Department of Economics - FEA/USP

ARIMA and LSTM: A Comparative Analysis of Financial Time Series Forecasting

JOÃO VITOR MATOS GONÇALVES

MICHEL ALEXANDRE

GILBERTO TADEU LIMA

WORKING PAPER SERIES Nº 2023-13

DEPARTMENT OF ECONOMICS, FEA-USP
WORKING PAPER Nº 2023-13

ARIMA and LSTM: A Comparative Analysis of Financial Time Series Forecasting

João Vitor Matos Gonçalves (joaovmatosg@usp.br)

Michel Alexandre (michel.alsilva@gmail.com)

Gilberto Tadeu Lima (giltadeu@usp.br)

Abstract:

This paper assesses the impact of time horizon on the relative performance of traditional econometric models and machine learning models in forecasting stock market prices. We employ an extensive daily series of Brazil IBX50 closing prices between 2012 and 2022 to compare the performance of two forecasting models: ARIMA (autoregressive integrated moving average) and LSTM (long short-term memory) models. Our results suggest that the ARIMA model predicts better data points that are closer to the training data, as it loses predictive power as the forecast window increases. We also find that the LSTM model is a more reliable source of prediction when dealing with longer forecast windows, yielding good results in all the windows tested in this paper.

Keywords: Finance; machine learning; deep learning; stock market.

JEL Codes: C22; C45; C53; G17.

ARIMA and LSTM: A comparative analysis of financial time series forecasting

João Vitor Matos Gonçalves - Universidade Federal Fluminense
Michel Alexandre - Universidade de São Paulo
Gilberto Tadeu Lima - Universidade de São Paulo

Abstract

This paper assesses the impact of time horizon on the relative performance of traditional econometric models and machine learning models in forecasting stock market prices. We employ an extensive daily series of Brazil IBX50 closing prices between 2012 and 2022 to compare the performance of two forecasting models: ARIMA (autoregressive integrated moving average) and LSTM (long short-term memory) models. Our results suggest that the ARIMA model predicts better data points that are closer to the training data, as it loses predictive power as the forecast window increases. We also find that the LSTM model is a more reliable source of prediction when dealing with longer forecast windows, yielding good results in all the windows tested in this paper.

Keywords: Finance; machine learning; deep learning; stock market.

1 Introduction

The use of data science techniques has become increasingly widespread with the massive growth in available data and the increasingly accessible data processing capacity. Scholarly research in the field of financial economics has been following this new trend, and it has been increasingly incorporating these techniques. The incorporation of these new techniques is intended to complement the techniques already established in the literature and to address the perceived limitations of traditional econometric techniques.

One of the data science techniques that has gained increasing prominence is the machine learning (ML) approach. Athey and Imbens (2019) outline various new techniques in this area, their applications, and their idiosyncrasies. This interesting review article demonstrates how these new techniques can be beneficial for research in economics by highlighting specific situations where a given technique can be employed. In particular, ML models outperform traditional econometric models in terms of out-of-sample predictive power. Goodell et al. (2021) develop a bibliometric analysis of the use of artificial intelligence and ML techniques employed in the field of finance. In addition to reviewing the literature, their paper points to recent trends in the use of these methodologies in finance.

Masini et al. (2023) present an extensive literature review on ML techniques applied specifically to economic and financial time series. Their paper also compares different techniques aiming at forecasting the realized variance in stock market indices. Giannone et al. (2021) analyze a group of different applications in finance, microeconomics, and macroeconomics to test whether economic models should be low dimensional, and the evidence presented shows low support for that assumption. Another interesting use of machine learning techniques is made in Gu et al. (2020), who explore high-dimensional data to predict stock market returns. Goulet Coulombe et al. (2022) provide evidence that the ability of ML models to deal with nonlinearities is very helpful in macroeconomic forecasting. Kotchoni et al. (2019) explore six classes of models to forecast different series, comparing the power of the prediction of traditional and ML models.

