International Journal of Innovative Research in Engineering and Management (IJIREM)
ISSN (Online): 2350-0557, Volume-11, Issue-1, February 2024
https://doi.org/10.55524/ijirem.2024.11.1.4
Article ID IJIR2496, Pages 15-19
www.ijirem.org

An Analysis Comparing the Predictions of Predictive Models with the Actual Number of New Cases during the COVID-19 Pandemic

Santosh C J[1], and Anurag Shakya[2]

[1]Institute of Business Management and Commerce, Mangalayatan University, Aligarh, India
[2]Institute of Business Management and Commerce, Mangalayatan University,Aligarh, India

Correspondence should be addressed to Santosh C J; 20201006_santosh@mangalayatan.edu.in

Received: 5 January 2024 Revised: 20 January 2024 Accepted: 30 January 2024

ABSTRACT- On January 3, 2020, Chinese health officials discovered a pneumonia outbreak in the Chinese city of Wuhan. The virus swiftly spread to most countries, infecting a substantial portion of the population. On September 28, 2020, about a million deaths were reported worldwide. The virus is typically transmitted through coughing or sneezing on others. Because there is no effective vaccine, the majority of governments have enforced lockdowns to slow the virus's spread. Other preventive measures, such as travel bans, social isolation, hand hygiene, and the use of face masks, were critical in restricting the virus's spread. However, it proved difficult to contain the virus, which had spread to over 200 nations with a population of over 7 billion people. Large data sets were acquired on a daily basis, and data analytics became critical for uncovering trends and establishing how the infection spread. Several studies have been conducted to develop a mathematical forecast for the pandemic since the disease's first cases in India. First examined with reference to India, these models differed significantly in their scope, underlying assumptions, and numerical forecasts. The objective of this research is to evaluate the predictive models' efficacy by comparing their forecasts with the actual number of new cases reported every day in India during the COVID-19 outbreak.

KEYWORDS- Predictive Models, COVID-19, Confirmed Cases, Python, AutoTS, Automatic Time Series Forecasting

I. INTRODUCTION

In January 2020, a contagious viral illness was first identified in Wuhan, China. Owing to its extremely high transmission rate, the COVID-19 infection rapidly expanded to every nation on Earth [1]. Lockdowns and travel restrictions were swiftly implemented by several nations. Furthermore, hand cleanliness, social distancing, isolation, and face masks became standard procedures. There have been reports of deaths, an unusually high hospitalization rate, and global unrest. Hand sanitizers, face masks, and hospital beds were said to be in short supply. Finding the most effective means to distribute resources around the world is currently one of the largest issues. Data analytics have to be used to overcome these challenges. Authorities used analytical techniques to forecast patterns and possible trends, which allowed them to identify high-risk areas and implement control measures.

Models for predicting daily new cases and deaths were developed using machine learning algorithms [2]. Numerous scholars have attempted to model this growing pandemic mathematically since the first instances occurred in India. Large discrepancies in the models' scope, assumptions, and numerical predictions, as well as the way the pandemic unfolded in India and the impact of different initiatives and health care services, were found during an initial review of these models [3]. A machine learning model is a program that analyzes previously recorded datasets to identify patterns or make conclusions. By using a sizable dataset to train, a machine learning model can easily carry out these tasks. The machine learning algorithm is tuned during training to identify specific patterns or outputs from the dataset, contingent on the task [4].

The goal of this study is to compare the number of new cases reported during the COVID-19 pandemic in India with predictions made by predictive algorithms. The findings of this study will allow authorities to put their trust in predictive algorithm forecasts in the future.

II. BACKGROUND

The fundamental components of machine learning models are machine learning algorithms [5]. Data that is labeled, unlabeled, or mixed is used to train these algorithms [6]. A machine learning algorithm is a mathematical method for finding patterns in a collection of data. The process of optimizing a machine learning algorithm to find particular patterns or outputs by applying it to a dataset referred to as training data is called model training. Once trained, the model can make predictions and reason over data that hasn't been seen before [7].

Supervised learning, unsupervised learning, and reinforcement learning are the three categories into which machine learning approaches fall [8]. An algorithm is given an input dataset in supervised machine learning, after which it is rewarded or optimized to achieve a

predetermined set of outputs [9]. Unsupervised machine learning involves giving an algorithm an input dataset, training it to categorize objects based on shared properties rather than rewarding or optimizing it for certain outputs. Through a series of trial-and-error exercises, the algorithm is trained to become self-sufficient in reinforcement learning [10]. When an algorithm continuously engages with its surroundings instead of depending solely on training data, reinforcement learning takes place [11].

