Prediction for the spread of COVID-19 in India using Machine Learning Methods

[1] Saridi Devendra Kumar [2] Prof D Lalitha Bhaskari
[1]M.tech in CST with Artificial Intelligence and Robotics,
Department of Computer Science and Systems Engineering,
Andhra University college of Engineering(A), Andhra University, Visakhapatnam,AP,India
[2]Professor, Department of Computer Science and Systems Engineering,
Andhra University college of Engineering(A), Andhra University,Visakhapatnam,AP,India.
Emails: [1]sarididevendra@gmail.com, [2]lalithabhaskari@yahoo.co.in

ABSTRACT: Nowadays, there is a very adverse impact on economic, cultural, social and almost all fields in the world because of Covid-19. The Covid-19 term is described as -'CO' for corona, 'VI' for virus, and 'D' for disease. It is an infectious disease caused by severe acute respiratory syndrome which is transmitted through respiratory droplets and contact routes. Since December 2019, corona-virus disease (COVID-19) has out-broke from the country China. The spread of COVID-19 in the whole world has put humanity at risk. The resources of some of the largest economies are stressed out due to the large infectivity and transmissibility of this disease. Due to the growing magnitude of number of cases and its subsequent stress on the administration and health professionals, some prediction methods would be required to predict the number of cases in future. In this paper, we have used data-driven estimation methods like Random Forest- regression, XGBoost, LightGBM and prophet for prediction of the number of COVID-19 cases in India: 15 days ahead and also compares India with the worldwide data and the effect of preventive measures like social isolation and lockdown and the number of tests being conducted on the spread of COVID-19. The prediction of various parameters (total number of cases, number of positive cases etc.) obtained by the proposed method is accurate within a certain range and will be a beneficial tool for administrators and health officials

KEYWORDS*:* Covid, CoronaVirus, Random Forest- regression, XGBoost, Prophet, LightGBM, evaluation metrics. Prediction

1. INTRODUCTION

On 31st December 2019, the novel Coronavirus, known as COVID-19 was reported in Wuhan, China for the very first time. Coronaviruses are infectious viruses which have adverse effects on the respiratory system of humans. The symptoms of COVID-19 may or may not be visual in infected individuals, therefore the spread rate can be faster. Till now, effective and well-tested vaccine against CoVID-19 has not been invented, only precautions are the safety measures.

World is moving through a very distressing stage by the spread of the novel coronavirus (SARS-CoV-2). It is a highly contagious disease and the World Health Organization (WHO) has declared it as a global public health emergency (L.-s. Wang et al., 2020). It originated in Wuhan, Hubei Province, People's Republic of China (PRC) in late December 2019, when a case of unidentified pneumonia was reported (Huang et al., 2020). PRC Centers for Disease Control (CDC) experts declared pneumonia as novel coronavirus pneumonia (NCP) as caused by a novel coronavirus and WHO officially named the disease COVID-19 (Huang et al., 2020). However, the International Committee on Taxonomy of Viruses (ICTV) named the virus as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). This is a class of β-coronavirus and has many potential natural hosts, intermediate hosts and final hosts as shown in Fig. 1. Due to these characteristics, there is a great challenge for prevention and treatment of the virus infection (Vellingiri et al., 2020). Despite of the large number of cases worldwide (as shown in Fig. 2 (Statista, 2020)) and low mortality rate (Liu et al., 2020) compared to SARS and the middle east respiratory syndrome (MERS) (as shown in Fig. 3 (Statista, 2020)), this virus has high infectivity and transmissibility. Preventive measures for COVID-19 include maintaining social distancing, washing hands frequently, avoiding touching the mouth, nose, and face (WHO, 2020).

