Applied Analytics Practicum - Enverus

# Introduction

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.

This project aims to investigate the methodologies utilized for predicting the performance of solar energy generation, with a specific focus on the geographical impact and modeling techniques employed. Our team will experiment on different models with distinct parameters to comprehensively compare and contrast various modeling scenarios. This project inspires us to combine multiple data sources and develop specific modeling techniques to explore the implications of predicting solar energy performance.

The problem statement at the core of this project revolves around comparing and contrasting different design options concerning utilizing location-specific source data and alternative modeling techniques. Enverus will contribute anonymized data sets encompassing input and target variables based on the data from Californio ISO toward the actual solar mega-watt generation (California ISO, 2023). The design matrix adopted for this study incorporates a variety of data sources, ranging from macro to meso and micro-regions in proximity to the area under investigation, as well as multiple time-series models, including Linear Regression, Random Forest, or XGBoost. The project's overarching objective is to be able to answer the following hypothesis questions.

Hypotheses:

* 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?

The findings are expected to gain valuable insights into the renewable energy industry and aid in developing more accurate and efficient modeling techniques. Ultimately, this project endeavors to facilitate informed decision-making in the planning and implementation of solar farms, promoting the adoption of sustainable energy solutions and furthering the advancement of the renewable energy sector.

# 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, in our data collection process, we don’t have access to photovoltahic cell specifications. 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).

# Methodology

For this section of the study, we have formed an extract, transform, and load (ETL) data pipeline for our data and processes in the graph below (Figure 1).

