Dialog Data Science Academy


Capstone Project Report

# MACHINE LEARNING MODEL DEVELOPMENT FOR PREDICTION AND FORECASTING OF SOLAR POWER GENERATION USING WEATHER DATA

by

Pasindu Liyanage

November 2021

# TABLE OF CONTENT



# LIST OF FIGURES

# 01. INTRODUCTION

## 1.1 Solar Power Generation

Solar power is a form of energy harnessed from the power and heat of the sun's rays which makes it one of the most important renewable sources of energy. Solar power generation can be defined as the conversion of energy from sunlight into electricity, either directly using photovoltaics (PV), indirectly using concentrated solar power, or a combination of both.

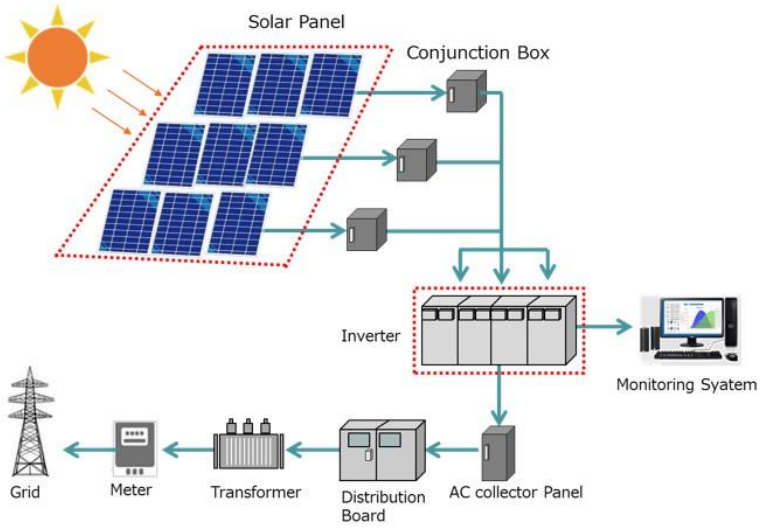


*Figure 1: Solar Power Generation Process*

At the present, solar power generation is becoming increasingly popular all around the world due the replenishing nature of non-renewable sources of power generation and hence countries are heavily investing in establishment of solar power plants to meet the energy demands of the future world.

## 1.2 Challenges with Solar Power Generation

The major challenge with solar power generation is the unpredictable nature of its output. Since the solar power is directly dependent on the power and heat of solar radiation, the production output could vary drastically depending on the weather conditions.

Hence, it has become an daunting task for authorities to effectively plan the demand supply models with solar and other power generation sources without credible information on solar power generation forecasting

## 1.3 Scope of the Project and Objectives

The main scope of this project is focused on the development of a Machine Learning algorithm to effectively predict the solar power generation with the usage of weather forecast information.

Key Objective of the Project

Development of a Machine Learning Model to accurately forecast the solar power generation with the usage of weather data to support the authorities in effectively planning the supply demand models.

# 02. DATASET

For the purpose of Model development, an open-source dataset was used from Kaggle.com

Link - https://www.kaggle.com/anikannal/solar-power-generation-data

The dataset has been gathered at two solar power plants in India over a 34-day period which consists of two pairs of files - each pair has one power generation dataset and one sensor readings dataset.

The power generation datasets are gathered at the inverter level where each inverter has multiple lines of solar panels attached to it. The sensor data is gathered at a plant level with a single array of sensors optimally placed at the plant.

Data Samples from Plant 01

| | DATE_TIME | PLANT_ID | SOURCE_KEY | DC_POWER | AC_POWER | DAILY_YIELD | TOTAL_YIELD |
|---|---|---|---|---|---|---|---|
| 0 | 15-05-2020 00:00 | 4135001 | 1BY6WEcLGh8j5v7 | 0.0 | 0.0 | 0.0 | 6259559.0 |
| 1 | 15-05-2020 00:00 | 4135001 | 1IF53ai7Xc0U56Y | 0.0 | 0.0 | 0.0 | 6183645.0 |
| 2 | 15-05-2020 00:00 | 4135001 | 3PZuoBAID5Wc2HD | 0.0 | 0.0 | 0.0 | 6987759.0 |
| 3 | 15-05-2020 00:00 | 4135001 | 7JYdWkrLSPkdwr4 | 0.0 | 0.0 | 0.0 | 7602960.0 |
| 4 | 15-05-2020 00:00 | 4135001 | McdE0feGgRqW7Ca | 0.0 | 0.0 | 0.0 | 7158964.0 |

