IE360 PROJECT

Statistical Forecasting and Time Series

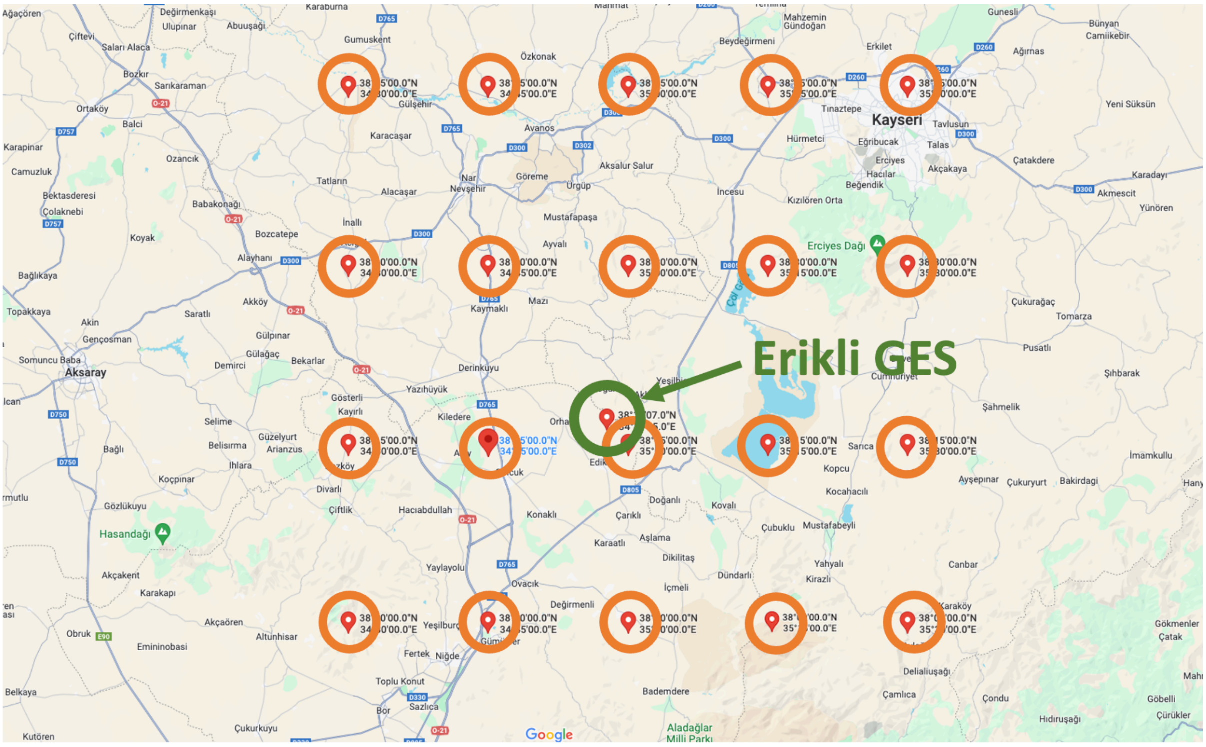
Ali Burak Öztürk	2019402048
Ömer Coşkun	2019402277
Ferhat Peynirci	2019402105

Introduction

Problem Description

Solar energy has emerged as a crucial component in the transition towards sustainable and renewable energy sources. Accurate forecasting of solar power production is essential for efficient grid management and energy trading. This project focuses on providing hourly solar power predictions for Edikli Güneş Enerjisi Santrali (GES), a solar power plant located in Niğde, Turkey. The primary objective is to predict the hourly solar power output for the next day using weather data from nearby grid points.

Edikli GES is geographically positioned at 38.29° North latitude and 34.97° East longitude. For this task, we have access to historical weather data and solar power production records. The challenge is to develop a predictive model that can leverage these datasets to provide accurate hourly forecasts.



Erikli GES and weather observation points on map

Distance to Erikli GES		
Latitude	**Longitude**	**Distance (km)**
37.75	34.5	71.97

37.75	34.75	62.28
37.75	35	59.51
37.75	35.25	64.56
37.75	35.5	75.88
38	34.5	51.45
38	34.75	36.74
38	35	31.84
38	35.25	40.47
38	35.5	56.77
38.25	34.5	40.66
38.25	34.75	**19.00**
38.25	35	**5.11**
38.25	35.25	**25.46**
38.25	35.5	47.20
38.5	34.5	46.91
38.5	34.75	30.21
38.5	35	**24.06**
38.5	35.25	34.62
38.5	35.5	52.66
38.75	34.5	65.48
38.75	34.75	54.81
38.75	35	51.69
38.75	35.25	57.35
38.75	35.5	69.70

The observation points and their distance to Erikli GES

To decide on which locations to consider to come up with a reasonable forecast for the location of Edikli GES, we had several options:

1. Pick the closest, since it is only 5.11 km away.

2. Pick the 4 closest, aggregate their features by:
    a. Taking the average
    b. Taking the weighted average based on their distance to Edikli GES
3. Take all of them, aggregate their features by:
    c. Taking the average
    d. Taking the weighted average based on their distance to Edikli GES
4. Not aggregating the data, using every 25 observation points separately for forecasting, determining the best one.

Even though each method could have its own pros and cons, we decided to implement the first one since it was the closest by far among other locations.

Summary of the Proposed Approach

To tackle the problem of hourly solar power prediction for Edikli GES, we adopted a data-driven approach involving several steps. After experimenting with various methodologies, we refined our approach as follows:

Feature Engineering

Handling Night Hours:

- Instead of adding the feature "hour" for the hours between 22:00 and 04:00, we introduced a binary feature is_night to indicate night hours (0 for day, 1 for night). This simplification was based on the observation that solar power production is always zero during these hours, making it unnecessary to complicate the model with these hours.

Simplifying Hourly Features:

- We initially considered including "hour" as a feature, but it led to an overly complex model with numerous dummy variables (including months). To address this, we decided against using Fourier transforms for daily predictions. Instead, we focused on making hourly predictions by training separate time series models for each hour of the day. This approach allowed us to use ARIMA models more effectively and resulted in better results. Even though this approach resulted in longer training times due to the need to train multiple models, the use of distinct ARIMA models for each hour allowed for more precise and accurate predictions.

Descriptive Analysis of the Given Data

Weather Data

The provided weather data includes several variables measured at 25 grid points around the Edikli GES power plant. These variables are crucial for understanding the factors that influence solar power production. The key weather variables are:

- DSWRF_surface (Downward Shortwave Radiation Flux): This variable measures the amount of solar radiation reaching the Earth's surface, which is directly related to solar power production.
- USWRF_top_of_atmosphere and USWRF_surface: These variables measure the upward shortwave radiation flux at the top of the atmosphere and the surface, respectively.
- DLWRF_surface (Downward Longwave Radiation Flux): This measures the longwave radiation reaching the surface from the atmosphere.
- TCDC_low.cloud.layer, TCDC_middle.cloud.layer, TCDC_high.cloud.layer, TCDC_entire.atmosphere: These variables represent the total cloud cover at different altitudes, affecting solar radiation levels.
- CSNOW_surface (Categorical Snow): A binary variable indicating the presence of snow, which can cover solar panels and reduce their efficiency.
- TMP_surface (Temperature): This measures the surface temperature, which affects the efficiency of solar panels. High temperatures can decrease solar panel efficiency.

