Assignment: Energy Consumption Forecasting and Uncertainty Quantification

Phase 1: Data Preparation & Time Series Exploration

1. **Loading data and cleaning dataset**

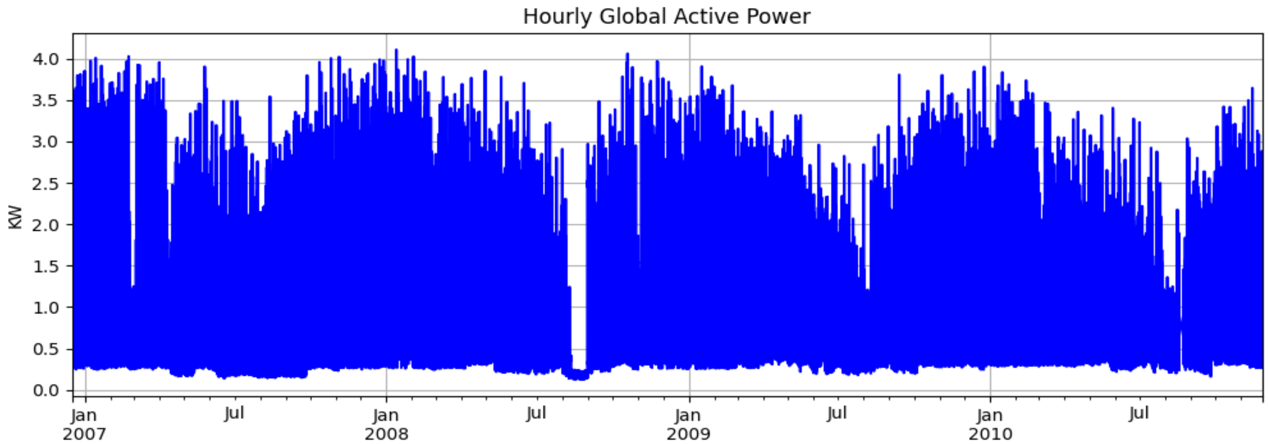
   - combining date feature with time feature to use it as index in future operation and for simplify the time series processes.
   - we noticed some values has “?” which indicates for no value, also we checked how many rows with partial empty or full empty regarding to their columns.
   - We have 25979 fully empty rows in this dataset; we did interpolation for the empty period rows with less than 1 hour and keep other empty rows as NaN. By doing this we keep regular time intervals and we don’t make interpolation for large gap periods avoiding bias in the dataset
   - We applied Z score to handle the outliers where it simple and fast then we interpolated the outliers when they were addressed

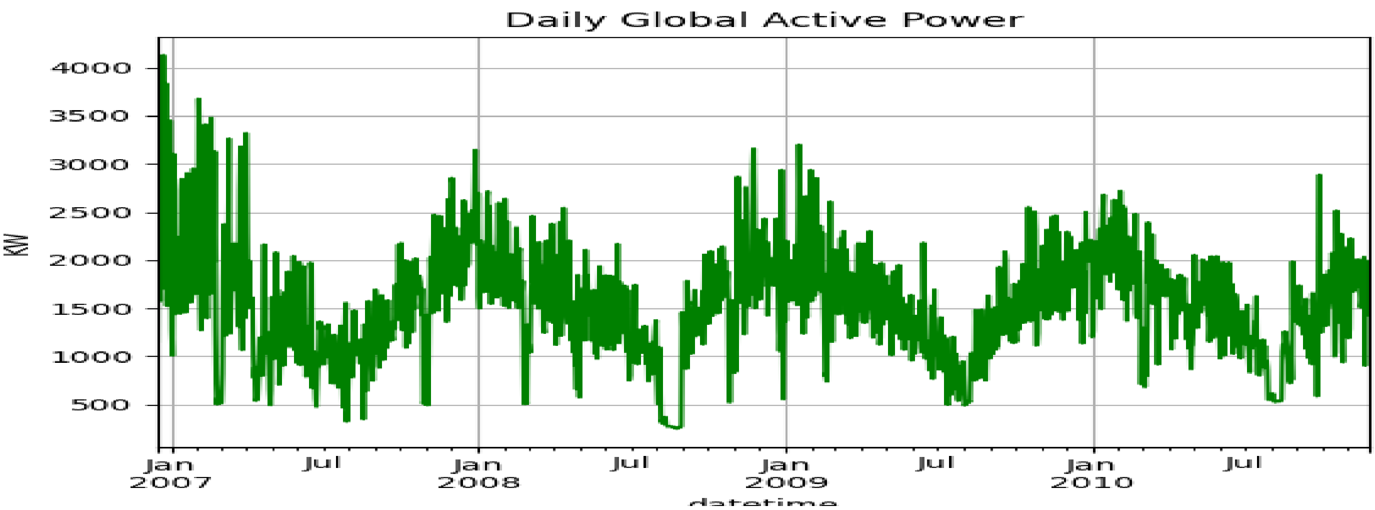
   I got these results:
   Number of values interpolated: 2016
   Number of gap blocks interpolated: 434
   Number of outliers detected and addressed: 217962