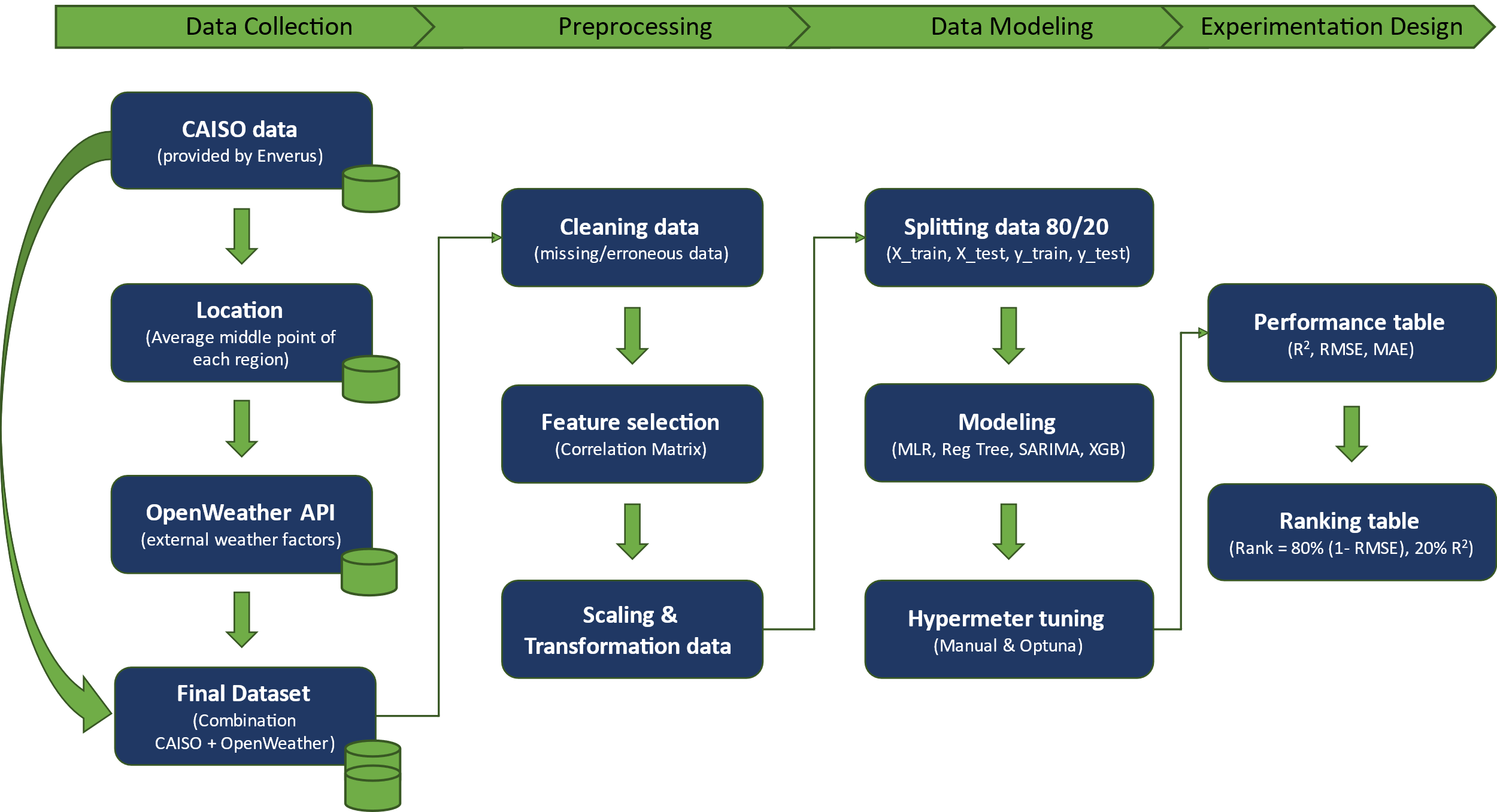


Figure 1. ETL Data Pipeline

## 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 the prediction of solar energy generation's performance.

Firstly, we will need to identify relevant data sources. As part of our collaboration with Enverus, we have been provided simplified datasets of timestamps and actual measure of mega-watts generated. 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 will have to determine a specific location to retrieve future data (such as latitude and longitude). In a second time, 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 data of the specific location from OpenWeather (OpenWeather, 2023).

Then, regarding the data collection, we will obtain the identified data from reliable sources, including Enverus proprietary data sets, publicly available data sets, academic research repositories, and industry reports. The data will be obtained in a structured format for efficient preprocessing and modeling.

Indeed, we will attempt to maintain a high level of data quality. 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. Any necessary data cleansing or corrections will be carried out to improve the overall quality of the dataset.

## 3.2 Preprocessing

Data preprocessing is a critical step in preparing the collected data for analysis and modelling for our study.

We will start with cleaning the data. Any missing, erroneous, or inconsistent data points identified during the data collection phase are 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 collected dataset to identify relevant features that contribute significantly to the prediction of solar farm performance. This step involves establishing a correlation matrix between variables, performing feature importance analysis, and using domain knowledge to select the most informative features for modelling.

Last but not least, we will proceed with feature scaling and transformation. Certain features may require scaling or transformation to ensure compatibility and optimal performance, as machine learning algorithm can interpret only numerical data. We will proceed with normalizing the numerical data and transforming the categorical data into dummy variable columns.

## 3.3 Data Modeling

The data modeling phase focuses on developing models to predict solar farm performance based on the collected and preprocessed data.

Based on the study's objectives and the literature review findings, we will select a range of modeling techniques to explore. These will include traditional statistical models like multi-linear regression and machine learning algorithms (e.g.,: random forests and extreme gradient boosting).

Then, we will focus on model development. We will develop and train the models using the preprocessed dataset for each selected modeling technique. This includes partitioning the dataset into training and test subsets, optimizing model parameters, and evaluating model performance using appropriate evaluation metrics (e.g., root mean squared error, R-squared).

Additionally, we will design an evaluation of the models. The developed models will be evaluated using cross-validation techniques to assess their generalization performance and robustness. We will analyze and compare the predictive capabilities of the models, considering factors such as accuracy, precision, recall, and computational efficiency.

## 3.4 Experimentation Design

The experimentation design phase focuses on setting up controlled experiments to compare and contrast different design options, taking into account geographical implications and modeling techniques. We will follow the processus from the below graph.

Firstly, we will need to start with the design matrix. We will construct a design matrix containing data combinations from California CAISO and OpenWeather. These data sources will include macro, meso and micro regions in the vicinity of solar location. Then, we will split the dataset into a train and test datasets for both the response and the dependent variables for our models, corresponding to 80% and 20% respectively.

Then we will proceed with modeling. Indeed, the model classes will consist of a variety of modeling techniques such as multi-linear regression, regression tree, Seasonal Autoregressive Integrated Moving Average (SARIMA), and eXtreme Gradient Boosting (XGBoost). Here, we will have an experimental set-up phase. The team will be testing different parameters within the models to optimize them.

