Energy Demand Forecasting

Brooklyn Chen

MS in Data Science, George Washington University

DATS 6313 Time series Analysis and Modeling

Dr. Reza Jafari

May 10, 2023

Table of Content

Table of figures and tables

Abstract

In this study, I aimed to forecast kW, a dependent variable, using the electricity-related variables: KWh, kVARh, kVAR, and kW. These variables measure energy, power, and reactive power and are commonly used in power system analysis. I utilized various time series analysis methods and models to predict kW, evaluated their performance, identified the optimal model, and developed the forecasting function for kW.

Introduction

Time series analysis and modeling process:

1. understand dataset
2. stationary test
3. time series decomposition
4. feature selection
5. predictions by different methods and models
6. diagnostic testing
7. final model selection
8. forecasting function

Description of Dataset

There are three independent variables in this dataset, including kWh, kVARh, and kVAR. kWh (kilowatt-hour) is used to measure electricity consumption. One kilowatt-hour is equal to the energy consumed by using one kilowatt of power for one hour. kVARh (kilovolt-ampere reactive hour) is a unit of reactive power of a power system. Reactive power is power generated by capacitors or inductors for regulating voltage and current. kVAR (kilovolt-ampere reactive) is a unit of reactive power used to measure the reactive power in a power system for maintaining stable voltage and preventing overload in the power system. The dependent variable in this dataset is kW (kilowatt), which is used to measure the rate at which electricity is used. One kilowatt is equal to a power consumption or production rate of 1000 watts per second.

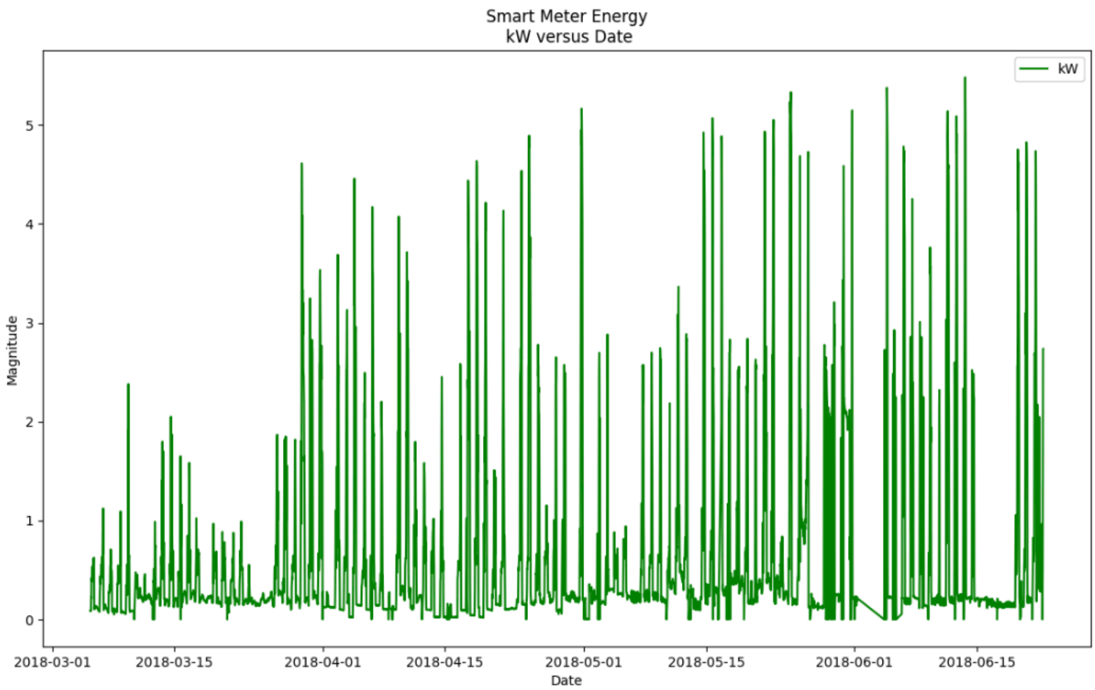
KWh and kW describe the actual power consumption and demand of the power system. kVARh and kVAR describe the reactive power in the power system, which is an important factor in regulating voltage and current. In power systems, it is sometimes necessary to adjust reactive power to maintain stable voltage and prevent overload.

The dataset consists of data collected every half hour. before analysis, I set the data type of Time_stamp as datetime, sorted the data and set time as index, and removed the serial column.

Table 1.
Smart Meter Energy(kW) – Raw Data

SERIAL	KWH	KW	KVARH	KVAR	TIME_STAMP
3098000032	10.854	0	7.814	0.002	2018/3/1 15:30
3098000032	10.75	0	7.813	0	2018/3/1 14:30
3098000032	12.325	0.086	8.302	0.076	2018/3/5 9:00
3098000032	12.372	0.094	8.345	0.086	2018/3/5 9:30
3098000032	12.415	0.086	8.386	0.082	2018/3/5 10:00
…	…	…	…	…	…
3098000032	6781.4	0.982	1721.24	0.406	2019/3/31 22:00
3098000032	6781.92	1.04	1721.32	0.46	2019/3/31 22:30
3098000032	6782.52	1.192	1721.39	0.498	2019/3/31 23:00
3098000032	6783.12	0	1721.47	0.436	2019/3/31 23:30
3098000032	6783.84	0	1721.55	0	2019/4/1 0:00

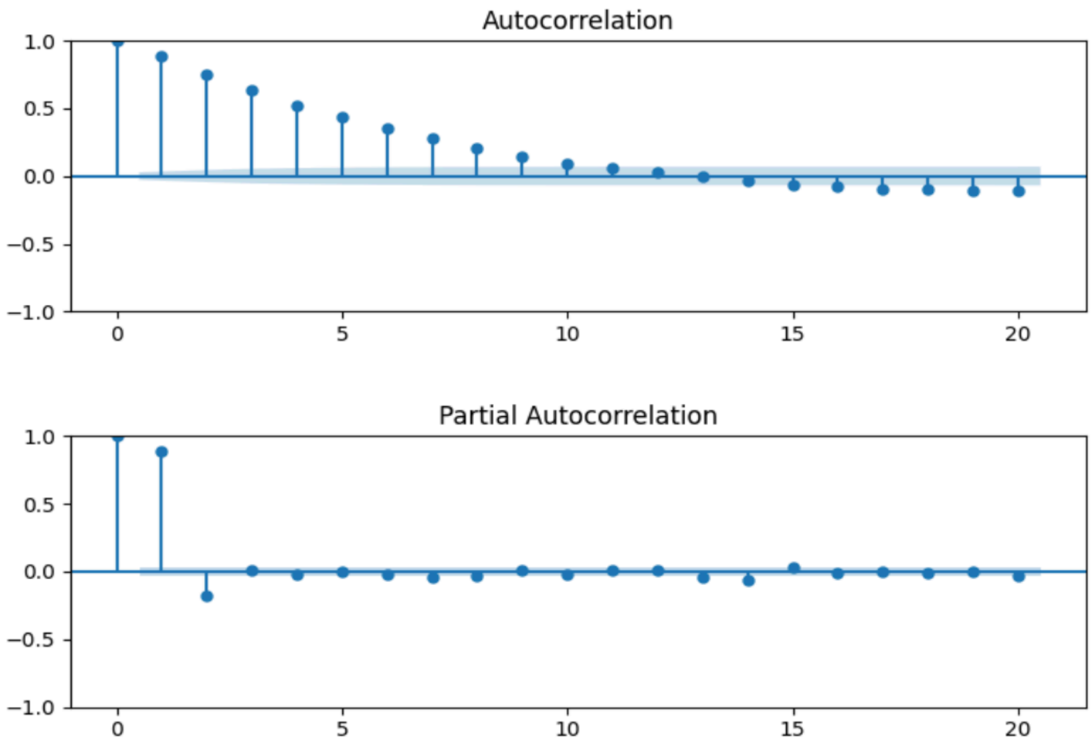
Figure 1.1
kW Versus Date



The data falls within the range of 0 to 5. No obvious trend or seasonality can be seen in this plot, which needs further analysis.

