Predicting Solar Energy Generation using Time Series and Machine Learning Models

Ayse B Sengul, PhD

Springboard Data Science Capstone Project

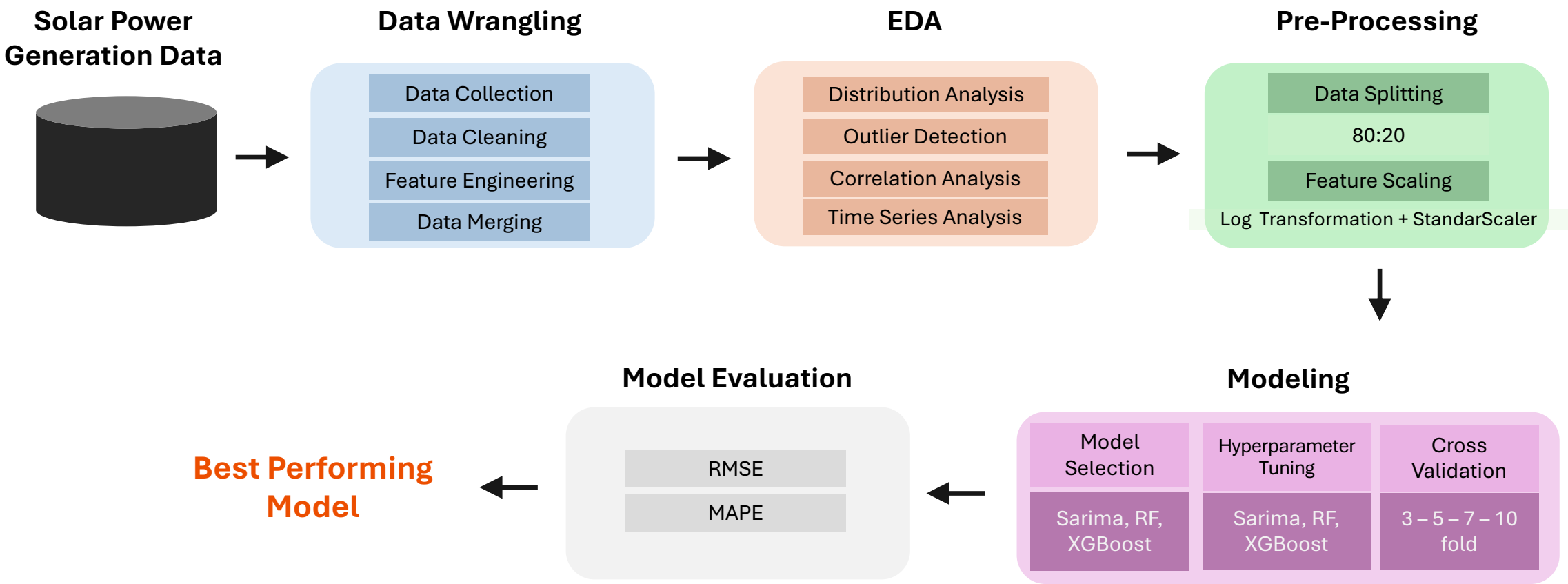
July 16, 2024

Problem Statement

- **Challenge:** Variability in environmental factors such as solar irradiance, temperature, and weather conditions makes accurate solar energy prediction difficult.
- **Importance:**
  - Efficient grid operation.
  - Economic benefits (1.56% reduction in general costs - $ 46.5 million).
  - Environmental impacts (32% reduction in CO_2 emissions).

Develop and evaluate advanced time series and machine learning models to improve the accuracy of solar energy generation prediction.

Flowchart of the Solar Energy Generation Prediction



RF: Random Forest, RMSE: root mean squared error, MAPE: mean absolute percentage error

Data Overview

- **Source:**
  - Solar Power Plant Data (Generation and Weather data)
  - 15 min intervals over 34 days.
  - May 15 – June 17, 2020.
- **Size:**
  - The generation data – (67698, 7)
  - The weather data – (3259, 6).
- **Key Features:**
  - 'DATE_TIME', 'AC_POWER', 'AMBIENT_TEMPERATURE', 'MODULE_TEMPERATURE', 'IRRADIATION

Data Wrangling

- **Converted** 'DATE_TIME' to datetime.
- **Extracted** 'DATE', 'TIME', 'MONTH', 'HOURS', and 'MINUTES'.
- **Dropped** redundant columns ('PLANT_ID' and 'SOURCE_KEY').
- **Merged** datasets on 'DATE_TIME'.
- Converted **categorical** 'SOURCE_KEY' to **numerical** 'INVERTER' and dropped the original 'SOURCE_KEY'.

Exploratory Data Analysis (EDA)

- **Histograms and KDE Plots**: Visualized numerical column distributions.
- **Box Plots**: Identified outliers.
- **Pair Plots and Correlation Matrix Heatmap**: Explored numerical feature relationships.
- **Time Series Plots**: Analyzed trends for AC_POWER, IRRADIATION, and MODULE_TEMPERATURE.
- **Daily Analysis Plots**: Examined total AC_POWER generation, IRRADIATION, and MODULE_TEMPERATURE by date.
- **Efficiency Analysis Plots**: Assessed inverter efficiency and trends over time.

Exploratory Data Analysis (EDA)