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 relevant evaluation metrics. This analysis will provide insight into the impact of different data sources and modeling techniques on solar performance prediction.

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 the prediction of 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.

# Results and Visualizations

## 4.1 Linear Regression Modeling

The first model built was simple linear regression model using only the scaled GHI values. The purpose of this model is to be a reference model for assessing the performance of the upcoming more complex models. The dataset was randomly split into 80% training and 20% testing. The simple linear regression model achieved overall RMSE of 1744 MW, and R2 value of 85.2% on the testing set. However, these results are optimistically misleading because approximately half of the records in the dataset are within night time with zero power generation. This makes both the training and testing sets imbalanced with roughly half of the target feature values being zero. To overcome this issue, we can eliminate night time hours (zero production hours) which can be easily identified as described during data exploration phase. The night time hours are between 22:00 p.m. to 5:00 a.m. inclusive. After excluding night time hours, the simple linear regression model was fitted again. The model achieved overall RMSE of 1997 MW, and R2 value of 80.2% on the testing set, which confirms our doubts that keeping zero production hours may result in optimistically misleading results. A scatter plot of the predicted values versus actual power generation on the testing set suggests that simple linear regression model is not a suitable model for the power generation forecast problem (Figure-2). In addition, the error distribution of the testing set does not appears normally distributed which support the earlier observation that simple linear regression model does not fit to the data properly. Figure-3 & 4 show a comparison between actual and modeled solar power generation for a winter and a summer week in 2021 and 2022.

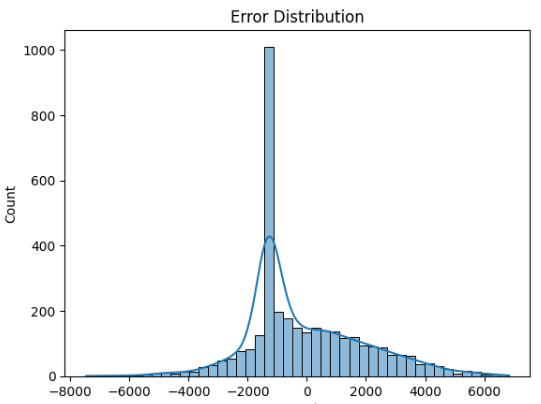
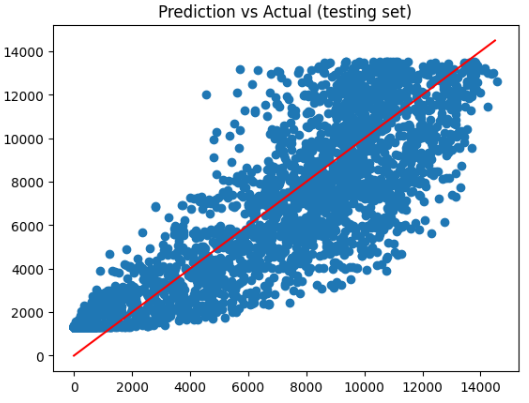


Figure-2. Prediction vs Actual; & error distribution for simple linear regression

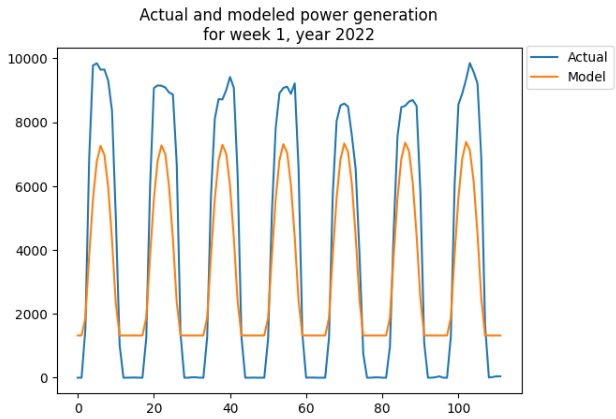
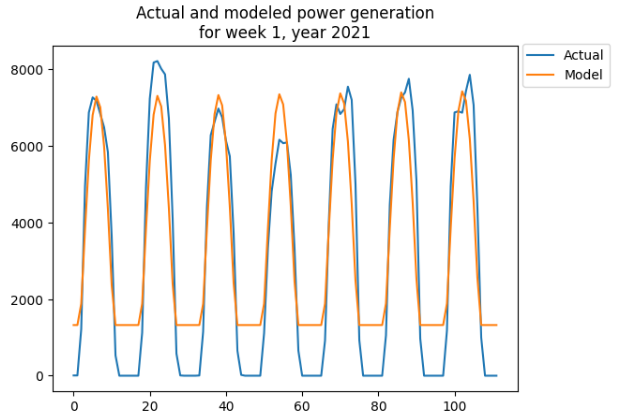


Figure-3. Actual and model power generation in a winter week in 2021 & 2022.

