Time Series Prediction Using Electricity Demand Data

Time series prediction refers to the process of forecasting future values of a variable based on its historical behavior over time. It is commonly used in various fields such as finance, economics, weather forecasting, and stock market analysis, among others. Time series data is characterized by the sequential nature of observations, where each data point is associated with a specific time stamp. The Autoregressive Integrated Moving Average (ARIMA) model is a popular method for analyzing and predicting future values based on historical data. ARIMA combines three key components: autoregression (AR), differencing (I), and moving average (MA).

Autoregression (AR): The autoregressive component (AR) considers the relationship between an observation and a certain number of lagged observations (previous values in the time series). It predicts future values based on linear regression of these lagged observations. The order of autoregression is denoted as "p" in ARIMA(p, d, q), indicating the number of lagged observations used for prediction.

Differencing (Integration - I): If the data is not stationary (we'll get to this concept below), differencing is applied to make it stationary. Differencing involves taking the difference between consecutive observations at a certain lag. This helps remove trends or seasonality and can be performed multiple times if necessary. The differencing order is denoted as "d" in ARIMA(p, d, q), where "d" represents the number of times differencing is applied.

Moving Average (MA): The moving average component (MA) accounts for the dependency between an observation and a residual error from a moving average model applied to lagged observations. It captures the impact of the shock or noise in the previous error terms. The order of the moving average is denoted as "q" in ARIMA(p, d, q), representing the number of lagged forecast errors used in the model.

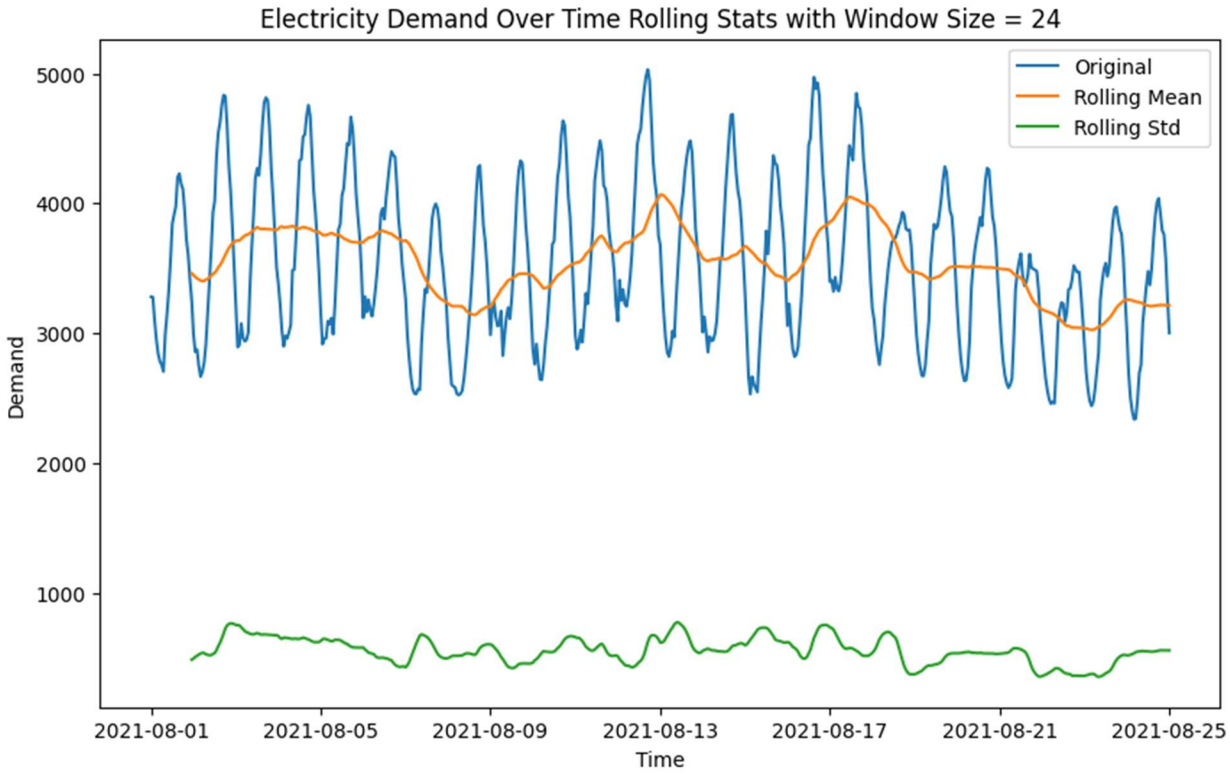
An important assumption of the ARIMA model is that the time series data is stationary, which means its statistical properties (mean, variance, autocorrelation) do not change over time. If the data is not stationary, it needs to be transformed to achieve stationarity. Stationarity can be assessed by examining the mean and variance over time or by performing statistical tests.