Figure 1.2

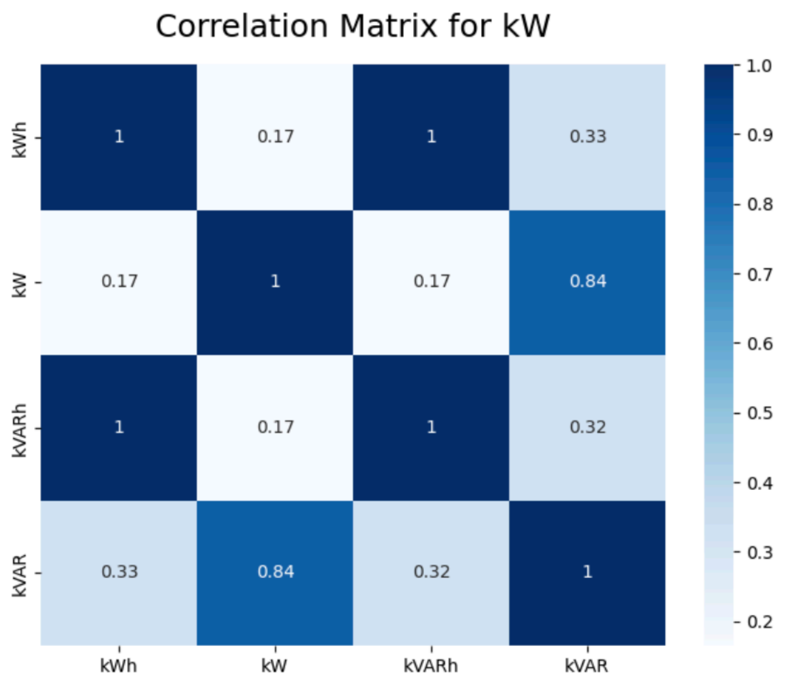
ACF/PACF of kW



The ACF shows a gradually decreasing trend while the PACF cuts immediately after one lag. Thus, the graphs suggest that ARMA (1, 0) or ARMA (2, 1) would be appropriate for the time series.

Figure 1.3

Correlation Matrix with the Pearson's Correlation Coefficient



The heatmap shows that ‘kVAR' highly correlate with the dependent variable kW’. Also, there is collinearity between ‘kVARh’ and ‘kWh’. They cannot independently predict the value of the dependent variable.

Stationary Test

Figure 2.1

ACF/PACF of kW

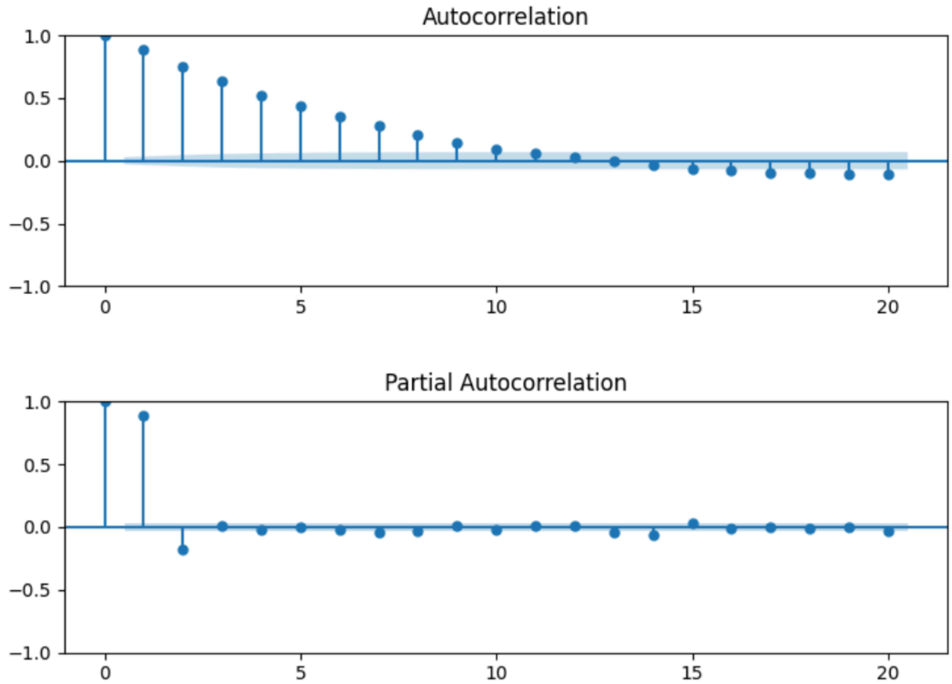


Figure 2.2

Rolling Mean and Variance of Raw Data

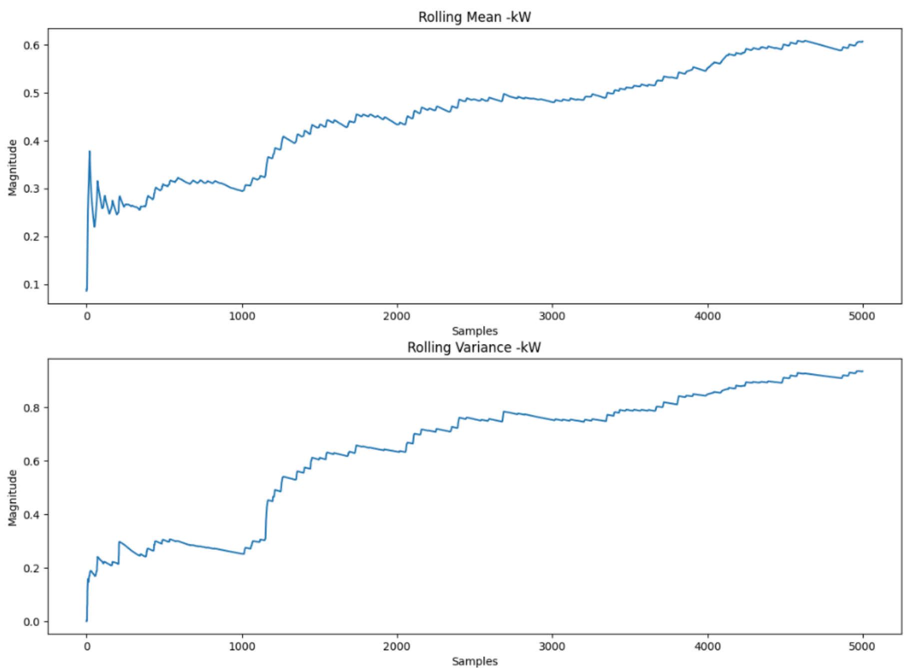


Figure 2.3

ADF and KPSS of Raw Data

```
ADF Statistic: -16.17
p-value: 0.00
Critical Values:
    1%: -3.43
    5%: -2.86
    10%: -2.57
Results of KPSS Test:
Test Statistic              2.108622
p-value                     0.010000
Lags Used                  37.000000
Critical Value (10%)        0.347000
Critical Value (5%)         0.463000
Critical Value (2.5%)       0.574000
Critical Value (1%)         0.739000
dtype: float64
```

The raw dataset is not stationary since the ACF shows a gradually decreasing trend. In addition, although the p-value of ADF Statistic is smaller than the critical value, the p-value of KPSS is also smaller than 0.05.