Though the continuous efforts are going on , the virus has managed to spread in most of the territories in the world and the World Health Organization (WHO) has announced COVID-19 as Pandemic. Most of the countries in the world are working cooperatively and openly to bring this situation under control.Data scientists and data mining researchers can play an important role during these types of situations. They can integrate the related data and technology to better understand the virus and its characteristics, which can help in taking right decisions and concrete plan of actions.As per the daily situation report of WHO, as on 15thJune 2020 the COVID-19 transmission scenario reports 78,23,289 confirmed cases with 4,31,541 deaths globally.Data mining is a technology, developing with database as well as artificial intelligence. It is a processing procedure of extracting credible and effective novel techniques and understandable patterns from the database [1]. Artificial intelligence (AI) is a field of programming building which gives PCs an ability to learn without being unequivocally modified [2]. AI models can be used for estimating and predicting spread rate, so AI is one of the beneficial tools to fight against pandemic likeCOVID-19

The first case of COVID-19 was reported in India on 30th January 2020 with origin from China (PIB, 2020). It spreads to the maximum of districts of the country. As on 9th April 2020 the total cases reported in India are 5734 with 472 recoveries and 166

deaths (Covid-19.in, 2020). However, the rate of infection is lower as compared to other countries.

There is a lot of stress on the part of administration and health officials for accommodating patients with possible symptoms of COVID-19. So, for that some prediction tools must be used to know about the number of cases in coming days for making preparations at the administrative level (Tobías, 2020; L. Wang et al., 2020; L.-s. Wang et al., 2020).

In this paper, we propose the data-driven LSTM method and the classical curve fitting method for the prediction of the number of patients to be accommodated in the subsequent days based on the data available. The proposed model can approximately predict the number of new COVID-19 cases, so the administration can make preparations accordingly to accommodate them.

II. PREDICTION MODELS

Algorithms

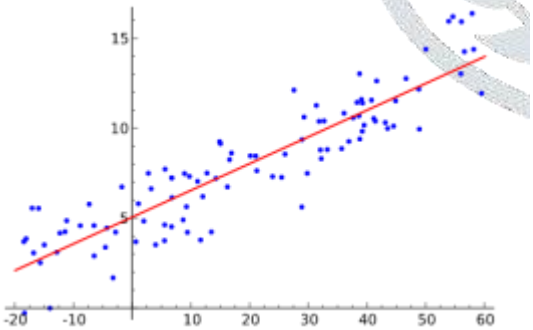
Regression algorithms:

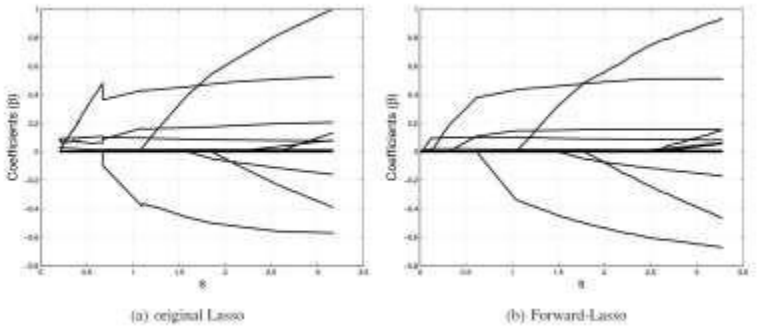
Regression algorithms fall under the family of Supervised Machine Learning algorithms which is a subset of machine learning algorithms. One of the main features of supervised learning algorithms is that they model dependencies and relationships between the target output and input features to predict the value for new data. Regression algorithms predict the output values based on input features from the data fed in the system. The go-to methodology is the algorithm builds a model on the features of training data and uses the model to predict the value for new data.

According to Oracle, here's a great definition of Regression – a data mining function to predict a number. Case in point, how regression models are leveraged to predict real estate value based on location, size and other factors. Today, regression models have many applications, particularly in financial forecasting, trend analysis, marketing, time series prediction and even drug response modeling. Some of the popular types of regression algorithms are linear regression, regression trees, lasso regression and multivariate regression

1. Simple Linear Regression model: Simple linear regression is a statistical method that enables users to summarise and study relationships between two continuous (quantitative) variables. Linear regression is a linear model wherein a model that assumes a linear relationship between the input variables (x) and the single output variable (y). Here the 'y' can be calculated from a linear combination of the input variables (x). When there is a single input variable (x), the method is called a simple linear regression. When there are multiple input variables, the procedure is referred to as multiple linear regression.

Application: some of the most popular applications of Linear regression algorithms are in financial portfolio prediction, salary forecasting, real estate predictions and in traffic when arriving at ETAs.



