EAI Endorsed Transactions
on Energy Web
Research Article EAI.EU

Key Technologies for a Regional Economic Development Index Based on Electricity Big Data

Wei Li[1,*], Wen Zhao[1], Xuefeng Bai[1], Xiaoming Cheng[1], Yongzhi Li[1], Jie Li[1]

[1]State Grid Shanxi Electric Power Company Lvliang Power Supply Company, Lvliang, China, 033000

Abstract

Electricity big data acts as a vital barometer of regional economic performance. In the transformation process of resource-based cities, traditional statistical indicators such as GDP and industrial added value suffer from obvious lagging effects and can hardly capture economic fluctuations promptly. In contrast, electricity big data, characterized by high-frequency and full-coverage attributes, can serve as an important support for the dynamic monitoring of the regional economy. Taking Lvliang City, a resource-based city, as the research object, this paper integrates multi-source time-series data covering power consumption, macro-monetary indicators, industrial product prices, and key-industry statistics. It constructs a candidate model library composed of regularized linear models and ensemble-learning algorithms, and establishes a closed-loop forecasting workflow of "data selection-correlation analysis-model evaluation-dynamic correction" to realize adaptive matching of "one-model-per-industry" for different industries. Prediction modeling is performed for Lvliang's macro-economy and its key industries, including coal, liquor, aluminum smelting, and iron and steel. Empirical results demonstrate that the model can achieve advanced prediction of economic trends for 3-6 months. The coefficients of determination (R^2) of the optimal Lasso regression models for the macro-economy, coal, liquor, and aluminum industries are 0.771, 0.764, 0.859, and 0.560 respectively, and the (R^2) value of the Elastic Net model for the iron-steel industry is 0.694. The average monthly prediction error for each industry stays within the 3-percentage-point limit, though the coal and aluminum smelting industries approach this error boundary, reflecting heterogeneous prediction performance across sectors. The model can characterize the differentiated-recovery features of various industries. This research can provide support for governments to implement anticipatory regulation and mitigate transformation-related risks, offer quantitative references for enterprise operation and optimal allocation of regional power resources, and facilitate the coordinated high-quality development of economy-energy-environment in Lyuliang.

Keywords: electricity big data, resource-based economy, economic forecasting, industrial transformation, high-quality development

Received on 29 August 2026, accepted on 19 October 2026, published on 19 November 2026

Copyright © 2026 Wei Li *et al.*, licensed to EAI. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi: 10.4108/ew.14804

1. Introduction

Electricity data are high-frequency, provide broad coverage, and are closely linked to economic activity. They can reflect regional production, business operations, and industrial activity in near real time. They also address the delay in conventional macroeconomic statistics. Economic monitoring and industrial activity forecasting based on electricity big data have therefore become important research topics at the intersection of energy economics and smart grids. Researchers in China and other countries have examined the economic value of electricity data, electricity–economy relationships, and quantitative forecasting methods. These studies provide theoretical and empirical support for regional economic monitoring based on electricity big data.

First, previous studies have examined the relationship between electricity data and economic activity. The main objective is to determine whether electricity indicators can serve as proxies for industrial activity and to clarify how

electricity consumption and load are coupled with macroeconomic and industrial indicators. Bale et al. [1] argued that industrial sensor data continuously record production activity, which supports the use of electricity data to represent industrial production. Yi et al. [2] showed that power systems are closely coupled with macroeconomic development. They also found that multidimensional external variables improve the completeness of economic assessment. Xu et al. [3] developed a quantitative framework linking electricity-use features to industrial value added. Liu et al. [4] used firm-level electricity data to show that micro-level electricity indicators can reflect industry-wide activity. Haggag et al. [5] found that changes in electricity time-series can represent medium- and long-term industrial trends. Wang et al. [6] combined firm-level electricity data to construct an industrial activity index. Their study directly demonstrated the feasibility of measuring industry conditions with electricity big data.

Second, researchers have studied the practical use of electricity big data in economic monitoring. Institutions and scholars have proposed several approaches for applying electricity data to macroeconomic and industrial monitoring. In 2018, the International Energy Agency (IEA) released a report that developed an electricity–economy linkage index from the relationship between electricity consumption and industrial output. The index enabled real-time tracking of industrial conditions in major economies. After 2020, the U.S. Federal Reserve incorporated electricity load into its economic recovery monitoring system. It used changes in manufacturing electricity consumption to assess supply-chain recovery. This practice demonstrated the value of electricity data for anticipating macroeconomic trends. Wang et al. [7] extended this application to electricity-market decisions. They used electricity big data to optimize trading decisions and assess market supply and demand in real time.

In 2021, State Grid Corporation of China launched the “Electricity-Based Economic Monitoring” pilot program. In industrial provinces such as Jiangsu and Zhejiang, the program combined growth in electricity consumption, the Purchasing Managers’ Index (PMI), and the Producer Price Index (PPI) to create a monthly electricity-based economic activity index. The index was used to anticipate trends in the value added of industrial enterprises above the designated size and produced a replicable provincial monitoring model. Most domestic empirical studies, however, focus on highly industrialized provinces. Targeted monitoring methods for resource-based prefecture-level cities remain scarce.

Finally, studies have investigated forecasting methods for electricity time-series data. Current methods fall into two broad groups: regularized statistical models and machine-learning or ensemble-learning algorithms. Many studies compare their application conditions, strengths, and limitations. Regularized models are interpretable and stable, so they are widely used as baselines in electricity–economy analysis. Raj et al. [8] and Bayu et al. [9] used linear regression to capture basic patterns in electricity indicators. Nieveen et al. [10] showed that ridge regression is robust for high-dimensional variable screening, multicollinearity control, and overfitting prevention. Zhang et al. [11] highlighted the ability of Lasso regression to select features and remove redundant dimensions in high-dimensional electricity data. Sadou et al. [12] extended the application range of conventional regression through multivariate non-linear regression. Linear regression, ridge regression, Lasso regression, and elastic net regression form a complete set of regularized baseline models. However, most studies select only one model or use these methods separately. Few studies compare all four models within one systematic benchmark. Conventional statistical models also have limited nonlinear fitting capacity. Their forecasting accuracy can therefore be constrained when electricity time-series show complex fluctuations.

To address these limitations, hybrid time-series models based on ensemble-learning and deep learning have become a major research area. Ullah et al. [13] combined ensemble machine learning with SHAP to improve forecast accuracy while reducing the black-box problem. Chauhan et al. [14] improved ensemble frameworks with intelligent optimization algorithms. Their approach reduced local-optimum problems and high computational costs. Kim et al. [15] developed an LSTM–XGBoost ensemble time-series model. Sim et al. [16] combined clustering and ensemble -learning to distinguish different electricity-use patterns. Melon et al. [17] proposed a dynamic evaluation framework that integrated time-series methods, ensemble-learning, and artificial intelligence.

