BHL

Prymalni Dualiści
Michał Piechota, Kacper Rodziewicz, Igor Rudolf, Gaspar Sekula

Solar Money
Project Report

Table of contents

Business objective

We would like to encourage as many potential clients to invest in green energy as possible. We are aware that financial reason is the most appealing argument when it comes to green energy transformation. Our idea would empower solar panel resellers and customers by providing accurate, user-friendly tools to assess the economic benefits of solar energy. Our innovate project aims to accelerate solar adoption, position resellers as leaders in sustainable energy solutions, and ultimately contribute to the global transition to renewable energy sources. We aim to provide explainable AI tool, so that our customers clearly understands, why the investment in solar panel would (or would not) be beneficial.

Data

The dataset contains meteorological data collected from the HI-SEAS weather station during September to December 2016. It includes the following fields:

- **UNIX time**: Seconds since January 1, 1970 (for sorting).
- **Date**: In yyyy-mm-dd format.
- **Local time**: In 24-hour hh:mm:ss format.
- **Numeric data**: Includes measurements such as solar radiation, temperature, humidity, wind direction, and wind speed.
- **Text data**: Includes sunrise and sunset times in Hawaii time.

Units of measurement:

- **Solar radiation**: Watts per meter² (target variable).

- **Temperature**: Degrees Fahrenheit.
- **Humidity**: Percent.
- **Wind direction**: Degrees.
- **Wind speed**: Miles per hour.
- **Barometric pressure**: Inches of Hg.

Data preprocessing and feature engineering

Detailed information is provided in the Appendix.

Models

Throughout the modeling process, a wide range of approaches was explored to identify the most effective strategy for predicting solar radiation:

1. **Feature Selection & Dimensionality Reduction:**
   Using the MRMR algorithm, features were ranked and various top subsets (5, 10, …, 35) were tested. Additional refinement through t-tests and AIC/BIC criteria ensured that chosen predictors met the theoretical assumptions of regression models (e.g., normal distribution of residuals).
2. **Model Variants:**
   A variety of models were explored, including classical Linear Regression, LASSO, RIDGE, tree-based methods such as XGBoost, and Neural Networks. Extensive hyperparameter tuning was performed for each model type to improve predictive performance.
3. **Evaluation Metrics & Validation:**
   Models were assessed using R^2 and RMSE, balancing predictive accuracy and model complexity. The best traditional approach achieved around 0.9 R^2.
4. **AutoML:**
   AutoGluon was employed for automated model selection and optimization, resulting in approximately 0.95 R^2, surpassing manually tuned models and underscoring the potential of automated, ensemble-based approaches.

Results

Our model provides tremendous results on independent validation datasets. The metrics reach

Inspired by paper Zaigrajew, V., Baniecki, H., Tulczyjew, Ł., Wijata, A. M., Nalepa, J., Longépé, N., & Biecek, P. (n.d.). *Red Teaming Models for Hyperspectral Image Analysis Using Explainable AI*. Warsaw University of Technology, University of Warsaw, Silesian University of Technology, KP Labs, and European Space Agency, we have experimented with number of features reduction. With only 15% most relevant predictors, we still had wonderful results: R^2 lost just 0.02 points and RMSE grew just 10%. This way our model could give almost-the-same results with just a fraction of original data.

Conslusions

Our application provides a simple and easy-to-use solution for both solar panels resellers as well as single users. We encourage people to choose eco-friendly energy by providing financial proofs. Our solar panel financial results calculator empowers solar panel resellers and customers by providing accurate, user-friendly

tools to assess the economic benefits of solar energy. Designed to support informed decision-making, the app calculates potential cost savings, return on investment, and long-term financial benefits tailored to individual user profiles. By simplifying complex energy data into clear insights, it enhances trust and transparency, enabling resellers to drive customer engagement and conversion rates. This innovation aims to accelerate solar adoption, position resellers as leaders in sustainable energy solutions, and ultimately contribute to the global transition to renewable energy sources. It provides explainable AI tool, so that our customers clearly understands, why the investment in solar panel would (or would not) be beneficial.

What is more, **our solution meets EU AI Act regulations**.

Appendix

Data preprocessing and feature engineering

During the preprocessing stage, the data was grouped by **Day**, **Month**, and **Hour** to compute aggregated statistics for each unique combination of these temporal features. It is important to note that the number of unique combinations of **Day**, **Month**, and **Hour** in the dataset is limited. After grouping, the resulting dataset consists of **1312 records**.

After performing data preprocessing on the initial dataset (`SolarPrediction.csv`), the following steps were taken, and the resulting dataset now contains the following variables:

Preprocessing Steps:

1. **Date Conversion and Extraction:**
    - The `Data` column was converted into a proper date format.
    - Two new columns, `Day` (day of the month) and `Month` (month of the year), were extracted from the `Data` column to simplify temporal analysis.
2. **Hour Extraction:**
    - The `Hour` column was derived from the `Time` column, capturing the hour of the day in numeric format.
3. **Sunrise and Sunset Times:**
    - The `TimeSunRise` and `TimeSunSet` columns were converted to time objects.
    - A new variable, `SunDurationMinutes`, was created to calculate the duration (in minutes) between sunrise and sunset.
4. **Removal of Redundant Column:**
    - The `UNIXTime` column, which was not necessary for the analysis, was removed.
5. **Aggregation:**
    - The data was grouped by `Day`, `Month`, and `Hour`.
    - For each group, the following statistics were computed for different variables:

        - **Temperature**: Mean, minimum, maximum, and standard deviation.
        - **Pressure**: Mean, minimum, maximum, and standard deviation.
        - **Humidity**: Mean, minimum, maximum, and standard deviation.
        - **Wind Direction**: Mean, minimum, maximum, and standard deviation.
        - **Wind Speed**: Mean, minimum, maximum, and standard deviation.
        - **Radiation**: Mean radiation was calculated to serve as the target variable.
        - **Sun Duration**: Mean duration of sunlight within each group.
        - **First TimeSunRise and TimeSunSet**: Retained the first occurrence of sunrise and sunset times within each group for reference.

