NEOMA
BUSINESS SCHOOL

Machine Learning and Artificial Intelligence : Group Project

MSC FBD

ENS Challenge Data 2023 : How can we explain the price of electricity ?

HADDAD Fayssal / AIT BRAHIM Hamza / GURSOY Onur

31 March 2023

Table des matières

1 Introduction of the Subject

In this study, we investigate the factors influencing daily electricity prices in France and Germany, focusing on the impact of meteorological, energy production, and commercial data. Our objective is to develop a model that accurately estimates the daily variation in electricity futures prices using these explanatory variables. Electricity futures are financial instruments that estimate the value of electricity at a specified future date based on current market conditions. We focus on short-term futures with a maturity of 24 hours. The dataset provided includes daily meteorological measurements (temperature, precipitation, and wind), energy production data (various primary energy sources), and electricity usage data (consumption, import/export, and exchanges between the two countries). The evaluation metric used is the Spearman correlation between the model's predictions and the actual futures price variations in the test dataset.

2 Data explanation

We will examine a comprehensive dataset consisting of several variables related to weather, energy production, and electricity usage in France and Germany. Meteorological variables include daily temperature, precipitation, and wind speed, which are known to affect both electricity production and demand. Energy production variables encompass daily price variations of primary energy sources such as natural gas, coal, and carbon emissions futures, reflecting the market dynamics of energy commodities. Additionally, we will study data on different types of energy production within each country, including natural gas, coal, hydropower, nuclear, solar, wind, and lignite. Electricity usage variables consist of total electricity consumption, residual load (electricity consumed after renewable energy sources are used), net imports and exports of electricity to and from Europe, and electricity exchanges between France and Germany. Analyzing these variables can provide insights into the complex relationships between weather, energy production, and electricity demand, ultimately helping us to better understand the factors that drive daily electricity prices in these two European countries.

Because there are many variables in the exercise, we may need to analyze the correlation matrix between each variable. The goal of studying a correlation matrix is to identify potential relationships between the features and the target variable as well as the relationships among the features themselves. A correlation matrix provides a visual representation of the linear relationships between pairs of variables, with correlation coefficients ranging from -1 to 1.

A correlation close to 1 or -1 indicates a strong positive or negative relationship, respectively, while a correlation close to 0 suggests a weak or no linear relationship. By examining the correlation matrix, we can gain insights into the relationships between variables and identify features that may have a significant impact on the target variable.

Additionally, the correlation matrix can help detect multicollinearity, which occurs when two or more features are highly correlated. Multicollinearity can lead to unstable estimates in regression models and make it difficult to interpret the importance of individual features. Identifying and addressing multicollinearity can improve the performance and interpretability of the model.

