Day-Ahead Price Forecasting for the Spanish Electricity Market

Álvaro Romero, José Ramón Dorronsoro, Julia Díaz *

Universidad Autónoma de Madrid and Instituto de Ingeniería del Conocimiento, Madrid (Spain)

Received 11 November 2017 | Accepted 20 February 2018 | Published 27 April 2018

Abstract

During the last years, electrical systems around the world and in particular the Spanish electric sector have undergone great changes with the focus of turning them into more liberalized and competitive markets. For this reason, in many countries like Spain have appeared electric markets where producers sell and electricity retailers buy the power we consume. All agents involved in this market need predictions of generation, demand and especially prices to be able to participate in them in a more efficient way, obtaining a greater profit. The present work is focused on the context of development of a tool that allows to predict the price of electricity for the next day in the most precise way possible. For such target, this document analyzes the electric market to understand how prices are calculated and who are the agents that can make prices vary. Traditional proposals in the literature range from the use of Game Theory to the use of Machine Learning, Time Series Analysis or Simulation Models. In this work we analyze a normalization of the target variable due to a strong seasonal component in an hourly and daily way to later benchmark several models of Machine Learning: Ridge Regression, K-Nearest Neighbors, Support Vector Machines, Neural Networks and Random Forest. After observing that the best model is Random Forest, a discussion has been carried out on the appropriateness of the normalization for this algorithm. From this analysis it is obtained that the model that gives the best results has been Random Forest without applying the normalization function. This is due to the loss of the close relationship between the objective variable and the electric demand, obtaining an Average Absolute Error of 3.92€ for the whole period of 2016.

Keywords

Machine Learning, Big Data, Electric Market, Predictive Analysis, Prices, Random Forest.

DOI: 10.9781/ijimai.2018.04.008

I. Introduction

ELECTRIC sector and in particular the Spanish electric market is highly complex but at the same time fundamental to be able to maintain the contemporary way of life. The market pool is where the energy that reaches our homes and industries is purchased and where the electricity produced in our power plants is sold.

The Spanish electricity system has undergone a process of transformation through a liberalization of it that began in 1997. All the tasks related to the supply of electricity such as generation, transport, distribution, retail and economic and technical management of the system, have been separated.

Specifically, the spot market for electricity, managed by the OMIE [20], provides participating agents with the possibility of contracting electricity in seven sessions: the first and main, the Daily Market, and six subsequent sessions, belonging to the so-called intraday market, distributed throughout the day. It is in the first of the sessions, the Daily Market, in which this paper will focus. In this market, a price per hour is established at which each MWh of energy will be sold and purchased.

The way to establish the price follows the algorithm Euphemia that emerged in the initiative "Price Coupling of Regions" (PCR) by seven European electricity markets, among which is the Spanish one.

This algorithm calculates the prices of electric energy efficiently, pursuing the maximization of welfare, which is defined as the surplus or profit, both of buyers and sellers, while optimizing the use of available capacity in interconnections.

For this welfare maximization, for both the daily and intraday markets, the Euphemia algorithm considers aggregate step curves.

In summary, the companies in charge of the generation make their offers (quantity of energy and price) and the companies in charge of retail, direct consumers, etc. demand the necessary energy at a certain price.

Once the bids are made, they are ordered according to price, in increasing order in the case of sales and decreasing order in the case of the purchase. The intersection of the supply and demand curves is called the matching point. In principle, this is the point that optimizes welfare and, therefore, establishes the price of energy for that particular hour.

All the energy offered and demanded at a price less than the matching point will be exchanged at that price, while the one with a higher price will not. This process is repeated for each of the 24 hours of a day.

II. Descriptive Analysis

The price of energy can be affected by many factors that are very complex and in some cases over which there are no data or reflect complex business strategies of companies that are not revealed to the general public.

* Corresponding author.

E-mail addresses: alvaro.romero@iic.uam.es (Á. Romero), jose.r.dorronsoro@iic.uam.es (J. R. Dorronsoro), julia.diaz@iic.uam.es (J. Díaz).

A. Time Series Analysis

The analysis of the time series of prices indicates a strong seasonality in the data due mainly to the effect of the demand. If we analyze the price in the days of the week we can see clearly in the histogram of Fig. 1 that the days with a different price distribution are the weekend days, in which the prices are usually lower. Every day they have most of the prices around €45 but in the case of working days there are a good number of hours with prices higher than those €45, while in the case of Saturdays and Sundays there is a number much more reduced of those hours. One can also see a greater weight of hours with prices below €30 on Saturdays and Sundays, although a little higher in the latter.

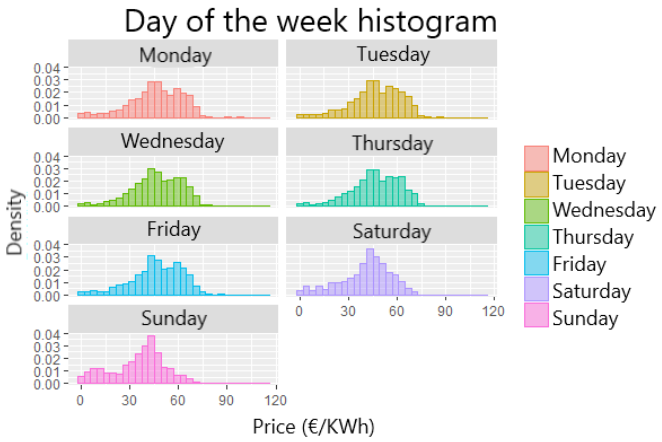


Fig. 1. Histogram by day of the week.

If we analyze the prices on an hourly basis, we can confirm the previous analysis that indicated that there were cheaper hours on weekends than on weekdays. As it happened in the daily case, in the hourly analysis the prices follow the trend of the electricity demand. So, during the night, there is less demand and, therefore, prices are lower. In Fig. 2, which has been made using the hourly average of the prices between 2014 and 2016, it can be seen that the drop in prices goes from 21:00 until approximately 5:00 where prices start to rise until 9:00, very pronounced on weekdays and until 10:00 and, not so pronounced, on weekends. The prices are approximately constant until the lunch hours where they begin to decrease. This valley comprises approximately from 13:00 to 17:00. Finally, prices reach their highest level around 20:00 – 21:00 on weekdays and 21:00-22:00 on weekends.

