ACEA Smart Water Analytics using Time Series Forecasting

Project Report

Soumyadeep Poddar

Introduction:

In this time series analysis project, we delve into the exploration and modeling of complex temporal data using Python. We begin by preprocessing the data, conducting visual analyses, and engineering features to capture patterns in the Aquifer Petrignano dataset. Through Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots, we identify potential autoregressive and moving average components. We then construct and train an ARIMA model and multi-layered LSTM recurrent neural network for time series forecasting.

Objective:

The objective of this project is to perform an in-depth exploration and analysis of time series data using advanced techniques. By applying Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) analyses, we aim to uncover inherent patterns and dependencies within the Aquifer Petrignano dataset. Furthermore, we strive to develop an ARIMA model and a multi-layered LSTM recurrent neural network for accurate time series forecasting.

Methodology:

1) Data Collection:
   - ✓ There are 9 different datasets taken from Kaggle .The ACEA Group deals with four different type of waterbodies: water springs(3), lakes(1), rivers(1) and aquifers(4). For our analysis, the Aquifer_Petrignano dataset is taken.
2) Data Preprocessing:
   - ✓ Handling missing values: Cleaned missing values by replacing them by nan values and filling them afterwards.
   - ✓ Smoothing data/ Resampling: Downsampling is done to extract additional information about the data.
   - ✓ Stationarity: Transformed the non-stationary dataset to stationary through methods like transformation and differencing.
3) Feature Engineering:
   - ✓ Encoding cyclical features:
   - ✓ Time series decomposition:
   - ✓ Lag:
4) Exploratory Data Analysis:

First, we plotted the seasonal components of the features and the correlation matrix of the core and the lagged features. We see the features are higher correlated in the case of shifted features (lagged ones) than the original ones.

- ✓ Autocorrelation Analysis: By plotting the autocorrelation and partial autocorrelation plots, we see that stationarity has been removed using the methods mentioned above.

5) Model Building and Evaluation:
   - ✓ First, we did Time series cross-validation which is a technique used to assess the performance of predictive models on time series data.
   - ✓ Then we performed, rolling window cross-validation which simulates how a model trained on historical data would perform when predicting future observations. By shifting the training and validation windows, this technique evaluates the model's performance over various time periods. It's particularly suitable for time series data where temporal patterns and trends matter.
   - ✓ Split the dataset into 85% for training and 15% for validation.
   - ✓ Prediction is done using ARIMA model and forecast is done for the next 90 steps.
   - ✓ Performed Auto-ARIMA to determine that best model is ARIMA (1,1,1).
   - ✓ Prediction is done using Multi-layered LSTM recurrent neural network model.
   - ✓ Evaluated RMSE for the two models.

Conclusion:

The project talks about stationarity and forecasting. We can see that LSTM performs better than ARIMA model.

Limitations:

It is important to note that this analysis is based on a specific dataset (Aquifer_Petrignano) and may not be representative of the other datasets on rivers, lakes etc.