Studies in Economics and International Finance
ISSN: 2583-1526
Vol. 4, No. 2, 2024, pp. 121-138

ARF INDIA
Academic Open Access Publishing
www.arfjournals.com

AN ESSAY ON CARBON PRICES

Stourou Maria

Department of Economics, Democritus University of Thrace, Komotini, 69100, Greece.
E-mail: maristou1@econ.duth.gr

ARTICLE HISTORY
Received : 26 August 2024
Revised : 23 September 2024
Accepted : 10 October 2024
Published : 10 December 2024

TO CITE THIS ARTICLE
Stourou Maria (2024). An Essay on Carbon Prices. *Studies in Economics & International Finance,* Vol. 4, No. 2, pp. 121-138.

Abstract: The purpose of this study is to predict carbon prices traded in the European exchange-traded fund market (EU ETF). In particular, using daily data from January 2021 to March 2022, we separately examine the relationship between carbon EU ETF and coal prices, the economic sentiment Index (ESI), stock, oil, natural gas prices, carbon dioxide (CO2) emissions, economic policy uncertainty for the United Stated and a geopolitical risk index. In order to further investigate how the aforementioned variables affect carbon prices, we use the Support Vector Regression (SVR), Artificial Neural Network (ANN), Decision Trees and Random Forests from the field of Machine Learning. The results consistently highlight that there is a positive correlation between carbon and coal prices, the price of DAX40, the oil, the natural gas prices and the price of geopolitical risk index, whereas there is a negative correlation between economic sentiment and economy policy uncertainty index. CO2 emissions and economy policy uncertainty index are also statistically significant. Specifically, the correlation between carbon EU ETF price and CO2 emissions is significant. Moreover, the results clearly indicate that machine learning methods adapt better to the phenomenon than traditional ones. Furthermore, the method that best adheres to carbon prices evolution is the random forest.

Keywords: carbon EU ETF; machine learning;

1. INTRODUCTION

One of the key mechanisms in reducing and controlling the negative impact of environmental pollution and motivating investment in more environmentally friendly and efficient alternatives, is carbon pricing. A bulk of literature examines on what drives the carbon prices market in order to better comprehend the above

mechanism and the factors that can affect it directly or indirectly. Emissions trading is a market-based approach to mitigate pollution by providing financial incentives to reduce emissions. Founded in 2005, the European (EU) Emissions Trading System (ETS) is the world's first international emissions trading market. The EU ETS is by far the most established market for carbon trading. It is noteworthy that 22% of all global emissions are covered by carbon pricing. The EU Emissions Trading Scheme (ETS) and EU carbon emissions allowances (EUSs) stand out, accounting for 40% of the block's emissions, issuing around 1.4 billion allowances per year [1].

In recent years, climate change, which according to researchers is becoming more and more threatening, is forcing Europe to conclude to various summits, but also to sign various agreements aimed at reducing the effects of climate change. For example, in 2015, the Paris Agreement deals with the reduction of gas emissions, their adjustment and financial details. Another example is the holding of a summit in April 2022, during which key EU players pledged to contribute to green growth through sustainable projects and clean energy. These efforts are on the rise as the international community works to reduce greenhouse gas emissions and promote green transformation.

Carbon pricing is a very flexible and cost-effective approach to mitigate the effects of climate change. For many years, the price of carbon was traded well below 20 euros per ton on the EU ETS. But since the Covid-19 pandemic in 2020, the price of carbon has risen abruptly to 90 euros per ton. Another contributory factor is that many funds have been created in order for people who are environmentally aware. Global assets managed by mutual funds, and stock exchanges related to the climate change almost tripled to $177 billion in 2019. Thus, many researchers have been led to study the variables that affect the price of carbon so as to both predict its future and to better understand the mechanism of its pricing.

Analyzing the theoretical basis of the carbon price formation and the carbon price transmission mechanism from the perspective of the agents that affect carbon price, carbon price is driven by marginal abatement cost (MAC), price elasticity of demand and other factors that affect the supply and demand of the quotas. MAC and price elasticity of demand are the key factors when enterprises consider the carbon price, and several factors affect supply and demand of the quotas resulting in the carbon price volatility [2].

European Allowances (EUA) prices exhibit a statistically significant and positive correlation to the stock index returns. For instance, a 1% decrease of the Economic Sentiment Index (ESI)[1] is associated with a decrease in the EUA price of approximately 1.2% [3]. When it comes to DAX40, and coal price, they have negative effects on carbon prices [4, 5]. The relationship between oil price and

carbon price is slightly complex. In the short term, oil price has a negative effect on carbon price; however, in the long term, it has a positive effect.[5]. The prices of natural gas have positive non-statistically significant correlations with carbon price [6]. Brent, natural gas and coal prices are selected as being the main carbon price drivers [7].

Caldara and Iacoviello develop a geopolitical risk (GPR) index, which is a measure of adverse geopolitical events and associated risks. This index spikes around the Gulf War, after 9/11, during the 2003 Iraq invasion, during the 2014 Russia-Ukraine crisis, and after the Paris terrorist attacks [8]. The results of aforementioned research show that High geopolitical risk leads to a decline in real activity, lower stock returns, and movements in capital flows away from emerging economies and towards advanced economies. GPR and oil prices are significantly negatively correlated [9], specifically, during a bearish phase of the market [10]. Hence, the relationship between oil and carbon prices, the geopolitical risk index affects positively the caron prices. Examining the spillover effect of EPU on the carbon futures market under different market conditions, it is argued that under bearish market conditions, the economic policy uncertainty $(EPU)^2$ is negatively correlated with carbon futures price returns during the COVID-19 crisis. Under bullish markets, variations in EPU positively impacts on future returns in carbon prices[11].

