IE360 PROJECT

Statistical Forecasting and Time Series

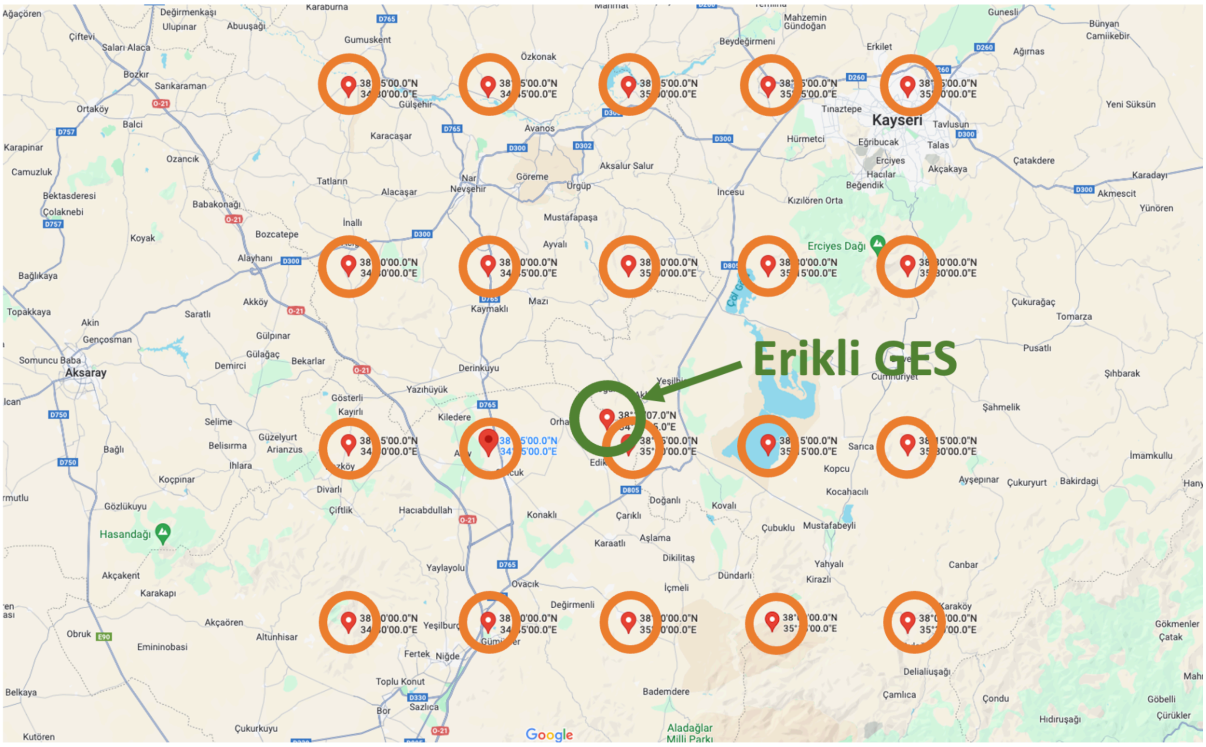
Ali Burak Öztürk	2019402048
Ömer Coşkun	2019402277
Ferhat Peynirci	2019402105

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.

