Forecasting Carbon Prices Using Machine Learning: A Comparative Study of ARIMA, Random Forest, and LSTM Models

Author: Erica Zhao

2024 Spring

Abstract

Abstract

1 Introduction

Introduce the topic, importance of forecasting carbon prices, and a brief introduction to the methods compared.

2 Literature Review

Discuss previous studies and their findings on forecasting carbon prices and similar economic indicators.

3 Methodology

3.1 ARIMA Model

ARIMA model has notable strengths for forecasting carbon prices, utilizing its robust capability to process time series data and capture the dynamic patterns in the volatile carbon prices. These features facilitate a comprehensive analysis, making ARIMA an ideal choice for short-term forecasting. However, we also need to acknowledge its potential limitations, such as the assumption of linearity and the need for data stationarity, which may cause potential issues.

Central to the ARIMA model is its ability to integrate autoregressive (AR) processes, differencing (I) to achieve stationarity, and moving average (MA) processes into a unified structure. The model is parameterized by three integers: p, d, and q, which denote the autoregressive order, the degree of differencing, and the moving average order, respectively. Mathematically, the ARIMA(p,d,q) model is expressed as:

$$x_t = u + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \alpha_p x_{t-p} - \beta_1 \varepsilon_{t-1} - \beta_2 \varepsilon_{t-2} - \ldots - \beta_q \varepsilon_{t-q} + \varepsilon_t \quad (1)$$

Here, x_t represents the forcasted carbon price at time t, and u is a constant term (which can be interpreted as the mean of the series if $d = 0$). The parameters $\alpha_1, \alpha_2, \ldots, \alpha_p$ represent the autoregressive (AR) terms, $\beta_1, \beta_2, \ldots, \beta_q$ represent the moving average (MA) terms, and $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-q}$ are the lagged forecast errors in the prediction equation.

In this study, we employed the ARIMA model to analyze fluctuations in carbon prices. Our approach commenced with an initial hypothesis based

on a preliminary examination of the dataset to identify any apparent trends and autocorrelations. Utilizing diagnostic tools such as autocorrelation and partial autocorrelation plots, alongside statistical tests, we determined the optimal differencing order and the appropriate orders for the autoregressive and moving average components of the model. This step also included assessments for the inclusion of seasonal adjustments. Subsequent to model identification, we engaged in parameter estimation through Maximum Likelihood Estimation methods to accurately fit the model to the historical data. The model was then validated to ensure that the residuals approximated white noise, indicating that the model captured the essential structures in the data. Following validation, we proceeded to utilize the ARIMA model for forecasting future carbon prices, leveraging its capability to analyze and interpret underlying trends in the data over various forecasting horizons. This methodological framework underpins our analysis and forecasts presented herein.

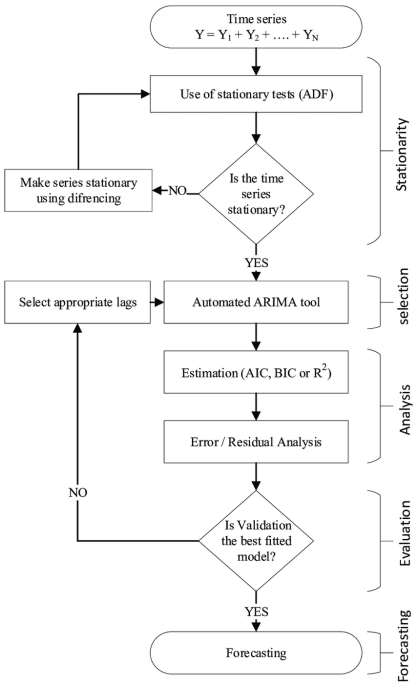

Figure 1: Figure 1. Flow chart of the ARIMA model selection process.

3.2 Random Forest

In this study, the Random Forest (RF) methodology was employed as the second ensemble-based predictive model for forecasting carbon prices. Random Forest is a form of the ensemble method known as Bagging, or Bootstrap Aggregating, which operates on the principle of creating multiple datasets from the original training set through bootstrap sampling. This sampling technique involves random selection with replacement, resulting in subsets where individual instances may repeat or be excluded.

