COVID-19: Analyzing and Predicting Spread and Resource Needs using Data-Driven Approaches

Bhavini Korthi
20110037

Kareena Beniwal
20110095

Karthikeya Narla
20110117

Zeeshan Snehil Bhagat
20110242

—Abstract—

The COVID-19 pandemic has impacted the world in an unprecedented way. In an effort to understand and mitigate its effects, data analysts across the world have been engaging with pandemic data. The pandemic has highlighted the importance of data analysis and prediction in identifying potential hotspots and resource needs. The COVID-19 pandemic has generated a vast amount of data from various sources, yet many crucial questions about its spread, potential hotspots and impact remain unanswered. This report aims to explore the various sources of COVID-19 data, including Kaggle, CoViD19-India, WHO, AnalyticsIndiaMag, the National Institute of Health, and Johns Hopkins University. The report will examine some possible questions that data analysts can ask about the COVID-19 data, such as identifying specific anomalies in the reported data, finding correlations between case growth and other health indicators at the district or block level, predicting the growth of the epidemic, and identifying areas that should be boosting their hospital beds soon.

In this project, we aim to analyze COVID-19 data available from various sources and apply data-driven approaches to identify patterns, correlations, and potential factors that influence the spread and resource needs.

Keywords and phrases COVID-19, spread, patterns, health indicators

Acknowledgements We want to thank Prof. Anirban Dasgupta, Computer Science and Engineering, IIT Gandhinagar for his help and support throughout the project.

1 Introduction

The COVID-19 pandemic has had a significant impact on the world, with over 150 million confirmed cases and 3 million deaths as of April 2023. The COVID-19 data is available from various sources, including Kaggle, CoViD19-India, WHO, AnalyticsIndiaMag, the National Institute of Health, and Johns Hopkins University.

The availability of such data provides an opportunity for data analysts to ask several questions about the pandemic. For example, they can identify specific anomalies in the reported data, such as discrepancies in the number of cases reported by different sources or inconsistencies in the reporting of deaths. They can also investigate the correlation between case growth and other health indicators at the district or block level, such as the number of hospital beds per capita, the prevalence of comorbidities, or the availability of healthcare resources.

In summary, the COVID-19 pandemic has provided a unique opportunity for data analysts to apply their skills and expertise to an urgent and pressing global issue. By leveraging the available data sources and applying sophisticated analytical methods, they can help to identify patterns, trends, and potential solutions that can inform policy decisions and public health interventions.

Since another outbreak of covid-19 is more likely to happen again in the future, we use the past data to identify the potential factors affecting the spread in certain geographic regions. This also includes working on the demographic data. Addressing these questions is critical to gain a comprehensive understanding of the pandemic and make informed policy decisions to manage and mitigate its effects effectively. The project has the potential to contribute to the understanding of the pandemic and inform policymakers on potential interventions to control the spread of the virus.

Our project focused on several key tasks to analyze COVID-19 data and gain insights into the pandemic.

First, we used the ***SIR (Susceptible-Infected-Removed)*** model to predict the spread of COVID-19 cases in different geographic regions. This model allowed us to understand the dynamics of the pandemic and predict the future trends in cases.

Second, we used ***XGBoost***, a powerful machine learning algorithm, to analyze the relationship between COVID-19 confirmed cases and health indicators such as hospital resources, demographics, and other health indicators. This analysis allowed us to identify potential factors that influence the spread of the virus and inform policy decisions.

Third, we used ***autoregression*** models to study the trends in confirmed cases and predict future trends. These models allowed us to understand the underlying patterns in the data and make informed predictions about the spread of the virus.

Finally, we used ***Z-score*** to identify potential anomalies in the reported data. This analysis allowed us to identify outliers and anomalies that could be indicative of significant changes in the underlying patterns of the data.

Overall, our project demonstrated the potential of data-driven approaches to gain insights into the COVID-19 pandemic and inform policy decisions. By using a range of analytical techniques, we were able to identify potential factors that influence the spread of the virus, predict future trends, and identify anomalies in the reported data.

2 Covid-19 data analysis

The entire code for this project can be found in this repository:
https://github.com/Karthikeyanarla/Data_Science_Project

Problem Statement

Given different datasets, analyze COVID-19 data available from various sources and apply data-driven approaches to identify patterns, correlations, and potential factors that influence the spread and resource needs.

2 Methodology

2.1 Dataset

The datasets are mostly taken from Kaggle, and WHO website. The dataset for different countries consists of characteristics like:

confirmed cases
deaths
recovered people

We took datasets of nearly 100 countries and nearly all of their states/provinces.

It also contains data for India which includes state-wise:

total samples
negative tested people
positive tested people

The health indicators data include:

cardiovascular diseases (%)
Cancers (%)
Diabetes, blood, & endocrine diseases (%)
Respiratory diseases (%)
HIV/AIDS and tuberculosis (%)
Nutritional deficiencies (%)
population growth rate
birth rate
net migration rate

To analyze the dataset, we divided the dataset into two parts: the trained dataset and the testing dataset. Some columns are being dropped while predicting rows corresponding to other columns.

The data set (surrounding the Indian Covid Cases) involved in prediction through autoregressive models and finding outliers through z-score includes features such as:

Date
Time
State/UnionTerritory
Confirmed
Cured
Deaths

Other irrelevant columns have been dropped.