Economic development is one of the major factors affecting the long-term trend of carbon price. For each 1% rise in the stock index, DAX40, the carbon price will rise 2.15% [12]. The carbon market is mainly affected by the coal, electricity and stock markets. Oil price, DAX index, coal price, gas price and carbon emissions are the main influencing factors. There is indeed a linear/polynomial relationship between the five afore-mentioned variables and the carbon price[13].

This study investigates the relationship between coal prices, the ESI, the stock market, the oil price, the natural gas price, the CO2 emissions, the US EPU, the geopolitical risk index and the carbon EU ETF price using machine learning methods and compare the results of these with the traditional Ordinary Least Square (OLS) regression, using the Root Mean Square Error (RMSE) and the Mean Absolutely Percentage Error (MAPE) loss metrics.

2. METHODOLOGY AND DATA

2.1. The Data

This study investigates the potential existence of a relationship between indices and variables in shaping carbon quota prices. We use daily data from January 4, 2021 to April 28, 2022. We collect the data for carbon EU ETF prices from the global energy think tank, Ember. Coal prices data are from the Nasdaq database,

oil and natural gas prices from the Federal Reserve Bank of St. Luis database while data on DAX40 are from Yahoo Finance. All data on EPU stem from the Federal Reserve Economic Data (FRED) of St. Luis. The data for GPR are derived from the Federal Reserve database.

In figure 1, we present the distribution of carbon prices.

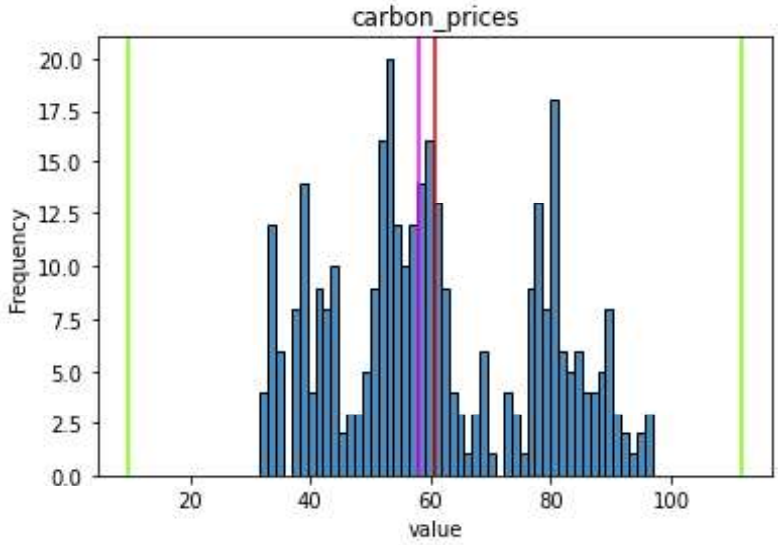

Figure 1: Carbon EU ETF prices. The red line corresponds with the mean, the purple line is the median and the green line denote 3 standard deviations from the mean.

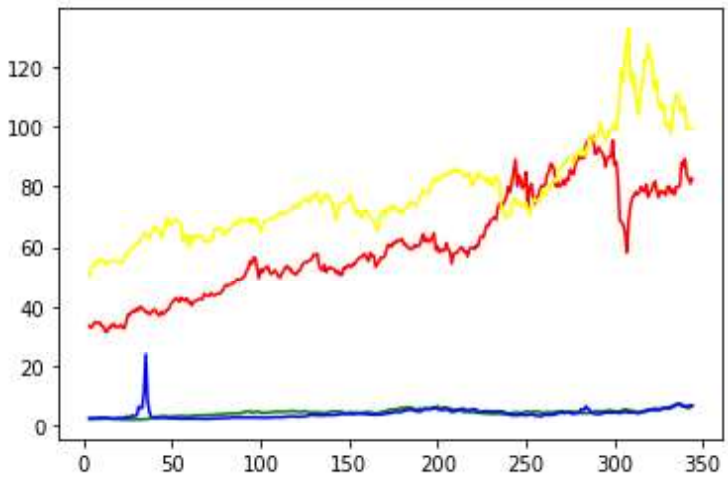

Figure 2: Carbon EU ETF prices, which are correspond with red line. The yellow line is the price of oil price. The blue line depicts the price of coal and the natural gas prices is depicting with green line. The x-axis is the days and y-axis is the price of variables

As we observe EU ETF prices deviate from the normal distribution, specifically, follow an asymmetry and platykurtic distribution with a negative kurtosis, since the median is different from the sample average value. There are no values that exceed 3 standard deviations (indicating no outliers).

In figure 2, we observe that there is a higher variance in the carbon EU ETF prices and oil price than the prices of the coal and natural gas. The graph shows that there is a negative correlation between carbon EU ETF prices and oil price, in contrast with the rest of variables, which are not influenced by the movement of the two other variables. Furthermore, the prices of natural gas and coal move together.