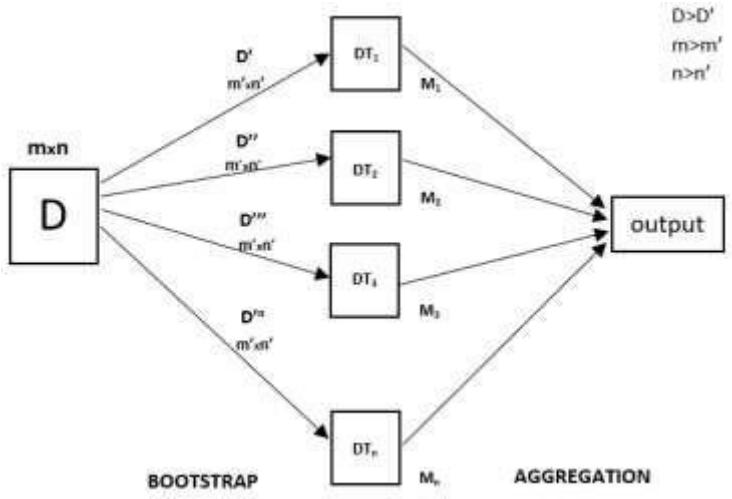


(a) original Lasso (b) Forward-Lasso

Random Forest

Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data and hence the output doesn't depend on one decision tree but multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs.

This part is Aggregation.



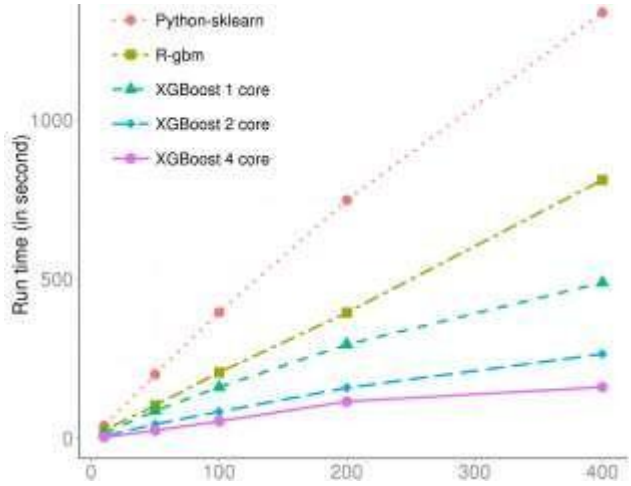
A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.
Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

We need to approach the Random Forest regression technique like any other machine learning technique

- Design a specific question or data and get the source to determine the required data.
- Make sure the data is in an accessible format else convert it to the required format.
- Specify all noticeable anomalies and missing data points that may be required to achieve the required data.
- Create a machine learning model
- Set the baseline model that you want to achieve
- Train the data machine learning model.
- Provide an insight into the model with test data
- Now compare the performance metrics of both the test data and the predicted data from the model.
- If it doesn't satisfy your expectations, you can try improving your model accordingly or dating your data or use another data modeling technique.
- At this stage you interpret the data you have gained and report accordingly.

XGBoost

XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.



In this post you will discover XGBoost and get a gentle introduction to what is, where it came from and how you can learn more.

After reading this post you will know:

- What XGBoost is and the goals of the project.
- Why XGBoost must be a part of your machine learning toolkit.
- Where you can learn more to start using XGBoost on your next machine learning project.

XGBoost is an open source library providing a high-performance implementation of gradient boosted decision trees. An underlying C++ codebase combined with a Python interface sitting on top makes for an extremely powerful yet easy to implement package.

The performance of XGBoost is no joke — it's become the go-to library for winning many Kaggle competitions. Its gradient boosting implementation is second to none and there's only more to come as the library continues to garner praise.

