Deep Learning and ARIMA modelling analysis for SBI stock price prediction: A Comparative study for Risk Management and Investment decision

*Harshit Kumar Singhal
**Dr. Gaurav Chopra

ABSTRACT

Stock Price Prediction is a crucial and challenging task for investors as well as traders because it involves volatile market conditions, government policy, financials results and news related to company. Forecasting of stock prices helps investors to make better investment decision and reduces their financial risk. In this research paper, we are using machine learning algorithms and tools to compare past two year data of stock by applying time series analyses techniques namely deep learning (LSTM) and ARIMA Modelling for predicting stock prices. Results of our study shows that the (1) Deep learning models predicted more accurately then ARIMA Modeling because LSTM can capture non-linear relationship between the stock prices and learn from the past by observing the historical patterns (2) Number of layers in LSTM model has significant impact on predicted value.

Key Words: ARIMA Modeling, Long & Short-Term Memory, News Sentiment analysis

Introduction

India is witnessing its growth story and banking sector is leading Indian economy to achieve its target of 5 trillion-dollar economy by 2030. Banking sector is considered as the Backbone of an economy because directly or indirectly every business or industry depends upon banks for its smooth operations.

The public banks which were earlier regarded as most stagnant and laggard because of high NPAs, low loan growth and more operating expenses than income, this deteriorates their balance sheet. But now those banks are regarded as the most trust worthy and prominent banks because it can be seen that the profits of PSUs banks are surpassing the overall banking sector. Additionally, in 2023 SBI (State Bank of India) the largest government bank became the No.1 company in India with highest ever profit surpassing the Reliance industries with Rs. 19222 crores (Business Standard).

All thanks to digitalization and strong government rebound policies that made their balance sheet strong. The State Bank of India plays a major role in the economy in the following ways:

1. It acts on the behalf of both central as well as state governments.
2. They receive and deposit money on the behalf of the government and all government transactions take place through the State bank of India
3. All government employees' salary is given through the State Bank of India and their public provident fund is also deposited in the State Bank of India.
4. All direct and indirect tax is collected in the State Bank of India.
5. All government loans and advances are given by the State Bank of India

Stock Price Prediction is a crucial and challenging task for investors as well as traders because it involves volatile market conditions, government policy, financials results and news related to company. Forecasting of stock prices helps investors to make better investment decision and reduces their financial risk. As we have so many machine learning algorithms for forecasting. But, for forecasting it is necessary to understand how to accurately predict the prices and decide which model is best on the basis of accuracy.

Therefore, the main objective for this research is to find out the answer of the following questions:

1. Can we accurately predict the future stock price?
2. Which Model is better LSTM or ARIMA Modelling?
3. Is the number of layers in the model affects its predicted value?

From various available machine learning algorithms, we have chosen two model ARIMA Model (Auto Regressive

*Student of BBA, IMS Unison University
**Assistant Professor, IMS Unison University, Dehradun

Integrated Moving Average Model) and LSTM (Long & Short-term Model)

ARIMA Modelling (AR, I, MA)

It is Auto Regressive Integrated Moving Average Model that means auto function automatically depicts the order for auto regressive, differencing and moving average by looking at the lowest value of AIC (Akaike Information Criteria) that measures the predictor to the statistically goodness fit of the model (Khan & Gupta, 2020). It captures the following key elements -

1. AR - it is the auto regression model that uses the dependencies between an observation and a number of lagged observations.
2. I - Integrated is used to make the time series stationary by measuring the differences of observations at different time.
3. MA - Moving average takes into account the dependency between observations and the residual error terms when a moving average model is used to lagged observations.

Before stepping into the building algorithm for ARIMA modelling, we use auto and partial correlation factor analysis on time series data to check whether data is having auto correlation or not (i.e., its next value is depended upon its previous value or not) (Chujai et al., 2013). For this purpose, we look at two blue dotted lines that signify significant levels if spikes are within these ranges, then it depicts data is not correlated and if spikes cross these significant levels, then it depicts data is auto-correlated and if data is correlated then we have to remove it.

And finally, we use ADF test (Augmented Dickey-Fuller test to know whether the time series data of the state bank of India share price is stationary or not which means the time series data does not have any trend and have constant variance over time then it is known as stationary data series and if time series data follows a trend and it time-dependent that does not follow constant variance over time then it is known as non-stationary data. For this purpose, we look at p - the value and if p - value is less than 0.05 then we reject the null hypothesis and accept the alternate hypothesis and come to the conclusion that time series data is stationary whereas if p - value is greater than 0.05 then we accept the null hypothesis and come to the conclusion that time series data is non-stationary. For forecasting, if time series data is non-stationary then it needs to convert to stationary data series by using differencing (Alghamdi et al., 2019).