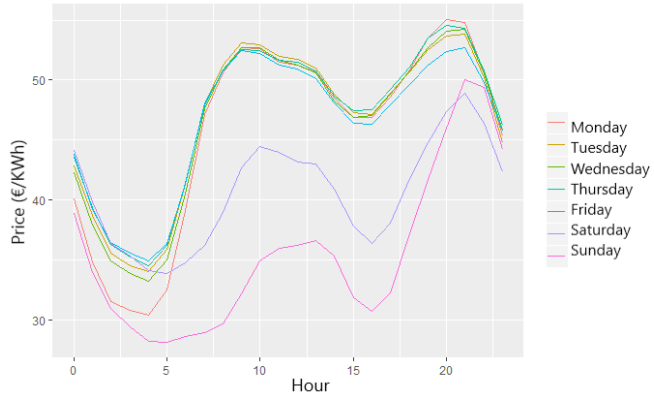


Fig. 2. Hourly mean by day of the week.

If we analyze the prices on a monthly basis as in Fig. 3, we see that in the summer months the prices reduce their hourly volatility and that the differences between weekends and midweek are maintained throughout the different months. The effect of the reduction of hourly volatility in the summer months is probably caused by the effect of refrigeration systems, the effect of tourism and the reduction of production in the industrial sector. Finally, in this price study it is interesting to observe what happens in the holidays since the patterns of electricity consumption in these days in general are clearly different from the working days.

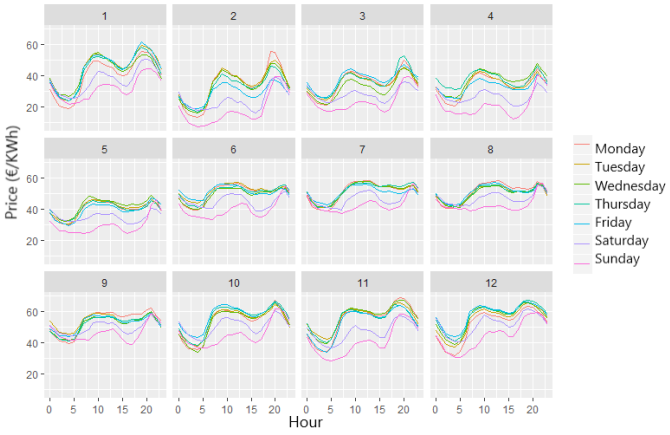


Fig. 3. Hourly Mean by week day and month.

To complete this analysis, only the national and Autonomous Community holidays, which are published in BOE (the Spanish Official Gazette), have been taken into account, but not those related to the individual cities because the effect of a holiday in small regions disappears when it is aggregated with the rest of the Spanish population. As one of the products of this analysis, the holiday coefficient for day i has been defined as the ratio of population in Spain on holidays on day i. In other words,

$$CF_i = \frac{\sum_{C \in \Omega_{fest_i}} P_C}{\sum_{C \in \Omega} P_C} \quad (1)$$

where Ω is the set of Spanish Autonomous Communities, Ω_{fest_i} is the Communities which are on holiday on day i and P_C the population of the Autonomous Community C. This coefficient allows us to see, in Fig. 4, the differences between holidays ($CF \neq 0$), working days ($CF = 0$ and day of the week other than Sunday) and Sundays (in black). It is observed that holidays have lower prices than working days and that in general the higher the holiday coefficient (therefore, the greater percentage of the Spanish population is on vacation) the lower the price.

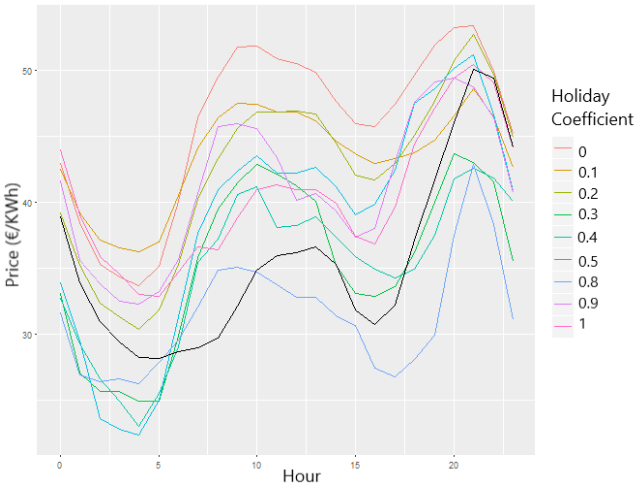


Fig. 4. Comparison of the price per hour among different holiday coefficients.

B. Variable Analysis

The variables for the price prediction problem most used in the literature [1] can be divided into the following categories: generation of renewables, electric demand, exports/imports, other countries' price, weather variables, economic variables, day type and hour.

Renewables generation. Renewables in Spain participate in the market at price 0 hence, they always lower the prices. Therefore, it is important to take them into account when predicting the price. Now, we will not have the real data for the day for which we have to give the prediction. Therefore, it is necessary to make predictions of production that will be those that enter as variables to the model. The two most important renewable generation technologies in Spain and which, therefore, should be taken into account as input variables for the model are:

- Wind energy.
- Solar energy (photovoltaic and thermosolar)

Demand. As well as the production of renewables, demand is a variable that directly affects energy prices since the more demand there is, the higher purchase offers and, therefore, the price rises.

Exports/Imports. Exports and imports are the interchange of electric energy between two countries, in the Spanish case, mainly with France. This interchange affects the price because there is more or less energy in the market and they occur due to the price of electricity in other countries.

