Stock Forecasting Using Prophet vs. LSTM Model Applying Time-Series Prediction

Arnav Sabharwal
Department of Artificial Intelligence and Data Science
Maharaja Agrasen Institute of Technology
Sector 22, PSP Area, Delhi, 110086, India
{arnavsab2@gmail.com)

***Abstract*— Forecasting and time-series modelling are critical steps in the data analysis process. Time series are frequently used in analytics and data science. Forecasting stock prices is a fascinating and essential topic in financial and academic studies. Due to the lack of key standards for estimating or projecting a stock price in the stock market, the stock market is an unstructured place for forecasting. As a consequence, forecasting stock prices is a time-series problem that is tricky to solve. A variety of methods and applications of machine learning are useful for carrying out stock price forecasting, such as technical analysis, fundamental analysis, time series analysis, and statistical analysis. This paper will guide readers on how to use prophet and LSTM models to execute stock price predictions and research. This approach and task are extremely difficult and unclear. Although the stock price is not anticipated due to its undetermined field, this research seeks to predict stocks using the principles of forecasting and data analysis.**

***Index Terms*— Predicting; Modelling; Analysis; Machine Learning; Time-series; Stock price; data analysis, Long Short-Term Memory (LSTM), forecasting.**

I. INTRODUCTION

Arithmetic analysis of historical data is used to predict various aspects of people's lives. Forecasting illness, changes in stock market activity, and weather can all be done if a pattern in previous data is due to time. It can be daily, weekly, monthly, or annually, for example. This type of forecast is frequently referred to as Time Series Forecasting. Observations were made in a time sequence, which is known as a time series[13]. The increasing accessibility of historical data, combined with the need for the production of forecasting, has caught the attention of Time Series Forecasting (TSF), which provides a sequence for predicting future values, particularly in light of traditional forecasting's limitations, such as complexity and time consumption[12]. A lot of variables influence the stock price[2]. Forecasting involves numerous aspects, including physical factors in opposition to psychological, rational, and irrational behavioural patterns. These factors make stock prices unforeseeable and hard to anticipate with great accuracy. According to one popular theory, stock prices are entirely arbitrary and their value cannot be predicted. This argument begs the question of why huge corporations use quantitative analysts to create prediction models[1]. Is machine learning effective in forecasting stock prices? To predict prices, this paper employs machine learning models, prophet, and Long Short-Term Memory (LSTM). The work is carried out using an earlier data set for the stock price of a publicly traded business (HDFC). One machine learning algorithm will be developed employing imaginative and popular methodologies to anticipate the company's future stock price; the name is prophet. The corporation may become exposed to market movements beyond its control, such as market sentiment, economic conditions, or sector developments. The hypothesis for the present study is that LSTMs will outshine other techniques and provide greater in-depth insight into the validity of the technical analysis.

II. LITERATURE REVIEW

This section offers foundations and basic working definitions. Fundamental and technical analysis, which are non-machine learning methodologies for stock valuation as well as machine learning approaches, provide an overview of the key aims and concepts. Forecasting time series of financial data has always been a vital subject and a compelling study area with multiple uses in business, economics, finance, and computer science.

A. Fundamental Analysis

Traditional methods of stock market assessment and forecasting stock prices include fundamental analysis, which looks at the stock's performance and the company's overall credibility, and statistical analysis, which primarily deals with multiplying numbers and identifying patterns in stock price variation[1]. In general, the fundamental analysis aims to examine some of the company's macro aspects. According to its notions, market value tends to gravitate towards the real deal or intrinsic value.

B. Technical Analysis

Technical analysis is nearly the inverse of fundamental analysis. Technical analysis is a means of forecasting and assessing prior market data, prices, and volume. In most cases, traders that employ this approach base their trading strategy on technical indicators that are calculated based on price, volume, and time [4]. Past stock price data is the only input to technical analysis. The technical analyst claims that the stock's previous pattern indicates prospective designs and prices [2].