The process of manually creating a machine learning model requires a number of phases and requires knowledge to non-ML specialists, increase ML productivity, and hasten ML research [13].

The field of automated machine learning, or AutoML, has been expanding recently and offers customers the ability to fully automate the construction of high-performance pipelines for classification or regression. In a variety of applications, autoML has been shown to automatically generate competitive and high-quality models, frequently surpassing manually tweaked models [14].

Some of the automated machine learning models include AutoWEKA, Auto-sklearn and Auto-PyTorch. Many automated machine learning algorithms were employed to forecast future cases during the COVID-19 pandemic. While some models and their results were incredibly precise, others weren't. To determine how well these models predict and to make sure they are trustworthy, it is vital to examine their predictions and compare them with COVID-19 real data.

III. METHODS

For analytics, we used data from Our World in Data, which can be found at the URL: hppts://ourworldindata.org/covidcases [15]. The information on daily confirmed cases, new fatalities, freshly provided vaccination, newly done COVID-19 tests, positive rate, reproduction rate, and stringency index was downloaded in CSV form. The data is compiled using Microsoft excel. We have considered using Auto TS library from python for this research.

of the relevant domain, mathematical skills, and computer science knowledge [12]. Automated machine learning helps organizations maximize the return on investment from data science while reducing time to value by utilizing the deeply entrenched expertise of data scientists without requiring them to devote time or money in building these skills themselves. Automated machine learning research includes a broad spectrum of tools and methods designed for both end users and researchers. Automated machine learning (ML) offers tools and platforms that make ML accessible

Automatic Time Series or Auto TS forecasting is a Python-based automated machine learning library that was created to automate time series forecasting. Time series forecasting is the process of predicting future values of a measurable variable based on the current and historical time series [16].

A time series is characterized by several mathematical models, each of which uses a particular set of parameters. Some of the models are ARIMA, GLS, UnivariateMotif, Theta, UnivariateRegression, FBProphet etc. We have considered ARIMA and PBProphet models for predicting new cases. The model is set to predict daily cases from 1st Jan 2022 till 31st Jan 2022. Post forecasting, we will compare the predictive model's forecast with the actuals recorded during COVID-19 pandemic in the month of Jan 2022 in India. From the results we can determine the efficacy of predictive models forecasting ability.

IV. THEORY

A. Actuals

Figure [1] shows the actuals recorded during COVID-19 in India in the month of January 2022. Figure [2] shows the visualization of the data set. Though the data set has information on different parameters, we have considered a few variables which are required for this research. This includes location, date, new cases, new deaths, reproduction rate, new tests, rositive rate, and new vaccinations.

Location	Date	New cases	New deaths	Reproduction rate	New tests	Positive rate	New vaccinations
India	01-01-2022	22775	406	2.24	1110855	0.0129	5204380
India	02-01-2022	27553	284	2.37	1082376	0.016	2516432
India	03-01-2022	33750	123	2.49	878990	0.02	7939936
India	04-01-2022	37379	124	2.6	1478119	0.0251	7804099
India	05-01-2022	58097	534	2.67	1388647	0.0335	9873694
India	06-01-2022	90928	325	2.67	1488509	0.0437	9787059
India	07-01-2022	117100	302	2.61	1513377	0.0567	13600889
India	08-01-2022	141986	285	2.5	1651831	0.0674	2415504
India	09-01-2022	159632	327	2.37	1563566	0.0788	9697867
India	10-01-2022	179723	146	2.24	1540827	0.0862	6517469
India	11-01-2022	168063	277	2.12	1579928	0.0981	13619197
India	12-01-2022	194720	442	2.01	2119100	0.1055	6455760
India	13-01-2022	247417	380	1.9		0.1141	9622002
India	14-01-2022	264202	315	1.8		0.1207	2558832
India	15-01-2022	268833	402	1.7	1613740	0.1302	9160005
India	16-01-2022	271202	314	1.61	1736014	0.1347	5057459
India	17-01-2022	258089	385	1.54	1313444	0.143	5167283
India	18-01-2022	238018	310	1.48	1649143	0.1494	143011
India	19-01-2022	282970	441	1.42	2010225	0.1565	16445190
India	20-01-2022	317532	491	1.37	1935180	0.163	7882289
India	21-01-2022	347254	703	1.32	2182108	0.165	9279228
India	22-01-2022	337704	488	1.26	1960954	0.1654	7475099
India	23-01-2022	333533	525	1.19	2020688	0.1655	5122818
India	24-01-2022	306064	439	1.13	1474753	0.1648	1208921
India	25-01-2022	255874	614	1.07	1807100	0.1631	6892485
India	26-01-2022	285914	665	1	1769745	0.1637	5871190
India	27-01-2022	286384	573	0.94	1594070	0.1606	2929184
India	28-01-2022	251209	627	0.9	1582307	0.1601	6181288
India	29-01-2022	235532	871	0.86	2026150	0.1511	5167292
India	30-01-2022	234281	893	0.82	1615993	0.1482	6900866
India	31-01-2022	209918	959	0.77	1607115	0.1392	7423888