To sum up, existing studies mostly focus on macro-scales at the provincial level and above, and there is a lack of power-economy forecasting solutions tailored for prefecture-level resource-based cities such as Lyuliang. Meanwhile, they generally suffer from shortcomings including insufficient data integration, single-model adoption, and poor industry-specific pertinence. Resource-dependent cities face prominent challenges in industrial restructuring and economic fluctuation monitoring under the dual pressure of resource depletion and low-carbon transition. A stream of existing studies has investigated the economic development, industrial transformation, and policy-driven transition of resource-based cities in China, revealing the heterogeneous impacts of resource reliance on urban economic growth and the role of industrial diversification [18-20]. Meanwhile, recent deep-learning advances represented by Temporal Fusion Transformers (TFT), and Informer have promoted the development of interpretable, multi-horizon, and long-sequence time-series forecasting, providing powerful tools for economic early warning with multi-source data [21-23]. Nevertheless, most of the above deep learning methods are data-hungry and lack good interpretability in practical economic scenarios. Few tailored forecasting frameworks have been designed for prefecture-level resource-based cities undergoing industrial transformation, especially using electricity big data to track industry cycle turning points. Accordingly, supported primarily by electricity big data from Lyuliang City, this paper develops a dynamic prosperity forecasting tool adapted to local multi-pillar industries to address the delay in economic judgment caused by the lag of traditional statistical indicators such as GDP and industrial added value. The main innovations and incremental contributions of this study are

summarized as follows: (1) Previous regional economic forecasting studies based on electricity big-data mostly adopt fixed single-model frameworks and lack closed-loop adaptive update mechanisms, especially for prefecture-level resource-based cities. To fill this gap, this paper constructs a closed-loop forecasting workflow of "data selection- correlation analysis- model evaluation- dynamic correction". Combined with chronological rolling cross-validation, dynamic factor iteration is introduced for parameter updating, which realizes 3- 6-month advance prediction of the regional economy and achieves the "one-model-per-industry" adaptive matching mechanism for different industries. (2) Existing empirical works mostly focus on aggregate macroeconomic forecasting, and seldom carry out separate targeted modeling for multiple pillar industries of resource-based cities, failing to reveal industry-specific coupling patterns between electricity indicators and economic prosperity. Different from previous aggregate-level analysis, this paper conducts separate modeling for four local pillar industries, including coal, baijiu, aluminum smelting, and steel. It quantitatively reveals the heterogeneous coupling relationships between electricity big-data features and industrial prosperity. The average monthly prediction error is kept below 3 percentage points, which enables the model to capture the differentiated-recovery characteristics of various industries. (3) Many existing electricity-economy forecasting studies prioritize prediction accuracy but neglect model interpretability for policy-oriented decision-making. This work balances forecasting accuracy and the economic interpretability of regression-based models. Beyond simple numerical prediction, it outputs industrial prosperity inflection-point signals together with quantitative forecasting results. It provides quantitative decision-making evidence for local government anticipatory regulation, enterprise operation reference, and regional power-resource optimal allocation for resource-based cities.

2. Model construction and experimental design

2.1. Candidate Models: Principles, Strengths, and Limitations

Electricity big data and regional economic indicators show both linear relationships and complex nonlinear fluctuations. A single model may therefore fail to provide both a good fit and strong generalization. This study develops a parallel comparison and selection framework that includes linear regularized models and ensemble-learning models. The framework supports interpretable analysis of linear relationships and flexible fitting of nonlinear patterns. Seven widely used algorithms are included in two groups. The first group consists of four conventional regularized statistical models: linear regression, ridge regression, Lasso regression, and elastic net regression. These models can identify interpretable linear relationships between electricity consumption and economic indicators. They can also select important electricity features and quantify variable effects. The second group includes random forest, gradient boosting, and XGBoost. These ensemble methods use multiple decision trees to capture nonlinear interactions and abrupt changes. They are therefore suitable for stage-specific changes in industrial activity in resource-based cities.

Regularized Linear Models Group

(i) Linear regression

$$\hat{y} = X\beta + \beta_0 \quad (1)$$

Where, $X \in R^{n\times p}$ denotes the electricity-consumption-related and derived feature matrix; β denotes the feature coefficient vector; β_0 is the intercept. The loss function for least-squares is $L_{lr} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Linear regression [24] constructs a linear mapping of $y = X\beta + \beta_0$, and adopts least squares as the loss function to minimize the sum of squared residuals between observed values and model-predicted values, to fit and predict the dependent variable. Ordinary least squares yields feature coefficients without regularization constraints. Nevertheless, when the feature dimension is high and multicollinearity exists among variables, problems such as overfitting and unstable coefficient estimation tend to occur. Accordingly, regularized linear models including Ridge and Lasso have been developed. By adding penalty terms to constrain the parameter space, these models realize parameter shrinkage and variable selection while maintaining prediction performance, and have been widely applied in empirical modeling scenarios such as electricity consumption forecasting.

(ii) Ridge Regression

$$L_{Ridge} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha\|\beta\|_2^2, \alpha > 0 \quad (2)$$

Where, α is the hyperparameter for regularization strength.

Ridge regression [25] is an L2 regularized shrinkage method. It introduces a squared L2 norm penalty term for the coefficient vector based on the sum of squared residuals loss. The regularization strength is controlled by a hyperparameter to continuously shrink regression coefficients, thereby alleviating the high variance problem of coefficient estimation caused by multicollinearity. Ridge regression can improve the prediction stability of the model, yet it cannot shrink feature coefficients to zero and thus lacks the function of automatic variable selection.

(iii) Lasso Regression

$$L_{Lasso} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha\|\beta\|_1, \alpha > 0 \quad (3)$$

Where, α is the hyperparameter for regularization strength.

Lasso regression [25] is an L1-regularized model, which introduces an L1 norm penalty term for the coefficient vector

based on the sum of squared residuals from least squares. Unlike ridge regression, the L1 penalty can compress some feature coefficients to zero, enabling automatic feature selection while mitigating overfitting and yielding a sparse regression model. Nevertheless, it suffers from unstable selection issues for highly correlated features.

(iv) Elastic Net

$$L_{EN} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha(\rho\|\beta\|_1 + (1-\rho)\|\beta\|_2^2), \alpha > 0 \quad (4)$$

Where, $\beta \in [0,1]$ denotes the mixing proportion of L1-L2, and α is the regularization strength, which simultaneously achieves feature selection and multicollinearity suppression.

Elastic Net [25] is a hybrid regularization model combining L1 and L2 penalties. Its loss function introduces both L1 norm and L2 norm penalty terms based on the sum of squared residuals, and balances the contributions of the two regularizers via the mixing parameter. This model combines two strengths: ridge regression for multicollinearity control and Lasso regression for automatic feature selection. It mitigates Lasso's unstable-selection issue when handling high-dimensional, highly-correlated features. However, users must optimize two sets of hyperparameters at the same time, raising the computational cost of model tuning.