B. LSTM

It is long short-term memory. This model learns from its previous values and treats its previous values as input for forecasting the next value. It is specially designed for handling sequential data or time series data. It is a special kind of Recurrent Neural Network with additional feature to memorize the data (Shah et al., 2018). In Recurrent Neural Network, a memory cell is introduced that act as a container for holding information for a period of time in order to make Long short-term memory model. These memory cell act as a passing line for transporting and storing past data for future use (Shumway et al., 2017). Owing the large amount of sequential data, it is very important to retain only selective or important data and to discard the irrelevant information from the network of cell (Moghar & Hamiche, 2020). So, for this purpose, three types of gates are involved in LSTM to control such data or information:

1. Forgot Gate - in this input is passed through an activation function which gives a binary value either 0 or 1 in order to understand which information is not useful in the cell and have to remove out of it and which information is useful to kept in the cell, respectively.
2. Input Gate - in this, new data or useful information is added or stored to the cell. Using the sigmoid function all the information is regulated and filtered to choose which values to be modified and finally using a tanh function a new vector is created for all possible values that gives an output from -1 to +1
3. Output Gate - in this, all the relevant data or information is presented from the cell state on the basis of the sigmoid and tanh function or we can say Input Gate (Cipiloglu Yildiz et al., 2022).

Research Methodology -

1. The following study aims at analyzing the past 1-year share price data of State Bank of India in order to forecast the price for the next 3 months. So, the following study involves use of descriptive analytics, predictive analytics and Quantitative analytics
2. It is Empirical Research type study that is based on or concerned with, or verified by observation or experience rather than theory or pure logic
3. All the research study is taken place by using R programming for ARIMA Modelling and Python programming for Deep Learning Model and for analyzing data we are using time series analysis.
4. We are using secondary source of data that is already available and published on various websites say yahoofinance.com or nseinida.com.
5. The duration of the data is from 01-01- 2021 to 31-08-2023.

Dataset:

The snapshot of the dataset that contains the share Price of State Bank of India along with Date, Opening Price of day, High Price of day, Low price of day, Closing price of day, Adjusted Closing Price and Volume of the day is shown in figure 1.

	Date	Open	High	Low	Close	Adj.Close	Volume
1	01-01-2021	274.90	280.00	274.40	279.40	267.3938	24531791
2	04-01-2021	281.85	283.90	277.75	281.05	268.9729	31450608
3	05-01-2021	278.05	282.45	277.00	281.75	269.6428	27393072
4	06-01-2021	283.00	289.15	281.40	285.05	272.8010	40765708
5	07-01-2021	289.00	291.80	287.00	287.70	275.3372	30546600
6	08-01-2021	290.10	291.40	285.20	286.00	273.7102	32543357
7	11-01-2021	288.00	288.20	279.60	282.50	270.3606	29703173
8	12-01-2021	280.00	293.85	277.90	292.50	279.9309	44307145
9	13-01-2021	296.00	308.00	294.50	306.80	293.6164	78042133
10	14-01-2021	306.70	309.25	303.80	307.25	294.0471	40123498
11	15-01-2021	306.80	310.90	301.30	303.85	290.7932	35307601
12	18-01-2021	303.50	308.65	292.20	294.45	281.7971	54969750
13	19-01-2021	297.65	302.50	296.40	298.60	285.7688	35803901
14	20-01-2021	298.80	304.70	296.85	302.55	289.5490	33759304
15	21-01-2021	304.00	305.15	291.50	294.85	282.1799	29995203
16	22-01-2021	295.50	298.00	282.40	283.70	271.5091	44440810
17	25-01-2021	284.50	288.00	277.05	280.95	268.8772	34034630
18	27-01-2021	280.20	284.85	272.70	275.65	263.8050	39211107
19	28-01-2021	271.90	283.35	269.50	282.35	270.2170	48717297
20	29-01-2021	285.60	290.65	279.10	282.10	269.9778	42300984
21	01-02-2021	285.10	313.90	282.75	310.70	297.3488	103414068
22	02-02-2021	315.00	337.40	315.00	333.10	318.7863	121434166
23	03-02-2021	333.75	339.90	325.80	335.95	321.5138	66420168
24	04-02-2021	334.40	358.00	331.10	355.10	339.8409	145203439
25	05-02-2021	387.00	408.35	385.15	393.10	376.2080	214965688
26	08-02-2021	398.00	407.80	394.75	397.05	379.9883	79389205
27	09-02-2021	396.75	403.50	389.00	394.80	377.8349	63869990
28	10-02-2021	394.05	397.90	388.00	392.25	375.3945	54163507
29	11-02-2021	391.80	396.95	388.25	390.15	373.3847	36392708
30	12-02-2021	391.00	399.35	385.55	393.15	376.2558	53589755