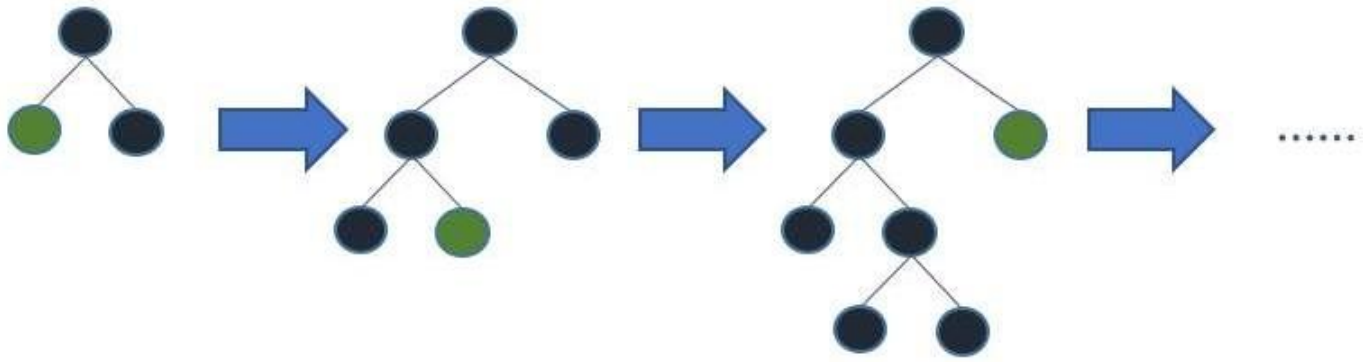
In this post we're going to go through the basics of the XGBoost library. We'll start with a practical explanation of how gradient boosting actually works and then go through a Python example of how XGBoost makes it oh-so quick and easy to do it.

LightGBM

Light GBM is a fast, distributed, high-performance gradient boosting framework based on a decision tree algorithm, used for ranking, classification and many other machine learning tasks.

Since it is based on decision tree algorithms, it splits the tree leaf wise with the best fit whereas other boosting algorithms split the tree depth wise or level wise rather than leaf-wise. So when growing on the same leaf in Light GBM, the leaf-wise algorithm can reduce more loss than the level-wise algorithm and hence results in much better accuracy which can rarely be achieved by any of the existing boosting algorithms. Also, it is surprisingly very fast, hence the word 'Light'.

Before is a diagrammatic representation by the makers of the Light GBM to explain the difference clearly.



Leaf-wise tree growth

Leaf wise tree growth in Light GBM.

Leaf wise splits lead to increase in complexity and may lead to overfitting and it can be overcome by specifying another parameter max-depth which specifies the depth to which splitting will occur.

Advantages of Light GBM

- Faster training speed and higher efficiency: Light GBM uses histogram based algorithm i.e it buckets continuous feature values into discrete bins which fasten the training procedure.
- Lower memory usage: Replaces continuous values to discrete bins which result in lower memory usage.
- Better accuracy than any other boosting algorithm: It produces much more complex trees by following leaf wise split approach rather than a level-wise approach which is the main factor in achieving higher accuracy. However, it can sometimes lead to overfitting which can be avoided by setting the max_depth parameter.
- Compatibility with Large Datasets: It is capable of performing equally well with large datasets with a significant reduction in training time as compared to XGBOOST.

III. Related Works

Upendra Kumar Tiwari & Rizwan Khan have tried to use the machine learning to analyse the current situation created by covid-19 and what may be its impact in future days. They have analyzed that the case of covid-19 inIndia is going to be the same as in Italy or South Korea. India might be going to face its worst days in future if we look at the pattern of these countries and India[5].

Herlawati tries to use a soft computing algorithm to predict the pattern of the COVID-19 pandemic in Indonesia. Support Vector Regression was used in Google Interactive Notebook with some kernels for comparison, i.e. radial basis function, linear and polynomial [6].

Dutta,Shawni , Samir Kumar Bandyopadhyay ,Tai-Hoon kim attempted to use Machine Learning Approach to build up model which will help clinical doctors for verification of disease within short period of time and also the paper attempts to predict growth of the disease in near future in the world. Experimental results indicate that the combined CNN-LSTM approach outperforms well over the other model [7].

Rustam et al. demonstrated the capability of ML models to forecast the number of upcoming patients affected by COVID-19 which is presently considered as a potential threat to mankind. Four standard forecasting models, such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES) have been used in this study to forecast the threatening factors of COVID-19. Three types of predictions are made by each of the models, such as the number of newly infected cases, the number of deaths, and the number of recoveries in the next 10 days. The results produced by the study prove it a promising mechanism to use these methods for the current scenario of the COVID-19 pandemic [8].