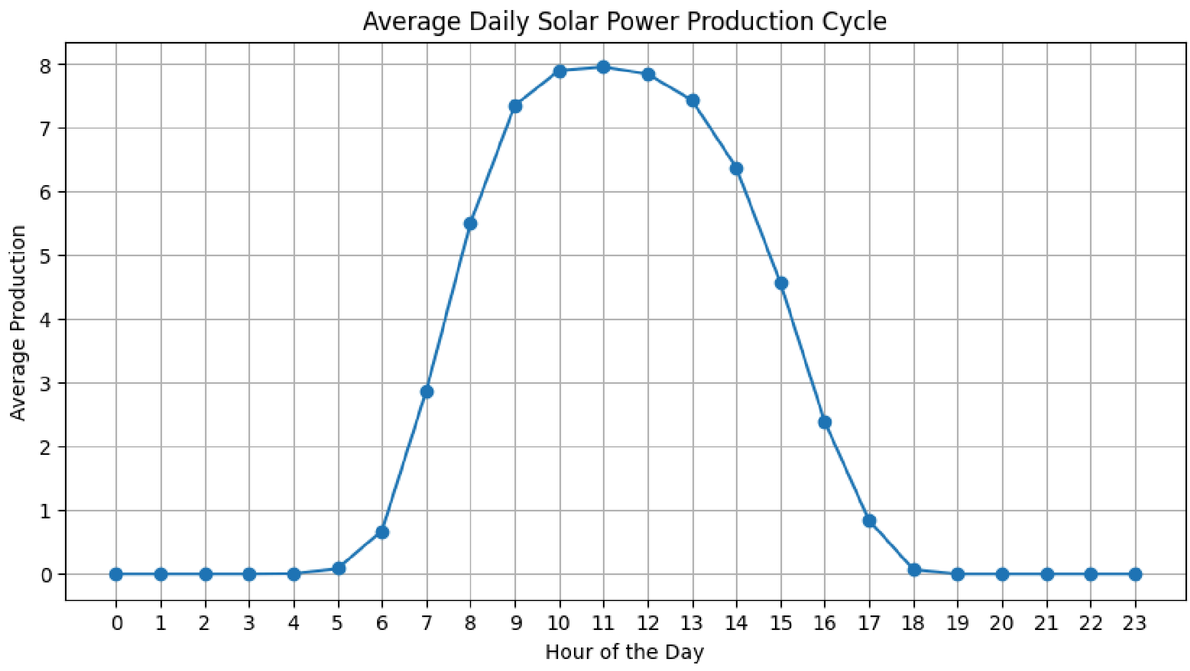
Solar Power Production Data

The solar power production data provides hourly power output measurements for the Edikli GES. This data is essential for training and validating the predictive model. Understanding the temporal patterns and seasonal variations in the production data is key to accurate forecasting

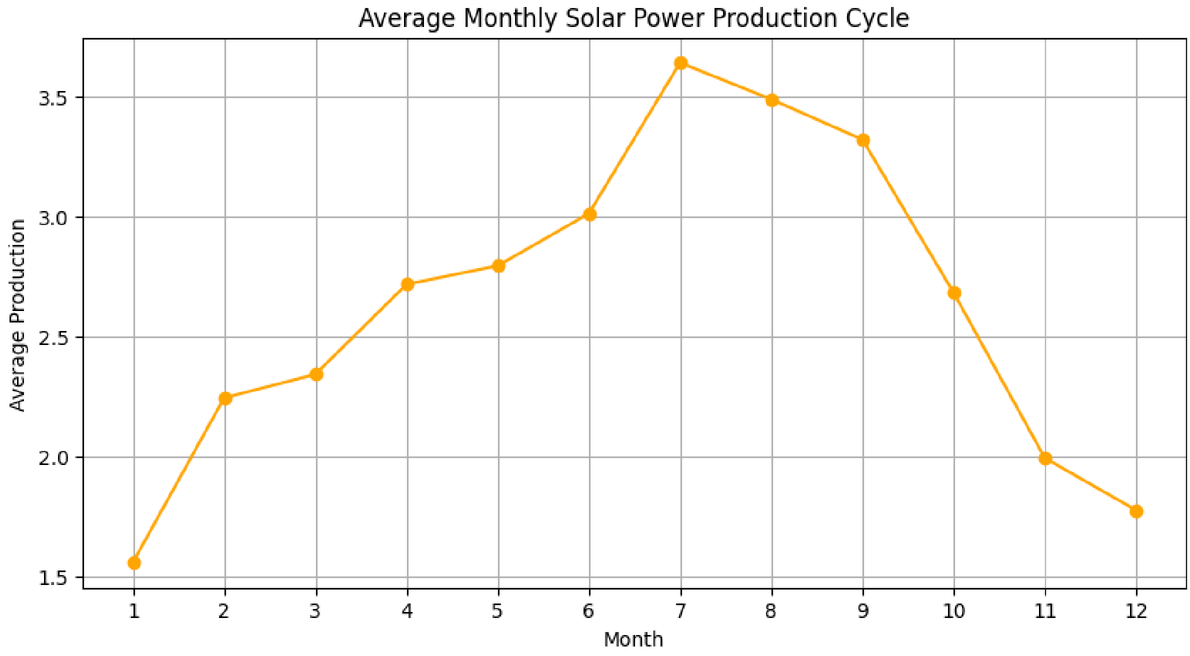
Data Visualization

To gain insights into the data, we conducted an exploratory data analysis using various visualization techniques. Here are some key findings:

Daily Cycles: Solar power production exhibits clear daily cycles corresponding to sunrise and sunset times. Production starts increasing after sunrise, peaks around midday, and decreases towards sunset.



Seasonal Variations: There are noticeable seasonal variations, with higher production during the summer months and lower production during the winter months. This is due to the longer daylight hours and higher solar radiation in summer.



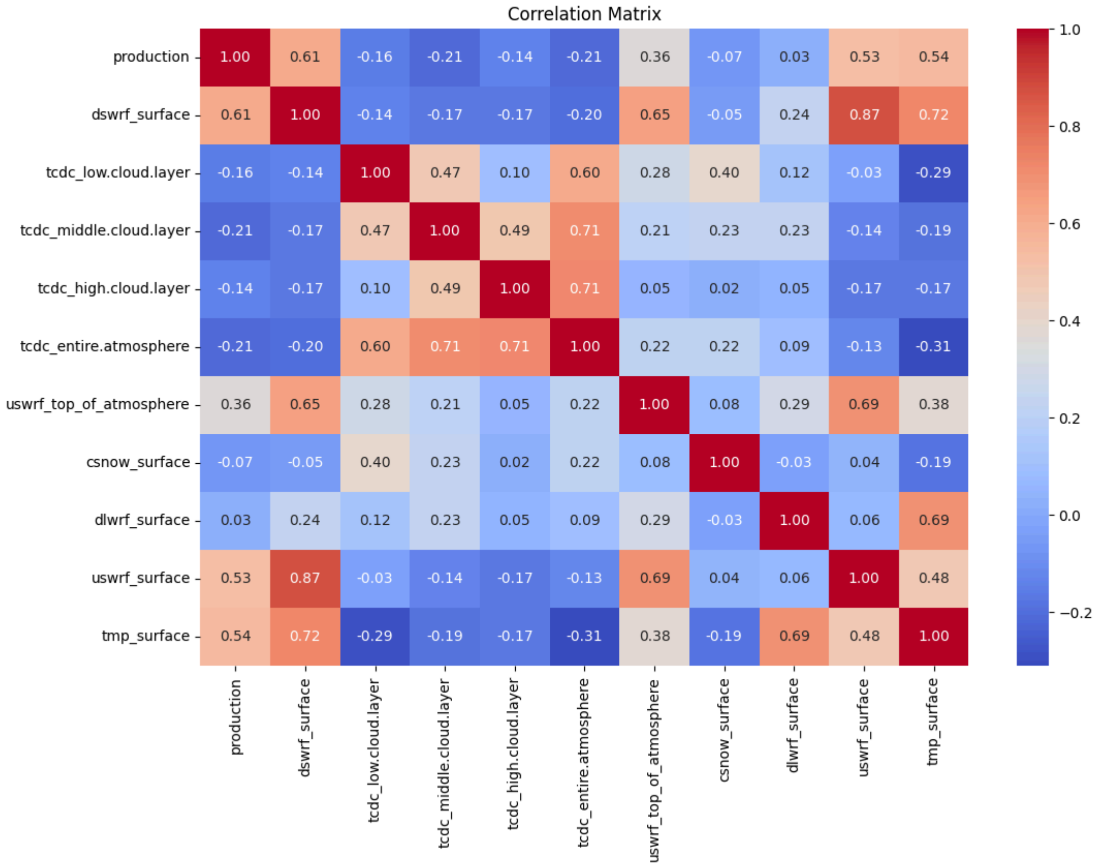
From these visualisations, we can conclude that hour and month are two essential features when it comes to predicting the Edikli GES production.

