Forecasting of Steam Coal Price Based on Robust Regularized Kernel Regression and Empirical Mode Decomposition

Xiangwan Fu[1†], Mingzhu Tang[1*], Dongqun Xu[2], Jun Yang[3], Donglin Chen[1] and Ziming Wang[4†]

[1]School of Energy and Power Engineering, Changsha University of Science and Technology, Changsha, China, [2]China Datang Corporation Ltd, Beijing, China, [3]Hunan Datang Xianyi Technology Co., Ltd., Changsha, China, [4]School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, China

Aiming at the problem of difficulties in modeling the nonlinear relation in the steam coal dataset, this article proposes a forecasting method for the price of steam coal based on robust regularized kernel regression and empirical mode decomposition. By selecting the polynomial kernel function, the robust loss function and L2 regular term to construct a robust regularized kernel regression model are used. The polynomial kernel function does not depend on the kernel parameters and can mine the global rules in the dataset so that improves the forecasting stability of the kernel model. This method maps the features to the high-dimensional space by using the polynomial kernel function to transform the nonlinear law in the original feature space into linear law in the high-dimensional space and helps learn the linear law in the high-dimensional feature space by using the linear model. The Huber loss function is selected to reduce the influence of abnormal noise in the dataset on the model performance, and the L2 regular term is used to reduce the risk of model overfitting. We use the combined model based on empirical mode decomposition (EMD) and auto regressive integrated moving average (ARIMA) model to compensate for the error of robust regularized kernel regression model, thus making up for the limitations of the single forecasting model. Finally, we use the steam coal dataset to verify the proposed model and such model has an optimal evaluation index value compared to other contrast models after the model performance is evaluated as per the evaluation index such as RMSE, MAE, and mean absolute percentage error.

Keywords: the steam coal price forecasting, kernel function, empirical mode decomposition, Huber loss function, L2 regular term

INTRODUCTION

Accurate forecasting of the steam coal price can provide a certain basis for enterprises related to steam coal to formulate the procurement plan. The steam coal price indicates the general performance of supply and demand on the steam coal market. China is a big consumer of coal (Xiong and Xu, 2021; Wang and Du, 2020). Forecasting the steam coal price accurately can help in the analysis of the steam coal market, grasp the implied law in the steam coal market, and improve the steam coal market's efficiency.

In recent years, many references have proposed many different methods for coal price forecasting. The time series model can mine the implicit law in the time series. Matyjaszek et al. removed the effect of abnormal fluctuations in prices on the forecasting model by using the transgenic time series (Matyjaszek et al., 2019). Ji et al. effectively improved the forecasting accuracy of the forecasting model by using the ARIMA model and neural network model (Ji et al., 2019). Wu et al. decomposed the price series into several components, used the ARIMA model and SBL model for coal price forecasting, and added up the forecasted values of all the components as final forecasting result. Compared to the contrast model adopted, this model can effectively improve the model forecasting precision (Wu et al., 2019). Chai et al. combined STL decomposition method with ETS model. The experimental results show that it has the best forecasting performance compared with benchmark models and neural network (Chai et al., 2021). It is difficult to learn the implicit nonlinearity law in the data by using the time series model, which is sensitive to the abnormal value in the data and only considers a single variable factor other than other influence factors.

A neural network model can learn the nonlinearity law in the data which is studied based on the interconnection between nerve cells. Alameer et al. effectively improved the forecasting accuracy of coal price based on LSTM model and DNN model (Alameer et al., 2020). Lu et al. adopted the full empirical mode to decompose and preprocess the original dataset, and then chose the radial basis function neural network model for model training and forecasting. The results show higher stability (Lu et al., 2020). Yang et al. adopted the improved whale optimization algorithm to optimize the decomposition and LSTM combined model based on the improved integration empirical model, which has a better model forecasting performance compared to other reference models (Yang et al., 2020). Zhang et al. decomposed the original data series by multi-resolution singular value decomposition method and forecasted the coal price by using MFO-optimized ELM model. Experimental results show the forecasting performance of the proposed model was superior to that of the contrast model (Zhang et al., 2019). However, the neural network model is a black box model which is difficult to interpret.

The steam coal market is a complex nonlinear system, containing influence factors such as economy, steam coal transportation, steam coal supply, and steam coal demand. The influence factors involve a wide range and many feature data and contain some noise data. This method improves the model interpretability by using linear model and reduces the adverse impact of noise data on the forecast model by using Huber loss function (Gupta et al., 2020). We use the kernel function to mine the implicit nonlinearity law in the steam coal data (Li and Li, 2019; Vu et al., 2019; Ye et al., 2021). The combined model can improve the model performance based on the advantages of the sub-model (Wang et al., 1210; Zhou et al., 2019; Wang et al., 2020a; Wang et al., 2020b; Qiao et al., 2021; Zhang et al., 2021). This method can decompose the forecasting error of the forecasting model into multiple modal components by using the EMD method (Yu et al., 2008; Xu et al., 2019; Wang and Wang, 2020; Xia and Wang, 2020), build the ARIMA model (Conejo et al., 2005; Karabiber and Xydis, 2019) for each modal component for forecasting, and add up the

forecasted values of all the modal components to compensate error for the original forecasting model.

For the problem of difficulties in modeling the nonlinear relation in the steam coal dataset, this article proposes a forecast method for the price of steam coal based on robust regularized kernel regression and empirical mode decomposition. The second part introduces the used algorithm theory content; the third part states the data preprocessing steps, the selection of features, and the whole process of model training and forecasting; the fourth part shows the model comparison and experimental results; and the fifth part contains conclusion and prospect.

METHODOLOGY

Huber–Ridge Model

The Huber function (Huber et al., 1992) has great robustness, which can effectively reduce the negative influence of abnormal data on model performance. The Huber loss function is shown in **Eq. 1**:

$$\phi_{\text{hub}}(u) = \begin{cases} u^2 & |u| \le M \\ M(2|u| - M) & |u| > M \end{cases}, \tag{1}$$

where u is the residual value and M is the threshold value of the Huber function. The Huber function imposes the punishment which is larger than the threshold value residual to effectively lower the influence of abnormal sample points on the model training.

The Ridge model is added with L2 penalty term based on the objective function of the linear regression model. The objective function of the model is shown in **Eq. 2**:

$$\min_{\omega} \frac{1}{2} \|X\omega - Y\|_2^2 + \alpha_{l2} \|\omega\|_2^2, \tag{2}$$

where X refers to the set of feature parameters, ω refers to the weight coefficient vector, Y refers to the forecasted target quantity, $\|\omega\|_2^2$ refers to the L2 penalty term, and α_{l2} refers to the regular coefficient.

L2 regular term compresses the feature weight value adversely to the model forecasting and makes it approximate to 0 in order to reduce the impact of features with low correlation. When the regular coefficient of α_{l2} is large, the L2 regular term makes more parameters' weight in the parameter weight vector approximate to 0 to screen out main features and mitigate the degree of model overfitting to some extent.

