Comparative Study of Pre and Post COVID impact on Stock Markets

Ashutosh Nagaria[1], Chirag Jain[2], Himadri Vipat[3], Darshan Mehta[4], Danish Sheikh[5]
Medi-Caps University, Indore
(anashutosh9[1], avied300[2], darshanmehta006[3], danish22shiekh[4], himadrivipat1109[5])@gmail.com

****

Abstract:

The Stock Market is over a century old concept used around the globe to raise money. It is a very volatile industry and predicting it can be particularly hard for the investors. The stock price predictor helps the investors to make educated guesses and hence manage risk efficiently by devising a diverse portfolio. The stock price prediction can be done effectively and accurately by using machine learning and deep learning algorithms.

Recently the world was hit by Coronavirus and the disease it causes it highly contagious. Due to which the entire operations of all countries have been shut down which had a hard impact on the economy. The economies of even the developed countries seems to be going down affecting the markets and eventually the Stock Markets.

In this paper, we have attempted to perform a comparative study between the pre and post covid impacts on stock prices of various industries viz. Automobile industry, IT industry, Pharmaceutical industry and the Indices of Indian stock market. We have used the ARIMA ie Auto Regressive Integrated Moving Average as our flagship model for all the predictions and discussed its accuracy and efficiency over other algorithms.

Keywords — **Machine Learning, Time-Series Analysis, ARIMA Model, LSTM, Stock price prediction, Indian Stock Market**

****

I. INTRODUCTION

The primary task for stock prediction is to predict the adjusted close price for the upcoming session of the market in the delivery trading order of the stocks. We have taken into consideration three industries Automobile, Pharmaceutical, IT and the Indices of Indian Stock Market NIFTY and SENSEX. The best algorithm for the task is the "Auto Regressive Integrated Moving Average" time series model also known as ARIMA. It has a particularly high accuracy in predicting the prices.

The datasets are taken from Yahoo! Finance website and they are highly accurate and fidel. The Stock Market does not function everyday and has regular breaks in time. This makes it a bit difficult for simple time series models to perform operations as data gets broken or discrete.

There are some factors which affect the stock market:

- Natural Disasters : Natural calamities like floods, droughts and situational slowdowns like recession of 2009, Covid 2020, which is particularly our interest causes a huge blow to the economy and hence the stock markets. Least can be done when we are having this kind of situation to revive the market.
- Entry or withdrawal of a large cap company: Failures or entrances of large cap companies definitely makes a huge difference. For example, the crash of Lehman Brothers and eventually filing bankruptcy led to a huge blow in NASDAQ. Similarly, Reliance India launching JIO had a significant impact on Indian stock market and economy.
- Controversial Statements and Tweets: Controversial tweets from powerful people like President, Prime Minister etc also affect the markets. For example, US President, Mr.Trump's statement on war with China may cause turbulence in markets.

II. DATA SET

For the stock price prediction we will use the historical data available on Yahoo! Finance.It provides all the relevant news regarding the stock market such as press releases , financial reports as well as the historical data of the companies..

The dataset on the yahoo finance contains seven columns which are named as:date, high , low ,close ,adj close , volume.

	Date	Open	High	Low	Close	Adj Close	Volume
0	2015-01-01	1428.400024	1435.800049	1424.099976	1429.199951	1390.314331	119605.0
1	2015-01-02	1435.000000	1442.500000	1428.050049	1432.300049	1393.329956	247162.0
2	2015-01-05	1434.150024	1441.650024	1425.250000	1430.300049	1391.384399	211101.0
3	2015-01-06	1425.000000	1428.949951	1386.300049	1395.849976	1357.871582	405597.0
4	2015-01-07	1395.849976	1401.599976	1365.349976	1377.599976	1340.118164	483850.0

Fig 1.The first five records of the dataset

For the stock price prediction we will explore three different sectors of the market which are as follows:-

- IT sector:- TCS and Infosys
- Pharmaceutical sector:- Cipla and Lupin
- Automobile sector:- TATA Motors and Force Motors

The dataset is collected with a time span of five years from 2015 to april 2020 for the comparative study of the sectors named above.The target variable for our prediction is the adj close value and the feature variables are open,high,low and volume.

