise condition. The null hypothesis states that the residuals are uncorrelated, while the alternative rejects it. The test results demonstrated that the residuals for all stock markets are white noise with a 95 % confidence.

Table 2: Selected ARIMA Models for Argentina, Brazil, Mexico, and Poland Markets.

| Stock Market | Model | AIC |
|---|---|---|
| Argentina | ARIMA (9,1,9) | 1882.25 |
| Brazil | ARIMA (7,1,7) | 18502.98 |
| Mexico | ARIMA (7,1,7) | 2946.46 |
| Poland | ARIMA (6,1,6) | 11890.18 |

The datasets were split into two samples during the cross-validation section; the training set corresponds to 80 %, and the testing set corresponds to the remaining 20 %. The opening stock prices were forecasted for 15 and 60 days, respectively. Visual representative cases (Argentina and Brazil markets) for the short-term forecasting are shown in Figure 2. The root mean square error (RMSE) is also calculated to evaluate each model's performance. Table 3 displays these values.

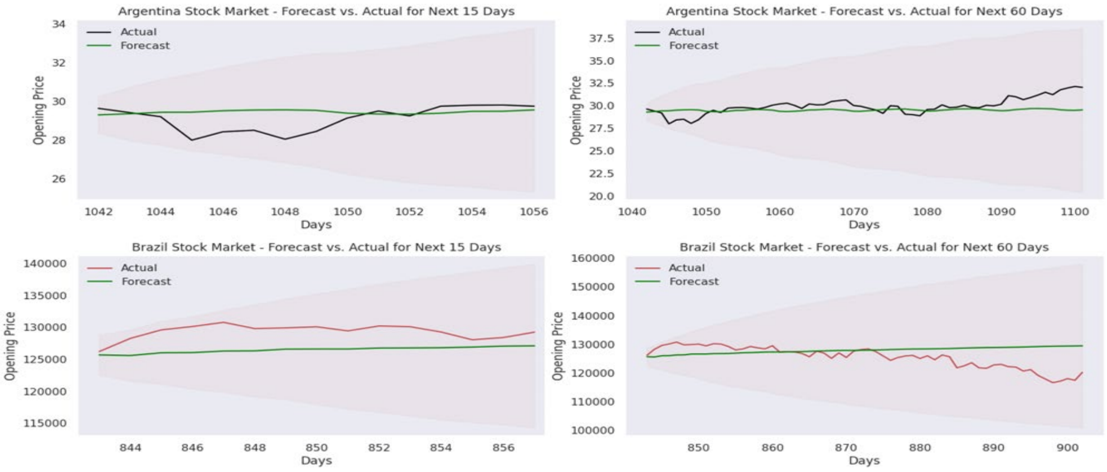

Fig. 2: Forecasting Results for the Next 15 and 60 Days from the Argentina and Brazil Markets.

Table 3: RMSE Forecasting Results for the ARIMA-based Models

| Stock Market | Model | 15 Days Forecasting | 60 Days Forecasting |
|---|---|---|---|
| Argentina | ARIMA (9,1,9) | 2.57 % | 3.28 % |
| Brazil | ARIMA (7,1,7) | 2.33 % | 4.18 % |
| Mexico | ARIMA (7,1,7) | 5.80 % | 5.83 % |
| Poland | ARIMA (6,1,6) | 2.12 % | 6.51 % |

Based on above results, Brazil, and Poland markets present a remarkable increment of the RMSE when forecasting 60 instead of 15 following days. For the Mexico, the RMSE smoothly increases when forecasting for the next 60 days compared to the next 15 days. In the case of the Argentina, the RMSE increases when incrementing the forecasting at 45 days. Overall, the Argentinian model presents the best forecasting performance for both periods. Figure 2 displays two market's short-term predictions (15 and 60 days) when applying the classic ARIMA-based model. Here the forecasted opening prices are plotted alongside the actuals. The Argentinian model well-captured the trend. For this model, the forecasted line's slope presented a similar inclination to the actual prices. This slope difference suggests that the Argentinian ARIMA (9,1,9) model could capture the trend well. However, the forecasted line from the Argentinian model presented null fluctuations (volatility aspects) compared to the actual values.

### 4.2. Boosting-based Model

In order to understand the trend and volatile behavior stock markets present, it was essential to create financial indicators, such as daily returns and daily, monthly, and annual volatility. Table 4 presents these metrics' results for two specific markets: Brazil and Brazil-Agro. In both markets, the volatility tends to increment for daily, monthly, and annual periods. However, a significant difference is more noticeable in the Brazil-Agro compared to the Brazil market. For this study, it is more interesting to analyze a stock with high levels of this condition. Therefore, based on the highest volatility values shown by the Brazil-Agro market, this stock is analyzed by applying two boosting techniques. The inflation and CPI covariates are gauged during the feature engineering section. An association analysis evaluates the strong relation of these covariates with the closing price when applying the Boosting regression, resulting in high levels of correlation with the target variable to be forecasted.

