ML Project Progress Report

Project Colab link

Contents

Abstract
Steps Taken
Results
— EDA
— Training
— Testing

Abstract

This project addresses the critical challenge of solar power output forecasting by developing and evaluating a Long Short-Term Memory (LSTM) neural network model.

Utilising historical generation and environmental sensor data from two distinct solar plants, a comprehensive data pipeline w

Steps Taken

Data Loading and Initial Processing

- Generation and weather data for Plant 1 and Plant 2 were loaded from their respective CSV files.
- Merging plant and weather data respectively while standardising time formats
- Handling missing values

Exploratory Data Analysis (EDA) Pre-processing

- Null values and duplicate entries (based on `DATE_TIME` and `SOURCE_KEY`) were removed.
- IQR-based outlier filtering was applied to 'generating' data (where AC_POWER and IRRADIATION > 0).
- Outliers were identified but not removed as they constituted less than 20% of the generating data.
- Plotted frequency distribution graphs, correlation matrices, box plots, etc.

LSTM Model Data Preparation

- For each plant, inverter-level data was aggregated by `DATE_TIME` to create plant-level average signals.
- Data was resampled to a consistent 15-minute frequency, and missing values were filled using time-based interpolation.
- Cyclical time features (`hour_sin`, `hour_cos`) were extracted from the `DATE_TIME` index.
- Features (`AMBIENT_TEMPERATURE`, `MODULE_TEMPERATURE`, `IRRADIATION`, `hour_sin`, `hour_cos`) and the target variable (`AC_POWER`) for Plant 1 were scaled using `MinMaxScaler`.
- Time series data was transformed into sequences (windows) suitable for LSTM, using `N_LOOKBACK_STEPS` of 12 (3 hours of 15-minute intervals).
- **Train-Test Split**: The Plant 1 dataset was split into 80% training and 20% testing sets.

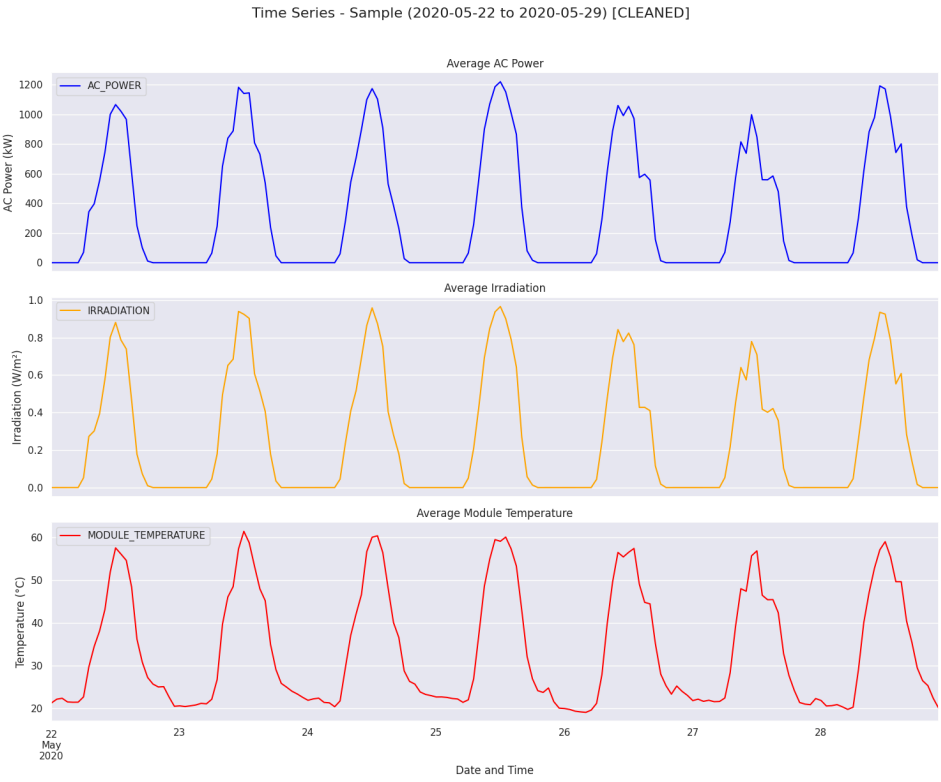
LSTM Model Building, Training, and Evaluation

