Phase-2 Submission – Data Analytics

Student Name: THARANI V
Register Number: 822423104082
Institution: M R K Institute of Technology
Department: Computer science and Engineering
Date of Submission: 29/04/2025
GitHub Repository Link: https://github.com/tharani1928/smartgrid-energy-forecasting?tab=readme-ov-file - smartgrid-energy-forecasting

---

1. Problem Statement

- The focus of this project is to predict energy consumption patterns using time series forecasting techniques for smart grids.
- In today's energy sector, efficient demand forecasting is crucial for maintaining grid stability, optimizing resource allocation, and minimizing operational costs.
- This analysis aims to address the operational need for accurate, data-driven energy demand predictions.
- By analysing historical consumption data, we can uncover patterns and trends that enable utility companies and grid operators to make informed decisions.
- These include balancing supply and demand, scheduling maintenance, reducing energy waste, and improving customer service through dynamic pricing models.
- The relevance of this analysis is rooted in its real-world applications: more accurate forecasting leads to more resilient and efficient energy systems, supports the integration of renewables, and enhances the sustainability of energy production and consumption.
- This project primarily falls under **predictive analytics** but also incorporates elements of **descriptive analytics**.
- Initially, we describe historical patterns to understand the structure of the data, and then we use predictive models to forecast future energy consumption based on these patterns.

2. Project Objectives

- The primary goal of this project is to develop an effective time series forecasting model to predict future energy consumption patterns within smart grids.
- By leveraging historical consumption data, the project aims to enhance the operational efficiency, reliability, and sustainability of energy systems.
- Key questions to be addressed through this analysis include:
    1. What are the underlying trends, seasonal patterns, and anomalies present in historical energy consumption data?
    2. How accurately can future energy demand be predicted using time series forecasting methods?
    3. What factors significantly influence variations in energy consumption over time?
    4. How can forecasting insights be utilized to support operational decision-making, such as load balancing and demand-side management?

- The expected deliverables from this project include:
    1. Identification and analysis of key consumption trends and seasonal effects.
    2. Development and validation of predictive models for short-term and long-term energy demand forecasting.
    3. Visualization of historical patterns and future projections to support strategic planning.
    4. Actionable recommendations for grid operators and policymakers based on forecasted consumption patterns.
- Initially, the objective was centred solely on forecasting future consumption. However, after a deeper exploration of the data, the objectives expanded to include a comprehensive analysis of historical consumption behaviours to improve model accuracy and to provide additional strategic insights.
- Thus, the project evolved to encompass both an analytical understanding of past patterns and the development of robust predictive models.

3. Flowchart of the Project Workflow

Problem Definition

↓

Data Acquisition

↓

Data Cleaning and Preprocessing

↓

Exploratory Data Analysis (EDA)

↓

Feature Engineering

↓

Model Selection and Development

↓

Model Training and Testing

↓

Forecast Generation

↓

Insights and Trend Extraction

↓

Visualization of Results

↓

Reporting and Recommendations

4. Data Description

For this project on predicting energy consumption patterns using time series forecasting in smart grids, a historical energy consumption dataset has been utilized.

- **Dataset Name and Source**:
  The dataset used is the "Household Electric Power Consumption" dataset, publicly available from the UCI Machine Learning Repository.

- **Data Type**:
  The dataset is **structured**, consisting of clearly defined rows and columns representing time-stamped energy consumption measurements.

- **Number of Rows and Columns**:
  The dataset contains approximately **2,075,259 rows** and **9 columns**. Each row corresponds to a minute-level measurement of power consumption over several years.

- **Static or Dynamic Dataset**:
  The dataset is **static.**

- **Key Fields or Attributes Relevant to the Problem**:
  - **Date**: The date when the measurement was recorded.
  - **Time**: The specific time of day corresponding to the measurement.
  - **Global active power**: The total active power consumed by the household (kilowatts).
  - **Global reactive power**: The reactive power consumed (kilowatts).
  - **Voltage**: The voltage measured at the time of consumption (volts).
  - **Global intensity**: The current intensity (amperes).

- **Sub_metering_1, Sub_metering_2, Sub_metering_3**: Energy measurements for specific household appliances or areas.

These attributes are critical for understanding overall consumption patterns, seasonal and daily load variations, and developing accurate forecasting models.

5. Data Preprocessing

To ensure the reliability and accuracy of the forecasting models, the dataset underwent several important pre-processing steps:

- **Handling Missing Values**:
  The dataset contained missing values, particularly in the power consumption and voltage fields. These missing entries were addressed by two strategies:
  - For small gaps, linear interpolation was applied to estimate missing values based on neighbouring timestamps.
  - For larger gaps, records were either imputed using domain knowledge or removed if reliable imputation was not feasible.
- **Removing Duplicates**:
  Duplicate entries were checked by identifying repeated timestamps. Any duplicated rows were removed to prevent bias or inaccuracies during model training.
- **Formatting and Parsing Data**:
  The **Date** and **Time** columns were combined into a single **Datetime** field, which was then parsed into proper datetime format using Python's pandas library. This step was crucial for setting the **Datetime** as the index for time series analysis.
- **Encoding Categorical Variables**:
  Since the dataset primarily consists of numerical data, no categorical variables required encoding. However, if any day-of-week or holiday indicators were later derived, these were encoded appropriately using one-hot encoding or cyclical encoding techniques (for example, encoding months and weekdays in a cyclical manner to reflect seasonal patterns).

