Solar Production Analysis and Forecasting

Santiago Martín Serrano

TABLE OF CONTENTS

1. INTRODUCTION
    1.1. Problem and Motivation.
    1.2. Characteristics of the problem.
    1.3. Strategy and Data (3 csv files).

2. ANALYSIS AND MODEL BUILDING
    2.1. Preprocessing and merge.
    2.2. Model Building.
    2.3. Graphics and Results

3. DIFFICULTIES AND LIMITATIONS
4. CONCLUSION

1.1. Solar Production Forecast. Why it's important?

-> Operative costs: Electricity operators (**RTE** in France, or **Red Eléctrica** in Spain) must maintain a **balance** between **generation** and **consume** each second.

- If the model fails by 500 MW at 12am, the operator has to activate a cyclic-combined central (gas) of emergency. **More expensive than programmed energy**. A good forecast could save millions in operative costs.

-> Reduction of Fines, for plant owners: One owner could say: I will produce 100 MW tomorrow at 12am. But, if it only obtains 80 MW, there is a fine to pay, due to energy deviation.

-> Solar panels in houses: A good prediction helps save money in electricity expenses…
- Help finding the best time to sell your excess of energy to the grid.
- A smart house could advise you to use more electricity at a certain hour.

1.2. Characteristics of the problem.

-> Meteo forecasting vs Solar forecasting. Meteo is a lot more complex**.** Differential equations of fluid dynamics, in supercomputers to predict pression, temperature…

-> Restricted-space regression problem. The production will never be negative, or higher than clear_sky.

-> Solar is a post-processing problem. <u>We rely on the meteo data</u>. If it fails, we fail.

-> Double cyclicality. "Loops" each day (24 h) and each year (seasons).

-> We will use **solar irradiance** variables, **meteo** variables, and **past solar productions**, to predict future solar productions (1 day).

- **Nowcasting**, complex mathematical predictions from the variables.
- **Short term forecasting** with satellite images.
- **+6h**, output from numerical data (what we are doing).

1.3. Strategy and Data.

GOAL: Predict solar production in France, 2 days in advance.

We will use **3 csv** files. From 30/01/**2022** to 31/12/**2024**. We have **1 hour per row** in every dataset.

-> **First dataset**: Past solar productions in france. (1h/row). At night -> 0's… **Target variable**.

-> Second dataset: Solar irradiance variables from a specific spot (Nouvelle-Aquitaine). These are Measures.

-> Third dataset: Weather Predictions, by hours. These represent a prediction, maybe something different happened in real life.

1.3. Strategy and Data.

$$\overset{\text{FRANCE}}{y(t+2)} = \overset{\text{(1) FRANCE}}{\underbrace{y(t)}_{\text{SOLAR PROD.}}} + \overset{\text{(2) VILLE}}{\underbrace{ax(t)}_{\text{SOLAR MEASURES}}} + \overset{\text{(3) VILLE}}{\underbrace{cz(t+2)}_{\text{WEATHER PREDICTIONS (2 days in advance)}}}$$

Predict solar production in France 2 days in advance.

-> **First dataset**: Past solar productions in france

-> Second dataset: Solar irradiance variables.

-> Third dataset: Weather predictions, by hours.

1.3. Strategy and Data.

$$\overset{\text{FRANCE}}{y(t+2)} = \overset{\text{1 FRANCE}}{\underbrace{y(t)}_{\text{SOLAR PROD.}}} + \overset{\text{2 VILLE}}{\underbrace{ax(t)}_{\text{SOLAR MEASURES}}} + \overset{\text{3 VILLE}}{\underbrace{cz(t+2)}_{\text{WEATHER PREDICTIONS (2days in advance)}}}$$

SOURCES:

-> **1º dataset**: Datagouv (République française).

-> 2º dataset: CAMS solar radiation time-series.

-> 3º dataset: Open-meteo (Weather Forecast API)

2.1. Analysis and Model. Preprocessing.

If we took a look at the Colab code…

The 3º dataset (weather prediction variables) **is first addressed.**

Then the 2º dataset (solar irradiance)**, and it's merged with the previous.**

Then the 1º dataset (solar production in France) **is treated and merged with the previous two.**

And then: **TRAINING AND TEST SPLITTING** (the test starts given a specific date)

2.1. Analysis and Model. Preprocessing.

And then: **TRAINING AND TEST SPLITTING** (the test starts given a specific date)

Also, **the 3º dataset**, with weather predictions, has been modified.