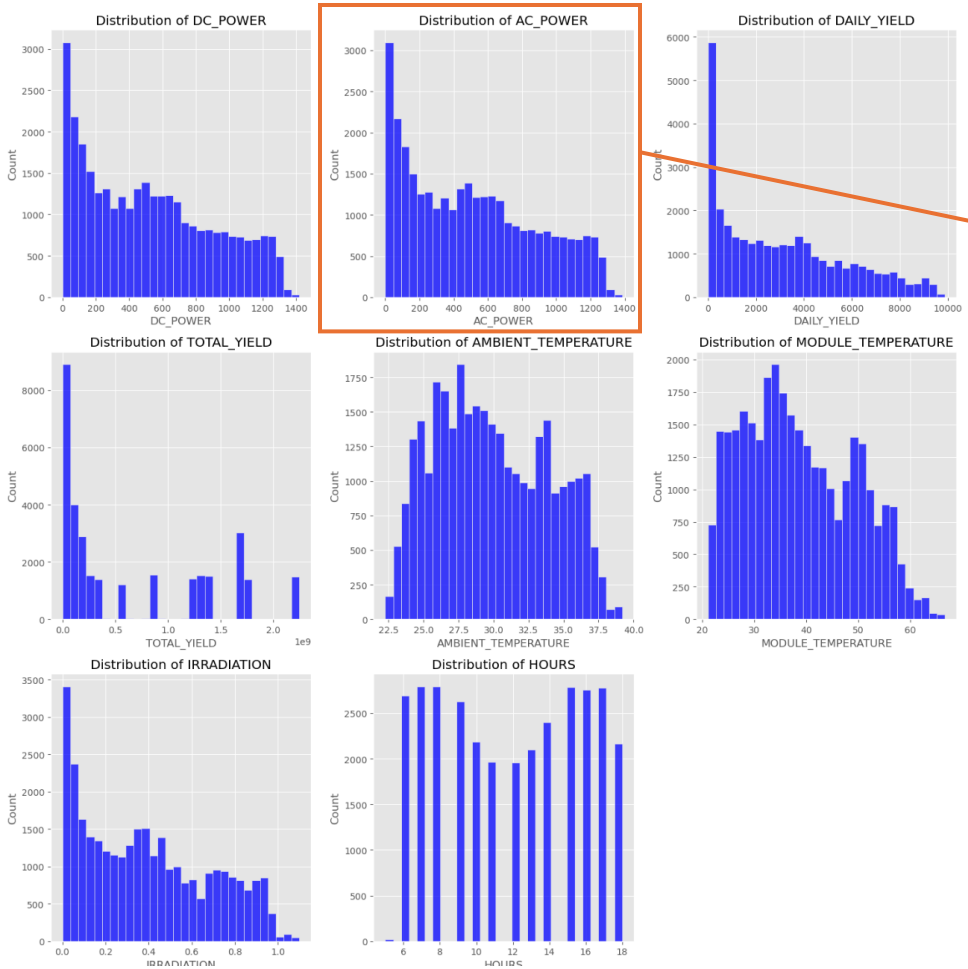
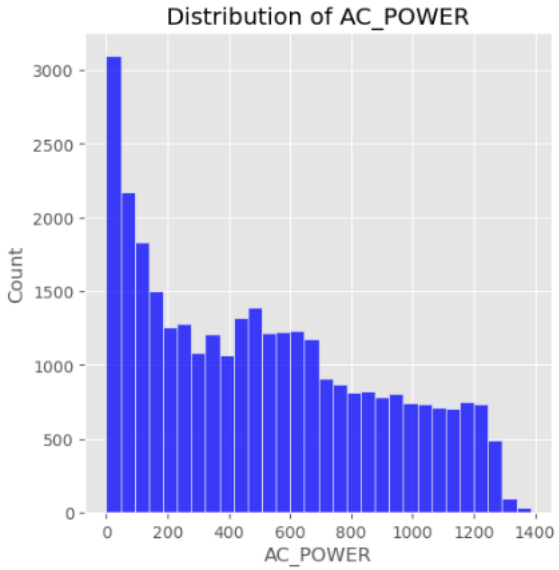


Figure 1. Distribution of numerical columns after filtering out zero values.



AC_POWER

AC_POWER values spanned from 0.2 – 1420 kW.

Most data btw 400 – 600 kW.

Mean value of AC_POWER – 521 kW.

Exploratory Data Analysis (EDA)

AC_POWER has a very strong positive correlation with **DC_POWER** , a high positive correlation with **IRRADIATION** and, **MODULE_TEMPERATURE**, and moderate correlation with **AMBIENT_TEMPERATURE**.

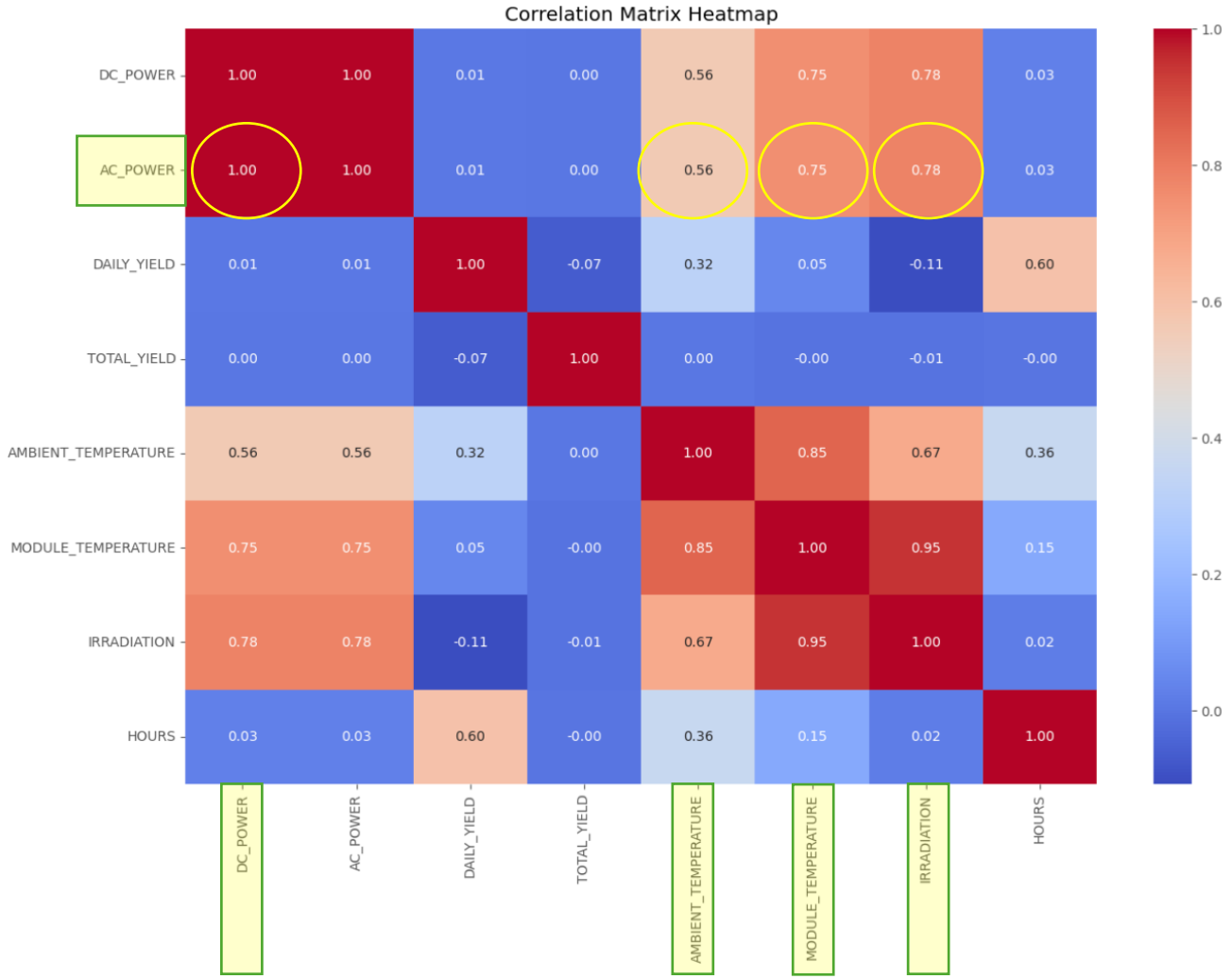


Figure 2. Correlation Heatmap of Numerical Features.

