uc3m | Universidad Carlos III de Madrid

Master Degree in Big Data Analytics

2022 - 2023

Master Thesis

"Prediction of matching prices in electricity markets through curve representation"

Daniel Foronda Pascual

Andrés M. Alonso Fernández

Logroño, junio 2023

AVOID PLAGIARISM

The University uses the **Turnitin Feedback Studio** for the delivery of student work. This program compares the originality of the work delivered by each student with millions of electronic resources and detects those parts of the text that are copied and pasted. Plagiarizing in a TFM is considered a **Serious Misconduct**, and may result in permanent expulsion from the University.

SUMMARY

In the Spanish electricity market, after the daily market is held in which prices are set for the next day, the secondary and tertiary markets take place, which allow companies to more accurately adjust the electricity they are able to offer. The objective of this Master thesis is to predict the final price reached in these markets by previously predicting the supply curve, which is the aggregate of what companies offer. We follow a Machine Learning approach being in our case Histogram Based Gradient Boosting the algorithm that gives the best results.

Keywords: Electricity Secondary Market, Electricity Supply Curves, Multivariate time series forecasting.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Spanish electricity market

1.1.1. History

Prior to 1997, the Spanish electricity sector was dominated by a state-owned utility, which held a monopoly over electricity production and distribution. However, in 1997, the Spanish Electricity Act was established, which initiated a process of liberalization and deregulation, aiming to introduce competition and promote a more efficient and dynamic electricity market.

In the early 2000s, Spain experienced a boom in renewable energy, particularly wind power. Government incentives and favorable policies attracted significant investments, making Spain one of the global leaders in wind energy capacity. In 2009 the Spanish government introduced a feed-in tariff system to promote renewable energy that guaranteed fixed prices for electricity generated from renewable sources attracting further investments. However, due to the rapid growth and higher costs, the government later reduced these incentives to mitigate the impact on consumers' electricity bills. (energiaysociedad.es, n.d.-a)

In recent years, Spain has continued to prioritize the expansion of renewable energy and the transition towards a more sustainable and decarbonized power sector. It has set ambitious targets for renewable energy penetration, aiming to reach 100% renewable electricity by 2050.

1.1.2. Agents

The main participants in the Spanish electricity market can be classified into several key players: (wikipedia.org, 2023)

Generation Companies: These companies are responsible for producing electricity. They own and operate power plants, including conventional thermal power plants, nuclear power plants, and renewable energy installations such as wind farms and solar power plants. Some of them are Iberdrola, Endesa, Gas Natural Fenosa (Naturgy), Acciona Energía.

Transmission System Operator (TSO): The TSO, known as Red Eléctrica de España (REE), manages and controls the high-voltage transmission grid. REE ensures the reliable and secure transmission of electricity throughout the country, maintaining the balance between supply and demand.

Distribution Companies: These companies operate the local distribution networks

and are responsible for delivering electricity to end consumers. They maintain and manage the distribution infrastructure, including power lines, transformers, and substations. Some examples are Endesa Distribución, Naturgy Distribución, Viesgo Distribución.

Retail Suppliers: Retail suppliers purchase electricity from the wholesale market and sell it to end consumers. They offer various pricing plans, manage customer relationships, and handle billing and customer services. Examples: Endesa Energía, Naturgy, Iberdrola, Repsol Electricidad y Gas, Holaluz.

Market Operator: The market operator, known as Operador del Mercado Ibérico de Energía (OMIE), oversees the operation of the wholesale electricity market. OMIE facilitates the trading of electricity between generation companies and retail suppliers, ensuring fair and transparent market conditions.

Regulator: The electricity market in Spain is regulated by the National Commission of Markets and Competition (CNMC). The CNMC ensures compliance with market rules, promotes competition, and regulates tariffs and prices to protect consumer interests.

Consumers: Consumers play a crucial role in the electricity market. They include residential, commercial, and industrial users who purchase electricity for their own consumption. Consumers have the option to choose their preferred retail supplier and participate in demand response programs to optimize their energy usage.

1.1.3. Daily, Intraday, Secondary, and Tertiary Markets

The Spanish electricity market operates through a three-tier system consisting of the daily, secondary, and tertiary markets. Each market serves a specific purpose and contributes to the determination of the final electricity price. Here's an explanation of how each market works: (endesa.com, 2022)

Daily Market (Mercado Diario): The daily market, also known as the spot market or the day-ahead market, is where electricity is traded for delivery on the following day. In this market, generation companies submit their offers to supply electricity based on their production costs and availability. At the same time, retailers and large consumers submit their bids for purchasing electricity. The market operator, Operador del Mercado Ibérico de Energía (OMIE), matches the offers and bids to determine the market matching price, also known as the marginal price. The market matching price is the price at which the demand for electricity matches the available supply. This price is used to settle the transactions in the daily market.

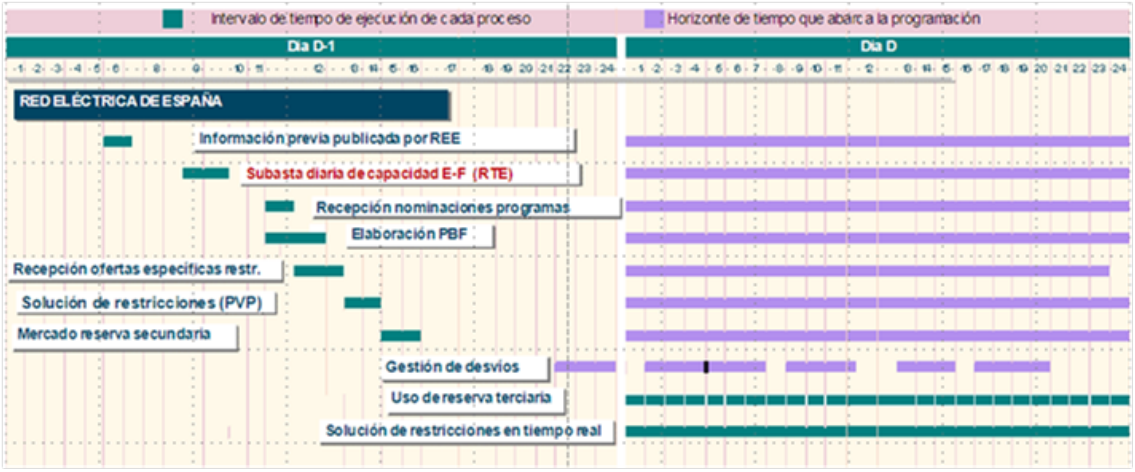
Secondary Market (Mercado Secundario): The secondary market takes place after the Daily Market enabling market participants to adjust their positions and make corrections to balance their portfolios. It provides flexibility for market participants to manage unexpected changes in supply or demand. There are two modalities: subir (rise) to increase the electricity supply of a generation company, and bajar (reduce) to decrease it.

Intraday Market (Mercado Intradiario): The intraday market, also known as the real-time market, is an additional segment where electricity is traded and balanced in real-time. It allows market participants, such as producers and consumers, to adjust their energy schedules and make last-minute trades and ensures the efficient utilization of electricity resources.

Tertiary Market (Mercado Terciario): The tertiary market, also referred to as the imbalance settlement market, addresses any imbalances between the contracted and actual consumption or generation of electricity. Market participants who deviate from their contracted positions during the delivery period can buy or sell imbalances in the tertiary market. The prices in the tertiary market are set based on the costs associated with balancing the system, including penalties for imbalances. The market operator, Red Eléctrica de España (REE), calculates the final settlement prices based on the imbalances and applies them to the relevant market participants' invoices.

