Assignment Report

A Comparative Case Study on Modelling and Predictive Power of ARIMA and LSTM Models on Time Series Data

Aaron Modiyil Joseph
22018497

University of Hertfordshire
01 May 2024

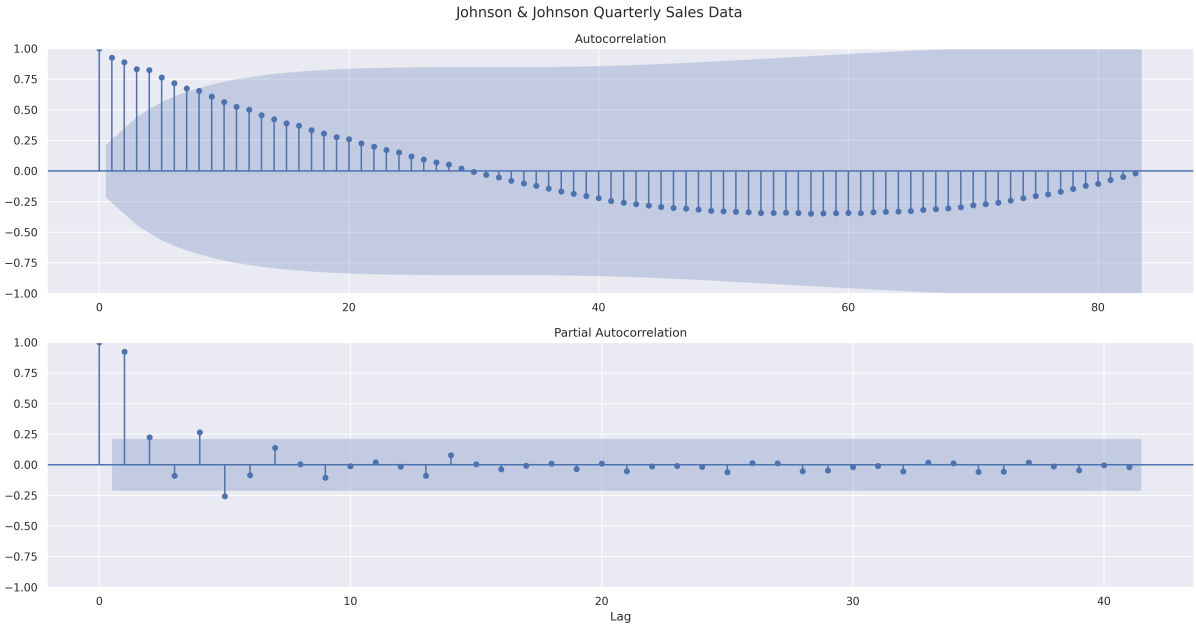

Figure 1: ACF and PACF plots of Johnson & Johnson Quarterly Sales Data

1 Introduction

Time series analysis and forecasting play a crucial role in understanding and predicting patterns in various domains, from finance and economics to meteorology and engineering. This study aims to explore the modeling capabilities and predictive power of two widely used techniques: Autoregressive Integrated Moving Average (ARIMA) models and Long Short-Term Memory (LSTM) neural networks. By applying these methods to real-world datasets from two prominent companies, Johnson & Johnson and Amazon, we strive to gain insights into the strengths and limitations of each approach. The comparative analysis will not only highlight the models' performance in capturing trends and seasonality but also shed light on their suitability for different types of time series data. Ultimately, this investigation seeks to inform future research and practical applications, contributing to the ongoing quest for accurate and reliable forecasting methods in an ever-changing and data-driven world.

2 Data Overview and Preprocessing

This assignment investigates two distinct datasets from different companies. The first dataset examines Johnson & Johnson's historical quarterly sales data from 1960 to 1980, comprising 84 data points. Each point corresponds to a specific quarter. The dataset, stored in a CSV file named "jj.csv", sheds light on the company's financial trajectory over two decades.

In contrast, our second study focuses on Amazon, a major e-commerce player. This more granular dataset includes stock market metrics such as opening price, high price, low price, closing price, adjusted close share price, and trading volume. For our analysis, we emphasize the closing price, which encapsulates daily trading information. Spanning from February 20, 2018, to February 17, 2023, this dataset consists of 1259 data points and is recorded in the "AMZN.csv" file, reflecting the dynamic nature of the modern stock market.