Table 1: Descriptive statistics

	Carbon ETF	*Coal*	*ESI*	*DAX40*	*Oil*	*Natural Gas*	*US EPU*	*CO2 emissions*	*GPR*
obs	341.00	341.00	341.00	341.00	341.00	341.00	341.00	341.00	341.00
mean	60.78	4.54	- 0.01	15106.08	78.24	4.18	136.63	60.56	128.63
std	16.99	1.07	0.09	754.49	16.11	1.64	58.28	16.93	87.86
min	31.62	2.15	- 0.22	12831.50	50.37	2.43	39.00	31.53	23.62
25%	49.45	4.17	- 0.06	14461.41	68.00	2.96	93.89	49.32	73.00
50%	58.16	4.63	0.00	15370.25	74.25	4.02	124.74	57.96	100.81
75%	77.95	5.05	0.06	15673.63	84.42	4.93	167.78	77.75	155.99
max	96.93	7.36	0.18	16271.75	133.18	23.86	399.87	96.41	539.58

In table 1, the first row depicts the total number of observations, which is the same for all variables. My interest is focused on comparing the value of the sample medium with the value of the second quadrant, because in this way the symmetry in the data is ensured.

The standard deviation (std) indicates the variability in a dataset. The standard deviation of carbon EU ETF is lower than that of DAX40, oil price, US EPU and GPR and higher than that of coal price, ESI, natural gas price and CO2 emissions. This shows us that the data points of carbon EU ETF price are clustered closer to the mean and the values in the dataset are relatively consistent. The high value of std shows that data values become more dissimilar and extreme values become more likely to appear. The interquartile range of carbon EU ETF is equal to 28,5 and this value shows that the middle value cluster more tightly.

It is important the data preparation in order to boost the model's prediction capabilities. Hence, it is essential to clean for outliers and standardize or normalize the data. Due to the fact that the histogram of Carbon EU ETF tends to follow the normal distribution, we use the log transformation.

2.2. Methodology

Cross validation (CV) method is used to calculate the accuracy of the model, deriving from random selection of training and validation dataset. According to this method, a part of dataset, k, and each single subset of it is used as the validation dataset and the rest of subsets, k-1, are combined to create a training dataset. This method is very useful in selecting hyperparameters for all models with CV (layers, number of neurons, C, e etc.).

The methodologies applied in the specific study are Support Vector Regression (SVR), Artificial Neural Networks (ΑΙΙ), decision trees, Random Forest (RF) and Ordinary Least Square (OLS). When it comes to the machine learning methods, the models are built in two steps, train and test. The various machine learning methods for regression, in contrast with linear regression is capable of "learning" not only linear relationship between target Y and features X, but also more complex non-linear relationships. The artificial neural network (ANN) attempts to solve a statistical problem by using several simultaneously working functions called "neurons", organized in layers. Every neuron accepts as input the value of the previous . The neurons of each layer are connected to neurons of other layers (usually only with previous and following). The first layer is called the input layer, while the final is called the output layer. Intermediates layers are called hidden layers. Specifically, ANN models follow the following mechanism. Each time a neuron is activated, it calculates a value using a transfer function g, comparing it with a threshold value. More specifically, activation function contributes to helping the neurons to "learn" the complex relationship that exists between the features and the target. The transfer function, which is a sigmoid or tanh activation, is:

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} \quad or \quad \tan h_\theta(\chi) = \frac{e^{\theta^T} x - e^{-\theta^T} x}{e^{\theta^T} x + e^{-\theta^T} x} \tag{1}$$

And the value defined as threshold is given by the function:

$$h_\theta\left(x^{(i)}\right) = g\left(\Sigma_{j=0}^{n} \theta_j^{(i)} x_i^{(i)}\right) = \begin{cases} 1 \; if > threshold \\ 0 \; if \leq threshold \end{cases} \tag{2}$$

or

$$h_\theta\left(x^{(i)}\right) = g\left(\Sigma_{j=0}^{n} \theta_j^{(i)} x_i^{(i)}\right) = \begin{cases} 1 \; if > threshold \\ -1 \; if \leq threshold \end{cases} \tag{3}$$

If the value obtained at the output is greater than the value set as the threshold, then that value is passed as input to the next neuron. Afterward, the values of the

weights (θ) are calculated, after first defining their initial values. Finally, we take a value as an output from the following:

$$output = f\left(\Sigma_{t=0}^{m} x_i w_i + b\right) \tag{4}$$

In a final stage, the generalization of the model is evaluated by using one part of the data as test set. The structure of an ANN model is depicted in the following figure 3.