Below we can see a diagram representing when each market occurs (Spanish: *Intervalo de tiempo de ejecución de cada proceso*, in green) and when each market applies (Spanish: *Horizonte de tiempo que abarca la programación*, in purple) from energiaysociedad.es, n.d.-b

Figure 1.1
Spanish electricity market scheme



1.2. Objective

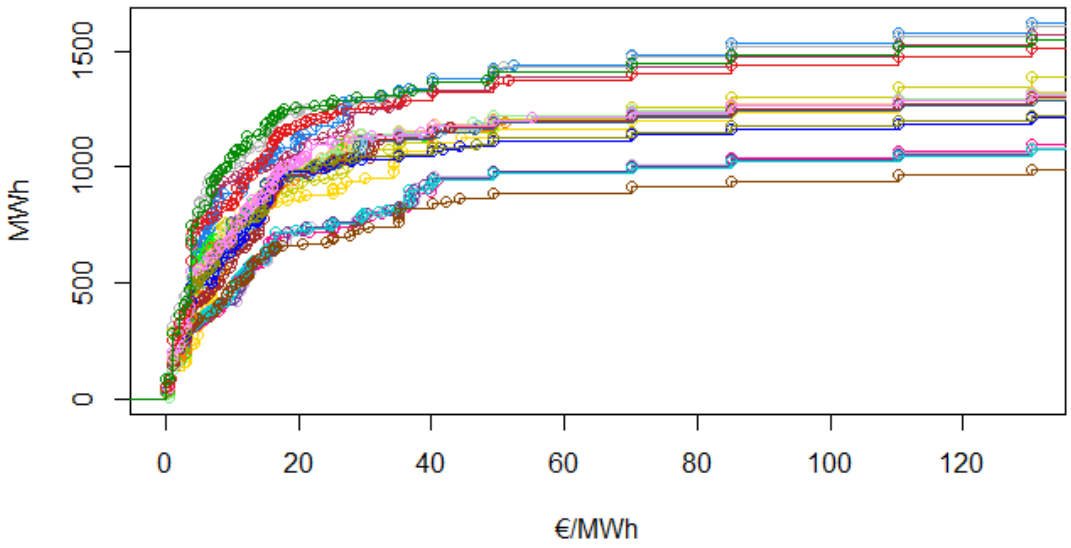
In this project, we will study prediction methods for the day ahead matching prices at the secondary electricity market in Spain by firstly predicting the supply curves and matching them later with the requirements to obtain the price.

In Figure 1.2 we can see an example of several supply curves for the rising secondary market for a given day (there is one curve for every hour).

For each hour there is a requirement. By intersecting the curve with the requirement

Figure 1.2

Examples of supply curves for one day



we obtain the matching price. The publication of the requirements takes place an hour and a quarter before the companies send the offers (see Figure 1.3 taken from BOE, 2020).

Figure 1.3

Publications and submission of requirements and offers for secondary and tertiary markets

Concepto	Hora límite de publicación (D-1)
Puesta a disposición de los PM y del OM de los resultados de la subasta de capacidad de contratos bilaterales con entrega física efectuada, en caso de congestión, en las interconexiones sin procedimiento coordinado de asignación de capacidad.	14:45 horas.
Publicación PDVP por el OS.	14:45 horas (en todo caso, hasta 75 min tras publicación PDBF).
Requerimientos de banda de regulación secundaria.	14:45 horas.
Presentación de ofertas de banda regulación secundaria.	16:00 horas (en todo caso, hasta 75 minutos tras la publicación del PDVP).
Asignación de banda de regulación secundaria.	16:30 horas (en todo caso, hasta 30 minutos tras el cierre de presentación de ofertas de regulación secundaria).
Requerimientos de reserva de regulación terciaria.	21:00 horas.
Presentación de ofertas de regulación terciaria.	23:00 horas.

1.3. State of the art

The main objective of this work is to predict the matching price in the secondary market. This could be attempted by treating that price as a time series taking into account some other exogenous variables. However instead of following this approach we will follow the

idea explained on Ziel and Steinert, 2016 which takes into account that the matching price is the intersection point between the supply and demand curves trying first to predict these curves in order to predict the matching price. In this way the authors are able to make more accurate predictions than those coming from techniques that were previously used based on the matching price series. This information on the origin of the matching price as the intersection of two curves seems to enable a more accurate prediction that also has potential applications in the bidding strategies of electricity companies.

In order to predict these supply and demand curves, one possible approach is to perform dimensionality reduction and therefore lose some information. To avoid this loss Mestre et al., 2020 use functional regression for the same purpose through a model based on a double-seasonal functional SARMAHX capable of capturing daily and weekly seasonality of the time series, also including exogenous variables.

On the other hand, Shah and Lisi, 2020 use both parametric and nonparametric functional autoregressive models (FAR and NPFAR) showing that nonparametric models lead to a statistically significant improvement in the forecasting accuracy compared to previous studies. A Non-Parametric Functional Autoregressive (NPFAR) model is a flexible approach used to analyze functional data and make predictions without assuming a specific functional form. It captures the dependence between variables over time by considering a functional response variable and its past values which allows for capturing complex patterns and dynamic characteristics of functional data.

In this thesis, instead of using techniques like those mentioned above, we will try to predict the curves using models based on machine learning algorithms.

2. METHODOLOGY

2.1. Software tools

To develop this project we used both R and Python using, among others, the sklearn, matplotlib, and torch packages. The code developed for this research can be consulted at: https://github.com/dani1717/supply-curve-forecast

2.2. Approximations

Supply curves are non-decreasing step functions functions. A first problem that we encounter when trying to use time series prediction methods on them is that the steps are located on different abscissas, as we have seen in Figure 1.2. To solve this problem, the first step we take is to establish a fixed grid on the x-axis (price) and approximate each supply curve to another increasing step function that has the steps in that grid. In this section we follow the procedure described in Alonso and Li, 2022.

2.2.1. Choosing the grid for prices

In order to accurately reflect the steps that occur most frequently in the supply functions we calculate the empirical cumulative distribution function (ecdf) of these steps and use certain evenly distributed percentiles of it. However, since there is a large concentration of different steps close to zero, we will establish a filter selecting only the prices whose supplied quantity (q, measured in MWh) is above a threshold q_0. That is, we will use n homogeneously distributed percentiles of the function $\hat{F}(p|q \geq q_0)$ being $\hat{F}(p|q) = N^{-1} \sum_{j=1}^{N} I(p \leq p_j, q \leq q_j)$ where the pairs (p_j, q_j) are observed bids in the curves. Then, the prices in the grid are obtained by:

$$p^i_{n,q_0} = \hat{F}^{-1}\left(\frac{i}{n} \,|\, q \geq q_0\right) \quad \text{given } n \text{ and } q_0. \tag{2.1}$$

Therefore there are two parameters that will determine the grid and that will influence the precision of the approximations: the size of the grid (n) and the minimum quantity (q_0) that we consider to take into account the prices in the ecdf. To tune these parameters and see which ones give the best results we first need to define a curve approximation method to be able to measure the error generated by each grid.

2.2.2. Approximations using $\mathcal{L}_2$ loss function

To assess the goodness of the approximations we need a loss function. We will use the following with r=2 as proposed by Alonso and Li, 2022., in which case we can analytically

obtain the minimum.

$$\mathcal{L}_r = \left\|C_t - \hat{C}_{t,n}\right\|_r^r = \int_0^{+\infty} \left| C_t(p) - \sum_{i=1}^{n} c_{t,i,n}\phi_{i,n}(p) \right|^r W(p)dp,$$

where

$$\phi_{i,n} = \begin{cases} 0 & \text{if } p < p_i \\ 1 & \text{if } p \geq p_i \end{cases}$$

and $W(p)$ is a non-negative weight function such that $\lim_{p\to+\infty} W(p) = 0$ in order to ensure the convergence of the above integral.

Since we want the approximations to be more accurate in the areas where the match with the requirement usually occurs, we have considered different candidates for the W function:

- The fit of the final prices in a train set with an exponential (W_finalPrices_exp), logNormal (W_finalPrices_logNormal), Cauchy (W_finalPrices_Cauchy), normal distribution (W_finalPrices_Normal), or the exponential of the latter (W_finalPrices _normal_exp).
- The fit of all prices in the training set with an exponential function (W_allPrices_exp).
- The fit of all unique prices in the training set using an exponential function (W_unique Prices_exp).

2.2.3. Finding the best parameters

In the first place, we will find the weight function W that gives the best results in terms of minimizing the difference between the matching price using the original curve and the approximate curve. For this we provisionally adopt a grid with $n = 45$ and $q_0 = 337$ (first quartile of the quantities) and we take a sample of size 1000 curves.

