Times Series Project Report

Ankush Dey
MDS202108

Aniket Santra
MDS202106

April 23, 2023

Abstract

Time series analysis has become an increasingly important tool in various fields such as finance, economics, and meteorology, to name a few. In this report, we present a comparison of three popular time series models, namely SARIMA, XGBoost, and LSTM, and evaluate their performance on a real-world dataset. We have used 3 types of metrics for e.g RMSE, MAE, and R-squared value. Our aim is to measure the electric power consumption in one household with a one-minute sampling rate over a period of almost 4 years.

Dataset

source:- https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption
It contains 2075259 rows and 7 columns
Columns Description:

- date: Date in format dd/mm/yyyy
- time: time in format hh:mm:ss
- global_active_power: household global minute-averaged active power(in kilowatt)
- global_reactive_power: household global minute-averaged reactive power (in kilowatt)
- voltage: minute-averaged voltage (in volt)
- global_intensity: household global minute-averaged current intensity (in ampere)
- 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)
- 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.

- 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.

Trend:- We have used seasonal decompose method to extract the trend, seasonality, and residual component. By fitting a linear regression line we can conclude that there is a downward trend in the data.

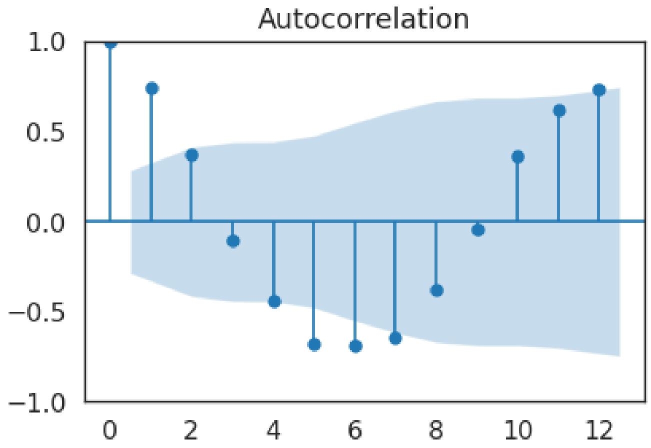
Stationarity After performing an augmented dickey-fuller test on the data, we got an p-value 0.7563. Which means the data is highly non-stationary. So to make it stationary, we use the successive difference method. So differencing it for 1 time we are getting p-value of 0.00. which means the data is now stationary.

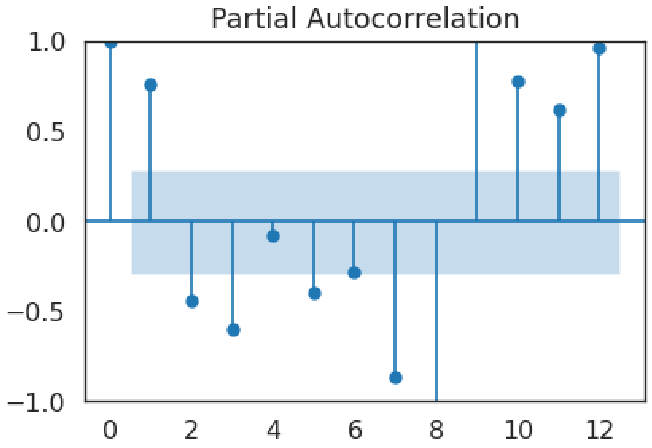
Seasonality After plotting the seasonality component we can see that the data is repeating its' pattern in every 12 months. So it is seasonal.
Hence we will use SARIMA(p,d,q)(P,D,Q,m) model.