The Huber loss function and L2 regular term are combined to construct the Huber–Ridge model (Owen, 2006), improving the model robustness and lowering the overfitting risk. Its objective function is shown in **Eq. 3**:

$$\widehat{w}_j = \operatorname{argmin}_w \left(\phi_{\text{hub}}(u) + \frac{\lambda}{2} \sum_{j=1}^{k} (w_j)^2 \right). \tag{3}$$

Polynomial Kernel Huber–Ridge Model

The T.M. Cover theorem (Cover, 1965) points out that the data in the high-dimensional space can show the linearity law more

easily. The kernel function maps the vector of low-dimensional feature space to the high-dimensional feature space, and transforms the nonlinearity law in the low-dimensional feature space into linearity law in the high-dimensional space to learn the linearity law in the high-dimensional space by using the linear model and indirectly learn the nonlinearity law in the original feature space based on the model. Due to the high-dimensional feature space having high dimensionality, the dimension disaster may happen if the model is directly used for fitting in the high-dimensional space. The introduced kernel function can effectively solve the above problem, and the kernel function can represent inner product value in the high-dimensional space with the inner product value in the low-dimensional space. Thus, it can avoid the inner product calculation in the high-dimensional space and greatly reduce the calculation of the model.

The regular risk functions have a unified expression mode (Schölkopf et al., 2001), as shown in **Eq. 4**:

$$f(x) = \sum_{i=1}^{n} w_i k(x, x_i) + b. \tag{4}$$

The kernel function is introduced to the Huber–Ridge model (Jianke Zhu et al., 2008). Thus, the model can learn the implicit nonlinearity law in the data. The objective function of Huber–Ridge kernel model is shown in **Eq. 6**:

$$u = y_i - f(x), \tag{5}$$

where y_i is the actual value, $f(x)$ is the forecasting value, and u is the forecasting error

$$\phi_{\text{hub}}(y_i, f(x)) = \begin{cases} M(2u - M) & A_1 = \{x | u > M\} \\ u^2 & A_2 = \{x | - M \le u \le M\}, \\ -M(2u - M)A_3 & A_3 = \{x | u > M\} \end{cases} \tag{6}$$

$$\widehat{w}_j = \underset{w}{\text{argmin}} \sum_{j=1}^{n} \phi_{\text{hub}}\left(y_i, \sum_{i=1}^{n} k(x_i, x_j)w_j + b\right)$$
$$+ \lambda \sum_{i,j=1}^{n} w_i w_j k(x_i, x_j), \tag{7}$$

$$T = \underset{w}{\text{argmin}} \sum_{i=1}^{n} \phi_{\text{hub}}\left(y, \sum_{i=1}^{n} k_i^T w + b \cdot I\right) + \lambda w^T K w. \tag{8}$$

Eqs. 9–12 can be obtained, respectively, by getting the partial derivative of w and b:

$$\frac{\partial T}{\partial w} = 2\left(\lambda K w + K I^0 K w + K q\right) = 0, \tag{9}$$

$$q = -I^0 y + me + b \cdot I^0, \tag{10}$$

$$e_i = \begin{cases} 1 & x_i \in A_1 \\ 0 & x_i \in A_2, \\ -1 & x_i \in A_3 \end{cases} \tag{11}$$

$$\frac{\partial T}{\partial b} = 2\left(I^0 K w + q\right) = 0, \tag{12}$$

where I^0 is a diagonal matrix of shape (n, n), the values of its elements in the A domain are 1, and the remaining values are 0.

The basic Newton method is used to iteratively update w and b, as shown in **Eq. 13**:

$$\begin{bmatrix} w' \\ b' \end{bmatrix} = \begin{bmatrix} w \\ b \end{bmatrix} - \gamma H^{-1} \nabla, \tag{13}$$

where H^{-1} is the first derivative of the gradient matrix as shown in **Eq. 14**, γ is the length of the step, usually with a value of 1, and ∇ is a Hessen matrix as shown in **Eq. 15**:

$$H^{-1} = \frac{1}{2} \begin{bmatrix} \lambda K + K I^0 K & I^0 K \\ K I^0 & I^0 \end{bmatrix}^{-1}, \tag{14}$$

$$\nabla = 2 \begin{bmatrix} \lambda K w + K I^0 K w + K q \\ I^0 K w + q \end{bmatrix}. \tag{15}$$

Simplify **Eq. 13** to obtain the final computational equation of objective function of the kernel Huber–Ridge model, as shown in **Eq. 16**:

$$\begin{bmatrix} w' \\ b' \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix} + \begin{bmatrix} \lambda I + I^0 K & 1 \\ 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} q \\ 0 \end{bmatrix}. \tag{16}$$

The polynomial kernel is a commonly used kernel function, and the polynomial kernel function is shown in **Eq. 17**:

$$\mathbf{K}(\mathbf{x}, \mathbf{x}) = (\mathbf{x} \cdot \mathbf{x} + 1)^d, \tag{17}$$

where $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ is the feature vector and d is the kernel parameter. The polynomial kernel function (17) is substituted into **Eq. 16** to obtain the final computational equation of polynomial kernel Huber–Ridge model objective function.

EMD Model

The empirical mode decomposition is a signal decomposition technology and decomposes the original signal into a series of components which are the intrinsic mode functions. The empirical mode decomposition is often used to handle the time series data and decomposes the original time series into a series of different components to explore the implicit law in the time series data.

The intrinsic mode function should meet the two conditions below:

1. In the data interval, difference between numbers of extreme points and zero points is at most one.
2. The average value of the upper envelope and the lower envelope is zero.

The EMD model is adaptive and can decompose the original series for a time series data without the number of components specified till the standard of stopping decomposition is met. The relationship between the original series and the decomposed components is shown in **Eq. 18**:

$$X(t) = \sum_{i=1}^{n} imf_i + r, \tag{18}$$

where $X(t)$ refers to the original time series, $\sum_{i=1}^{n} imf_i$ refers to the sum of the components, and r refers to the residual. When the residual series is a monotonic function, the decomposition stopped.

The decomposition step of empirical mode decomposition is shown as follows:

STEP 1: Identify all the maximum points and minimum points in the time series, and fit the upper envelope e_u and the lower envelope e_l by using the cubic spline finite difference method according to maximum points and minimum points.

STEP 2: Calculate the average value of the upper envelope e_u and the lower envelope e_l, and obtain the mean envelope e_{mean} $\left(e_{mean} = \frac{e_u + e_l}{2}\right)$.

STEP 3: Calculate the difference between the original series $X(t)$ and the mean envelope e_{mean}, and obtain the intermediate time series ($e_i = X(t) - e_{mean}$).

STEP 4: Judge whether the intermediate time series e_i can be an intrinsic mode function according to the constraint condition of intrinsic mode functions. If satisfied, the intermediate time series shall be used as the imf_i component. If unsatisfied, such intermediate time series shall be used as the basis to execute the steps 1–4.