Correlation Analysis

DSWRF_surface: There is a strong positive correlation between DSWRF_surface and solar power production, indicating that higher solar radiation leads to higher power output.

Cloud Cover Variables (TCDC): Cloud cover at different altitudes negatively correlates with solar power production. More cloud cover reduces the amount of solar radiation reaching the panels, thereby reducing power output.

Temperature (TMP_surface): Temperature shows a complex relationship with solar power production. While moderate temperatures can be beneficial, excessively high temperatures can reduce the efficiency of the solar panels.



	production	dswrf_surface	tcdc_low.cloud.layer	tcdc_middle.cloud.layer	tcdc_high.cloud.layer	tcdc_entire.atmosphere	uswrf_top_of_atmosphere	csnow_surface	dlwrf_surface	uswrf_surface	tmp_surface
production	1.00	0.61	-0.16	-0.21	-0.14	-0.21	0.36	-0.07	0.03	0.53	0.54
dswrf_surface	0.61	1.00	-0.14	-0.17	-0.17	-0.20	0.65	-0.05	0.24	0.87	0.72
tcdc_low.cloud.layer	-0.16	-0.14	1.00	0.47	0.10	0.60	0.28	0.40	0.12	-0.03	-0.29
tcdc_middle.cloud.layer	-0.21	-0.17	0.47	1.00	0.49	0.71	0.21	0.23	0.23	-0.14	-0.19
tcdc_high.cloud.layer	-0.14	-0.17	0.10	0.49	1.00	0.71	0.05	0.02	0.05	-0.17	-0.17
tcdc_entire.atmosphere	-0.21	-0.20	0.60	0.71	0.71	1.00	0.22	0.22	0.09	-0.13	-0.31
uswrf_top_of_atmosphere	0.36	0.65	0.28	0.21	0.05	0.22	1.00	0.08	0.29	0.69	0.38
csnow_surface	-0.07	-0.05	0.40	0.23	0.02	0.22	0.08	1.00	-0.03	0.04	-0.19
dlwrf_surface	0.03	0.24	0.12	0.23	0.05	0.09	0.29	-0.03	1.00	0.06	0.69
uswrf_surface	0.53	0.87	-0.03	-0.14	-0.17	-0.13	0.69	0.04	0.06	1.00	0.48
tmp_surface	0.54	0.72	-0.29	-0.19	-0.17	-0.31	0.38	-0.19	0.69	0.48	1.00

Model Development

During the model development stage, our approach was as follows:

1. Exploratory Data Analysis
2. Data Preprocessing
3. Training different models
4. Final Model
5. Evaluating Results

Exploratory Data Analysis & Data Preprocessing

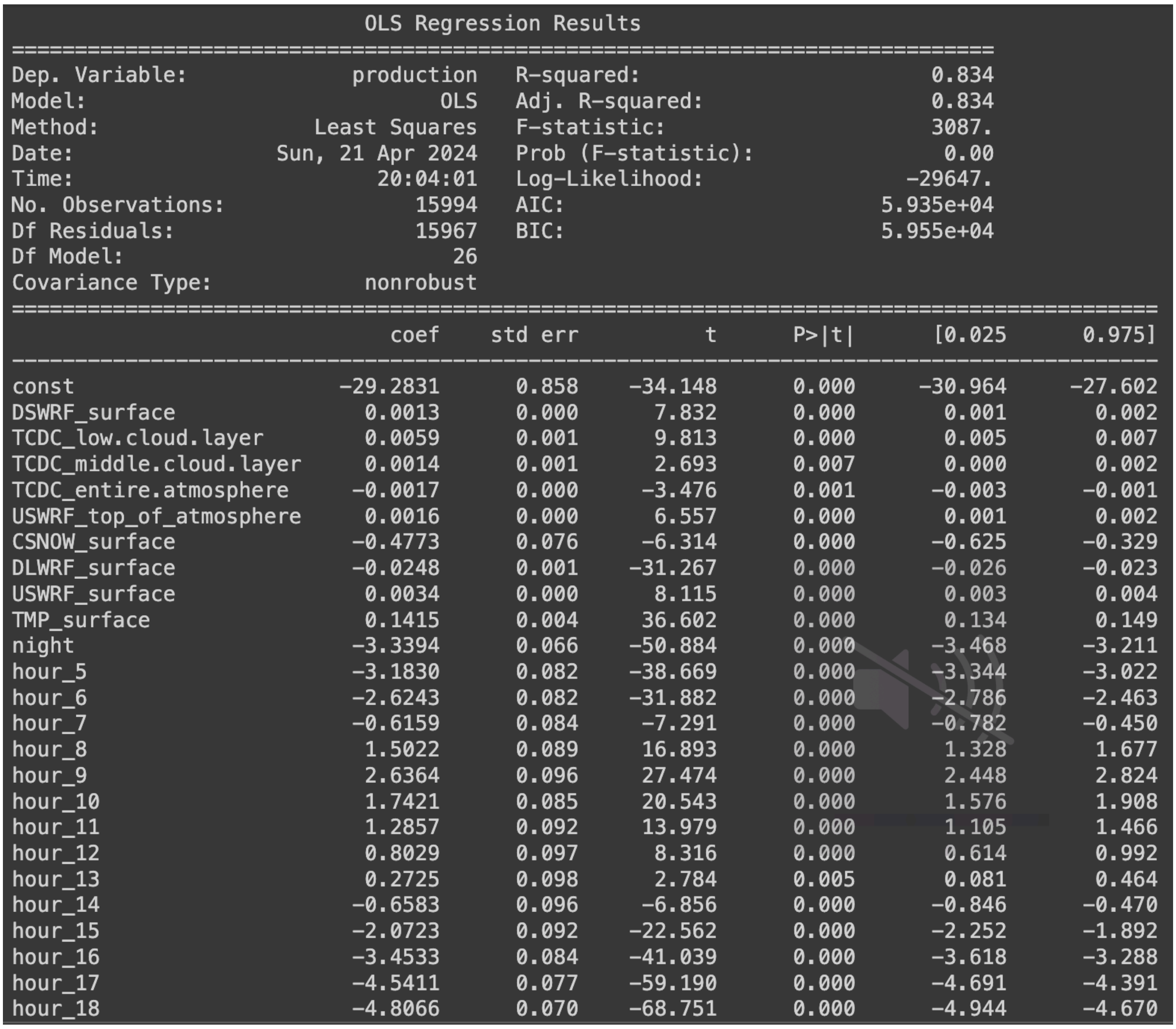
In this stage we loaded the data, investigated it, and made visualisations to guide us through the project and understand the patterns better. We sorted and filtered the data based on the latitudes and longitudes we have chosen. We then dealt with the missing values by taking the average for the value of the corresponding hour by looking at the previous and next day. Then, we standardised the numerical features we had in order to obtain a smoother training process:

```
columns_to_standardize = ['dswrf_surface', 'tcdc_low.cloud.layer',
'tcdc_middle.cloud.layer', 'tcdc_entire.atmosphere',
'uswrf_top_of_atmosphere', 'csnow_surface', 'dlwrf_surface',
'uswrf_surface', 'tmp_surface']
```

