LUT School of Engineering Sciences
Computer Vision and Pattern Recognition
BM20A6100 - Advanced Data Analysis and Machine Learning
Professor – & Teaching Assistant: Lasse Lensu and Akseli Suutari

Forecasting the electric power consumption for a house

Week 2 - Initial exploratory analysis

Saturday 8th November, 2025
Umme Tanjuma Haque
Chamath Wijerathne
Nada Rahali

Contents

List of Figures

1 Introduction

The goal of this work is to perform an initial exploratory data analysis, which will act as a foundation for the project's predictive modelling in the future. In this part, we are working on visualizing the data, analyzing patterns, and planning how to partition it for model calibration and validation.

For this task, we selected the Electric Power Consumption timeseries Dataset from the UCI Machine Learning Repository, which records minute-level household electricity usage from December 2006 to November 2010.

2 Initial Exploratory Analysis

As mentioned, the data set used for this project is the Electric Power Consumption Dataset from the UCI repository, which contains measurements at one-minute intervals of household electricity usage between December 2006 and November 2010. The data includes variables such as Voltage, Sub-metering 1-3, Global Active Power, and reactive power.

The single DateTime index was constructed by combining the Date and Time columns to create a continuous time series. Missing measurement values (1.25% of all timestamps) were handled through linear interpolation, since the timestamps themselves were complete and evenly spaced at one-minute intervals.

A new target variable was defined as:

$$y = \frac{\text{Global Active Power} \times 1000}{60} - (\text{Sub\_metering\_1} + \text{Sub\_metering\_2} + \text{Sub\_metering\_3})$$

representing the energy used by appliances not covered by the three sub-meters.

2.1 Observations

The minute-level data appeared noisy, as shown in Figure 1 below.

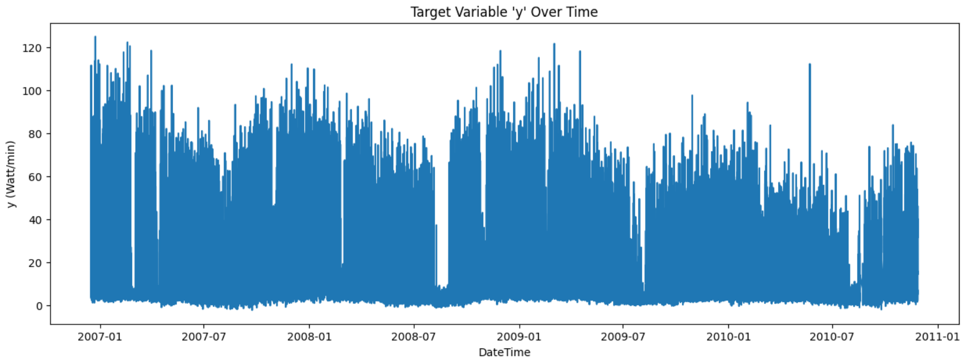

Figure 1: Minute-level energy consumption data

To better observe the patterns, the hourly-averaged data was plotted as shown in Figure 2, where some higher peaks can be seen during the morning or evening.

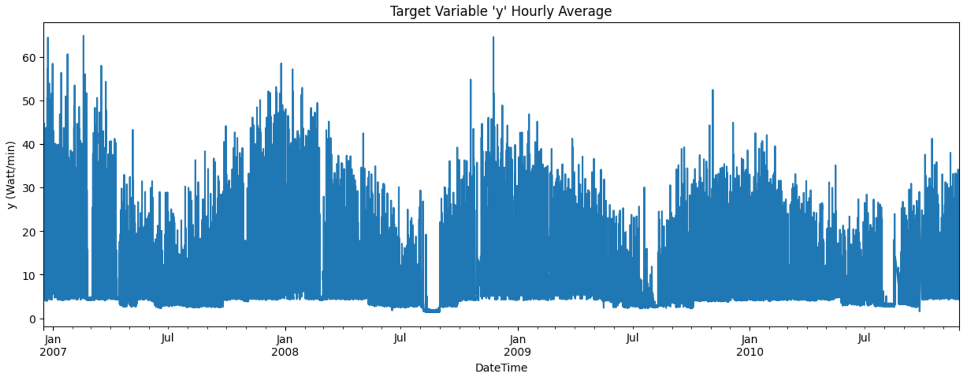

Figure 2: Hourly Average Energy Consumption Over Time

The daily averages, as seen in the plot in Figure 3, reach about 20- 30 Wh/min,and are highest in the winter (i.e., December to February, where likely the energy use was up due to winter heating, etc.) and lower in the summer months.

