Household Power Consumption

Final Presentation of 2023 Spring Time Series

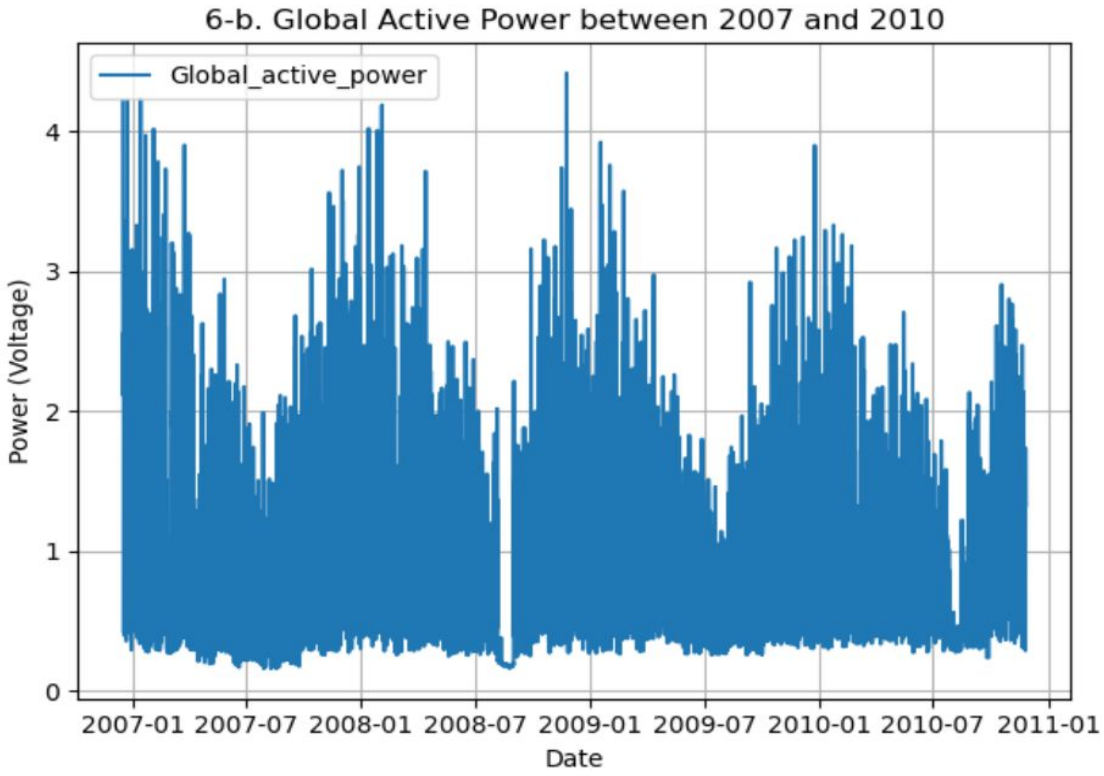
HaeLee Kim

Introduction

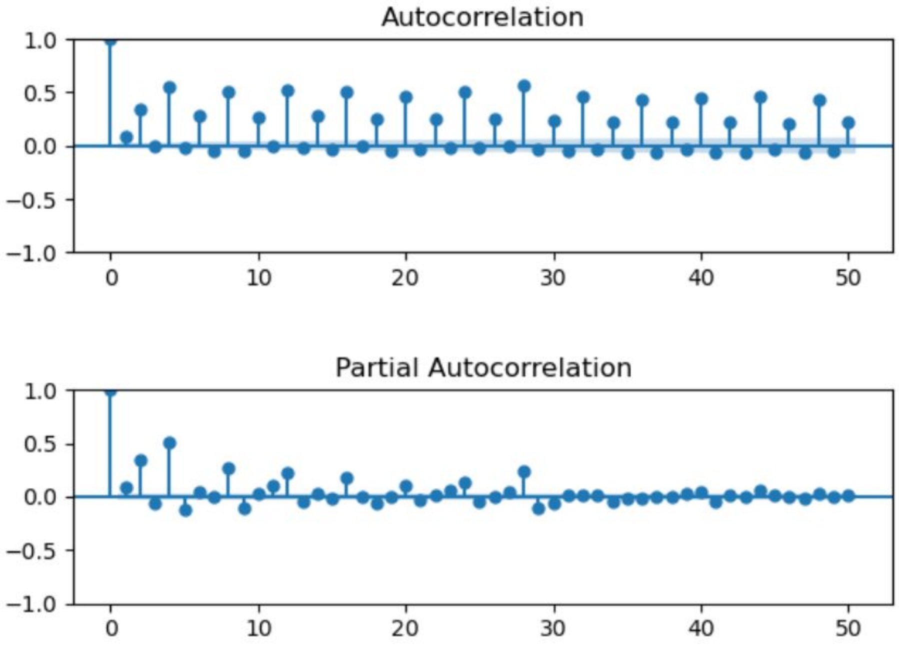
- Goal: Accurate predictions of the household power consumption
- Dataset: Individual household electric power consumption Data Set
- Source: https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption
- Contains 2,075,259 data points with 7 features (type: float)
- Collected in a local area - Sceaux (7km of Paris) - of France
  between December 2006 and November 2010 (47 months), with a data point recorded every minute.
- The missing values filled with the mean of that specific hour
- Resampled Dataset :
  Step 1. resampled hourly first
  Step 2. resampled every 6 hours, resulting in one day having 4 data points. (5766x7)
- Time Series Models: Base Models - Average, Naive, Drift, SES (Simple Exponential Smoothing), Multiple Linear Regression, ARMA, SARIMA, LSTM