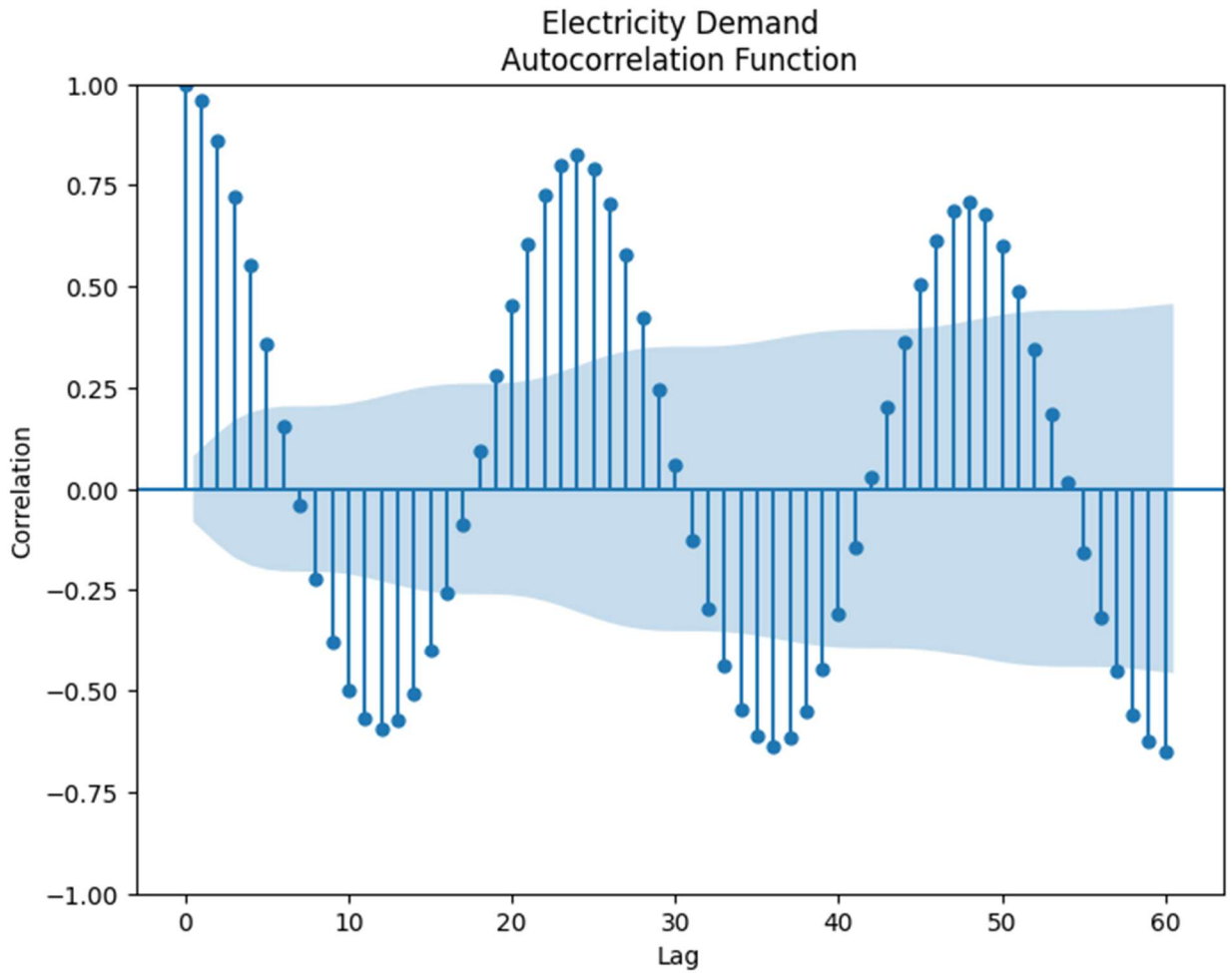
Mean Stationarity: A stationary time series has a constant mean over time. This is crucial because it allows us to make predictions based on the assumption that the future behavior of the series will be similar to its past behavior. If the mean is not stationary, the predictions may be biased and inaccurate.

