Tomislav Gelo, PhD

Full Professor
University of Zagreb
Faculty of Economics and Business
E-mail: tgelo@net.efzg.hr
Orcid: https://orcid.org/0000-0002-4804-4315

Marko Družić, PhD

Associate Professor
University of Zagreb
Faculty of Economics and Business
E-mail: mdruzic@net.efzg.hr
Orcid: https://orcid.org/0000-0002-6436-663X

THE UTILITY OF MACHINE LEARNING IN THE ANALYSIS OF THE CLEAN ENERGY TRANSITION: THE CASE OF GERMANY

UDC / UDK: 620.91:504.06
JEL classification / JEL klasifikacija: Q42, Q48, Q56
DOI: 10.17818/EMIP/2025/11
Original scientific paper / Izvorni znanstveni rad
Received / Primljeno: October 28, 2024 / 28. listopada 2024.
Accepted / Prihvaćeno: December 23, 2024 / 23. prosinca 2024.

Abstract

One of the main components of the clean energy transition process in the EU are its liberalized electricity markets. Since most of the electricity is traded in day-ahead closed auctions, reliable and accurate electricity price prediction has become a question of paramount importance. This has led to the extensive use of machine learning algorithms, which have become increasingly powerful in the last decade, in predicting the movement of key economic variables in the energy sector. However, their use is currently for the most part limited to producing black-box predictions, with no attempt to produce explanations or economic insight. The purpose of this paper is to attempt to see whether a bridge can be built between the disconnected realms of economic analysis and machine learning. We use decision tree-based techniques to analyse the variability of hourly prices in the German electricity market from 2015-2020. We then compare the results with coefficient magnitudes from a linear regression framework. Our results indicate that the two approaches end up in substantial agreement on variable importance. We conclude that this is an area worth exploring further, since it can lead to expanding the energy sector analysis toolkit, which could lead to more informed energy policy.

Keywords: machine learning, regression, random forest, day ahead electricity price, variable importance

1. INTRODUCTION

Climate change is the result of greenhouse gas emissions, while greenhouse gas emissions are the result of fossil fuel consumption. Today, 82 percent in the World (and 70 percent in European Union) of primary energy consumption is covered by fossil fuels (oil, coal and natural gas). Increased energy consumption levels are linked with higher CO2 levels (Zhigolli,and Fetai, 2024). Questions and problems related to energy have become more and more important in the last few years, and it is necessary to research them from different aspects (Šandrk Nukić, 2020). Considering the growing demand for energy, especially in developing countries, new models of economic growth and development based on renewable energy sources are required. The transition from fossil to renewable energy sources takes place through energy transition. Fossil fuel capacities are decreasing, while renewable capacities are increasing. This is accompanied by a digitalisation of the energy sector, through the establishment of the smart grid (Gelo, 2020). Along with China and the USA, the European Union is the leader of the energy tradition in the world. Significant changes are taking place in the EU electricity market in the process of energy transition.

The electricity market is liberalized and electricity exchanges are established in all countries. As a result, a new energy market model is formed in which an energy consumer is at the same time an energy producer, a "prosumer" (Kotilainen et al., 2016) while centralised energy production is replaced by distributed energy production. Even within the EU, there are still differences between the old and new EU members. Price and income elasticities are considerably higher in old member states (Arčabić et al., 2021). Prices on the electricity exchange are the result of supply and demand. The demand curve is in principle inelastic and it mostly remains the same regardless of the price because there are no substitutes. Consumption changes significantly during weekends and holidays. Small changes on the supply side of electricity can lead to major price changes. The addition of electricity production from renewable sources in combination with conventional production (from fossil fuels or nuclear power plants) affects the supply curve and thereby determines a new price as a result of market dynamics. Electricity price forecasting is essential for reliable and cost-effective operations in the power industry (Bozlak and Yaşar, 2024).

Since the creation of these electricity markets in the EU in the 1990s, the demand for more accurate forecasts of electricity demand, supply, and especially prices has steadily increased. With the growth of electricity production from renewable sources, there is a significant drop in prices and an increase in the volatility of electricity prices, and there are often negative prices on the market at weekends and holidays when consumption is lower. This is happening more often and could have the effect of destabilizing the electricity market and slowing down the energy transition. As the majority of electricity in the EU is traded in day-ahead closed auctions, reliable and accurate electricity price forecasting has become a question of paramount importance. This is a direct result of more precise forecasts correlating with less economic inefficiency (better risk management, less energy waste, lesser costs etc).

Owing to this increasing demand, there is a substantial and growing literature which leverages various different methods for forecasting purposes. Lago et al. (2018) categorize the literature according to the method used to (i) game theory models, (ii) fundamental methods, (iii) reduced-form models, (iv) statistical models, and (v) machine learning (ML) methods, with statistical and ML models being the most successful. ML models started coming to the fore approximately a decade ago (Weron, 2014) and since then there has been an explosion of studies using them for electricity price forecasts (EPF) and to a lesser extent to other applications such as load forecasting (Dudek, 2022; Zhao et al., 2022). The most used ML techniques are variants of neural networks (NN), support vector machines (SVM), and decision tree based techniques, of which random forest (RF) and gradient boosting (GB) dominate (Beaulne, 2021). Research in this area has mainly addressed the question of whether machine learning models can provide better results than standard statistical approaches (Vlah Jerić, 2023). The shape of the electricity price curve has been altered as some factors that underpinned the electricity price forecast (EPF) lost their importance and new influential factors emerged (Bâra, & Oprea, 2024).

