Investigation of Price-Feature Selection Algorithms for the Day-Ahead Electricity Markets

Radhakrishnan Angamuthu Chinnathambi, Mitch Campion, Arun Sukumaran Nair, Prakash Ranganathan
Department of Electrical Engineering
University of North Dakota
Grand Forks, ND, USA
prakash.ranganthan@und.edu

Abstract— **This paper investigates three types of feature selection techniques such as relative importance using Linear Regression (LR), Multivariate Adaptive Regression Splines (MARS), and Random forest (RF) to reduce the forecasts error for the hourly spot price of the Iberian electricity markets. Two pricing datasets of durations three and six months were used to validate the performance of the model. Three different set of features (17, 4, 2) for three and six months duration were used in this study. These selected features were applied to the two-stage hybrid model such as ARIMA-GLM, ARIMA-SVM, and ARIMA- RF. Finally, three variables (or features) that are commonly matched were selected and tested. Considerable reduction in Mean Absolute Percentage Errors (MAPE) values were observed for both three and six-month datasets.**

Index Terms-- **Feature Selection, Relative importance, MARS, Random forest, Day-ahead price forecast, Iberian Market.**

I. Introduction

Feature or variable selection [1] is an important step to improve the accuracy of the predictive model. It is widely used to reduce the computation time and discard the features that have a high correlation among them. Feature selection is also known as variable selection or attribute selection. It is the automatic selection of the features that are very relevant to our problem. Feature selection is different from dimensionality reduction. While feature selection includes or excludes the variables, dimensionality reduction creates a new combination of features. Some of the examples of dimensionality reduction methods are Principal Component Analysis (PCA), Singular Value Decomposition (SVD) and Sammon's mapping.

Feature selection methods act as a filter muting out important features that are highly irrelevant to the data. They help us to create an accurate predictive model by choosing features that give better accuracy whilst requiring fewer data samples. Feature selection helps us to find the unwanted, redundant features that do not contribute to the predictive model or in fact can reduce the performance of the model. Fewer variables are desired in the predictive model since it reduces the complexity by reducing the computation time. Model with fewer variables is easy to comprehend and explain. In short feature selection can be summarized as follows: improving the prediction performance of the predictors, selecting predictors that are faster and efficient.

In this paper, three different feature selection approaches are used to discard the irrelevant features for the day-ahead price forecasts of the Iberian electricity markets. In, [2], the authors have previously used 17 variables for the same three and six months of the dataset. The authors have tested a two-stage hybrid model (ARIMA-GLM) that performs better than the conventional techniques such as ARIMA, random forest (RF) and Support vector machines (SVM). In [3], the authors have used the same Iberian market dataset to predict the day-ahead load using deep neural networks (DNN). The results have indicated that the DNN models performed better than the conventional techniques such as ARIMA, RF and SVM. In [4], the authors have used Multi-stage hybrid model such as ARIMA-RF, ARIMA-SVM, and ARIMA-GLM to predict the day-ahead price for the same Iberian market. The result shows that the ARIMA-GLM combination performs better for longer duration periods, while ARIMA-SVM combination performs better for shorter duration periods. In [5], the authors have used the same Iberian market to predict the day-ahead price using DNN. The results indicated that for 17 variables, as the no of layers were increased in the model, it performed well for the complex relationship between the 17 independent and dependent variables. For 4 and 2 variables, as layers were increased, inconsistencies were observed.

The main motivation behind this paper is to improve the prediction accuracy by selecting important features/variables that are very relevant to the day-ahead electricity market. A 1% improvement in the MAPE (Mean Absolute Percentage Error) will save $300,000 in savings per year per GW peak load for the day-ahead price forecast.

In [6], multivariable mutual information is applied for feature selection to select the appropriate features for the price forecasting. In this paper, support vector regression (SVR) is applied and the experimental results showed that these feature selection methods perform accurate prediction. In [7], feature selection techniques are compared and analyzed. It is used as a filter prior to forecasting method. The popular search methods such as Best-First Search, Greedy-Step Wise Search, Exhaustive Search, Genetic Search, Random Search, and Ranker is compared with the proposed feature selection technique. In[8], new feature selection method is presented and a hybrid filter-wrapper approach is proposed.