STEP 5: Subtract the component imf_i from the original time series $X(t)$ and execute the steps 1–4 again. If the standard of stopping decomposition is satisfied, the decomposition process will end.

ARIMA Model

The autoregressive integrated moving average (ARIMA) model is defined in **Eq. 19**:

$$y^t = \varphi_1 y^{t-1} + \varphi_2 y^{t-2} + \cdots + \varphi_p y^{t-p} + \varepsilon^t - \theta_1 \varepsilon^{t-1} - \theta_2 \varepsilon^{t-2} - \cdots \\ - \theta_q \varepsilon^{t-q} + \theta_0,$$

(19)

where y^t and ε^t indicate the actual value and residual value at the time point t, respectively, and $\varphi = (\varphi_1, \varphi_2, \cdots, \varphi_p)$ and $\theta = (\theta_1, \theta_2, \cdots, \theta_q)$ refer to the weight vectors. p and q are the model orders. The historical time series data and historical white noise error data of the variable are used to forecast the current value.

The prerequisite of using ARIMA model is to use stationary data. The non-stationary data can be handled by combining the autoregressive integrated moving average (ARIMA) model and different methods (Gilbert, 2005). The ARIMA model has three parameters, (p, d, q), in which d refers to the differential order of the data series.

Evaluation Indexes

The model performance is evaluated by the mean absolute error (MAE) and the definition of MAE is shown in **Eq. 20**:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \widehat{y}_i|.$$

(20)

The root-mean-square error (RMSE) is used for model performance assessment and the definition of RMSE is shown in **Eq. 21**:

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

(21)

The mean absolute percentage error (MAPE) is used for model performance assessment and the definition of MAPE is shown in **Eq. 22**:

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \widehat{y}_i}{y_i} \right|,$$

(22)

where n is the number of samples in the set verified, $\widehat{y}_i$ is the forecasted value of the model, and y_i is the true value. The closer the MAE, RMSE, and MAPE values are to 0, the better the model performance will be.

The Training and Forecasting Process of Polynomial Kernel Huber–Ridge–EMD–ARIMA Model

The training and forecasting process for the forecast framework of polynomial kernel Huber–Ridge–EMD–ARIMA model (PK–Huber–Ridge–EMD–ARIMA) proposed is shown in **Figure 1**. Its process steps are as follows.

STEP 1: Data preprocessing and feature selection: Screen the correlation features according to Pearson correlation coefficient and Spearman correlation coefficient after the data preprocessing and divide training dataset and test dataset.

STEP 2: Model training: The model parameter M, α_{l2}, and training dataset are input to the polynomial kernel Huber–Ridge for model training.

STEP 3: Model forecasting: The test dataset is input to the trained polynomial kernel Huber–Ridge model. The model outputs the time series $\{y'_1, \cdots, y'_k, y'_{k+1}\}$ of coal price forecasting data.

STEP 4: Forecast the forecasting error of steam coal price at the next time point: The forecasting error series $\{\varepsilon_1, \cdots \varepsilon_k\}$ of coal price is input to the EMD model, and the EMD model outputs j modal components $(IMF_1, \cdots, IMF_j)$. Each modal component is subject to training and forecasting by the ARIMA model; the ARIMA model $(ARIMA_1, \cdots, ARIMA_j)$ corresponding to each modal component outputs the coal price forecasting error $(\varepsilon^1_{k+1}, \cdots, \varepsilon^j_{k+1})$ of each modal component at the next time point, respectively. The $(\varepsilon^1_{k+1}, \cdots, \varepsilon^j_{k+1})$ series accumulation is conducted, and the accumulation result $\{\varepsilon_{k+1}\}$ is used as the forecasted value of the coal price forecasting error at a time point k+1.

STEP 5: Obtain the final forecasted value of steam coal price at the next time point: The forecasted value $\{\varepsilon'_{k+1}\}$ of coal price forecasting error at the time point k+1 is used for $\{y'_{k+1}\}$ correction of the forecasted value of coal price at a time point k+1. The forecasted value $\{y''_{k+1}\}$ of coal price after correction is used as the forecasted value of coal price at the final time point k+1.

EMPIRICAL STUDY

Data Description

Qinhuangdao steam coal price data that are 4500-kilocalorie, 5000-kilocalorie, and 5500-kilocalorie steam coal exit price data of Qinhuangdao Port from Jan. 2017 to Jul. 2021, are used for experimental study. The sampling is conducted once a week, which is the weekly frequency data.
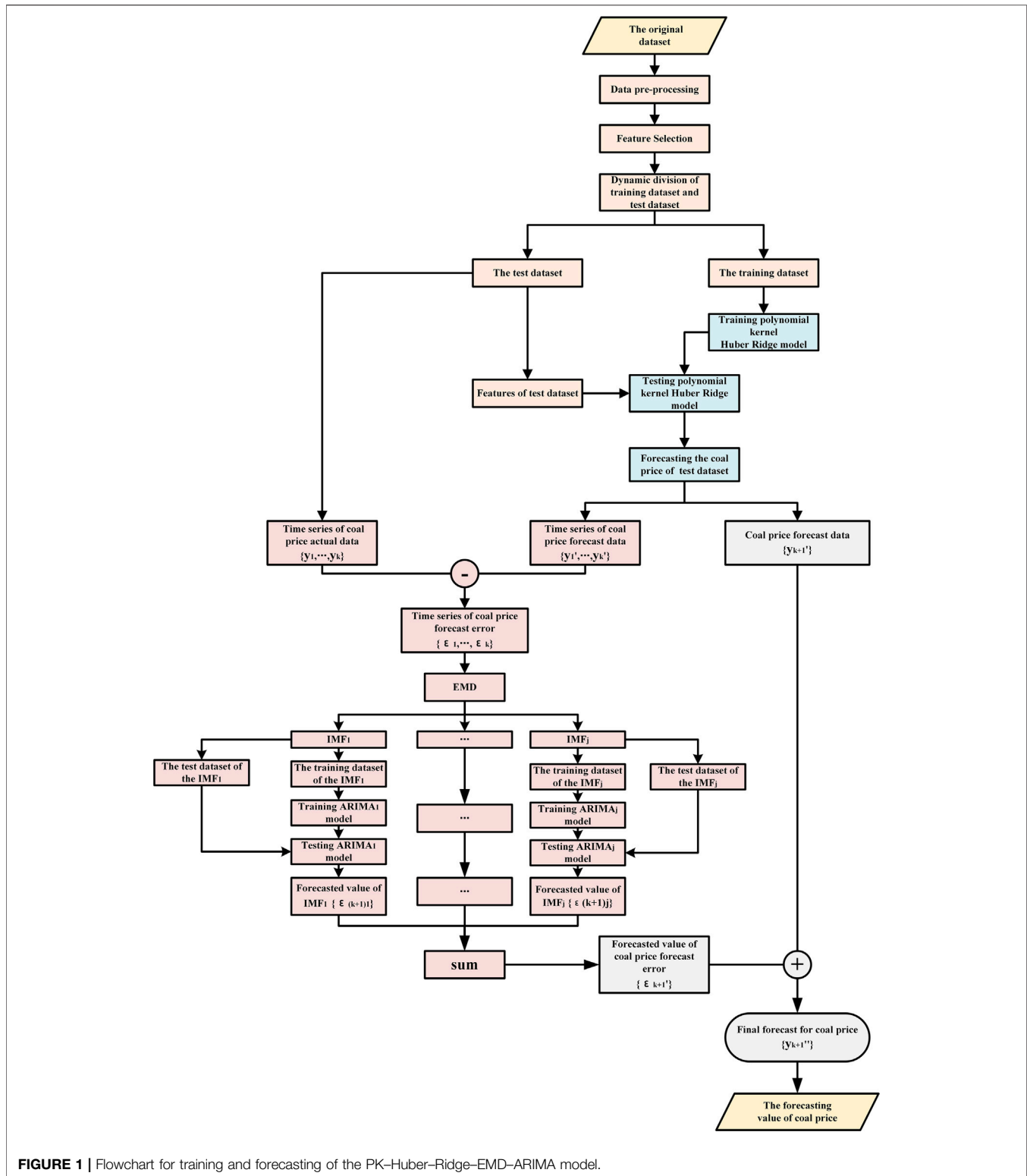
FIGURE 1 | Flowchart for training and forecasting of the PK–Huber–Ridge–EMD–ARIMA model.

