Forecasting Flu Cases with Time Series and Machine Learning

Introduction

Accurately forecasting flu cases is critical for public health planning and resource allocation. In this project, we analyze CDC ILITOTAL data—a time series dataset of scaled flu cases—to predict future trends. Our goal is to identify seasonal patterns, develop forecasting models, and provide actionable insights for stakeholders.

Problem Statement

The primary objective is to forecast flu activity to help public health officials anticipate seasonal peaks. By doing so, the project aims to:

- Identify periods of high flu incidence.
- Assist in timely allocation of medical resources.
- Enable proactive public health interventions.

Data Overview

- Data Source: CDC Flu Data (ILITOTAL)
- Time Period: Multiple years of weekly data
- Key Variable: ILITOTAL (Scaled Flu Cases)

The dataset was cleaned and preprocessed to ensure accurate modeling. Detailed data wrangling and exploratory data analysis (EDA) steps were documented in the Data Wrangling and EDA Notebook.

Data Wrangling and EDA

Data Wrangling:

- Datetime Conversion: Date columns were converted to a datetime index.
- Missing Value Handling: Incomplete records were addressed to maintain data integrity.
- Scaling: The ILITOTAL variable was normalized to ensure consistency across features.

Exploratory Analysis:

- Time Series Visualization: Plots revealed clear seasonal patterns, with peaks corresponding to typical flu seasons.
- Distribution Analysis: Histograms and boxplots were used to understand the variability and distribution of flu cases.
- Correlation Analysis: Additional features were examined (if applicable) to assess relationships with flu activity.
-

Feature Engineering and Preprocessing

Feature Engineering:

- Lag Features:
  - Created lag variables (e.g., `lag1`, `lag2`, and `lag52`) to capture both short-term and yearly seasonal patterns.
- Dummy Variables:
  - Generated categorical time indicators (e.g., week-of-year) to further model seasonality.
- Magnitude Standardization:
  - Applied z-score normalization to all features to ensure comparable scales.

For a detailed view, refer to the Preprocessing and Training Notebook.

Modeling Approaches

Three different models were developed to forecast flu cases:

SARIMA Model

- Approach:
  - Traditional Seasonal ARIMA model to capture trends and seasonality.
- Insights:
  - Effectively modeled seasonal peaks, but did not include external predictors.

SARIMAX Model

- Approach:
  - An extension of SARIMA that incorporates a dummy exogenous variable.
- Insights:
  - The dummy variable did not provide additional predictive power, resulting in performance similar to SARIMA.
- Model Summary:
  - Key parameters include AR(1) (0.3662), MA(1) (0.2804), and a significant seasonal MA term (-0.8029).

Linear Regression with Lag Features (ML Model)

- Approach:
  - Transformed the time series into a supervised learning problem using lag features.
- Performance:
  - RMSE: 0.286
  - MAE: 0.163

- Insights:
    - A simple model that provided competitive accuracy, serving as a strong baseline.

<u>Results and Evaluation</u>

Visualizations:

- Forecast Plots:
    - Historical data and forecasts from SARIMA and SARIMAX models are visualized with confidence intervals.
- Prediction Comparison:
    - A plot of actual vs. predicted values for the linear regression model highlights the model's performance.
- Residual Analysis:
    - Diagnostic plots confirm that the residuals are approximately white noise, indicating a good model fit.

Key Findings:

- SARIMA and SARIMAX models effectively capture seasonal patterns, as seen from the model summaries and diagnostic tests.
- The linear regression model, despite its simplicity, achieves low error metrics (RMSE = 0.286, MAE = 0.163) and offers a strong baseline.

<u>Future Work and Further Research</u>

Ideas for Further Research:

- Advanced Feature Engineering:
    - Explore additional lag features, interaction terms, or seasonal dummy variables.
- Integration of Exogenous Variables:
    - Acquire and integrate external datasets such as social mobility or real-time vaccination data.
- Model Experimentation:
    - Investigate more complex models such as TBATS, Facebook Prophet, or even neural network approaches (e.g., LSTM) for improved performance.
- Real-time Forecasting:
    - Develop a pipeline to regularly update the models with new data for near real-time forecasting.

Conclusion

This project demonstrates that effective forecasting of flu cases is achievable through careful data wrangling, exploratory analysis, and model comparison. The linear regression model with lag features, despite its simplicity, performs competitively and offers a practical tool for stakeholders. With further enhancements—especially the integration of meaningful exogenous variables—the forecasting system can be a valuable asset for proactive public health planning.