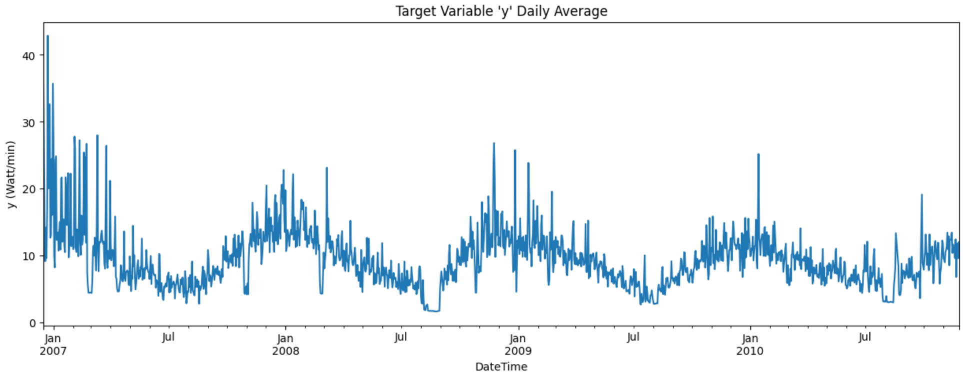

Figure 3: Daily average energy consumption over time

3 Time-Series Decomposition

To understand the structure of the data and patterns better, the Seasonal-Trend decomposition using the LOESS (STL) method was utilized on both the monthly and daily averaged data.

The decomposition follows:

$$y_t = T_t + S_t + R_t$$

where T_t is the long-term **trend**, S_t the **seasonal** component, and R_t the **residuals**.

3.1 Observations

- In terms of the trend, a slight downward trend is seen over the years, where the energy use decreased from about 11 to 9 Wh/min.
- In terms of the seasonality, there are clear patterns where, approximately, summers (hot weather) had lower energy use and winters (cold weather) had higher energy use, as also noted earlier.
- In terms of the residual component, there are irregular, small fluctuations showcasing random daily variations and occasional outliers.