Data Preprocessing

The original data about steam coal price and its features have the disadvantage of missing value and inconsistent data time sampling frequency; so, such original data shall be pre-processed and the data processed can be brought into the model for model training and forecasting.
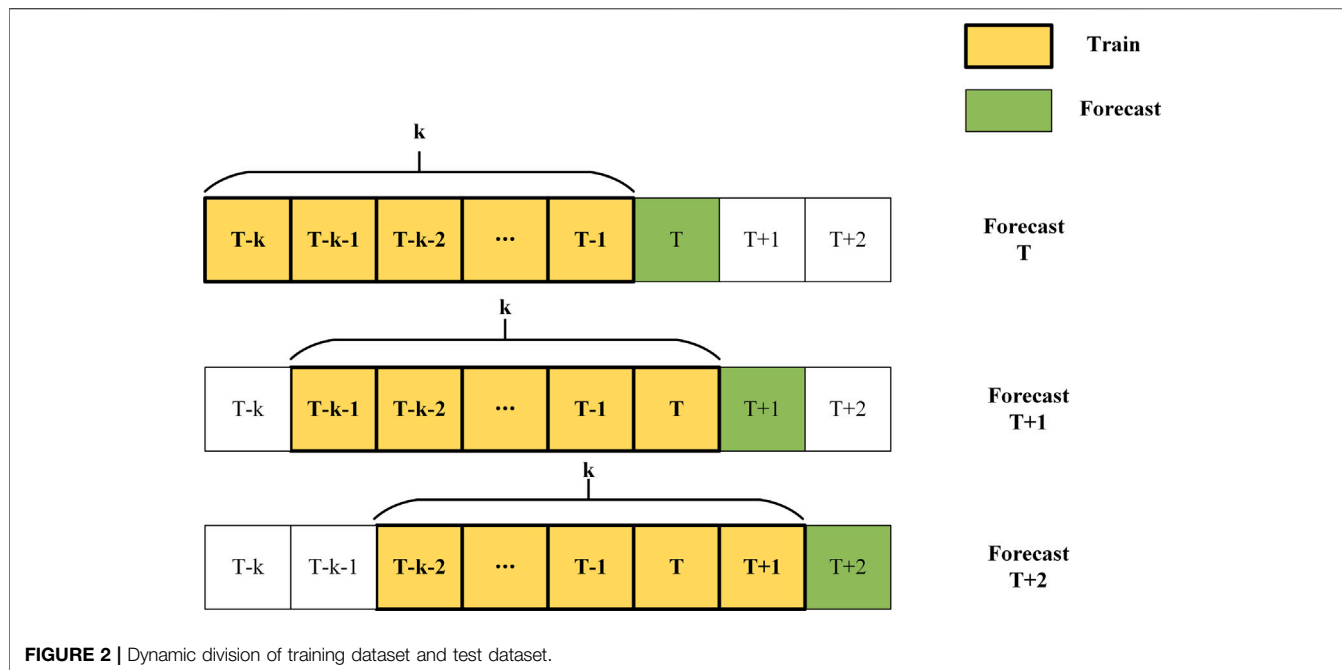
FIGURE 2 | Dynamic division of training dataset and test dataset.

TABLE 1 | Selection of steam coal feature data.

Types of the feature	Feature set
Coal production	Raw coal production of key state-owned coal mines, coal production of large coal enterprises, ..., and raw coal production of non–state-owned coal mines
Coal supply	National coal import quantity, coal inventory and dispatching data of Qinhuangdao Port, cargo ship ratio of four ports around Bohai Sea, and historical data of total coal storage in five ports around Bohai Sea
Coal transportation	Coal transportation quantity of Daqin Line, coal sales volume of national key coal mines, ..., and daily average number of coal railway loading vehicles
Coal consumption	National industrial power consumption, national social electricity consumption, power generation in coastal provinces, and coal consumption in power grid
Macroeconomic	Coal future price, added value of national secondary industry, GDP, consumer price index, the producer price index, and investment in fixed assets of the whole society

TABLE 2 | Parameters setting of the feature selection process.

Types of parameter	r_{mp}	r_{sp}	$delay_{max}$
Parameter value	0.5	0.5	5

TABLE 3 | Linear features with optimal delay.

Features	Optimal delay order (delay)	Pearson correlation coefficient (r_p)
Quantity of anchored vessels in Caofeidian Port	1	0.589,706,909
Volume of ships anchored at Caofeidian Port Phase II	1	0.594,892,547
Total coal stock in ports around the Bohai Sea	1	0.552,136,323
Power coal future closing price	1	0.895,980,863
. . . .	. . . .	. . . .
Total coal stock in Fangcheng Port	4	0.575,740,516

TABLE 4 | Nonlinear features with optimal delay.

Features	Optimal delay order (delay)	Pearson correlation coefficient (r_s)
Ship ratio of four ports around the Bohai Sea	3	0.74,183,513
Quantity of anchored vessels in Qinhuangdao Port	2	0.743,160,533
Quantity of anchored vessels in CIT Jingtang Port	2	0.552,341,897
Total coal stock in mainstream ports	5	0.625,224,659
………	………	………
Total coal stock at coastal ports	5	0.589,975,915

TABLE 5 | Value of hyper-parameters for five different models.

