Household Electric Power Consumption Analysis

SRI LAKSHMI SUDHA GANNI

2024-06-08

# 1. Summary/Abstract

The Individual Household Electric Power Consumption dataset from the UCI Machine Learning Repository comprises minute-by-minute measurements of electric power consumption in a single household in France, collected from December 2006 to November 2010.  
  
It includes over 2 million observations, recording metrics such as:

1. Global active power
2. Voltage
3. Current intensity (Current in ampere)
4. Energy sub-metering – representing energy consumed in kitchen, laundry room, HVAC system and rest of the household.

This extensive dataset is valuable for time series analysis and energy consumption behavior modeling, aiding in the development of energy efficiency strategies.

In this project, we aim to explore the dataset, clean it, and perform a preliminary analysis to understand the data better. We will also investigate the relationship between energy consumption and other variables, such as seasonality and sub components of different parts of the house to identify patterns and trends. These trends can be used to offer recommendations in terms of power production, appliance purchases, tax incentives for efficient HVAC systems, etc.

# 2. Introduction

## 2.1 General Background Information

Decarbonization of electric power grid is an important aspect of mitigating the impacts of global warming. As part of this effort governments around the world are focusing on both the addition of renewable energy sources such as solar and wind to the grid as well as reduction of household emissions. This project tries to address the second aspect of energy efficiency through analysis of household electricity usage. Understanding trends in this dataset is crucial for developing strategies to enhance energy efficiency and reduce environmental impact.

The Individual Household Electric Power Consumption dataset provides detailed, time series usage data from a single household with a resolution of up to a minute. This allows for in-depth analysis of consumption patterns. This dataset supports analyses aimed at identifying factors that influence energy use, such as time of day or appliance operation. Analyzing this data helps in forecasting future energy needs and informing both consumers and policymakers on effective energy management practices.

## 2.2 Description of data and data source

The dataset, sourced from the UCI Machine Learning Repository, consists of minute-by-minute records of electric power consumption in a single household in France from December 2006 to November 2010.

It features over 2 million data points, including measurements such as

1. Global active power
2. Global reactive power
3. Voltage
4. Current intensity
5. Energy consumption by different household appliances in kitchen laundry room and HVAC.

These output variables can be combined with input features such as time of day and year to identify the impact of seasonality and time of day on energy consumption

## 2.3 Questions/Hypotheses to be addressed

The research questions for this analysis are:

1.How does household energy consumption vary throughout the day and across different seasons?

2.What are the main drivers of variations in energy consumption within the household?

3.How effectively can machine learning models forecast short-term and long-term energy usage?

# 3. Methods

Methods Description Data : Data Cleaning:

The data cleaning process involves: Removing entries with missing values. Converting date and time entries into a unified datetime format. Standardizing numeric values and handling outliers using statistical thresholds.

Analysis Approaches:

The analysis will utilize:

Time Series Analysis: To model and forecast energy consumption trends.

Regression Analysis: To identify key factors influencing power usage.

Clustering: To categorize consumption patterns into distinct profiles.

Machine Learning Models: Regression techniques for time series analysis and k means clustering to identify categories of energy consumption by season. These methods ensure a comprehensive understanding of household energy dynamics and facilitates the development of tailored energy-saving strategies.

## 3.1 Schematic of workflow

|  |
| --- |
| Figure 1: A Flowchart of the Proposed Model. |

## 3.2 Data aquisition

The dataset was obtained from the UCI Machine Learning Repository, specifically from their archive of machine learning datasets. It is publicly accessible online, allowing for direct download and integration into data analysis workflows. This data was originally collected to facilitate research on energy consumption patterns in households, providing a detailed, minute-by-minute log of electricity usage over a period of nearly four years. The dataset’s comprehensive nature makes it ideal for developing predictive models and exploring energy usage dynamics.