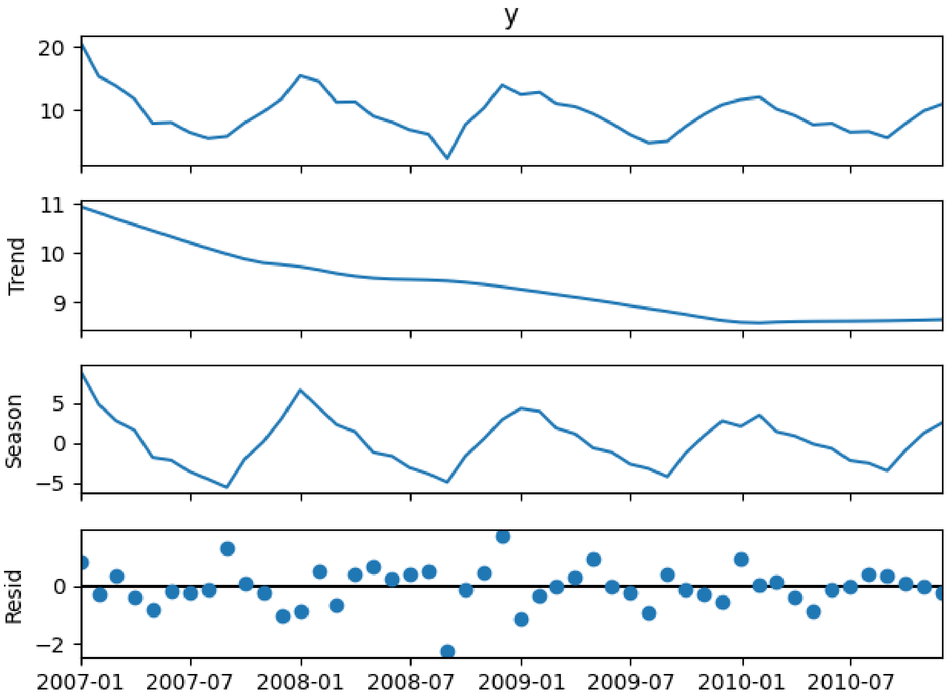

Figure 4: STL Decomposition of Daily Energy Consumption Time Series

4 Autocorrelation Analysis

This kind of analysis was done to essentially value how energy is being used currently in comparison to previous days.

The figure below illustrates 2 clusters:

- The smaller cluster showcases high-consumption days (approximately around 220–260 Wh/min)
- The larger cluster represents values 0 to 120, representing most of the data showing daily patterns.

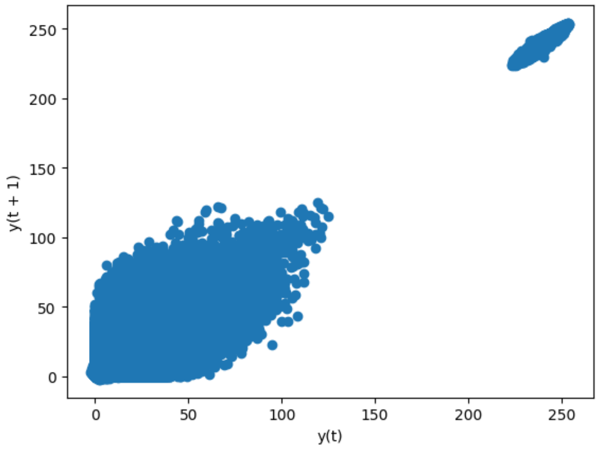

Figure 5: Autocorrelation Lag Plot for Daily Energy Consumption

The ACF (Autocorrelation Function) showed a significant positive correlation at lag = 1 (around 0.70), which gradually got degraded over longer lags, consistent with temporal dependence and repeating daily-seasonal behavior.

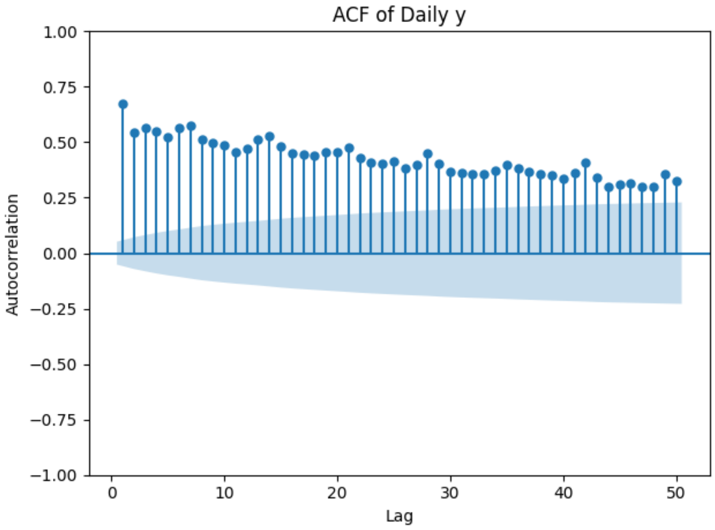

Figure 6: Autocorrelation of Daily y

5 Partitioning Strategy

Due to this dataset being a timeseries dataset, the partition strategy must follow a chronological order. This will help ensure the temporal order is maintained and data leakage is prevented.

- **Training set (60%)** - The earliest data (first 60% data) will be used for model calibration
- **Validation set (20%)** - The next 20% data will be utilized for hyperparameter tuning
- **Testing set (20%)** - The latest 20% data will be utilized for the final evaluation

This is also in alignment with how real-world forecasting works, where past data is used to make predictions for future data.