Stock Price Prediction using Hybrid ARIMA-LSTM model

1 Learnings from the project

This project gave exposure to various concepts in Statistics and Machine Learning.

1.1 Statistics

1.1.1 Fundamentals

This project introduced topics such as random variables, covariance, correlation, probability distribution functions, skewness, etc., which are fundamentals of statistics.

1.1.2 Hypothesis Testing

Hypothesis testing is used to check the overall significance of the regression model or the significance of each coefficient of the regression model.

- **F-Test**: Used to check the significance of the whole model.

$$F = \frac{(RSS_0 - RSS_1)/k}{RSS_1/(n-k-1)} \tag{1}$$

- **Z-Test**: Used to check the significance of the coefficient of the model.

$$z = \frac{\hat{W}_j}{SE(\hat{W}_j)} \tag{2}$$

1.1.3 Time Series Analysis

Time Series are used to model data over time like stock prices or sales of products.

Components of Time Series:

- **Trend**: Long-term progression of series.
- **Seasonality**: Regular and repeating patterns that occur at fixed intervals.
- **Cyclic**: Long-term oscillations that are not fixed in period.
- **Noise**: Random and unpredictable variations in data.

Moving Average: Used to estimate the trend of a series by taking simple or weighted averages of a small set of consecutive data.

Lag Operator: Shifts data by one time step.

$$LY_t = Y_{t-1} \tag{3}$$

$$L^kY_t = Y_{t-k} \tag{4}$$

Autocorrelation Function (ACF): Measures correlation between observations as different lags.

$$\rho_k = \frac{Cov(Y_t, Y_{t-k})}{\sigma^2} \tag{5}$$

Stationarity: A time series is stationary if its mean and variance are constant and covariance depends only on k.

1.1.4 Time Series Models

- **AR Model**: Current value of the series depends linearly on previous values.

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \epsilon_t \tag{6}$$

- **MA Model**: Current value is a linear combination of past errors.

$$X_t = \mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q} + \epsilon_t \tag{7}$$

- **ARMA Model**: Combination of AR and MA model.
- **ARIMA Model**: Adds differencing in ARMA model for removing non-stationarity. **PACF plot**: Used to find p. **ACF plot**: Used to find q.

1.2 Machine Learning

1.2.1 Regression Model

Implemented linear regression model and polynomial regression model.

1.2.2 Neural Networks

Neural Networks: Learns pattern from data.

Recurrent Neural Networks: Type of Neural Network used to work with sequential data.

LSTM: To solve the problem of vanishing or exploding gradient of RNNs.

ARIMA-LSTM Hybrid Model: Using LSTM network to predict residuals of ARIMA model predictions.

2 Implementation of Hybrid ARIMA-LSTM Model

This model uses the ARIMA model and the residuals left from the ARIMA model are used to train an LSTM model. The final Prediction would be the sum of ARIMA prediction and LSTM model prediction.

2.1 Data Preprocessing

Stock price data was taken from the yfinance library. Stocks used for prediction were TCS *(TCS.NS)* and NIFTY50 *(ˆNSEI)*. Data used was for a period of 10 years.

Data was split into 80:20 for training and testing.

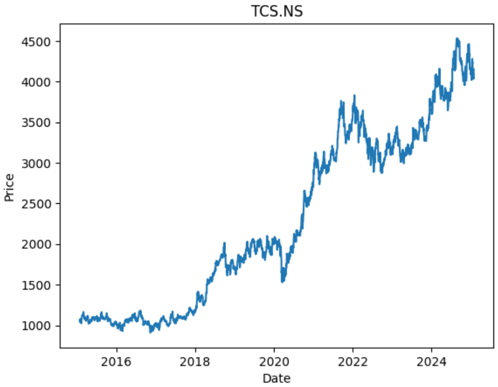

Figure 1: TCS

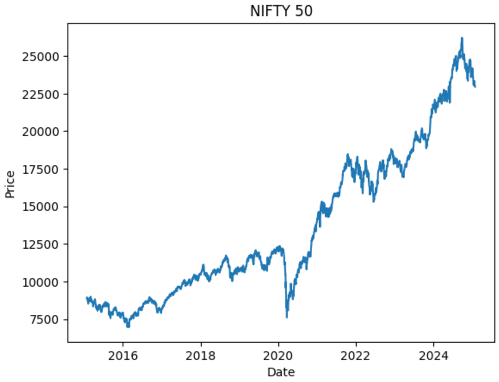

Figure 2: NIFTY50

2.1.1 Stationarity Check

Stationarity of a time series is checked by the augmented Dickey-Fuller test (ADF test).