Forecasting model	Value of hyper-parameters for forecasting model
Lasso	$\alpha_{l1} = 0.1$
Ridge	$\alpha_{l2} = 0.2$
Huber–Ridge	$M = 1.35; \alpha_{l2} = 0.2$
PK–Huber–Ridge	$d = 2; M = 1.35; \alpha_{l2} = 0.2$
PK–Huber–Ridge–EMD–ARIMA	$d_k = 2; M = 1.35; \alpha_{l2} = 0.2; p \in [1, 2, 3, 4];$ $d_a \in [1, 2, 3]; q \in [1, 2, 3, 4]$

TABLE 6 | Evaluation index value of forecasting results of five forecasting models.

Dataset	Forecasting model	MAE	RMSE	MAPE(%)
Dataset 1	Lasso	36.0138	53.6206	6.6257
	Ridge	28.6532	44.1649	5.1159
	Huber–Ridge	30.7005	50.0595	5.6320
	PK–Huber–Ridge	26.8503	40.3592	4.8223
	PK–Huber–Ridge–EMD–ARIMA	**19.2267**	**26.0293**	**3.4813**
Dataset 2	Lasso	38.6772	60.7722	6.2920
	Ridge	29.8993	47.9723	4.5484
	Huber–Ridge	32.3476	55.5871	5.1023
	PK–Huber–Ridge	30.4047	45.9770	4.6723
	PK–Huber–Ridge–EMD–ARIMA	**18.9126**	**26.3342**	**2.9432**
Dataset 3	Lasso	37.6168	63.9781	5.3979
	Ridge	34.8572	57.777	4.7187
	Huber–Ridge	34.0952	61.9462	4.8548
	PK–Huber–Ridge	33.828	56.22	4.7179
	PK–Huber–Ridge–EMD–ARIMA	**22.9183**	**37.6673**	**3.1237**

The data preprocessing step is shown as follows:

1. Unify the data sampling frequency: The data input to the model have the features of one-to-one relationship between coal feature data and coal price, that is, the sampling frequency of coal feature data and coal price data is the same. When sampling frequency of the original coal price data and related feature data is inconsistent, such original data shall be operated at the unified data sampling frequency; the frequency of the data higher than the specified sampling frequency shall be reduced and the frequency of the data lower than the specified sampling frequency shall be raised. The daily frequency data are reduced to weekly frequency data. The quarterly and monthly data are raised to the weekly data and the missing value arises after the low-frequency data are

raised to the high-frequency data. The raised data are processed by ascending order as per the date, and then the missing value is filled up by linear difference filling.

2. Fill up the missing value: There are some missing values and non-numerical parts in the original coal data which need to be filled up to better utilize the dataset. The missing part in the data is filled up by linear difference, and the non-numerical part is deleted and then the missing part deleted is filled up by linear difference. The equation of missing value between filling points (x_0, y_0) and (x_k, y_k) is shown in **Eq. 23**:
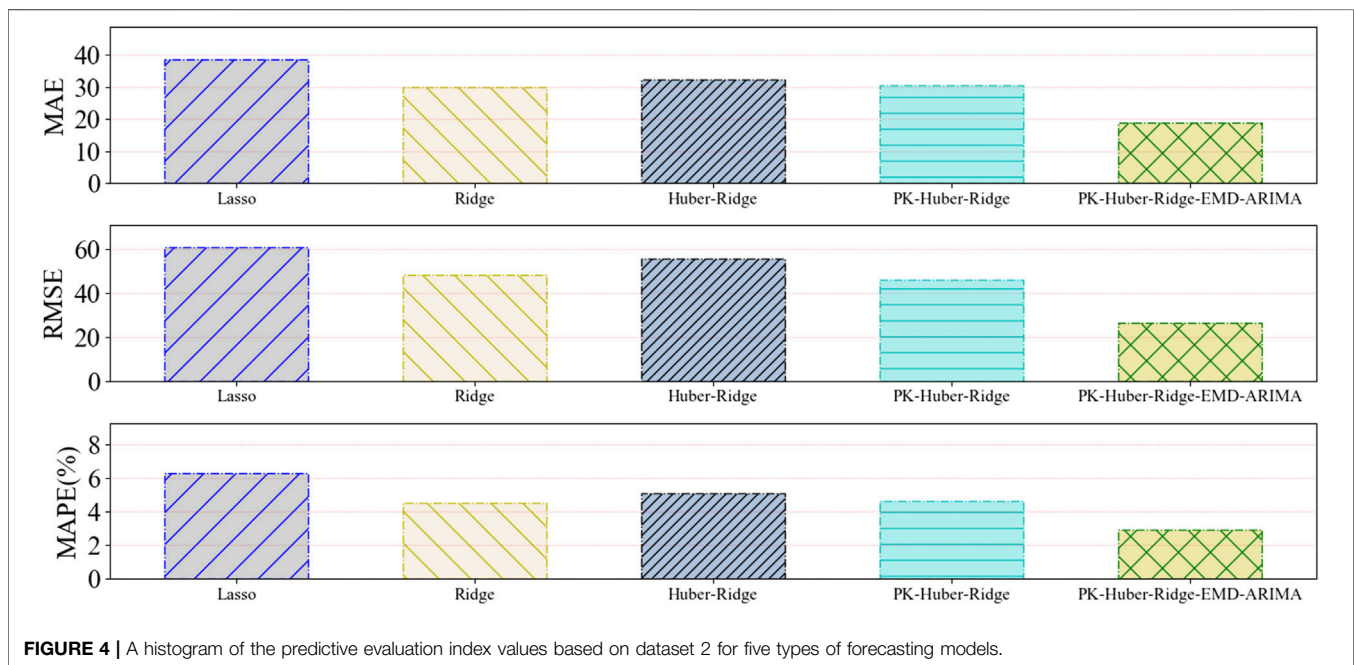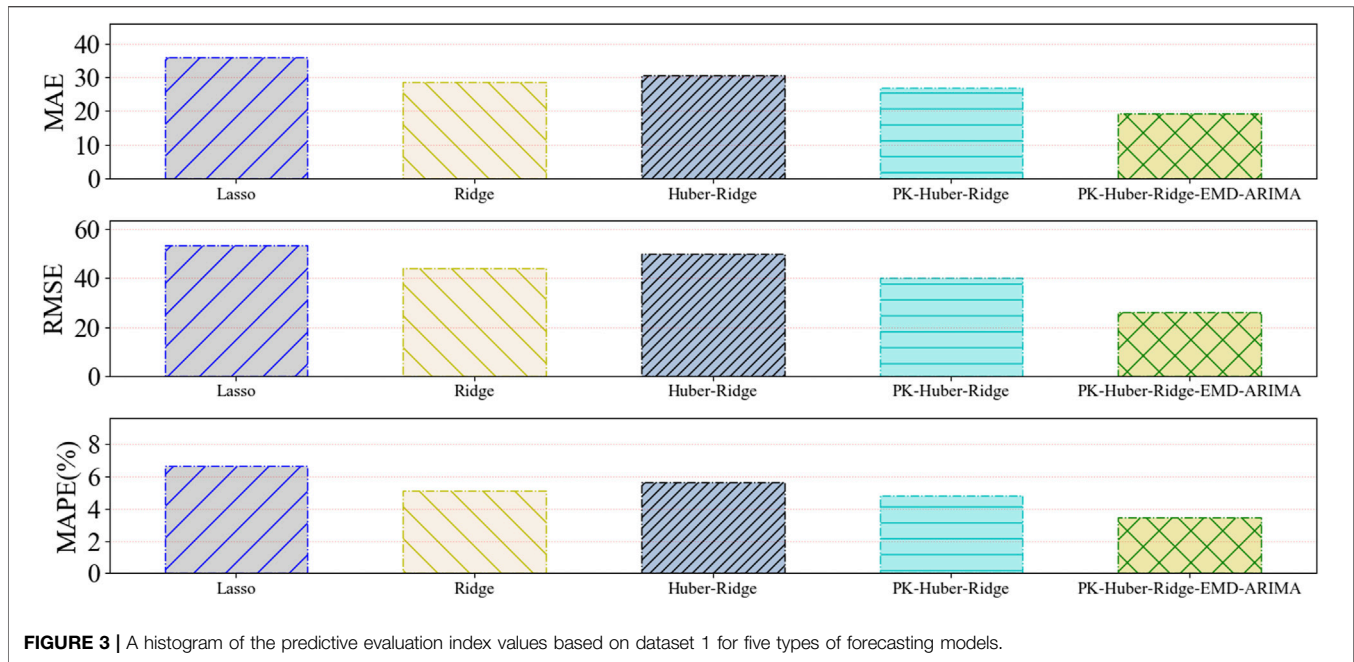
$$\phi(x) = \frac{x - x_k}{x_0 - x_k} y_0 + \frac{x - x_0}{x_k - x_0} y_k. \quad (23)$$

