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

Mohammed Ali Alshara

College of Computer and Information Sciences
Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, Saudi Arabia

Abstract
Forecasting and time series modelling plays a vital role in the data analysis process. Time Series is widely used in analytics & data science. Forecasting stock prices is a popular and important topic in financial and academic studies. A stock market is an unregulated place for forecasting due to the absence of essential rules for estimating or predicting a stock price in the stock market. Therefore, predicting stock prices is a time-series problem and challenging. Machine learning has many methods and applications instrumental in implementing stock price forecasting, such as technical analysis, fundamental analysis, time series analysis, statistical analysis. This paper will discuss implementing the stock price, forecasting, and research using prophet and LSTM models. This process and task are very complex and involve uncertainty. Although the stock price never is predicted due to its ambiguous field, this paper aims to apply the concept of forecasting and data analysis to predict stocks.

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

1. Introduction

Various imminent characteristics of people's lives rely on historical data arithmetic analysis. For example, prediction of illness, changes in stock market activities, weather prediction. can be forecasted if a pattern in historical data is due to time. It can be for example, daily, weekly, monthly, or annually. This form of the forecast is commonly called Forecasting of the Time Series. Observations were sequentially taking in time, usually called time series [13].
The increasing availability of historical data with the need for production forecasting has attracted the attention of Time Series Forecasting (TSF), which gives a sequence of predicting future values, especially with the limitations of traditional forecasting, such as complexity and time-consuming. [12]

Investment firms and even individuals use financial models better understand market behaviour and make profitable investments. [1] Stock price prediction attempts to determine the future value of a company's stocks or other publicly traded financial instruments. A successful forecast of the future price of the stock may yield a significant profit [3]. Predicting how the stock markets will perform is a difficult thing to do. [4] Countless factors move the stock price [2]. There are many factors involved in forecasting - physical factors Vs. Psychological, rational, irrational behaviour. These aspects make stock prices volatile and difficult to predict with a high degree of accuracy. One of the prevailing theories says that stock prices are completely random, and their value cannot be predicted. This theory raises the question of why large companies are employing quantitative analysts to build predictive models [1].
Is machine learning predicting stock prices effective? Investors make guesses calculated by analysing the data. [1, 12] Using some of the features such as the latest announcements about the organization, quarterly revenue results, read the news, study company history, industry trends, and many other variables. Machine learning technologies can discover patterns and insights that we do not see and have not seen before and can be used to make accurate, unerringly predictions.

This paper seeks to use machine learning models, prophet, and Long Short-Term Memory (LSTM) to predict prices. Work is done with a historical dataset for the stock price of a listed company (Google inc.). One machine-learning algorithm to predict the company's future stock price will be implemented using advanced and popular techniques; the name is a prophet.

The company may become vulnerable to market fluctuations outside of control, including market sentiment, economic conditions, or developments in the sector.
The hypothesis for this experiment is that LSTMs will demonstrably outperform other techniques as a prophet and provide more in-depth insight into the technical analysis's validity.

Manuscript received February 5, 2022
Manuscript revised February 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.2.24

2. Literature Review

In this section, foundations and basic working definitions are provided. An overview of the relevant purposes and concepts is fundamental and technical analysis, which are non-machine learning methods for stock valuation, and machine learning approaches.
Forecasting financial time series has always been an important topic and an exciting research area with many business, economics, finance, and computer science applications.

Time series analysis aims to study path observations of time series and build a model to describe the data structure and predict future time series values. Due to the importance of time series prediction in many applied science branches, it is necessary to build a useful model to improve prediction [5]. The prevailing traditional methods of dealing with the problem consist mainly of fundamental analysis and technical analysis. Simultaneously, there are more experiments to introduce new advanced techniques such as machine learning for forecasting in recent years [4]. Multiple data visualization is shown for different pattern creation discussed in [15] and Student performance prediction model is introduced in [16] applying data mining regression model approach and getting the outcomes via some study factors from the dataset. Another simulation prediction outcomes is discussed in [17] by COVID-19 data effects in Saudi Arabia.

2.1 Fundamental Analysis

Traditional methods of analysing the stock market and forecasting stock prices include fundamental analysis that looks at the stock's performance and the company's general credibility, and statistical analysis that is concerned only with multiplying numbers and identifying patterns in stock price variation [1].