2. Prediction of Covid 19 Cases using Autoregression Models and Finding Outliers

2.1 Dataset Preprocessing

1. **Handling NaN values**

   For the columns with NaN values, we either removed them if they might affect our model or technique or we replaced them with zeroes so that the code runs smoothly.

2. **Handling types after reading from CSV data**

   After reading from the CSV files, we found that the data present were mostly of string data type. So we had to typecast it to the necessary data types, such as date columns to DateTime objects, and the columns containing numbers to numeric types.

3. **Handling columns having similar meaning but one of them is filled at a time**

   For the Indian Covid data set, we found that initially for each State and India overall, the data set kept track of Individuals Vaccinated and after a certain date it kept track of doses administered. Though they are not the same, but they would be close numerically for the initial days when total doses were not tracked (for a given subset of population like male, female, transgender, age groups like 18-44 years, 45-60 years and 60+ years; total doses for the whole population is present in the data for all days). So, we filled the doses administered columns with their corresponding individuals vaccinated columns.

2.2 Prediction of Covid 19 Cases using Autoregression Models

2.2.1 Methodology

We used some autoregression models with a lag of 5 days and tried predicting the last 30, 90 and 150 days confirmed cases trend. To do so, we splitted the data set into two parts: training set and testing set (which corresponds to the data pertaining to the last required number of days that we are trying to predict). We tried to finetune some of the hyperparameters underlying some of the models so that they can give better results for some given parameters. The models used are as follows:

2.2.2 ARIMA

ARIMA is a time series modeling technique used for forecasting. It is a combination of three models: Autoregression (AR), Integration (I), and Moving Average (MA). The AR model uses lagged values of the dependent variable to predict future values, while the MA model uses the errors from previous predictions to predict future values. The I part of ARIMA is used to make the time series stationary by differencing the series. ARIMA models can be used to model both stationary and non-stationary time series data. The model is relatively simple to use and requires few parameters to be estimated. However, it assumes that the underlying data is linear and stationary, which may not be the case in some real-world scenarios. ARIMA models can be extended to incorporate seasonality and exogenous variables, resulting in more powerful forecasting models such as SARIMA and ARIMAX. ARIMA is widely used in time series forecasting because it can capture both short-term and long-term trends in the data.

2.2.3 SARIMAX

SARIMAX (Seasonal Autoregressive Integrated Moving Average with eXogenous factors) is an extension of the ARIMA model that includes seasonal components and exogenous variables. It is a popular time series modeling technique used for forecasting in various industries such as finance, economics, and epidemiology. SARIMAX models can capture both the temporal and seasonal patterns in the data and can be used to forecast future values of the time series. The model includes parameters for the autoregressive, differencing, and moving average components, as well as parameters for the seasonal autoregressive, differencing, and moving average components. Additionally, exogenous variables such as economic indicators or weather data can be included in the model to improve forecast accuracy. SARIMAX models are widely used in real-world forecasting applications due to their flexibility and ability to capture complex patterns in the data.

2.2.4 AutoReg

Autoregression (AR) is a time series modeling technique that uses past values of a variable to predict its future values. In an autoregressive model, the value of the variable at time t is regressed on its lagged values up to time t-1. The order of the AR model (p) determines the number of lagged values used in the model. AR models are widely used in time series analysis and forecasting because they can capture the autocorrelation structure of the data and can be used to forecast future values of the time series. However, AR models assume that the underlying data is stationary, and may not perform well when the data has a trend or seasonality. In such cases, ARIMA or SARIMA models may be more appropriate.

2.2.5 Manual AutoReg with Linear Regression

A manual Autoregressive (AR) model using linear regression involves treating lagged values of the dependent variable as separate predictor variables in a multiple linear regression model. This approach is useful when the order of the AR model is relatively small. The process involves choosing the order of the AR model, creating new columns for each lagged value up to the chosen order, splitting the dataset into training and test sets, fitting a multiple linear regression model to the training set, making predictions on the test set, and evaluating the performance of the model. However, this approach can become impractical for larger orders due to the large number of predictor variables involved. In such cases, specialized time series modeling techniques such as ARIMA or SARIMA may be more efficient.

2.2.6 Results and Conclusion

The results for all the autoregressive models upon the India Covid data set for prediction of confirmed cases are summarized in Table 2. We have used two metrics for comparison, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). We can infer the following from the results:

1. We can see that the AutoReg model performed the worst since it assumes that the data is stationary(mean and variance do not change) and tries to project the just immediate behaviour.