Variance Stationarity: Stationarity also implies that the variance of the time series remains constant over time. This is important because it ensures that the statistical properties of the series do not change systematically over different time periods. If the variance is not stationary, the uncertainty or volatility of the series can vary over time, making predictions more challenging and less reliable.

Autocorrelation Stationarity: Autocorrelation measures the relationship between a variable and its past values. Stationarity in autocorrelation means that the strength and pattern of this relationship remain constant over time. If the autocorrelation is stationary, it suggests that the future values of the time series can be predicted based on its historical behavior. On the other hand, non-stationary autocorrelation indicates that the patterns and dependencies in the series change over time, making prediction more difficult.

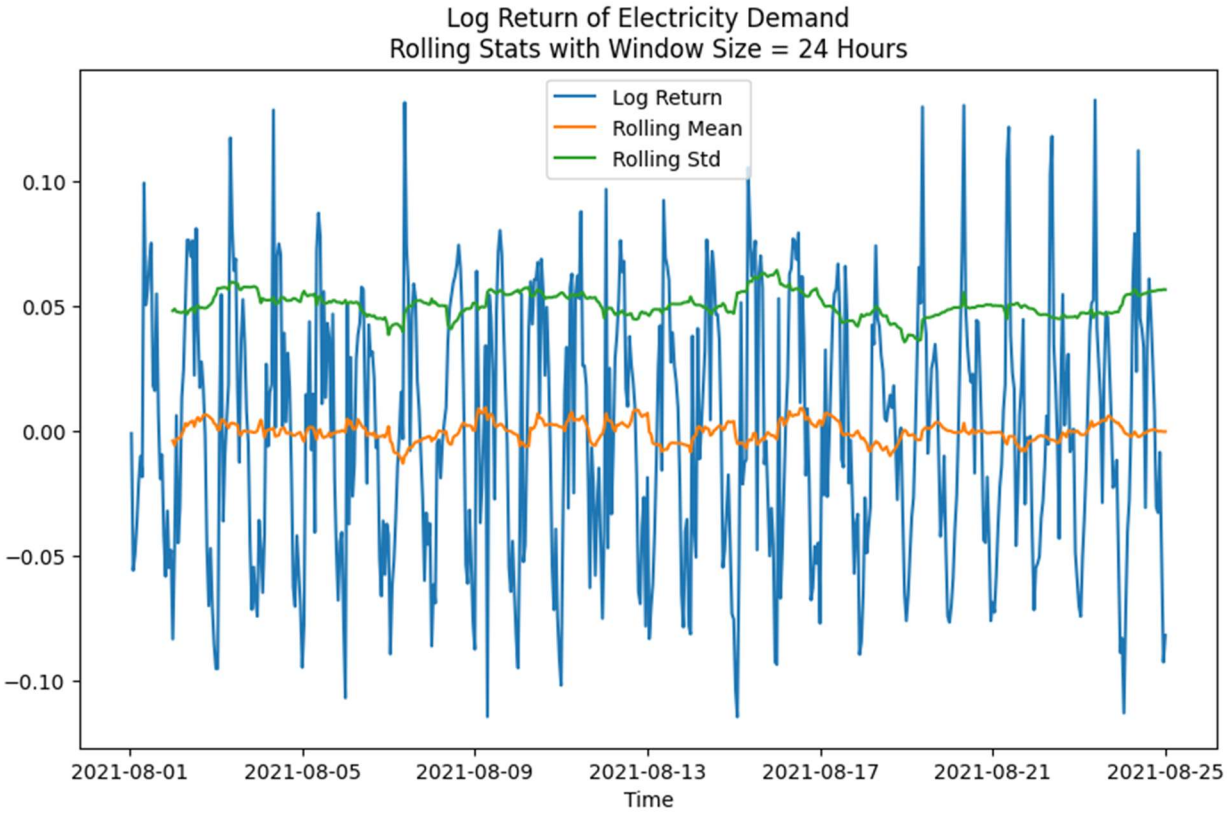
Let's see if the electricity demand time series data is stationary by first plotting the mean and variance then plotting the autocorrelation function.