*Figure 2: Solar Power Generation Data for Plant 01*

| | DATE_TIME | PLANT_ID | SOURCE_KEY | AMBIENT_TEMPERATURE | MODULE_TEMPERATURE | IRRADIATION |
|---|---|---|---|---|---|---|
| 0 | 2020-05-15 00:00:00 | 4135001 | HmiyD2TTLFNqkNe | 25.184316 | 22.857507 | 0.0 |
| 1 | 2020-05-15 00:15:00 | 4135001 | HmiyD2TTLFNqkNe | 25.084589 | 22.761668 | 0.0 |
| 2 | 2020-05-15 00:30:00 | 4135001 | HmiyD2TTLFNqkNe | 24.935753 | 22.592306 | 0.0 |
| 3 | 2020-05-15 00:45:00 | 4135001 | HmiyD2TTLFNqkNe | 24.846130 | 22.360852 | 0.0 |
| 4 | 2020-05-15 01:00:00 | 4135001 | HmiyD2TTLFNqkNe | 24.621525 | 22.165423 | 0.0 |

*Figure 3: Weather Data for Plant 01*

Information Related to Datasets

Solar Power Generation Data

```
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   DATE_TIME    68778 non-null  object
 1   PLANT_ID     68778 non-null  int64
 2   SOURCE_KEY   68778 non-null  object
 3   DC_POWER     68778 non-null  float64
 4   AC_POWER     68778 non-null  float64
 5   DAILY_YIELD  68778 non-null  float64
 6   TOTAL_YIELD  68778 non-null  float64
```

Weather Data

```
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   DATE_TIME           3182 non-null   object
 1   PLANT_ID            3182 non-null   int64
 2   SOURCE_KEY          3182 non-null   object
 3   AMBIENT_TEMPERATURE 3182 non-null   float64
 4   MODULE_TEMPERATURE  3182 non-null   float64
 5   IRRADIATION         3182 non-null   float64
```

# 03. METHODOLOGY

Solar prediction model development was broken down into 5 main stages including

1. Data pre-processing
2. Exploratory data analysis
3. Model Development
4. Model Evaluation
5. Model Saving and Testing

## 3.1. Data Pre-Processing

Data pre-processing is one of the most important parts of the machine learning model development which mainly focuses on preparing the data as per the requirements of the model development.

Below data pre-processing tasks were carried out with regards to solar prediction model development.

1. Handling null occurrences
2. Dropping unnecessary columns
3. Formatting timestamp as a datetime object
4. Creating separate additional columns for date and time
5. Replacing source keys with simple source IDs

## 3.2. Exploratory Data Analysis

Second stage of the model development was focused on exploratory data analysis where different parameters were evaluated numerically and graphically to better understand the variations and correlations.

Some of the exploratory data analysis evaluated includes,

1. Solar Power Generation based analysis
    a. DC and AC Power generation variation analysis
    b. Daily yield variation analysis
    c. Individual day-based power generation variation analysis
    d. Comparison of the output of different solar panels
2. Weather data-based analysis
    a. Ambient temperature variation analysis
    b. Module temperature variation analysis
    c. Irradiation variation analysis
3. Correlation analysis with solar power generation and weather parameters

## 3.3. Model Development

With the results and observation from exploratory data analysis, it was identified that there is a linear relationship between the AC output yield and some of the weather data. Hence it was decided to use '*Linear Regression'* to carry out the model development.

