Maria Regina Mucilli
Pablo Lagroue
Daniele Ferrara

Forecasting Electricity Prices in France and Germany

All commodities prices, including the one of electricity, are dependent on many factors, from supply and demand dynamics to market speculations, from macroeconomic and geopolitical events to technological innovations. In our analysis, we introduce a model, testing it against 3 challenger models. The objective of our work will be to predict the price of electricity over one day as a function of 35 input parameters. Our study focuses on the electricity markets in Europe.

Our model is also tested with other explanatory variables, such as other commodities' prices and production volumes, carbon emissions futures' performances, total daily exchange of electricity between France and Germany, export and import of electricity to and from Europe for the two countries.

The objective of the analysis is to assess whether the explanatory variables efficiently predict the electricity prices in Europe. To do so we isolated the explanatory variables from the other ones, and we tested the spearman correlation of every variable with the price of electricity, we found out that no one was truly related with the price, but the whole data was.

We are aware that events such as wars, introduction of tariffs, supply cuts for specific commodities (e.g. oil production cuts by OPEC) or other macroeconomic and/or geopolitical events can strongly affect, directly or indirectly, the electricity market; nevertheless, our idea is to investigate the relationship between electricity prices and all the columns we have and verify whether a trend can be identified or not.

Plan of Experimentation

To do so we firstly identify the response and explanatory variables to test in our Linear Regression. Given our dataset, our response variable is Futures daily price variation of 24h electricity baseload. Our explanatory variables are the daily return of gas, coal and carbon emissions futures, weather forecasts data (temperature, rain, wind), energy production for various commodities, different electricity consumption data in France and Germany.
Then, we run a Linear Regression using the above mentioned variables. Traditional Linear Regression is limited since variable cross-correlation is not taken into account in the analysis, and this could generate issues in the study of the explanatory power of some variables.

To overcome these obstacles, we test our champion model against three challenger models which represent an alternative to our first approach. Using the same explanatory variables as before, we run Lasso Regression Analysis (a), Ridge Regression (kernel: 'rbf') (b), Random Forest Regression (c). Once all the results from the champion and challenger models are collected, we assess the results both from a statistical and economic point of view, to understand if our variables and assumptions are realistic, and to verify the presence of a trend and the possible reason(s) behind it. Based on the specificity of the models and the related results, we can then evaluate which model is the most efficient in explaining the relation between electricity prices and the explanatory variables. Our evaluation is based on the quantitative results arising from the regressions and the cross-correlation measures among the variables, in order to rank the models and assess their efficiency.

The Dataset

The dataset we use is composed of 1,494 rows, with each row representing one calendar day (associated to a unique ID, DAY_ID). Our dataset covers around 4 years of data and consists of 35 columns that provide detailed and structured information. The first column, ID, serves as a unique row identifier linked to a specific day (DAY_ID) and a country (COUNTRY). The DAY_ID column represents the day identifier, with dates anonymized, ensuring that all data corresponding to a specific day remains consistent. The COUNTRY column identifies the country, with FR representing France and DE representing Germany. The choice of studying only these two countries is due to the dataset available, the reliability of the data found and the fact that they represent the two biggest electricity markets in Europe, representing a solid proxy for the wider European electricity market.

Subsequent columns capture daily commodity price variations, including GAS_RET for European gas, COAL_RET for European coal, and CARBON_RET for carbon emissions futures. Additionally, the datasets include daily weather measures specific to the country, such as x_TEMP for temperature, x_RAIN for rainfall, and x_WIND for wind.

There are also energy production measures recorded daily in the respective country, including x_GAS for natural gas, x_COAL for hard coal, x_HYDRO for hydro reservoir, x_NUCLEAR for daily nuclear production, x_SOLAR for photovoltaic energy, x_WINDPOW for wind power, and x_LIGNITE for lignite. Furthermore, the datasets provide metrics for daily electricity usage in the country, which include x_CONSUMPTON for total electricity consumption, x_RESIDUAL_LOAD for electricity consumption after utilizing all renewable energies, x_NET_IMPORT for imported electricity from Europe, x_NET_EXPORT for exported electricity to Europe, DE_FR_EXCHANGE for total daily electricity exchange between Germany and France, and FR_DE_EXCHANGE for total daily electricity exchange between France and Germany.

The output datasets, representing the response variable, consist of two columns: ID, which is the unique row identifier, associated with a day (DAY_ID) and a country (COUNTRY), and TARGET, which indicates the daily price variation for 24h electricity baseload futures.

