# Forecasting Solar Energy Production: Exploring Weather Impacts and Machine Learning Models for Grid Optimization

DS3000 Introduction to Machine Learning


Joao Santos
*Department of Electrical and Computer Engineering, Faculty of Engineering, Western University*
London, Canada
jlangl@uwo.ca

Khalid Altahan
*Department of Electrical and Computer Engineering, Faculty of Engineering, Western University*
London, Canada
kaltahan@uwo.ca

Nicholas Moniz
*Department of Electrical and Computer Engineering, Faculty of Engineering, Western University*
London, Canada
nmoniz5@uwo.ca

Ayush Sharma
*Department of Electrical and Computer Engineering, Faculty of Engineering, Western University*
London, Canada
ashar463@uwo.ca

Rohan Datta
*Department of Electrical and Computer Engineering, Faculty of Engineering, Western University*
London, Canada
rdatta8@uwo.ca



***Abstract* - This study assesses the effect of weather on solar energy output and forecasts Global Horizontal Irradiance (GHI) using machine learning algorithms. A multilayer perceptron (MLP), linear regression, and XGBoost regression were the three models that were created and evaluated. An $R^2$ of 0.92 was attained by the MLP and XGBoost, with the MLP performing better thanks to a lower RMSE of 14.41. On the other hand, dataset constraints limited the linear regression model. The aforementioned findings underscore the capacity of sophisticated machine learning methodologies to enhance grid dependability, maximize energy storage, and anticipate renewable energy.**



***Index Terms – Solar Energy, Global Horizontal Irradiance (GHI), XGBoost, Renewable Energy Forecasting, Energy Optimization***


## I. Introduction

The output of renewable energy is highly variable due to its reliance on weather conditions, which creates significant challenges for managing the electrical grid. This project aims to explore how weather affects the production of solar energy and to develop models that can forecast these changes. One of the features is specifically, the Global Horizontal Irradiance (GHI) index, included in the dataset, which measures the irradiance over a horizontal surface in W/s². Accurate forecasts are crucial for optimizing the storage of energy, enhancing its reliability, and ensuring it is distributed efficiently. The project will assess various machine learning models to determine the most effective strategies for predicting the output of renewable energy.

## II. Exploratory Data Analysis

The dataset measured GHI, energy delta (a measurement of energy pulled from the grid after a given interval of time), temperature, wind speed, pressure, humidity, as well as the presence of some environmental variables such as clouds, rain, and snow at intervals of 15 minutes over the span of 3 years [1]. An initial analysis after creating a correlation matrix between some relevant features revealed clear relationships, especially between GHI and the Energy Delta, which proved to have a strong linear relationship based on the correlation matrix we generated in Figure 1.

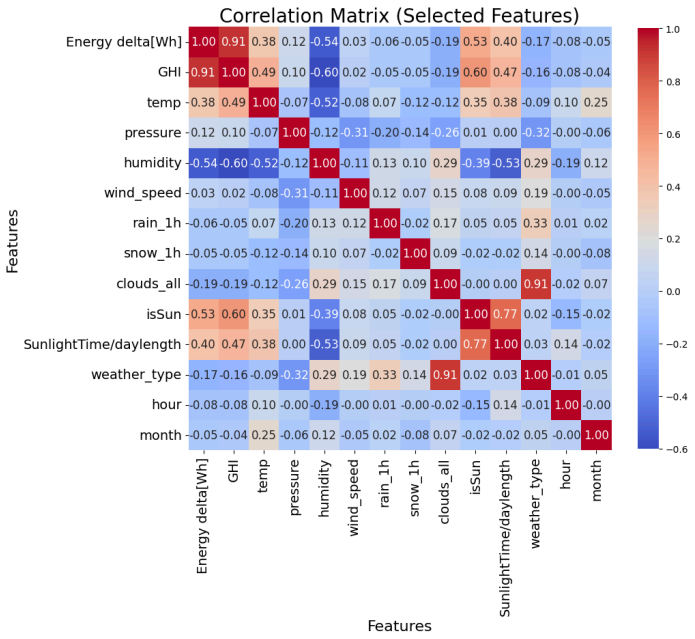


**Figure 1**. Correlation Matrix of All Features

From this information, two models were created with GHI being the target variable: a simple linear regression model using Energy Delta as a feature and an XGBoost regression model using some of the less correlated features. Both models were then evaluated using metrics such as Root Mean Squared Error (RMSE) and $R^2$ score on the training and test sets to ensure variance in the target variable could be explained by our selected features as well as to ensure our model didn't have too much overfitting.