Figure 2.4

ACF/PACF of First Order Differencing

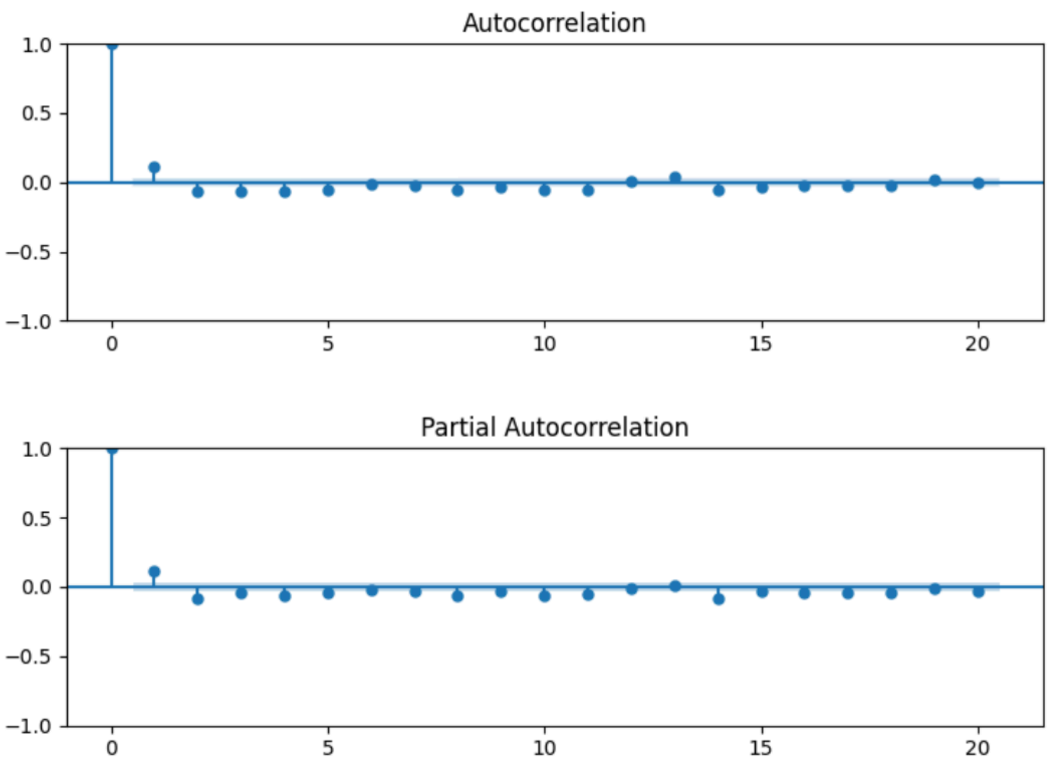


Figure 2.5

Rolling Mean and Variance of First Order Differencing

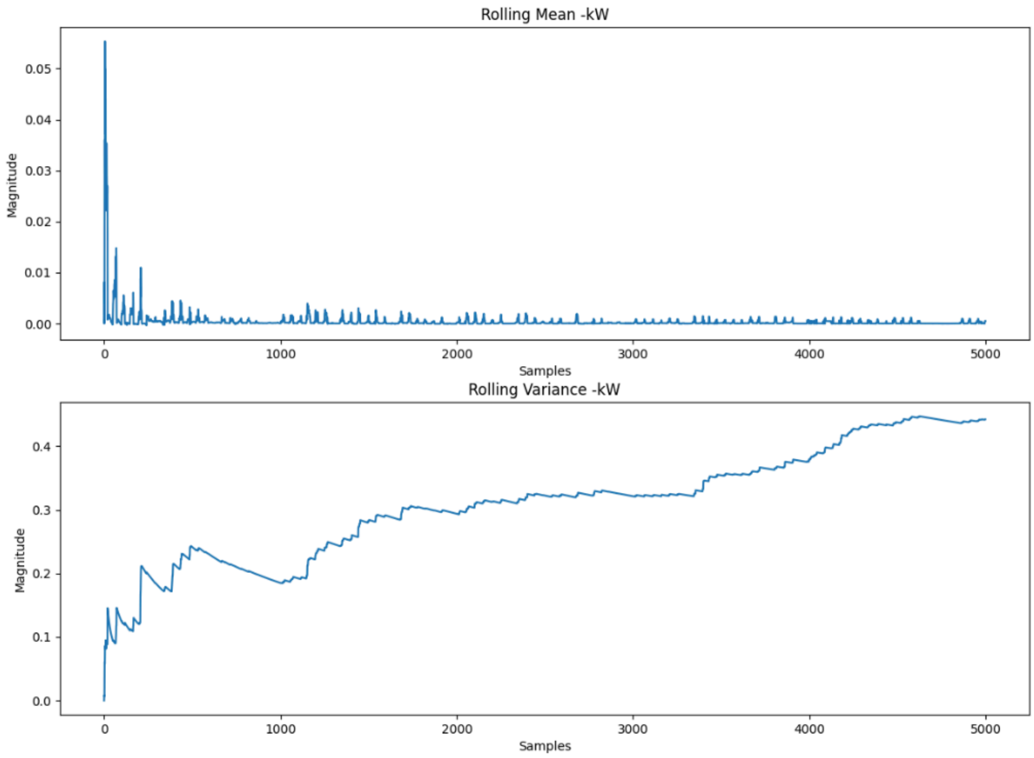


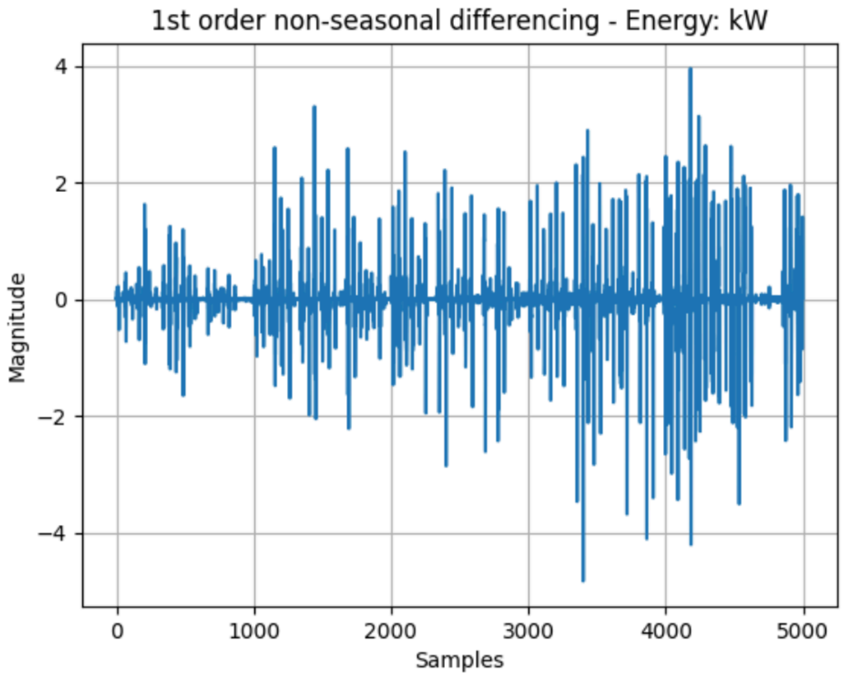
Figure 2.6

ADF and KPSS of First Order Differencing

```
ADF Statistic: -20.60
p-value: 0.00
Critical Values:
    1%: -3.43
    5%: -2.86
    10%: -2.57
Results of KPSS Test:
Test Statistic              0.007547
p-value                     0.100000
Lags Used                  35.000000
Critical Value (10%)        0.347000
Critical Value (5%)         0.463000
Critical Value (2.5%)       0.574000
Critical Value (1%)         0.739000
dtype: float64
```

Figure 2.7

1st Oder Differencing of kW

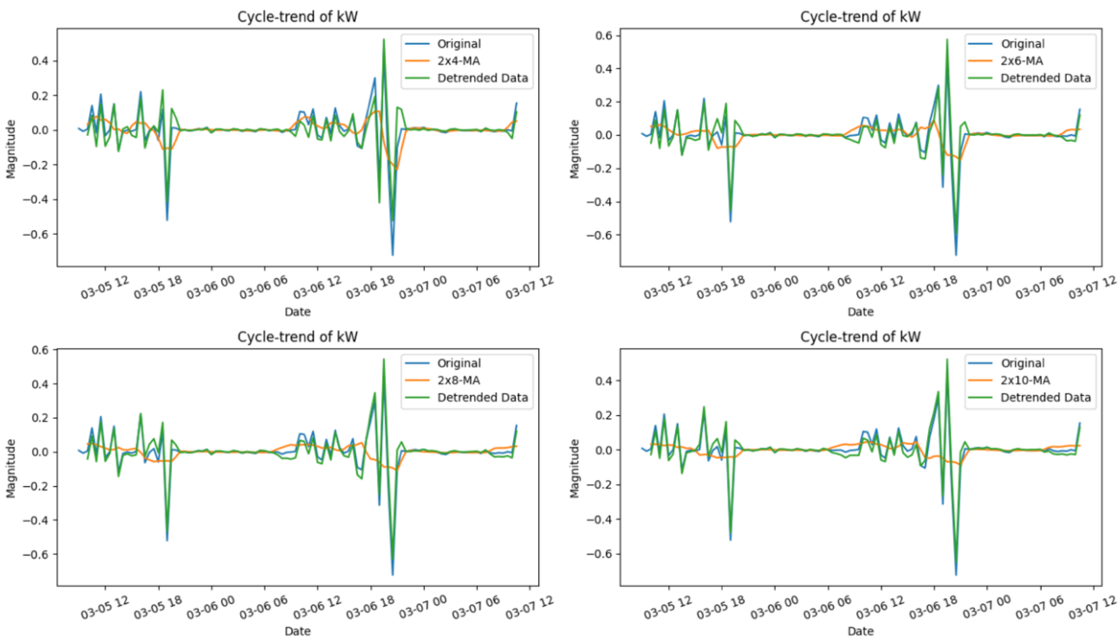


After performing the first difference transformation of the original raw dataset, the rolling mean and standard deviation are approximately horizontal. The p-value of ADF Statistic is smaller than the critical value, and the p-value of KPSS is greater than 0.05. The dataset is stationary after the transformation.