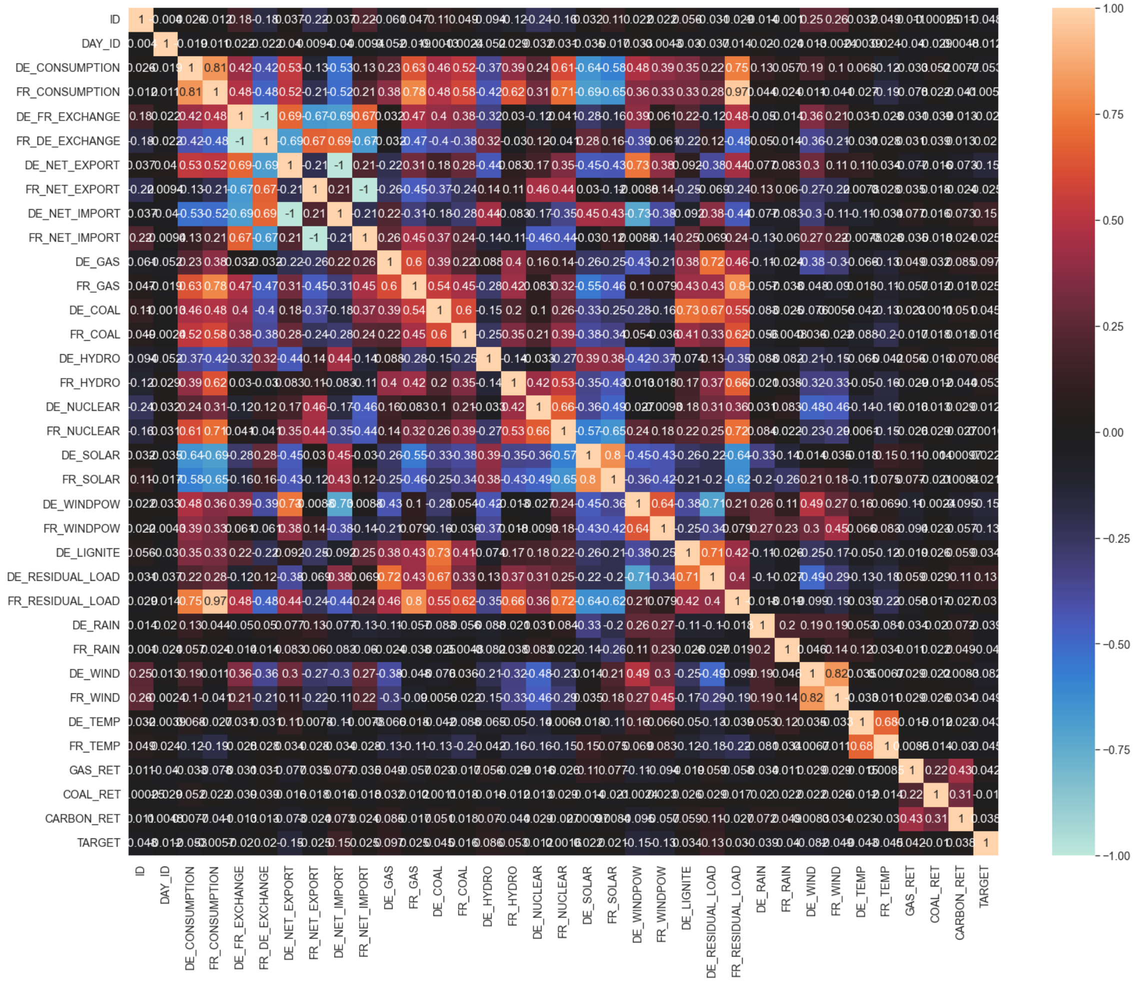

FIGURE 1 – Correlation Matrix : we can see strong multi-correlation between the features

Thanks to the correlation matrix, we observe that some variables are highly correlated between each others. It confirms that a simple linear regression analysis is not enough to catch all the informations that we get from our dataset to explain the variation of electricity futures contracts price.

3 Spearman Correlation

The Spearman correlation, also known as Spearman's rank correlation coefficient, is a non-parametric measure of the strength and direction of association between two variables. It is based on the ranks of the data rather than their actual values, making it less sensitive to outliers and more robust against non-linear relationships. In our analysis, we will utilize the Spearman correlation to assess the performance of our models. By comparing the Spearman correlation coefficients of the different methods we have modeled, we aim to select the model that most closely approximates the true variations in the price of electricity futures contracts. This approach will enable us to choose the most accurate and reliable model for understanding the factors influencing daily electricity prices in France and Germany.

Several models can be used in order to determine the best Spearman Correlation. If we get the best model for our dataset (with the analysis of Spearman Correlation), it is great to make an in-depth analysis of which variables explain in the best way the electricity daily price.

4 Target Variable Analysis

Before studying the reasons of the price variations of electricity futures, we need to analyze the Target Variable, which corresponds to the real price variations that we observe during our timeframe. We want to check about the normality of the Target Variable.

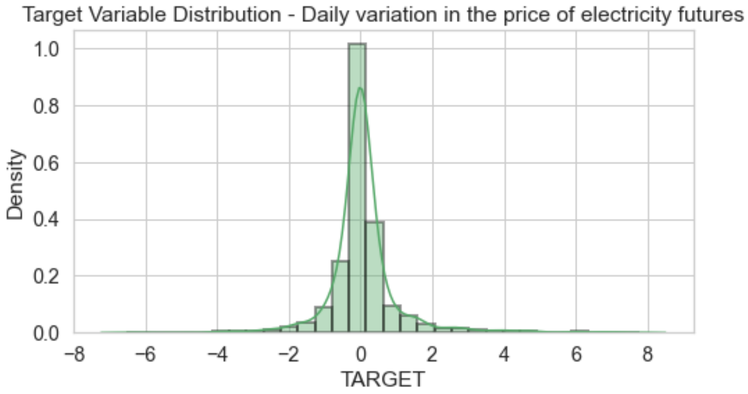

FIGURE 2 – Target Variable Distribution : The Target Variable seems to be be normally distributed

By plotting the Target Variable Distribution, we see that the data seem to be normally distributed. We have verified the normality hypothesis with a Shapiro-Wilk test. This test shows us that the target variable does not follow a normal distribution, as its p-value is under our threshold (0.05).

