PA II - PROJECT

FORECASTING ELECTRICITY POWER CONSUMPTION

Ayush Aggarwal, Rishabh Setty,
Vedant Paithane, Rohit Sharma

Github Link: https://github.com/ayush9818/Electricity-Power-Prediction/tree/main

Introduction to the Raw Data:

The dataset under examination originates from a CSV file titled 'electricity_data_resampled.csv', which catalogs the minutely electric power consumption within a household. This compilation of data comprises various metrics such as 'Global_active_power', 'Global_reactive_power', 'Voltage', 'Global_intensity', and three separate sub-metering readings, each indicative of energy consumption in distinct sections of the domicile. The data is time stamped with a 'DateTime' index, reflecting the moment each measurement was captured. With records commencing in 2006, the dataset provides an extensive temporal snapshot of energy usage, with each entry aggregating information over a specified duration.

Fitting Models to the Data Array:

The preprocessing journey involves several key transformations to render the data suitable for predictive modeling. The features were first normalized using 'MinMaxScaler' to bring all variables to a comparable scale, a crucial step for many machine learning algorithms. Further, a feature engineering step was incorporated to calculate 'sub_metering_rem'—a representation of the remaining energy consumption not captured by the existing sub-metering features. This was computed by converting 'Global_active_power' to total energy in watt-hours and subtracting the energy accounted for by 'Sub_metering_1', 'Sub_metering_2', and 'Sub_metering_3'.

Row-by-Row Explanation and Variable Roles:

Each row in the preprocessed dataset encapsulates the daily readings of electrical consumption, with variables providing insights into both overall and specific energy usage areas within the household. The 'Global_active_power' is indicative of the total energy used, minus the sub-metered sections, while 'Global_reactive_power' reflects the energy that does not perform any work but is necessary for the reactive components of the load. The 'Voltage' and 'Global_intensity' are indicative of the electrical supply quality and the overall electrical consumption intensity, respectively. The sub-metering variables offer a detailed breakdown of energy usage, with 'sub_metering_rem' now adding an additional dimension to the dataset by accounting for unspecified consumption.

Processed Data and Rationale Behind Train/Test Split:

The data preparation stage is meticulously carried out by the function 'prepare_data_for_training', which iterates through the dataset to generate overlapping seven-day sequences, paired with the corresponding target values. The outcome is a

three-dimensional array, which aligns with the input requirements for advanced time-series models. The train/test data division adheres to a weekly demarcation, allocating 70% of the weeks to training and reserving 30% for testing. This temporal division is strategic, aiming to evaluate the model's predictive prowess on data that emulates a real-world 'future' scenario, as opposed to data observed during the training phase. Such a temporal split is critical in preserving the chronological order and dependencies, which is fundamental for the integrity of time series forecasting. This methodology is a guard against overfitting and offers a more accurate assessment of the model's performance on previously unseen data.

In conclusion, the preprocessing steps undertaken are essential in shaping the raw data into a form that is not only model-ready but also retains the temporal patterns necessary for effective forecasting. The detailed feature engineering and careful train/test split underscore our commitment to creating a robust predictive model. The next stage, the Exploratory Data Analysis (EDA), will build upon this foundation to explore the data further, uncovering patterns, trends, and anomalies that will guide our modeling strategy.

Exploratory Data Analysis

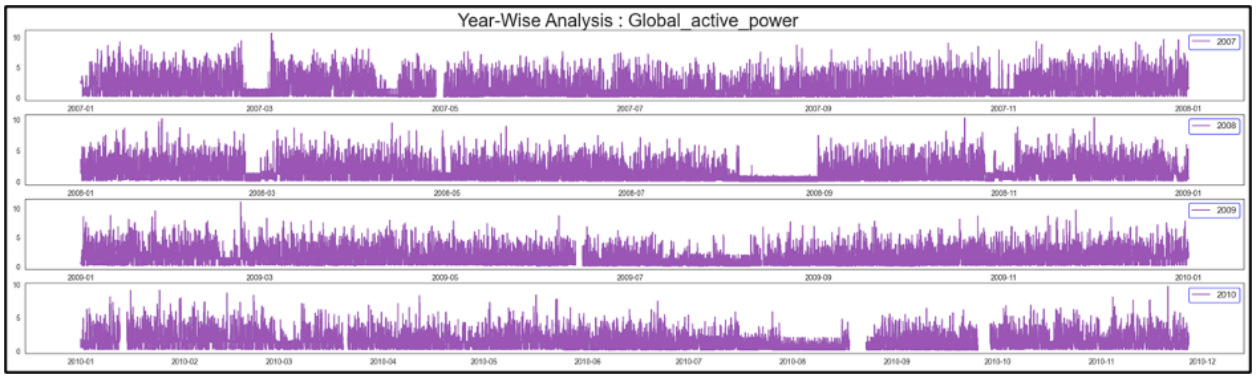
Analyzing electricity consumption patterns is essential for informed energy management and planning. Our exploratory data analysis sifts through years of electricity data to detect trends and anomalies across key power consumption metrics such as Global Active Power, Global Reactive Power, Voltage, Global Intensity, and Sub Metering. We aim to reveal seasonal variations and unique consumption events to aid in the creation of models that can accurately predict future energy needs, ensuring optimized energy distribution and use. This report navigates data complexities, clarifies trends, and underscores the importance of addressing missing data, all crucial for forecasting and meeting energy demands effectively.

Year wise visualization to detect seasonal patterns and trends:

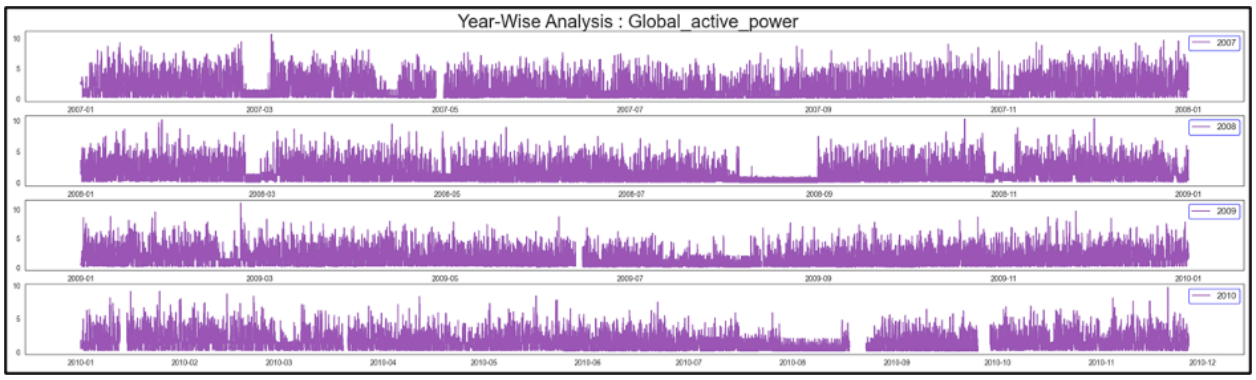
Our next step in the analysis hones in on year wise visualizations. This pivotal technique will dissect our dataset into individual years, providing a clearer picture of how electricity consumption ebbs and flows with the seasons. We'll not only discern patterns and trends but also confront the gaps—those stretches of missing data that could skew our understanding. The insights we derive from each variable's year-over-year performance will guide us in choosing the most appropriate methods for imputing these missing values. By doing so, we aim to restore the continuity of our data,

ensuring that our subsequent analyses and predictions are built on the most complete and accurate information possible.