   Challenges: this dataset contains 25979 fully empty rows and most of them shape a time gap which exceeds more than one hour . This will effect on the future operations, we will try to reduce their impact on the models training and evaluation

2. **Temporal Aggregation & Multi-Frequency Views**

   - We did hourly aggregation using mean function for all features to explore the average load, intraday patterns and behavior, for example see the different consuming power rates between the morning, afternoon, and evening.
   - We did daily aggregation using the sum function (1442 days) for all features to recognize energy usage during workdays and the weekend, except voltage and intensity we used mean function to give meaningful voltage and current trends
   - We did weekly aggregation using sum function (207 weeks) for all features to recognize the energy consuming patterns in different seasons and months, except voltage and intensity we used mean function to give meaningful voltage and current trends


Hourly Global Active Power
KW
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Jan
2007
Jul
Jan
2008
Jul
Jan
2009
Jul
Jan
2010
Jul

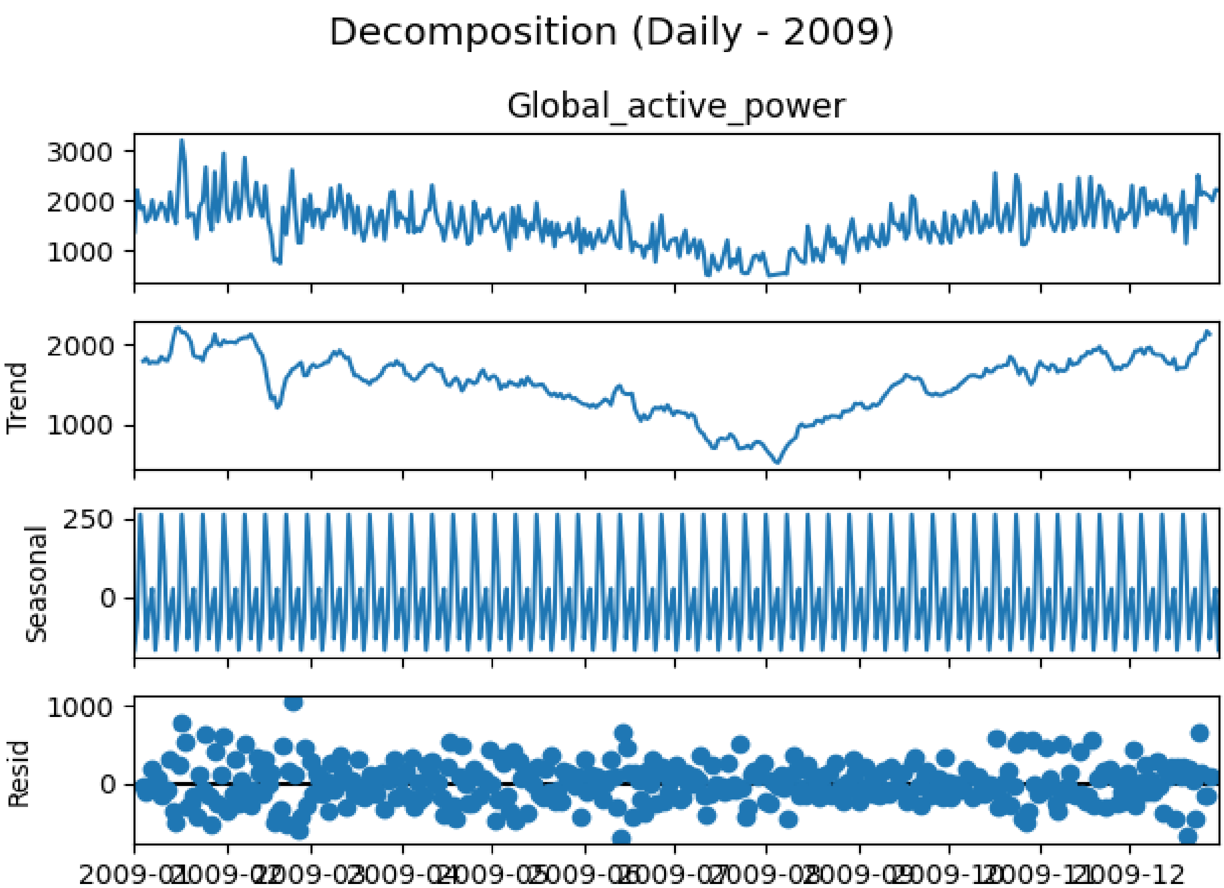

Daily Global Active Power
KW
4000
3500
3000
2500
2000
1500
1000
500
Jan
2007
Jul
Jan
2008
Jul
Jan
2009
Jul
Jan
2010
Jul

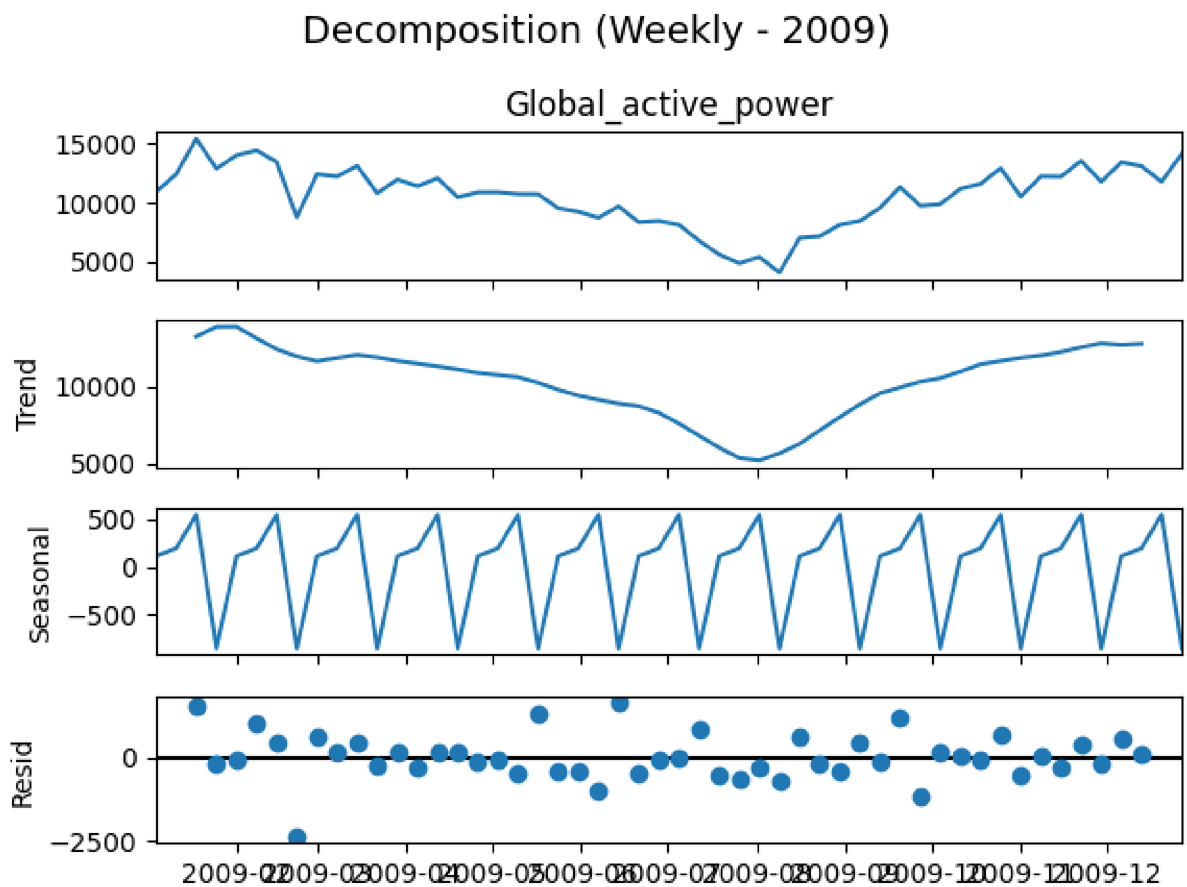

Weekly Global Active Power
KW
17500
15000
12500
10000
7500
5000
2500
Jan
2007
Jul
Jan
2008
Jul
Jan
2009
Jul
Jan
2010
Jul