In general, the fundamental analysis attempts to analyse some of the macro features that the company shows. It is based on its principles that the market value tends to move towards the real deal or intrinsic value [4]. The fundamental analysis refers to the stock's valuation, considering the company's information, related news, the general economy, and the specific economy of the company's sector, among other factors [2].

2.2 Technical Analysis

Technical analysis is almost the opposite of fundamental analysis. Technical analysis is an analysis methodology to predict and studying past market data, price, and volume.
In general, investors who use this approach formulate their trading strategy based on some technical indicators calculated according to price, volume, and time [4]. The only input to technical analysis is past stock price data. The technical analyst believes that the previous pattern in the stock indicates future designs and prices [2].

2.3 Analysis Based on Models

There are many different machine learning algorithms and approaches and finding the right method has proven challenging [2]. Time series models and machine learning models are independent of technical and fundamental analysis. They rely on mathematical theories and devise useful models by entering training data. The derived model can then be used to predict new data [4].

This paper proposed using a machine-learning model for predicting the price of a given stock.
This project's challenge accurately predicts the future closing value of a given stock across a given period in the future. This project was being used a prophet model and Long Short-Term Memory network, usually called "LSTMs," to predict Google's price in this paper and using a data set of past prices.

3. The Research Method

In this paper, Quantitative methods is applied using models and python code to analyse and visualize the data.

3.1 Analytic models:

Initially tried the prophet model to predict the stock prices using historical closing prices and visualize both the predicted price and values over time. The model predicts five years of data points based on the test data set. Then have used Long Short-Term Memory networks to predict Google's closing price using a data set of past prices.

This project used Root Mean Squared Error (RMSE) as a performance measure to calculate the difference of predicted and actual values of stock at the adjusted close price between the performance of the model (prophet) and model (LSTM).

3.2 Exploring the Stock Prices Dataset

The dataset used in this paper is of Google from October 7, 2015, to October 7, 2020. This type of data is a series of data points indexed in time order or a time series. The goal is to predict the price for any given date after training. For ease of reproduction and reusability, all data was pull from the Yahoo finance Python API.

There are multiple variables in the dataset: date, open, high, low, close, adj close, and volume.

- The columns *Open* and *Close* represent the starting and final price at which the stock is traded on a particular day.
- *High*, *Low,* and *close* represent the maximum, minimum, and last price of the day's share.
- The adjusted closing price amends a stock's closing price to reflect its value after accounting for any corporate actions [3].
- Volume- it is the amount of an asset or security that is subject to change during a specific period, often over a day [3].

The prediction must be making for the adjusted closing price. Yahoo finance already adjusts the closing prices; it just needs to make predictions for the "CLOSE" price. [1] The "Adjusted Close" variable is the only feature of financial time series to be fed into the prophet and LSTM models [10].

Setup starting by importing all necessary libraries (NumPy), (pandas), (matplotlib). Load the dataset and define the target variable for the problem. Then import the CSV file into Python using read_csv () from pandas. The dataset is of the following form (Table.1):

	Date	Open	High	Low	Close	Adj Close	Volum
0	2015-10-07	649.239990	650.609009	632.150024	642.359985	642.359985	209270
1	2015-10-08	641.359985	644.450012	625.559998	639.159973	639.159973	218210
2	2015-10-09	640.000000	645.989990	635.317993	643.609985	643.609985	164870
3	2015-10-12	642.090027	648.500000	639.010010	646.669983	646.669983	127520
4	2015-10-13	643.150024	657.812012	643.150024	652.299988	652.299988	180770

Table .1: Head of the data set

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

	Open	High	Low	Close	Adj Close	Volume
count	1259.000000	1259.000000	1259.000000	1259.000000	1259.000000	1.259000e+03
mean	1042.866910	1052.814744	1033.208519	1043.459510	1043.459510	1.685598e+06
std	241.203969	244.808889	238.216019	241.706492	241.706492	7.773399e+05
min	640.000000	644.450012	625.559998	639.159973	639.159973	3.475000e+05
25%	801.994995	805.299988	795.640015	800.355011	800.355011	1.212700e+06
50%	1058.069946	1070.920044	1047.099976	1057.790039	1057.790039	1.482000e+06
75%	1196.955017	1206.403991	1187.879028	1198.625000	1198.625000	1.912300e+06
max	1709.713989	1733.180054	1666.329956	1728.280029	1728.280029	6.653900e+06

Table.2: Mean SD, Max, and Min of the dataset.