Early application of decision tree based methods focused on the Spanish and New York electricity markets (Juárez et al., 2015; Mei et al., 2014) showcasing the power of Random Forests (RF). Later research found gradient boosting methods (GB) to be superior (Arya & Vijaya Chandrakala, 2021; Beaulne, 2021; Lucas et al., 2020; Naumzik & Feuerriegel, 2020; Poggi et al., 2023; Pop et al., 2021), with one exception (Visser et al., 2020) where the authors found that RF outperformed GB and SVM. However, there are also recent papers that conclude that various NN architectures and deep learning methods give the best results (Hillmann, 2022; Jin & Azuka, 2022; Lago et al., 2018). We learn from the literature that ML techniques outperform statistical techniques on average, with the leaders being XGBoost and specific forms of NNs (artificial neural networks, convolutional neural networks DNN's etc.). However, Lago et al. (2021) warn that much of the research done is non-reproducible and very difficult to compare because different datasets and timeframes are used (some time-frames are too short), state-of-the-art methods of one group are compared with simple models of a different group leading to unfair comparisons etc. They attempt to alleviate these problems by producing open-access benchmark models and datasets, which are gradually being used by other researchers (Tschora et al., 2022). The jury still seems to be out on the question of the best methods for prediction.

However, the use of ML methods when discussing energy policy has so far been limited to pure prediction, rather than explanation. This follows a general trend where ML has thus far been slow to penetrate Economics journals, which still seem to prefer the standard regression-based econometrics toolbox. This is not surprising considering that most ML methods (especially state-of-the art ones like deep learning and NNs) are essentially black boxes, offering no explanation as to how the forecasts were attained. There is however an exception. Decision tree-based methods, in addition to having variants which do very well prediction-wise, have from the start had a natural way of ranking variables by their importance (see section 2).

The main goal of the paper is to see whether these decision tree-based techniques can be used profitably as an *explanation* generating, rather than merely a forecasting generating process in energy policy. In other words, we want to find out whether ML can be used to help build economic intuition and understanding, rather than just churn out forecasts. The practical aims of the paper are relatively small, we attempt to see whether the way in which ML methods assign importance to variables conforms with economic intuition and with the more standard regression approach.

To test this, we take data from the German day-ahead electricity market on electricity consumption, wind power generation, solar generation, various time variables (hour, day, month etc.), and run a simple regression framework (the kind that is ubiquitous in economic research) on electricity prices. Then we do the same but with an ML decision tree framework. More precisely, we run simple versions of several classic decision tree methods to see which one gives the best prediction results, then we compare its variable importance ordering with our regression results. Our results indicate that the orderings are almost identical. Furthermore, we introduce a rolling window methodology to see whether the ML methodology can be used in a dynamic framework, explaining changes through time.

The rest of the paper is organized as follows: section 2 describes our Methodology, including a detailed description of the regression and ML framework used, section 3 our choice of data and variables, including a descriptive analysis of the general trends in the data, section 4 gives our data source and software libraries used to ensure full reproducibility. Section 5 includes our results and discussion, and section 6 concludes with final remarks and policy implications.

2. METHODOLOGY

This section consists of two parts: regression and machine learning methodology. Since our main goal is to see whether the two methods give similar results which are compatible with economic intuition, for the sake of simplicity we opt for simple versions of both. In the case of regression this means that we perform a simple linear regression with minimal data transformations to minimize the effects of seasonality and heterogeneity.

In the case of our ML methodology, we opt for the default software hyperparameters without tuning. We justify this with the fact that since the main purpose is not prediction, reducing a few percentage points of error at the cost of significantly increased computing time is not necessarily something that is desirable. The goal in economic analysis is often to have easy to execute reliable methods that can be quickly adapted to novel research ideas on the fly. This however does not mean that tuning has no place in variable importance analysis, indeed if the hyperparameters are badly set and the error is high this can have a profound effect on variable importance ordering. This is due to the fact that variable importance metrics only give information on how important a variable is to the

model, implying that if the model is bad its variable importance might also be biased. However, based on our results (see Section 5) and their alignment with intuition we do not think this is a problem here.

2.1. Regression methodology

The day-ahead electricity market is a notoriously difficult setting for regression techniques. As Dudek (2022) points out the time series in question exhibits a non-linear trend, high variable variability, pronounced seasonality on multiple levels (daily, weekly, monthly) and constant significant random disruptions.

To attempt to alleviate seasonality, we first transform the data from hourly to daily observations by taking the mean of all the numerical variables in each respective day. While this substantially reduces the number of observations it still leaves plenty to run an effective regression on (just below 3000) while reducing seasonality to more manageable levels. As an additional step to reduce variable variability we normalize the numeric variables according to the standard formula:

$$x_n = \frac{x - \overline{x}}{s} \tag{1}$$

where x_n is the normalized version of an observation, x is the observation in the original units (in our case number of MW), $\underline{x}$ is the mean and *s* is the standard deviation.

We then opt for a simple linear regression framework of the form:

$$Price_t = \beta_0 + \beta_1 X_{t-1} + \beta_2 Y_t + u_t \tag{2}$$

where t is the day for which prices are predicted, **X** is a vector of predictive numerical variables from the day before (*solar, wind, total_production, consumption*) and **Y** is a vector containing dummy variables for the *month, hour* and *weekday* respectively. To account for heteroscedasticity and autocorrelation in the error term we use the standard approach in time series by using the Newey-West estimator (Newey & West, 1987).

2.2. Machine learning methodology

For our Machine learning methodology, we consider 4 models: single decision trees, bagged ensemble, random forest, and extreme gradient boosting.

2.2.1. Decision trees (CART)

The basic methodology of decision trees, also known as CART (classification and regression trees), are described by Breiman (1984). In essence, the algorithm forms a series of "if-then" queries, by which it selects a variable and a value to iteratively split the dataset into parent and child nodes until it reaches a

stopping criteria. The criterion with which the algorithm chooses variables and values to split on is based on an impurity reduction metric. If the dependent variable is continuous (as is the case with electricity prices) the measure of impurity is usually variance (Breiman et al., 1984). More specifically, the algorithm first calculates the variance of the dependent variable then it seeks the variable/value pair that results in a split of the dataset which results in an overall greatest reduction of variance, according to the formula:

$$IR(X) = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} - \left(\frac{n_{st}}{n}\frac{\sum_{i=1}^{n_{st}}(st_i - \mu_{st})^2}{n_{st}} + \frac{n_{sf}}{n}\frac{\sum_{i=1}^{n_{sf}}(sf_i - \mu_{sf})^2}{n_{sf}}\right) \quad (3)$$

where IR stands for impurity reduction, X is the set of n observations of the parent node variable, ST and SF are the child nodes, (or more accurately the post-split sets of observations where the split condition is true and false respectively), n, n_{st}, and n_{sf} are the number of observations in each respective set, and μ, μ_{st}, and μ_{sf} are the means of each respective set.