- We saved these aggregation results in deferent csv files and visualize “Global_active_power” feature to make sure we have real variation in these deferent aggregations as follows:

3. Exploratory Time Series Analysis

- We choose daily aggregation data to decompose the time series to uncover trends, seasonality, and residuals, where it has high frequency fluctuations and we can use it for weekly cycles and trends that are meaningful for forecasting. While hourly data are heavy for decomposition over long periods.
- We use model='additive' and period=7 to explore weekly patterns on daily data and period = 4 for weekly data to explore the monthly pattern.
- We applied the previous composition on the year of 2009 as time window, we choose one year so we can visually notice the trend and seasonality





a. Trend Plot

- From trend plot for both 7 days (weekly) trends and 4 weeks (monthly) trends we can see the global active power consumption is gradually decreasing from week to week starting from January to July which has the lowest power consumption then it is gradually increasing from August to December. possibly due to higher occupancy in winter season where heating consuming more power or increased indoor activity, while in summer they may have vacation or have less appliance.

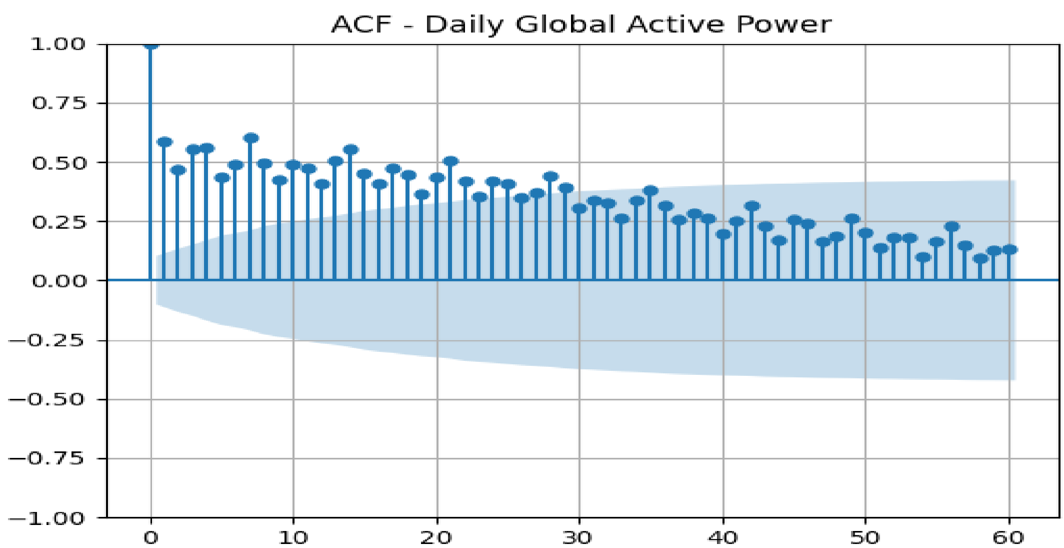
b. **Seasonal plot**

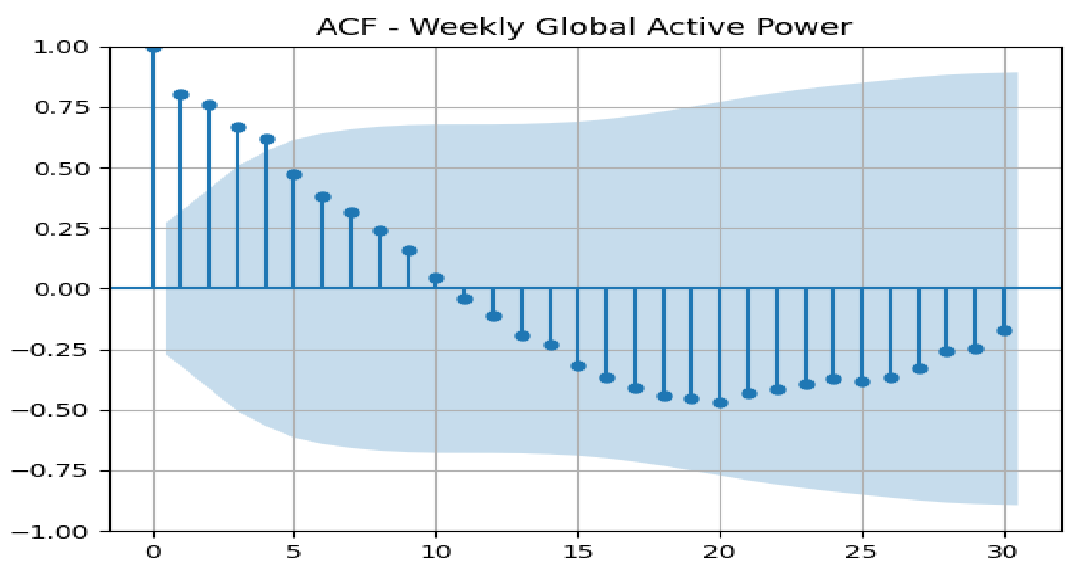
- From the 7- days plot There is a repeated weekly pattern, and when I returned to which days have the most peak according to the dates in the csv files they were Friday, Saturday, and Sunday. This give predictable results
- From the 4- weeks plot there is also a repeated monthly pattern, the amplitude is larger than weekly pattern which indicate for more important variation over the month. So the monthly pattern are stronger and may indicate for intramonth habits that repeated strongly.

c. Residual plot

- The 7- day (weekly) plot There is no constant pattern for outliers, they are scattered randomly. Some high residual at the beginning and the end of the year.
- In monthly residual result show less noisy and gathering around 0 , this show the monthly level is more stable than the weekly level.

