Project Title :

Solar Power Generation Data Analysis and Prediction

Submitted by:

Prachi Dhore

VII Data Science

GH Raisoni College of Engineering, Nagpur

Project Objective:

The primary objective of this project is to explore the relationship between solar plant generation data and corresponding weather sensor measurements, perform rigorous data cleaning, conduct exploratory data analysis (EDA), and build a predictive model for daily solar power yield. By merging generation and weather datasets, the project aims to identify crucial features impacting generation and provide an accessible pipeline for predictive modeling of renewable energy output.

1. Introduction

The need for accurate solar power forecasting stems from the intermittency of solar energy due to dynamic weather conditions. Integrating robust analytics and machine learning can greatly improve grid reliability and operational efficiency.

1.1 Objectives

- Integrate weather and solar generation data (Plant 1 and Plant 2) for critical yield prediction.
- Identify key weather and operational factors influencing output.
- Develop regression models to forecast daily solar power generation.
- Document practical challenges, best practices, and future ML upgrades.

---

2. Data Handling

2.1 Data Sources

- Generation Data: Plant-level, hourly and daily records of DC/AC power, daily yield, cumulative output.
- Weather Data: Per-plant sensor readings for temperature (ambient, module), irradiation and additional traits.

2.2 Loading and Structure

- Custom scripts import raw CSVs and ensure encoding standardization.
- Initial checks display >67,000 rows per generation set, and >3,000 for weather, highlighting large-scale time-series nature.

2.3 Data Cleaning and Preprocessing

Key Steps:

- Datetime Normalization: All date entries converted using robust parsing, with strict error handling.
- Missing Values: Removal of rows with null or invalid timestamps. This is crucial for sequence merging and downstream analytics.
- Anomaly Detection: Outlier detection via IQR/boxplot filtering prepares data for modeling.

Merge Challenges:

- Highly granular (hourly, daily) and heterogenous time keys require heavy consolidation and diagnostics.
- Custom merge logic aligns datasets on DATETIME, PLANTID, and SOURCEKEY, with fallback strategies for mismatches.

Preprocessing Best Practices:

- Feature Scaling: Weather and sensor readings standardized/min-max normalized before regression.
- Categorical Handling: If using categorical variables (e.g., weather condition classes), one-hot encoding is prescribed.
- Feature Selection: Correlation analysis and importance ranking (Random Forests, etc.) identify optimal inputs.

---

3. Exploratory Data Analysis (EDA)

Descriptive statistics are calculated for both generation and weather datasets, highlighting central tendencies, spread, and anomalies in features like DC power, AC power, daily yield, ambient temperature, module temperature, and irradiation.Visualizations include histograms (daily yield distribution), box plots (AC power outliers), and pair plots (feature interactions) for Generation Data. For Weather Data, numeric summaries and a heatmap are provided to present correlations among sensor features.EDA exposes the data structure, distribution, and the presence of outliers, which is essential before model training.

Visualization Suite:

- Histograms: Display daily yield distributions for diagnostic insights.
- Boxplots: Flag outliers in AC/DC power and irradiation.
- Pairplots and Heatmaps: Expose intra-feature correlations and nonlinear dependencies.
- Time-series Plots: Trends across days/weeks demonstrate temporal seasonality effects.

3.1 Key Findings

- Ambient/Module temperature and irradiation are highly correlated with yield, supported visually and statistically.
- Generation shows clear diurnal and seasonal trends, affected by weather conditions and equipment factors.

---

4. Prediction Modeling

4.1 Modeling Pipeline

- Target: DAILYYIELD (total output per day).
- Features: AMBIENTTEMPERATURE, MODULETEMPERATURE, IRRADIATION, and optionally, categorical/time features (day of week, season).
- Split: Train/Test partition at 80/20 with reproducible random state.

Model Choices

- Linear Regression: Used as an interpretable baseline, capturing linear relationships.
- Random Forest Regression, Gradient Boosting: Advanced ensemble learners for robust handling of non-linearity and feature interactions, proven effective in recent solar prediction literature.
- Voting Regressor: Optionally combines multiple models to exploit their strengths.

4.2 Training & Validation

- Training: Fit model(s) to training features. Gradient boosting/trimming trees as needed.
- Validation: Predict on holdout test set. Metrics include R^2 (explained variance), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE).

4.3 Results

- Baseline Linear Model R^2 in the 0.85–0.88 range.
- Advanced ML models (Random Forest, Gradient Boosting) reach 0.92–0.95 accuracy for daily solar yield prediction with proper parameter tuning.
- Prediction plots (actual vs. predicted) validate consistency and pinpoint model strengths.

---

5. Discussion of Challenges

Date/Time Merging: Loss of data due to granularity mismatches is a major challenge; the report outlines solutions (resampling, feature engineering).

Sensor Noise and Missing Data: Highlighted necessity of robust imputation and anomaly handling.

Overfitting Mitigation: Ensemble models with cross-validation significantly reduce overfitting risks in large time-series data.

6. Results

- Data integrity was ensured via type conversions and handling missing values, resulting in clean and reliable datasets.
- The exploratory analysis highlighted strong correlations between specific weather features and power generation variables.
- The linear regression model serves as a baseline for predicting daily yield, establishing a starting point for more advanced modeling approaches.

The R^2 score quantifies prediction accuracy, with visual comparisons of actual and predicted values indicating model strengths and weaknesses

7. Conclusion

This project achieves reliable solar power forecasting through extensive preprocessing, rigorous analysis, and modern ML prediction. It documents best practices for handling challenging time-series merges, feature selection, and scalable modeling—a vital blueprint for renewable energy data science going forward.