A Comparative study on the Prediction of Covid-19 cases in India using Regression Algorithms

B Aishwarya[1], N V Subba Reddy[2] ,B Ashwath Rao[3]
Department of Computer Science and Engineering, Manipal Institute of Technology,
Manipal Academy of Higher Education, Manipal, India
[1] aishwarya.b4@learner.manipal.edu
[2] nvs.reddy@manipal.edu
[3] ashwath.rao.b@gmail.com

Abstract. Forecasting real-time data using Machine learning has proved its significance in expecting better outcomes to improve future decision-making. ML models have been used in many application areas that require recognition and emphasization of harmful aspects. In this study we have tried to use Machine Learning models to forecast upcoming COVID-19 cases in India, considering the problem to be a burning topic at the current time. In general, three Machine Learning models, Random Forest Regressor, Long Short Term Memory (LSTM), and FB Prophet model have been discussed to forecast the number of Active cases for the next 30 days. The results produced by the study prove it is a good technique to use these models for the prediction of the COVID-19 pandemic.

Keywords: COVID-19, Random Forest Regressor, LSTM, FB Prophet, Time series forecast

1.Introduction

The novel coronavirus, also called COVID-19, is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The epidemic was first recognized in Wuhan, China, in December 2019, which has rolled out worldwide, becoming a major medical issue all over the globe[1]. Coronaviruses (CoV) are a vast class of infections that causes sickness and various other symptoms in humans usually led by the cold, for example, the Middle East respiratory syndrome coronavirus (MERS-CoV) and severe acute respiratory syndrome coronavirus (SARS-CoV) are the various syndromes related to coronavirus. The COVID-19 is another genus of the coronavirus family found in 2019[2], whose symptoms might vary from mild to moderate. Mild symptoms involve fever, cough, headache, fatigue, breathing difficulties, and loss of smell and taste. In contrast, severe symptoms include difficulty in breathing, chest pain, etc and it requires immediate hospitalization. COVID-19 usually gets transmitted from an infected person's mouth or nose in small liquid particles form when they cough, sneeze, speak, or breath. The risk of other people getting infected is higher when people are nearby distance within 1 meter for a long time, and even particles can be transmitted over longer distances, particularly in poorly ventilated closed rooms. There are various other causes of infection, such as touching contaminated surfaces and objects infected by coronavirus when touching eyes, nose, and mouth without washing hands. As per data received as of 25th June 2021 from MoHFW, Government of India, the total number of confirmed coronavirus cases in India is roughly 30,082,778 with 562,530 of them being active. The demand for medical equipment and beds is continuously increasing, becoming a threat to the medical system worldwide. Thus, the need for effective forecasting using the existing trend should be in place to take early measures to prevent further spread.

Machine learning (ML) has proven itself to be one of the prominent fields of study over the years by providing solutions to many real-world problems. It is applied in various domains such as healthcare, image recognition, self-driving cars, natural language processing (NLP), stock market trading, online fraud detection, etc. One of the important domains of Machine Learning is forecasting. Several ML algorithms have been used to guide researchers, government, and people in taking further steps in the domains such as climate forecasting, epidemic forecasting, stock market forecasting as well as disease prediction. Various regression and neural network models have been applied for multiple time series-related problems. ML algorithms[4] play an essential role in epidemiological analysis and prediction. They can help mitigate epidemic patterns when large amounts of epidemiological data are present to take an early step in preventing the spread of disease. We have attempted to develop a forecasting model for COVID-19, which forecasts the number of active cases for the next 30 days. In this work, some of the best regression models such as the Random Forest regressor, Long Short Term Memory

(LSTM), and the FB Prophet model are used to solve the problem of forecasting. The covid_19_india dataset is used for training and testing the model and has the statistics from 30th January 2020 to 8th June 2021. The model has been evaluated by various parametric measures such as R-squared score (R2 score) and root mean square error (RMSE).