Exploratory Data Analysis (EDA)

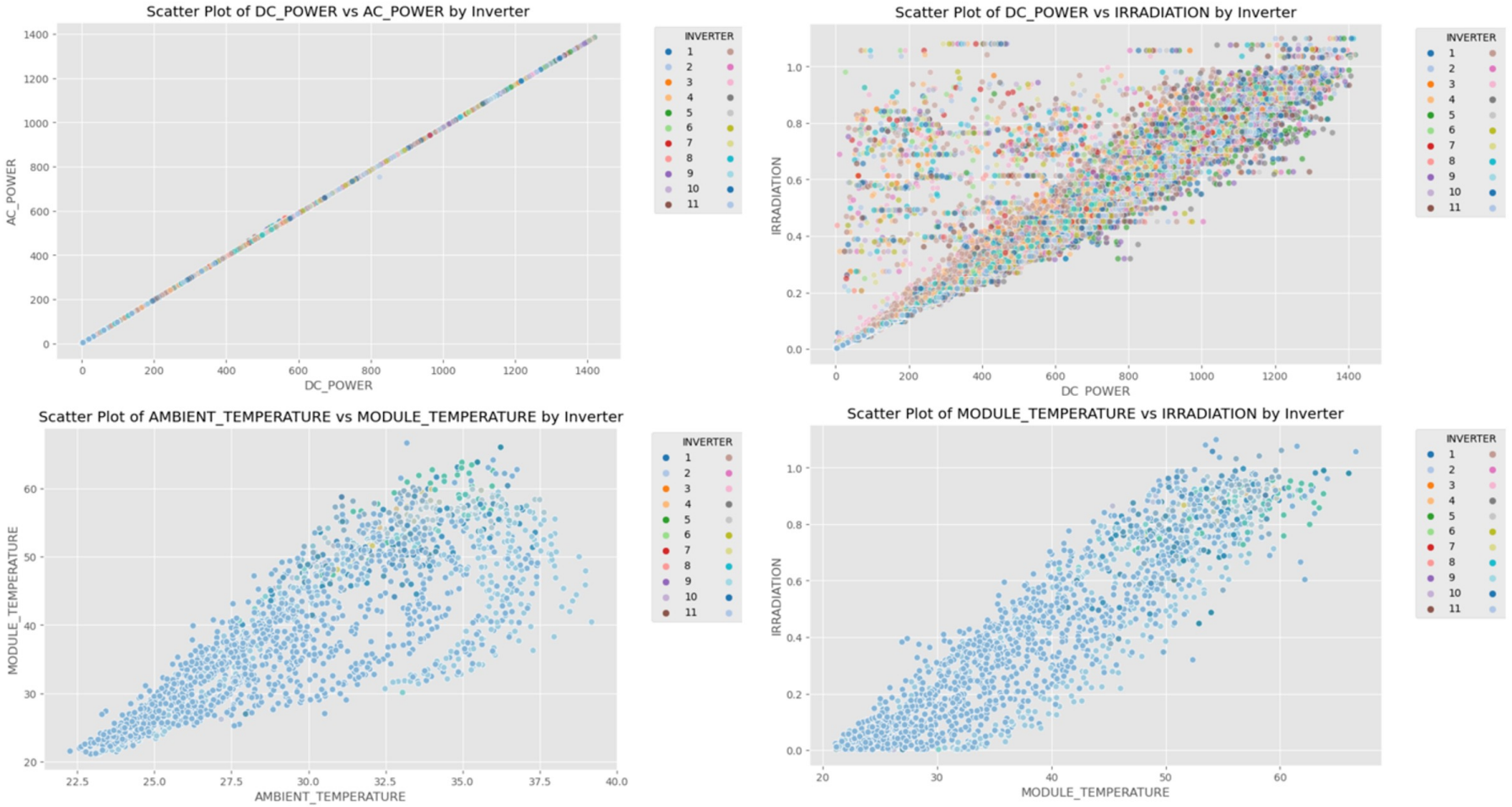


Figure 3. Scatter plot of strongly correlated features.

Exploratory Data Analysis (EDA)

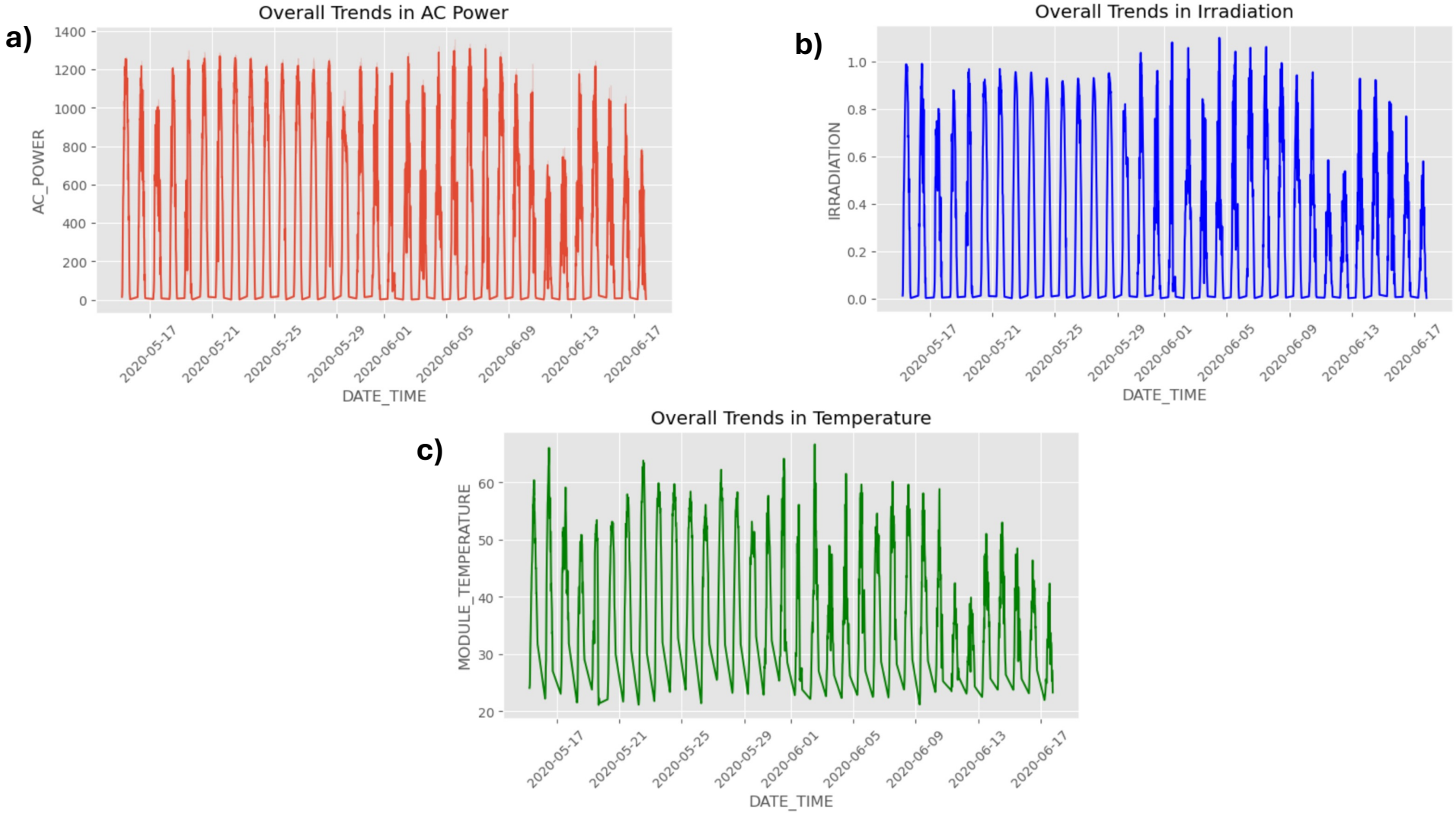
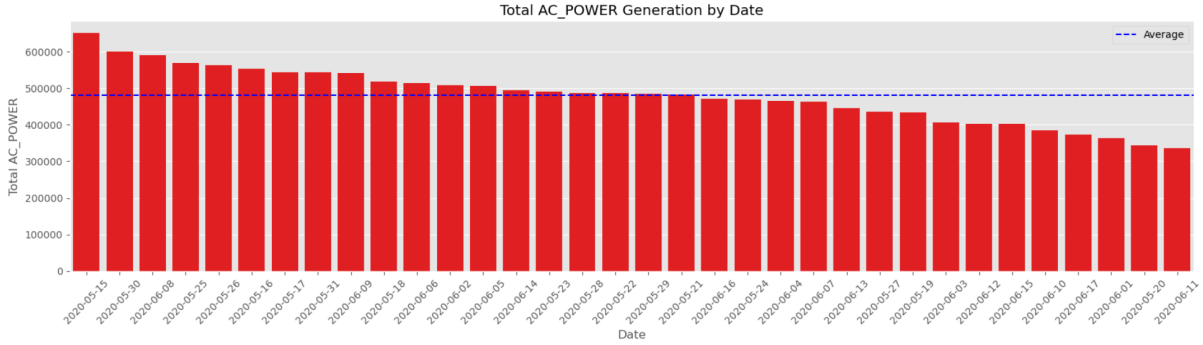