Before we delve into the intricacies of modeling, we undertake several crucial data preprocessing steps to ensure the integrity and appropriateness of our dataset. The data is initially loaded into a Pandas DataFrame, providing a structured and convenient format for manipulation and analysis. We meticulously check the dataset for any null values and duplicates. This step is vital to prevent any skewing of results or inaccuracies in the subsequent analysis. The date column is converted into a datetime format and is subsequently set as the index of the DataFrame. This conversion facilitates time-series analysis and ensures that temporal relationships are accurately represented. Utilizing the .describe() method, we obtain a comprehensive summary of the statistics for the datasets. To bet-

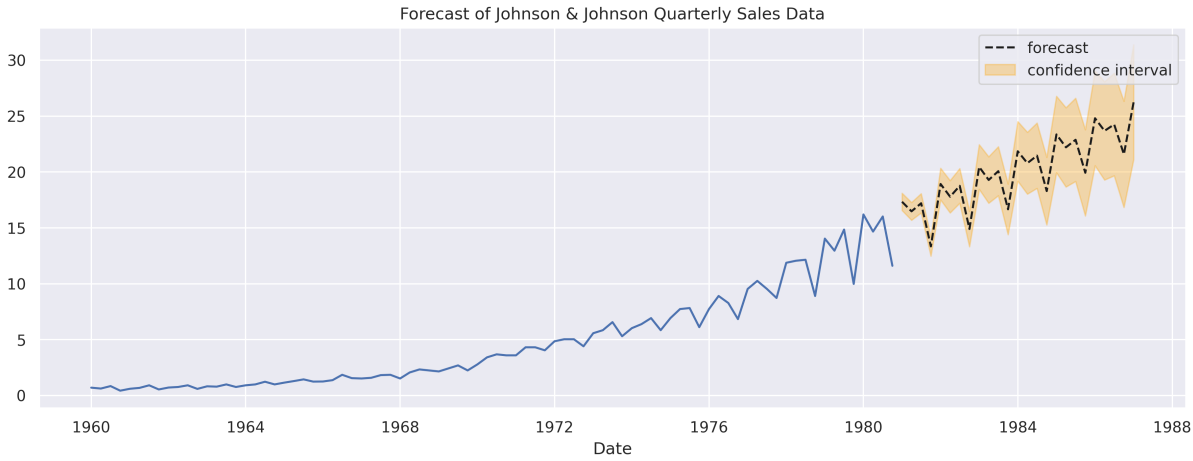

Figure 2: ARIMA Forecast of Johnson & Johnson Quarterly Sales Data

ter understand the trend and fluctuations over time, line plots are created. Histograms are also generated to provide insights into the distribution of values within the dataset.

3 Johnson & Johnson Quarterly Sales Data

3.1 Stationarity

Initially, our analysis began with an examination of the data's auto-correlation. We utilized the plot_acf_pacf function, which we previously defined, to generate the Auto-correlation Function (ACF) and Partial Auto-correlation Function (PACF) plots. As depicted in Figure 1, the ACF plot exhibits an exponentially decreasing trend, resembling a sine wave, while the PACF plot reveals a significant correlation at a lag of 1, followed by an exponential decay. These patterns suggest that the data is non-stationary, necessitating a first-order differencing to attain stationarity. To further substantiate the non-stationary nature of the data, we conducted an Augmented Dickey-Fuller (ADF) test using the adfuller_test function. The resulting p-value was 1.0, significantly exceeding the threshold of 0.05. This outcome provides weak support for the null hypothesis, indicating the presence of a unit root and confirming the data's non-stationarity.

In search of an appropriate transformation to stabilize the variance, we applied a Box-Cox transformation and determined the lambda value to be 0.05. This value, being close to zero, suggests that a logarithmic transformation would be suitable. Subsequently, we visualized the transformed data through line plots and histograms with the plot_line_hist function, which visually affirmed the Box-Cox transformation's efficacy. To stabilize the variance and scale the data, we implemented a logarithmic transformation. Post-transformation, the data was plotted, and the ACF and PACF were re-evaluated. The correlation persisted, as logarithmic transformations do not alter correlation structures.

We then applied a differencing operator to the log-transformed data to eliminate trends and seasonality, thereby achieving stationarity. The line plots, histograms, ACF, and PACF plots all indicated that the data had become stationary. This was further corroborated by a p-value of 0.0004 in the subsequent ADF test.

