IE 360

Project Report

Arda Ergene – 2020402117
Hacı Osman Salih Aydın – 2020402105
Buse Naz Koçali – 2020402180

05.06.2024

Introduction

Problem Description:

This project revolves around providing hourly solar power predictions for the Edikli GES solar power plant situated in Niğde, Turkey. The aim is to simulate real-world energy trading scenarios, where predictions are based on weather data available until the predicted day.

In solar power forecasting there are various factors, including weather variability, complex interactions between weather variables and solar power production, and spatial and temporal dynamics.

The provided dataset includes weather variables from 25 grid points surrounding the Edikli GES plant. These variables encompass downward shortwave radiation flux, solar radiation, cloud cover, snow presence, and temperature. Leveraging this data, participants are tasked with building predictive models that can forecast solar power production accurately and reliably.

Overall, the goal of this project is to predict the solar power energy generetion based data given in the most accurate way.

Summary of Approach:

To prepare the data for prediction, we implemented several modifications. Initially, we focused on transforming the data to meet the requirements of predictive modeling. This included reshaping the dataset into a suitable format for time series analysis, ensuring that each observation corresponded to a specific time interval (e.g., hourly) and incorporating relevant features for forecasting.

Next, we conducted exploratory data analysis (EDA) to gain insights into the dataset's structure and characteristics. This involved examining the distribution of variables, identifying outliers, and handling missing values through imputation or removal.

Following data preprocessing, we conducted a thorough analysis of the correlation matrix to assess the relationships between the target variable (production) and other predictor variables. This allowed us to identify potential predictors that significantly influence solar power production and prioritize them in our modeling approach.

To determine if the data exhibited stationarity, a crucial assumption for time series analysis, we employed techniques such as examining rolling mean and variance. Additionally, we conducted tests for stationarity, such as the KPSS (Kwiatkowski–Phillips–Schmidt–Shin) test, to validate our findings and guide further data transformations if needed.

In our quest to address seasonality, we employed differencing techniques to remove trends and periodic patterns from the data. We visualized autocorrelation (ACF) and partial autocorrelation (PACF) plots to identify the lag values for potential seasonal effects.

For modeling, we experimented with various techniques, starting with simple linear regression to establish baseline performance. Subsequently, we employed more sophisticated time series models such as ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal ARIMA), leveraging the autoARIMA function to determine optimal parameter configurations.

Finally, we evaluated the performance of each model using error metrics such as MAD, MSE, AIC and BIC. Based on these metrics, we selected the best-performing model to generate solar power predictions effectively.

In summary, our approach involved comprehensive data preprocessing, exploratory analysis, and model selection to develop accurate and reliable forecasts for solar power production.

Descriptive Data Analysis:

The descriptive analysis of the given data involved a thorough examination of the dataset's characteristics, including its structure, distribution, and relationships between variables. Here's an overview of the key aspects covered in the descriptive analysis:

- The dataset was structured to contain multiple variables related to weather conditions and solar power production.

- Each observation represented a specific time interval (e.g., hourly) and included measurements for various weather parameters such as solar radiation, cloud cover, temperature, and snow presence.

- A correlation matrix was constructed to quantify the relationships between different variables in the dataset. We found that the most correlated variable is dswrf_surface which is shortwave radiation flux.

- By checking the hourly production plots, one can easily observe that there is seasonality.

- Also, by checking the ACF and PACF plots saesonilty is obvious.

- Differencing is applied to remove seasonality and make the data stationary, facilitating more accurate modeling.

- To check stationartiy, we checked the rolling mean and variance plots and used KPSS Unit Root test.
- As a result of KPSS Unit Root test, p-value is lower than significant value which means data is not stationary.

- We observed the means of hourly production seperately, the hour between 0-5 and 21-23 approximately mean of production is 0. So, in prediction phase we used this results.

Approach:

Data Loading and Preparation:

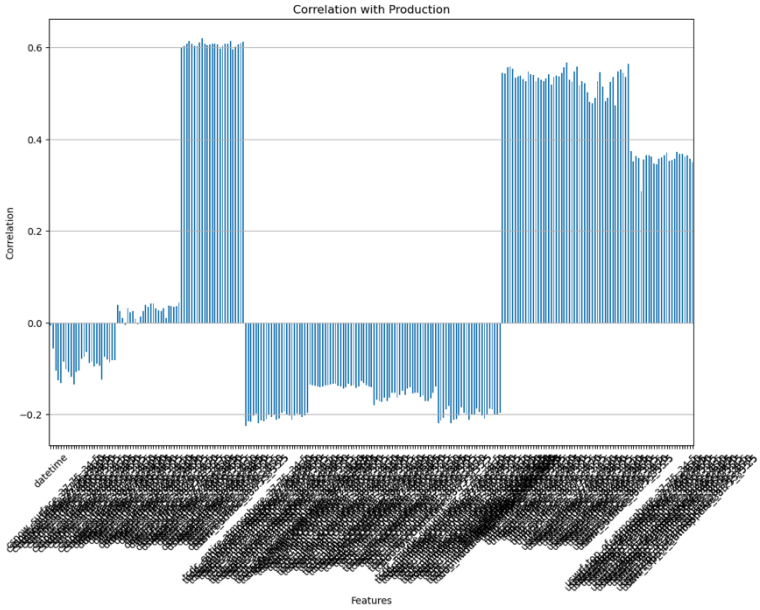
- We converted datasets' "date" column to datetime format and dropped the uncessary columns like 'date' and 'hour'. Reshaped the weather data to wideformat.