Figure 4. Time trend in a) AC_POWER, b) IRRADIATION, and c) MODULE_TEMPERATURE.

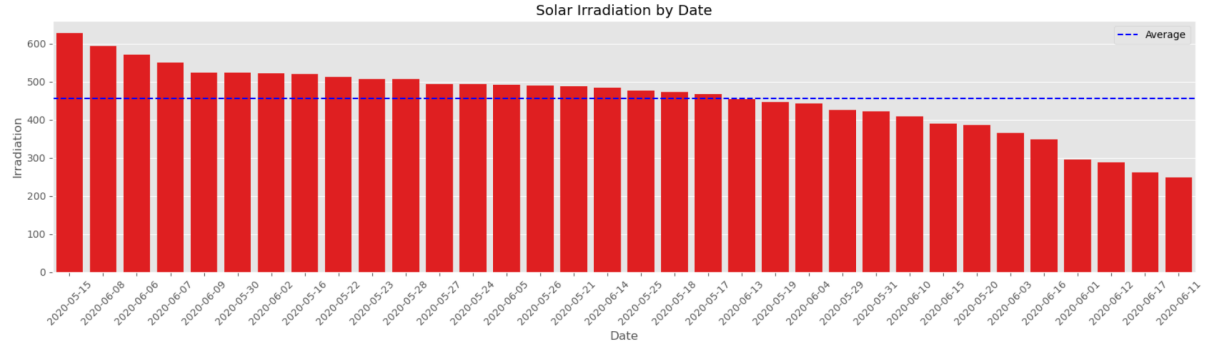
Exploratory Data Analysis (EDA)

a)



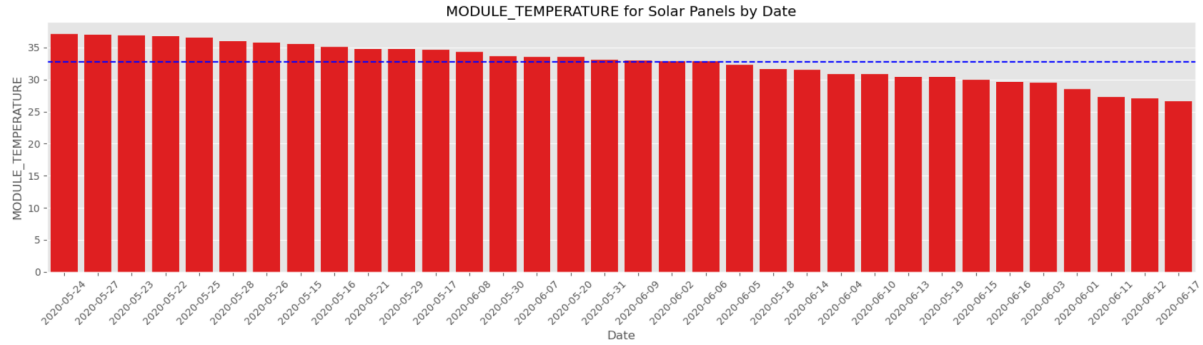
Highest: May 15
Lowest: June 11

b)



Highest: May 15
Lowest: June 11

c)



Highest: May 24
Lowest: June 17

Figure 5. a) AC_POWER generation, b) Solar IRRADIATION and c) MODULE_TEMPERATURE by date.

Exploratory Data Analysis (EDA)