C. Analysis Based on Models

There are numerous machine learning algorithms and methodologies, so picking the best way has proven difficult[2]. Technical and fundamental analyses have no bearing on time series models and machine learning models. They rely on math concepts to create meaningful models from training data. The model that emerges can then be used to forecast new data[4]. The use of a machine-learning model to predict the price of a specific stock was proposed in this research. The objective of this project is to reliably predict the future closing value of a specific stock over a given time period. In this paper, a prophet model and Long Short-Term Memory network, also known as "LSTMs," were used to anticipate HDFC's price using a data set of earlier prices.

III. The Research Method

In this work, quantitative methods are used to examine and visualise data using models and Python code.

A. Analytic models:

Initially, the prophet model was utilized, and then Long Short-Term Memory networks were used to estimate Google's closing price using a data set of previous values. Based on the prophet model's test data set, the model forecasts twenty years of data points.

Root Mean Squared Error (RMSE) was employed as a performance measure in this research to calculate the difference between predicted and actual stock values at the close price between the performance of the model (prophet) and model (LSTM).

B. Exploring the Stock Prices Dataset

The data set utilised in the present study is of HDFC and spans from January 3, 2000, through April 30, 2020. A time series is a collection of data points that are indexed in time order. After training, the goal is to figure out the price for any given day. All data was retrieved through Kaggle Data set for simplicity of replication and re-usability. The dataset contains several variables, including date, open, high, low, close, and volume.

- The columns Open and Close show the opening and closing prices of the stock on a given day.
- The peak, lowest, and final price of the day's share are represented by the words "high," "low," and "close."
- Volume- the amount of an asset or security that is liable to vary during a specified period of time, usually a day[3].

The prediction must be done using the Closing Price in the data set. The only feature that should be given to the financial time series being predicted using prophet and LSTM models is the "Close" variable [10]. Begin by importing all the required libraries (NumPy), (pandas), (seaborn),(sklearn),(tensorflow),(prophet) and (matplotlib). Load the data set and specify the problem's target variable. Then, import the CSV file into Python and read it using pandas' read_csv(). The data set is organised as follows (Table.1):

	Date	Symbol	Series	Prev Close	Open	High	Low	Last	Close	VWAP	Volume	Turnover	Trades	Deliverable Volume	%Deliverb
0	2000-01-03	HDFC	EQ	271.75	293.50	293.50	293.50	293.50	293.50	293.50	22744	6.675364e+11	NaN	NaN	N
1	2000-01-04	HDFC	EQ	293.50	317.00	317.00	297.00	304.00	304.05	303.62	255251	7.749972e+12	NaN	NaN	N
2	2000-01-05	HDFC	EQ	304.05	290.00	303.90	285.00	295.00	292.80	294.53	269087	7.925368e+12	NaN	NaN	N
3	2000-01-06	HDFC	EQ	292.80	301.00	314.00	295.00	296.00	296.45	300.14	305916	9.181669e+12	NaN	NaN	N
4	2000-01-07	HDFC	EQ	296.45	290.00	296.35	281.00	287.10	286.55	288.80	197039	5.690480e+12	NaN	NaN	N
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
5301	2021-04-26	HDFC	EQ	2497.35	2500.00	2534.10	2483.20	2502.00	2509.80	2508.07	3916088	9.821805e+14	121028.0	2440395.0	0.62

TABLE I

Head of the Data Set

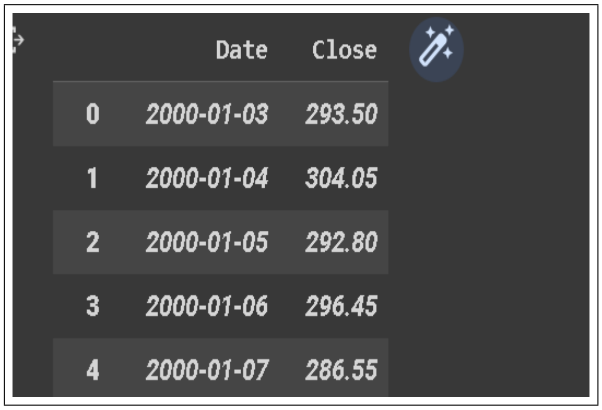
Mean, the standard deviation, maximum, and minimum of the data, as shown in (Table.2):

