CO_2 Factor Forecasting Using Renewable Energy Data

Data Collection and Cleaning

I collected hourly data from the Nationaal Energie Dashboard (NED) spanning 31 December 2020 to 31 December 2024, resulting in 35,064 observations for the CO_2 factor (kg CO_2/kWh) and four renewable energy features (all in kWh). After cleaning, I used this four-year dataset to better capture seasonal patterns, variability, and long-term trends compared to using just a single year.

Visualization and Descriptive Statistics

To better observe trends and patterns, I smoothed the hourly time series using a 24-hour moving average.
Figure 1 shows the CO_2 factor over time, revealing a clear downward trend which is likely due to the growing share of renewable energy replacing traditional sources.

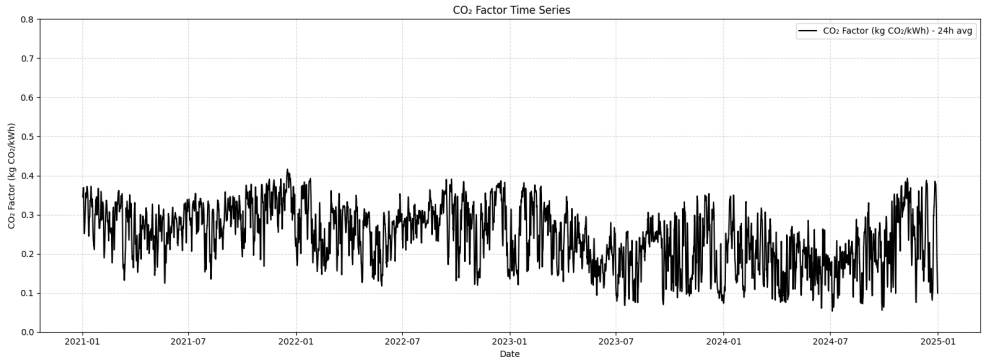

Figure 1: CO_2 factor (kg/kWh) Time Series

Figure 2 displays the energy production from the four renewable sources. Production from all sources, besides biomass, shows an overall increase over time. Seasonality in the data is evident as well.

