Progress Report II

1. Resampling Methods

In Chapter 6, we discussed two common resampling methods: cross-validation and bootstrapping. Since my dataset is a time series, I applied a resampling technique to aggregate the data into hourly intervals. This step was essential to simplify the dataset and focus on trends and patterns at a higher level of granularity:

- Original Dataset Shape: (19735, 28)
- After Resampling: (3290, 28)

By resampling, I reduced the dataset size significantly while preserving the key temporal patterns, making it more suitable for time series analysis and modeling.

2. Time Series Decomposition

To better understand the underlying patterns in the Appliances Energy Consumption time series, I applied a seasonal decomposition technique. This process breaks down the time series into three key components:

- Trend: Captures the long-term movement or direction in the data.
- Seasonality: Identifies repeating patterns or cycles hours
- Residuals: Represents the remaining noise or unexplained variation after removing the trend and seasonality.

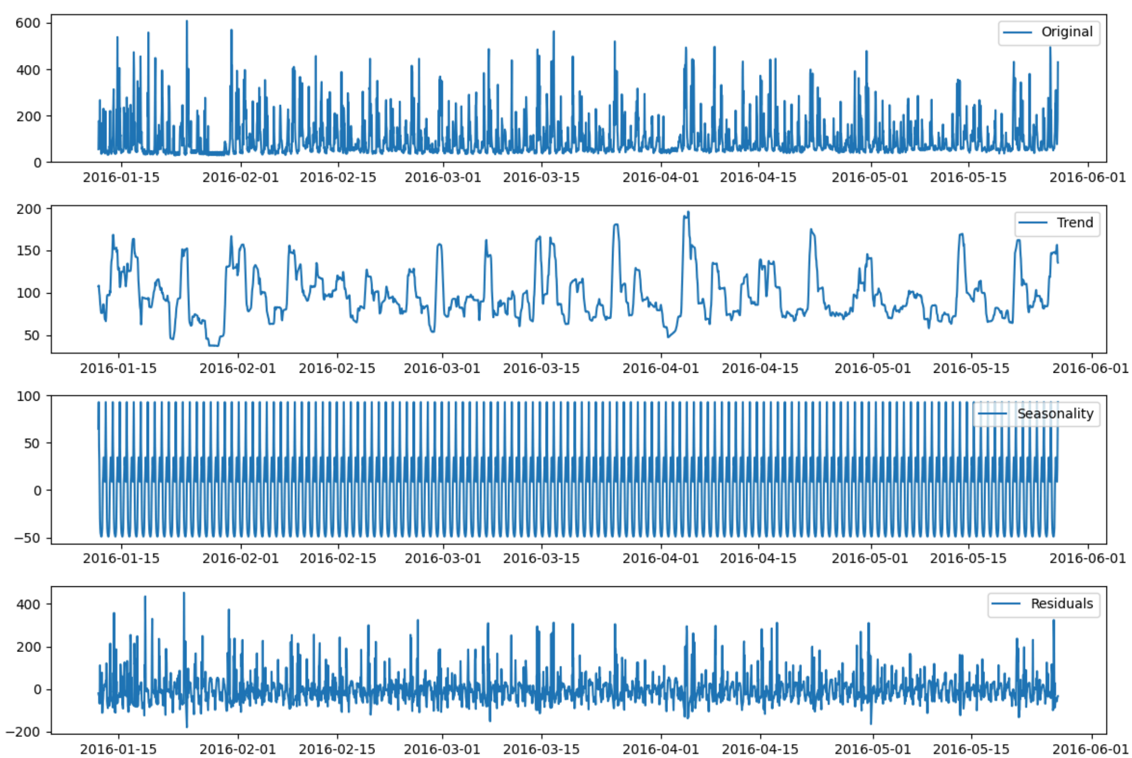

Figure 1: Appliance Decomposition

I used the seasonal_decompose function from the statsmodels library with an additive model. Since the data was resampled to hourly intervals, I specified a period of 24 to capture daily seasonality. The decomposition was performed as follows:

Code: “decomposition = seasonal_decompose(df_hourly['Appliances'], model='additive', period=24)”

Outcome:

- It shows a **slight up-and-down pattern**, with noticeable bumps during certain periods (e.g., around mid-March and April).

3. Check Stationarity

I applied the Augmented Dickey-Fuller (ADF) test to assess whether the Appliances time series is stationary.

```
ADF Statistic: -8.948888280256888
p-value: 8.833753129594426e-15
Critical Values:
1% -3.432357502010421
5% -2.862426994644342
10% -2.567242166152283
```

Interpretation:

- The ADF statistic is much lower than all critical values.
- The p-value is significantly less than 0.05, indicating strong evidence to reject the null hypothesis of non-stationarity.
- The Appliances time series is stationary, and no additional differencing is needed before modeling.

4. Split Data

I split the data into 80% for training and 20% for testing, following a standard practice. This allows me to train the model on the training set and evaluate its performance on the testing set.

