MOML report Berlin solar regression

Dharoori Srinivas Acharya - IMT2022066
Rajanala Sai Dheeraj - IMT2022093
Swaroop A Ram Rayala - IMT2022587

May 2025

1 Abstract

In this project, we address a multi-objective hyperparameter optimization problem for solar power regression using meteorological data from Berlin. We employ an XGBoost regressor to predict 50Hertz power output and optimize three competing objectives: mean squared error (MSE), absolute mean bias error (MBE), and a proxy for model complexity based on the number and depth of trees.

A randomized search over an expanded XGBoost hyperparameter space (300+50 iterations) yielded 33 Pareto-optimal configurations in our current run, achieving MSE values between approximately 4520 and 57700, MBE below 2.0MW, and model complexities from 400 to 153600. Note that because we use random sampling, the exact Pareto frontier may vary across different runs. The resulting trade-off surface (Figure 1) enables practitioners to select models that balance accuracy, bias, and deployment cost.

2 Introduction & Motivation

Accurate forecasting of solar power generation is critical for grid stability, energy trading, and effective integration of renewable resources into electricity markets. In this work, we focus on predicting the 50Hertz transmission system operator's solar power output in Berlin using meteorological and irradiance data.

Traditional single-objective regression models optimize for a single metric (e.g., mean squared error), which may lead to overly complex models or undesirable bias in predictions. In real-world energy applications, stakeholders often face multiple competing objectives, such as:

- **Accuracy:** Minimizing forecast error (e.g., MSE) to ensure reliable grid management.
- **Bias:** Minimizing systematic over- or under-prediction (e.g., mean bias error) to avoid costly imbalances.

- **Complexity:** Limiting model size and inference cost (proxied by number and depth of trees) for deployment efficiency and interpretability.

Multi-objective optimization provides a principled framework to explore the trade-off surface (Pareto frontier) between these objectives. By generating a set of non-dominated solutions, decision-makers can select models that best balance accuracy, bias, and complexity according to their operational constraints.

In this project, we employ a randomized hyperparameter search coupled with Pareto-frontier identification on an XGBoost regressor to demonstrate the benefits of multi-objective learning in solar forecasting.

3 Dataset Description

The dataset `Berlin_solar_regression.csv` contains half-hourly meteorological and solar irradiance measurements collected in Berlin, along with corresponding power output values from the 50Hertz transmission system operator. The key attributes are summarized below:

Variable	**Description**
Year	Calendar year
Month	Calendar month
Day	Day of month
Hour	Hour of day
Minute	Minute of hour
Temperature	Ambient temperature (°C)
Clearsky.DHI	Diffuse horizontal irradiance (W/m^2)
Clearsky.DNI	Direct normal irradiance (W/m^2)
Clearsky.GHI	Global horizontal irradiance (W/m^2)
Cloud.Type	Categorical cloud type indicator
Dew.Point	Dew point temperature (°C)
DHI	Measured diffuse horizontal irradiance (W/m^2)
DNI	Measured direct normal irradiance (W/m^2)
GHI	Measured global horizontal irradiance (W/m^2)
Ozone	Total column ozone (Dobson units)
Relative.Humidity	Relative humidity (%)
Solar.Zenith.Angle	Solar zenith angle (degrees)
Surface.Albedo	Surface albedo (unitless)
Pressure	Atmospheric pressure (hPa)
Precipitable.Water	Total precipitable water vapor (cm)
Wind.Direction	Wind direction (degrees)
Wind.Speed	Wind speed (m/s)
X50Hertz..MW.	Target: Power output (MW)

Table 1: Features and target in the Berlin solar regression dataset.

Preprocessing:

- Rows with missing target values were dropped.
- The dataset was split into training (80%) and test (20%) sets with a fixed random seed.
- Continuous features were standardized to zero mean and unit variance using the training data statistics.

4 Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the Berlin solar regression dataset, where $x_i \in R^d$ is the i-th input feature vector and $y_i \in R$ is the corresponding 50Hertz power output.