978-1-5386-5419-4/18/$31.00 ©2018 IEEE

The main contribution of the paper is the use of feature selection approaches such as Relative importance using Linear Regression (LR), Multivariate Adaptive Regression Splines (MARS), and Random forest (RF) for multi-stage hybrid forecasting techniques with ARIMA. Features selected from these approaches were used to predict the day-ahead price for the Iberian electricity market using different hybrid techniques such as ARIMA-RF, ARIMA-SVM, and ARIMA-GLM. These techniques were investigated for various duration of datasets such as three months (3M) and six months (6M).

The remainder of this paper is organized as follows: section 2 discusses the features/variable selection methods used in this paper. Section 3 focuses on features selected from different approaches for different datasets. Section 4 provides feature selection results and discussion. Section 5 presents the concluding remarks.

II. Feature selection methods

Finding the best feature selection approach that best explains the variance in the dependent variable or response variable is the key for building a high-performance predictive model.

A. Relative Importance by Linear regression

A linear regression can be used to identify the key variables in the model. This method identifies the important variables as a relative percentage. The first step involves fitting the linear regression model using the given set of predictor variables. Then, by using 'calc.relimp' function in 'R' software, the relative percentage can be computed [9].

Relative importance refers to the measure of the contribution of the individual regressors in the multiple regression model. The assessment of the relative importance in the model is simple as long as the regressors are uncorrelated. It gives the total proportion of the variance explained by the model with all the variables.

It also gives the individual contribution of each predictor variable to the overall R^2

R^2 Measures the proportion of the variance in the dependent variable that is explained by the regressors in the model.

B. Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines can be used for variable selection. The earth package in 'R' software estimates the variable importance based on the generalized cross-validation (GCV), nsubsets (a number of subset models the variable occurs) and residual sum of squares (RSS) [9].

There are three statistics that can be used to measure the importance of the variable in the MARS model. They are GCV, nsubsets and RSS. MARS model includes a backward elimination features selection which estimates the variable Importance based on the reduction in the error of the generalized cross-validation. It tracks the changes in the model statistics for each predictor variable and adds the reduction in the statistics such as GCV when new features are added to the model. This total reduction gives the measure of the variable importance.

i) Generalized cross-validation (GCV)

Generalized cross-validation is a model validation technique for assessing the statistical results. It tells how well the predictive model performs in practice. The main goal of cross-validation is to limit problems like overfitting by testing the dataset in the training phase. It also gives an insight on how well the predictive model performs to a particular dataset (unknown dataset).

ii. *Residual sum of squares (RSS)*

In statistics, the Residual Sum of Squares (RSS), also known as the Sum of Squared Residuals (SSR) or the Sum of Squared Errors of prediction (SSE) is the sum of the square of the residuals. Residuals are deviations from the actual values of data. It helps to estimate the discrepancy between the observed and the actual data. Smaller the RSS, better the predictive performance of the model. It indicates how well the model generalizes to future results. It provides the measure to aid in model selection and parameter selection.

C. Random forest for variable selection

Random forest can be very effective in determining the best set of predictors by best explaining the variance in the response or independent variable [9].

Identifying the significant variable selection process in random forest [10]*:*

1. For each tree in the model, it calculates the number of votes
2. Then, it performs a random permutation of the predictor's value (let's say variable-k) in the dataset and check the number of votes for the correct class.
3. Subtract the number of votes for the correct class in the permuted data from the number of votes for the correct class in the original dataset.
4. The average of the value in all the trees is the variable importance score. This score is normalized by computing the standard deviation.
5. Variable having the large values are ranked more important than the other variables.

III. Feature selection Test results

Three different feature selection approaches were used in this work. All these combinations were tested with the same three months, six months dataset for the Iberian electricity market. Maximum of seventeen predictor variables from [11] were used as shown in Table I.

Table I. Seventeen decision variables

Variable No.	Description
1,2	Hourly Price D, Hourly Price D-6
3,4	Hourly Power Demand (P.D) D-1 & D-6
5,6	Hourly Hydropower Generation D-1 & D-6

7,8	Hourly Solar Power D-1 & D-6
9,10	Hourly Coal Power Gen (C.P) D-1 & D-6
11,12	Hourly Wind Power Generation D-1 & D-6
13, 14	Hourly Combined Cycle Power Generation (C.G) D-1 & D-6
15,16,17	Hourly Temperature, Wind speed, Radiation D+1

Table II shows the features selected from these approaches for three and six months dataset. These features were applied to the different hybrid models which are explained in the next section clearly. These features are ranked according to their importance.