5 Principal Component Analysis

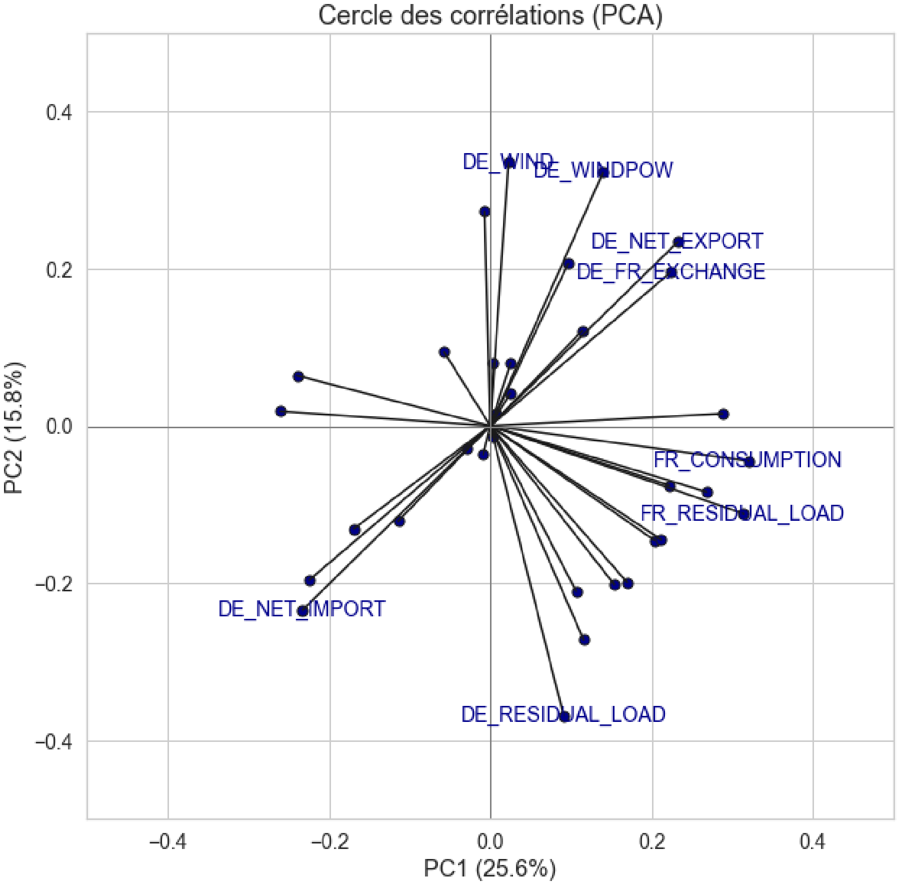

FIGURE 3 – Circle of the Correlation based on PCA

We can fix these multicollinearity we saw previously with techniques like Variance Inflation Factor (VIF) or PCA decomposition.

Thanks to the Correlation Circle of the Principal Component Analysis, we can figure out which variables have a high impact on our target variable (positively or negatively). The further the point is from the center of the circle, the greater the impact on the target variable. Our first dimension (PC1) explains 25.6% of our model, our second dimension (PC2) explains 15.8% of our dataset.

We have decided to print the 8 most impactful variables based on PCA. We see that the Exchange Between France and Germany, the Germany Wind, the Germany Windpow, the French Consumption of electricity, the French and German consumption of energy after using renewable energy, the net import and export of Germany are the most important to explain the electricity price based on PCA.

But these informations are not significant enough. Indeed, if we plot a bar chart of the variance explained by the PCA, we have this result :

