Time Series Regression Assignment

Ryoji Takahashi

December 1, 2024

1 Description of Assignments and Objectives

The time series regression assignments are

1. Choose an open multivariate time series dataset that is not financial and where the variable to be predicted does not have seasonality (reference source).
2. Check that the time series is non-seasonal
3. Select, train and test the machine learning/time series model(s) that allow: Predicting 100 time periods in the future with the particularity that from the moment where the prediction begins, the value of the regressor variables are not available
4. Graphically represent the results and evaluate the goodness of the model using the appropriate KPIs
5. Draw conclusions from the exercise carried out.
6. Prepare a deliverable with all the necessary files to reproduce the analysis and put the trained model(s) into production (DevOps integration).

1.1 Dataset

1. Dataset. household power consumption was selected for this study. It contains 2075259 rows x 8 columns of active power recodes from 16/12/2006 to 26/11/2010 by minutes, and I have selected "Global active power' measure by "date time" as the time series target. As in the code, data was resample to hourly data to reduce noise, and missing values were interpolated.

1.2 Identification of non-seasonal time series

To determine whether time series data is seasonal or non-seasonal, I have examined the following criteria:

- The seasonal component in the decomposition plot is flat or absent, it further confirms the dataset is non-seasonal
- The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots do not show repeating patterns at regular intervals,
- The Augmented Dickey-Fuller (ADF) test checks if the series has a unit root. A p-value > 0.05 indicates the series is non-stationary, but this does not directly imply seasonality.

Since there are cases that are stationary but it can be non-seasonal, the last ADF test is not directly address to the question of non-seasonality. However, it is worth to look at it.

First, as in Figure 1, the decomposition plot shows flat, and

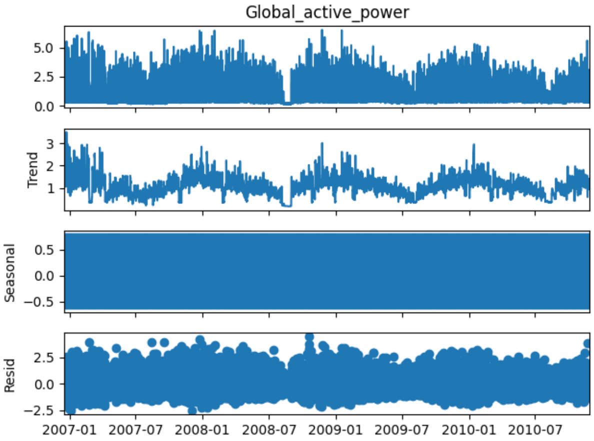

Figure 1: Seasonal Decompose

Figure 2 shows each plot does not have perfect repeating patterns at 240 intervals.

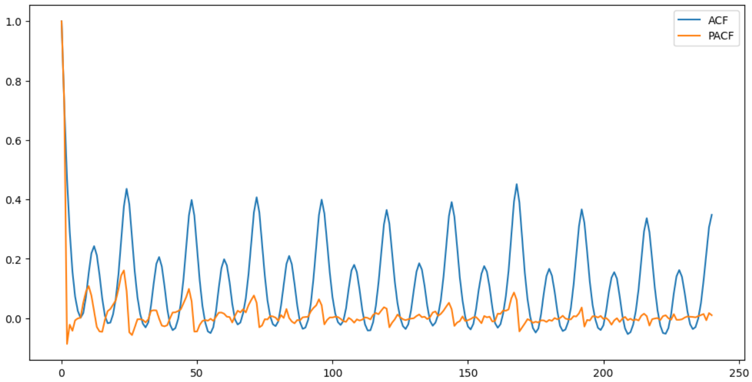

Figure 2: ACF, PACF

Finally, ADF test shows that ADF Statistic: -14.382566338396966 and P-value: $9.101197279691604e^{-27}$. These results confirm that the dataset is stationary and non-seasonal.