Other Countries' Price. Because the Euphemia algorithm takes into account the price of other countries this variable is very important. Specifically, for the Spanish case the one that affects most the Spanish price is the French one, which is the main exporter of electricity to Spain. For this reason, as we will see later, there is a great correlation between the French and Spanish prices.

Weather variables. The meteorological variables affect both the production of renewables and the demand and, therefore, it will be necessary for the price predictions the usage of meteorological variables.

Economic variables. There are economic variables that affect the price of energy because they are related to demand and the production of non-renewable energy such as GDP, the price of gas and oil, etc. The difficulty of finding these variables for the necessary period of study and with sufficient granularity have meant that they are left for future work. In addition, these variables usually explain longer-term components, since they affect periods of months or even years and in this work we are more focused on the short term.

Type of day and time. The price depends on the type of day we face. This can be intuited easily, because the habits of electricity consumption are not the same in winter than in summer, or in different hours, etc. That is why all this must be taken into account when creating the model. In this study the following variables will be used:

- Hour.
- Weekday.
- Day of the month.
- Month.

In addition to the primary variables discussed above, other variables that may be of great interest have been generated in this work. In the first group of these variables are the past values of the price of electricity in Spain and France. These variables represent what happened at the same time one, two and three days before. These values are very important because of the hourly seasonal component that the price has and that was detailed before. There is a second group in which temperature appears. The temperature was available in the capitals of each province of Spain. To have a single value for the whole country, a weighted average based on population has been created so that the temperature of bigger provinces such as Madrid or Barcelona will have more relevance in the variable than temperatures in smaller ones such as Teruel or Soria.

The temperature has an effect on consumption that is not linear. Therefore, in this work the following temperature definitions have been studied:

$$\boldsymbol{T_2} = |T - \bar{T}| \tag{2}$$

$$\boldsymbol{T_3} = \begin{cases} 0 & T < 10 \\ T - 10 & T \geq 10 \end{cases} \tag{3}$$

$$\boldsymbol{T_4} = \begin{cases} -T + 15 & T < 15 \\ 0 & 15 \leq T < 20 \\ T - 15 & T \geq 20 \end{cases} \tag{4}$$

$$\boldsymbol{T_5} = \begin{cases} -T + 10 & T < 10 \\ 0 & 10 \leq T < 15 \\ T - 15 & T \geq 15 \end{cases} \tag{5}$$

$$\boldsymbol{T_6} = \begin{cases} -T + 5 & T < 5 \\ 0 & 5 \leq T < 20 \\ T - 20 & si\ T \geq 20 \end{cases} \tag{6}$$

$$\boldsymbol{T_7} = \begin{cases} -T + 5 & T < 5 \\ 0 & 5 \leq T < 25 \\ T - 25 & T \geq 25 \end{cases} \tag{7}$$

Here T is the single value of temperature explained above and $\bar{T}$ the average of this temperature during the whole period.

With the analysis of correlations, we can highlight the greater importance of some variables that we have commented previously. In descending order, the variables with the highest linear correlation are: the price in Spain 24, 48 and 72 hours before, wind generation, the price in France and electricity demand.

III. Predictive Analysis

This paper intends to apply some of the most used techniques in regression problems, more specifically, those that have been recently used in the field of predicting the price of energy with certain changes in the treatment of data. This publication serves as a modern benchmark to be measured against, since electricity markets have changed enormously in recent times.

As can be seen in [1] and in [2] the approaches taken by different authors in the past for the problem of predicting the price of energy have been diverse. These methods range from Game Theory to Computational Intelligence, through Simulation Models and Time Series.

In our study we will focus mainly on techniques related to Time Series and Computational Intelligence because these techniques are those that have been supported by researches such as [3] where it is indicated that the Economic and Game Theory methods are a good approximation but certainly insufficient in case of wanting to make precise short-term predictions.

A. Time Series

Time Series are successions of values spaced in constant periods of time; that is, the phenomenon is observed in moments taken regularly.

The analysis. of time series aims to model the underlying temporal

structure in the observations taken in a certain period of time. Once the modeling is done, these algorithms serve to understand the time series but also to predict their behavior in the future. In the bibliography there are many references in which these models are used for price prediction such as [4] which uses Leipzig Power Exchange data for their experiments and ARMA models with some modifications. The work in [5], one of the first predictions of the price of electricity in Spain after the regulatory change, makes use of a Seasonal ARIMA model with different parameters for the Spanish and Californian market. In [6] and in [7] an improvement is proposed using a wavelet transformation prior to the use of the ARIMA model to reduce the volatility of the time series that is applied to electricity price data; In general, this transformation offers better results than a traditional ARIMA. There are also numerous studies that combine (S) ARIMA models with other prediction models, that use exogenous variables [8] or that make a model per day of the week or even per hour.

B. Machine Learning Models

Machine Learning is a field of Artificial Intelligence whose definition is complex and in which the different authors do not agree but we could define it as the subject that studies the techniques and algorithms that allow machines to adapt to dynamic situations and, therefore, somehow learn to predict the future, from discovering underlying patterns in the data.

The focus of this work is on supervised algorithms (data are fully labelled); more specifically, the case of price prediction belongs to a type of supervised models called regression models because the label to predict is a real value that goes, in this case, from 0 up to €180,3 (which are the minimum value and the maximum value for the prices in the Spanish market).

Some of the most used techniques for regression problems are Multiple Linear Regression, Decision Trees, K-Nearest Neighbours, Support Vector Machines, Neural Networks and ensembles that use some of the above models together to get better predictions.

1) Ridge Regression

Traditional Linear Regression has problems when there is not independence among the variables. Specifically, when there is collinearity Linear Regression does not work correctly. To avoid this, or somehow eliminate these collinearities with there are several techniques. For example, the method of Analysis of Principal Components, dimensionality reduction technique that generates orthogonal variables [9]. Another widely used resource is the use of Ridge Regression, proposed by Hoerl and Kennard [10] which introduces a regularization term in order to avoid overfitting and underfitting.