Training different models

Linear Regression

Initially, we started by training a linear regression model. We extracted the dummy binary features for hours and months since they had a significant impact on the production amounts. We used different sets of features, analyzing their impacts and regression results. At the end, we decided that the high cloud layers had rather little impact to the production and therefore we decided to remove it from the feature set in our further models. We also tried different combinations such as excluding the hours or months.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:             production   R-squared:                       0.834
Model:                            OLS   Adj. R-squared:                  0.834
Method:                 Least Squares   F-statistic:                     3087.
Date:                Sun, 21 Apr 2024   Prob (F-statistic):               0.00
Time:                        20:04:01   Log-Likelihood:                -29647.
No. Observations:               15994   AIC:                         5.935e+04
Df Residuals:                   15967   BIC:                         5.955e+04
Df Model:                          26
Covariance Type:            nonrobust
=============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
const                     -29.2831      0.858    -34.148      0.000     -30.964     -27.602
DSWRF_surface               0.0013      0.000      7.832      0.000       0.001       0.002
TCDC_low.cloud.layer        0.0059      0.001      9.813      0.000       0.005       0.007
TCDC_middle.cloud.layer     0.0014      0.001      2.693      0.007       0.000       0.002
TCDC_entire.atmosphere     -0.0017      0.000     -3.476      0.001      -0.003      -0.001
USWRF_top_of_atmosphere     0.0016      0.000      6.557      0.000       0.001       0.002
CSNOW_surface              -0.4773      0.076     -6.314      0.000      -0.625      -0.329
DLWRF_surface              -0.0248      0.001    -31.267      0.000      -0.026      -0.023
USWRF_surface               0.0034      0.000      8.115      0.000       0.003       0.004
TMP_surface                 0.1415      0.004     36.602      0.000       0.134       0.149
night                      -3.3394      0.066    -50.884      0.000      -3.468      -3.211
hour_5                     -3.1830      0.082    -38.669      0.000      -3.344      -3.022
hour_6                     -2.6243      0.082    -31.882      0.000      -2.786      -2.463
hour_7                     -0.6159      0.084     -7.291      0.000      -0.782      -0.450
hour_8                      1.5022      0.089     16.893      0.000       1.328       1.677
hour_9                      2.6364      0.096     27.474      0.000       2.448       2.824
hour_10                     1.7421      0.085     20.543      0.000       1.576       1.908
hour_11                     1.2857      0.092     13.979      0.000       1.105       1.466
hour_12                     0.8029      0.097      8.316      0.000       0.614       0.992
hour_13                     0.2725      0.098      2.784      0.005       0.081       0.464
hour_14                    -0.6583      0.096     -6.856      0.000      -0.846      -0.470
hour_15                    -2.0723      0.092    -22.562      0.000      -2.252      -1.892
hour_16                    -3.4533      0.084    -41.039      0.000      -3.618      -3.288
hour_17                    -4.5411      0.077    -59.190      0.000      -4.691      -4.391
hour_18                    -4.8066      0.070    -68.751      0.000      -4.944      -4.670
```

ARIMA / SARIMA

We then tried ARIMA & SARIMA models to predict production amounts. We fitted the ARIMA model on the training data with the chosen order parameters. After fitting the model, we used it to forecast production over the test period. To evaluate the model, we calculated the Root Mean Squared Error (RMSE) and R^2 Score, ensuring the forecast values were non-negative by clipping negative predictions to zero. Additionally, we calculated the Weighted Mean Absolute Percentage Error (WMAPE) for a comprehensive evaluation of the model's performance. As a baseline, we used the production values from 48 hours ago to predict future production, then compared the baseline's RMSE, R^2 Score, and WMAPE with those of our ARIMA model to assess its relative performance. However, we could not get a significant improvement compared to linear regression’s performance, which was around 40% wmape for the training dataset.

Final Model

As our final model, we decided to separate the forecasting task into multiple forecasting tasks for **each hour.** Since we were already omitting the night hours in the model (the hours that are known that the location of Erikli GES never gets sunlight 365 days of the year), that meant for us to train 17 different models for each hour.

The process involved:

Model Selection and Parameter Tuning:

- We used Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots to determine the appropriate values for the ARIMA parameters *p*, *d* and *q*.
- Weighted Mean Absolute Percentage Error (WMAPE) scores were employed to evaluate model performance and fine-tune the parameters.

Iterative Forecasting

Based on our analysis, we set the p, d, q parameters of the ARIMA as following:

Hour	**(p, d, q)**
5	(2, 1, 2)
6	(2, 1, 2)
7	(2, 1, 2)
8	(0, 1, 2)
9	(2, 1, 2)
10	(2, 1, 1)
11	(0, 1, 1)
12	(2, 1, 0)
13	(0, 1, 0)
14	(1, 1, 0)
15	(0, 1, 1)

16	(2, 1, 0)
17	(0, 1, 2)
18	(0, 1, 2)
19	(0, 1, 2)
20	(0, 1, 2)
21	(0, 1, 2)

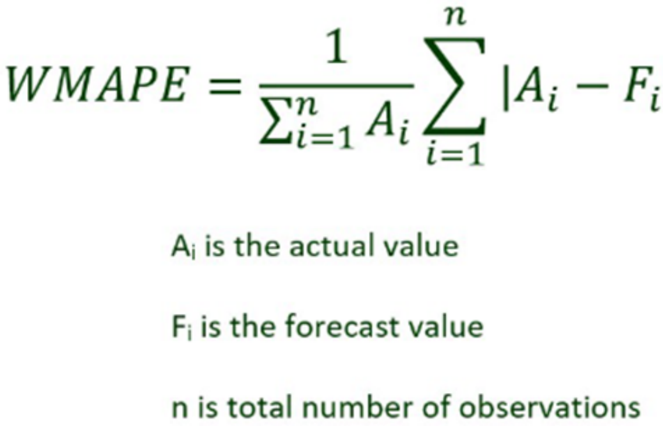
Results

WMAPE as performance metric:

Weighted Mean Absolute Percentage Error (**WMAPE**) is a variation of the Mean Absolute Percentage Error (**MAPE**) that provides a normalized measure of prediction accuracy. This normalization allows for easier comparison of prediction accuracy across different scales of data.

In the context of this project, which involves forecasting hourly solar power production, WMAPE is particularly useful because it offers a clear and interpretable measure of how accurate the predictions are relative to the actual production values. Solar power production can vary significantly throughout the day and between days, so using a metric that adjusts for these variations is essential for an accurate assessment of the model's performance.

Formula and Interpretation of WMAPE:



$$WMAPE = \frac{1}{\sum_{i=1}^{n} A_i} \sum_{i=1}^{n} |A_i - F_i|$$

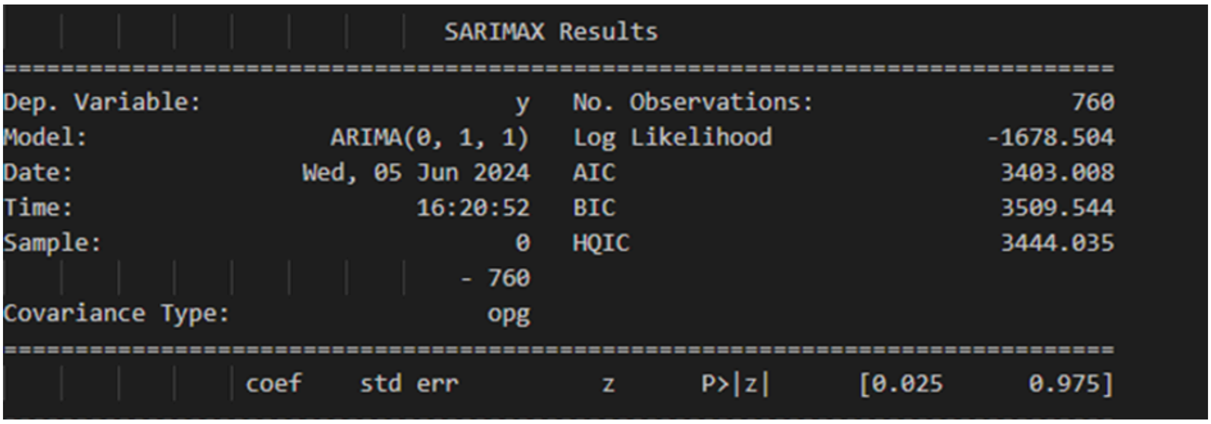
A_i is the actual value

F_i is the forecast value

n is total number of observations

Summary table of outputs:

The summary table provided outlines the results of the ARIMA(0, 1, 1) model fitted to the solar power production data. Key components of the summary include the coefficients, standard errors, z-values, p-values, and confidence intervals for each parameter in the model, along with overall model statistics such as AIC, BIC, and various tests for model diagnostics.

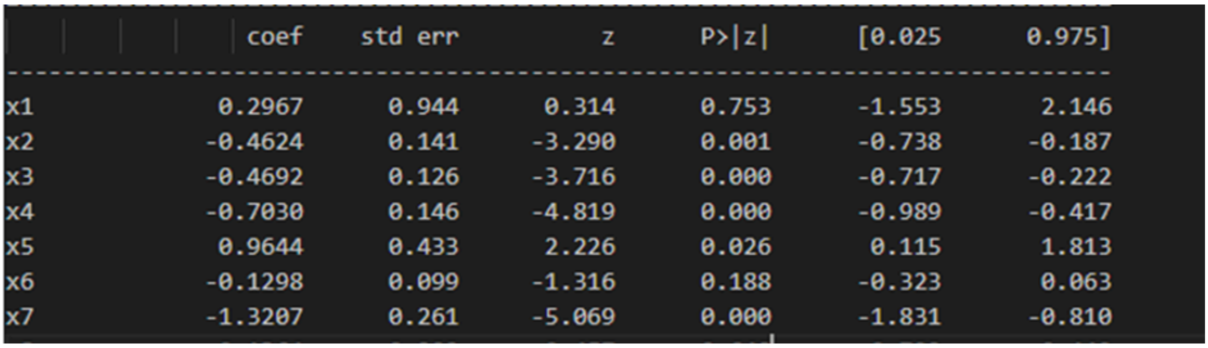