This process is iteratively repeated until a desired stop criteria is achieved (tree depth/number of splits, variance of the remaining nodes etc.). As a final step the output is given as the mean of the values in a respective node. There is a very useful by-product of decision-tree based techniques: their potential interpretability. Since the main split criteria is impurity reduction, the splits at the top of the tree are judged by the algorithm to be more important than later ones, essentially producing a ranking of variable importance (Louppe, 2015).

2.2.2. Bagging and Random forest

There is a weakness of basic decision trees in that slightly different datasets can result in different feature selection for initial splits, which leads to radically different branching and consequently potentially very volatile prediction results (overfitting). To deal with this issue, several different approaches were taken. One is to instead of building a single tree, a "forest" or "tree ensemble" is built instead, consisting of many decision trees in which each tree is trained on a bootstrapped sample of the data (Breiman, 1996), also known as *bagging*. Another is the *random forest* approach, where multiple trees are trained on a random sub-sample the data and of the features (Breiman, 2001). In both cases the final prediction is gained by taking the mean of each decision tree prediction by following the equation:

$$\hat{f}_{rf}^{K}(x) = \frac{1}{K}\sum_{k=1}^{K} t_i(x) \quad (4)$$

Where $\hat{f}_{rf}^{K}(x)$ is the output of the final model, $t_i(x)$ is the prediction result of a single *i*-th decision tree, and K is the total number of decision trees.

2.2.3. Gradient boosting

A third approach is to allow the trees to learn from each other, by iteratively assigning more importance to misclassified observations for each subsequent tree and thereby enabling the trees to learn from the misclassifications of their predecessors. These are referred to as GBDT (*gradient boosting decision trees*), and closely follow the pioneering work of Friedman (2000; 2001). The gradient boosting decision trees methodology operates by minimizing an objective function, which consists of an error metric (loss function) with an added term to penalize complexity (Ertuğrul et al., 2022). The procedure follows the general equation:

$$\Phi = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \quad (5)$$

where n is the number of data samples, $l(y_i, \hat{y}_i)$ is the *loss function* and $\gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$ the regularization parameter where γ is the penalty associated with introducing an additional node, λ is a hyperparameter which determines the severity of the penalization, ω_j^2 is the Euclidian norm of node j weights, where $j \in \{1, \dots, T\}$ and T is the total number of nodes.

Allowing each tree to learn from its predecessor, and approximating the function by a second order Taylor expansion, for the k-th iteration we get:

$$\Phi_k = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(k-1)} + g_i f_k(x_i) + \frac{1}{2} h_i f_k^2(x_i)\right) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \quad (6)$$

where $g_i = \delta_{\hat{y}^{(k-1)}} l(y_i, \hat{y}_i^{(k-1)})$ and $h_i = \delta^2{}_{\hat{y}^{(k-1)}} l(y_i, \hat{y}_i^{(k-1)})$ are the respective first and second order gradient values of the loss function.

Of these types of models the most powerful current iterations are XGBoost, CatBoost, and LightGBM which are currently filling top seats in machine learning competitions (Sagi & Rokach, 2021). In our analysis we use XGBoost.

2.2.4. Performance metrics and variable importance

The gains in prediction power that come from more complex models come with the cost of becoming far less interpretable than single decision trees. This is the result of the impracticality of going through the decision process of each of the hundreds or thousands of trees involved in the final decision of each model. One of the earliest and still widely used methods for aggregating the decision process to human readable output are permutation-type variable importance calculations proposed by Breiman (2001), which is the one used in our analysis. The process consists of permuting each variable while keeping the others constant and recording the change in the error rate that the permutation produces. The more negative the rate of change, the more important the variable is to the model.

The error rate /performance metrics used are given in equations (7)-(10):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (7)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (8)$$

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \quad (9)$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (10)$$

3. DATA

3.1. Variable and time period selection

Our variables of choice are *day_hour* (hour of the day from 1-24), *consumption_day_before* (number of consumed MW 24 hours prior), *planned_production_wind* (day ahead forecast of wind power generation), *planned_production_solar* (day ahead forecast of solar power generation), *planned_production_total* (day ahead forecast of total electricity generation), *month* (month of the year), *weekday* (day of the week 1-7), and *price_24* (electricity price 24 hours prior).

Our dataset consists of the period from 2015-2020 worth of hourly data for all of the aforementioned variables, which when removing missing values adds up to 48360 observations. While additional variables are sometimes used in the literature (oil and gas prices, grid load etc.) these are harder to obtain on an hourly basis and often proprietary, so we stick with publicly available data. While the data for 2021 and 2022 is available we decided to drop them from the dataset and focus on the more stable 2015-2020 period, on account of price variations in the market starting to become significantly more pronounced in late 2021 and 2022, corresponding to energy sector upheavals following the epidemic of COVID-19 and the war in Ukraine. We apply the standard ML methodology of splitting the data into a *train* and *test* set. No randomization is required since we are dealing with a time series, so we simply use the data from the years 2015-2019 as the training set, and 2020 as the test set (giving 43800 observations for the training set, and 8784 for the test set). After removing missing data we are left with 39578 observations in the training set and 8782 in the test set.