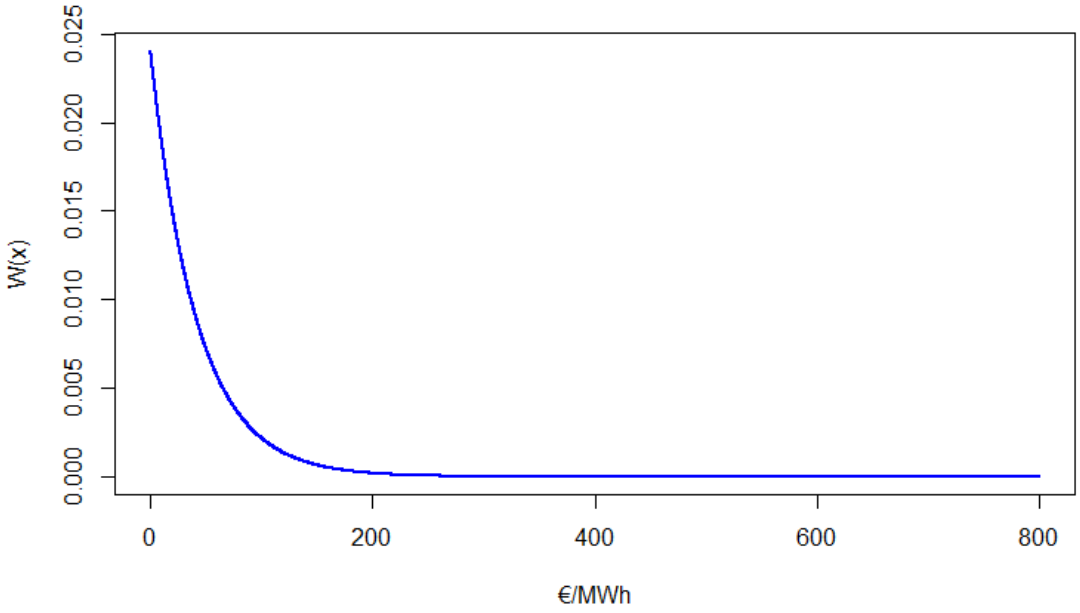
Weight function	avgError	median	sd	Q1	Q3	P90	P95	P99
W_uniquePrices_exp	0.37	0.23	1.92	0.1	0.41	0.66	0.95	2.56
W_finalPrices_logNormal	0.37	0.23	1.76	0.1	0.41	0.67	0.98	2.64
W_allPrices_exp	0.37	0.23	1.69	0.1	0.41	0.67	1.00	2.40
W_finalPrices_cauchy	0.39	0.23	2.19	0.1	0.41	0.70	1.00	2.77
W_finalPrices_exp	0.39	0.23	2.35	0.1	0.41	0.69	1.00	2.75
W_finalPrices_Normal	0.40	0.22	2.29	0.1	0.41	0.68	1.00	2.61
W_finalPrices_Normal_Exp	0.47	0.23	2.78	0.1	0.42	0.69	1.00	2.77

Table 2.1

Approximation errors for the final price depending on W

After this experiment, we select the exponential fit of the unique prices, W_uniquePrices_exp, as weight function since in this way we obtain an error of less than €1/MWh in 95% of the curves. We can visualize the function in the following plot.

Figure 2.1
Weight function for approximation error



After choosing W we can refine the selection of n and q_0. As it is logical, a larger n will increase the accuracy of the approximations but it will subsequently slow down the prediction methods and could lead to unnecessary redundancies. For q_0 we consider two possibilities: zero or the first quartile of the quantities (337 MWh). The following tables show the approximation errors for the final price depending on these two parameters with a sample size of 1000 curves.

n	avgError	median	sd	Q1	Q3	P90	P95	P99
35	0.5494965	0.30	1.6812160	0.12	0.60	1.00	1.4800	4.2216
40	0.4692460	0.27	1.0790573	0.10	0.52	0.90	1.2900	3.7800
45	0.4405317	0.24	2.3306209	0.10	0.47	0.77	1.1200	3.1803
50	0.4045010	0.23	2.0040731	0.10	0.42	0.75	1.0605	2.9901
55	0.3392000	0.20	0.7780733	0.08	0.40	0.67	0.9500	2.4405

Table 2.2
Approximation errors for the final price depending on n for $q_0 = 0$

n	avgError	median	sd	Q1	Q3	P90	P95	P99
35	0.4863605	0.32	2.273142	0.13	0.53	0.83	1.2000	3.0604
40	0.4634894	0.27	2.772642	0.12	0.47	0.77	1.1205	2.9603
45	0.3796765	0.23	2.197812	0.10	0.42	0.67	0.9900	2.5701
50	0.3974545	0.20	3.526804	0.09	0.39	0.60	0.9100	2.5101
55	0.3438410	0.20	2.847417	0.09	0.36	0.57	0.8000	2.1102

Table 2.3

Approximation errors for the final price depending on n for $q_0 = 337$

It is worth mentioning that there are some errors that can exceed 100 €/MWh. This is because sometimes the supply curve does not reach the required quantity and therefore they do not intersect while there is a final price assigned. This makes the difference between that price and its approximation very large and the median to be a more valuable measure than the mean.

As a conclusion we have chosen $n = 50$ and $q_0 = 337$ as the parameters that guarantee a good balance between precision and simplicity. Clearly $q_0 = 337$ works better than $q_0 = 0$ while n=50 somewhat improves the median of the errors while increasing it to $n = 55$ would hardly improve it.

To check that the approximations are good we can compare them with two naive estimators. The first one consists of approximating each curve with the previous day's curve and using it to calculate the matching price. The second naive estimator simply consists of estimating the final price of a day as the same as the previous day. The table 2.4 shows that both methods are much worse than using the approximations we have obtained.

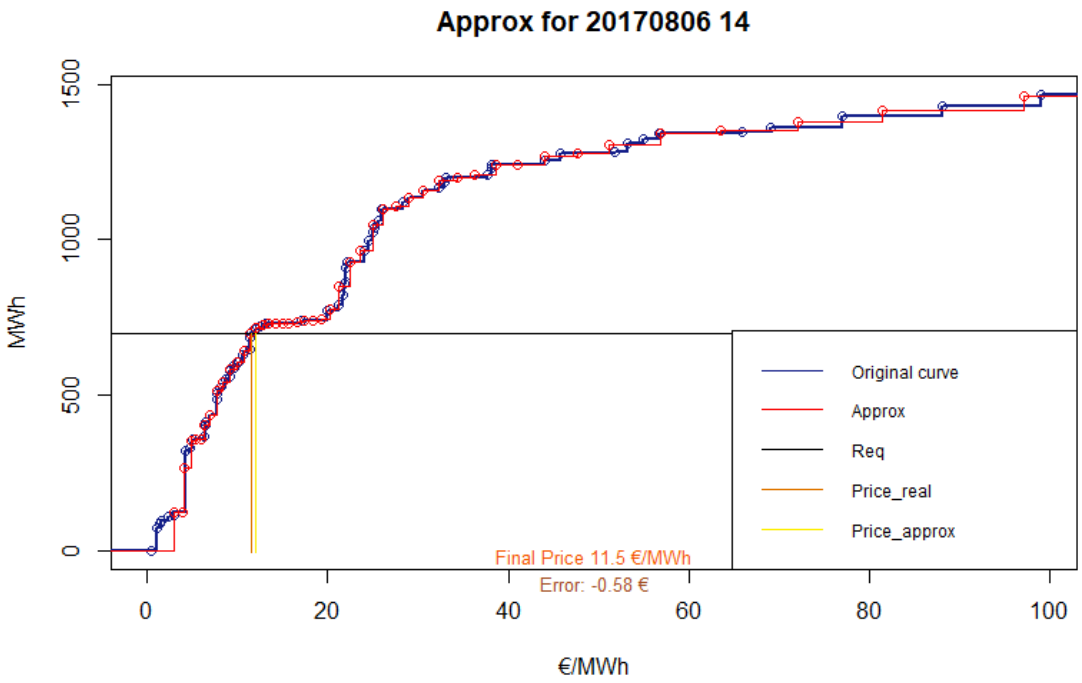
naive_estimator	avgError	median	sd	Q1	Q3	P90	P95	P99
previousDayCurve	7.03	3.60	12.19	1.38	8.35	16.41	24.33	50.56
previousDayPrice	6.20	3.36	8.70	1.30	7.68	14.65	21.22	43.10

Table 2.4

Prediction errors for the final price with naive estimators

Finally, in figure 2.2 we can see an example of an original supply curve (in blue) together with its approximation (in red) and with the requirement (horizontal black line), the true matching price and the matching price obtained through the approximation.

Figure 2.2
Example of curve's approximation



2.3. Models

2.3.1. Metrics and naive estimators

Our procedure has two steps, first the prediction of the curves is performed and then the prediction of the matching price is obtained by intersecting the requirement and the predicted curve. We need two metrics and two naive estimators, one for each prediction step we carry out. Firstly, to measure the error and effectiveness of our supply curve forecasts, we will use the mean absolute error (MAE) and the root mean squared error (RMSE) as metrics, and the previous day curve as naive estimator. On the other hand, with regard to the predictions of the matching price, we use the MAE as a metric and the matching price of the previous day as a naive estimator. For both cases we tried other naive estimators such as the previous week's curve and prices but they give worse results.

