PERFORMANCE EVALUATION OF DIFFERENT MACHINE LEARNING MODELS TO PREDICT COVID-19 CASES FROM TIME SERIES DATA.

Introduction:

This project is about predicting the number of COVID-19 cases based on the data from the previous two years. Our goal of this project is to conduct an extensive analysis on the time series data of the dataset curated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). To achieve this goal of predicting the spread of COVID-19 cases in the near future, we want to apply different machine learning prediction models. We intend to perform the comparative performance analysis of the various applied models.

Problem description:

The goal of this project is to predict the number of confirmed COVID-19 cases based on the time series analysis of the reported confirmed cases.

Methodology:

In the following section we are going to report the data set analysis methods and description of the Machine Learning models used.

Dataset Exploration:

The dataset contains 825 columns. Each column refers to a date and the values indicate the total number of confirmed COVID-19 cases till that date. The first 11 columns consist of different descriptors about the source of the data like the state, county, latitude, longitude and so on. The rest of the columns contain the cumulative number of confirmed cases till that day. The dataset was analyzed state wise for the United States only.

First, we calculated the individual number of confirmed cases for a given day. Figure 1 shows the comparative analysis of different states of the country. We can see that California, Texas and Florida are the most affected states in the US. They constitute 11%, 8% and 7% of the total number cases in the country respectively. Figure 2 shows the histograms for the top 10 states with respect to the number of cases confirmed daily and figure 3 for the lowest 10 states for the same.