Before starting with the Champion Model (Linear Regression), the data are cleaned by substituting the missing values (NaN) with 0s and dropping the "Country" column, since we apply the same methodology to French and German data and our target is the general European electricity market, for which we use both France and Germany's markets as a proxy.

Our Champion Model: Analyzing the Energy Market with Linear Regression

Linear Regression, particularly using the Ordinary Least Squares (OLS) method, is a statistical technique employed to model the relationship between a dependent variable and multiple independent variables. In our analysis of daily price variations for 24h electricity baseload futures (refereed as the TARGET variable), the model aims to predict these price changes based on various factors, including commodity prices, weather conditions, and energy production and consumption metrics. The primary objectives of the model are to provide accurate predictions, estimate the strength of relationships between variables, and identify significant predictors that impact electricity pricing. To ensure the validity of the model, key assumptions must be met, including linearity, independence of residuals, homoscedasticity, normality of residuals, and the absence of multicollinearity among independent variables.

The Dependent Variable: TARGET

The TARGET variable represents the daily price variation for electricity baseload futures and serves as the focal point of our predictions. This can also be intended as the return of such financial instruments. Our champion model is a traditional Linear Regression, using the following independent variables.

Independent Variables

To predict the TARGET variable, we integrate various independent variables that capture essential aspects of the energy market:

Commodity Prices: We include daily returns for key commodities, such as GAS_RET (European gas prices), COAL_RET (European coal prices), and CARBON_RET (carbon emissions futures). These prices directly impact electricity production costs and represent an alternative source of energy to electricity.

Weather Conditions: Variables like x_TEMP (temperature), x_RAIN (rainfall), and x_WIND (wind measurements) are crucial, as weather significantly affects both energy demand and renewable energy production.

Energy Production Metrics: We analyze different energy sources, including x_GAS (natural gas), x_COAL (hard coal), x_HYDRO (hydro), x_NUCLEAR (nuclear), x_SOLAR (solar), x_WINDPOW (wind), and x_LIGNITE (lignite). These metrics provide insights into overall electricity availability and production capabilities.

Electricity Consumption Metrics: Metrics such as x_CONSUMPTION (total electricity consumption) and x_RESIDUAL_LOAD (consumption after utilizing renewable energy) are included, along with x_NET_IMPORT and x_NET_EXPORT, which reflect electricity flow between countries. DE_FR_EXCHANGE and FR_DE_EXCHANGE indicate daily electricity exchanges between Germany and France, highlighting demand and supply dynamics.

Model Implementation and Insights

In order to implement the model, the dataset is divided into two sets: 90% for training (X_train, Y_train) and 10% for testing (X_test, Y_test) to evaluate the model on unseen data.

The scoring function used is Spearman's correlation, which measures the monotonic relationship between the predicted and actual daily price changes over the testing dataset sample.

A k-fold cross-validation with k=5 is implemented as a validation technique, which involves splitting the training data into 5 subsets, training the model on 4 of them, and then evaluating it, using Spearman's correlation, on the remaining one. This process is repeated for every possible combination.

Using Linear Regression, we analyze the relationships between the TARGET variable and the independent variables. The model estimates coefficients that reveal the impact of each factor on electricity prices. For instance, a positive coefficient for x_TEMP may indicate that higher temperatures lead to increased electricity prices, while certain commodity prices may negatively influence prices based on production adjustments.

The insights generated from this model enhance our understanding of energy market dynamics, allowing stakeholders to optimize production strategies and develop effective pricing models.

In summary, the application of Linear Regression in analyzing daily price variations in electricity baseload futures provides a structured framework for understanding the complex interactions within energy markets. By leveraging a comprehensive dataset that includes commodity prices, weather conditions, and consumption metrics, we can derive a model that can be efficient in predicting European electricity prices.

Challenger models: Ridge, Lasso Regression and Random Forest Regression

In order to deepen the analysis and compare the Linear Regression Model with more complex models we implement two Ridge Regression models, with different kernels and the Lasso model.

The first challenger model implemented is the Ridge Regression with radial basis function kernel which is used for non-linear relationships between features. In the wake of being more efficient we use a GridSearch Cross Validation method which helps us to find the best hyperparameters to train the model. The hyperparameters are ‘alpha’, regularization parameter and ‘gamma’ which is the parameter for the radial basis function (RBF) kernel; it controls the influence of a single training point. As a validation technique 5-folds cross validation is implemented.