In order to provide the model “weather forecasting” for the next two days, we needed to **shift the dataset** - 2 days.

2.1. Analysis and Model. Preprocessing.

Preprocessing operations like, renaming cols, casts to datetime or numerical types, are used… At the end, a **final merge of all 3 datasets**:

```
Current nº of rows in the dataset: 25607
                    datetime  solaire_mw  toa  clear_sky_ghi  clear_sky_bhi  \
0 2022-01-30 00:00:00+00:00         0.0  0.0            0.0            0.0
1 2022-01-30 01:00:00+00:00         0.5  0.0            0.0            0.0
2 2022-01-30 02:00:00+00:00         0.0  0.0            0.0            0.0
3 2022-01-30 03:00:00+00:00         0.0  0.0            0.0            0.0
4 2022-01-30 04:00:00+00:00         0.0  0.0            0.0            0.0

   clear_sky_dhi  clear_sky_bni  ghi  bhi  dhi  ...  shortwave_radiation  \
0            0.0            0.0  0.0  0.0  0.0  ...                  0.0
1            0.0            0.0  0.0  0.0  0.0  ...                  0.0
2            0.0            0.0  0.0  0.0  0.0  ...                  0.0
3            0.0            0.0  0.0  0.0  0.0  ...                  0.0
4            0.0            0.0  0.0  0.0  0.0  ...                  0.0

  diffuse_radiation  direct_normal_irradiance  global_tilted_irradiance  \
0               0.0                       0.0                       0.0
1               0.0                       0.0                       0.0
2               0.0                       0.0                       0.0
3               0.0                       0.0                       0.0
4               0.0                       0.0                       0.0

   dew_point_2m  surface_pressure  sunshine_duration  \
0           2.7            1015.6                0.0
1           2.9            1015.4                0.0
2           2.3            1016.0                0.0
3           2.0            1016.1                0.0
4           1.7            1016.9                0.0
```

```
[5 rows x 30 columns]
```

2.2. Model Building.

To build the model, we have tried several options… obtaining the best performance with a *XGBoost* model (a model based on Random Forest, that learns based on the errors of the previous trees).