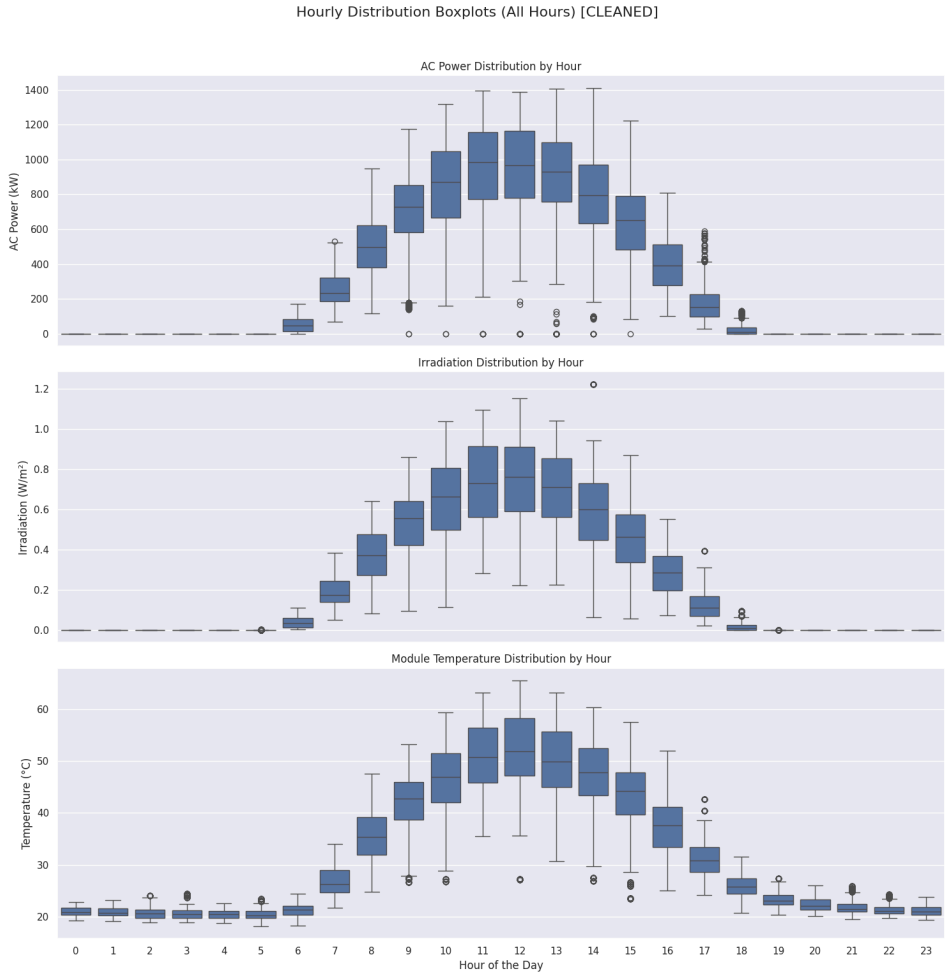
- A Sequential LSTM model was built, consisting of:
  - LSTM layer (50 units, 'relu' activation)
  - Dense output layer (1 unit - correlating to target variable)