## III. Approach

The research process started with cleaning and preprocessing the data to ensure its accuracy and consistency. Certain features, like isSun, sunlightTime, and dayLength, were removed in favor of other metrics such as SunlightTime/dayLength, which provided a more useful and meaningful metric of time than the other features [1]. The team created a set of pair plots in Figure 2 to help visualize relationships as well as help us select the most important features that had a strong linear relationship with GHI.

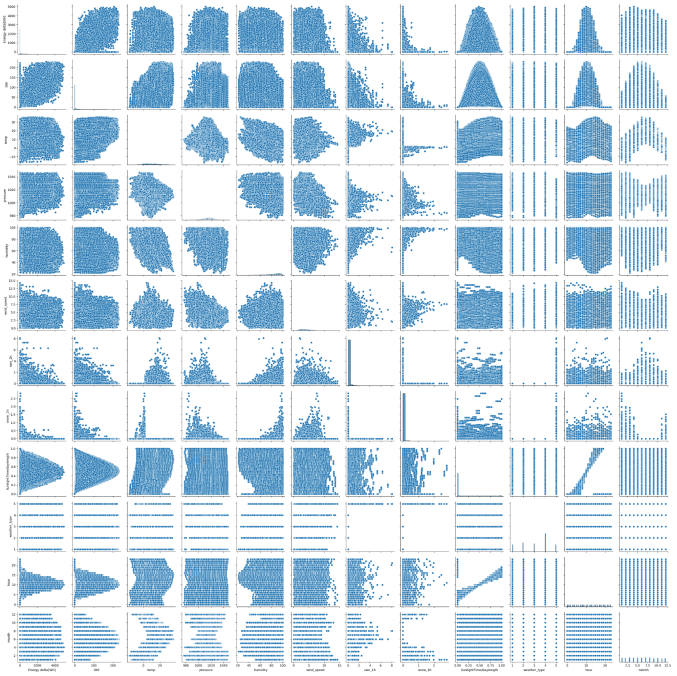

**Figure 2**. Pair Plots of all features vs. GHI

Based on relationships as seen in Figure 1 and Figure 2, the team decided on a simple linear regression model using Energy Delta as the predictor for the GHI since Energy Delta had a high correlation coefficient. The team also discussed the potential of creating a multiple linear regression model and if it would be practical. After further analysis of the heatmap (see Figure 1), the highest correlation coefficient was 0.6, which is a relatively weak relationship. Given this info, the team decided that a multiple linear regression model would not be useful or practical for the project.

The team agreed that an XGBoost regression model would be beneficial, and we decided to incorporate highly correlated variables like temperature, humidity, wind speed, pressure, and clouds, as well as the time of day to improve its performance. Lastly the team also agreed on developing a multi layer perceptron model. The data of all the models was divided into training and testing sets (80-20), and the predictions made by the models were assessed for their accuracy. The performance and results for both models can be seen in the findings section below.

## IV. Findings

A plot for the simple linear regression model between Energy Delta and GHI can be seen in Figure 3 below.

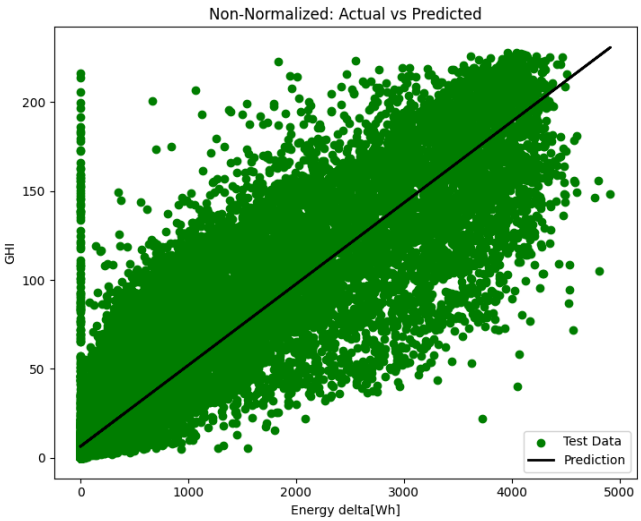


**Figure 3**. Linear Regression of GHI

This model yielded an RMSE value of 21.04 and an $R^2$ of 0.84 for the test set and an RMSE value of 21.11 and an $R^2$ of 0.84 on the training set. This suggests that the model performed relatively well, and around 84% of variance in GHI could be explained by Energy Delta. Since the evaluation metrics were almost identical in both the training and test sets we can also say that this model does not have overfitting and performs well on unseen data. However, upon further analysis of the dataset, the team noticed that while Energy Delta has a strong linear relationship with GHI, Energy Delta is not a reliable way of predicting GHI. This was due to the fact that when GHI is 0, Energy Delta will always be 0. GHI is 0 during hours when the sun isn't out, so this is what created the strong linear relationship. The model, however, struggles to predict GHI whenever the sun is out because Energy Delta is only related to how much energy was consumed, not how much was generated. Overall, while this model seemed to have performed well, outliers in the dataset created a false linear relationship that we determined after constructing the model. While the simple linear regression model didn't perform up to standards, the XGBoost model performed exceptionally well. Below in Figure 4 is a chart on how important each selected feature was in the model.