Figure 1: New cases - India - January 2022

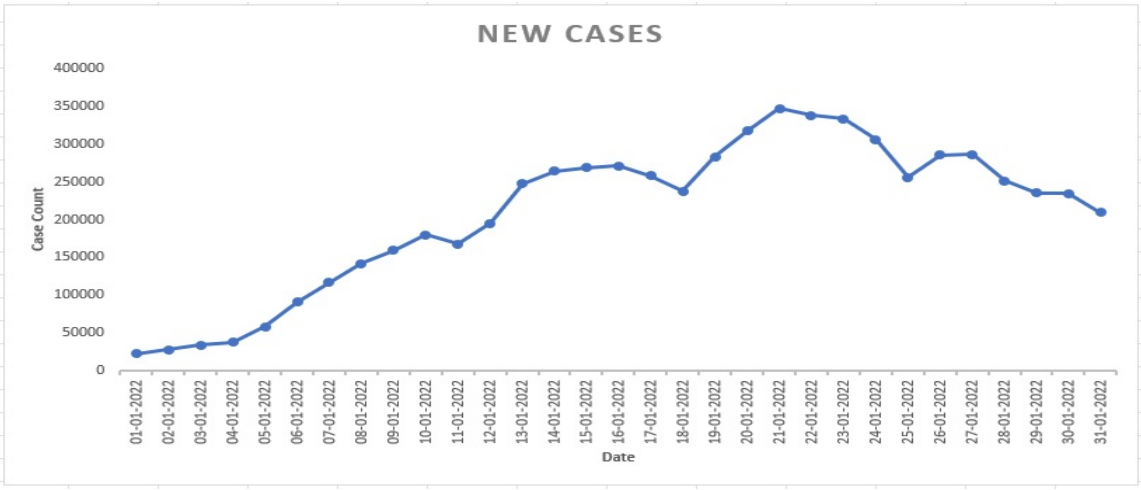

Figure 2: COVID-19 Cases – India

B. Model Prediction

After installing Anaconda, launch Jupyter Notebook, begin by importing the necessary Python libraries and dataset for the task.

```
#Import required packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from autots import AutoTS
import warnings
warnings.filterwarnings('ignore')

#Load CSV file
data = pd.read_csv('../documents/owid-covid-data.csv')
```

For this research, we have considered COVID-19 data recorded in India. A filter is added to set location as India. For training we have consider the data set with date range between 1st Jan 2021 and 31st Dec 2021.

In the next step, we will install the AutoTS library and prepare the data set.

```
#Add filters
data = data.loc[data['location'] ==  'India']

data = data.loc[(data['date'] >= '2021-01-01') &
(data['date'] < '2022-01-01')]
\end{python}

The data set is then visualized using python's
Matplotlib  library.

\begin{python}
#Visualize data set required for this research
fields = data[['date', 'new_cases']]
fields.plot()
plt.show()
```

We will define the model and pass the parameters in this phase. Some of the parameters are discussed below.

```
#Install AutoTS
!pip install autots

#Prepare data set
train = data[['date', 'new_cases']]
train['new_cases'].iloc[0] = train['new_cases'].iloc[1]

#Convert date format
train['date'] = pd.to_datetime(train['date'])

#Check for missing values
train.isna().sum()

#preprocessing data
from sklearn.preprocessing import StandardScaler
scaler_newcases = StandardScaler()
scaler_newdeaths = StandardScaler()
train['new_cases']                                    =
scaler_newcases.fit_transform(train[['new_cases']])
```

