Capstone Project

Global COVID-19 Trends and New Case Prediction for the USA: Final Report

1. **Problem Statement:**

The COVID-19 pandemic has profoundly impacted public health and the economy, necessitating forecasting and risk prediction models to help decision-making. Accurate forecasting of COVID-19 cases can assist policymakers, healthcare professionals, and researchers in optimizing resource allocation, implementing timely interventions, and mitigating the spread of the virus.

**Goal:** The primary goal of this project is to explore the COVID-19 trends around the globe and to develop a machine learning model to predict new cases in the United States. This will involve: • Time Series Forecasting: Predicting the number of new cases over the next few weeks in the USA.

1. **Model Outcomes or Predictions:**

This problem is a regression-type problem using historical labeled data (supervised learning). The model learns from this data to minimize prediction errors (e.g., MAE, RMSE). The following algorithms are utilized:

• ARIMA (Autoregressive Integrated Moving Average) — captures trends and patterns.

• SARIMA (Seasonal ARIMA) — incorporates seasonality & trends.

• Exog SARIMA — extends SARIMA by adding external factors (e.g., mobility, vaccinations).

• Random Forest Regressor — handles time-lagged features and complex relationships.

• XGBoost — gradient boosting approach.

• LSTM — neural networks designed for sequential data.

The model aims to predict weekly new COVID-19 cases in the USA.

1. **Data Acquisition:**

Primary Data Source: WHO COVID-19 global dataset from Kaggle, providing comprehensive time-series data on reported cases and deaths globally.

• Dataset: WHO COVID-19 dataset

• Source: Kaggle

• Time Period: 2020 to 2023

• Target Variable: New COVID-19 cases reported weekly

• Data Size: 57,840 entries and 8 columns

1. **Data Preprocessing:**

• Handled missing values (imputed missing cases/deaths as zeros).

• Resolved encoding issues in country names.

• Dropped irrelevant columns (e.g., country codes).

• Removed duplicates, negatives, and zeros (for analysis).

• Feature engineering: calculated 7-day moving average, growth rate, and reproduction number (R₀).

1. **Modeling:** Algorithms applied:
   1. ARIMA
   2. SARIMA
   3. Exog SARIMA
   4. Random Forest Regressor
   5. XGBoost
   6. LSTM
2. **Model Evaluation:**

Each model was evaluated using MAE, MSE, RMSE, AIC, and BIC metrics.

**Model Evaluation Observations:**

1. ARIMA vs. SARIMA:

• SARIMA outperforms ARIMA with lower MAE and RMSE.

• AIC/BIC values slightly favor ARIMA, but SARIMA’s improved accuracy outweighs this.

1. EXOG SARIMA:

• Incorporating exogenous factors improved MAE, MSE, and RMSE compared to ARIMA/SARIMA.

• AIC/BIC values indicate a slightly worse fit but are offset by improved prediction accuracy.

1. Random Forest:

• Achieved significantly lower MAE than ARIMA and SARIMA, indicating strong absolute error reduction.

1. XGBoost:

• Higher MAE and RMSE suggest it struggles more with this dataset compared to Random Forest.

1. LSTM (Training vs. Test):

• LSTM (Training) performed decently, but LSTM (Test) yielded the best overall performance.

• LSTM (Test) demonstrates strong generalization, outperforming statistical methods and other machine learning models.

**Key Takeaways:**

• SARIMA and EXOG SARIMA provide strong baseline forecasts.

• Random Forest minimizes MAE better than ARIMA and SARIMA but shows higher RMSE.

• LSTM (Test) outperforms all other models, making it the top-performing model.

• XGBoost shows weaker performance, likely needing tuning or different feature engineering.

Final Model Performance Comparison:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **MAE** | **MSE** | **RMSE** | **AIC** | **BIC** |
| **ARIMA** | 134,360.86 | 2.48e+10 | 157,527.53 | 3380.63 | 3397.55 |
| **SARIMA** | 97,100.10 | 1.64e+10 | 128,000.76 | 3462.81 | 3474.15 |
| **EXOG SARIMA** | 79,852.14 | 8.67e+9 | 93,150.65 | 3669.55 | 3689.84 |
| **Random Forest** | 68,322.60 | 6.6e+9 | 81348.77 | N/A | N/A |
| **XGBoost** | 122,012.04 | 3.28e+10 | 181259.3 | N/A | N/A |
| **LSTM (Training)** | 81,126.04 | 3.43e+10 | 185,258.85 | N/A | N/A |
| **LSTM (Test)** | 38,242.36 | 5.6e+7 | 56,711.65 | N/A | N/A |

The final report concludes that LSTM (Test) is the most effective model for predicting new COVID-19 cases in the USA.

**Link for the Jupyter Notebook:**

<https://github.com/preetikumar20/Capstone_Covid19/blob/a871861f8f05c0c51227f05e58bcdd50e33ce2d5/final_capstone_preetidubey.ipynb>