Ensemble-Learning-Model Group

(i) Random-Forest

Random Forests [25] is an ensemble-learning algorithm based on Bagging bootstrap sampling. It trains multiple independent regression trees. $\{h(x;\theta_m)\}_{m=1}^{M}$ through random sampling of samples and random feature sampling at nodes. For regression tasks, the arithmetic mean of the prediction outputs from all individual trees is taken as the final output.

$$\hat{y}_{RF} = \frac{1}{M}\sum_{m=1}^{M}\{h(x;\theta_m)\} \quad (5)$$

Where $\{h(x;\theta_m)\}_{m=1}^{M}$ denotes the sample subset and feature subset obtained by random sampling for the m-th tree.

(ii) Gradient Boosting Decision Tree

Gradient Boosting Decision Tree (GBDT) [25] is an additive ensemble model. It is trained in a serial-iterative manner. In each iteration, a weak regression tree is trained to fit the negative gradient (pseudo-residual) of the loss function of the preceding model. A learning rate is introduced to shrink the output of each tree, and the overall model is updated by cumulative summation step-by-step. Compared with Random Forests, GBDT focuses on bias reduction and exhibits strong fitting capacity for nonlinear relationships. Nevertheless, its serial training scheme is sensitive to noisy samples and more prone to overfitting.

GBDT adopts serial-iterative training, where the t-th tree fits the residuals of the preceding model, with the negative gradient of the loss function as the fitting target

$$F_t(x) = F_{t-1}(x) + \eta \cdot h_t(x) \quad (6)$$

where $\eta \in (0,1]$ denotes the learning rate, and $h_t(x)$ represents the weak regression tree at the t-th iteration.

(iii) Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) [20] is an engineered improvement of GBDT. It employs second-order Taylor expansion to solve the loss function and incorporates a tree-complexity regularization term into the objective function. The total objective function reads

$$L_{XGB} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T}(\gamma T_t + \frac{1}{2}\lambda\|\omega_t\|^2) \quad (7)$$

where $l(\cdot)$ is the sample-wise loss, T_t the number of leaf nodes of the t-th tree, ω_t the leaf-node weight, γ the leaf-split penalty, and λ the weight. It supports automatic missing value processing, feature binning, pre-sorting, and parallel computing.

All models are trained on the same samples and evaluated under a common standard. This design uses the different strengths of each algorithm. Linear models are used to identify stable relationships between electricity indicators and economic trends. They meet the interpretability needs of macroeconomic monitoring and policy analysis. Ensemble-learning models are used to improve forecasts during short-term fluctuations and turning points. They address the limited ability of linear models to describe nonlinear shocks. After training, K-fold cross-validation is used to calculate error metrics for each model. Model fit and generalization are compared across algorithms. The best model is then selected separately for the citywide economy and for the coal, liquor, aluminum smelting, and steel industries. The objective is to combine economic interpretation with accurate and dynamic forecasting. Table 1 summarizes the core concepts, objectives, strengths, and limitations of the models. It is worth noting that such differentiated model selection originates from distinct data characteristics of different industry datasets. The steel-industry dataset integrates multiple interrelated monetary indicators and electricity-consumption-related variables, resulting in severe multicollinearity among input features. Faced with high multicollinearity, standalone Lasso regression may yield unstable feature-selection outcomes and risk discarding practically meaningful predictive variables. By contrast, Elastic Net combines L1-norm feature screening and L2-norm multicollinearity suppression, which alleviates the limitations of pure Lasso regression under highly correlated inputs and delivers superior performance for steel-industry forecasting. For the macroeconomy, coal, liquor, and aluminum-smelting datasets, the multicollinearity degree among predictor variables is relatively mild. Therefore, Lasso regression can realize stable feature selection and achieve satisfactory forecasting accuracy. These dataset-level differences ultimately produce different optimal models for individual industries.

Table 1. Comparison of candidate forecasting models

Algorithm	Core concept	Objective	Strengths	Limitations
Linear Regression [25]	Fits a straight line or hyperplane to the data and assumes a linear relationship between predictors and the target.	Minimizes the sum of squared differences between predicted and observed values by ordinary least squares.	1. Simple to understand and implement 2. Fast computation 3. Strong interpretability	1. Sensitive to outliers 2. May overfit, especially when predictors outnumber samples 3. Cannot capture nonlinear relationships
Ridge Regression [27]	A regularized form of linear regression that adds an L2 penalty (α times the sum of squared coefficients) to the loss function.	Minimize squared error while penalizing large coefficients. Coefficients shrink toward zero but are usually not set exactly to zero.	1. Reduces overfitting 2. Handles multicollinearity 3. More robust than ordinary linear regression	Does not perform feature selection. All predictors are retained, although their coefficients become smaller.
Lasso Regression [24]	A regularized form of linear regression that adds an L1 penalty (α times the sum of absolute coefficient values) to the loss function.	Minimizes squared error while penalizing coefficients. Coefficients of unimportant predictors can shrink to zero.	1. Reduces overfitting 2. Performs feature selection 3. Automatically removes unimportant predictors	1. May select one predictor arbitrarily when predictors are highly correlated 2. Less stable than ridge regression
Elastic Net [28]	Combines ridge and Lasso regression. The loss function contains both L1 and L2 penalties.	Combines the advantages of L1 and L2 regularization to perform feature selection and manage multicollinearity.	1. Balances feature selection and multicollinearity control 2. More robust than Lasso or Ridge alone	1. Requires tuning of two hyperparameters: mixing ratio and regularization strength 2. More complex tuning process
Random Forest [29]	Builds multiple decision trees and averages their predictions for regression or uses voting for classification.	Reduces variance through sample and feature subsampling and improves model generalization.	1. High predictive accuracy 2. Strong performance with high-dimensional data 3. Resistant to overfitting 4. Provides feature-importance estimates	1. Less interpretable than a single decision tree or a linear model 2. May not be optimal for sparse data
Gradient Boosting [30]	Builds decision trees sequentially. Each new tree learns from the residuals of the preceding model.	Add models iteratively so that each new model corrects errors made by earlier models.	1. Often more accurate than random forest 2. Supports several loss functions	1. Sequential training can be slow 2. Sensitive to hyperparameters 3. Can overfit
XGBoost [26]	An efficient implementation of gradient boosting with several optimizations to gradient-boosted decision trees (GBDT).	Improves traditional GBDT to increase computational efficiency and model performance.	1. Strong predictive performance 2. Fast training 3. Supports custom loss functions 4. Handles missing values internally 5. Supports parallel computing	1. Many parameters 2. Complex tuning process

2.2 Model Selection and Evaluation Framework

Traditional single-model forecasting has limitations. To improve prediction accuracy, stability, and generalization for macroeconomic and key-industry trends, this study puts forward an adaptive model-selection method driven by intelligent algorithms. Rather than predefining a fixed forecasting model, the approach builds a multi-model candidate library with respect to sample size, fluctuation features, variable correlations, nonlinearity, and historical prediction errors of datasets from different industries. Intelligent optimization algorithms are utilized to automatically conduct model screening, parameter tuning, and performance assessment, achieving adaptive matching of optimal models for diverse forecasting targets.

