Applied Analytics Practicum - Enverus

Team 5

hdupouy3@gatech.edu
oalmansour3@gatech.edu
jgarcia340@gatech.edu

Abstract **-** This project presents a study to evaluate the impact of various locations and modeling techniques on predicting solar farm energy generation's performance. This research project was undertaken in collaboration with Enverus, an energy technology company renowned for its innovative software, data, and services in the energy market. This study compares and contrasts different modeling options by incorporating location-specific source data and alternative modeling techniques. This paper discusses the investigation's problem statement, objectives, and desired goals towards solar energy generation.

Table of Contents

Table of Figures

Table of Tables

1.INTRODUCTION

1.1. Enverus

Enverus, an energy technology company, is the driving force behind this research endeavor. Since its establishment in 1999, Enverus has emerged as a prominent provider of energy market data, analytics, and technology solutions (Enverus, 2023). With a commitment to optimizing operations and fostering a deeper understanding of energy markets, Enverus offers innovative software, data, and services to facilitate informed decision-making within the energy sector. Indeed, Enverus gives energy firms the platforms, tools, and applications they need to be adaptable and thrive in a challenging and changing market environment. Additionally, Enverus provides invaluable services such as expert guidance, data analysis, and market intelligence, further solidifying its position as a leader in the field.

1.2. Objective

1.2.1. Project objectives

This project's primary objective is to study the effect of incorporating location-specific source data and various modeling techniques to predict energy generation performance from solar farms. We will compare and contrast several models in an effort to improve energy generation in a specific location. This will be achieved through extensive research and experimentation. Enverus will provide us with California ISO solar output dataset. This dataset contains the target variable and timestamps extracted from California ISO toward the actual solar megawatt generation (California ISO, 2023). This research will contribute to the existing literature on the prediction of the energy generation performance of solar farms, and it will also assist energy companies like Enverus in making informed decisions in order to thrive in the ever-changing energy market.

1.2.1. Value and Magnitude

This project holds substantial value for the company by generating valuable insights and knowledge concerning the prediction of energy generation. By integrating location-specific source data and advanced modeling techniques, the project aims to enhance solar farm performance models' accuracy and predictive performance. By improving the precision and reliability of these models, energy companies can make more informed decisions regarding their operations, resource allocation, and identification of regions with high solar energy generation potential. This optimization

of processes and resource utilization can lead to cost savings and overall performance improvements, enabling companies to stay competitive in a rapidly changing market.

The magnitude of this project is noteworthy due to its comprehensive approach and incorporation of multiple elements. This project requires external data and time-series forecasting models to produce predictions. Based on the combination of the data provided and retrieved, it needs to leverage machine learning models and visualization libraries. Indeed, the scale of the project involves analyzing a substantial amount of data, with up to 50,000 rows and 20 input variables per data source. Furthermore, advanced modeling techniques, such as Random Forest and eXtreme Gradient Boosting (XGBoost), capture complex relationships and patterns within the data. The project also includes a thorough review of relevant literature on solar generation forecasting to incorporate best practices and industry knowledge. Given the comprehensive nature of the project and its potential impact on decision-making processes within energy companies, the magnitude of this endeavor is significant.

1.3. Problem statement

Existing methods for forecasting solar farm performance have limitations due to a lack of consideration for location-specific data and reliance on traditional modeling tools. This leads to inaccurate and unreliable predictions. To address this issue and improve the accuracy and foresight of solar farm performance models, evaluating and comparing the outcomes obtained by utilizing location-specific data sources and advanced modeling methodologies is necessary. By doing so, energy providers can enhance operational efficiency, optimize resource allocation, and identify areas with the highest potential for solar power production.

Research questions:

- When comparing generic data sources and traditional modeling methods to solar energy-generating performance prediction, how does integrating location-specific data sources and sophisticated modeling techniques affect the accuracy and predictive performance?
- Moreover, how might these enhancements help energy firms maximize operations, allocate resources, and pinpoint areas with the most significant potential for solar power generation?

1.4. Hypotheses

Our first hypothesis mentioned that location and modeling methods affect solar farm performance. Using location-specific source data, which accounts for physical closeness to the place under investigation, the models should reflect various locales' distinctive traits and variances. Solar irradiance, weather patterns, topographical characteristics, and environmental factors may be included. This study's design matrix includes data from nearby macro, meso, and micro-regions. This study aims to solve research questions and understand solar generation performance aspects.

Indeed, this hypothesis also predicts that sophisticated algorithms like Random Forest, XGBoost, and others can capture complicated linkages and nonlinearities in data. These methods may reveal patterns and relationships that linear regression models miss. By merging advanced modeling approaches with location-specific data sources, solar farm performance projections will be more precise and dependable. This helps energy businesses optimize operations, allocate resources, and find solar energy-generating zones.

The study will investigate numerous data sources and model classes to test our notion. If considerable gains are found, the models' accuracy, robustness, and generalizability will demonstrate the suggested approach's usefulness.

2.LITERATURE REVIEW

In our specific study, the geographical location under investigation is within the California region. However, it is important to acknowledge that the exact geographical location within California plays an important role in the performance of solar farms, as pointed out by Miao, Ning, Gu, Yan, and Ma (2018). These authors highlight the variability of solar radiation across different locations and propose a comprehensive framework that considers factors such as latitude, longitude, and weather patterns to assess their impact on energy generation. The results of their research highlight significant differences in performance depending on the specific geographical location, further emphasizing the need for accurate modeling techniques (Miao, Ning, Gu, Yan, & Ma, 2018). Therefore, for our project, we will strive to identify a precise location within California that allows for robust data analysis and the use of valued data science techniques for predicting time series data.

Solar power generation is influenced by several environmental factors, as highlighted by Singh and Singh (2021). These factors include solar irradiance, temperature, humidity, dust, shading, and wind speed. In addition, the technical design characteristics of photovoltaic cells, such as the materials used in their manufacture, also affect the power generation capabilities of the cells (Chikate & Sadawarte, 2015). Therefore, we don't have access to photovoltaic cell specifications in our data collection process. However, we aim to collect additional data points in order to obtain a comprehensive overview of all significant factors affecting the performance of solar energy generation.

Similar studies in the field of solar energy generation have consistently emphasized the need for robust modeling techniques. Hobbs et al. (2022), in their research of probabilistic solar prediction using the probabilistic Watt-Sun model, underlined the need for a reliable and accurate model despite the inherent uncertainties associated with prediction predictors. Similarly, Aksoy and Genc (2023) proposed an ensemble model that combines several basic models, such as Random Forests and Gradient Boosting Machines, to improve the accuracy of energy production forecasts. Their research demonstrates the effectiveness of ensemble learning in capturing complex relationships between input variables and energy production (Aksoy & Genc, 2023).

3.METHODOLOGY

For this study section, we have implemented an extract, transform, and load (ETL) data pipeline, as depicted in Figure 1. The ETL process is crucial for collecting, preparing, and organizing the data to ensure its suitability for analysis and modeling.

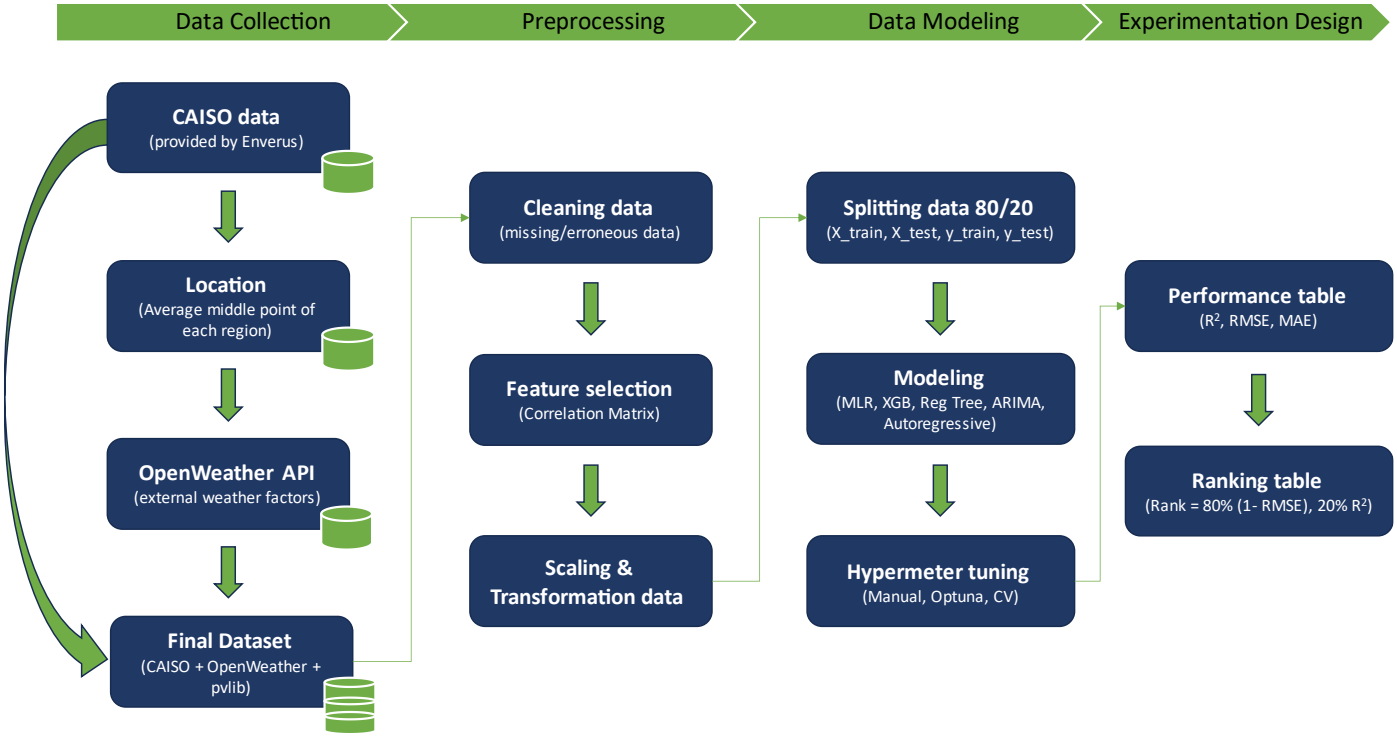


Figure 1. ETL Data Pipeline

3.1. Data Collection

The data collection process for this project involves gathering relevant information to assess the impact of geographical locations and modeling techniques on predicting solar energy generation's performance. We will construct a design matrix containing California CAISO and OpenWeather data combinations. These data sources will include macro, meso, and micro-regions in the vicinity of solar location.

Firstly, we will need to understand the supplied data. As part of our collaboration with Enverus, we have been provided a simplified dataset of timestamps and an actual measure of megawatts generated (Appendix A). This 'Actual' measure is based on an average of the three main regions of California. Consequently, as this measure is an average location of the three main regions, we must determine a specific location to retrieve future data. In this case, we will take the middle location for each region based on latitude and longitude and will make an average of those coordinates to

obtain our central point of California. This process will allow us to coordinate with the data provided by Enverus.

After defining California's middle point location, we will identify and select appropriate data sources that provide environmental factors towards weather factors (temperature, humidity, dust, shading, and wind speed). In this regard, we will use the specific location data from OpenWeather to join it to Enverus' provided data (OpenWeather, 2023)(Appendix A).

Additionally, as part of the requirement from Enverus, we retrieved the measures of Global Horizontal Irradiance (GHI), Direct Normal Irradiance (DNI), and Diffuse Horizontal Irradiance (DHI) from pvlib (Sandia National Laboratories & pvlib python Development Team, 2023). GHI measures the total amount of solar radiation received on a horizontal surface at a given location (HOMER Pro, 2023). It includes both direct solar radiation and diffuse radiation scattered by the atmosphere. The GHI is critical to solar power generation because it provides valuable information about the available solar resources at a given location. By understanding the GHI, solar power system designers can estimate the potential energy output of solar panels and optimize their placement and orientation. GHI data helps determine the feasibility and efficiency of solar projects, allowing for better decision-making when using solar energy to generate electricity.

Indeed, we will attempt to maintain high data quality in this process. To ensure the reliability and integrity of the collected data, we will perform quality assurance with checks. This includes verifying data sources, checking for missing or erroneous data points, and addressing any data inconsistencies or outliers. Indeed, this full dataset represents 26185 rows and 31 columns. Any necessary data cleansing or corrections will be carried out to improve the overall quality of the dataset.

3.2. Preprocessing

The process of data preprocessing is essential in ensuring that the data collected for our study is properly prepared for analysis and modelling.

We will start the explanatory data analysis by cleaning the data. Any missing, erroneous, or inconsistent data points identified during the data collection phase will be addressed. Missing values may be imputed using appropriate techniques, while erroneous or inconsistent data points are corrected or removed based on predefined criteria.

