Solar Irradiance Forecasting Using Machine Learning

Vaibhav

July 3, 2025

1 Project Overview

This project develops a machine learning framework for forecasting Global Horizontal Irradiance (GHI) using meteorological and temporal features. The goal is to predict solar irradiance values with high accuracy for renewable energy applications.

2 Data Preprocessing and Feature Engineering

2.1 Data Cleaning

The dataset underwent systematic preprocessing:

- Removed features with more than 65% missing values: `horizontal_radiation_2`, `incident_radiation_2`, `reflected_radiation_2`
- Eliminated highly correlated features (correlation greater than 0.90) through iterative correlation analysis
- Imputed `module_temperature_3` missing values using median imputation
- Applied IQR-based outlier capping to preserve data integrity

2.2 Feature Engineering

Comprehensive feature engineering was performed:

- **Temporal Features**: Hour, month, day, day-of-week, and part-of-day (dawn/morning/afternoon/evening/night)
- **Lag Features**: GHI lag-1 and GHI difference-1 to capture temporal dependencies
- **Categorical Encoding**: One-hot encoding for part-of-day feature

3 Exploratory Data Analysis

3.1 Target Variable Distribution

GHI exhibits a right-skewed distribution typical of solar irradiance data, with frequent zero values during nighttime hours and peak values during midday.

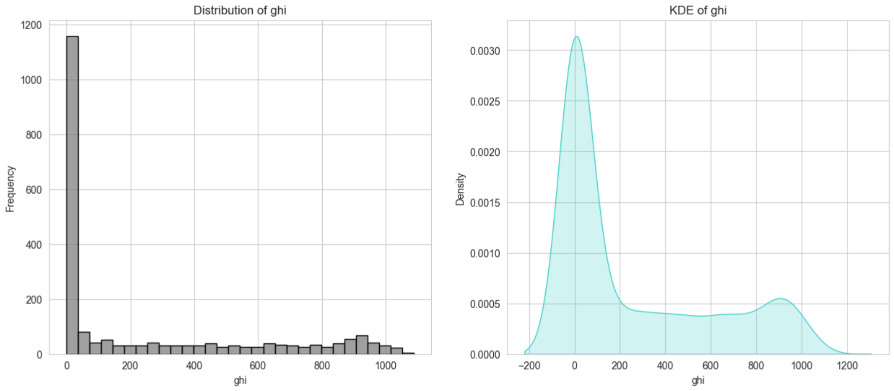
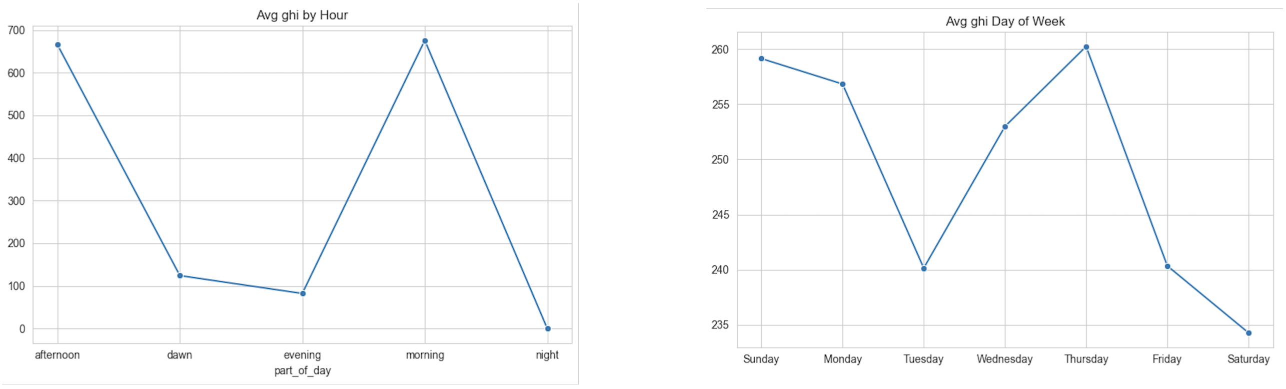


Figure 1: GHI Distribution

3.2 Temporal Patterns

Analysis revealed clear diurnal and weekly patterns in GHI values, with highest irradiance during morning and afternoon periods.



(a) GHI by Part of Day

(b) GHI by Day of Week

Figure 2: Temporal GHI Patterns

3.3 Feature Relationships

Correlation heatmap analysis identified key predictive features: hour (strong diurnal pattern), ambient temperature (positive correlation), and GHI lag-1 (sequential dependency).