Looks like both the mean and variance are changing a little over time. However, the autocorrelation function is relatively stationary.

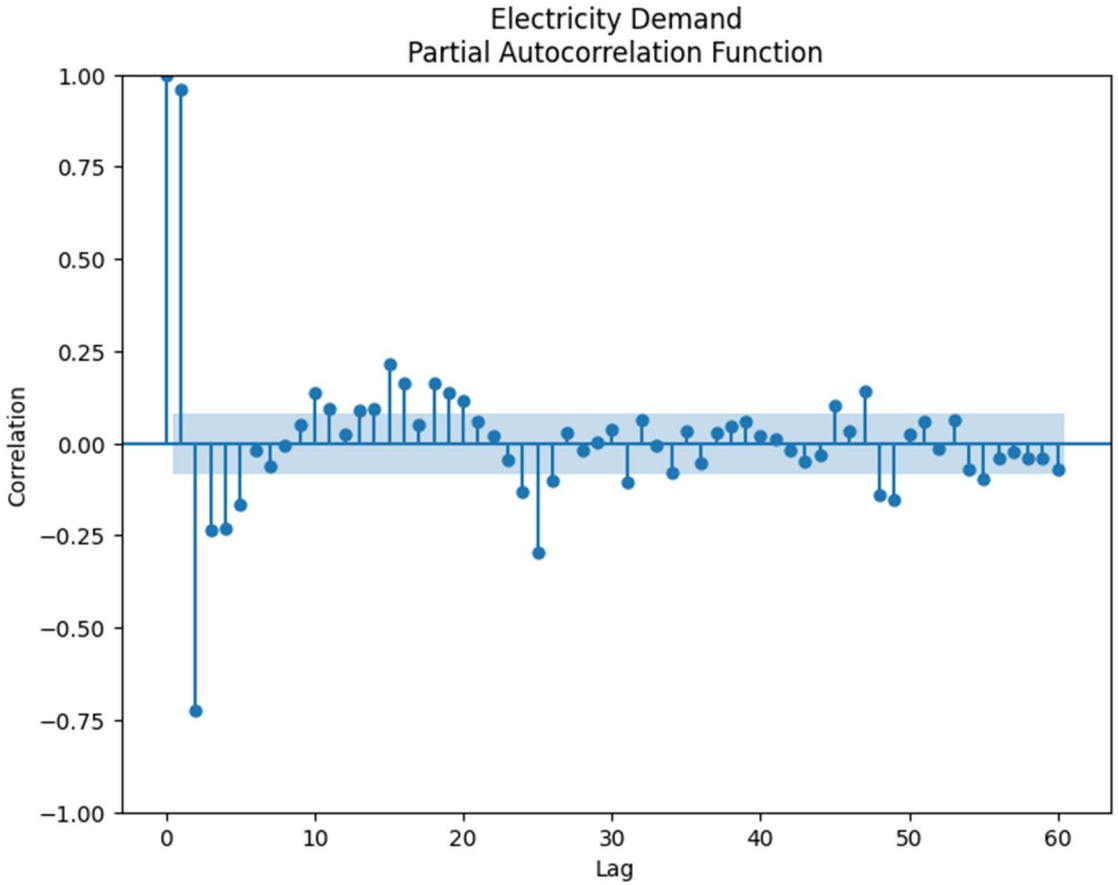
We can try to make our data more stationary by taking the log and the first order difference of values. This is often called the log return. Let's see what our data looks like after taking the log return.



Looks like the mean and variance are more stationary now.