2.Materials and Methods

2.1 Dataset

The purpose of this work is the future forecasting of the COVID-19 layout in India, mainly focusing on predicting active cases for the next 30 days. The dataset is obtained from the Ministry of Health and Welfare, Government of India website. The covid_19_india.csv file contains a daily trend of the covid cases in India, including the number of confirmed cases, deaths, and cured cases sorted by state.

2.2 Predictive models used

This study compares the predictions made by three regression models-Random Forest Regressor, LSTM, and the FbProphet model. A brief description of the models is given below:

2.2.1 Random Forest Regressor Model: Tin Kam Housing proposed the first random decision forest algorithm in 1995 using the random subspace approach, which in Ho's formulations, is a means to achieve the"stochastic discrimination", a path to the classification introduced by Eugene Kleinberg. Random forest is a supervised learning algorithm that uses ensemble learning methods to solve regression and classification problems. In the case of regression, it works by creating a pile of decision trees at the training phase and producing the mean/mode of prediction of the individual trees as the output. The theory behind random forest is the wisdom of crowds where a huge amount of uncorrelated models performing as a representative will exceed any of the specific constituent models. There won't be any interaction among the trees of a random forest..A random forest can perform like an estimator algorithm that aggregates the outcome of many decision trees and produces the most optimal outcome as the output.

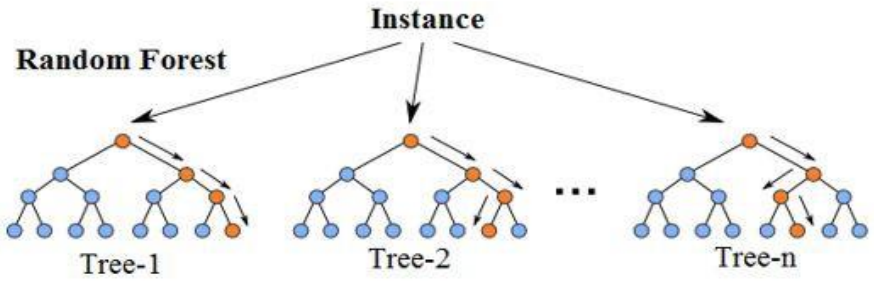

Figure 1. Random Forest Regressor model

2.2.2.Long Short-Term Memory (LSTM): Long Short-Term Memory networks, usually referred to as "LSTMs", is a sort of RNN, first introduced by Hochreiter and Schmiduber in 1997. These are used in various fields such as identification of speech, sentiment analysis, and text prediction. Signals flow through memory cells which are controlled by input, forget, and output gates. These gates monitors stored data, read and written on the cell. Google, Apple, and Amazon use LSTM for voice recognition.LSTM is an improvement over RNN. During the training phase of RNN, the data flows in the loop again and again which in turn results in considerable updates to the weights of the neural network. The cause for this is the collection of error gradients during an update, resulting in an unbalanced network. There might be cases in RNN where the values of weights become so large, resulting in overflow and NaN values. The above drawback in the RNN has led to the invention of LSTM. LSTM can solve this issue as it uses gates to control the memorizing process.LSTM learns from the sequence input file and develops the models by considering context and previous states. The cell block of LSTM reserves relevant information of earlier forms. The new information going into the cell is
managed by input, forget and output gates present within the cell, and those cell values are utilized by the output of the LSTM block for the calculation purpose.

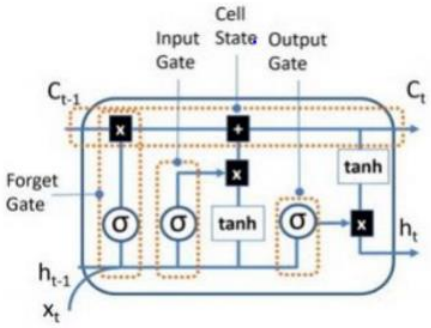

Figure 2. An LSTM unit

2.2.3 FBProphet Model: Facebook Prophet is open-source software released by Facebook's Core Data Science team. This model is used for time series forecasting. Data is based on an additive model where random values are fitted with annually, weekly, and regular seasonality and holiday effects. Prophet is vigorous to lost data and variates in the course, and it handles outliers well. It uses a decomposable time series model with three main components: annual, trends, holidays or events effect and error which are merged into this equation:

$$f(x) = g(x) + s(x) + h(x) + e(t) \quad (1)$$

FBProphet bids to fit several linear and nonlinear functions of time, using time as a regressor as components to the model. By default, FBProphet will fit the data using a linear model but it can be changed to the nonlinear model (logistics growth) by its parameters.