3. Standardize the dataset: There is a dimensional difference between different types of data. To avoid the dimensional error and lower the model performance, the standardized equation is used to transform the data distribution into standard distribution with the mean value of 0 and variance of 1. The standardized equation is shown as follows:

$$x_{ki} = \frac{X_{ki} - \overline{X_i}}{\sigma_i} \sigma_i = \left(X_{1i} - \overline{X_i} \right)^2 + \frac{\left(X_{2i} - \overline{X_i} \right)^2 + \ldots + \left(X_{ni} - \overline{X_i} \right)^2}{n}. \quad (24)$$

After transformation as per **Eq. 24**, the distribution of original feature data is transformed into a standard normal distribution with the mean value of 0 and variance of 1. x_{ki} indicates the kth numerical value of the ith feature index. $\overline{X_i}$ indicates the mean value of the ith feature index data, σ_i indicates the standard deviation of the ith feature index, and n indicates the sample size of the ith feature index.

4) Divide training dataset and test dataset: The training dataset and test dataset are not fixed, but they change dynamically. In the energy market, the influencing factors of energy indicators change with time (Liang et al., 2019). The coal data and related feature data at the time point within the sliding window of k width are used as the training dataset for

FIGURE 3 | A histogram of the predictive evaluation index values based on dataset 1 for five types of forecasting models.

FIGURE 4 | A histogram of the predictive evaluation index values based on dataset 2 for five types of forecasting models.
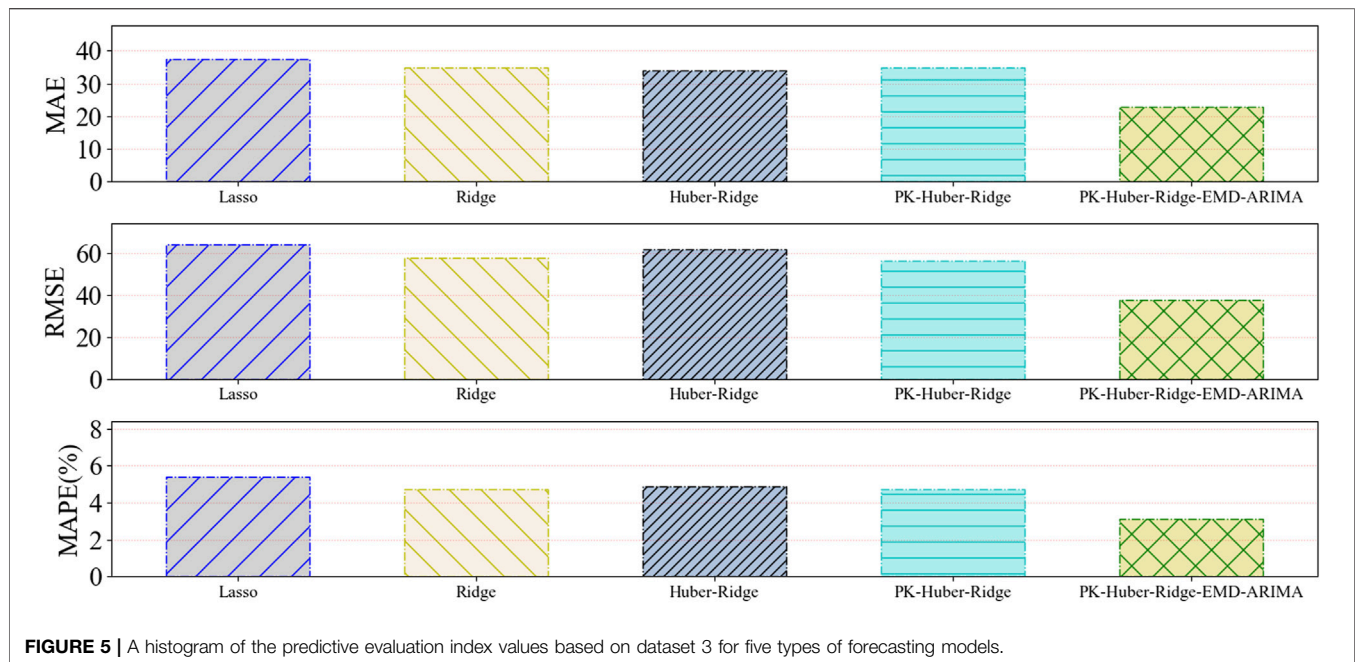
training the price forecasting model. The price forecasting model outputs the forecasted coal price at a time point according to the coal-related feature data after sliding the window. The corresponding original price data and relevant feature data are used as the test dataset to verify the model forecasting performance. The dynamic division of the training dataset and test dataset progresses over time, as shown in **Figure 2**.

Feature Selection

Selecting comprehensive and relevant features can greatly improve the performance of the forecasting model. All feature data are presented as a data matrix, and the optimal feature variable is chosen according to the feature type and feature optimal time interval.