The *planned_production_total* variable presents a problem, since it effectively includes *planned_production_wind* and *planned_production_solar*. While multicollinearity is not an important problem for decision tree models it is for regression. To attempt to deal with this issue we ran both the ML and econometric approach subtracting wind and sun production from *planned_production_total*. While the regression coefficients do turn out to be slightly different the basic ordering is preserved, however curiously the ML algorithms perform worse with this artificially constructed variable than with the original planned production. Since the approach in the end made no substantial difference to the regression framework, but negatively impacted the ML framework we opted to keep *planned_production_total* unchanged.

3.2. Descriptive analysis

The German economy is the leading economy of the European Union with a share of about 25 percent in the total gross domestic product of the European Union in 2022 [database](Eurostat, GDP and main components). Germany also has the highest consumption of electricity with a share of 20 percent in the total consumption of the European Union [database](Eurostat, Supply, transformation and consumption of electricity). The German electricity market is the reference market of the European Union because of its size. It is three times larger than the French market. Due to its geographical location, it is very well connected to neighbouring electricity markets. Surplus electricity from production from renewable energy sources can be exported to these markets. Germany is also among the leading countries in the European Union in terms of capacity and production of electricity from renewable sources. Germany has 37 percent of the European Union's wind and solar power generation capacity and 34 percent of the EU's wind and solar power generation in 2020 [database] (Eurostat, Supply, transformation and consumption of electricity). The share of electricity production from renewable sources in the total electricity consumption in Germany was 51 percent in 2020. This is a growth of 33 percent compared to 2015, when the share was 36 percent. The largest growth in the same period was the production of electricity from wind (64%), and sun (28%). Production from other sources (fossil sources such as gas and coal and nuclear energy) decreased by 28% (Figure 1).

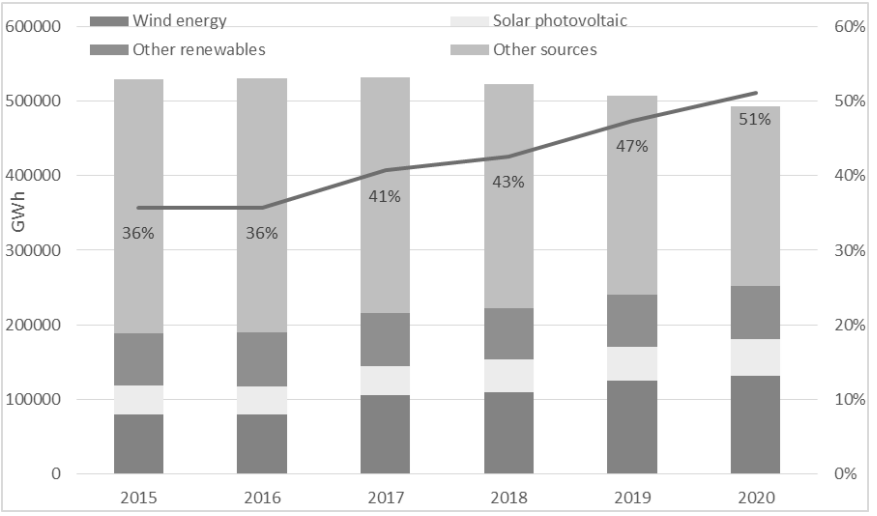

Figure 1 Structure of electricity consumption in Germany according to sources from 2015 to 2020

Source: authors analyses (based on: Eurostat, Supply, transformation and consumption of electricity, https://ec.europa.eu/eurostat/databrowser/view/NRG_CB_E/default/table?lang=en: 05.09.2023.; IRENA (2023), Renewable energy statistics 2023, International Renewable Energy Agency, Abu Dhabi, www.irena.org/Publications: 04.09.2023.).

In the analysed period from 2015 to 2020, there is a noticeable seasonal trend in electricity production from renewable sources. Production from solar energy is highest in the period from April to September, when the days are the longest and sunniest, while production from wind energy is highest in the period from October to March. Planning production from solar energy is much simpler. Production from these two renewable sources is supplemented seasonally.

The growth of electricity production from renewable sources has an impact on prices on the electricity exchanges. Figure 2 shows the impact of renewable energy production from wind and sun in Germany on day-ahead prices at the German electricity exchange (EPEX DE). The year 2019 was analysed because it is a year with stable electricity consumption (before the COVID 19 pandemic).

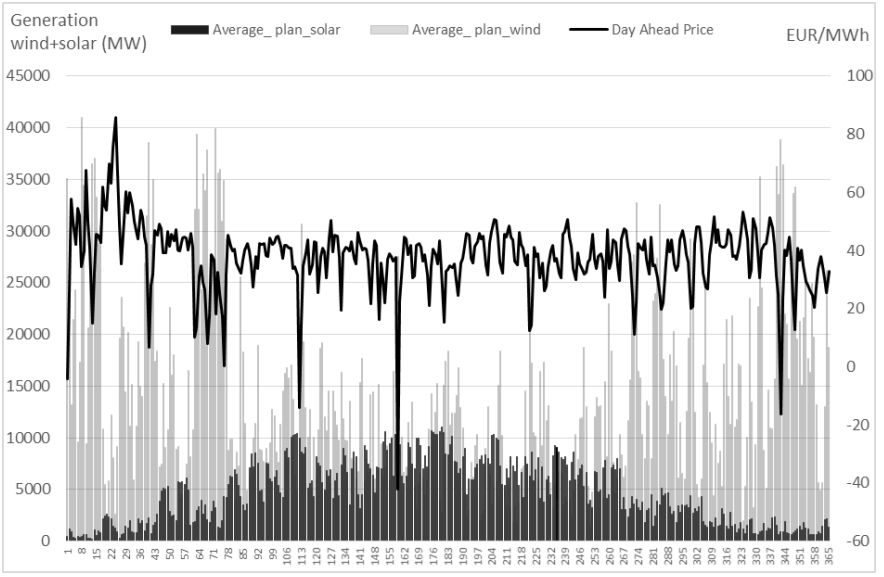

Figure 2 The impact of renewable energy production in Germany on day-ahead prices at EPEX DE in 2019