Resulting Variables in the Dataset:

1. **Temporal Features:**
    - Day: Day of the month (1–31).
    - Month: Month of the year (1–12).
    - Hour: Hour of the day (0–23).
2. **Feature's aggregation statistics as described above**
3. **Target Variable:**
    - Radiation_mean: Average solar radiation within the group, which will serve as the target variable for prediction.
4. **Sunlight Features:**
    - SunDurationMinutes_mean: Average duration of sunlight in minutes within the group.
    - TimeSunRise_first: First recorded sunrise time for the group.
    - TimeSunSet_first: First recorded sunset time for the group.

At the end of the preprocessing pipeline, the dataset was split into **training** and **test sets** to facilitate model training and evaluation. The split was performed with 80% of the data allocated to the training set and the remaining 20% allocated to the test set.

Feature engineering

As part of the data preprocessing pipeline, feature engineering was applied to derive additional informative variables and enhance the performance of the regression model predicting Radiation_mean. Below is a summary of the steps performed and their rationale:

1. Proportion of Day (Proportion_Of_Day)

- **Description**: Represents the relative position of the hour within the daylight period.
- **Formula**:
  (Hour - TimeSunRise_first) / (TimeSunSet_first - TimeSunRise_first)
- **Rationale**: Solar radiation typically changes with the position of the sun. This feature provides a normalized measure of how far the current hour is into the daylight period.

2. Sinusoidal Transformation of Hour (`Hour_Sin`, `Hour_Cos`)

- **Description**: Encodes the cyclic nature of the hour variable.
- **Formulas**:
  `Hour_Sin = sin(2 * pi * Hour / 24)`
  `Hour_Cos = cos(2 * pi * Hour / 24)`
- **Rationale**: Since the hour is a cyclic variable (e.g., 23:00 is close to 00:00), sinusoidal transformations ensure this cyclicity is represented in the model.

3. Temperature Amplitude (`Temp_Amplitude`)

- **Description**: Measures the difference between the maximum and minimum temperatures during the day.
- **Formula**:
  `Temperature_max - Temperature_min`
- **Rationale**: Large temperature variations throughout the day often indicate changes in solar radiation.

4. Temperature Ratio (`Temp_Ratio`)

- **Description**: Represents the relative average temperature compared to the maximum temperature.
- **Formula**:
  `Temperature_mean / Temperature_max`
- **Rationale**: This feature indicates how close the average temperature is to the maximum, reflecting the distribution of temperatures throughout the day.

5. Wind Energy (`Wind_Energy`)

- **Description**: Models the kinetic energy of wind based on its speed.
- **Formula**:
  `Speed_mean^2`
- **Rationale**: Wind energy is proportional to the square of wind speed and may influence cloud cover, which affects solar radiation.

6. Wind Direction Encoding (`WindDir_Sin`, `WindDir_Cos`)

- **Description**: Encodes the cyclic nature of wind direction.
- **Formulas**:
  `WindDir_Sin = sin(WindDirection_mean * pi / 180)`
  `WindDir_Cos = cos(WindDirection_mean * pi / 180)`

- **Rationale**: Wind direction is cyclic (0° and 360° are equivalent), and these transformations help the model capture this relationship.

7. Humidity Ratio (`Humidity_Ratio`)

- **Description**: Represents the ratio of average humidity to the maximum humidity during the day.
- **Formula**:
  `Humidity_mean / Humidity_max`
- **Rationale**: Highlights the variability of humidity throughout the day and its potential impact on atmospheric conditions.

8. Time to Noon (`Time_To_Noon`)

- **Description**: Measures the distance from solar noon based on sunrise and sunset times.
- **Formula**:
  `abs((Hour - TimeSunRise_first) - (TimeSunSet_first - TimeSunRise_first) / 2)`
- **Rationale**: Solar radiation is typically highest around noon, making this feature potentially predictive.

9. Wind-Humidity Interaction (`Wind_Humidity_Interaction`)

- **Description**: Models the interaction between wind speed and humidity.
- **Formula**:
  `Speed_mean * Humidity_mean`
- **Rationale**: The combination of wind and humidity may affect cloud cover and, consequently, solar radiation.

10. Temperature-Time Interaction (`Temp_Time_Interaction`)

- **Description**: Captures the interaction between average temperature and the relative time of day.
- **Formula**:
  `Temperature_mean * Proportion_Of_Day`
- **Rationale**: Reflects how temperature variations throughout the day interact with the time of day, impacting solar radiation.

After performing feature engineering and fitting an initial linear regression model, we identified two features, `TimeSunSet_first` and `Temp_Amplitude`, as linearly dependent on other predictors in the dataset. This linear dependence caused the coefficients for these variables to become undefined (`NA`) in the regression model.

To resolve this issue and improve the stability of the model, we removed these two features from the dataset. This step ensures that the model only includes independent predictors, avoiding multicollinearity and enhancing the interpretability of the results.