Secondly, we implement the Lasso Regression model. The only hyperparameter tuned is ‘alpha’ which controls the strength of the penalty on the coefficients.

Last, we implement the RandomForest Regression, this method helps to reduce the risk of overfitting and improves the robustness of the model. In essence, it's a powerful tool for regression tasks that can handle complex and non-linear data. In the wake of being more efficient we also use a GridSearch Cross Validation method which helps us to find the best hyperparameters to train the model. The only hyperparameter we have is n_estimators: it specifies the number of trees in the forest. Essentially, it determines how many decision trees the model will build and average to make its final prediction.

As a validation technique 5-folds cross validation is always implemented.

Analysis of the results from a data science point of view

Linear Regression Model Results

The reported Spearman Correlation Scores from cross-validation are [13.96, 13.12, 22.01, 27.97, 21.79] with mean correlation score across the 5 folds 19.77%. These values represent the monotonic correlation (scaled to a percentage) between the linear model's predictions and the actual target values across the 5 folds of the K-Fold validation. The scores vary between approximately 13% and 28%, indicating that the model has limited but non-negligible ability to capture a monotonic relationship between the input features and the target. The average of the model's overall performance during validation indicates that the linear regression model explains a small portion of the monotonic relationship in the data.

Then, evaluating the model on the test sample, the Spearman Correlation is 22.4% which is slightly higher than the cross-validation average (19.77%), indicating that the model performs consistently between the validation and test sets.
The model generalizes reasonably well to unseen data. However, a correlation of ~22% still indicates a weak predictive performance.

Ridge Regression with radial basis function kernel

From the grid of hyperparameters we have the best combination of hyperparameters for the problem, based on cross-validation performance which is ‘alpha’ = 5 and ‘gamma’ = 0.001.

After finding the best hyperparameters, the cross-validation scores are computed using Spearman correlation as the scoring metric. The average result is 21.23%. Comparing this value with the one given by Linear Regression, we notice that it is slightly higher than the previous. This indicates that this Kernel Ridge Regression with the tuned parameters has better predictive performance, but still explains a small portion of the target variability. On the test set we have a result for the Spearman's correlation of 20.2%.

Lasso Regression Model Results

The best results of GridSearchCV is 'alpha' = 0.01 resulting in the best performance (highest Spearman correlation).

Lasso regression, compared to the other methods, shows a notably better performance both on the train set (21.40%) and test set (22.7%). Lasso has the property of introducing sparsity into the coefficients, which may make the model more robust and less prone to overfitting. Lasso has also a good generalization ability on the test set and is particularly effective at identifying the most influential variables.

Random Forest Regression Model Results

The best results of GridSearchCV is 'n_estimators' = 95 resulting in the best performance (highest Spearman correlation).

Random Forest shows that the correlation on the train set is lower than in other models with an average correlation 15.56%, but the performance on the test set is the best of all, 24.9%. This may indicate that the model has learned some complex structures in the data that other models failed to capture.

Random Forest tends to avoid overfitting, as evidenced by its lower train set correlation, and its ability to generalize to test data is superior. This suggests that Random Forest effectively models data variability and adapts well to non-linear relationships between variables.

Comparative Analysis

Evaluating the models implemented in terms of accuracy on the test set, Random Forest Regressor shows the best results followed by Lasso Regression.

In terms of overfitting, Ridge Regression with RBF Kernel and Lasso Regression show similar performance across the train and test sets, suggesting a good balance between learning from training data and generalizing to test data. Linear Regression has a slight difference between train and test results, but the gap is not significant.

Another aspect which differs among the implemented models is the interpretability. Indeed Linear Regression and Lasso are the most interpretable models. Linear regression is a simple and easy-to-understand model, while Lasso, although introducing sparsity, remains relatively transparent compared to complex models like Random Forest. Ridge Regression with RBF Kernel and Random Forest, being non-linear models, are less interpretable and more challenging to understand compared to linear models.

Finally, Random Forest offers the best performance on the test data and demonstrates strong generalization ability but with limited interpretability; while Lasso Regression seems to be the most balanced model, with good performance on both the train and test sets, and relatively high interpretability. Ridge Regression and Linear Regression provide similar results, with Ridge slightly improving upon linear regression when regularization is introduced, but overall yielding lower performance compared to Lasso.