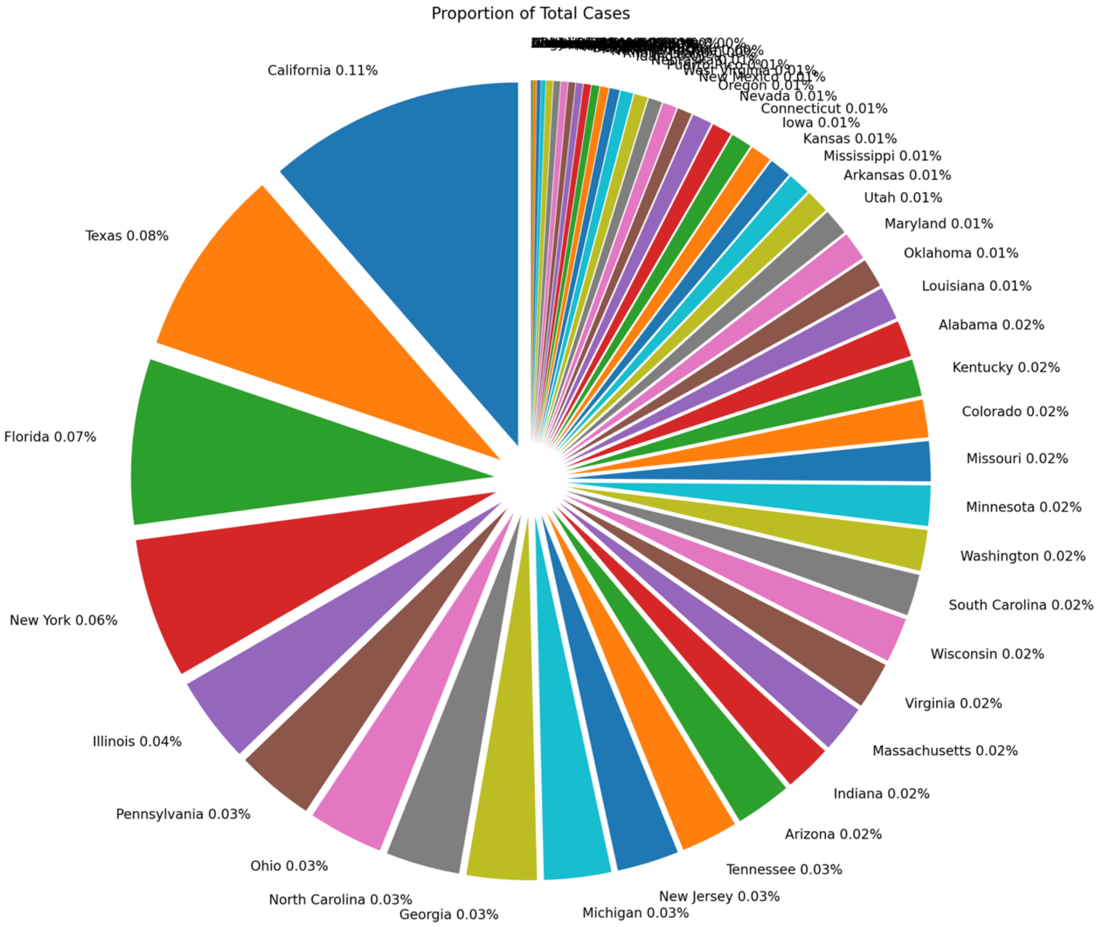

Figure 1: A pie chart showing the proportion of different states’ total COVID-19 cases

We also report the histogram showing the total number of confirmed cases throughout the top 10 and lowest 10 states with respect to confirmed cases in Figure 5 and Figure 6 respectively.