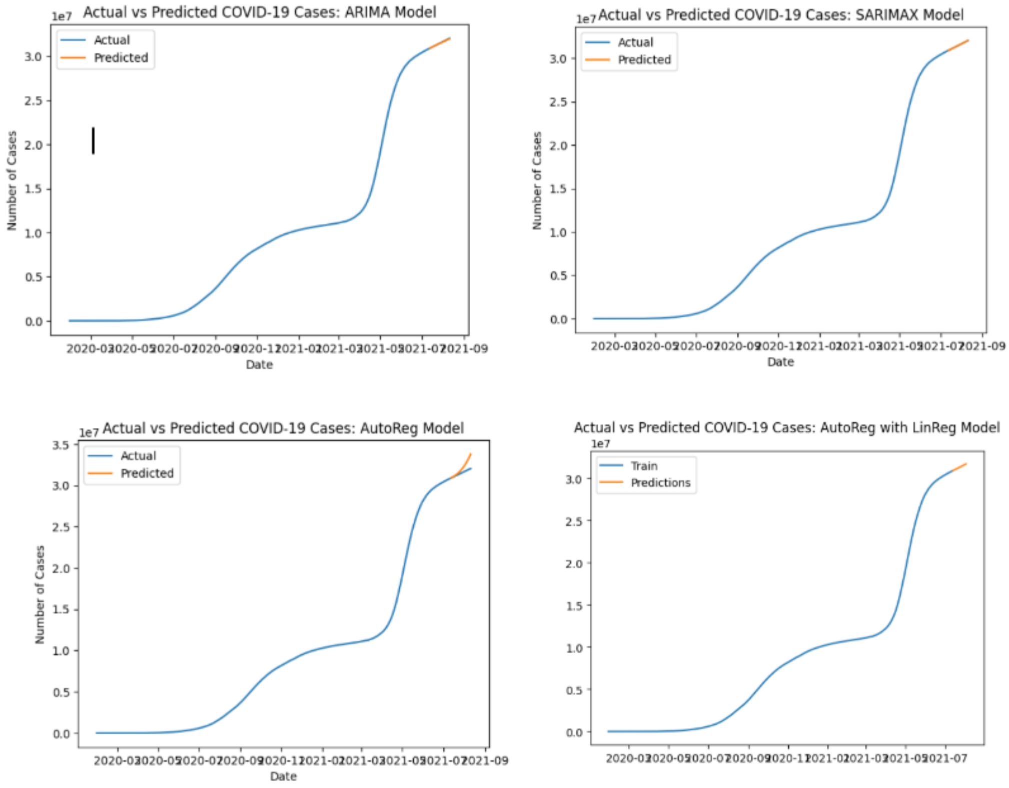
2. The Auto Regression (AR) model with linear regression performed very well. One possible reason for this could be that the AutoReg with Linear Regression model is simpler and more straightforward, while still being able to capture the internal correlation among the structure of the data. It also does not require as much data preprocessing or parameter tuning as more complex models such as SARIMAX. Additionally, the effect of exogenous variables may not be significant enough in this specific case of COVID-19 forecasting, which makes the SARIMAX model less suitable.

2. We observed that the ARIMA and SARIMAX did fairly well for 30 days but degraded later on. This could be possibly due to them not being able to capture the data’s trend and essence due to non optimal hyperparameters. So they can give improved results if we choose the right hyperparameters.

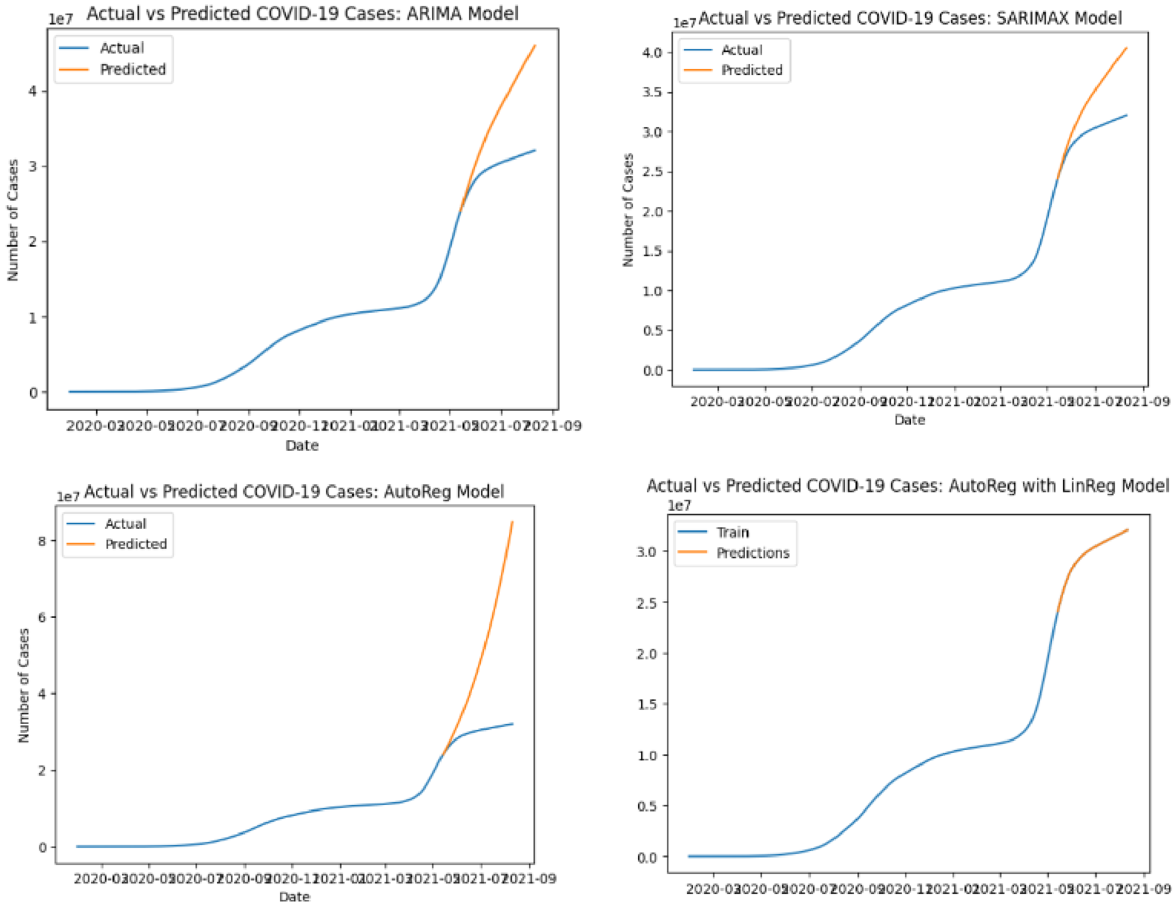
Overall, the SARIMAX and ARIMA model might become better if we finetune it to enable them to capture the underlying trends. However, for now, we can see that the manual AutoReg model with Linear Regression is the most suitable for forecasting the COVID-19 cases, as it is relatively simpler and captures the fine details of the data. Also, the data may have a lot of outliers as the mean and variance of the data fluctuates significantly, so some of the models failed since the real life scenario was a lot more unpredictable than expected.