Infer to the data set that date, high and low values are not essential features of the data. The features High, Low, Volume important, but it observed that Open and Close prices have a direct relation. It matters the opening price of the stock and closing prices of the stock. If have higher closing prices than the opening prices that have some profit otherwise see losses.

The volume of stocks is also essential. The rising market should see rising volume, i.e., increasing price and decreasing volume show a lack of interest and warning of a potential reversal. A price drop (or rise) on large volumes is a stronger signal that something in the stock has fundamentally changed [1].

The upcoming sections explore these variables and use different techniques as a prophet to predict the stock's daily closing price. Hence, removed high, low, close the volume features from the data set during the processing step (fig.1).

```
Date
2015-10-09    643.609985
2015-10-12    646.669983
2015-10-13    652.299988
2015-10-14    651.159973
2015-10-15    661.739990
Name: Adj Close, dtype: float64
```

Fig.1: data after removing High and low features

The mean, standard deviation, maximum, and minimum of the processed data was found to be following (fig.2)

```
count     1259.000000
mean      1045.474054
std        241.875405
min        642.609985
25%        801.837494
50%       1060.619995
75%       1200.299988
max       1728.280029
Name: Adj Close, dtype: float64
```

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

3.3 Exploratory Visualization to Visualize The Data:

In this paper used the Matplotlib python package for the initial graphing of the data set. (Fig.3) show the hysterical data plotted in scale.

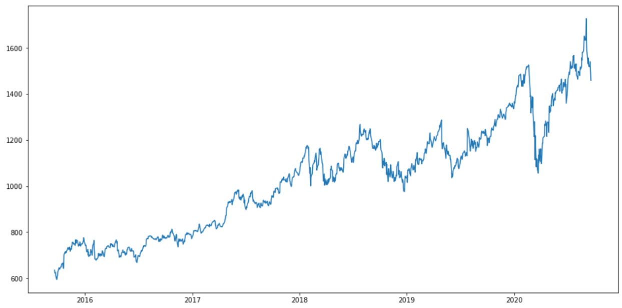

Fig.3: Visualization of processed hysterical data fetched from the API

The closing price of a stock usually determines the profit or loss calculation for the day; hence, it will consider the target variable's adj closing price. So, plot the target variable to understand how it shapes up in the data (fig.2).

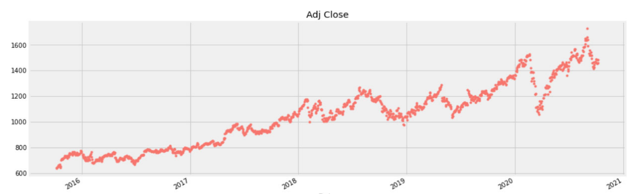

Fig.2: Representing of Google Stocks Adjusted Closing Values.

Correlation is a measure of the correlation between two features: how much Y will vary with a variation in X. The correlation method that used is name as the Pearson Correlation. Coefficient is a popular to measure correlation, as the range of values ranges from -1 to 1. In mathematics terms, it can be understanding as if two features are positively correlated. They are directly proportional, and if they share a negative correlation, then they are inversely proportional. [6] Not all text is understandable, so visualize the correlation coefficient (fig.3).

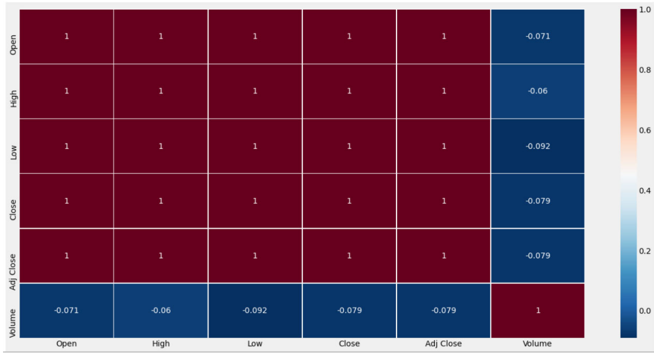

Fig.3 Correlation Map

The Dark Maroon zone denotes the highly correlated features.