- The reshaped weather data is merged with the production data based on the 'datetime' column using an inner join.

- We splitted the merged data into two subsets:
  Available Data: Contains records with non-null values for the 'production' column.
  To-be-Forecasted Data: Contains records with null values for the 'production' column.

Data Analysis:

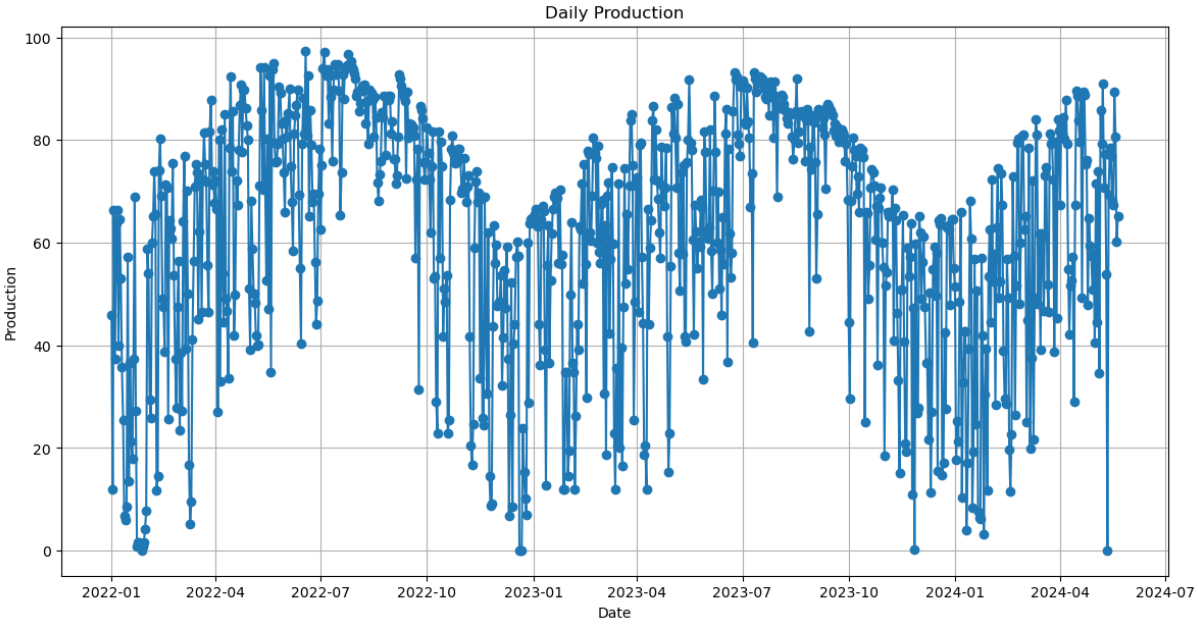
- Our analysis commenced with the creation of a correlation matrix to explore the interrelationships between 'production' and other pertinent variables within the dataset.



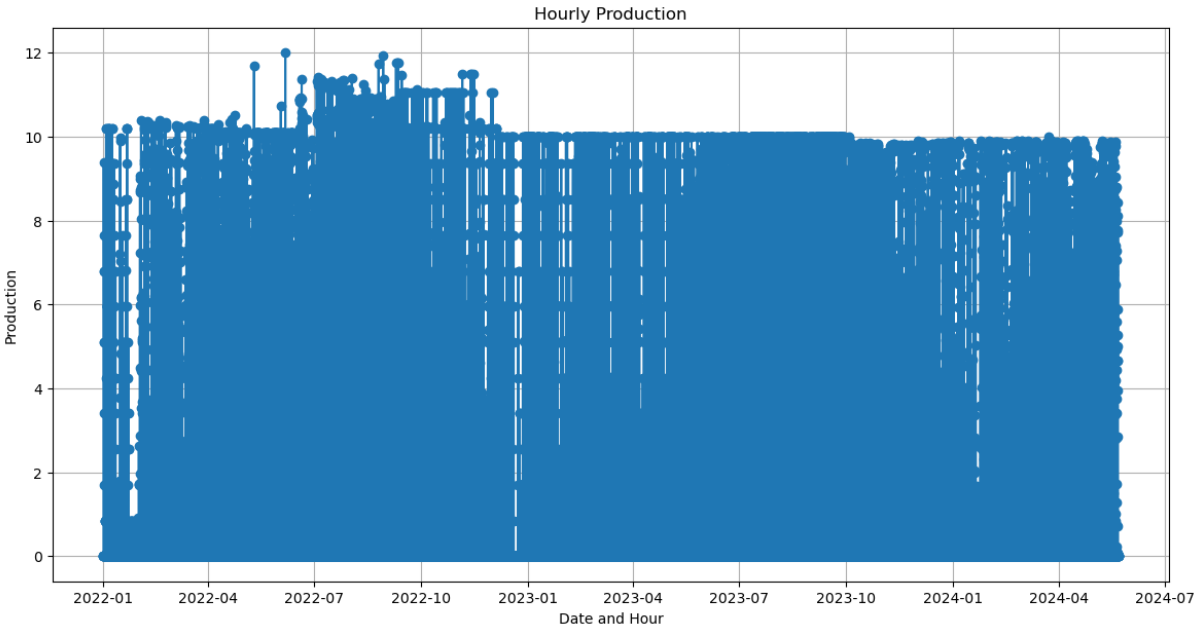
Correlation Matrix

- The correlation matrix served as a foundation for understanding the degree of association between 'production' and various factors, ultimately pinpointing 'dswrf_surface_38.0_35.25' as exhibiting the strongest correlation.