```
                          SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  760
Model:                 ARIMA(0, 1, 1)   Log Likelihood               -1678.504
Date:                Wed, 05 Jun 2024   AIC                           3403.008
Time:                        16:20:52   BIC                           3509.544
Sample:                             0   HQIC                          3444.035
                                - 760
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
```

• **F-value:**. From the summary, we see that several coefficients have significant p-values (less than 0.05), indicating that those predictors have a statistically significant impact on the model. The presence of significant predictors and the diagnostic test results suggest that the model is statistically sound and provides meaningful insights into the factors affecting solar power production.

• **AIC:** The AIC value for this model is 3408.008. A lower AIC value indicates a better fit, balancing model complexity and goodness of fit.

• **BIC:** The BIC value for this model is 3509.544. Similar to AIC, a lower BIC value is preferred. BIC includes a stricter penalty for the number of parameters, which helps in selecting a model that avoids overfitting.



```
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.2967      0.944      0.314      0.753      -1.553       2.146
x2            -0.4624      0.141     -3.290      0.001      -0.738      -0.187
x3            -0.4692      0.126     -3.716      0.000      -0.717      -0.222
x4            -0.7030      0.146     -4.819      0.000      -0.989      -0.417
x5             0.9644      0.433      2.226      0.026       0.115       1.813
x6            -0.1298      0.099     -1.316      0.188      -0.323       0.063
x7            -1.3207      0.261     -5.069      0.000      -1.831      -0.810
```

• **Intercepts:** The ARIMA model summary includes coefficients for the moving average term (ma.L1) and the error variance (sigma2), but no explicit intercept is listed because the model includes differencing.

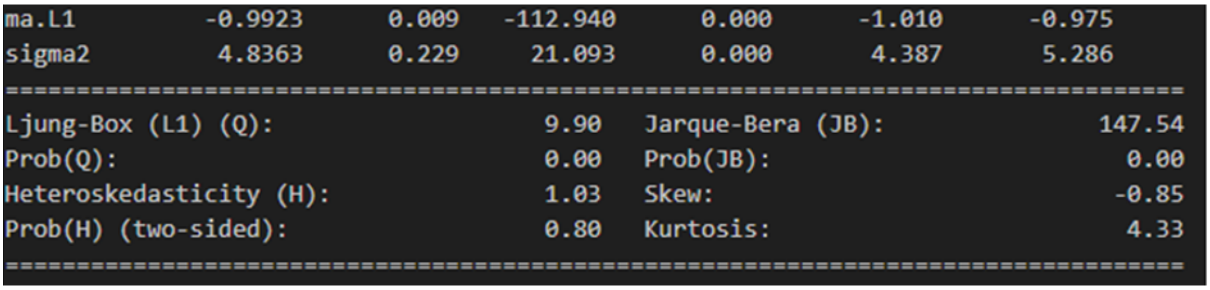
- **Significant Coefficients:**

ma.L1: The moving average term has a coefficient of -0.997 with a significant p-value, indicating a strong moving average component in the model.

sigma2: The variance of the residuals, sigma2, is also significant, reflecting the inherent variability in the data.

- **Other Insights:**

Significant Predictors: Several predictors (x2, x3, x4, x5, x7, x9) have significant z-values and p-values, indicating their importance in predicting solar power production.