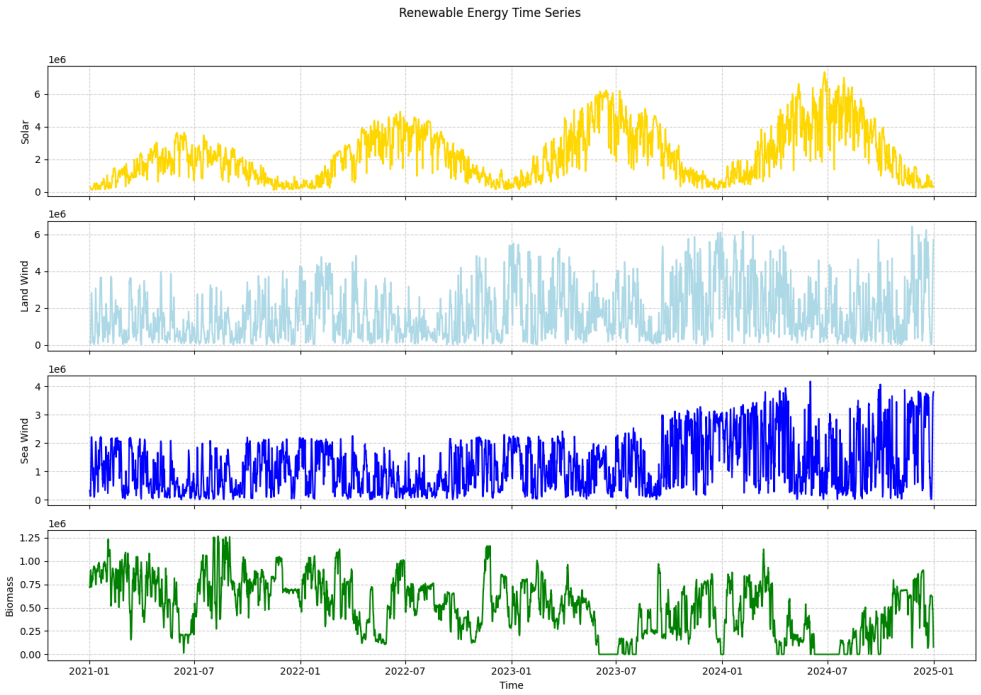

Figure 2: Renewable Sources Volume (kWh) Time Series

Figure 3 shows scatter plots, histograms, and density estimates of the variables. CO_2 factor is negatively correlated with solar, land wind, and sea wind, but positively correlated with biomass. Most features (except biomass) are right-skewed, reflecting many low or zero generation hours.

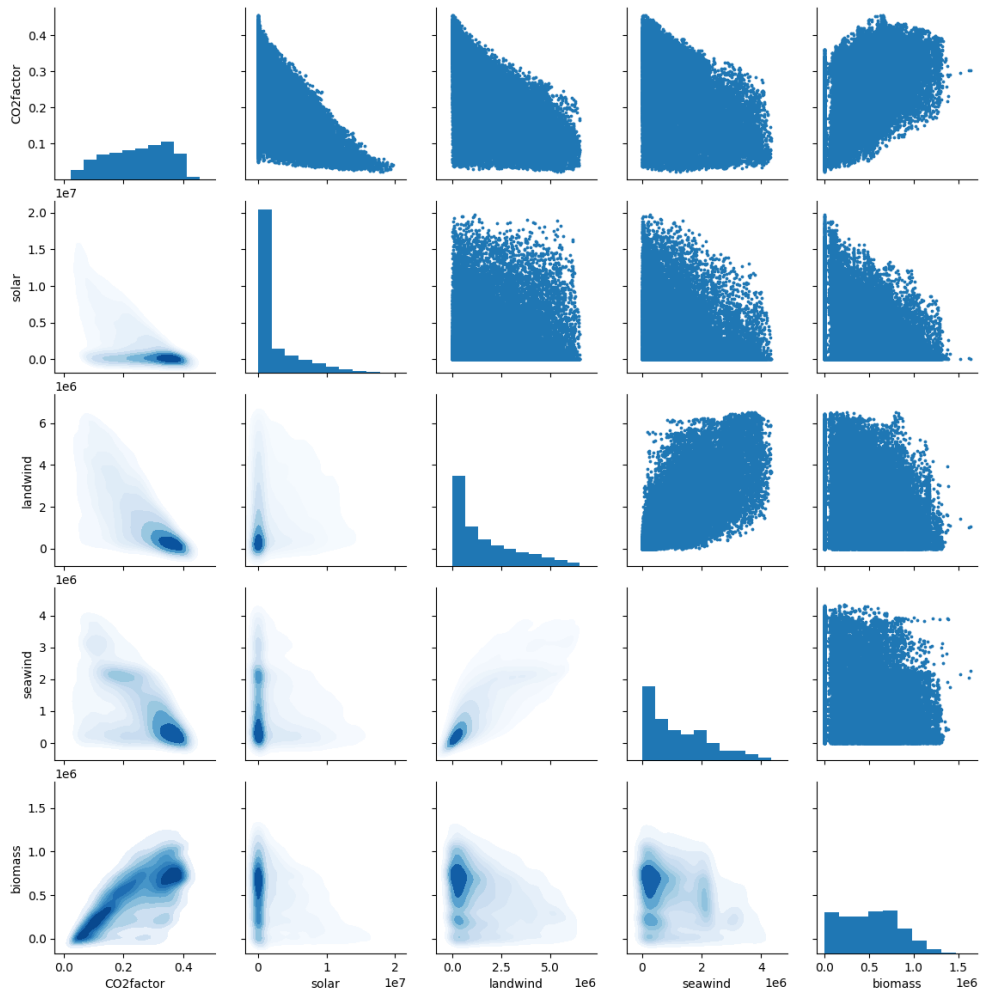

Figure 3: Relationships and Distributions of Variables

The aforementioned relationships are also shown quantitatively by the correlation matrix in Figure 4.