4. Algorithms and Techniques Used

This paper aims to study time-series data and explore as many options as possible to accurately predict the Stock Price.

4.1 Prophet model:

Several time series techniques can be implemented on a stock prediction dataset, but most of these techniques require much pre-processing of the data before constructing the model. [3]

The prophet is an open-source library designed for forecasts for time-series datasets. It is easy and designed to automatically find a good set of hyperparameters for the model to make skilful forecasts for data with trends and seasonal structure by default. [8]

The prophet is an additive model with the following components: $y(t) = g(t) + s(t) + h(t) + \epsilon_t$

The prophet is an algorithm to build forecasting models for time series data. It is unlike the traditional approach as it tries to fit additive regression models. Moreover, it is very flexible when it comes to the data that is fed to the algorithm. [9]

Prophet only takes data as a data frame with a "ds" (date stamp) and "y" (value want to forecast). Therefore, the data had been converted to the appropriate format by adding the dates and value to the new attribute "ds," "y." The ds (date stamp) column should be a format expected by pandas; it can be of any format like YYYY-MM-DD HH:MM: SS for a timestamp and YYYY-MM-DD for a date. The y column must be numeric, and it should represent the measurement or attribute which needs to forecast.

Then create the data frame by use data.frame() function. then Fit prophet class prophet () into a new instance named "m." Prophet follows the sklearn model API. The instance of the Prophet class is created and then call its fit and predict methods. [11] The functions in the list below were use in the model, which are part of the prophet library:

- cross_validation () to apply a cross-validation test for testing the accuracy of the prophet model before use.
- performance_metrics () to compute the performance MAPE metric on the output of our cross-validation.
- prophet () to apply for the prophet forecast.
    - prophet_plot_components () to plot components of a prophet forecast, which will print with the trend, weekly, yearly.

Using cross cross_validation () determines the period, which is the number of times between cut-off dates, a horizon, the number of days, and the initial, which is the first training period. The result will be a data frame with the forecast "yhat," actual value" y," and the cut-off date.
Using performance_metrics () will get a table with various prediction performance metrics, as shown in Figure.5.

	horizon	mse	rmse	mae	mape	mdape	coverage
0	36 days	13042.647865	114.204413	105.120995	0.097685	0.098940	0.01
1	37 days	13634.049787	116.764934	107.967091	0.100340	0.101792	0.00
2	38 days	14168.303044	119.030681	110.462488	0.102670	0.104190	0.00
3	39 days	14421.574735	120.089861	111.944796	0.104140	0.105407	0.00
4	40 days	14961.205142	122.316005	114.372477	0.106592	0.110460	0.00

Fig.5 prophet Accuracy Results.

Use plot_cross_validation_metric () to plot RMSE as shown is Figure 6.

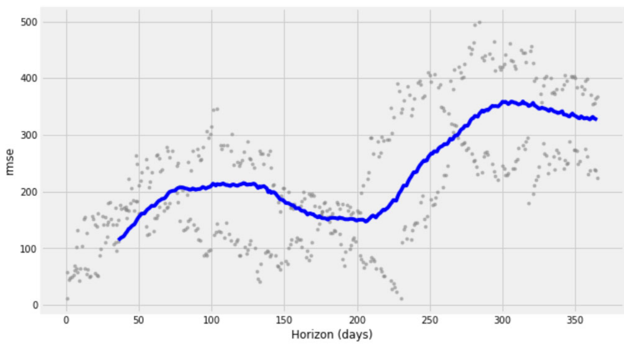

Fig. 6. RMSE plot

Apply the forecasting on the dataset using make_future_dataframe (). To store the data frame forecast and make prediction predict () function had been call. Calling forecast () to see the predictions and inspect the data frame and print the prediction's value. Forecasting the prophet model showed the prediction which predicted that stocks would go up as shown in Fig. 7