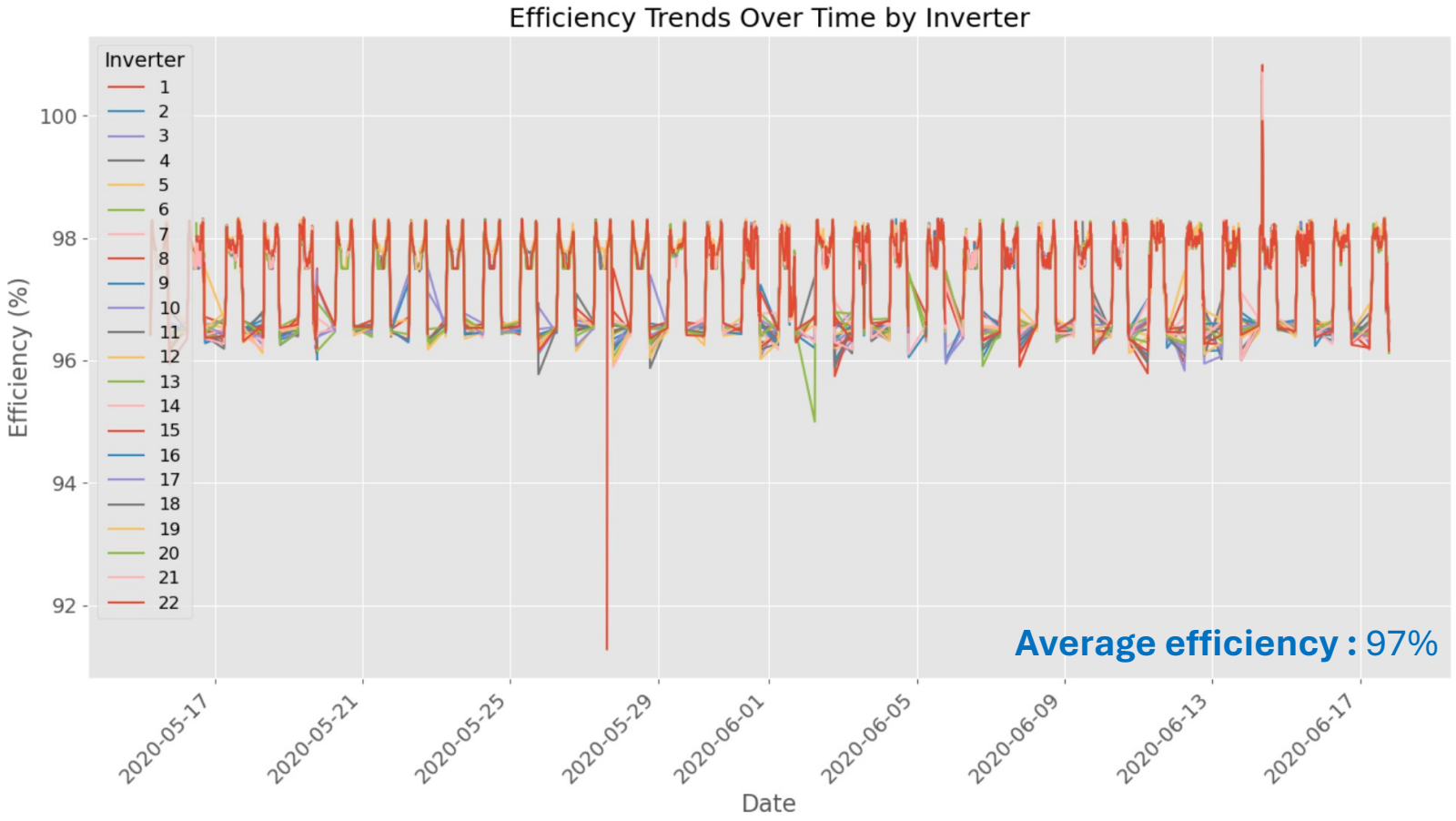


Figure 6. Efficiency Trends over Time.

Pre-Processing

- **Converted** 'DATE_TIME' to datetime and set as index.
- Applied sine and cosine **transformations** to 'HOURS' and 'MINUTES' .
- **Dropped** redundant columns ('DATE', 'TIME', 'HOURS', 'MINUTES').
- **Aggregated** data by taking the mean of selected columns per timestamp.
- **Resampled** data to 15-minute intervals
- **Interpolated** missing values .

Pre-Processing

Figure 7. Decomposition of time series data.

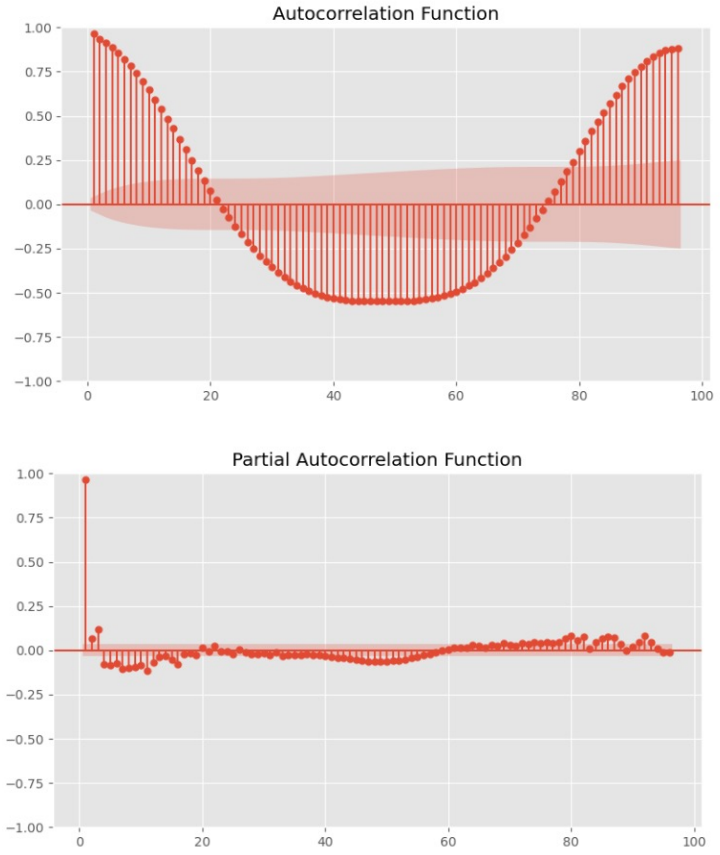


Figure 8. ACF and PACF plot.

Modeling

- **Models**: SARIMA, Random Forest, XGBoost
- **Evaluation metrics**: RMSE, MAPE.
- **Data splitting:** 80-20 ratio
- **Feature scaling**: Log transformation and StandardScaler
- **Hyperparameter tuning**: Grid Searhing
  - SARIMA: p, d, q (range of 0 to 2)
  - Random Forest: 'n_estimators': [100, 200, 300], 'max_depth': [None, 10, 20], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4], 'max_features': ['sqrt', 'log2']
  - XGBoost: 'n_estimators': [100, 200, 300], 'max_depth': [3, 5, 7], 'learning_rate': [0.01, 0.05, 0.1]
- **Cross-validation**: 3-fold, 5-fold, 7-fold, 10-fold.

Modeling – SARIMA

Initial Sarima
order=(1, 0, 0) seasonal_order=(1, 1, 0, 96)

Tuned Sarima
order=(1, 1, 1) seasonal_order=(1, 1, 0, 96)

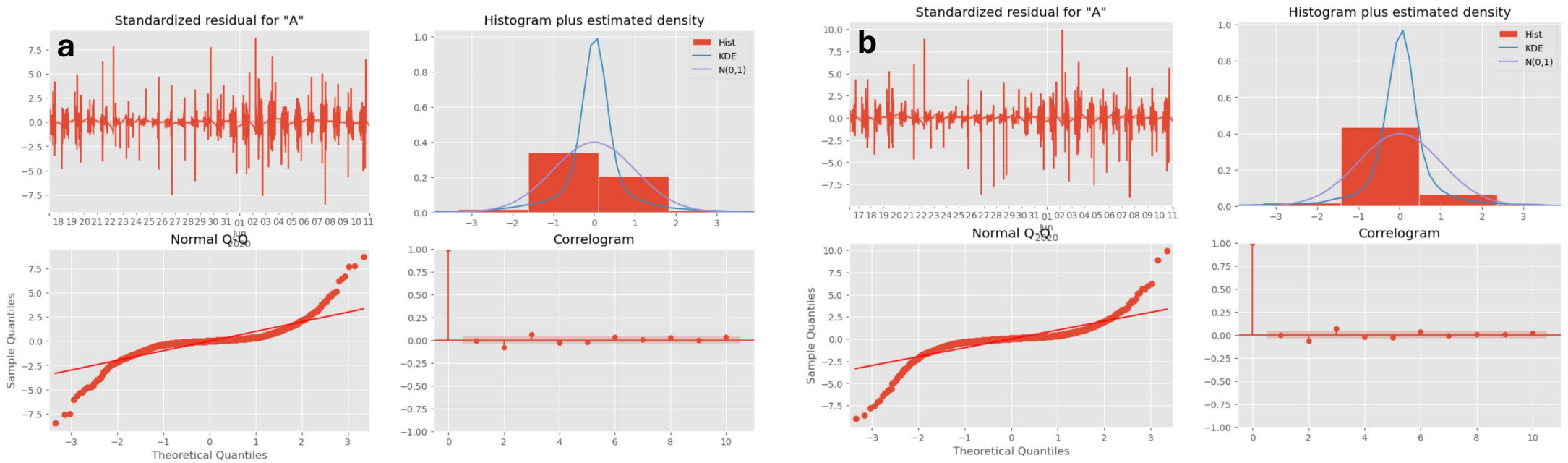


Figure 9. The diagnostic plots of the a) initial and b) tuned SARIMA model.

