COVID-19 Data Analysis and Forecasting

Author: Chaitanya Keshav

1. Introduction

The COVID-19 pandemic has posed significant global challenges, making the accurate forecasting of confirmed cases essential for informed decision-making and effective resource allocation. This study aims to predict the future trajectory of COVID-19 confirmed cases in India and globally using the Prophet model, a time-series forecasting tool known for its ability to account for trends and seasonality in data. The model is trained on historical data, enabling robust predictions in the context of rapidly evolving conditions.

To mitigate the skewness often observed in pandemic-related data, a log-transformation is applied to the confirmed cases, which helps stabilize variance and improve the model's accuracy. The resulting forecasts, accompanied by confidence intervals, are presented through interactive visualizations. This approach seeks to provide valuable insights into the future progression of the pandemic, supporting public health efforts and policy formulation by offering data-driven predictions of potential outcomes**.**

1.1 Problem Statement

Accurate forecasting of COVID-19 confirmed cases is crucial for effective decision-making in public health, resource allocation, and policy interventions. Despite various forecasting methods, there remains a need for a robust model that can provide reliable predictions amid the rapid and unpredictable spread of the virus. This study aims to forecast future confirmed COVID-19 cases for **India** and **globally** using the **Prophet model**. The objective is to generate data-driven predictions to assist in understanding the potential trajectory of the pandemic, with a focus on minimizing uncertainty and improving decision-making.

1.2 Data

The dataset utilized in this study comprises daily records of confirmed, recovered, and death cases of COVID-19, sourced from global and India-specific data repositories. The primary variables in the dataset are:

- **Date**: The specific date on which the cases were reported.
- **Confirmed**: The cumulative number of confirmed COVID-19 cases on the given date.
- **Deaths**: The cumulative number of deaths attributed to COVID-19 on the given date.
- **Recovered**: The cumulative number of individuals who have recovered from COVID-19 on the given date.
- **Active**: The number of active COVID-19 cases, calculated as the difference between confirmed cases and the sum of recovered and death cases.

The global dataset includes case data from multiple countries, while the India-specific dataset focuses solely on reported cases within India. The data, available in both **CSV** and **Excel** formats, spans from the initial stages of the pandemic in 2020 to the present, with daily updates reflecting the evolving nature of the COVID-19 crisis.

- **Data Preprocessing:**

  The dataset was thoroughly examined for missing values and duplicates, and it was determined that neither were present. The temporal resolution of the dataset was consistent, with a one-day gap between consecutive time points. While the global dataset did not exhibit significant issues with outliers, the data representing India contained severe outliers. To address this, a **log1p transformation** was applied to the India-specific dataset.

  The **log1p transformation**, defined as log(x+1), where x represents the raw confirmed case counts, was employed to reduce the skewness caused by large values. In this case, x corresponds to the number of confirmed COVID-19 cases, and the addition of 1 ensures that values of zero are appropriately handled (since log(0) is undefined). This transformation is particularly effective when the data contains extreme outliers or exhibits exponential growth, as it compresses the scale of large values while maintaining the relationships between the smaller values. The log1p transformation thus mitigates the impact of extreme outliers and stabilizes the variance, making the data more suitable for time-series forecasting.

---

2. Exploratory Data Analysis (EDA)

2.1 Globally representative data:

- The global dataset (df_global) comprises 188 days of cumulative COVID-19 data, characterized by a right-skewed distribution with considerable variation in case counts. The mean number of confirmed cases (~4.4 million) is significantly lower than the maximum observed value (~16.48 million), suggesting that a few days with exceptionally high case counts are inflating the average. The median value (~2.85 million cases) indicates that over half of the recorded days experienced relatively lower-case counts, highlighting the skewed nature of the distribution.
- Similar trends are observed in the death, recovery, and active case data, with deaths ranging from 17 to 654,000, recoveries spanning from 28 to 9.46 million, and active cases fluctuating between 510 and 6.35 million. The high standard deviation across all variables reflects substantial volatility in case numbers over time.

