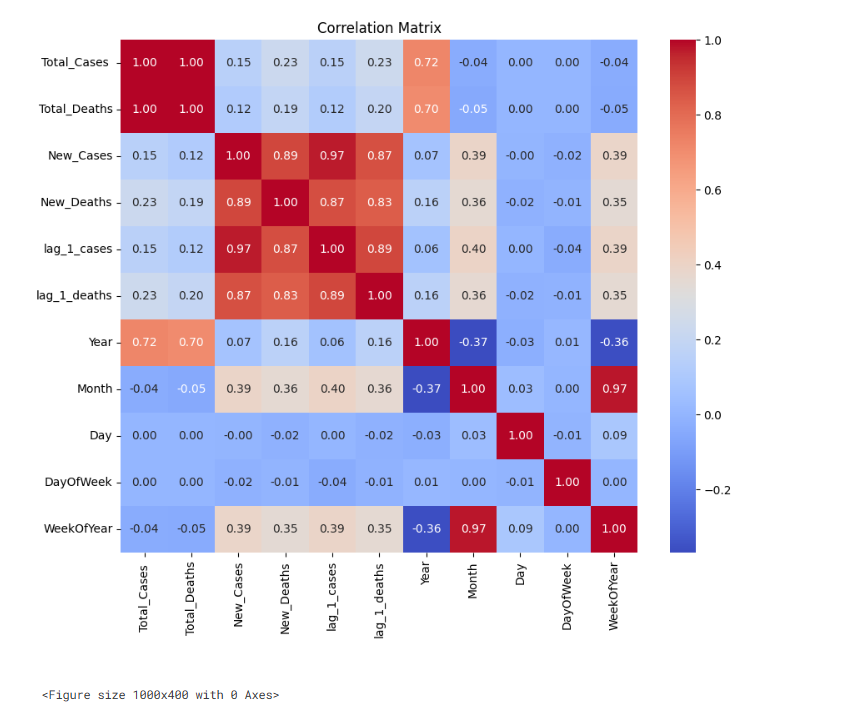
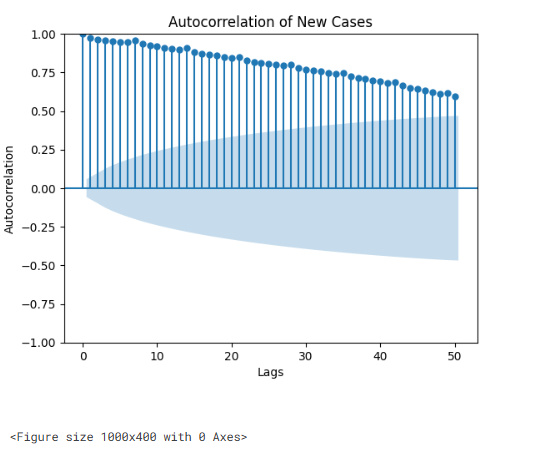
1. It plots a **correlation matrix** using a heatmap to show relationships between variables.

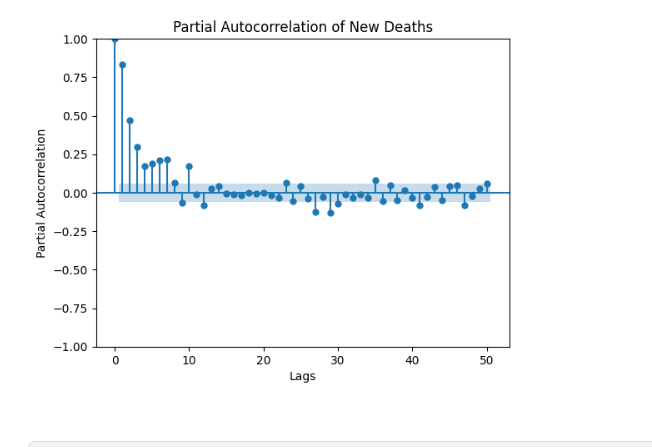
2. It creates **autocorrelation plots** for New\_Cases and New\_Deaths to analyze how each value correlates with its past values over a lag period of 50.

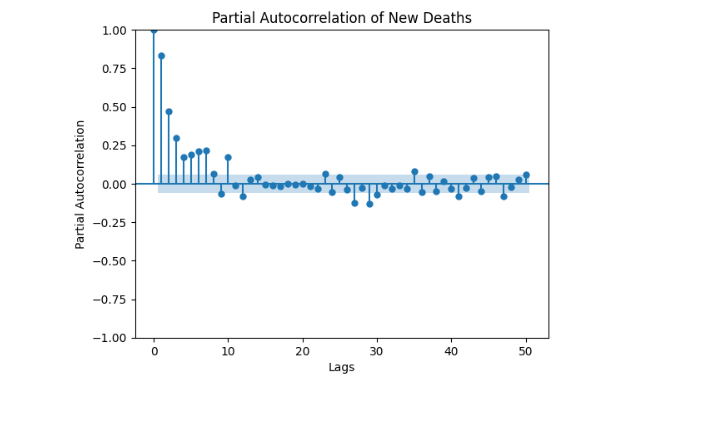




1.This **plots the partial autocorrelation (PACF)** for New\_Cases and New\_Deaths, showing the correlation of each value with its prior values, controlling for intermediate lags.

2.The **ywm method** is used for PACF estimation, suitable for large datasets, with 50 lags analyzed to identify significant relationships over time.





**XGBOOST:**

**Preprocessing:** -Convert 'Date' to datetime and set as index

-Lag features (1, 2, 7-day lags).

- Rolling statistics (7-day mean and std).

- Add time-based features (year, month, day, day of the week, week of the year

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | RMSE | MSE | MAE | R2 |
| New cases | 108.34 | 11737.06 | 46.80 | 0.98 |
| New deaths | 1.86 | 3.45 | 0.86 | 0.77 |

|  |  |  |
| --- | --- | --- |
| MODEL | Preprocessing | Key parameter |
| XGBOOST | Log transformation, Standard Scaling, PCA, Feature Importance selection | Objective: 'reg:squarederror', Random\_state |
| RandomForest | Log transformation, Standard Scaling, PCA, Feature Importance selection. | objective: 'reg:squarederror', Random\_state: 42 |
| GradientBoosting | Log transformation, Standard Scaling, PCA, Feature Importance selection. | Random\_state: 42 |
| LightGBM | Log transformation, Standard Scaling, PCA, Feature Importance selection, Bayesian Hyperparameter | Bayesian Hyperparameter Tuning with Grid Search, Random\_state: 42 |
| StackingRegressor | Combination of all individual model predictions (with log transform, scaling, and PCA applied). | Uses XGBoost, RF, GB, and LGBM predictions with Random Forest as final estimator. |