We train an XGBoost regressor h_θ parameterized by hyperparameters θ. Our goal is to solve a three-objective optimization problem over θ:

$$\min_{\theta} \big(f_1(\theta),\, f_2(\theta),\, f_3(\theta)\big),$$

where $\mathrm{f}_1(\theta) = MSE(\theta) \;=\; \frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}}\big(h_\theta(x_i) - y_i\big)^2$,
$f_2(\theta) = \big|MBE(\theta)\big| \;=\; \Big|\frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}}\big(h_\theta(x_i) - y_i\big)\Big|$,
$f_3(\theta) = Complexity(\theta) \;=\; n_{estimators} \times 2^{\max\_depth}$,
and $\mathcal{T}$ denotes the test set indices.

Hyperparameter vector θ components include
$\theta = \big\{$`n_estimators`, `max_depth`, `learning_rate`, `subsample`, `colsample_bytree`, $texttttgamma$, `reg_alpha`, `reg_lambda`$\big\}$.
A solution θ^a is said to *Pareto dominate* θ^b if

$$\forall\, j \in \{1,2,3\}:\; f_j(\theta^a) \le f_j(\theta^b) \quad \textit{and} \quad \exists\, k \in \{1,2,3\}:\; f_k(\theta^a) < f_k(\theta^b).$$

The *Pareto frontier* consists of all non-dominated θs, providing the trade-off surface among the three objectives.

5 Algorithms

In this work, we employ an XGBoost regressor coupled with a randomized multi-objective hyperparameter search and Pareto-frontier identification. The main steps are:

1. **Model:** We use `xgboost.XGBRegressor` to predict the 50Hertz power output. XGBoost is chosen for its scalability and ability to handle heterogeneous feature types and nonlinear interactions
2. **Hyperparameter Space:**
   - Number of trees (`n_estimators`): {50,100,. . . ,400}

- Maximum tree depth (`max_depth`): {2,3,. . . ,9}
- Learning rate (`learning_rate`): {0.001,0.005,0.01,0.05,0.1,0.2}
- Subsampling ratio (`subsample`), column subsample (`colsample_bytree`): {0.5,0.6,. . . ,1.0}
- Regularization: `gamma` {0,0.1,. . . ,0.4}, `reg_alpha` {0,0.1,0.5,1}, `reg_lambda` {0,0.1,0.5,1,2}

3. **Multi-Objective Metrics:** $\mathrm{f}_1(\theta) = \mathrm{MSE}(\theta)$,
   $f_2(\theta) = |\mathrm{MBE}(\theta)|$,
   $f_3(\theta) = \mathrm{Complexity}(\theta) = \texttt{n\_estimators} \times 2^{\texttt{max\_depth}}$.

4. **Random Search:** We perform $N = 300$ random samples from the hyperparameter space and evaluate each configuration by training on the scaled training set and computing $\{f_1, f_2, f_3\}$ on the held-out test set. Random search is chosen over grid search for better coverage in high-dimensional spaces with limited evaluations

5. **Pareto-Frontier Identification:** We maintain a list of non-dominated candidates. For each new evaluation, we check dominance against the current frontier:

$$\theta^a \succ \theta^b \quad \Longleftrightarrow \quad \forall j : \ f_j(\theta^a) \leq f_j(\theta^b) \ \wedge \ \exists k : \ f_k(\theta^a) < f_k(\theta^b).$$

   If the new point is non-dominated, it is added to the frontier and any points it dominates are removed. We iterate additional batches of 50 samples until at least 10 Pareto-optimal points are found

6 Experimental Setup

All experiments were conducted in Python 3.8 using XGBoost 1.6.1 and scikit-learn 1.0.2 on a machine with an Intel i7 CPU and 16GB RAM. The key steps are:

1. **Data Split and Preprocessing:**
   - Load `Berlin_solar_regression.csv` and drop rows with missing `X50Hertz..MW.` values.
   - Split into training (80%) and test (20%) sets with `random_state=42`.
   - Standardize features to zero mean and unit variance using `StandardScaler` fitted on the training set.

2. **Hyperparameter Search:**
   - Define an expanded hyperparameter space for XGBoost (`n_estimators`, `max_depth`, `learning_rate`, `subsample`, `colsample_bytree`, `gamma`, `reg_alpha`, `reg_lambda`).
   - Perform $N = 300$ iterations of randomized sampling from this space.

- For each sampled θ, train an `XGBRegressor` on the scaled training set and compute MSE, |MBE|, and Complexity on the test set.