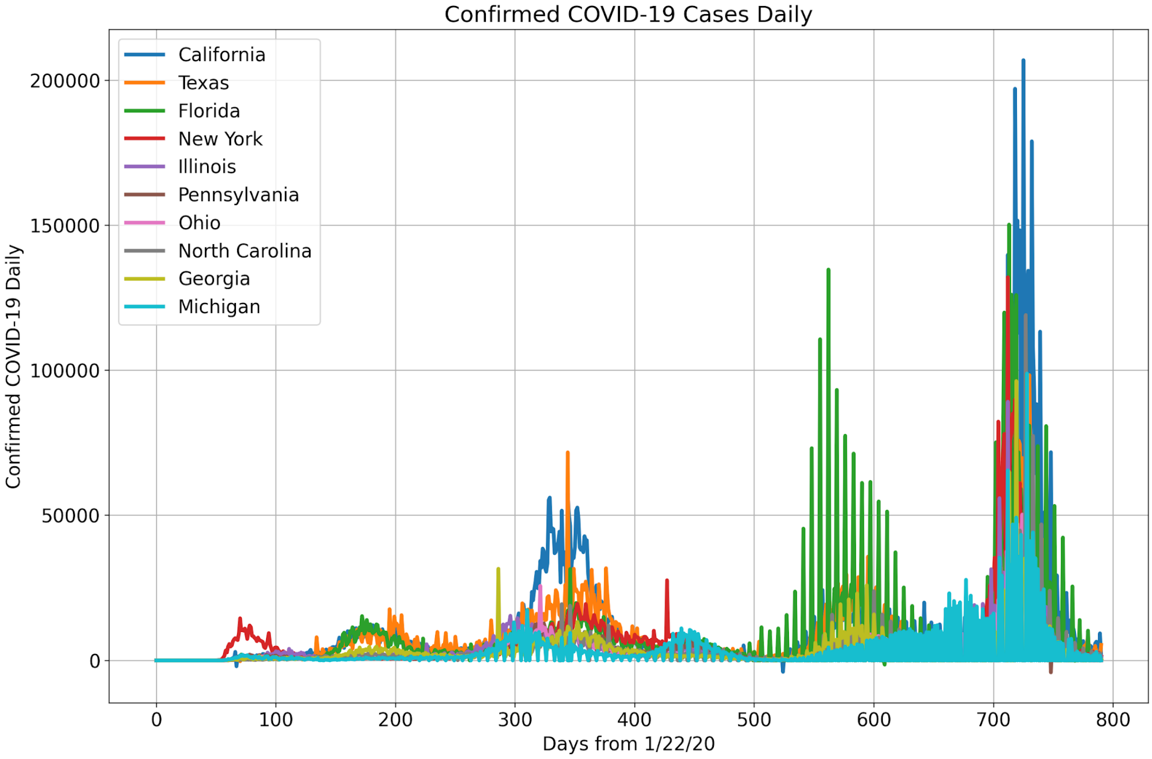

Figure 2: Top-10 States with average daily confirmed cases

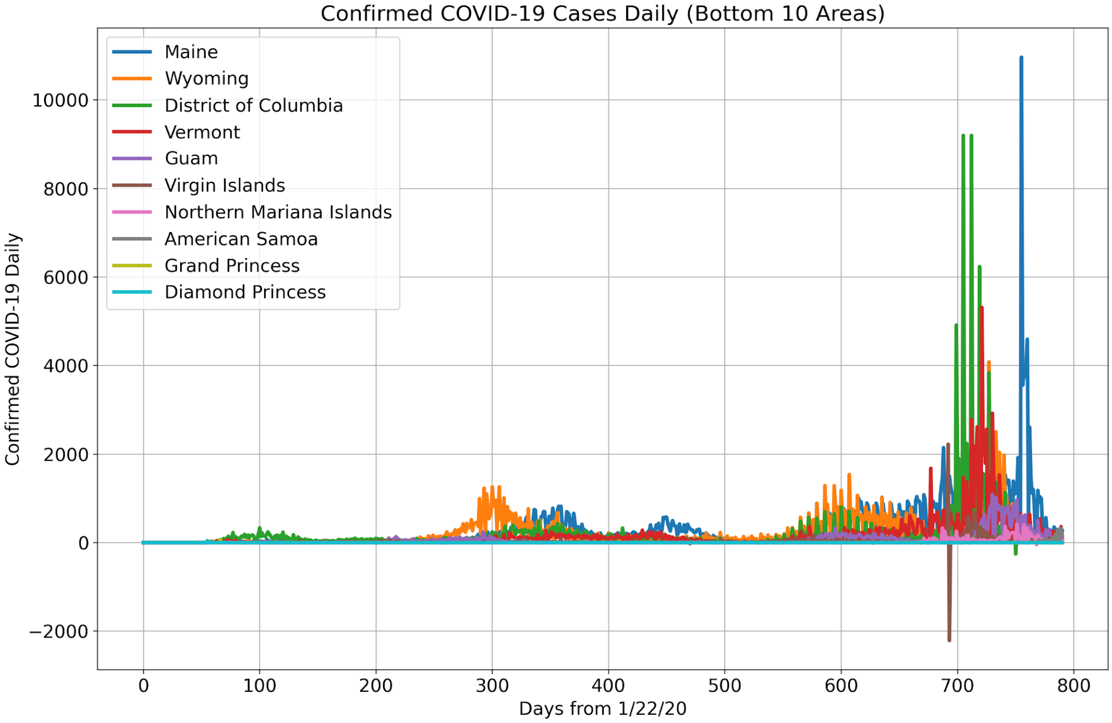

Figure 3: Bottom-10 States with average daily confirmed cases

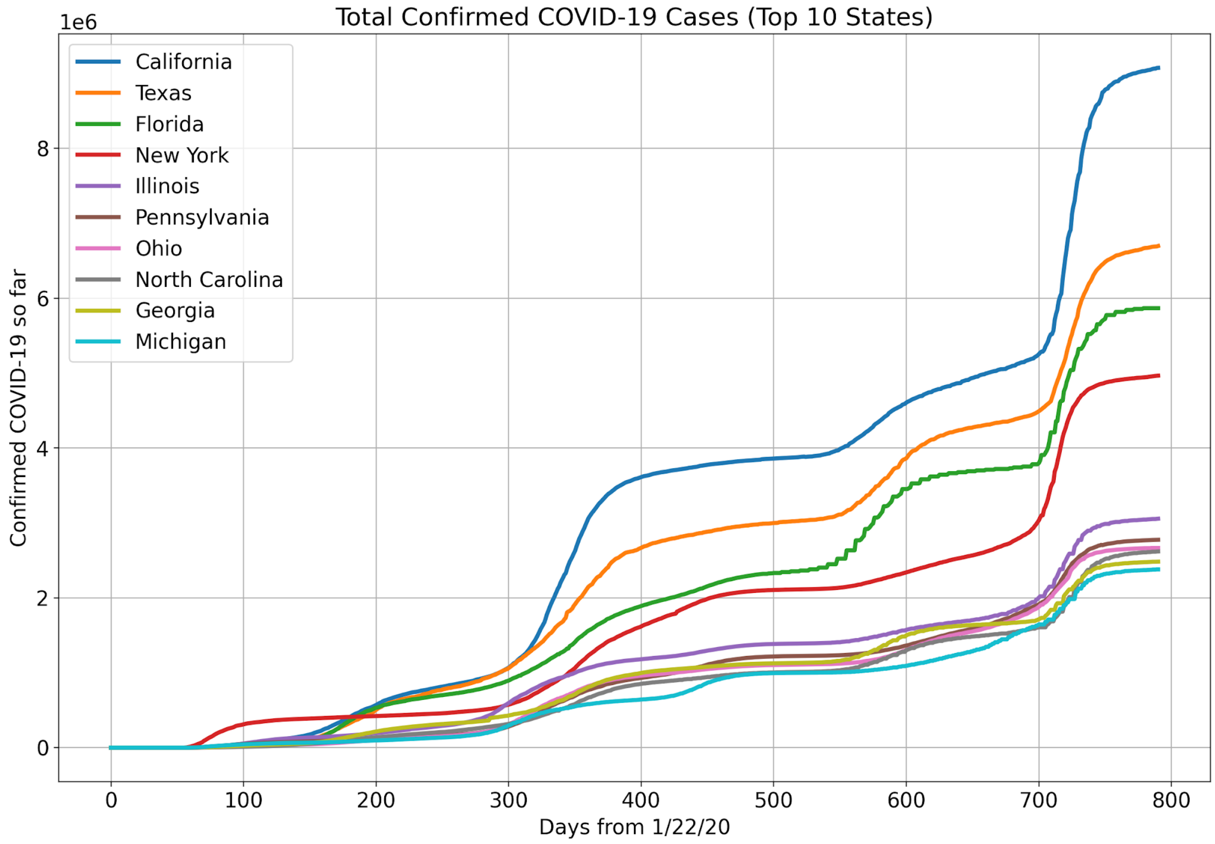

Figure 4: Top-10 States with total confirmed cases so far

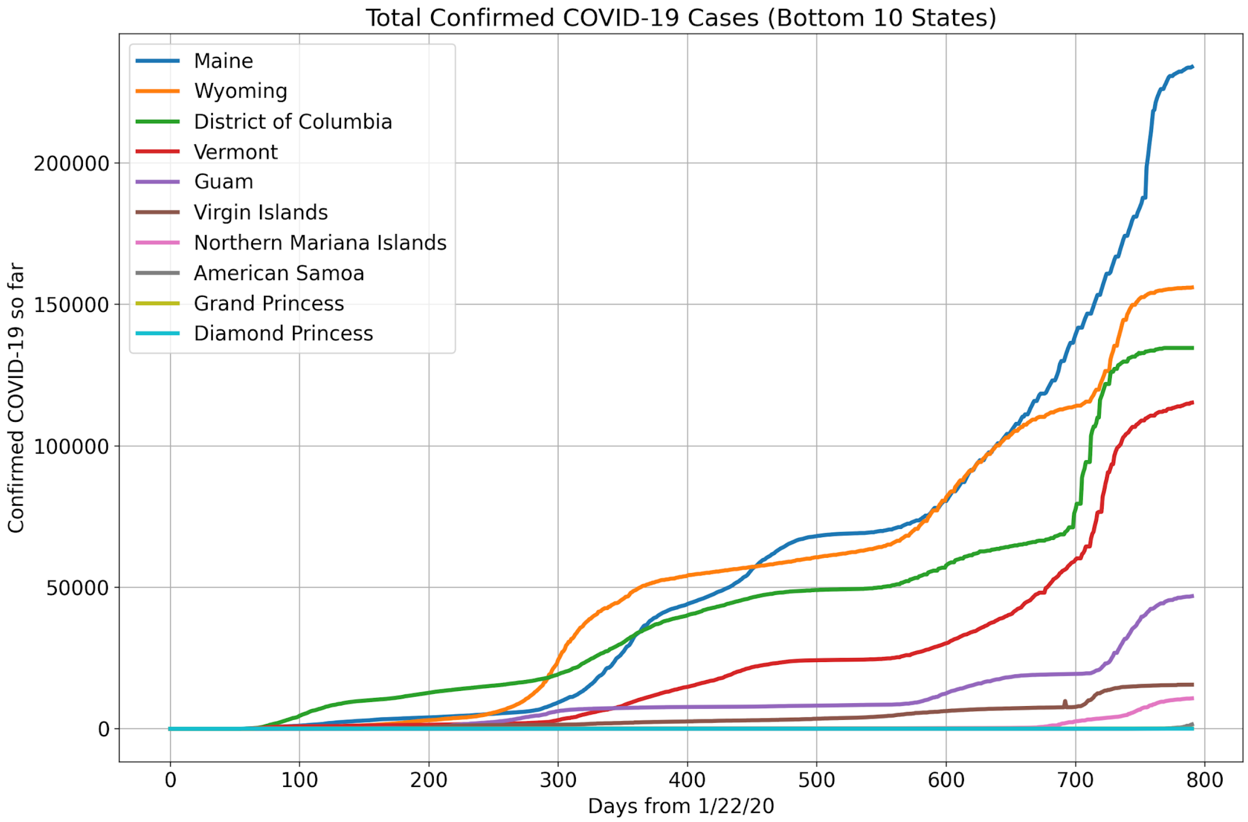

Figure 5: Bottom-10 States with total confirmed cases so far

Models Used for Prediction:

We used the following four models for prediction.

- Persistence Model
- AutoRegression Model
- Linear Regression Model
- ARIMA model

The persistence model was used as the baseline model. We used the data from the previous day to predict the next day value for the persistence model. Autoregression model is a linear regression model that uses lagged variables as input variables. Following is the equation for the regression model.

$$Y=b_0 + b_1 * X(t-1) + b_2 * X(t-2) + \ldots\ldots\ldots.$$

The ARIMA model uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Results:

The following figures contain results for the different prediction models for the state of Michigan to compare their performance. We predicted results for the next 30 days for all the models.

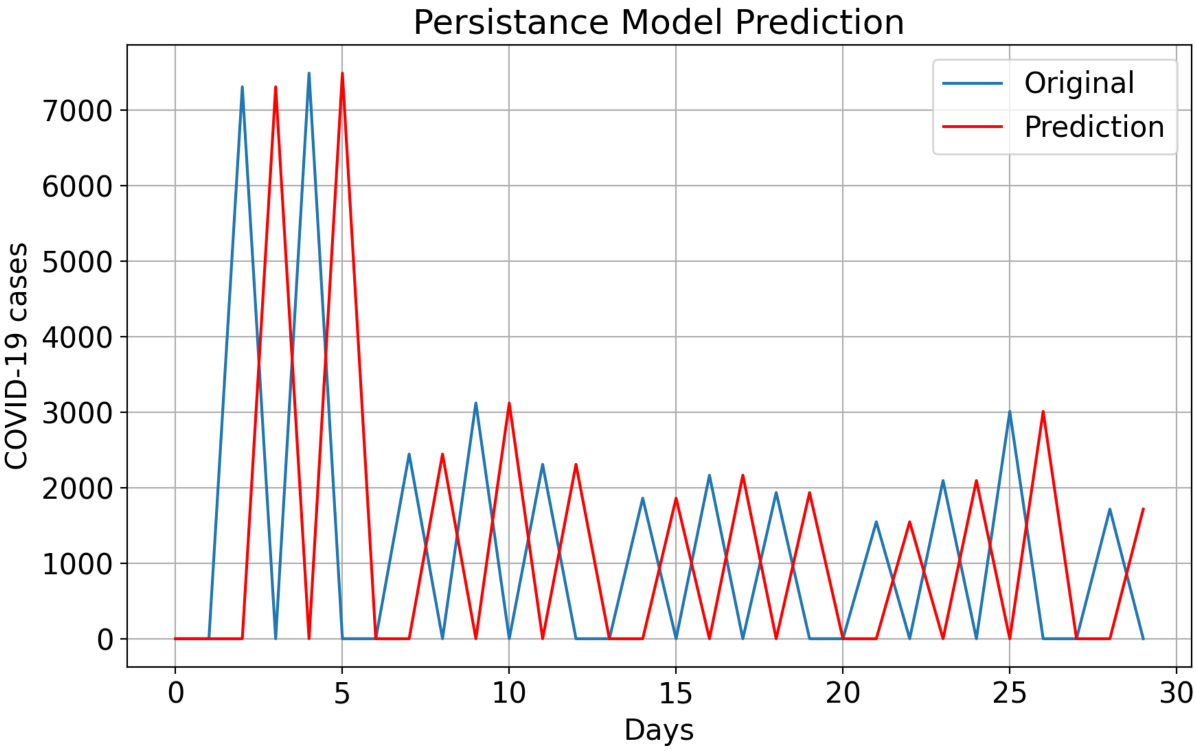

Figure 6: Prediction of COVID-19 cases for the next 30 days using the Persistence model.

