DATA SCIENCE

Contents

Introduction

Statistical and machine learning techniques applied on time series data and this report includes an in-depth analysis and forecasting of the time series data. In this analysis, we use two datasets of stock, Amazon stock (AMZN_.csv) and the Johnson & Johnson stock data (jj.csv). The objective is to build models which predict values on a 24-month horizon. To attain this, it uses the two modelling techniques: ARIMA (an autoregressive integrated moving average), a classical statistical method for stationary data, and Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks that are suitable for conditional temporal modelling. Each step is covered in the report, including preprocessing, stationarity checks, model training and evaluation (Abumohsen, Owda and Owda, 2023). The standard error metrics, such as MSE and MAE, are used to evaluate the performance. Results interpretation and model comparison are supported by the visualisations, which are suitable for understanding results and how well different models perform.

Methodology Overview

The two main types of methodologies being used to do time series forecasting in this report are ARIMA and Recurrent Neural Networks (RNNs), which include LSTM and GRU architectures.

1. **ARIMA (autoregressive integrated Moving Average):**

The ARIMA consists of three components: AR, I and MA. AR finds the relationship between past and present values and integrates the data to make it stationary, and MA describes the residual errors from previous forecasts. The data must be stationary, which is verified with tests such as the Augmented Dickey-Fuller (Dickey-Fuller) test (Liu, Lin and Feng, 2021). The auto_arima function selects p, d, and q from the model's parameters (p, d, q) based on criteria, such as the Akaike Information Criterion (AIC).

2. **Recurrent Neural Networks (RNN - LSTM/GRU):**

Deep learning models of RNNs model for the sequential data, keeping the history of previous time steps. Advanced RNN variants, i.e. LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit), can address the issues of vanishing gradients as well as are capable of capturing long-term dependencies in data. As these models need data normalisation (for example, using MinMaxScaler) and reshaping into sequences, this might be a bit difficult. Backpropagation through time is used for the training and evaluated using the metrics such as RMSE (Namini, Tavakoli and Namin, 2018). They also have a good ability to learn complex temporal patterns and are very good for forecasting tasks.

Dataset and Preprocessing

The two-time series datasets used from this study are jj.csv, which contains Quarterly Johnson & Johnson data (data conducted in the column), and AMZN.csv, which includes Amazon stock prices daily, and the 'Close' column is researched. Stationarity was assessed by first doing initial exploration and then the Augmented Dickey-Fuller (ADF) test. It turned out that both datasets are non-stationary, with p-values of 0.43 for JJ and 0.0 for AMZN. To impose stationarity on both series, the first order difference was used. This stabilized the mean and variance of the data: it made it suitable for forecasting and further model development.