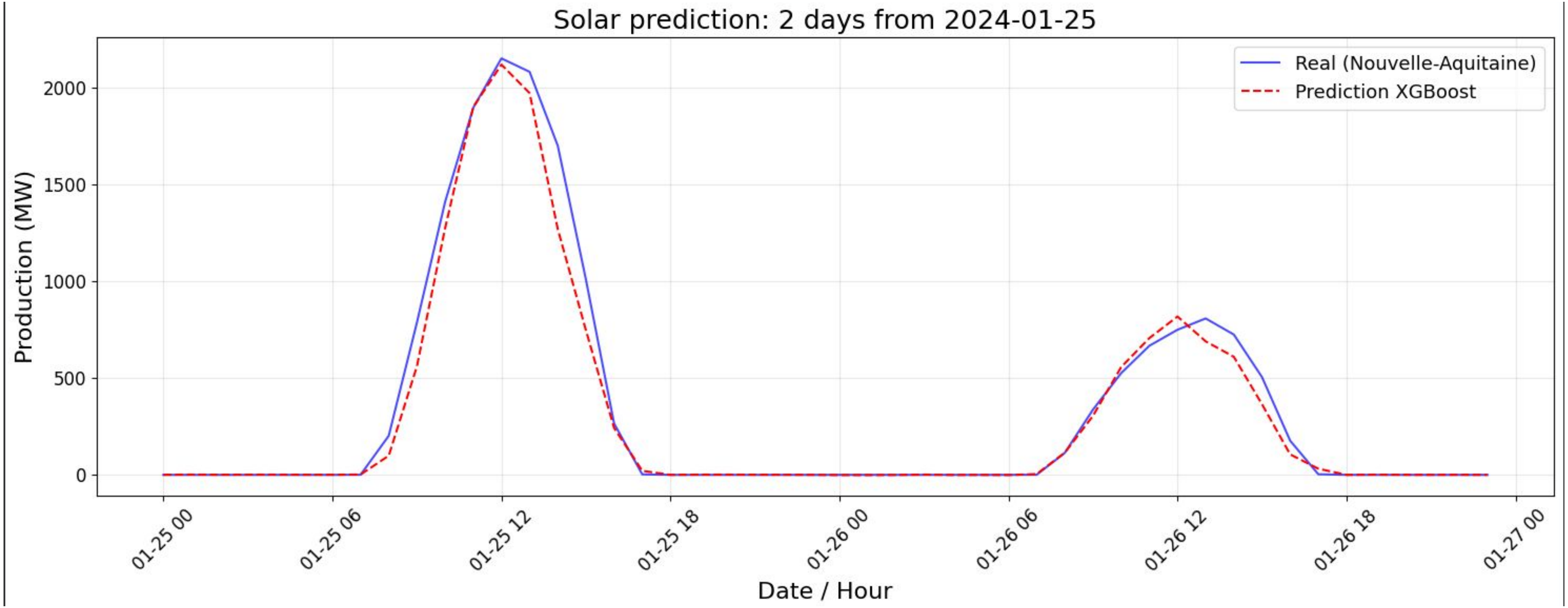
```python
# --- "ROBUST & BALANCED" CONFIGURATION, Constraints applied ---
model_xgb = xgb.XGBRegressor(
    n_estimators=1000,          # More trees but with smaller impact per tree
    learning_rate=0.03,         # Smaller steps for finer convergence
    max_depth=6,                # Lower depth to reduce overfitting/memorization
    # --- MODEL CONSTRAINTS (Regularization) ---
    colsample_bytree=0.5,       # EACH TREE can only see 50% of the features.
                                # This forces the model to learn from secondary variables!
    subsample=0.7,              # Uses only 70% of rows per iteration for diversity
    reg_alpha=10,               # Strong L1 regularization (filters out weak variables)
    reg_lambda=1,               # L2 regularization (prevents feature weights from exploding)
    # ------------------------------------------
    random_state=42
)
```

```python
model_rf = RandomForestRegress
    n_estimators=300,          #
    max_depth=20,              #
    min_samples_leaf=4,
    max_features='sqrt',
    bootstrap=True,
    random_state=42
)
```

2.3. Graphics and Results.

Real Data VS XGBoost

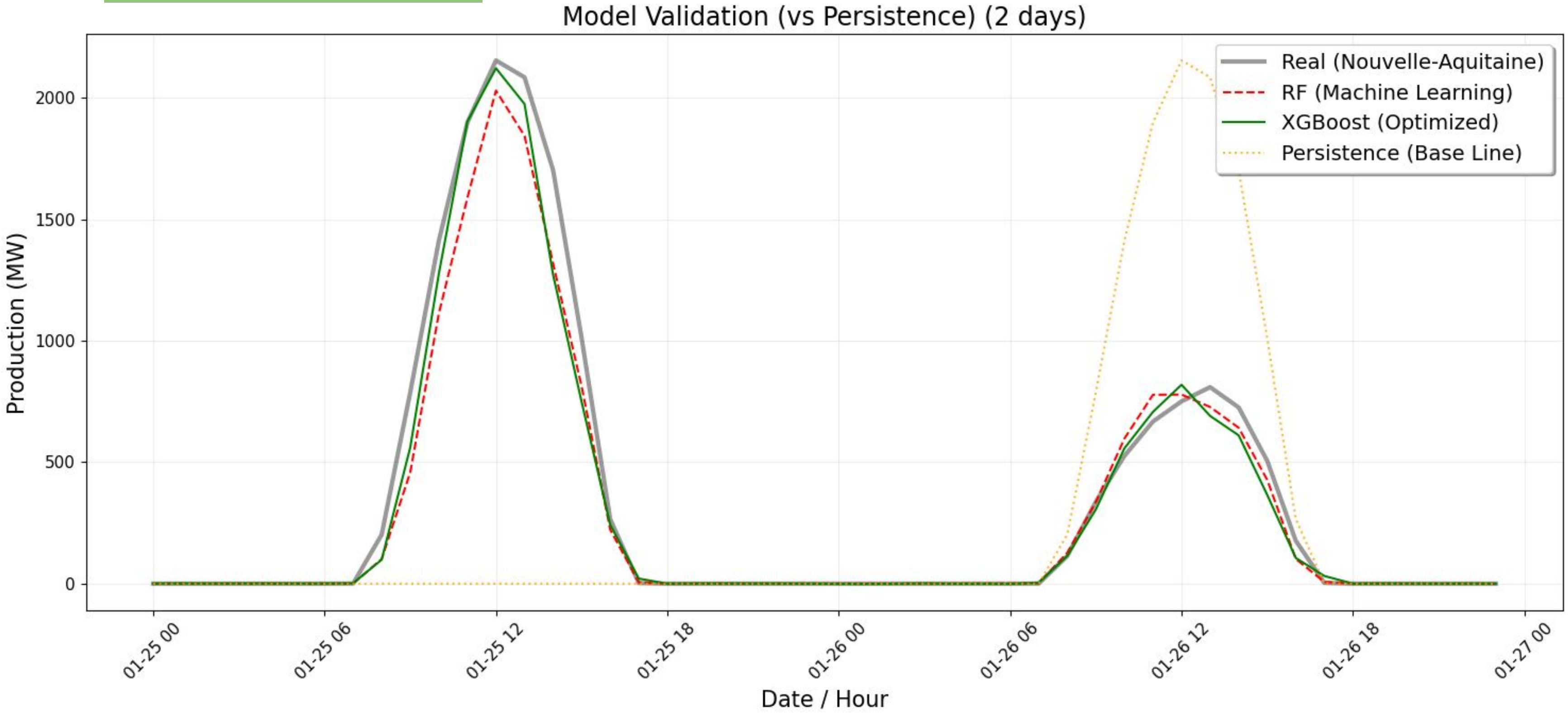
```
Metrics in the test period of (2 days):
  R2 Score: 0.9754
  MAE:      41.96 MW
  nMAE:      1.95%
```