Data Analysis and Interpretation:

ARIMA Modelling Algorithm

Auto ARIMA is based on the linear regression model that is best suited for simple forecasting that does not require to manually type the order of auto regressive, differencing and moving average. As on the basis of the lowest value AIC (Akaike information Criteria) it automatically depicts the order of AR, I and MA.

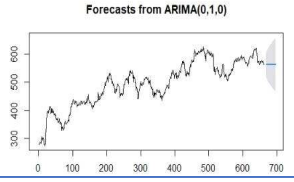
Using this, we've uploaded data into R studio using read.csv function and store it in a variable called SBI. This SBI variable helps to read the data stored in the excel sheet that contains date, opening price, closing price, High price, low price and Volume for a particular date.

Then, we have converted the excel sheet data into time series data using tseries library in which we have taken closing price as basis for further analysis and data is started from 01-January-2023 to 31-August- 2023.

```
Fitting models using approximations to speed things up...

ARIMA(2,1,2) with drift         : 4651.657
ARIMA(0,1,0) with drift         : 4649.82
ARIMA(1,1,0) with drift         : 4652.109
ARIMA(0,1,1) with drift         : 4651.07
ARIMA(0,1,0)                    : 4649.61
ARIMA(1,1,1) with drift         : 4653.414

Now re-fitting the best model(s) without approximations...

ARIMA(0,1,0)                    : 4653.821

Best model: ARIMA(0,1,0)
```

In this figure, we have used auto arima function to build model for forecasting and it has also given the order value of AR, I and MA and suggest which model is best on the basis of lowest AIC value that is 4653.8

```
Forecasted_values <- forecast(Sbi_auto_arima, level=c(95),h=30) Plot(Forecasted_values)
```



Using this algorithm, we have forecasted values for the next 30 days at 95% confidence interval using sbi.auto.arima model and finally plot it to show the forecasted values along with all previous values.

LSTM Algorithm

LSTM is based on the recurrent neural network that has the capability to learn and store past information to forecast the future values. In this analysis we are using 90% of the data for training and 10% of the data for testing. Before developing algorithm for LSTM we need to install following libraries such as pandas, numpy, matplotlib, sklearn, Sequential, Dense, LSTM, Dropout. Here is the algorithm for developing LSTM Model -

```
import pandas as pd import numpy as np
import matplotlib.pyplot as plt from sklearn import metrics
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense,LSTM,Dropout
df = pd.read_csv('/content/drive/MyDrive/SBI.csv',parse_dates=True,index_col='Date') sbi_close = df['Close']
sbi_close.reset_index(drop=True,inplace=True)
sbi_close = pd.DataFrame(sbi_close)
```

Using this algorithm, we have imported data into google colab using read_csv function from the google drive. And load the sbi closing price into sbi_close variable

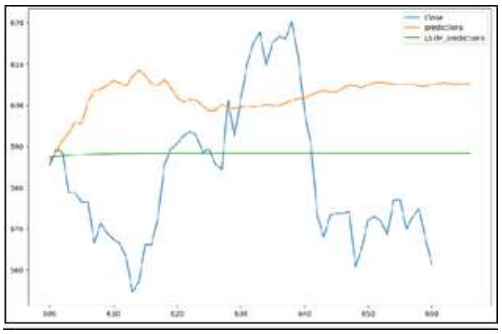
```
train = sbi_close.iloc[:600] test = sbi_close.iloc[600:]
scaled_train = scaler.transform(train) scaled_test = scaler.transform(test)
```

Using this algorithm, we have divided the closing price into training and test data that contains 90% of data and 10% of data respectively, remember it is important to scale these training and testing data using scaler function.