	Prev Close	Open	High	Low	Last	Close	VWAP	Volume
count	5306.000000	5306.000000	5306.000000	5306.000000	5306.000000	5306.000000	5306.000000	5.306000e+03
mean	1283.666114	1284.393074	1304.269732	1263.297842	1283.885017	1284.071005	1283.664578	1.848187e+06
std	709.395090	709.703665	721.308080	697.450309	709.250204	709.430515	709.109622	2.991387e+06
min	271.750000	284.000000	290.500000	273.250000	282.850000	283.850000	283.600000	2.919000e+03
25%	668.650000	669.712500	677.512500	660.000000	669.000000	668.662500	668.265000	3.034970e+05
50%	1136.275000	1135.400000	1156.725000	1119.000000	1135.000000	1136.675000	1136.720000	1.337788e+06
75%	1811.475000	1813.812500	1835.000000	1783.075000	1812.000000	1811.787500	1811.680000	2.732310e+06
max	3180.150000	3148.000000	3262.000000	3100.550000	3178.000000	3180.150000	3166.580000	1.584141e+08

TABLE II

Mean SD, Max, and Min of the dataset.

Assume from the data set that the date, high and low values are not crucial characteristics of the data. The characteristics High, Low, and Volume are vital, yet it was discovered that Open and Close Prices are directly related. It is important to know the stock's opening and closing prices. If the closing prices are higher than the opening prices, there will be some profit; otherwise, there will be losses. The quantity of stocks is also critical. A growing market should see increased volume; conversely, rising prices and declining volume indicate a lack of interest and warn of a probable reversal. A price decline (or gain) on huge volumes is a stronger indicator that something fundamentally happened in the stock. The following sections investigate these characteristics and employ various methodologies to forecast the stock's daily closing price. As a result, during the processing stage, the high, low, volume and other non-essential features have been eliminated from the data set (Fig.1).



	Date	Close
0	2000-01-03	293.50
1	2000-01-04	304.05
2	2000-01-05	292.80
3	2000-01-06	296.45
4	2000-01-07	286.55

Fig. 1. Data set after removing non-essential attributes

The mean, standard deviation, maximum, and minimum values of the processed data were discovered to be as follows (Fig.2).

	Close
count	5306.000000
mean	1284.071005
std	709.430515
min	283.850000
25%	668.662500
50%	1136.675000
75%	1811.787500
max	3180.150000

Fig. 2. mean SD, Max, and Min of the dataset.

C. Exploratory Visualization to Visualize The Data:

For the initial graphing of the data set in this study, the Matplotlib package was utilised. The historical data is shown in scale (Fig.3):

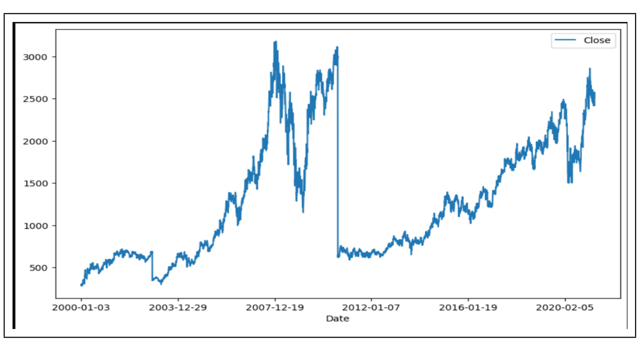


Fig. 3. Visualization of processed historical data.

The closing price of a stock usually dictates the profit or loss computation for the day; so, it will consider the closing price of the target variable. So, plot the desired variable to see how it appears in the data (Fig.4).