Modeling – SARIMA

Initial Sarima
order=(1, 0, 0) seasonal_order=(1, 1, 0, 96)

Tuned Sarima
order=(1, 1, 1) seasonal_order=(1, 1, 0, 96)

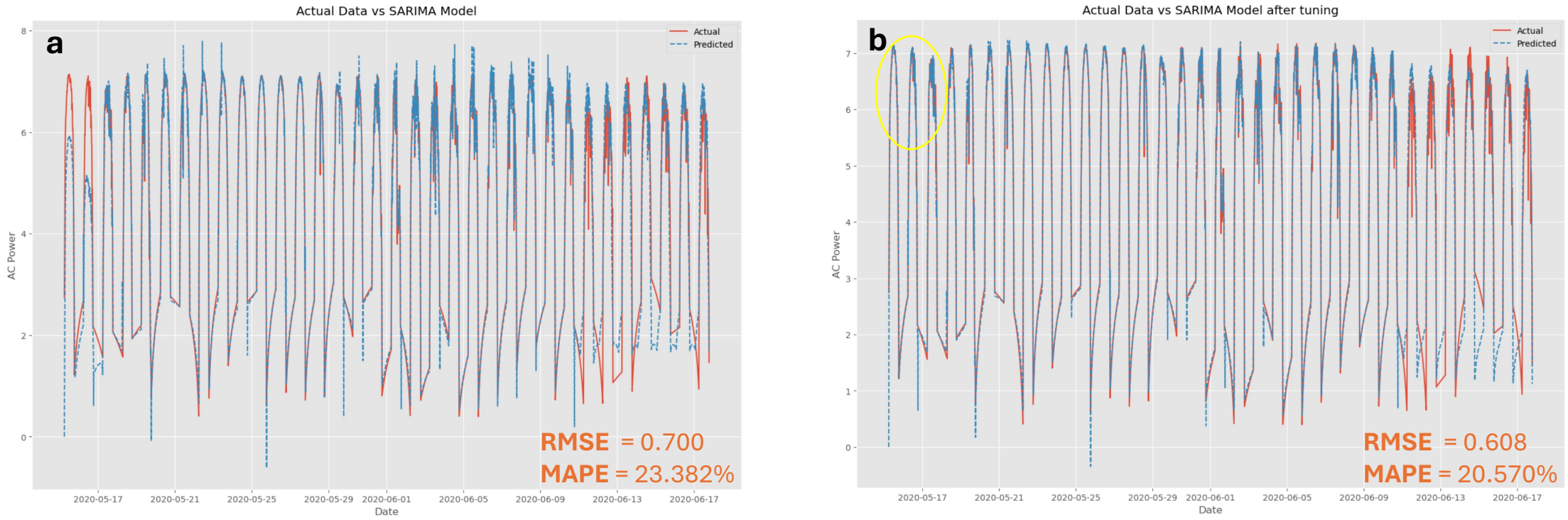


Figure 10. Actual vs. predicted values for the a) initial and b) tuned SARIMA model.

Modeling – Random Forest

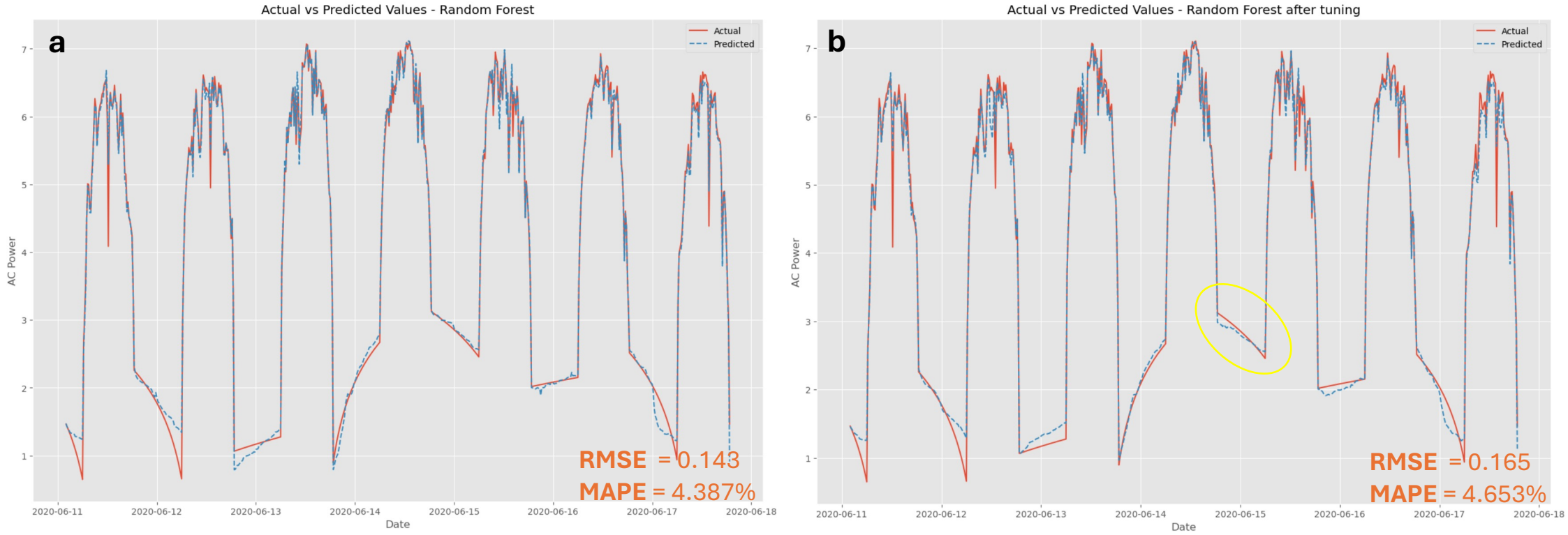


Figure 11. Actual vs. predicted values for the a) initial and b) tuned Random Forest model.

Modeling – Random Forest

Table 1. Random Forest Model Cross-Validation Results.