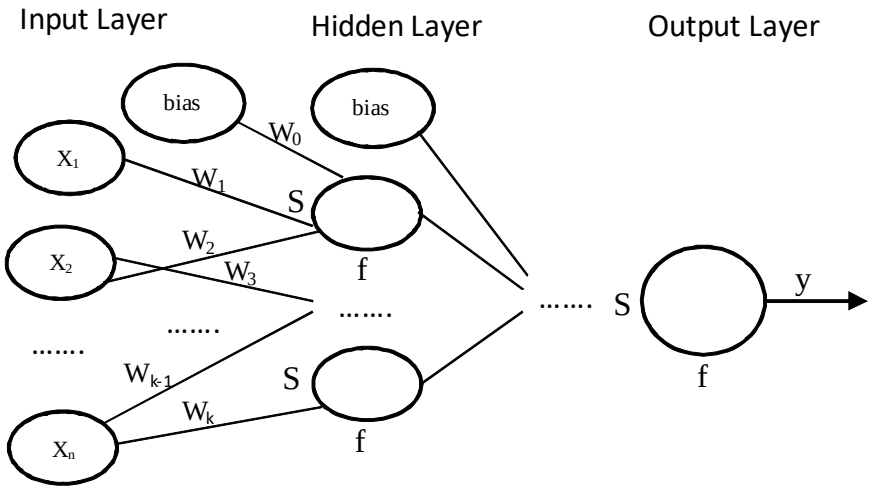

Figure 3: Structure of an ANN model

Support Vector Regressions (SVR) were proposed by Cortes and Vapnik (1995) [14]. The scope of this machine learning method is to find a hyperplane in an n-dimensional space. Specifically, the SVR model is an extension of the Support Vector Machine (SVM) algorithm. Its main difference from OLS estimators of linear regression is that it minimizes the Euclidean norm of the coefficient vector, w_i

$$\min \frac{1}{2} \| w \|^2 \tag{5}$$

A tolerance margin (ε) is defined and considering the following constraint $|y_i - w_i x_i| \leq \varepsilon_i$, while in OLS estimators the main goal is to minimize the square error. For example, in the case of simple linear regression, the following applies:

$$L_\theta(\varepsilon) = \min_{w_i} \Sigma_{i=1}^{n} (y_i - w_i x_i)^2 \tag{6}$$

Where, y_i is the target value and xi is the characteristic under study. Other extensions of the above equation are Lasso, Ridge and ElasticNet. Going into a

deeper analysis of the SVR model we can say that it consists of two main pillars, as it happens in most machine learning methods. The first pillar concerns the train set, during which the model is trained. The training starts by running the model and having minimized the quantity $\min \frac{1}{2} \| w \|^2$. A significant part of the data remains outside the support vector limits (these are the points that have occurred through the process of minimizing and which determine the two limits for our tolerance to error). To overcome this problem, we introduce in quantity $\min \frac{1}{2} \| w \|^2$. slack variables with which we will be able to expand the limits of our tolerance to error taking into account the deviations between the support vector and the other points, which are located near the support vector. The slack variables are variables that express the distance of a point outside the boundaries with the support vectors and is denoted as ζ_i. Therefore, the quantity $\min \frac{1}{2} \| w \|^2$. is transformed as follows:

$$\min \frac{1}{2} \| w \|^2 + C \Sigma_{i=1}^{n} | \zeta_i | \tag{7}$$

with the following constraint: $| y_i - w_i x_i | \leq \varepsilon + \Sigma_{i=1}^{n} | \zeta_i |$..

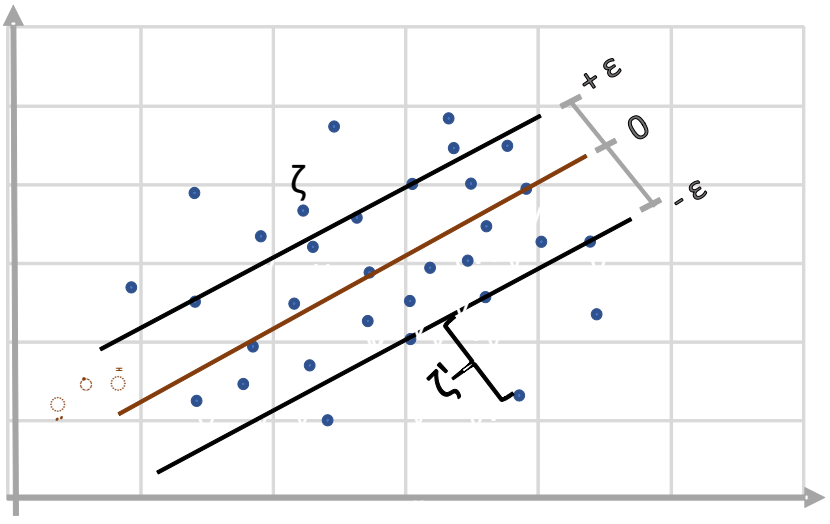

Figure 4: Illustration of an SVR regression function

The constant C is a positive control parameter. As the constant C increases, the percentage of data entering the model increases and consequently the number of deviations ζ_i. During the first training, the constant C receives a negligible value such as the value 10^{-4}. Moreover, the model is then re-trained, giving C a new value, higher than the previous one. And as consequence the value of C increases the cross validation accuracy. This process stops finding the appropriate constant C, which controls the degree of punishment of samples beyond the error ε.

In the second pillar, the test set, the model is evaluated by using the remaining data, which are not used during training to test the generalization ability of the model.

Kernel methods are pattern recognition methods which allow us to construct algorithms in dot product spaces. The advantage of these methods is that it characterizes the function class used for estimation via the representer theorem. These methods are much widespread in the field of machine learning, specifically, they are mostly applied by the SVM (or SVR) algorithm. The SVM (or SVR) kernel functions are a set of mathematical functions, which define inner products or similarity in a transformed space. We evaluate the linear and the RBF kernel.

Linear kernel:

$$K_1(x_1, x_2) = x_1^T x_2 \tag{8}$$

RBF kernel:

$$K_1(x_1, x_2) = e^{-\gamma \|x_1 - x_2\|^2} \tag{9}$$

γ is kernel parameter.