Stationarity check showed that both the time series were not stationary. Differencing of order 2 made both of them stationary in the acceptable range.

2.1.2 ACF and PACF

ACF and PACF plots of the differenced data were used to find values of parameters p and q. Both the time series had similar plots where ACF had the third spike in the acceptable range and PACF plot had 4th spike near the acceptable range.

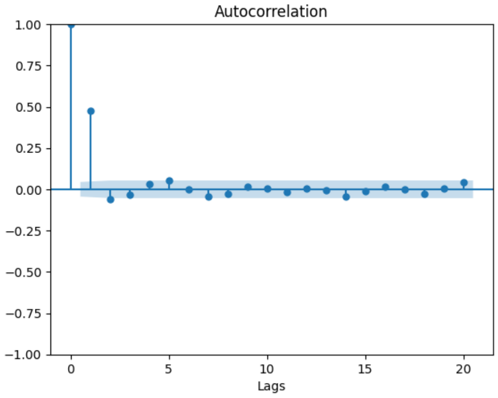

Figure 3: TCS

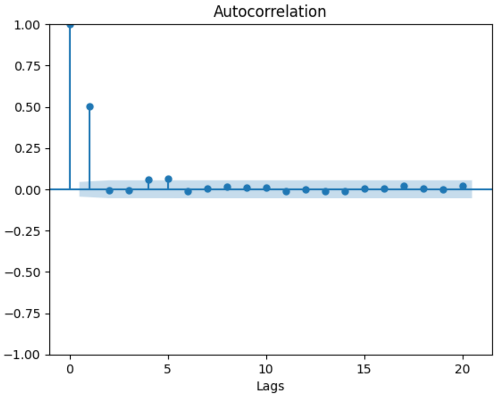

Figure 4: NIFTY50

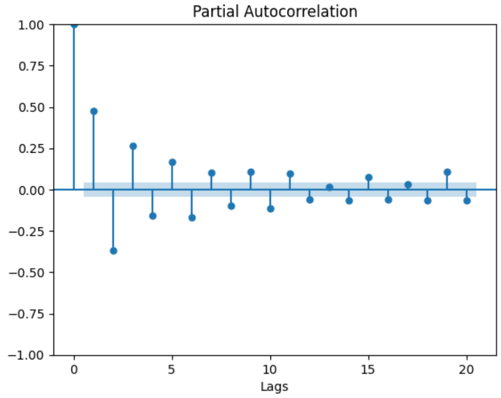

Figure 5: TCS

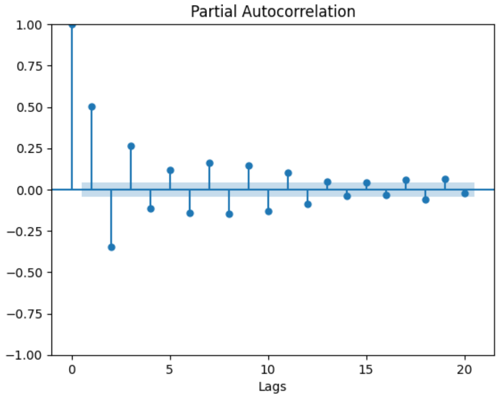

Figure 6: NIFTY50

2.2 ARIMA model

ARIMA model was used from *statsmodel* library, whose parameters were calculated before. SARIMA model performed better due to it considering seasonality in its prediction. The model was fitted with train data and predictions were made on test data and forecasted values till 31-12-2025.