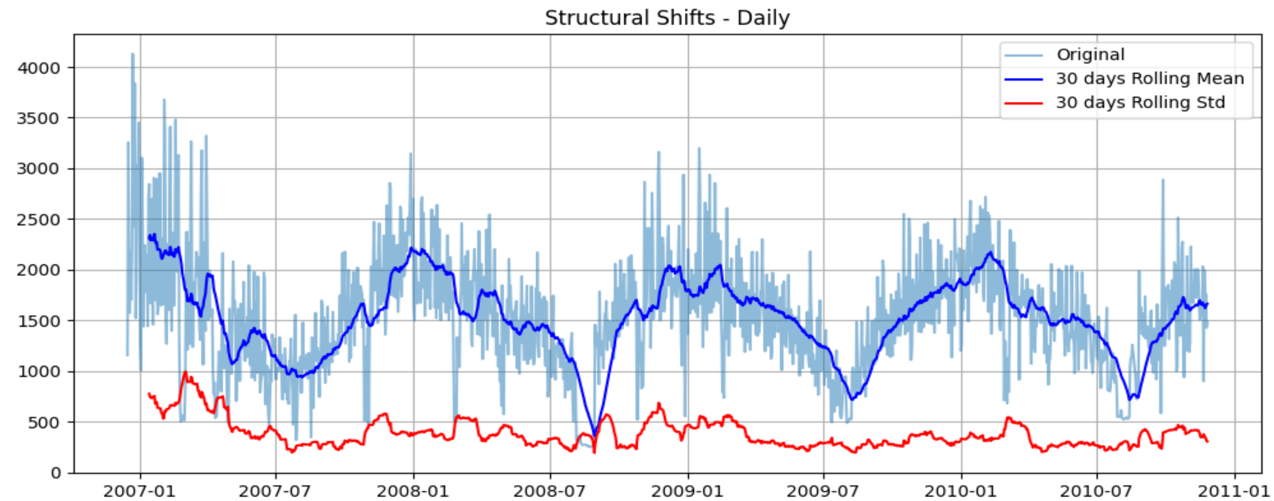
d. **Autocorrelation**

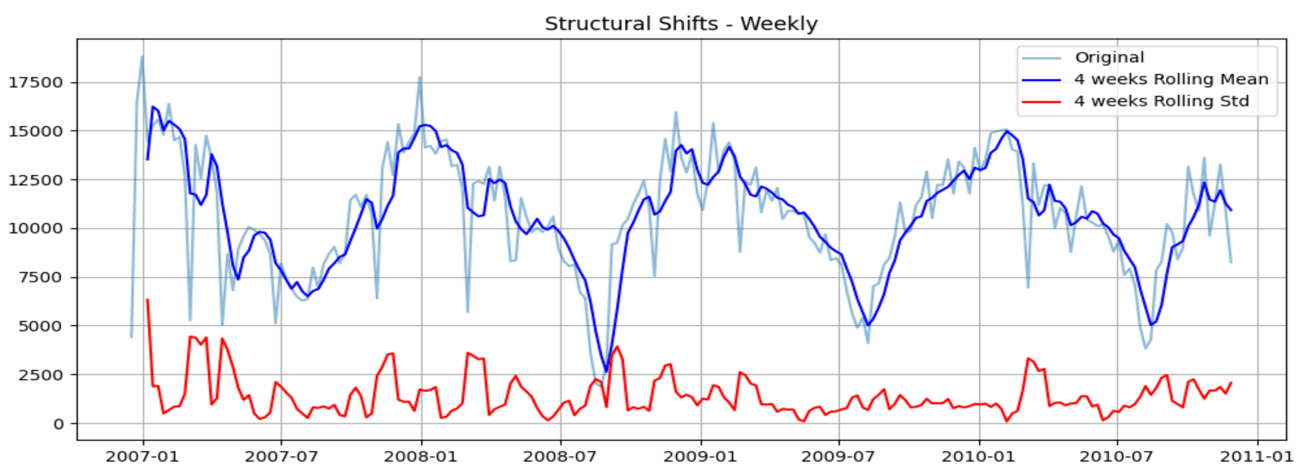




- Daily ACF indicate to have autocorrelated across many lags where each day is influenced by the previous several days.
- Weekly ACF also indicate to seasonal patterns with longer cycle

e. Structural shift





- for the structural shift for both daily and weekly in 2009, the rolling mean decreasing until reach august then start increasing to reach almost the same amplitude. It has U-shaped consumption pattern in 2009 Indicates seasonal structural shift , and it refer to Lower energy use in summer and Higher use in winter because of heating or end-of-year activity

- The rolling std is fluctuating up and down along the year within the same range. almost the consumption variability is stable throughout the year. predictable fluctuations
- the daily and weekly original have fluctuation, but it follows rolling mean

f. Forecasting relevance

- daily forecasting and monthly forecasting are practical and will give meaningful results according to the previous analysis and noticeable patterns
    - models like SARIMA or SARIMAX with seasonal order of 7 or 28
    - Prophet (auto-detects multiple seasonalities)
    - LSTM/XGBoost with lag and seasonal feature engineering
- For weekly forecasting anomaly detection and external features are considered, while monthly modeling should perform reliably, where it less affected by outliers.
- For lags of up to 7–14 days it good to use ARIMA/SARIMA, while monthly forecasting using seasonal decomposition like SARIMA or Prophet with monthly seasonality. Also LSTM, XGBoost can be used where no sharp cutoff in ACF