Morgan and Sonquist used in 1963 a machine learning method, namely decision trees for regression [16]. They use it as a complement and alternative to regression in order to analyze survey data. The primary goal of decision trees is to map all possible decision paths in the form of the tree. They consist of branch and nodes, namely root node, which is the first tree's node, internal node and leaf nodes that is a terminal node and assigns a classification. Each branch suggests a dichotomous decision and corresponds to the result of the test. The structure of decision tree is depicted in figure 5.

A random forest (RF) consists of a collection of decision trees that are different one from another due to the fact that they are trained through bagging and random variable selection. The forests which become separated with an oblique hyperplane can achieve accuracy while they grow without suffering from overfitting [15]. The general technique of bootstrap aggregating applied with respect to training the model. In the first step, RF algorithm constructs each tree for different sample set from the dataset. In the second step, each tree for different sample set is trained

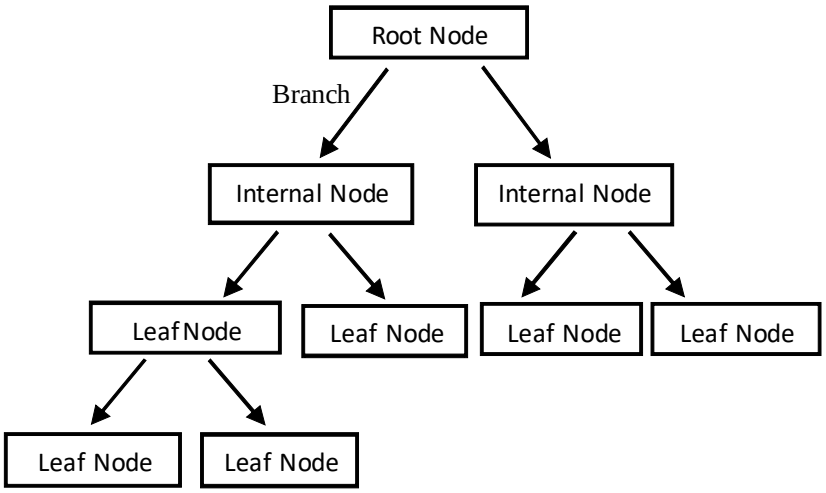

Figure 5: Structure of decision tree. This constitutes from root node, which is the first tree's node, internal node and leaf nodes that is a terminal node and assigns a classification and branch suggests a dichotomous decision and corresponds to result of the test.

by the algorithm; RF has as input the independent and dependent variable of the train set. In the third step, the dependent variable is predicted and the forecasting model is evaluated. The final output of the RF regression is the average of the outputs of all decision trees. Thus, the value of the dependent variable is equal to

$$\frac{1}{k}\Sigma_{i=1}^{k} Y_k \tag{10}$$

with k is a sample set.

3 EMPIRICAL RESULTS

We notice that there is a positive correlation between coal price, the price of DAX40, oil price, natural gas price and the price of geopolitical risk index. On the contrary, the price of the economic sentiment index and the economy policy uncertainty index are negatively correlated with the carbon EU ETF price. Finally, there is an almost perfect correlation between CO2 emissions and carbon EU ETF prices.

Table 2. Shows the correlation between all variables and price of carbon ETFs and the result by calculating a Pearson correlation coefficient and p-value. The CO2 emissions and the US EPU are statistically significant. There is a positive correlation between coal price, price of DAX40, oil price, natural gas price and price of geopolitical risk index with carbon EU ETF. There is a negative correlation between ESI and US EPU with carbon EU ETF. Finally, CO2 emissions and carbon EU ETF prices are almost perfect correlated.

	Pearson Correlation Coefficient	*P-value*
Carbon EU ETF - Coal	0.6195	> 0.05
Carbon EU ETF - ESI	- 0.2797	> 0.05
Carbon EU ETF – DAX40	0.2742	> 0.05
Carbon EU ETF - Oil	0.7486	> 0.05
Carbon EU ETF – Natural Gas	0.3536	> 0.05
Carbon EU ETF - US EPU	-0.1571	< 0.05
Carbon EU ETF – CO2 emissions	0.9988	< 0.05
Carbon EU ETF - GPR	0.4812	> 0.05

The results are shown in table 2. The p-value for all variables is bigger than the 0.05 significance level, so the results do not lead to statistical significance, apart from the CO2 emissions and the US EPU. The p-value for CO2 emissions and the US EPU are equal to 0.036 and 0.0, respectively.

In order to predict the price of carbon EU ETFs, we evaluate the forecasting horizon of one, five, ten and thirty days ahead. The analysis of my forecast model using the lagged prices of the independence variable begins by dividing the set of data into two parts. The first 80% as a train set and the last 20% as a test set.

The study begins with developing an autoregressive (AR) model, which is defined as follows:

$$Y = Y_{t-1} + Y_{t-2} + \ldots + Y_{t-n} + \varepsilon_i \tag{11}$$

The probabilistic statistical measures can be used for model selection. In this study, they are used with respect to finding the best lag which minimizes the Bayesian Information Criterion (BIC). The different lag orders are presented in table 3.