Ranjan, Rajesh used susceptible-infected-recovered (SIR) models based on available data to make short and long-term predictions on a daily basis. Based on the SIR model, it is estimated that India will enter equilibrium by the end of May2020[9].

Fong et al. demonstrated an optimized forecasting model that is constructed from a new algorithm, namelypolynomial neural network with corrective feedback (PNN+cf) is able to make a forecast that has relatively the lowest prediction error. The results showcase that the newly proposed methodology and PNN+cf are useful in generating acceptable forecast upon the critical time of disease outbreak when the samples are far from abundant [10].

Petropoulos, Fotios & Makridakis, Spyros introduced an objective approach to predicting the continuation of the COVID-19 using a simple, but powerful method to do so. Assuming that the data used is reliable and that the future will continue to follow the past pattern of the disease, their forecasts suggest a continuing increase in the confirmed COVID-19 cases with sizable associated uncertainty. The risks are far from symmetric as underestimating its spread like a pandemic and not doing enough to contain it is much more severe than overspending and being over careful when it will not be needed. This paper also describes the timeline of a live forecasting exercise with massive potential implications for planning and decision making and provides objective forecasts for the confirmed cases of COVID-19 [11].

Zheng N, Du S, Wang J, et al. proposed a hybrid artificial-intelligence (AI) model for COVID-19 prediction. The experimental results on the epidemic data of several typical provinces and cities in China showed that individuals with coronavirus have a higher infection rate within the third to eighth days after they were infected, which is more in line with the actual transmission laws of the epidemic [12].

Heni Bouhamed used a Deep Learning nested sequence prediction model with Long Short-Term Memory (LSTM) architecture for the continuous monitoring of the infection and recovering processes. This model was built based on the epidemic data evolution of 79 countries between the date of their first case and March 13, 2020. The data is based on 12 variables for cumulative case number prediction and 13 variables (among which the cumulative number of cases) for cumulative recoveries number prediction [13]

IV EVALUATION METRICS

Data Collection : Time series data collected from Kaggle has been used for the experimental result analysis. The time period of data is from 22/01/2020 to 15/06/2020. The data includes the cumulative count of confirmed, death and recovered cases of COVID-19 from different countries. However, this paper focuses only on India's data for analysis and forecasting of COVID-19 confirmed cases.