Source: Salopek, 2021

There is significant volatility between prices on electricity exchanges and production from renewable energy sources. When production from renewable sources is large (primarily from wind), prices on the electricity exchange are lower and even negative. The production of electricity from the sun, primarily in the summer months, in combination with wind also affects prices on the electricity exchange, although it seems that the impact of wind on prices is more significant. In 2019, there were several times lower consumption and higher production, which resulted in negative prices on the electricity exchange. In the case of negative

prices, the producer is forced to pay the consumer to take over the electricity due to a lack of his own production flexibility. The average share of electricity production from wind and solar energy in total consumption on working days is 32% and on non-working days about 40% (because consumption is lower). The average hourly price for working days on EPEX DE was €41.20/MWh, and the average hourly price for weekends (Saturday and Sunday) and non-working days was €30.01/MWh. The correlation is expectedly negative. The growth of the share of electricity production from renewable sources in the total consumption of electricity affects the drop in the price of electricity on the electricity exchange and vice versa (Salopek, 2021).

4. DATA AVAILABILITY AND SOFTWARE LIBRARIES

All of the data used in the analysis was collected from ENTSOE Transparency platform (https://transparency.entsoe.eu). The analysis is done in the R programming language, within the *tidymodels* and *tidyverse* ecosystems of packages. Specifically, for data wrangling and visualization we use *dplyr*, *lubridate*, *ggplot2 and vip*. Our *decision tree* analysis is done with *rpart*, bagged ensemble with *baguette*, the *random forest* with the *ranger* package, and *extreme gradient boosting* with the *xgboost* package (Table 1).

Table 1 Software libraries

Type of analysis	Software packages
data preparation	*dplyr, lubridate*
descriptive statistics	*dplyr, ggplot2, vip*
Decision trees	*rpart*
Bagged ensemble	*baguette*
Random Forest	*ranger*
XGB	*xgboost*

5. RESULTS AND DISCUSSION

Our results are grouped into a prediction and an interpretation section. In Section 5.1. we use the prediction results of the various Machine learning models presented in Section 2 to find the best model using the metrics described in Section 2.2.4. As a second step in Section 5.2. we compare the variable importance ordering our chosen model gives to our regression framework results.

5.1. Prediction results

Our prediction results are given in table 2. All four metrics agree on the ordering where basic decision trees perform the worst, followed by bagged trees then xgboost, with random forest reporting the smallest error. There is not much

surprise in the ordering except maybe in the fact that random forest outperformed xgboost, which is a somewhat rare occurrence in the literature (see Section 1). We leave it to further research to see if perhaps this ordering changes when tuning is introduced, if it does the same variable importance method can be used with xgboost.

Table 2 Prediction results

metric	dt	Bagged	rf	xgboost
rmse	12.62	9.39	8.41	9.02
mae	8.89	6.46	5.64	6.13
smape	37.03	31.22	27.66	29.92
rsq	0.48	0.74	0.77	0.75

There is one additional advantage of random forest over xgboost, and that is computing time. Since the random forest algorithm consists of simultaneously building multiple decision trees and averaging their results, it can benefit from parallelization. On the other hand, xgboost consists of building every subsequent tree from information delivered by the previous one, making the process much slower. Additionally, xgboost has 3-4 impactful hyperparameters to tune, while RF has less ((Dudek, 2022) argues there is actually only one) which also contributes to RF being the faster algorithm.[1] We conclude that while properly tuned xgboost might prove the superior model when sufficient computing time is allowed, RF is suitable for our purposes of determining the viability of ML for economic analysis, both in regard to precision and computing time.

5.2. Interpretation and variable importance

This section consists of two parts: the economic interpretation and variable importance of the regression methodology discussed in Section 2.1., and the interpretation of the results given by the RF algorithm.

5.2.1. Regression analysis

Our regression results are given in table 3. While the dummy/categorical variables are included in the regression (weekday, month, hour) to avoid omitted variable bias, we report only the numeric variable signs to make the table easier to read and since those are straightforward to compare with RF variable importance (the full table of coefficients is available upon request). The issue is that the regression framework produces a coefficient sign and statistical significance for

[1] As a test, the authors ran a tuning algorithm on xgboost with random values of three hyperparameters, the calculations took several hours and the result was worse than with default settings reported here (results not included but available upon request).

every category in a categorical variable (i.e. for every of the 12 months in the month variable etc.) while the RF methodology produces a single number for each categorical variable making the comparison difficult.

The results give an adjusted R squared of 0.73 which seems reasonable when considering the nature of the dataset discussed in Section 3. The coefficients give expected signs with planned production of renewables having negative signs, and lagged price and consumption having positive signs as expected.

Table 3 Regression results 2015-2019

	Dependent variable:
	price
planned_production_total	-0.104**
planned_production_sun	-0.024
planned_production_wind	-0.439***
price_day_earlier	0.537***
consumption_day_before	0.336***
weekday, month, hour	
Observations	2,016
R^2	0.735
Adjusted R^2	0.732
Residual Std. Error	0.518 (df = 1993)
F Statistic	251.268^{***} (df = 22; 1993)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

The interesting part is the coefficient magnitudes. Their absolute values are visualized in Figure 3. Our regression methodology gives the smallest relative importance to electricity produced by sunlight, which can partly be explained by the predictability and consistency of the energy source. In fact, the Newey-West standard errors make the variable statistically insignificant in explaining electricity price variations, while the rest of the variables are all significant according to the standard significance levels. The variables with the largest effect are the lagged price and planned wind production, followed by lagged consumption.