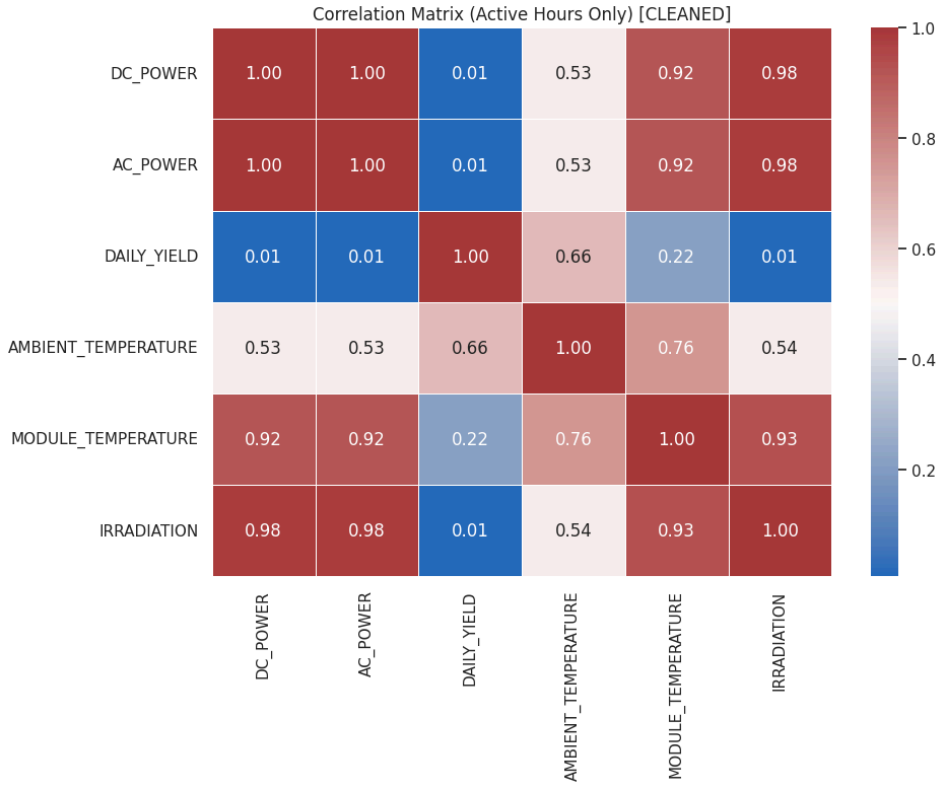
- The model was compiled using the Adam optimizer with a learning rate of 0.001 and `mean_squared_error` as the loss function.
- The LSTM model was trained on the Plant 1 training data for 50 epochs, with the Plant 1 test set used for validation.
- The trained model was evaluated on the Plant 1 test set, and the Root Mean Squared Error (RMSE) was calculated and displayed.
- Plant 2 data was loaded, preprocessed, and scaled using the same scalers fitted on Plant 1. The model's performance on this external dataset was then evaluated, and its RMSE was calculated and displayed.

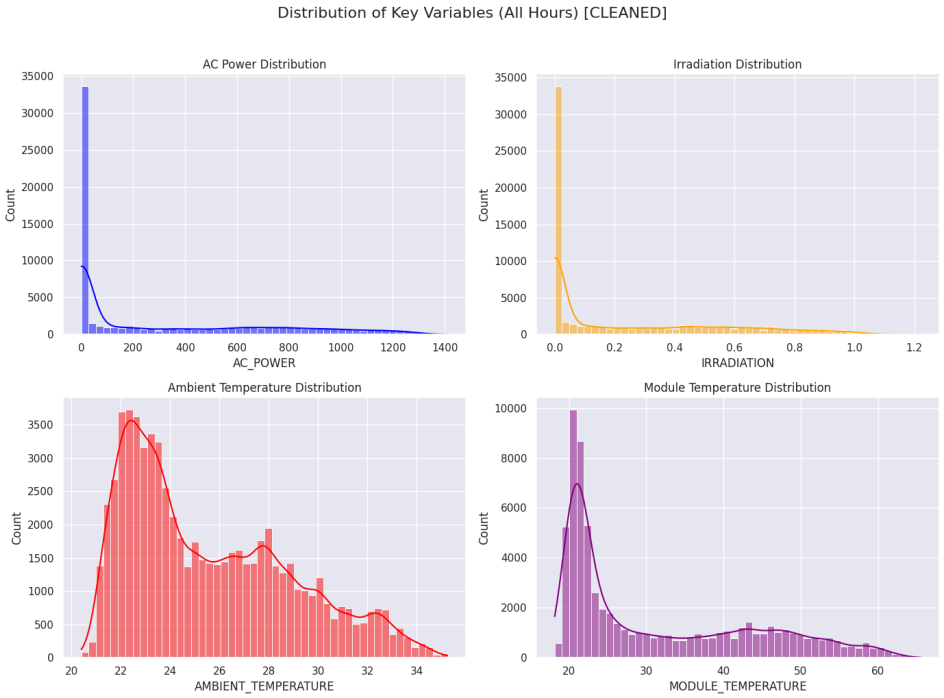
Results

EDA




Hourly Distribution Boxplots (All Hours) [CLEANED]
AC Power Distribution by Hour
AC Power (kW)
0
200
400
600
800
1000
1200
1400
Irradiation Distribution by Hour
Irradiation (W/m²)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Module Temperature Distribution by Hour
Temperature (°C)
20
30
40
50
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour of the Day