Then, we will proceed with feature selection. We will analyze the full dataset to identify relevant features that contribute significantly to predicting solar energy performance. This step involves establishing a correlation matrix between variables,

checking the numerical variables' density, inspecting the categorical variables' distribution, and using domain knowledge to select the most informative features for modeling.

Last but not least, we will proceed with imputation, feature scaling, and transformation. For our missing values, we will proceed with the imputation of the mean value in the numerical variables to avoid the instabilities in the predictions and increase our prediction accuracy (Thomas & Rajabi, 2021). Then, our features will require scaling or transformation to ensure compatibility and optimal performance, as machine learning algorithms can interpret only numerical data (Rebala, Ravi, & Churiwala, 2019). We will normalize the numerical data and transform the categorical data into dummy variable columns with the library Scikit-learn.

3.3. Data Modeling

Based on the preprocessed collected data, the data modeling phase focuses on developing models to predict solar energy generation performance. It includes partitioning the dataset into training and test subsets and optimizing model parameters.

Firstly, we will develop and train the models using the preprocessed dataset for each selected modeling technique. Consequently, we will split the preprocessed data into a train and a test set. The test set will represent the last four months of data, while the train set will be the large remaining part. Then, we will split the dataset with the response variable (y = Actual) and the dependent variables (X). Finally, we will have four data frames corresponding to X_train, y_train, X_test, and y_test.

Furthermore, based on the study's objectives and the literature review findings, we will explore a range of modeling techniques. These models will include traditional statistical models to more advanced machine learning models. Indeed, we will use the following models: multi-linear regression (MLR), random forests (RF), eXtreme Gradient Boosting (XGBoost), autoregressive integrated moving average (ARIMA), and Autoregressive model.

We will have an experimental setup phase with the hyperparameters in this next step. Finding the ideal collection of hyperparameters for a particular machine-learning model is known as hyperparameter optimization (Yang & Shami, 2020). Hyperparameters, such as learning rate, number of layers, or number of units in a neural network, are parameters defined before the training process and learned from the input (Yang & Shami, 2020). Finding the best settings manually can be a time-consuming and tedious process, but it can significantly influence a model's performance. In order to allow to have some flexibility, we will use different techniques to obtain the best

parameters. In this way, our most complete model would be XGBoost with the use of Optuna and cross-validation. Optuna will automatically optimize the hyperparameters by setting a search space dynamically and pruning strategies (Akiba, Sano, Yanase, Ohta, & Koyama, 2019).

3.4. Experimentation Design

The experimentation design phase is centered around the establishment of controlled experiments for the purpose of comparing and contrasting various options for metrics.

In this section, we will be evaluating model performance using appropriate evaluation metrics with root mean squared error (RMSE) based on the mean square error (MSE), R-squared (R^2), and accuracy. To obtain these metrics, we will use the Scikit-learn library to generate them directly. However, to understand fully the meaning of each of these metrics, we will show their equation:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

The average difference between predicted and actual values is measured using the RMSE (Root Mean Square Error) metric in regression tasks (Botchkarev, 2019). In order to assess the average error between predicted and observed values in the same unit as the target variable, the square root of the mean of the squared differences is computed (Botchkarev, 2019).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

A statistical metric called R-squared (coefficient of determination) shows how much of the variance in the dependent variable can be accounted for by the independent variables in a regression model (Botchkarev, 2019). A better fit of the model to the data is indicated by higher values ranging from 0 to 1.

$$\text{Accuracy} = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} * 100$$

Our last performance metric for classification models is accuracy. It determines the proportion of all forecasts that were accurate. The result is calculated by dividing the proportion of accurate predictions by the total number of predictions, multiplied by one hundred, and it represents the model's capacity to categorize cases properly (Yin, Wortman Vaughan, & Wallach, 2019).

Lastly, we will design a performance comparison table based on the requirements from Enverus. The performance of the developed models will be compared and contrasted based on the relevant evaluation metrics. Indeed, we will analyze and compare the predictive capabilities of the models based on the equation provided by Enverus. The final ranking of our models will be based on the following equation to determine the best model:

$$Final\_Rank = (0.8 \cdot \text{Accuracy\_Rank}) + (0.2 \cdot \text{Explanability\_Rank})$$

Through this comprehensive methodology, which includes data collection, preprocessing, data modeling, and experimental design, we aim to provide valuable insights into the impact of geographical locations and modeling techniques on predicting solar energy generation performance. This approach allows systematic analysis and comparison of different design options, contributing to the advancement of accurate and reliable methodologies in the field of solar energy prediction. The following section will present our findings from the explanatory data analysis.

4.EXPLANATORY DATA ANALYSIS

In this section of the study, we will present the results of the explanatory data analysis conducted on the collected data. This analysis includes an examination of general trends as well as a correlation matrix. The findings provided valuable insights into the data set, allowing us to gain a better understanding of the trends, patterns, and relationships that exist within it.

4.1. General trends

To understand the data provided by Enverus, we have plotted the "Actual" (average power generation in MW) for the three regions over the entire time period in Figure 2. In this graph, we can clearly see a pattern for each year. The energy generation tends to be at its maximum in July for the years 2020 to 2022 included. Moreover, this pattern tends to increase each year, which would indicate a real need to focus more on this energy from Enverus for the coming years.

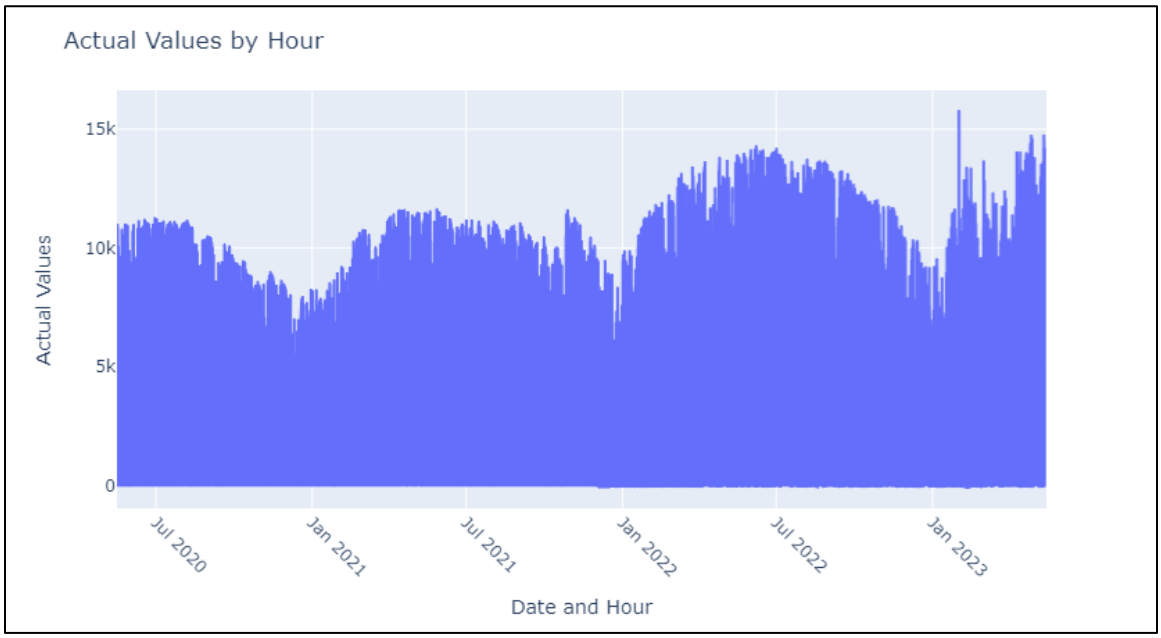


Figure 2. Actual value by hour over the entire period

We then looked at the "Actual" measurement per hour during the day in Figure 3. Here, we confirmed that energy generation would generally occur during the daylight hours from 6 AM to 9 PM.

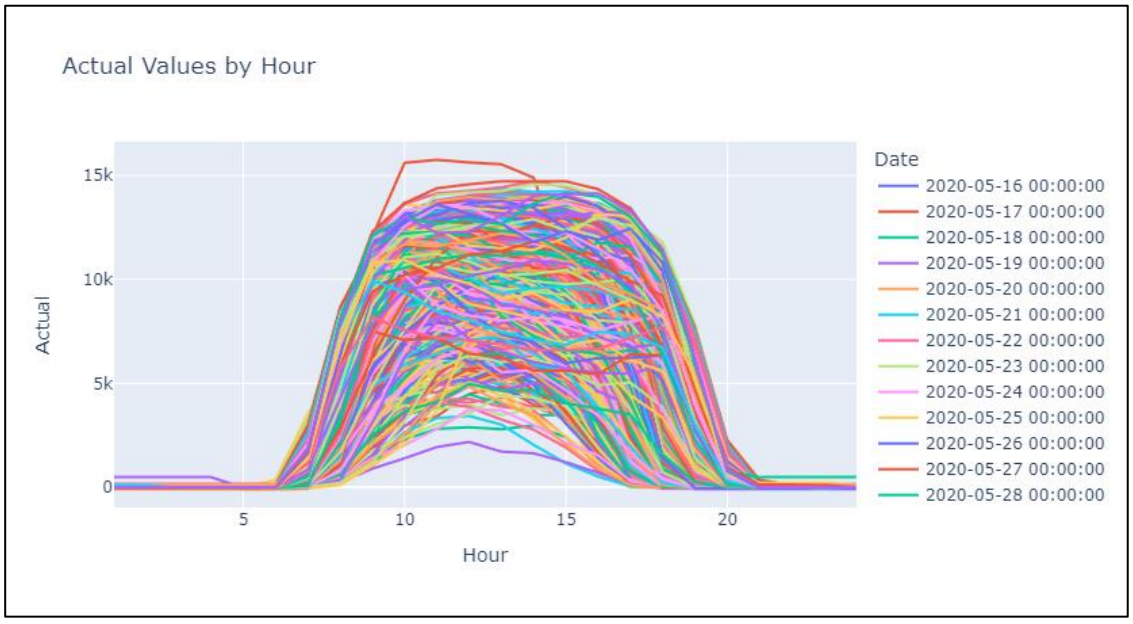


Figure 3. Actual value by hour through the day

Furthermore, we were wondering to confirm that the generation is generally greater during summer with the long daylight hours by comparing a week in summer (week 26) and another week in winter (week 51). Based on Figure 4, we can assert that the generation is greater in summer. In this regard, we can even see a similar trend between summer and winter for the generation of GHI on the right side of Figure 4.

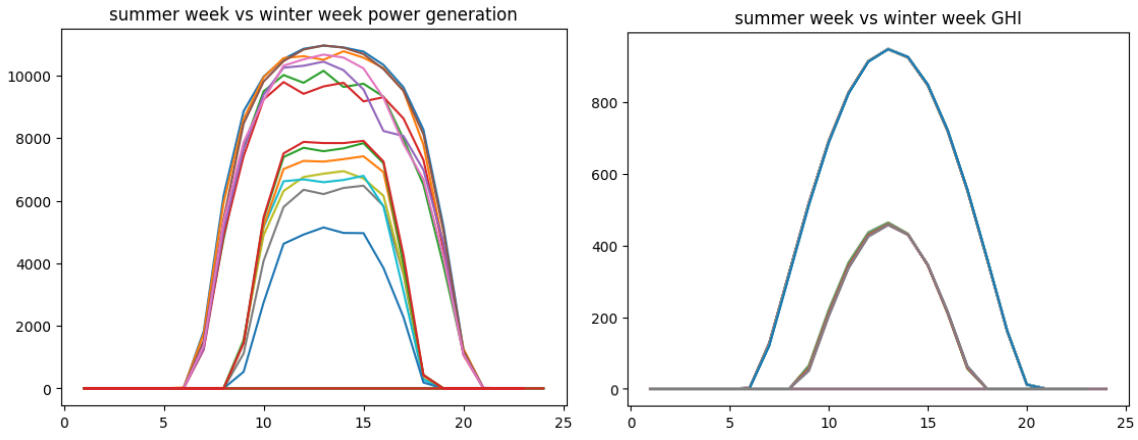


Figure 4. Comparing a week of Actual value in summer and winter

Furthermore, we determined the exact location of the middle point of California to be of latitude of 37.051548 and a longitude of -120.699371. Indeed, the exact address would be CA 152, CA 33, Merced County, California, United States. Determining the exact location helped us unlock further information. Moreover, we illustrated the Sun's path for our specific location over the period given in Figure 5. Based on Figure 5, we can also see the hourly solar zenith at hour 13 (1 PM). Also, the boundary of the region of the sky that the Sun traverses during a year is marked by the solstices. According to this diagram, there is never a time of the year when the Sun is directly overhead (zenith=0).