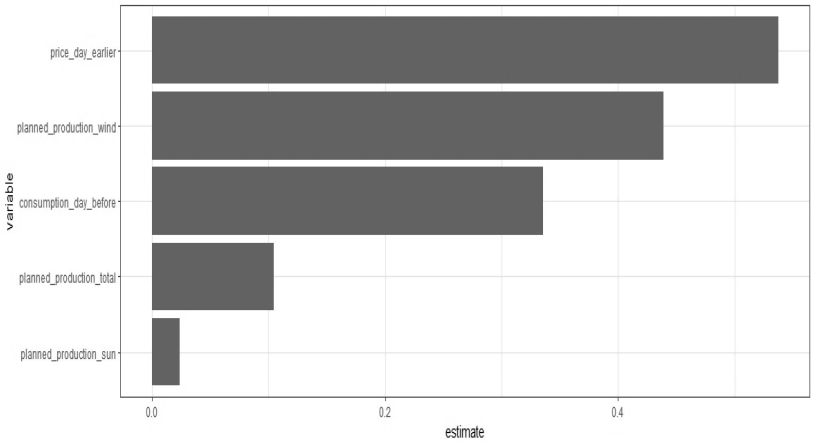

Figure 3 Absolute values of regression results 2015-2019

5.2.2. Random forest variable importance

The variable importance ranking of our RF algorithm is given in Figure 5. Here we present the categorical variables which are shown to be relatively less important than the numerical ones which are our primary focus. No direct comparison can be made since the values are presented on different scales, however what is striking is the ordering similarity with our regression results. The ordering of the numerical variables is identical except in the case of the first two (*planned_production_wind* and *price_day_earlier*). Furthermore, since the lagged price variable has no interpretative value from an economic perspective (there is not much to be gleaned from explaining electricity prices with electricity prices), the ordering of the numerical variables is for all intents and purposes the same. This is important since it gives credence to the assertion that the random forest model can be profitably employed to gain economic insights, not just for prediction purposes.

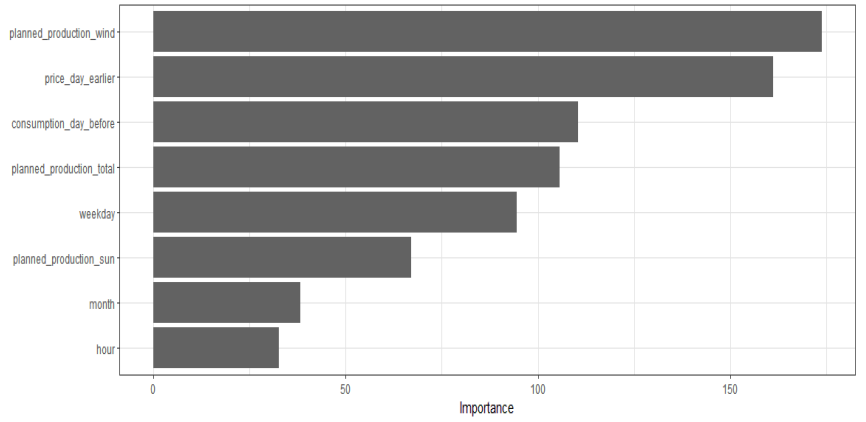

Figure 4 Random forest variable importance ranking 2015-2019

Additionally, we attempt to give a dynamic picture of the changing relationship between the variables and their relative importance over time. We run the same methodology with a one year rolling window. The results are given in Figure 5.

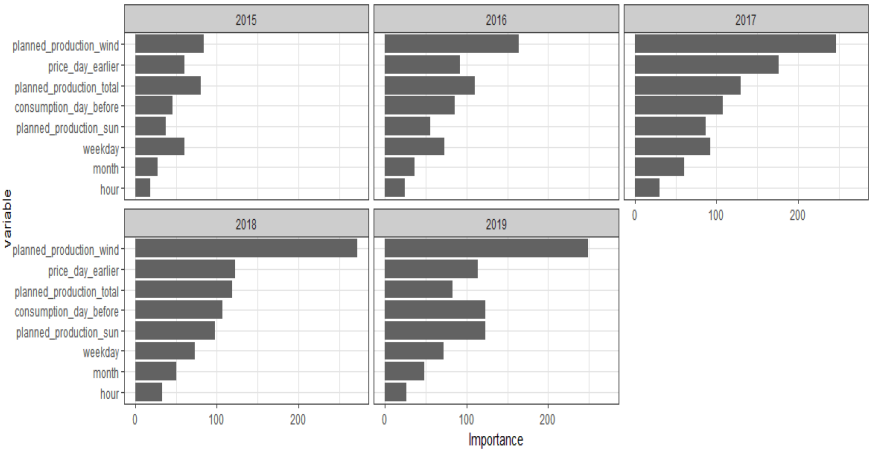

Figure 5 Variable importance of one year rolling window Random Forest from 2015-2019

The results align with economic intuition and are in accordance with Figure 1 (Structure of electricity consumption in Germany). With the growth of electricity production from wind (growth of 56 percent from 2015 to 2019), its relative importance as a variable in relation to other variables in the model from 2015 to 2019 increases. In the model, production from wind is always in first place in terms of importance in each year. As we can see the importance of wind as a variable in forecasts increases due to its volatility. The importance of the solar production variable is also increasing in the model. From 2015 to 2019, production increased by 15 percent, on the other hand in terms of importance, in 2015 it was in 6th place in 2019 it comes close to second place. These findings in our opinion further corroborate the viability of RF variable importance metrics for policy analysis.

6. CONCLUSIONS AND POLICY IMPLICATIONS

Since most of the electricity in the EU is traded in day-ahead closed auctions, reliable and accurate electricity price prediction has become a question of paramount importance. This is a direct result of more precise predictions correlating with less economic inefficiency (better risk management, less energy waste, lesser costs etc). This has led to the extensive use of machine learning algorithms, which have become increasingly powerful in the last decade, in predicting the movement of key economic variables in the energy sector (electricity price, grid-load etc). However, their use is currently for the most part limited to

producing black-box predictions, with no attempt to produce explanations or interesting new hypotheses about causal connections between variables of interest.