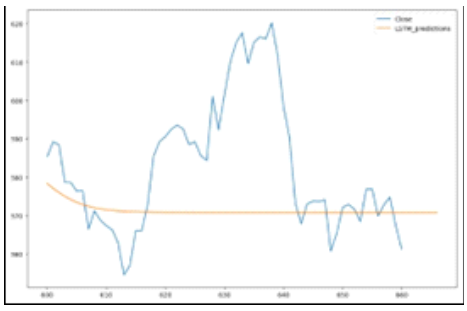
In this algorithm, we have build LSTM model with 55 layers and drop out ratio of 0.2 that means it will not consider 1 value out 5 values for forecasting and suggest forecasting after consider loss on the basis of mean squared error and optimizing using Adam optimizer.

```
length = 66
generator = TimeseriesGenerator (scaled_train, scaled_train, length= length,batch_size=1) validation_generator = TimeseriesGenerator(scaled_test,scaled_test,length=length,batch_size=1) model = Sequential()
model.add(LSTM(55,input_shape=(length,n_features)))
model.add(Dropout(rate=0.2))
model.add(Dense(1))
model.compile(optimizer='adam',loss='mse')
test_prediction = []
first_eval_batch = scaled_train[-length:]
current_batch = first_eval_batch.reshape ((1,length,n_features)) for i in range(len(test)) :
current_pred = model.predict(current_batch)[0]
test_prediction.append(current_pred)
current_batch = np.append(current_batch [:,1:,:], [[current_pred]],axis=1)

true_prediction2 = scaler.inverse_transform(test_prediction) test['LSTM_predictions'] = true_prediction2 test.plot(figsize=(12,8))
```

This plot shows the closing price of state bank of India along with forecasted value using LSTM model with 55 layers and it is showing the forecasted value ranges between Rs. 585 to 588.



This plot shows the closing price of state bank of India along with forecasted value using LSTM model with 80 layers and it is showing the forecasted value ranges between Rs. 573 to 578.



Limitations of the research study:

Well, we cannot deny some limitations while using technology or machine learning algorithms for forecasting the future possible price because of the following reasons-

1. There is always presence of some error range between 5% - 8%.
2. It does not apply to stocks which are more volatile during trading hours.
3. Stock Market events are majorly affected by macro-economic events.
4. The coding is sophisticated and difficult to understand for the non tech retail investors.
5. It is very time consuming to decide the exact number of layer for LSTM to get desired results.

Results & Findings:

Deep learning LSTM model is more accurate as compared to ARIMA modelling because ARIMA modelling has more statistical error than ARIMA modelling. The number of layers in LSTM model affects the forecasted values. Also one cannot accurately predict the price of the share price due to the presence of macro events and demand & supply factors but we can accurately predict the trend of the price.

Conclusion:

Time Series analysis on State bank of India's share price helps to forecast the future values on the basis of their past behavior by analyzing data in a more statical manner. Both the model ARIMA and LSTM had predicted values but with presence of some degree of error so we should not completely rely on these predicted values and we should only take it an advice or to identify the trend of the share price for managing our risk and taking investing decision.

References:

- Khan, F. M., & Gupta, R. (2020). ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. Journal of Safety Science and Resilience, 1(1), 12-18.
- Chujai, P., Kerdprasop, N., & Kerdprasop, K. (2013, March). Time series analysis of household electric consumption with ARIMA and ARMA models. In Proceedings of the international multiconference of engineers and computer scientists (Vol. 1, pp. 295-300). Hong Kong: IAENG.
- Alghamdi, T., Elgazzar, K., Bayoumi, M., Sharaf, T., & Shah, S. (2019, June). Forecasting traffic congestion using ARIMA modeling. In 2019 15th international wireless communications & mobile computing conference (IWCMC) (pp. 1227-1232). IEEE.
- Shumway, R. H., Stoffer, D. S., Shumway, R. H., & Stoffer, D. S. (2017). ARIMA models. Time series analysis and its applications: with R examples, 75-163.

- Shah, D., Campbell, W., & Zulkernine, F. H. (2018, December). A comparative study of LSTM and DNN for stock market forecasting. In 2018 IEEE international conference on big data (big data) (pp. 4148-4155). IEEE.
- Moghar, A., & Hamiche, M. (2020). Stock market prediction using LSTM recurrent neural network. Procedia Computer Science, 170, 1168-1173.
- Cipiloglu Yildiz, Z., & Yildiz, S. B. (2022). A portfolio construction framework using LSTM?based stock markets forecasting. International Journal of Finance & Economics, 27(2), 2356-2366.
- https://www.business-standard. com/companies /news/sbi-beats-reliance-industries-to-become-india-s-most-profitable-company-123080700804_1.html