Newer techniques such as data scraping and sentiment analysis can also be found in contributions such as that made by Duarte et al. (2021), who utilize news information for stock price prediction. Meanwhile, Saurabh and Dey (2020) present a series of ML forecasting models to evaluate the correlation between stock market prices and market reactions, as captured via tweets related to the theme. Liu et al. (2023) also use sentiment analysis from tweets, aiming to predict stock market returns. News sentiment is also explored by Souma et al. (2019). The authors define news sentiment according to the variation of the one-minute average of the stock returns right after the release of the news article. Applying deep learning techniques, they show that the performance of their methodology

depends on the score of the news selected to create the training dataset.

The related literature also includes contributions that compare traditional techniques with ML techniques. Agrawal et al. (2023) present a comparison of Bitcoin value prediction using an ARIMA model against ML techniques such as decision trees, long short-term memory, random forests, and support vector machines. Bezerra and Albuquerque (2019) also compare these two groups of techniques in terms of the daily returns of the American and Brazilian stock markets. Meanwhile, Araujo and Gaglianone (2023) compare a range of 50 different forecasting methods for Brazilian inflation and conclude that ML models tend to outperform more traditional methods (such as the ARIMA model). However, the authors also conclude that there is not a clearly superior method, the definition of which depends on the data and the metric used to evaluate predictions.

Similarly, Garcia et al. (2017) use high-dimensional models to predict Brazilian inflation. They conclude that this type of model, especially the complete subset regression, performs very well for different horizons of prediction. In stock market prediction using ML, Pierdzioch and Risse (2018) test the rational expectations hypothesis by making predictions on the American stock market using boosted regression trees and by comparing the results for different windows of analysis.

An element that is usually pointed out by the literature as having an important impact on the performance of predictive models is the time horizon. Applying ML techniques for the prediction of the NASDAQ index, Cervelló-Royo and Guijarro (2020) show that technical indicators have poor predictive power for the next trading day but that the performance significantly improves when the time horizon is expanded to forecast the 10-trading-day trend. Moreover, the relative performance of predictive models is driven by the time horizon under consideration. Comparing the performance of three methods for predicting the direction of financial time series, Ersan et al. (2020) show that increasing the time window size helps only for hourly data and up to a certain extent. Kotchoni et al. (2019) show that the ARMA model is the best at forecasting inflation changes in the short run, while regularized data-rich model averaging (RDRMA) dominates in the case of longer horizons. Faust and Wright (2013) perform a horse race among different forecasting models for inflation and find that the best predictive model depends on the time horizon being considered.

This paper aims to contribute to this burgeoning literature financial time series forecasting by exploring different prediction scenarios — defined by different time horizons — to identify the most suitable technique for each scenario. We apply two representative models — autoregressive integrated moving average (ARIMA) and long short-term memory (LSTM) models — to predict the daily closing prices of the IBX50, the index composed of the 50 most traded stocks in the Brazilian stock market. The performance of the models is assessed under different time horizons: 5 days, 15 days, 30 days, and the whole out-of-sample

series. Our results show that the ARIMA model outperforms the LSTM model in the 5-day time horizon, but the LSTM model is a better predictive model than the ARIMA model under longer time horizons. Therefore, we provide robust evidence that the vaunted advantage of ML models over traditional econometric tools for forecasting tasks depends on the time horizon under consideration.

The remainder of this paper is organized as follows. The dataset and the methodological issues involved are described and discussed in Section 2, while the results are reported and discussed in Section 3. Finally, Section 4 concludes.

2 Data and methodology

The series that are analyzed in this paper consist of the daily closing price of the IBX50 from December 12, 2013, to December 29, 2022. The IBX50 is the Brazil 50 index composed of the 50 most traded stocks, and for that reason, it is highly representative of the Brazilian stock market. The data used in this analysis were retrieved from Yahoo Finance via its Python API. The data have a total of 2140 observations. To conduct the analysis, approximately 90% of the data were used for training, and the remaining 10% were used for testing.