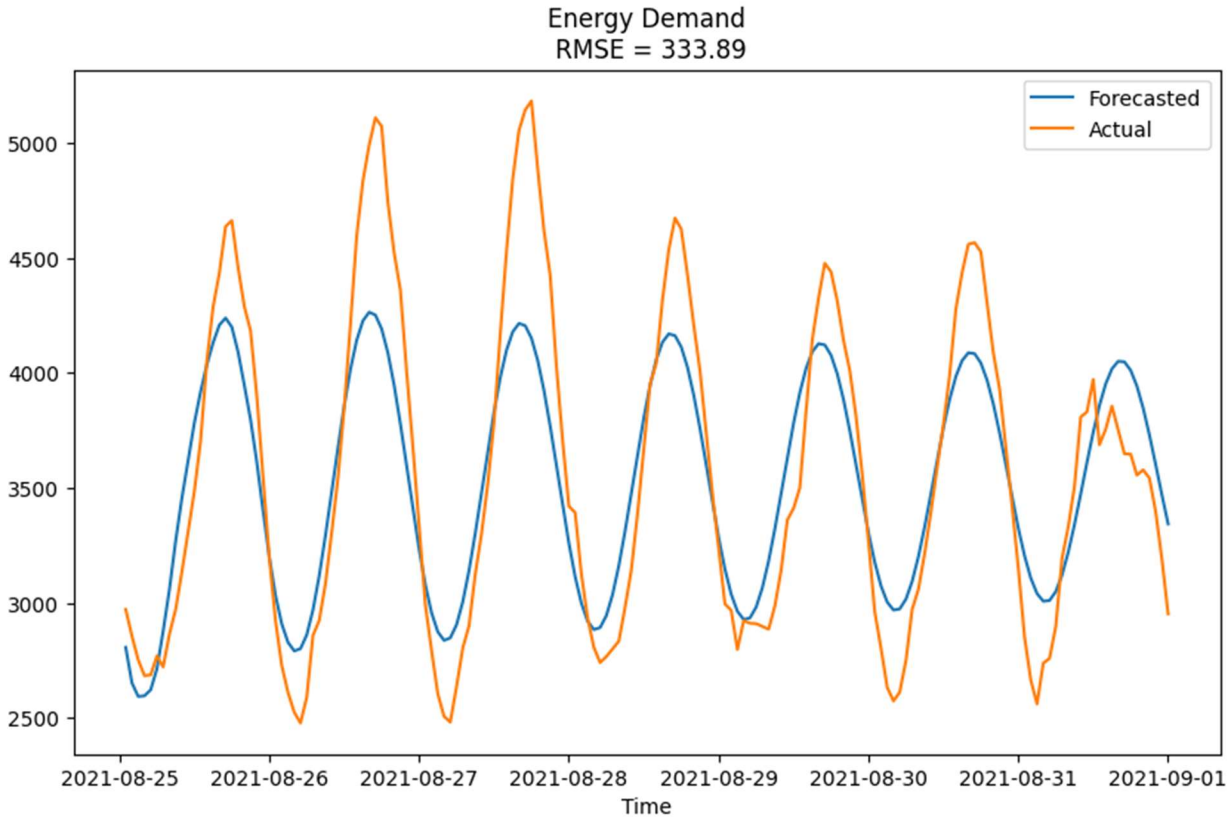
To determine suitable values for p, d, and q in ARIMA, we can analyze the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the differenced series. These plots provide insights into the potential values for p and q. The Partial Autocorrelation Function (PACF) is a tool used in time series analysis to identify the direct relationship between a variable and its lagged values while controlling for the influence of intermediate lags. It helps to determine the optimal lag order for autoregressive (AR) models, which are commonly used in time series prediction.

Here is an example of the PACF on the raw electricity demand data.