2.3. Graphics and Results

```
Results for 2 days of test:
Metrics Random Forest   : R2=  0.9634 | MAE=   53.84 MW | nMAE=  2.50%
Metrics XGBoost         : R2=  0.9754 | MAE=   41.96 MW | nMAE=  1.95%
Metrics Persistence     : R2= -0.5913 | MAE=  383.82 MW | nMAE= 17.82%
```

Real Data VS Models



2.3. Graphics and Results

```
Results for 2 days of test:
Metrics Random Forest  : R2=  0.9634 | MAE=   53.84 MW | nMAE=  2.50%
Metrics XGBoost        : R2=  0.9754 | MAE=   41.96 MW | nMAE=  1.95%
Metrics Persistence    : R2= -0.5913 | MAE=  383.82 MW | nMAE= 17.82%
```

```
Mean metrics per day:
     datetime        MAE        R2  Max_Real
0  2024-01-25  56.567560  0.974536    2154.0
1  2024-01-26  27.361564  0.969380     809.0
```

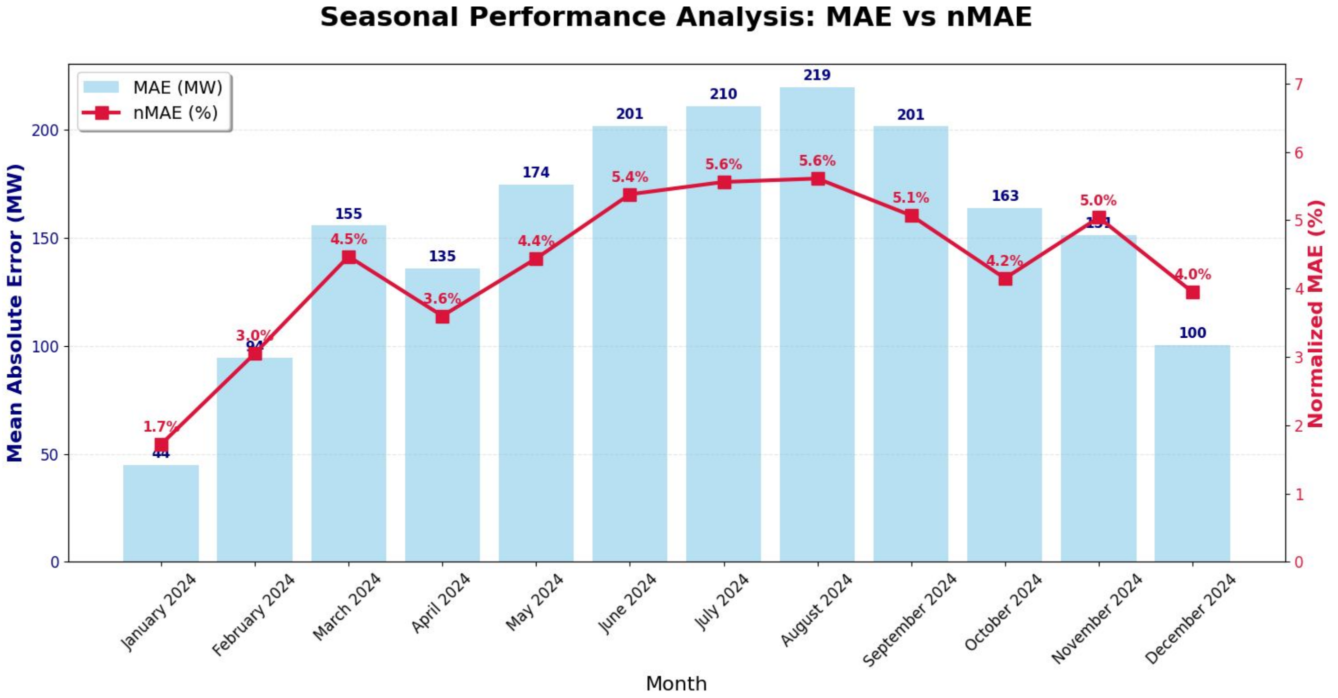
← metrics per day (separated)

```
 COMPARING PERFORMANCE BY HORIZON:
   Horizon (Days)  MAE (MW)  R2 Score  nMAE (%)
0               1     56.57    0.9745      2.63
1               7    115.97    0.8629      4.93
2              30     96.71    0.8748      3.54
```

2.3. Graphics and Results

MAE error increases as the total production increases.

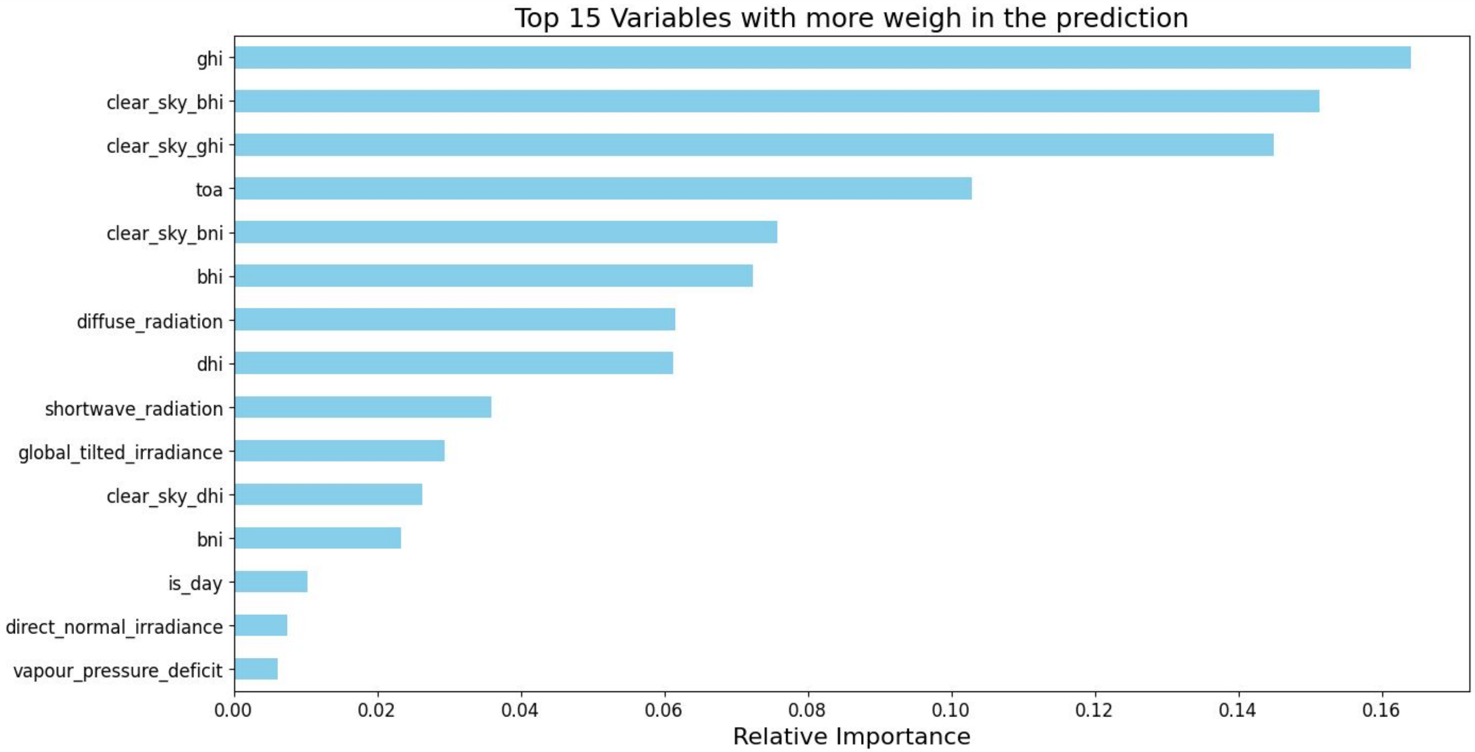
nMAE (MAE/max), gives us the real error in relationship to the amount of production (max).