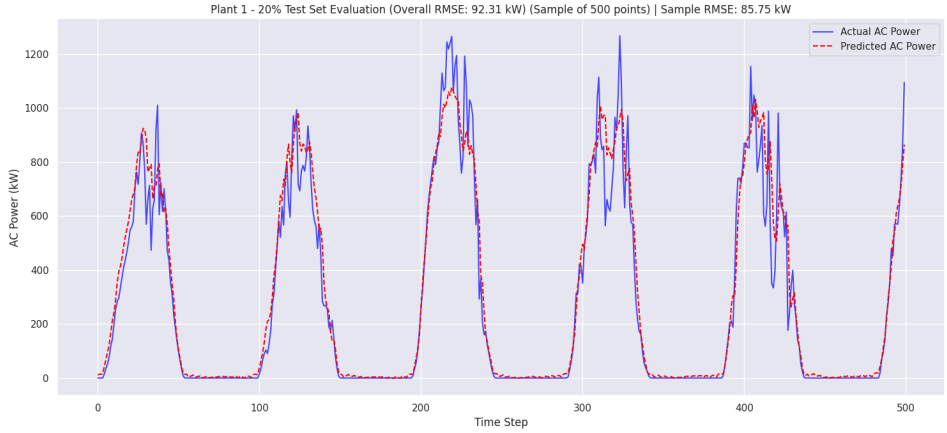

Correlation Matrix (Active Hours Only) [CLEANED]
DC_POWER
AC_POWER
DAILY_YIELD
AMBIENT_TEMPERATURE
MODULE_TEMPERATURE
IRRADIATION
1.00 1.00 0.01 0.53 0.92 0.98
1.00 1.00 0.01 0.53 0.92 0.98
0.01 0.01 1.00 0.66 0.22 0.01
0.53 0.53 0.66 1.00 0.76 0.54
0.92 0.92 0.22 0.76 1.00 0.93
0.98 0.98 0.01 0.54 0.93 1.00
1.0
0.8
0.6
0.4
0.2

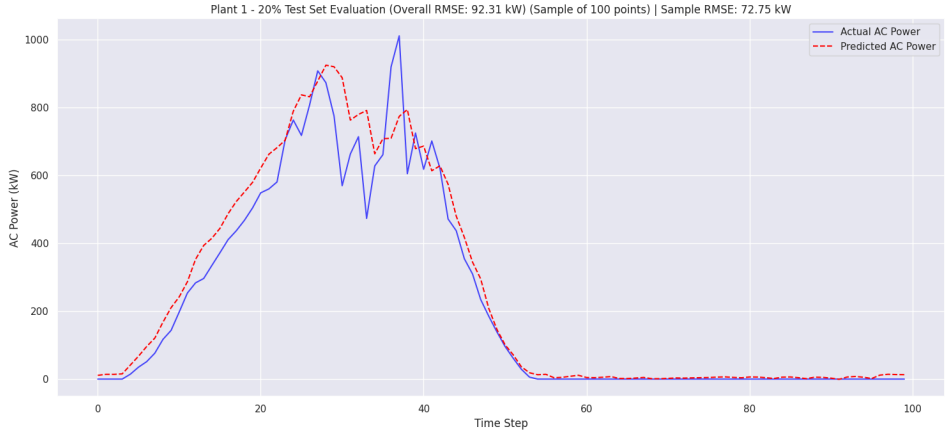

Distribution of Key Variables (All Hours) [CLEANED]
AC Power Distribution
AC_POWER
Irradiation Distribution
IRRADIATION
Ambient Temperature Distribution
AMBIENT_TEMPERATURE
Module Temperature Distribution
MODULE_TEMPERATURE
Count

Training


Model Training & Validation Loss
Training Loss
Validation Loss
Loss (MSE)
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0
10
20
30
40
50
Epoch

Testing


Plant 1 - 20% Test Set Evaluation (Overall RMSE: 92.31 kW) (Sample of 500 points) | Sample RMSE: 85.75 kW
Actual AC Power
Predicted AC Power
AC Power (kW)
1200
1000
800
600
400
200
0
0
100
200
300
400
500
Time Step


Plant 1 - 20% Test Set Evaluation (Overall RMSE: 92.31 kW) (Sample of 100 points) | Sample RMSE: 72.75 kW
Actual AC Power
Predicted AC Power
AC Power (kW)
1000
800
600
400
200
0
0
20
40
60
80
100
Time Step