The main objective of this paper is to evaluate how two types of representative methodologies perform under the same circumstances and to understand the scenarios in which each methodology prevails. To meet our objective, two representative methodologies were selected, the first of which is the ARIMA model, which represents a more traditional approach in finance and economics.[1] To represent a methodology in the ML and deep learning group, the long short-term memory (LSTM) model was chosen because of it performs well in time series prediction scenarios.

2.1 ARIMA

ARIMA stands for "autoregressive integrated moving average". This family is a generalization of ARMA models, which were first introduced by Box et al. (2015). ARIMA models are used in series that have an autoregressive and a moving average component, but unlike ARMA models, they can work with time series that are nonstationary by differencing the series as many times as needed to eliminate the nonstationarity. The ARIMA process is composed of three parts: the autoregressive (AR), the integrated (I), and the moving average (MA). The autoregressive model is generically represented by $AR(p)$, where p represents the

[1] Other models, such as ARMA, SARIMA, and SARIMAX models, were considered for this analysis, but due to the behavior of the series being used, the ARIMA model was ultimately chosen.

order of the autoregressive. In this case, the value of X_t is a combination of past values of p (Peixeiro, 2022):

$$X_t = C + \phi_1 X_{t-1} + ...\phi_p X_{t-p} + \epsilon_t. \tag{1}$$

Likewise, for the moving average process, the value of $MA(q)$ is a combination of past information, but this time, not the value itself but the q past errors. Therefore, this time, X_t is represented as follows:

$$X_t = \mu + \epsilon_1 \theta_{t-1} + ...\epsilon_p \theta_{t-q} + \epsilon_t, \tag{2}$$

where μ is the mean of the series and θ quantifies the impact of past errors on the present value. The integrated part is represented by the letter d, which represents how many times the series will be differentiated until it becomes stationary. Therefore, when all three elements are combined, we have a process represented by $ARIMA(p, d, q)$, where p, d and q denote the autoregressive, the integration, and the moving average orders of the model, respectively.

2.2 LSTM

Long Short-Term Memory, developed by Hochreiter and Schmidhuber (1997), is a type of recurrent neural network (RNN) that processes data in sequence, where the order of the data matters, similar to a time series. This methodology was created to address a problem that RNN models suffer, which is gradient disappearance. RNN models ignore past information that is too far in the sequence.

The LSTM model is composed of three layers, namely, the forget gate, input gate, and output gate, in that order. The forget layer is the first layer, and it decides what information from the sequence is retained in the network. Information is then passed to the input gate, where it is decided which information is important to the current element of the sequence. Finally, information is passed to the output gate, where past information processes the current value in the sequence.

2.3 Dataset

In this empirical study, we used the Brazilian stock market to illustrate the forecasting capacity of these two representative methodologies. Figure 1 shows the evolution of the closing prices of the IBX50 in a daily frequency from the end of 2012 until the last available prices in 2022. The complete series has a total of 2,483 data points. As expected, the series shows some volatility with an upward trend.

As is usual in this type of empirical study, we divided the dataset into two parts: the training dataset and the test dataset. The training dataset was used to

develop both models, using approximately 90% of the original series, thus representing the first 2,229 points of the series. The test dataset was used to assess the performance of the models (more details are provided in Section 3). In Figure 1, the unshaded (shaded) part of the plot corresponds to the training (test) dataset.

Figure 1: Historical closing price of the IBX50 at a daily frequency: training (unshaded) and test (shaded) datasets.

3 Results and discussion

To appropriately evaluate and compare the forecasting capacity of these two representative methodologies (traditional econometrics and ML), the first step is to create models that are adequate to the type of series that is being analyzed. Therefore, the next two subsections explore how the two models were created using the same training dataset and show the test that was performed to assess the performance of the predictions of the models. Subsection 3.3 then shows how the two models that were created in the preceding subsections perform under different forecasting scenarios.

