Predicting Solar Energy with Explainable AI: A Gradient Boosted Regression Tree Approach

Shovan Jana, Student, IITH, Satyajeet Singh, Student, IITH,

Link of the online blog in Medium: https://medium.com/@shovanj.dakshana17/predicting-solar-energy-with-explainable-ai-a-gradient-boosted-regression-tree-approach-a51d3f1bfad9

Abstract—Solar energy integration into the power grid necessitates accurate prediction of daily solar influx. This project leverages Explainable AI (XAI) to enhance the credibility and interpretability of a Gradient Boosted Regression Tree (GBRT) model for solar energy prediction at a specific Oklahoma Mesonet site. Data from the AMS 2013-2014 Solar Energy Prediction Contest was used to train the GBRT model. Feature importance analysis via permutation importance and Recursive Feature Elimination (RFE) identified key features influencing solar energy, including downward short-wave radiation flux, precipitable water, and temperature. Hyperparameter tuning further optimized the model's performance. The final XAI-enhanced GBRT model achieved an R-squared score of 0.7934, demonstrating its effectiveness in solar energy prediction. This project highlights the importance of XAI in building trust and understanding in machine learning models for real-world applications, particularly in the domain of renewable energy integration.

Index Terms—Solar Energy, xAI, GBRT, Hyperparameter, Feature Importance.

I. Introduction

THE ever-growing demand for clean energy has propelled renewable sources like solar power to the forefront. However, integrating solar energy into the electrical grid presents a hurdle: its inherent variability due to weather. Utility companies require precise forecasts of solar energy production to maintain a delicate balance between renewable and fossil fuel sources. Inaccurate forecasts can translate to significant costs for utilities, either through excess fuel consumption or emergency purchases of electricity. Traditionally, numerical weather prediction models have been the cornerstone of power forecasting. Today, a powerful combination is emerging, statistical and machine learning techniques are being harnessed alongside these models to generate even more accurate predictions. But for wider adoption, there's a crucial need to unveil the inner workings of these complex machine learning models – a concept known as Explainable AI (XAI). This study delves into this very concept. By leveraging a simple Gradient Boosted Regression Tree (GBRT) model, we aim to predict the daily solar energy influx at a specific Oklahoma Mesonet site. The data used in this study originates from the AMS 2013-2014 Solar Energy Prediction Contest, where the American Meteorological Society challenged participants to predict solar energy for various Mesonet sites. Our project goes beyond just prediction. Various XAI techniques were employed to shed light on the features and decision-making processes within the GBRT model. This transparency not only enhances the model's credibility but also provide valuable insights for optimizing solar energy integration into the power grid.

II. Data and Preprocessing

The data used in this project originates from the AMS 2013-2014 Solar Energy Prediction Contest. The relevant data includes:

gefs_train.zip : It had 15 NetCDF4 files, each containing one model variable in a multidimensional array. The first dimension being date of model run (1994-01-01 to 200712-31), second, ensemble member (GEFS has 11 ensemble members), third dimension corresponds to forecast hour (runs from 12 to 24 hours in 3 hours increment). The fourth and fifth dimension were latitude and longitude of the grid. Data included weather variables like precipitation, radiation fluxes, air pressure, humidity, cloud cover temperature etc.

train.csv: Contains the total daily incoming solar energy (J/m²) for the Mesonet sites. This is the label for the whole dataset.

station_info.csv: Provides the latitude, longitude, and elevation of the Mesonet stations.

Our analysis focused on a particular Mesonet site named 'WEAT' (35 lat 262 long). Preprocessing the data involved first loading each of the netCDF data file into python multiindex data-frame for easy handling and then selecting the required data for the particular location (35 latitude 262 longitude). Now to combine all the time stamps, depending on the variable, either mean or sum of the values of each of the 5-time stamps were taken to calculate the daily amount of the variable. For example, for the temperature data taking a mean to get daily mean temperature is acceptable but for daily precipitation data, precipitation recorded during each time stamp needs to be added. But again, this was done for 11 ensemble members so to get one value for a day it was averaged over all the ensemble members. Finally, all these values were compiled into a single data frame with each column being a feature variable. This was a well curated data for a competition, so no missing values were found after inspection for the particular site.

S. Jana is with the Department of Sustainable Engineering, Indian Institute of Technology, Hyderabad, Telangana, India e-mail:gs23mtech11108@iith.ac.in

S. Singh is with the Department of Sustainable Engineering, Indian Institute of Technology, Hyderabad, Telangana, India email:gs23mtech11107@iith.ac.in.