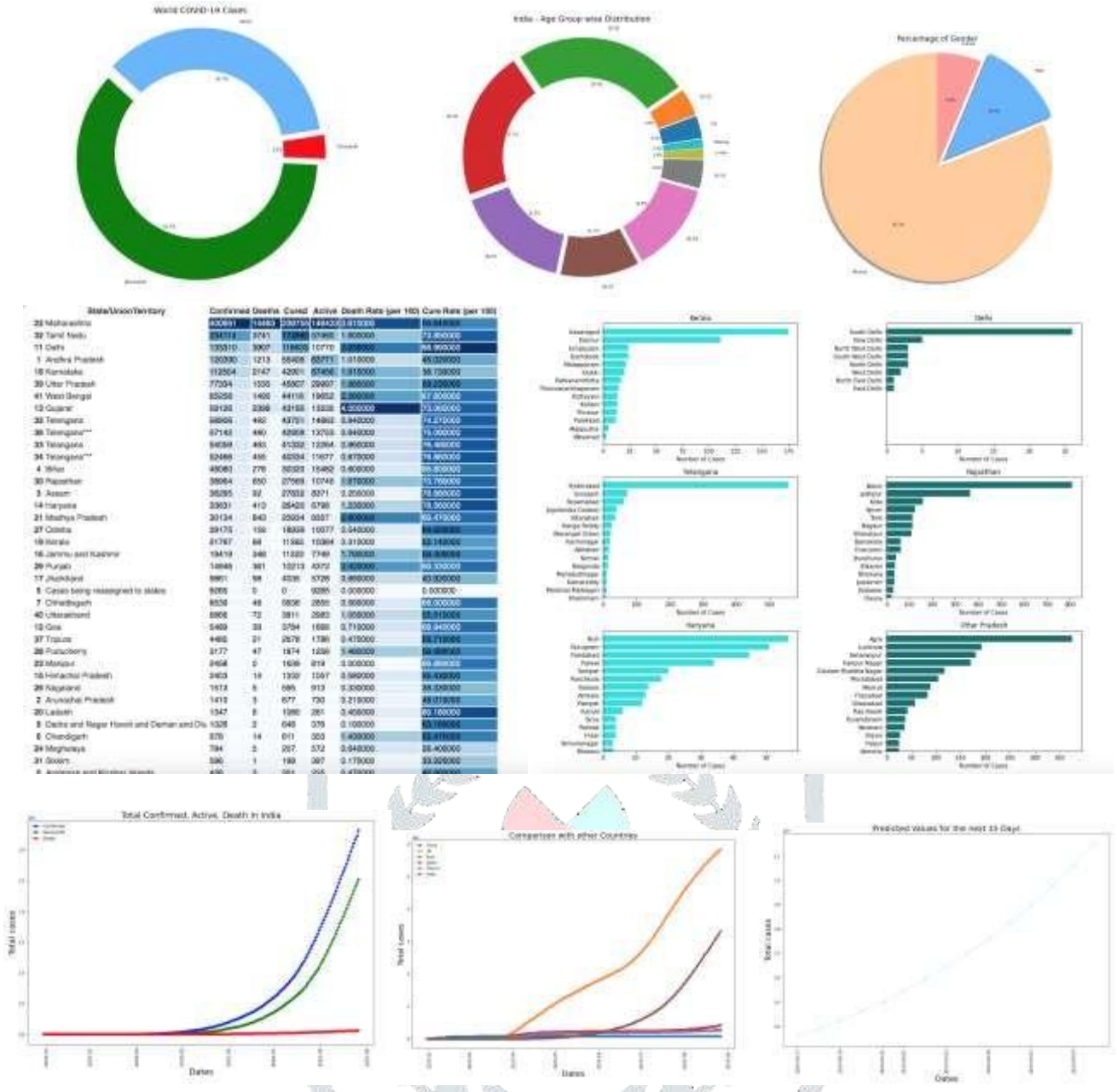
Data Preprocessing: This collected data undergoes various steps of pre-processing which makes it more sensible. Data is pre-processed by eliminating missing values, irrelevant values.

Prediction Algorithm: In machine learning, various time series forecasting models are available like ARIMA, SARIMA, GARCH, Dynamic linear models, TBATS, Prophet, LSTM, etc.Here we are using Prophet. The Prophet is a forecasting model which allows dealing with multiple seasonalities. It is open source software and is released by Facebook's Core Data Science team. The prophet model assumes that time series can be decomposed as follows:

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

The three terms g(t), s(t) and h(t) correspond respectively to trend, seasonality and holiday. The last term is the error term.

Data Analysis and Outbreak Prediction: For analysis and forecasting of the number of COVID-19 patients in India, the framework shown in figure 1 is used. The data is collected for the duration of 22ndJanuary 2020 till 15thJune 2020 from Kaggle. This dataset has everyday level data on the number of influenced cases, recovery, and deceased. It has a total of 38107 rows and 10 columns. These are the columns in the dataset: Province/State, Country/Region, Lat, Long, Date, Confirmed, Deaths, Recovered, Active and WHO Region



CONCLUSION

In this paper, a data-driven forecasting/estimation method has been used to estimate the possible number of positive cases of COVID-19 in India for the next 15 days The number cases has also been estimated by using Random Forest- regression, XGBoost, LightGBM. The effect of preventing measures like social isolation and lockdown has also been observed which shows that by these preventive measures, spread of the virus can be reduced significantly. We conclude that numbers of Covid-19 confirmed cases will increase rapidly. The idea behind this work is to make predictions about the numbers of cases in the near future. Prophet predictive analytics algorithm, Kaggle dataset is used for making predictions. The predictions show that the confirmed COVID-19 infected cases would be approximately 8.2 million cases (growing by an Average growth factor of 1.0705944599385337) by the end of October 2020. We hope that these predictions will be helpful in different sectors to take necessary actions. This study will be enhanced in the future. We plan to explore and use the most accurate and appropriate Machine Learning methodologies for forecasting real-time data.