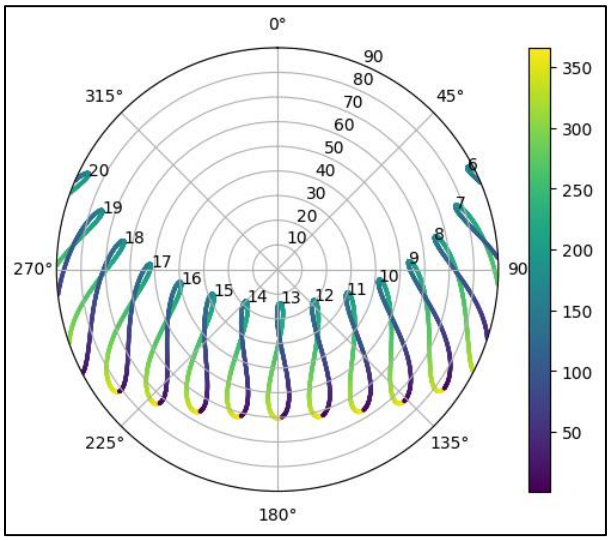


Figure 5. Diagram of the analemma

In addition, we wanted to explore the impact of GHI and cloud cover on the solar output to recognize potential correlations. However, since GHI and solar output are temporal data with daily cycles, it is more suitable to explore the trends at one point in time to eliminate any temporal dependancies. For this reason, we created a 3D plot with GHI, cloud cover, and solar output at 1:00 PM (Figure 6). The plot revealed that, the solar output at 1:00 PM (at given time point) is positively correlated with GHI, but negatively correlated with cloud cover. This observation is sensible, because the more cloudy the sky is, the less solar irradiace reach to photovoltaic cells, and hence, the lower the solar output is. In addition, looking at Figures 7 & 8, we can see the solar output over 3 weeks from October 11th 2021 to November 1st 2021 and the cloud cover over the same period. As cloud cover becomes heavier in the middle week (orange), the solar output peak tends to decreases, affirming the significance of clould cover on PV cells solar output.

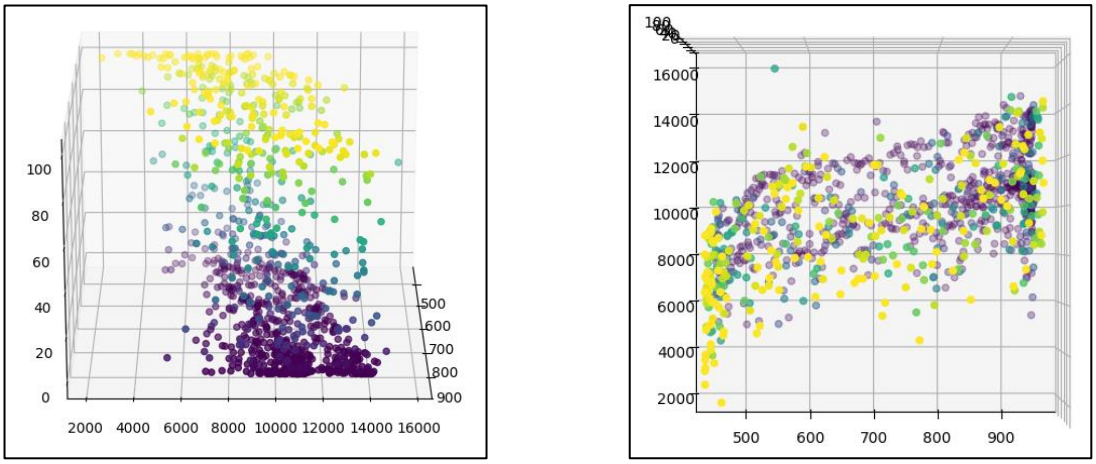


Figure 6. Cloud cover vs Solar output at 1:00 PM (left). Solar output at 1:00 PM. vs GHI (right)

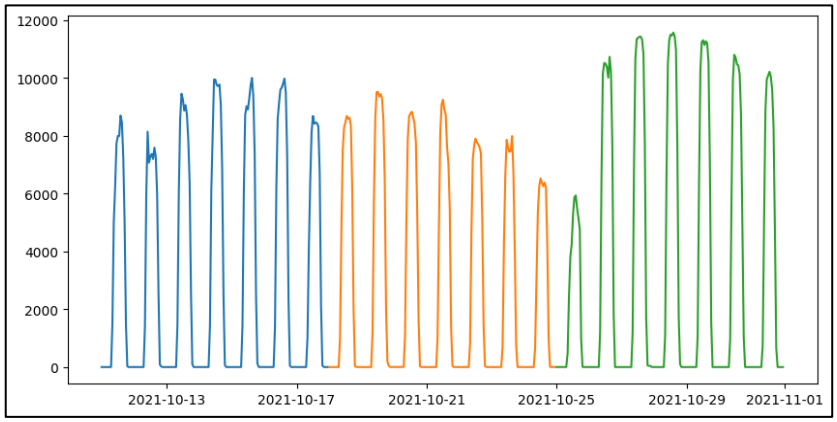


Figure 7. Actual output over 3 weeks

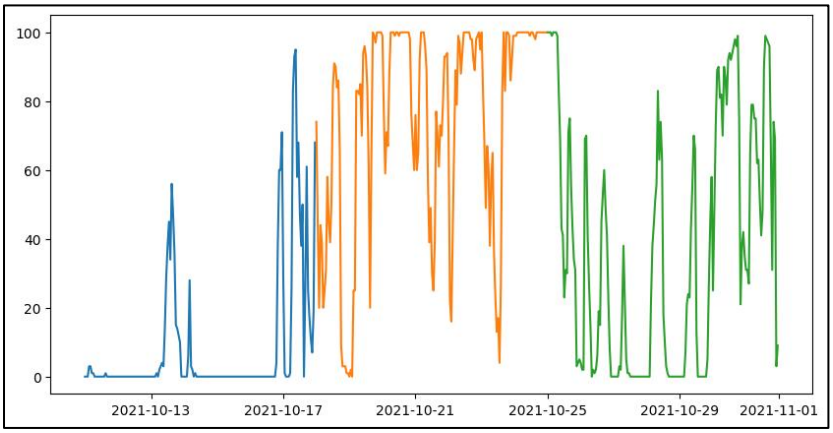


Figure 8. Cloud cover over 3 weeks

4.2. Feature selection

During this phase of the explanatory data analysis, our objective was to identify and select the most relevant columns that will be utilised in our model. To accomplish this, it is necessary to analyse the numerical data followed by the categorical data.

To understand the relationships between the numerical data, we generated its correlation matrix in Figure 9. By calculating the correlation coefficients, we learned more about how strongly one attribute was related to another, such as the different temperature metrics. These strong linear relationships would indicate a high correlation, which could be a sign of redundant or powerful characteristics. Consequently, we will filter out these numerical values.

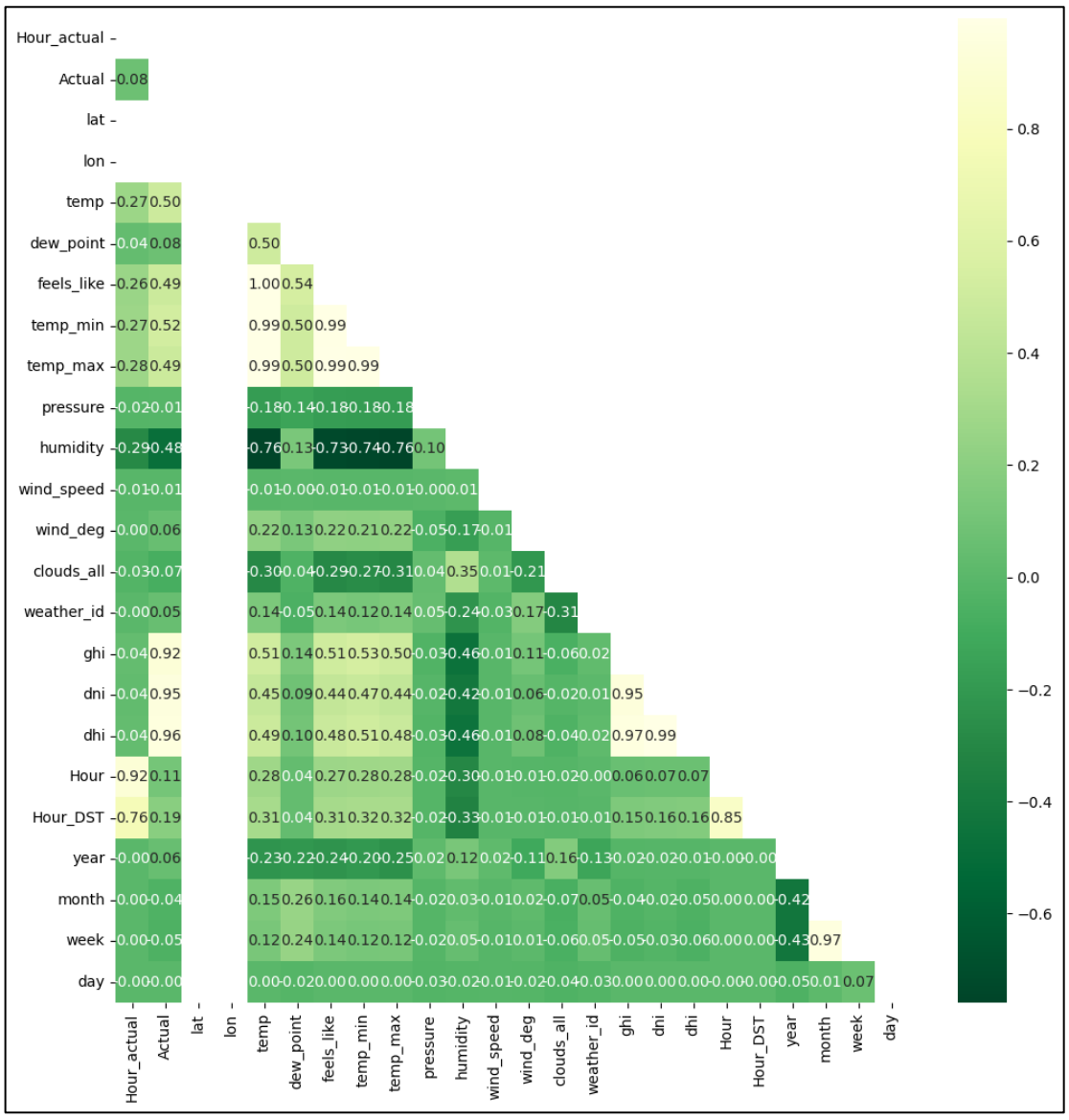


Figure 9. First Correlation Matrix

After filtering out the high correlation values, we end up with a lower correlation among all numerical values. In Figure 10, we only kept the 'ghi' (GHI), which has a high correlation of 92% with our 'Actual' values, because it is a must-have for Enverus and an anticipated correlation for energy generation.

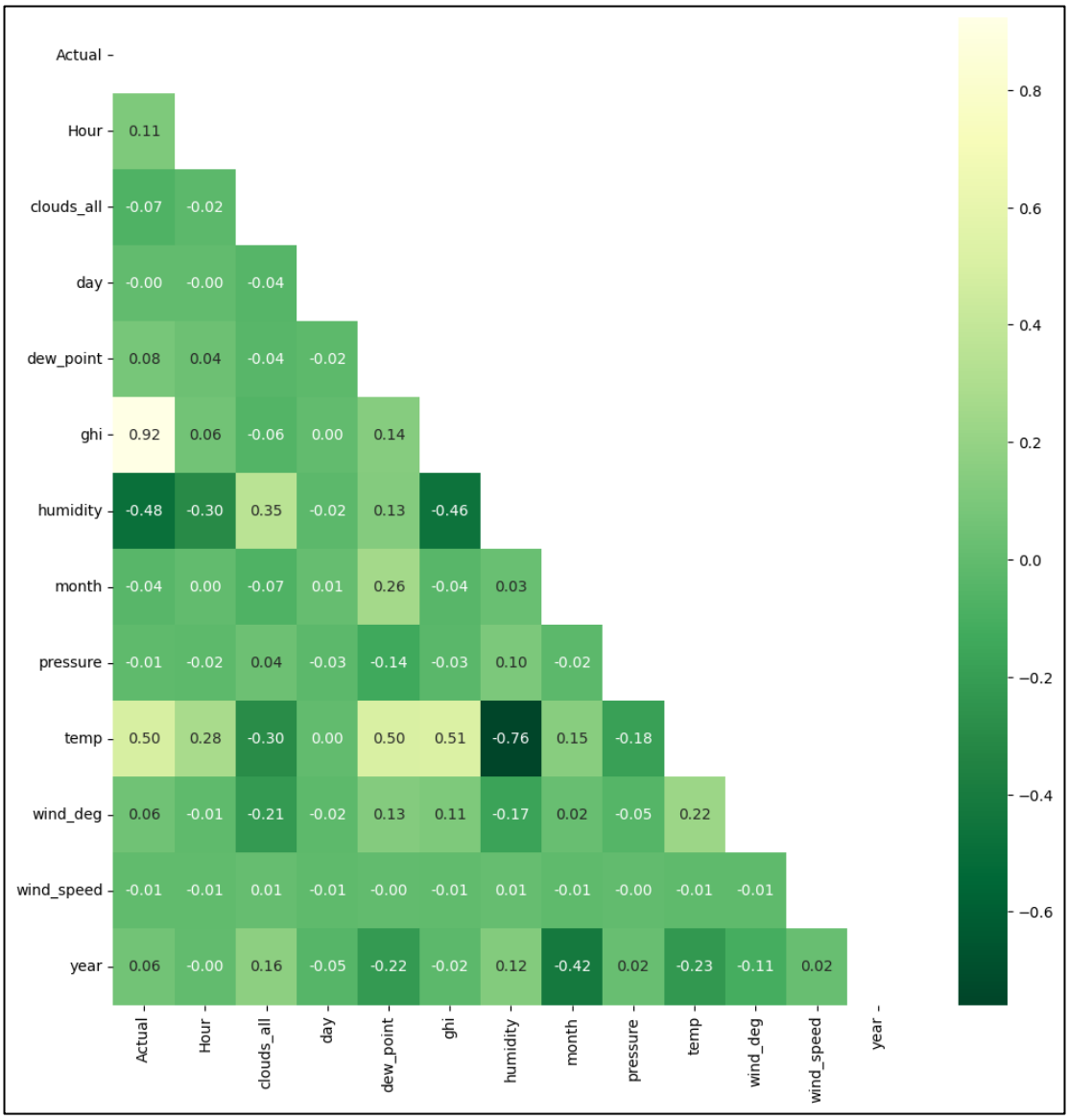


Figure 10. Final Correlation Matrix

In order to identify potential outliers or skewness, an initial analysis was conducted on the distribution of the numerical values. Figure 11 displays the line graphs representing the distributions of each numeric characteristic that were calculated. The ability to assess the central tendency and spread of the data and identify potential outliers was facilitated by this..