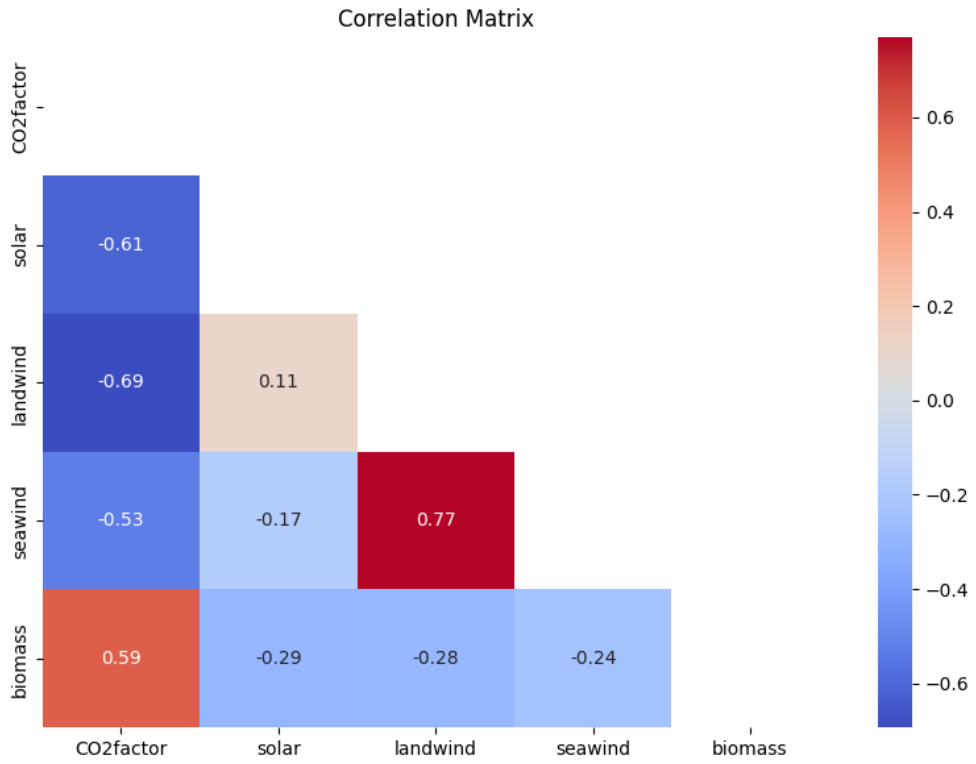

Figure 4: Correlation Matrix - Heatmap

Table 1 shows the descriptive statistics for each variable. The features are on a much larger scale than the CO_2 factor, indicating the need for scaling before modeling. Solar energy stands out with both the highest magnitude and the greatest variability among the energy sources.

Table 1: Descriptive Statistics of CO_2 Factor and Renewable Energy Features

Statistic	CO_2 Factor	Solar	Land Wind	Sea Wind	Biomass
Observations	35,064	35,064	35,064	35,064	35,064
Mean	0.24	2,202,041	1,688,767	1,206,620	507,932
Std Dev	0.10	3,532,621	1,627,761	1,009,525	316,022
Min	0.02	0	0	0	0
25%	0.16	0	328,458	332,749	241,010
50%	0.25	71,453	1,123,035	959,000	518,188
75%	0.33	3,271,873	2,702,556	1,952,250	741,465
Max	0.45	19,736,230	6,518,273	4,342,999	1,637,997

CO_2 Factor is measured in kg/kWh, Renewable Sources of energy are measured in kWh).

Defining Y and Feature Engineering

For multi-step forecasting, I use a multi-output regression approach, where the 168 future values of the CO_2 factor (from hour t to t+167) serve as targets. This results in 168 independent models—one for each forecast step. The feature matrix X, containing the four main features (solar, land wind, sea wind, and biomass), is standardized to have zero mean and unit variance.