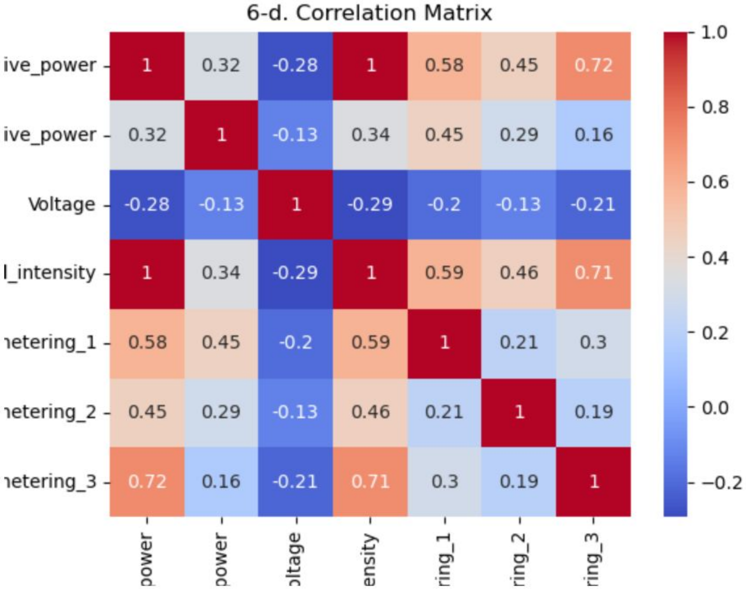
1. Description of the dataset (a) (b)

#	Column	Non-Null Count	Dtype
0	Global_active_power	5766 non-null	float64
1	Global_reactive_power	5766 non-null	float64
2	Voltage	5766 non-null	float64
3	Global_intensity	5766 non-null	float64
4	Sub_metering_1	5766 non-null	float64
5	Sub_metering_2	5766 non-null	float64
6	Sub_metering_3	5766 non-null	float64



1. Description of the dataset (c) (d) (e)

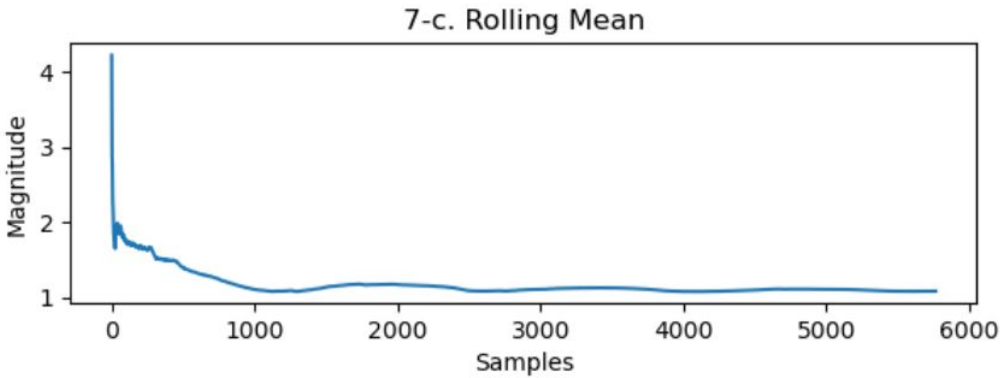




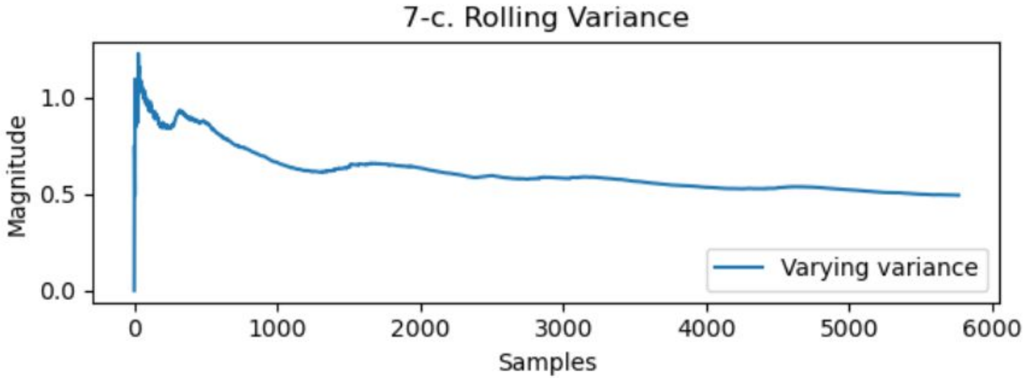
e. split the dataset into train set (80%) and test set (20%)