- Forecast_length - Desired length of the prediction period.
- Frequency - The training data's frequency (day, month, year, hour, minute, seconds)
- Prediction interval - confidence intervals concept
- Ensemble - It is the process of combining two or more models that have been trained using shared data to produce the desired outcome.
- model_list - With the help of model_list, we can choose from a variety of alternatives to choose a pool of models to train on and can quickly identify which option is the best option among the others.

```
model_list = ['ARIMA', 'FBProphet']
model = ['ARIMA', 'FBProphet']

#Create the Model
mod_newcases = AutoTS(
    forecast_length=31,
    frequency='infer',
    prediction_interval=0.9,
    ensemble='all',
    model_list=model,
    max_generations=5,
    num_validations=2,
    validation_method='even',
    n_jobs = -1
)
```

This stage involves putting our data through several models and determining which one fits our data the best. The best model details will be published in print.

```
#Fit The Model
mod_newcases = mod_newcases.fit(train, date_col='date',
                value_col='new_cases', id_col=None)
print(mod_newcases)
```

Using the Matplotlib toolkit, forecasting and forecast visualization are created in this step

```
forecast_newcases =
scaler_newcases.inverse_transform(forecast_newcases)
plt.figure(figsize=(12,5))
plt.plot(forecast_newcases,color='black', linestyle='dashed',
linewidth = 1, marker='o', markerfacecolor='skyblue',
markersize=10);
plt.xlabel('Date')
plt.ylabel('Number of cases')
plt.title('India, COVID-19, New Cases')
plt.show()
```

V. RESULTS

The total number of recorded cases in the month of Jan 2022 is 64,63,636 cases and the predicted total number of cases is 65,86,338, with a residue of 1,22,702 cases.
Figure [3] shows the charts model prediction results and the actuals recorded in India. From the charts we can conclude both the actuals and the prediction model shows a gradual increase in number of cases at the beginning of the month and the peak values are recorded a little before the end of the month.

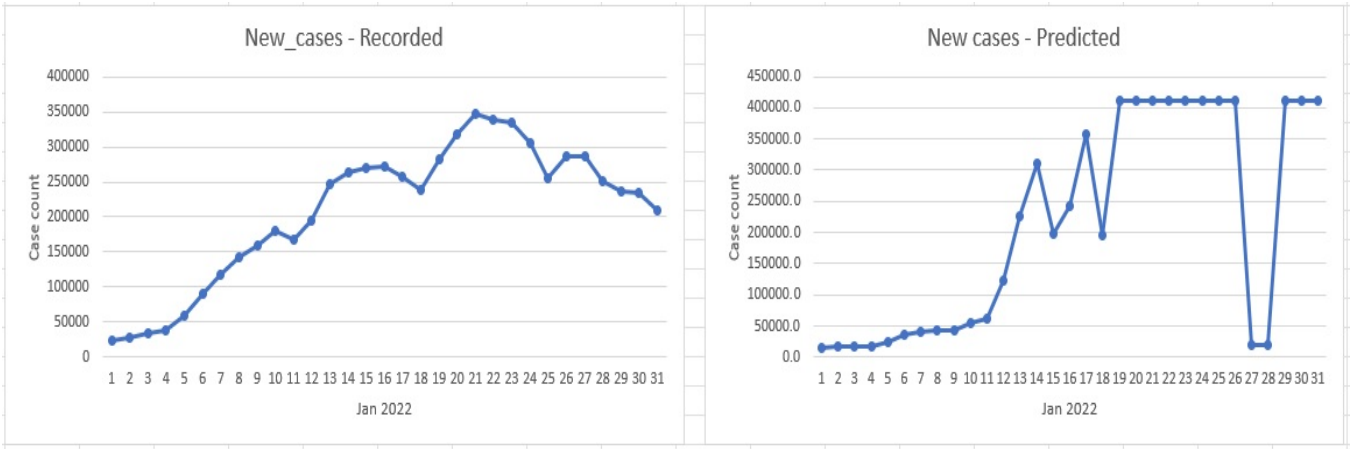

Figure 3: New cases recorded and Model forecasting

Figure [4] shows the residuals. The part of the validation data that the model is unable to explain is represented as residuals. The residual for each observation is the difference between predicted values and observed values. The sum of residuals is 1,22,702 which is 1.86\% increase from the sum of actual COVID-19 cases recorded in India in the month of January 2022. From the figure it is clearly evident the model prediction is reliable and can be improved by enhancing the training process by feeding different sets of COVID-19 data.