3.2 ARIMA

Our methodology for determining the optimal order (p, d, q) for the ARIMA model encompassed two distinct approaches. The first approach involved iterating over a range of values for p, d, and q (specifically, from zero to 8 for p and q, and fixed at 1 for d). For each combination, we computed the Akaike Information Criterion (AIC) and identified the model with the lowest AIC value as the most suitable. The second approach utilized the auto_arima function from the pmdarima library, which automates the selection process.

The comparative analysis yielded two potential models: ARIMA(6,1,3) from the first approach and ARIMA(4,1,1) from the second. Upon defining and fitting both models to the actual data, and plotting actual vs fitted data, we observed that ARIMA(6,1,3) model provided a marginally bet-

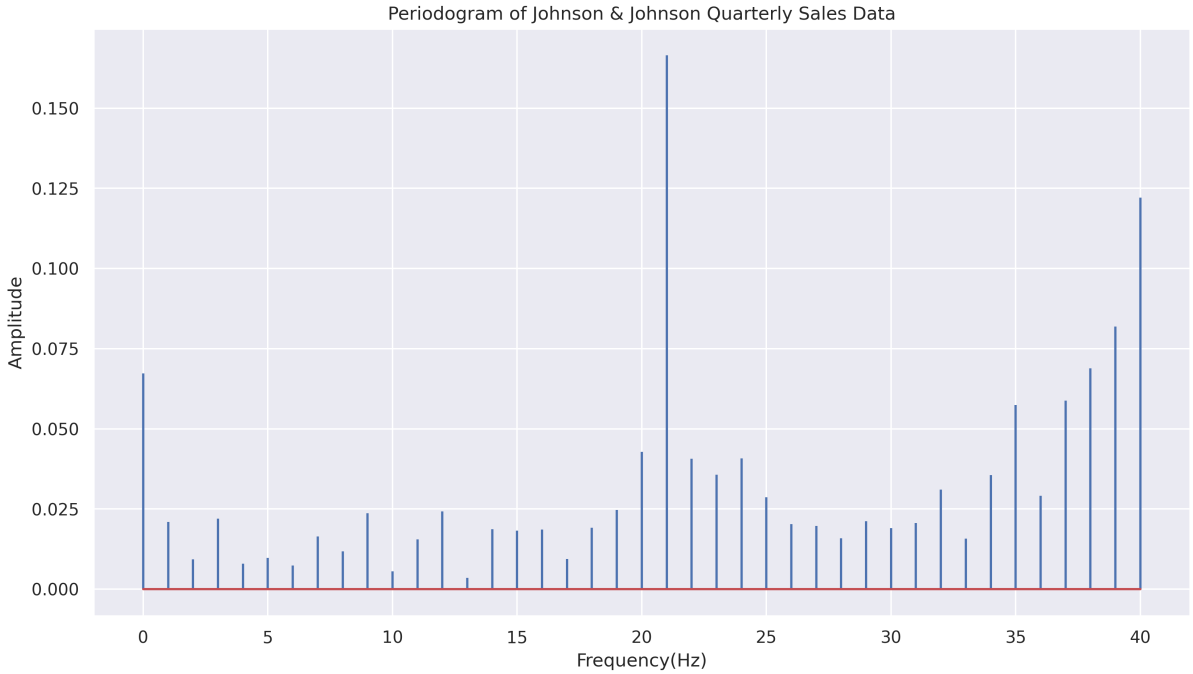

Figure 3: Periodogram of Johnson & Johnson Quarterly Sales Data

ter fit. This observation is supported by the Root Mean Squared Error (RMSE) values, where the RMSE for ARIMA(6,1,3) was 0.4, lower than the RMSE of 0.443 for ARIMA(4,1,1). These findings suggest that the ARIMA(6,1,3) model has superior predictive accuracy and is thus the preferred model for our analysis. We use RMSE as our metric throughout this study as it is the most popular benchmark in time series analysis.

The forecast generated using the ARIMA(6,1,3) model is presented in Figure 2. The projection indicates that the quarterly sales are expected to continue their growth trajectory, consistent with the seasonality previously observed in the data.

3.3 Fourier transforms

The periodogram, as illustrated in Figure 3, provides a spectral analysis of the Johnson & Johnson quarterly sales data. Notable peaks at specific frequencies, particularly at $21Hz$ and $40Hz$, indicate the existence of periodic components within the data. The most pronounced peak at $21Hz$ is indicative of the dominant sales cycle. This frequency corresponds to the most significant regular fluctuation in sales figures, suggesting a strong seasonal pattern.