**Their evaluation matrixmis given below**:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Model | rmse | mse | mae | R2 |  |
| NEW cases | xgbost | 111.26 | 12378.5 | 53.54 | 0.98 |  |
| RandomForest | 100.27 | 10054.89 | 50.67 | 0.979 |  |
| GradientBOOsting | 113.67 | 12918.63 | 53.82 | 0.97 |  |
| LIghtGBM | 114.44 | 13097.37 | 55.82 | 0.97 |  |
| Stacking | 103.77 | 10769.1 | 51.72 | 0.98 |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| New deaths | Model | Rmse | mse | mae | R2 |
| Xgboost | 785.21 | 616550.29 | 369.21 | -36700.26 |
| RandomForest | 779.29 | 603002.21 | 365.51 | -36700.26 |
| GradientBOOsting | 776.53 | 607284.59 | 365.51 | -36441.45 |
| LIghtGBM | 769.08 | 591498.52 | 361.59 | -35746.23 |
| Stacking | 1.31 | 1.75 | 0.71 | 0.89 |
|  |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | MODEL | preprocessing | Rmse | mse | mae | R2 |
| NEW CASES | TPOT | **PCA**: Dimensionality reduction to retain the 10 most important components.  **Log Transformation**: Target variable transformed to address skewness. | 108.67 | 11815.08 | 53.72 | 0.98 |
| NEW cases | LSTM | **MinMaxScaler**: Scaled data to the range [0, 1] to help with convergence.  **Lag Features**: Historical values used to create windows for sequence modeling.  **Rolling Statistics**: Trends captured with moving averages and standard deviations.  **Interaction Features**: Featurecases\_deaths\_interaction created to capture interactions between cases and deaths.  **Windowing for Time-Series Data**: Data prepared into windows of size 60 (n\_steps). | 749.35 | 561530.32 | 328.393 | -0.16 |
| **New deaths** | **LSTM** | **1.72** | **2.96** | **0.85** | **0.81** |

XGBOOST Regressor:

|  |  |  |
| --- | --- | --- |
| preprocessing | details |  |
| 1.Feature Scaling | StandardScaler is used to scale input features with mean = 0 and std = 1 for consistency. | Best RMSE Score: 0.559 |
| 2.Log-Transformation | Applied to the target variable (New\_Cases) to reduce skewness and stabilize variance. |
| Feature engineering | - **Lag Features**: Use of previous values of New\_Cases and New\_Deaths.  - **Rolling Statistics**: Moving averages and standard deviations to capture trends.  - **Interaction Features**: Product of New\_Cases and New\_Deaths. |
| Engineering 4.Dimensionality Reduction | PCA is applied to reduce the feature set to the top 10 components, enhancing performance and avoiding overfitting.. |

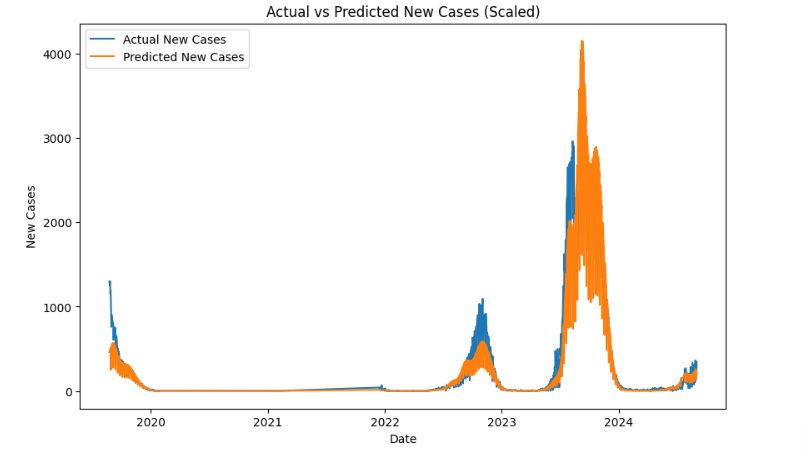
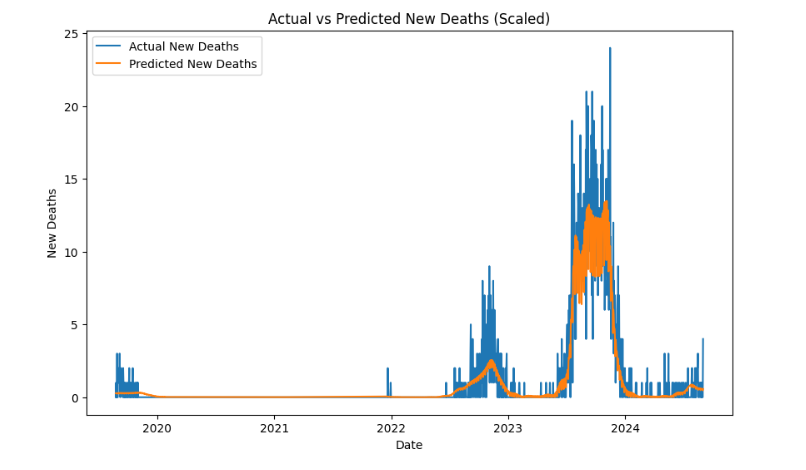
**RandomForestRegressor:**

|  |  |  |
| --- | --- | --- |
| preprocessing | details |  |
| 1.Feature Scaling | - **Lag Features**: Captures historical patterns by introducing previous values of New\_Cases and New\_Deaths.  - **Rolling Statistics**: Calculates moving averages and standard deviations to model trends and variability. | Average RMSE: 584.47 |
| 2.Log-Transformation | pplied to the target variable to address skewness and stabilize variance. |
| Scaling | **StandardScaler**: Scales input features to mean = 0 and std = 1, ensuring that the optimization converges faster. |
| 4.Dimensionality Reduction | **PCA (Principal Component Analysis)**: Reduces feature dimensions to 10 components, helping avoid overfitting. |