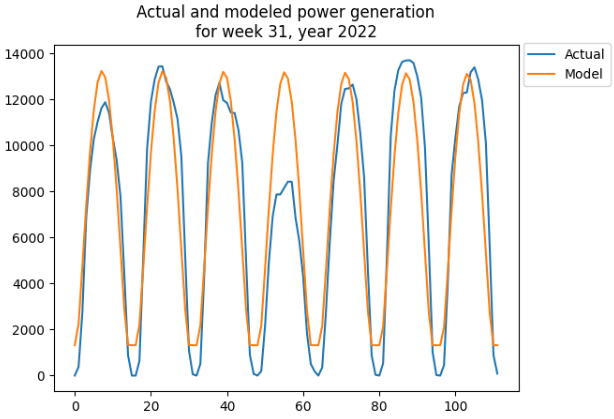
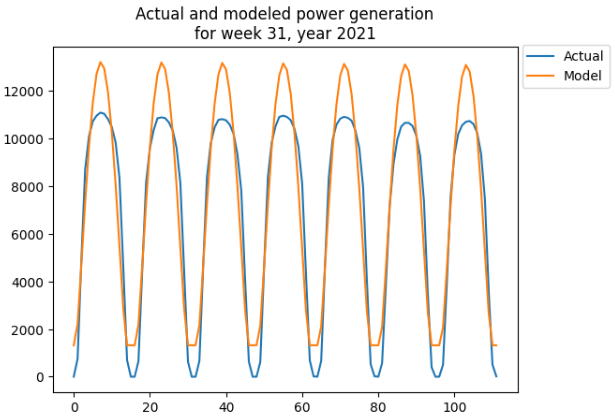


Figure-4. Actual and model power generation in a summer week in 2021 & 2022.

Next, we explored multiple linear regression model using the scaled numeric features of the whole dataset. Categorical features were excluded as their information is already captured in some of the numeric features. Also, we excluded night time hours as described earlier to avoid misleading results. 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 RMSE of 1492 MW, and R2 value of 89.2%, while the final model utilizing only GHI and DNI was able to achieve RMSE of 1606 MW, R2 value of 87.2%. This suggests that both GHI and DNI are significant for forecasting power generation. A scatter plot of the predicted values versus actual power generation on the testing set suggests that MLR model yet is not a suitable model for the power generation forecast problem (Figure-5). In addition, the error distribution of the testing set does not appears normally distributed which support the earlier observation that MLR model does not fit to the data properly, although the model’s predictive performance may be fair. Figure-6 & 7 show a comparison between actual and modeled solar power generation for a winter and a summer week in 2021 and 2022. Since regression models do not appear suitable, random forest models will be explored next.

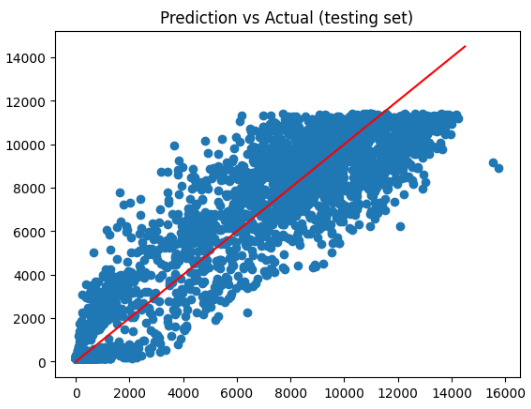
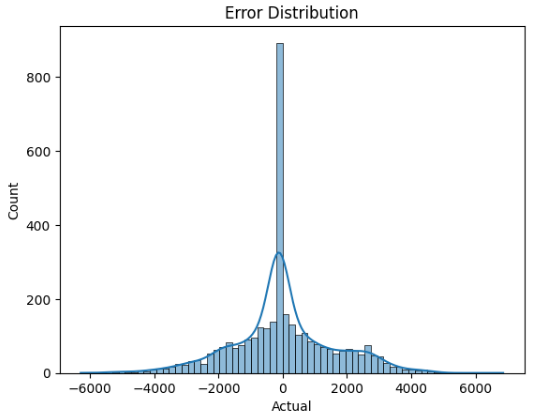


Figure-5. Prediction vs Actual; & error distribution for MLR model.