Model Selection

The main criterion for model selection via intelligent optimization algorithms is determining the most appropriate model by comparing the Average Cross-Validation Score during model training. As a robust model-evaluation technique, cross-validation splits the original dataset into several non-overlapping subsets (folds). In each iteration, one subset acts as the validation set while the rest form the training set for score calculation. The average of all validation scores is taken as the final metric for model performance. Its merits lie in more reliable and stable performance assessment, better indication of generalization capability, higher data utilization, and facilitation of model selection and hyperparameter tuning.

The average cross-validation score is calculated as follows:

$$\mathrm{P} = \frac{1}{K}\sum_{i=1}^{K} S_i \tag{8}$$

Where P is the average cross-validation score, K is the number of folds (usually 5 or 10), and Si is the score from the ith validation run, such as accuracy or mean squared error.

To quantify forecast error, model fit, and explanatory power, this study uses four standard regression metrics: mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). Table 2 defines these metrics and their interpretation. In this study, MAE reflects the average monthly prediction deviation of year-on-year growth (YoY) industrial output-value growth. All the above metrics are calculated on year-on-year growth rates of industrial output. It should be noted that local prediction bias may occur at business-cycle turning points, including cyclical peaks and troughs, which represents the practical application boundary of the established models. For all industries in this study, the average monthly prediction error is controlled below 3 percentage points.

Table 2. Definitions and interpretation of model evaluation metrics

	Definition	Interpretation
MSE	The mean of the squared differences between observed and predicted values. Squaring gives greater weight to large errors, so MSE is sensitive to major forecast deviations.	A lower MSE indicates better model fit and forecast performance. MSE = 0 represents a perfect prediction with no error.
RMSE	The square root of MSE. It restores the unit of the original economic indicator and directly represents the typical forecast error.	A lower RMSE indicates a smaller overall forecast error and higher predictive accuracy.
MAE	The mean of the absolute differences between observed and predicted values. All errors receive equal weight, so the metric is less affected by extreme values.	A lower MAE indicates a smaller average deviation between predicted and observed values.
R^2	Measures how well predictors, such as electricity consumption and macroeconomic variables, explain year-over-year changes in industrial value added. Its range is $(-\infty, 1]$.	A value closer to 1 indicates stronger explanatory power. A value near 0 indicates weak explanatory power. A negative value means that the model performs worse than the sample mean.

Feature Construction

(i) Rolling Statistical Features (Window Size = 3)

Several rolling statistical features with a window size of 3 are constructed in this paper, including the 3-period rolling mean and 3-period rolling standard deviation, as summarized in Table 3.

Table 3. Rolling Statistical Features

Feature Name	Mathematical Formula	Corresponding Code Implementation
3-period rolling mean	$rolling_mean_3(t) = \frac{1}{3}\sum_{k=0}^{2} y(t-k)$	Rolling(window=3).mean（）
3-period rolling standard deviation	$rolling_std_3(t) = \sqrt{\frac{1}{2}\sum_{k=0}^{2}(y(t-k) - rolling_mean_3(t))^2}$	rolling（window=3）.std（）

(ii) Input Feature Splicing Rules

Construction logic for the feature vector of a single sample:

$$X_t = [y(t-10), y(t-9), \ldots, y(t-1), rolling_mean_3(t), rolling_std_3(t), f_1(t), f_2(t), \ldots, f_m(t) \quad (9)$$

where $y(t-10) \sim y(t-1)$ is the set of the latest ten historical target values, and $f_1(t) \sim f_m(t)$ stands for the newest non-date, non-target-variable features at time t.

(iii) Data Standardization

Feature-scaling formula:

$$X_{scaled} = \frac{X-\mu}{\sigma} \quad (10)$$

where μ is the feature mean of the training set, and σ is the feature standard deviation of the training set.

(iv) Objective Function of Linear Regression Model

Ridge Regression

$$\min_{\boldsymbol{w}} \left[\frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{X}_i \cdot \boldsymbol{w})^2 + \alpha \cdot |\boldsymbol{w}|_2^2\right] \quad (11)$$

$$\|w\|_2^2 = \sum_{j=1}^{p} w_j^2 \quad (12)$$

Lasso Regression

$$\min_{\boldsymbol{w}} \left[\frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{X}_i \cdot \boldsymbol{w})^2 + \alpha \cdot |\boldsymbol{w}_j|\right] \quad (13)$$

Elastic Net：

$$\min_{w} \frac{1}{2n}\sum_{i=1}^{n}(y_i - X_i \cdot w - b)^2 + \alpha \cdot \left(l_1^{ratio} \cdot \sum_{j=1}^{p} |w_j| + \frac{1-p_1^{ratio}}{2} \cdot \sum_{j=1}^{p} w_j^2\right) \quad (14)$$

Selection Procedure

Adaptive model selection based on intelligent algorithms mainly includes the following procedures:

Step 1: Data Preprocessing. Macroeconomic indicators, power data, industrial indicators, and external variables are processed through missing-value treatment, outlier detection, standardization, and time-frequency unification, enabling multi-source data to be utilized under a unified forecasting framework.

Step 2: Preliminary Feature Screening. The K-L information quantity method, time-difference correlation analysis, and variable correlation analysis are employed to filter indicators with strong relevance to the forecasting target and generate an initial feature set. Industry-specific feature libraries are built for each sector to avoid disturbances caused by irrelevant variables during model training.

Step 3: Candidate-Model Training. Filtered features are imported into the candidate model library for preliminary training of diverse models, and rolling-time-window cross-validation is used for performance assessment. Given the time-sequential characteristics of economic and industrial datasets, random shuffling is excluded from validation. Recursive validation following chronological order is implemented to make evaluation results more aligned with real-world application scenarios.

Step 4: Intelligent-Algorithm-Driven Parameter Tuning. With the comprehensive evaluation function as the optimization objective, intelligent optimization algorithms automatically search for optimal parameter combinations for

alternative models. Specifically, the regularization coefficient is tuned for ridge regression; kernel parameters and penalty coefficients for support vector regression; tree depth, learning rate, subsample ratio, and number of weak learners for Random Forests and gradient-boosting models; hidden-layer size, learning rate, and iteration epochs for neural networks.

Step 5: Comprehensive Scoring of Candidate Models. Each candidate model is comprehensively scored based on RMSE, MAE, MAPE, R^2, error volatility, and model complexity over the validation set, and the model with the lowest MSE is selected for forecasting. Even if a model presents high fitting accuracy, its priority will be downgraded in comprehensive assessment when it exhibits high validation-error volatility or excessive complexity.