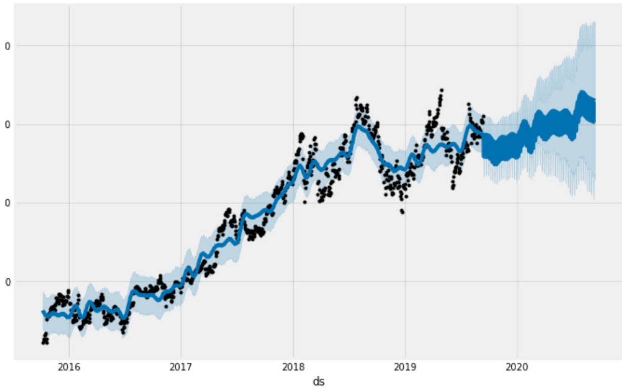

Fig.7 forecasting prophet

plot_components () function had been calling to inspect the forecast components, as shown in Figure. 8.

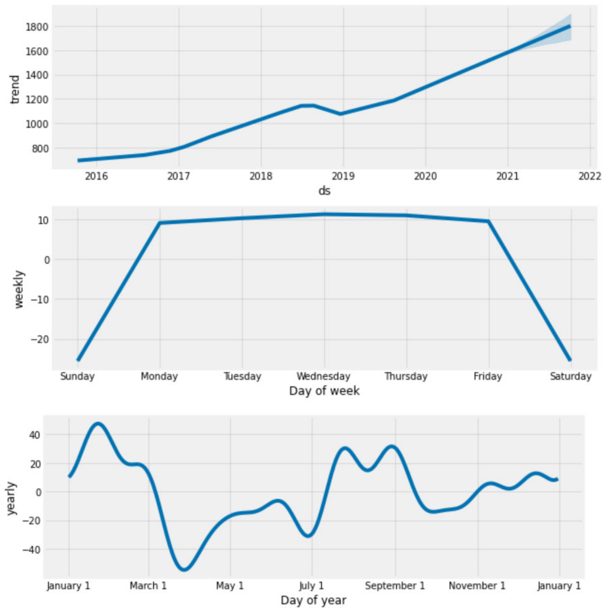

Fig. 8. Forecast components.

The values are what will take on consider here the RMSE. The model's accuracy and result are depicted as follows using the root mean square error function.
#rmse

```
forecast_valid = forecast['yhat'] [987:]
rms=np.sqrt(np.mean(np.power((np.array(valid['y']) RMS
```

The Root Mean Square Error (RMSE) is a measure frequently used for assessing the accuracy of prediction obtained by a model. [10] the accuracy results by calculating the Mean Absolute Percentage Error RMSE showed on Fig. 9

```
228.93653882755595
```

Fig. 9. Prophet Model RMSE result.

Prophet (like most time series forecasting techniques) tries to capture the trend and seasonality from past data. This model usually performs well on time-series datasets but fails to live up to its reputation in this case. fig.10

Fig.10 RMSE

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

4.2 LSTM model

A neural network is an architecture for processing distributed and parallel information that consists of processing elements called neurons, interconnected, and unidirectional signal channels called connections. Each processing element branches into as many output connections as desired and carries signals known as a neuron output signal. The neuron output signal can be of any mathematical type desired. [14]

Recurrent Neural Nets have vanished Gradient descent problem, which does not allow it to learn from past data. This problem has been solving by using long-term memory networks, usually referred to as LSTMs. [1]

Long short-term memory was first introduced by Hoch Reiter and Schmid Huber in 1997 to address the problems. Long-short term memory tackles learning to remember information over a time interval by introducing memory cells and gate units in the neural network architecture. A typical formulation involves the use of memory cells, each of which has a cell state that stores previously encountered information. Every time an input is passed into the memory cell, the output is determined by a combination of the cell state (representing the previous information), and the cell state is updated. When another input is passed into the memory cell, the updated cell state and the new input can compute the new output [7].

The algorithm implements by Keras library along with Theano were install on a cluster of high-performance computing centre 10]. The algorithm defines a function called "fit" to build the LSTM model. It takes the training dataset, the number of epochs and the number of times a given dataset is fit, and the number of neurons and the number of memory units or blocks. When the network is built must be compiled to comply with the mathematical notations used in Theano. When compiling a model, the loss function must be defined together with the optimization algorithm.