- Seasonal structural shifts require that model consider seasonality, specially mid-year.
- The stable variance allows us to use ARIMA or SARIMA.

Phase 2: Feature Engineering & Forecasting

1. Feature Design

- We chose the daily dataset for feature engineering
- We extracted new features relating to time-based to capture daily and weekly patterns
    - day_of_week → 0 = Monday - 6 = Sunday,
    - month → 1-12
    - day → 1-31
- we added the lag features to predict the future where time series has temporal dependencies
- we added Rolling Statistics (mean, and std) to capture short term trend and reduce noise
- we derived Interaction Terms using day_of_week * month to make the model learn conditional patterns
- we added Holiday Flags where power consumption has big variation in holidays
- we added weather data for every day for the city of Sceaux, France, where it was downloaded from this site https://open-meteo.com/en/docs for the period that covered in the daily aggregation dataset. We added three type of temperature (min, max, mean).
- We saved the final results in featured_energy_data.csv

2. Train-Test Strategy

- we have four years of daily data, I chose time-based split with 3 month forecast horizon, because it is a good period to be covered and predicted based on our data type which is the power consumption. this period give clear idea for power consumption for the next season like fall and winter and put the business in less risk when we could predict this time of period. In the same time, it is not too long and too short and it is suitable for modeling based on the previous analysis of the trend and seasonality.
- Model will be trained on data from 12/16/2006 to 8/31/2010 to cover the period from Sep to Nov 2010 as test set.