**Time Series:**

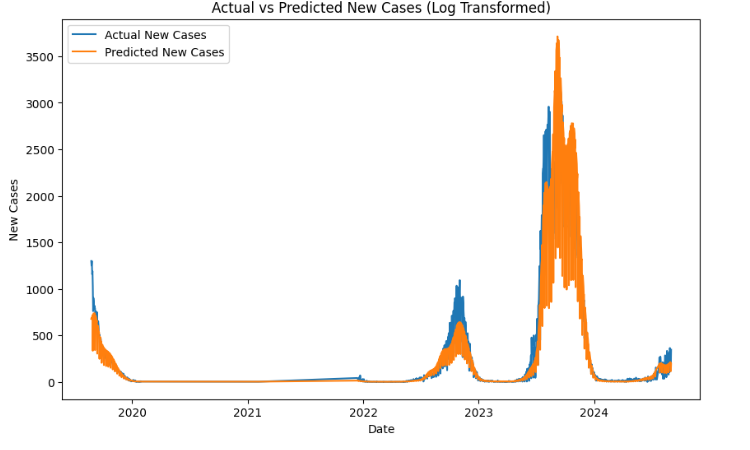
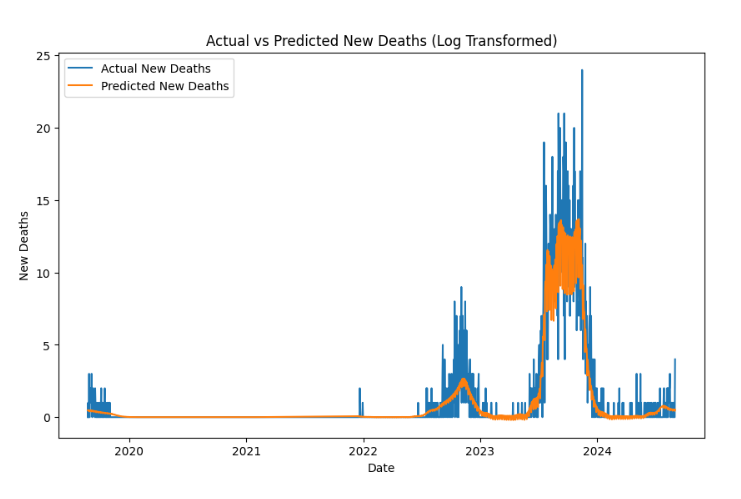
Prophet:(Actual vs predicted)🡪Scaled

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| preprocessing | details |  | RMSE | MSE | MAE | | R2 |
| Date Parsing | The Date column is converted to datetime format to align with Prophet’s requirements. | New cases | 274.22 | 75198.86 | 116.06 | 0.833 | |
| Log Transformation | - Applied to New\_Cases and New\_Deaths using log1p to reduce skewness and avoid negative predictions.  - This ensures the model works better with non-negative targets.  - Predictions are inverse-transformed using expm1 to return to the original scale. |
| Feature Preparation | he dataset is renamed to meet Prophet’s input requirements (Date → ds, Log\_New\_Cases → y). | New deaths | 1.61 | 2.61 | 0.78 | | 0.82 |
| Clipping Predictions | Negative predictions are clipped to 0 to ensure the forecast makes practical sense (no negative cases or deaths). |  |  |  | |  |

**LOG transformed**: same preprocessing in Prophet scaled

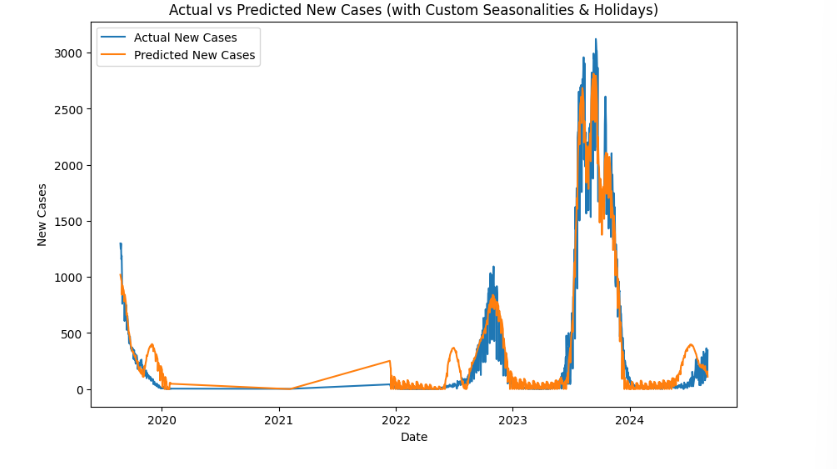
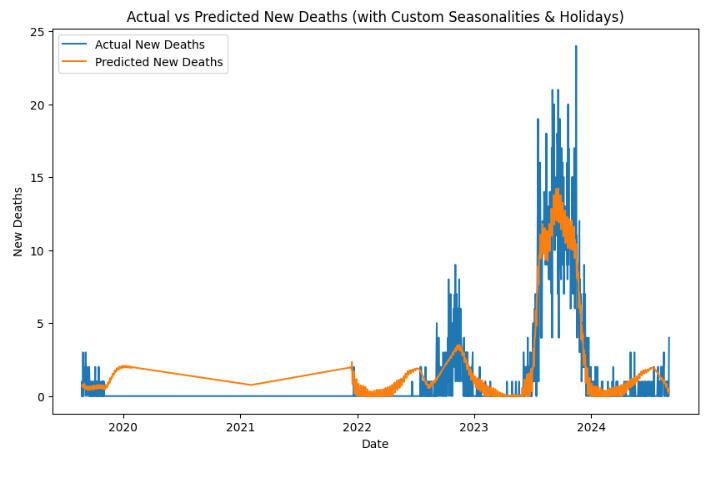
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | rmse | msse | mae | R2 |
| New cases | 231.61 | 53539.56 | 100.56 | 0.89 |
| New deaths | 1.59 | 2.53 | 0.78 | 0.83 |

**Prophet:(Forecasting new cases and new deaths)**

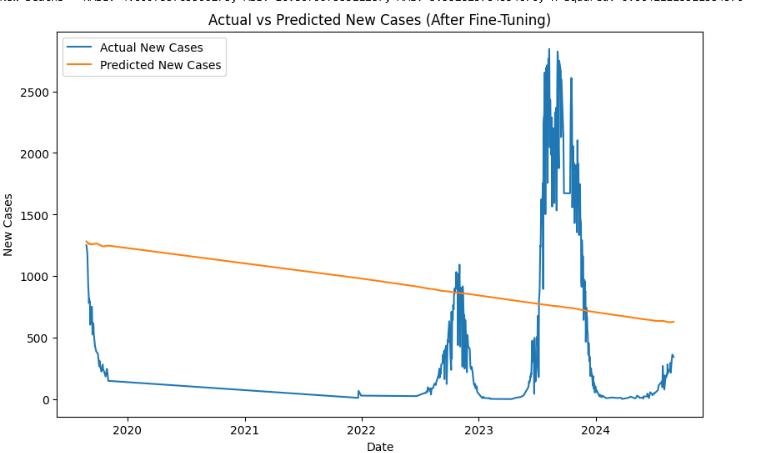
With custom seasonalities and holiday:

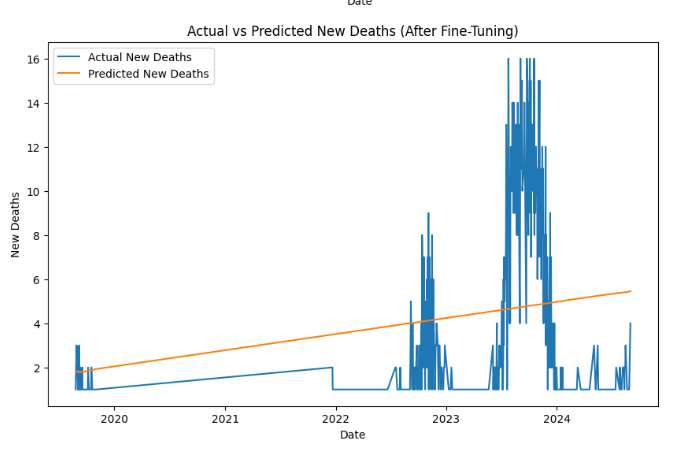
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| preprocessing | details |  | RMSE | MSE | MAE | | R2 |
| Date Parsing | The Date column is converted to datetime format to align with Prophet’s requirements. | New cases | 114.56 | 20896.35 | 96.02 | 0.95 | |
| Scaling | - **StandardScaler**: Scales New Cases and New Deaths to standardize the input data with mean = 0 and std = 1, ensuring better convergence during training.  - Inverse transformation is applied to predicted values to bring them back to the original scale.. |
| Adding Floor Constraint | A **floor of 0** is added to prevent negative predictions, ensuring practical forecast values (no negative cases or deaths). | New deaths | 1.68 | 2.84 | 1.09 | | 0.81 |
| Custom Holidays and Seasonalities | - Custom **bi-weekly seasonality** (14-day period) is introduced to capture recurring patterns.  - **Health intervention holidays** are added to model their impact. |  |  |  | |  |

**After fine tuning**: same preprocessing in Prophet Seasonalities and holiday

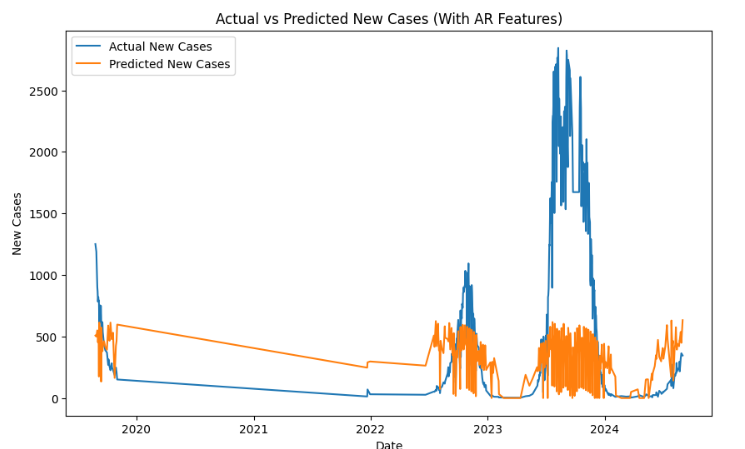
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | rmse | mse | mae | R2 |
| New cases | 833.56 | 694810.13 | 697.98 | -0.089 |
| New deaths | 4.1 | 16.81 | 3.35 | 0.064 |

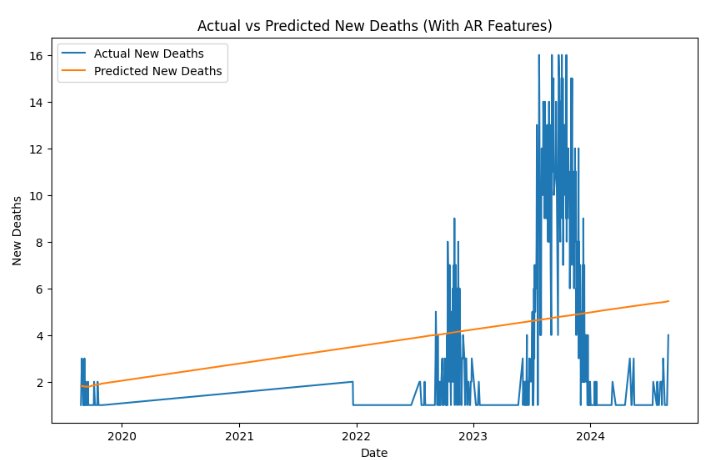




With AR FEATURE:

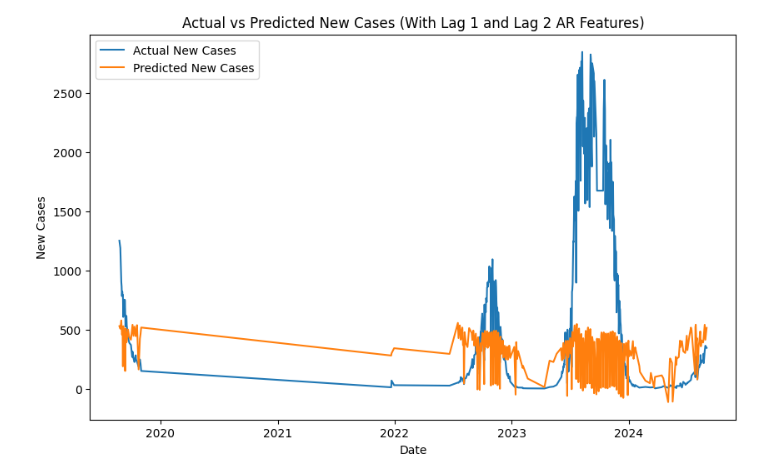
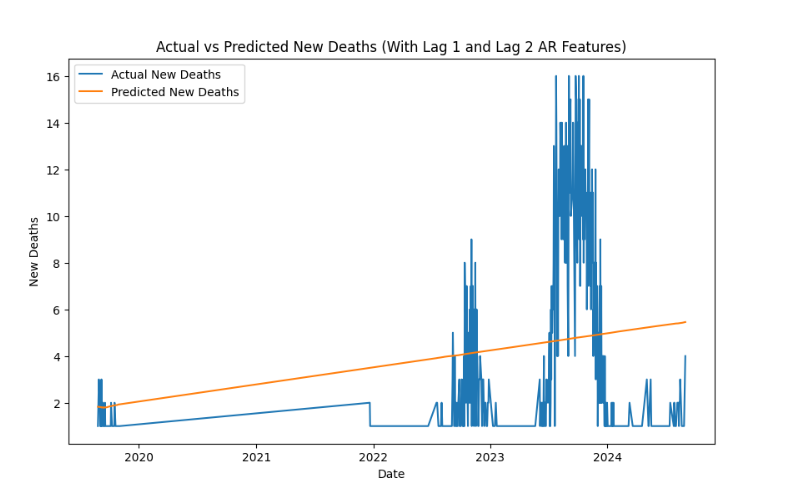
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| preprocessing | details |  | RMSE | MSE | MAE | | R2 |
| Date Parsing | Converts the Date column to **datetime** format, a requirement for Prophet. | New cases | 897.57 |  | 611.28 | -0.25 | |
| Scaling | - **StandardScaler**: Standardizes New Cases and New Deaths to mean = 0 and std = 1 for better convergence.  - **Inverse transformation** applied to bring predictions back to the original scale. |
| Autoregressive Features (Lag) | Adds **previous day’s values** as features (lag\_1\_cases, lag\_1\_deaths), enabling the model to leverage the temporal dependencies. | New deaths | 4.09 |  | 3.35 | | 0.06 |





