Assignment

Task:

Building a predictive model to forecast electricity demand.

DataSet Used:

<https://archive.ics.uci.edu/dataset/235/individual+household+electric+power+consumption>

1.date: Date in format dd/mm/yyyy

2.time: time in format hh:mm:ss

3.global\_active\_power: household global minute-averaged active power (in kilowatt)

4.global\_reactive\_power: household global minute-averaged reactive power (in kilowatt)

5.voltage: minute-averaged voltage (in volt)

6.global\_intensity: household global minute-averaged current intensity (in ampere)

7.sub\_metering\_1: energy sub-metering No. 1 (in watt-hour of active energy). It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered).

8.sub\_metering\_2: energy sub-metering No. 2 (in watt-hour of active energy). It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.

9.sub\_metering\_3: energy sub-metering No. 3 (in watt-hour of active energy). It corresponds to an electric water-heater and an air-conditioner

Business Logic:

Electricity demand forecasting is a crucial aspect of energy management for utilities, grid operators, and energy providers. Accurate forecasting enables better planning, operation, and optimization of electricity supply, ensuring reliability and cost-effectiveness.

Helps in:

* Load Balancing: Forecasting helps in predicting the load on the power grid, allowing utilities to balance the supply and demand.
* Optimal Resource Allocation: Utilities can allocate resources more efficiently based on demand predictions.
* Energy Storage Management: Accurate forecasts enable better management of energy storage systems, such as batteries.
* Market Participation: In deregulated markets, accurate demand forecasting is essential for participating in electricity markets. It helps utilities and energy providers make informed bids and offers, enhancing their competitive position.

Research papers referred:

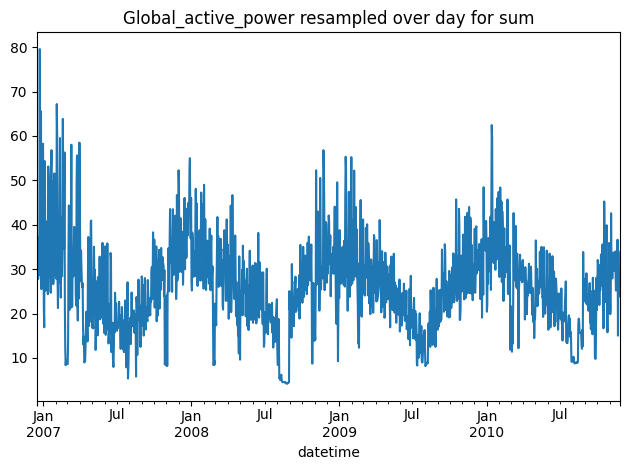
<https://www.mdpi.com/1996-1073/17/3/630>

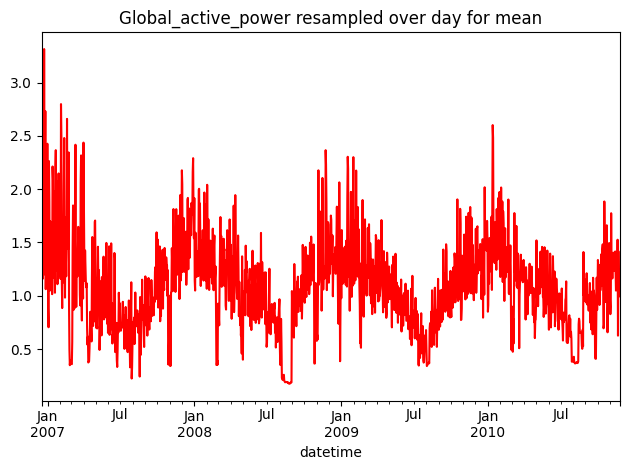
Preprocessing:

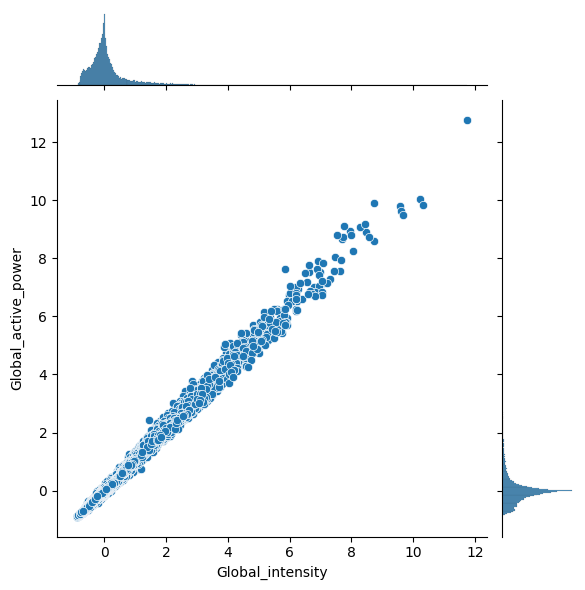
* Resampling the data on an hourly frequency basis to make it less noisy for us to work on.
* Made a function to identify null cells and then fill it with the mean of neighboring 10 values.
* Find out the total number of null values.
* Ran the fill missing value function, checked for any null values, and dropped any persisting ones.

Feature Engineering:

* I segregated the time date column into hours days and months.
* Made a new column that will store the consumption exactly 24 hours ago, as it will get us a good measure for accurate prediction.
* Made a column for rolling mean and standard deviation to smooth out short-term fluctuations( like sudden spikes).
* Electricity consumption may not always have a linear relationship with individual features like voltage and intensity. By creating an interaction term between voltage and intensity, we can capture non-linear effects that arise when these two features interact







Strong Positive Correlation:

* There is a strong positive correlation between Global\_intensity and Global\_active\_power. As the global intensity (which represents the current intensity) increases, the global active power (which represents the power consumption) also increases.

Linear Relationship:

* The relationship appears to be approximately linear for most of the data points. This indicates that Global\_intensity is a good predictor of Global\_active\_power and suggests that simple linear models might perform well for this dataset.

Outliers:

* There are a few points that are scattered away from the main trend line. These outliers could be due to sudden spikes or anomalies in power usage and intensity, or they could indicate measurement errors. These points should be investigated further to understand their cause.

Models Used:

Linear Regression

* Reasoning: Linear regression is a simple and interpretable model that assumes a linear relationship between the input features and the target variable. It is useful for understanding the basic trends and relationships in the data.

Decision Trees

* Reasoning: Decision trees are non-linear models that split the data based on feature values to make predictions. They can handle complex relationships and interactions between features, making them suitable for datasets with non-linear patterns.

Random Forest

* Reasoning: Random Forest is an ensemble method that builds multiple decision trees and averages their predictions. This reduces overfitting and improves generalization by leveraging the wisdom of the crowd.

Gradient Boosting

* Reasoning: Gradient Boosting is another ensemble method that builds trees sequentially, each one correcting the errors of the previous ones. This approach results in a strong predictive model that can capture intricate patterns in the data.

Support Vector Regression (SVR)

* Reasoning: SVR aims to find a function that deviates from the actual observed values by a value less than a specified threshold. It is effective in high-dimensional spaces and can handle non-linear relationships using kernel functions.

Used Mean Absolute Percentage Error (MAPE) as a performance metric.

Linear Regression:

MAPE: 0.0187

Linear regression shows a low MAPE, indicating that it performs quite well on this dataset. Given the linear relationship observed between Global\_intensity and Global\_active\_power, it makes sense that linear regression would be effective.

Decision Tree:

MAPE: 0.0244

The decision tree has a higher MAPE compared to linear regression. Decision trees can sometimes overfit the training data, especially if not properly pruned, which might be contributing to its lower performance.

Random Forest:

MAPE: 0.0173

Random forest, an ensemble method, has the lowest MAPE among all the models. This indicates that the model captures the complexities in the data better than a single decision tree, thanks to its ensemble nature, which reduces overfitting and improves generalization.

Gradient Boosting:

MAPE: 0.0190

Gradient boosting also performs well, with a MAPE close to that of linear regression. It is generally good at handling various patterns in the data due to its iterative boosting approach. However, it is slightly outperformed by random forest in this case.

Support Vector Regression (SVR):

MAPE: 0.0731

SVR has the highest MAPE, indicating it performs the worst among the models tested. This could be due to its sensitivity to the choice of kernel and parameters, and it may not be as well-suited for this particular dataset or problem compared to the other models.