Time series Decomposition

Figure 3.1

Cycle-Trend of kW



There seems no trend-cycle in the data after differencing.

Figure 3.2

STL applied to kW (1)

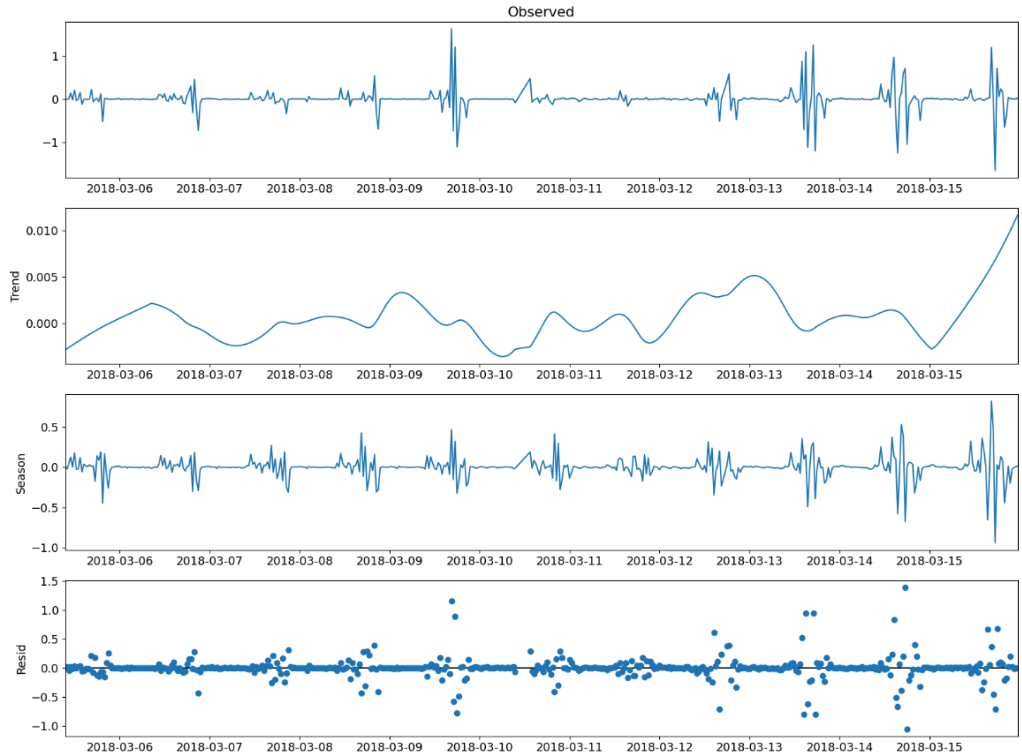
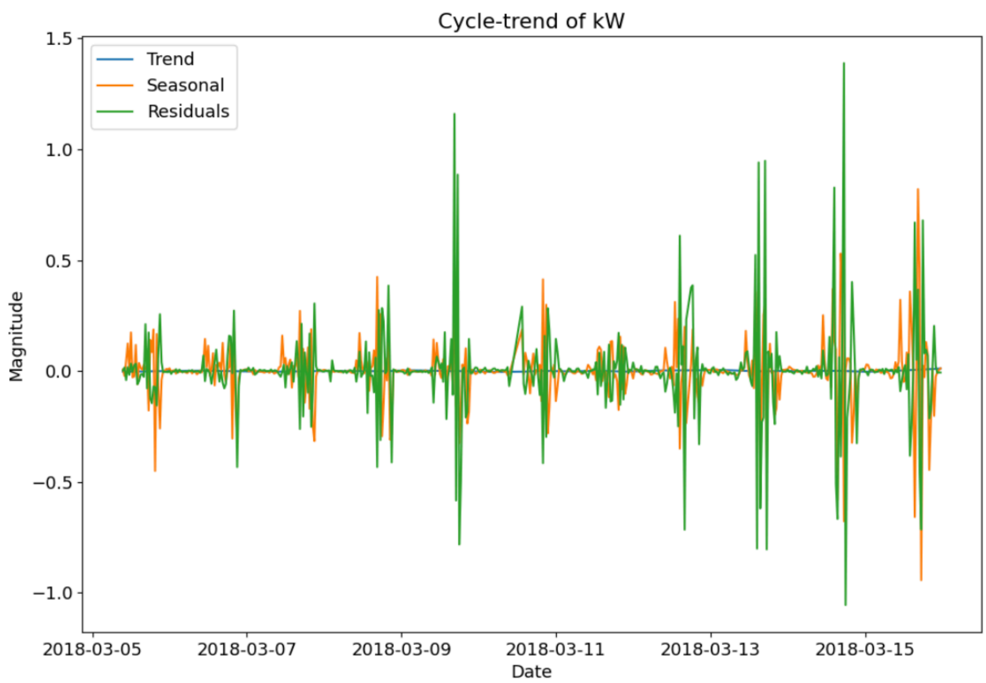


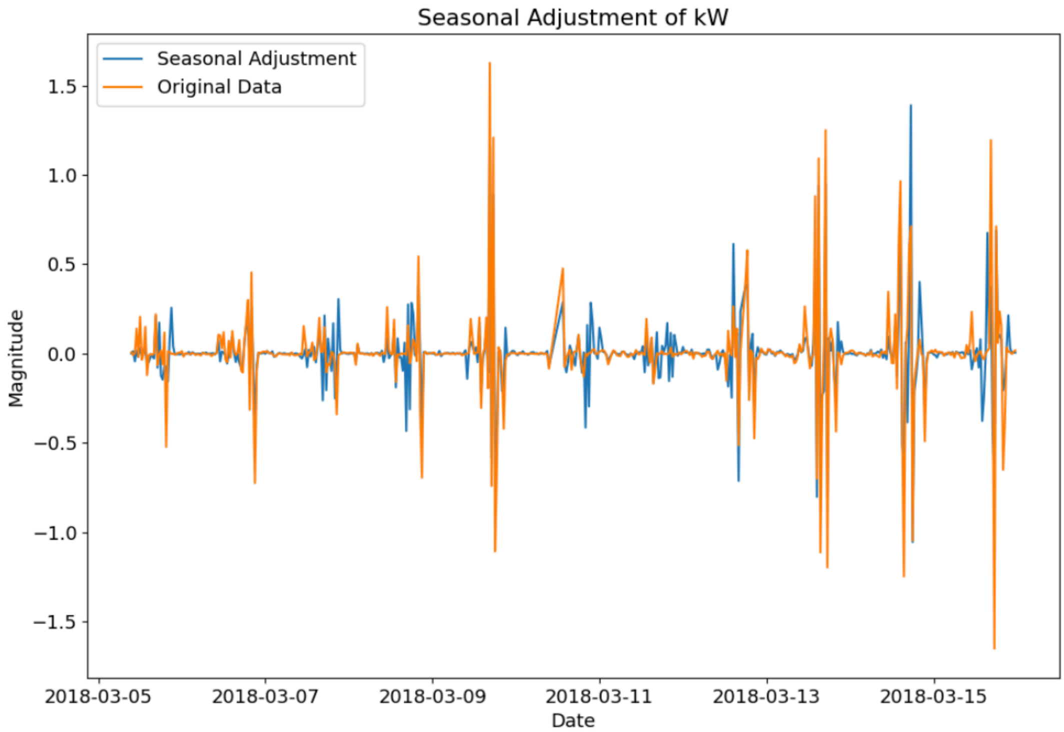
Figure 3.3

STL applied to kW (2)



The STL decomposition plots show that the seasonal component repeated patterns in the data that occur at fixed intervals and changed slightly over time. On the other hand, the trend component appears to oscillate around a fixed level. This suggests that there is little to no trend in the data.