2) Decision Trees

Decision trees are a nonparametric supervised method that can be used for both classification and regression. It is a method widely used and described in depth in different references as, for example, [11].

The objective of this algorithm is to create a model that predicts the value of the objective variable by learning basic rules inferred from the variables of the data and that define regions whose edges are always parallel to the axes. Within each region a simple function is assigned, sometimes a constant.

3) K-Nearest Neighbours

The method of the K-Nearest Neighbours (K-NN) is based on inferring the variable to predict using the K cases in the training set that are more similar to the new data. The number K of neighbours to use in training can be defined by the user and, therefore, must be hyperparametrized since changing the number of neighbours to use can improve or worsen the results of the algorithm.

4) Neural Networks

Artificial Neural Networks are a type of Machine Learning algorithms that are inspired by the neuronal functioning of living beings. A Neural Network contains several processing units that connect to each other forming different architectures. Each unit or artificial neuron simulates the functioning of a neuron: it is activated if the total amount of signal it receives exceeds its activation threshold. In this case, the node is activated and emits a signal to the rest of the adjacent neurons. Therefore, each unit becomes a transmitter of the signal that can increase or decrease said signal.

Neural Networks are widely used for the problem of price prediction. The most relevant papers for this study are:

- [12], which is one of the first studies on the subject and in which an architecture is used with 15 input parameters, 15 hidden units and 1 output to predict the price for the Victorian Power System.
- [13] where data from the Spanish market are used to make a comparison between several models; in particular, one of the proposed is a multilayer perceptron with an architecture of one hidden layer and making use of wind and demand as predictor variables.
- [14] is also relevant where they use a combination of networks to predict the maximum, minimum and average value that is finally provided to 5 main neural networks to predict the price.

5) Support Vector Machines

The SVM is a model that started being used for classification and that generates a hyperplane that separates the two classes in an optimal way. In the case of regression, it is usually called SVR and its basic idea is to map the training data to a high dimension feature space through a non-linear mapping where we can perform a linear regression using a special loss function called ε-intensive loss.

SVMs are used for the prediction of prices in a large variety of publications, as in [15] which makes a comparison between Neural Networks and SVM or as in [16] that uses them to make accurate predictions and also to provide a confidence interval. There are several proposals for hybrid models using SVMs as in [17] that makes a hybrid model with SVM to capture non-linear patterns and ARIMA; this hybrid model is called in this publication SVRARIMA.

6) Random Forests

The Random Forest algorithm is a very effective ensemble of trees and is widely used today. Each tree is generated by random sampling with replacement of the original train set. The algorithm for Random Forest can be written as follows [18]:

1. Subsamples of the original data are chosen by bootstrapping.
2. For each one of the sub-samples, a regression tree is built without pruning, but in each node, instead of choosing the best possible cut among all the attributes, a subset of them is chosen randomly.
3. New data are predicted by the aggregation of all individual predictions using a mean in the case of regression.

For this method there are not many references for price prediction and most are very modern; in particular, [19] can be highlighted and it uses Random Forest for the prediction of the price in the Electricity Market of New York and uses as predictors the temporary series of prices itself lagged 3, 24, 168 and 720 hours, the demand, the temperature and a day of the week indicator.

IV. Training and Prediction Process

In order to obtain the most accurate model possible for predicting the price of the Spanish market, the following procedure was

applied in this work: a normalization of the target variable and a normalization of the rest of the variables for those models that need it, a hyperparameterization of the selected models, a study of the errors and a deeper study of the best model to diminish its errors.

A. Data Normalization

Because with the study of the time series we discovered a great difference between working days and non-working days, a normalization of the price is proposed to achieve a reduction in variance. For this we must define two types of standardization: normalization of non-working days and hourly normalization. On the one hand, the normalization of non-working days conceptually aims to eliminate the effect of a decrease in the price of non-working days caused by the decrease in work activity. Mathematically normalization of non-working days can be defined as:

$$p'_{lh} = p_{lh} - \overline{p_{lh}} \tag{8}$$

$$p'_{sh} = p_{sh} - \overline{p_{sh}} \tag{9}$$

$$p'_{dh} = p_{dh} - \overline{p_{dh}} \tag{10}$$

$$p'_{fh} = p_{fh} - \overline{p_{fh}} CF_h \tag{11}$$

where p_{lh} is the hourly price of the working days, p_{sh} is the hourly price of the Saturdays, p_{dh} is the hourly price of the Sundays, p_{fh} the hourly price of the holidays and CF is the holiday coefficient. And Hourly normalization is then defined as:

$$p'_h = p_h - \overline{p_h} \tag{12}$$

where p_h is the price at hour h and $\overline{p_h}$ is the mean of every price at hour h.

For some algorithms is also important to consider the normalization of the rest of the variables. This normalization has been carried out by tuning the normalization function.

Therefore, for all the models in which it is necessary to normalize the attributes in some way, the normalization function has been hyperparametrized, always doing a grid search on the following four types of normalization:

- **MaxAbs scaler**. Scale each attribute by the maximum value that attribute can take.
- **Robust scaler**. This method uses robust statistics. Therefore, it subtracts the median and use the interquartile range to scale the data.
- **Standard scaler**. It substracts the mean and scales to obtain unit variance.
- **MinMax scaler**. It is usually used as an alternative to the previous one and is mathematically defined as follows:

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)} \tag{13}$$

B. Parameter tuning

In the process of choosing the best model there are two jobs; on the one hand, the selection of the prediction algorithm and, on the other hand, we need to find the best values for the several hyperparameters that the different algorithms might receive.

To carry out the selection of the best parameters, the hyperparametrization process is applied, which basically performs trainings with a given set of train data and predictions for the validation set. In this case, because we are working with temporal data, we have opted for the use of a temporary validation following the scheme of Fig. 5.