The "mean squared error" and "ADAM" are used as the loss function and the optimization algorithm.

After compilation, it is time to fit the model to the training dataset. Since the network model is stateful, the network's resetting stage must be managed and controlled, especially when there is more than one epoch.

Furthermore, since the objective is to train an optimized model using earlier stages, it is necessary to set the shuffling parameter to false to improve the learning mechanism.

A small function is created to call the LSTM model and predict the next step in the dataset. The algorithm's active part starts where an LSTM model is built with a given training dataset, the number of the epoch, and neurons. Furthermore, the forecast is taking place for the training data. Then use the built LSTM model to forecast the test dataset and report the obtained RMSE values fig.11.

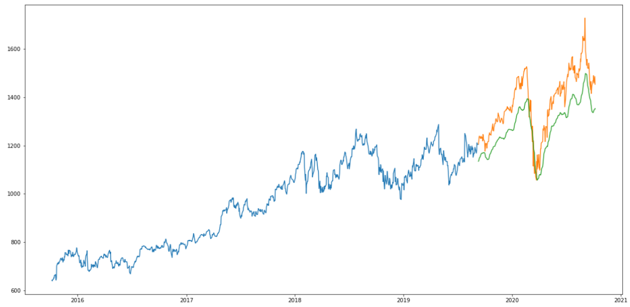

Fig. 11 forecast LSTM

The data related to the Google stock market show that the average Rooted Mean Squared Error (RMSE) using LSTM models are 78.831. Fig.12.

```
78.83160285504756
```

Fig.12. LSTM Accuracy Result.

5. Results and Discussion

Recalling the ideas of technical analysis in stock price for pattern prediction [15] shows that with the use of LSTMs, it can nearly correctly predict a future stock price. Consider these results to be very favourable and can serve as a baseline for future work. The RMSE calculating showed that the accuracy of forecasting the two models must value. The LSTM model showed better accuracy than the prophet. The prediction of Google stocks on LSTM showed continuity in value, where this prediction to the next year 2021/22, there will be a significant increase in the value of stocks.

Prophet algorithm was not as robust as an LSTM implementation. Considering that our only data input was previous stock prices as training data, to predict the next year of future stock price movement, which high accuracy shows the prowess of LSTMs and recurrent neural networks.

6. Conclusion

The research used Google stocks historical data for the past five from October 7, 2015, to October 7, 2020, to compare the prophet model and LSTM models' results. After several tests, LSTM showed accurate results in its calculating values, which showed the potential of using the LSTM model on time series data to accurately predict stock data, which will help investors in stocks in their investment decisions. The forecasting of Google stocks on LSTM showed continuity in value. This research compared the results and calculated the accuracy based on two models. Future work will compare more than two models and calculate the accuracy to reach the most accurate one.

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. 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. 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 https://machinelearningmastery.com/time-series-forecasting-with-prophet-in-python/

[9]. Kan Nishida. An Introduction to Time Series Forecasting with Prophet in Exploratory. 2017. 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. 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. 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]. S. Khan, "Data Visualization to Explore the Countries Dataset for Pattern Creation", *Int. J. Onl. Eng.*, vol. 17, no. 13, pp. pp. 4–19, Dec. 2021.
[16]. Shakir Khan (2021). "Study Factors for Student Performance Applying Data Mining Regression Model Approach", IJCSNS International Journal of Computer Science and Network Security, Vol. 21 No. 2, pp. 188-192. https://doi.org/10.22937/IJCSNS.2021.21.2.21
[17]. S. Khan, "Visual Data Analysis and Simulation Prediction for COVID-19 in Saudi Arabia Using SEIR Prediction Model", *Int. J. Onl. Eng.*, vol. 17, no. 08, pp. pp. 154–167, Aug. 2021.

Mohammed Ali Alshara received BSc from Imam Mohammad Ibn Saud Islamic University in 2002 and MSc in 2008 and PhD in 2016 from University of North Texas. He is currently working as Assistant Professor and Vice Dean for Quality and Development in College of Computer and Information Sciences (CCIS), Imam Mohamad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia. His research interest includes Data Analytics, Knowledge Management Systems, and Information Retrieval.