3.4 LSTM

To prepare our dataset for machine learning, we transformed the time series data into a supervised learning structure. This was achieved by using the timeseries_to_supervised function, which created lagged versions of the differenced logarithmic Johnson & Johnson quarterly sales data. The resulting array was then split into training and test sets, with the latter comprising approximately 20% of the data. We scaled the data to fit within the range of -1 to 1. This normalization ensures that the LSTM model receives data within a consistent range, facilitating better learning and convergence.

Given the sensitivity of LSTM forecasts to initial conditions, we conducted 30 repeats of the experiment to ensure robustness. For each iteration, we first fitted the LSTM model with the scaled training data, then reshaped the training data to match the LSTM's expected input format and used it to build up the model's state for forecasting and lastly, performed walk-forward validation on the test data, making one-step forecasts, inverting the scaling and differencing to obtain predictions in their original scale. After each repeat, we calculated the Root Mean Squared Error (RMSE) to evaluate the model's performance. These RMSE values were recorded and later summarized to assess the overall predictive accuracy of the LSTM model. We

Figure 4: ARIMA Forecast of Amazon's Closing Share Price

compiled the RMSE values into a DataFrame and generated descriptive statistics to summarize the results. A boxplot was created (Figure 5) to visualize the distribution of RMSE values, providing a clear indication of the model's performance variability. The mean RMSE across all repeats was 0.196. This indicates a high level of precision in the model's forecasts. Additionally, we plotted the actual and predicted values of the Johnson & Johnson quarterly sales data to visually compare the LSTM model's forecasts against the true data (Figure 7).

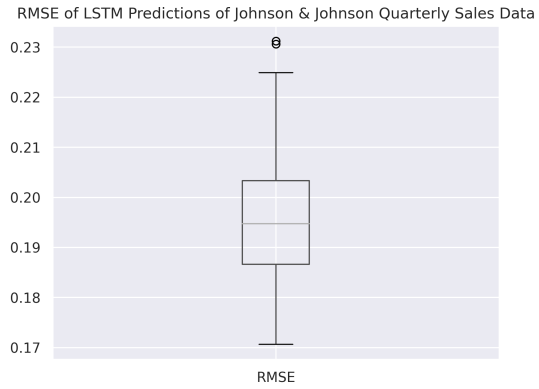

Figure 5: RMSE of LSTM Predictions of Johnson & Johnson Quarterly Sales Data'

4 Amazon Stock Market Share Price

In analyzing the Amazon dataset, we adhered to the same steps as those applied to the Johnson & Johnson data, with a few key differences. We talk about the closing share price when we say amazon data in this report as we use only this for this assignment as mentioned before.

4.1 Stationarity

Similar to the Johnson & Johnson dataset, the ACF plot for the Amazon data exhibits an exponentially decreasing trend, resembling a sine wave. Additionally, the PACF plot reveals a significant correlation at a lag of 1, followed by an exponential decay. These patterns collectively suggest that the data is non-stationary, necessitating a first-order differencing to attain stationarity. ADF test results a p-values of 0.453, reconfirming the non-stationarity of the data.

Unlike the Johnson & Johnson dataset, we opted not to use a logarithmic transformation for the Amazon data. Instead, we directly utilized the Box-Cox transformed data. This decision was based on the nature of the data and the Box-Cox transformation's ability to stabilize variance and make the data more normally distributed and the lambda value is -0.37, which is not that close to zero. Post-transformation, the data was plotted, and the ACF and PACF were re-evaluated. The correlation persisted, as the transformations do not alter correlation structures.

Just like Johnson & Johnson data, We then applied a differencing operator to the log-transformed data to eliminate trends and seasonality, thereby achieving stationarity. The line plots, histograms, ACF, and PACF plots all indicated that the data had become stationary. Additionally, the subsequent ADF test yielded an impressively low p-value of $7.63e^{-26}$.