Figure 3.4
Seasonal Adjusted of kW



After first-order differencing, the strength of trend is 0.0, and the strength of seasonality is 0.349. Both are very weak.

Feature Selection

Table 2
Feature Selection

	BACKWARD STEPWISE REGRESSION	VIF	PCA	SIMPLE LINEAR REGRESSION
Selected Features	kVAR, kWh	kVAR, kVARh	kWh, kVAR	kVAR
AIC	3943.71	4055.25	-7105.85	7313.22
BIC	3968.88	4067.83	-7086.30	7326.26
Adjusted R-squared	0.78	0.84	0.72	0.71

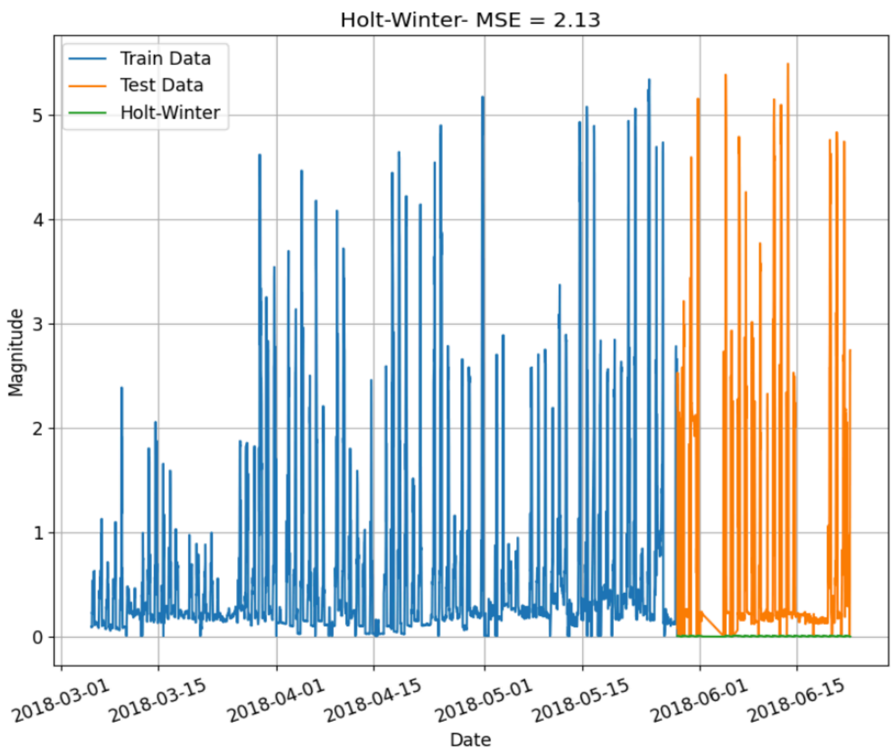
SVD Analysis (Singular Values)	[5.73950677e+04 1.51333983e+01]	[2.32236685e+04 1.53012682e+01]	[5.73950677e+04 1.51333983e+01]	NA
Condition Number (k)	3792.6093	1517.761	3792.6093	1
Collinearity	v	v	v	x

The features selected by Backward Stepwise Regression and PCA were [kVAR, kWh], and the other combination selected by VIF is [kVAR, kVARh]. However, since there is collinearity between the features, I selected only one feature, kVAR, to fit a Simple Linear Regression model and evaluate its performance using the adjusted R-squared metric. Regarding Linear Regression, more detailed information will be provided. The results showed that the model performed reasonably well, explaining 71% of the variation in the data.

Holt-Winters Method

Figure 4.1

Holt-Winter Method



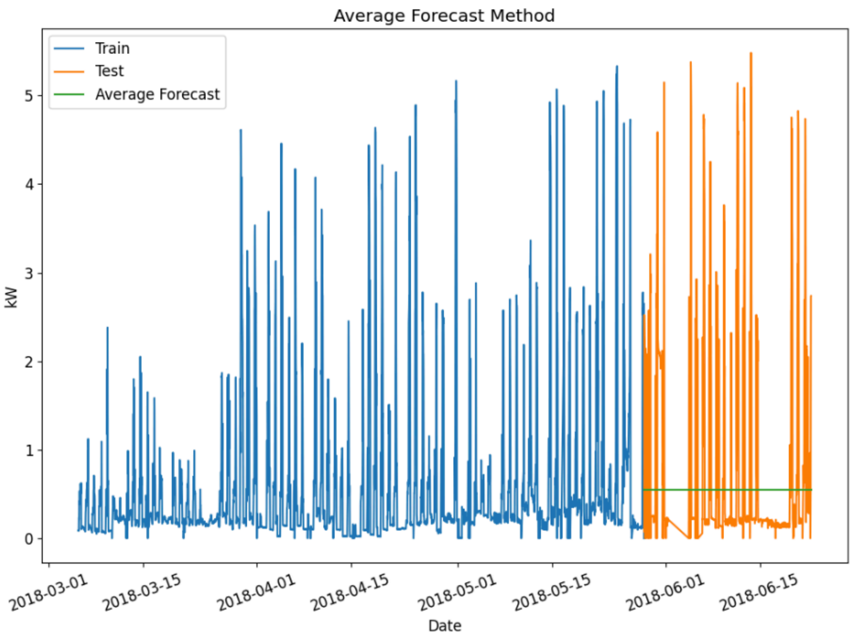
Ten different analysis methods were used in this project. The first is Holt-Winters method. The method involves decomposing a time series into three components: the trend, the seasonal component, and the residual or error component. It then applies a smoothing function to each component to forecast future values. The

MSE of Holt-Winters method is 2.13. The residual mean is 0.84. The residual variance is 1.43.