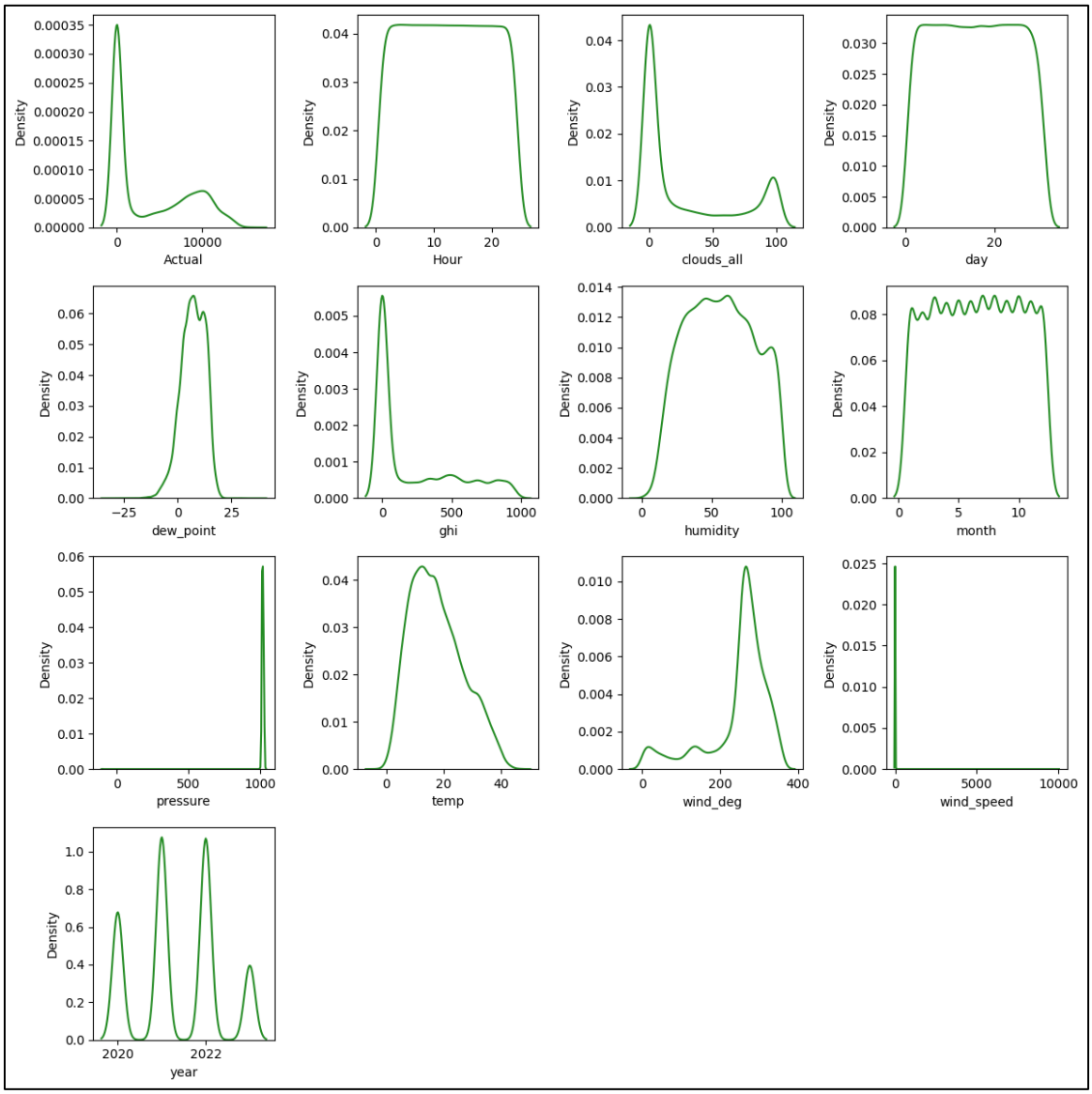


Figure 11. Density of the numerical variables

In fact, the 'Actual' and 'GHI' tend to be slightly skewed to the right. While 'pressure' tends to have a high density in the 1000s, this would suggest fewer data points in that range, indicating that the values in that range are rarer or occur less frequently. Also, "win_speed" has a high density in the zero mean, which is a distribution of numerical values centered on zero. It indicates that the majority of the data points are clustered around the center and deviate from the mean with relatively slight variance.

Regarding the categorical features, we analyzed their distributions with the bar charts in Figure 12. It allowed us to understand the distribution of classes within each categorical feature and identify potential class imbalances or rare categories. Here, we have some highly unbalanced features, like 'sky is clear' in 'weather_description' and '01n' &' 01d' in 'weather_icon', that could introduce bias into the model. At the same time, the rare categories might not provide enough information to make accurate predictions.

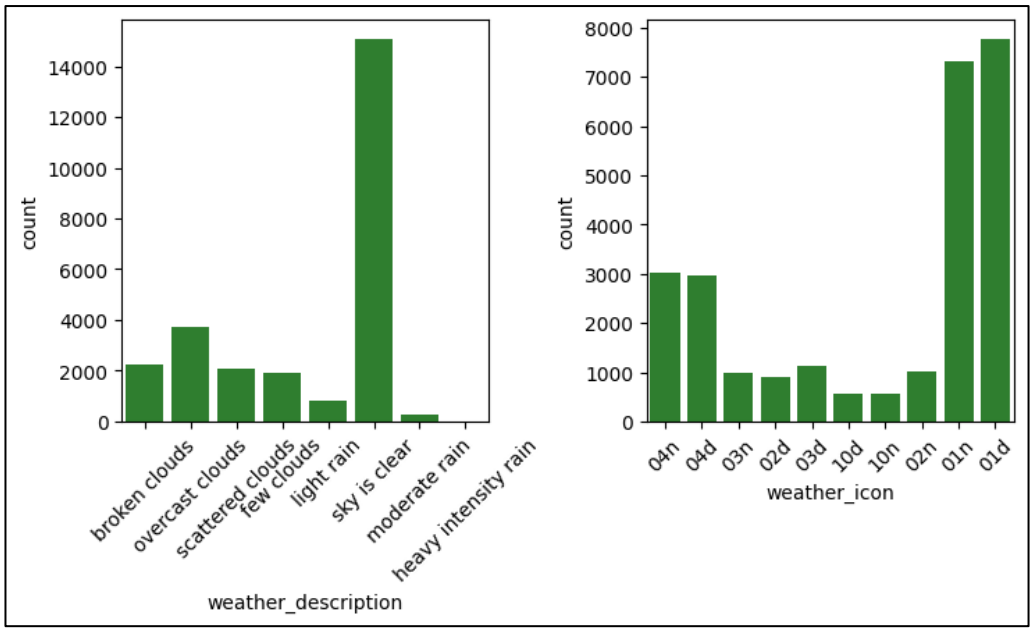


Figure 12. Distribution of the categorical variables

In our final dataset, we have chosen to retain twelve numerical features and two categorical features in addition to the response variables. The final dataset consists of 26,185 rows and 16 columns. The decision to remain at the feature selection stage was influenced by two factors: the small number of features and the presence of feature selection within certain models. In this section, the regression models will be applied to the final dataset in order to generate predictions.

5.RESULTS AND VISUALIZATIONS

5.1. Linear Regression Modeling

The initial step involved constructing a basic linear regression model solely utilizing the scaled GHI values. The purpose of this model is to serve as a benchmark for evaluating the performance of future, more intricate models.

The simple linear regression model was evaluated through 5-fold cross-validation and achieved an overall R^2 value of 85.4%. However, these results are optimistically misleading because approximately half of the records in the dataset are within nighttime with zero power generation. This imbalanced training and testing set with roughly half of the target feature values being zero. To overcome this issue, we can eliminate nighttime hours (zero production hours), easily identified as described during the data exploration phase. The nighttime hours are between 10:00 PM to 5:00 AM inclusive.

After excluding nighttime hours, the simple linear regression model was reassessed through 5-fold cross-validation, and the model achieved an overall R2 value of 79.8%. This observation confirms our doubts that keeping zero production hours may result in optimistically misleading results. Then, the dataset was split into training and testing, allocating the last four months for testing while the earlier part of the data was for training. After excluding nighttime hours, the simple linear regression model was trained using only GHI data and achieved an RMSE of 2379 MW, an R^2 value of 74.1% on the testing set. Our scatter plot of the predicted values versus actual power generation on the testing set suggests that the simple linear regression model is not suitable for the power generation forecast problem (Figure 13). In addition, the testing error distribution of the testing set does not appear normally distributed, which supports the earlier observation that the simple linear regression model does not properly fit the data. Figure 14 compares actual and modeled solar power generation for a 7-day period within the testing set.

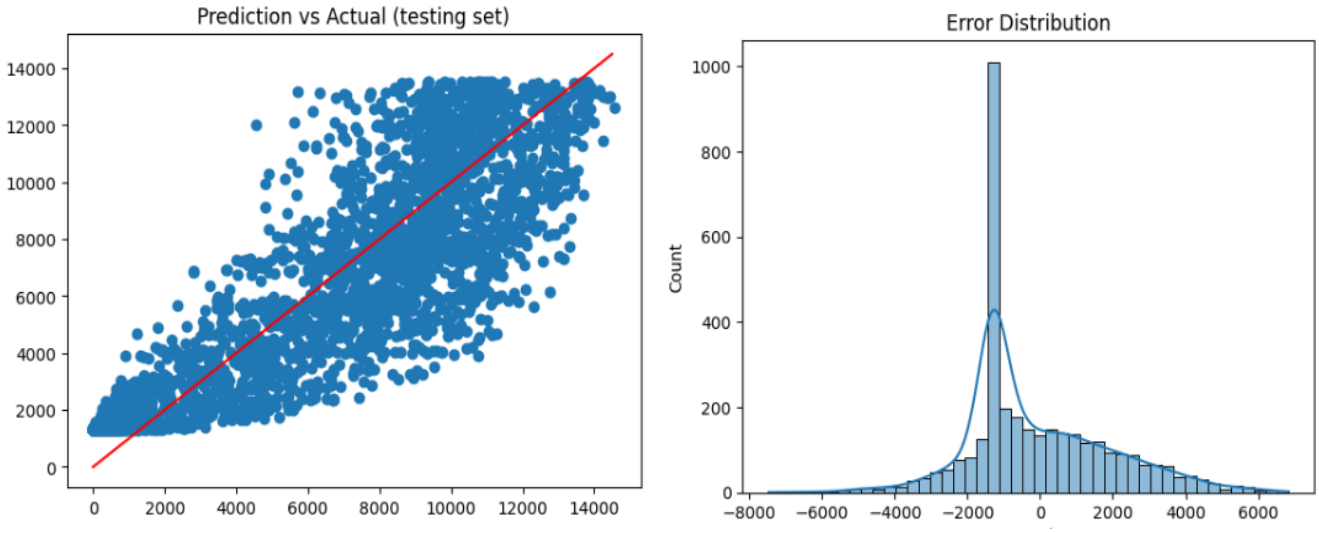


Figure 13. Prediction vs Actual; & error distribution for simple linear regression