2. Stationarity - ADF / KPSS / Rolling Mean & Variance

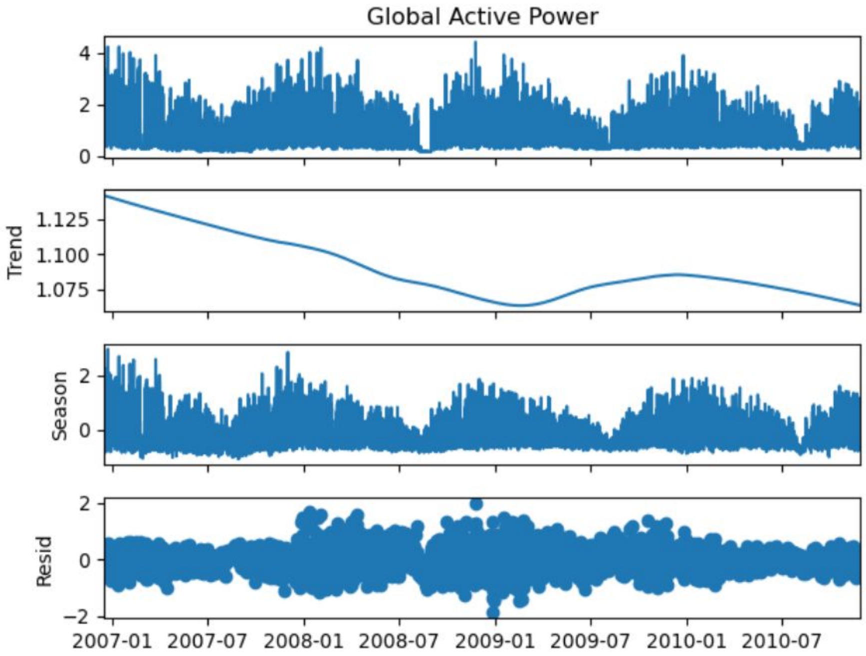
```
ADF Statistic: -5.844490
p-value: 0.000000
Critical Values:
    1%: -3.431
    5%: -2.862
    10%: -2.567
```

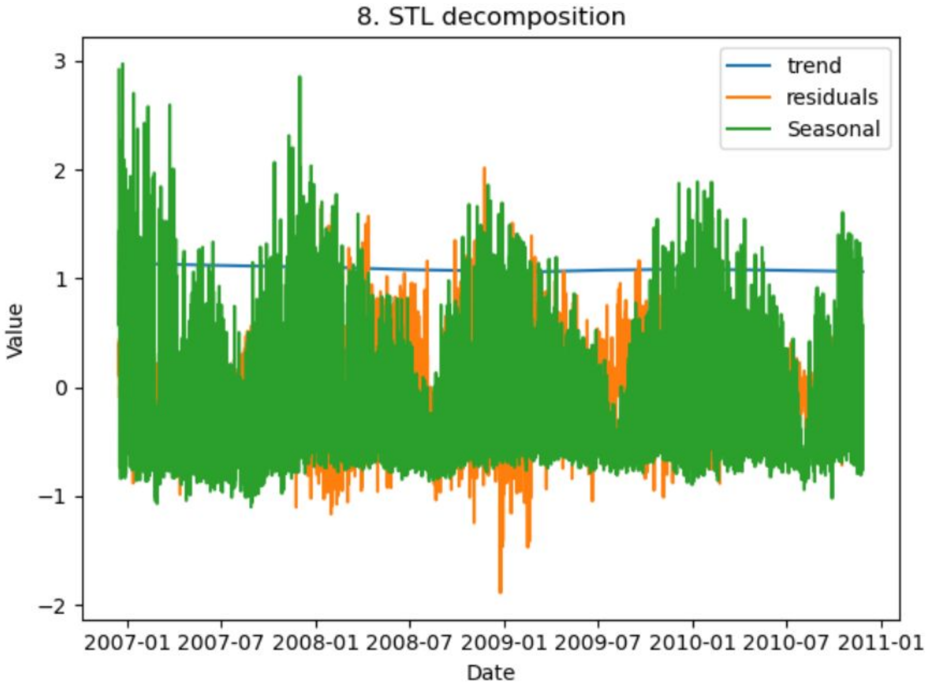



```
Results of KPSS Test:
Test Statistic              0.443840
p-value                     0.058259
Lags Used                  39.000000
Critical Value (10%)        0.347000
Critical Value (5%)         0.463000
Critical Value (2.5%)       0.574000
Critical Value (1%)         0.739000
```



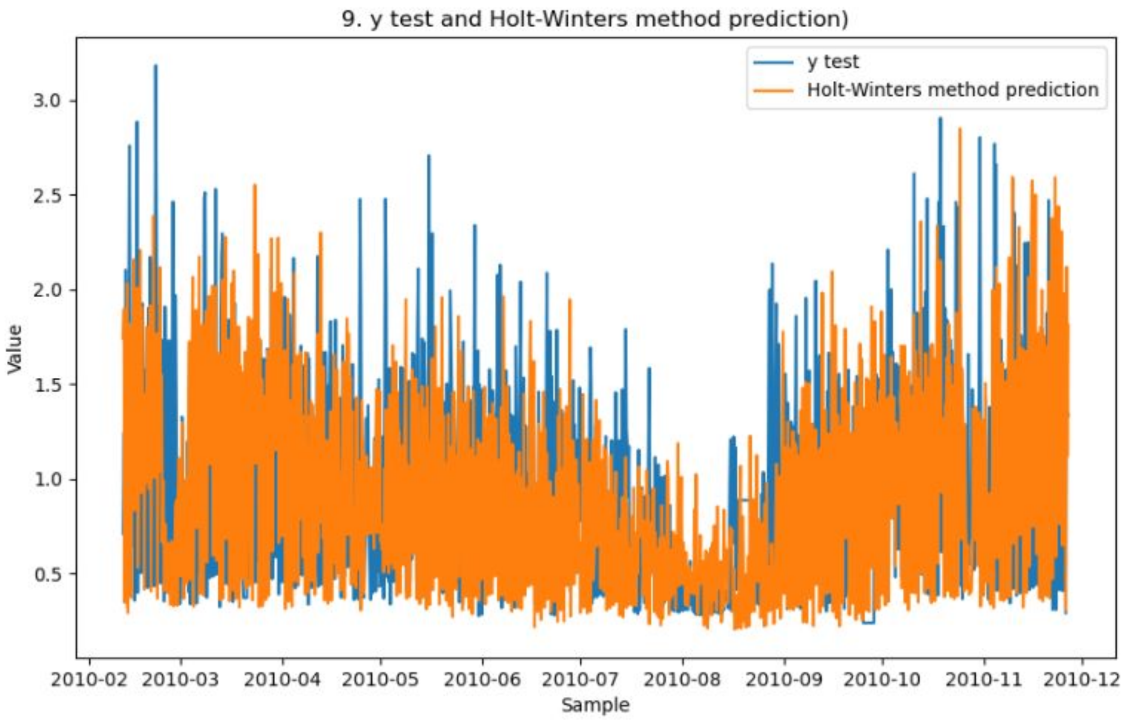
3. Time Series Decomposition






```
8. The strength of trend for this data set is 0.005443627734045942
8. The strength of Seasonality for this data set is 0.799827216323149
```