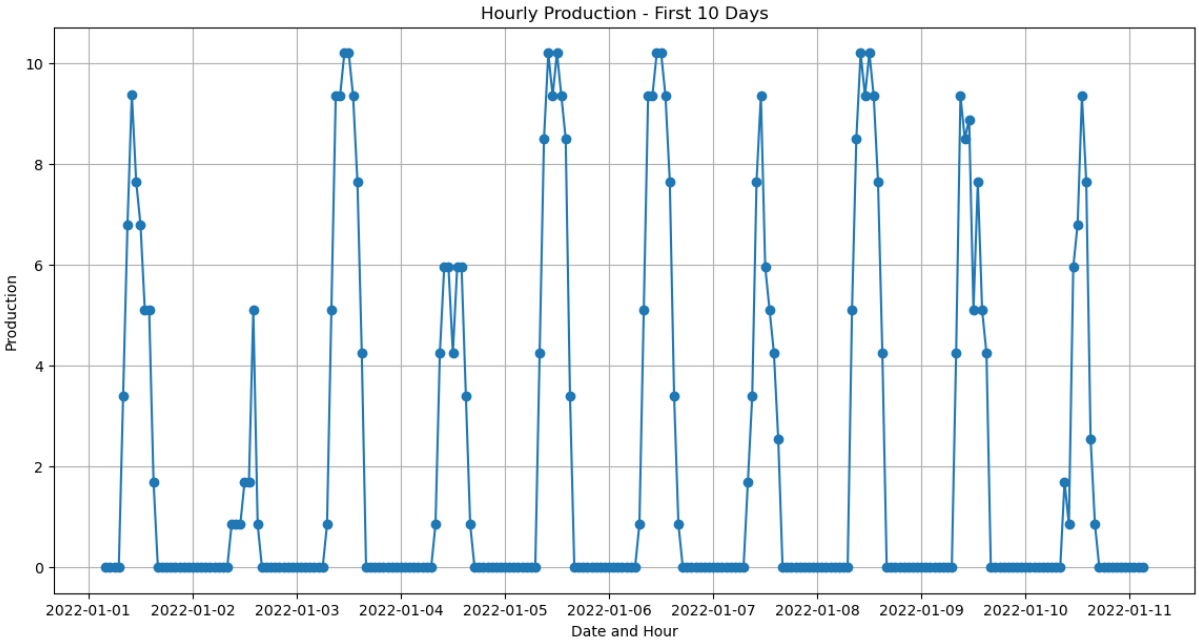
- Following this initial step, we delved deeper into the data by visualizing both hourly and daily production patterns. This involved plotting the first and last 10 days of hourly production data to elucidate any discernible trends or recurring patterns. Our observations revealed a conspicuous daily seasonality, characterized by fluctuations in production levels, albeit without a clear overarching trend.