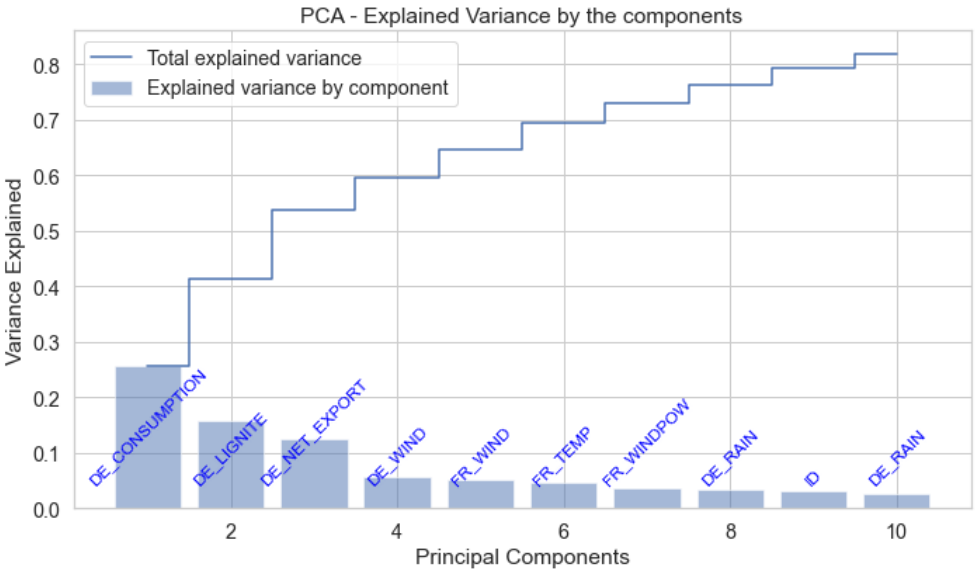

Figure 4 – Bar Chart reprensting Variance explanation of PCA

We see that 80% of the variance is explained by 10 components, which is really high. It means that, for this model, the components are not impactful enough. However, after calculating the R^2 (coefficient of determination) of the PCA model, we see that the PCA model is not accurate at all. So, we have figured out that the Random Forest model was a better one to try to explain the variation of electricity price (as he has a score R^2 of 35%, while PCA has a score of 5%)

6 Spearman Correlation calculation

As we want to explain electricity futures contracts price variations (daily timefreame), we need to calculate the spearman correlation with different models. The best Spearman correlation is the highest one.

We have calculated the Spearman Correlation with several models :

- Linear Regression
- Random Forest
- Ridge Regression
- Lasso Regression
- ElasticNet Regression
- Neural Network

— Principal Component Analysis
Here are our results :

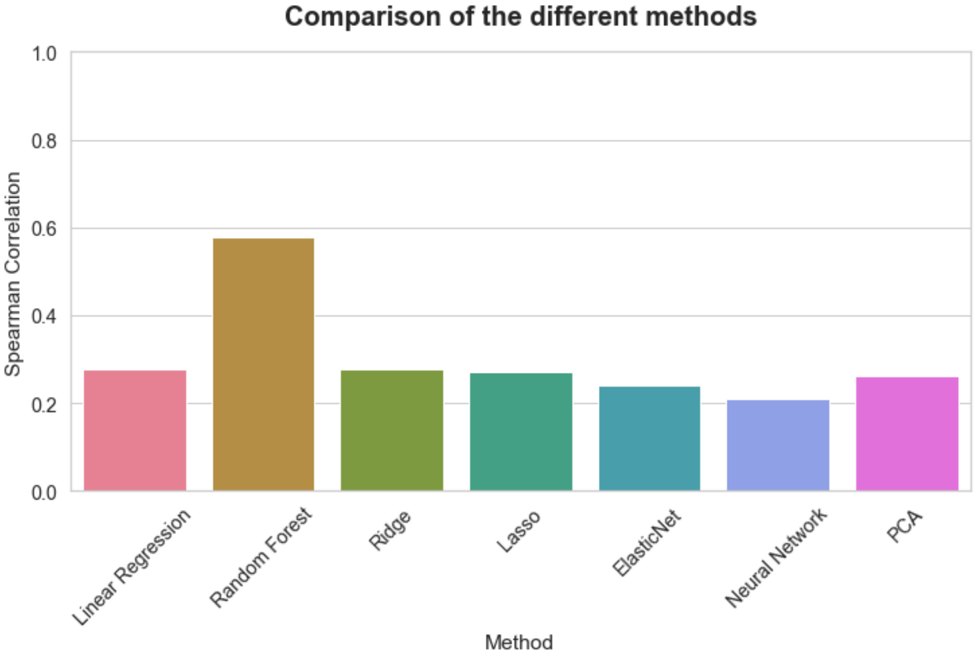

FIGURE 5 – Comparison of the different results

We see that the Random Forest and the Principal Component Analysis models seem to be the best for our context, regarding the Spearman Correlation. So we will focus on these two models to try to explain the variation of electricity futures contract prices.

7 Random Forest model

The Random Forest model seems to be a good choice for this exercise for several reasons :

Given the multifactorial nature of electricity prices, which are influenced by weather, energy production, geopolitics, and trade, the Random Forest model is able to capture complex and non-linear relationships between the features and the target variable.

Random Forest is an ensemble method that combines multiple decision trees, which makes it more robust and less prone to overfitting compared to single decision tree models. This is important for this exercise, as the target variable, electricity prices, can be influenced by various factors and their interactions.

Feature Importance : Random Forest models can provide feature importance rankings, which can help identify the most relevant variables in explaining electricity prices. This is valuable in this exercise, as there are numerous variables related to weather, energy production, and trade.