4. Holt-Winters method



Mean Squared Error: 0.2565

5. Feature Selection / Elimination

o 1st step:

VIF = [('Global_reactive_power', 10.17), ('Voltage', 10.43), ('Global_intensity', 13.16), ('Sub_metering_1', 2.36), 'Sub_metering_2', 1.75), ('Sub_metering_3', 5.60)]

Singular Values = [3.34852823e+08 2.01072139e+05 4.46070182e+04 2.44847237e+04 8.30242560e+03 9.90714491e+00]

Condition Number = 5813.70

o 2nd step:

VIF = [('Global_reactive_power', 10.16), ('Voltage', 9.31), ('Sub_metering_1', 1.71), ('Sub_metering_2', 1.40), ('Sub_metering_3', 2.73)]

Condition Number = 5811.26

o 3rd step:

VIF = [('Voltage', 2.53), ('Sub_metering_1', 1.44), ('Sub_metering_2', 1.33), ('Sub_metering_3', 2.73)]

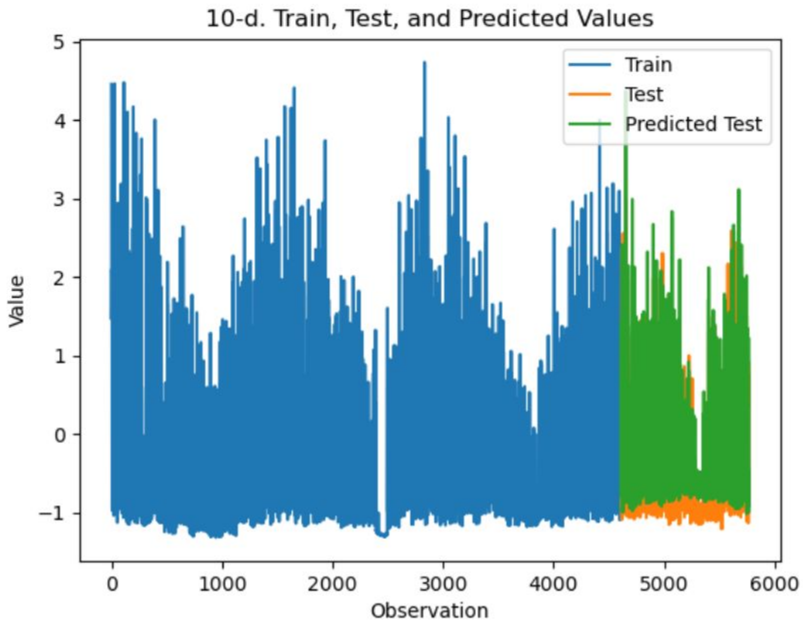
Condition Number = 5811.26

o 4th step:

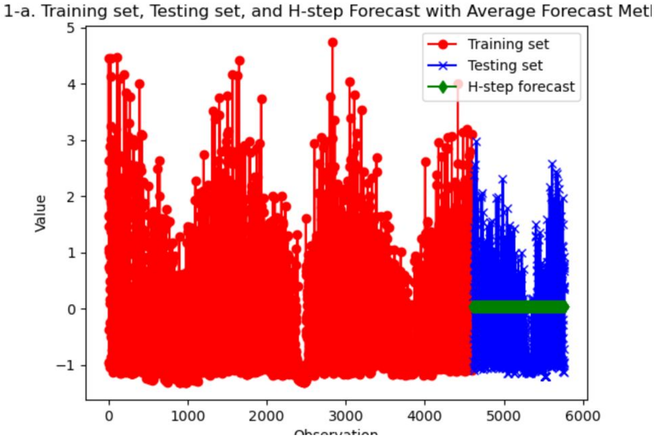
Backward Stepwise Regression

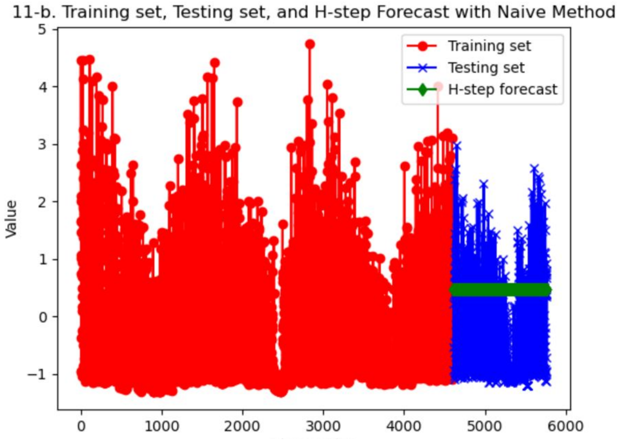
5. Feature Selection / Elimination

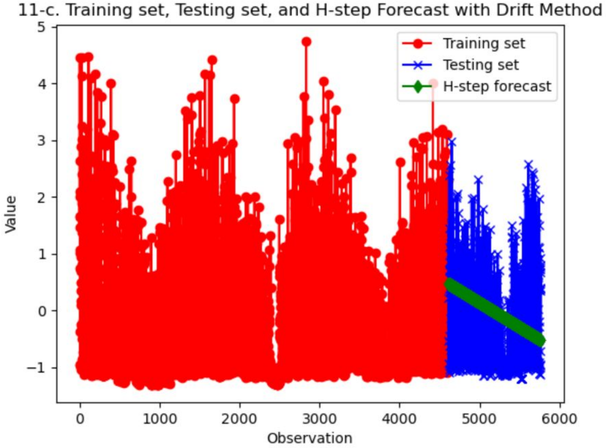
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.734
Model:                            OLS   Adj. R-squared:                  0.734
Method:                 Least Squares   F-statistic:                     3178.
Date:                Tue, 09 May 2023   Prob (F-statistic):               0.00
Time:                        16:34:02   Log-Likelihood:                -3688.4
No. Observations:                4612   AIC:                             7387.
Df Residuals:                    4607   BIC:                             7419.
Df Model:                           4
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const              0.0237      0.008      2.979      0.003       0.008       0.039
Voltage           -0.0457      0.008     -5.818      0.000      -0.061      -0.030
Sub_metering_1     0.3568      0.008     42.999      0.000       0.341       0.373
Sub_metering_2     0.2596      0.008     32.915      0.000       0.244       0.275
Sub_metering_3     0.5756      0.008     67.832      0.000       0.559       0.592
==============================================================================
Omnibus:                     1789.002   Durbin-Watson:                   1.393
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7697.182
Skew:                           1.876   Prob(JB):                         0.00
Kurtosis:                       8.097   Cond. No.                         1.56
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
10-d. Features to keep: ['Voltage', 'Sub_metering_1', 'Sub_metering_2', 'Sub_metering_3']
10-d. Features to eliminate: []
```

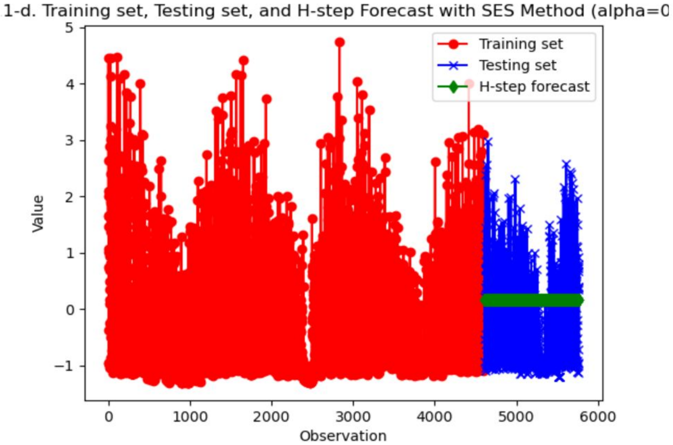


6. Base Models - average, naïve, drift, simple and exponential smoothing









7. Multiple Linear Regression Model

a. MSE - One-step Ahead Prediction : 0.289

MSE - Forecasting: 1.650

b. AIC: 7386.812

c. BIC: 7418.994

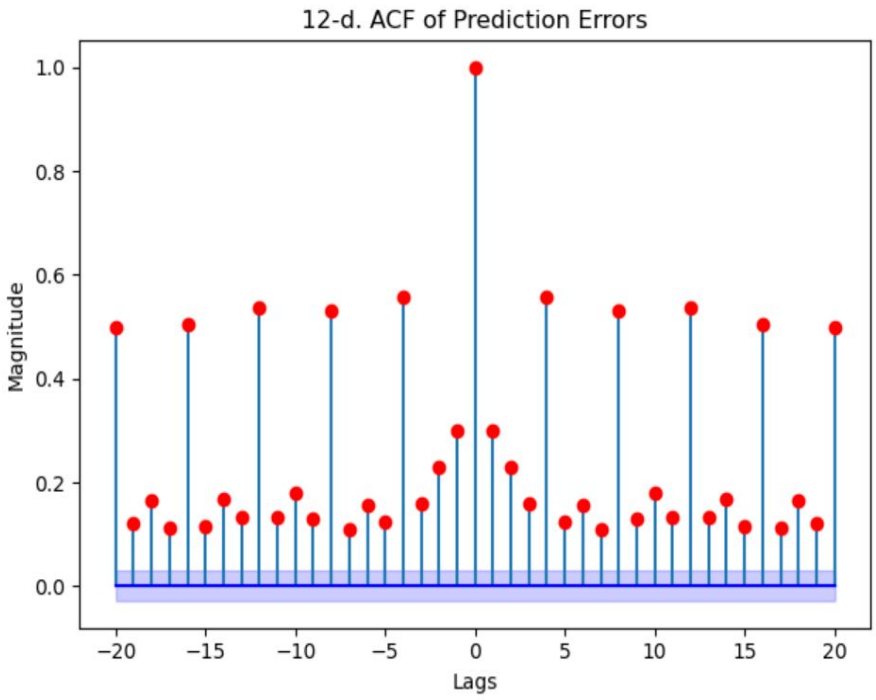
d. RMSE: 0.538

e. Adjusted R-squared: 0.733

f. Q value is 8325.847

g. Mean of the residuals is 5.145

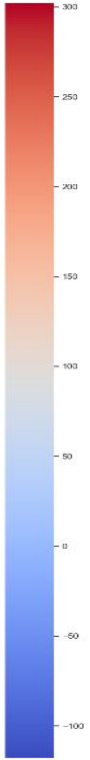
h. Variance of the residuals is 0.289

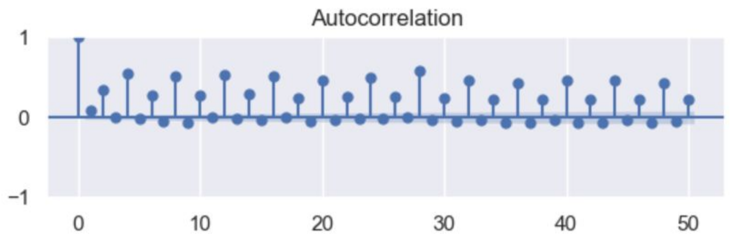


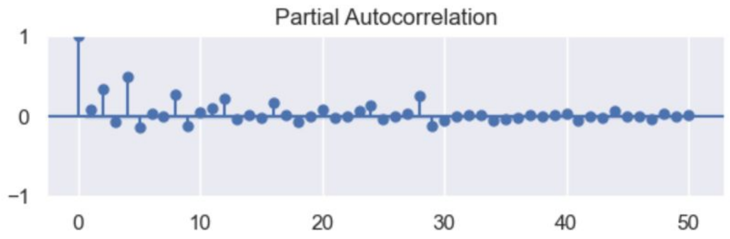
8. ARMA / ARIMA / SARIMA model order determination (a) (b) (c)

13-a. Generalized Partial Autocorrelation(GPAC) Table: ARMA

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0.09	0.34	-0.06	0.50	-0.13	0.04	-0.01	0.27	-0.12	0.04	0.11	0.23	-0.04	0.02	-0.02
1	3.91	0.36	2.64	0.49	0.03	0.02	1.83	0.27	-0.02	0.34	0.02	0.25	0.08	-0.01	0.20
2	-0.01	1.58	-0.19	0.49	-0.41	-1.10	0.82	0.27	3.93	0.34	-3.09	0.25	0.18	2.32	0.32