2.3.2. Preprocessing of data

After the approximation of the supply curves to a fixed grid of prices, we have a table of dimensions $(n_{data}, grid_{size})$ where the $grid_{size}$ is 50. To try to predict future supply curves we will add lags and transform this table by reducing its dimensionality through Principal Component Analysis (PCA), we also incorporate exogenous variables, others related to calendar effects and finally doing feature engineering to create new variables. These are

the steps we have taken:

- **Lag 24**: In each row we added 50 new columns with the values of the previous day's curve on which we will later apply dimensionality reduction.
- **Calendar variables**: we have included the following dummy variables using one-hot-encoding: hour of the day, day of the week, month, quarter, a binary variable on whether that day is a national holiday or not.
- **Exogenous variables**: We have included an indicator of the wind speed and solar diffuse radiation of each capital of province in Spain obtained from https://open-meteo.com/ (104 variables) on which we will later apply dimension reduction. We have also included a column with the Dutch TTF gas price, the reference price in the European market, from www.investing.com. Lastly, we have included a column with the matching price reached in the daily market and another one with the amount of MWh assigned.
- **Train and test split and dimensionality reduction**: Before reducing the dimensions of the variable input we perform a train and test split reserving the years from 2014 to 2018 for training and 2019, 2020 and 2021 for test. Next we perform a principal component analysis on the 50 columns of the lag24 of each curve. As expected, the 50 columns are highly correlated and selecting the first five principal components we have an explained variance of 98.51%. Something similar happens with the wind and radiation variables in each province. From the 52 solar radiation variables, by selecting just the first two we obtain a 89.66% of explained variance while for the wind with 10 principal components we explain 79.82% of the variance. This process helps us to greatly reduce the number of input variables and therefore the processing time. All this process is carried out in the training set and reproduced with the same parameters in the testing set.
- **Customized features**: Perhaps the lag 24 for each curve is not enough information about the evolution of the offers in the last days/weeks. For this reason we added some more variables that enable the algorithm to detect this trajectory. After trying different options we added information in two ways. On the one hand, to try to provide some data on the evolution of the supply for this same hour over the last few weeks, we have taken the first principal component of each supply curve for that same hour in the last 12 weeks (84 columns). Through PCA we have reduced this information to 15 components. We have also added information relative to all the hours of the previous four days, taking the first principal component for each curve and then doing PCA to select the first eight principal components that explain this trajectory.

Of course before performing each PCA we scaled the data. Subsequently, for the training of some methods such as neural networks, we have also scaled and normalized the

input variables. On the other hand, these parameters that we mentioned like 12 previous weeks of information on the same hour and four previous days of information on all hours are the values that we found after hyper-parameter optimization (HPO) with the method that worked best with the data. Initially for the tests they had different values.

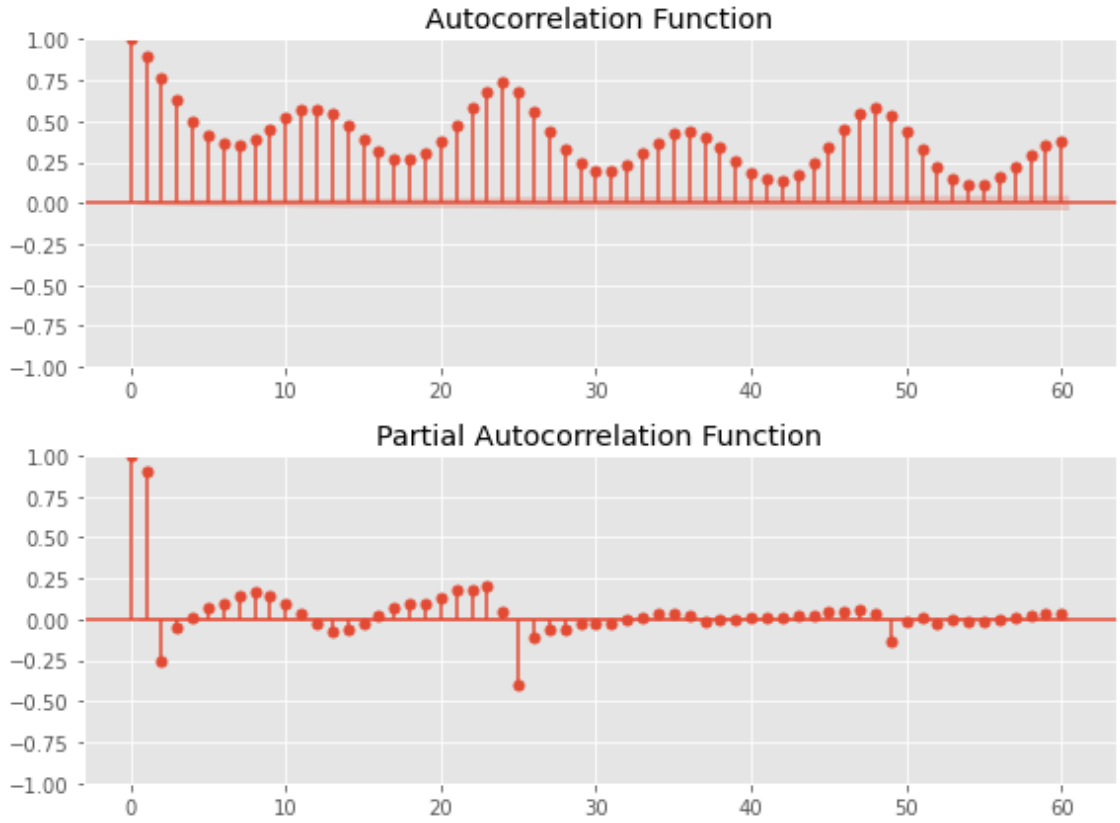
Finally we count with 91 input variables which are the following ones:

- Lag 24 (*5 variables*): The first five principal components of the supply curve for the same hour in the day before.
- Hour (*24 columns*)
- Weekday (*7 columns*)
- Month (*12 columns*)
- Quarter (*4 columns*)
- Holiday (*1 column*)
- Solar radiation (*2 columns*): The first two PCs of the diffuse solar radiation in Spanish capitals of province
- Wind speed (*10 columns*): The first ten PCs of the wind speed in Spanish capitals of province
- Gas price (*1 column*): Price of Dutch TTF gas
- Daily market (*2 columns*): The matching price and quantity assigned in the daily market of the same day.
- Same hour evolution (*15 columns*): Information on the evolution of the first principal component of each curve for the same hour in the last 12 weeks.
- All hours evolution (*8 columns*): Information on the evolution of the first principal component of each curve for all hours in the last four days.

2.3.3. ARIMA model

The first model that we will try will be a Seasonal Autoregressive Integrated Moving Average (SARIMA) with a 24-hour period in order to have a reference (Hamilton, 1994). A brief analysis of the autocorrelations of one of the 50 series (in this plot Q30) yields the following results.

Figure 2.3
ACF and PACF of a series



To fit a SARIMA model to the series, we first reduced its dimension by means of PCA, taking only the first two principal components that explain 95.39% of the variance. Then, to predict a 24-hour horizon, we use the auto_arima procedure from statsmodels with training data from the previous month. The results are only slightly better than those of the naive estimator.

2.3.4. Machine Learning models

To try to predict the 50 series, different machine learning algorithms were tried, including Random Forest (RF), Histogram-based (HB) Gradient Boosting, Dense Neural Networks (DNN) and Long-Short-Term Memory (LSTM).

Four different alternatives have been tested with **Random Forest**: 1) A single RF model considering only the time series. In other words, the input variables being the first 5 PCs of the previous day curve. 2) A single RF model with all the input variables, endogenous and exogenous. 3) 24 RF models, one for each hour, considering only the time series (first 5 PCs of the previous day curve). 4) 24 RF models, one for each hour, considering all the input variables.

We tested these models in the years 2019-2021 in which there is a problem with irregularities in the electricity market due to the pandemic. For this reason, in the table at the

end of this section, in addition to the results for that period, we present also the error for the year 2019 itself, which allows us to see how the model predicts the curves for the next year, since perhaps an interval of three years, especially in these conditions, might be too long.

As it is observed in the table a single model works better in this case. This is probably due to the fact that by dividing the data into 24 groups, they do not have a sufficient number of observations for an efficient fit of the model.

Histogram Based Gradient Boosting works better than Random Forest giving better results with a single model than 24 different ones, probably for the same reason. To try to mitigate this problem of having few observations in each group, we attempted to cluster the hours, clustering them into four different groups and therefore fitting four models and not 24. The results, which also appear in the table, are no better than a single HB Gradient Boosting model so we have chosen the latter as our model.

A **DNN** model was also tried performing HPO on the number of layers, the size of each layer, dropout, etc. Finally the best option of DNN was that of 24 different models, each one being a neural network of one hidden layer of 1443 neurons. However the results did not improve the HB Gradient Boosting. On the other hand, a model with **LSTM** also was created, with very poor results.