Model	**No Of Last Days Predicted**	**RMSE**	**MAE**
Auto Regression (Lag =5) with Linear Regression	30	6018.154843093456	4881.723224785427
	90	8085.183857921982	6369.832521559629
	150	34855.12554396177	26868.8014399228
ARIMA (Lag=5)	30	40267.89843598383	29639.961549902582
	90	7972533.359278885	6729461.565703996
	150	5860973.065450959	5149825.310761608
SARIMAX (Lag=5, Seasonal Lag=3 for No of periods=7)	30	4944101.718500594	4212217.159496969
	90	41.390	21.367
	150	7010294.227637742	6076937.509550317
AutoReg (Lag=5)	30	562998.1659926238	427453.9016800161
	90	24848716.295093376	19168409.518106822
	150	13299784.971222378	11254562.003304629

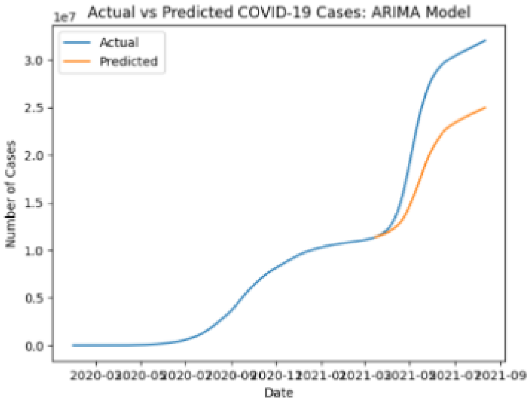
Table 1 Comparison of different models

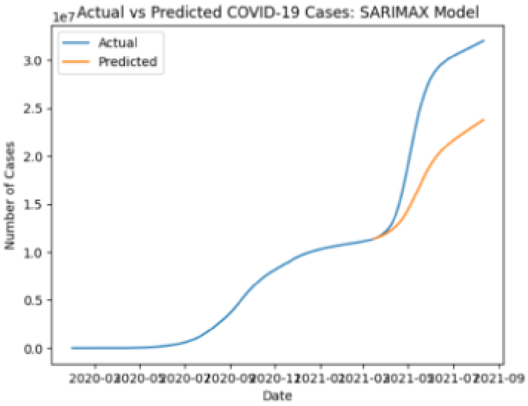


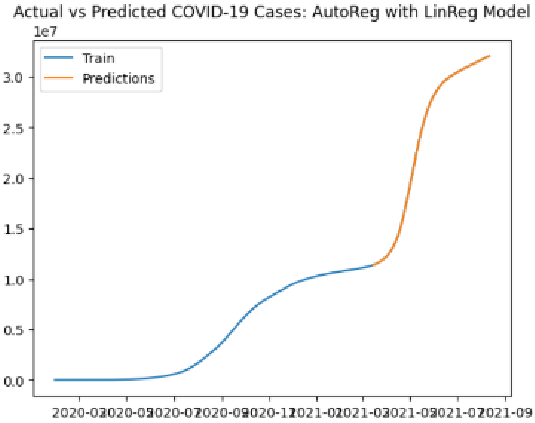
No Of Last Days being predicted = 30

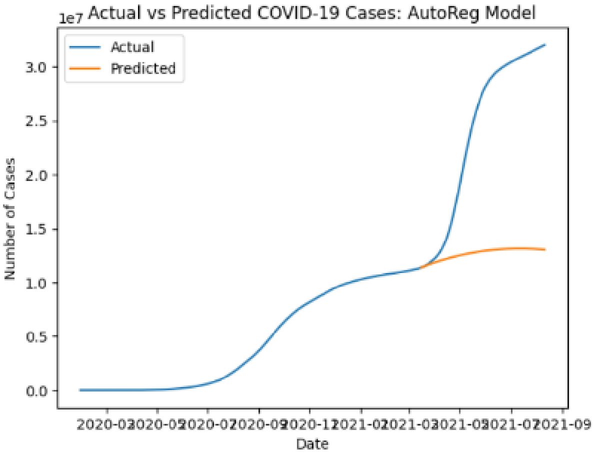


No of Last Days being predicted = 90









No of Last Days being predicted = 150

2.3 Finding Outliers/Anomalies in the Covid Dataset

2.3.1 Methodology

We used some models/techniques to analyze the confirmed cases and the deaths. We tried to finetune some of the hyperparameters underlying some of the models so that they can give better results for some given parameters. The models used are as follows:

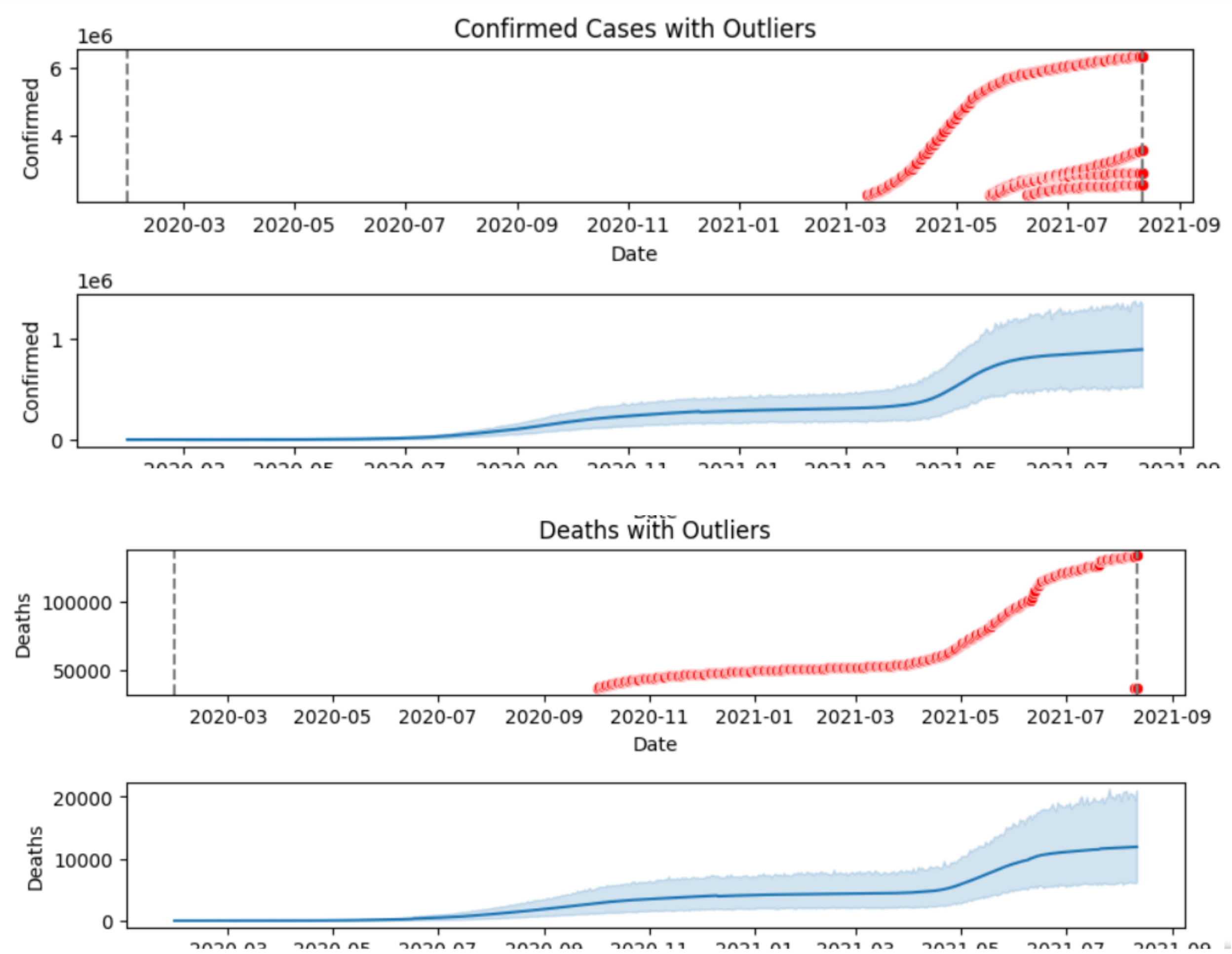
2.3.2 Z-Score

Z-score is a statistical measure that is commonly used to identify outliers in a dataset. It involves calculating the standard deviation of a data point from the mean of the dataset and then normalizing it by dividing the difference by the standard deviation. This results in a standard score or Z-score, which indicates how many standard deviations away from the mean a data point is.

In general, data points with a Z-score greater than 3 or less than -3 are considered outliers, as they are more than three standard deviations away from the mean. However, the threshold for identifying outliers can be adjusted based on the characteristics of the dataset or the specific problem requirements.

Z-score is a simple and widely used method for identifying outliers. However, it may not be suitable for datasets with non-normal distributions or for datasets with a large number of dimensions, as it assumes that the data is normally distributed and independent.

2.3.3 Results and Conclusion



Z-Score Plots with Outliers

In the case of COVID-19 confirmed cases and deaths, it appears that Z-score was able to identify outliers in the latter part of the trend. These outliers could be indicative of significant changes in the underlying patterns of the data, such as sudden spikes or drops in cases or deaths. It could also be that the trend's mean and variance in the latter part of the trend have become such that the data values with respect to them are very large or small, giving rise to outliers, which indicates that the trend is becoming unpredictable.