3	-117.91	1.59	4.23	0.48	-17.68	-15.34	1.02	0.33	-0.73	1.22	0.72	0.24	4.06	-3.14	1.09
4	-0.03	0.51	-0.08	0.94	-0.12	-0.10	-0.20	-0.14	-0.07	0.09	0.22	-0.02	-0.04	0.18	-0.17
5	-18.05	0.51	6.11	0.95	-0.89	0.14	-0.12	0.01	-0.22	0.25	0.22	-0.48	-0.13	0.14	-0.04
6	-0.16	1.78	-0.07	0.89	-1.85	-1.65	-0.11	-2.80	-0.25	-0.12	0.21	0.99	1.19	0.13	0.44
7	-11.36	1.79	22.89	0.88	-1.00	-0.05	1.66	-0.02	-0.38	-0.77	0.24	0.22	0.04	-0.08	-0.06
8	-0.12	0.53	0.15	1.14	-1.02	-32.02	1.65	-37.51	-0.40	-0.08	0.14	0.18	0.37	-0.12	-0.30
9	-4.54	0.55	-3.86	1.10	-0.28	1.42	-1.92	1.84	-0.78	-0.67	0.19	0.14	-0.12	0.73	0.17
10	0.01	1.97	0.37	1.07	3.78	1.02	1.02	0.99	0.69	0.86	-0.00	0.18	1.04	0.93	-1.80
11	302.09	1.97	-5.86	1.03	-0.42	-0.03	-0.50	0.65	-1.61	0.86	38.37	0.17	-0.51	0.39	-0.50
12	-0.04	0.53	-0.03	0.94	-0.49	8.30	-0.55	0.24	-0.44	0.24	1.77	4.05	-0.32	0.01	-0.09
13	-14.18	0.53	17.35	0.94	-3.09	-2.16	-0.93	-0.37	-0.34	2.60	1.25	0.92	-0.29	-3.98	-0.10
14	-0.12	1.62	-0.07	1.02	0.35	-0.48	-1.53	1.80	-0.96	0.97	0.68	1.22	-1.62	0.28	-1.99







9. Levenberg Marquardt Algorithm

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      0   No. Observations:                 4612
Model:                 ARIMA(4, 0, 4)   Log Likelihood               -5186.447
Date:                Tue, 09 May 2023   AIC                          10392.894
Time:                        18:23:51   BIC                          10457.258
Sample:                             0   HQIC                         10415.545
                               - 4612
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0662      0.074      0.894      0.371      -0.079       0.211
ar.L1         -0.0087      0.002     -3.827      0.000      -0.013      -0.004
ar.L2         -0.0064      0.002     -2.856      0.004      -0.011      -0.002
ar.L3         -0.0084      0.002     -3.588      0.000      -0.013      -0.004
ar.L4          0.9890      0.002    412.210      0.000       0.984       0.994
ma.L1          0.1149      0.008     13.982      0.000       0.099       0.131
ma.L2          0.1291      0.009     14.925      0.000       0.112       0.146
ma.L3          0.0947      0.008     11.752      0.000       0.079       0.110
ma.L4         -0.8180      0.008    -97.656      0.000      -0.834      -0.802
sigma2         0.5545      0.009     64.535      0.000       0.538       0.571
===================================================================================
Ljung-Box (L1) (Q):                  74.83   Jarque-Bera (JB):              1321.23
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.61   Skew:                             0.79
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.09
===================================================================================
```