Table II. Feature test results using different methods for different datasets

Dataset	Feature Selections	Features
Three Months	LR	Price D-6, Price D, P.D D-6, C.G D-6
	MARS	Price D, Price D-6, P.D D-6, Wind speed
	RF	Price D-6, P.D D-6, Price D, C.P D-6
Six Months	LR	Price D-6, Price D, P.D D-6, C.G D-6
	MARS	Price D-6, Price D, P.D D-6, C.G D-6
	RF	Price D-6, Price D, C.G D-6, C.P D-1

IV. Numerical Results & Discussion

Three different feature selection approaches were used in this work as mentioned earlier. Variables selected from these approaches were applied to the different hybrid models such as ARIMA, ARIMA-GLM, ARIMA- RF, and ARIMA-SVM, All these combinations were tested with the same three months (3M), six months (6M).

Table III shows the Comparison of MAPE for different hybrid models such as ARIMA, ARIMA-GLM, ARIMA- RF, and ARIMA-SVM using different feature selection approaches such as LR, MARS and RF for 3M dataset for (17,4,3,2) Variables.

TABLE III. Comparison of MAPE for hybrid models using feature selection approaches for 3M dataset

Parameter	ARIMA	ARIMA-GLM	ARIMA -SVM	ARIMA - RF
17 Variables	3.29	3.22	3.29	3.10
4 Variables (LR)	3.05	2.83	3.00	3.027
4 Variables ((MARS))	3.10	3.10	3.02	3.53
4 Variables (RF)	3.10	2.90	3.04	3.06
3 Common Variables	3.07	2.85	3.008	3.23
2 Variables (LR)	2.27	2.27	2.28	2.78
2 Variables (MARS)	2.27	2.27	2.28	2.78
2 Variables (RF)	2.27	2.27	2.28	2.78

Case 1: Applying Hybrid models to 3M (months) / 17V (variables)

In this case, three months dataset with all 17 variables were used to predict the day-ahead price of the electricity market. 17 variables used in this case were clearly mentioned in the table-1. For 17 variables, we can see that all hybrid models performed better. Clearly, ARIMA-RF outperforms other hybrid methods in this case as shown in Fig.1.

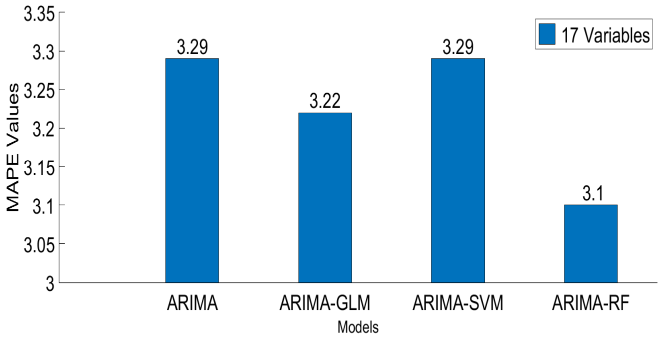

Figure 1: Comparison of MAPE for 90 days (May 01, 2015 to July 30, 2015) for 17 variables to predict day-ahead price (July 31, 2015)

Case 2a. Applying Hybrid models to 3M/4V using LR

In this case, three months dataset with four variables taken from Relative importance approach were used to predict the day-ahead price of the electricity market. Four variables taken from this approach are as follows: Price D-6, Price D, P.D D-6, and C.G D-6. For 4 variables, we can see that all hybrid models performed better. Clearly, ARIMA-GLM outperforms other hybrid method in this case. Also, there was a significant reduction in MAPE in almost all the hybrid models in this case than using all the 17 variables as shown in Fig.2.