Table 3: I apply different lag orders and I find this lag which minimizes the Bayesian Information Criterion. The asterisk denotes the lowest value of BIC

LAG ORDERS	*BIC*
1	1.160
2	1.119 *
3	1.136
4	1.159
5	1.175
6	1.172
7	1.180
8	1.202
9	1.212
10	1.228

We found that when the lag order equals 2, the BIC is 1.119, that is minimized. Thus, the AR model is the following:

$$Y = Y_{t-1} + Y_{t-2} + \varepsilon_i \tag{12}$$

A common method in evaluating the generalization ability of a model is testing it in an out-of-sample forecasting exercise. In order to evaluate the SVR model having applied two different kernels, we use the Root Mean Square Error (RMSE) and the Mean Absolutely Percentage Error (MAPE) criterion, defined as follows:

$$MAPE = \frac{100}{n} \Sigma_{i=1}^{n} \frac{\hat{y}_i - y_i}{y_i} \tag{13}$$

$$RMSE = \sqrt{\frac{1}{n} \Sigma_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2} \tag{14}$$

The following Table 4 depicts the result of the evaluation of the autoregressive model of carbon EU ETF price with two lags.

Table 4: Shows the results of evaluation the AR model with two lags, when I use the 80% of the data in the train set and 20% of the data in the test set. The asterisk denotes the lowest error

	RMSE		*MAPE*	
Autoregressive model	*In-sample*	*Out-of-sample*	*In-sample*	*Out-of-sample*
OLS	0.177	0.229	0.973	1.140
SVR-linear	0.179	0.234	0.831	0.937
SVR-RBF	0.104*	0.181*	0.307	0.270
Neural Network	2.197	1.657	0.021*	0.021*
Decision Tree	3.247	3.853	0.039	0.045
Random Forest	1.461	2.363	0.124	0.166

With respect to predicting EU ETFs prices, we apply under the same specifications the same methods for different lags; specifically, one, five, ten, fifteen and thirty lags and evaluate them. Additionally, we develop a simple autoregressive SVR with two different kernels, linear and RBF. We apply a 5-fold cross validation to both. We use cross validation to train the model so as to find the best structure for the neural network model that will achieve the lowest forecasting error. Moreover, we develop a decision tree and a random forest which are more successfully adapted to the forecasting model. The best decision tree has a maximum depth of 5 and the minimum number of samples required to be at

a leaf node is 0,1. When it comes to random forest, there is unlimited number of leaf nodes and the number of trees in the forest is 10.

For the selection of the optimum cost (C), we follow a 5-fold cross validation procedure for the daily dataset. Specifically, the dataset is separated randomly into five subsets and each model is trained and validated 5 times. Table 5 depicts the forecasting accuracy of the SVR models for each of the two kernels and comparing them with traditional statistic method OLS. Table 5 shows the value of MAPE and RMSE for different kernels of SVR and comparing them with the traditional OLS method. The empirical findings demonstrate that the kernel of SVR which adapt better in my data is the linear one. Furthermore, we come to the conclusion that machine learnings methods adapt better than the traditional statistic methods.

When it comes to the neural network model, the input consists of the prices of independent variable, carbon EU ETF prices using two lags, i.e. one and two days back, the ESI, DAX40, the coal price, the oil price, the natural gas price, the CO2 emissions, the US EPU and the geopolitical risk index for lags as a feature. The number of neurons included in the hidden layers are determined based on a 5-fold cross validation training scheme. We find that the best artificial neural network models, that adapt better in my forecast model with different lags, has 5 layers, but they differ on activation functions and training algorithms. The result of evaluating the different models, which result by applying various hidden layers size, activation functions and trained algorithms.

The results show, for one lag, the layers of the best neural network contain 8, 10, 30, 10, 1 neurons and with the same activation function Rectified Linear Unit (ReLU). The neural network model with one lag is trained using Limited-memory BFGS (lbfgs) algorithm for 178 epochs. The best neural network, with five lags, has as activation function Hyperbolic Tangent Function (Tanh) and uses Limited-memory BFGS (lbfgs) as train algorithm for 114 epochs. Its layers contain the same neurons as the afore-mentioned neural network. The ones with fifteen and thirty lags use Adam for training and Tanh as activation function for 522 and 161 epochs, respectively. Their layers contain 8, 10, 8, 5, 1 neurons. The input for all different ANN models consists of 8 features.

When it comes to the decision tree for different lags, the decision tree model with one lag has five nodes and the root node of the decision tree is produced by selecting oil prices as the best input from the set of inputs that are available. In contrast, the decision tree models with five, fifteen and thirty lags have previous carbon EU ETF prices as a root node. The decision tree models for fifteen and thirty lags have five nodes, whereas the one with five lags has four nodes. The random forest with one lag has as input of first node the emissions of carbon dioxide. For five and fifteen lags, the first node of random forest is the prices of

carbon EU ETF with two lags. Conversely, the prices of carbon EU ETF with one lag are the input of random forest with thirty lags.

Comparing the value of RMSE, we conclude that the method of machine learnings that adapt better in forecasting the price of carbon EU ETFs is the random forest. Moreover, when we forecast the value of carbon EU ETFs over the time, we notice that the accuracy of our forecasting model is decreasing. Thus, our conclusions in the short term are more accurate than those in the long term.