ARMA(4, 0):

Train MSE (Prediction) = 0.71, Test MSE (Forecasting) = 0.64

ARMA(4, 4):

Train MSE (Prediction) = 0.56, Test MSE (Forecasting) = 0.60

ARMA(8, 0):

Train MSE (Prediction) = 0.64, Test MSE (Forecasting) = 0.64

ARMA(8, 4):

Train MSE (Prediction) = 0.54, Test MSE (Forecasting) = 0.67

10. Diagnostic Analysis

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      0   No. Observations:                 4612
Model:                 ARIMA(4, 0, 4)   Log Likelihood               -5186.447
Date:                Tue, 09 May 2023   AIC                          10392.894
Time:                        18:23:51   BIC                          10457.258
Sample:                             0   HQIC                         10415.545
                               - 4612
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0662      0.074      0.894      0.371      -0.079       0.211
ar.L1         -0.0087      0.002     -3.827      0.000      -0.013      -0.004
ar.L2         -0.0064      0.002     -2.856      0.004      -0.011      -0.002
ar.L3         -0.0084      0.002     -3.588      0.000      -0.013      -0.004
ar.L4          0.9890      0.002    412.210      0.000       0.984       0.994
ma.L1          0.1149      0.008     13.982      0.000       0.099       0.131
ma.L2          0.1291      0.009     14.925      0.000       0.112       0.146
ma.L3          0.0947      0.008     11.752      0.000       0.079       0.110
ma.L4         -0.8180      0.008    -97.656      0.000      -0.834      -0.802
sigma2         0.5545      0.009     64.535      0.000       0.538       0.571
===================================================================================
Ljung-Box (L1) (Q):                  74.83   Jarque-Bera (JB):              1321.23
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.61   Skew:                             0.79
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.09
===================================================================================

MSE of the SARIMA model (Training: Prediction): 0.5599
MSE of the SARIMA model (Test: Forecasting): 0.5730
```

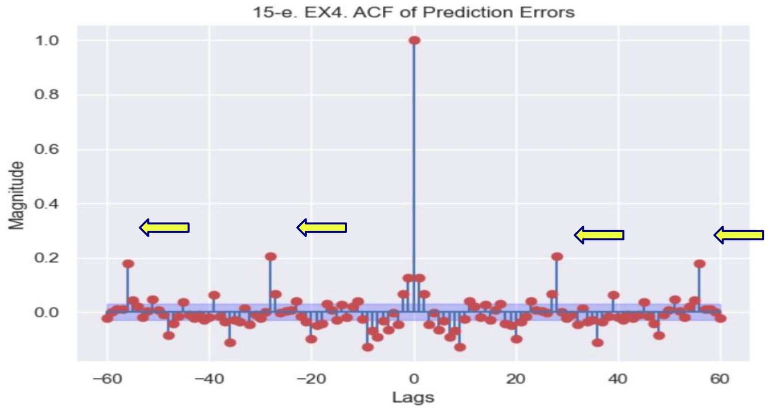
a. Poles - AR roots:

[-9.99935888e-01+0.j -6.77291587e-05+0.99885101j -6.77291587e-05-0.99885101j 9.91381616e-01+0.j]

b. Zeros - MA roots:

[-0.97227237+0.j -0.00465345+0.98527024j -0.00465345-0.98527024j 0.86666145+0.j]

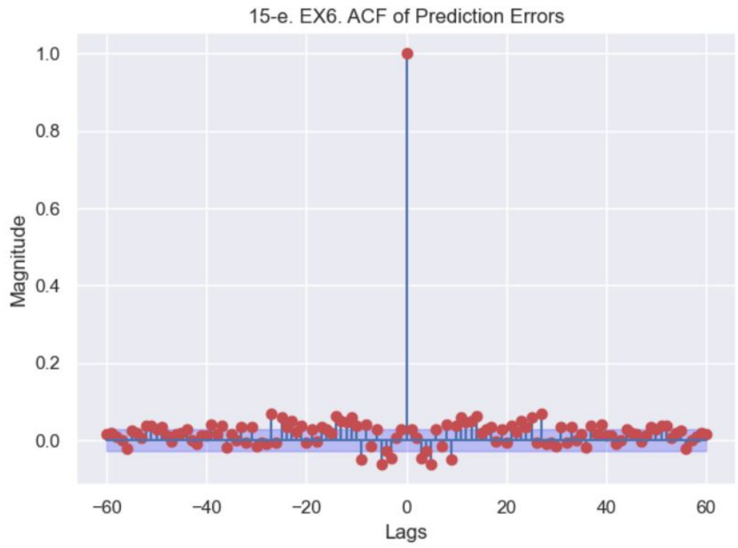
c. Estimated variance of the error for ARMA(4,4): 1.0



10. Diagnostic Analysis - Finding Better Models

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                           0   No. Observations:          4612
Model:          SARIMAX(4, 0, 4)x(1, 0, [1], 28)   Log Likelihood        -4829.440
Date:                     Tue, 09 May 2023   AIC                    9680.879
Time:                             18:50:46   BIC                    9751.680
Sample:                                  0   HQIC                   9705.796
                                    - 4612
Covariance Type:                       opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.1402      0.017     -8.129      0.000      -0.174      -0.106
ar.L2         -0.1004      0.019     -5.354      0.000      -0.137      -0.064
ar.L3         -0.1242      0.019     -6.444      0.000      -0.162      -0.086
ar.L4          0.8012      0.017     47.262      0.000       0.768       0.834
ma.L1          0.3580      0.023     15.292      0.000       0.312       0.404
ma.L2          0.3382      0.027     12.671      0.000       0.286       0.391
ma.L3          0.2913      0.026     11.310      0.000       0.241       0.342
ma.L4         -0.5031      0.024    -21.105      0.000      -0.550      -0.456
ar.S.L28       0.9697      0.004    218.533      0.000       0.961       0.978
ma.S.L28      -0.8283      0.011    -73.737      0.000      -0.850      -0.806
sigma2         0.4744      0.007     68.398      0.000       0.461       0.488
===================================================================================
Ljung-Box (L1) (Q):                   4.11   Jarque-Bera (JB):              1508.06
Prob(Q):                              0.04   Prob(JB):                         0.00
Heteroskedasticity (H):               0.63   Skew:                             0.74
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.38
===================================================================================
```

```
MSE of the SARIMA model (Training: Prediction): 0.4771
MSE of the SARIMA model (Test: Forecasting): 0.6063
```