Analysis of the results from a business point of view

Economically speaking, finding a model with a high predictive power, would allow us to arbitrage on the prices of electricity futures, by selling or buying such financial instruments based on the results of our prediction model. Even in a competitive market as the electricity one, an efficient model would grant a high return, especially in the first period of implementation of the strategy; after some months, or even weeks or days, the arbitrage would become of public domain, meaning that other agents in the market will start to use the same, or a similar, prediction model, and to make the same operations on the market. This would impact the supply and demand dynamics and it would make it impossible to profit from the predictions of electricity futures, since everyone would move in the same direction, immediately inflating or defaulting prices when the model predicted so. Therefore, even a model with a very high predictive power, in the medium-long run, could become inefficient, and other variables which were not taken into account before could become crucial in explaining futures' prices.

Now, focusing on the 4 models we presented in this project (our champion model and the 3 challenger models) and the variables we used, we were not able to strongly predict the prices of electricity futures contracts, since the performance of our models, expressed in terms of the Spearman correlation, were in the range 20-25%, meaning that through pure analysis we were able to only marginally explain which are the drivers of electricity prices in Europe.This can be partially explained

by the complexity of the datasets and regardless of the model, it is complicated to get a high correlation between the explanatory variables and the response variable (electricity prices). A solution could have been to split the dataset in 2 sub-datasets, one for France and one for Germany, but this would have made the comparison between models impossible.

Analysing commodities markets and their dynamics, in particular the electricity one, we can spot some factors which play a relevant role in determining energy prices, such as macroeconomics events (e.g. wars or embargos), supply cuts (e.g. OPEC cuts on oil production) or market speculation (heavy buying or selling by hedge funds and other markets agents in order to profit from the abovementioned factors as well as from others). Our models do not take into account macroeconomic factors, which, however, are very difficult, if not impossible, to measure through variables - it is a hard challenge to predict political games, wars, tariffs or other such events and therefore to identify a proper associated variable. Regarding supply cuts and market speculation, our models partially address the dynamics of supply and demand in the electricity market, but only in terms of exchanges of electricity between France and Germany, and hence considering the two countries as a closed system not influenced by other countries or by the exchange market. Even though France and Germany represents the two biggest economies in Europe and have many similarities, using data collected from both the countries, considering them as a big single market, to run an analysis on the electricity market is simplistic and does not take into account the different sources of electricity for the two countries, which can be affected differently by macroeconomic events. For example, after the invasion of Ukraine, electricity prices, and hence electricity futures, in Germany were impacted more strongly than the ones in France, since Germany was much more dependent on Russian gas for electricity production, while France is more reliant on Nuclear plants. These differences are not considered in our analysis.
Moreover, market speculations and investment strategies by markets' agents are not at all introduced in our models, which is an excessive simplification from reality.

The lack of variables which measure such factors is probably the reason for our models poor performances.

Conclusion

The models implemented in this project, including a champion linear regression model and three challenger models, Lasso Regression, Ridge Regression, and Random Forest Regression, were not able to strongly predict the prices of electricity futures contracts.. The performance of the models, measured by the Spearman correlation, was in the range of 20-25%. This indicates that the analysis was only able to marginally explain the drivers of electricity prices in Europe.

Among the factors which contribute to the limited predictive power of the models are the complexity of the data, the exclusion of macroeconomic factors, such as wars or embargos, supply cuts, or market speculation, which significantly impact energy prices. Furthermore, the data act as a simplification of the market; indeed the analysis treats France and Germany as a single, closed market, which does not fully capture the different sources of electricity and their varying sensitivities to macroeconomic events. For instance, Germany's greater reliance on Russian gas compared to France's nuclear power dependence was not considered. Although the models address supply and demand dynamics through electricity exchanges between France and Germany, they do not consider the influence of other countries or the broader exchange market. Market speculation and investment strategies are also not included .

Among the models, this project reveals that Random Forest Regression shows the best performance on the test data, indicating a strong generalization ability, but it has limited interpretability; Lasso Regression appears to be the most balanced model, with good performance on both training and test sets, and relatively high interpretability, while Ridge Regression and Linear Regression provided similar results, with Ridge showing a slight improvement when regularization was introduced.

From a business perspective, an efficient model could allow for arbitrage on electricity futures prices. However, the effectiveness of such models may diminish over time as market participants adopt similar strategies. In the medium to long term, other factors not initially considered could become critical in explaining futures' prices.

In conclusion, while this project provides valuable insights into energy market dynamics, the predictive power of the models is limited by the complexity of the market, and the exclusion of significant macroeconomic and market factors . To improve the models' performance, future studies should consider including macroeconomic variables and expanding the geographical scope of the analysis.