With lag1 and lag2 feature

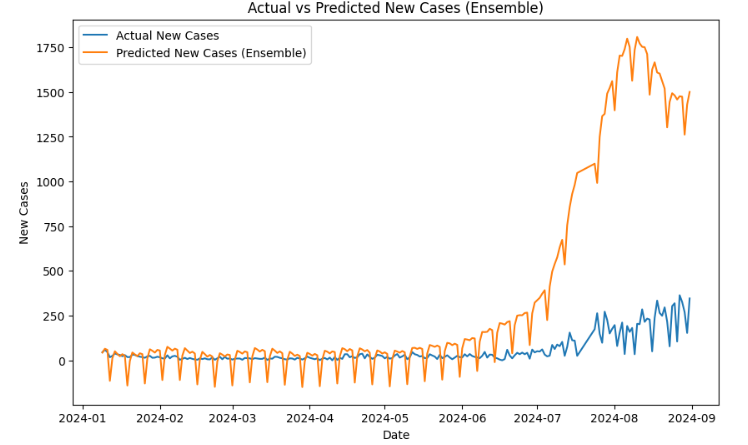
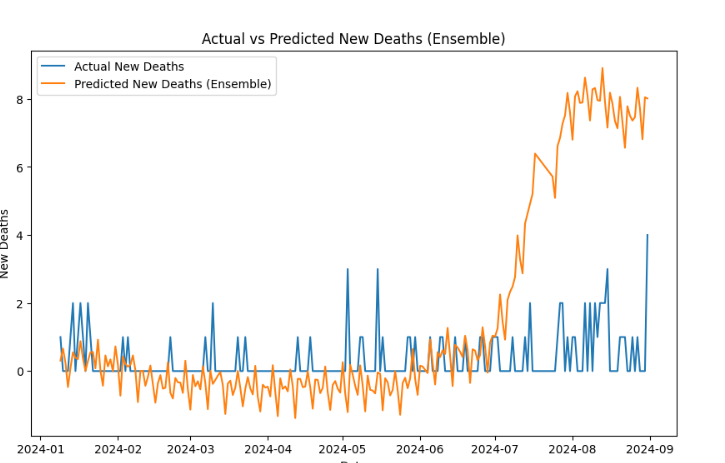
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| preprocessing | details |  | RMSE | MSE | MAE | | R2 |
| Date Parsing | Converts the Date column to **datetime** format, which is required for Prophet to understand time-based data | New cases | 919.25 | 845027.43 | 630.57 | -0.31 | |
| Scaling | - **StandardScaler**: Standardizes New Cases and New Deaths to mean = 0 and std = 1 for better convergence.  - **Inverse transformation** applied to bring predictions back to the original scale. |
| AR feature(lagged value) | Adds **lag\_1** and **lag\_2** days of values as predictors (lag\_1\_cases, lag\_2\_cases, lag\_1\_deaths, lag\_2\_deaths). These help the model understand the sequential dependencies in the data. | New deaths | 4.09 | 16.81 | 3.35 | | 0.06 |

**Ensumble with prophet**:

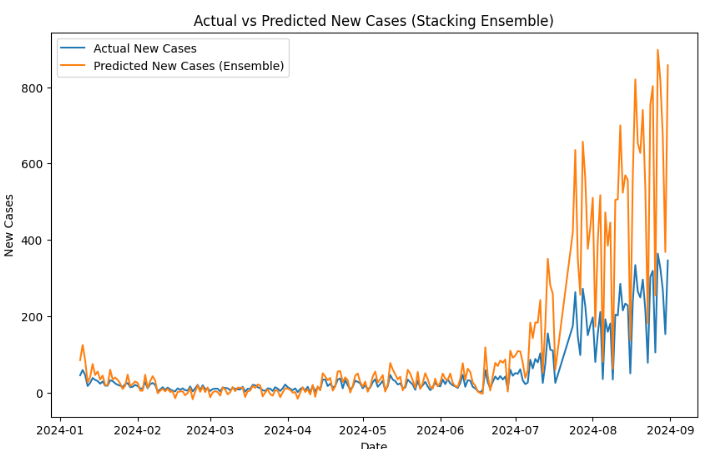
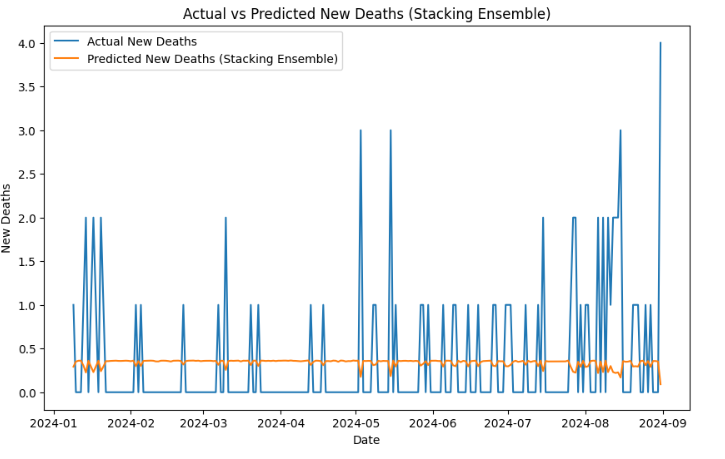
(using three model prophet,sarima,xgboost)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| preprocessing | details |  | RMSE | MSE | MAE | | R2 |
| Date Parsing | The Date column is converted to **datetime** format to ensure Prophet correctly interprets the time-series data. | New cases | 580.59 | 337089.16 | 310.78 | -0.52.2 | |
| Lag Features Creation | Two lag features are created for both New\_Cases and New\_Deaths (i.e., lag\_1 and lag\_2). These features capture temporal dependencies and are used as inputs for **XGBoost**. |
| Handling Missing Values | Rows with missing values (e.g., after shifting for lag features) are **dropped** to avoid issues during model training and prediction. | New deaths | 3.02 | 9.15 | 1.78 | | -1.8 |
|  |  |  |  |  |  | |  |