Table 5. Depicts the in and out of sample RMSE and MAPE, with respect to evaluating the model which uses as input the prices of carbon EU ETF prices with two lags, the ESI, the DAX40,the coal prices, the oil prices, the natural gas prices, the CO2 emissions ,the US EPU and the geopolitical risk index with one, five, ten, fifteen and thirty lags, when I use the 80% of the dataset for training and the remaining 20% for testing. The results show that the best SVR kernel is the linear and the best machine learning methods in order to forecast the price of carbon EU ETF is the random forest. The asterisk denotes the lowest error

	RMSE		*MAPE*	
	In-sample	*Out-of-sample*	*In-sample*	*Out-of-sample*
Panel A:		t-1		
OLS	0.051	0.059	0.064	0.083
SVR-linear	0.052	0.059	0.057	0.074
SVR-RBF	0.010	0.413	0.100	0.937
Neural Network	0.014	0.014	0.002	0.002
Decision Tree	0.044	0.047	0.008	0.009
Random Forest	0.008	0.018	0.035	0.054
Panel B:			t-5	
OLS	0.050	0.062	0.062	0.077
SVR-linear	0.051	0.062	0.061	0.075
SVR-RBF	0.010	0.328	0.165	0.887
Neural Network	0.025	0.031	0.004	0.005
Decision Tree	0.049	0.045	0.009	0.008
Random Forest	0.015	0.033	0.050	0.074
Panel C:			t-15	
OLS	0.101	0.166	0.382	0.784
SVR-linear	0.102	0.169	0.420	0.452
SVR-RBF	0.010	0.443	0.057	2.081
Neural Network	0.024	0.045	0.004	0.007
Decision Tree	0.044	0.051	0.008	0.009
Random Forest	0.016	0.033	0.052	0.078
Panel D:		t-30		
OLS	0.129	0.125	0.502	0.210
SVR-linear	0.130	0.124	0.379	0.208
SVR-RBF	0.010	0.480	0.051	1.627
Neural Network	0.202	0.218	0.036	0.040
Decision Tree	0.047	0.051	0.008	0.009
Random Forest	0.014	0.042	0.048	0.083

In order to optimize the prediction model accuracy, using linear regression and random forest, we find the variables that are the most influential to the accuracy of the forecasting model. The empirical findings show that coal prices, oil prices, the CO2 emissions and the US EPU are the most important variables that determine carbon EU ETF prices. Hence, the form of forecasting model is the following:

$$Y_t = Y_{t-1} + Y_{t-2} + \log X_{coal,t-1} + \log X_{oil,t-1} + \log X_{CO_2,t-1}) + \log X_{EPU,t-1}\ \varepsilon_t \quad (15)$$

The results are almost the same as before and are depicted in table 6.

Table 6: Depicts the in and out of sample RMSE and MAPE, with respect to evaluating the model which uses as input the prices of carbon EU ETF prices with two lags, the ESI, the DAX40,the coal prices, the oil prices, the natural gas prices, the CO2 emissions ,the US EPU and the geopolitical risk index with one, five, ten, fifteen and thirty lags, when I use the 80% of the dataset for training and the remaining 20% for testing. The results show that the best SVR kernel is the linear and the best machine learning methods in order to forecast the price of carbon EU ETF is the random forest. The asterisk denotes the lowest error

	RMSE		*MAPE*	
	In-sample	*Out-of-sample*	*In-sample*	*Out-of-sample*
Panel A:		t-1		
OLS	0.051	0.059	0.064	0.083
SVR-linear	0.052	0.059	0.057	0.074
SVR-RBF	0.010	0.413	0.100	0.937
Neural Network	0.014	0.014	0.002	0.002
Decision Tree	0.044	0.047	0.008	0.009
Random Forest	0.008	0.018	0.035	0.054
Panel B:			t-5	
OLS	0.050	0.062	0.062	0.077
SVR-linear	0.051	0.062	0.061	0.075
SVR-RBF	0.010	0.328	0.165	0.887
Neural Network	0.025	0.031	0.004	0.005
Decision Tree	0.049	0.045	0.009	0.008
Random Forest	0.015	0.033	0.050	0.074
Panel C:			t-15	
OLS	0.101	0.166	0.382	0.784
SVR-linear	0.102	0.169	0.420	0.452
SVR-RBF	0.010	0.443	0.057	2.081
Neural Network	0.024	0.045	0.004	0.007
Decision Tree	0.044	0.051	0.008	0.009
Random Forest	0.016	0.033	0.052	0.078

contd. table 6

	RMSE		*MAPE*	
	In-sample	*Out-of-sample*	*In-sample*	*Out-of-sample*
Panel D:		t-30		
OLS	0.129	0.125	0.502	0.210
SVR-linear	0.130	0.124	0.379	0.208
SVR-RBF	0.010	0.480	0.051	1.627
Neural Network	0.202	0.218	0.036	0.040
Decision Tree	0.047	0.051	0.008	0.009
Random Forest	0.014	0.042	0.048	0.083

Under the same circumstances, we use the same machine learning methods for predicting and comparing with OLS regression. Comparing the results of all models, we notice that for both the models that use all variables and the models that use only the most important ones, the random forest is better for prediction than the other machine learning methods. Moreover, it follows that the best model for forecasting the prices of carbon EU ETF is that with all independent variables:

$$Y_t = Y_{t-1} + Y_{t-2} + \log X_{coal,t-1} + \log X_{oil,t-1} + \log X_{CO_2,t-1} + \log X_{EPU,t-1} + \log X_{DAX,t-1} + \log X_{natural\ gas,t-1} + \log X_{ESI,t-1} + \log X_{GPR,t-1} + \varepsilon_i \quad (16)$$

We observe that the characteristics of neural network models with various lags are the same with the previous neural network models. The same thing happens with the rest of the machine learning methods, support vector machine, decision tree, random forest. This motivates us to conclude that dependent variable, carbon EU ETF prices, is closely related to the previous value with tow lags. Furthermore, comparing RMSE and MAPE, as they are presented in the table 5 and 6, we notice that the forecast is the most accurate in the short term (has a smaller RMSE) than the long time. Moreover, the empirical findings show that the best SVR model is the one with the linear kernel and the machine learning methods adapt more closely to the phenomenon than the OLS models.