III. Methodology

A. Model Selection:XGBoost

This project utilizes XGBoost, a gradient boosting framework firstly because this model is known for its capability for handling data of mixed type, being robust to outliers and also since it is nonparametric, and it has high predicting power. XGBoost trains multiple decision trees sequentially, with each tree focusing on improving the model's performance on previously misclassified examples. And also, all the winners of the original contest used the Gradient Boosted Regression Tree method to get the best results. A preliminary Gradient Boosted Regression Tree model was fitted using all the 15 features from the feature data frame mentioned earlier. The model was fitted using XGBoost's cross-validation technique for robust implementation. Then it was updated taking feedback from Feature Importance and Hyperparameter Tuning to get the subsequently optimised model.

B. Feature Importance

Understanding which features contribute most to the model's predictions is crucial. Three methods were employed for feature importance analysis:

- XGBoost's Built-in Feature Importance: XGBoost provides an intrinsic feature importance score based on how often a feature is used in a tree and the gain it brings to the model's split decisions. This can give a preliminary idea about which features might be more important than others.
- Permutation Combination: This technique randomly shuffles a feature's values, disrupting its relationship with the target variable. Training with an important feature in shuffled condition is theorised to decrease the model performance. The extent of decrease in model performance indicates the hierarchy of feature's importance.
- Recursive Feature Elimination (RFE): This method iteratively removes the least important feature based on a chosen metric (e.g., feature importance score) and retrains the model. This process continues until a desired number of features remain.

C. Hyperparameter Tuning

Hyperparameters are parameters of the learning algorithm that are not learned during training. Tuning these parameters is essential for optimal model performance. This project explores two hyperparameter tuning techniques:

- Grid Search: This exhaustive search method evaluates a predefined grid of hyperparameter combinations and selects the one yielding the best performance on a validation set.
- Hyperband: This is a more efficient approach that iteratively evaluates a series of hyperparameter configurations, focusing on promising regions while discarding less performing ones.

IV. Results and Discussion

The results obtained from preliminary model fitting using all the 15 variables had an R2 Score of 0.7829. The MAE and RMSE values were 2390521.44 and 3313316.32 respectively.

The XGBOOST inbuilt feature importance library was used to identify which of the features were contributing most for the model prediction. The results obtained are shown in figure

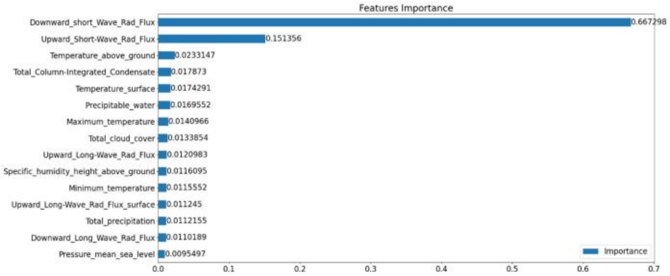

Fig. 1. XGBoost Feature Importance for Preliminary Model

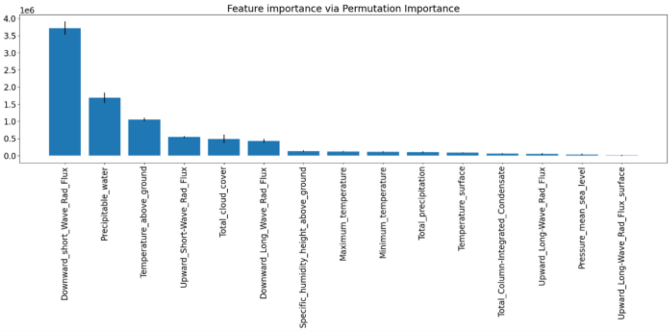

Fig. 2. Permutation Test Feature Importance for Preliminary Model

Downward Short-Wave Radiation flux was given the highest importance followed by Upward Short-Wave Radiation Flux. Combined they contributed to around 81Now to xAI method of Permutation Test was used to understand the feature importance in our model predictions. The results from the permutation test are shown in

The permutation test gave highest importance to Downward Short Wave Radiation Flux followed by Precipitable Water, Temperature above Ground, Upward Short Wave Radiation Flux and Total Cloud Cover etc. This result was not a coincidence, it can be explained physically too. The important features and their significance are:

Downward Short Wave Radiation Flux: As most of the solar energy coming to earth is contributed from short wave radiation so it signifies amount solar radiation reaching the earth's surface. More value indicates more potential solar energy. So, it is indeed the most contributing factor.

Precipitable Water: It represents the amount of water vapour present in the atmosphere. Water vapour has absorptive as well as reflective properties towards solar radiation. So it is an important factor towards the amount of solar radiation reaching the earth.

Temperature above ground: Also called Air Temperature. Different molecules in air may vibrate in different wavelengths at different temperatures causing absorption and scattering of solar radiation.