2.3 Performance Evaluation Parameters

The learning models are evaluated using R-square (R2) score and root mean square error (RMSE). Regression models are estimated using R-square which is a numerical method for evaluating the model. The statistics show the variance rate of the dependent variable that collectively describes the independent variable. The relation between the variable and model is computed on a 0-1 scale. Once the training is done goodness of fit can be measured by R2. Higher the R2 score indicates the goodness of the trained model.

$$R2 = Variance\ explained\ by\ model \div Total\ variance \quad (2)$$

Root mean square error is stated as the standard deviation of the prediction errors. Prediction errors, also called residuals represent the distance from the most suitable line to the original data points. RMSE is hence an estimate of how focused the original data points are around the best fit line. It is given as follows.

$$RMSE = 1 - (1 - \mathrm{R\^{}2})\frac{n-1}{n-(k+1)} \quad (3)$$

3.Methodology

The COVID-19 pandemic has proved to be a hazard to human life. It has caused crores of deaths and the death rate is increasing daily all over the world. The death rate has been growing daily throughout the world with many countries facing consequent waves of the pandemic. In this study, we have attempted to develop a model to perform future forecasting on the number of active cases for the next 30 days using historical data so that we can contribute a bit to the epidemic. The forecasting has been done by using three regressors that are suitable for these circumstances. The dataset used for this study consists of a daily time series table, consisting of the number of confirmed cases, number of death cases, and cured cases from the time when the pandemic began. Firstly, the preprocessing of data is done to compute the daily number of active cases using the data provided regarding the deaths, confirmed cases, and cured. The resultant data consisting of the number of Active cases in India concerning the dates have been used for fitting the models. The dataset is split into two subdivisions: a training set (346 days) to train the models and a test set (149 days). The different training models discussed above have been trained on

the test and training sets to predict and forecast the number of active cases for the next 30 days. The models have been evaluated based on parameter measures such as R2 -score and RMSE. The proposed method applied in the research has been shown in the form of a block diagram in Figure 3.