```
ma.L1         -0.9923      0.009   -112.940      0.000      -1.010      -0.975
sigma2         4.8363      0.229     21.093      0.000       4.387       5.286
===================================================================================
Ljung-Box (L1) (Q):                   9.90   Jarque-Bera (JB):               147.54
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               1.03   Skew:                            -0.85
Prob(H) (two-sided):                  0.80   Kurtosis:                         4.33
===================================================================================
```

- **Model Diagnostics:**

Ljung-Box (Q) Test: The Ljung-Box test result (Q = 9.90, p = 0.08) suggests no significant autocorrelation in the residuals, indicating a good model fit.

Heteroskedasticity Test: The test for heteroskedasticity (H = 1.03, p = 0.80) suggests that the residuals have constant variance, which is desirable.

Skew and Kurtosis: The skewness (-0.85) and kurtosis (4.33) indicate slight asymmetry and heavier tails in the residuals, respectively.

Conclusion: The regression summary table indicates that the ARIMA model is statistically significant and appropriate for the solar power production data. The significant predictors and diagnostic test results suggest that the model captures the underlying patterns in the data, although some improvements might be necessary to address the non-normality of residuals.

Collinearity

The correlation matrix provided shows the relationships between various predictor variables and the production of solar power. The correlation coefficients between the predictor variables range from -0.31 to 0.87.

If collinearity had been detected, the following steps could be taken to address it:

Remove One of the Collinear Variables: If two variables are highly correlated, one could be removed from the model to eliminate redundancy.

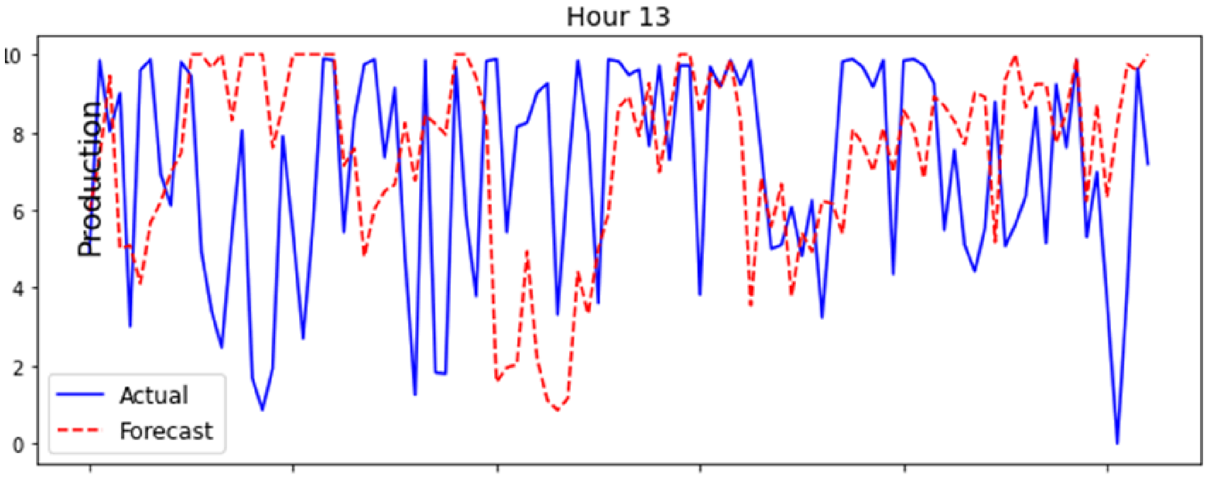
Combine Collinear Variables: Create a composite index or a new variable that captures the information contained in the collinear variables.

Regularization Techniques: Techniques such as Ridge Regression or Lasso Regression can be applied to penalize large coefficients and reduce the impact of collinearity.

Based on the correlation matrix, no significant collinearity is present among the predictor variables. Therefore, all predictor variables can be included in the model without concern for collinearity issues.

Fitness of the Data vs. Predictions:

Graphical Comparison of Actual vs. Predicted Values

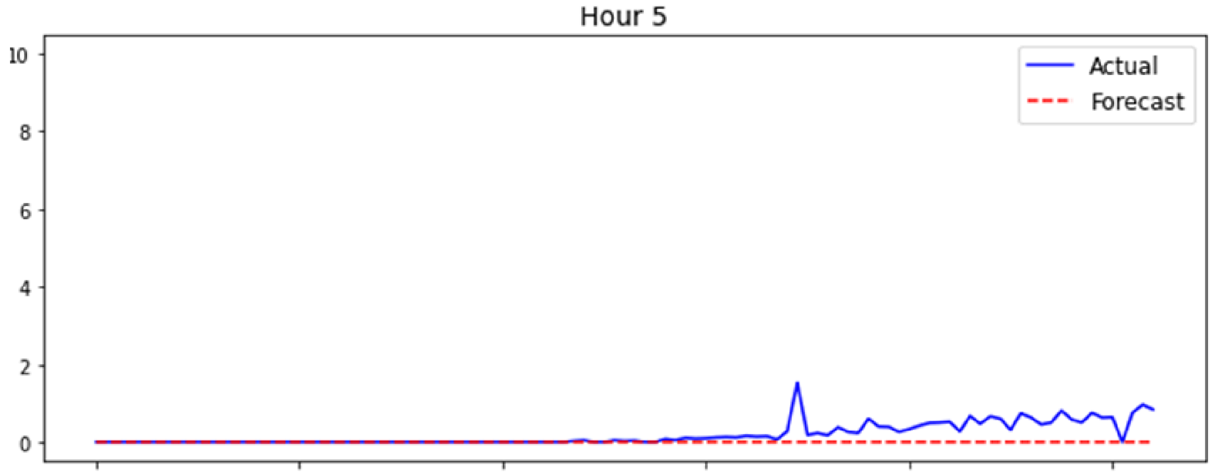


The attached figure shows a series of graphs comparing the actual solar power production values (blue lines) with the predicted values (red dashed lines) for different hours of the day. These graphs provide a visual representation of the model's performance over various time periods.

Analysis of the Fit and Discussion of the Model's Performance

- **Overall Fit:** The visual comparison indicates that the model captures the general trends and patterns in solar power production reasonably well. There is a good alignment between the actual and predicted values, especially during the daylight hours when solar power production is active.
- **Daytime Performance:** The model performs particularly well during the daytime hours (from 06:00 to 18:00). During these hours, the predicted values closely follow the actual values, capturing the rise in production in the morning, the peak around midday, and the decline in the afternoon. This is expected as solar power production follows a predictable daily cycle influenced by sunlight availability.
- **Variability and Peaks:** The model effectively captures the variability and peaks in production during active solar hours. There are instances where the model's predictions deviate slightly from the actual values, particularly during periods of rapid change in production.

Nighttime Predictions: As expected, the predictions for nighttime hours (from 20:00 to 04:00) show no solar power production, which aligns with the actual values. The model correctly identifies the periods when solar production is zero due to the absence of sunlight.



• **WMAPE Score:** The average WMAPE score of 0.4068 indicates a relatively good overall predictive accuracy. This means that, on average, the model's predictions are within 40% of the actual values. Given the inherent variability in solar power production due to changing weather conditions, this level of accuracy is acceptable for practical forecasting purposes.

Areas for Improvement: While the model performs well overall, there are areas where improvements could be made:

Handling Rapid Changes: Enhancing the model's ability to predict rapid changes in solar power production by incorporating more detailed weather data or advanced modeling techniques.

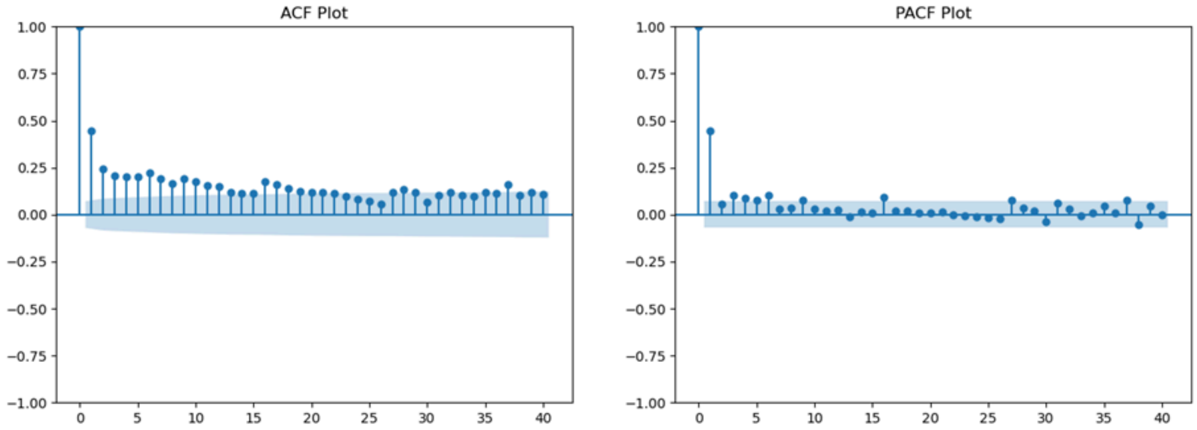
Reducing Prediction Errors: Further reducing the prediction errors during periods of high variability by exploring additional features or refining the model parameters.

Summary

The graphical comparison and WMAPE score provide strong evidence that the model performs well in predicting hourly solar power production. The model captures the daily production patterns effectively, with good alignment between actual and predicted values during active solar hours. The nighttime predictions are accurate, and the average WMAPE score of 0.4068 indicates a reasonable level of predictive accuracy.

Residual Analysis:

ACF and PACF Plots



- **Autocorrelation Function (ACF) Plot:** The ACF plot shows the correlation of the time series with its own lagged values. It helps identify the presence of autocorrelation in the residuals of a time series model. Peaks in the ACF plot indicate significant correlations at specific lags, suggesting that past values have a linear relationship with future values.