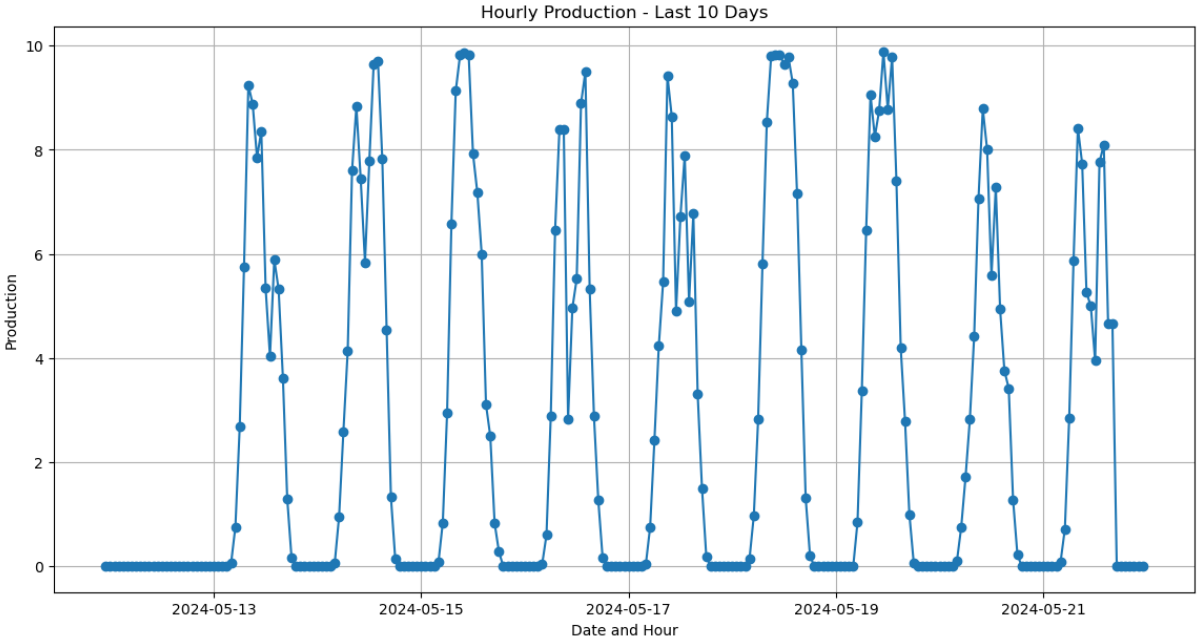
Daily Production Plot



Hourly Production Plot

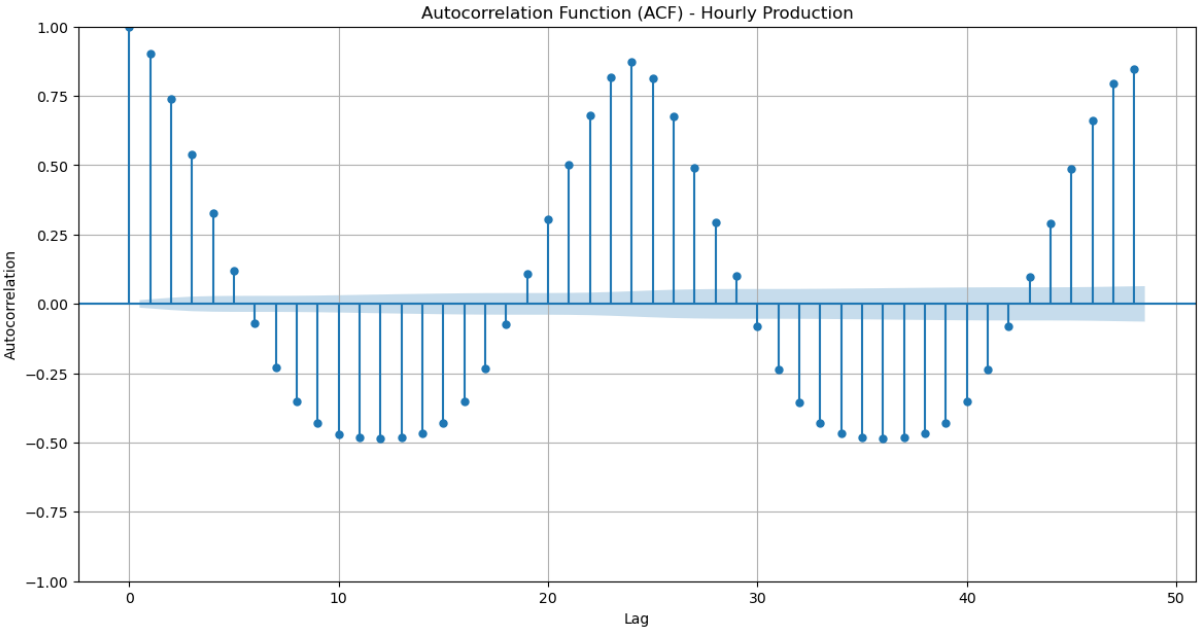


Hourly Production Plot for First 10 Days



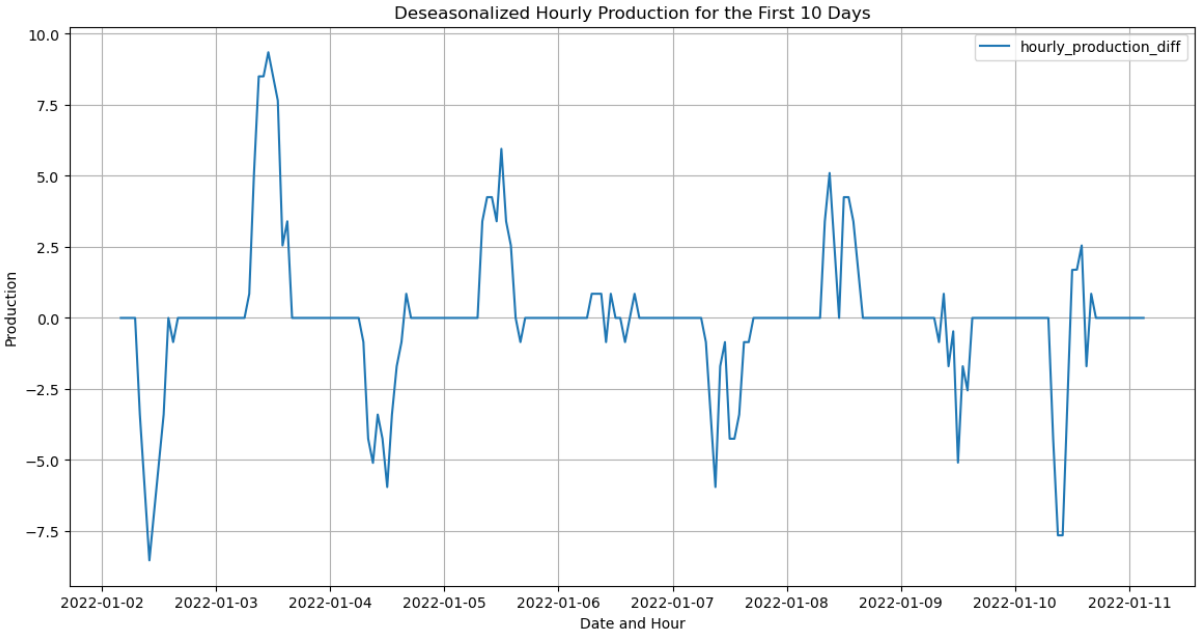
Hourly Production Plot for First 10 Days

- Additionally, we conducted an indepth examination of the Autocorrelation Function (ACF) plot for the production data. This analysis revealed a sinusoidal pattern, indicative of underlying seasonality, with notable autocorrelation observed at lag 1 and 24.