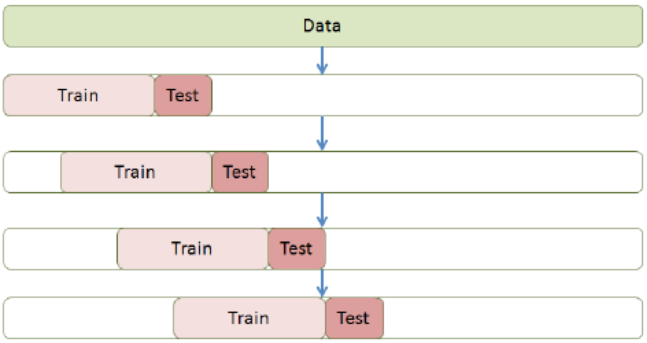


Fig. 5. Validation Workflow of this work.

In this scheme a cross validation is carried out in a continuous way so that it is impossible to take values from the future to predict past values as it could happen in a usual cross-validation as we see in Fig. 6.

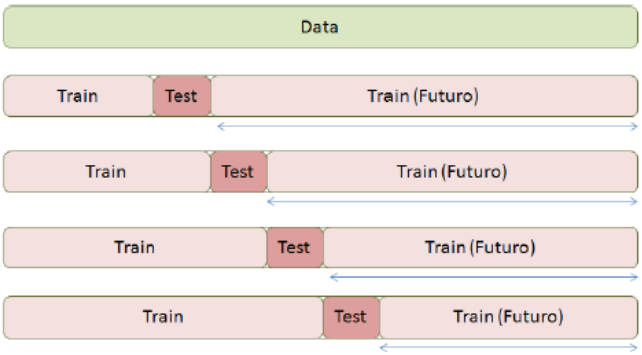


Fig. 6. Traditional cross validation.

Also, it should be noted that the test period is always a full day and the number of days taken for each train period has been hyperparametrized, going from 10 days at least up to 1 year, to then choose the most appropriate one. In general, the results show that it usually affects the chosen train period, being optimal for 7-9 months approximately. This is due to the fact that if one does not consider enough cases to train the model, it does not have enough information to generalize but one cannot take a very long period back because it would find cases that may have been produced by past macroeconomic facts that we are not taking into account in this work.

V. Results and Discussion

In this section we will discuss in detail the results of all the tests created for the prediction of the price in Spain. The period of validation is 2016 and the process for every model has been the same, training and predicting every day. For the benchmark, more than 20,000 models have been created with the different parameters discussed above.

A. Benchmark

In particular, the best results of each of the models are presented next: In the case of the Ridge Regression, the parameters that have given the best result have been $\alpha = 1.01$, with 90 days for training and normalizing the data with maxAbs normalization. With this configuration we see that the average absolute error in validation is around €5.73 (see Table I).

TABLE I.
MAE BY MONTH AN ANNUAL FOR THE BEST RIDGE REGRESSION MODEL

Jan	Feb	Mar	Apr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dec	Total
5.79	7.07	5.40	5.87	6.67	5.49	5.92	5.54	4.65	5.90	4.98	5.56	5.73

With the K-NN we find the best conjuration makes use, in the same way as Ridge Regression, of 270 days for the training and the robust function for the normalization of the data and 20 neighbours. With these parameters a reduction of the MAE with respect to the previous value is achieved, obtaining €5.30 in validation (see Table II).

TABLE II. MAE BY MONTH AN ANNUAL FOR THE BEST K-NN MODEL

Jan	Feb	Mar	Apr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dec	Total
5.70	7.71	5.35	5.41	5.21	6.19	4.85	4.41	3.49	5.34	5.53	4.57	5.30

In the case of the Multilayer Perceptron, using the regularization parameter $\alpha = 1$, the robust normalization and, as has been said before, two hidden layers with 50 neurons, an error of €5.29 in validation is obtained (see Table III).

TABLE III. MAE BY MONTH AN ANNUAL FOR THE BEST MLP MODEL

Jan	Feb	Mar	Apr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dec	Total
4.73	6.01	5.19	5.19	5.58	4.83	5.23	5.94	4.35	5.40	5.89	5.20	5.29

In the case of Support Vector Machines for Regression, an error minimization of €4,92 in validation is reached (see Table IV), when C takes the value of 100 $\varepsilon = 1$, $\gamma = 1$, with normalization minMax and with 210 days for training, where C is the penalty parameter of the error term, γ is the kernel coefficient and ε specifies the epsilon-tube within which no penalty is associated with points predicted within a distance ε from the actual value.

TABLE IV. MAE BY MONTH AN ANNUAL FOR THE BEST SVR MODEL

Jan	Feb	Mar	Apr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dec	Total
5.63	6.60	4.06	4.75	4.77	4.60	4.44	4.00	5.22	4.62	5.69	4.63	4.92

Finally, the model that has given the best result has been the Random Forest with n_estimators = 710, min_samp_leaf = 2, max_feat = 0.4 and 270 training days, where n_estimators. Is the number of trees, min_samp_leaf is the minimum number of samples to be in a leaf node and max_feat is the number of features that considers in each split. With all this, a MAE of €4,44 in validation has been achieved.

B. Best Model Analysis

One of the most interesting properties of the best model, Random Forest, is that it allows us to calculate the relevance of the variables. This relevance is taken as the number of times that attribute appears in each of the created trees. Therefore, in our case, as we do a daily training, in each one we can calculate the variable frequency of appereance in each tree. Throughout the validation year, we have 365 frequencies for each variable, so if we take the average of all that series we have the average frequency of each variable. With these means we have a good measure of how important each variable is throughout the entire year. In Fig. 7, Fig. 9 and Fig. 12 the variables are numbered as follows:

0. Demand
1. Solar production.
2. Wind Production.
3. Price in Spain 24 hours before.
4. Price in France 24 hours before.
5. Price in Spain 48 hours before.
6. Price in France 48 hours before.
7. Price in Spain 72 hours before.
8. Price in France 72 hours before.
9. Holiday coefficient.
10. Month.
11. Day of the month.
12. Day of the week.
13. Time.
14. Week of the year.
15. Type of day. That takes the values of 1 if it is Saturday, 2 if it is Sunday, 3 if it is a holiday and 0 in other days.
16. Working. That takes the values of 0 if it is Saturday, Sunday or holiday and 1 the other cases.
17. T
18. T2
19. T3
20. T4
21. T5
22. T6
23. T7

1) RF with Normalized Price

Using the technique described above to calculate the relevance of the variables, we can see that the most important ones are wind energy, the price in Spain 24 hours before and the temperature in the T5 version. The low relevance that the demand obtains is surprising because the intersection point between demand and generation will be higher or lower depending on the total amount of energy demanded. This is clearly due to the normalization explained before made to the target variable; in this normalization the effect of the hours and the days of the week is eliminated and if it is not applied in the same way to the demand it causes the demand to stop being related to the prices.

