An ARIMA-LSTM model for predicting stock return prices with random forest Technique

Divyanshi Sharma

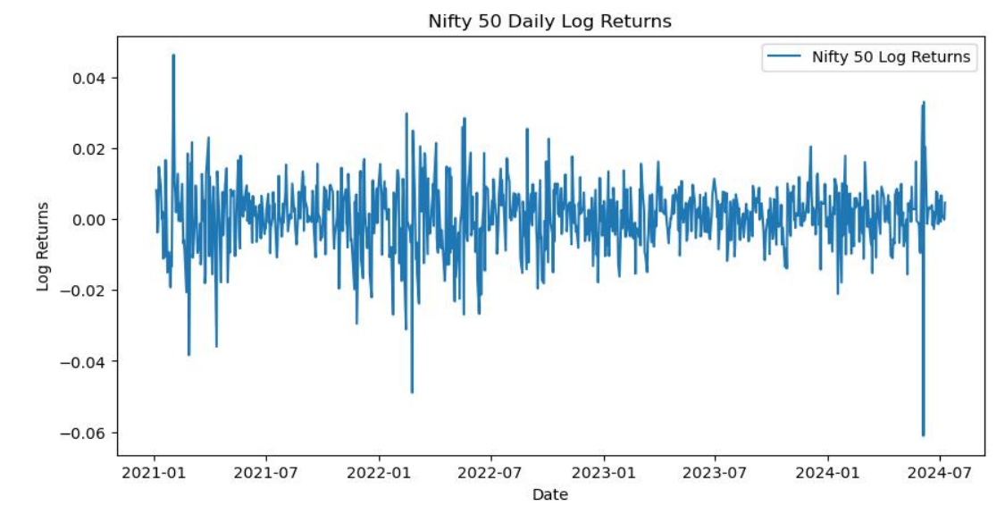
Guided By: Prof. Sanjiv Kumar, Department of Economic Sciences

Introduction

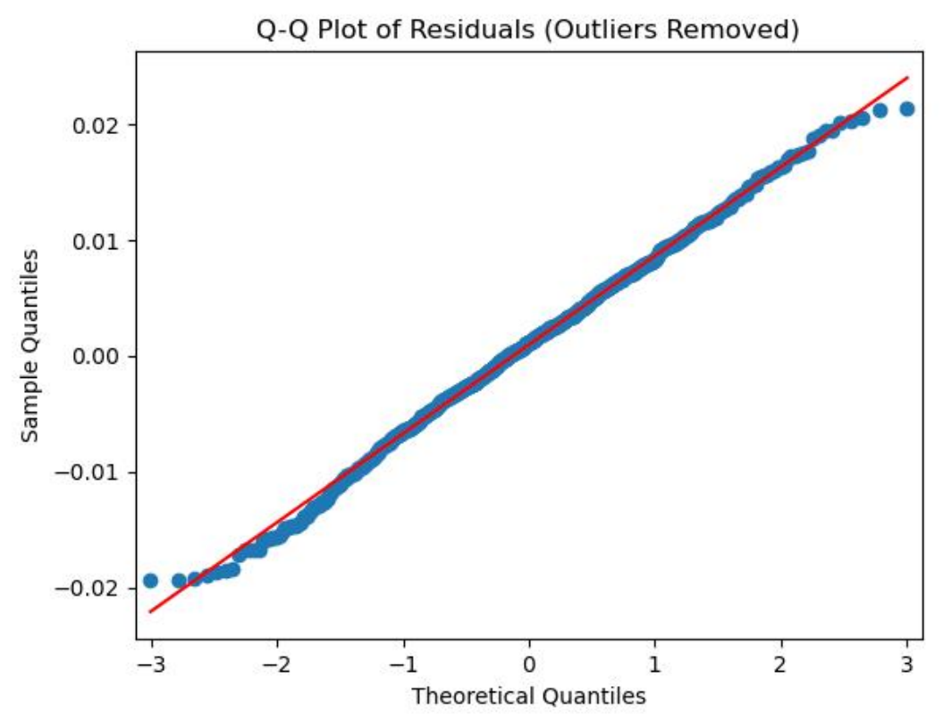
Machine learning mechanism is establishing itself as a promising area for modelling and forecasting complex time series over conventional statistical models. In this article, focus has been made on presenting a machine learning algorithm with special attention to deep learning model in form of a potential alternative to statistical models such as Autoregressive Integrated Moving Average (ARIMA) and ARIMA-Generalised Autoregressive Conditional Heteroscedasticity (GARCH) models. Further, an improved hybrid ARIMA-Long Short-Term Memory (LSTM) model based on the random forest lag selection criterion has been introduced. ARIMA model has been used to estimate the mean effect and the GARCH model is employed with the residuals obtained from the ARIMA model to estimate the volatile behaviour of the series. ARIMA-GARCH models act as superior statistical models over ARIMA models based on the lowest AIC and BIC values. LSTM model is employed on all normalised training data series. After which we built a comparison scenario independently between ARIMA, ARIMA-GARCH, LSTM and ARIMA-LSTM models on forecasting accuracy.