Global Active Power: Global active power consumption consistently demonstrates seasonal rhythms, indicating a cyclical pattern likely tied to climatic variations affecting heating and cooling needs. While the data hints at possible underlying trends of power usage changes across years, these aren't immediately distinguishable and would benefit from additional analysis techniques, such as trend decomposition or moving average calculations.



Global Reactive Power Seasonal variations in global reactive power mirror those of active power, suggesting they share common driving factors. However, unlike active power, reactive power usage doesn't reveal a definitive long-term trend in the data visualization, despite its periodic highs and lows, which implies seasonally dependent usage of reactive power-dependent devices.



Other variables that shows similar fluctuations:

Voltage: The stability of the voltage is visible in the minor fluctuations around a constant average, lacking any obvious seasonal or long-term trends. This stability is anticipated since power systems are designed to maintain voltage within a safe and consistent operational range for protection and compatibility purposes.

Global Intensity: Global intensity's variability closely aligns with the global active power pattern, implying a proportional relationship between power consumption and the intensity of electric current draw. Although clear seasonal peaks and valleys are present, discerning a long-term trend from the data presented would require a deeper analytical approach.

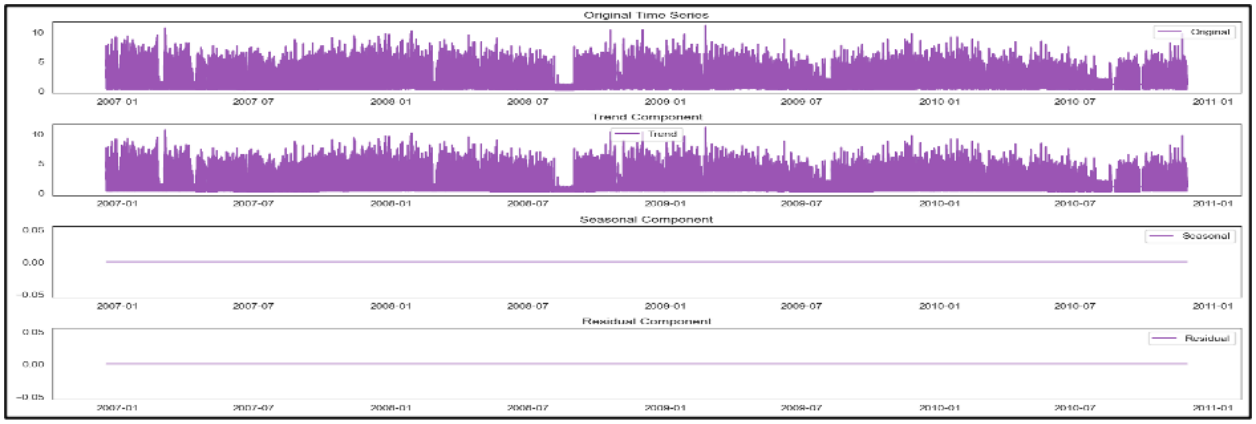
Submetering: In our analysis of electricity usage as recorded by the three sub-metering systems, we notice distinctive patterns. Sub_metering_1 and Sub_metering_2 show erratic usage with many spikes, suggesting that the appliances or systems they track are used only occasionally. There's no regular pattern of use across seasons or a consistent increase or decrease over the years, pointing to their use in response to immediate needs rather than habitual or seasonal behavior. Sub_metering_3 shows more consistent usage changes, suggesting adjustments in household appliance use or occupancy.

While we don't see a strong seasonal pattern, there are notable fluctuations: for instance, air conditioner use was steady in 2007, spiked in the hotter months of 2008 and 2009, and then was consistently high from 2010 onwards. These observations suggest that while some appliances are used as needed, others may be influenced by external factors like weather. We believer understanding these patterns will be crucial when filling in the gaps where data is missing, ensuring that we get an accurate forecasting model.

Imputation techniques for handling missing values:
From these graphs, we see steady trends in the data from year to year, indicating that the yearly mean could be a good way to fill in the roughly 25,000 missing values for each feature. Before we commit to this method, a closer look at the data's visual trends will help confirm that using the yearly average won't skew our overall analysis. This step will ensure our imputation strengthens the dataset's reliability for future forecasting.

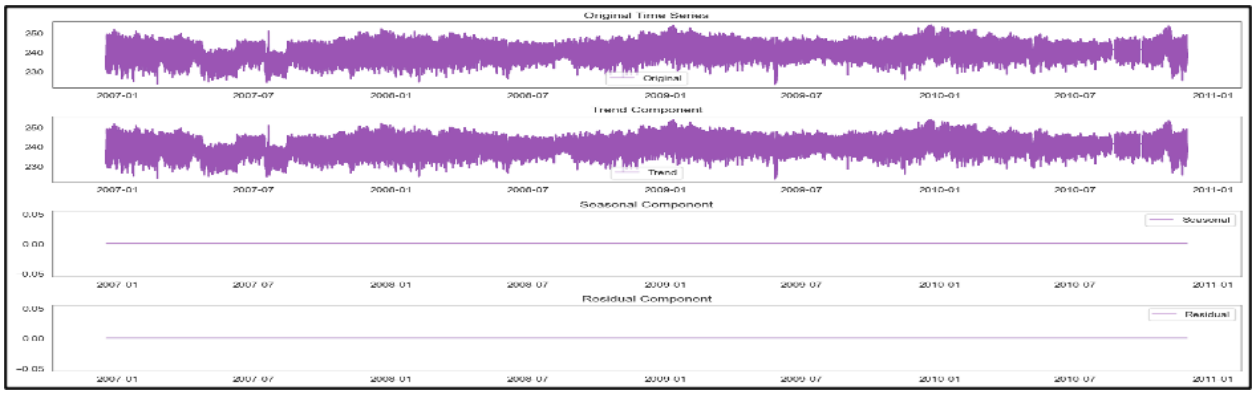
Time series decomposition to separate trends and seasonal effects.

Global Active Power: The seasonality is evident, with clear patterns of consumption peaks and troughs throughout the years, which could correspond to seasonal changes in weather and associated heating or cooling needs.