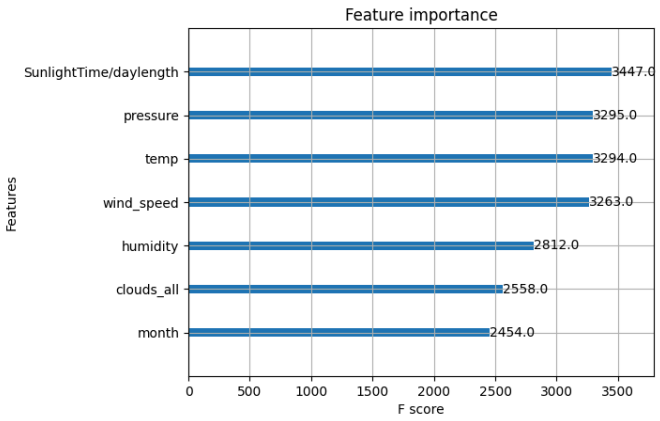


**Figure 4**. XGBoost Feature Importance for GHI

The XGBoost model yielded an RMSE value of 14.77 and an $R^2$ 0.92 on the test set. Considering the highest GHI value in our dataset is 229, our RMSE value is very good, and our $R^2$ suggests that 92% of variation in GHI can be explained by features seen in Figure 4. The following parameters below in Figure 5 were used to build the XGBoost model.

```
model = xgb.XGBRegressor(
    n_estimators=200,      # Number of trees
    learning_rate=0.2,     # Step size shrinkage
    max_depth=7,          # Maximum depth of a tree
    subsample=0.8,         # Row sampling
    colsample_bytree=0.8, # Feature sampling
    early_stopping_rounds = 10, # If model isn't improving stop
    reg_lambda = 1000, # Penalty
    random_state=42
)
```

**Figure 5.** XGBoost Model parameters

We discovered that it is possible to change both the RMSE value and $R^2$ by fine-tuning the parameters. Various trials were performed with different parameters in order to maximize the model's performance. The team had the option of creating a model with a low RMSE with mild overfitting or a model with a slightly higher RMSE with little to no overfitting. In general, everyone agreed that it was better to have less overfitting so our model would perform better on unseen data, so we structured the model's parameters to use a high penalty and a small number of trees. In order to test for overfitting, we calculated the RMSE and $R^2$ for the training set, which ended up being very close to the values in our test set. The training set's RMSE value was 14.11, and the $R^2$ was 0.93.

Overall we are very confident that this model will be able to predict GHI relatively accurately on unseen data. The Multi-Layer Perceptron (MLP) model was the last one to be

deployed, and it performed marginally better in predicting Global Horizontal Irradiance (GHI) than the Multi-Linear Regression and XGBoost models. With training metrics of $R^2$ = 0.9239 and RMSE = 14.37, the MLP obtained an $R^2$ of 0.9243 and an RMSE of 14.43 on the test set. The MLP model's capacity to generalize effectively to unknown data made the most noticeable difference, despite the modest overall performance improvement. This was demonstrated by the reduced disparity between the $R^2$ and RMSE values for the test and training sets. The MLP model's training and validation loss per epoch is shown in Figure 6.

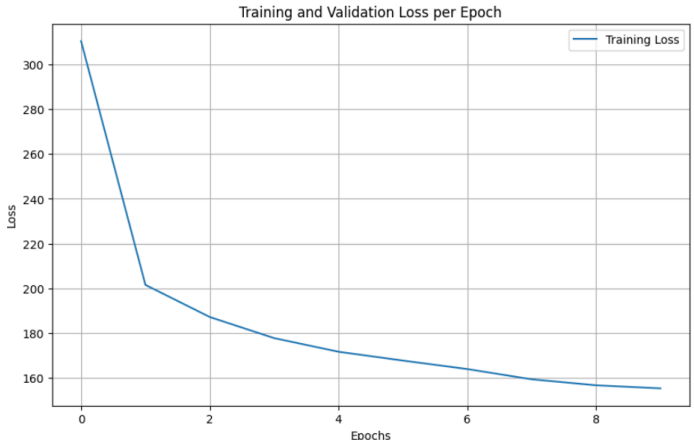


**Figure 6.** Training Loss per Epoch

The model's robustness is further demonstrated by the validation loss, which stays very near to the training loss with few variations, and the training loss, which gradually drops across the epochs, demonstrating efficient learning of the data's patterns. The MLP is the best-performing model in this study because of its unique ability to strike a balance between stable generalization and minimal error.