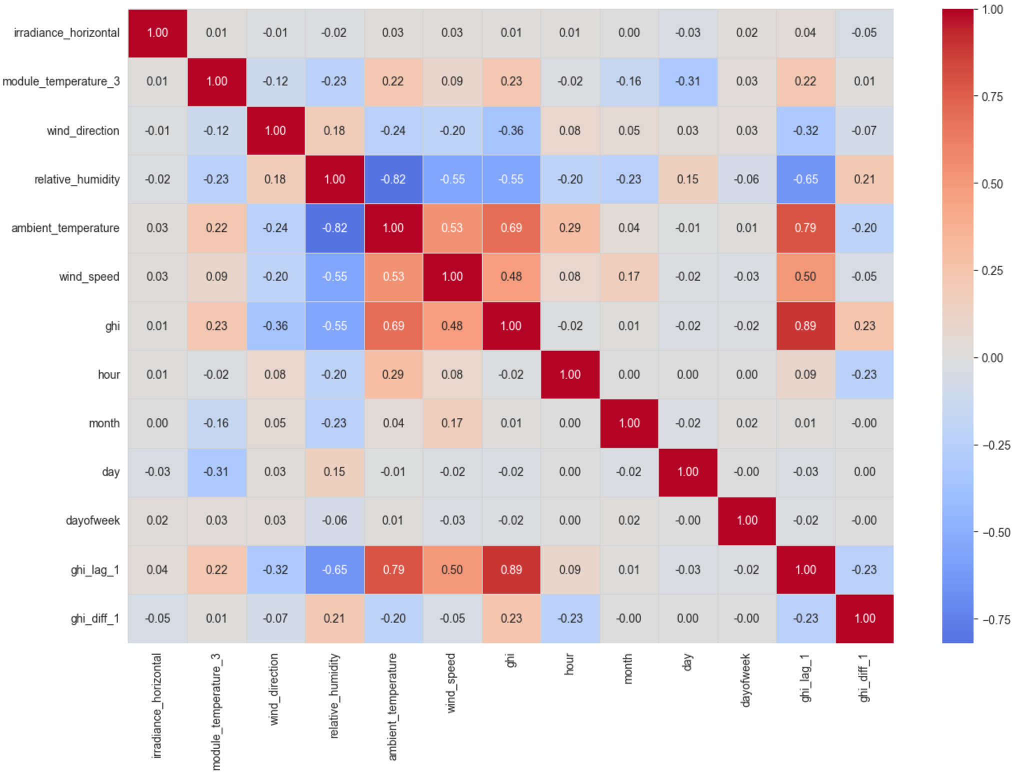


Figure 3: Final Correlation Heatmap

4 Model Architecture and Implementation

4.1 Approach

Three regression models were implemented and compared:

4.1.1 Linear Regression

Baseline model with standardized features. Simple linear relationships assumed between features and target.

4.1.2 Random Forest Regressor

Ensemble of decision trees with bootstrap aggregation. Handles non-linear relationships and provides feature importance rankings.

4.1.3 XGBoost Regressor

Gradient boosting framework with tree-based learners. Advanced regularization techniques and efficient optimization for high predictive accuracy.

4.2 Model Configuration

- **Data Split**: 60% training, 20% validation, 20% test
- **Scaling**: StandardScaler applied to all features
- **Hyperparameter Tuning**: GridSearchCV with 3-fold cross-validation

5 Hyperparameter Tuning

Optimal parameters identified through grid search:

Model	**Key Parameters**	**Optimal Values**
Random Forest	n_estimators, max_depth	200, 20
XGBoost	n_estimators, learning_rate, max_depth	200, 0.1, 6

Table 1: Optimal Hyperparameters

6 Evaluation Method

6.1 Custom MAPE Implementation

A specialized MAPE function was developed to handle solar data characteristics:

$$\text{Custom MAPE} = \frac{1}{n}\sum_{i \in S}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

where $S = \{i : y_i \notin (-1, 1)\}$

This excludes near-zero values that occur during nighttime, focusing evaluation on meaningful solar irradiance periods.