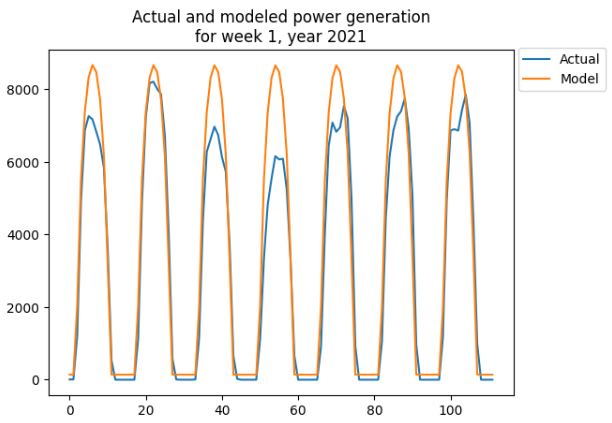
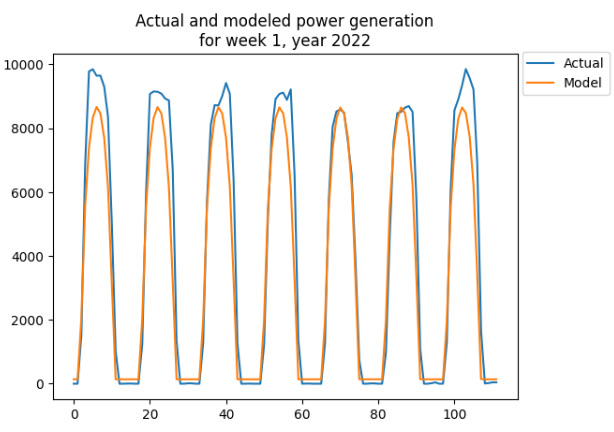


Figure-6. Actual and model power generation in a winter week in 2021 & 2022.

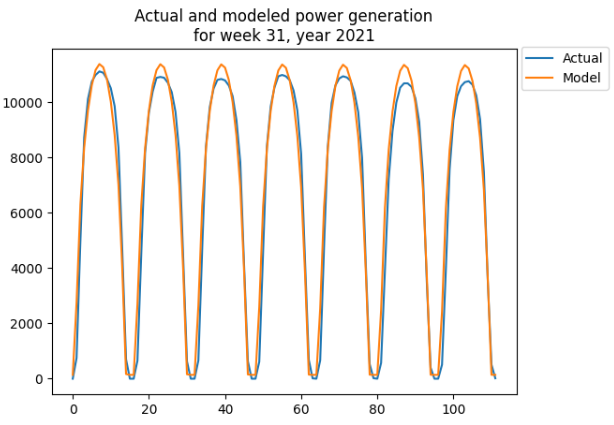
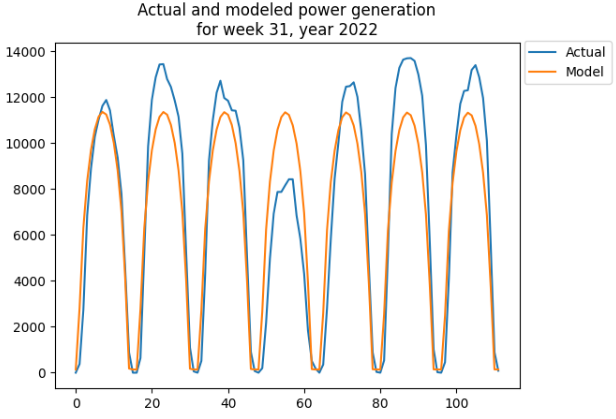


Figure-7. Actual and model power generation in a summer week in 2021 & 2022.