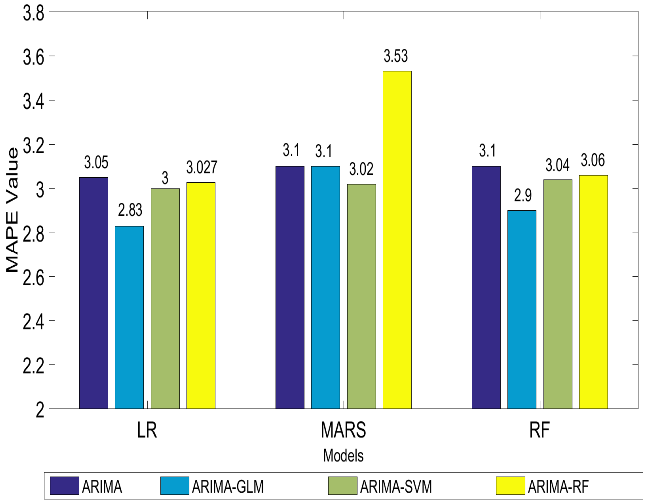

Figure 2: Comparison of MAPE for 90 days (May 01, 2015 to July 30, 2015) for 4 variables selected from variable selection approaches to predict day-ahead price (July 31, 2015)

Case 2b. Applying Hybrid models to 3M/4V using Multivariate Adaptive Regression Splines (MARS

In this case, three months dataset with four variables taken from MARS approach were used to predict the day-ahead price of the electricity market. Four variables taken from this approach are as follows: Price D, Price D-6, P.D D-6, Wind speed. From using 4 variables in this approach, we can see that there is an increase in MAPE in almost all the hybrid models.

Clearly, ARIMA-SVM outperforms other hybrid method in this case as shown in Fig.2. It can be inferred that these predictor variables selected from this variable selection approach do not help in reducing the MAPE.

Case 2c. Applying Hybrid models 3M/4V using RF

In this case, three months dataset with four variables taken from Random forest approach were used to predict the day-ahead price of the electricity market. Four variables taken from this approach are as follows: Price D-6, P.D D-6, Price D, and C.P D-6. From using 4 variables in this approach, we can see that there is a reduction in MAPE in almost all the hybrid models than the previous case using MARS model.

Clearly, ARIMA-GLM outperforms other hybrid method in this case as shown in Fig.2. It can be inferred that these predictor variables from this variable selection approach are better than MARS model but comes second in performance when compared to the relative importance approach.

Case 3. Applying Hybrid models to 3M /2V dataset using LR, MARS & RF approach for 2 variables

In this case, three months dataset with two variables taken from Relative importance approach, Multi-variate Adaptive Regression Splines (MARS), Random forest were used to predict the day-ahead price of the electricity market. Two variables taken from these approaches are as follows: Price D-6, Price D. One important inference for using two variables is that all these feature selection approaches gave the same features. For 2 variables, we can see that all hybrid models performed better than 17 & 4 variables taken from different approaches as shown in Fig.3.

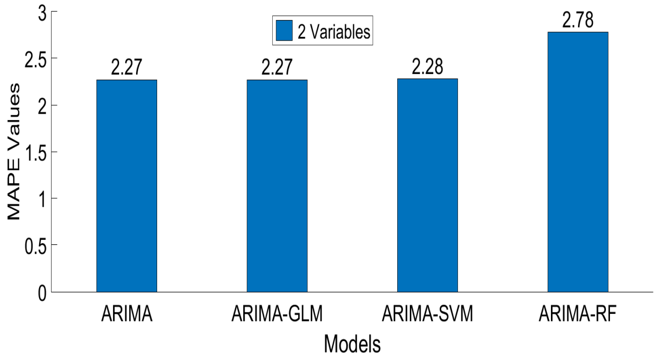

Figure 3: Comparison of MAPE for 90 days (May 01, 2015 to July 30 2015) for 2 variables selected from variable selection approaches to predict day-ahead price (July 31, 2015)

Case 4. Applying Hybrid models to six months dataset using all 17 variables

In this case, six months dataset with all 17 variables were used to predict the day-ahead price of the electricity market. 17 variables used in this case were clearly mentioned in the table-1. For 17 variables, we can see that all hybrid models performed better. Clearly, ARIMA-GLM outperforms other hybrid method in this case as shown in Fig.4.