6.2 Additional Metrics

- **RMSE**: Root Mean Square Error for magnitude-sensitive evaluation
- **MAE**: Mean Absolute Error for robust performance assessment
- **R²**: Coefficient of determination for explained variance

7 Results and Performance

7.1 Model Comparison

Model	**Validation MAPE (%)**	**Test MAPE (%)**	**Status**
Linear Regression	4.45e-12	1348.56	Overfitting
Random Forest	18.99	19.2	Good
XGBoost	16.66	**14.53**	Excellent

Table 2: Model Performance Comparison

7.2 Final Model Performance

The optimized XGBoost model achieved:

- **Test MAPE**: 14.53% (excluding near-zero values)
- **Strong Generalization**: Consistent performance across validation and test sets
- **Robust Predictions**: Effective handling of temporal patterns and weather dependencies

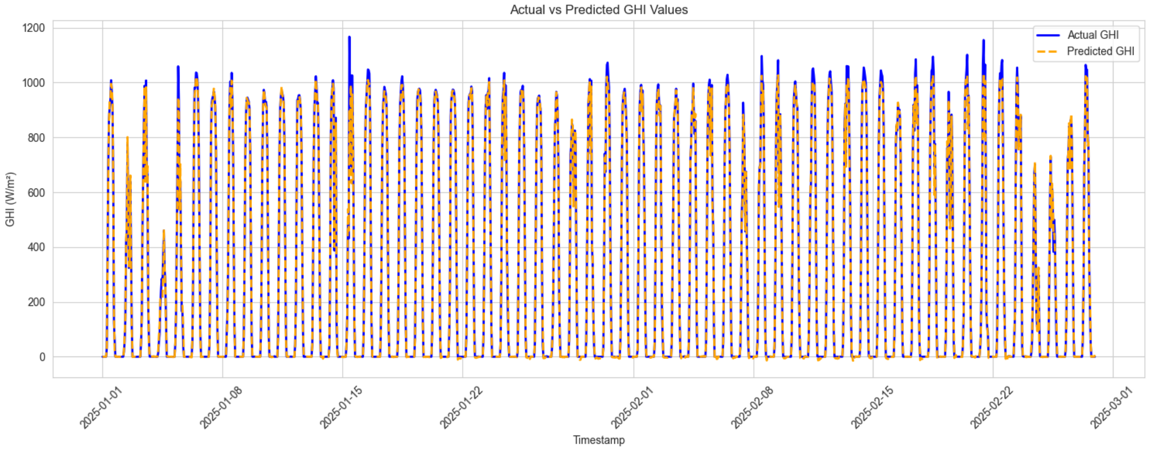


Figure 4: Actual vs Predicted GHI Values - Time Series Comparison

The time series plot demonstrates the XGBoost model's ability to accurately capture the diurnal patterns of solar irradiance, with predicted values closely following the actual measurements across different weather conditions and seasonal variations. The model effectively handles both peak irradiance periods and low-light transitions, validating its robustness for practical solar forecasting applications.

8 Key Insights

8.1 Feature Importance

Most influential features for GHI prediction:

1. **Temporal Features**: Hour and part-of-day showed strongest predictive power
2. **Lag Features**: GHI lag-1 captured essential temporal dependencies
3. **Meteorological Features**: Ambient temperature contributed significantly

8.2 Model Behavior

- **XGBoost Superiority**: Gradient boosting effectively captured non-linear solar patterns
- **Ensemble Benefits**: Random Forest provided robust baseline performance
- **Linear Limitations**: Simple models inadequate for complex temporal relationships

9 Challenges Faced

9.1 Data Quality Issues

- **Missing Data**: Significant gaps in radiation measurements required careful feature selection
- **Multicollinearity**: High correlation between measurements needed systematic elimination
- **Outliers**: Extreme meteorological values required robust treatment strategies

9.2 Evaluation Complexity

- **Zero Values**: Nighttime GHI values complicated standard MAPE calculations
- **Custom Metrics**: Required specialized evaluation function for meaningful assessment
- **Temporal Dependencies**: Complex time-series patterns challenging to capture

9.3 Computational Considerations

- **Hyperparameter Tuning**: Extensive grid search computationally intensive
- **Feature Engineering**: Iterative correlation analysis time-consuming
- **Model Complexity**: Balance between accuracy and interpretability

10 Conclusion

This project successfully developed a robust machine learning framework for solar irradiance forecasting, achieving a test MAPE of 14.53% with the optimized XGBoost model. Key contributions include:

- Custom MAPE metric suitable for solar data evaluation
- Comprehensive feature engineering with temporal and lag features
- Systematic correlation-based feature selection
- Robust outlier treatment preserving data integrity

The final XGBoost model demonstrates strong predictive performance and generalization capability, making it suitable for practical solar energy forecasting applications. The methodology provides a solid foundation for renewable energy prediction systems and can be adapted to other forecasting domains.