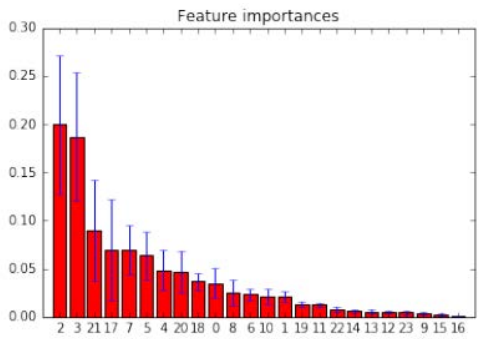


Fig. 7. Feature importance with price normalized.

If we observe the comparison between the hourly chart of demand and the price when we normalize in the manner described before, we see in Fig. 8 that any relationship between both variables has been lost. Therefore, it is necessary to consider what happens when we normalize both variables following the same method, analysis that is carried out in the following subsection.

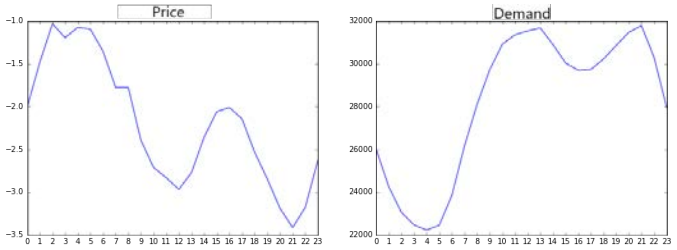


Fig. 8. Comparison between the hourly averaged Demand and the hourly averaged Normalized Price.

2) RF with Demand and Price Both Normalized

If we re-execute the hyperparametrization by normalizing the demand, we see that the results improve somewhat with respect to the previous case, not only in total, but also month by month, as we can see in Table V.

TABLE V. MAE BY MONTH FOR RFR WITH DEMAND AND PRICE NORMALIZED

Jan	Feb	Mar	Apr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dec	Total
5.03	6.19	4.00	4.66	5.29	4.29	3.90	3.45	2.99	3.90	4.22	3.55	4.28

In addition, if we look at Fig. 9, where the most important variables of the best model are calculated, we observe that the demand is now in sixth position and the wind generation and the price 24 hours before are still in the lead.

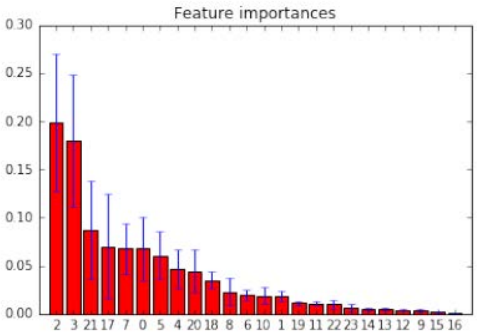


Fig. 9. Feature importance with normalization on both demand and price variables.

Somehow, by normalizing both variables in the same way, we are better preserving the common structure that they have. In addition to checking the importance, we again review the comparison of the two new demand and price variables; in Fig. 10, we observe that although the time relation between both is better preserved, this one is improvable if we compare it with the existing relation between the two variables without normalizing as we see in Fig. 11.

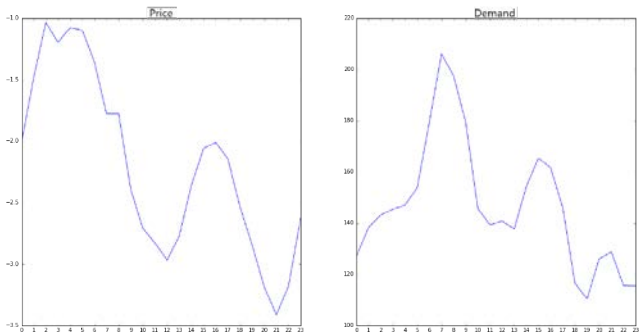


Fig. 10. Comparison between the hourly averaged demand and the hourly averaged price, both normalized.

3) RF without Normalization

When not normalizing, in the comparison between the hourly evolution of both variables of Fig. 11, a strong relation between both is observed.

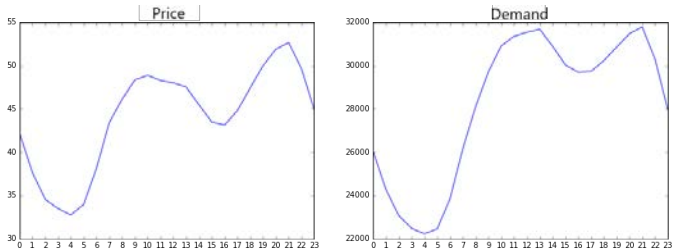


Fig. 11. Comparison between the hourly averaged demand and price.

With the hyperparametrization of Random Forest using the data without normalizing one can observe a remarkable improvement. Therefore, we can conclude that the best option is not to normalize when using Random Forest not to lose any underlying relationship between the different attributes and the price even though we have a greater variance in the target variable. Regarding the importance of the variables, they have changed radically. The demand is now in first place followed by the price 24 hours before and the wind generation, and the others have much less relevance, as we can see in Fig. 12.