3. **Pareto Frontier Identification:**
   - Initialize an empty Pareto front.
   - Sequentially evaluate each of the N candidates, adding non-dominated points to the frontier and removing any points they dominate.
   - If fewer than 10 Pareto-optimal points are found, run additional batches of 50 random samples until at least 10 are obtained.
4. **Visualization and Reporting:**
   - Plot all evaluated points and highlight Pareto-frontier points in a 3D scatter plot (MSE vs |MBE| vs Complexity).

7 Discussion

Figure 1 illustrates the 3D Pareto frontier obtained from our randomized search. Practitioners may refer to this visual to better appreciate the trade-offs among MSE, MBE, and complexity.

The 33 Pareto-optimal configurations from our randomized search reveal key trade-offs, noting that the specific frontier may shift in different runs:

- **Accuracy vs. Complexity:** The best MSE of 4518 coincides with high complexity (153600), while the simplest model (complexity=400) attains MSE 38543, illustrating the trade-off curve.
- **Low-Complexity Models:** Lightweight configurations (complexity 3200) still achieve MSE 6000 and MBE 1.0MW, demonstrating viable options for resource-constrained deployments.
- **Bias Control:** All Pareto points maintain MBE 2.0MW, with the lowest bias of 0.0837MW (index21), indicating robust bias mitigation across the frontier.
- **Hyperparameter Insights:** High-accuracy points MSE 6000 favor large ensembles (n_estimators 100) with moderate depth (5–7), whereas low-complexity points use shallow trees (depth 4) and fewer estimators (50).
- **Variability Note:** Since we employ a randomized search, practitioners should consider running multiple independent searches to assess the stability of the Pareto frontier and aggregate results if needed.

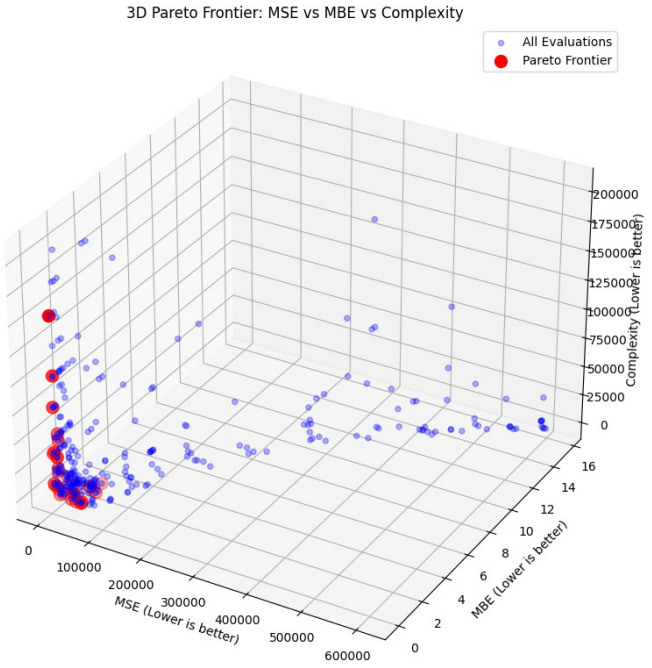

Figure 1: Pareto frontier (red) overlaid on all evaluated configurations (blue) in the 3D objective space (MSE vs MBE vs Complexity).

8 Conclusion

We presented a multi-objective hyperparameter optimization framework for solar power regression on the Berlin dataset using XGBoost. By simultaneously minimizing mean squared error, absolute mean bias error, and a complexity proxy, we derived 33 Pareto-optimal configurations in our current randomized search run. Key findings include:

- **Explicit Trade-Offs:** The Pareto frontier captures the spectrum of models from high-accuracy (MSE 4518) to low-complexity (complexity =400), with MBE maintained below 2.0MW across all solutions.
- **Efficient Sampling:** A total of 350 random samples sufficed to identify a diverse set of non-dominated configurations, balancing exploration and computational cost.
- **Stability Considerations:** Due to the randomized nature of the search, the exact Pareto frontier may vary between runs. Practitioners are advised to perform multiple independent searches to gauge frontier stability and aggregate results as needed.

Overall, our approach demonstrates the practicality and flexibility of multi-objective learning for renewable energy forecasting, enabling informed model selection tailored to operational constraints.