Z-score is a simple and widely used method for identifying outliers in a dataset, and can be particularly useful for detecting significant changes in the underlying patterns of the data. However, it is important to keep in mind that the method assumes a normal distribution of the data, which may not always be the case in real-world scenarios.

Overall, the outliers identified by Z-score in the latter part of the COVID-19 confirmed and death trends could be useful for further analysis and investigation, and may help inform public health policies and interventions. However, it is important to consider the limitations of the method and to use it in conjunction with other analytical tools to gain a more comprehensive understanding of the data.

3. The SIR Model for Spread of Disease - The Differential Equation Model

3.1 About:

We identify the independent and dependent variables as the first step in the modeling process. The independent variable is time t, measured in days. We are considering two sets of dependent variables related to each other.

The first set of dependent variables counts people in each of the groups, each as a function of time:

S = S(t) is the number of susceptible individuals,
I = I(t) is the number of infected individuals, and
R = R(t) is the number of recovered individuals.

The fraction of the total population is represented by the second set of dependent variables in each of the three categories. So, if N is the total population, we have:

s(t) = S(t)/Nthe susceptible fraction of the population,
i(t) = I(t)/N the infected fraction of the population, and
r(t) = R(t)/N the recovered fraction of the population

The governing differential equations are as follows:
The SIR model is described by a system of three ordinary differential equations that govern the rate of change of each category (S, I, R) over time. The equations are:

$$\frac{dS}{dt} = -\beta SI$$

$$\frac{dI}{dt} = \beta SI - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

where:

S is the number of susceptible individuals
I is the number of infected individuals
R is the number of recovered individuals

β is the rate of transmission (how quickly the disease spreads from infected to susceptible individuals) γ is the recovery rate (how quickly infected individuals recover and become immune to the disease)

- The first equation describes the rate of change of the susceptible population, which decreases as individuals become infected. The term -βSI represents the rate at which susceptible individuals become infected.

- The second equation describes the rate of change of the infected population, which increases as individuals become infected and decreases as they recover. The term βSI represents the rate at which susceptible individuals become infected, while the term γI represents the rate at which infected individuals recover.

- The third equation describes the rate of change of the recovered population, which increases as individuals recover from the disease. The term γI represents the rate at which infected individuals recover and become immune to the disease.

- The SIR model is a simple but effective way to model the spread of infectious diseases in a population, and has been used to inform public health policy and decision-making during outbreaks.

3.2 Assumptions:

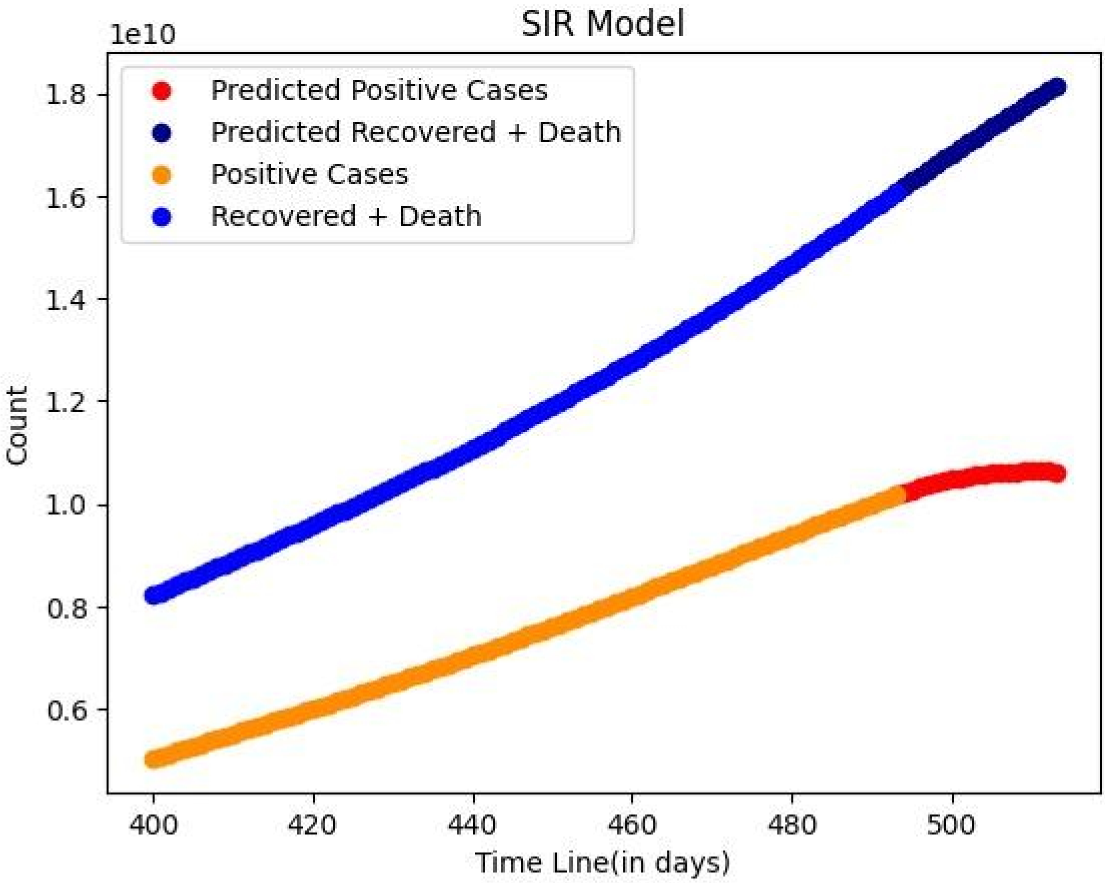
1. The population is constant - There is no migration, birth, or death of individuals in the population during the duration of the epidemic.
2. Homogeneous mixing - Every individual has an equal chance of coming into contact with every other individual in the population, regardless of location or behavior.
3. Infectivity is constant - Every infected individual has the same level of infectiousness throughout the duration of the disease.
4. Recovery is permanent - Individuals who have recovered from the disease are immune and cannot be infected again.
5. No latent period - Individuals become infectious immediately after being infected.
6. No vaccination or intervention measures - The model assumes that there is no intervention, such as vaccination or treatment, that could modify the transmission dynamics of the disease.