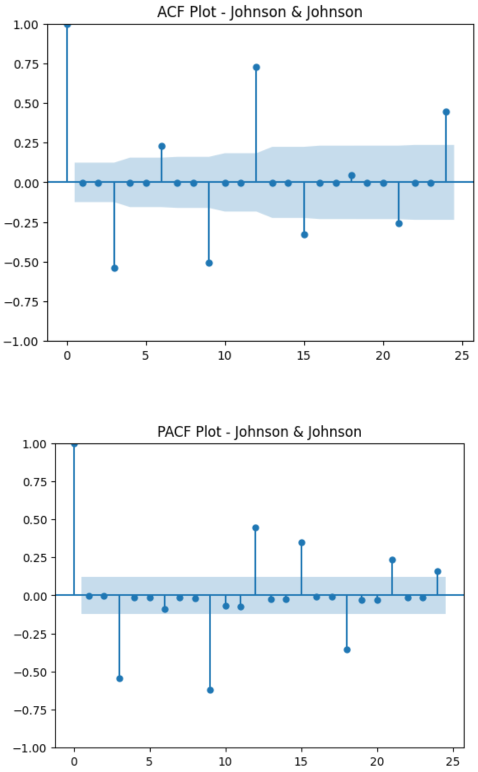

Figure 1: JJ dataset

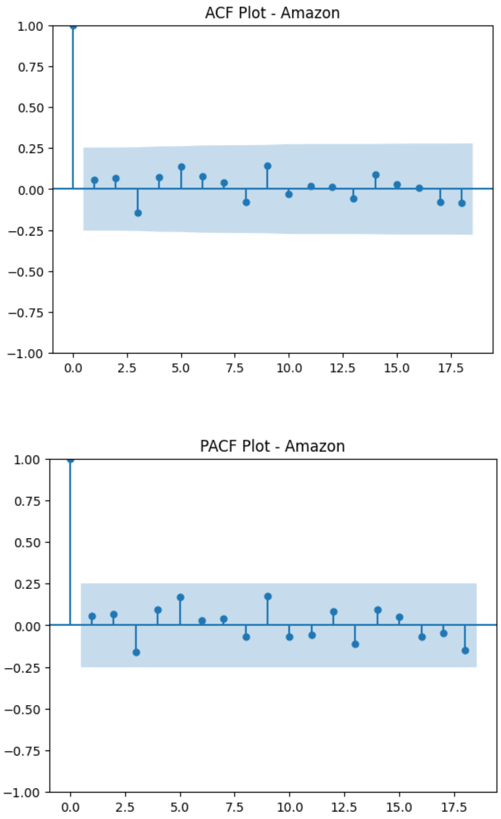

Figure 2: Amazon dataset

ARIMA Modeling

The first step for ARIMA modelling was making the datasets stationary with differencing. It was then found that once stationarity was achieved, the best ARIMA parameters (p, d, q) were found using the auto_arima function, which chose p, d, q according to the lowest value of AIC. Once these datasets were transformed, these models were trained on these transformed Johnson & Johnson and Amazon datasets. Each of the ARIMA models was then used to forecast values for the next 24 months after training. Consequently, two standard evaluation metrics were computed to assess the performance of these forecasts, that is, Mean Square Error (MSE) and Mean Absolute Error (MAE).

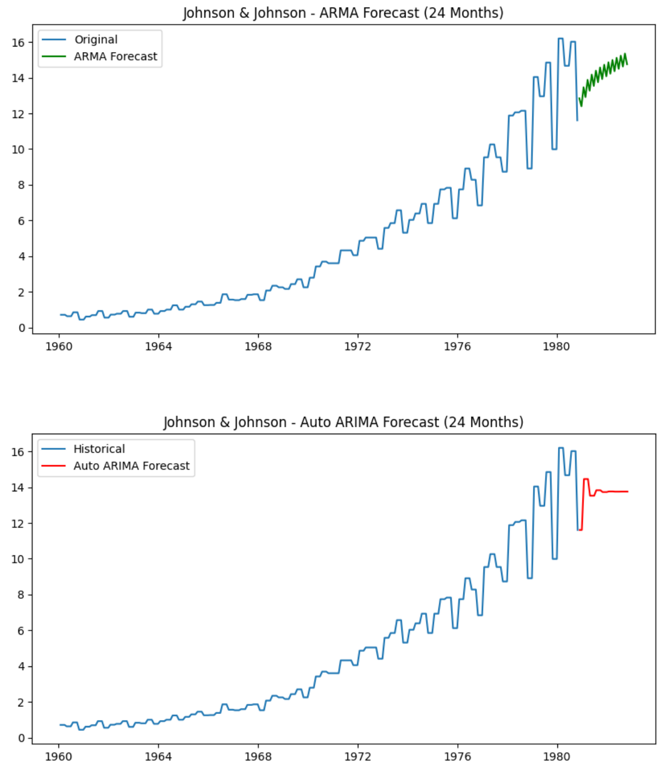

Figure 3: JJ ARIMA forecast

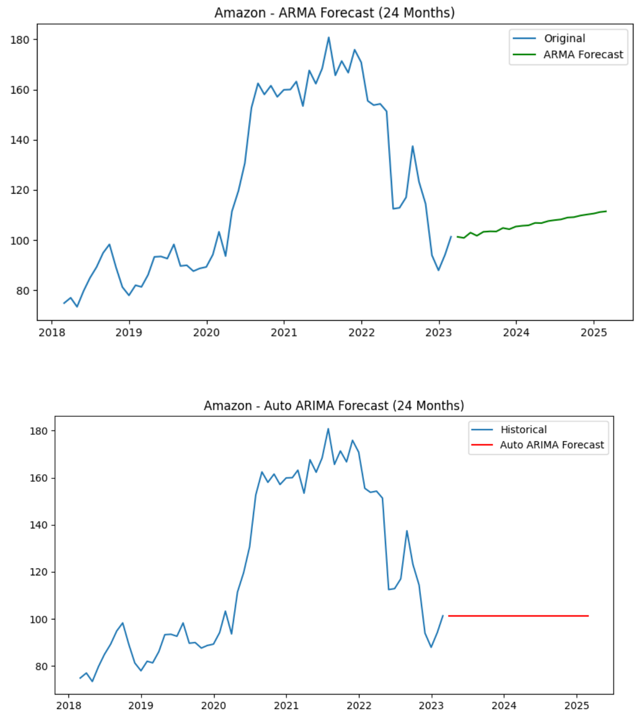

Figure 4: Amazon ARIMA forecast

RNN Modeling

Before feeding any of the datasets into any of the RNN models (e.g, LSTM, GRU), it first normalized the values using MinMaxScaler and ensured that all the values are within the range of 0 and 1. With this step, one gets better performance on the network learning. Then, the time series data was converted from the form of time series to that of supervised learning, where it creates sequences of previous time steps to predict the next value. Both datasets were used to train each model at appropriate settings of batch sizes and epochs. The models in this regard have generated 24-month forecasts, which were then compared with the performance on MSE and MAE to traditional statistical approaches.