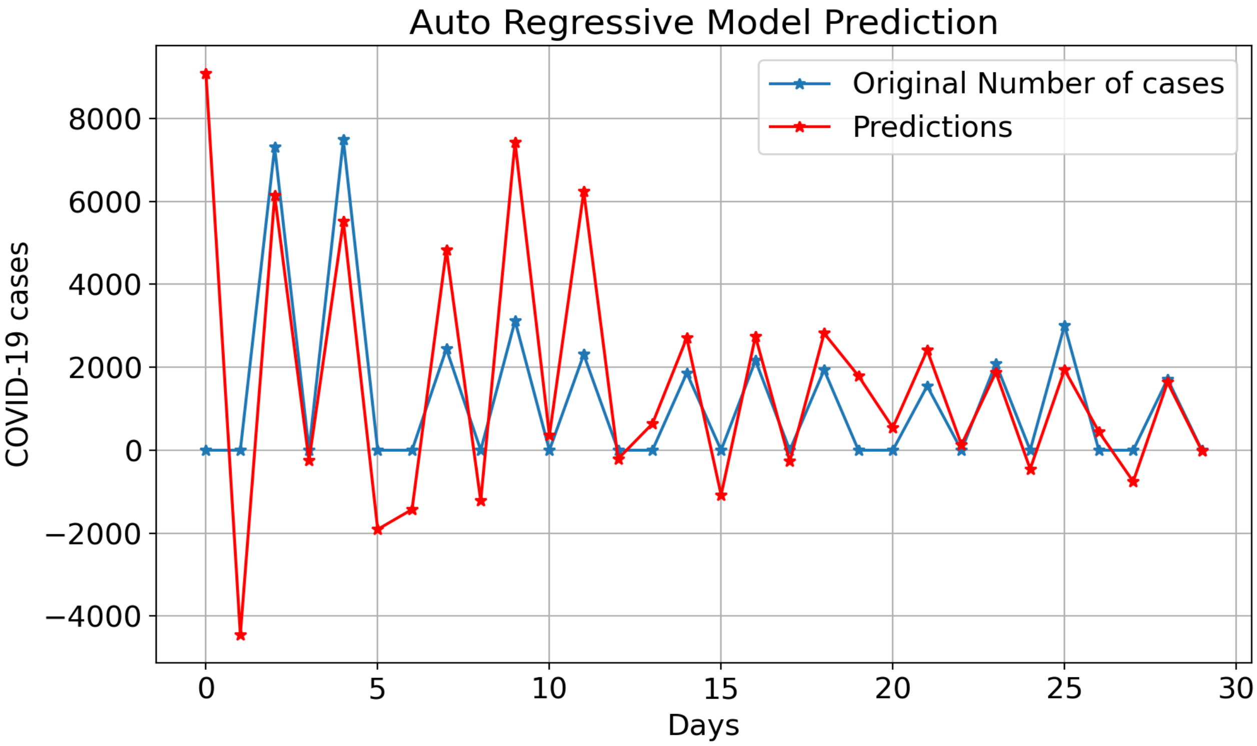

Figure 7: Prediction of COVID-19 cases for the next 30 days using the Autoregression model