To capture temporal effects observed earlier, I introduce hourly, weekly, and seasonal indicators. Recognizing the cyclical nature of time, I encode hours using sine and cosine functions instead of binary indicators. While indicators work better for multiple regression models, the sine and cosine functions work better for the machine learning models. This is expected since the machine learning models are tree based and it is preferable to have lower dimensions of X for optimal performance (this is linked to feature importance).

Models

I use three different modeling approach to derive insights. I begin with multiple linear regression because it is easy to interpret. It gives a ba-

sic idea of the data patterns but by construction fails to capture the data nonlinearities. To address this, I adopt more flexible machine learning models: Random Forest (RF) and Extreme Gradient Boosting (XGBoost). Several variants of each of the models are tested (See Python code), but only the best-performing version of each is reported in Table 2. Models are trained without data shuffling to respect the time series structure.

Table 2: Average Performance Metrics for Forecasting Models

Model	R^2	**RMSE**	**MAE**
Multiple Linear Regression	0.3495	0.0887	0.0742
Random Forest	0.3727	0.0868	0.0710
Extreme Gradient Boosting	0.3909	0.0856	0.0710

The performance metrics are averaged over the 168 forecasted steps. R^2 is the average R^2 over all forecasted steps. RMSE is the average Root Mean Squared Error over all forecasted steps. MAE is the average Mean Absolute Error over all forecasted steps.

Linear regression works better with binary time features. Nonlinear models work better when we use the cyclical nature of seasons, days and hours. This is expected since the nonlinear models are tree based and it is preferable to have lower dimensions of X for optimal performance.

I present now the relative graphs of the best performing model (XGBoost):

Figure 5 shows the Predicted vs Actual forecasted values at each time step. It is evident that the model gets worse at later period forecasts as they get harder and harder to predict.

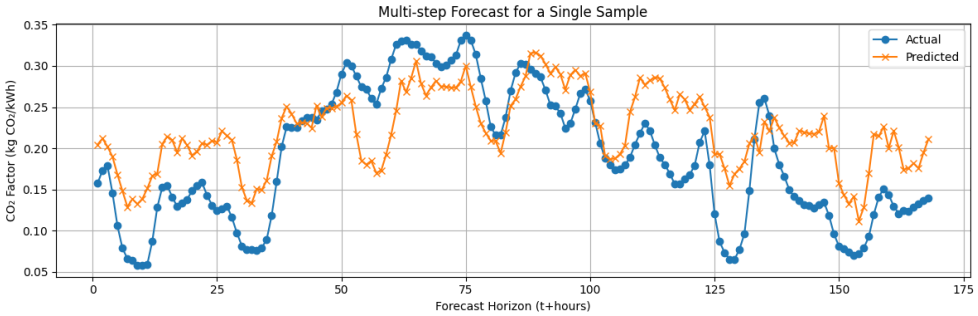

Figure 5: Predicted CO_2 factor vs Actual CO_2 factor

Figure 6 shows the RMSE per forecasting step. It confirms that the model gets worse over time as the RMSE starts low but increases steadily over time.

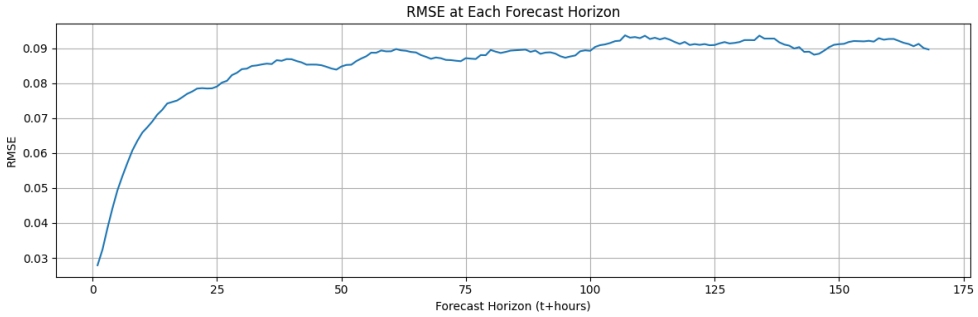

Figure 6: RMSE for each forecast step

Figure 7 shows the feature importance of XGBoost model. Hours have the most critical role in predicting CO_2 factor. From the renewable sources features, solar energy is the most important in predicting CO_2 factor.