## 3.3 Data import and cleaning

Brief Overview of Data Cleaning Steps:

1.Data Import: Load the dataset into R using fread for efficient handling, especially given the large size of the dataset.

2.Handling Missing Values: Identify and replace or remove missing entries, typically represented as “?” in this dataset.

3.Data Type Conversion: Convert all pertinent columns to their appropriate data types, such as numeric for energy readings and datetime for timestamps.

4.Outlier Detection and Removal: Identify statistical outliers in energy consumption readings and decide on a strategy for handling these, either through removal or capping.

#Loading the libraries:  
  
library(data.table)  
library(ggplot2)

#Load the Data:  
data <- fread("household\_power\_consumption.txt", sep=";", na.strings="?")  
# Convert Date and Time into POSIXct  
#data[, datetime := as.POSIXct(paste(Date, Time), format="%d/%m/%Y %H:%M:%S")]  
# Convert other columns to numeric, handling NAs  
cols <- c("Global\_active\_power", "Global\_reactive\_power", "Voltage", "Global\_intensity", "Sub\_metering\_1", "Sub\_metering\_2", "Sub\_metering\_3")  
data[, (cols) := lapply(.SD, as.numeric), .SDcols = cols]  
  
# Remove rows with NA values  
data <- na.omit(data)  
  
# Remove outliers based on Global\_active\_power  
qnt <- quantile(data$Global\_active\_power, probs=c(.25, .75), na.rm = TRUE)  
caps <- quantile(data$Global\_active\_power, probs=c(.01, .99), na.rm = TRUE)  
iqr <- IQR(data$Global\_active\_power)  
data <- data[Global\_active\_power > (qnt[1] - 1.5\*iqr) & Global\_active\_power < (qnt[2] + 1.5\*iqr) & Global\_active\_power >= caps[1] & Global\_active\_power <= caps[2]]

## 3.4 Statistical analysis

For the Individual Household Electric Power Consumption dataset, statistical analyses could include:

Descriptive Statistics: Computing mean, median, mode, variance, and standard deviation to understand the central tendencies and dispersion in energy consumption.

Time Series Analysis:To model and forecast energy consumption trends.

Regression Analysis: To identify key factors influencing power usage.

Clustering: To categorize consumption patterns into distinct profiles.

Machine Learning Models: Regression techniques for time series analysis and k means clustering to identify categories of energy consumption by season. These methods ensure a comprehensive understanding of household energy dynamics and facilitates the development of tailored energy-saving strategies.

summary\_stats <- summary(data)  
print(summary\_stats)

Date Time Global\_active\_power  
 Length:1935294 Length:1935294 Min. :0.1100   
 Class :character Class :character 1st Qu.:0.3060   
 Mode :character Mode :character Median :0.5360   
 Mean :0.9438   
 3rd Qu.:1.4600   
 Max. :3.3560   
 Global\_reactive\_power Voltage Global\_intensity Sub\_metering\_1   
 Min. :0.0000 Min. :224.0 Min. : 0.400 Min. : 0.0000   
 1st Qu.:0.0480 1st Qu.:239.2 1st Qu.: 1.400 1st Qu.: 0.0000   
 Median :0.1000 Median :241.2 Median : 2.400 Median : 0.0000   
 Mean :0.1212 Mean :241.0 Mean : 4.002 Mean : 0.3695   
 3rd Qu.:0.1900 3rd Qu.:243.0 3rd Qu.: 6.000 3rd Qu.: 0.0000   
 Max. :1.2400 Max. :254.2 Max. :16.400 Max. :52.0000   
 Sub\_metering\_2 Sub\_metering\_3   
 Min. : 0.000 Min. : 0.000   
 1st Qu.: 0.000 1st Qu.: 0.000   
 Median : 0.000 Median : 1.000   
 Mean : 0.736 Mean : 6.106   
 3rd Qu.: 1.000 3rd Qu.:17.000   
 Max. :52.000 Max. :31.000