## 4.2 Random Forest

Random forest model was built using the scaled numeric features of the whole dataset. Categorical features were excluded as their information is already captured in some of the numeric features. The initial model was built with 100 trees and using the scaled numeric features. Next, multiple model fit iterations while excluding one feature at a time were implemented and assessed through R2 value and RMSE to recognize the significant features. The assessment revealed that GHI, DNI, cloud cover, humidity, pressure, and the week number are significant features. Next, the number of trees for the random forest model was optimized based on RMSE and R2 value as shown in Figure-8 below. The optimal number of trees was found to be 20, with overall RMSE of 1269 MW, and R2 value of 92.2% on the testing set. A scatter plot of the predicted values versus actual power generation on the testing set suggests that random forest model is a suitable model for the power generation forecast problem (Figure-9). In addition, the error distribution of the testing set appears approximately normally distributed which support the earlier observation that random forest model has an acceptable fit to the data. Figure-10 & 11 show a comparison between actual and modeled solar power generation for a winter and a summer week in 2021 and 2022, signifying the reasonable performance of the model.

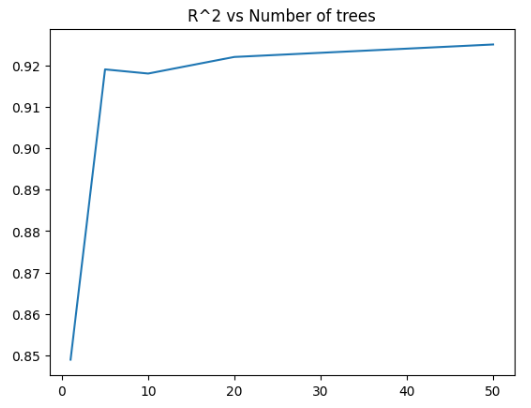
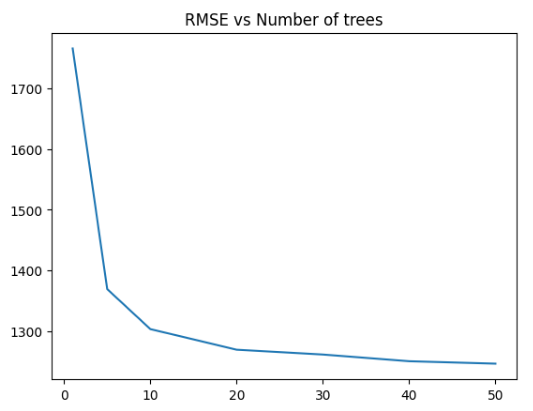


Figure-8. RMSE & R2 vs Number of Trees.

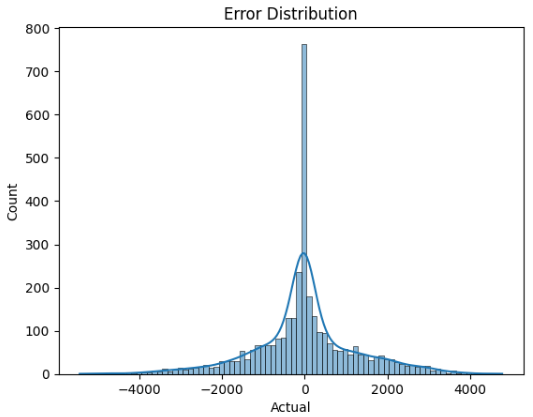
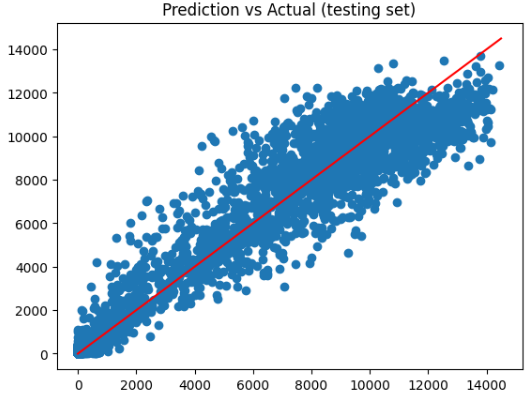


Figure-9. Prediction vs Actual; & error distribution for random forest model.