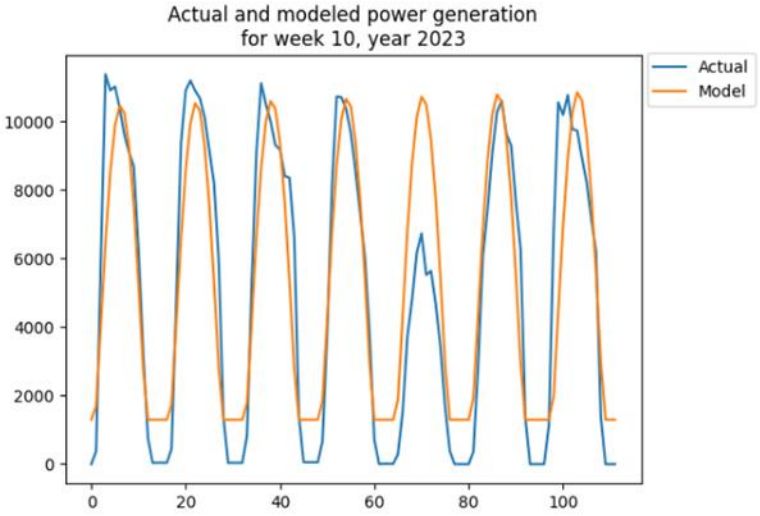


Figure 14. LR - Actual and model power generation in week 10 in 2023

In addition, we conducted an analysis on multiple linear regression models by utilizing the scaled numeric features of the entire dataset. The exclusion of categorical features was based on the understanding that their information is already represented in certain numeric features. Nighttime hours were excluded as previously mentioned in order to prevent any potential misleading outcomes. Multiple model fit iterations were implemented while excluding one feature at a time to recognize the significant features. The first model utilizing all numeric features achieved an RMSE of 1870 MW, and an R^2 value of 84.0%, while the final model using only GHI and DNI reached an RMSE of 1897 MW, an R^2 value of 83.6%. It suggests that both GHI and DNI are significant for

forecasting power generation. Our scatter plot of the predicted values versus actual power generation on the testing set suggests that the MLR model is not suitable for the power generation forecast problem yet (Figure 15). In addition, the testing error distribution of the testing set does not appear normally distributed, which supports the earlier observation that the MLR model does not fit the data properly. However, the model's predictive performance may be fair. Figure 16 shows a comparison between actual and modeled solar power generation for a 7-day period within the testing set. Since regression models do not appear suitable, random forest models will be explored next.

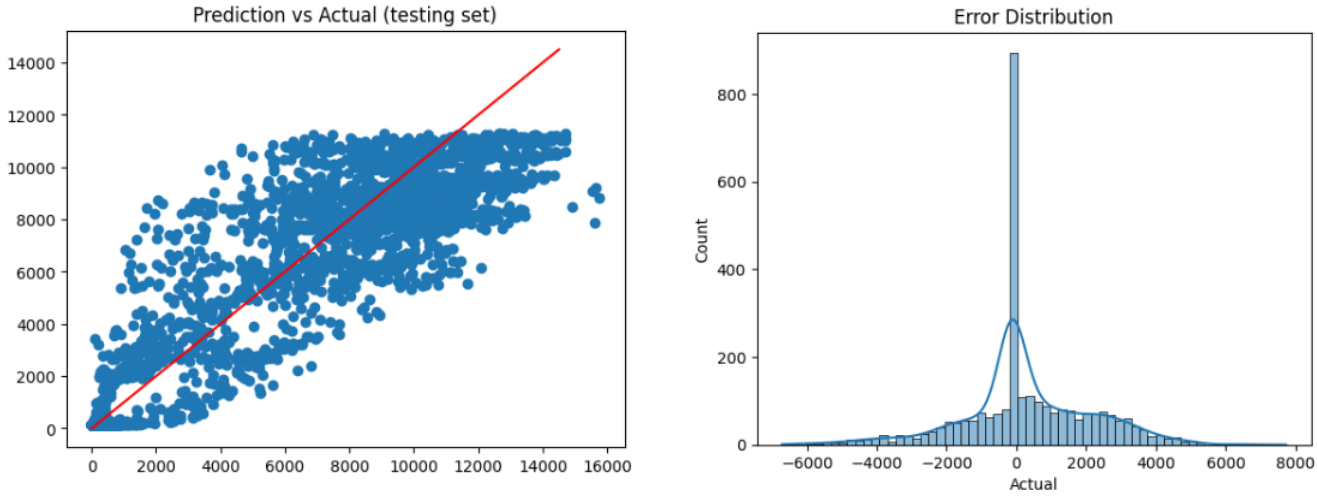


Figure 15. LR scaled - Prediction vs. Actual; & testing error distribution for MLR model

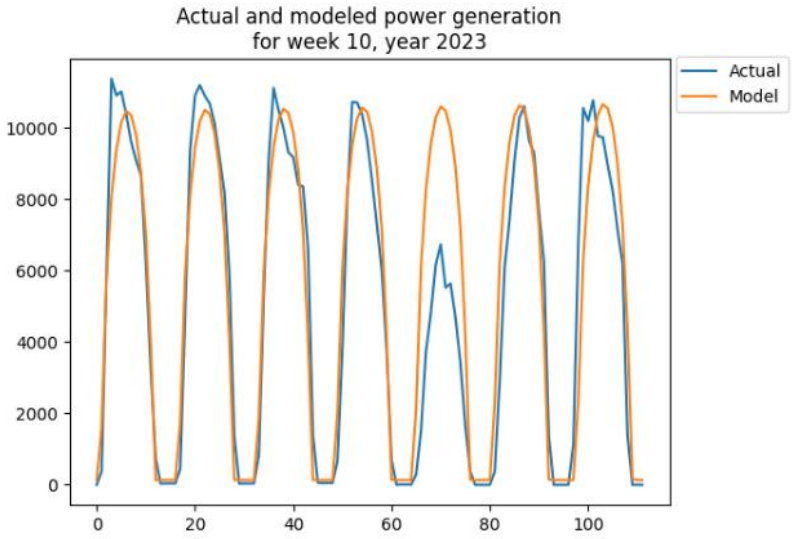


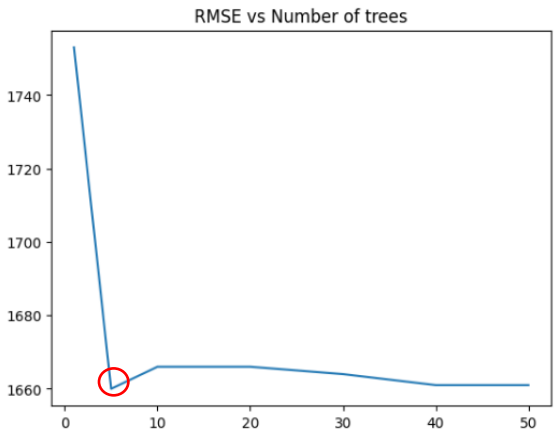
Figure 16. LR scaled - Actual and model power generation in a winter week in 2021 & 2022

5.2. Random Forest

The entire dataset was used to construct a random forest model, with the numeric features scaled beforehand. The exclusion of categorical features was based on the understanding that their information is already represented in certain numeric features. The initial model was constructed using 100 trees and incorporating the scaled numeric features. In order to identify the significant features, a series of models were created and evaluated using R2 value and RMSE. These models were fitted with multiple iterations, with each iteration excluding one feature at a time. The assessment revealed significant features of GHI, DNI, cloud cover, humidity, pressure, the week number, and the hour of the day. Next, the number of trees for the random forest model was optimized based on RMSE and R2 values, as shown in Figure 17 below. The optimal number of trees was found to be five, with an overall RMSE of 1666 MW, and an R^2 value of 87.3% on the testing set.

Similarly, the optimal number of samples per leaf was determined to be eleven. Although the model's predictive performance is relatively good, our scatter plot of the predicted values versus actual power generation on the testing set suggests that random forest may not be a suitable model for the power generation forecast problem (Figure 18). In addition, the testing error distribution of the testing set does not appear customarily distributed, which supports the earlier observation that the random forest model does not have an acceptable fit to the data. Figure 19 compares actual and modeled solar power generation for a 7-day period within the testing set. The model appears to be consistently underestimating peak hours' energy output.

For this reason, we explored a different approach by creating two random forest models: one for peak hours and another for off-peak hours. However, this approach resulted in an RMSE increase to 2492 and an R^2 value decrease to 71.6%, suggesting that the aggregating model approach may not be optimal. For this reason, time series modeling will be explored afterward.



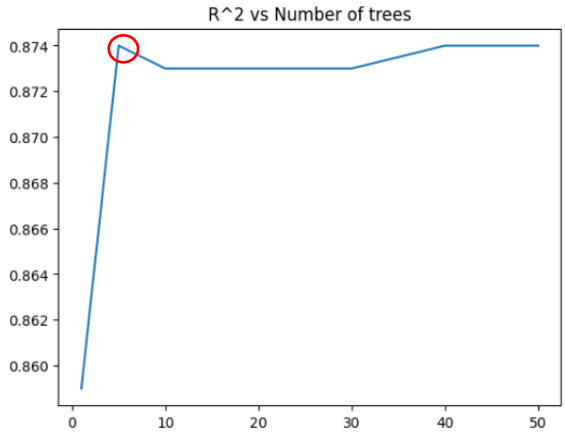


Figure 17. RF - RMSE & R2 vs Number of Trees

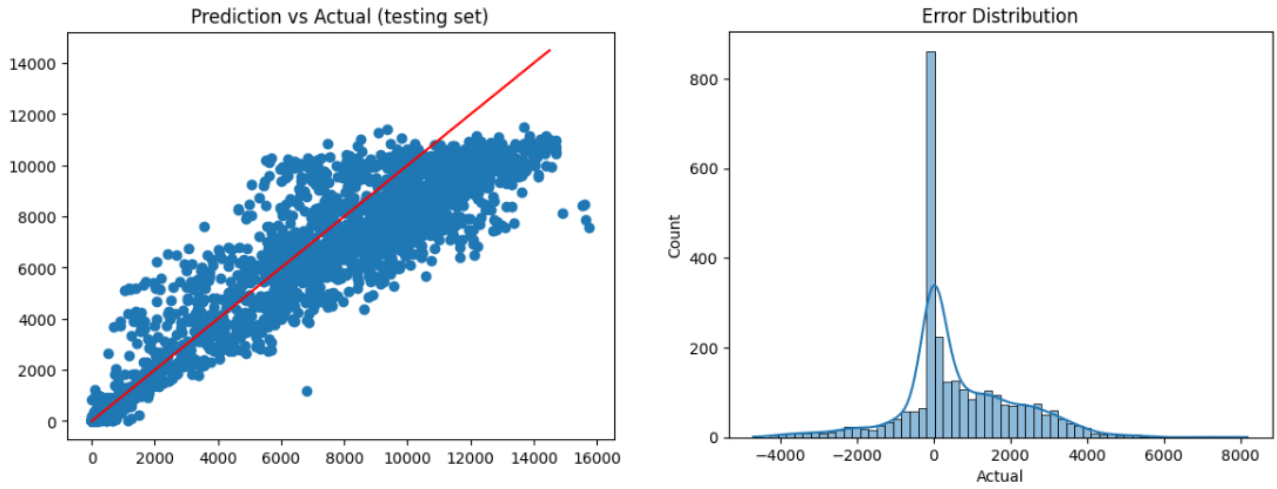


Figure 18. RF - Prediction vs Actual; & error distribution for random forest model

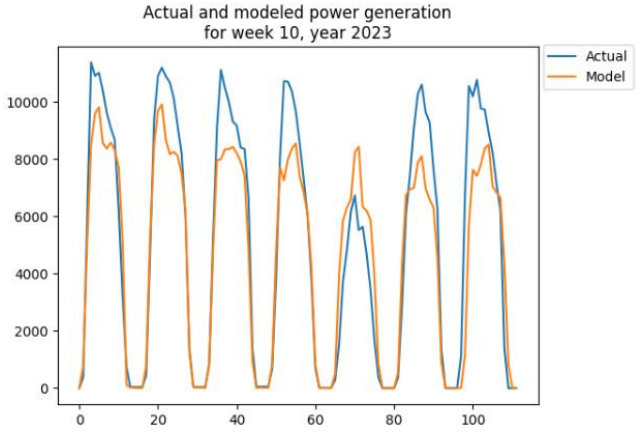


Figure 19. RF - Actual and model power generation in a summer week in 2021 & 2022

5.3. Time Series Modeling

Time series models were explored on the actual solar power generation without excluding nighttime. The auto_arima function was utilized to optimize the parameters p, d, and q, comprising the order of the model. The function auto_arima suggested that the optimal order of Arima is (1,1,2). However, using this model in order to forecast one week revealed that it was poor, as shown in Figure 20. Upon exploration, we noticed that only the autoregressive part of the ARIMA model was the significant parameter, meaning that the ARIMA model simplifies to the autoregressive model. To identify the optimal number of lags in the autoregressive model, we plotted the partial autocorrelation in the whole dataset, as shown in Figure 20, and determined the optimal number of lags to be 50 hours. The autoregressive model was trained on 720 hours to estimate the parameters associated with each of the 50 lags to forecast the future 72 hours (3 days). The autoregressive model achieved an overall RMSE of 1900 MW, with a relatively low R^2 value of 76.1%. However, upon inspection, we noticed that the model forecast tends to lag ahead of the actual output, as shown in Figure 21.