5. Model Train

There are various techniques available for training time series models. I chose SARIMAX because:

- Stationarity: SARIMAX is well-suited for stationary data, and as shown above, I confirmed that my dataset is stationary using the Augmented Dickey-Fuller (ADF) test.
- Seasonality: SARIMAX can handle seasonal patterns effectively, which is crucial for my dataset that exhibits daily seasonality.
- Exogenous Variables: SARIMAX allows the inclusion of external predictors (exogenous variables), which can enhance the model's accuracy if additional features are available.

Considering these factors, I determined that SARIMAX was a suitable choice for my time series analysis and forecasting. While I have not yet explored other techniques to enhance the model, I plan to fine-tune its parameters to improve its performance. Although the model is learning, its current performance is not satisfactory.

SARIMAX Results

```
==============================================================================
==================
Dep. Variable:                    Appliances   No. Observations:               2632
Model:           SARIMAX(1, 1, 1)x(1, 1, 1, 24)   Log Likelihood            -14442.018
Date:                       Wed, 23 Apr 2025   AIC                       28894.035
Time:                               05:01:03   BIC                       28923.315
Sample:                           01-11-2016   HQIC                      28904.648
                                - 04-30-2016
Covariance Type:                         opg
==============================================================================
======
             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         0.4820      0.012     41.305      0.000       0.459       0.505
ma.L1        -1.0000      0.715     -1.398      0.162      -2.402       0.402
ar.S.L24      0.0035      0.011      0.325      0.745      -0.018       0.025
ma.S.L24     -0.9627      0.006   -149.926      0.000      -0.975      -0.950
sigma2     4150.3549   2986.627      1.390      0.165   -1703.326      1e+04
==============================================================================
===========
Ljung-Box (L1) (Q):                  32.34   Jarque-Bera (JB):              9719.20
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.72   Skew:                             2.08
Prob(H) (two-sided):                  0.00   Kurtosis:                        11.55
==============================================================================
===========
```

Figure 2: Output SARIMAX Model

6. Prediction Model

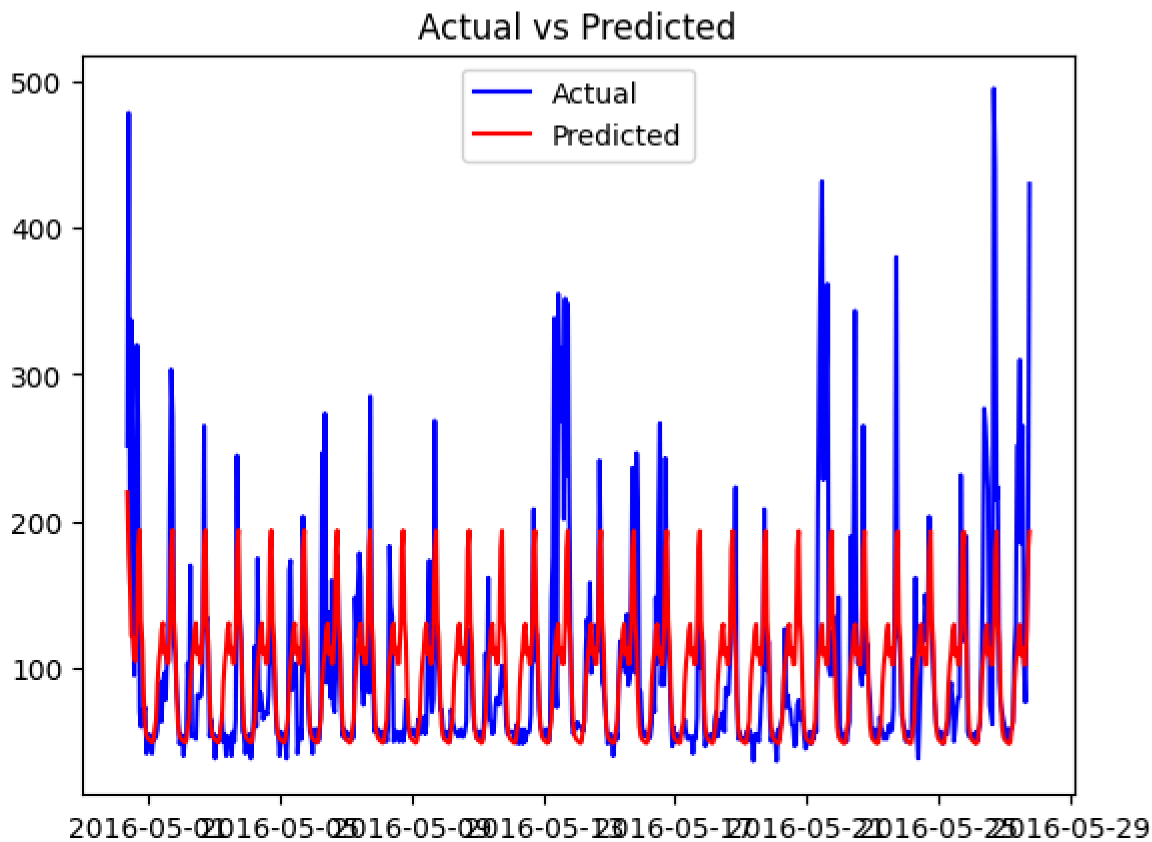

Figure 3: Prediction output