Step 6: Output of the Optimal Model. The highest-scoring model is designated as the final model for the corresponding forecasting object, outputting future predictions, prediction error intervals, model evaluation metrics, and key influencing variables. Diverse optimal models can be generated for different industries, enabling adaptive forecasting following the principle of "one-model-per-industry, one scheme for one forecasting object".

2.3 End-to-End Modeling Workflow

The proposed forecasting workflow has four sequential stages: data preprocessing, feature engineering, model training and evaluation, and forecasting and interpretation (Figure 1).

Stage 1: Data preprocessing. Data import, format conversion, missing-value imputation, chronological sorting, outlier detection, and elimination are performed to improve the data quality of multi-source power-economy datasets.

Stage 2: Feature engineering. Rolling-window statistical features and time-series derived indicators are constructed, followed by feature standardization to generate final model inputs. Preliminary feature screening is conducted to retain variables strongly correlated with forecasting targets.

Stage 3: Model training and evaluation. Regularized linear models and ensemble-learning algorithms in the candidate library are trained in parallel. Considering the time-series nature of economic data, chronological rolling cross-validation rather than random-shuffled K-fold cross-validation is adopted for performance assessment. Multiple evaluation metrics are used for comprehensive scoring to realize "one-model-per-industry" adaptive matching for different forecasting objects.

Stage 4: Forecasting, dynamic correction and interpretation. The selected optimal model generates multi-step rolling forecasts. Dynamic factor iteration is implemented for parameter updating to realize closed-loop dynamic correction. Feature-importance analysis and residual analysis are applied to interpret prediction outputs, providing quantitative decision-making support. In this workflow, dynamic correction mainly relies on dynamic factor iteration to realize parameter updating of the established optimal model, rather than automatic feature updating or complete model reselection. When new monthly time-series observations are accumulated, regression coefficients and hyperparameters of the industry-specific optimal model will be iteratively refreshed, which enables the forecasting model to adapt to the time-varying coupling relationship between electricity-consumption features and industrial prosperity.

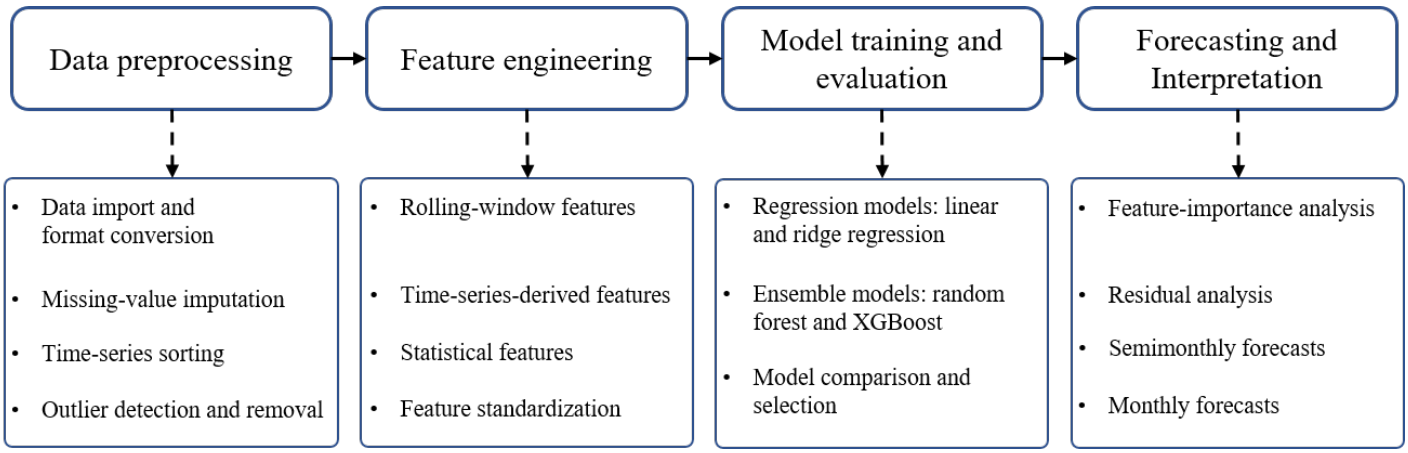


Figure 1. End-to-end technical workflow for regional economic forecasting based on electricity big data

Note that dynamic correction implements parameter iteration updating for trained models, excluding automatic feature updating and full model reselection.

3. Results and Analysis

3.1. Macroeconomic Forecasting for Lüliang City

Data for this analysis were obtained from the Lüliang Municipal Bureau of Statistics. Monthly gross domestic product (GDP) data were unavailable. Because Lüliang is an industrial city, the value added of industrial enterprises above a designated size was used as a proxy for GDP and as the prediction target. The predictor variables included total electricity consumption and its cumulative year-on-year (YoY) growth; the end-of-period balances and YoY growth rates of broad money (M2), narrow money (M1), and currency in circulation (M0); the Producer Price Index for Industrial Products (same month of the previous year = 100); and cumulative YoY growth in electricity consumption in the tertiary sector, tobacco products manufacturing, electrical machinery and equipment manufacturing, transportation, storage and postal services, real estate, wine, beverage and refined tea manufacturing, and cultural, educational, sports and recreational goods manufacturing.

Variable Correlation Analysis

From the correlation heatmap in Figure 2, the variables most strongly correlated with the cumulative year-on-year growth industrial output-value growth are ranked as follows: Producer Price Index (0.75), cumulative year-on-year growth electricity consumption for wine, beverage and refined-tea manufacturing (0.63), end-period M1 money supply (0.51), cumulative year-on-year growth electricity consumption (0.47), and year-on-year growth rate of M2 money supply (0.46). Notably, the PPI correlation coefficient exceeds 0.7, indicating a strong positive correlation. This result shows that fluctuations in industrial-product prices act as a core driver of industrial output value in Lüliang City. With a correlation coefficient of 0.63, liquor-sector electricity consumption acts as a secondary core variable, mirroring the driving effect of the local pillar liquor industry on macro-industrial prosperity. The remaining monetary and electricity-consumption variables show correlation coefficients mostly within 0.4–0.5, corresponding to moderate correlation, and collectively constitute multi-dimensional driving factors for macroeconomic performance.