Model	2019-2021		2019	
	MAE	RMSE	MAE	RMSE
Naive	172	229	181	240
SARIMA	167	220	174	229
Random Forest w/o exogenous vars	151	195	153	196
Random Forest w/ exogenous vars	157	194	141	180
Random Forest 24 models w/o exogenous vars	160	204	162	206
Random Forest 24 models w/ exogenous vars	168	216	150	192
HB Gradient Boosting	151	195	137	175
HB Gradient Boosting 24 models	159	204	143	183
HB Gradient Boosting 4 models	155	200	137	176
DNN 24 models	169	233	138	178

Table 2.5

Errors metrics in curve prediction, 2019 and 2019–2021

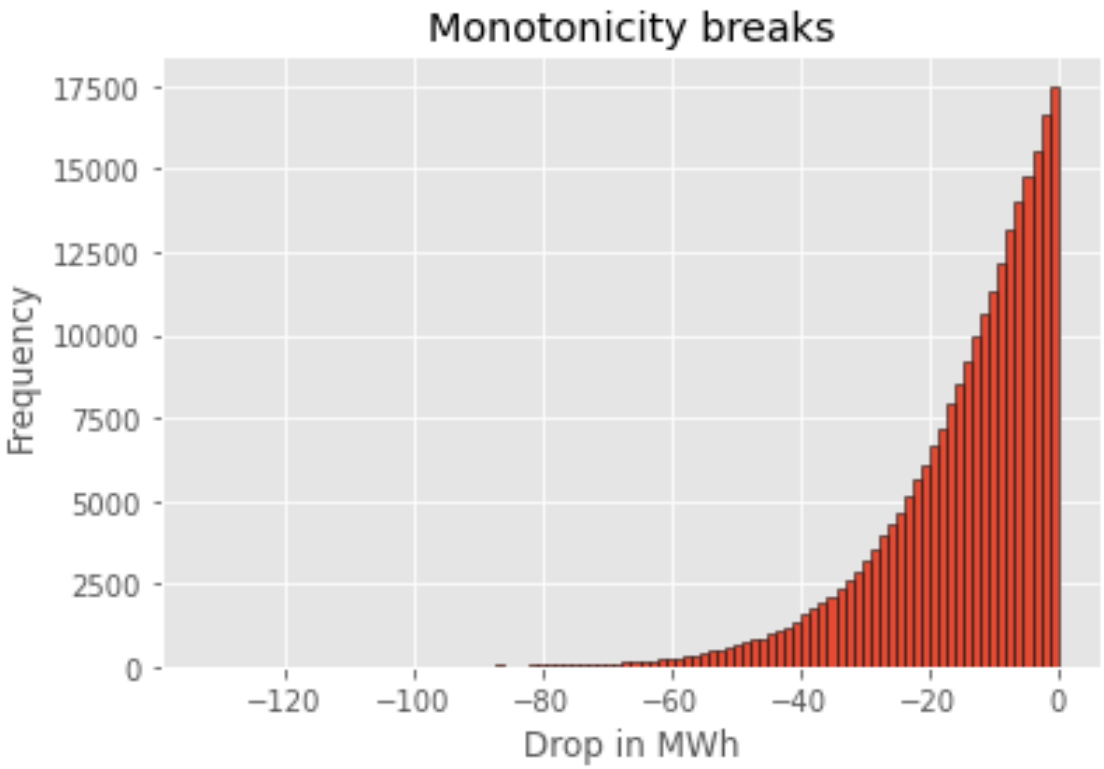
From these results, trying to reduce the error, we tried to create an **ensemble** with the models that worked best using a linear regressor and Random Forest as metamodels without improving the results in any case.

2.3.5. Monotonicity restoration

Each model predicts the future values of 50 time series, however it must be taken into consideration that these 50 series make up different supply curves that must always be non-decreasing functions. In practice, we observe that the models do not exactly respect this characteristic, producing small mismatches that break the monotonicity between the 50 values. Below we can observe the frequency and size of the monotonicity breaks in the outputs from HB Gradient Boosting.

Figure 2.4

Histogram of monotonicity breaks before correction



To solve this problem a method has been proposed to restore monotonicity. It consists of the following procedure: if in a curve there is a point followed by more than one point of lesser value before the curve rises up again, then we consider this local maximum as an error and we decrease it. Whereas if it's only a single point that has a lesser value before the curve rises again, then we increase the low point. In this way, iterating several times we manage to restore the monotony in the curves and these new corrected curves are closer to the original ones. In the following tables it is possible to see how this process reduces the errors of the predictions with respect to the real curves.

MAE	2019 I	2019 II	2019 III	2020 I	2020 II	2020 III	2021 I	2021 II	2021 III
Before correction	131.27	123.85	153.29	163.23	136.05	136.54	136.85	147.38	146.01
After correction	131.19	123.57	152.86	162.74	135.49	136.20	136.89	147.32	145.58
Difference (%)	-0.06	-0.23	-0.28	-0.30	-0.41	-0.25	0.03	-0.03	-0.29

Table 2.6

MAE before and after monotonicity restoration

RMSE	2019 I	2019 II	2019 III	2020 I	2020 II	2020 III	2021 I	2021 II	2021 III
Before correction	167.20	158.05	195.89	209.50	171.04	174.37	176.54	188.90	189.13
After correction	167.04	157.69	195.33	208.90	170.43	173.89	176.58	188.89	188.65
Difference (%) %	-0.10	-0.23	-0.29	-0.29	-0.36	-0.27	0.03	-0.00	-0.25

Table 2.7

RMSE before and after monotonicity restoration

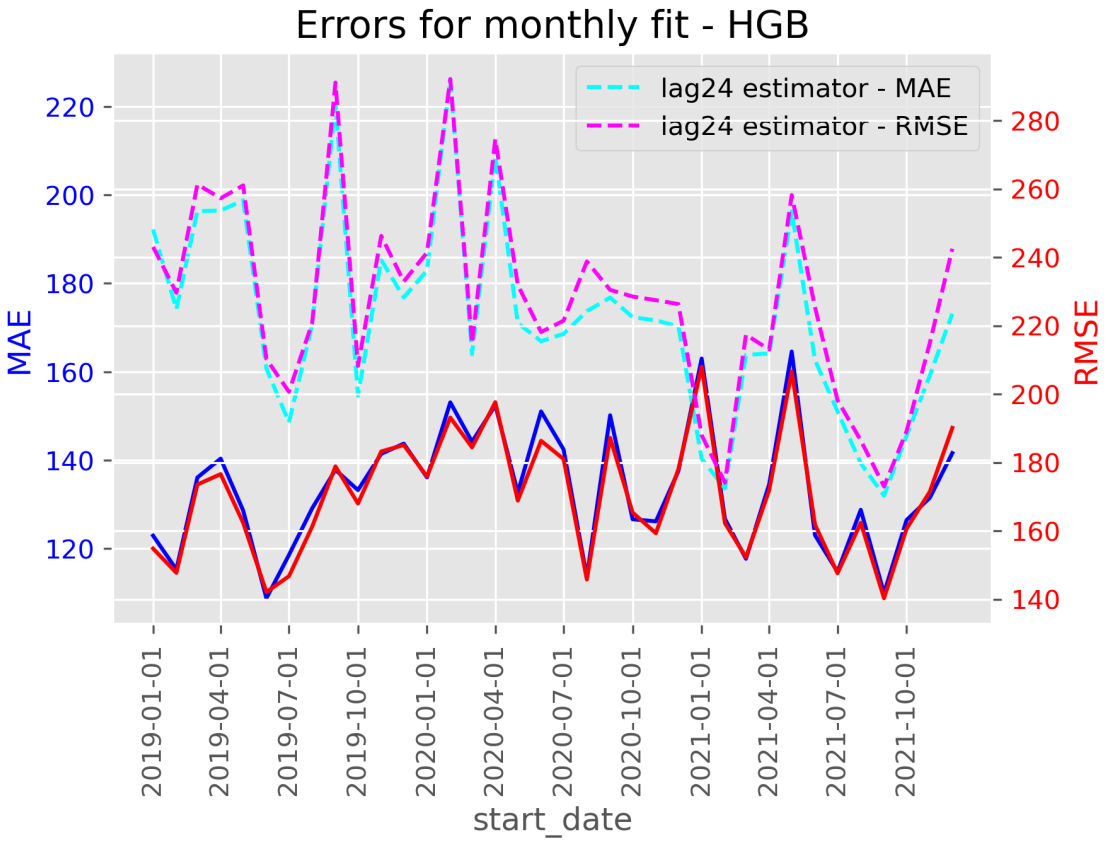
We can see that the improvement is small, which tells us that monotonicity problems were not large in the prediction results of the chosen model. In any case, it is preferable to apply this correction so that the predictions satisfy the non-decreasing monotonicity constraint.