For this reason, we manually decided to shift the forecast one hour earlier. This simple trick resulted in outstanding improvement of the autoregressive model, leading RMSE to decline to 1055 MW and R2 increase to 89.5%. Figure 22 shows the same forecast period used in Figure 11 after applying the shifting trick. In addition, the testing error distribution for the model appears approximately normally distributed, as shown in Figure 23. Consequently, autoregressive models may be suitable for the power generation forecast problem.

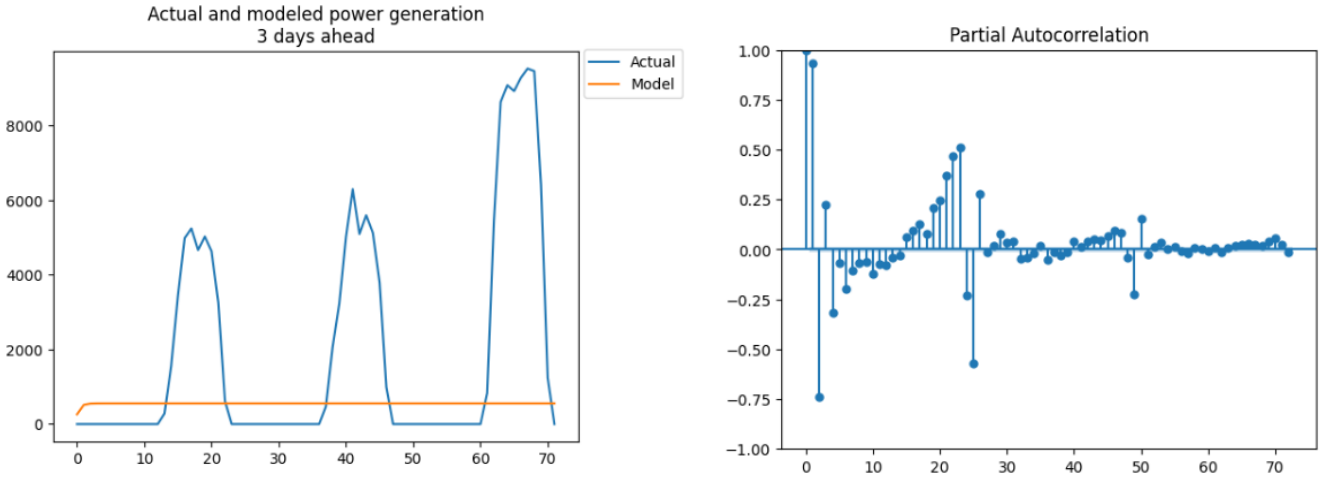


Figure 20. ARIMA - Actual and model power and Autoregressive partial autocorrelation

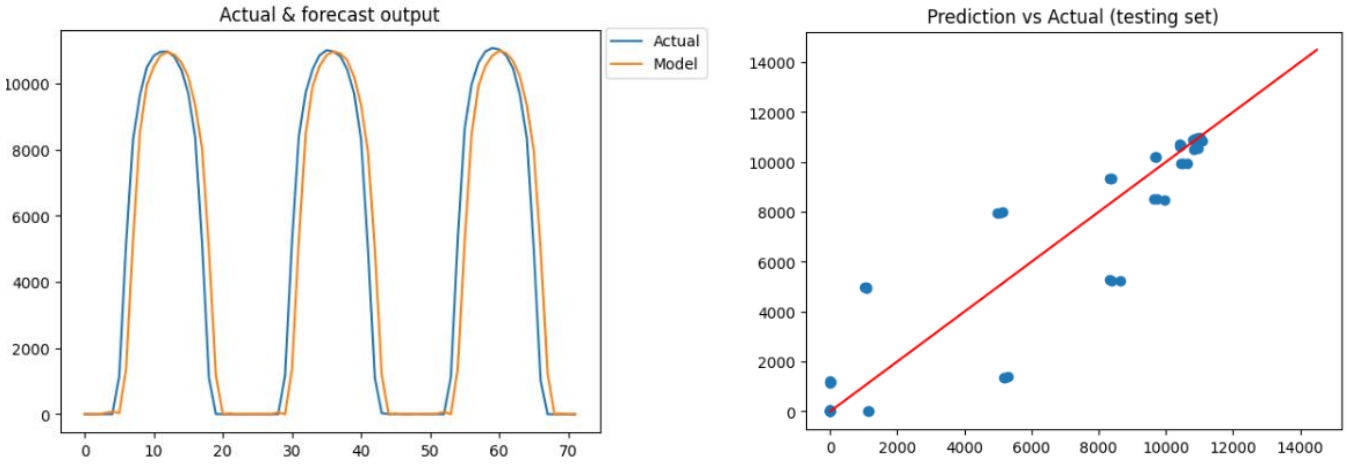


Figure 21. Autoregressive - Autoregressive model forecast over three days.

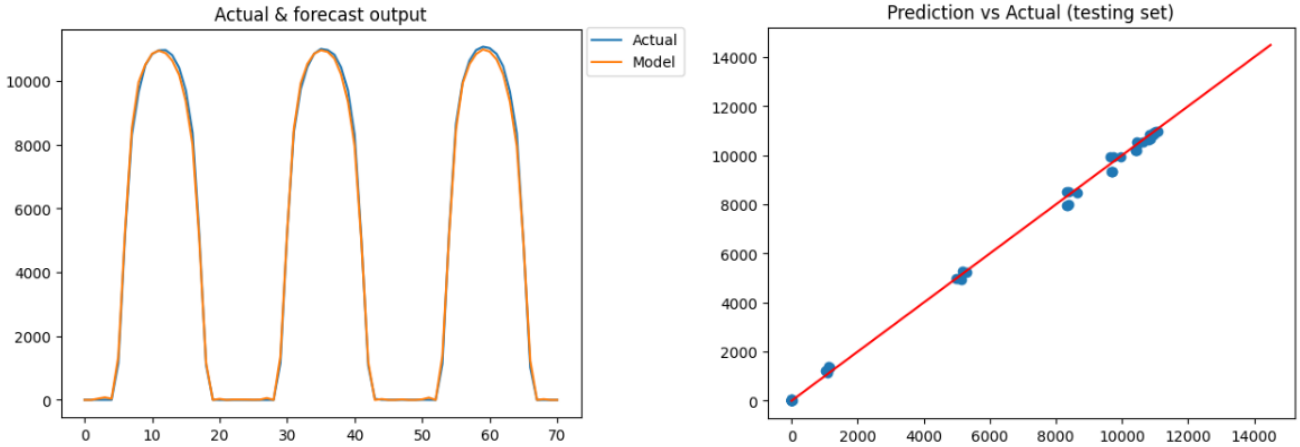


Figure 22. Autoregressive - Autoregressive model forecast over three days w/ shifting trick.

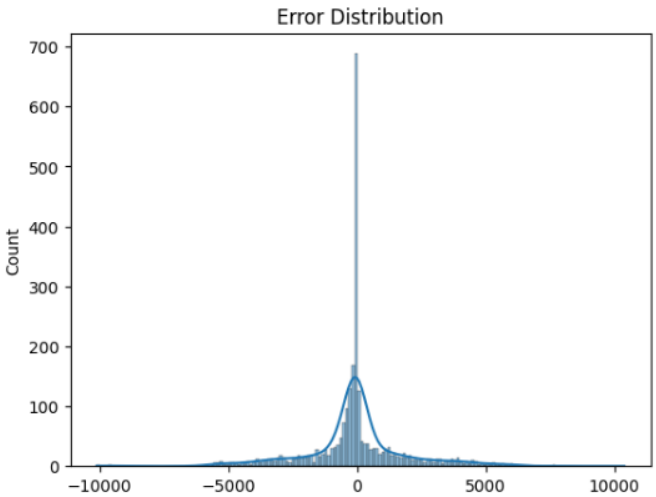


Figure 23. Autoregressive - Testing error distribution for shifted autoregressive model.

5.4. XGBoost

For our last model, we are using XGBoost because it is well known for its ability to handle complicated data sets and deliver powerful results. Its widespread use is due to the sophisticated ensemble learning approach, which combines the predictions of multiple separate models or weak learners to produce a robust final model (Srinivas & Katarya, 2022). XGBoost effectively handles various feature types thanks to parallel processing and tree-based learning techniques, making it suitable for a wide range of fields. Also, the regularization algorithms used by XGBoost ensure reliable predictions by preventing overfitting and improving generalization (Srinivas & Katarya, 2022).

Firstly, we ran the model without any optimization on the split final dataset. In Figure 24, we have the feature importance displayed. As expected, 'ghi' and 'day' are the top ones due to their strong correlation. The remaining features have some level of significance, although their importance may vary following the optimization of the hyperparameters.

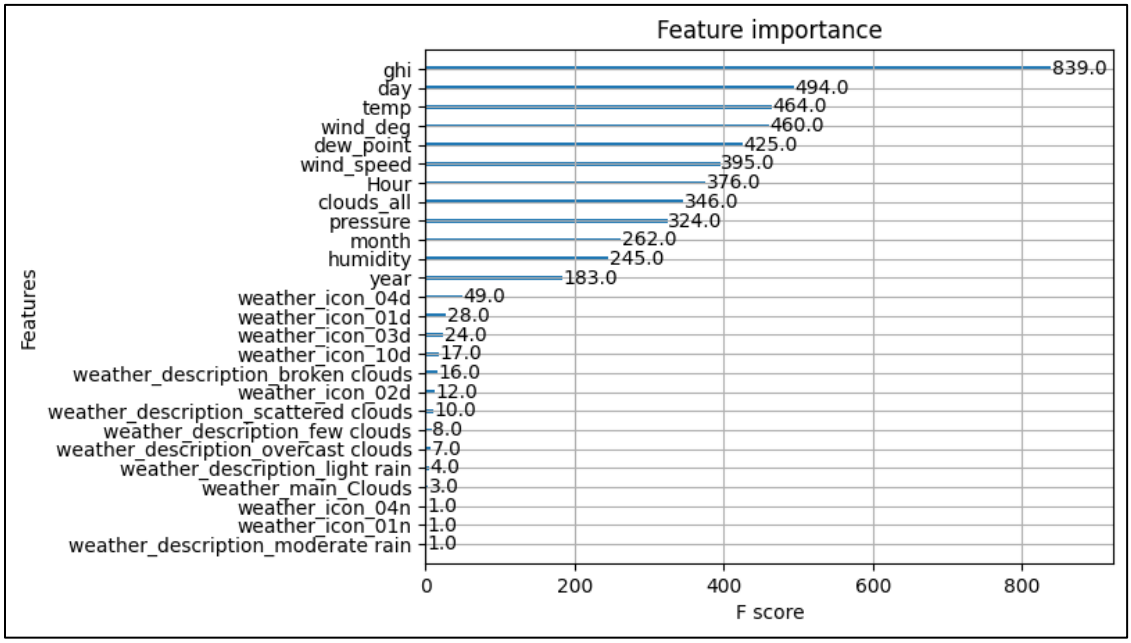


Figure 24. XGB - Feature importance on non-optimized model

However, no model promises perfect results under all circumstances. Proper evaluation, tuning, and handling potential problems such as overfitting are critical. The handling of class imbalances and hyperparameter tuning influences model performance. To effectively search the hyperparameter space and determine the ideal configuration for XGBoost, Optuna uses state-of-the-art optimization methods, including Bayesian

optimization and tree-structured Parzen estimators (Akiba, Sano, Yanase, Ohta, & Koyama, 2019; Srinivas & Katarya, 2022). It effectively searches for hyperparameters that lead to improved performance by judiciously balancing exploration and exploitation.

We can automate the hyperparameter tuning process for XGBoost using Optuna's capabilities, which will save a lot of time and effort. Optuna intelligently samples hyperparameter combinations and evaluates them using specified objective functions or evaluation metrics instead of manually tweaking hyperparameters and evaluating model performance. Because Optuna provides a simple API to specify the search range for hyperparameters and an interface to communicate with the XGBoost training process, its integration with XGBoost is smooth (Akiba, Sano, Yanase, Ohta, & Koyama, 2019). The fact that it supports both discrete and continuous hyperparameters makes it possible to perform a thorough search over a wide range of values.