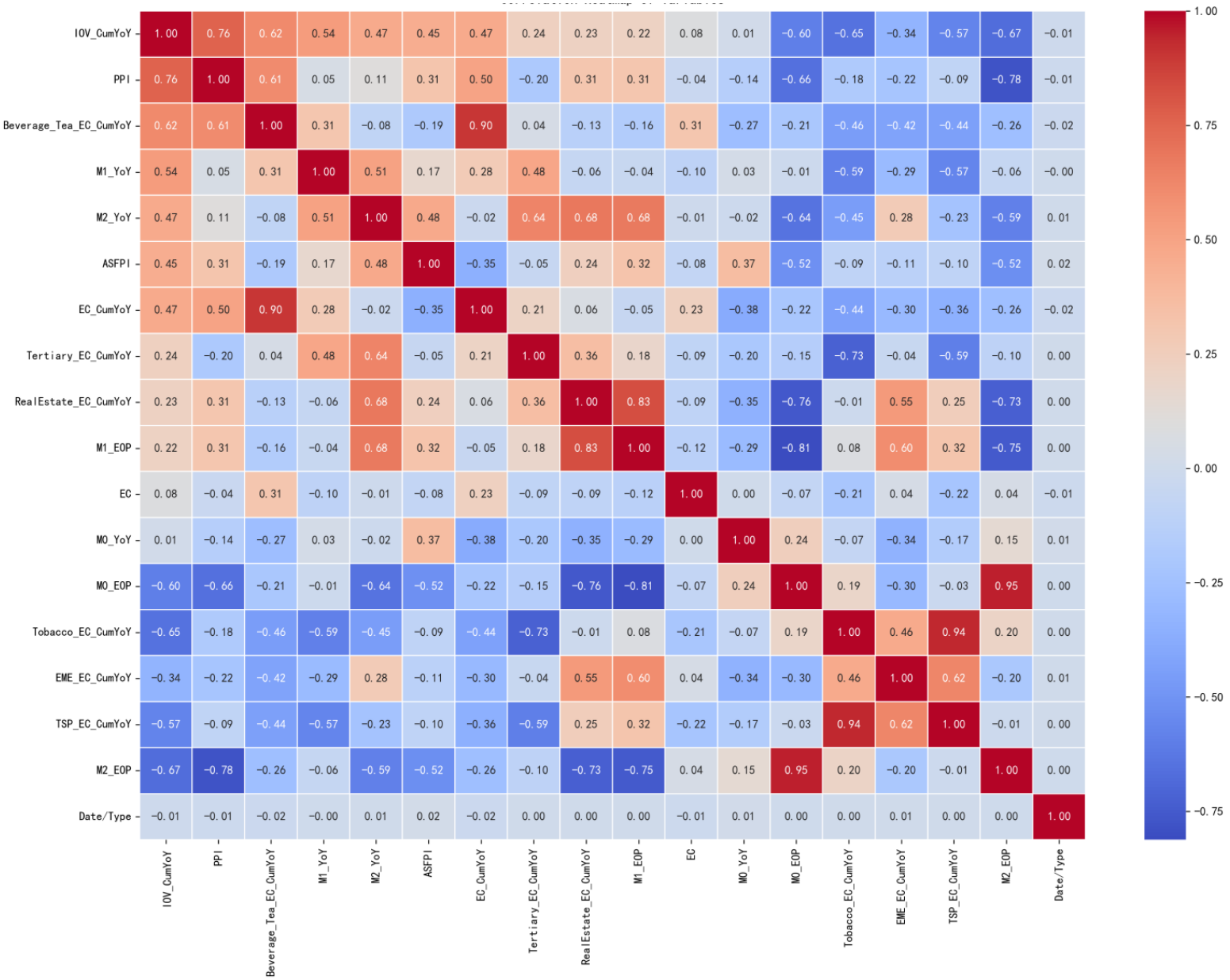


Figure 2. Correlation heatmap for the variables used in the Lüliang macroeconomic forecasting model

Model Evaluation

As illustrated in Figure 3, which compares cross-validation outcomes, lower MSE indicates higher model accuracy; hence, Lasso is identified as the optimal model. Quantitative comparison shows that Lasso regression achieves an MSE as low as 0.000344. It reduces the error by 78.5% relative to Ridge regression (MSE = 0.0016), 83.6% relative to Random Forest (MSE = 0.0021), and 90.4% relative to XGBoost (MSE = 0.0036), presenting a substantial error decline. The optimal Lasso model yields a coefficient of determination of 0.771, implying that 77.1% of industrial output-value fluctuations can be interpreted by the model. The mean absolute error MAE = 0.0141 demonstrates that the average prediction bias for monthly year-on-year growth in industrial output value is only 1.41 percentage points. Such forecasting accuracy meets the demands for governmental monthly economic condition judgment. Comprehensive model evaluation metrics are provided in Table 4. The MAE indicates that the average monthly prediction deviation of year-on-year output value growth is 1.41 percentage points, which meets the requirement for industry-cycle monitoring.

Table 4. Evaluation metrics for the selected Lasso regression model

Metric	Value	Metric	Value
MSE	0.000344	RMSE	0.018559
MAE	0.014054	R^2	0.771228

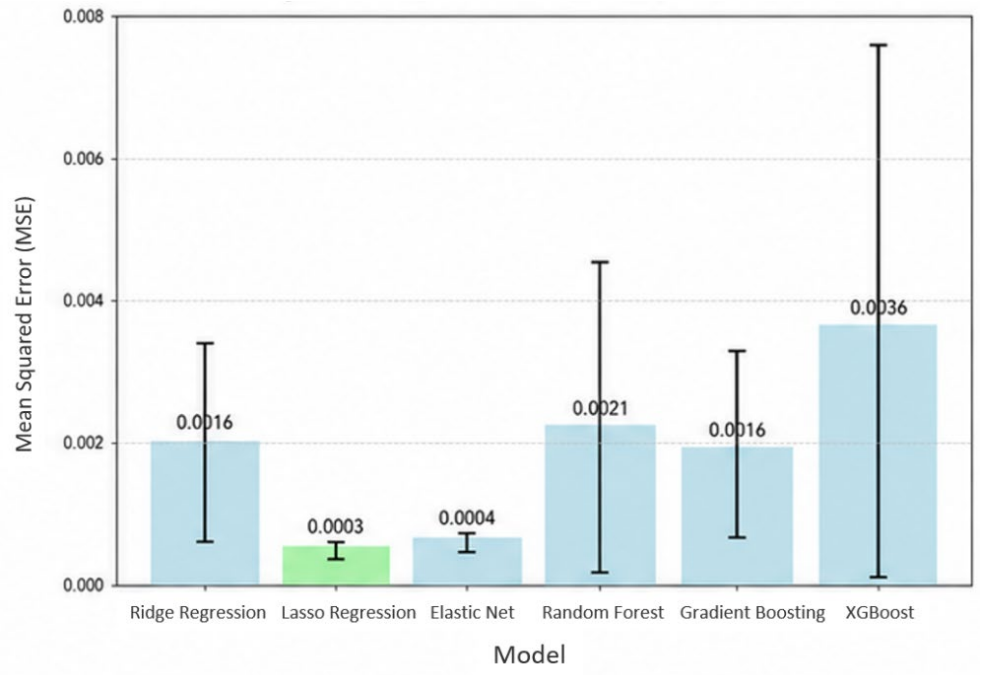


Figure 3. Comparison of five-fold cross-validation results across candidate models (k = 5)

Model Forecasting Results

For an intuitive comparison of time-series fitting performance among seven models on the test set, comparison curves are generated by overlaying each model's predicted curve onto the real time-series of cumulative year-on-year growth in industrial output value growth, as presented in Figure 4. Visual inspection of the curve trends reveals that Ridge Regression, Random Forest, XGBoost, and other algorithms produce considerable deviations at periodic troughs and abrupt turning points. By contrast, the Lasso regression prediction curve achieves better alignment with the real-time series over the entire time horizon, free from severeover-- or under-prediction. It delivers superior overall fitting stability relative to the remaining six models.