Table 1. shows the summary of evaluation metrics for all the different models that were developed.

Table 1. Summary of Results

| Model | Test $R^2$ | Test RMSE | Training $R^2$ | Training RMSE |
|---|---|---|---|---|
| MLR | 0.84 | 21.04 | 0.84 | 21.11 |
| XGB | 0.92 | 14.81 | 0.93 | 14.13 |
| MLP | 0.92 | 14.43 | 0.92 | 14.44 |

Overall the team is very content with the performance of the XGB and MLP models and we can say with confidence that these models can accurately predict GHI relatively accurately on unseen data. If you are curious or want more info on the results, see the Google Colab link attached in the references [2] or the Github repository [3].

## V. Future Directions

Further data analysis would aim at classifying the predictions into decisions regarding optimal energy production times and minimizing power loss. This would involve the creation of a model and prediction of an optimal threshold for the classification. Some models to be experimented with would be a Random Forest, which provides a simpler implementation but no direct regularization techniques or XGBoost, which is better optimized for complex data sets like the Renewable Power problem, but requires more tuning.

## VI. Conclusion

This study emphasizes how weather, specifically GHI, has a major influence on solar energy production. High-precision predictions of renewable energy generation have been proven to be possible through the use of machine learning algorithms, supported by our XGBoost investigation. These results demonstrate how data-driven strategies can help optimize renewable energy sources and electrical grid dependability, accelerating the transition to sustainable energy.

## VII. References

[1] Afroz, “Renewable Power Generation and weather Conditions,”Kaggle.com,2024. https://www.kaggle.com/datasets/pythonafroz/renewable-power-generation-and-weather-conditions/data (accessed Dec. 06, 2024).

[2] “Google Colab,” Google.com, 2019. https://colab.research.google.com/drive/15yHGdO3_d3TK5-xW7OnzfQhwRVNZRtp8 (accessed Dec. 06, 2024)

[3] Nico-M7, “GitHub - Nico-M7/ds3000-group22,” GitHub, 2024. https://github.com/Nico-M7/ds3000-group22 (accessed Dec. 06, 2024).

## VIII. Appendix A: Contribution

This section gives an outline for the contributions that each individual team member made towards this project throughout the semester. The distribution of responsibilities made throughout this project displays the collaborative effort that led to the completion of this study.

*Joao Santos* - Worked on finding an initial dataset that fits the requirements for the project, brainstormed methodologies that could be used. Completed a list of analysis methods that would be applied to the dataset in preparation for the coding phase. Helped out with the final presentation recordings to ensure a full description of our project was completely explained. Made the GitHub repository to submit all the code to. Completed the conclusion slide and content for the final presentation.

*Khalid Altahan* - Worked on creating the proposal report that outlined exactly what the group intended to accomplish throughout this project, helped complete the brainstorming presentation content and slides to ensure all sections were complete and explained well. Was responsible for managing the Google Colab document (environment used to run all of the code). Completed the results explanation and the content of the final presentation.

*Ayush Sharma* - Worked on finding an initial dataset that fits the requirements for the project, created the brainstorming and final presentation slides and filled in the content for all the sections. Did work on the introduction and methodologies sections of the final presentation. Played an important role in completing the introduction for both the draft and final report.

*Nicholas Moniz* - Worked on understanding the dataset to explain to others, brainstormed specific methods that could be used to do analysis with as well as data modelling methodologies. Also completed the voice-over for the brainstorming presentation. Programmed methods and imported libraries into the project code to conduct the data modelling. Did work on the introduction and methodologies sections of the final presentation. Completed an integral role in the report draft and final project report with filling in the sections for Methodologies and Findings.

*Rohan Datta* - Worked on finding the dataset that was used and fits the requirements for the project, studied the dataset in order to explain it to others, brainstormed some methodologies that could be used for analysis on the dataset. Completed a list of analysis methods that would be applied to the dataset in preparation for the coding phase. Worked on the code of the project, specifically on the data exploration aspect. Completed the content for the implementation in the final presentation. Ensured that the references and formatting for the reports were consistent with IEEE formatting as well as captioning all of the figures/tables.