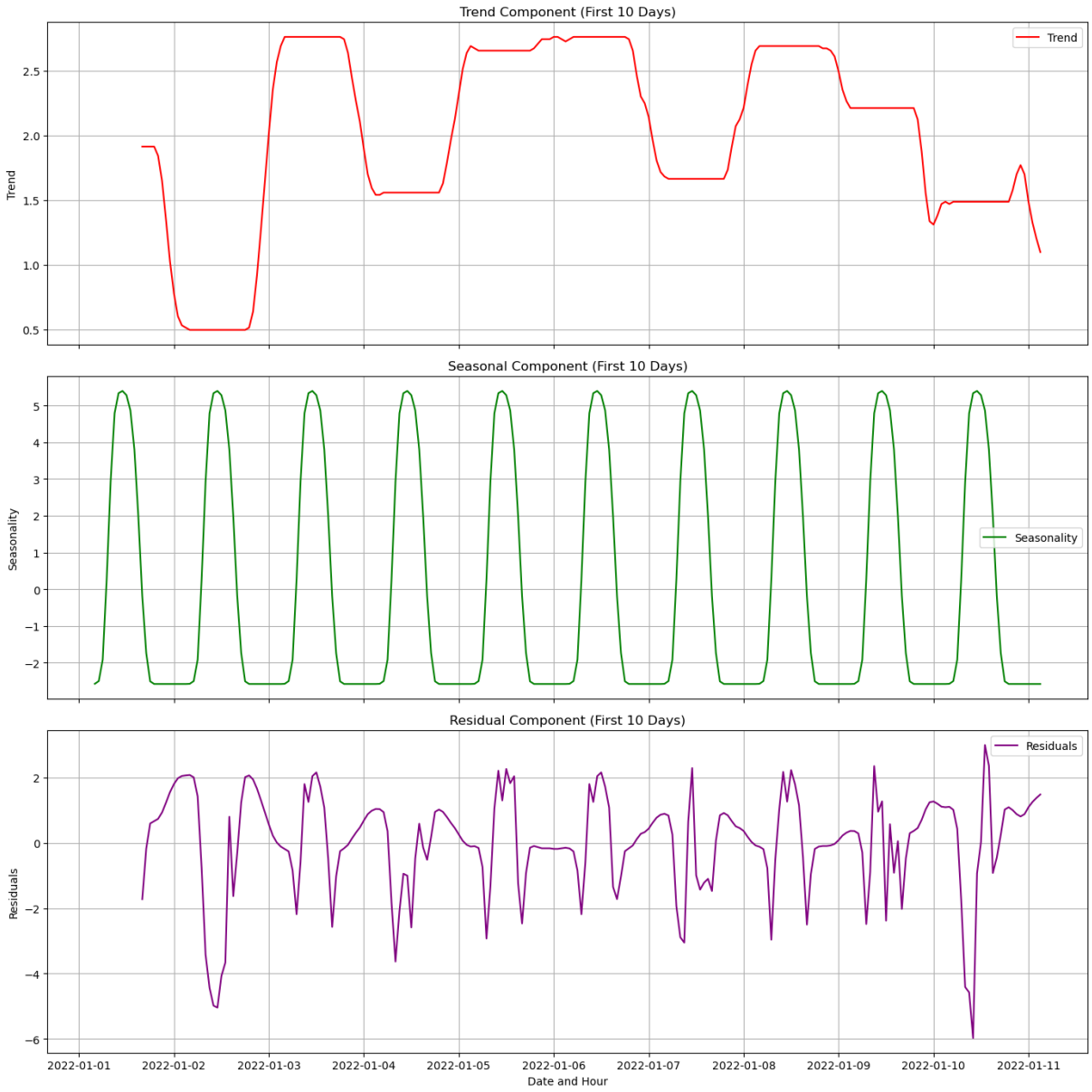


ACF Plot for Hourly Production

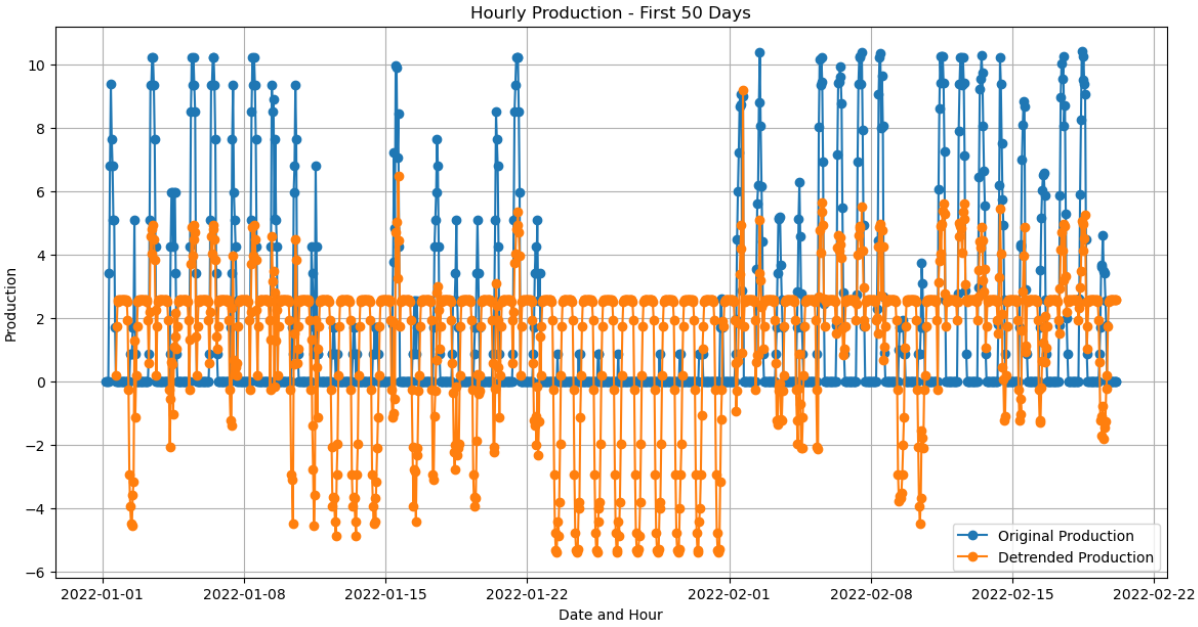
- To gain further insights, we proceeded to decompose the data into its constituent trend and seasonal components. By plotting the decomposed data, we aimed to ascertain whether seasonality or trend persisted postdecomposition. Our findings indicated successful deseasonalization of the data.