2.1 Exploratory Data Analysis

The dataset of each company comprises 1035 observations and seven characteristics.Only float and integer values are found in it and no columns have null or missing values.

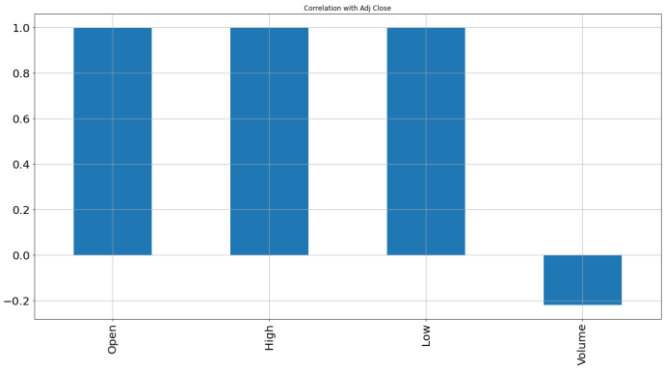


Fig2. Correlation of adj close variable with other variables

From the above table we find that the correlation of adj close variables with open, high, low variables is around 1 which is significant for the model building and they are highly interdependent.

	Open	High	Low	Close	Adj Close	Volume
count	1305.000000	1305.000000	1305.000000	1305.000000	1305.000000	1.305000e+03
mean	571.818008	578.614983	563.681378	570.736973	562.525815	2.114420e+06
std	69.833611	70.164178	68.864973	69.259748	66.132204	2.138062e+06
min	370.000000	390.750000	355.299988	374.700012	374.700012	1.233630e+05
25%	528.599976	534.000000	521.000000	526.500000	518.703247	1.053561e+06
50%	570.200012	577.450012	562.900024	570.049988	561.574219	1.553260e+06
75%	619.000000	628.500000	611.099976	617.900024	608.648865	2.339657e+06
max	744.950012	752.849976	730.250000	739.599976	724.793030	2.720771e+07

Fig 3. Descriptive Analysis of the dataset

As you can see from the above figure ,there is a big difference between min values and max values in the adj close variable, it suggests that there will be outliers in it.

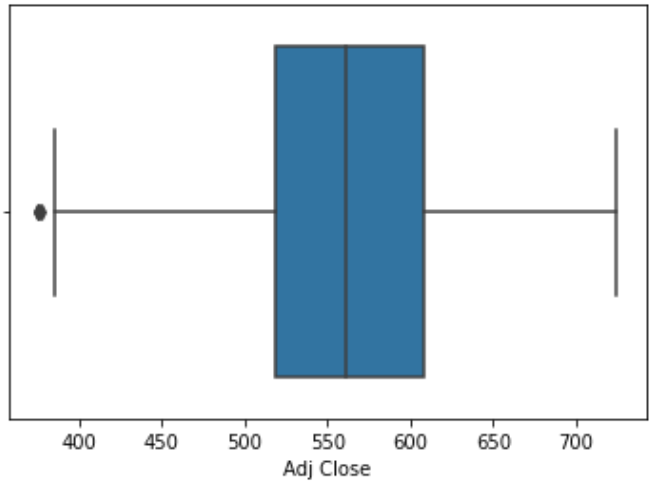


Fig4 Boxplot of the adj close variable

The following outliers will be removed from the datasets and have been normalized for further use.

III. METHODOLOGY

The main methods employed in this paper are the Long Short Term Memory approach and time series Auto Regressive Integrated Moving Average, ARIMA model.

Autoregressive integrated moving average (ARIMA): Proposed by George Box and Gwilym Jenkins in 1970, ARIMA models are among the most popular linear models. In ARIMA models, the future value of a variable is obtained through a linear function of some past observations of the variable and some random errors. The process that generates the time series has the form of:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

where yt is the variable that will be explained at time t; c is the constant or intercept;

φi(i = 1, 2, ...p) and θj(j = 1, 2, ...q) are the model parameters; p and q are integers and are often referred to as AR and MA orders of the model, respectively;

and et is the error term.