ARIMA Model

Before the development of the ARIMA model, it was deemed necessary to ensure the stationarity of all data series. The Augmented Dickey-Fuller (ADF) test was employed for this purpose. The results indicated that all data series become stationary at their level point after taking their log. Based on this evidence, the order of d was set to be 1 in the ARIMA model. The best-fit ARIMA models, based on the estimated parameters and standard errors, were found to be ARIMA(2,1,2). The autocorrelation of the residuals at zero at different lags was confirmed by the p-value (greater than 0.05) of the Ljung-Box Q test. This indicated that the residuals followed a normal distribution with zero mean and constant variance, which was further confirmed through the visualisation of histograms and Q-Q plots.



```
ADF Statistic: -14.88356682271712
p-value: 1.595933773432541e-27
```

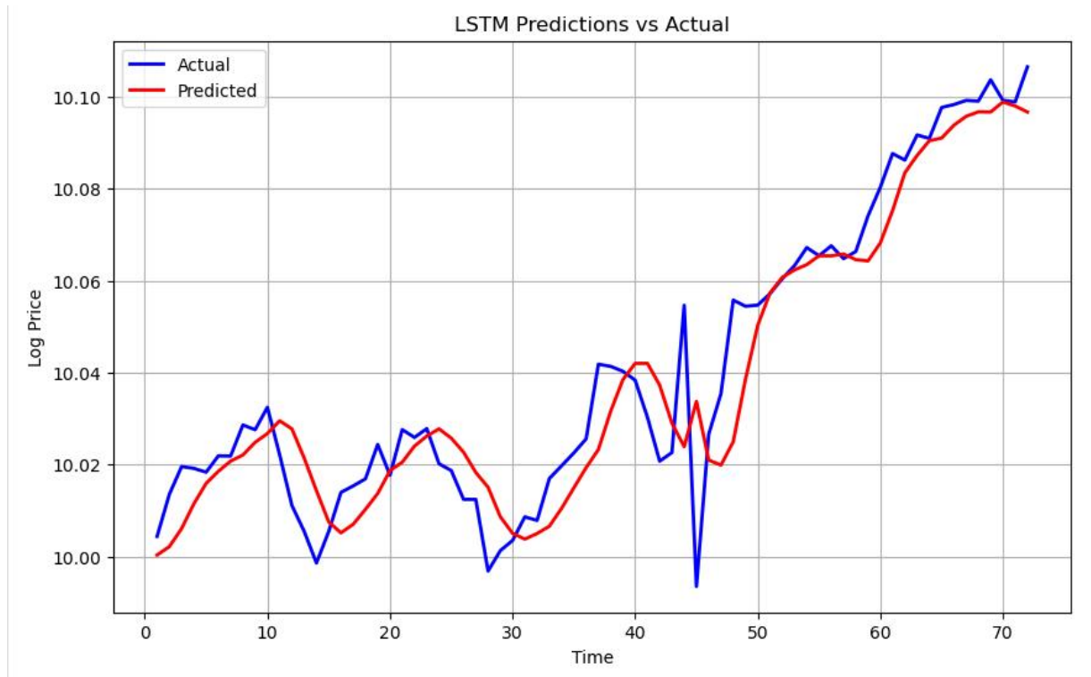


ARIMA-GARCH Model

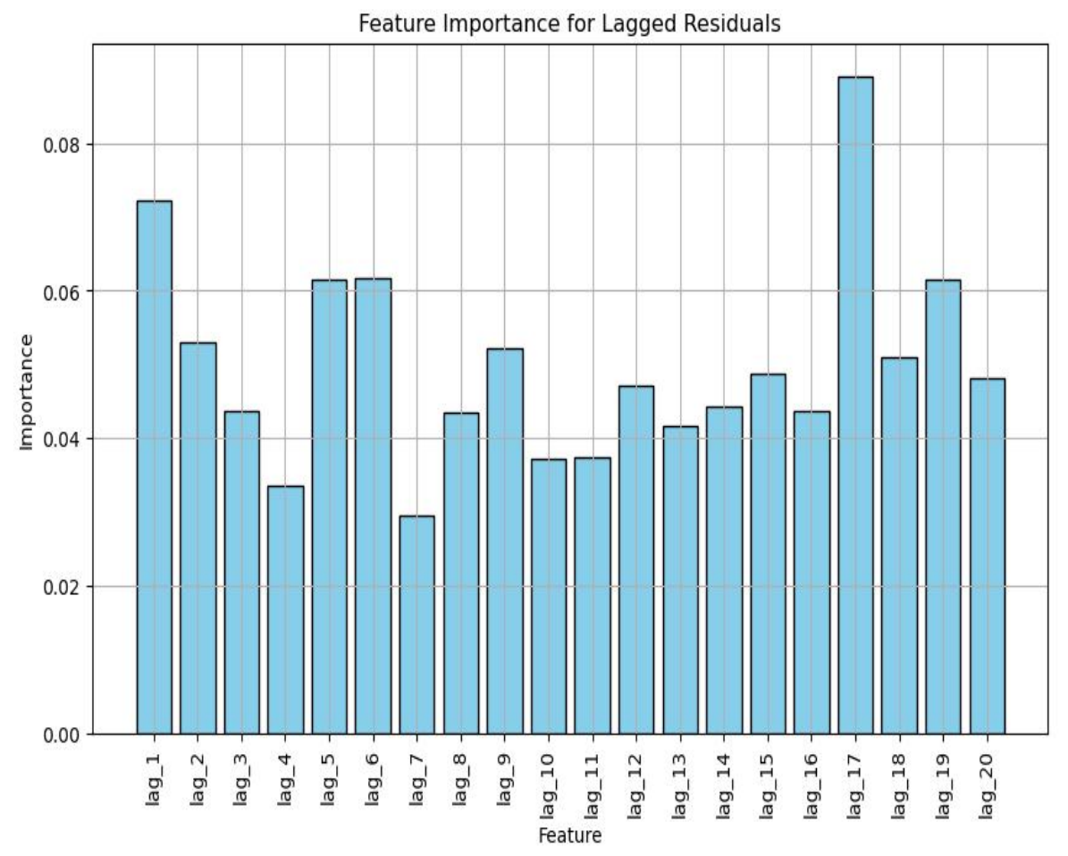
An ARCH LM test was performed on the residuals obtained from the ARIMA model to detect the presence of volatility. All the p-values obtained were less than 0.05 from the test at different lags (10, 15 and 20), which confirmed the presence of volatility in the residuals. So the GARCH model was employed for the trained residuals of the ARIMA model. The model was developed as the ARIMA-GARCH model. Best GARCH parameters based on AIC: p_garch: 5.0, q_garch: 2.0, with the normality of the residuals again confirmed by the Ljung-Box Q test as p-value was greater than 0.05. The mean effect of the series (ARIMA) and volatile effect (ARIMA-GARCH) were compared with model selection criteria, the lowest value of AIC. From the table, ARIMA-GARCH models obtained lower AIC value than the ARIMA models for all the series in training data. This confirmed that the ARIMA-GARCH model was superior for capturing volatility leading to a better prediction for the test data series.