2.3. Graphics and Results

RF was more balanced.

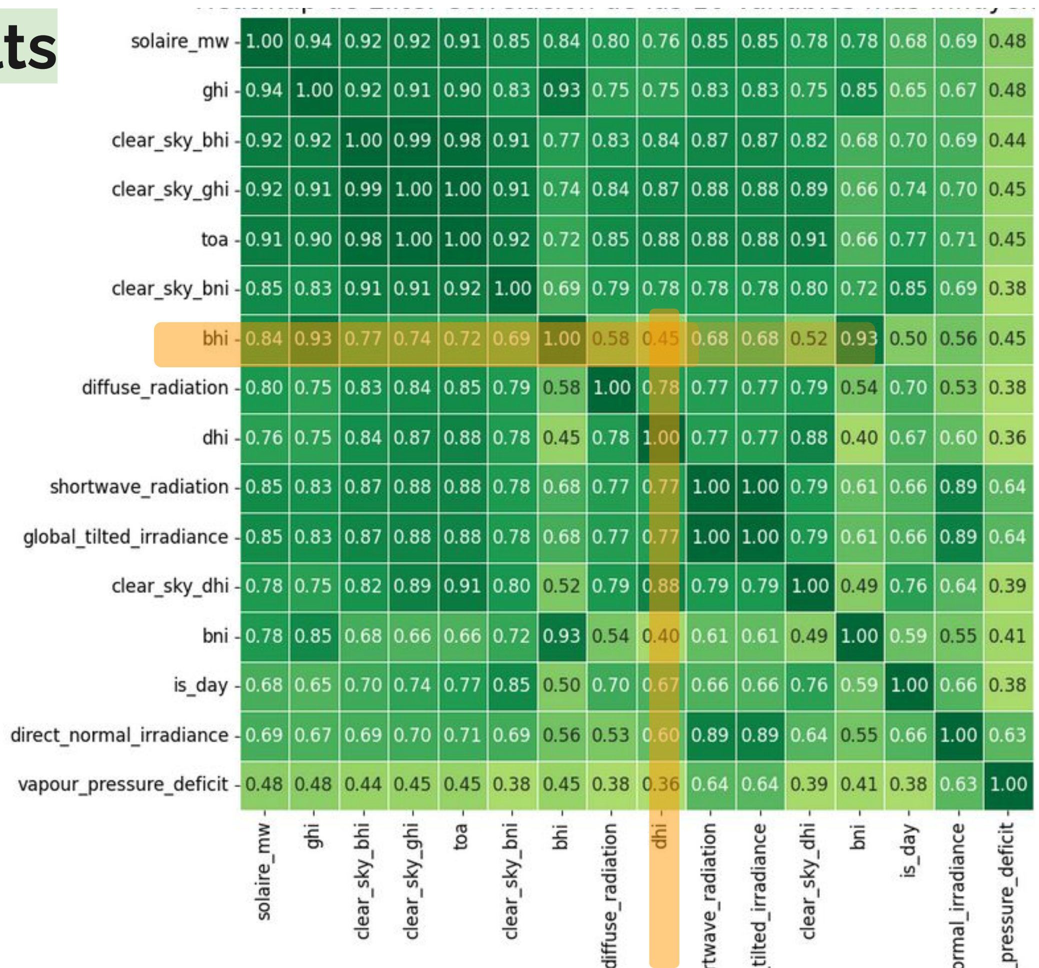
In **XGBoost, only a few variables dominated.** Not a solid model. **So, we needed to apply more regularisation.**



2.3. Graphics and Results

Heatmap of the top 15 variables

ghi clear_sky_bhi toa bhi dhi.