FUTURE SCOPE

Due to the pandemic of Coronavirus and COVID-19, all countries are looking to-wards mitigation plans to control the spread with the help of some modeling techniques. This research aims to understand the complete medical perspective of this COVID-19 pandemic and how predictive analytics will empower the predictions. Analysis of various predictive analytics methods available in the literature is presented in this chapter. We have also discussed and presented the comparative analysis of various predictive analytics models and algorithms by suggesting more appropriate use cases for application. Our study indicates that there is a need for a thorough assessment of these predictive analytics algorithms based on the type of question to be answered. Application of Prophet predictive analytics algorithm on Kaggle dataset and its predictions are also presented in this chapter.

Simulation result of this model shows that the confirmed COVID-19 infected cases would be approximately 8.2 million cases (growing by an Average growth factor of 1.0705944599385337) by the end of October 2020. We hope that these predictions will be also helpful to pharmaceutical companies to manufacture drugs in faster rate

REFERENCES

[1]R.S. Walse, G.D. Kurundkar, P. U. Bhalchandra, "A Review: Design and Development of Novel Techniques for Clustering and Classification of Data," International Journal of Scientific Research in Computer Science and Engineering, Vol.6, Issue.1, pp.19-22,2018

[2] Hemant Kumar Soni, "Machine Learning â€“ A New Paradigm of AI," International Journal of Scientific Research in Network Security and Communication, Vol.7, Issue.3, pp.31-32,2019

[3] Amogha A.K., "Load Forecasting Algorithms with Simulation & Coding," International Journal of Scientific Research in Network Security and Communication, Vol.7, Issue.2, pp.15-20, 2019

[4]K. Krishna Rani Samal, Korra Sathya Babu, Santosh KumarDas, Abhirup Acharaya, ”Time Series based Air Pollution Forecasting using SARIMA and Prophet Model” , In the Proceedings of ITCC 2019: International Conference on Information Technology and Computer Communications, pp 80-85,2019.

[5] Upendra Kumar Tiwari & Rizwan Khan, ” Role of Machine Learning to Predict the Outbreak of Covid-19 in India”, Journal of Xi'an University of Architecture & Technology, Vol.12, Issue.4, pp. 2663-2669,2020.

[6] Herlawati ,“COVID-19 Spread Pattern Using Support Vector Regression”, PIKSEL : Penelitian Ilmu Komputer Sistem Embedded and Logic Journal , Vol.8, Issue.1, pp. 67-74,2020

[7] Dutta,Shawni , Samir Kumar Bandyopadhyay ,Tai-Hoon kim , ”CNN-LSTM Model for Verifying Predictions of Covid-19 Cases”, Asian Journal of Computer Science and Information Technology, Vol.5, Issue.4, pp. 25-32,2020

[8] Rustam, Furqan & Reshi, Aijaz & Mehmood, Arif & Ullah, Dr. Saleem & On, Byungwon & Aslam, Waqar & Choi, Gyu Sang, “COVID-19 Future Forecasting Using Supervised Machine Learning Models”, in IEEE Access, Vol. 8, pp. 101489-101499,2020

[9]R. Ranjan, “Predictions for COVID-19 outbreak in India using Epidemiological Models”medRx 10.1101/2020.04.02.20051466,2020

[10] Simon James Fong, Gloria Li, Nilanjan Dey, Rubén González Crespo, Enrique Herrera-Viedma, “ Finding an Accurate Early Forecasting Model from Small Dataset: A Case of 2019-nCoV Novel Coronavirus Outbreak”, International Journal of Interactive Multimedia and Artificial Intelligence,Vol.6,Issue.1, pp. 132-140 , 2020

[11] Petropoulos F, Makridakis S, “Forecasting the novel coronavirus COVID-19” , PLOS ONE journal , March 31, 2020. https://doi.org/10.1371/journal.pone.0231236 ,2020

[12] Zheng N,Du S, Wang J, Zhang H, Cui W, Kang Z, et al. , “Predicting COVID-19 in China Using Hybrid AI Model”, IEEE Trans Cybern, https://doi.org/10.1109/TCYB.2020.2990162, 2020

[13] Heni Bouhamed, "Covid-19 Cases and Recovery Previsions with Deep Learning Nested Sequence Prediction Models with Long Short-Term Memory (LSTM) Architecture," International Journal of Scientific Research in Computer Science and Engineering, Vol.8, Issue.2, pp.10-15,2020