Proposed Model (ARIMA-LSTM based on random forest)

Based on the objective of the study, the hybrid ARIMA-LSTM model was developed by estimating the input lag length selection using traditional ACF and random forest approaches. ARIMA model squared residuals were used to estimate input lag through ACF and random forest technique. Based on these two techniques, we chose the relevant input lag and then implemented the LSTM model for residual prediction. It is noteworthy that for each series the number of lags selected from the random forest technique was smaller as compared to that of ACF. This directly reduces the number of parameters to be estimated and the computational time as well.



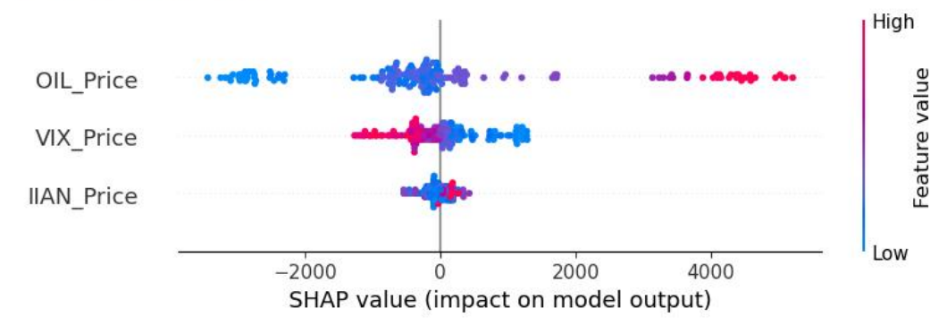
```
R2 Score for LSTM: 0.8871736088919062
MAE for LSTM using ACF: 0.007780776463036703
MSE for LSTM using ACF: 0.00011512731308893486
RMSE for LSTM using ACF: 0.010729739656158246
MASE for LSTM using ACF: 0.523480432994946
```