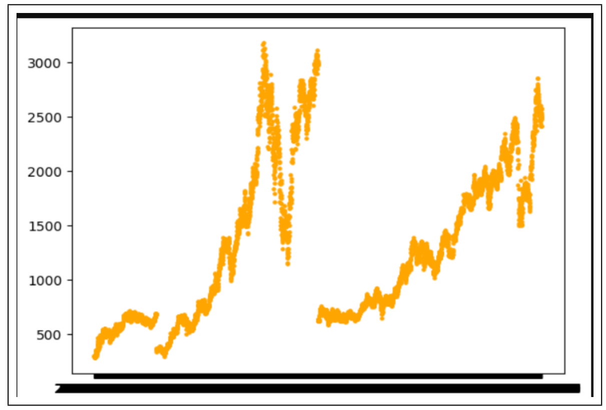


Fig. 4. Representing of HDFC Stocks Closing Values

Correlation is an indication of the association between two features: how much Y changes when X changes. Pearson Correlation is the term assigned to the correlation approach that was implemented. The coefficient is frequently used to quantify correlation because its values range from -1 to 1. It can be interpreted mathematically as if two qualities are positively connected. They are directly proportional if they have a positive correlation, and inversely proportional if they possess a negative correlation[6]. Because not all material is understandable, depict the correlation coefficient (Fig.5).

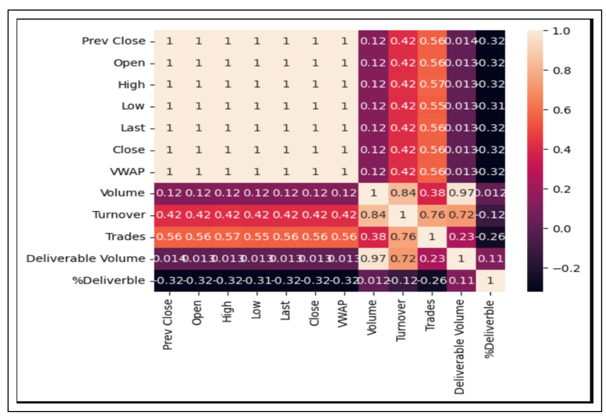


Fig. 5. Correlation Map

The dark zones portray highly correlated features.

IV. Algorithms and Techniques Used

This research intends to investigate time-series data and explore as many choices as possible in order to effectively anticipate the Stock Price.

A. Prophet model:

Many time series techniques can be used for a stock prediction dataset, but the majority of these approaches necessitate extensive data pre-processing before building the model[3]. The prophet is an open-source prediction library for time-series datasets. It is simple and aimed at identifying a good collection of hyperparameters for the model that produce accurate forecasts for data with trends and seasonal structure by default. [8]. The prophet is an additive model made up of the following elements:

$y(t) = g(t) + s(t) + h(t) + \epsilon_t$

It differs from the standard technique in that it attempts to fit additive regression models. Furthermore, it is quite adaptable in terms of the data provided to the algorithm. [9] Prophet only accepts data in the format of a data frame with a "ds" (date stamp) and a "y" (value to forecast). As a result, the data was modified to the proper format by appending the dates and values to the new attributes "ds," "y." The ds (date stamp) column should be in a format that pandas expect, such as YYYY-MM-DD HH:MM:SS for a timestamp and YYYY-MM-DD for a date. The y column must be numeric and indicate the measurement or attribute to be predicted. The data frame is then created with the data.frame() function. then Construct a new instance of prophet class prophet () called "prophet_model." Prophet adheres to the sklearn model API. The Prophet class object is created, and then both the fit and predict methods are called[11]. The functions listed below, which are part of the prophet library, were used in the model:

- prophet() to apply a prophet forecast.
- cross_validation() is used to perform a cross-validation test on the prophet model before it is used.
- performance_metrics() is used to construct the performance MAPE measure based on the results.
- prophet_plot_components() plots the components of a prophet forecast, which prints the trend, weekly, and yearly.