3. **Modelling**

We are the following models

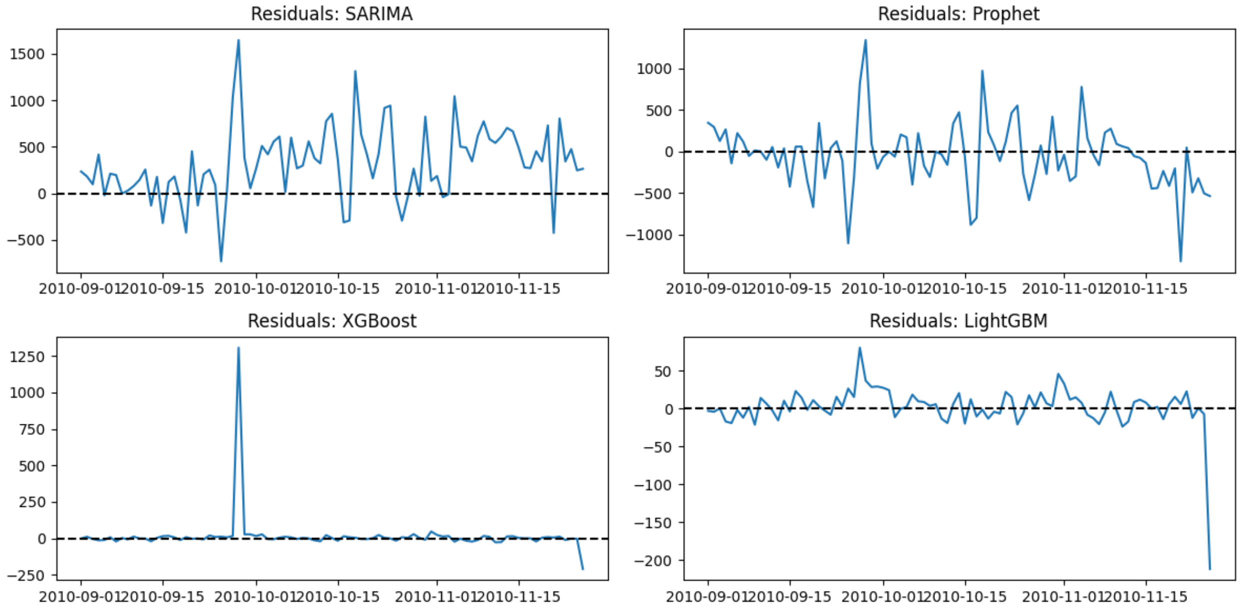
1. SARIMA as statistical method to capture seasonality for global active power
2. XGBoost as ML model to learn from the features engineered like temperature, lag, rolling statics)
3. Gradient Boosting as Ensemble method to combined models to improve the performance.
4. Prophet as probabilistic method to forecast with uncertainty estimation

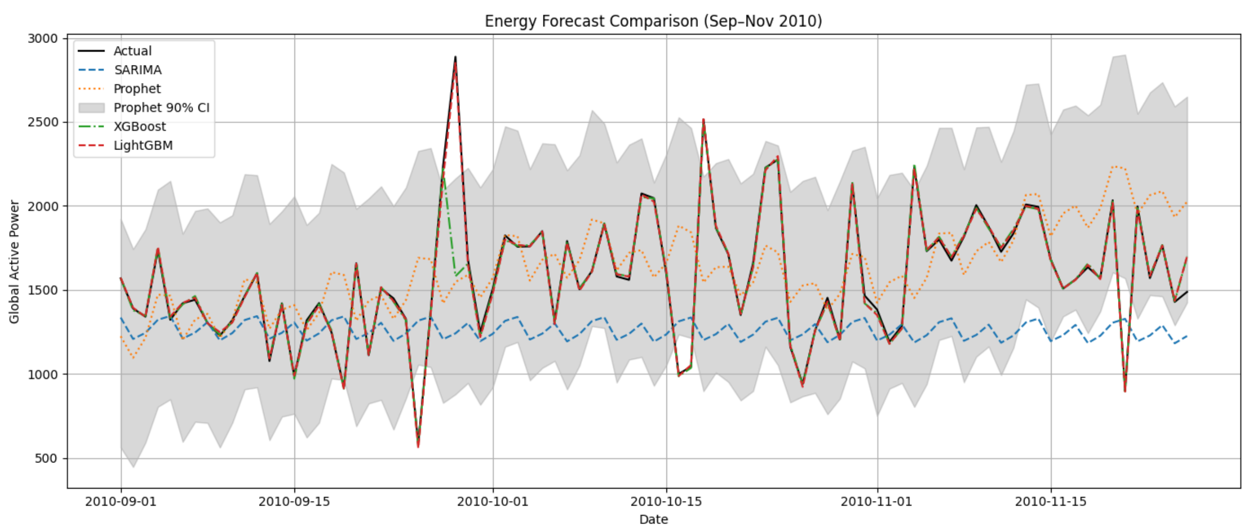
4. **Evaluation**

a. **Evaluation Results** metrics to assess forecast accuracy and interval quality, they were printed in the terminal