Base-models

Figure 4.2

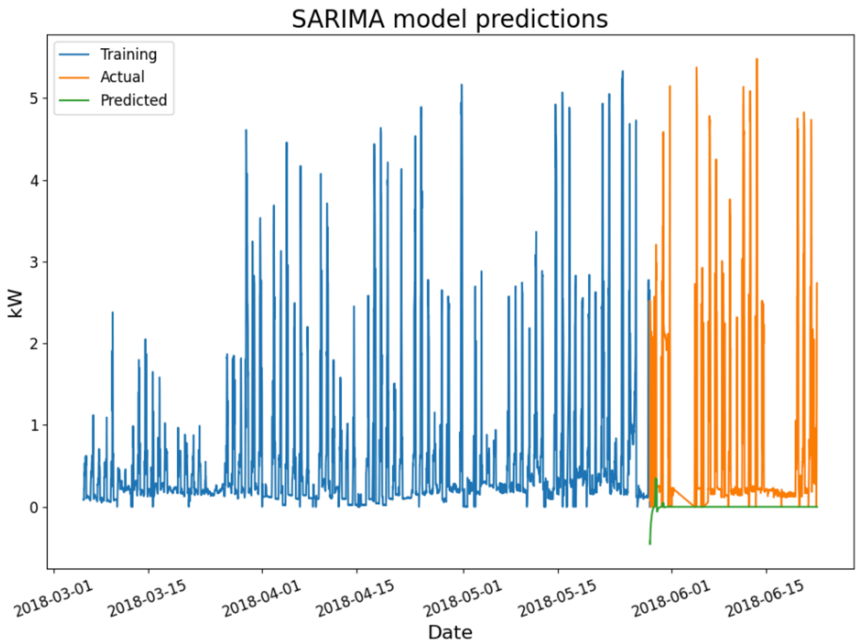
Average Forecast Method



RMSE: 1.23 Residual Mean: 0.28 Residual Variance: 1.43

Figure 4.3

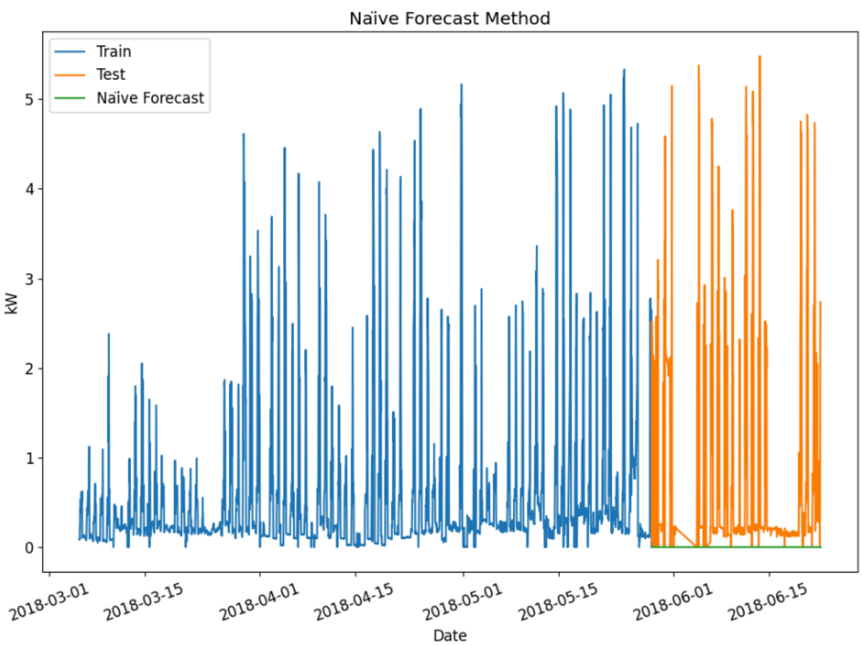
ARIMA(2, 0, 2) Model



RMSE: 0.42 Residual Mean: 0.00 Residual Variance: 0.179

Average Forecast Method uses the historical average of the time series data which are the training data to predict future values. A smaller RMSE indicates that the model is making better predictions, while a larger Residual Mean suggests that the model is not able to capture all the patterns in the data. The statistics shows us the ARIMA model performed better in terms of accurately predicting the outcome variable.

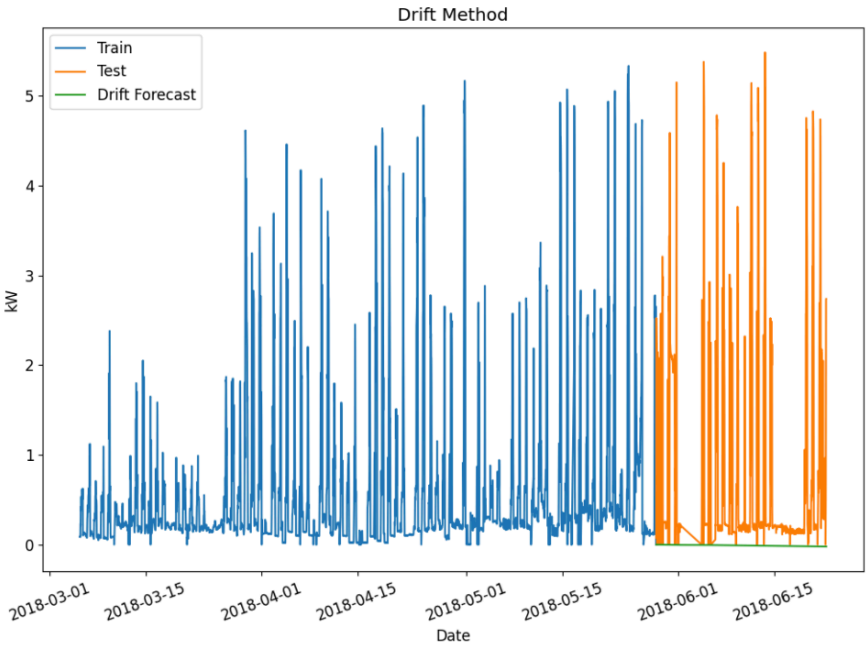
Figure 4.4
Naïve Forecast Method



RMSE: 1.46 Residual Mean: 0.83 Residual Variance: 1.43

Naïve Forecast Method assumes that there is no trend or seasonality in the data and that the future values will be the same as the most recent past value in the training set. ARIMA model also performed better than Naïve method.

Figure 4.5
Drift Method

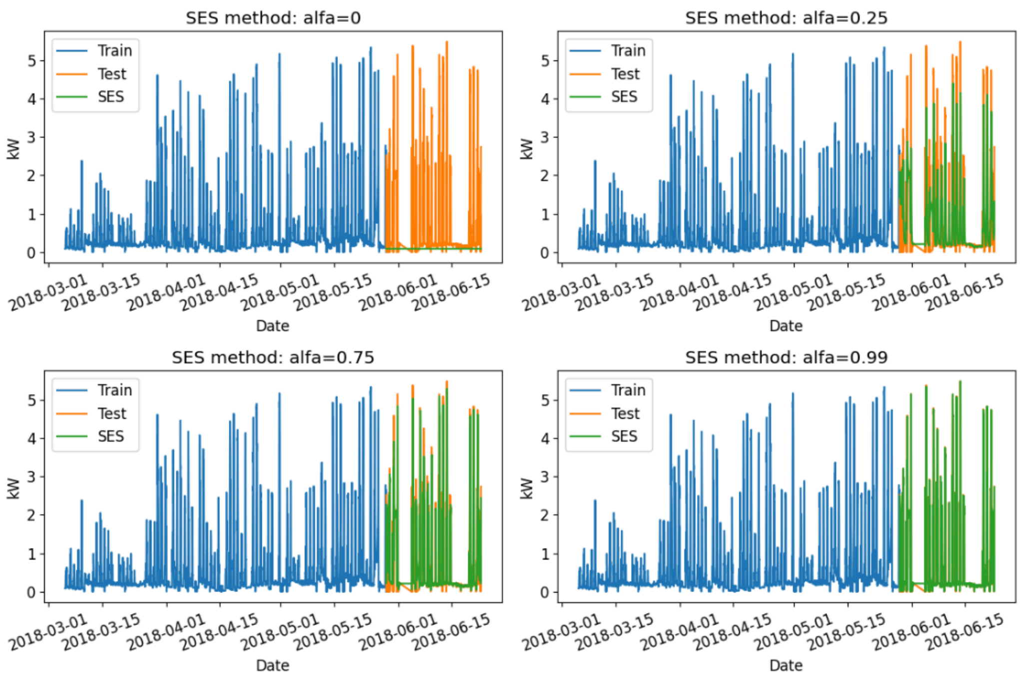


RMSE: 1.46 Residual Mean: 0.84 Residual Variance: 1.43

Drift Method assumes that the future value of a time series will be equal to the last observed value plus the average change observed in previous time periods. It is particularly helpful when forecasting time series data with an increasing or decreasing trend. In this case, ARIMA model performed better in terms of accurately predicting than drift method.

Figure 4.6

SES Method



RMSE: 1.41~1.69 Residual Mean: -0.001~0.75 Residual Variance: 1.43~2.85

SES method assumes that the time series has no trend or seasonality and uses a weighted average of past observations to predict future values. The weights decay exponentially as the observations get older, with more weight being given to recent observations. The value of α determines the rate at which the weights decay and has a range of 0 to 1. When α is closer to 1, the method gives more weight to recent observations and reacts more quickly to changes in the data. ARIMA model still performed better in terms of accurately predicting.

Linear Regression

Figure 5.1

OLS Regression Results (1)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     kW   R-squared:                       0.780
Model:                            OLS   Adj. R-squared:                  0.780
Method:                 Least Squares   F-statistic:                     7104.
Date:                Tue, 09 May 2023   Prob (F-statistic):               0.00
Time:                        00:33:24   Log-Likelihood:                -1985.1
No. Observations:                4000   AIC:                             3976.
Df Residuals:                    3997   BIC:                             3995.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.1046      0.012     -9.042      0.000      -0.127      -0.082
kVAR           3.9198      0.033    117.245      0.000       3.854       3.985
kVARh         -0.0012   5.05e-05    -24.434      0.000      -0.001      -0.001
==============================================================================
Omnibus:                     1596.327   Durbin-Watson:                   0.319
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13804.747
Skew:                           1.667   Prob(JB):                         0.00
Kurtosis:                      11.469   Cond. No.                     1.25e+03
==============================================================================
                             Test for Constraints
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
c0            -0.1046      0.012     -9.042      0.000      -0.127      -0.082
c1             3.9198      0.033    117.245      0.000       3.854       3.985
c2            -0.0012   5.05e-05    -24.434      0.000      -0.001      -0.001
==============================================================================
<F test: F=7104.386860063783, p=0.0, df_denom=4e+03, df_num=2>
```

RMSE: 0.83 Residuals variance: 0.53 Residuals mean: 0.40

The sixth method is linear regression. Although we have known that there is a collinearity problem when doing feature selection, I still used the regression model to confirm it. The output shows the results of the Ljung-Box test for autocorrelation of the residuals. The null hypothesis is that there is no autocorrelation up to 20 lags. The p-value is 0, indicating that there is strong evidence against the null hypothesis, and therefore, there is autocorrelation in the residuals. Also, the condition number is large. This indicates that there is strong multicollinearity.

Figure 5.2

OLS Regression Results (2)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     kW   R-squared:                       0.748
Model:                            OLS   Adj. R-squared:                  0.748
Method:                 Least Squares   F-statistic:                 1.185e+04
Date:                Tue, 09 May 2023   Prob (F-statistic):               0.00
Time:                        00:36:25   Log-Likelihood:                -2263.5
No. Observations:                4000   AIC:                             4531.
Df Residuals:                    3998   BIC:                             4544.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.2687      0.010    -26.589      0.000      -0.288      -0.249
kVAR           3.6109      0.033    108.838      0.000       3.546       3.676
==============================================================================
Omnibus:                     1249.495   Durbin-Watson:                   0.256
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             8080.724
Skew:                           1.329   Prob(JB):                         0.00
Kurtosis:                       9.436   Cond. No.                         5.19
==============================================================================
                             Test for Constraints
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
c0            -0.2687      0.010    -26.589      0.000      -0.288      -0.249
c1             3.6109      0.033    108.838      0.000       3.546       3.676
==============================================================================
<F test: F=11845.80580827514, p=0.0, df_denom=4e+03, df_num=1>
```