3.3 Data Analysis Procedure:

1. Importing necessary libraries.
2. Preprocessing:
    1. Data is cleaned by dropping redundant columns like Province/State, Country/Region, Last Update, etc.
    2. Data is grouped by date and aggregated to sum the number of confirmed cases, deaths, and recoveries .
    3. The cumulative sum of confirmed cases, deaths, and recoveries is calculated.
3. Training and further working:

The SIR model is a compartmental model that divides the population into three compartments: susceptible, infected, and recovered. It assumes that individuals can move between these compartments based on certain rates. We first extract the confirmed, recovered, and death cases from input data and calculate the number of susceptible individuals. Then calculate the transmission rate (beta) and the recovery rate (gamma) from the data. The reproduction number (R0) is calculated as the ratio of beta to gamma.

The next step is to use a Ridge regression model to predict the future transmission and recovery rates. We train two Ridge regression models on the past 10 days of transmission and recovery rates and predict the next 10 days of rates. The predicted rates are then used to predict the future number of susceptible, infected, and recovered cases using the SIR model.



Plot : the actual and predicted number of infected and recovered cases over time

The plot includes the actual data up to the current day and the predicted data for the next 20 days. The predicted data is generated using the predicted transmission and recovery rates. The plot also shows the expected number of confirmed cases, which is the sum of predicted infected and recovered cases.

We can see that deaths are increasing approximately linearly and the confirmed cases are stabilizing. This may be in adherence to the SIR model which assumes homogeneity of the population and its constant count. We can say that as more people got infected then there are less chances of other people getting infected, and the people who got infected earlier are dying.

4. XGBoost Regression:

Decision trees are employed by XGBoost as weak learners, and the method trains a group of trees with each tree attempting to fix the mistakes committed by the one before it. The approach minimises a loss function, which evaluates the difference between the anticipated and actual values, using a gradient descent optimisation technique.

Any arbitrary differentiable loss function and the gradient descent optimisation procedure are used to fit the models. Gradient boosting gets its name from this because as the model is fitted, it minimises the loss gradient, much like a neural network.

XGBoost improves on the conventional gradient boosting technique in a number of ways.

4.1 Assumptions:

1. As there is no direct connection between the Health Indicators like Cancer diseases, Respiratory diseases, and other such indicators, we believe these features will be having the lower feature importance while we train the model using EXTREME Gradient Boosting.
2. Features like First 10_day, First 50_day confirmed cases and other such features should have more importance than health indicators.