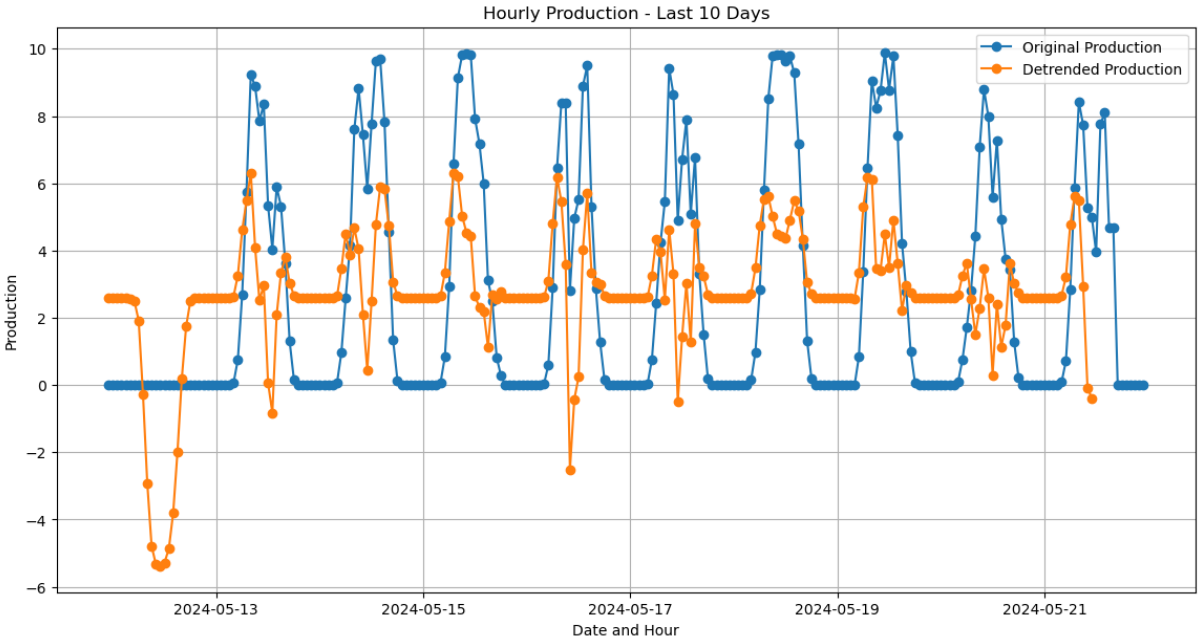
Deseasonalized Hourly Production Plot for First 10 Days



Decompositon of Production Data



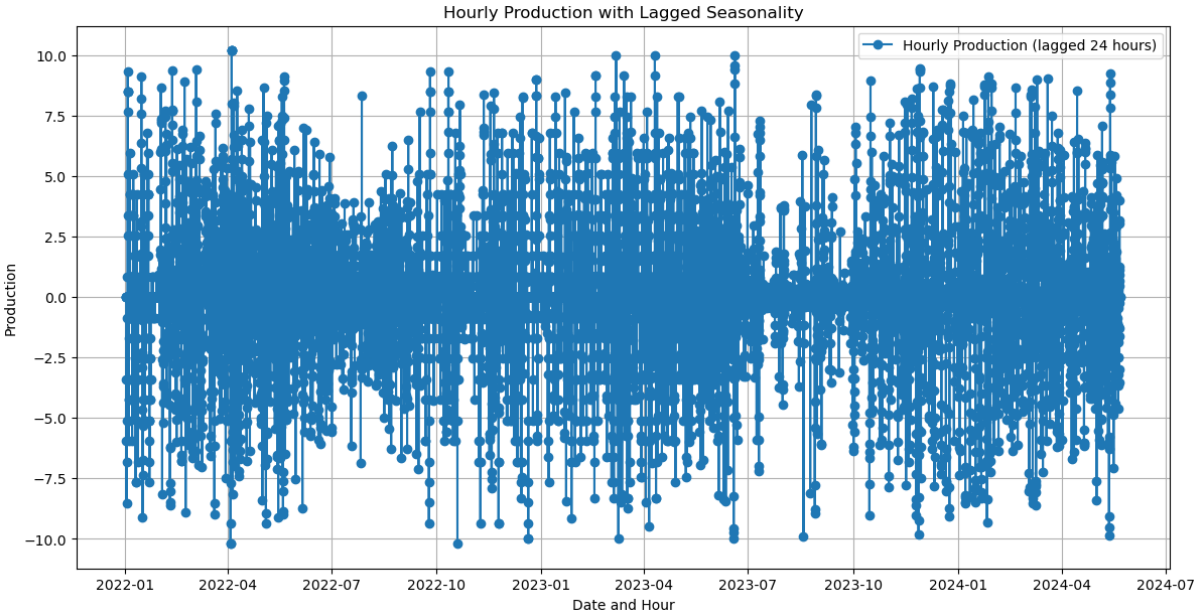
Comparison of Original and Detrended Production for First 50 Days