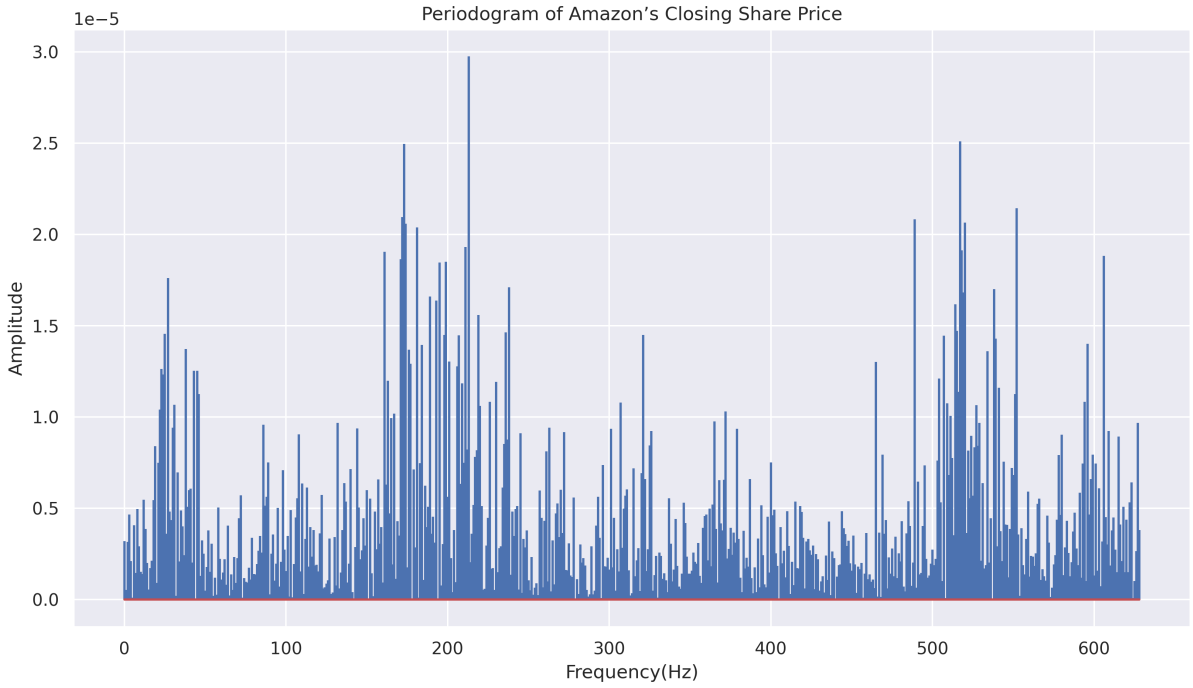

Figure 6: Periodogram of Amazon's Closing Share Price

4.2 ARIMA

In case of amazon data, both the manual iteration approach and the auto_arima function from the pmdarima library gave ARIMA(2,1,2) as the best model. However, the resulting RMSE value was considerably high at 3.438.

The forecast generated using the ARIMA(2,1,2) model is presented in Figure 4. The prediction is an almost straight line.

4.3 Fourier transforms

The periodogram, as illustrated in Figure 6, provides a spectral analysis of Amazon's closing share price. There are several prominent peaks at various frequencies, indicating significant periodic components in the time series data. The presence of these peaks underscores the complexity of Amazon's stock price dynamics. Stock data is inherently volatile, and Amazon's closing share price is no exception.

4.4 LSTM

With the LSTM model, we deviated from the repeated experiment methodology. The initial RMSE obtained was remarkably low at 0.005, suggesting an excellent fit to the data. Given this high level of accuracy, we deemed further repetitions unnecessary.

5 Comparitive Analysis and Future Works

For Johnson & Johnson data, the ARIMA model, with an order of (6,1,3), yielded an RMSE value of 0.4. This model adequately captured the data's seasonality, reflecting the cyclical nature of the sales figures in its forecasts. In contrast, the LSTM model demonstrated a superior fit to the data, with a mean RMSE of 0.196. Despite its more accurate trend predictions, the LSTM model did not fully replicate the seasonality present in the actual data(Figure 7). This discrepancy may be attributed to the model's complexity relative to the dataset's size, which consists of only 84 data points. Given the limited dataset, a simpler neural network architecture, such as the Gated Recurrent Unit (GRU), could be explored in future work. The GRU has the advantage of using less memory and being faster than the LSTM, which may make it more suitable for smaller datasets. This potential avenue for research could provide a balance between model complexity and the ability to capture the essential features of the data, including its seasonality (Pal and Prakash, 2017).