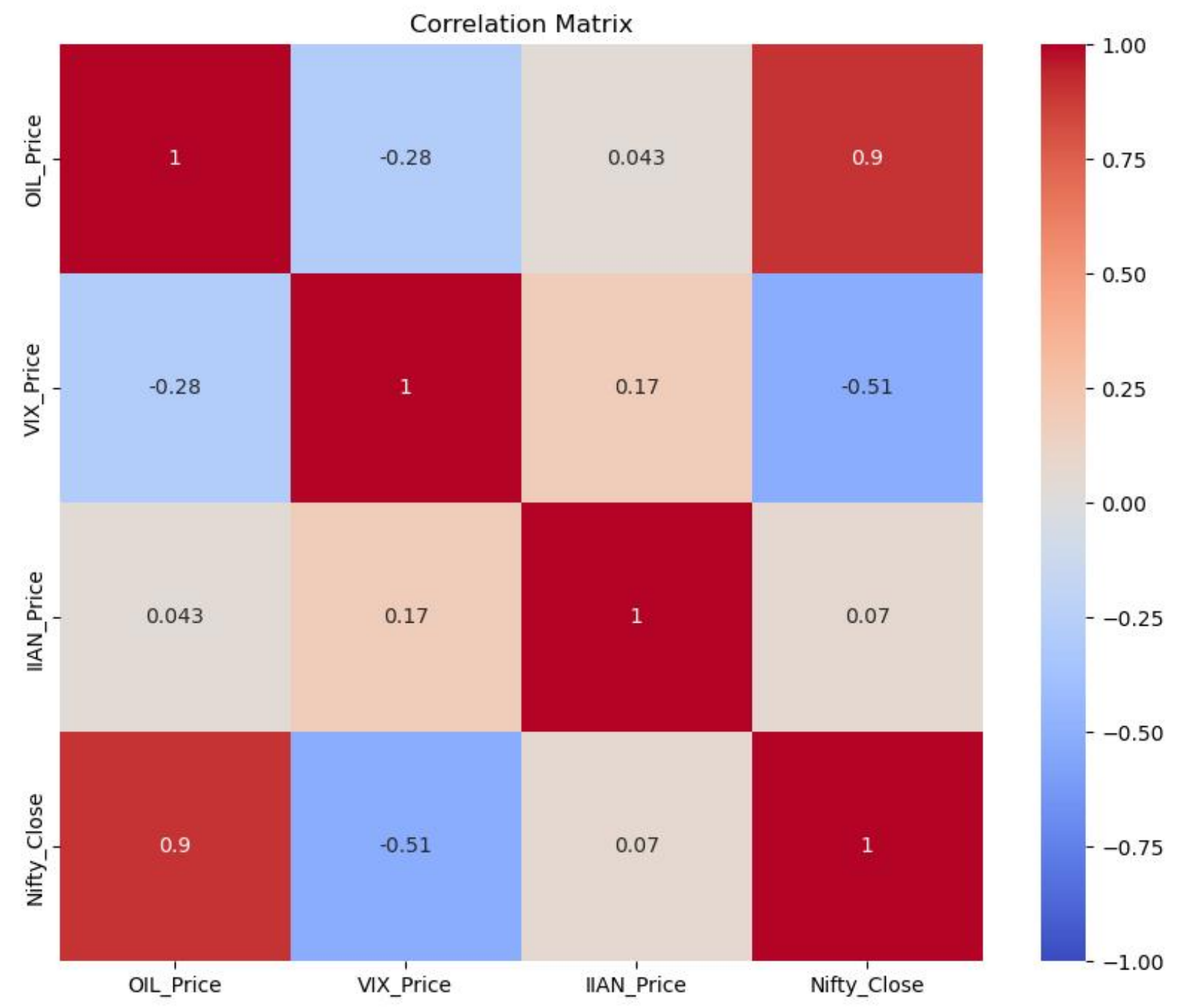
Gradient Tree Boosting

In my analysis, I applied Gradient Boosting, a powerful machine learning technique known for its superior predictive performance. This model was chosen because it effectively incorporates and learns from independent variables to improve accuracy. For this analysis, I selected the Oil Volatility Index, the Volatility Index, and the Indian Energy Exchange as the independent variables (x), while the dependent variable (y) was the Nifty50 stock closing prices. By leveraging the strengths of Gradient Boosting, which iteratively builds and optimises the model, I achieved an impressive accuracy of 98.7%, significantly outperforming other models in capturing the complex relationships within the data.

```
Mean Squared Error: 76410.50193832375
R2 Score: 0.9870255867108301
```



```
             OIL_Price  VIX_Price  IIAN_Price  Nifty_Close
OIL_Price     1.000000  -0.278330    0.042705     0.899364
VIX_Price    -0.278330   1.000000    0.174016    -0.510420
IIAN_Price    0.042705   0.174016    1.000000     0.069825
Nifty_Close   0.899364  -0.510420    0.069825     1.000000
```



Conclusion

In this present investigation, we have introduced a novel and efficient random forest-based ARIMA-LSTM hybrid model. The models selected were ARIMA, ARIMA-GARCH, LSTM and ARIMA-LSTM owing to their wide acceptability in literature for satisfactorily modelling and forecasting financial time series. We identify ARIMA-GARCH as better suited to model the data sets than ARIMA due to lower AIC values. For forecasting volatility of the price series, ARIMA-GARCH model performed uniformly superior to its competitor ARIMA model in terms of lower RMSE, MAPE and MASE values. Further, the results obtained from LSTM were compared with that of the statistical models. Lastly, Gradient Tree boosting showed superior results as it incorporates and learns from independent variables to improve accuracy.