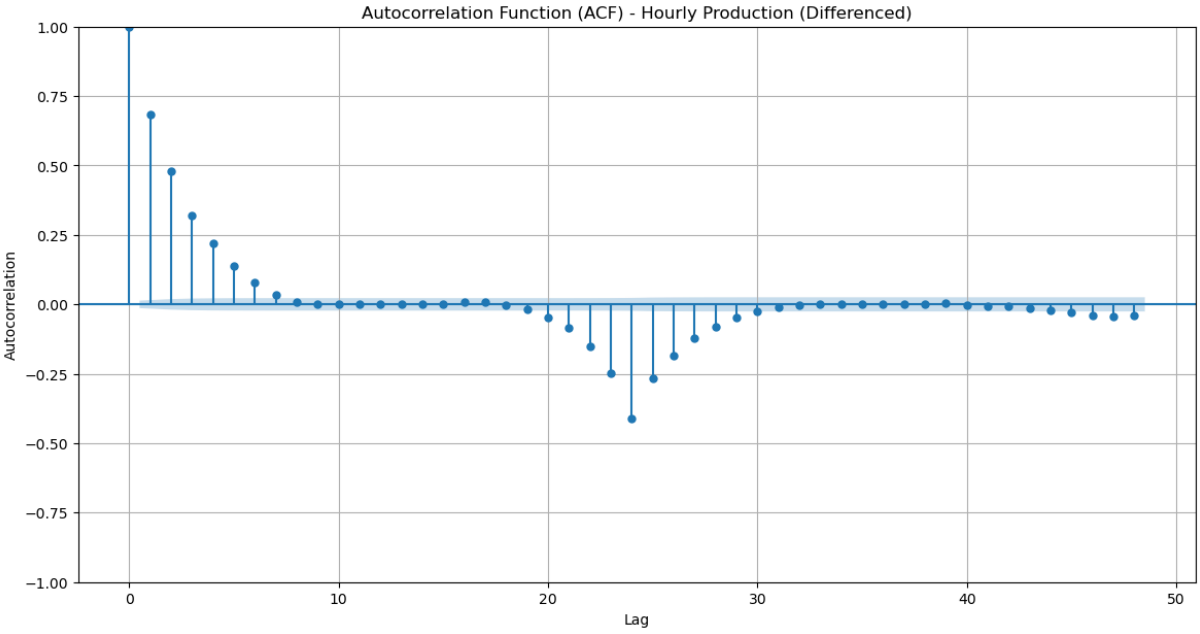
Hourly Production for Last 10 Days

- Furthermore, we utilized the ACF plot on the deseasonalized data to confirm the absence of seasonality.

- To mitigate the effects of seasonality, we employed an alternative approach involving the addition of lag diff columns at intervals of 1 and 24. This technique effectively dampened the impact of seasonality on the dataset.



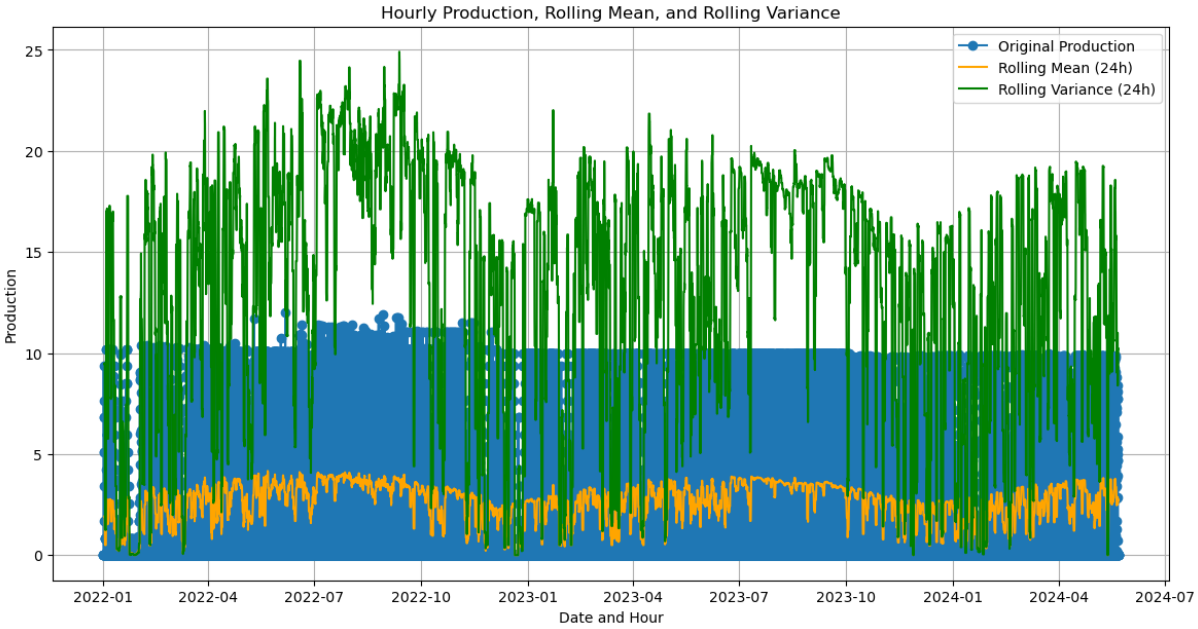
Hourly Production Plot for Lagged Data



ACF Plot for Differenced Data

- Subsequently, we subjected both the original production data and the data with lag differencing to the KPSS Unit Root test. The results revealed the original data's nonstationary nature, whereas the differenced data demonstrated stationarity.