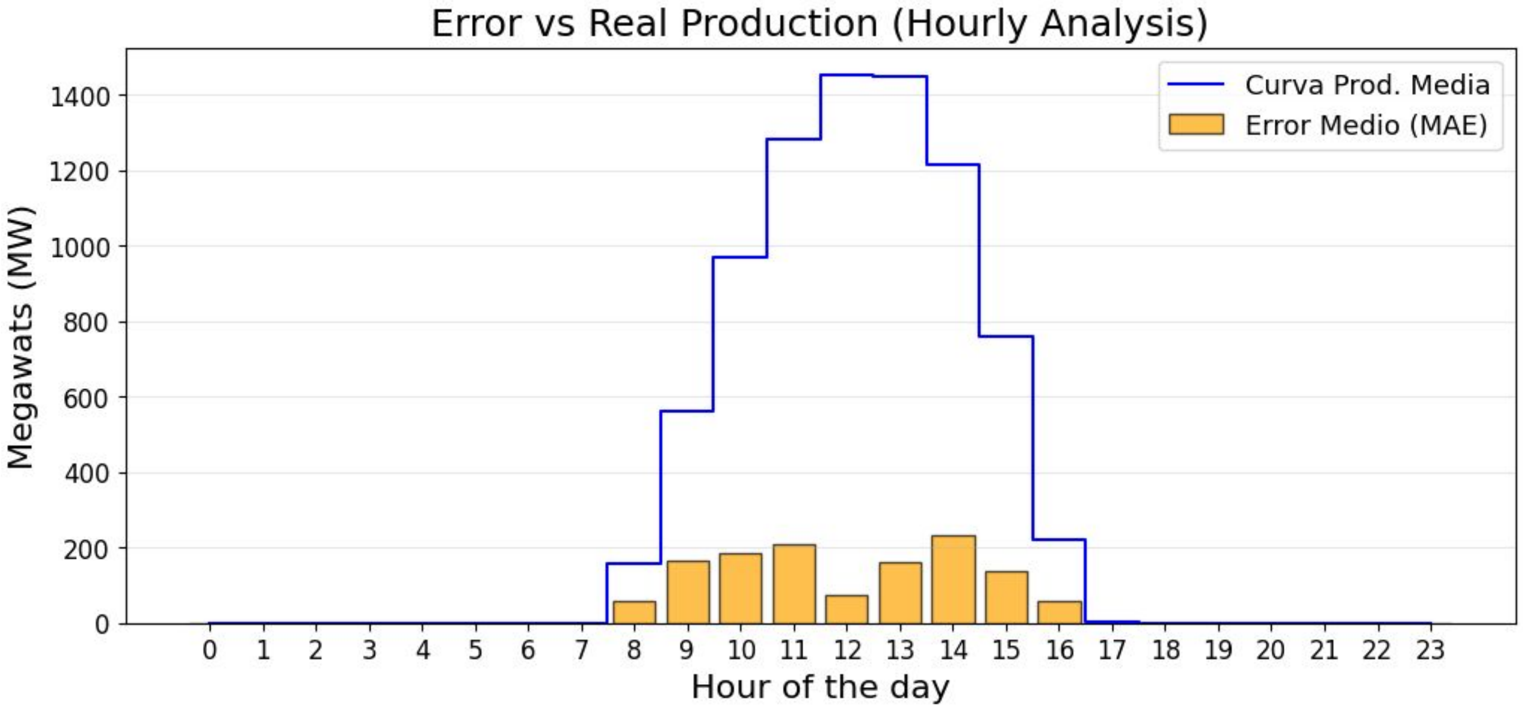


2.3. Graphics and Results

```
Metrics in the test period of (2 days):
  R2 Score: 0.9754
  MAE:      41.96 MW
  nMAE:      1.95%
```