Table 4: Volatility Values for the Brazil and Brazil-Agro Markets.

| Stock Market | Daily Volatility | Monthly Volatility | Annual Volatility |
|---|---|---|---|
| Brazil | 1.76 % | 8.06 % | 27.94 % |
| Brazil - Agro | 3.04 % | 13.94 % | 48.31 % |

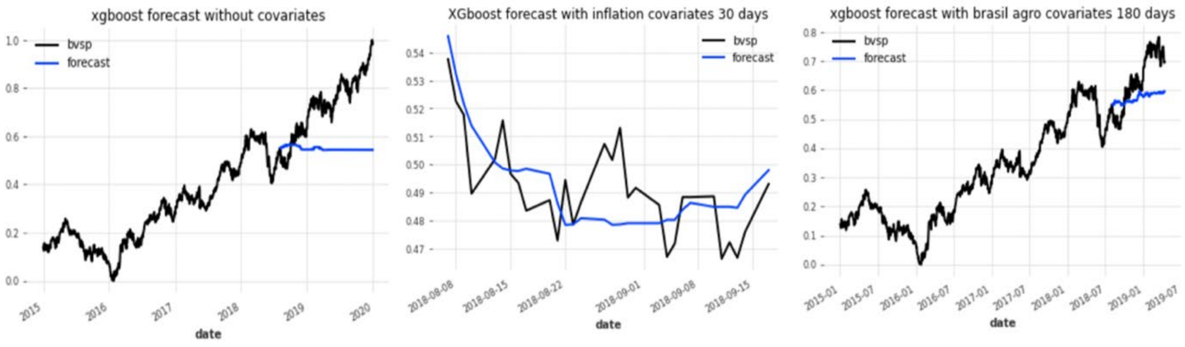

Fig. 3: Forecasting Results for the Next 30 and 180 Days from the Brazil-Agro Market Using XGBoost-based Algorithm.

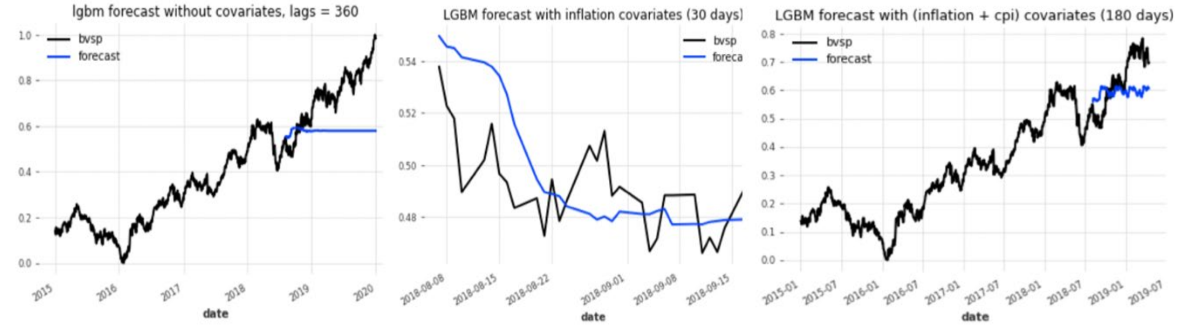

Fig. 4: Forecasting Results for the Next 30 and 180 Days from the Brazil-Agro Market Using LightGBM-based Algorithm.

Table 5: RMSE and MAPE Forecasting Results for the Boosting-based Models

| Features | RMSE (XGBoost) | RMSE (LightGBM) | MAPE (XGBoost) | MAPE (LightGBM) |
|---|---|---|---|---|
| Without Inflation Covariates | 0.22 % | 0.19 % | 24.83 % | 21.38 % |
| With Inflation Covariates (30 Days) | 0.01 % | 0.02 % | 2.32 % | 3.51 % |
| With Inflation Covariates + CPI (180 Days) | 0.10 % | 0.10 % | 13.19 % | 12.72 % |

Comparing RMSE values of the XGBoost model for the Brazil-Agro stock market price prediction, the results reveal that those considering inflation covariates and CPI perform better than models without considering them. The MAPE also significantly improves when these covariates are considered in the algorithm, with very good performances for the short-long-term forecasting. The XGBoost technique efficiently predicted the closing prices from the Brazil-Agro stock market, capturing the volatility and associated trend very well. When considering inflation covariates and CPI, the RMSE values for the LightGBM models are lower than those that did not consider these financial features. Thus, adding these covariates enhanced the performance of the studied models. The MAPE values also show that the models that consider inflation covariates and CPI outperformed very well compared to those that did not. These results indicate that the LightGBM technique is highly effective for short-long-term stock price predictions as it captures the crashes produced by fluctuations and the trend direction. Overall, the inclusion of inflation covariates and CPI impacted the model's performance in both algorithms positively.