Using cross cross_validation (), the period, which is the number of times between the cut-off dates, a horizon, the number of days, and the initial, which is the first training period, are all determined. The output will be a data frame containing the forecast "yhat," the actual value "y," and the cut-off date. The performance_metrics () function returns a table containing several prediction performance metrics, as seen in Figure.5.

	ds	yhat	yhat_lower	yhat_upper	y	cutoff
0	2002-02-11	646.435403	622.891662	671.418864	635.95	2002-02-10
1	2002-02-12	646.650723	623.066245	671.884316	643.50	2002-02-10
2	2002-02-13	645.748603	621.958836	671.471239	650.95	2002-02-10
3	2002-02-14	647.247068	621.287358	670.973468	690.30	2002-02-10
4	2002-02-15	647.944469	624.177850	673.357151	668.40	2002-02-10

TABLE III
CROSS_VALIDATION TABLE

Use plot_cross_validation_metric() to plot RMSE as displayed in Figure 6.

	horizon	mse	rmse	mae	mape	mdape	smape	coverage
0	36 days	148093.198758	384.828791	250.917779	0.245839	0.106257	0.267879	0.512418
1	37 days	149707.880248	386.921026	251.633118	0.247053	0.105619	0.268505	0.511432
2	38 days	150713.098540	388.217849	252.601238	0.248400	0.105476	0.271269	0.509091
3	39 days	151281.405390	388.949104	253.275370	0.249736	0.105619	0.272683	0.506779
4	40 days	152288.137502	390.241127	254.104633	0.251217	0.106146	0.274170	0.503015

TABLE IV
PERFORMANCE_METRICS() TABLE

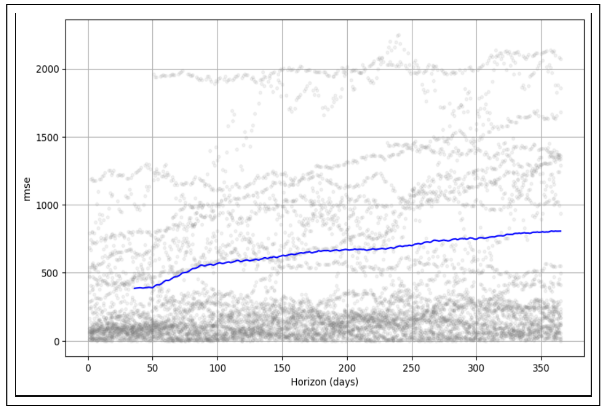


Fig. 6. RMSE PLOT

Using make_future_dataframe(), perform forecasting on the dataset. The predict() function was used for saving the data frame forecast and making predictions. Calling forecast() to review the predictions, inspect the data frame, and print the value of the prediction. Forecasting the prophet model yielded a prediction that stocks would rise, as seen in Fig.7.

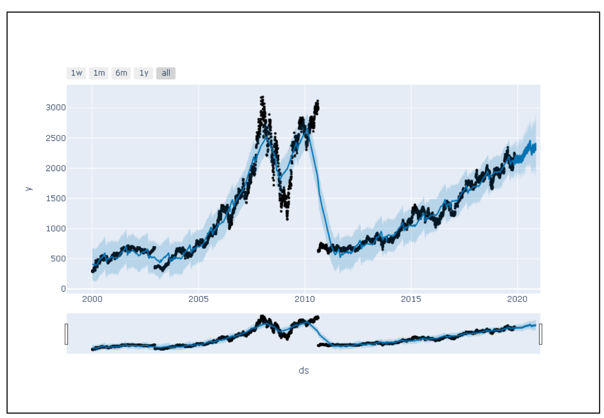


Fig. 7. Prophet Forecasting

plot_components() function was employed to inspect the forecast components, as shown in Figure. 8.which are essential in analysing the trend.

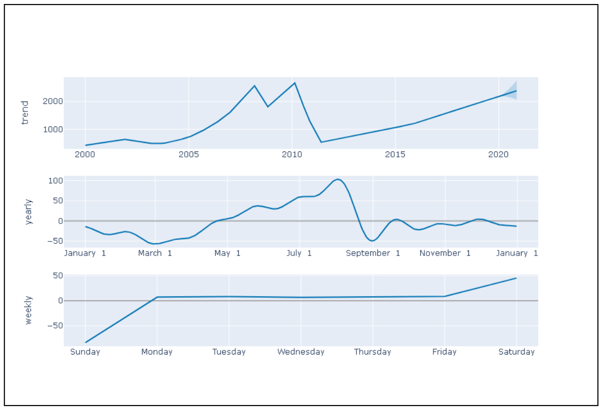


Fig. 8. Forecast Components

The Root Mean Square Error (RMSE) is a popular metric for analysing the accuracy of a model's prediction.Fig.9