The Random Forest models can adapt well to different data distributions and variable scales, making them suitable for this exercise, which involves diverse features.

After fitting the random forest model with our data, we wanted to plot the 10 best important features of the model. As we saw previously, there is a huge difference between the Spearman Correlation of Random Forest and the other model : The Random Forest model seems to be the best one according to our exercise (based mainly on Spearman correlation).

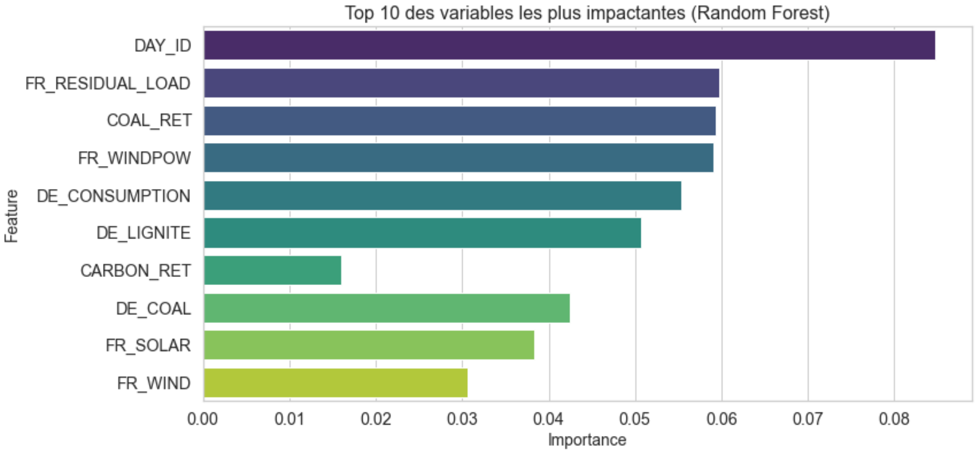

Figure 6 – The 10 variables that explain the most the Electricity Price variation

Based on the random forest, we see that the Day ID is the most important feature, following by the energy consumption after consumption of renewable energy in France, the Coal in Europe, the FR Windpow (which corresponds to wind turbines), the DE lignite production, the DE Coal, etc...

This seems logical : The date can influence electricity prices due to seasonal trends in demand and production. For instance, colder months might increase the demand for heating, while warmer months might lead to increased use of air conditioning. Moreover, some energy sources are more abundant in certain seasons, such as hydroelectric power during periods of heavy rainfall or snowmelt.

Coal is a significant source of energy production in Europe. As a result, fluctuations in coal prices can directly impact electricity prices. Higher coal prices increase the cost of electricity production, leading to higher electricity prices, while lower coal prices decrease the cost, resulting in lower electricity prices.

The Lignite is a type of low-quality coal and is a quite significant energy source in Germany. Changes in lignite production can affect the overall electricity supply in the country, and consequently, the electricity prices. An increase in lignite production might lead to lower electricity prices due to higher supply, while a decrease could result in higher prices because of reduced supply.

The Solar energy is an important renewable energy source in France. When solar production is high, it can help meet the electricity demand and reduce reliance on more expensive or less sustainable energy sources, leading to lower electricity prices. Conversely, lower solar production may require increased use of other energy sources, potentially raising electricity prices.

Finally, the wind is another key renewable energy source in France. Windy weather conditions can lead to increased wind power production, which can lower electricity prices by reducing the need for other, more expensive energy sources. On the other hand, less windy conditions might reduce wind power production, potentially leading to higher electricity prices as other energy sources are used to meet the demand.

8 Conclusion

On our python script, you can see that, for our test set, we had a really low Spearman Correlation for all the models. It shows that the machine learning methods are not enough powerful to explain the future price of electricity, it is hard to predict it with these methods and these variables, as electricity price is a complex science to study. In conclusion, explaining electricity prices is a complex task due to the multitude of factors involved. The Random Forest model has shed light on some key variables that significantly impact electricity prices, such as the date, coal prices in Europe, German lignite production, French solar production, and French wind conditions. These factors reflect the importance of seasonal trends, energy source availability, and the interplay between different types of energy production in determining electricity prices.

But we have to admit that the electricity market is influenced by more than just the parameters we have studied, it is not enough to totally explain the electricity prices. Geopolitical events, technological advancements, regulatory changes, and market dynamics also play a significant role in shaping electricity prices. To gain a deeper understanding of it, further research could explore the impact of these additional factors and how they interact with the variables identified by the Random Forest model.

By recognizing the multifaceted nature of electricity pricing, we can better anticipate and respond to fluctuations in the market, ultimately benefiting consumers, producers, and the environment.