Upward Short-Wave Radiation Flux: Amount of shortwave radiation reflected from the Earth's surface. Lower albedo (reflectivity) of the ground surface translates to less reflection and potentially more energy available on earth.

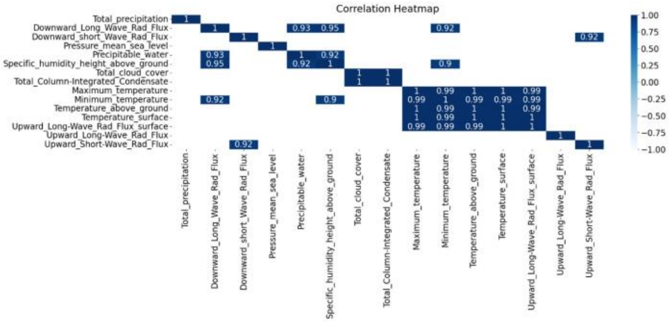

Fig. 3. Correlation Heatmap

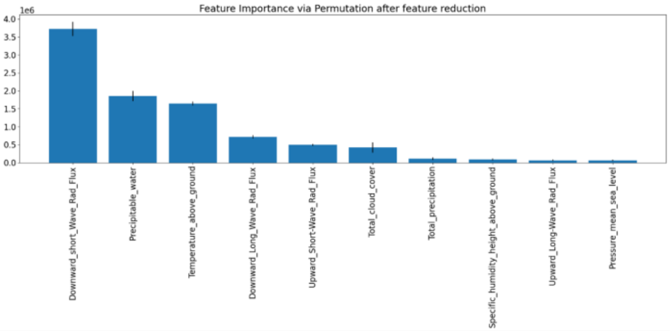

Fig. 4. Permutation Feature Importance after Feature Reduction

Along with important features both the tests showed that some of the features' contribution were very negligible. So, a correlation matrix was plotted to understand the correlation between all 15 variables. These correlations give us an idea of relation of one feature with another. Combining understanding from the importance tests and this matrix, some of the less important features were discarded which were highly correlated (coloured blue in the heat map) with more important features, and this decreases overall model calculations and potentially increase performance. The new features which were considered are 'Total precipitation', 'Downward Long Wave Radiation Flux','Downward Short Wave Radiation Flux','Pressure at Mean Sea Level','Precipitable Water', Specific Humidity at 2m above Ground', 'Total Cloud Cover', 'Temperature above Ground', 'Upward Long Wave Radiation Flux', and 'Upward Short Wave Radiation Flux'.

A new model was trained again using the reduced new features. Scores from Reduced Features Model were as follows:

R2: 0.7846 (Change in R2 Score 0.0017)

MAE: 2392578.70 (Change in MAE 2057.264)

RMSE: 3299903.57 (Change in RMSE -13412.740)

The results obtained from the reduced features for the model training had a R2 Score increased by 0.0017 depicting a slight increase in the model performance. Also, there was a significant decrease in RMSE score which depicts enhances the model performance. The Recursive Feature Elimination technique was implemented on this model to verify the feature selections. The result obtained from Recursive Feature Elimination backed up the selection to be optimum by giving all the new features important ranking and didn't eliminate any of the features.

Hyperparameters influences the complexity as well as learning of an ML model. By adjusting them the model can be fine tuned to achieve the best possible performance i.e. find the balance between underfitting and overfitting. The selection of best hyperparameters results in improving the model performance. Hence, to find the best hyperparameter for model training, Grid search and Hyperband Analysis was performed using "GridSearchCV" function from "sklearn" library and "hyperopt" library respectively. The best parameters from the Grid search and Hyperband analysis were found to be following:

Best Parameters from Grid Search:
learning_rate: 0.01 max_depth: 3
n_estimators: 2000

Best Parameters from Hyperband Analysis:
learning_rate: 0.01 max_depth: 3
n_estimators: 1900 colsample_bytree:
0.630 gamma: 4.07

The model trained using the parameters from grid search gave the best results. The model gave an R2 Score of 0.7934 depicting an improvement of 0.0087 from the previously trained R2 score. Even the MAE and RMSE values decreased further by 18853.49 and 67382.20 indicating a significant improvement in the model's performance.