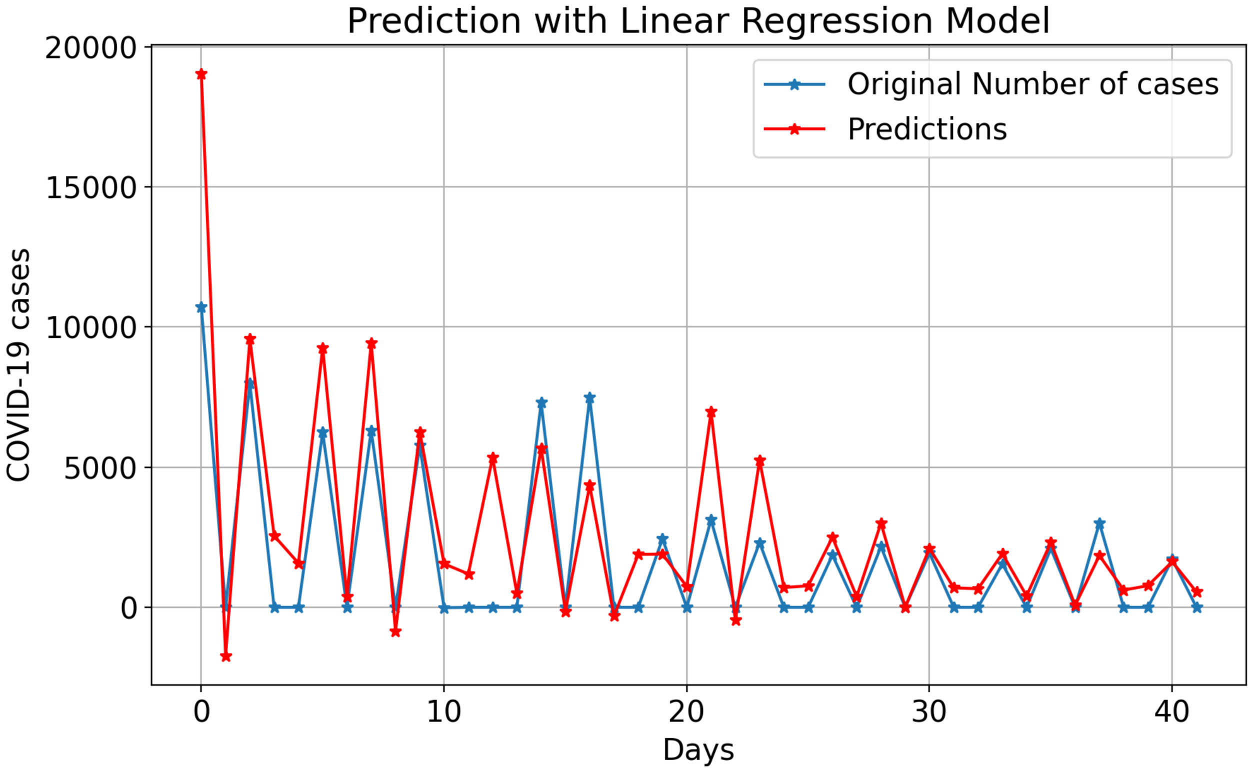

Figure 8: Prediction of COVID-19 cases for the next 30 days using the Linear Regression model.