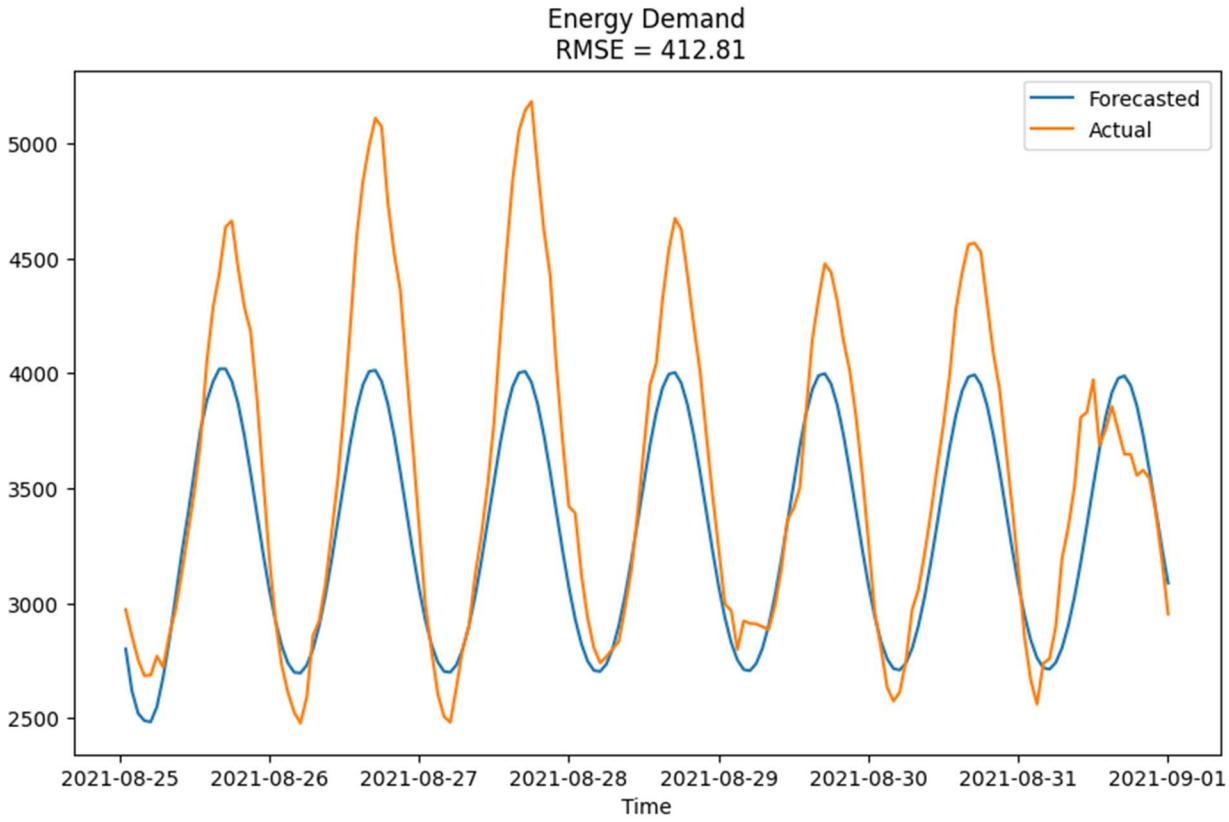
If the PACF value at a particular lag is significantly different from zero, it indicates a direct relationship between the current value and that specific lagged value. This suggests that including that lag in an autoregressive model could be beneficial for capturing the dependencies in the time series. If the PACF value at a specific lag is close to zero, it suggests that the direct influence of that lag on the current value is relatively weak or non-existent. The PACF values beyond a certain lag tend to become small and less significant, indicating diminishing direct influences as the lag increases. By analyzing the plot and identifying the significant PACF values, we can determine the appropriate lag order for an autoregressive model.

Now let's run the model and look at our predictions. First, we will predict using the raw data with no log transformation or differencing. This means we are just using an ARMA model here. We will set p=6, d=0, and q=24. We will evaluate our predictions using root mean squared error(RMSE). RMSE is often used to quantify the differences between values predicted by a model and the observed values. RMSE is the square root of the average of squared errors. The effect of each error on RMSD is proportional to the size of the squared error so larger errors have a disproportionately large effect on RMSD.



That looks pretty good. The model is not predicting the highest and lowest demand very well but otherwise is getting close to the actual values.

Now let's run the model using log transformation and first order differencing. We will log transform the series and set p=6, d=1, and q=24.



Interestingly, we got a higher RMSE with ARIMA and log transformed data than when we used the ARMA with the raw data.