3.1 ARIMA Model

As noted in Section 2, the ARIMA process is composed of three parameters (p, d, and q). The first to be defined is the parameter d, which indicates how many times the series will be differentiated until it becomes stationary. To access this information, the augmented Dickey–Fuller test was performed to check whether the series is stationary. The null hypothesis of this test is that a unit root is present

in the series; that is, the series is nonstationary (Cheung and Lai, 1995). As shown in Table 1, the original series does not reject the null hypothesis of stationarity, but the first-order difference has a p value small enough to ensure that it does not have a unit root. Given that the first-difference series is stationary, we no longer need to differentiate the series, and we can define parameter d as being equal to one.

Table 1: Augmented Dickey-Fuller test.

	ADF	**P-value**
Level	-1.20	0.6735
First Differences	-8.33	<0.0001

Having defined the first parameter, we plotted the ACF and PACF to verify whether there was a clear pattern to identify the p and q parameters, but the result was not clear. Thus, the strategy chosen was to set different values for the p and q parameters and to set the d parameter equal to one.[2] The model with the lowest AIC (Akaike information criterion) value was selected: the model selected was the $ARIMA(5, 1, 5)$ because it had the lowest AIC value; thus, it can be considered the model with the best fit to the data. Table 2 shows the values for the ten models with the lowest AIC values.

Table 2: Augmented Dickey-Fuller test with different time lags.

Time lag	Statistic	P-value
1	0.219209	0.639643
2	0.238613	0.887536
3	0.252818	0.968643
4	0.857370	0.930596
5	1.389218	0.925489
6	1.502233	0.959346
7	2.258313	0.944172
8	2.351062	0.968315
9	2.626203	0.977295
10	3.102355	0.978912

Once the parameters of our ARIMA model were selected, we evaluated whether the model residuals were stationary. To do so, four plots were made so that we could visually analyze the behavior of the residuals. Figure 2 shows the four plots

[2]We varied the values of the p and q parameters from 0 to 5 with step 1. Then, 36 different models were created.

that were created to check whether the residuals are a stationary process. The top-left plot shows that the residuals have no trends, while the top-right plot shows that the distribution of the residuals resembles a normal distribution. The Q–Q plot shows a straight line for most of the series except for the extreme values and the last value, while the bottom-right has the ACF plot with no autocorrelation after time lag 0. In addition to all the plots shown in Figure 2, the residuals passed the Ljung–Box test, which shows that none of the ten first lags of the residuals were correlated.