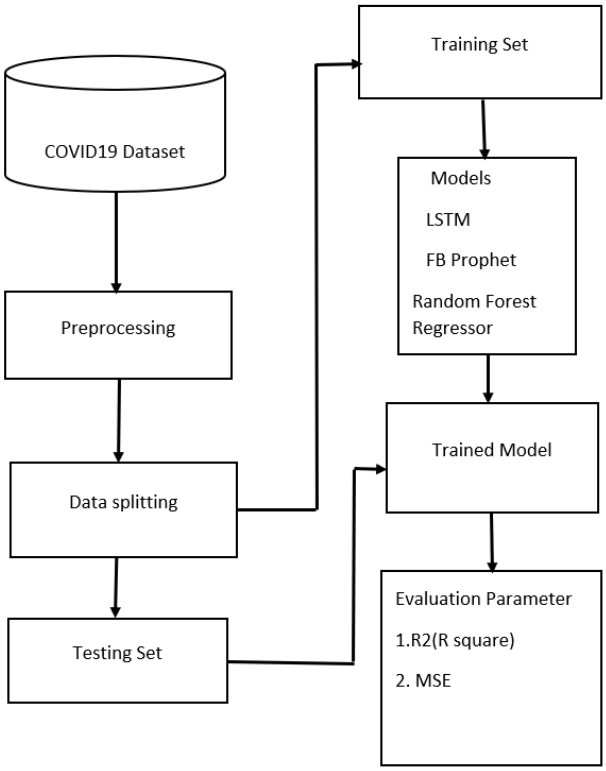

Figure 3. Proposed workflow

4.Results

This research tries to develop a model for forecasting the number of COVID-19 active cases in India using various machine learning techniques. The CSV file used for this research contains data about the everyday records of the number of newly confirmed cases, the number of cured cases, and the number of deaths due to COVID-19 in India. This study tries to forecast the number of active COVID-19 cases for the upcoming 30 days. The results have been tabulated by comparing the performance metrics of the three regressors that have been used to fulfill the requirements mentioned above. The dataset used in this model is the covid_19_india dataset containing the confirmed, death, and cured cases from 30th January 2020 to 8th June 2021 ordered by states and union territories.

	Sno	Date	Time	State/UnionTerritory	ConfirmedIndianNational	ConfirmedForeignNational	Cured	Deaths	Confirmed
0	1	2020-01-30	6:00 PM	Kerala	1	0	0	0	1
1	2	2020-01-31	6:00 PM	Kerala	1	0	0	0	1
2	3	2020-02-01	6:00 PM	Kerala	2	0	0	0	2
3	4	2020-02-02	6:00 PM	Kerala	3	0	0	0	3
4	5	2020-02-03	6:00 PM	Kerala	3	0	0	0	3
...	...	...	...	...	...	...	...	...	...
15801	15802	2021-06-08	8:00 AM	Telangana	-	-	564303	3394	593103
15802	15803	2021-06-08	8:00 AM	Tripura	-	-	49579	572	56169
15803	15804	2021-06-08	8:00 AM	Uttarakhand	-	-	313566	6731	334419
15804	15805	2021-06-08	8:00 AM	Uttar Pradesh	-	-	1662069	21333	1699083
15805	15806	2021-06-08	8:00 AM	West Bengal	-	-	1388771	16362	1432019

15806 rows × 10 columns

Figure 4. Dataset

The study performs predictions on active cases and the performance of the models has been recorded based on the r2-score and the RMSE as shown in Table 1. The results show that the Random Forest regressor outperforms the other model for the given dataset.

Table 1. Performance metrics on predictions made

Model	**R-square score**	**RMSE on test data**
Random Forest Regressor	0.998725607	29382.54
LSTM	0.997627202	61665.34
FBProphet	0.777258495	32986.84

Figures 5, 6 and 7 show the plots of the performance of Random Forest regressor, LSTM and FBProphet models respectively in the form of graphs. The graphs show the actual active cases vs the predicted values.