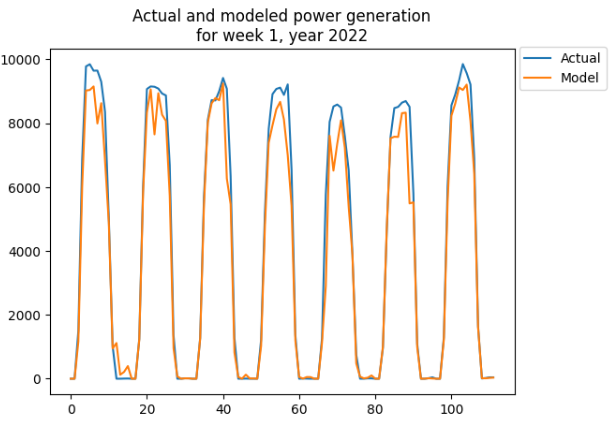
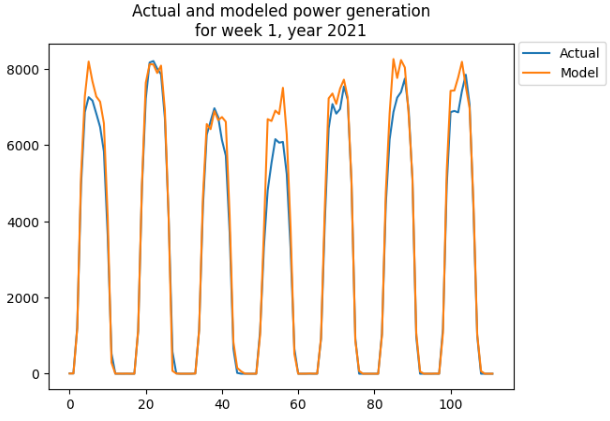


Figure-10. Actual and model power generation in a winter week in 2021 & 2022.

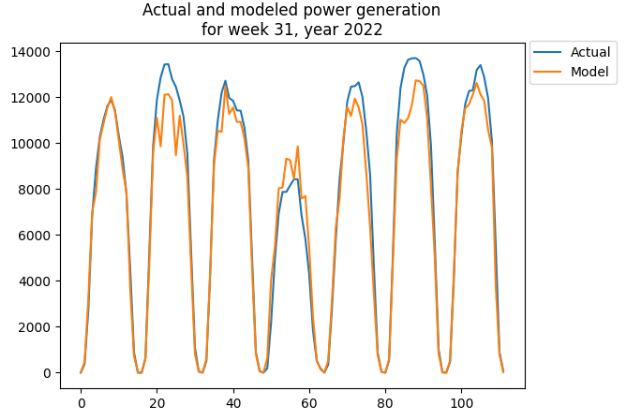
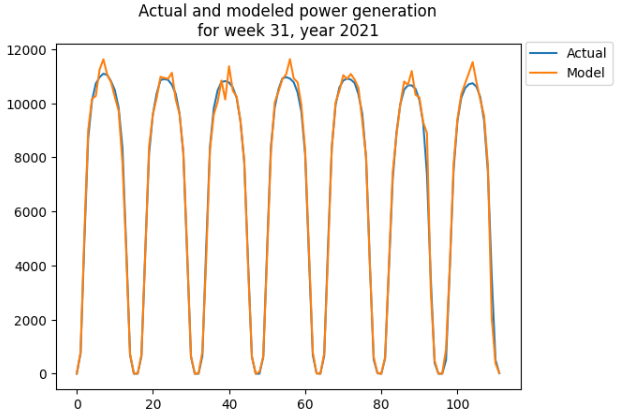
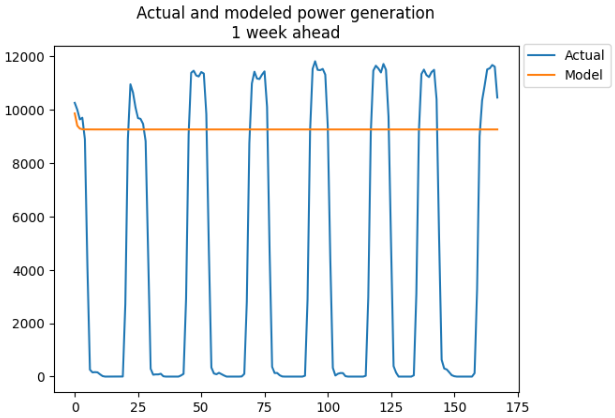