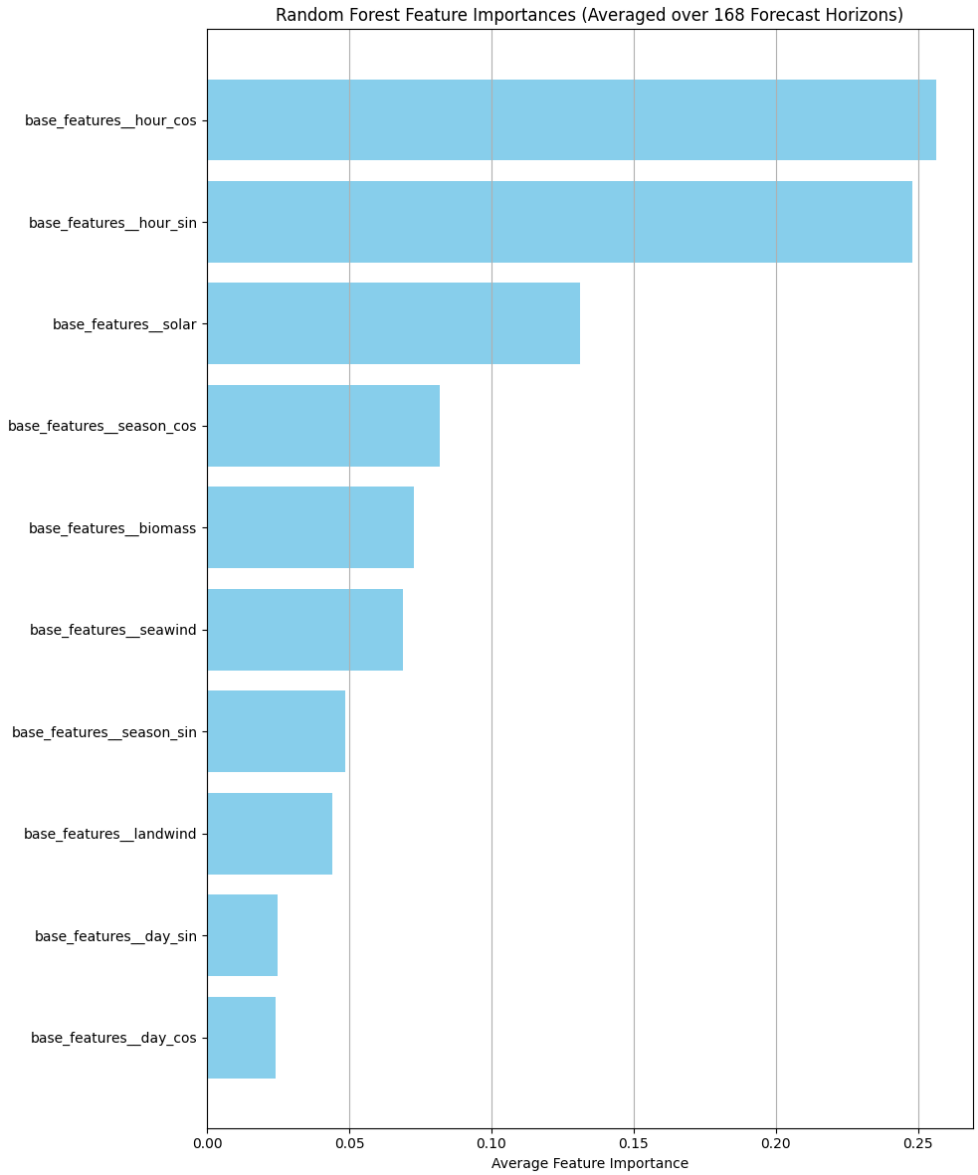

Figure 7: Predicted CO_2 factor vs Actual CO_2 factor