We were obtaining good **nMAE** and **MAE** results but…
The night was a major factor.
The worst value is actually: **200MW error**

```
MAE Daylight (Only sun hours): 128.98 MW
nMAE Daylight: 5.99%
```



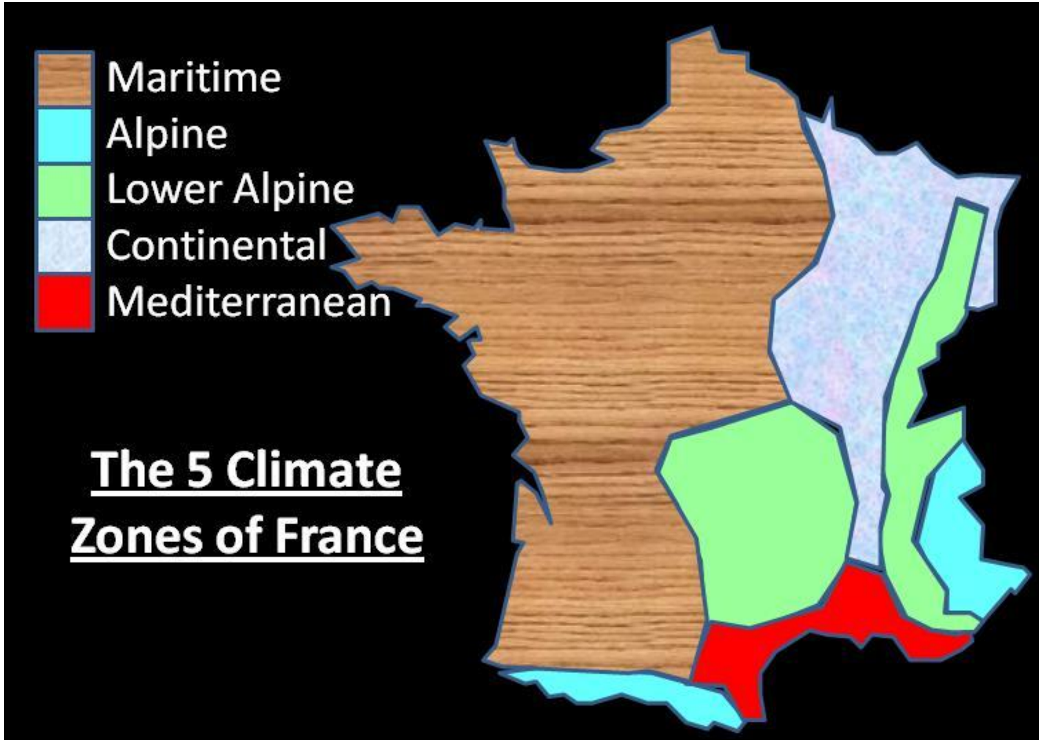
3. DIFFICULTIES & LIMITATIONS

1. **FINDING GOOD DATASETS. And stick to one plan (small size project is better than nothing)**
   a. **Limitation:** 2° and 3° datasets are only from a specific point in France.

2. **DATA LEAKAGE WHEN TESTING THE MODEL.**
   a. **Tried a lag column** for solaire_mw (1° dataset), but the model should know nothing about that!

3. **HIDDEN PROBLEMS**
   a. **Night time** made nMAE lower, and measures were inaccurate.

3. DIFFICULTIES & LIMITATIONS



4. CONCLUSION

We have provided a

```
Metrics in the test period of (2 days):
  R2 Score: 0.9754
  MAE:      41.96 MW
  nMAE:      1.95%
```

model for 2 days forecast.

→ It is not the ideal approach. However, solar production and weather variables have shown compatibilities, with the creation of a “precise” XGBoost model.

→ It is a wide range of study, and more complex models and variables are taken into account, just to predict 2-6 hours.

→ We have seen the valuable insights we can obtain from 3 datasets of solar production, solar variables and weather forecasting.

BIBLIOGRAPHY

- https://en.wikipedia.org/wiki/Solar_power_forecasting
- https://onlinelibrary.wiley.com/doi/10.1155/2022/7797488
- https://www.sciencedirect.com/science/article/pii/S2352484723011228
- https://github.com/carmenabans/Solar-energy-production-forecasting-with-ML
- https://www.sciencedirect.com/science/article/abs/pii/S0038092X12001429
- Gemini AI was used for the purpose of better understanding the texts written and code of the project.