Model Development Configuration

| | |
|---|---|
| Type of Model | Linear Regression model from scikit learn |
| Independent Variables [X] | Irradiation & Ambient Temperature<br>Since these two are parameters have the highest correlation with AC power and these two can be captured from weather forecast |
| Dependent Variable [Y] | AC Power<br>Since AC Power is the measured output from the solar plant |
| Train Test Split | Train Data:Test Data – 80:20 |

## 3.4. Model Evaluation

Evaluation of the accuracy of the model was done using Mean Squared Error, Root Mean Squared Error and Coefficient of Determination Values

## 3.5. Model Saving and Testing

Finalized model is saved using JobLib library and tested with a sample data frame to check if the desired output is achieved.

# 04. RESULTS

Linear regression model development for solar power generation was successfully completed after several rounds of fine tuning the model.

Results of the finalized model is shown below.

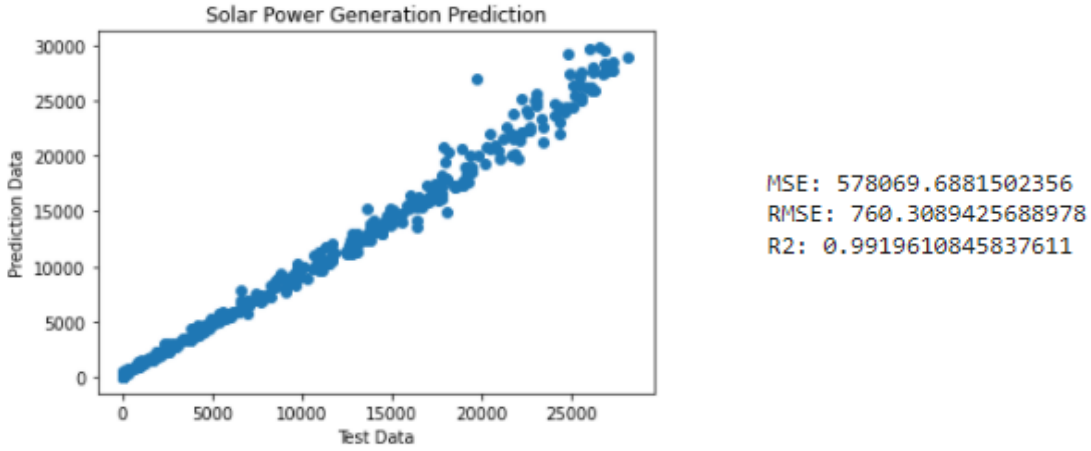


*Figure 4: Test data vs predicted data for AC power generation*

The final model was able to achieve a very high accuracy of over 99% with regards to the $R^2$ Value.

# 05. DISCUSSION

The solar power generation prediction model development was carried out using a set of data collected for 34 days at a solar power generation plant in India. The data set was available for two different solar power plants with solar power generation data and weather data separately. The usability of data set was at a very optimum level with no null occurrences in any of the datasets.

At the initial stage of the model development, different data parameters were evaluated with respect to time and other factors to understand the operation of the power plant.

Some of the key observations made during the data analysis are as follows,

1. For both plants 01 and 02, the inverters produce power from 5:30 to 18:30 since the sunlight is available only during that time period.
2. Power Generation graphs for individual days are very similar to each other despite some fluctuations between 11:00 to 14:00.
3. As per the output of the correlation analysis
    a. DC Power and AC Power are perfectly linearly related
    b. DC Power and AC Power generated has a direct relationship with Irradiation
    c. Module temperature is increasing proportional to the ambient temperature and Irradiation

The main objective of the model development is to predict and forecast the solar power generation for the next few days.

Even though time series forecast is an option to predict the solar generation, due to the unreliability of weather, the prediction will be more accurate if its based-on weather forecast data.

As per the observations made during the analysis, it was evident that the AC Power will be directly dependent on irradiation, ambient temperature, and module temperature. However, since the predictions have to be done on weather forecasting module temperature will not be available for the predictions. Hence the final model prediction was done using the irradiation and ambient temperature as input parameters.

# 06. CONCLUSION

After several rounds of testing and fine tuning, a machine learning model was successfully developed using linear regression algorithms for the prediction of solar power generation using weather forecast information.

Model will be tested with Sri Lankan data in the coming days and will be integrated with a solar energy prediction application be used for Sri Lankan energy demand planning operations.