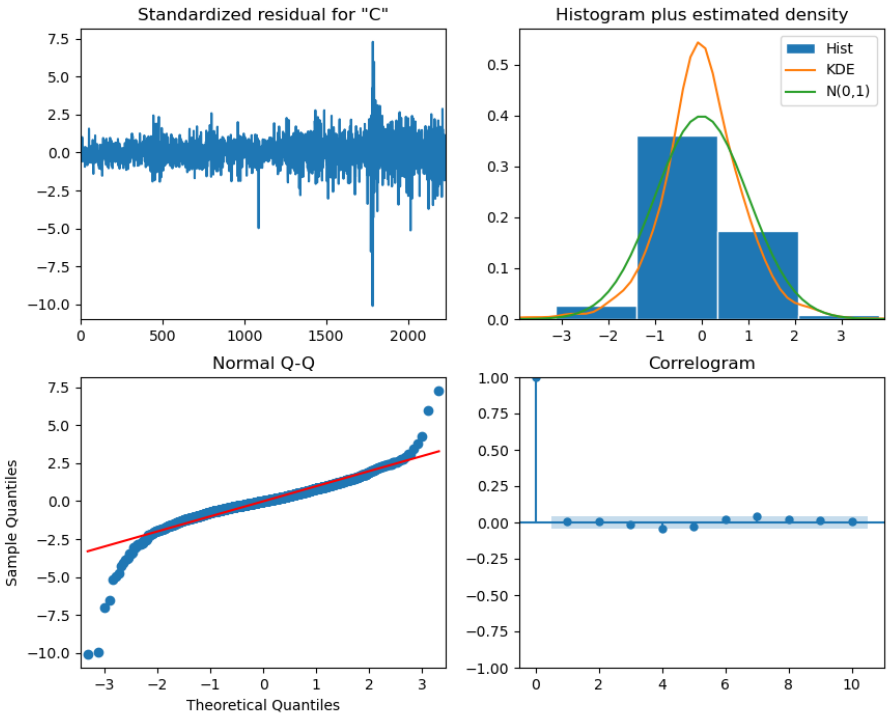

Figure 2: Residual analysis. From left to right in the first row, we have the residuals and the histogram of the residuals, while in the second row, we have the Q-Q plot and the ACF plot.

After we were able to confirm that the model is valid, we measured the fit of the model with the training data and calculated the mean average percentage error (MAPE) and root mean square error (RMSE), with the respective values shown in Table 3.

Table 3: Error metrics for the ARIMA model.

MAPE	RMSE
0.0112	202.4418

3.2 LSTM Model

As with most deep learning models, the LSTM model does not require too many steps and tests to create a model. However, some tuning is needed to ensure that the model is adequate to the reality of the considered data. The first parameter to be defined is the size of the sequence, given that the LSTM model evaluates data as a sequence. Since we are dealing with a daily time series, we chose the sequence of 7 days as representative of a week.

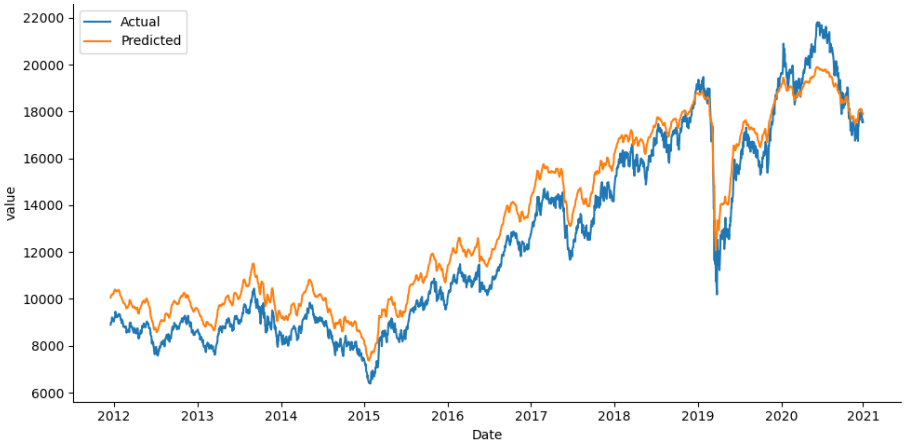

Figure 3: Actual and predicted closing prices for the LSTM model.

The result was a model with four layers, one input layer, two bidirectional LSTM layers, and one output layer, as shown in Figure 3. Note that the difference between the predicted and actual values is never too large. This behavior is a sign of a good model since it has a good adjustment without suffering from overfitting. This conclusion is supported by the calculation of some metrics for the predictions presented in Table 4.

Table 4: Error metrics for the LSTM model.

MAPE	RMSE
0.0898	1043.8081

Now that we have created the two models using the training dataset, we can

compare how they perform in different forecasting windows. Subsection 3.3 will show the results of this comparison.

3.3 Comparing forecasts

The forecasts that are being performed for the test dataset use the information in the training dataset, as we are making individual predictions for different time horizons. Therefore, if the test dataset ends at point t, we will predict from point $t+1$, the start of our test dataset, until we reach the end of our time series at point T. Equation 3 illustrates this process, where $\hat{x}_{t+i}$ is the predicted value of x in $t+i$, which is a function of all the observed values of x from time zero until t.

$$\hat{x}_{t+i} = f(x_0, \ldots, x_t) \tag{3}$$

To compare the two models, we use different metrics of errors to check whether one model is superior to the other in more than one metric. In addition, we use different forecasting windows. First, the models predict the first k days of the test dataset, where $k = \{5, 15, 30\}$. Thus, unlike Kotchoni et al. (2019), we use not *rolling* windows but *increasing* windows. Then, we evaluate how the two models perform when we use the whole test dataset. By doing so, we can access the capacity of the considered models to forecast farther into the future. To carry out this analysis, we employ the RMSE and MAPE as our metrics of the quality of the forecasting.