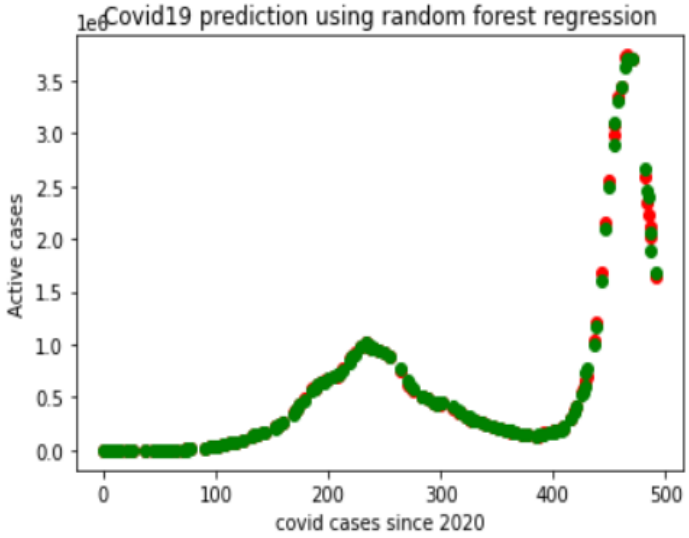

Figure 5. Actual v/s Predicted values using Random Forest

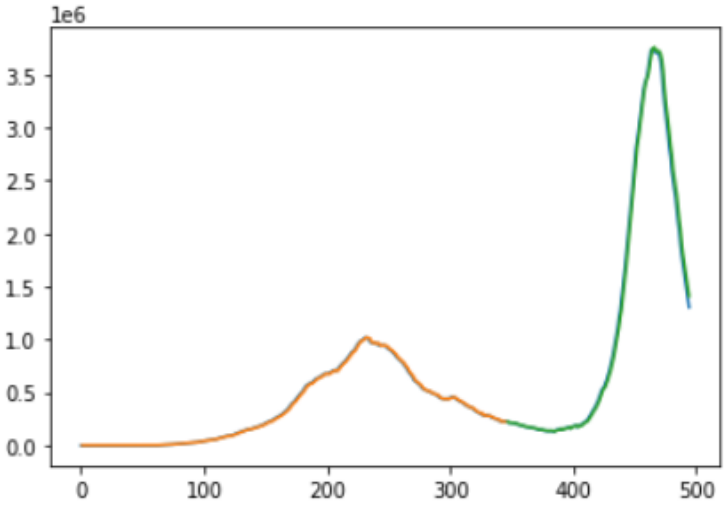

Figure 6. Actual v/s Predicted values using LSTM

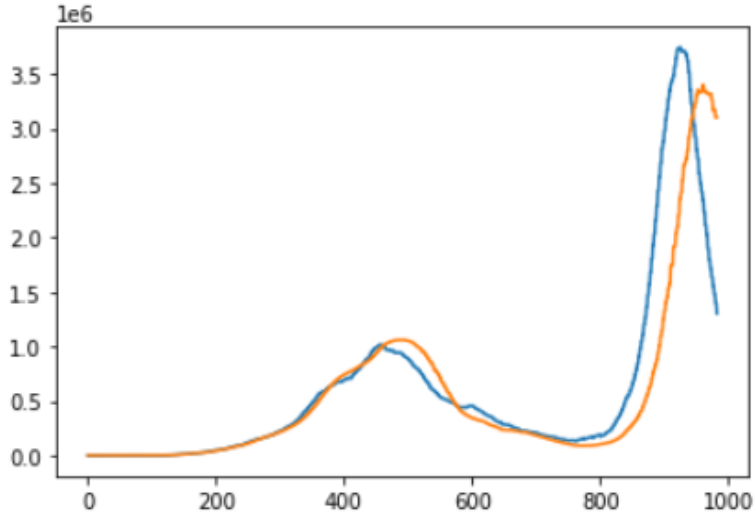

Figure 7. Actual v/s Predicted values using FBProphet

The plots and the r2 score obtained from the three models show that Random Forest and LSTM give a better fit on the actual vs the predicted values compared to FBProphet. The forecast for the next 30 days was performed using LSTM and FBProphet. Table 2 shows the results obtained for the active cases from 9th June to 8th July using FBProphet and LSTM. The values obtained from FBProphet show an upward trend in the number of instances while LSTM shows a variable trend.

5.Conclusion and Future Work

The increase of COVID-19 cases can cause a huge global disaster. Researchers and government firms worldwide have predicted that this pandemic can significantly impact the world population for the next couple of years, leading to a human crisis. In this work, ML based model has been designed for foretelling the risk of COVID-19 outburst in India. The model outlines the dataset which contains the daily reported cases and advances predictions for the forthcoming 30 days using machine learning algorithms. The study results prove that the Random Forest regressor performs comparatively better than other models for the given dataset.