1.3 ML Models

I have trained from the beginning of 80 %, and tested and validated for the last 20 %, and deployed The Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX), XGBoost with optuna, and Long Short Term Memory (LSTM). SARIMAX model is a powerful time series forecasting technique that extends the traditional ARIMA model to account for seasonality and external factors. As I have seen the better performance in XGBoost with optuna, and also very quick to calculate, I have applied in here as well. LSTM is known to provide greater accuracy for demanded forecasts.
Figure 3 shows the prediction results, and I have utilized Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-Squared (R2) as KPIs.

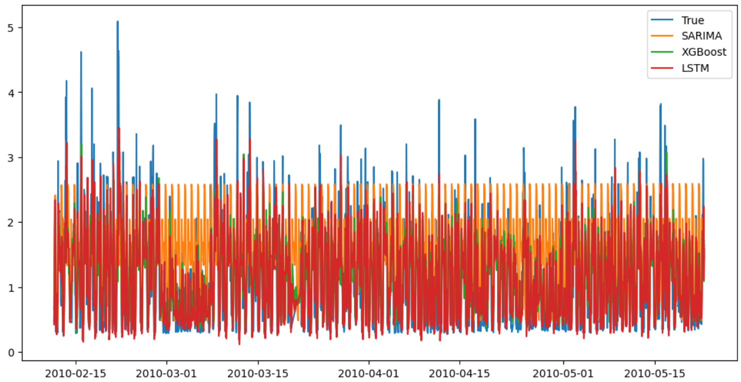

Figure 3: Prediction results of 100 time (24h) periods in the future

As in the assignment, I plotted 100 time 24 hours periods. It is hard to see the details. Therefore, I showed all results in the table 1. As I can see, XGBoost is performing slightly better than LSTM. This reason could be that, due to hardware limitations, I did not perform any hyper-parameter tuning in LSTM. If I performed hyper-parameter tuning, I would likely observe more accurate results than XGBoost.
Here, Figure 4 provides a clearer view compared to Figure 3.

Model Performance Metrics			
Model \ KPI	MSE	MAE	R^2
SARIMAX	0.629	0.598	-0.061
XGBoost	0.266	0.358	0.551
LSTM	0.290	0.376	0.510

Table 1: Comparison of Model Performance Metrics 100 time 24-hour periods

As shown, XGBoost, LSTM predictions are performing better than SARIMAX.

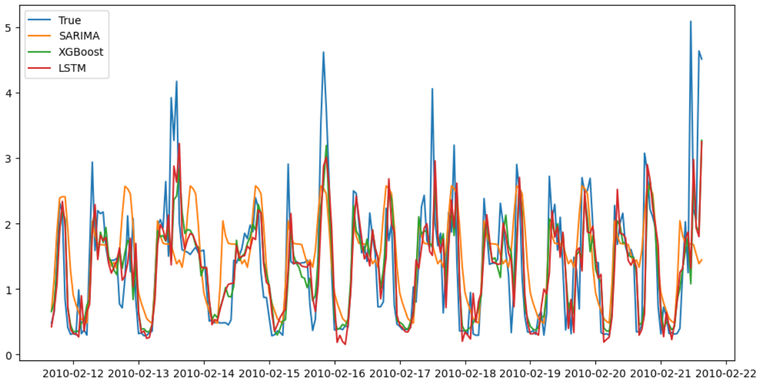

Figure 4: Prediction results of 10 time (24h) periods in the future

Model Performance Metrics			
Model \ KPI	MSE	MAE	R^2
SARIMAX	0.599	0.543	0.344
XGBoost	0.425	0.417	0.535
LSTM	0.459	0.439	0.497

Table 2: Comparison of Model Performance Metrics (10 times 24-hour periods).

I would like to point out that due to less data points, XGBoost, LSTM accuracies are slightly worse than previous results. However, SARIMAX is slightly better than the previous ones. This indicates that SARIMAX is for smaller time series datasets.
Finally, for delivery and reproducing, as in the code, I saved training models to pkl and keras files so that one can load these files and easily run tests.

2 Conclusion and Further Discussion

Time series analysis is continually evolving, driven by advancements in technology and the increasing availability of data. I aim to continue exploring automated model selection, nonlinear models, and machine learning with large datasets. Especially, real-time forecasting using generative AI is an intriguing area for further research.