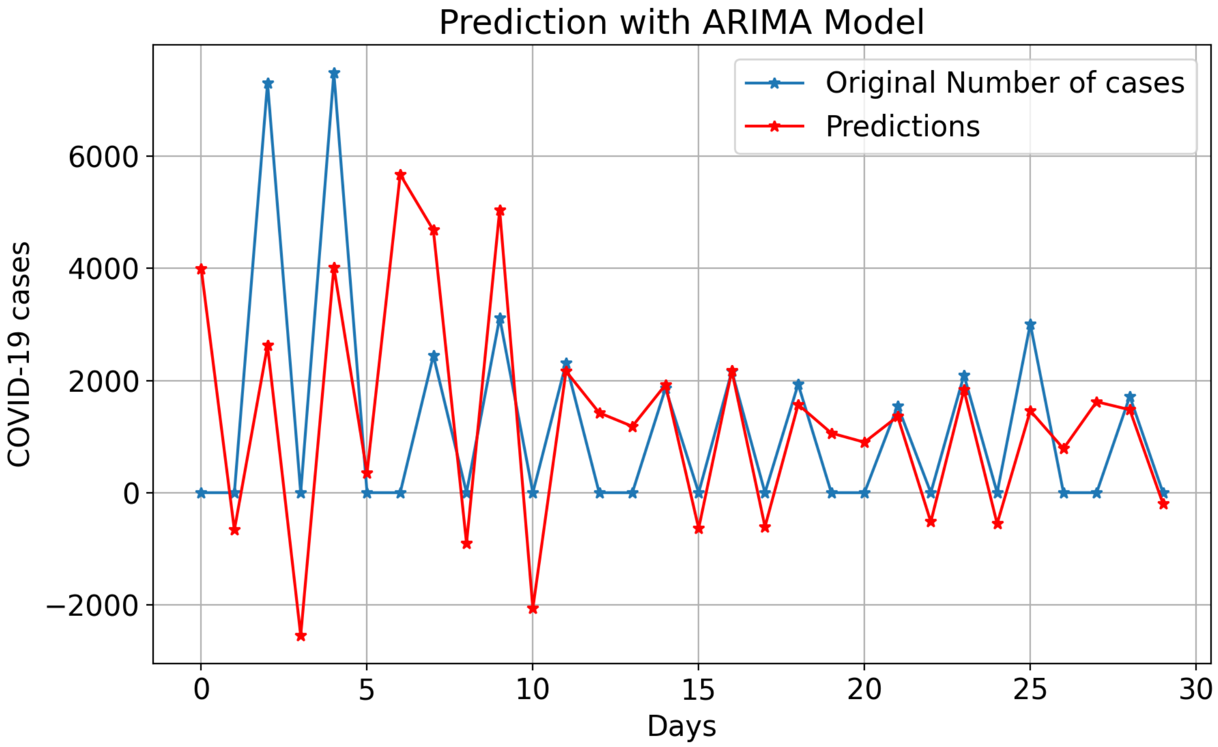

Figure 9: Prediction of COVID-19 cases for the next 30 days using the ARIMA model.

Figure 10 depicts the comparison of the four models used for the prediction based on the RMSE error.

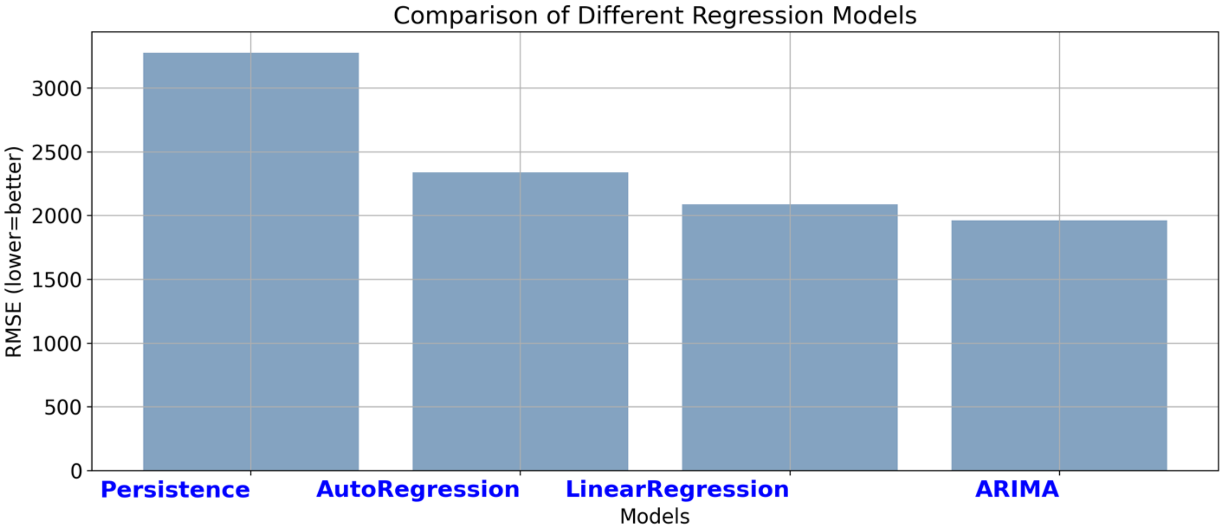

Figure 10: Performance comparison of the four prediction models

Conclusion and Future Works:

We see that among the four models, the AIMA model outperformed the other models. However, in the Project, we only used the top 10 and below 10 sates with respect to the confirmed cases of the United States. As an extension of the analysis, the states can be clustered in several groups based on the pattern of the daily cases. Separate models can be trained for different cluster of states with similar COVID-19 cases pattern to predict more accurately. Impact of different weather attributes can also be analyzed to get a better prediction of COVID-19 cases.