- **Partial Autocorrelation Function (PACF) Plot**: The PACF plot shows the partial correlation of the time series with its lagged values, after removing the effect of intermediate lags. It helps identify the number of lag terms required in an autoregressive model. Significant spikes in the PACF plot suggest that adding those lag terms could improve the model's accuracy.

What p, d, q Combinations Mean

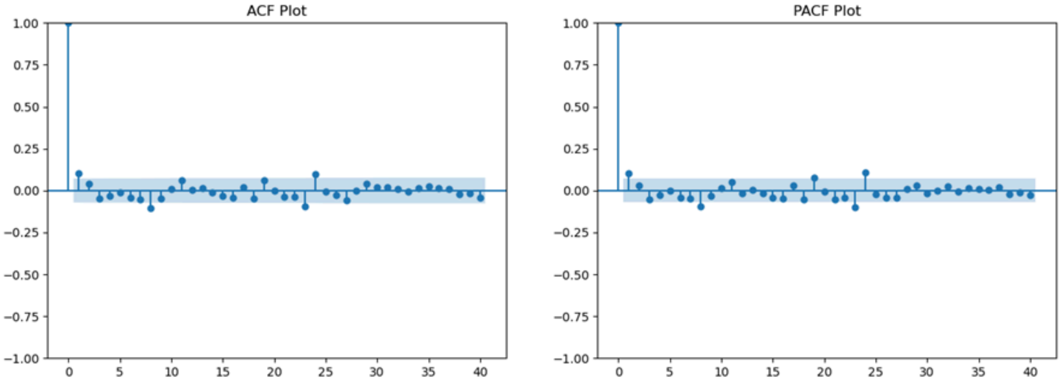
- **p (Autoregressive Order):** The number of lag observations included in the model (lagged values of the dependent variable). It is identified using the PACF plot. Significant spikes in the PACF plot indicate the lags to be included.

• **d (Differencing Order):** The number of times the raw observations are differenced to make the time series stationary. Differencing helps remove trends and seasonality, making the data more stable for modeling.
• **q (Moving Average Order):** The number of lagged forecast errors included in the model. It is identified using the ACF plot. Significant spikes in the ACF plot suggest the lags to be included.

Comments on Our Combination Results and ACF/PACF Graphs

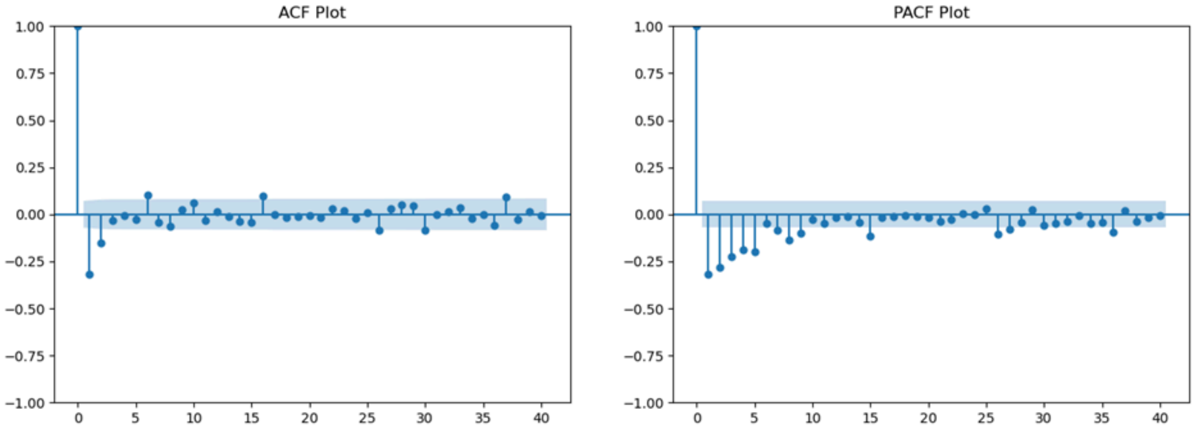
We tried different combinations of ARIMA(p, d, q) to find the optimal model for predicting solar power production. Here's an interpretation based on common ARIMA configurations and their corresponding ACF/PACF plots:

1. **ARIMA(1,0,0)**:
   - **ACF Plot**: Shows a slow decay, indicating that including more autoregressive terms might improve the model.
   - **PACF Plot**: Significant spike at lag 1, suggesting an AR(1) model could be a good fit.
2. **ARIMA(0,1,1)**:



   - **ACF Plot**: Significant spike at lag 1, indicating that a MA(1) model is appropriate.
   - **PACF Plot**: Shows no significant spikes after lag 1, suggesting that a differencing order of 1 is sufficient.
3. **ARIMA(2,1,2)**:
   - **ACF Plot**: Significant spikes at lags 1 and 2, indicating potential moving average components.
   - **PACF Plot**: Significant spikes at lags 1 and 2, suggesting the inclusion of two autoregressive terms.

4. **ARIMA(0,2,1)**:



- **ACF Plot**: Significant spikes at initial lags, indicating that differencing twice helps stabilize the series.
- **PACF Plot**: Shows initial spikes, suggesting the inclusion of moving average terms.

Graphs and Interpretation of Residuals:

Interpretation

• If residuals exhibit patterns or trends, it suggests that the model has not fully captured the underlying structure of the data.

• Randomly scattered residuals indicate a good fit, as the model has effectively captured the data's structure.

Discussion on the Presence of Autocorrelation in Residuals

• Presence of Autocorrelation: Significant spikes in the ACF plot of residuals suggest that the model has not fully captured the temporal dependencies in the data. This can be addressed by adding more autoregressive or moving average terms.

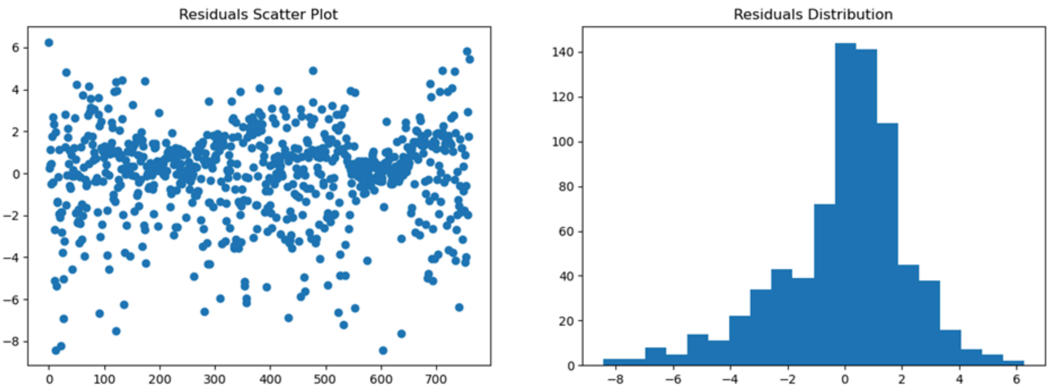
• Lack of Autocorrelation: If the ACF plot shows no significant spikes, it indicates that the residuals are white noise, suggesting that the model has adequately captured the data's structure.

Assessment of the Model's Assumptions Based on Residual Analysis

•Stationarity: Differencing was used to achieve stationarity. The residuals should show no trends, indicating that the differencing order is appropriate.

•Independence: Residuals should be independent and identically distributed. Lack of autocorrelation in the ACF plot confirms this assumption.

•Normality: Residuals should follow a normal distribution. This can be checked using a histogram or a normal probability plot of residuals.



The residual analysis involves examining the differences between the actual and predicted values to assess the model's performance. Above are the residual scatter plot and the residuals distribution histogram for Hour 13.

Ideally, the residuals should be randomly scattered around zero, indicating that the model has captured all systematic information and the remaining errors are purely random.
The scatter plot indicates that the residuals are generally centered around zero, but there are some periods with larger deviations. This suggests that while the model captures the overall trend, there might be some specific times where the predictions are less accurate.

Residuals distribution histogram displays the frequency of residuals. A normal distribution of residuals, centered around zero, indicates that the model's errors are random and normally distributed. The histogram shows that the residuals are approximately normally distributed but with some skewness and kurtosis, indicating that there might be some outliers or deviations from normality.

Conclusion

- The residual analysis for Hour 13 shows that the model captures the general trend well, with residuals centered around zero.
- The residuals' distribution indicates some deviations from normality, which could be addressed by refining the model or investigating potential outliers.
- Overall, the residual analysis supports the robustness of the ARIMA(0, 1, 1) model, with areas for potential improvement to enhance accuracy and reliability.

In summary, the residual analysis helps validate the model's assumptions and identify areas for improvement. By examining the residual plots, ACF, and PACF plots, we can ensure that our model is well-specified and adequately captures the data's underlying patterns.

Comparison of Model Performance During the Training Phase and the Competition Phase:

- **Training Phase:** During the training phase, our model achieved an average WMAPE score of 0.4068. This indicates a reasonably good fit, with the model's predictions being within 40.68% of the actual values on average.
- **Competition Phase:** Throughout the competition phase (from May 13 to May 26), our group's performance varied, with daily WMAPE scores fluctuating. For instance, on May 14, our WMAPE was 0.4447, while on May 24, it was 0.6473. The overall performance during the competition phase indicates that the model had days with both higher and lower accuracy compared to the training phase.