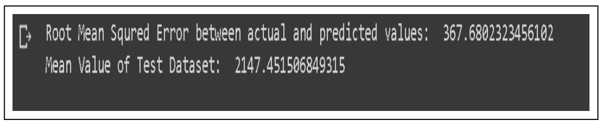


Fig. 9. prophet Model RMSE Result

Prophet, like most time series forecasting algorithms, attempts to extract trend and seasonality from historical data.Fig.10

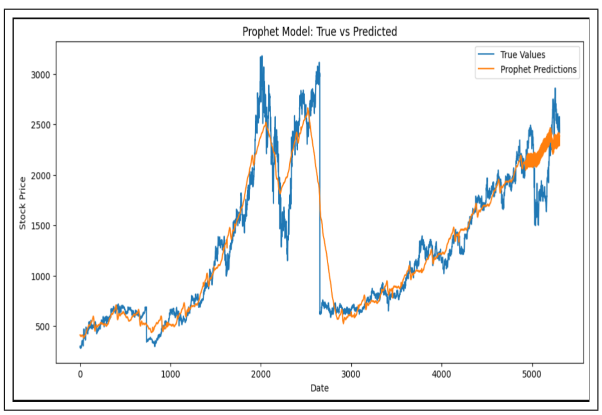


Fig. 10. prophet Model RMSE Result

Prophet offered promising results. There was a moment when the prediction (in orange) intersected with the actual price (blue).

B. LSTM model

A neural network is an architecture for processing dispersed and parallel information that comprises of interlinked and unidirectional signal channels called connections and processing components called neurons. Each processing element may possess as many output connections as it wants and emits signals known as neuron output signals. The neuron output signal can be of any desired mathematical form[14]. Recurrent Neural Nets have solved the Gradient Descent problem, which prohibits them from learning from previous data. Long-term memory networks, also known as LSTMs, have been used to solve this problem[1]. LSTM addresses the problem of learning to recall information over time intervals by incorporating memory cells and gate units into Neural network architecture. A typical formulation employs memory cells, each of which has a cell state that preserves previously encountered information. When an input is sent into a memory cell, the output is decided by a combination of the cell state (which represents prior information) and the cell state is updated. When another input is introduced into the memory cell, the updated cell state along with the new input can be used to compute the new output[7]. The algorithm implements by Keras library and uses "mean squared error" and "ADAM" as the loss function and the optimization algorithm.

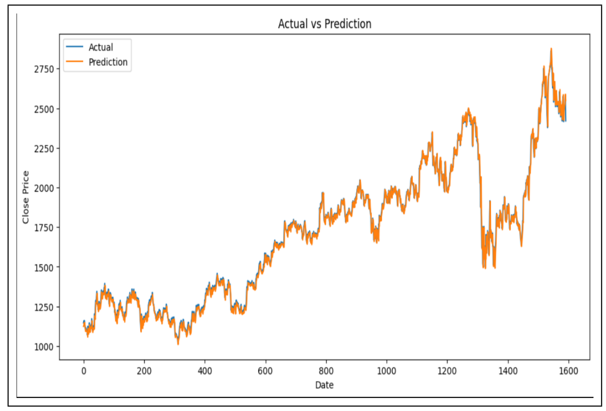


Fig. 11. LSTM Forecast

The data relating to the HDFC stock market depicts that the Rooted Mean Squared Error (RMSE) using LSTM models is 35.42093704. Fig.12.

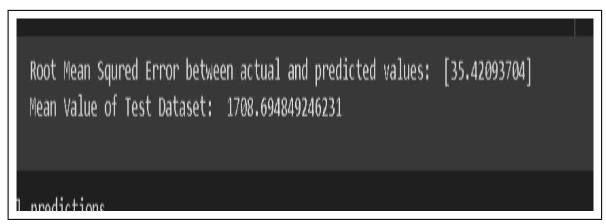


Fig. 12. LSTM RMSE

V. Results and Discussion

Revisiting the ideas of technical analysis in stock price for pattern prediction [15], it can nearly effectively predict a future stock price using LSTMs. Find these results to be highly positive and useful as a starting point for future investigation. The RMSE calculation revealed that the two models' forecasting accuracy must be appraised. The LSTM model outperformed the prophet in terms of accuracy. The LSTM projection of HDFC stocks demonstrated consistency in value, with this prediction to the next year 2021/22 indicating an enormous rise in stock value. Prophet algorithm was not as robust as an LSTM implementation.