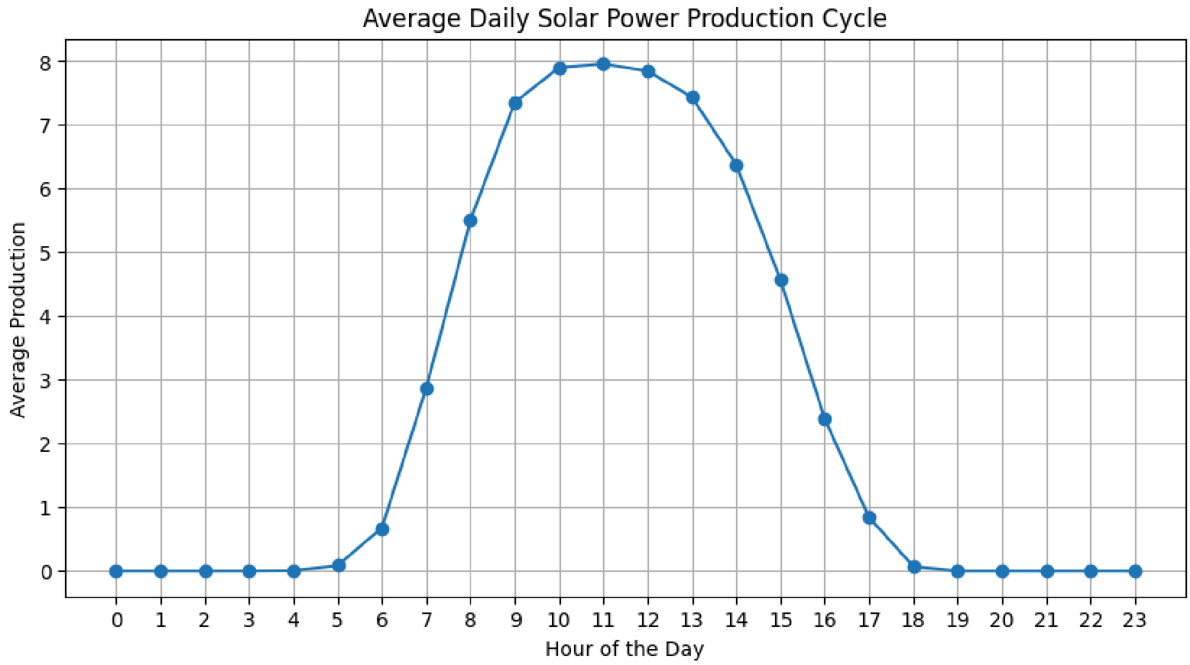
Solar Power Production Data

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

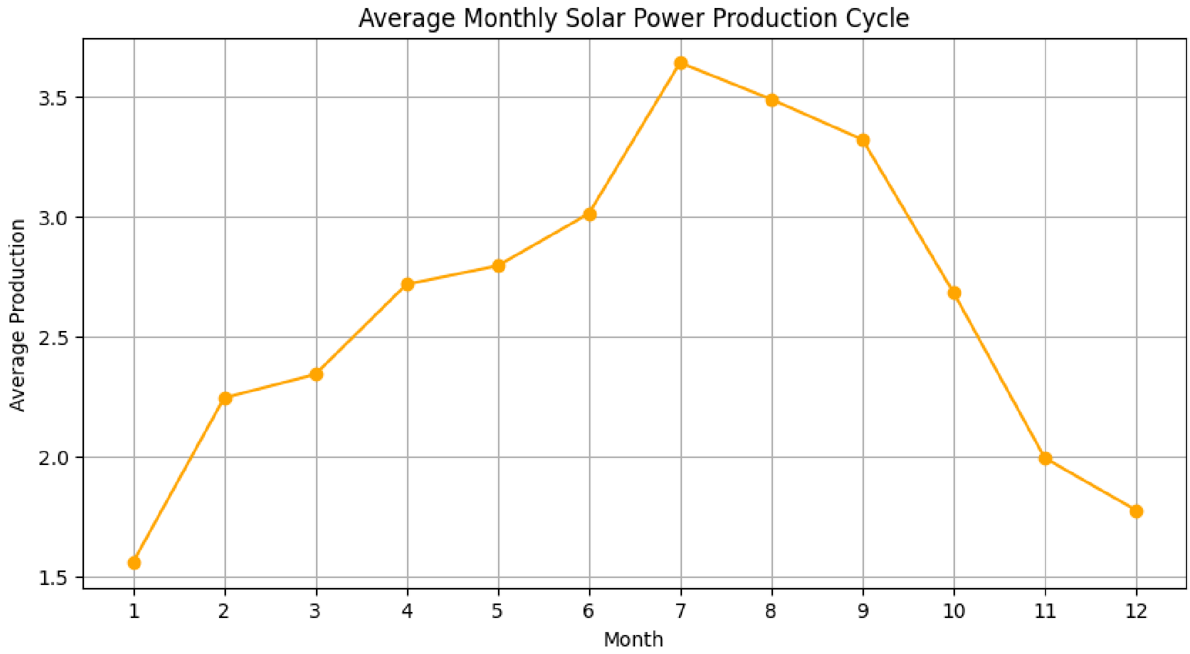
Data Visualization

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.