Cross-Validation	Max Depth	Max Features	N_Estimators	Best RMSE Score	Mean CV on Training Set	Mean CV on Testing Set	Standard Deviation on Testing Set
3-fold	20.0	sqrt	200	0.1114	0.1131	0.1481	0.0308
5-fold	NaN	sqrt	300	0.1078	0.1088	0.1453	0.0342
7-fold	20.0	log2	300	0.1048	0.1055	0.1384	0.0433
10-fold	NaN	sqrt	300	0.1035	0.1034	0.1326	0.0405

Modeling – Random Forest

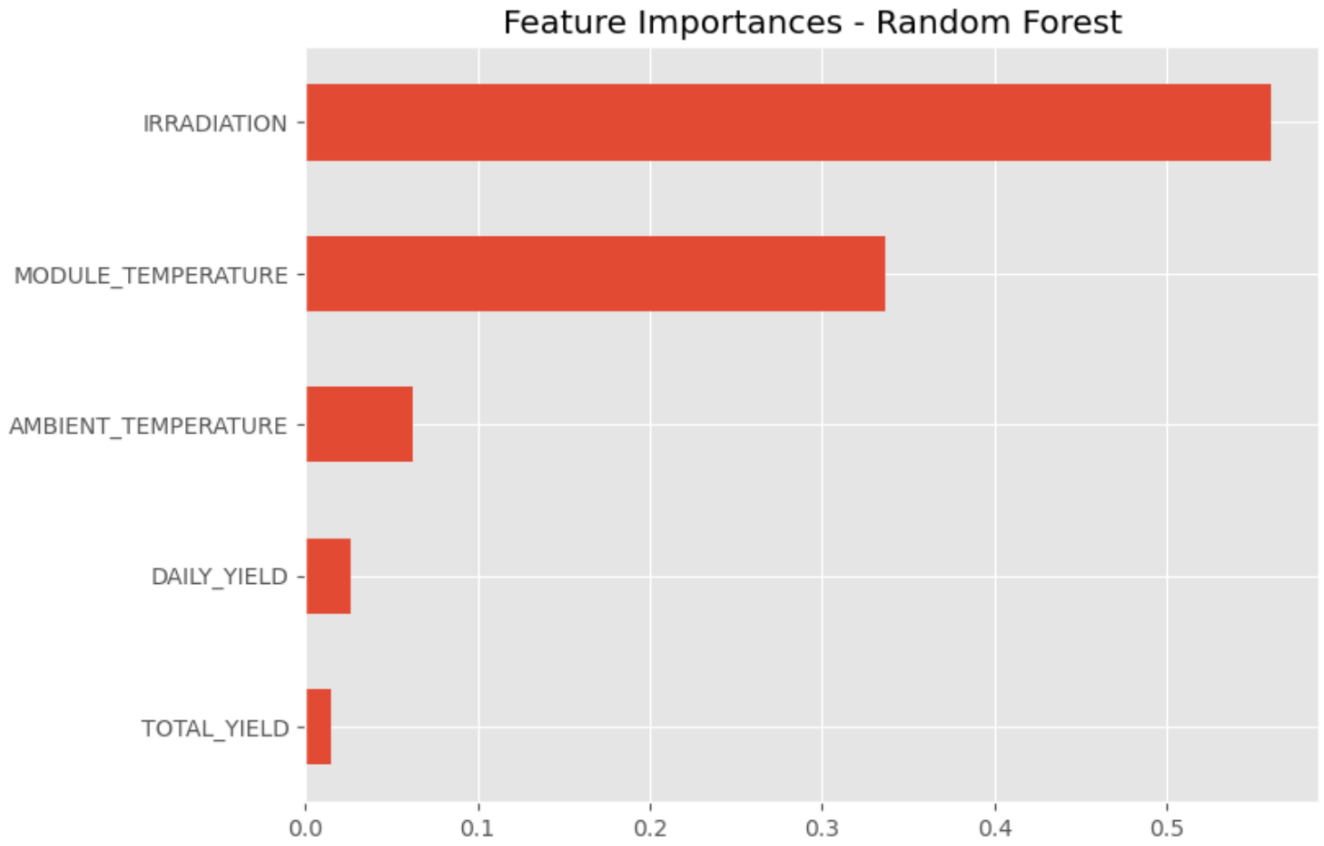


Figure 12. Feature importance of the Random Forest model.

Modeling – XGBoost

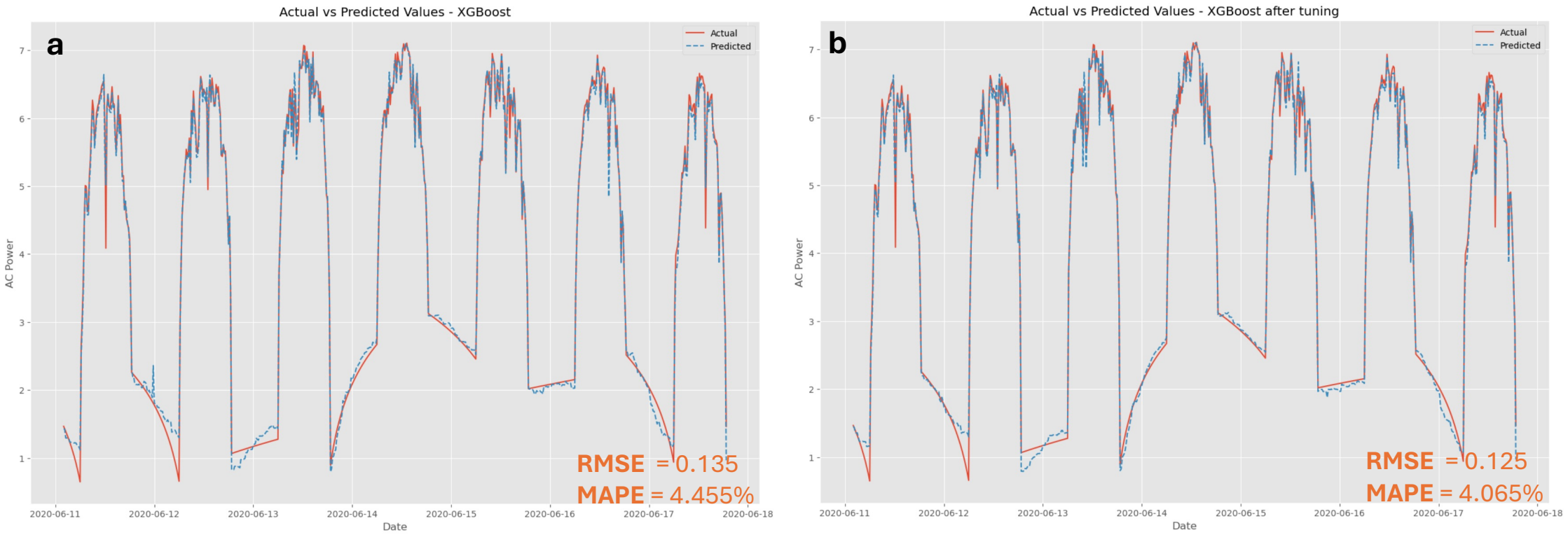


Figure 13. Actual vs. predicted values for the a) initial and b) tuned XGBoost model.

Modeling – XGBoost

Table 2. XGBoost Model Cross-Validation Results.