3. FINAL RESULTS

3.1. Final model for curve prediction

Finally, the model used to predict the supply curves is a HB Gradient Boosting. In order to obtain the best possible results when estimating the efficiency of the model, a monthly retraining was carried out, that is, the model is retrained with all the data, from the first observation to last day of the previous month, to predict the new month. We also tried with a fixed time window (for example, the last two years) when predicting each month, but the results were better if all the data prior to that date is considered for the training. To predict each month, therefore, all the data is preprocessed again each iteration, the model is retrained and the curves for the next month are forecasted. In this way we obtain the errors in the prediction of the curves.

Figure 3.1
Errors in curve forecast per month

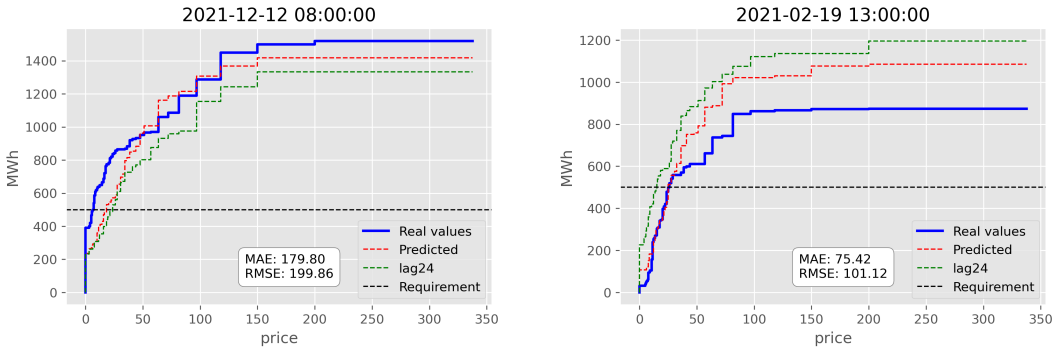


We can see that the errors in the test set are always lower with the final model than with the naive estimator except in one month, January 2021. We can look at two examples of predictions for supply curves. We remark that it is particularly important how the curve is

predicted in the neighborhood near the intersection with the requirement (black horizontal line).

Figure 3.2

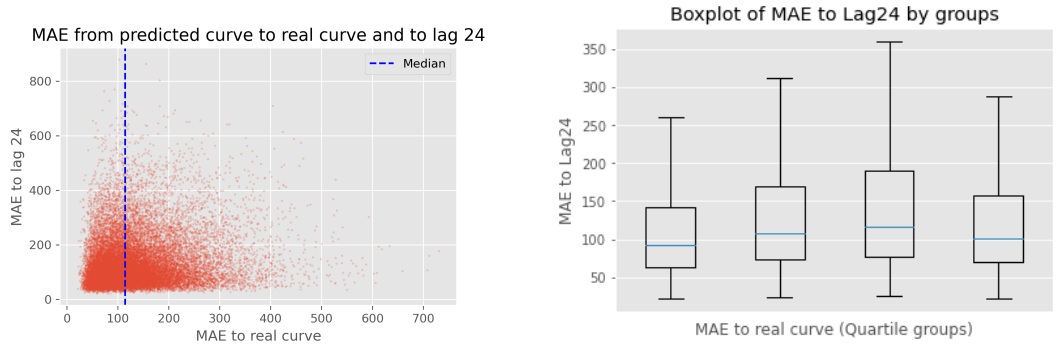
Examples of curve forecasts



Looking at some examples of predictions, the importance of the previous day's curve for the new prediction becomes clear, as will be confirmed in the next section. Therefore, it may be interesting to study its influence with greater detail. We can illustrate this fact with the first of the following plots that shows on the horizontal axis the distance from a predicted curve to the actual curve, and on the vertical axis the distance from the predicted curve to the previous day's curve.

Figure 3.3

Influence of lag 24 on predicted curves



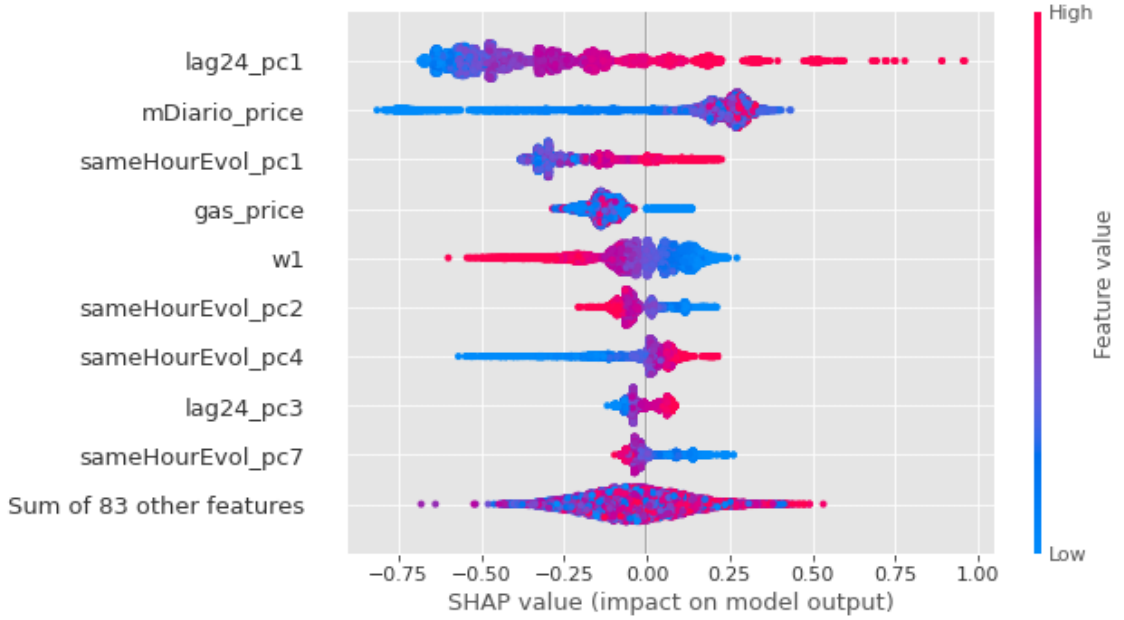
As can be seen, the distance from the predicted curve to the previous day's curve follows a right skewed distribution. It is interesting to note that 25% of the best predicted curves have a shorter distance from the previous day's curve than other groups, which means that in many of the curves with a more accurate prediction, the real value closely matches the lag 24 so in those cases they may not be very difficult to predict. In the opposite case, however, something similar happens, 25% of the worst predictions have a smaller distance from previous day's curve than in groups of quartiles 2 and 3, so it might happen that in many cases the algorithm is predicting the curve very similarly to that of the previous day but the real value is quite different. In these cases, perhaps the algorithm

is not correctly using certain information to disassociate the prediction from the lag 24, or perhaps the input data is not providing the key information that causes the supply curve to be different on those days from that of the previous day.

3.1.1. Interpretability

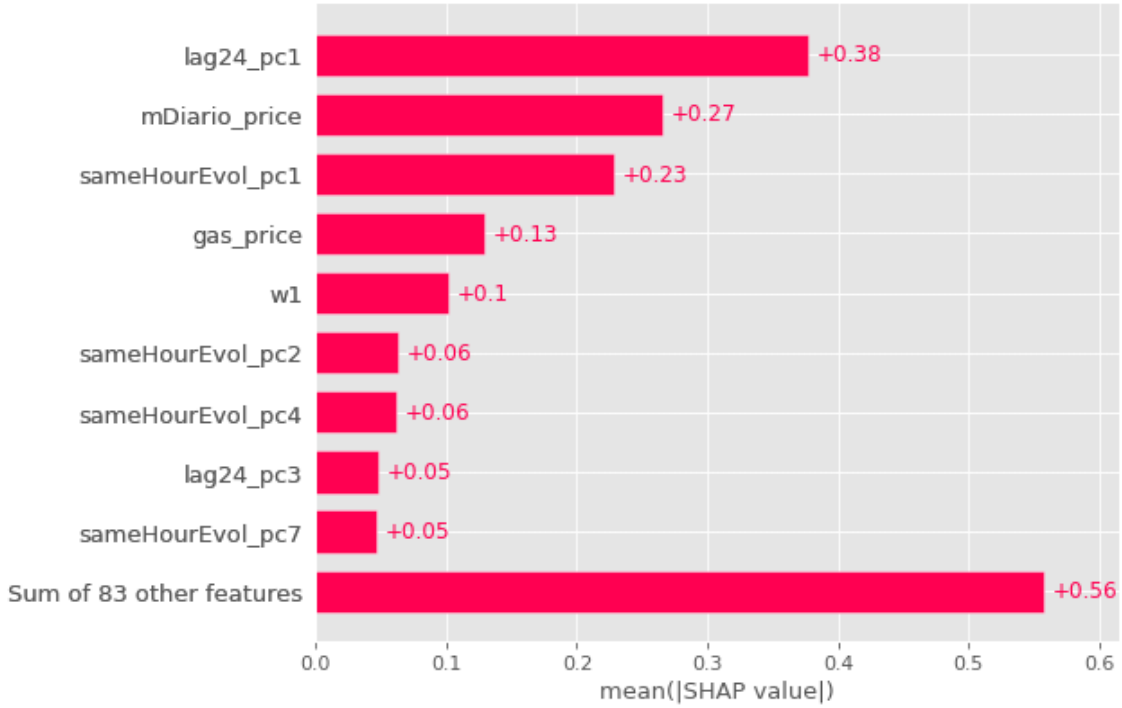
Histogram based Gradient Boosting is an optimized variant of the Gradient Boosting algorithm especially useful when working with large or high-dimensional data sets like ours. Unfortunately, it does not provide feature importances that might allow us to know which input variables are more decisive in the predictions. However we can use the shap package that makes use of the Shapley values (Shapley and Shubik, 1954). Shapley values are a technique derived from game theory used to fairly allocate each player's contribution to the overall outcome of a cooperative game. When applying them to the interpretation of machine learning models, the central idea is to evaluate how the prediction changes when a specific feature is included or excluded, considering all possible feature combinations. In this way we can obtain the following results for the prediction of a time series, in this case Q30. (w1 represents the first principal component of the wind speed)