In addition, Optuna provides features such as early pausing, trimming, and parallel execution that further speed up and increase the effectiveness of the hyperparameter optimization process. It intelligently terminates unproductive experiments early and efficiently distributes computing resources. As shown in Figure 25, Optuna quickly found the best objective value for our model.

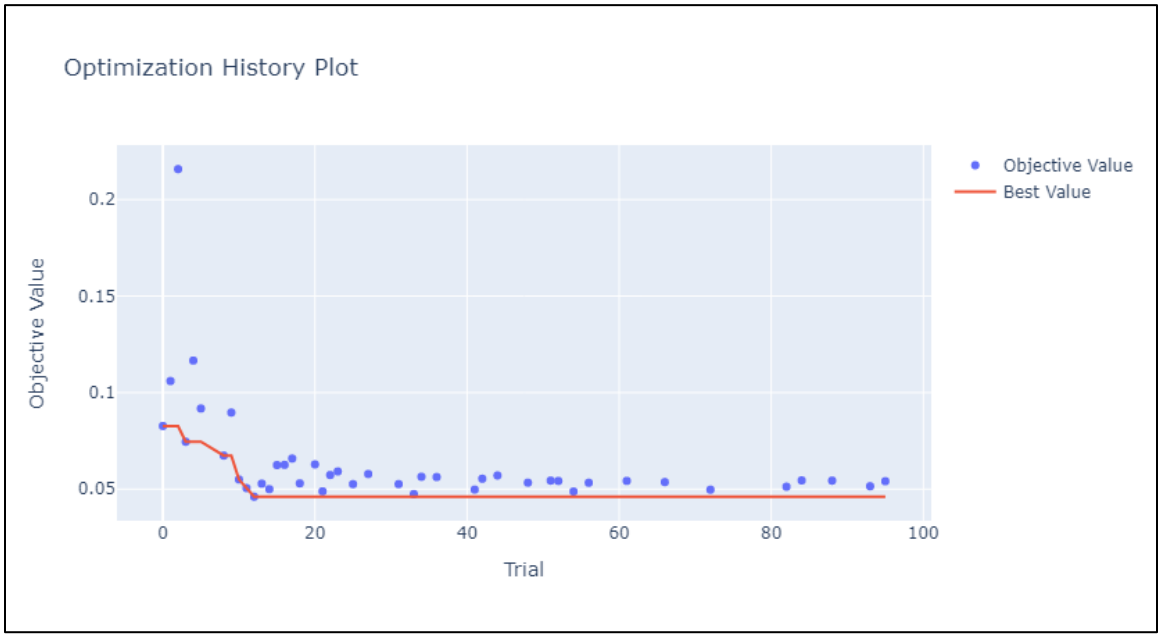


Figure 25. XGB Optuna - Optimization History Plot

The Parallel Coordinate plot offered by Optuna assists in the interpretation and assessment of the outcomes of hyperparameter tuning (Akiba, Sano, Yanase, Ohta, & Koyama, 2019). The tool offers a straightforward and concise viewpoint, facilitating the identification of effective hyperparameter configurations. It achieves this by representing each trial as a line and visually representing the hyperparameters on multiple axes for our model. Indeed, Figure 26 clearly shows a repetitive pattern that would work best for our dataset.

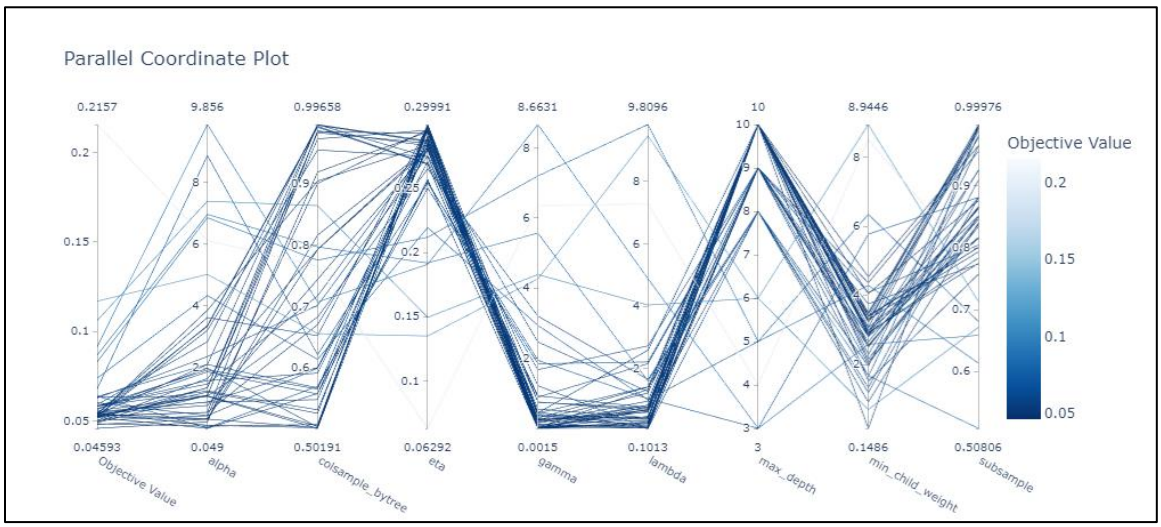


Figure 26. XGB Optuna - Parallel Coordinate Plot

As shown in Figure 27, the final hyperparameters would highly rely on the eta parameter with a significance of 67%. The shrinkage applied to each tree's contribution to the ensemble is controlled by the eta parameter, which impacts the model's convergence rate and overall performance (Akiba, Sano, Yanase, Ohta, & Koyama, 2019).

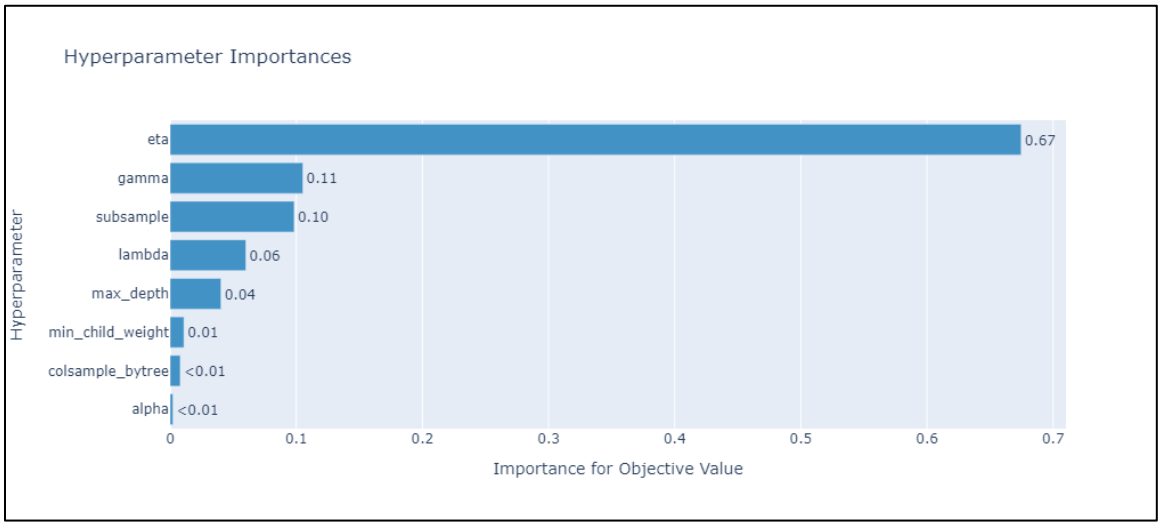


Figure 27. XGB Optuna - Hyperparameter Importance

After using Optuna for our hyperparameter tuning, we can observe a shift in the feature importance (Figure 28). 'ghi' and 'day' remain high, but 'month' finally increased, which would be logical as this is also a timestamp and indicator of the season too for energy generation. Indeed, Figure 28 illustrates a better importance of the features when modeling for predicting energy generation.

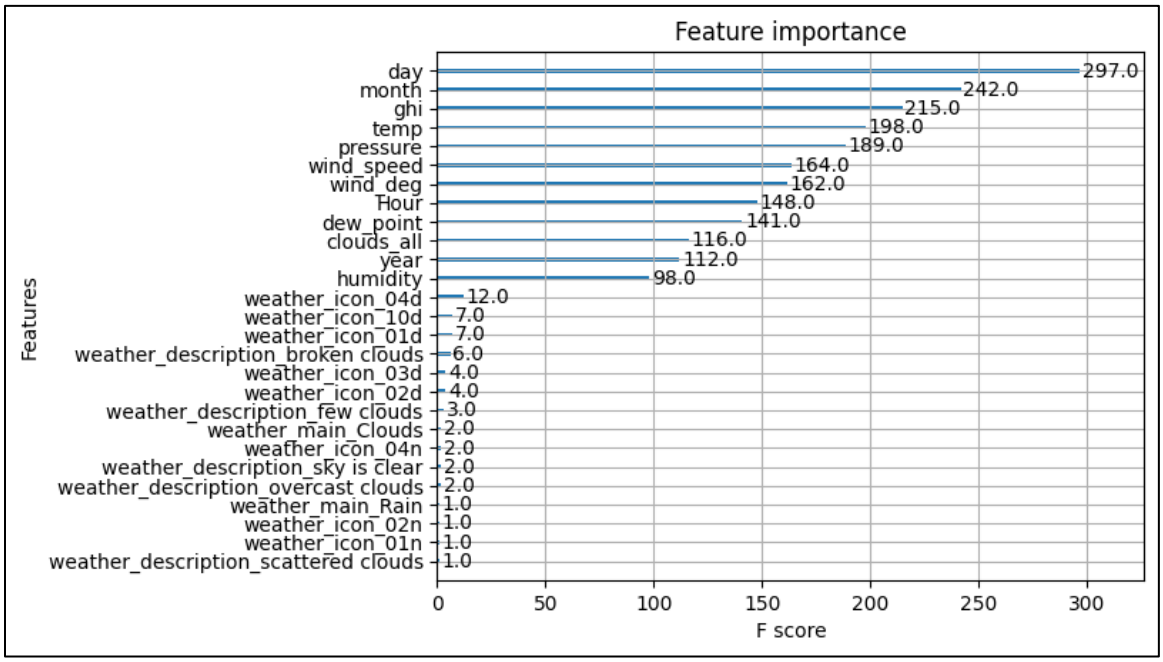


Figure 28. XGB Optimized - Feature importance on optimized model

Then, we plotted our predictions against the actuals in Figure 29. We can observe that our model is pretty conservative compared to what happened.

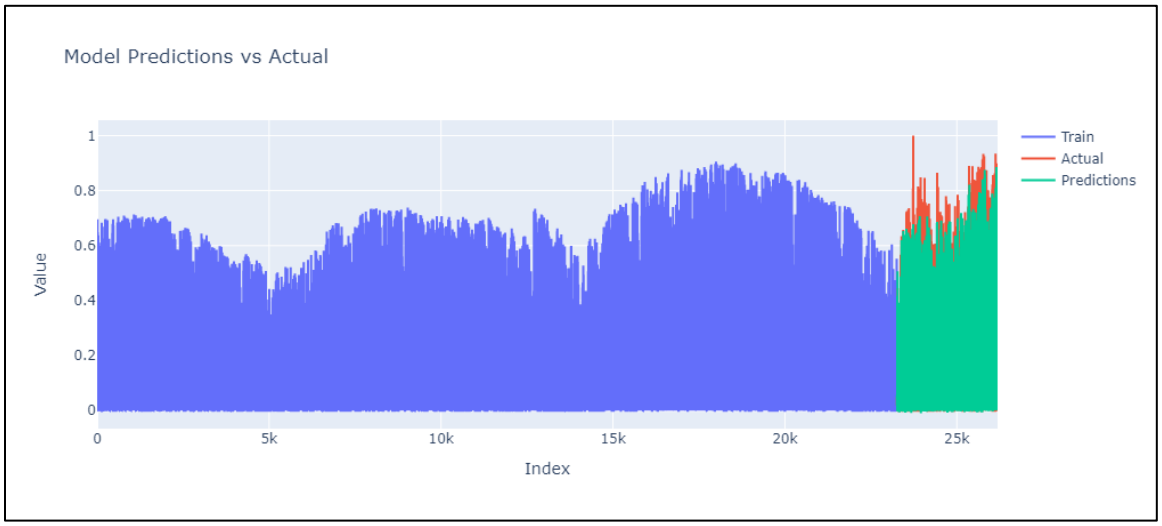


Figure 29. XGB Optimized - Model Predictions vs Actual

Now, let's focus solely on the predictions in order to gain a better understanding of what transpired. Figure 30 illustrates the data from the past four months, during which we conducted tests to evaluate the accuracy of our predictions. The visibility of predictions can be enhanced by partially reducing the prominence of the actual data. data. However, it is worth noting that there are some deviations between the model's predictions and the real-world data.