SARIMA

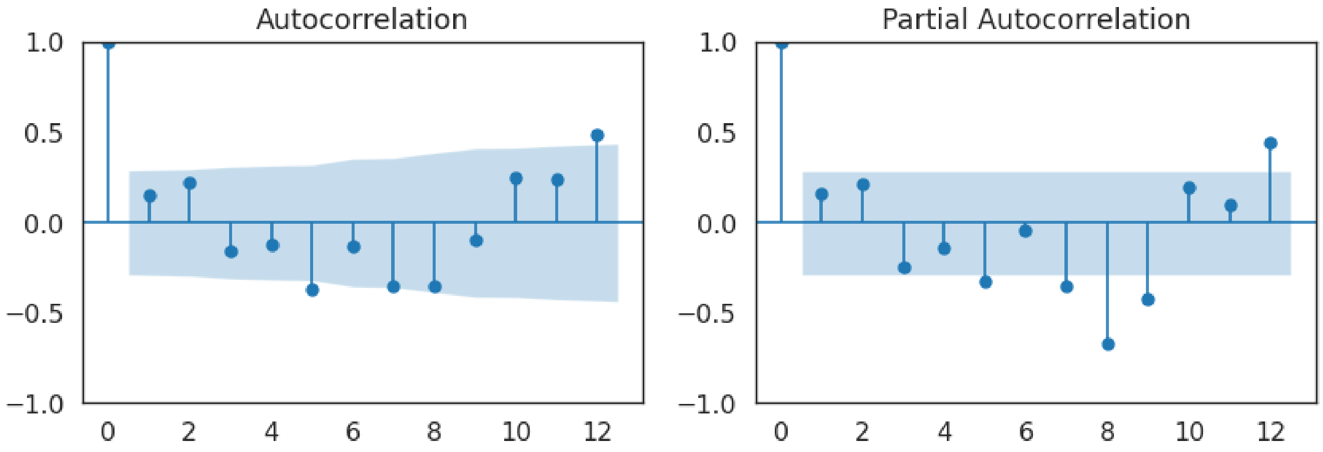
Autoregressive Integrated Moving Average, or ARIMA, is one of the most widely used forecasting methods for univariate time series data forecasting.
Although the method can handle data with a trend, it does not support time series with a seasonal component.
An extension to ARIMA that supports the direct modeling of the seasonal component of the series is called SARIMA.
Since the original data is too large for fitting into SARIMA model, we are shrinking the data into a monthly pattern.

To evaluate the parameter of the model we need to analyse the acf and pacf plot of the data. Now 1st we have Analyse the ACF and PACF plots of the seasonality component. From PACF plot we can see that the value of the 1st lag is significant, though the value of the 2nd lag is outside the blue region its not as significant compared to 1st lag, so we can roughly say P = 1, from the ACF plot we can clearly conclude that Q=1 since only 1st lag is significant. and since we using succesive differences upto order 1. So D =1 .



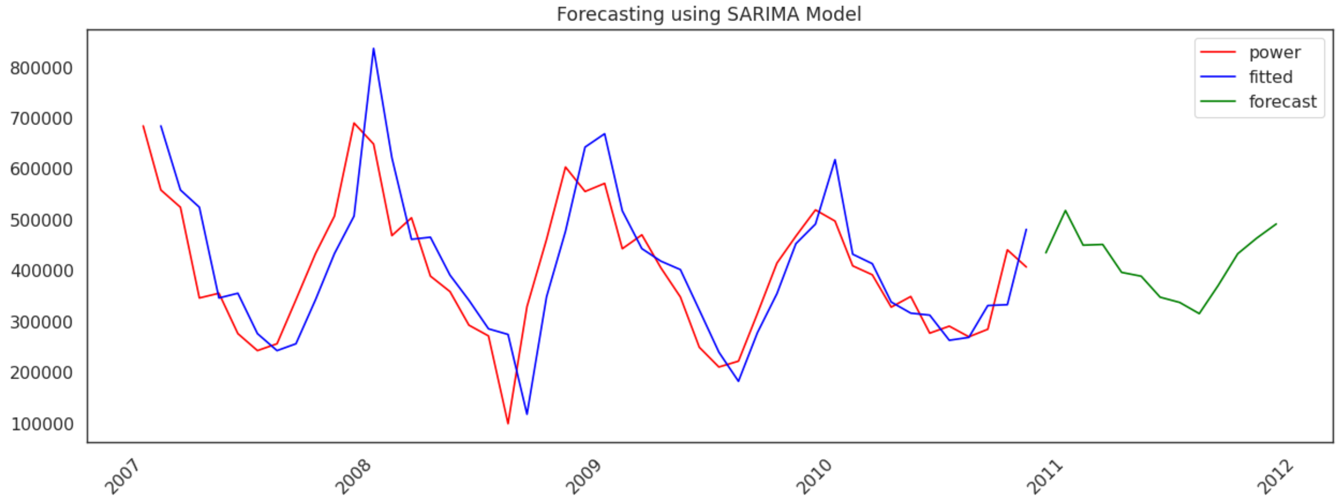


Now from the acf, pacf plot of the differenced data we got that in both cases values from the 1st lag are getting inside the blue region. which mean the ar and ma components are either 1 or 0. To finalize the values we have used the grid search method based on their aic values and we got that the values are minimum for p=0 and q=0.



so our final model is SARIMA(0,1,0)(1,1,1,12).