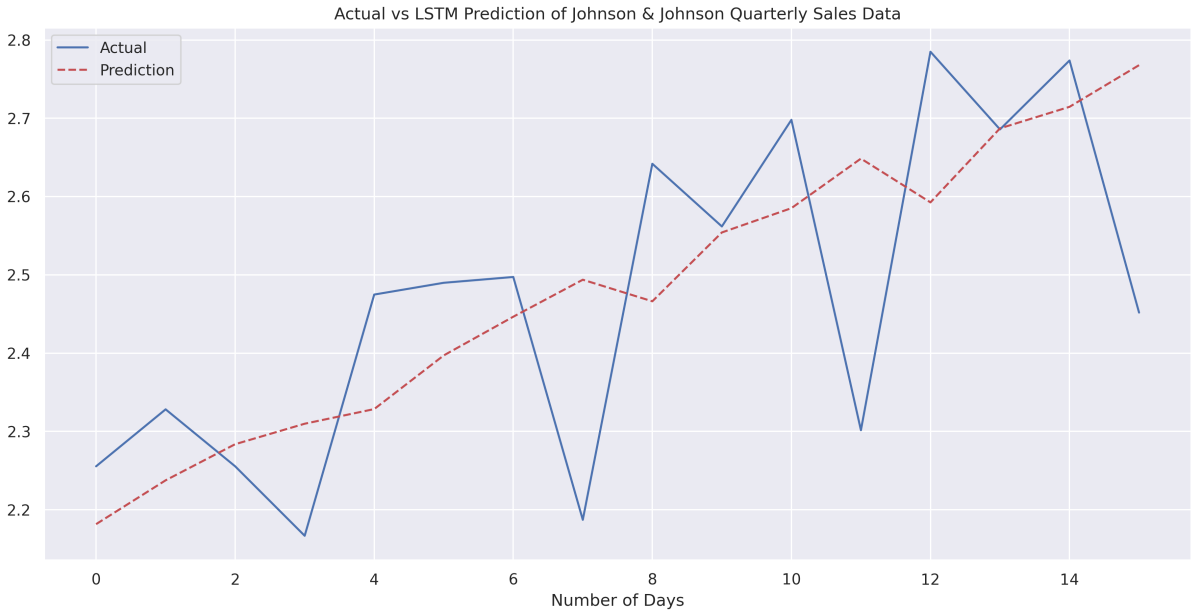

Figure 7: Actual vs LSTM Prediction of Johnson & Johnson Quarterly Sales Data

For Amazon dataset, The ARIMA model yielded an RMSE of 3.438, which is relatively high and indicates a less than satisfactory fit to the data. This suggests that the ARIMA model's capabilities may be insufficient for capturing the intricate patterns and volatility inherent in stock data, resulting in forecasts that do not align well with the actual data. Incoparating exogenous variables (external factors) could perhaps enhance model performance. Conversely, the LSTM model demonstrated exceptional performance, achieving an RMSE of 0.005 on the first attempt. This low RMSE value signifies a highly accurate model that fits the data well. Given the size and complexity of the dataset, the LSTM's intricate neural network architecture is well-suited to model and predict such volatile time series data. While the basic LSTM model used in our study was highly effective, there is potential for further improvement. A more sophisticated LSTM model, tailored and tuned specifically for stock market data, could potentially yield even more accurate forecasts. Such a model would take into account the intricate patterns and rapid fluctuations characteristic of stock data, providing a powerful tool for predictive analysis (Pal and Prakash, 2017).

6 Conclusion

The analysis of Johnson & Johnson's quarterly sales data revealed that while the ARIMA(6,1,3) model can adequately captured the seasonality present in the data, the LSTM model achieves a lower mean RMSE of 0.196. Conversely, for the more volatile Amazon stock market share price data, the LSTM model outperformed the ARIMA(2,1,2) model by a substantial margin, with an exceptionally low RMSE of 0.005, showcasing its ability to handle complex, non-linear patterns. While ARIMA models may excel in capturing well-defined seasonal patterns, LSTM networks excel at modeling intricate, non-stationary time series, making them a powerful tool for applications such as stock market forecasting.

References

Brockwell, P.J. (2004), *Introduction to Time Series and Forecasting*, Springer, Dordrecht. Available at: https://ebookcentral.proquest.com (Accessed: 01 May 2024).

Cahuantzi, R., Chen, X. and Güttel, S. (2021) *A comparison of LSTM and GRU networks for learning symbolic sequences*. CoRR abs/2107.02248. Available at: https://arxiv.org/abs/2107.02248 (Accessed: 01 May 2024).

Pal, A. and Prakash P.K.S. (2017). *Practical time series analysis: master time series data processing, visualization, and modeling using Python*. Packt Publishing Ltd. Available at: https://learning.oreilly.com (Accessed: 01 May 2024).

Petneházi, G. (2019). *Recurrent Neural Networks for Time Series Forecasting*. CoRR abs/1901.00069. Available at: https://arxiv.org/abs/1901.00069v1 (Accessed: 01 May 2024).