SARIMA | MAE: 394.340 | RMSE: 500.452
Prophet | MAE: 288.596 | RMSE: 403.506
Prophet 90% CI Coverage: 89.66%
XGBoost | MAE: 28.244 | RMSE: 142.569
LightGBM | MAE: 15.405 | RMSE: 28.709

b. **Visualize forecasts and analyze residuals.**





c. Model comparisons

- **Forecast comparisons**
    - SARIMA: it has MAE: 394.3 and RMSE: 500.5 which indicate to have lowest accuracy between the models where it underfits spikes and seasonality it has high residual return to high bias or variance.
    - Prophet: it has high residual also but less than SARIMA with MAE: 288.6 and RMSE: 403.5, it has uncertainty estimation with 89.66% CI, but it misses sharp changes to capture spikes and dips.

- XGBoost is better than SARIMA and Prophet regarding the residual stability and accuracy where it has MAE: 28.244 and RMSE: 142.569, the residual plot shows that it has one outlier but it could be sensitive to the outliers.
- LightGBM: best accuracy overall with MAE: 15.405 and RMSE: 28.709, and it has low and stable residual which makes it great for high performance tasks. It follows values even at spikes and dips

- **Summarizing trade-offs between accuracy, interpretability, and uncertainty**
  - SARIMA and Prophet have high interpretability with low accuracy. SARIMA consider as statistical model where we can see its components , it is not black box, and its parameters have clear meanings which represents a specific type of temporal dependency. In addition, we can manually monitoring coefficients to understand how model responds to pas values. Prophet also interpretable, where it is additive model components and we can see the trend and seasonal effects over the time. On the other hand, XGBoost and LightGBM has higher accuracy because they learn from the additional features more nonlinear relationships that increase the performance in the models, in other words it has more ability to see hidden patterns. But XGBoost and LightGBM are difficult to explain and they have low interpretability where they are need external tools to understand either for power analyst or the stakeholders like SHAP. Regarding the uncertainty, only Prophet model can handle uncertainty to identify the range of future values that has for example 90 % confidence value, this feature actually increase the level of interpretability to help power analyst for more future options which reflects on risk management, but accuracy still not the best characteristic in this model.

d. Actionable Insights & Risk Assessment

- For energy demand management LightGBM has the highest forecast accuracy, so it will be suitable for daily operational planning and purchasing and we can optimize generators scheduling based it prediction. Which also could reduce the operational costs
- Prophet provides uncertainty interval which make it better choice for risk awareness, for example upper confidence value could be used for worst case of power consumption coverage. SARIMA also with prophet can provie better resource allocation and planning
- Regarding Model Reliability XGBoost has good performance but can be sensitive for large outliers. So we can use alert system when we have low forecast confidence. We can use Prophet with XGBoost as hybrid system where LightGBM for short term and Prophet fro long term planning.

e. Professional Communication

- **For non-technical audiences**

For example, we can say that "*to predict energy consumption for the next three months we used AI models we have model called LightGBM can predict power usage with lowest error, this model is very good to plan accurately for fuel volume, cost and how many KW we need". also, we have other forecasting model called "Prophet" it can provide us with how much uncertain level of the forecasts do we have, this* *helps us to build our expectation on worst scenario and when do we need to change our supply strategy to reduce or avoid shortfalls or blackouts"*.

- **For technical audiences**
  According to the evaluation results, *LightGBM perform better than other models with MAE =15.4 and RMSE = 28.7 which make it fit* for daily operational and cost planning *while prophet provide* uncertainty estimation with 98.66% CI coverage make it better for risk awareness. From the residual analysis SARIMA shows underfits, while XGBoost has large outliers and affected more. So we can build multi hybrid system based on the analysis results to cover different requirements regarding forecasts, uncertainty, risk management, operational planning, power load usage, and reducing cost, considering different time intervals to get better recommendations and power management performance.