Figure 3.4
Shap values of features for Q30 (beeswarm plot)



In the previous plot there is one point for each predicted hour. The shap value is proportional to the influence that each of the features has on the prediction of Q30, so if the majority of points are far from the center (0) that variable will have great importance in the predictions. We can summarize this information in the following plot.

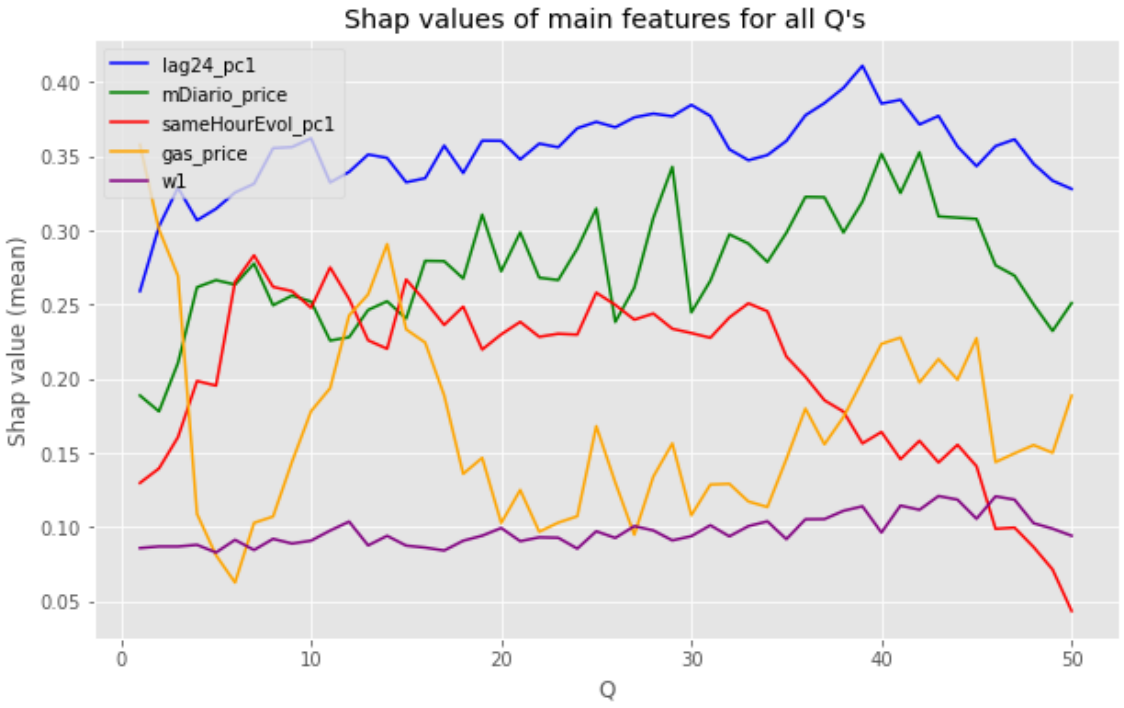
Figure 3.5
Shap values of features for Q30 (barplot)



From this information we can obtain a graph of the shap values of the most important features for all the series like the one appearing below. It shows that the most influential variable in the prediction of a curve is the curve of the previous day. This influence is also more or less constant along the diferent Q's, although it shows a slight increasing trend up to Q39.

Figure 3.6

Shap values of the most relevant features for all series



The second most important variable for predicting the supply curve is the daily market matching price for that day. Next, the evolution of the first principal component of the curve for that same hour over the last 15 weeks is the third most important variable, although it has less weight when predicting the values at the end of each curve. Lastly, the price of TTF gas, which mainly affects the series between Q12 and Q16 and between Q39 and Q44 approximately, and the first component of wind speed. We can remark that the shap values do not add up to one and that the importance in the series Q1 to Q5 is not very indicative since their values are very close in all cases.

3.2. Matching prices prediction

Once we have solved the prediction problem for the supply curves, we only have to calculate their intersection with the requirement for that hour/day to get a prediction for the matching price. After that we compare these predictions with the true final prices obtaining the following table.

	HGB	Naive price	Naive curve
mean	4.97	6.36	7.08
std	8.48	10.34	21.00
min	0.00	0.00	0.00
25%	0.96	1.03	1.00
50%	2.28	2.83	2.80
75%	5.38	7.38	7.40
90%	11.63	15.20	15.75
95%	18.33	23.71	24.79
99%	43.49	52.87	58.00
max	208.20	209.62	592.87

Table 3.1

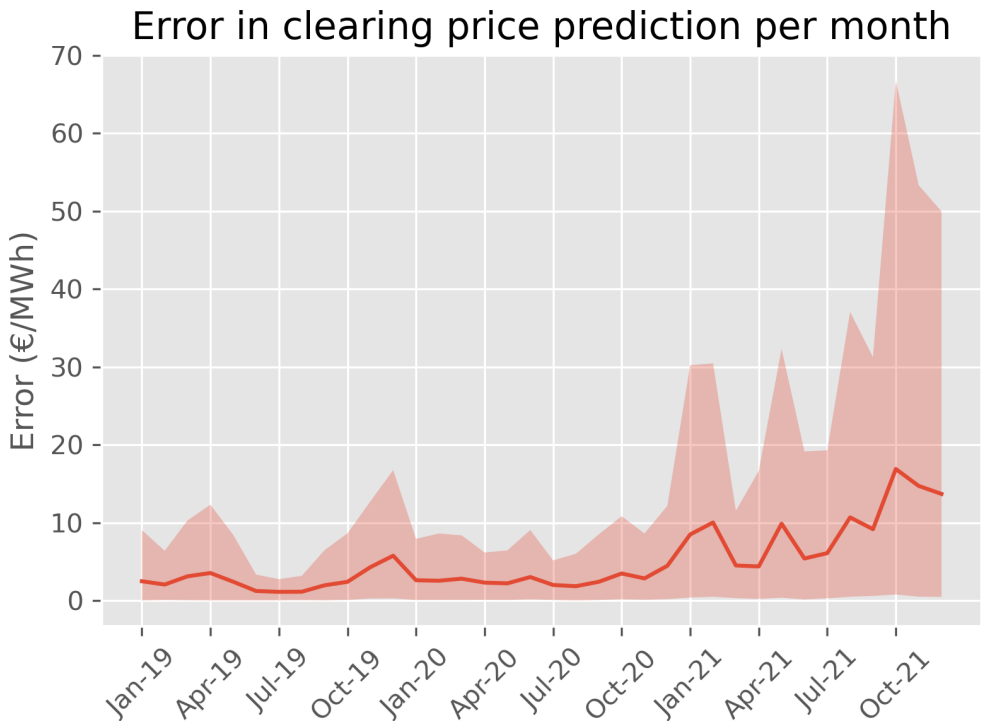
Summary of the absolute value of the prediction errors for matching prices (€/MWh)

In the table, we compare the prediction error with the HB Gradient Boosting (HGB) with the ones obtained with two naive methods: (1) the previous day matching price (Naive Price) and (2) the intersection of the previous day's supply curve with the actual requirement (Naive curve).

Also it's possible to study the error month by month as in the following graph, which shows the interval between the 5th and 95th percentile as well.