Table 2.Forecast of Active cases for upcoming 30 days

Dates	LSTM active cases prediction for next 30 days	FBProphet active cases prediction for next 30 days
09-06-2021	12339190.00	3040952.00
10-06-2021	9641394.00	3067020.00
11-06-2021	8286262.00	3093089.00
12-06-2021	7443921.00	3119157.00
13-06-2021	6859218.00	3145225.00
14-06-2021	6424821.00	3171294.00
15-06-2021	6086870.00	3197362.00
16-06-2021	5815042.00	3223430.00
17-06-2021	5590814.00	3249499.00
18-06-2021	5402167.00	3275567.00
19-06-2021	5240916.00	3301635.00
20-06-2021	5101278.00	3327703.00
21-06-2021	4979036.00	3353772.00
22-06-2021	4871031.00	3379840.00
23-06-2021	4774850.00	3405908.00
24-06-2021	4688612.00	3431977.00
25-06-2021	4610824.00	3458045.00
26-06-2021	4540290.00	3484113.00
27-06-2021	4476031.00	3510181.00
28-06-2021	4417246.00	3536250.00
29-06-2021	4363266.00	3562318.00
30-06-2021	4313531.00	3588386.00
01-07-2021	4267567.00	3614455.00
02-07-2021	4224968.00	3640523.00
03-07-2021	4185388.00	3666591.00
04-07-2021	4148527.00	3692660.00
05-07-2021	4114124.00	3718728.00
06-07-2021	4081953.00	3744796.00
07-07-2021	4051813.00	3770864.00
08-07-2021	4023528.00	3796933.00

LSTM also performs well in fitting the predicted values to the actual values but has a higher RMSE than Random forest. FBProphet model doesn't perform well enough for this dataset since it doesn't provide the best fit for the given dataset. Future forecast for the next 30 days was performed using both LSTM and FBProphet and the results obtained by both the models varied by a minimum of 1,00,000 cases. By this, we can say that prediction using Random Forest and LSTM performs well for the current circumstances. It could be helpful to use this model to understand the upcoming situation. This work could be a great help for government authorities to take an early step to avoid the layout of the Covid-19 pandemic. Since India has proactively started vaccinating its citizens in the past few months, we could extend the study by considering the vaccination rate as one of the input features for prediction. This study will be further improved by working on the vaccination rate across the states. Additionally, we have planned to survey the prediction methodology using the updated dataset and other suitable ML methods for forecasting.

References

1. Stoecklin, S.B., Rolland, P., Silue, Y., Mailles, A., Campese, C., Simondon, A., Mechain, M., et al.: First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020. Euro Surveill. 25(6), 2000094 (2020)

2. Prasad, S., Potdar, V., Cherian, S., Abraham, P., Basu, A.: Transmission electron microscopy imaging of SARS-CoV-2. Indian J. Med. Res. 151, 241–243 (2020)

3. Jia, L., Li, K., Jiang, Y., Guo, X.: Prediction and analysis of coronavirus disease 2019. arXiv preprint arXiv:2003.05447 (2020)

4.Pandey, S., Solanki, A.: Music instrument recognition using deep convolutional neural networks. Int. J. Inf. Technol. 13(3), 129–149 (2019)

5. Ajay Shrestha and Ausif Mahmood, "Review of Deep Learning Algorithms and Architecture", IEEE Access, Volume 7, May 2019, DOI: doi.org/ 10.1109/ACCESS.2019.2912200

6. Arun Solanki and Tarana Singh, "COVID-19 Epidemic Analysis and Prediction Using Machine Learning Algorithms"

7. Sean J Taylor, Benjamin Letham, "Forecasting at scale"

8. J. Lupon, H. K. Gaggin, M. de Antonio, M. Domingo, A. Galan, E. Zamora, J. Vila, J. Penafiel, A. Urrutia, E. Ferrer, N. Vallejo, J. L. Januzzi, and A. Bayes-Genis, ''Biomarker-assist score for reverse remodeling prediction in heart failure: The ST2-R2 score,'' Int. J. Cardiol., vol. 184, pp. 337–343, Apr. 2015.

9. J.-H. Han and S.-Y. Chi, ''Consideration of manufacturing data to apply machine learning methods for predictive manufacturing,'' in Proc. 8th Int. Conf. Ubiquitous Future Netw. (ICUFN), Jul. 2016, pp. 109–113.

10. G. Bontempi, S. B. Taieb, and Y.-A. Le Borgne, ''Machine learning strategies for time series forecasting,'' in Proc. Eur. Bus. Intell. Summer School. Berlin, Germany: Springer, 2012, pp. 62–77