Using sarima,prophet,xgboost,stacking regressor:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| preprocessing | details |  | RMSE | MSE | MAE | | R2 |
| Date Parsing | The Date column is converted to **datetime** format to ensure Prophet correctly interprets the time-series data. | New cases | 138.36 | 19144.512 | 66.58 | -2.02 | |
| Lag Features Creation | **Lag 1 and Lag 2 features** are created for both New\_Cases and New\_Deaths.  - These features help XGBoost capture temporal dependencies. |
| Handling Missing Values | Rows with missing values (e.g., created due to shifting for lag features) are **dropped** to ensure smooth model training.. | New deaths | 0.72 | 0.53 | 0.54 | | -0.12 |

**NEW Ensumble:**

Prophet:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| preprocessing | details |  | RMSE | MSE | MAE | | R2 |
| Date Parsing | The Date column is converted to **datetime** format, required for Prophet to interpret the time-series.. | New cases | 533.72 |  |  |  | |
| Lag Features Creation | - Adds **lag\_1** and **lag\_2** features for both New\_Cases and New\_Deaths.  - These lag features capture short-term dependencies, though they are not directly used by Prophet but can help in other models if combined.. |
| New deaths | 2.82 |  |  |  | |

SARIMA:

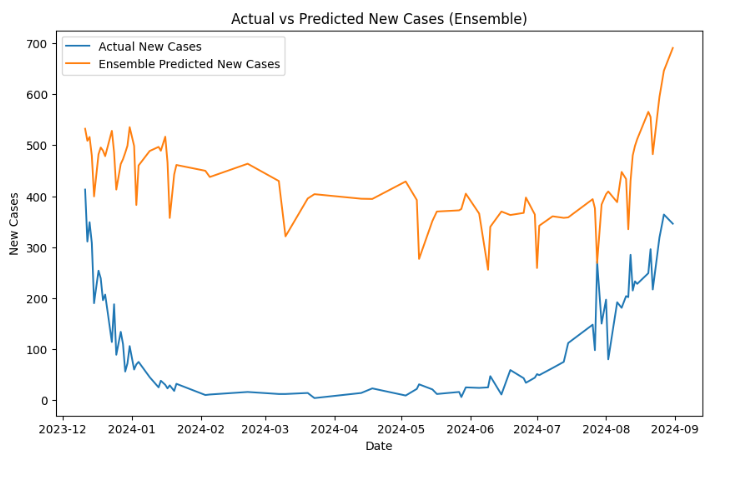
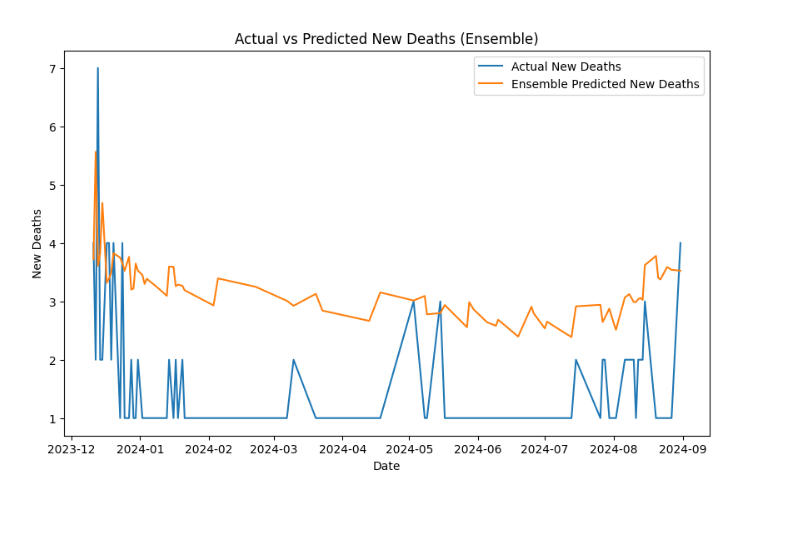
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| preprocessing | details |  | RMSE | MSE | MAE | | | R2 |
| Handling Missing Values | Any missing values are handled before applying SARIMA to avoid prediction error | New cases | 356.33 |  |  | |  | |
| Seasonal Patterns | Since the data has **seasonal components**, SARIMA uses seasonal=True to capture recurring patterns. |
| New deaths | 1.61 |  |  |  | | |
| Prediction Horizon | Forecasts are made for the **length of the test set** (n\_periods=len(test\_data)) to evaluate the model’s accuracy. |  |

XGBOOST regressor:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| preprocessing | details |  | RMSE | MSE | MAE | | | R2 |
| Lag Features Creation | - Two lag features are created: lag\_1 and lag\_2 for both **New Cases** and **New Deaths**.  - These lag features capture temporal dependencies, helping XGBoost learn patterns based on recent past values. | New cases | 64.37 |  |  | |  | |
| Handling Missing Values | Missing values from shifting operations are **filled with 0** to avoid training issues. |
| New deaths | 1.43 |  |  |  | | |
| Grid Search for Hyperparameter Tuning | A **Grid Search with Cross-Validation** (CV = 5) is used to find the optimal hyperparameters for XGBoost.  The parameters explored include:  - **n\_estimators**: Number of boosting rounds (50, 100)  - **learning\_rate**: Controls the contribution of each tree (0.01, 0.1)  - **max\_depth**: Maximum depth of trees (3,5) |  |

Ensemble of Prophet(0.4), SARIMA(0.3), and XGBoost(0.3):