Renowned for its robustness, the Random Forest is inherently adept at handling large, high-dimensional datasets and mitigating overfitting—a common pitfall in complex models. The ensemble nature of Random Forest, where multiple decision trees vote on the outcome, reduces variance and enhances performance stability. Nonetheless, it may incur considerable computational expense, particularly with a large number of trees, and can be less interpretable due to its ensemble complexity.

The RF algorithm constructs a multitude of decision trees $(T_1, T_2, \ldots, T_N)$ during the training process. For regression tasks, the RF employs bootstrap aggregating, where each tree T_i is trained on a bootstrap sample D_i, drawn with replacement from the complete training dataset D. During tree construction, a random subset of m features is selected at each node for splitting. Given a new input x, each tree T_i yields a prediction Y_i. The RF's final prediction $\hat{Y}$ for input x is the average of these predictions:

$$\hat{Y}(x) = \frac{1}{N} \sum_{i=1}^{N} Y_i(x) \tag{2}$$

Within each tree, the learning objective is to minimize the prediction error, quantified by the mean squared error (MSE), following the least squares approach:

$$\mathrm{Err}(t) = \frac{1}{n_t} \sum_{D_t} (y_i - k_t)^2 \tag{3}$$

where n_t represents the number of instances in node t, and k_t is the average of instances in each node.

Implementation within this study followed a systematic approach: An ensemble of decision trees was constructed, each predicated on a distinct bootstrap sample. Key parameters, such as the number of trees (denoted as n_estimators in machine learning frameworks), maximum depth, and the

number of features considered at each split, were meticulously calibrated. The decision trees were trained individually and the Random Forest aggregated outputs from individual trees in the prediction phase. The ensemble's prediction corresponds to the average of the outputs, thereby converging towards a central tendency.

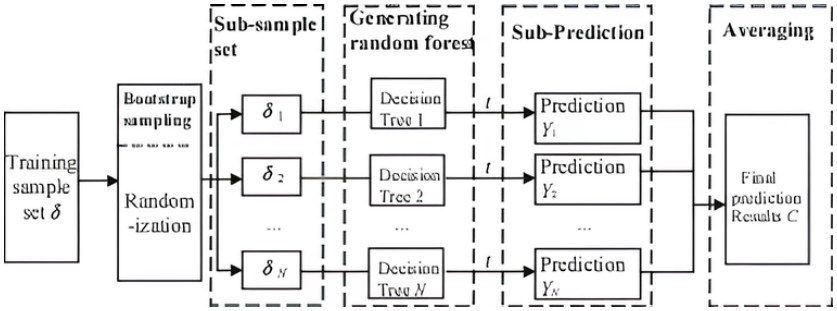

Figure 2: Figure 2. Flow chart of the RF model selection process.

3.3 LSTM Model

Long Short-Term Memory (LSTM) networks, the third model integrated into our analysis, represent an evolution of recurrent neural networks (RNNs) specifically engineered to predict carbon prices. LSTMs excel in capturing and leveraging long-term dependencies within time-series data, a capability crucial for the accurate forecasting of dynamic economic indicators such as carbon prices. Their architecture is intricately designed to address the vanishing gradient problem—a significant challenge in traditional RNNs—thereby enhancing their learning capacity and predictive performance.

LSTM units include memory cells that preserve states over time, enabling sustained information retention throughout the sequence. The mathematical operations of an LSTM cell are governed by the following equations:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{4}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{5}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{6}$$

$$c_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{7}$$

$$h_t = o_t * \tanh(c_t) \tag{8}$$

The forget gate f_t, input gate i_t, and output gate o_t within each cell dictate the preservation and incorporation of information, ensuring the model's adaptability to the intricacies of the data.
In our implementation, LSTM networks were trained on historical carbon price data, carefully prepared to facilitate effective learning. The architecture of the LSTM was designed to reflect the complexity of the data, with the optimization of hyperparameters guided by validation performance. Training involved backpropagation through time to allow the network to capture the temporal dynamics inherent to the carbon market. The trained LSTM model was then deployed to forecast future carbon prices, utilizing its ability to extrapolate observed trends into actionable insights.