Cross-Validation	Max Depth	N Estimators	Learning Rate	Best RMSE Score	Mean CV Score on Training Set	Mean CV Score on Testing Set	Standard Deviation on Testing Set
3-fold	5	300	0.1	0.0870	0.08690	0.1021	0.0041
5-fold	5	300	0.1	0.0803	0.08031	0.0968	0.0227
7-fold	5	300	0.1	0.0854	0.08540	0.0887	0.0221
10-fold	5	300	0.1	0.0789	0.07890	0.0899	0.0318

Modeling – XGBoost

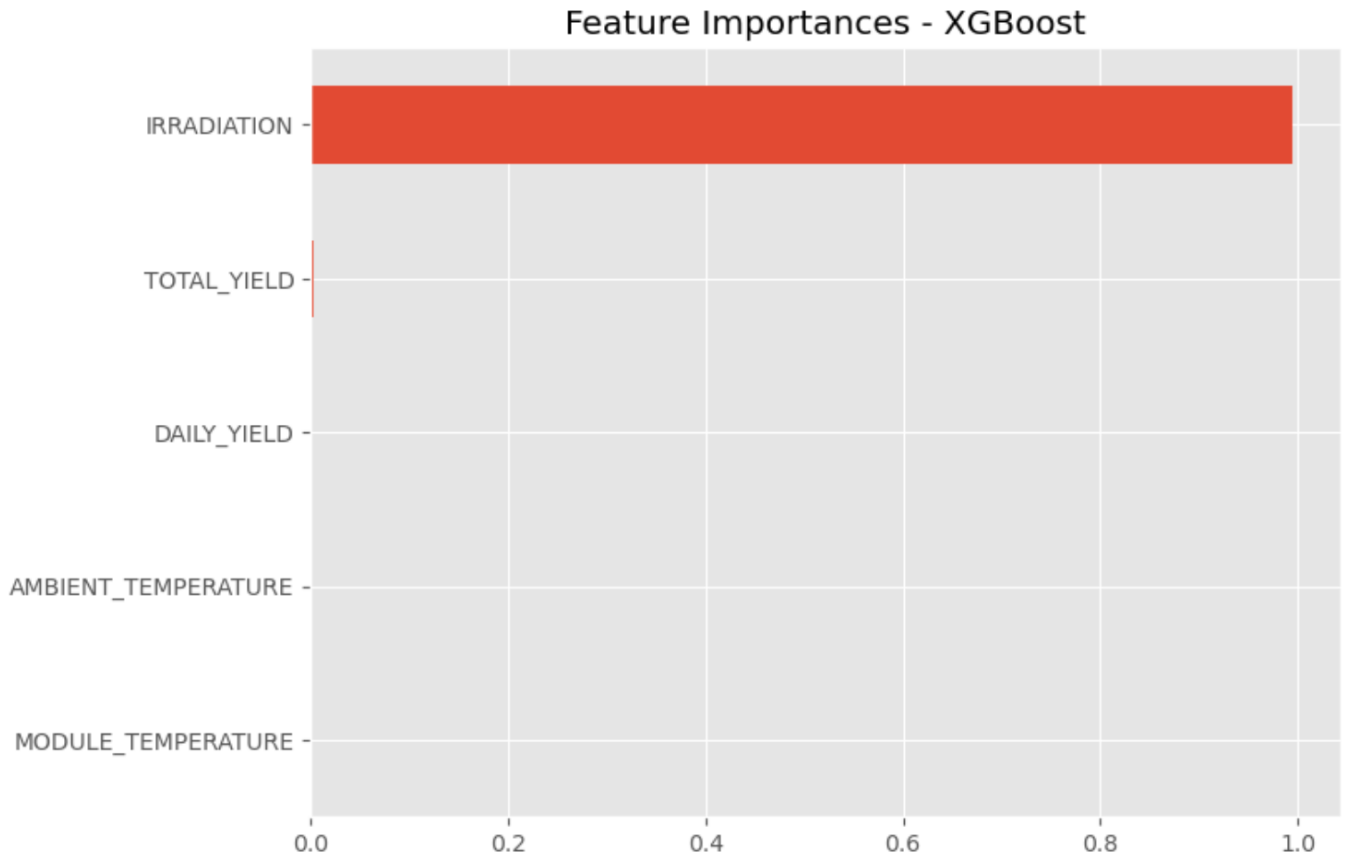


Figure 14. Feature importance of the XGBoost model.

Results

Table 3. Comparison of models.

Model	RMSE	MAPE	CV RMSE Score on Testing Set	Training Time (minutes)	Best Parameters	CV Standard Deviation
SARIMA	0.608	20.570	-	143	order=(1,1,1), seasonal_order=(0,1,1,96)	-
Random Forest	0.165	4.665	10-fold: 0.1035	6	max_features=sqrt, n_estimators=300	0.0405
XGBoost	0.125	4.065	10-fold: 0.0789	1	max_depth=5, learning_rate=0.1, n_estimators=300	0.0318

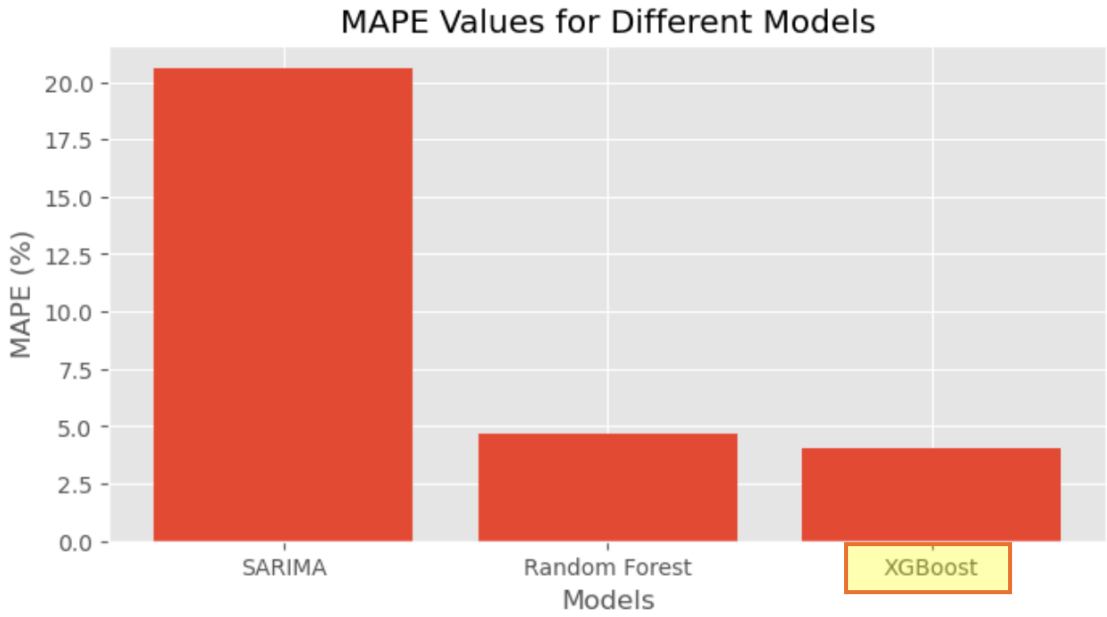


Figure 15. Comparison of MAPE values for different models.

Conclusion

- **Best Model:** XGBoost (Lowest RMSE: 0.125, MAPE: 4.065%)
- **Efficiency:** XGBoost (1 minute training time)
- **Random Forest:** Good performance, slight overfitting
- **SARIMA:** Least effective, high RMSE and MAPE, long training time

Future Work

- Incorporate additional environmental factors (**humidity, wind speed, cloud cover**) to improve model accuracy.
- Enhance feature engineering techniques (**cyclical features, lag features, rolling statistics**).
- Explore alternative time series models like **Prophet** for better performance.
- Investigate deep learning models such as **LSTM** networks for time series forecasting.

Questions?