## 5. Conclusion

In this study, we used ARIMA-based, XGBoost, and Light-GBM boosting methods when forecasting short-long-term prices of specific stock markets with high-frequency and volatility conditions. The results for the simple ARIMA-based models indicated that the models generally had better performances for the short-term prediction than for the long-term. When using the ARIMA-based models, the volatile stock price's dynamic could not be captured well by them when forecasting the short-long term. This behavior is explained by the regression method used when building the ARIMA time series. Since the ARIMA-based models are built by enforcing a non-stationary and linear condition, and not considering aspects of dynamics (volatility), this highly impacts an underfitting condition on stock prices. In the case of the Boosting models, it is clear that including inflation covariates, the CPI indicator, and lags helped the models learn better the market patterns (trend direction, volatility, and crashes) during the training process. The reason is that the decision tree algorithm used by Boosting continuously learns from past errors at each iteration, thus minimizing these residuals. Therefore, the Boosting algorithm can reduce the underfitting modeling condition for emerging market price predictions if compared with what is demonstrated by the ARIMA-based modeling cases.

## References

[1] M. A. M. Bhuiyan, "Predicting stochastic volatility for extreme fluctuations in high frequency time series," The University of Texas at El Paso, 2020.

[2] K. A. El-Wassal, "Understanding the growth in emerging stock markets," *Journal of Emerging Market Finance,* vol. 4, no. 3, pp. 227-261, 2005.

[3] A. Chari and P. B. Henry, "Firm-specific information and the efficiency of investment," *Journal of Financial Economics,* vol. 87, no. 3, pp. 636-655, 2008.

[4] N. Gupta and K. Yuan, "On the growth effect of stock market liberalizations," *The Review of Financial Studies,* vol. 22, no. 11, p. 4715–4752, 2009.

[5] M. Shangchen, "Predicting the SP500 Index Trend Based on GBDT and LightGBM Methods," in *In E3S Web of Conferences*, Nanjing, 2000.

[6] P. Liu, "Time Series Forecasting Based on ARIMA and LSTM," in *In 2022 2nd International Conference on Enterprise Management and Economic Development (ICEMED 2022)*, Dalian, 2022.

[7] T. Mashadihasanli, "Stock Market Price Forecasting Using the Arima Model: an Application to Istanbul, Turkiye," *Journal of Economic Policy Researches,* vol. 9, no. 2, pp. 439-454, 2022.

[8] Z. Zhu and K. He, "Prediction of Amazon's Stock Price Based on ARIMA, XGBoost, and LSTM Models," *Proceedings of Business and Economic Studies,* vol. 5, no. 5, pp. 127-136, 2022.

[9] A. Teller, U. Pigorsch and C. Pigorsch, "Short-to Long-Term Realized Volatility Forecasting using Extreme Gradient Boosting," *SSRN,* vol. 4267541, 2022.

[10] N. Minhaj, R. Ahmed, I. A. Khalique and M. Imran, "A Comparative Research of Stock Price Prediction of Selected Stock Indexes and the Stock Market by Using Arima Model," *Global Economics Science,* pp. 1-19, 2023.

[11] S. Saha, "Blog: Machine Learning Tools," 27 March 2023. [Online]. Available: https://neptune.ai/blog/xgboost-vs-lightgbm. [Accessed March 2023].

[12] B. Boehmke and B. M. Greenwell, Hands-on machine learning with R, CRC press, 2019.

[13] R. H. Shumway and D. S. Stoffer, Time series analysis and its applications, vol. 3, Springer, 2000.

[14] R. J. Hyndman and G. Athanasopoulos, Forecasting: principles and practice, OTexts, 2018.

[15] F. Hadavimoghaddam, M. Ostadhassan, M. A. Sadri, T. Bondarenko, I. Chebyshev and A. Semnani, "Prediction of water saturation from well log data by machine learning algorithms: boosting and super learner," *Journal of Marine Science and Engineering,* vol. 9, no. 6, p. 666, 2021.

[16] K. L. C. Thai, "How to Calculate the Daily Returns And Volatility of a Stock with Python," 25 June 2022. [Online]. Available: https://blog.devgenius.io/how-to-calculate-the-daily-returns-and-volatility-of-a-stock-with-python-d4e1de53e53b. [Accessed March 2023].

[17] J. Fernando, "Consumer Price Index (CPI) Explained: What It Is and How It's Used," 14 March 2023. [Online]. Available: https://www.investopedia.com/terms/c/consumerpriceindex.asp. [Accessed April 2023].