The purpose of this paper is to attempt to see whether machine learning tools can be used in this way, i.e. the same way as standard econometric tools in order to generate new economic insights. To do this, we use a sub-sample of machine learning methods, namely decision tree-based techniques, to analyse the variability of hourly prices in the German electricity market from 2015-2020. Unlike other popular machine learning algorithms, decision trees have the advantage of having built-in variable importance measures. Of the four methods tested, our results indicate that the best performing simple version is the Random Forest algorithm. We compare its variable importance metrics with coefficient magnitudes from a simple linear regression framework of the kind that is more traditionally used by economists. Our results indicate that the two approaches end up in substantial agreement on which variables carry the most explanatory power when explaining electricity price variation. Additionally, we run a one year rolling window Random Forest algorithm which yield variable importance metrics which also significantly align with economic intuition.

We conclude that this is an area worth exploring further, since decision trees are non-parametric, making them more suitable to capture non-linear connections in complex datasets such as those produced by the energy sector. This opens the door for the possibility that they will perform better than standard econometric tools in some cases, which can lead to more informed energy policy.

Author Contributions: Conceptualization, T.G. and M.D.; Methodology and Software, M.D.; Validation, T.G and M.D; Formal Analysis, M.D.; Investigation, T.G. and M.D.; Resources: T.G.; Data Curation, M.D., Writing – Original Draft Preparation, T.G. and M.D.; Writing – Review & Editing, T.G. and M.D.; Visualization, T.G. and M.D.

Funding: The research presented in the manuscript did not receive any funding from external funding sources.

Conflict of Interest: None.

REFERENCES

Arčabić, V., Gelo, T., Sonora, R. J., & Šimurina, J. (2021). Cointegration of electricity consumption and GDP in the presence of smooth structural changes. *Energy Economics*, 97, 105196. https://doi.org/10.1016/j.eneco.2021.105196

Arya, K., & Vijaya Chandrakala, K. R. M. (2021). Machine Learning Based Prediction and Forecasting of Electricity Price During COVID-19. *2021 IEEE International Power and Renewable Energy Conference (IPRECON)*, 1-6. https://doi.org/10.1109/IPRECON52453.2021.9640701

Bâra, A., & Oprea, S.-V. (2024). Predicting Day-Ahead Electricity Market Prices through the Integration of Macroeconomic Factors and Machine Learning Techniques. *International Journal of Computational Intelligence Systems*, 17, 10. https://doi.org/10.1007/s44196-023-00387-3

Beaulne, A. (2021). *European day-ahead electricity price forecasting*. https://papyrus.bib.umontreal.ca/xmlui/handle/1866/25095

Bozlak, B. Ç., & Yaşar, F. C. (2024). An optimized deep learning approach for forecasting day-ahead electricity prices. *Electric Power Systems Research*, 229. https://doi.org/10.1016/j.epsr.2024.110129.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24* (2), 123-140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random Forests. *Machine Learning*, *45* (1), 5-32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Taylor & Francis.

Dudek, G. (2022). A Comprehensive Study of Random Forest for Short-Term Load Forecasting. *Energies*, *15* (20), Article 20. https://doi.org/10.3390/en15207547

Eurostat. GDP and main components. https://ec.europa.eu/eurostat/databrowser/view/NAMA_10_GDP/default/table?lang=en

Eurostat. Supply, transformation and consumption of electricity. https://ec.europa.eu/eurostat/databrowser/view/NRG_CB_E/default/table?lang=en

Ertuğrul, H. M., Kartal, M. T., Depren, S. K., & Soytaş, U. (2022). Determinants of Electricity Prices in Turkey: An Application of Machine Learning and Time Series Models. *Energies*, *15* (20), Article 20. https://doi.org/10.3390/en15207512

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29* (5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, *28* (2), 337-407. https://doi.org/10.1214/aos/1016218223

Gelo, T. (2020) Energy Transition of the European Union. In Družić, G., & Gelo, T. (eds.), Conference Proceedings of the International Conference on the Economics of Decoupling (ICED) (pp. 211-234). Croatian Academy of Sciences and Arts and Faculty of Economics and Business University of Zagreb, Zagreb.

IRENA (2023). Renewable energy statistics 2023. International Renewable Energy Agency, Abu Dhabi. www.irena.org/Publications

Hillmann, S. M. (2022). *Time Series Electricity Price Forecast on the German Day-ahead Market* (Master thesis). Retrieved from: https://run.unl.pt/handle/10362/140860

Jin, K., & Azuka, C. (2022). *Modeling Electricity Prices in the German Energy market – With Applications to Renewables*. http://lup.lub.lu.se/student-papers/record/9084692

Juárez, I., Mira-McWilliams, J., & González, C. (2015). Important variable assessment and electricity price forecasting based on regression tree models: Classification and regression trees, Bagging and Random Forests. *IET Generation, Transmission & Distribution*, *9*. https://doi.org/10.1049/iet-gtd.2014.0655

Kotilainen, K., Sommarberg, M., Järventausta, P., & Aalto, P. (2016). Prosumer centric digital energy ecosystem framework. *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, 47-51. https://doi.org/10.1145/3012071.3012080

Lago, J., De Ridder, F., & De Schutter, B. (2018). Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. *Applied Energy*, 221, 386-405. https://doi.org/10.1016/j.apenergy.2018.02.069

Lago, J., Marcjasz, G., De Schutter, B., & Weron, R. (2021). Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293, 116983. https://doi.org/10.1016/j.apenergy.2021.116983

Louppe, G. (2015). *Understanding Random Forests: From Theory to Practice* (arXiv:1407.7502). arXiv. http://arxiv.org/abs/1407.7502

Lucas, A., Pegios, K., Kotsakis, E., & Clarke, D. (2020). Price Forecasting for the Balancing Energy Market Using Machine-Learning Regression. *Energies*, *13* (20), Article 20. https://doi.org/10.3390/en13205420