RMSE: 0.74 Residuals variance: 0.54 Residuals mean: -0.04

To address the issue of collinearity among independent variables, I attempted to sum the variables and treat them as one independent variable. However, the condition number, k, was still above 1000. Despite this, I noticed that the adj. R square for kVAR was above 70%. As a result, I decided to conduct a simple linear regression analysis using kVAR as the independent variable. Although there is no collinearity. The RMSE also lower than the previous multiple linear regression. The residual is NOT white, indicating that there is autocorrelation in the residuals.

ARMA, ARIMA and SARIMA Model

I use three methods to estimate the ARMA order. First, for the autocorrelation function, it shows a gradually decreasing trend while the PACF cuts immediately after one lag. Thus, the graphs suggest that ARMA (1, 0) or ARMA (2, 1) would be appropriate for the time series. Second, by plotting the AIC and BIC of different ARMA order. We can see the order (2, 0, 2) has the smallest AIC and BIC. So, this order would be used to build the ARIMA/SARIMA model. Third, by GPAC table, we can estimate the ARMA order. It presents a pattern in the second column of the table.

Figure 6.1

AIC and BIC of ARIMA Model

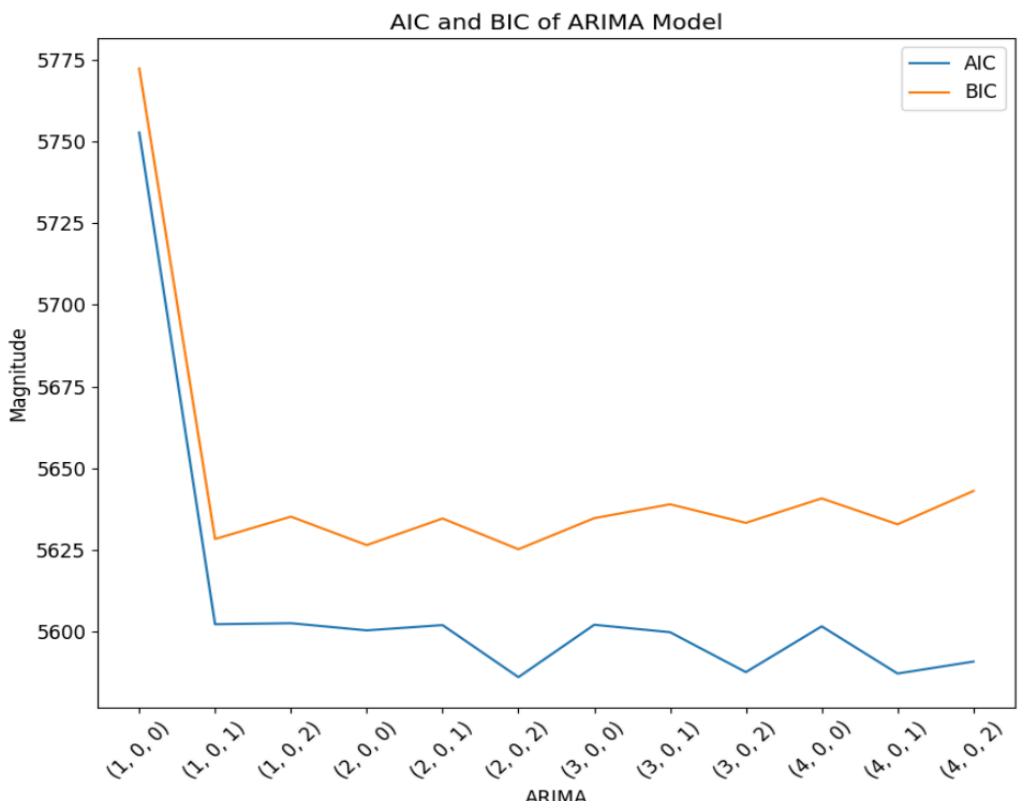


Figure 6.2

GPAC Table

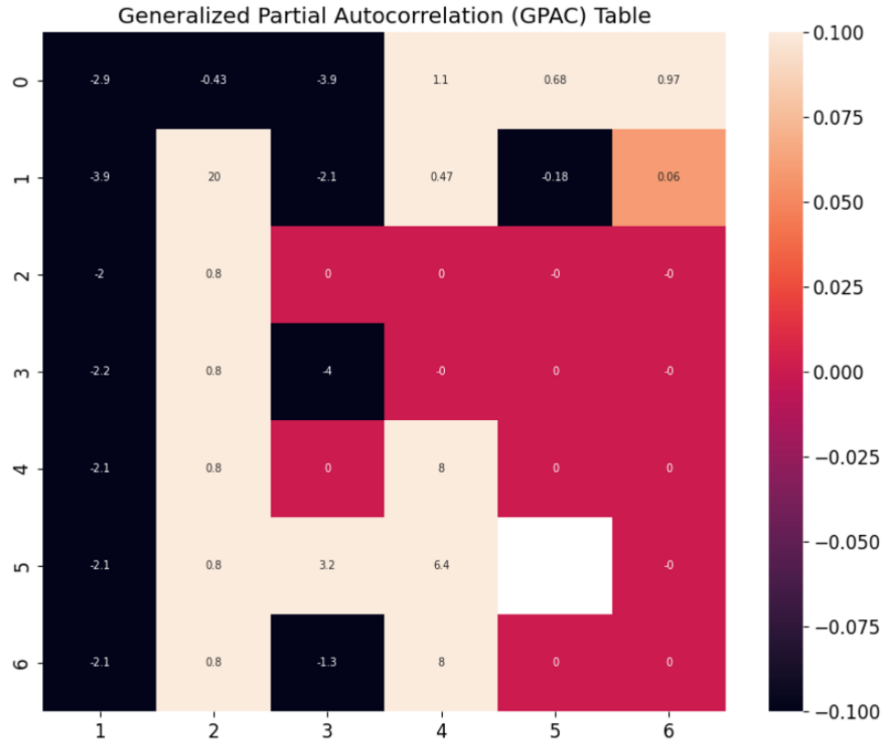


Figure 6.3

ARMA Model Summary

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                     kW   No. Observations:                 5000
Model:                 ARIMA(2, 0, 2)   Log Likelihood               -2787.078
Date:                Tue, 09 May 2023   AIC                           5586.156
Time:                        03:17:25   BIC                           5625.260
Sample:                             0   HQIC                          5599.861
                               - 5000
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.6087      0.064      9.577      0.000       0.484       0.733
ar.L1          1.7778      0.039     45.997      0.000       1.702       1.854
ar.L2         -0.7970      0.032    -24.645      0.000      -0.860      -0.734
ma.L1         -0.7352      0.039    -18.673      0.000      -0.812      -0.658
ma.L2         -0.1495      0.012    -12.562      0.000      -0.173      -0.126
sigma2         0.1785      0.001    124.765      0.000       0.176       0.181
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):             95221.64
Prob(Q):                              0.94   Prob(JB):                         0.00
Heteroskedasticity (H):               4.60   Skew:                             0.56
Prob(H) (two-sided):                  0.00   Kurtosis:                        24.35
===================================================================================
```

ARMA(2,2) AIC: 559 BIC: 563 RMSE: 0.42 Residual Mean: 0.00 Residual Variance: 0.179

Figure 6.4

ARIMA Model Summary

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                     kW   No. Observations:                 5000
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -2794.049
Date:                Tue, 09 May 2023   AIC                           5598.097
Time:                        03:18:29   BIC                           5630.682
Sample:                             0   HQIC                          5609.518
                               - 5000
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.0011      0.045     22.284      0.000       0.913       1.089
ar.L2         -0.1409      0.040     -3.549      0.000      -0.219      -0.063
ma.L1         -0.9573      0.044    -21.782      0.000      -1.043      -0.871
ma.L2         -0.0411      0.044     -0.935      0.350      -0.127       0.045
sigma2         0.1789      0.001    150.598      0.000       0.177       0.181
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):             94831.67
Prob(Q):                              0.97   Prob(JB):                         0.00
Heteroskedasticity (H):               4.66   Skew:                             0.47
Prob(H) (two-sided):                  0.00   Kurtosis:                        24.32
===================================================================================
```