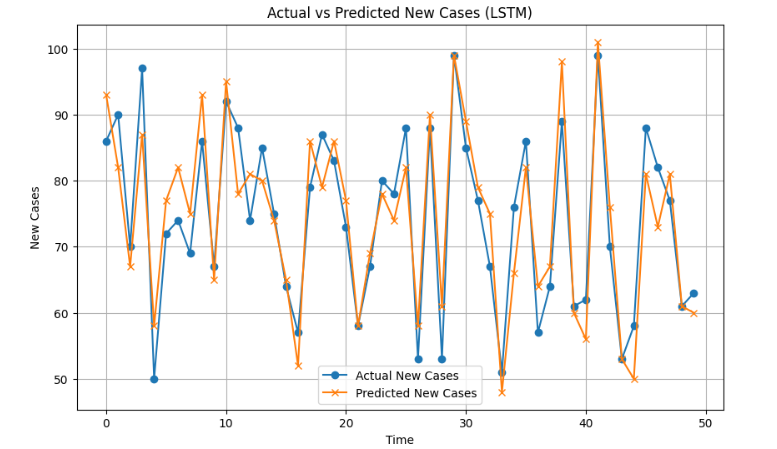
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| preprocessing | details |  | RMSE | MSE | MAE | | | R2 |
| Date Parsing | - Converts the Date column to **datetime** format, required for Prophet to interpret the time-series data. | New cases | 329.38 |  |  | |  | |
| Lag Features Creation | Creates **lag\_1 and lag\_2 features** to capture recent trends for both New\_Cases and New\_Deaths. These are used by **XGBoost** for predictive modeling. |
| New deaths | 1.85 |  |  |  | | |
| Grid Search for XGBOOST Tuning | Performs hyperparameter tuning on XGBoost using **GridSearchCV** to optimize the model’s performance. |  |
| Handling Missing Values | Missing values created by shifting operations are **filled with 0** to ensure smooth model training. |  |  |  |  |  | | |
| Model Training | Trains **Prophet, SARIMA, and XGBoost models** independently on the training set. |  |  |  |  |  | | |

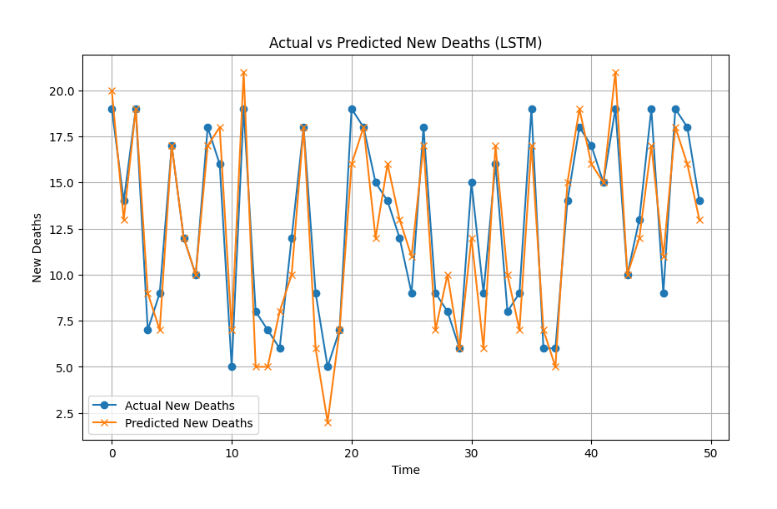
 

***Deep learning way***:

LSTM:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| preprocessing | details |  | RMSE | MSE | MAE | | | R2 |
| Date Parsing | - Converts the Date column to **datetime** format to ensure proper handling of time-series data. | New cases | 0.29 |  |  | |  | |
| Scaling with Min-Max Scaler | Uses **MinMaxScaler** to scale New\_Cases and New\_Deaths between **0 and 1**. LSTM models perform better with normalized data. |
| New deaths | 0.77 |  |  |  | | |
| Creating LSTM Dataset | Converts the time-series data into **sequences** using a defined **time step** (e.g., 7 days for cases and 10 days for deaths).  Each input sequence consists of the last few days, and the target is the next day’s value.. |  |
| Reshaping Data for LSTM | The input data is reshaped to **[samples, time steps, features]**, as required by LSTM.  For example, (samples, 7, 1) means 7 time steps and 1 feature per time step. |  |  |  |  |  | | |
| Train-Test Split | Splits data into **80% training** and **20% testing** sets to evaluate model performance.  - Shuffling is avoided to maintain the sequence in time-series data. |  |  |  |  |  | | |

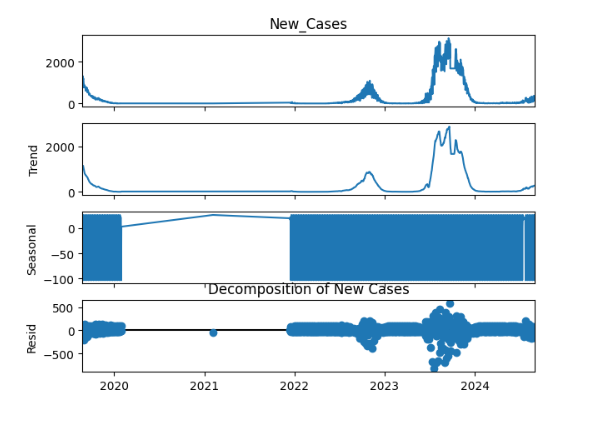


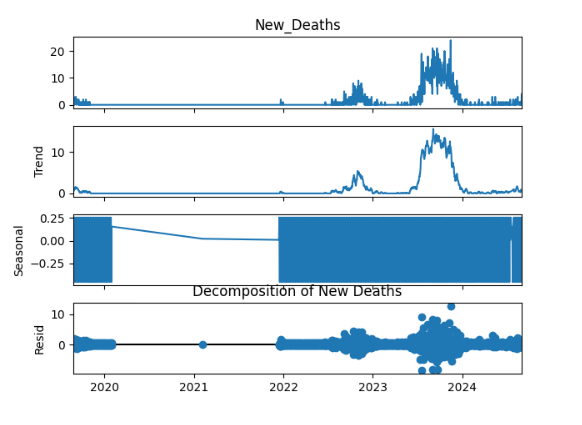


**Hybrid prpht, Xgiboost Sarima,Lstm with prophet:**

Model Applied: Seasonal Decomposition using Seasonal Decompose (Statsmodels)

