A COMPARATIVE ANALYSIS OF TIME SERIES MODELS FOR PREDICTING THE S&P SL 20 INDEX OF THE COLOMBO STOCK EXCHANGE (CSE)

M. Vithushan*[*] *and G. A. P. Kethmi

Department of Finance, University of Kelaniya, Sri Lanka

The performance of stock markets is influenced by various factors, and understanding these dynamics is crucial for investors and policymakers. This research centers on Sri Lanka's capital market, with particular attention to the Colombo Stock Exchange (CSE), and analyzes the influence of the Standard & Poor's Sri Lanka 20 (S&P SL 20) index, which represents the top 20 leading companies listed on the CSE. Specifically, the primary objective of this research is to compare the effectiveness of traditional time series models with machine learning and deep learning models in predicting the S&P SL 20 index. These models, developed using computerized programs, are evaluated based on their predictive performance within the context of the CSE. The study will use daily S&P SL 20 stock index data obtained from the CSE data library enclosing the period 2010 to 2018. This methodology compares Autoregressive Integrated Moving Average (ARIMA), which is a traditional time series model and Long Short-Term Memory (LSTM), which is a recurrent neural network model. In this research, the Python language will be employed for analysis. The ARIMA and LSTM models are evaluated using three performance metrics: MAE, MAPE, and RMSE. ARIMA slightly outperforms LSTM in MAE (233.96 vs. 249.37) and RMSE (269.57 vs. 269.86), which essentially indicates better overall accuracy in absolute and squared error terms. However, LSTM achieves a marginally lower MAPE (6.96% vs. 7.04%), showing fewer relative percentage errors. All in all, both models offer similar performances, with minor differences depending on the metric. Both ARIMA and LSTM show strengths in predicting the S&P SL 20 Index. ARIMA excels in minimizing absolute errors (MAE), ideal for linear trends. LSTM's lower MAPE highlights its ability to capture nonlinear patterns. With similar RMSE values, both handle overall errors well. ARIMA's constant predictions in some periods reveal its limitations with limited or weak trend data. The choice depends on the forecasting goal: ARIMA for linear trends and LSTM for complex patterns.

Keywords: time series, Colombo Stock Exchange, autoregressive integrated moving average, long short-term memory

[*]*Corresponding Author: vithush-fe19078@stu.kln.ac.lk*

ISSN 3093-6470

A COMPARATIVE ANALYSIS OF TIME SERIES MODELS FOR PREDICTING THE S&P SL 20 INDEX OF THE COLOMBO STOCK EXCHANGE (CSE)

M. Vithushan **[*]** ***and G. A. P. Kethmi***
Department of Finance, University of Kelaniya, Sri Lanka

INTRODUCTION

The S&P SL 20 index is an index that represents the top 20 leading companies listed in the CSE. This study aims to evaluate and contrast the predictive capabilities of a traditional time series model, Auto-Regressive Integrated Moving Average - ARIMA and a machine learning approach, Long Short-Term Memory-LSTM for forecasting the index, which were built by computerized programs while analyzing the performances. The S&P SL index displays both linear and nonlinear fluctuational patterns influenced by capital market volatility. Specifically, ARIMA accurately captures short-term linear trends, while LSTM captures nonlinear dependencies. Hence, a hybrid ARIMA-LSTM approach effectively combines both methods to enhance the forecasting accuracy of the S&P SL 20 index. The foundation of time series forecasting lies in models like ARIMA, which use historical data to predict future trends. Kumar and Thenmozhi (2020) looked at using ARIMA for intraday stock price forecasting, highlighting its usefulness for analyzing high-frequency data. This research shows that while methods continue to improve, there’s a strong focus on refining forecasting techniques to make predictions more accurate. Over time, the field of prediction has undergone a significant change with the introduction of machine learning and deep learning techniques such as Support Vector Machines (SVM) and Random Forests (RF), along with deep learning models like Recurrent Neural Networks (RNN) and LSTM, which have become powerful tools. Ojo et al. (2019) employed LSTM exclusively as their predictive method in forecasting stock market behavior to compare the changes over the fluctuations. Kulshreshtha and Vijayalakshmi (2020) proposed a novel ARIMA-LSTM hybrid model for stock prediction, demonstrating superior performance compared to Prophet, with an MSE of 3.03, an RMSE of 1.74, and a 99% model fit. Additionally, they developed a user interface for hosting the prediction system based on the ARIMA-LSTM hybrid. Integration ensures that ARIMA and LSTM form the best combination. In view of making forecasts of stock prices for those using both traditional and machine learning models (ARIMA and LSTM, respectively), one should be very sensitive with respect to the research objectives and context. It is thus justified to formulate and test the following hypotheses:
H1: There is a significant difference in the accuracy of time series models to predict the stock price of the S&P 20 index.