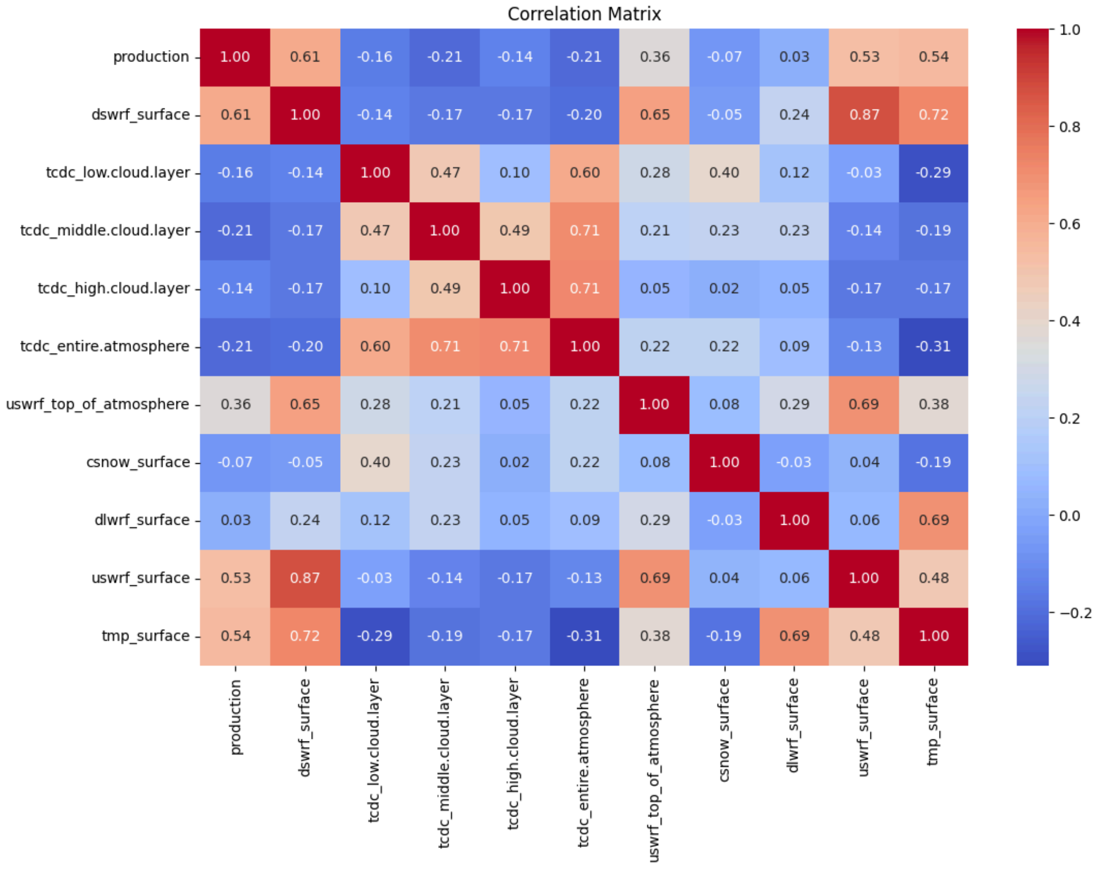
From these visualisations, we can conclude that hour and month are two essential features when it comes to predicting the Edikli GES production.