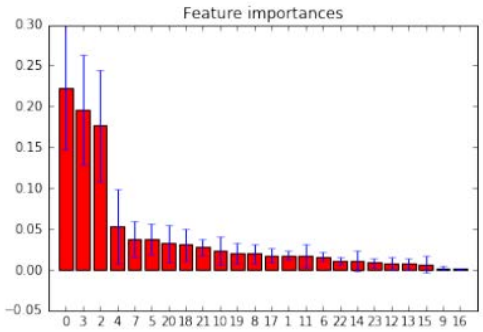


Fig. 12. Feature importance with no normalization.

By adding this improvement, we can observe a greater precision in each of the hours of the months, appreciable in the summary of the whole year that we have in Fig. 13, where softer colors can be observed in the case where the demand is normalized and even softer in the heat map corresponding to the data without normalizing. In addition, with these figures we can understand that in general there are some hours and months for which it is more difficult to predict. For example, in the summer months, especially in August and September, errors are much lower for all hours of the day than the rest of the year.

The errors of the best resulting model, Random Forest without normalization, are found summarized in Table VI.

TABLE VI. MAE BY MONTH FOR RFR WITHOUT NORMALIZATION

Jan	Feb	Mar	Apr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dec	Total
4.67	6.67	3.80	3.75	4.32	4.53	3.00	2.34	2.83	4.01	4.24	2.93	3.92

They are difficult to compare with those obtained by other authors in part because a large number of publications are old and the market has profoundly changed as it happens with [17], [6] or [5]. In fact, the appearance of participating agents has increased in great quantity in the last two or three years.

On the other hand, in many publications measures of certain months, weeks or even days are taken. For example, [6] and [5] measure the

error using weeks of different periods. For its part, [19] measures the error on certain days of June. Even other publications such as [15], [17] make predictions for markets other than Spanish that are difficult to compare.

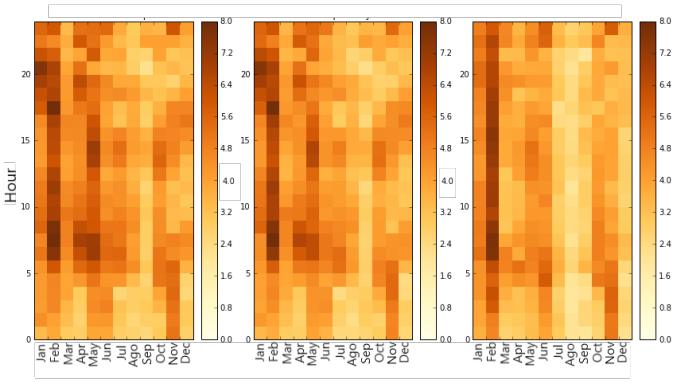


Fig. 13. Summary of errors by month and hour.

C. Results with Test Data of RF without Normalization

As the test data, the available period of 2017 has been reserved, that is, from January to July, both included. In it, the use of the best model, Random Forest has been tested without normalizing, so that it is trained every day and predicts for the next day. Following this methodology, the result is somewhat worse than for the evaluation period as is usual in this type of problem. This is due to the fact that in some way when performing the hyperparametrization the model is being adjusted to learn the validation set.

More in detail, the test error is €4.50 and broken down by months, we can see it in Table VII, where we observe that the error is degraded, especially since April. Despite this, in the actual operation of a product like this, there are two maintenance elements of the model that would improve their predictions. In the first place, there would not be a single model, but there would be several that are in follow-up and that periodically could change because there is a model that is offering better results than the one that is in production. Secondly, that starting in April, the model gets worse. It may be due to a need for re-hyperparametrization.

TABLE VII.
TEST MAE BY MONTH FOR RFR WITHOUT NORMALIZATION IN 2017

Jan	Feb	Mar	Apr	May	Jun	Jul	Total
3.30	2.92	4.60	6.28	4.60	8.01	3.25	4.50

VI. Conclusion and Future Work

In this work, different models have been implemented with the final objective of solving the problem of price prediction in the electricity market. For this, several stages have been followed.

In the first place it has been necessary to review exhaustively the different models and variables present in the bibliography that have been used for price prediction. It should be noted the difficulty in the comparison with the publications cited in this paper because of the temporal and geographic difference of the publications regarding this work.

For the choice of the model, a benchmark has been carried out between different types of models and parameters among which are Ridge Regression, K-Nearest Neighbours, Multilayer Perceptron, Support Vector Machines and Random Forests. Of all the combinations of models and parameters, the most precise has been Random Forest and for it a more detailed study has been carried out, including an analysis of the most relevant variables, a comparison of different types of normalization and an exploration of errors by months and hours.

In this work we have also seen that if we normalize the time series of prices to reduce their variance, it can cause a loss of information about the underlying patterns in the data that are very useful for prediction. In particular, it has been observed how normalization blurred the strong relationship that the price has with the demand, which is partially solved by applying the same standardization treatment to both variables. Despite this improvement, the results provide an unfavourable outcome to normalization since without it, the best results are obtained.

The best predictive model achieved, has managed to obtain a MAE in validation of 3.92 and of 4.50 in test, a result that is already useful for all the agents that participate in the market.

On the other hand, a possible future line of research is the influence of economic factors, among which one can include the price of the most used raw materials for obtaining energy (coal, oil and gas), the Gross Domestic Product, etc.

Another very important aspect to consider is the interconnections with the bordering countries. In the case of Spain, the most important connection is the French one and, therefore, the price in France was included but in addition to this price, the technical capacity of exports and imports as well as the quantities imported/exported are very important.

Labour is a key aspect in the prediction of prices due to the effect of holidays, Saturdays and Sundays on the demand for electricity. Therefore, this is a key aspect that should be further investigated. In this work we have tried to solve it through a normalization that has been proven as not very useful. Therefore, other lines of research in this regard are, on the one hand, to perform a post-process of key days (holidays, Christmas and summer time, bank holidays, etc.) and, on the other, to make separate models for those days, with the problem of the scarcity of data.