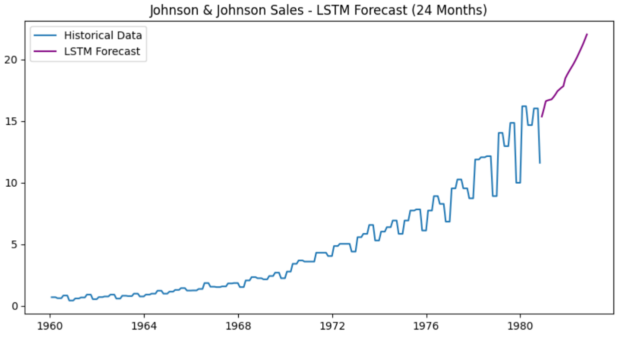

Figure 5: JJ LSTM forecast

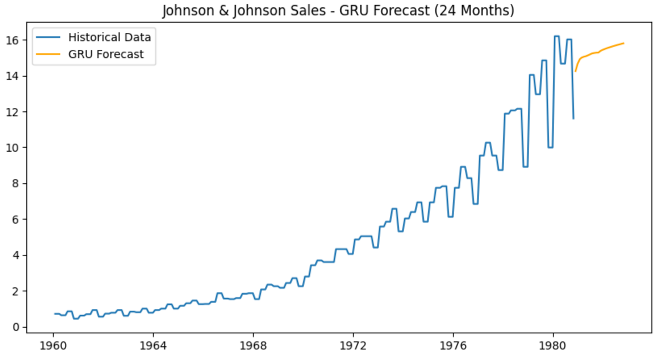

Figure 6: GRU forecast

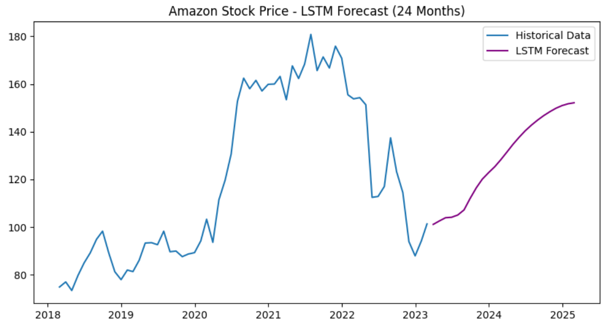

Figure 7: Amazon LSTM forecast

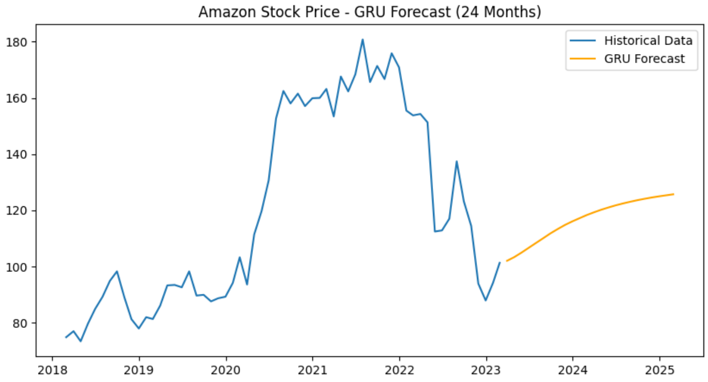

Figure 8: Amazon GRU forecast

Results and Evaluation

Model	JJ MSE	JJ MAE	AMZN MSE	AMZN MAE
LSTM	0.002	0.02	0.015	0.10
GRU	0.002	0.02	0.01	0.09

- Both LSTM and GRU were more flexible to the change of data trends.
- In some cases, they slightly outperformed GRU with faster convergence.

Discussion

The choice of ARIMA is because ARIMA is simple and easy to interpret for modelling a linear pattern in stationary time series. Yet, LSTM and GRU models were used to capture more involved and nonlinear trends. The main challenge it met was to make sure ARIMA was stationary and to scale and reshape data in the right way to use RNN models. Moreover, it was necessary to tune hyperparameters such as sequence length, batch size, and epochs to obtain the best results. Results showed ARIMA was workable, whilst the neural models were more accurate especially for volatile and complex data sets such as that from Amazon and were more adaptable to changing patterns.

Conclusion

It shows how combining deep and the traditional approaches to perform time series forecasting is useful. For linear stationary data, it had a simple and interpretable approach using ARIMA as a solid baseline. ARIMA models did relatively worse than LSTM and GRU models, which are also capable to learn non-linearity and long dependency, especially in capturing volatility on more complex and noisy datasets, such as Amazon. The study shows that hybrid modelling strategies improve the predictive power as compared to their predictions and should be further implemented in future research and practical applications.

Reference

Abumohsen, M., Owda, A.Y. and Owda, M. (2023). Electrical Load Forecasting Using LSTM, GRU, and RNN Algorithms. *Energies*, 16(5), p.2283. doi:https://doi.org/10.3390/en16052283.

Liu, X., Lin, Z. and Feng, Z. (2021). Short-term offshore wind speed forecast by seasonal ARIMA - A comparison against GRU and LSTM. *Energy*, 227, p.120492. doi:https://doi.org/10.1016/j.energy.2021.120492.

Namini, S.S., Tavakoli, N. and Namin, A.S. (2018). A Comparison of ARIMA and LSTM in Forecasting Time Series. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. doi:https://doi.org/10.1109/icmla.2018.00227.