Feature type: The coal price pertains to many factors; there are many factors influencing coal market price, and the main

FIGURE 5 | A histogram of the predictive evaluation index values based on dataset 3 for five types of forecasting models.

influence factors cover coal supply coal consumption, coal transportation, and economic factor. The feature indexes chosen are shown in **Table 1**.

There are many initial feature indexes, so the feature screening needs to be performed. The feature variable at the same time point as the forecasted target variable does not necessarily have the highest correlation, and it is necessary to find out the optimal time interval, that is, optimal delay order, for each feature variable. The optimal delay linear feature and the optimal delay nonlinear relevant features are screened as per Pearson correlation coefficient and Spearman correlation coefficient.

The value range of Pearson correlation coefficient r_p Spearman and correlation coefficient r_s is [-1, 1]. The closer the absolute value of r_p and r_s is to 1, the stronger the correlation will be; the closer the absolute value of r_p and r_s is to 0, the weaker the correlation will be.

Given the corresponding threshold values of the Pearson correlation coefficient and Spearman correlation coefficient are r_{mp} and r_{sp}, respectively, the feature indexes whose correlation coefficient exceeds the threshold value r_{mp} and r_{sp} are screened. Given the maximum delay order of the feature is $delay_{max}$, the parameter setting is shown in **Table 2**.

The chosen feature variables and steam coal price are input to the forecasting model mentioned, and the model outputs the steam coal price at the next time point. **Table 3** and **Table 4** show the selection of the optimal delay feature variable when forecasting the steam coal price on Jul. 6, 2021. The feature variable selected as per this method changes over time.

Experiment Result

In this article, Lasso, Ridge, Huber–Ridge, PK–Huber–Ridge, and PK–Huber–Ridge–EMD–ARIMA models are used for comparison. One-step forecasting is used for empirical test.

Qinhuangdao thermal coal data and feature data at the first 120 time points are used as the data variables of the forecasting model, and the forecasting model outputs the thermal coal price data at the 121st time point.

The set values of hyperparameters of Lasso, Ridge, Huber–Ridge, PK–Huber–Ridge, and PK–Huber–Ridge–EMD–ARIMA models are shown in **Table 5**. Here, α_{l1} is the coefficient of L1 regular term; α_{l2} is the coefficient of L2 regular term; M is the threshold of Huber loss function; d_k is the kernel parameter of the polynomial kernel and represents the order of the polynomial; p is the autoregressive order of ARIMA model; d_a represents the difference order; and q represents the moving average order. BIC criterion (Burnham and Anderson, 2004) is used to select the optimal ARIMA model hyperparameters p, d_a, and q.

The forecasting model is used to forecast 4500-kilocalorie steam coal price data (Dataset 1), 5000-kilocalorie steam coal price data (Dataset 2), and 5500-kilocalorie steam coal price data (Dataset 3) of Qinhuangdao Port from March 17, 2020 to July 6, 2021. **Table 6** and **Figure 3** and **Figure 4** and **Figure 5** show the evaluation index results of the forecasting model.

Through the comparison of the experimental results of five thermal coal price forecasting models, the following conclusions can be obtained.

Compared with the single model, the proposed combination model has a better forecasting performance. In dataset 1, dataset 2, and dataset 3 experiments, the forecasting performance of PK–Huber–Ridge–EMD–ARIMA model is better than the Lasso model, Ridge model and Huber–Ridge model, and PK–Huber–Ridge model. The thermal coal price dataset is complex, and the forecasting performance of a single forecasting model is very limited. The combination model can better deal with complex datasets. PK–Huber–Ridge–EMD–ARIMA model adopts the method of decomposition integration and time series

forecasting model to compensate for the error of single model. We consider the residual error rule of model forecasting to complement the hidden rules that the original single model does not learn.

Compared with the ordinary model, the robust kernel function model has better performance. In dataset 1, dataset 2, and dataset 3 experiments, the forecasting performance of PK–Huber–Ridge–EMD–ARIMA model is better than the Lasso model, Ridge model, and Huber–Ridge model. Thermal coal dataset has nonlinear law. PK–Huber–Ridge–EMD–ARIMA model uses polynomial kernel function to map nonlinear features into high-dimensional space, so that the linear model can learn the nonlinear law in the original feature space, so as to further improve the forecasting performance of the forecast model.

RESULT AND DISCUSSION

For the nonlinearity law in the steam coal dataset and limitations of the single forecasting model, this article proposes the forecast method for the price of steam coal based on robust regularized kernel regression and empirical mode decomposition. The robust regularized kernel regression model learns the nonlinearity law in the original data by using the kernel function. This model selects the Huber loss function to enhance the robustness of the forecasting model. We select the L2 regular term to lower the risk of model overfitting. The combined model based on EMD and ARIMA is used for error compensation against the Huber–Ridge polynomial kernel model, further improving the forecasting performance of the forecasting model. Compared to Lasso, Ridge, Huber–Ridge, and PK–Huber–Ridge, the proposed forecasting model (PK–Huber–Ridge–EMD–ARIMA) has the minimum value of MAE, RMSE, and MAPE.

The influence factors of steam coal price are complex which are easily affected by national policies. How to quantify policy factors and input them into the forecasting model for model training and model forecasting is the next work.

DATA AVAILABILITY STATEMENT

The data analyzed in this study are subject to the following licenses/restrictions: The data used to support the findings of this study are currently under embargo, while the research findings are commercialized. Requests to access these datasets should be directed to tmz@csust.edu.cn.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

REFERENCES

Alameer, Z., Fathalla, A., Li, K., Ye, H., and Jianhua, Z. Multistep-ahead Forecasting of Coal Prices Using a Hybrid Deep Learning Model. *Resour. Pol.* 65, 2020.

Burnham, K. P., and Anderson, D. R. (2004). Multimodel Inference. *Sociological Methods Res.* 33 (2), 261–304. doi:10.1177/0049124104268644

Chai, J., Zhao, C., Hu, Y., and Zhang, Z. G. (2021). Structural Analysis and Forecast of Gold price Returns. *J. Manag. Sci. Eng.* 6 (2), 135–145. doi:10.1016/j.jmse.2021.02.011

Conejo, A. J., Plazas, M. A., Espinola, R., and Molina, A. B. (2005). Day-ahead Electricity price Forecasting Using the Wavelet Transform and ARIMA Models. *IEEE Trans. Power Syst.* 20 (2), 1035–1042. doi:10.1109/tpwrs.2005.846054

Cover, T. M. (1965). Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *IEEE Trans. Electron. Comput.* EC-14 (3), 326–334. doi:10.1109/pgec.1965.264137

Gilbert, K. (2005). An ARIMA Supply Chain Model. *Manag. Sci.* 51 (2), 305–310. doi:10.1287/mnsc.1040.0308

Gupta, D., Hazarika, B. B., and Berlin, M. (2020). Robust Regularized Extreme Learning Machine with Asymmetric Huber Loss Function. *Neural Comput. Applic* 32 (16), 12971–12998. doi:10.1007/s00521-020-04741-w

Huber, P. J. (1992). "Robust Estimation of a Location Parameter," in *Breakthroughs in Statistics: Methodology and Distribution*. Editors S. Kotz and N. L. Johnson (New York, NY: Springer New York), 492–518. doi:10.1007/978-1-4612-4380-9_35

Ye, J., He, L., and Jin, H. (2021). A Denoising Carbon price Forecasting Method Based on the Integration of Kernel Independent Component Analysis and Least Squares Support Vector Regression. *Neurocomputing* 434, 67–79.