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

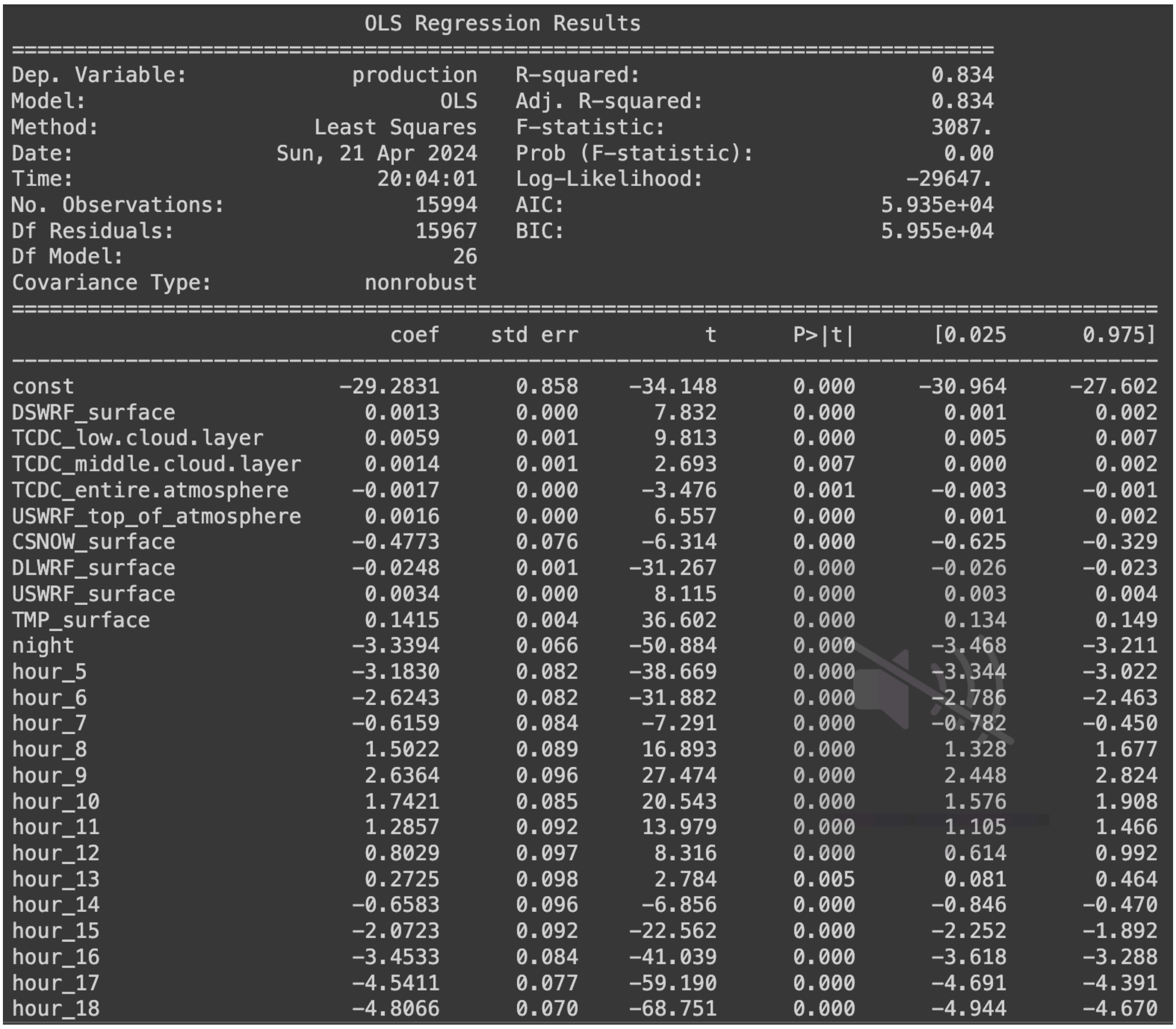
In this stage we loaded the data, investigated it, and made visualisations to guide us through the project and understand the patterns better. We sorted and filtered the data based on the latitudes and longitudes we have chosen. We then dealt with the missing values by taking the average for the value 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

Iterative Forecasting

To validate our models, we employed an iterative forecasting approach during the test period from February 1st to May 15th:

1. *Five-Day Predictions:*
    - We made five-day-ahead predictions using the trained models.
    - After each five-day period, the predictions were added to the training data.
2. *Efficiency and Accuracy:*
    - This approach allowed us to balance the need for faster testing with maintaining model accuracy, ensuring that the model was constantly updated with the latest data without the need for daily retraining.

Conclusion and Future Work