- To analyze temporal trends and variability, we computed and plotted rolling mean and variance values for the production data. These analyses provided insights into the presence or absence of discernible trends over time, indicating a lack of significant trends in mean or variance data for production.

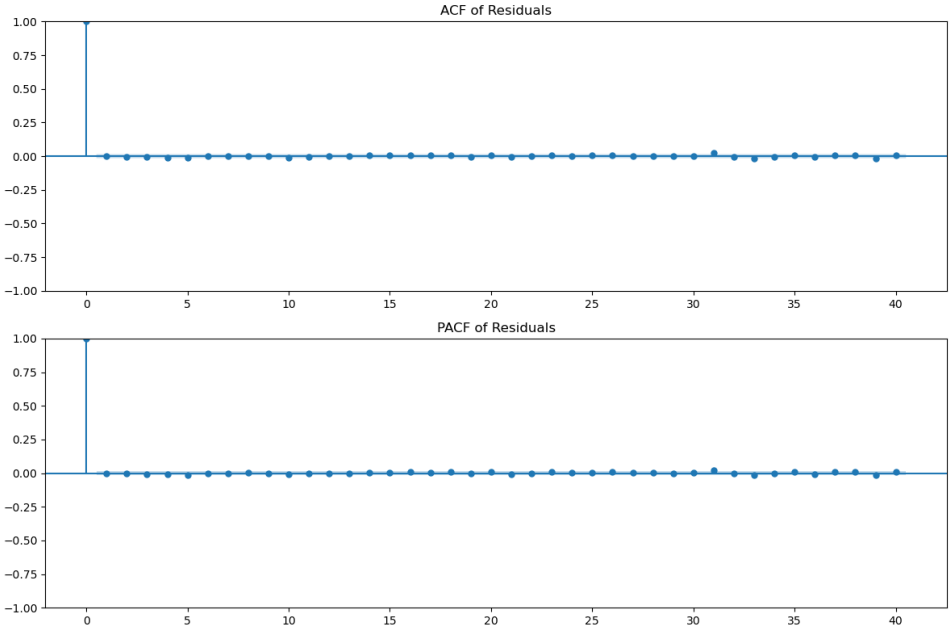


Comparison of Hourly Production, Rolling Mean and Rolling Variance

- Moving forward, we applied Ordinary Least Squares (OLS) regression modeling to both the original and differenced datasets to assess their predictive capabilities.

- The linear regression model with original data yielded an R2 of 0.759, signifying a moderate degree of explanatory power. In contrast, models incorporating differenced data, particularly those integrating lag1 and lag24, exhibited improved R2 values, with lag24 outperforming lag1.

- Residuals analysis, conducted via the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), revealed residual autocorrelation, prompting further consideration of ARIMA models.



ACF and PACF Plot for Residuals

- To identify optimal parameters for ARIMA modeling, we employed the autoarima function, which systematically evaluates various parameter combinations to minimize evaluation metrics such as AIC or BIC.

- Although the autoARIMA initially suggested (0,0,0) parameters, closer examination of ACF and PACF plots revealed spikes at PACF, leading us to select p (autocorrelation parameter) as 2 and d (differencing order) as 1.

- Given the presence of daily seasonality, we opted for SARIMA (Seasonal ARIMA) models over traditional ARIMA models.

- We experimented with different SARIMA parameter combinations, including SARIMA(2,1,0)(0,1,1)24, to identify the most suitable model configuration.

- Finally, we conducted a thorough assessment of error metrics such as Mean Absolute Deviation (MAD), Mean Squared Error (MSE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) to ascertain the performance of each model configuration and determine the most effective forecasting model.

Results:

OLS Model Metrics:
MAD: 1.337007515831581
MSE: 3.2833649003980296
AIC: 84886.21193687376
BIC: 86881.2156297267

ARIMA Model Metrics:
MAD: 3.216482704736275
MSE: 13.843958551407557
AIC: 91152.20552184162
BIC: 91175.3805775368

SARIMA Model Metrics:
MAD: 3.2237961315392547
MSE: 13.870484850606617
AIC: 91186.92106939963
BIC: 91233.26256783317

By comparising the metrics we decided that the best model is OLS model.

Conclusions and Future Work:

In conclusion, our approach to forecasting solar power production at the Edikli GES involved a comprehensive analysis of weather variables and production data. We began by reshaping the data and conducting a descriptive analysis to understand its characteristics. Utilizing

correlation matrices, rolling statistics, and seasonal decomposition, we identified key patterns and trends. We implemented various models, including linear regression, ARIMA, and SARIMA, leveraging the autoARIMA function for parameter selection. Our approach emphasized data-driven methods and model evaluation, ensuring robustness and accuracy in forecasting. Through continuous refinement and experimentation, we achieved promising results, showcasing the effectiveness of our methodology in addressing real-world energy forecasting challenges. Moving forward, further enhancements and extensions, such as incorporating additional features and refining model architectures, hold potential for advancing our forecasting capabilities.

To improve results and have a better forecasts, we can use machine learning algorithms such as Random Forest Regressor, Gradiant Boosting Regression or Decision Regression Tree. Another way to get better results may be using an hybrit model like spllitting hours into sets and using different models for different sets.

Code:

https://github.com/BU-IE-360/spring24-ardaergene/blob/main/IE%20360%20GROUP%20PROJECT/360%20PROJECT%20GROUP.ipynb