Mei, J., He, D., Harley, R., Habetler, T., & Qu, G. (2014). A random forest method for real-time price forecasting in New York electricity market. *2014 IEEE PES General Meeting, Conference & Exposition*, 1-5. https://doi.org/10.1109/PESGM.2014.6939932

Naumzik, C., & Feuerriegel, S. (2020). Forecasting electricity prices with machine learning: Predictor sensitivity. *International Journal of Energy Sector Management*, *15* (1), 157-172. https://doi.org/10.1108/IJESM-01-2020-0001

Newey, W. K., & West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, *55* (3), 703-708. https://doi.org/10.2307/1913610

Poggi, A., Di Persio, L., & Ehrhardt, M. (2023). Electricity Price Forecasting via Statistical and Deep Learning Approaches: The German Case. *AppliedMath*, *3* (2), Article 2. https://doi.org/10.3390/appliedmath3020018

Pop, C. B., Chifu, V. R., Cordea, C., Chifu, E. S., & Barsan, O. (2021). Forecasting the Short-Term Energy Consumption Using Random Forests and Gradient Boosting. *2021 20th RoEduNet Conference: Networking in Education and Research (RoEduNet)*, 1-6. https://doi.org/10.1109/RoEduNet54112.2021.9638276

Sagi, O., & Rokach, L. (2021). Approximating XGBoost with an interpretable decision tree. *Information Sciences*, 572, 522-542. https://doi.org/10.1016/j.ins.2021.05.055

Salopek, V. (2021). Utjecaj proizvodnje električne energije vjetra i sunca na trenutne cijene na burzama električne energije (završni specijalistički rad), Sveučilište u Zagrebu, Ekonomski fakultet, Zagreb. Retrieved from: https://urn.nsk.hr/urn:nbn:hr:148:084691

Šandrk Nukić, I. (2020). Sustainable Management of Energy Consumption in Public Buildings as a Determinant of Sustainable Economy. *Ekonomska misao i praksa*, *29* (1), 247-268. Retrieved from: https://hrcak.srce.hr/239599

Tschora, L., Pierre, E., Plantevit, M., & Robardet, C. (2022). Electricity price forecasting on the day-ahead market using machine learning. *Applied Energy*, 313, 118752. https://doi.org/10.1016/j.apenergy.2022.118752

Visser, L., AlSkaif, T., & van Sark, W. (2020). The Importance of Predictor Variables and Feature Selection in Day-ahead Electricity Price Forecasting. *2020 International Conference on Smart Energy Systems and Technologies (SEST)*, 1-6. https://doi.org/10.1109/SEST48500.2020.9203273

Vlah Jerić, S. (2023). Analysis of the Financial Performance of Machine Learning Models for Predicting the Direction of Changes in CEE and ESS Stock Market Indices with Different Classification Evaluation Metrics. *Ekonomska misao i praksa*, *32* (2), 533-545. https://doi.org/10.17818/EMIP/2023/2.12

Weron, R. (2014). 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. https://doi.org/10.1016/j.ijforecast.2014.08.008

Zhao, X., Li, Q., Xue, W., Zhao, Y., Zhao, H., & Guo, S. (2022). Research on Ultra-Short-Term Load Forecasting Based on Real-Time Electricity Price and Window-Based XGBoost Model. *Energies*, *15* (19), Article 19. https://doi.org/10.3390/en15197367

Zhigolli, G., & Fetai, B. (2024). The Relationship Between CO2 Emissions and GDP per Capita, Energy Consumption, Industrial Production in the Case of Western Balkan Countries. *Ekonomska misao i praksa*, *33* (2), 539-553. https://doi.org/10.17818/EMIP/2024/2.10

Dr. sc. Tomislav Gelo

Redoviti profesor
Sveučilište u Zagrebu
Ekonomski fakultet
E-mail: tgelo@net.efzg.hr
Orcid: https://orcid.org/0000-0002-4804-4315

Dr. sc. Marko Družić

Izvanredni profesor
Sveučilište u Zagrebu
Ekonomski fakultet
E-mail: mdruzic@net.efzg.hr
Orcid: https://orcid.org/0000-0002-6436-663X

KORISNOST STROJNOG UČENJA U ANALIZI PRIJELAZA NA ČISTU ENERGIJU: SLUČAJ NJEMAČKE

Sažetak

Jedna od glavnih sastavnica procesa tranzicije čiste energije u EU jesu liberalizirana tržišta električne energije. Budući da se većinom električnom energijom trguje na aukcijama zatvorenim za dan unaprijed, pouzdano i točno predviđanje cijene električne energije postalo je pitanje od iznimne važnosti. To je dovelo do opsežne upotrebe algoritama strojnog učenja koji su postali sve moćniji u posljednjem desetljeću u predviđanju kretanja ključnih ekonomskih varijabli u energetskom sektoru. Međutim, njihova je upotreba trenutačno najvećim dijelom ograničena na stvaranje predviđanja crne kutije, bez pokušaja davanja objašnjenja ili ekonomskog uvida. Svrha je ovog rada pokušati vidjeti može li se izgraditi most između odvojene domene ekonomske analize i strojnog učenja. U radu se koriste tehnike koje se temelje na stablu odlučivanja za analizu varijabilnosti cijena po satu na njemačkom tržištu električne energije od 2015. do 2020. godine. Dobiveni rezultati uspoređuju se s veličinama koeficijenata iz okvira linearne regresije. Oni pokazuju da se dva pristupa u značajnoj mjeri slažu u pogledu važnosti varijable. Zaključujemo da je ovo područje vrijedno daljnjeg istraživanja jer može dovesti do proširenja alata za analizu energetskog sektora, što bi moglo dovesti do bolje energetske politike.

Ključne riječi: strojno učenje, regresija, slučajna šuma, cijena električne energije dan unaprijed, važnost varijable.

JEL klasifikacija: Q42, Q48, Q56.