4.2 Data Analysis:

1. Preprocessing:

1. Dropping the "Province_State" column

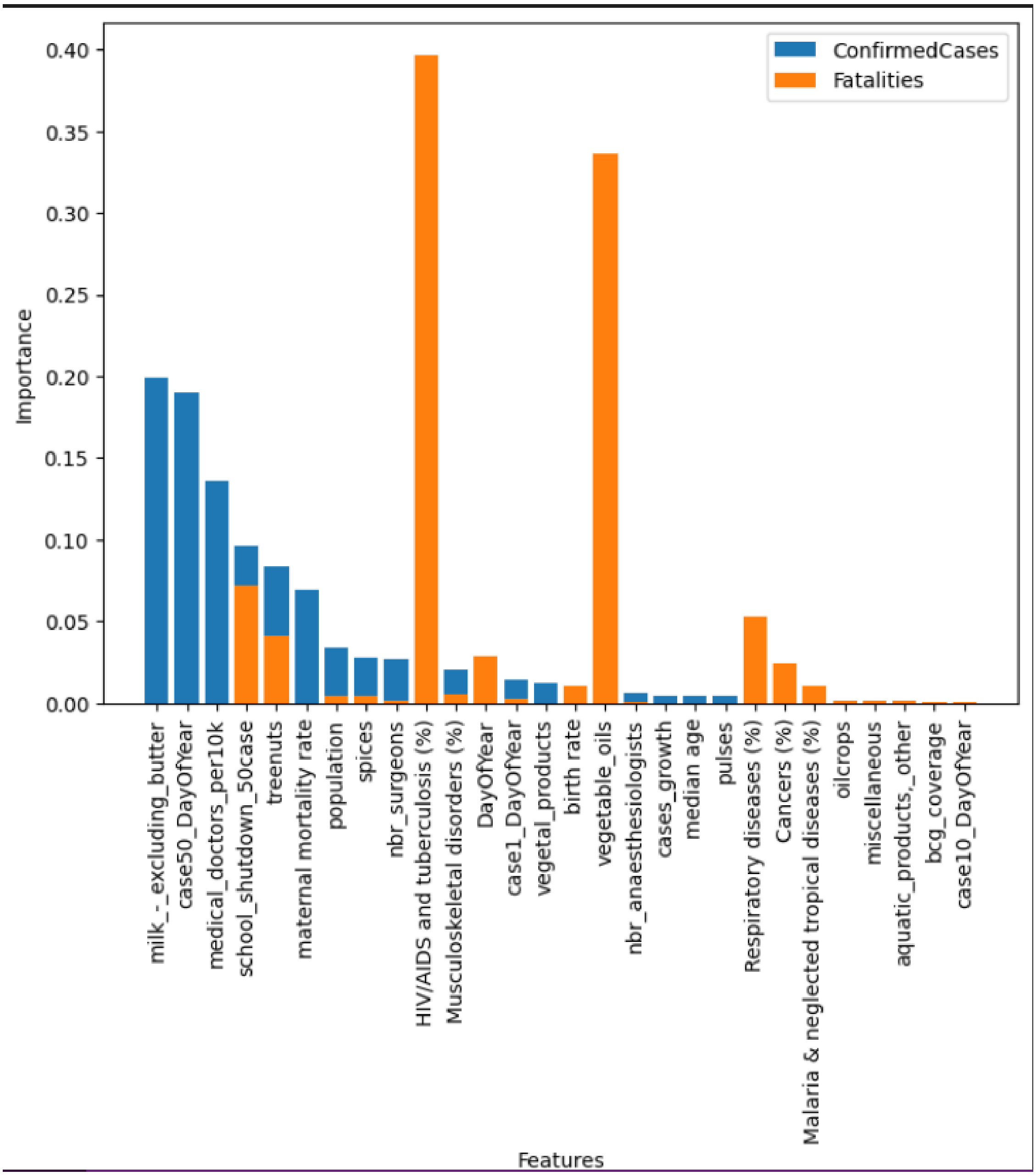
2. If the DataFrame contains a "ConfirmedCases" column, we calculate the cumulative maximum of confirmed cases for each location. This ensures that the "ConfirmedCases" column only contains the maximum number of confirmed cases for each location up to the current date.

Similarly, if the DataFrame contains a "Fatalities" column, we calculate the cumulative maximum of fatalities for each location. This ensures that the "Fatalities" column only contains the maximum number of fatalities for each location up to the current date.

If the DataFrame does not contain a "DayOfYear" column, we create one using the dayofyear attribute of the Pandas datetime object. This column represents the day of the year for each date in the "Date" column.

2. Fit an XGBoost model to the training data and return the root mean squared logarithmic error (RMSLE) and the trained model object.

3. The top 20 most important features for predicting both ConfirmedCases and Fatalities are plotted.



Note: The fatalities are only due to COVID-19, not by other diseases or causes.

The above plot shows the role of features such as tuberculosis, cancers, doctors per 10K of the population, etc in people getting infected by covid and deaths caused by it.

It can be seen that people having HIV/AIDS and tuberculosis, who used vegetable oils, etc. happened to be more fatal from COVID-19 compared to other health indicators such as cancer, respiratory diseases etc. This seems to be unusual and deviated from our hypothesis.

The reason may be an anomaly in the dataset. The dataset may happen to contain a larger amount of samples for these indicators and hence may be biased.

5. Conclusion:

In conclusion, our project aimed to analyze COVID-19 data from various sources and apply data-driven approaches to identify patterns, correlations, and potential factors that influence the spread and resource needs of the pandemic. We used various machine learning techniques such as the SIR model, XGBoost, autoregression model, and Z-score to analyze and predict COVID-19 cases and trends, along with health indicators that affect the spread of the virus.

Our findings can contribute to a better understanding of the pandemic and inform policymakers on potential interventions to control the spread of the virus. We identified potential factors that influence the spread of the virus, such as population density, health indicators, and demographic factors. We also used machine learning techniques to predict the future trends in COVID-19 cases and identify anomalies in the reported data.

The Auto Regression (AR) model with linear regression performed very well. ARIMA and SARIMAX did fairly well for 30 days but degraded later on.

Overall, our project demonstrates the potential of data-driven approaches to inform policy decisions and manage the effects of the COVID-19 pandemic. As the pandemic continues to affect millions of people globally, our work can contribute to a better understanding of the virus and inform interventions to control its spread and mitigate its effects.

—References—

https://github.com/Yu-Group/covid19-severity-prediction
https://github.com/rv20197/COVID-19-Analysis-and-Prediction-Using-AI
https://github.com/AaronWard/covidify
https://github.com/paulvangentcom/python_corona_simulation
https://github.com/tirthajyoti/Covid-19-analysis
https://github.com/CityOfLosAngeles/covid19-indicators
https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge
https://www.kaggle.com/code/nxpnsv/logistic-xgb-hybrid#Prepare-submission