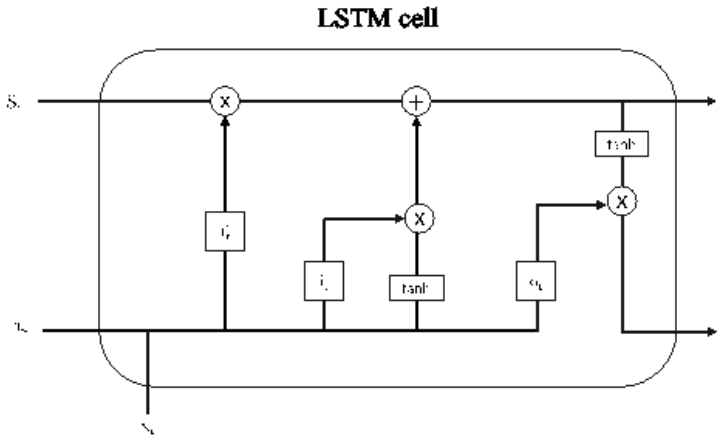

Figure 3: Figure 3. Flow chart of the LSTM model selection process.

4 Data and Descriptive Statistics

In the analysis of carbon pricing trends within the European Union's Emissions Trading System (EU ETS), this study draws upon a meticulously compiled dataset encompassing the period from January 1, 2021, to January 31, 2024. During this time frame, a total of 763 data points were collected, providing a granular view of carbon price movements. The carbon price data, sourced from Investing.com, is critical for understanding the economic implications of regulatory shifts and market dynamics under the EU's climate policy framework.

This timeframe is deliberately aligned with the initiation of Phase 4 of the European Union's Emissions Trading System (EU ETS), marking a significant transition from the system's previous phases. The focus on this specific period is due to the substantial regulatory shifts introduced in Phase 4, which differ markedly from the earlier phases in terms of emissions reduction targets and allowance allocation mechanisms.

Notably, in Phase 3 of the EU ETS (2013-2020), the Union-wide cap for stationary installations was subject to an annual decrease by a linear reduction factor of 1.74%, with the 2013 cap determined based on the average total quantity of allowances issued annually during 2008-2012. Conversely, Phase 4 (2021-2030) introduces a more aggressive approach to emissions reduction, with the cap on emissions decreasing annually at an increased linear reduction factor of 2.2%. This heightened ambition reflects the EU's commitment to achieving more substantial emissions reductions and necessitates an in-depth analysis of carbon pricing trends under the new regulatory framework. The study aims to isolate the impacts of these changes on carbon pricing, underscoring the importance of understanding the dynamics introduced by Phase 4's enhanced regulatory measures.

The descriptive statistics for the dataset, which includes carbon pricing trends within the European Union's Emissions Trading System (EU ETS) from January 1, 2021, to January 31, 2024, are presented in Table 1. These statistics provide insights into the behavior of the carbon market during Phase 4 of the EU ETS. The carbon price exhibits a mean of €72.98 and a standard deviation of €16.45, reflecting market volatility over the 763 observations in the dataset.

Table 1: Descriptive Statistics of the Dataset

Variable	Mean	Standard Deviation	Maximum	Minimum
Carbon	72.98	16.45	98.01	31.96
Coal	176.47	93.46	439.00	63.75
Crude Oil	83.82	14.26	127.98	51.09
Natural Gas	72.45	54.61	339.20	15.53
Fit_for_55	0.83	0.38	1	0
Law	0.22	0.41	1	0
Compromise	0.25	0.43	1	0
EPU	276.88	69.80	418.62	141.39
DAX	14980.43	1082.54	16972.34	11975.55
SPGTCLNT	1699.21	295.68	2720.79	1088.34
FTMIB	25692.05	2414.81	30744.24	20352.98
CAC40	6722.63	526.14	7677.47	5399.21
AEX	732.69	44.97	827.57	630.58
STOXX	448.01	22.92	494.35	382.89
Temperature	12.94	6.64	31.1	-3.2
Precipitation	1.53	4.05	36.03	0.0
Solar Energy	6.75	5.88	27.1	0.1
Wind Speed	11.39	3.82	30.7	3.8

The time series plot of the carbon price, depicted in Figure 4, emphasizes the fluctuations over the observed period. Furthermore, the study will explore the patterns in related energy commodities, policy indicators, European stock indices, and weather variables, all of which could influence emission levels and subsequently, carbon market dynamics.