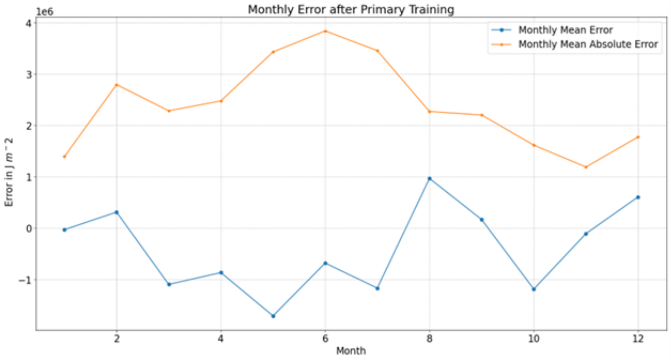

Fig. 5. Monthly Error after Primary Training

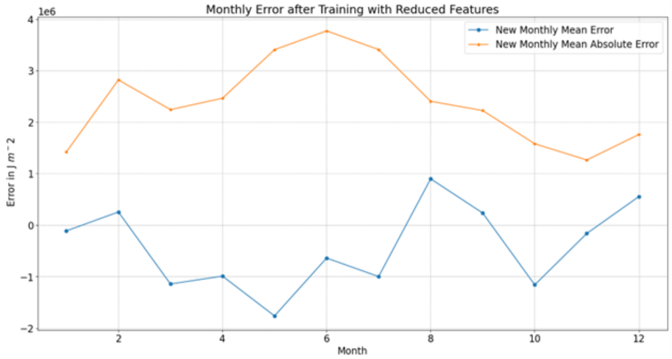

Fig. 6. Monthly Error with Reduced Feature Model

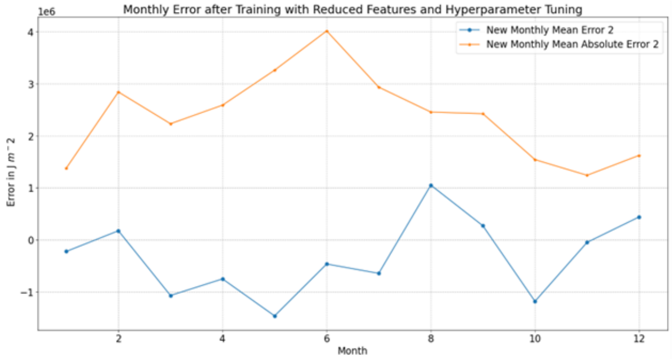

Fig. 7. Monthly Error with Reduced Feature and Hyperparameter Tuning

Month wise MAE (Mean Absolute Error) and Mean Error were plotted to understand the monthly variations in errors and improvements across all the three models.

From the graphs plotted it can be observed that there are the highest absolute errors but the least mean errors in the months of April, May, June, and July. It means that during these months there are large positive as well as negative deviation from the actual value such that the opposite signs nullify each other in case of mean error but gets added up for absolute error. This can be explained physically as follows:

- Spring and summer months in Oklahoma typically experience more frequent cloud cover compared to fall and winter.
- Increased convective activity, leading to afternoon thunderstorms and showers.
- Have higher humidity levels compared to drier fall and winter months. Complex interaction between humidity and solar radiation.
- Steeper change in the solar zenith angle (angle between the sun and a location on Earth).

These rapid weather changes are very unpredictable and cannot be modelled easily so the large deviation is observed during these months.

The final model prediction using the most important features and the best hyperparameters is plotted for the train and the test dataset.

It can be observed that the model predictions fit and captures the pattern reasonably well.

V. Conclusion

This project shows the application of Gradient Boosted Regression Trees (via XGBoost) for short-term solar energy prediction at a single Mesonet site while demonstrating importance of high complexity climate data preprocessing and impact of correlated features and potential of feature selection and engineering to enhance model performance. By employing feature importance techniques like permutation importance, recursive feature elimination and hyperparameter tuning methods like grid search hyperband it outlined the systematic approach to effective application of ML for solving Climate related requirements. It also showed how physical interpretations add valuable context to the model and also provides credibility to the predictions for real world application. Future work could involve expanding the analysis to multiple Mesonet sites and investigating the impact of incorporating additional features of time series like ‘lagged features’ on solar energy data prediction. Also, alternative machine learning algorithms can be explored to have a comparative understanding of the field. By continuously improving prediction models, we can contribute to a more reliable and efficient utilization of solar energy as a renewable energy source.

References

[1] A. McGovern, D. J. Gagne, J. Basara, T. M. Hamill, and D. Margolin, “Solar Energy Prediction: An International Contest to Initiate Interdisciplinary Research on Compelling Meteorological Problems,” Bulletin of the American Meteorological Society, vol. 96, no. 8, pp. 1388–1395, Aug. 2015, doi: 10.1175/BAMS-D-14-00006.1.

[2] “AMS 2013-2014 Solar Energy Prediction Contest.” Accessed: May 04, 2024. [Online]. Available: https://kaggle.com/competitions/ams-2014solar-energy-prediction-contest

[3] “XGBoost Documentation — xgboost 2.0.3 documentation.” Accessed: May 04, 2024. [Online]. Available: https://xgboost.readthedocs.io/en/stable/