- **Identifying and Treating Outliers**:
  Outliers were detected by visual inspection using box plots and by applying statistical methods such as the Z-score and IQR (Interquartile Range) method.
  - Extremely high or low energy readings inconsistent with household consumption patterns were flagged.
  - Some outliers resulting from recording errors were corrected when possible, or otherwise removed to maintain data quality.
- **Documented Transformations and Their Reasons**:
  - **Datetime Indexing**: Enabled time series-specific operations such as resampling and rolling analysis.
  - **Resampling**: In cases where minute-level granularity was too detailed for forecasting goals, the data was resampled to hourly or daily averages.
  - **Normalization**: Some forecasting models (e.g., LSTM) required normalization or standardization of input features to improve convergence and performance.
  - **Feature Extraction**: Additional time-based features such as hour of the day, day of the week, and month were extracted to capture seasonal and cyclic trends.

Each preprocessing step was carefully documented to maintain data integrity and ensure the reproducibility of the forecasting pipeline.

6. Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) phase was conducted to gain a comprehensive understanding of the historical energy consumption patterns and to identify trends, seasonality, and anomalies relevant to the forecasting task.

Univariate Analysis

- **Distribution of Single Variables**:
  - **Global Active Power**: A histogram and density plot were created to examine the distribution of active power consumption. The distribution was right-skewed, indicating that most households consumed moderate amounts of power with occasional high spikes.
  - **Voltage**: A histogram showed that voltage values were normally distributed around a mean of approximately 240V, with minor variations.
  - **Sub-metering Variables**: Each sub-metering measurement was visualized individually using line plots to understand appliance-level consumption patterns over time.
- **Seasonality and Trends**:
  - Time series decomposition was performed to separate overall trends, seasonality, and residuals from the **Global Active Power** variable, revealing strong daily and weekly seasonal components.

Bivariate/Multivariate Analysis

- **Correlation Heatmap**:
  - A correlation matrix heatmap was generated to visualize the relationships among features such as Global Active Power, Voltage, and Sub-metering values.
  - Results showed a strong positive correlation between Global Active Power and Global Intensity, and weaker, inverse correlations between Voltage and consumption metrics.
- **Pair Plots**:
  - Pair plots were used to explore the relationships between different energy attributes, particularly between the sub-meterings and overall power consumption.

- **Grouped Bar Charts and Line Plots**:
  - Energy consumption was aggregated by hours, days of the week, and months to identify cyclical usage patterns.
  - Grouped bar charts illustrated that consumption peaks typically occurred in the evenings and during winter months.

Analysis of Key Metrics or KPIs

- **Average Daily Consumption**:
  - The mean daily consumption was computed, highlighting weekdays with higher energy use compared to weekends.
- **Peak Load Periods**:
  - Analysis identified peak consumption hours (early morning and evening), which are crucial for load forecasting and grid balancing.
- **Seasonal Variations**:
  - Seasonal patterns showed that winter months generally had higher energy usage, likely due to increased heating needs.

Summary of Insights and Patterns Identified

- Energy consumption patterns exhibited strong **daily and weekly seasonality**.
- **Higher consumption** was recorded during evenings and weekends.
- **Voltage levels** remained relatively stable but showed minor dips during peak load periods.
- **Sub-metering analysis** revealed that certain appliances contributed significantly to peak loads.
- **Correlation analysis** indicated which features could be most useful for improving forecasting models.

- Some **outliers** in consumption were tied to unusual events (e.g., holidays), suggesting that external factors should be considered in model development.

These insights provided critical direction for selecting appropriate time series models and designing features to enhance forecasting accuracy.

7. Tools and Technologies Used

To perform the analysis, forecasting, and visualization of energy consumption patterns for smart grids, a combination of programming tools, libraries, and development environments were utilized. The key technologies employed in this project include:

- **Programming Language**:
  - **Python** was chosen for its extensive ecosystem of libraries supporting data analysis, time series forecasting, and machine learning.

- **Notebook/IDE**:
  - **Google Colab** was used for cloud-based development, allowing for easy access to computational resources and collaborative editing.
  - **Jupyter Notebook** served as the local environment for interactive coding, data exploration, and visualization.
- **Libraries**:
  - **pandas**: Used for data manipulation, cleaning, and transformation, particularly important for time series operations.
  - **numpy**: Applied for numerical computations and handling large datasets efficiently.
  - **matplotlib**: Utilized for creating basic plots such as line graphs, histograms, and bar charts to visualize consumption trends.
  - **seaborn**: Provided advanced statistical plotting tools, such as correlation heatmaps and pair plots.

- **plotly**: Enabled interactive and dynamic visualizations, including time series plots and dashboards to enhance data interpretation.

➢ **Optional Automation Tools**:

- **pandas-profiling**: Assisted in generating automated, comprehensive reports for preliminary data exploration, including distributions, correlations, and missing value analysis.

These tools collectively supported the end-to-end workflow, from data pre-processing and exploratory analysis to model development, evaluation, and visualization.

8. Team Members and Contributions

Name	Contribution
VAISHNAVI V	Data pre-processing
VITHYASHASINI S	Exploratory data analysis
YOGALAKSHMI V	Handling tools and technologies