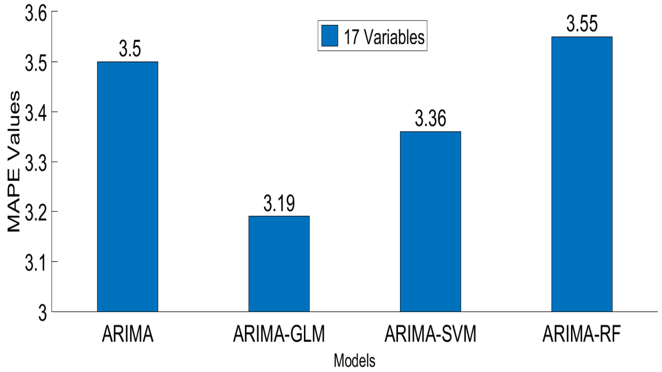

Figure 4: Comparison of MAPE for six months (Feb 01, 2015 to July 30 2015) for 17 variables to predict day-ahead price (July 31, 2015)

TABLE IV. COMPARISON OF MAPE FOR HYBRID MODELS USING FEATURE SELECTION APPROACHES FOR 6M DATASET

Parameter	ARIMA	ARIMA-GLM	ARIMA -SVM	ARIMA - RF
17 Variables	3.50	3.19	3.36	3.55
4 Variables (LR)	3.11	2.94	3.05	2.97
4 Variables ((MARS)	3.11	2.94	3.05	2.97
4 Variables (RF)	2.20	2.22	2.22	1.99
3 Common Variables	3.10	2.89	3.02	3.36
2 Variables (LR)	2.38	2.39	2.40	2.83
2 Variables (MARS)	2.38	2.39	2.40	2.83
2 Variables (RF)	2.38	2.39	2.40	2.83

Case 5a: Applying Hybrid models to six months dataset using Relative importance(LR) approach for 4 variable

In this case, six months dataset with four variables taken from Relative importance approach were used to predict the day-ahead price of the electricity market. Four variables taken from this approach as follows: Price D-6, Price D, P.D D-6, C.G D-6. For 4 variables, we can see that all hybrid models performed better than using 17 variables.

Clearly, ARIMA-GLM outperforms other hybrid method in this case. Also, there was a significant reduction in MAPE in almost all the hybrid models in this case than using all the 17 variables as shown in Fig.5.

Case 5b. Applying Hybrid models to six months dataset using Multivariate Adaptive Regression Splines (MARS) for 4 variables

In this case, six months dataset with four variables taken from MARS approach were used to predict the day-ahead price of the electricity market. Four variables taken from this approach as follows: Price D-6, Price D, P.D D-6, C.G D-6. These variables are same as the previous set of variables taken from the relative importance method. For 4 variables, we can see that all hybrid models performed better than using 17 variables as shown in Fig.5.

Case 5c. Applying Hybrid models to six months dataset using Random forest for 4 variables

In this case, six months dataset with four variables taken from Random forest approach were used to predict the day-ahead price of the electricity market. Four variables taken from this approach are as follows: Price D-6, Price D, C.G D-6, and C.P D-1. For 4 variables, we can see that all hybrid models performed better than the previous set of 4 variables using Relative importance and MARS approach. Clearly, ARIMA-RF outperforms other hybrid method in this case. Also, there was a significant reduction in MAPE in almost all the hybrid models in this case than using all the 17 variables and other 4 variables as shown in Fig.5.

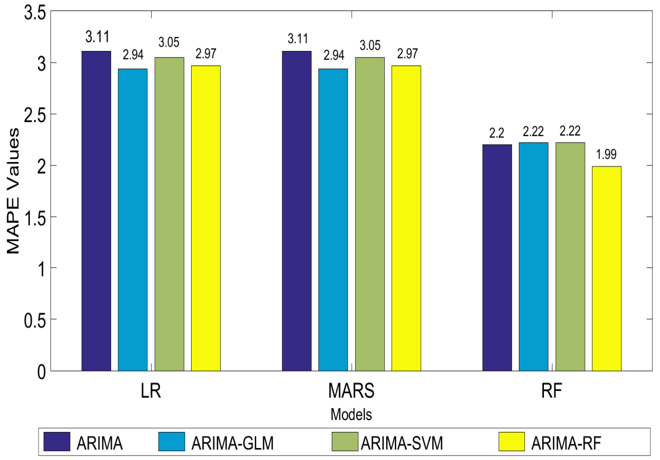

Figure 5: Comparison of MAPE for six months (Feb 01, 2015 to July 30 2015) for 4 variables selected from variable selection approaches to predict day-ahead price (July 31, 2015)

Case 6. Applying Hybrid models to six months dataset using LR, MARS & RF approach for 2 variables

In this case, six months dataset with two variables taken from LR, MARS & RF approach were used to predict the day-ahead price of the electricity market. Two variables taken from these approach are as follows: Price D-6, Price D. For 2 variables, we can see that all hybrid models performed better than 17 & 4 variables taken from the same approach. Also, there was a significant reduction in MAPE in almost all the hybrid models in this case than using all the 17 & 4 variables from the same approach as shown in Fig.6.