METHODOLOGY

The study will employ daily S&P SL 20 stock index data from the CSE data repository covering the years 2010 to 2018. Data beyond that time frame is not taken into consideration because of the unusual circumstances that happened in Sri Lanka, which may have had a significant impact on the event. ARIMA can handle a number of model types, including regression analysis with ARMA errors, interrupted time series, and seasonal models. ARIMA (p, d, q) has the letters stand for the three main elements of the model, which are the number of terms for moving average, differences, and autoregression (Table 1). The LSTM approach used the command of logical gates to function. The LSTM model is built with multiple input layers. The LSTM layers capture sequential dependencies and generate the predictions. Training is improved using the Adam optimizer, with Mean Squared Error (MSE) employed as the loss function to minimize forecast errors. This architecture of LSTM layout is designed to handle long-term dependencies. These are unique gates that use a sigmoid function, which is a mathematical formula that returns a value between 0 and 1. This regulates how much data may flow through, with zero being prohibited and one being fully allowed. Three types of gates function within each LSTM unit to manage cell states:

- The Forget Gate generates values between 0 and 1, with 1 indicating complete retention and 0 implying total disregard.
- The Memory Gate: An initial sigmoid layer—the "input gate layer"chooses which values to update, while a tan layer has a vector of possible new values that may be added to the state.
- Output Gate: The output gate regulates the final information that is transmitted from each cell. This output is also fed from the cell state with some filtered and newly added data.

RESULTS AND DISCUSSION

Application of the ARIMA Model for Time Series Forecasting: Steps, Rationale, and Detailed Findings:

The ARIMA model requires stationarity, meaning that the statistical properties (mean, variance) of the series should remain consistent over time. Three different methods were applied to check for stationarity in the time series. Hypotheses of the ADF Test (Figure 1):

Null Hypothesis (H_0): The time series has a unit root (it is non-stationary).
Alternative Hypothesis (H_1): The time series does not have a unit root (it is stationary).

ARIMA models operate under the assumption that the time series is stationary, meaning its statistical properties, such as mean and variance, remain constant over time. To assess whether the data meets this condition, three techniques were employed, one of which was the Augmented Dickey-Fuller (ADF) test. The null hypothesis (H_0) indicates that the time series is non-stationary due to the presence of a unit root, whereas the alternative hypothesis (H_1) proposes that the series is stationary. Applying the ADF test to the log-transformed data yielded a p-value of 0.05108, slightly above the 0.05 threshold, indicating that the series was non-stationary. (Table 2: Sarimax Results). To determine model parameters, ACF and PACF plots were examined. The ACF showed a sharp cutoff, and the PACF tapered off gradually, suggesting an autoregressive (AR) process. Based on these patterns, the ARIMA (2,1,0) model was selected: p = 2 (from significant lags in the PACF), d = 1 (to achieve stationarity), and q = 0 (due to lack of significant moving average terms). This model was applied to a univariate series of 2160 observations of the target variable, Indexes, and further diagnostic tests were conducted to assess model performance. (Table 3: Diagnostic Tests)

Application of the LSTM Model for Time Series Forecasting: Steps, Rationale, and Detailed Findings:
The data is split into training and testing subsets based on specific date intervals:

- The training set spans from 2010 to 2016, allowing the model to learn patterns and trends in historical data.
- The testing set spans from 2017 to 2018, serving as unseen data to evaluate how well the model generalizes.