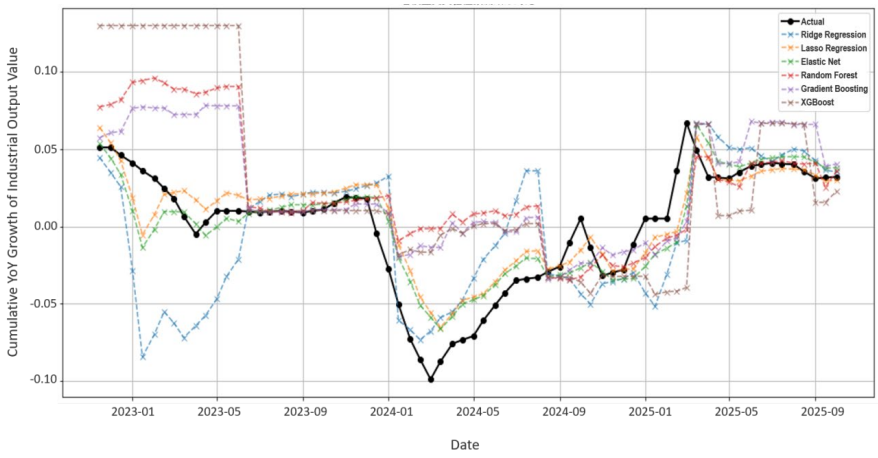


Figure 4. Comparison of predictions generated by the candidate models

According to the optimal selection results from cross-validation, the Lasso regression model with the best comprehensive performance is adopted for separate time-series fitting. The contrast between model predictions and actual year-on-year growth industrial output values is illustrated in Figure 5. Referring to the aforementioned quantitative evaluation metrics, the model yields an MAE of only 0.014054, corresponding to an average prediction bias of 1.41 percentage points for monthly year-on-year growth output-value growth. Minor under-prediction is observed only in the industrial trough of Q1 2024, whereas errors remain within reasonable bounds for the rest of the time span. Accordingly, it meets the precision demands for monthly regional economic judgement.

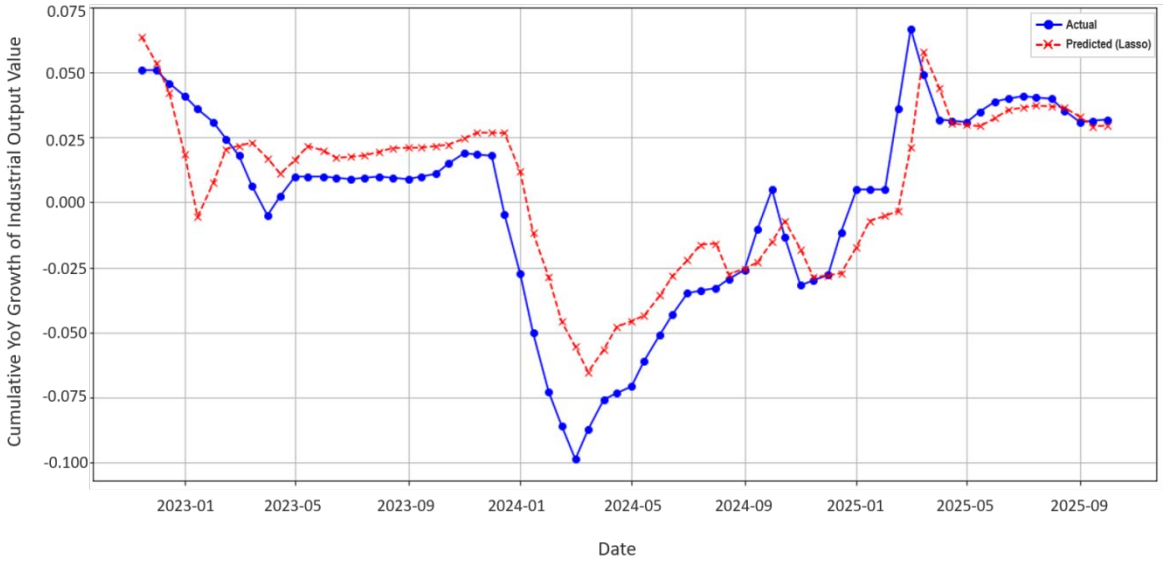


Figure 5. Comparison of Lasso regression predictions and observed values

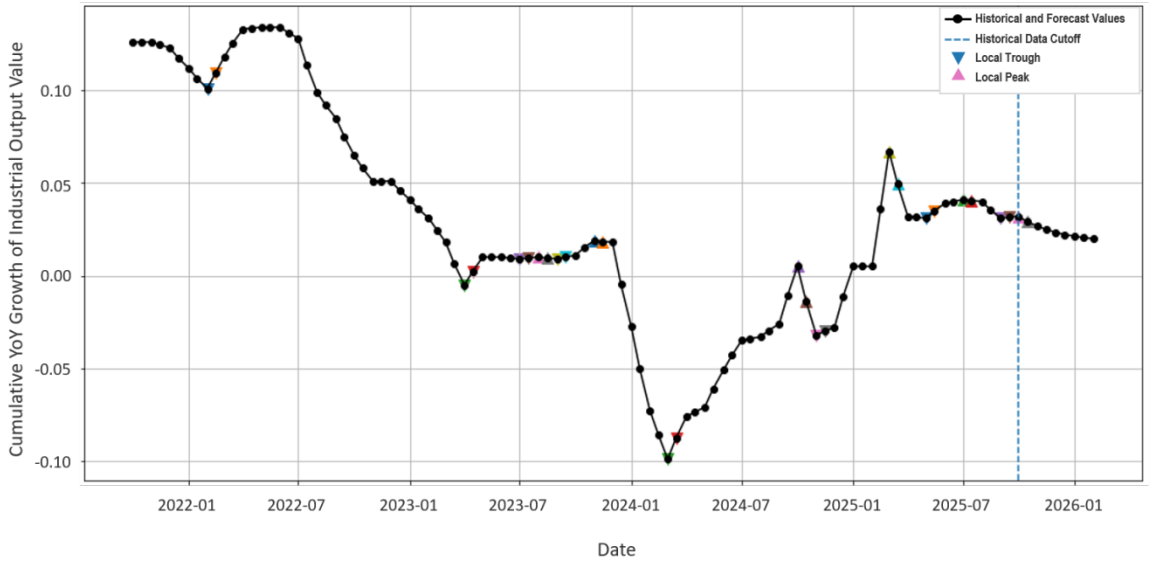


Figure 6. Future multi-step forecast and predicted turning points

Multi-step rolling prediction with eight horizons is performed based on the trained Lasso regression model. Historical observed time-series and future prediction intervals are integrated, with annotations for local prosperity