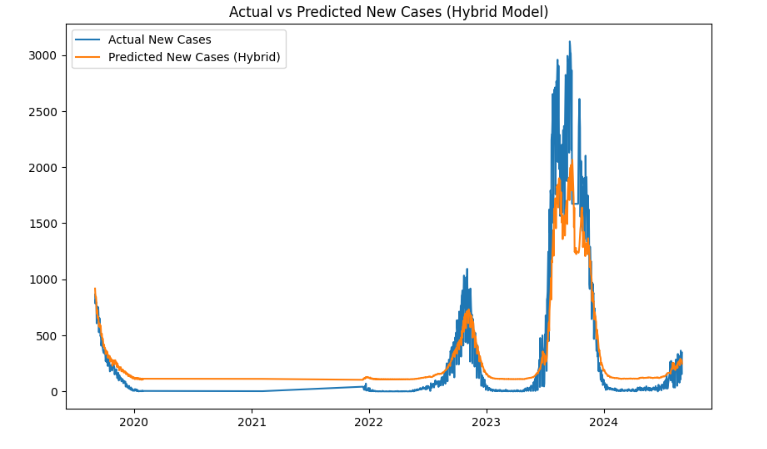
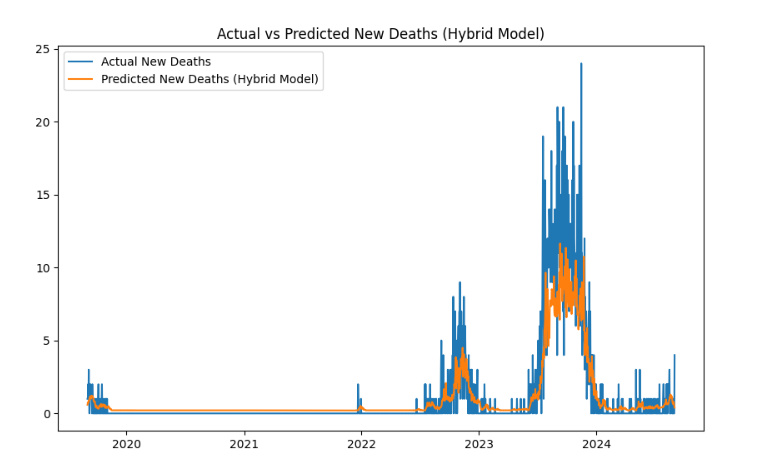
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| preprocessing | details |  | RMSE | MSE | MAE | | R2 |
| Date Parsing | Converts the Date column to **datetime** format, allowing the data to be indexed by date for time-series analysis | New cases |  |  |  |  | |
| Setting the Index | Sets the **Date** column as the index, ensuring the data is structured as a time series.. |
| Handling Missing Values | Although not explicitly handled here, any missing values in the time series could disrupt the decomposition and should be filled or removed beforehand. | New deaths |  |  |  | |  |
| Seasonal Decomposition | Decomposes the **New Cases** and **New Deaths** time series separately to analyze **trends, seasonality, and residual** |  |  |  |  | |  |





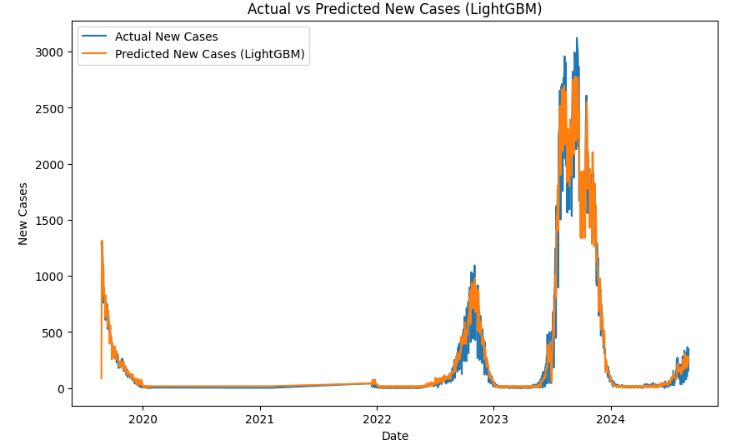
**There use multiple models(sarima,lstm,xgboost,ensambling average):**

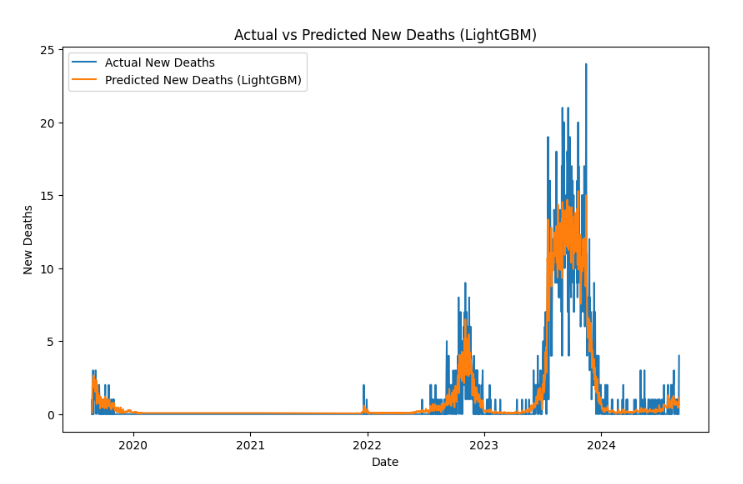
|  |  |  |
| --- | --- | --- |
| preprocessing | details |  |
| . Adding Time-based Features | - Converts the Date column to **datetime** format to ensure proper handling of time-series data. |
| Rolling Statistics | Computes **rolling mean and rolling standard deviation** with a 7-day window to smooth the data and capture short-term trends. |
| Differencing | Differencing is used to achieve **stationarity** in the data by subtracting the previous day’s value. This helps in reducing trends for SARIMA. |
| Scaling with MinMaxScaler | Scales New\_Cases and New\_Deaths to the range [0,1] for LSTM input. LSTM models perform better with scaled data. | Final RMSE New Cases: 271.47  Final RMSE New Deaths: 2.087 |
| Creating LSTM Dataset | Converts the scaled data into sequences with a **time step of 7 days** for input into the LSTM model. |
| Ensuring Prediction Alignment | Predictions from SARIMA, LSTM, and XGBoost are **truncated** to the minimum length to align them for ensemble averaging. |
| Training SARIMA Models | Fits **SARIMA models** to both New\_Cases and New\_Deaths to capture seasonal patterns and forecast future values. |
| Training LSTM Models | LSTM models are trained using **historical sequences** of 7 days to predict the next day’s value for both New\_Cases and New\_Deaths. |
| Training XGBoost Models | XGBoost models are trained on **lagged features (day of the week, month, rolling mean, rolling std)** to predict cases and deaths. |
| Ensemble Averaging | Final predictions for both New\_Cases and New\_Deaths are obtained by **weighted averaging** of the predictions from SARIMA, LSTM, and XGBoost. |

**** ****

LIGHTGBM:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | rmse | Mse | mae | R2 |
| New cases | 100.47 |  | 48.67 | 0.98 |
| New deaths | 1.26 |  | 0.66 | 0.88 |

****

****

**TFT:error dekhay..**