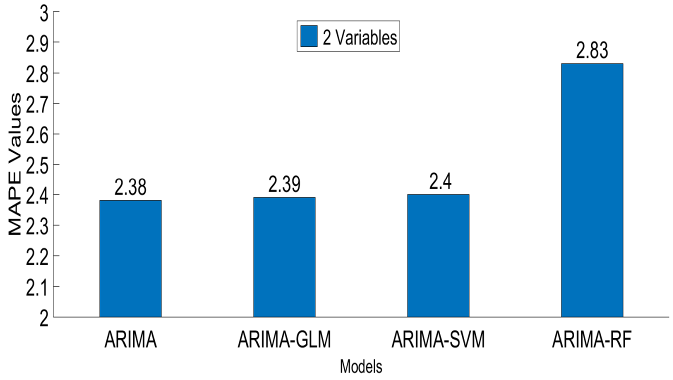

Figure 6: Comparison of MAPE for six months (Feb 01, 2015 to July 30 2015) for 2 variables selected from variable selection approaches to predict day-ahead price (July 31, 2015)

V. Conclusion

This paper investigated the performance of feature selection models for the day-ahead Iberian electricity markets. Features selected using LR, MARS and random forest were applied to different hybrid models to predict the day-ahead price. Two datasets (3/6 months) were used to validate the performance of the model. Three different set of variables (17, 4, 2) with ARIMA parameters were used in this problem. Finally, three common variables selected from these feature selection approaches were tested with all these datasets. MAPE was reduced considerably using 4, 3 and 2 variables selected from these feature selection approaches. Our future work includes investigating other datasets to test the scalability of the results

ReferenceS

[1] "Introdcution to Feature Selection." [Online]. Available: https://machinelearningmastery.com/an-introduction-to-feature-selection/.

[2] R. A. Chinnathambi, "Investigation of forecasting methods for the hourly spot price of the Day-Ahead Electric Power Markets," in *IEEE International Conference on Big Data (Big Data)*, 2016, pp. 3079–3086.

[3] T. Hossen, S. J. Plathottam, R. K. Angamuthu, P. Ranganathan, and H. Salehfar, "Short-Term Load Forecasting Using Deep Neural Networks (DNN)."

[4] R. Angamuthu Chinnathambi, A. Mukherjee, M. Campion, H. Salehfar, T. Hansen, J. Lin, and P. Ranganathan, "A Multi-Stage Price Forecasting Model for Day-Ahead Electricity Markets," *Forecasting*, vol. 1, no. 1, p. 3, 2018.

[5] R. Angamuthu Chinnathambi, S. J. Plathottam, T. Hossen, A. S. Nair, and P. Ranganathan, "Deep Neural Networks (DNN) for Day-Ahead Electricity Price Markets," in *IEEE Canada Electrical Power and Energy Conference (EPEC 2018) (Accepted)*, 2018.

[6] Z. Qiu, "Mutivariable mutual information based feature selection for electricity price forecasting," *2012 Int. Conf. Mach. Learn. Cybern.*, pp. 168–173, 2012.

[7] A. Mohamed and M. E. El-Hawary, "Effective input features selection for electricity price forecasting," *Can. Conf. Electr. Comput. Eng.*, vol. 2016–Octob, 2016.

[8] O. Abedinia, N. Amjady, and H. Zareipour, "A New Feature Selection Technique for Load and Price Forecast of Electrical Power Systems," *IEEE Trans. Power Syst.*, vol. 32, no. 1, pp. 62–74, 2017.

[9] "Feature selection approaches." [Online]. Available: http://r-statistics.co/Variable-Selection-and-Importance-With-R.html.

[10] "How Variable Importance in Random Forest works." [Online]. Available: https://www.listendata.com/2014/11/random-forest-with-r.html.

[11] C. Monteiro, L. A. Fernandez-Jimenez, and I. J. Ramirez-Rosado, "Explanatory information analysis for day-ahead price forecasting in the Iberian electricity market," *Energies*, vol. 8, no. 9, pp. 10464–10486, 2015.