Figure 3.7

Prediction Errors of matching price by month using HBM



In the graph we can see that in the last year the errors are notably higher than in the two previous ones. In fact, if we calculate the errors per year, the results are as follows.

datetime	mean	min	q1	median	q3	p90	p95	p99	max
2019	2.65	0.00	0.70	1.45	3.10	6.55	9.70	16.33	59.00
2020	2.74	0.00	0.85	1.90	3.65	6.26	8.58	13.37	43.21
2021	9.52	0.00	2.12	5.06	11.61	23.03	35.06	66.71	208.20

Table 3.2

Errors in matching price predictions by year (€/MWh)

As can be seen, the predictions for 2021 are much worse than those of previous years, and could even be below the performance of naive estimators. However, if we disaggregate the errors made by the naive estimators and show only those for 2021, we will see how they are worse than HB Gradient Boosting.

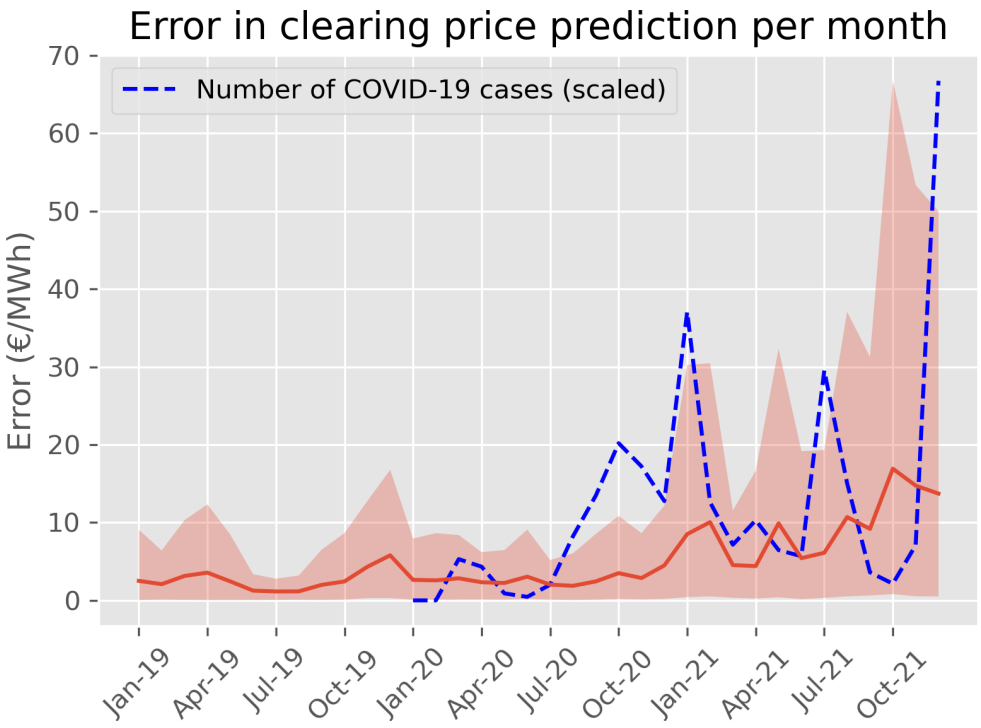
2021	**HGB**	**Naive price**	**Naive curve**
mean	9.52	11.68	13.02
std	12.82	15.17	28.78
min	0.00	0.00	0.00
25%	2.12	2.59	2.53
50%	5.06	6.51	6.40
75%	11.61	14.36	14.64
90%	23.03	28.27	29.76
95%	35.06	43.30	45.00
99%	66.71	74.98	82.84
max	208.20	209.62	585.39

Table 3.3

Errors in matching price predictions - 2021 (€/MWh)

A possible explanation for this irregularity can be attributed to the measures implemented in response to the COVID-19 pandemic in 2021. However, it is worth noting that strong measures were also implemented in 2020, yet the errors do not appear to show such a pronounced effect. When comparing the month-to-month error with the number of detected COVID-19 cases, considering the limited resources available for accurate case detection in the early months of the pandemic, we observe the following graph.

Figure 3.8
Error in matching price prediction and Covid-19



The pandemic could be a factor that contributed to greater difficulty in predicting the results. But also during that year there is an important change in the electricity market legislation (BOE, 2021) that modifies the minimum and maximum prices as well as various support measures to address energy poverty such as the expansion of the Electricity Social Bonus or the strengthening of the Supply Guarantee Fund. These reasons can serve to justify, at least in part, the irregularity of the year 2021 and therefore the lower efficiency of the predictions.

4. CONCLUSIONS

The conclusions that we can draw are divided into parts: those related to the prediction of supply curves and those related to the prediction of price clearing.

First of all, despite the complexity of the problem, we have been able to predict the supply curves by significantly improving the naive estimator except in one month. For this we used a multioutput Histogram Based Gradient Boosting model fitting it with all the previous data to predict each single month. The information of the previous day's curve and the matching price of the daily market of the same day have been the factors with the greatest influence, although the rest of the factors have a very important weight in the prediction. However, as we can see in figure 3.1, the improvement of these predictions with respect to the naive estimator is larger in the years 2019 and 2020 than in 2021. In general, as we saw in figure 3.3, when the prediction of the curve is not good, there is a tendency for the predicted curve to get closer to the previous day's curve. This fact could indicate that there is some reason why the real curve differs from that of the previous day that might be absent in the input features and whose inclusion among them could improve the performance of the algorithm.

The abnormality on the year 2021 is amplified when estimating the matching prices. For the years 2019 and 2020 we were able to predict the final price with an accuracy greater than 10€/MWh in 95% of the cases. However, 2021 is a year that apparently behaves much more irregularly or at least, not following the same patterns and where the results, although improving the naive estimators, were not as good as in the previous ones. The expansion of the pandemic and the measures taken to stop it could be related with this fact, making 2021 an especially difficult year to predict.

BIBLIOGRAPHY

Alonso, A. M., & Li, Z. (2022). Approximation of supply curves. *Preprint*.

BOE. (2020). No. 335 del jueves 24 de diciembre de 2020, Sec. III. Pág. 120122–120317, Resolución de 10 de diciembre de 2020, de la Comisión Nacional de los Mercados y la Competencia, por la que se aprueba la adaptación de los procedimientos de operación del sistema a las condiciones relativas al balance aprobadas por Resolución de 11 de diciembre de 2019. https://www.boe.es/eli/es/res/2020/12/10/(7)

BOE. (2021). No. 120, del jueves 20 de mayo de 2021. Sec. I, Pág. 61443–61605. Resolución de 6 de mayo de 2021, de la Comisión Nacional de los Mercados y la Competencia, por la que se aprueban las reglas de funcionamiento de los mercados diario e intradiario de energía eléctrica para su adaptación de los límites de oferta a los límites de casación europeos. https://www.boe.es/boe/dias/2021/05/20/pdfs/BOE-A-2021-8362.pdf

endesa.com. (2022). https://www.endesa.com/es/la-cara-e/sector-energetico/como-funciona-el-mercado-electrico-en-espana

energiaysociedad.es. (n.d.-a). https://www.energiaysociedad.es/manual-de-la-energia/1-2-historia-de-la-electricidad-en-espana/

energiaysociedad.es. (n.d.-b). https://www.energiaysociedad.es/manual-de-la-energia/6-5-mecanismos-de-ajuste-de-demanda-y-produccion

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Mestre, G., Portela, J., Muñoz San Roque, A., & Alonso, E. (2020). Forecasting hourly supply curves in the Italian day-ahead electricity market with a double-seasonal SARMAHX model. *International Journal of Electrical Power & Energy Systems*, *121*, 106083. https://doi.org/10.1016/j.ijepes.2020.106083

Shah, I., & Lisi, F. (2020). Forecasting of electricity price through a functional prediction of sale and purchase curves. *Journal of Forecasting*, *39*, 242–259. https://doi.org/10.1002/for.2624

Shapley, L., & Shubik, M. (1954). A method for evaluating the distribution of power in a committee system. *American Political Science Review*, *48*(3), 787–792.

wikipedia.org. (2023). https://es.wikipedia.org/wiki/Mercado_el%C3%A9ctrico_de_Espa%C3%B1a

Ziel, F., & Steinert, R. (2016). Electricity price forecasting using sale and purchase curves: The X-model. *Energy Economics*, *59*, 435–454. https://doi.org/10.1016/j.eneco.2016.08.008

APPENDIX I - WORK SCHEDULE

Below it's the gantt chart concerning the development of this thesis. It must be taken into account that the amount of time available for the thesis is not the same at the beginning as it is at the end, given the asymmetrical distribution of the classes of other subjects.

Figure 4.1

Gantt chart of the work schedule