4. CONCLUSIONS AND POLICY RECOMMENDATIONS

This study examines the relationship between carbon EU ETFs prices and stock market, CO2 emissions, oil prices, natural gas prices and the economic sentiment and economy policy uncertainty index. We apply machine learning methods, specifically, support vector regression, artificial neural networks, decision trees and random forests for prediction. We examine the correlation between carbon EU ETFs prices and the independent variables. The empirical findings show that

carbon EU ETF prices with economic sentiment index and economy policy uncertainty index have a negative correlation. In contrast, carbon EU ETF prices and CO2 emissions are perfectly correlated. All others independent variables and carbon EU ETF prices are positively correlated.

Comparing machine learning methods with the OLS regression, it can be seen that the machine learning methods are better than the regression for carbon EU ETF prices forecasting. For the support vector machine, the best kernel is the linear. The best forecasting method is the random forest. This method adapts better in my forecasting model from the other ones. Finally, due to the fact that the RMSE in and out of sample gets bigger while the duration for prediction gets higher, we come to the conclusion that a short-term prediction is more accurate than a long-term forecast.

Overall, the selection of CO2 emissions and US EPU variables, as the ones that lead to the most accurate model, suggests that an increase in carbon dioxide emissions can rise carbon prices. Thus, the study recommends that the European council should be careful about the measures it takes, because they can lead to a decrease in carbon prices. Furthermore, if the US EPU index is high, the carbon prices are decreased, for instance, a volatility in stock markets provokes a reduction in carbon prices. The paper acknowledges further studies on the positive correlation between CO2 emissions with the carbon price. Furthermore, this study uses limited data for training, thus, the future ones can utilize a larger amount of data for more accurate outcomes. Additionally, the proposed prediction model can be furtherly developed and improved towards better performance and more accurate prediction results, by using more variables that may affect the carbon price but are yet to be examined.

NOTES

1. ESI is a measure that calculates the confidence levels between manufacturers (which constitute the 40% of the index), service providers (which constitute the 30% of the index), consumers (which constitute the 20% of the index), retailers (which constitute the 5% of the index) and constructors (which constitute the 5% of the index).
2. US EPU is an index which counts the number of newspaper articles that include the terms "uncertain" or "uncertainty", "economic" or "economy", and one or more policy-relevant terms.

REFERENCES

Gordon, J. (2022). Carbon ETFs: Driving real word impact. Available at: https://www.etfstream.com/news/carbon-etfs-driving-real-world-impact/ (Accessed: 20 January 2022).

Ji, C.; Hu, Y.; Bao-Jun, T.; Research on carbon market price mechanism and influencing factors: a literature review. Nat Hazards. 2018, 92.

Koch, N.; Fuss, S.; Grosjean, G; Edenhofer, O; Causes of the EU-ETS price drop: Recession, CDM, renewable policies or a bit of everything?-New evidence. Energy Policy. 2014, 73.

Jiang, Y.; Lei, Y.; Yang, Y.; Wang, F.; Factors affecting the pilot trading market of carbon emissions in China. Petroleum Science. 2018, 15.

Ji, C.; Hu, Y.; Tang, B.; Qu, S.; Price drivers in the carbon emissions trading scheme: Evidence from Chinese emissions trading scheme pilots. Journal of Cleaner Production. 2021, 278.

Zenga, S.; Nana, X.; Liua, C.; Chena, J.; The response of the Beijing carbon emissions allowance price (BJC) to macroeconomic and energy price indices. Energy Policy. 2017, 106.

Chevallier, J.; A model of carbon price interactions with macroeconomic and energy dynamics. Energy Economics. 2011, 33.

Caldara, D.; Iacoviello, M.; Measuring geopolitical risk. FRB International Finance Discussion Paper. 2018, 1222.

Mitsas, S.; Golitsis, P.; Khudoykulov, K.; Cogent Economics and Finance. 2022, 10.

Qin, Y.; Hong, K.; Chen, J.; Zhang, Z.; Asymmetric effects of geopolitical risks on energy returns and volatility under different market conditions. Energy Economics. 2020, 90.

Dou, Y.; Li, Y.; Dong, K.; Ren, X.; Dynamic linkages between economic policy uncertainty and the carbon futures market: Does Covid-19 pandemic matter? Resources Policy. 2022, 75.

Zhu, B.; Ye, S.; Han, D.; Wang, P.; He, K.; Wei, Y.; Xie, R.; A multiscale analysis for carbon price drivers. Energy Economics. 2019, 78.

Li, J.; Macro carbon price prediction with support vector regression and Paris accord targets. Available at: https://arxiv.org/abs/2212.11787 (Accessed: 30 November 2022).

Cortes, C.; Vapnik, V.; Support-Vector Networks. Machine Learning. 1995, 20.

Ho T.; The random subspace method for constructing decision forests. IEEE Transaction on pattern analysis and machine intelligence. 1998, 20.

Morgan, J., Sonquist, J.; Problems in the Analysis of Survey Data, and a Proposal. Journal of the American Statistical Association. 1963, 58.