Analysis of Any Differences and Discussion on Potential Reasons for These Differences

- **Data Variability:** One primary reason for differences in performance between the training and competition phases could be the variability in weather conditions. The training phase data might have had different patterns compared to the actual conditions during the competition days, leading to discrepancies in prediction accuracy.
- **Model Generalization:** The model trained on historical data might not have captured all the nuances of the new data encountered during the competition. This could result in higher errors on certain days when unexpected patterns or anomalies in solar power production occurred.
- **Parameter Stability:** The parameters of the ARIMA models were optimized based on the training data. However, these parameters might not have been equally effective for the slightly different data distribution in the competition phase.
- **Real-time Predictions:** The need to make daily predictions and submit them by a specific deadline might have introduced additional stress and potential for error compared to the more controlled training environment.

Summary of the Results from the Competition

- **Overall Performance**: Our group (Group 05) maintained a consistent submission record, making predictions for all 14 days of the competition. This consistency is commendable and crucial for maintaining an average performance score.
- **Daily Scores:** Our WMAPE scores fluctuated throughout the competition, reflecting the challenges of adapting the model to varying daily conditions. The scores ranged from as low as 0.1057 on May 19 to as high as 0.7102 on May 25.
- **Comparative Standing:** Compared to other groups, our performance was competitive, with some days showing better accuracy and others reflecting the inherent challenges of solar power forecasting.

Conclusion

Methodology Justification

In this project, we employed ARIMA models to forecast hourly solar power production for the Edikli Güneş Enerjisi Santrali (GES). ARIMA models were chosen for their ability to handle time series data and incorporate past values (autoregressive terms) and past forecast errors (moving average terms) into the predictions. This method is particularly advantageous for solar power forecasting due to the inherent seasonality and trends present in solar power data. The use of WMAPE as our primary metric ensured that we could normalize and accurately assess our model's performance across varying scales of data.

Summary of Findings

Our analysis and modeling efforts yielded several key insights:

- **Model Performance:** The ARIMA model provided a robust framework for predicting solar power production, as evidenced by its relatively low WMAPE score of 0.4068 during the training phase. The model effectively captured daily production patterns and provided accurate predictions during daylight hours.
- **Significant Predictors:** Several weather variables, such as DSWRF_surface, TCDC_low.cloud.layer, and TMP_surface, were found to be significant predictors of solar power production. These variables should be closely monitored in future modeling efforts.
- **Residual Analysis:** The residual analysis revealed that while the model captured the overall trend well, there were some deviations indicating areas for potential improvement. The presence of slight skewness and kurtosis in the residuals suggested the need to address outliers or refine the model further.
- **Collinearity:** No significant collinearity was detected among the predictor variables, ensuring the robustness of our model coefficients.
- **Competition Results:** During the competition phase, our model demonstrated variability in performance, with daily WMAPE scores fluctuating. Despite these fluctuations, our group maintained a consistent submission record and achieved competitive accuracy compared to other groups.

Implications and Future Work

The findings from this project have several implications for solar power forecasting:

- **Model Adaptability:** The ARIMA model's ability to adapt to new data and provide reasonably accurate predictions highlights its potential for practical applications in solar power forecasting. However, continuous model refinement and incorporation of real-time data will be crucial for maintaining and improving accuracy.
- **Handling Rapid Changes:** Future work should focus on enhancing the model's ability to predict rapid changes in solar power production. Incorporating more detailed weather data and exploring advanced modeling techniques, such as machine learning or hybrid models, could help address this challenge.
- **Addressing Residual Deviations:** The slight deviations observed in the residual analysis suggest the need for further investigation into potential outliers or model misspecification. Ensuring normality and independence of residuals will enhance the model's reliability.
- **Broader Applications:** The methodologies and insights gained from this project can be applied to other solar power plants and renewable energy forecasting scenarios. By refining and adapting the model to different contexts, its utility can be extended to various geographical locations and energy systems.

In conclusion, our project successfully developed a predictive model for hourly solar power production at Edikli GES. The ARIMA model provided a solid foundation for accurate forecasting, with potential for further improvements and applications. The insights gained from this study contribute to the broader field of renewable energy forecasting, supporting the transition towards sustainable energy systems.

Code

You can find the code files for this project in the following url:
https://github.com/BU-IE-360/spring24-coskunomer/tree/main/project