As shown in Table 5, the ARIMA model starts with good performance metrics in the 5-day window, but as the window becomes larger, the results yielded by the ARIMA model begin to worsen. In fact, for the four windows under consideration, the 30-day window is the one for which the ARIMA model yields the worst results. With the ARIMA model, how quickly the RMSE deteriorates is noteworthy: for the 5-day window, the ARIMA model performs better than the LSTM model. Regarding the LSTM model, the results have a different behavior: for the MAPE statistic, the results become slightly worse as the windows grows larger and the RMSE statistic becomes better, achieving its peak at the 30-day window. In Figure 4, it can be observed that the LSTM model outperforms the ARIMA model under longer time horizons.

Table 5: Forecasting metrics for different windows

Window	Model	MAPE	RMSE
5 days	ARIMA	0.0108	223.24
	LSTM	0.0401	741.24
15 days	ARIMA	0.0367	777.54
	LSTM	0.0167	438.81
30 days	ARIMA	0.0616	1295.37
	LSTM	0.0141	359.30
Complete series	ARIMA	0.0527	1191.97
	LSTM	0.0251	581.06

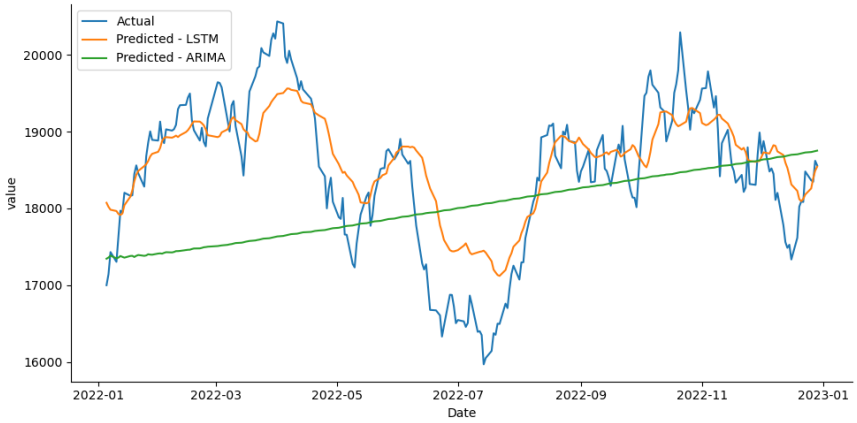

Figure 4: Actual closing prices and predictions for the two models in the test dataset.

4 Conclusion

The increasing use of ML models in several fields in economics is mainly explained by their better predictive power compared to that of more traditional econometric tools. However, the specialized literature on the subject points out that the time horizon impacts both the absolute and the relative performance of different predictive models.

Against this backdrop, this paper developed a comparative analysis of the predictive power of two financial time series forecasting models considering different time horizons. We chose one model of each type under consideration, namely, a more traditional model (ARIMA) and a representative ML model (LSTM), both of which are suitable for time series analysis.

We performed a forecasting exercise using the Brazilian IBX50, an index composed of the 50 most traded stocks in the Brazilian stock market, applying both models. We considered four different time horizons: the first k observations of the out-of-sample series, with $k = \{5, 15, 30\}$, and the whole out-of-sample series. While the ARIMA model proved to be a better predictive model than the LSTM model in the 5-day time horizon, the latter outperformed the former under longer time horizons.