2.2.1 Result

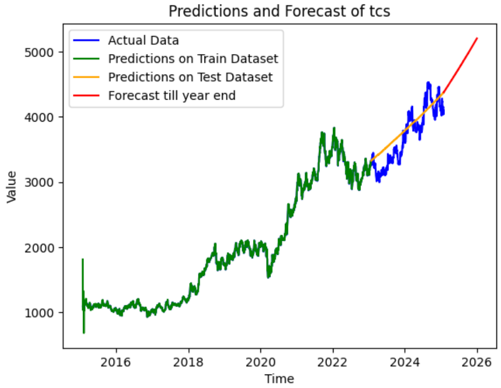

Figure 7: TCS

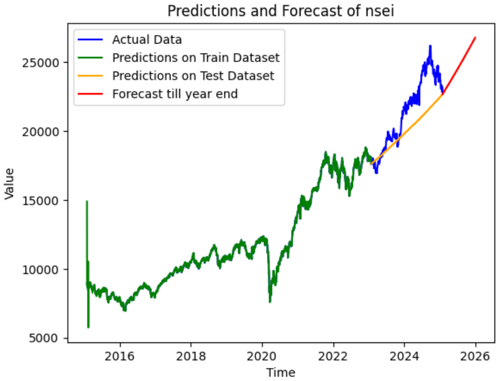

Figure 8: NIFTY50

RMSE, *MASE*, and *MAPE* were calculated on predictions of test data.

	RMSE	MASE	MAPE
TCS	230.22	82.14	5.60
NIFTY50	1854.59	130.18	6.43

Table 1: Errors from ARIMA model.

2.3 Hybrid ARIMA-LSTM model

The LSTM model was trained on a sequence of residuals left from the ARIMA model to predict the next residual in the sequence.

The LSTM model was made using *PyTorch* having 1 input, 1 output, and 15 hidden dimensions.

2.3.1 Residuals Processing

Residuals were calculated as the difference of predictions of the ARIMA model and test data values.

$$Residuals = Data - Predictions \tag{8}$$

Residuals were normalized by min-max scaling.

$$Residuals\ Normalized = \frac{Residuals - min(Residuals)}{max(Residuals) - min(Residuals)} \tag{9}$$

Residuals were also split into smaller sequences of length 10 and randomly shuffled and 10% of sets were used for training the LSTM model.

2.3.2 Model implementation

The model was trained for 20 epochs. For predictions on test data residuals were calculated according to the formulae above. For forecasting, residuals were previous predictions of the LSTM model itself.

The final prediction was the sum of the ARIMA model prediction and LSTM model prediction.

$$Final\ Prediction = ARIMA\ Prediction + LSTM\ Prediction \tag{10}$$

2.3.3 Result

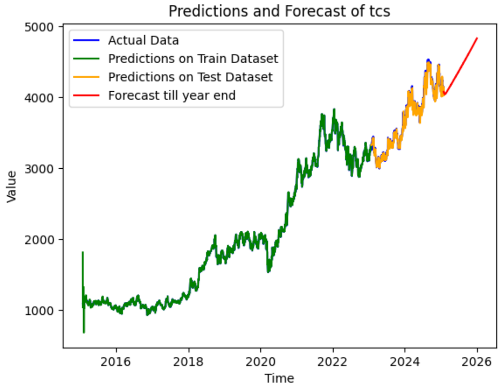

Figure 9: TCS

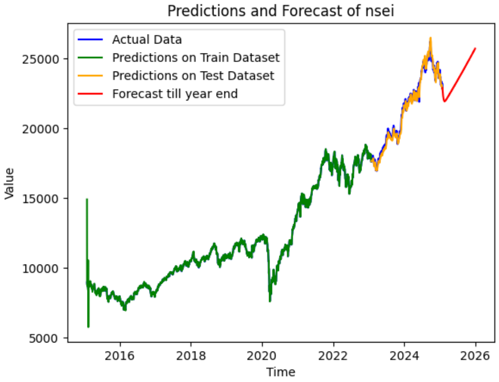

Figure 10: NIFTY50

	RMSE	MASE	MAPE
TCS	53.57	1.08	1.03
NIFTY50	245.11	2.36	0.92

Table 2: Errors from Hybrid model.

