Time Series Regression Assignment

Ryoji Takahashi

December 3, 2024

1 Description of the Assignment and Objectives

The objectives of this time series regression assignment are:

1. To choose an open multivariate time series dataset (not financial) where the target variable does not exhibit seasonality (reference source).
2. To confirm the dataset's non-seasonality.
3. To select, train, and test machine learning/time series models that can predict 100 future time periods, assuming no regressor variable values are available beyond the prediction point.
4. To graphically present the predictions and evaluate the model performance using relevant KPIs.
5. To draw conclusions from the conducted analysis.
6. To deliver all necessary files for reproducing the analysis and deploying the trained models (DevOps integration).

1.1 Dataset Selection and Description

The Population Time Series Data was selected for this study. The dataset contains 816 rows and 3 columns of population records from 01/01/1952 to 01/12/2019. The target variable is the "value," representing the population size over time.
Given the dataset's small size, it was resampled to daily intervals, and missing values were interpolated. This preprocessing increased the sample size to 24,807, providing sufficient data for building prediction models.

1.2 Non-Seasonality Confirmation

To confirm non-seasonality, the following steps were undertaken:

- Seasonal decomposition: A decomposition plot was analyzed to ensure the seasonal component is flat or negligible.
- Augmented Dickey-Fuller (ADF) test: This test was applied to check for stationarity. A p-value > 0.05 indicates the series is non-stationary, though this does not directly confirm non-seasonality.

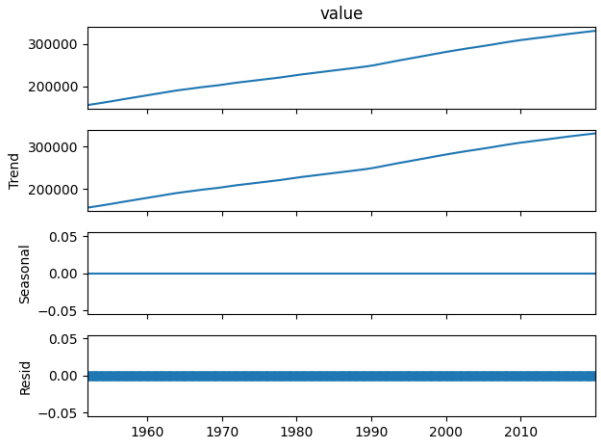

Figure 1: Seasonal Decomposition

Figure 1 shows the decomposition plot. The seasonal component is flat, and the ADF test results (ADF Statistic $= -0.742$, P-value $= 0.835$) confirm that the dataset is non-stationary and non-seasonal. Note that yearly interpolated data was used to generate the decomposition plot.

1.3 Trained Models and Methodology

The dataset was split into training (80%) and testing (20%) subsets. The following models were trained and evaluated:

- **SARIMAX (Seasonal ARIMA with Exogenous Variables):** A time series forecasting technique extending ARIMA to include seasonal effects and external regressors. It is well-suited for smaller datasets.
- **XGBoost with Optuna:** A gradient boosting framework optimized using the Optuna library for hyperparameter tuning. XGBoost is effective for structured data and handles missing values efficiently. While XGBoost has shown strong results in seasonal datasets, it struggled in this non-seasonal dataset. Further investigation is required to explore possible reasons, including XGBoost's sensitivity to non-seasonal data structure.
- **Long Short-Term Memory (LSTM):** A recurrent neural network architecture tailored for sequence predictions. Its ability to learn long-term dependencies makes it powerful for time series forecasting.

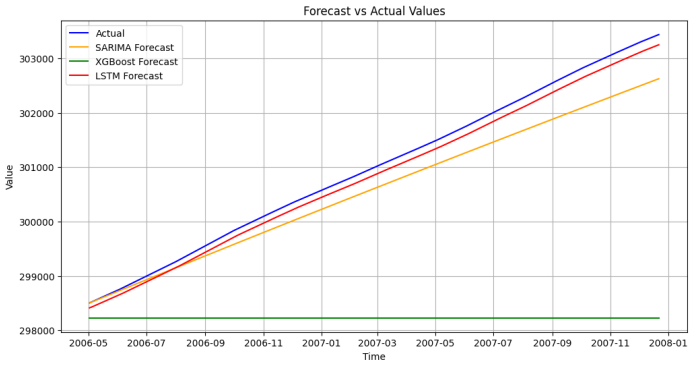

Figure 2: Prediction Results for 100 Future Periods (6 Days Each)

Figure 2 illustrates the predictions for 100 future time periods (6 days each).

Model	MSE	MAE	R^2
SARIMAX	218267.59	404.41	0.892
XGBoost	9682372.40	2766.88	-3.777
LSTM	20921.49	141.99	0.990

Table 1: Performance Metrics of Trained Models

As shown in Table 1, LSTM outperformed SARIMAX and XGBoost in accuracy but might have overfitted due to its complex architecture. SARIMAX displayed consistent performance, while XGBoost struggled. Further investigation is needed to determine how XGBoost can be adapted for non-seasonal datasets.

2 Conclusion and Future Work

The analysis highlighted that SARIMAX and LSTM are better suited for non-seasonal datasets, with SARIMAX being computationally efficient and LSTM excelling in accuracy for larger datasets.
Future improvements could involve:

- Addressing overfitting in LSTM models by incorporating dropout layers and early stopping.
- Investigating XGBoost's poor performance, particularly in non-seasonal data scenarios, while leveraging its proven effectiveness with seasonal data.
- Exploring real-time forecasting methods and generative AI for time series analysis.

Time series forecasting remains an evolving field, with exciting opportunities in automated model selection and real-time prediction for large-scale applications.