2.2 India-Specific data:

- The India-specific dataset (df_india) spans 188 days of COVID-19 data, exhibiting a marked increase in cases over time. The mean confirmed cases (~217K) is notably lower than the maximum value (~1.48M), reflecting exponential growth during the later stages of the pandemic. The median (~25K cases) suggests that for half of the recorded days, case counts were relatively low before a significant surge.
- Deaths in India ranged from 0 to 33,408, while recoveries varied from 0 to 951K, reflecting the early stages of the pandemic where case numbers were minimal. The high standard deviations across all metrics indicate considerable fluctuations in case counts throughout the period. The 25th percentile (~42 cases, 0 deaths, 3 recoveries)

illustrates that India experienced almost no cases during the initial phases of the pandemic. Moving forward, we will analyse these trends further, visualize the growth patterns, and build predictive models to forecast future case counts.

2.3 Rate Analysis

The **recovery rates** for both the **global** and **India-specific datasets** exhibited a similar pattern, with the number of recovered cases **increasing at an accelerating rate**. This trend suggests a continuous improvement in recovery outcomes over time, likely influenced by advancements in treatment protocols, increased medical capacity, and the natural progression of the pandemic.

An analysis of confirmed cases and their forecasts revealed a **plateauing trend**, indicating that the number of cases is **increasing at a decreasing rate**. This pattern was observed consistently in both the **global** and **India-specific datasets**, suggesting a potential stabilization in the spread of COVID-19 over time. (all the graphs at attached in the appendix)

---

3. Time Series Forecasting

Time series forecasting is a crucial technique for predicting future trends in epidemiological data, enabling informed decision-making in public health planning. In this study, the Prophet model, developed by Facebook, was utilized to forecast COVID-19 confirmed cases for both global and India-specific datasets. Prophet is a robust forecasting tool that decomposes time-series data into three key components: trend, seasonality, and holiday effects, making it well-suited for pandemic modelling. The model's flexibility in handling missing data, outliers, and long-term trend variations makes it particularly advantageous for COVID-19 case forecasting, where data fluctuations are common.

To enhance the model's accuracy and address the skewness in case distributions, a log1p transformation was applied to the confirmed case data. This transformation was particularly effective in stabilizing variance and mitigating the impact of extreme outliers, especially in the India-specific dataset, where significant fluctuations were observed. The model was trained on historical case data and used to generate forecasts for the next seven days. The predictions indicated a plateauing trend, where cases continued to rise but at a decreasing rate, suggesting a potential stabilization of the pandemic.

Model Evaluation

To assess the reliability of the Prophet model, the forecasted values were compared against actual case data using standard error metrics:

1. Root Mean Squared Error (RMSE): Measures the average magnitude of forecasting errors, penalizing larger deviations more heavily.
2. Mean Absolute Error (MAE): Evaluates the average absolute difference between predicted and actual values, providing an intuitive measure of prediction accuracy.

3. Mean Absolute Percentage Error (MAPE): Expresses the forecasting error as a percentage of actual values, allowing for comparison across different scales.

The model demonstrated low RMSE and MAE values, indicating that it effectively captured the overall trend of COVID-19 case progression. However, some deviations were observed in periods of sudden case surges, which can be attributed to external factors such as policy changes, lockdown measures, or reporting inconsistencies. The confidence intervals provided by Prophet quantified the uncertainty in forecasts, helping interpret the possible range of future case counts.