Our results provide further robust corroborating evidence of the strong predictive power of ML models for time series analysis. However, our results also show that the predictive performance of a ML model depends on the time horizon under consideration. In the particular case of financial time series explored in this paper, the representative ML model (LSTM) outperformed the traditional econometric model (ARIMA) only in the case of larger time windows. These results highlight the importance of always comparing the predictive power of different models by performing a rigorous statistical horse race between them. Moreover, this statistical horse race should include more traditional models and cutting-edge techniques, with the time horizon involved being a variable to be taken into due consideration.

References

A. Agrawal, M. Mani, and S. Varshney. Bitcoin forecasting performance measurement: A comparative study of econometric, machine learning and artificial intelligence-based models. *Journal of International Commerce, Economics and Policy*, 14(2):2350008, 2023.

G. S. Araujo and W. P. Gaglianone. Machine learning methods for inflation forecasting in Brazil: New contenders versus classical models. *Latin American Journal of Central Banking*, 4(2):100087, 2023.

S. Athey and G. W. Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725, 2019.

P. C. S. Bezerra and P. H. M. Albuquerque. Volatility forecasting: The support vector regression can beat the random walk. *Economic Computation & Economic Cybernetics Studies & Research*, 53(4):115–126, 2019.

G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

R. Cervelló-Royo and F. Guijarro. Forecasting stock market trend: A comparison

of machine learning algorithms. *Finance, Markets and Valuation*, 6(1):37–49, 2020.

Y.-W. Cheung and K. S. Lai. Lag order and critical values of the augmented Dickey–Fuller test. *Journal of Business & Economic Statistics*, 13(3):277–280, 1995.

J. J. Duarte, S. Montenegro González, and J. C. Cruz. Predicting stock price falls using news data: Evidence from the Brazilian market. *Computational Economics*, 57:311–340, 2021.

D. Ersan, C. Nishioka, and A. Scherp. Comparison of machine learning methods for financial time series forecasting at the examples of over 10 years of daily and hourly data of DAX 30 and S&P 500. *Journal of Computational Social Science*, 3:103–133, 2020.

J. Faust and J. H. Wright. Forecasting inflation. In *Handbook of economic forecasting*, volume 2, pages 2–56. Elsevier, 2013.

M. G. Garcia, M. C. Medeiros, and G. F. Vasconcelos. Real-time inflation forecasting with high-dimensional models: The case of Brazil. *International Journal of Forecasting*, 33(3):679–693, 2017.

D. Giannone, M. Lenza, and G. E. Primiceri. Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437, 2021.

J. W. Goodell, S. Kumar, W. M. Lim, and D. Pattnaik. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, 32:100577, 2021.

P. Goulet Coulombe, M. Leroux, D. Stevanovic, and S. Surprenant. How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, 37(5):920–964, 2022.

S. Gu, B. Kelly, and D. Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

R. Kotchoni, M. Leroux, and D. Stevanovic. Macroeconomic forecast accuracy in a data-rich environment. *Journal of Applied Econometrics*, 34(7):1050–1072, 2019.

Q. Liu, W.-S. Lee, M. Huang, and Q. Wu. Synergy between stock prices and investor sentiment in social media. *Borsa Istanbul review*, 23(1):76–92, 2023.

R. P. Masini, M. C. Medeiros, and E. F. Mendes. Machine learning advances for time series forecasting. *Journal of economic surveys*, 37(1):76–111, 2023.

M. Peixeiro. *Time Series Forecasting in Python*. Manning Publications, 2022.

C. Pierdzioch and M. Risse. A machine-learning analysis of the rationality of aggregate stock market forecasts. *International Journal of Finance & Economics*, 23(4):642–654, 2018.

S. Saurabh and K. Dey. Unraveling the relationship between social moods and the stock market: Evidence from the United Kingdom. *Journal of Behavioral and Experimental Finance*, 26:100300, 2020.

W. Souma, I. Vodenska, and H. Aoyama. Enhanced news sentiment analysis using deep learning methods. *Journal of Computational Social Science*, 2(1):33–46, 2019.