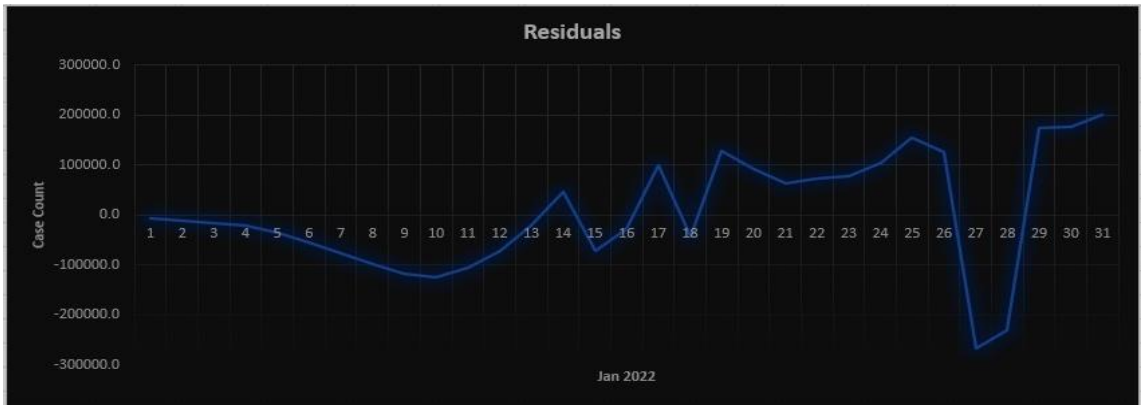

Figure 4: Residuals

VI. DISCUSSION

Using multiple prediction models results in significant variations in the predictions. Particularly when employing a time series forecasting technique that is automatic. The variance can be attributed to the predictive capacity of the mathematical models, their construction methods, or the parameters that these models employ to forecast future occurrences.

VII. CONCLUSION

We can infer from this study that automatic time series forecasting models are trustworthy and can be used to anticipate future events. A few dependencies exist, including those on environment variables, configurations, libraries, frameworks, and data sources. These dependencies may reduce the predictive models' efficacy. We demand the employment of automatic time series prediction models to predict unexpected disasters before they occur, and we propose the creation of a uniform procedure for data collecting during pandemics.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] Dhama, K., Khan, S., Tiwari, R., Sircar, S., Bhat, S., Malik, Y.S., Singh, K.P., Chaicumpa, W., Bonilla-Aldana, D.K. and Rodriguez-Morales, A.J., 2020. Coronavirus disease 2019–COVID-19. Clinical microbiology reviews, 33(4), pp.10-1128.

[2] Peng, Y., Li, C., Rong, Y., Pang, C.P., Chen, X. and Chen, H., 2021. Real-time prediction of the daily incidence of COVID-19 in 215 countries and territories using machine learning: model development and validation. Journal of Medical Internet Research, 23(6), p.e24285

[3] Kotwal, A., Yadav, A.K., Yadav, J., Kotwal, J. and Khune, S., 2020. Predictive models of COVID-19 in India: a rapid review. Medical journal armed forces India, 76(4), pp.377-386.

[4] Mahesh, B., 2020. Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9(1), pp.381-386.

[5] Bonaccorso, G., 2017. Machine learning algorithms. Packt Publishing Ltd.

[6] Chawla, N.V. and Karakoulas, G., 2005. Learning from labeled and unlabeled data: An empirical study across techniques and domains. Journal of Artificial Intelligence Research, 23, pp.331-366.

[7] Yang, L. and Shami, A., 2020. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415, pp.295-316.

[8] Muhammad, I. and Yan, Z., 2015. SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. ICTACT Journal on Soft Computing, 5(3).

[9] Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160(1), pp.3-24.

[10] Hahne, F., Huber, W., Gentleman, R., Falcon, S., Gentleman, R. and Carey, V.J., 2008. Unsupervised machine learning. Bioconductor case studies, pp.137-157.

[11] Barto, A.G., 1997. Reinforcement learning. In Neural systems for control (pp. 7-30). Academic Press.

[12] Sarker, I.H., 2021. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci 2: 160.

[13] CUROTTI, T., 2019. Automated machine learning: competence development, market analysis and tools evaluation on a business case.

[14] Tetteroo, J., Baratchi, M. and Hoos, H.H., 2022. Automated machine learning for COVID-19 forecasting. IEEE Access, 10, pp.94718-94737.

[15] Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E. and Roser, M., 2020. Coronavirus pandemic (COVID-19). Our world in data.

[16] Peixeiro, M., 2022. Time series forecasting in python. Simon and Schuster.