VI. CONCLUSIONS

The study compared the results of the prophet model with the LSTM models using HDFC stock historical data from January 3, 2000 to April 30, 2020. After multiple tests, the LSTM model produced accurate outcomes in its calculating values, demonstrating the potential of utilizing the LSTM model on time series data to reliably anticipate stock data, which will aid stock investors in their investment decisions. After comparing the two models, the researchers compared the outcomes and calculated the accuracy. Future studies will compare more than two models and calculate their accuracy to determine which one is the most accurate.

REFERENCES

[1] Ashutosh Sharma, Sanket Modak, Eashwaran Sridhar. Data Visualization and Stock Market and Prediction. International Research Journal of Engineering and Technology (IRJET). Volume: 06 Issue: 09, 2019

[2] Frank Saldivar, Mauricio Ortiz. Stock Market Price Prediction Using Various Machine Learning Approaches. 2019

[3] Stock market prediction.Available: https://en.wikipedia.org/wiki/Stock_market_prediction.

[4] Xin-Yao Qian. Financial Series Prediction: Comparison Between Precision of Time Series Models and Machine Learning Methods. 2017

[5] . SIMA SIAMI NAMIN, AKBAR SIAMI NAMIN. FORECASTING ECONOMIC AND FINANCIAL TIME SERIES: ARIMA VS.LSTM.2018

[6] PETER FOY. Machine Learning for Finance: Price Prediction with Linear Regression. 2019.Available:https://www.mlq.ai/price-prediction-with^linear-regression/

[7] CHAU Tsun Man, SUEN Heung Ping, TO Cheuk Lam, WONG Cheuk Kin. Stock Price Prediction App using Machine Learning Models Optimized by Evolution. 2019.

[8] Jason Brownlee. Time Series Forecasting with Prophet in Python 2020 Available: https://machinelearningmastery.com/time-series-forecasting^with-prophet-in-python/

[9] Kan Nishida. An Introduction to Time Series Forecasting with Prophet in Exploratory. 2017. AvailableL: https://blog.exploratory.io/an^introduction-to-time-series-forecasting-with-prophet-package-in^exploratory-129ed0c12112

[10] SIMA SIAMI NAMIN, AKBAR SIAMI NAMIN.FORECASTING ECONOMIC AND FINANCIAL TIME SERIES: ARIMA VS. LSTM. 2018

[11] Ashish Vishwakarma, Alok Singh, Avantika Mahadik, and Rashmita Pradhan. Stock Price Prediction Using Sarima and Prophet Machine Learning Model. International Journal of Advanced Research in Science, Communication, and Technology (IJARSCT). Volume 9, Issue 1, September 2020

[12] Shakir Khan and Hela Alghulaiakh, "ARIMA Model for Accurate Time Series Stocks Forecasting." International Journal of Advanced Computer Science and Applications (IJACSA), 11(7), 2020. Available: http://dx.doi.org/10.14569/IJACSA.2020.0110765

[13] Shakir Khan and Amani Alfaifi, "Modelling of Coronavirus Behaviour to Predict it is Spread" International Journal of Advanced Computer Science and Applications (IJACSA), 11(5), 2020.Available:http://dx.doi.org/10.14569/IJACSA.2020.0110552

[14] Abu Sarwar Zamani, Nasser Saad Al-Arifi and Shakir Khan,. Response Prediction of Earthquake motion using Artificial Neural Networks. International Journal of Applied Research in Computer Science and Information Technology. 2012. Vol. 1, No. 2, pp. 50-57.

[15] Available: https://koreascience.kr/article/JAKO202213042317199.pdf.