3 Conclusion

3.1 Predictions

There was a significant improvement in the predictions of the model considering only 10% of residual data was used for training the LSTM model.

	Model	**RMSE**	**MASE**	**MAPE**
TCS	*Hybrid*	53.57	1.08	1.03
	ARIMA	230.22	82.14	5.60
NIFTY50	*Hybrid*	245.11	2.36	0.92
	ARIMA	1854.59	130.18	6.43

Table 3: Comparison between models

3.2 Forecasting

Since residuals from the ARIMA model consist of random noise, the LSTM model worked well when real data was given to it. However, since real data was available for forecasting, the residuals predicted by the model would eventually converge to a specific value. This created just a shift between predictions of the ARIMA model and the Hybrid model.

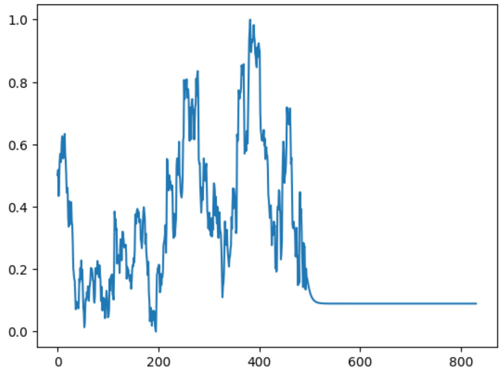

Figure 11: TCS Residuals

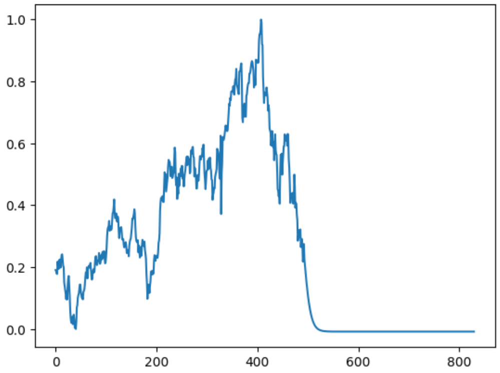

Figure 12: NIFTY50 Residuals

Residuals worked well only for a short period in forecasting where real data had a bit of influence on the model.

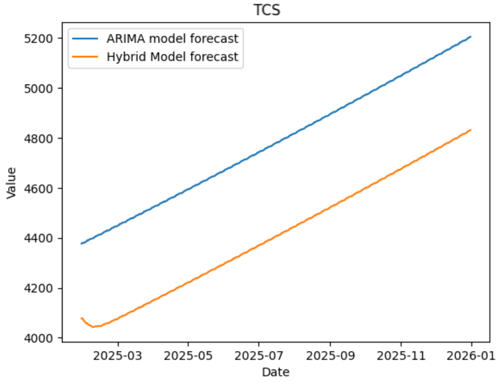

Figure 13: TCS

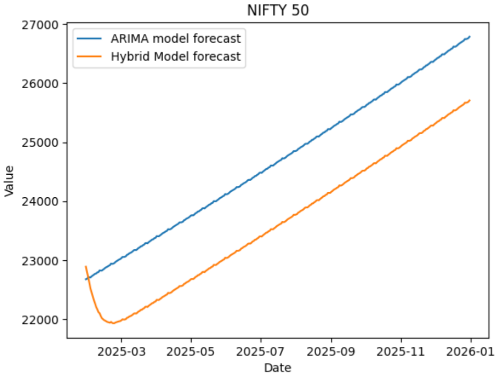

Figure 14: NIFTY50

Forecast Analysis: Both the stocks will increase till the year-end. TCS is expected to reach nearly 4800 and NIFTY50 to cross 25000 in value. Therefore, considering their expected growth, buying these stocks could be a strategic investment decision.