ARIMA(2,0,2) AIC: 560 BIC: 563 RMSE: 0.42 Residual Mean: 0.01 Residual Variance: 0.18

Figure 6.5

SARIMA Model Summary

```
                                     SARIMAX Results
=========================================================================================
Dep. Variable:                                kW   No. Observations:                 5000
Model:             SARIMAX(2, 0, 2)x(1, 0, [], 48)   Log Likelihood               -2819.123
Date:                           Tue, 09 May 2023   AIC                           5650.247
Time:                                   03:19:03   BIC                           5689.350
Sample:                                        0   HQIC                          5663.952
                                          - 5000
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.3727      0.503      0.741      0.459      -0.613       1.358
ar.L2          0.4538      0.444      1.023      0.306      -0.416       1.323
ma.L1          0.6762      0.504      1.342      0.180      -0.311       1.664
ma.L2          0.0730      0.088      0.833      0.405      -0.099       0.245
ar.S.L48       0.1241      0.008     15.725      0.000       0.109       0.140
sigma2         0.1807      0.001    159.442      0.000       0.179       0.183
===================================================================================
Ljung-Box (L1) (Q):                   1.34   Jarque-Bera (JB):             98867.05
Prob(Q):                              0.25   Prob(JB):                         0.00
Heteroskedasticity (H):               4.63   Skew:                             0.22
Prob(H) (two-sided):                  0.00   Kurtosis:                        24.78
===================================================================================
```

SARIMA(2,0,2) D = 48 AIC: 565 BIC: 569 RMSE: 0.43 Residual Mean: 0.05 Residual Variance: 0.18

What can be seem from the summaries of ARMA, ARIMA, and SARIMA models is that the AIC, BIC, RMSE, and Residual Mean of ARMA are smaller than the values of ARIMA and SARIMA. Which indicates that the ARMA (2,2) model is a better fit for the data.

Levenberg Marquardt algorithm

Table 3

ARMA model by Levenberg Marquardt Algorithm

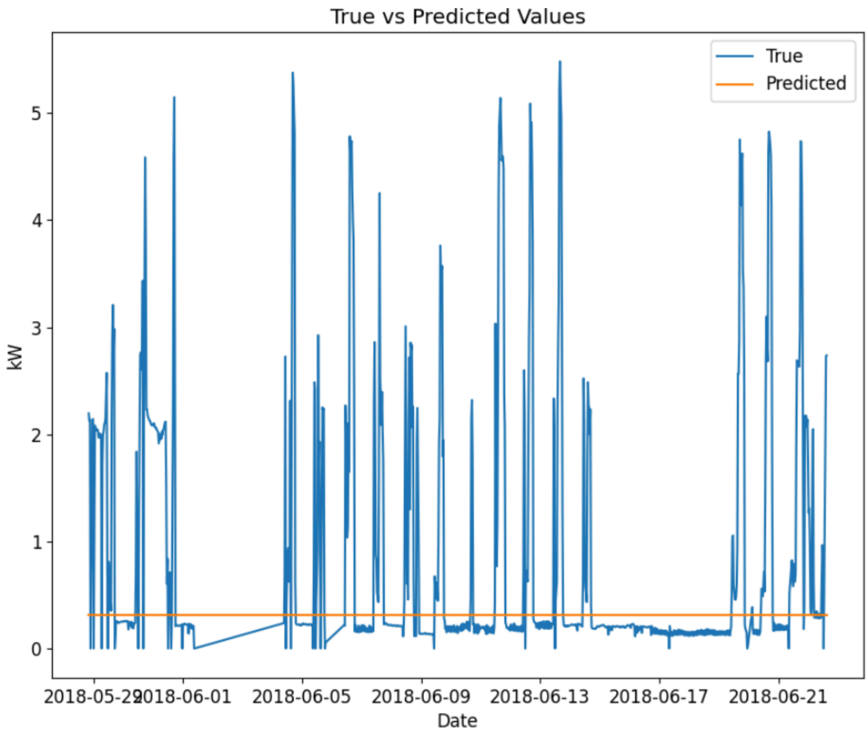
	Parameter Estimates	Standard Deviation	Confidence Interval 0	Confidence Interval 1
const	0.609	0.064	0.484	0.733
ar.L1	1.778	0.039	1.702	1.854
ar.L2	-0.797	0.032	-0.860	-0.734
ma.L1	-0.735	0.039	-0.812	-0.658
ma.L2	-0.149	0.012	-0.173	-0.126
sigma2	0.178	0.001	0.176	0.181

Above table shows the estimate ARMA model parameters using the Levenberg Marquardt algorithm, including the parameter estimates, the standard deviation of the parameter estimates and the confidence intervals.

Deep Learning Model

Figure 7.1

Multivariate LSTM Model



RMSE: 1.31 Residual Mean: 0.54 Residual Variance: 1.43

Multivariate LSTM model is based on a Long Short-Term Memory algorithm, which is a type of recurrent neural network (RNN) used for processing sequential data. The output of the model is for 5 step ahead prediction.

Final Model selection

Table 4 Final Model Selection

	Holt-Winters method	Average	Naïve	Drift	SES	ARIMA (2, 0, 2)	ARIMA (2, 1, 2)	SARIMA (2, 0, 2)48	LSTM	Linear Regression
RMSE	2.13	1.2284	1.4575	1.4631	1.4101	0.42	0.42	0.43	1.299	0.7376
Residual Mean	0.8359	0.2835	0.8342	0.8449	0.7482	0	0.0088	0.0533	0.5382	-0.0393
Residual Variance	1.4276	1.4286	1.4286	1.4266	1.4286	0.1785	0.1789	0.1779	1.4260	0.5425

RMSE (Root Mean Squared Error) is a measure of the accuracy of a model's predictions, therefore, a smaller RMSE value means that the predicted values are closer to the actual values. The residual mean represents the average difference between the predicted values and the actual values, so a smaller residual mean indicates that the model is making more accurate predictions on average. The residual variance represents the spread or variability of the errors, so a smaller residual variance indicates that the errors are more consistent or predictable. In general, a good model will have a combination of a low RMSE, small residual mean, and small residual variance. As a result, the best model between these 10 different methods is ARMA(2, 2) and the forecast function is shown below.

Forecast Function

ARIMA(2,0,2):

$\hat{Y}(t) = 0.6087 + 1.7778y(t-1) - 0.7970y(t-2) - 0.7352e(t-1) - 0.1495e(t-2)$

Summary

To sum up, ARMA(2, 2) model fits the best in this dataset. Also, I found that the performances of base-model are similar in this case. However, there are some limitations. For example, there are too few independent variables in the dataset, which may limit its predictive power. Moreover, there is a problem of collinearity among the independent variables, which can lead to unstable and unreliable estimates of the regression coefficients, making it difficult to interpret the effect of each individual predictor variable on the outcome variable. In addition, collinearity can lead to overfitting of the model, which can result in poor predictions on new data. Addressing these issues can lead to improved performance of the model.

Reference

Zahid, A. (2020, January 24). Smart Meter Energy(kW) demand forecasting. Kaggle. Retrieved May 4, 2023, from https://www.kaggle.com/datasets/asimzahid/smart-meter-energykw-demand-forecasting