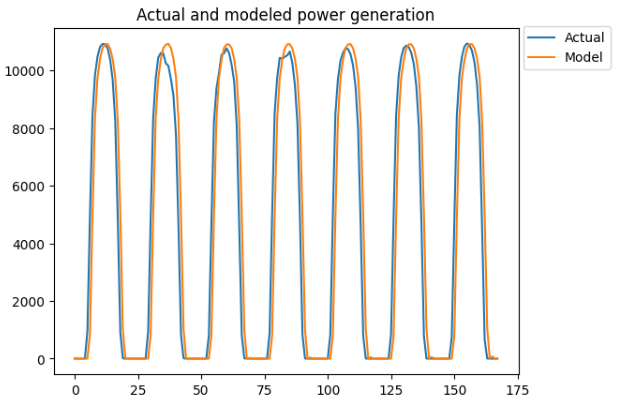
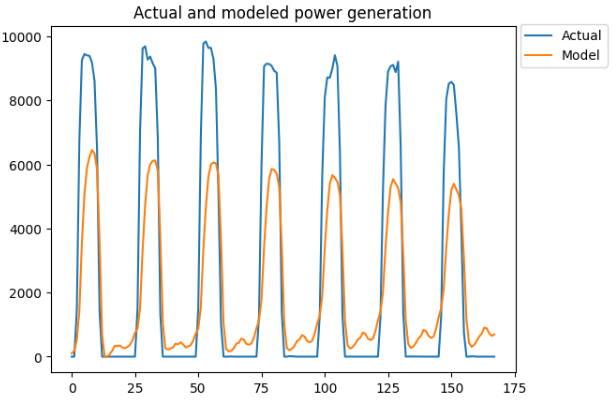
Figure-11. Actual and model power generation in a summer week in 2021 & 2022.

## 4.3 Time Series Modeling

Time series models were explored on the actual solar power generation without excluding night time. The auto\_arima function was utilized to optimize the parameters p, d, 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 order to forecast one week ahead revealed that it was poor as shown in Figure-12. Upon exploration, we noticed that only the autoregressive part of ARIMA model was the significant parameter, meaning that ARIMA model simplifies to autoregressive model. To identify the optimal number of lags in autoregressive model, we plotted the partial autocorrelation in the whole dataset as shown in Figure-13, and determined the optimal number of lags to be 74 hours. In order to maintain 80-20 train test split, the autoregressive model should be trained on 840 hours, to estimate the parameters associated with each of the 74 lags, in order to forecast the future 168 hours. The autoregressive model achieved an overall RMSE of 2015 MW, and R2 value of 76.9% which is the reference simple regression model. In addition, time series models lack explainability, which is important for the forecast problem. Consequently, autoregressive models or in general, time series models may not be suitable for the power generation forecast problem. Figure-14 shows the forecast performance of the autoregressive model in 2 different weeks, signifying the weak performance of the model.



*Figure-11. Actual and modeled power generation. Figure-12. Autoregressive partial autocorrelation.*

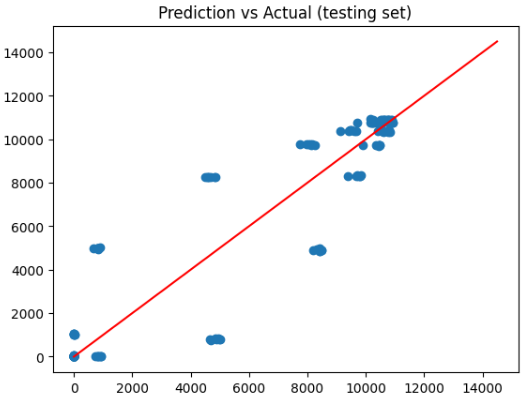
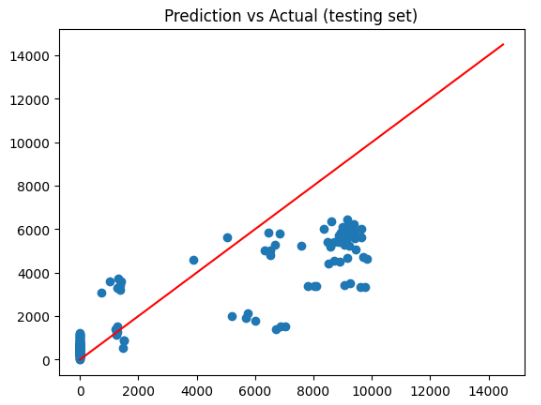
 

Figure-13. Autoregressive model performance in 2 different weeks.

# Conclusion and discussion

# References

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

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

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

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