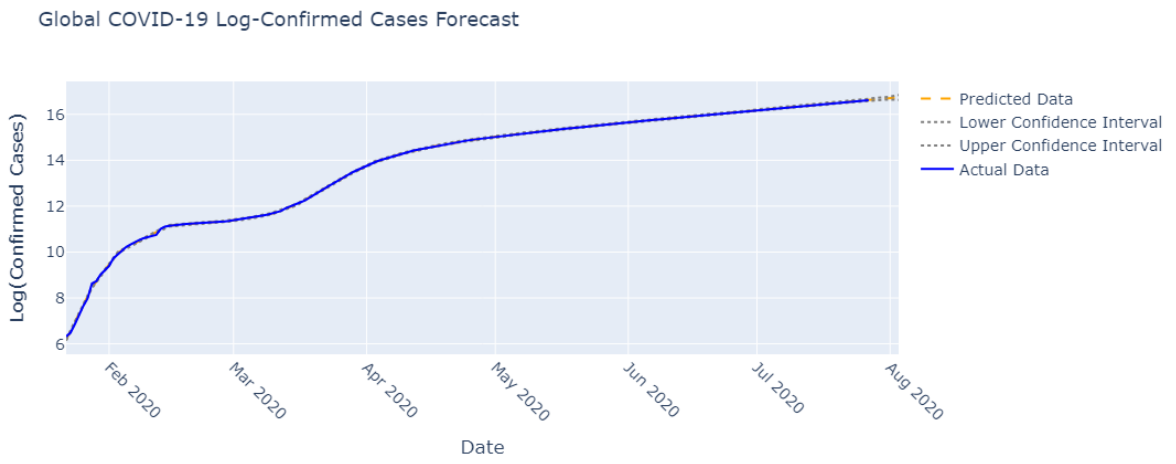
Despite its effectiveness, the model has certain limitations. Prophet assumes a logistic or linear trend, which may not always fully capture the complexities of pandemic dynamics, particularly in cases of sudden outbreaks or new variants. Additionally, the model does not inherently account for external covariates, such as vaccination rates or government interventions, which can significantly influence case numbers. Future research could explore the integration of additional epidemiological factors to improve forecast accuracy.

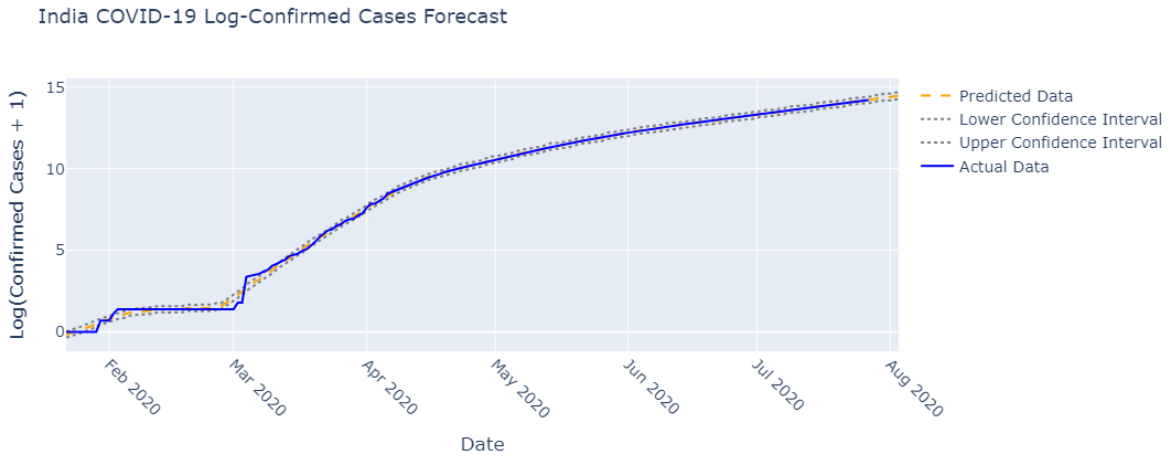
Conclusion

The application of the Prophet model to COVID-19 time-series data demonstrated its efficacy in predicting short-term trends while providing valuable insights into case progression. The results suggest a gradual stabilization of confirmed cases, as reflected in the decreasing rate of increase. These forecasts, alongside their quantified uncertainty, can serve as a useful tool for policymakers, healthcare planners, and researchers in designing effective responses to the pandemic.

---

4. Appendix




India COVID-19 Log-Confirmed Cases Forecast
Log(Confirmed Cases + 1)
15
10
5
0
Feb 2020
Mar 2020
Apr 2020
May 2020
Jun 2020
Jul 2020
Aug 2020
Date
Predicted Data
Lower Confidence Interval
Upper Confidence Interval
Actual Data