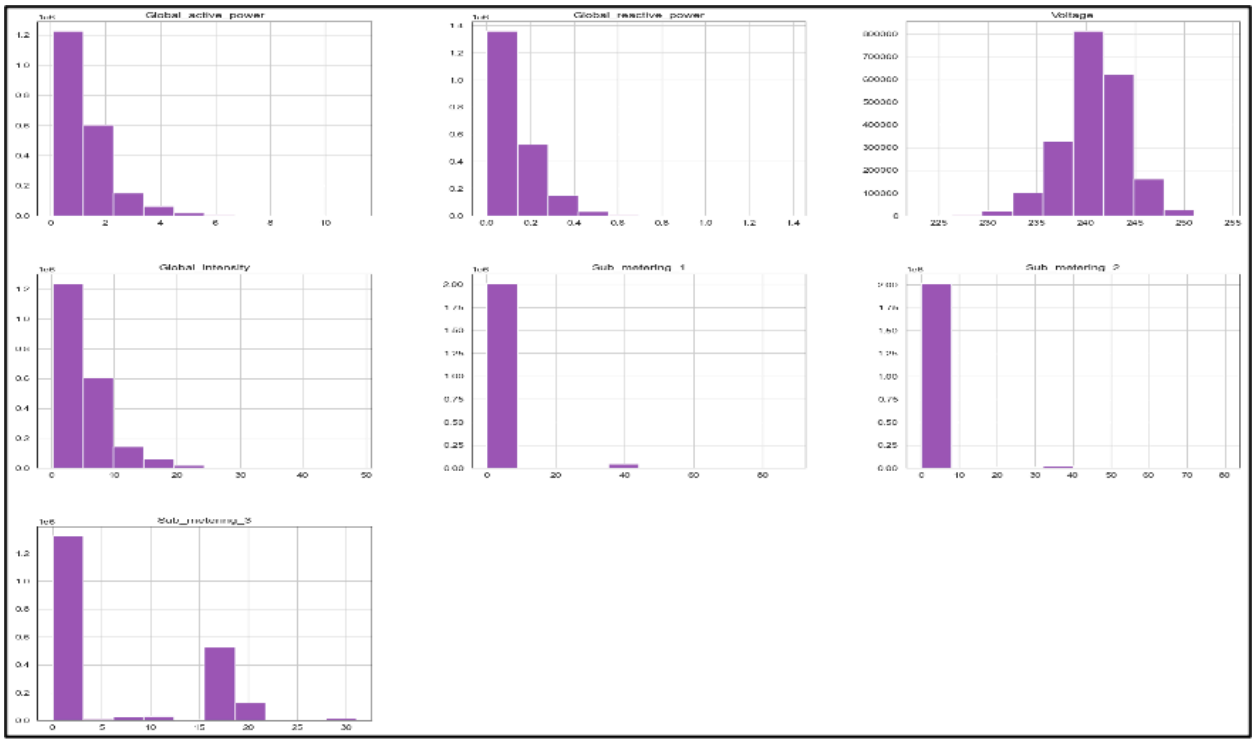


Global Reactive Power and **Global Intensity** both exhibit seasonal patterns, with the former having a less pronounced trend, suggesting diverse influencing factors compared to Active Power. Global Intensity closely mirrors the seasonality of Active Power, indicating a direct correlation between overall power consumption and the intensity of use throughout the year.

Voltage: The trend component appears quite flat, indicating voltage levels are generally stable over time, which is expected in a well-regulated power system. Seasonality is not clearly evident, which aligns with the expectation that voltage should remain constant barring any systemic changes.



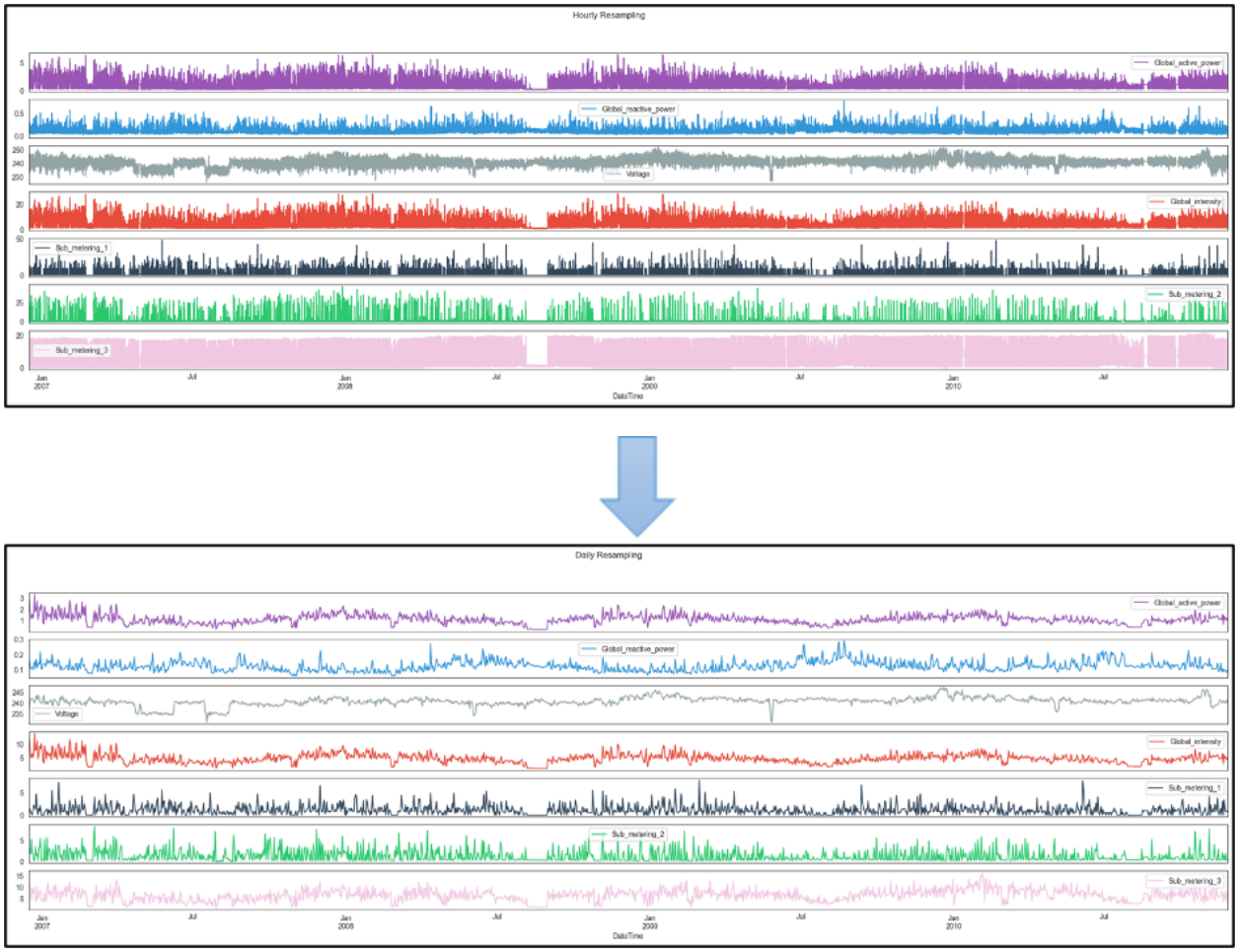
Submetering 1,2&3: The decomposition of Sub_metering_1 and Sub_metering_2 reveals a pattern of sporadic spikes without a discernible trend or seasonality, indicating intermittent and non-seasonal use of certain appliances. In contrast, Sub_metering_3 demonstrates more consistent usage patterns with a slightly more pronounced trend, hinting at long-term changes in appliance use or occupancy. However, similar to the other sub-meterings, there is no distinct seasonal influence evident.



Now having dissected the electricity usage data through time series decomposition, our next analytical step is to employ resampling techniques. This will allow us to scrutinize the data at varying frequencies, providing a lens through which we can view and understand consumption patterns over different time frames. Equally important is the role of correlation analysis; by exploring the interconnections between different consumption metrics and potential predictors, we can uncover significant relationships.

Resampling for data aggregation and distribution analysis.