Lastly, the most important and at the same time the most difficult aspect to include in the model is the strategy of the agents in the market. Despite its difficulty, there are data that are published by the OMIE that could be used to solve this problem.

Acknowledgments

With partial support from Spain's grants TIN2016-76406-P and S2013/ICE-2845 CASI-CAM-CM. Work supported also by project FACIL--Ayudas Fundación BBVA a Equipos de Investigación Científica 2016 and the UAM--ADIC Chair for Data Science and Machine Learning.

References

[1] Sanjeev Kumar Aggarwal, Lalit Mohan Saini, and Ashwani Kumar. Electricity price forecasting in deregulated markets: A review and evaluation. International Journal of Electrical Power & Energy Systems, 31(1):13-22, 2009.

[2] Rafal Weron. Electricity price forecasting: A review of the state-of-the-art with a look into the future. International journal of forecasting, 30(4):1030-1081, 2014.

[3] Derek W. Bunn. Forecasting loads and prices in competitive power markets. Proceedings of the IEEE, 88(2):163-169, 2000.

[4] Jesús Crespo Cuaresma, Jaroslava Hlouskova, Stephan Kossmeier, and Michael Obersteiner. Forecasting electricity spot-prices using linear univariate time-series models. Applied Energy, 77(1):87-106, 2004.

[5] Javier Contreras, Rosario Espinola, Francisco J. Nogales, and Antonio J. Conejo. Arima models to predict next-day electricity prices. IEEE transactions on power systems, 18(3):1014-1020, 2003.

[6] Antonio J. Conejo, Miguel A. Plazas, Rosa Espinola, and Ana B. Molina. Day-ahead electricity price forecasting using the wavelet transform and arima models. IEEE transactions on power systems, 20(2):1035-1042, 2005.

[7] Chang-il Kim, In-Keun Yu, and YH Song. Prediction of system marginal price of electricity using wavelet transform analysis. Energy Conversion and Management, 43(14):1839-1851, 2002.

[8] Rafal Weron, Adam Misiorek, et al. Forecasting spot electricity prices with time series models. In Proceedings of the European Electricity Market EEM-05 Conference, pages 133-141, 2005.

[9] Manuel Gurrea. Análisis de componentes principales. Proyecto e-Math Financiado por la Secretaría de Estado de Educación y Universidades (MECD), 2000.

[10] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1):55-67, 1970.

[11] Christopher M. Bishop. Pattern recognition and machine learning. Springer, 2006.

[12] BR Szkuta, L. Augusto Sanabria, and Tharam S. Dillon. Electricity price short-term forecasting using artificial neural networks. IEEE transactions on power systems, 14(3):851-857, 1999.

[13] Alberto Cruz, Antonio Muñoz, Juan Luis Zamora, and Rosa Espínola. The effect of wind generation and weekday on Spanish electricity spot price forecasting. Electric Power Systems Research, 81(10):1924-1935, 2011

[14] Raquel Gareta, Luis M. Romeo, and Antonia Gil. Forecasting of electricity prices with neural networks. Energy Conversion and Management, 47(13):1770-1778, 2006.

[15] Damien C. Sansom, Tom Downs, Tapan K. Saha, et al. Evaluation of support vector machine based forecasting tool in electricity price forecasting for Australian national electricity market participants. Journal of Electrical & Electronics Engineering, Australia, 22(3):227, 2003

[16] Jun Hua Zhao, Zhao Yang Dong, Zhao Xu, and Kit Po Wong. A statistical approach for interval forecasting of the electricity price. IEEE Transactions on Power Systems, 23(2):267-276, 2008.

[17] Jinxing Che and Jianzhou Wang. Short-term electricity prices forecasting based on support vector regression and auto-regressive integrated moving average modeling. Energy Conversion and Management, 51(10):1911-1917, 2010.

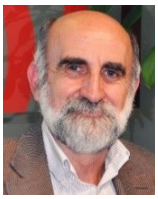
[18] Andy Liaw, Matthew Wiener, et al. Classification and regression by random forest. R news, 2(3):18-22, 2002.

[19] Jie Mei, Dawei He, Ronald Harley, Thomas Habetler, and Guannan Qu. A random forest method for real-time price forecasting in New York electricity market. In PES General Meeting| Conference & Exposition, 2014 IEEE, pages 1-5. IEEE, 2014.

[20] "OMIE" http://www.omie.es/inicio

Álvaro Romero Miralles

Álvaro Romero Miralles holds a Master's Degree in Computer Engineering and a Master's Degree in ICT Research and Innovation in Computational Intelligence a Degree in Mathematics and a Degree Computer Engineering from Universidad Autónoma de Madrid. Currently he works as Data Scientist and Project Manager at Health and Energy Predictive Analytics group of Instituto de Ingeniería del Conocimiento (IIC), a private Big Data R&D institution. He has experience in fraud detection, predictive maintenance, optimization problems among others. He collaborates as a professor in different business schools such as MBIT School and ENAE Business School.



José Dorronsoro

José Dorronsoro (PhD, Washington University in St Louis; USA) is Professor of Computer Engineering at the Universidad Autónoma de Madrid. He has authored more than 100 scientific papers in mathematical analysis, machine learning and applications and has directed a large number of research and innovation projects. Dr Dorronsoro is also a senior scientist at the Instituto de Ingeniería del Conocimiento (IIC), where he works on research and innovation on renewable energy.

Julia Díaz

Julia Díaz is Ms Degree in Mathematics, PhD in Computer Science both from Universidad Autónoma de Madrid (UAM-Spain) and General Management Program from IESE-Universidad de Navarra (Spain). At present she is Senior Innovation Manager in a private R&D+i institution named Instituto de Ingeniería del Conocimiento (IIC-UAM) dedicated to extracting knowledge on the basis of high volumes of heterogeneous data (Big Data) and optimizing business processes in areas such as healthcare and energy. She also is Part Time PhD Professor in Computer Sciences in the UAM and Professor in Big Data & Data Sciences Master in UAM and ESADE.