After date conversion, only the Date and Value columns are retained. This ensures that the model focuses exclusively on time and target values, discarding irrelevant columns that might introduce noise. The model's performance metrics indicate its effectiveness in capturing temporal patterns. The MAE of 249.37 suggests an average deviation of 249.37 units between predictions and actual values, reflecting a relatively low error given the dataset's scale. The MAPE of 7.04% indicates that predictions are, on average, within 7% of the actual values, demonstrating strong accuracy, particularly valuable in financial forecasting. The RMSE of 269.57, which emphasizes larger errors, is close to the MAE, implying the model handles most data well with minimal impact from outliers or extreme fluctuations. (Table 4). Even though LSTM captures the long nonlinear functions, here the ARIMA outperformed LSTM. This could be due to several reasons of the nature of CSE, which is highly influenced by policy and political decisions of the government, such as limited trend/seasonality in training data, overfitting, data scarcity, and the nature of financial data, highlighting the challenges in time series forecasting. However, due to these discrepancies in research, ARIMA's predictions are found to be more accurate than LSTM in this context. (Figure 2)

CONCLUSIONS/RECOMMENDATIONS

The results suggest that both ARIMA and LSTM have various strengths and weaknesses in stock price prediction. ARIMA performs comparatively better in terms of MAE, showing it is more suitable for the type of data where prediction must consider the linear relationship. On the other side, LSTM slightly outperforms MAPE, suggesting its ability for generalization, since the model is proportionally correct, likely due to its capacity to capture non-linear and complex dependencies in the data. Similar RMSE values mark that both models are almost equally capable in managing error magnitudes across the dataset. However, the marginal differences in MAE and MAPE suggest the model choice depends on the specific objectives of the forecasting task (Figure 2). For example:

- If minimizing absolute prediction errors is the priority, ARIMA might be preferred.
- If proportional accuracy and handling non-linear trends are more important, LSTM may be more suitable.

The S&P SL20 has significant implications for investors in emerging economies. Particularly for institutional and individual investors making decisions regarding asset allocation and portfolio diversification, this ARIMA-LSTM model offers more accurate forecasts of market movements.

REFERENCES

Kumar, M., & Thenmozhi, M. (2006). *Forecasting stock index movement: A comparison of support vector machines and random forest* (Indian Institute of Capital Markets 9th Capital Markets Conference Paper). SSRN. https://doi.org/10.2139/ssrn.876544

Ojo, S. O., Owolawi, P. A., & Mphahlele, M. (2019, November). *Stock market behaviour prediction using stacked LSTM networks*. In *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)* (pp. 1–6). IEEE.https://doi.org/10.1109/IMITEC45504.2019.9015840

Kulshreshtha, S., & Vijayalakshmi, A. (2020). *An ARIMA-LSTM hybrid model for stock market prediction using live data*. *Journal of Engineering Science and Technology Review, 13*(4), 117–123. https://doi.org/10.25103/jestr.134.11

Ariyo, A. A., Adewumi, A. O., & Ayo, C. K. (2014, March). *Stock price prediction using the ARIMA model*. In *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation* (pp. 106–112). IEEE. https://doi.org/10.1109/UKSim.2014.67

Figure 1: ACF and PACF plot results after differencing

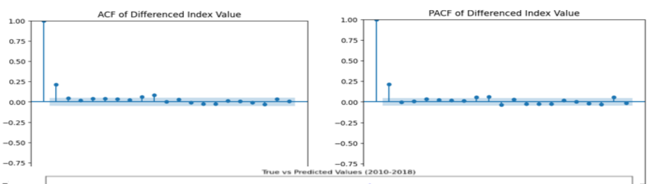

Figure 2 … TM model

Research method	Quantitative research
Research philosophy	Positivism philosophy
Research logic	Deductive approach
Main data analysis technique	ARIMA,LSTM
Software for main data analysis	Python

Table 1: Research Introduction

Table 2: Sarimax Results

	Co.Ef	Std err	z	P> |z|	[0.025	0.975]
ar.L1	0.2137	0.013	16.518	0.000	0.188	0.239
Ar.L2	-0.0001	0.017	-0.008	0.994	-0.033	0.033
Sigma2	646.4658	11.323	57.092	0.000	624.273	668.659

Table 3: Diagnostic Tests

Ljung-Box (L1) (Q)	0.00	Jarque-Bera (JB)	2374.34
Prob (Q)	0.98	Prob (JB)	0.00
Heteroskedasticity (H)	0.52	Skew	-0.05
Prob(H) (two-sided)	0.00	Kurtosis	8.83

Table 4: Metrics comparison

Metrics	ARIMA	LSTM
MAE	233.96	249.37
MAPE	7.04%	6.96%

RMSE	269.57	269.86