Hourly Sampling vs Daily Sampling:

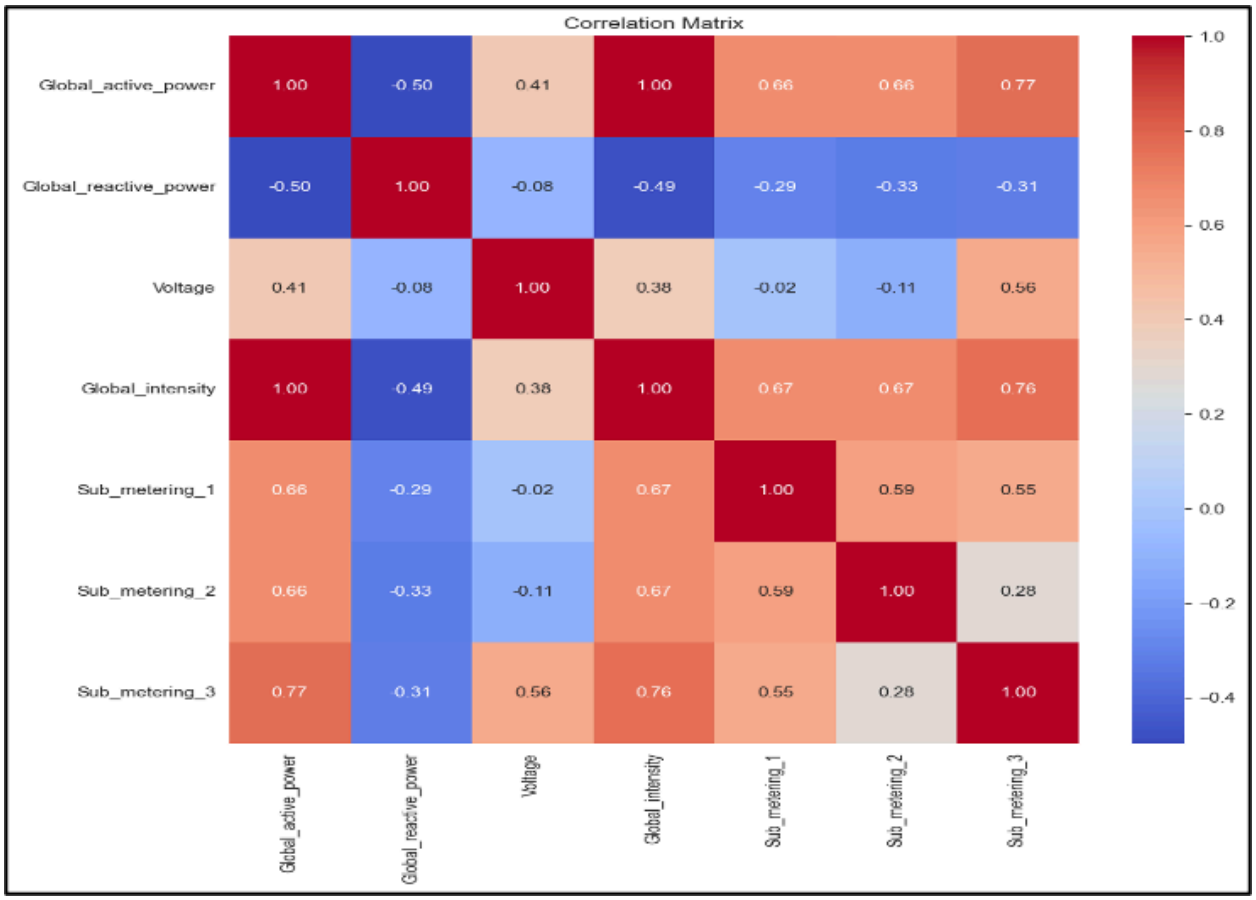


From the plot we can see that the daily resampling effectively smoothens the dataset while closely retaining the shape and patterns of the original data. The inherent trends and seasonal fluctuations are preserved through this process, which is crucial for maintaining the integrity of the data's narrative when moving towards predictive modeling. This smoothing effect achieved by daily resampling helps to filter out the noise and short-term spikes that are prominent in the hourly data, thereby providing a clearer view of the consumption behavior over time.

It simplifies the data while preserving the essential trends and seasonality, offering a more interpretable and manageable dataset for modeling. The distributions of daily resampled variables seem to follow a more normal distribution,

Correlation Analysis:

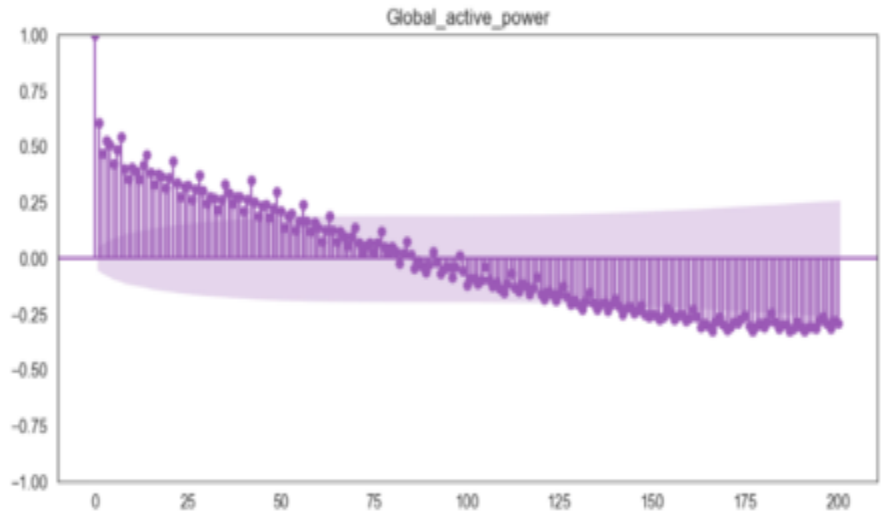
The correlation matrix (hourly, daily & monthly) reveals a compelling insight: Global Active Power is perfectly correlated with Global Intensity. This could lead to multicollinearity and hence potentially skewing the results and making it difficult to discern the individual impact of each variable. Therefore, we decided to drop Global Active Power which helped us in a more accurate and reliable forecasting.