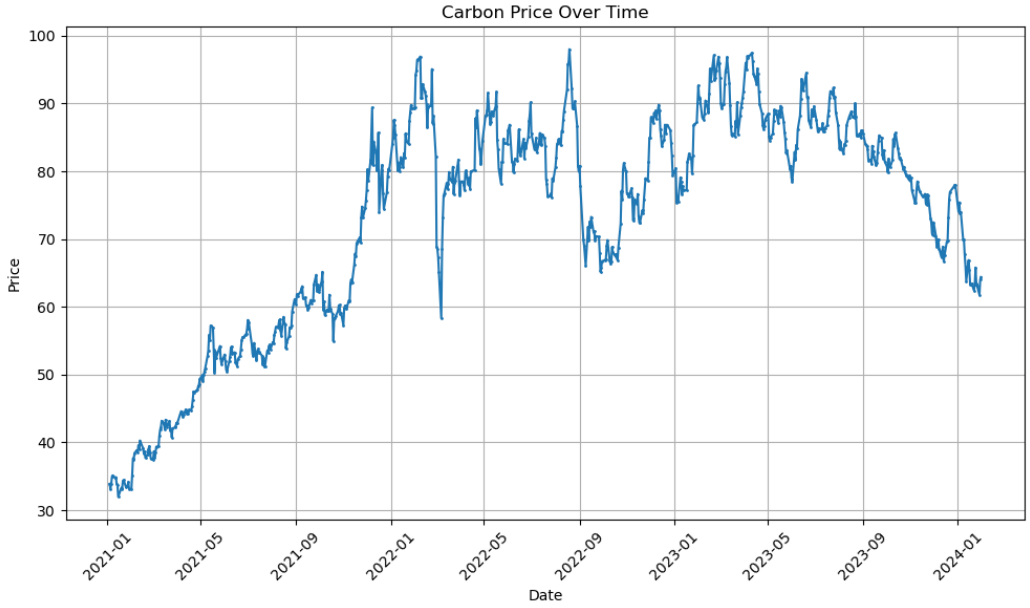

Figure 4: Time Series Plot of Carbon Price from January 2021 to January 2024.

5 Forecasting of Carbon Price

5.1 Evaluation Criteria

Evaluating the accuracy and reliability of our predictive model is crucial for ensuring its practical applicability in forecasting carbon prices. In this study, we have utilized several statistical measures to assess model performance. The primary metric we considered is the Root Mean Squared Error (RMSE), which provides an estimate of the model's prediction error magnitude, offering an objective performance measure less sensitive to outliers. The Mean Absolute Percentage Error (MAPE) is used as a supplementary metric to understand the relative size of the forecast errors in percentage terms. These metrics are formally defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \hat{x}_t)^2} \tag{9}$$

$$\text{MAPE} = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{x_t - \hat{x}_t}{x_t} \right| \tag{10}$$

where x_t represents the actual observed value, $\hat{x}_t$ denotes the forecasted value, and n is the number of observations.

The RMSE serves as the principal measure due to its ability to capture the absolute errors in the model's predictions. Conversely, the MAPE provides additional context by illustrating the error magnitude relative to the actual values, which is particularly insightful when considering the proportional accuracy of the model across different price levels.

To provide a comprehensive evaluation, the model's performance is assessed not only through RMSE and MAPE but also by analyzing the Mean Absolute Error (MAE), Mean Square Error (MSE), and the Coefficient of Estimations. These metrics together offer a holistic view of the model's predictive capabilities.

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^{n} |x_t - \hat{x}_t| \tag{11}$$

$$\text{MSE} = \frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{x}_t)^2 \tag{12}$$

5.2 Parameter Determination

Discuss how parameters were determined for each model.

5.2.1 ARIMA Model

5.2.2 RF Regression

5.2.3 LSTM Model

5.3 Performance and Results

5.3.1 In Sample Performance

5.3.2 Out Sample Performance

Present the results of each model.

5.4 Discussion

Discuss the results, implications, and any observed trends or anomalies.

6 Conclusion

Summarize the findings, their implications for