The assumption regarding the random errors εt is that they are independently and identically distributed with a mean zero and constant variance of σ2. This model involves a three-step iterative process of identification, estimation, and diagnostic checking. The identification step involves specifying a tentative model by deciding the order of the AR (p) and MA (q) terms. Once a tentative model is specified, the parameters of the model must be estimated, in such a way that the overall measure of errors is minimized, which is generally done with a nonlinear optimization procedure. After the estimation of parameters, diagnostic checking for the adequacy of the model must be done, which involves testing whether the model assumptions about the errors εt are satisfied. If the model is adequate, one can proceed to forecast; if not, a new tentative model must be identified following the parameter estimation and model verification.This process with three steps must be repeated until a satisfactory model is selected to forecast the data.

LSTM:

Long Short Term Memory: Introduced by Hochreiter & Schmidhuber (1997), and were refined and popularized by many people in following work.The key to LSTMs is the cell state, the horizontal line running through the top of the diagram.The cell state is kind of like a conveyor belt. It runs straight down the entire chain, with only some minor linear interactions. It's very easy for information to just flow along it unchanged.The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates.

Gates are a way to optionally let information through. They are composed out of a sigmoid neural net layer and a pointwise multiplication operation.The sigmoid layer outputs numbers between zero and one, describing how much of each component should be let through. A value of zero means "let nothing through," while a value of one means "let everything through!". An LSTM has three of these gates, to protect and control the cell state.

The equations for the gates in LSTM are:

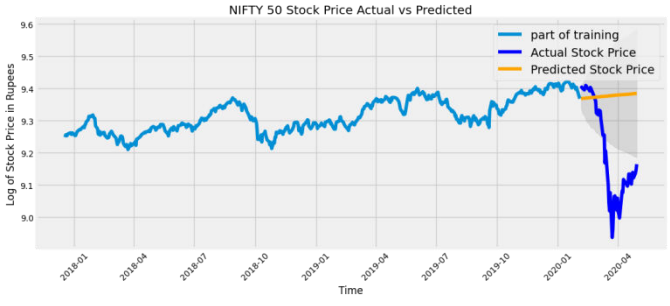
$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i)$$
$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f)$$
$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o)$$

Equation of Gates

$i_t \rightarrow represents\ input\ gate.$
$f_t \rightarrow represents\ forget\ gate.$
$o_t \rightarrow represents\ output\ gate.$
$\sigma \rightarrow represents\ sigmoid\ function.$
$w_x \rightarrow weight\ for\ the\ respective\ gate(x)\ neurons.$
$h_{t-1} \rightarrow output\ of\ the\ previous\ lstm\ block(at\ timestamp\ t-1).$
$x_t \rightarrow input\ at\ current\ timestamp.$
$b_x \rightarrow biases\ for\ the\ respective\ gates(x).$

- First equation is for the Input Gate which tells us what new information we're going to store in the cell state.
- Second is for the forget gate which tells the information to throw away from the cell state.
- Third one is for the output gate which is used to provide the activation to the final output of the lstm block at timestamp 't'.

Finally,we filter the cell state and then it is passed through the activation function which predicts what portion should appear as the output of the current lstm unit at timestamp t.



From the following image, it can be clearly seen that there has been a sharp drop in the prices of NIFTY50 companies. The shaded region in the graph (at the end) illustrated the range the industry is likely to fall into. But due to COVID, the stock has been out of the range. Also, the minima of the curve can be noticed in the end of March when the lockdowns were initiated in India due to the huge widespread of the pandemic. But it can be seen recovering from the phase and is likely to be running smoothly in the coming future.

Some important Ratios :

Mean Absolute Percentage Error (MAPE) is the sum of the individual absolute errors divided by the demand (average of the percentage errors). Mean absolute percentage error is commonly used as a loss function for regression problems and in model evaluation, because of its very intuitive interpretation in terms of relative error.

Mean Absolute Error (MAE) is simply the mean of the absolute errors (it is not scaled to the average demand). Mean absolute percentage error is commonly used as a loss function for regression problems and in model evaluation, because of its very intuitive interpretation in terms of relative error.

Root Mean Squared Error (RMSE) is defined as the square root of the average squared error. (Just like MAE, RMSE is not scaled to the demand). RMSE is a measure of how spread out these residuals are.

Mean Squared Error is the mean squared error or mean squared deviation of an estimator measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. MSE is a risk function, corresponding to the expected value of the squared error loss.