After fitting the model we got the below prediction curve:

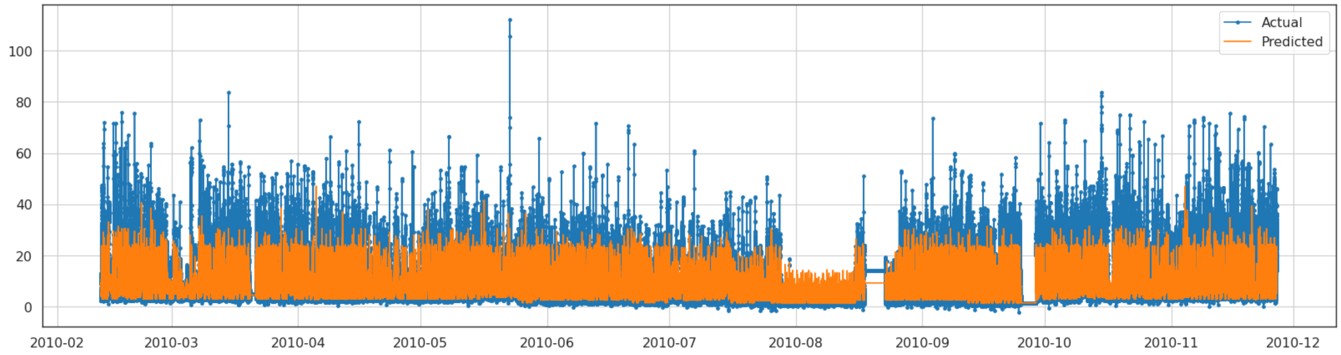


XGboost

It’s a machine learning algorithm used for supervised learning problems, especially for regression and classification tasks. It is an extension of the classic Gradient Boosting algorithm.
It works by combining several weak decision tree models into a single strong model. It is based on the principle of ensemble learning, where multiple weak learners are combined to create a single powerful model.
Since it is a supervised algorithm, we have created input variables and target variables, where the input contains the ”Global Reactive power’, ”Global intensity”,and ”Voltage”. where the target

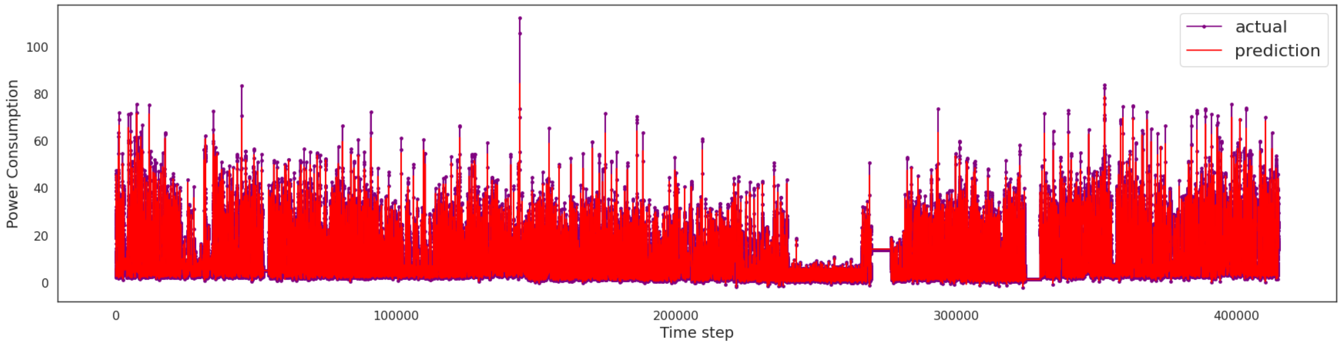
variable contains the power consumption.

below the image of the fitted data is given:



LSTM

It is a type of recurrent neural network (RNN) that is designed to handle the issue of vanishing gradients in traditional RNNs.
LSTM model will learn a function that maps a sequence of past observations as input to an output observation. As such, the sequence of observations must be transformed into multiple examples from which the LSTM can learn.
since it is sequential data, we have taken 1st 30 values as input and the 31st value as their target. So after preparing the data in this way, we have fitted it in LSTM model.
below the image of the fitted data is given:



Model Comparison

Model	RMSE	MAE	R^2 score
XGboost	4.9758	2.7720	0.47
LSTM	2.1500	1.1119	0.901

Model	RMSE	MAE	R^2 score
SARIMA	88879.006	70089.006	0.62

So our final conclusion is that LSTM is the best model while working on large datasets.