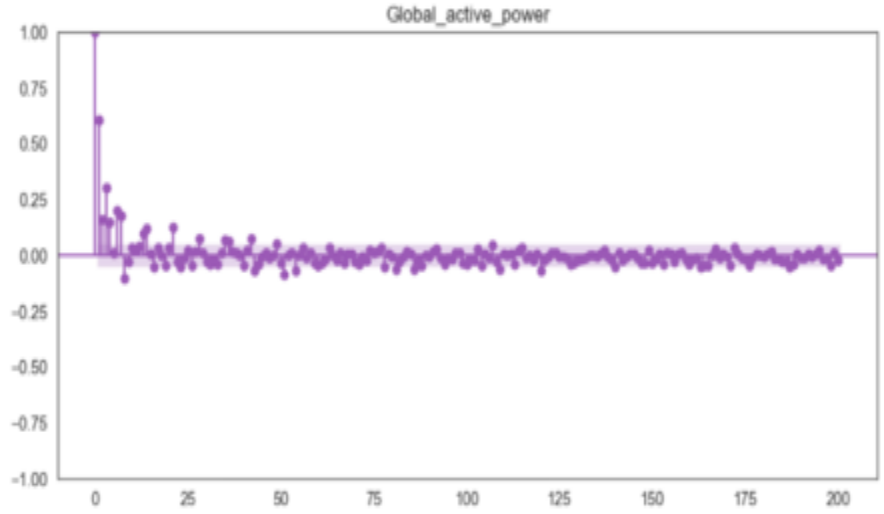


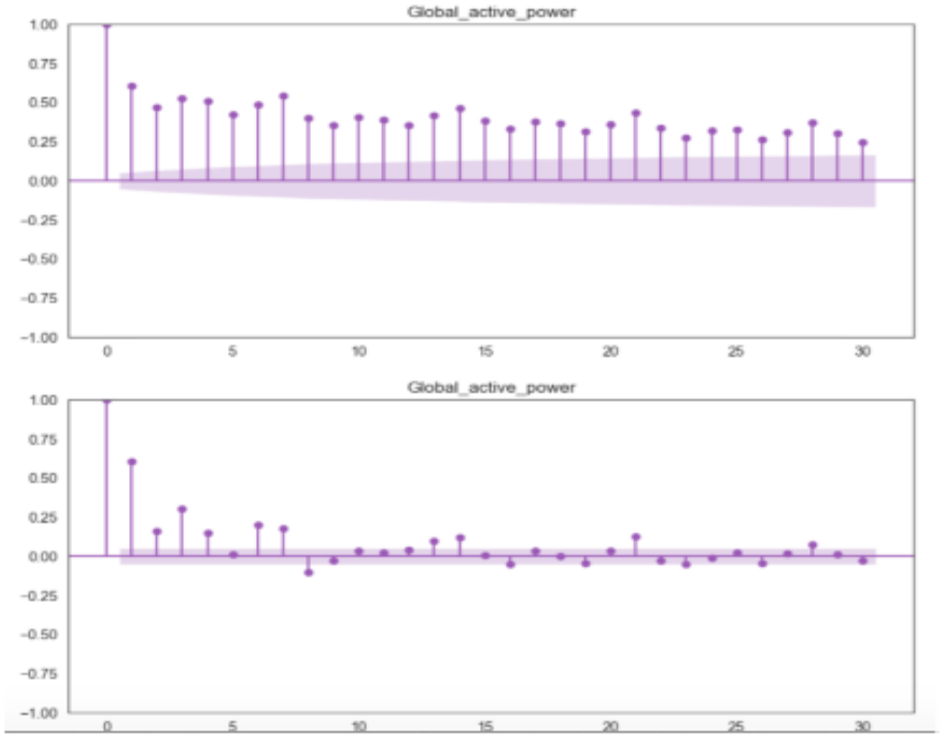
Stationarity tests & Autocorrelation Analysis

Stationary Check: The results from the stationarity checks on the dataset, conducted using the Augmented Dickey-Fuller test, indicate that all variables under consideration are stationary. With p-values well below the 0.05 threshold, we can confidently assert that the time series data does not exhibit any trends or seasonality that would typically require differencing or transformation for most time series forecasting models

Autocorrelation Analysis:







The autocorrelation plots for Global Active Power exhibit a clear pattern of declining correlation as the lag increases. Initially, there is a strong autocorrelation at short lags, which indicates that the power usage is closely related to its immediate past values. As the lag extends beyond one week, the correlation drops significantly, which suggests that the influence of past usage on future consumption diminishes after this point. This pattern of declining correlation becomes particularly notable after 7 days, implying that past data beyond this timeframe has little to no effect on current values.

This autocorrelation analysis is similar for all other variables as well. Therefore, we can conclude that a history of 7 days is sufficient to capture the most significant temporal dependencies. Therefore, for forecasting purposes, using the past week's data to predict the upcoming week's power consumption is likely to be effective, ensuring that our models are informed by the most relevant and impactful data points.

Predictive modeling

The models we used for our Global Active power prediction were:

1) An AR(Autoregressive) model
2) An MA(Moving Average) model
3) An ARIMA(Autoregressive Integrated Moving Average) model
4) Multivariate SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Inputs)model and
5) An LSTM model

Models:

1) **AR model**: An autoregressive (AR) model is a time series model that predicts future values based solely on the past values of the series itself. It assumes that the future value of a variable depends linearly on its previous values. In our given problem, an AR model could be effective because it captures the inherent autocorrelation present in energy consumption data. Energy consumption patterns often exhibit persistence over time, meaning that current consumption levels are influenced by recent consumption levels. By incorporating this autocorrelation, an AR model can effectively capture short-term trends and patterns in energy consumption, making it a suitable choice for forecasting energy usage in the absence of external factors or additional explanatory variables. However, this model assumes a linearity in the trend of values, which might be an erroneous assumption. Further, it cannot capture trends based on seasonality or a window of time present in the data.

2) **MA model**: A Moving Average (MA) model is a time series model that predicts future values based on the average of past prediction errors. Unlike autoregressive models, MA models focus on the past errors rather than the past values of the series itself. For our given problem, we considered that an MA model might be useful for capturing short-term fluctuations or noise in the data. Energy consumption patterns can be influenced by various factors, including random fluctuations due to changes in weather, human behavior, or system dynamics. An MA model can help in smoothing out these fluctuations and identifying underlying trends in energy consumption data.

3) **ARIMA model**: The Autoregressive Integrated Moving Average (ARIMA) model is a time series forecasting technique that combines autoregression, differencing, and moving average components. ARIMA models are effective for capturing the underlying structure and patterns in time series data, making them suitable for

energy consumption forecasting. For our problem statement, we chose ARIMA because the model can account for both the autocorrelation within the data (captured by the autoregressive and moving average components) and the trend and seasonality (handled through differencing). By analyzing historical energy consumption data, an ARIMA model can identify patterns such as daily, weekly, or seasonal fluctuations and use them to make accurate predictions of future consumption.

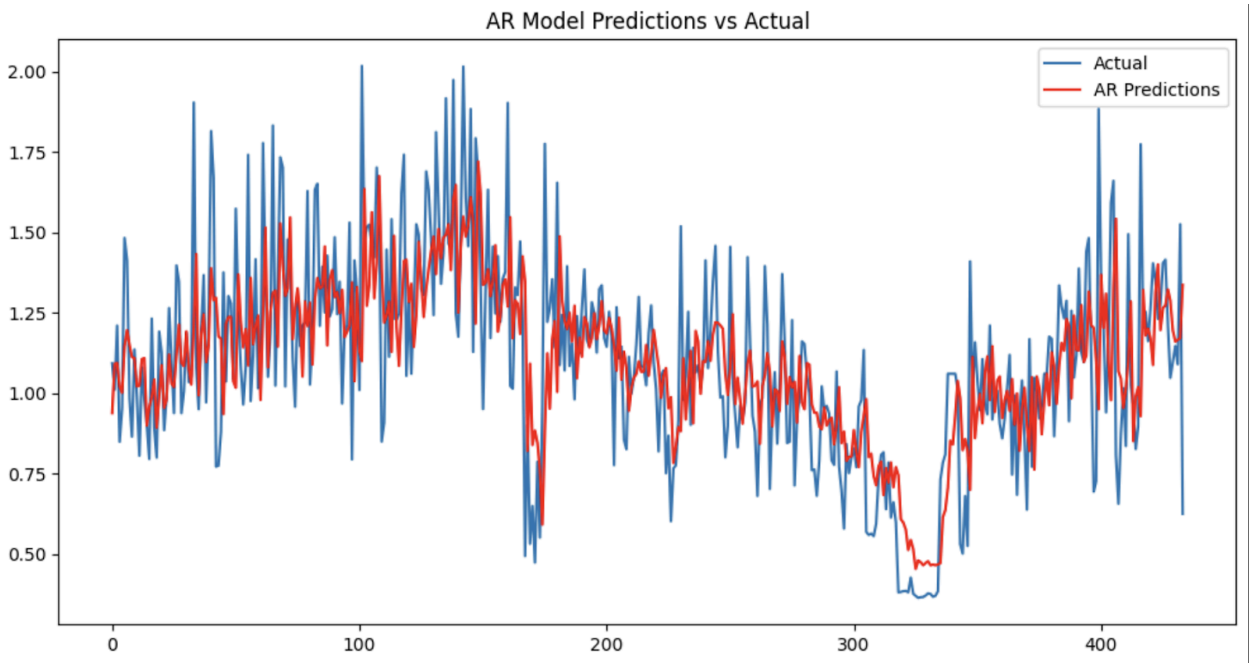
4) **SARIMAX model**: The Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) model is an advanced forecasting technique tailored for analyzing energy consumption patterns, with "Global_active_power" as the primary variable of interest and the remaining variables, including "Global_reactive_power," "Voltage," "Global_intensity," and the three sub-metering variables ("Sub_metering_1," "Sub_metering_2," "Sub_metering_3"), serving as exogenous variables. By incorporating autoregressive, differencing, and moving average components with these exogenous variables, SARIMAX models provide a comprehensive approach to analyzing time series data while accounting for external influences. The inclusion of exogenous variables enables the model to capture factors like temperature, day of the week, or holidays that may impact energy usage patterns beyond the inherent seasonal trends captured by the model. This integration enhances the forecasting accuracy by considering the influence of external factors on energy consumption dynamics. Leveraging historical energy consumption data alongside relevant exogenous variables, SARIMAX models yield more precise and insightful forecasts, which is why we decided to include this model in our modeling approach.