COMPARATIVE STUDY

After the implementation of the algorithms, we found out the following table and prepared the corresponding table:

Current Situation of Stocks amid COVID Table

Stock/Performance	MSE	MAE	RMSE	MAPE
TCS	0.0264419	0.132137	0.162609	0.0176574
Infosys	0.042139875815 85948	0.1628303787590 3122	0.2052799937058 1507	0.0253228975612 30462
Sunpharma	0.016627172025 254263	0.1023914012746 5213	0.1289463920598 5664	0.017268735132 2991
Lupin	0.018355598517 76083	0.1108377362106 7547	0.1354828347716 4488	0.0170927605380 06565
Tata Motors	0.385820655044 0799	0.5183733751730 956	0.6211446329511 991	0.1189565434917 8475
Eicher Motors	0.088109372249 36603	0.2522935829631 763	0.2968322291284 5236	0.0263184774930 64606
NIFTY	0.051591698577 05023	0.1826346132517 5695	0.2271380606086 3123	0.0200742526742 2393
SENSEX	0.051249854711 05446	0.1818939064843 8733	0.2263843075636 0844	0.0176126425405 36315

The table above represents the state of the parameters RMSE, MSE, MAE, MAPE for the various industries. .

As per the table, we can see that the pharmaceutical industry is having particularly low values of these errors which makes complete sense in these times of COVID. Both the companies are dedicated to making protection equipment like masks and sanitizers. The demand of medicine in an economy is minimally dependent on the state of markets. Whereas in the automobile sector, large value of errors shows high variance in the prices. The COVID has completely ceased the production, supply and obviously the sales of any automobile and hence the stock price is constantly going down.

Meanwhile the IT industry has shifted its operation inside doors ie. work from home, the amount of work hasn't been largely reduced but a significant drop has been noticed in the number of projects as no projects are being given to any IT company in India or the world. The hig value of MAE and RMSE in Infosys clearly shows the disturbance and the turbulence industry is going through.

The indices of stock market the NIFTY and the SENSEX have also followed a down path as the Indian economy is not performing well and as per reports, India is about to go into recession phase due to COVID 19.

IV. CONCLUSIONS

The following conclusions can be drawn from the corresponding study:

- The pharmaceutical sector has been performing well and is likely to outperform other industries. Hence, this is the best industry to invest in at the moment to maximise the returns. They also have very low MSE error.
- The automobile industry has faced a huge blow due to the pandemic and is likely to worsen in the near future. Any investment in the industry must be made very cautiously.
- The IT industry has also faced a slowdown but is likely to recover in a few quarters and is likely to be running smoothly in near future.
- The indices, which represent the health of the industry are also low due to the pandemic. The growth path will follow once the markets are open and businesses start flooding the markets with cash.
- The ARIMA model has higer accuracy over LSTM model by 5.88%.

ACKNOWLEDGMENT

I would like to express my very great appreciation to Dr Ravi Changle and Mr Rajat K Jain for their valuable and constructive suggestions during the planning and development of this research work. Their willingness to give his time so generously has been very much appreciated.

I would also like to thank Tata Consultancy Services for the opportunity to work on this paper.

REFERENCES

[1] Stock Price Prediction Using the ARIMA Model (2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation) 2014

[2] K. Raza, "Prediction of Stock Market performance by using machine learning techniques," 2017 International Conference on Innovations in

Electrical Engineering and Computational Technologies (ICIEECT), Karachi, 2017, pp. 1-1.

[3] M. Usmani, S. H. Adil, K. Raza and S. S. A. Ali, "Stock market prediction using machine learning techniques," 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, 2016, pp. 322-327.

[4] Adebiyi AA, Oluinka A (2014) Comparison of ARIMA and artificial neural network models for stock market prediction. Journal of Applied Mathematics. https://doi.org/10.1155/2014/614342

[5] Comparative Study of ARIMA Methods for Forecasting Time Series of the Mexican Stock Exchange Javier A. Rangel-González, Juan Frausto-Solis, J. Javier González-Barbosa, Rodolfo A. Pazos-Rangel and Héctor J. Fraire-Huacuja © Springer International Publishing AG 2018 Castillo et al. (eds.), Fuzzy Logic Augmentation of Neural and Optimization Algorithms: Theoretical Aspects and Real Applications, Studies in Computational Intelligence 749, https://doi.org/10.1007/978-3-319-71008-2_34