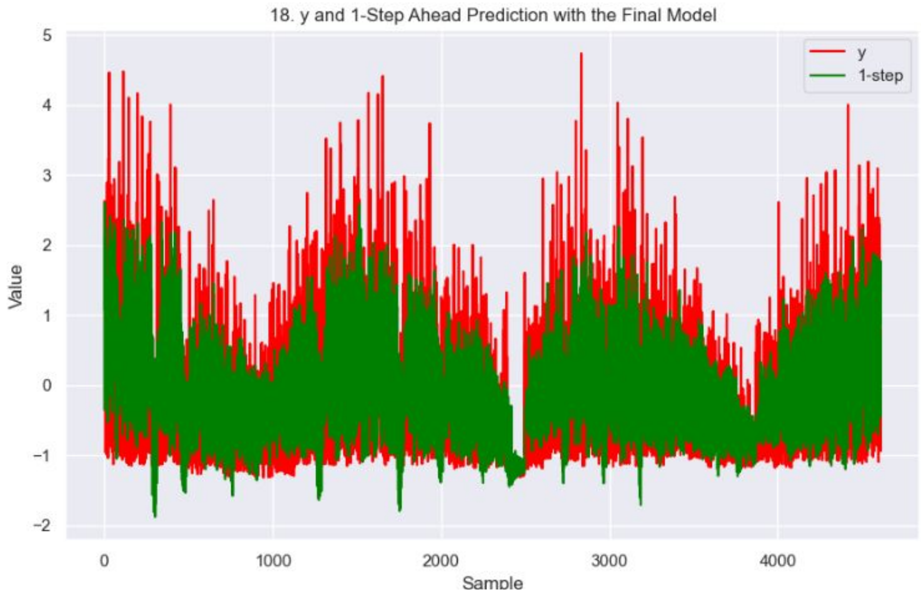
11. Deep Learning Model

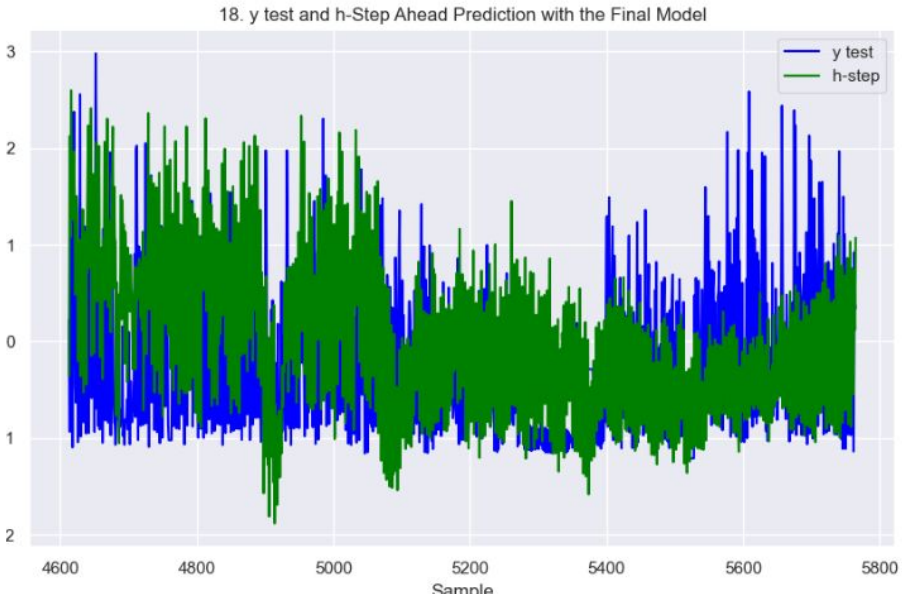
Mean squared error (MSE) of LSTM is 1.5568

12. Final Model Selection

Models	MSE	AIC	BIC
Average	0.6480		
Naive	0.9933		
Drift	0.6838		
SES	0.2008		
Multiple Linear Regression	1.6502		
ARIMA (4,0,4)	0.6032	10392.894	10457.258
SARIMA (4,0,4) x (1,0,1,28)	0.6063	9680.879	9751.680
LSTM	1.5568		

13. Forecast Function & H-step Ahead Prediction





Summary / Conclusion

- aimed to provide accurate predictions of household power consumption using Time Series models
- implemented Time Series models, including Average, Naive, Drift, SES, Multiple Linear Regression, ARMA, SARIMA, and LSTM.
- evaluated the performance of these models using the MSE, AIC, and BIC.
- found that the SARIMA (4,0,4) x (1,0,1,28) model provided the best fit to the data with an MSE of 0.6063.


Thank you!