5) **LSTM**: Long Short-Term Memory (LSTM) models, distinct from SARIMAX models, are recurrent neural networks known for their effectiveness in modeling and forecasting energy consumption patterns. Unlike SARIMAX models, which rely on statistical methods and predefined assumptions about the data's structure, LSTM models leverage their unique architecture to capture complex temporal dependencies and long-term patterns directly from the data. In contrast to SARIMAX's explicit modeling of autoregressive, moving average, and seasonal components, LSTM models are capable of learning from historical energy usage data without relying on these predefined components, making them more flexible and adaptable to a wider range of data patterns. Their ability to handle sequence data efficiently enables LSTM models to capture seasonality, irregularities, and trends present in energy consumption data, even in the presence of noisy or missing data. Thus, we decided to include LSTMs in our

approach, mainly to compare a different modeling approach to the more mathematical models above.

Model performance:

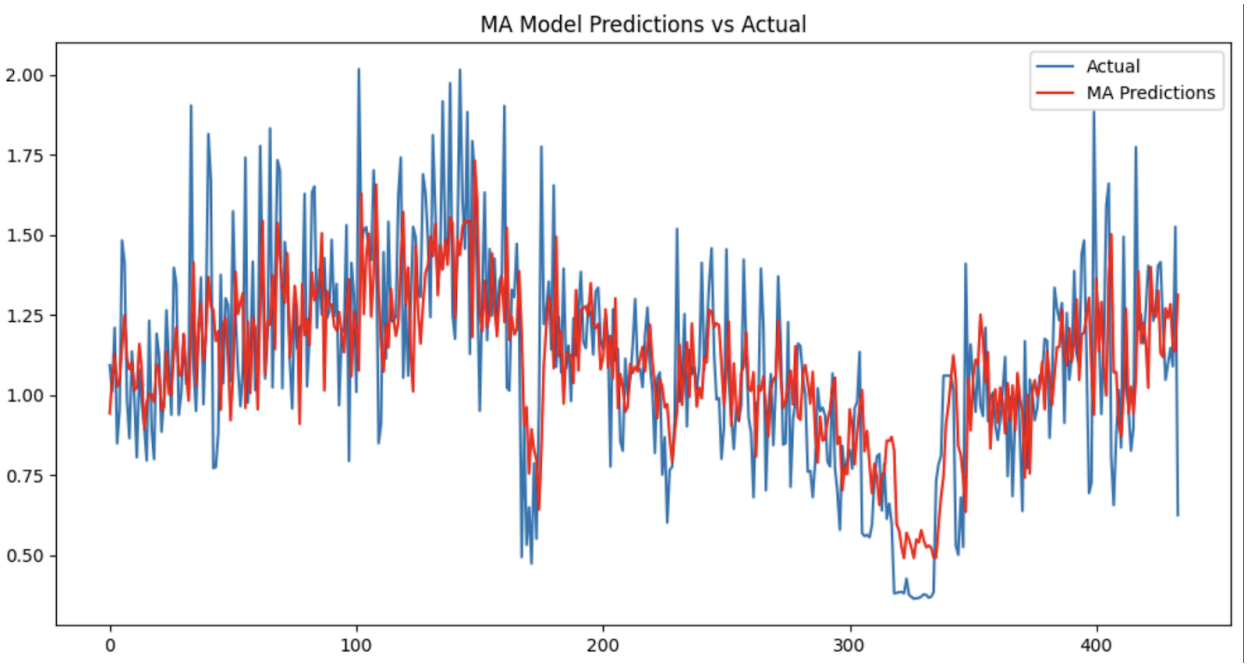
We will compare the models on their prediction MSE, as well as look at the plot of the predictions of each model vs the actual trend of the Global Power. All the models are trained over a dataset of 1000 points, and a test set of 400 points is used for prediction.
1) AR model: The AR model has an average test data MSE of 0.061. The BIC (Bayesian Information Criterion) was used to identify the best number of terms for this model, and that turned out to be 8. The coefficient summary for this model tells us that except Lag 3, all other terms are significant, with p-value<0.05.
Looking at the prediction plot:



As we can see, the model does fairly well in following the pattern of the data. It can identify the general trend of the overall Global Power. However, it is not able to handle the peak values and any short term fluctuations. To look into those, we look at the MA model.

2) MA model: The MA model has an average test MSE of 0.063. Using the BIC to find the best parameters, we find that using 21 lagged terms works the best for the model, with a BIC score of 583. The coefficient summary for this model tells us that other than Lag 16, all other terms are significant, with p-value < 0.05.

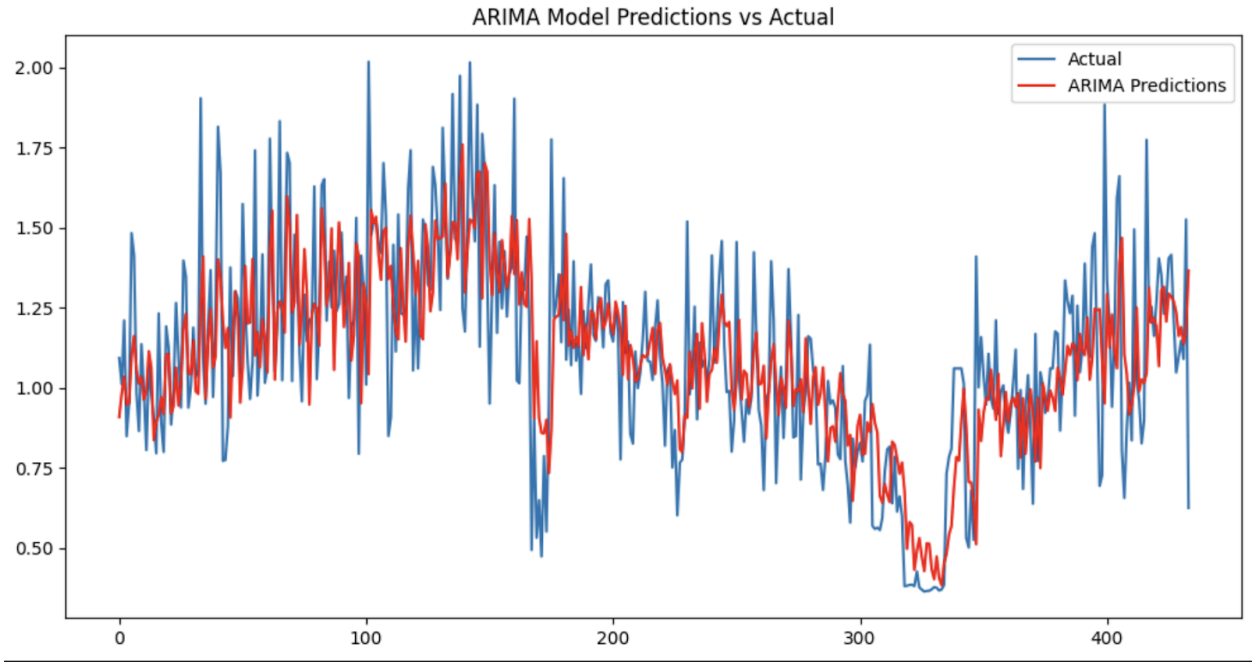
Next, we look at the prediction plot:



As we can see, the MA model can capture the peaks better than the AR model. However, in many instances it spikes out from the range of the actual values, thus not following the overall trend as closely as the AR model. To combine both these features, we look at the ARIMA model.

3) ARIMA model: The ARIMA model has an overall test MSE of 0.058, which is much better than the individual AR and MA models by themselves. Checking BIC to find the best parameters, we see that the best parameters are AR of order 7, differencing order of 1 and MA of order 19. These are quite similar values to the individual models, showing that the values agree with each other.

Checking the coefficient summary for this model tells us that unlike the previous models, there are a lot of terms here with P-value > 0.05. Most of the terms are not very significant, but that is probably because of multicollinearity and the significance of one lagging variable over another, as well as the effect of the differencing step that can introduce additional noise into the data, leading to more variability and potentially reducing the significance of AR and MA terms.
Finally, we look at the prediction plot:

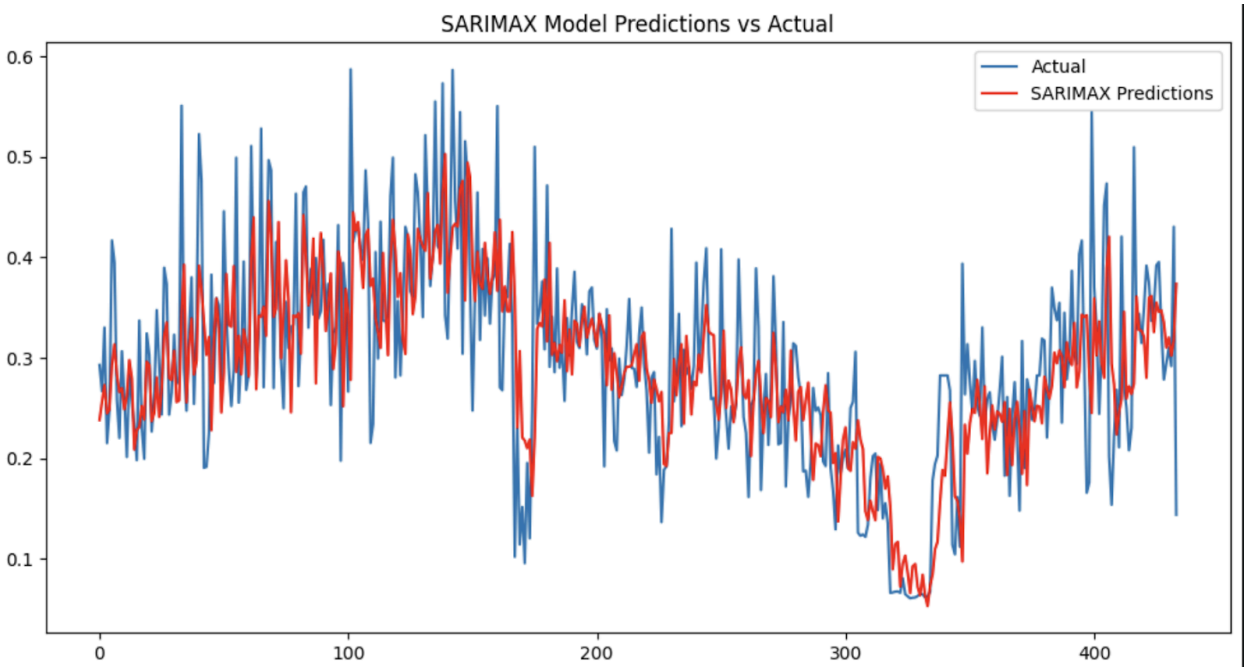


As we can see, this plot is much better than the MA and AR plot. It is following the trend and not overshooting like the MA plot was, and it is following the peaks and troughs much more closely than the AR plot was. Thus, it can produce the best out of both these models. However, this model does not take into account the information stored in other variables, such as Voltage levels as well as the power consumption of the individual appliances. It only identifies the trend in Global Active Power. So, to take into consideration these other variables as well, we look at the SARIMAX model.

4) SARIMAX model: So, we trained the SARIMAX model, using Global Active Power as our Endogenous variable and our Sub meterings, Global Intensity, Reactive Power and Voltage taken as Exogenous variables. The ARIMA components of AR, Differencing term and MA were taken as 7, 1 and 19 respectively, using the components of the previous best ARIMA model. Finally, for the seasonal part, we tried a variety of combinations, focusing on reducing our testing RMSE, finally deciding on the combination of (2,0,0,8), which meant 2 autoregressive terms and a seasonal prediction over 8 time-periods.

We trained the model till it converged, which took 49 iterations. The final objective function at convergence had a value of -4.5, which is a pretty low value, showing that the model fit well to the data. Overall, the test MSE for this model was 0.006, which is way lower than the previous three models. This shows that using the seasonal trends along with the ARIMA model is helping, and that the low convergence value is not a case of just overfitting.

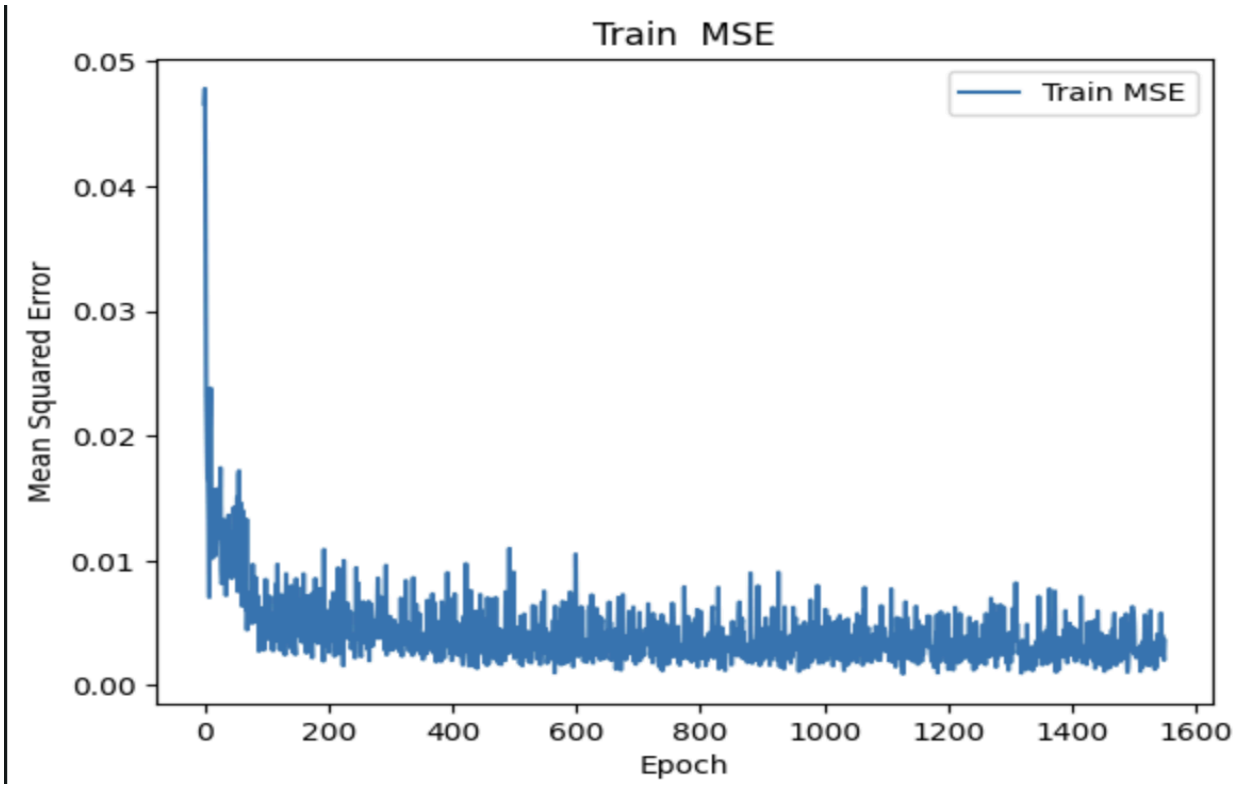
Our prediction plot looks like:



As we can see, this model does pretty well, both in terms of capturing the overall trend of the data, as well as predicting the sharp peak values. It does especially well in the plateau region around the 300 time period mark, which previous models failed to capture sufficiently well. We can see that the model's overall trend is not affected by sudden spikes in power usage, but also does pretty well in reflecting on these sudden spikes without overfitting. Thus, it seems to be the best model for this dataset so far.
So far, all our methods have been very statistical and are based on a statistical study of the time series. They rely on linear dependency of the time layers, and the regressive and moving average parts. So, to try a more non-linear approach,we decided to try an LSTM, a Neural network model that is more suited to look at any non-linearity in the data.

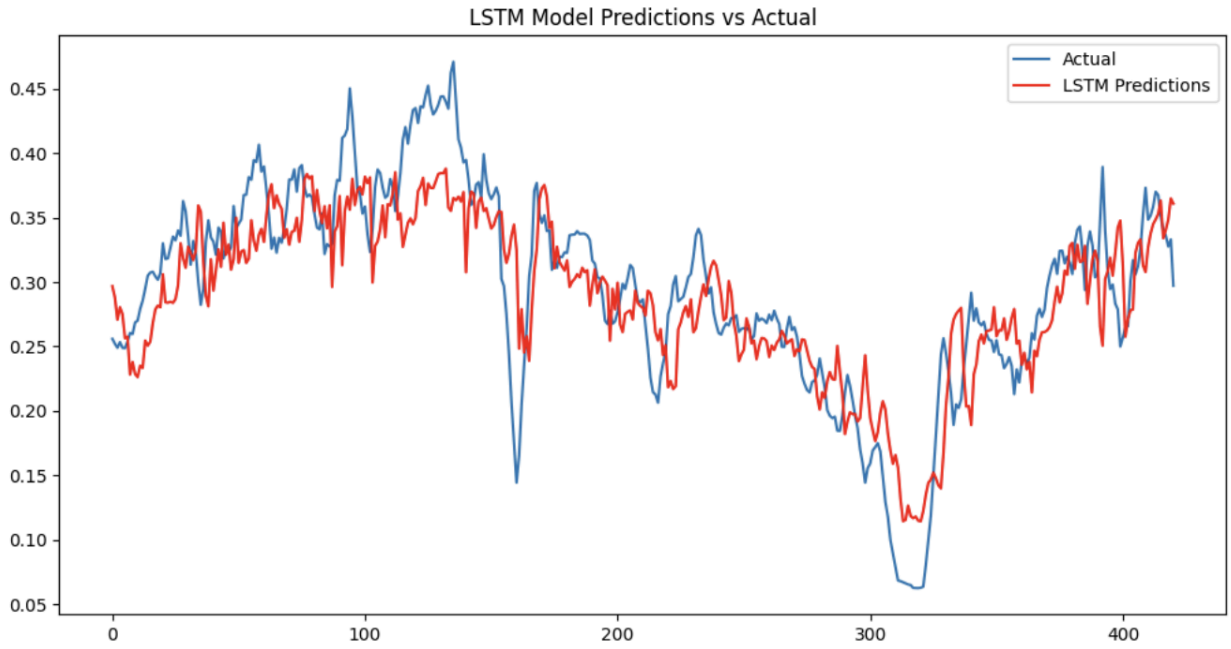
5) LSTM model: Finally, we decided to train the LSTM model. We built the model using the PyTorch lightning library, feeding our 7 predictor variable features and keeping Global Active Power as the output variable. Trying to minimize the training RMSE for the model, while deciding on the number of hidden nodes and layers, we decided upon an LSTM model with 4 hidden layers, with 200, 100, 150 and 50 neurons in each layer. The output was obviously size 1, and we trained the model for 50 epochs.

To ensure the model is converging within 50 epochs, we plotted the training RMSE vs the epochs. The graph was as follows:



As we can see, the model is converging very quickly, and around 50 itself the training RMSE drops below 0.01. So, there is no reason to train the model for a large number of epochs.

Finally, the test RMSE for this model was around 0.002. This is the best Test RMSE so far that we have achieved. It shows that the model is suitable for this problem and is not overfitting. One drawback for this model is that being a Neural Network model, we can't see how the time periods influence the output or how the lag variables affect it. It is a very black box model, giving very little insights about the behavior of the predictor variables.Looking at the prediction plot, we can see that although the model does well with the predictions, the predictions seem to be lagging a bit. This is possibly because the LSTM is just using the previous temporal inputs and is overfitting on this dataset, thus giving a good Test RMSE, since generally the n-1 value is going to be a good prediction for the current timestamp value. So, clearly this model is too complicated for a dataset which has very few features to compare.



Model Comparison:

Models	AR	MA	ARIMA	SARIMAX	LSTM
Type of model	Statistical	Statistical	Statistical	Statistical	Neural Net
Test MSE	0.061	0.063	0.058	0.006	0.002
Pro	Follows trendline	Can handle sudden fluctuations	Best of both AR and MA	Accounts for other variables as well	Accounts for non-linearity in data
Con	Can't accurately predict edge cases.	Doesn't always follow the trendline and overshoots.	Does not account for the information stored in other variables.	Reliance on linearity in data can be an issue.	Very black box model, overfits and not ideal for simpler datasets.

Final Model & Recommendation:

Overall, looking at the test RMSE and the test plots, we decide that the SARIMAX model is the best for predicting Global Active power. Not only does it account for seasonal changes in the data, it also involves the other predictor variables such as Voltage and individual appliance consumption. Further, since this dataset does not have that many features, a statistical model like SARIMAX could be a much better choice, since it won't overfit, an issue that occurs with a much more complicated model like LSTM. So, our final recommendation is a SARIMAX model, with ARIMA of 7,1,19 and seasonal combination of 2,0,0 and 8.

Link to dataset:
https://www.kaggle.com/datasets/uciml/electric-power-consumption-data-set