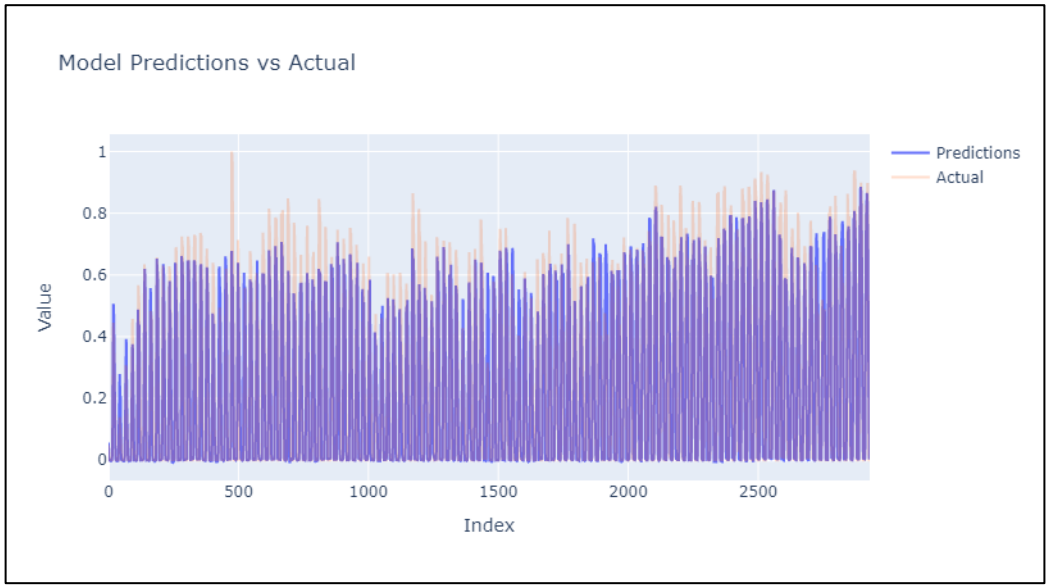


Figure 30. XGB Optimized - Model Predictions vs. Actual focus on Predictions period

Based on Figure 31, we observed that our predictions against the actual still clearly form a line trend which would indicate the validity of this model.

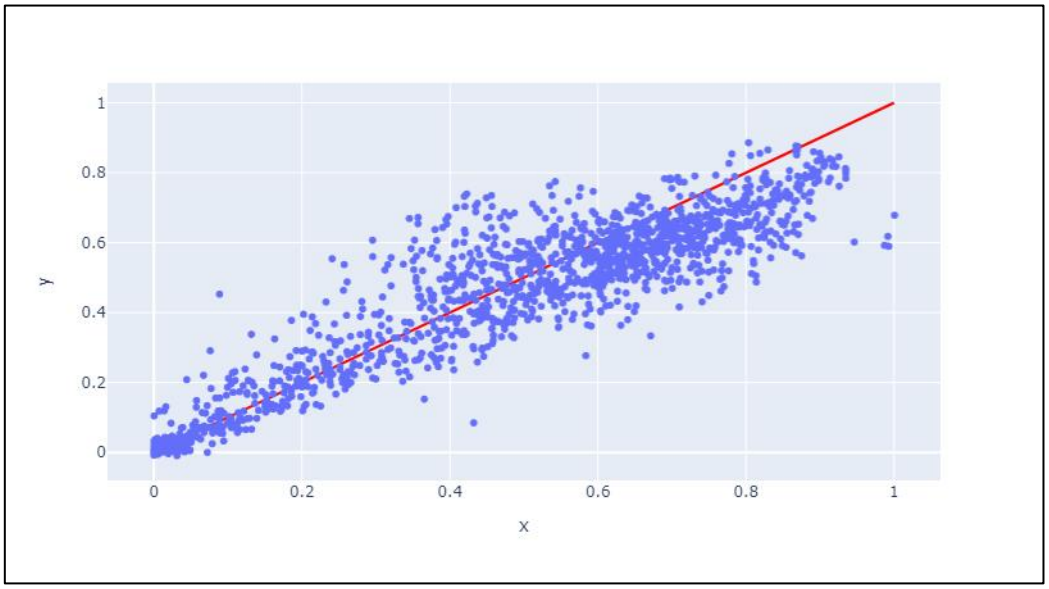


Figure 31. XGB Optimized - Model Predictions vs Actual

6.CONCLUSION

6.1. Results

Our research project was conducted in partnership with Enverus and aimed to improve the prediction of solar energy generation's performance by combining location data and a variety of modeling methodologies. The goal of the research was to improve operational efficiency and decision-making within the energy industry. The study aims to provide insight into accurate forecasts and viable regions for solar power generation by analyzing and contrasting multiple models.

Our literature review emphasized the need to consider geographic location, as solar radiation varies from place to place. Variables such as latitude, longitude, and weather patterns are essential in energy production. Indeed, the performance of solar generation is also affected by environmental variables such as temperature, wind speed, dust, shade, and humidity. Previous research has emphasized the use of ensemble models and robust modeling techniques to capture complicated interactions.

Based on Table 1 below, the modeling and data analysis results indicated significant differences in the effectiveness of the models. At the top, the shifted Autoregressive model was the most accurate, with an accuracy of 0.8722501. The fact that it outperformed all other models in terms of accuracy suggests that it can be used to predict power generation performance.

MODEL	ACCURACY	N. RMSE	R2	MAE	RMSE_SCORE	R2_SCORE	SCORE	FINAL_RANK
SHIFTED AUTOREGRESSIVE	0.8722501	0.0669808	0.8951669	671.82246	1	2	1.2	1
XGBOOST	0.83003413	0.0787941	0.9501566	698.32203	2	1	1.8	2
RANDOM FOREST	0.70102389	0.1054079	0.8739532	1105.0111	3	3	3	3
AUTOREGRESSIVE	0.726	0.1206196	0.761	1189	4	5	4.2	4
LINEAR REGRESSION (GHI, DNI)	0.65494881	0.1207227	0.8346655	1330.7575	5	4	4.8	5
LINEAR REGRESSION (GHI)	0.60511945	0.150738	0.7422309	1919.8666	6	6	6	6
TWO RANDOM FORESTS	0.315	0.1582021	0.716	2195	7	7	7	7
ARIMA	0	1	0	9999	8	8	8	8

Table 1. Models' results summary

For the rest of our models, The XGBoost and Random Forest models also showed potential for making accurate predictions.

These findings provide Enverus and energy companies with insightful information that helps them improve operations, manage resources wisely, and identify locations with the most significant potential for solar power generation. It is important to note that these results are unique to this study's data set and experimental design.

6.2. Discussion

Furthermore, regarding the results of our models, we could say that the shifted autoregressive model is a bit too holistic as it takes into consideration only the response variable. In comparison, XGBoost and the rest of our models were based on the indicators provided by pvlib and OpenWeather, which would make them more adaptable and closer to reality.

Future studies should investigate how these models perform and whether they are generalizable to other scenarios and data sets. Future studies could focus on adding more characteristics, such as solar cell specifications and other environmental variables, to increase the accuracy and predictive ability of the models.

This research study results have paved the way for improved operational efficiency and decision-making in the energy sector by highlighting the importance of site-specific data and advanced modeling techniques in evaluating the energy generation performance of solar farms.

7.REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631. doi:https://doi.org/10.1145/3292500.3330701

Aksoy, N., & Genc, I. (2023). Predictive models development using gradient boosting based methods for solar power plants. *Journal of Computational Science, 67*(101958), 1-10. doi:https://doi.org/10.1016/j.jocs.2023.101958

Botchkarev, A. (2019). Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology. *Interdisciplinary Journal of Information, Knowledge, and Management*, 045-076. doi:https://doi.org/10.48550/arXiv.1809.03006

California ISO. (2023). *California ISO Open Access Same-time Information System.* Retrieved from California ISO: http://oasis.caiso.com/mrioasis/logon.do?reason=application.baseAction.noSession

Chikate, B. V., & Sadawarte, Y. (2015). The factors affecting the performance of solar cell. *International journal of computer applications, 1*(1), 1-5. doi:0975 – 8887

Enverus. (2023). *Intelligent Connections*. Retrieved from Enverus: https://www.enverus.com

Hobbs, B. F., Zhang, J., Hamann, H. F., Siebenschuh, C., Zhang, R., Li, B., . . . Zhang, S. (2022). ISO, Using probabilistic solar power forecasts to inform flexible ramp product procurement for the California. *Solar Energy Advances, 2*(100024), 1-11. doi:https://doi.org/10.1016/j.seja.2022.100024

HOMER Pro. (2023). *Global Horizontal Irradiance (GHI)*. Retrieved from HOMER Pro: https://www.homerenergy.com/products/pro/docs/3.11/global_horizontal_irradiance_ghi.html

Miao, S., Ning, G., Gu, Y., Yan, J., & Ma, B. (2018). Markov Chain model for solar farm generation and its application to generation performance evaluation. *Journal of Cleaner Production, 186*(1), 905-917. doi:https://doi.org/10.1016/j.jclepro.2018.03.173

OpenWeather. (2023). *Weather API.* Retrieved from OpenWeather: https://openweathermap.org/api

Rebala, G., Ravi, A., & Churiwala, S. (2019). *Machine Learning Definition and Basics. In: An Introduction to Machine Learning.* Worldwide: Springer, Cham. doi:https://doi.org/10.1007/978-3-030-15729-6_1

Sandia National Laboratories & pvlib python Development Team. (2023). *User Guide*. Retrieved from pvlib: https://pvlib-python.readthedocs.io/en/stable/user_guide/index.html

Singh, A. K., & Singh, R. R. (2021). An overview of factors influencing solar power efficiency and strategies for enhancing. *Innovations in Power and Advanced Computing Technologies (i-PACT), 1*(1), 1-6. doi:10.1109/i-PACT52855.2021.9696845

Srinivas, P., & Katarya, R. (2022). hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost. *Biomedical Signal Processing and Control, 73*(1), 103456. doi:https://doi.org/10.1016/j.bspc.2021.103456

Thomas, T., & Rajabi, E. (2021). A systematic review of machine leaning-based missing value imputation techniques. *Data Technologies and Applications, 55*(4), 558-585. doi:https://doi.org/10.1108/DTA-12-2020-0298

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing, 415*(1), 295-316. doi:https://doi.org/10.1016/j.neucom.2020.07.061

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the Effect of Accuracy on Trust in Machine Learning Models. *Association for Computing Machinery*, 1–12. doi:https://doi.org/10.1145/3290605.3300509

8.APPENDIX A

Data provided:

CAISO Total System Solar Power		
Analysis Period:		
Date	Hour	Actual
16-05-20	1	0
16-05-20	2	0
16-05-20	3	0
16-05-20	4	0
16-05-20	5	0
16-05-20	6	4
16-05-20	7	1535
16-05-20	8	5772
16-05-20	9	8539
16-05-20	10	9490

OpenWeather data:

DateTime	16-05-20 1:00	DateTime: Date and time of the observation
Date	16-05-20	Date: Date of the observation
Hour	1	Hour: Hour of the observation
Actual	0	Actual: Actual value (unspecified in the extract)
city_name	Dos Palos Y	city_name: Name of the city (Dos Palos Y)
lat	37.051548	lat: Latitude coordinate of the location (37.051548)
lon	-120.699371	lon: Longitude coordinate of the location (-120.699371)
temp	25.61	temp: Temperature
dew_point	9.85	dew_point: Dew point temperature
feels_like	25.19	feels_like: Perceived temperature
temp_min	24.99	temp_min: Minimum temperature
temp_max	28.18	temp_max: Maximum temperature
pressure	1014	pressure: Atmospheric pressure
humidity	37	humidity: Humidity level
wind_speed	4.6	wind_speed: Wind speed
wind_deg	321	wind_deg: Wind direction in degrees
rain_1h		rain_1h: Precipitation in the last hour (unspecified in the extract)
clouds_all	73	clouds_all: Cloud coverage in percentage
weather_id	803	weather_id: Weather condition ID
weather_main	Clouds	weather_main: Main weather category (e.g., Clouds)
weather_description	broken clouds	weather_description: Description of the weather condition (e.g., broken clouds)
weather_icon	04d	weather_icon: Icon representing the weather condition
ghi	0	ghi: Global horizontal irradiance (unspecified in the extract)
dni	0	dni: Direct normal irradiance (unspecified in the extract)
dhi	0	dhi: Diffuse horizontal irradiance (unspecified in the extract)
DateTime_DST	2020-05-16 01:00:00-07:00	DateTime_DST: Date and time of the observation adjusted for daylight saving time (DST)
Hour_DST	1	Hour_DST: Hour of the observation adjusted for daylight saving time (DST)
year	2020	year: Year of the observation
month	5	month: Month of the observation
week	20	week: Week number of the year
day	16	day: Day of the month