Ji, L., Zou, Y., He, K., and Zhu, B. (2019). Carbon Futures price Forecasting Based with ARIMA-CNN-LSTM Model. *Proced. Comp. Sci.* 162, 33–38. doi:10.1016/j.procs.2019.11.254

Jianke Zhu, J., Hoi, S., and Lyu, M. R. T. (2008). Robust Regularized Kernel Regression. *IEEE Trans. Syst. Man. Cybern. B* 38 (6), 1639–1644. doi:10.1109/tsmcb.2008.927279

Karabiber, O. A., and Xydis, G. (2019). Electricity Price Forecasting in the Danish Day-Ahead Market Using the TBATS, ANN and ARIMA Methods. *Energies* 12 (5). doi:10.3390/en12050928

Li, Y., and Li, Z. (2019). Forecasting of Coal Demand in China Based on Support Vector Machine Optimized by the Improved Gravitational Search Algorithm. *Energies* 12 (12). doi:10.3390/en12122249

Liang, T., Chai, J., Zhang, Y. J., and Zhang, Z. G. (2019). Refined Analysis and Prediction of Natural Gas Consumption in China. *J. Manag. Sci. Eng.* 4 (2), 91–104. doi:10.1016/j.jmse.2019.07.001

Lu, H., Ma, X., Huang, K., and Azimi, M. Carbon Trading Volume and price Forecasting in China Using Multiple Machine Learning Models. *J. Clean. Prod.* 249, 2020.

Matyjaszek, M., Riesgo Fernández, P., Krzemień, A., Wodarski, K., and Fidalgo Valverde, G. (2019). Forecasting Coking Coal Prices by Means of ARIMA Models and Neural Networks, Considering the Transgenic Time Series Theory. *Resour. Pol.* 61, 283–292. doi:10.1016/j.resourpol.2019.02.017

Owen, A. B. (2006). A Robust Hybrid of Lasso and ridge Regression. *Contemp. Math.* 443, 59–72.

Qiao, W., Liu, W., and Liu, E. A Combination Model Based on Wavelet Transform for Predicting the Difference between Monthly Natural Gas Production and Consumption of U.S. *Energy* 235, 2021.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001).A Generalized Representer Theorem. In Computational Learning Theory. Berlin, Heidelberg, 416–426. doi:10.1007/3-540-44581-1_27

Vu, D. H., Muttaqi, K. M., Agalgaonkar, A. P., and Bouzerdoum, A. (2019). Short-Term Forecasting of Electricity Spot Prices Containing Random Spikes Using a Time-Varying Autoregressive Model Combined with Kernel Regression. *IEEE Trans. Ind. Inf.* 15 (9), 5378–5388. doi:10.1109/tii.2019.2911700

Wang, B., and Wang, J. Energy Futures and Spots Prices Forecasting by Hybrid SW-GRU with EMD and Error Evaluation. *Energ. Econ.*, 90, 2020.

Wang, J., Cao, J., Yuan, S., and Cheng, M. Short-term Forecasting of Natural Gas Prices by Using a Novel Hybrid Method Based on a Combination of the CEEMDAN-SE-And the PSO-ALS-Optimized GRU Network. *Energy* 233, 121082.

Wang, J., Lei, C., and Guo, M. Daily Natural Gas price Forecasting by a Weighted Hybrid Data-Driven Model. *J. Pet. Sci. Eng.* 192, 2020.

Wang, J., Zhou, H., Hong, T., Li, X., and Wang, S. A Multi-Granularity Heterogeneous Combination Approach to Crude Oil price Forecasting. *Energ. Econ.* 91, 2020.

Wang, K., and Du, F. (2020). Coal-gas Compound Dynamic Disasters in China: A Review. *Process Saf. Environ. Prot.* 133, 1–17. doi:10.1016/j.psep.2019.10.006

Wu, J., Chen, Y., Zhou, T., and Li, T. (2019). An Adaptive Hybrid Learning Paradigm Integrating CEEMD, ARIMA and SBL for Crude Oil Price Forecasting. *Energies* 12 (7). doi:10.3390/en12071239

Xia, C., and Wang, Z. Drivers Analysis and Empirical Mode Decomposition Based Forecasting of Energy Consumption Structure. *J. Clean. Prod.* 254, 2020.

Xiong, J., and Xu, D. (2021). Relationship between Energy Consumption, Economic Growth and Environmental Pollution in China. *Environ. Res.* 194.

Xu, W., Hu, H., and Yang, W. (2019). Energy Time Series Forecasting Based on Empirical Mode Decomposition and FRBF-AR Model. *IEEE Access* 7, 36540–36548. doi:10.1109/access.2019.2902510

Yang, S., Chen, D., Li, S., and Wang, W. (2020). Carbon price Forecasting Based on Modified Ensemble Empirical Mode Decomposition and Long Short-Term Memory Optimized by Improved Whale Optimization Algorithm. *Sci. Total Environ.* 716, 137117. doi:10.1016/j.scitotenv.2020.137117

Yu, L., Wang, S., and Lai, K. K. (2008). Forecasting Crude Oil price with an EMD-Based Neural Network Ensemble Learning Paradigm. *Energ. Econ.* 30 (5), 2623–2635. doi:10.1016/j.eneco.2008.05.003

Zhang, H., Yang, Y., Zhang, Y., He, Z., Yuan, W., Yang, Y., et al. (2021). A Combined Model Based on SSA, Neural Networks, and LSSVM for Short-Term Electric Load and price Forecasting. *Neural Comput. Applic* 33 (2), 773–788. doi:10.1007/s00521-020-05113-0

Zhang, X., Zhang, C., and Wei, Z. (2019). Carbon Price Forecasting Based on Multi-Resolution Singular Value Decomposition and Extreme Learning Machine Optimized by the Moth–Flame Optimization Algorithm Considering Energy and Economic Factors. *Energies* 12 (22). doi:10.3390/en12224283

Zhou, J., Huo, X., Xu, X., and Li, Y. (2019). Forecasting the Carbon Price Using Extreme-Point Symmetric Mode Decomposition and Extreme Learning Machine Optimized by the Grey Wolf Optimizer Algorithm. *Energies* 12 (5). doi:10.3390/en12050950