peaks and trough turning points, as shown in Figure 6. Exact predicted values for the subsequent eight time-steps are listed in Table 5. Quantitative outputs indicate that Lvliang's city-wide cumulative year-on-year growth industrial output value will maintain a downward trend, gradually dropping from 0.029054 (15 October 2025) to 0.019937 (1 February 2026). The cumulative growth rate declines by 0.9117 percentage points over eight periods, with an average monthly drop of about 0.13 percentage points, verifying the moderate short-term slow-growth trend of Lvliang's macro-industry. As reflected by the turning-point labels in Table 6, October 2025 marks the local prosperity peak of this cycle, and regional industrial prosperity thereafter enters a sustained downward trajectory.

Table 5. Multi-step forecast results of Lüliang macroeconomy

No.	Forecast Date	Predicted Value
1	2025-10-15	0.029054
2	2025-11-01	0.026714
3	2025-11-15	0.024854
4	2025-12-01	0.023376
5	2025-12-15	0.022201
6	2026-01-01	0.021268
7	2026-01-15	0.020526
8	2026-02-01	0.019937

Table 6. Local peak-trough turning-point records of Lüliang macroeconomic prosperity

Date	Type	Value
2025-10-01	Local peak (transition from increase to decrease)	0.032000
2025-10-15	Local peak (transition from increase to decrease)	0.029054

3.2 Coal Industry Forecasting Analysis

Data for this analysis were obtained from the Lüliang Municipal Bureau of Statistics and the CCTD China Coal Market Network. The predictor variables included coal-industry electricity consumption and its cumulative YoY growth; the end-of-period balances and YoY growth rates of M2, M1, and M0; the Producer Price Index for Industrial Products (same month of the previous year = 100); national raw coal output, output value, and their YoY growth rates; coal import volume and value; and cumulative YoY growth in electricity consumption in non-metallic mineral products manufacturing, fabricated metal products manufacturing, and the mining industry.

Variable Correlation Analysis

As can be seen from the correlation heatmap Figure 7, the variables with the highest correlation with the cumulative year-on-year growth rate of industrial output value are coal import volume (0.70), cumulative year-on-year output growth rate (0.58), national raw coal output (year-on-year) (0.52), and cumulative year-on-year electricity consumption of the non-metallic mineral products industry (0.47), in sequence. This indicates that the development of the coal industry in Luliang City is closely related to national raw coal output, the PPI index, and the M0 index.

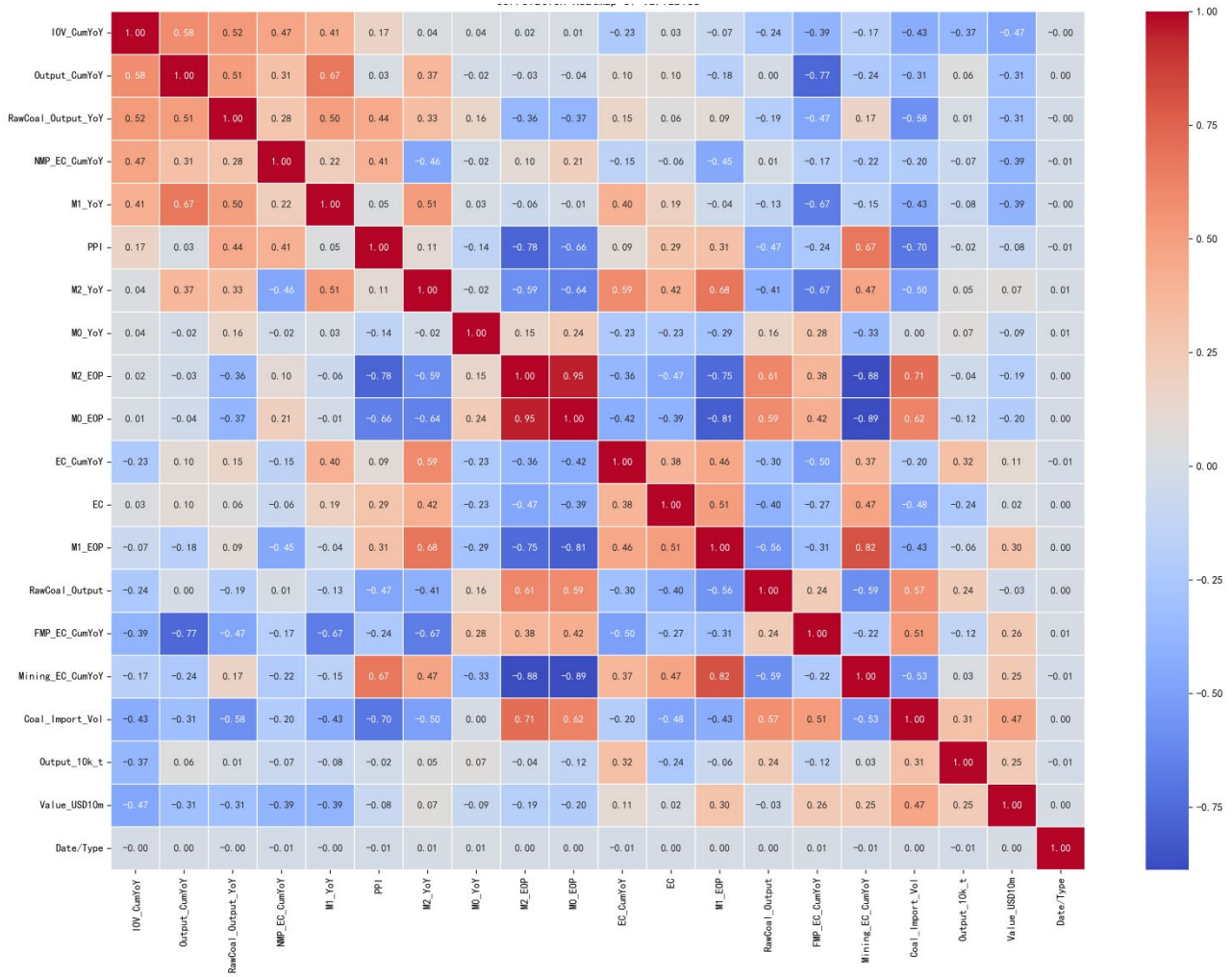


Figure 7. Variable correlation heatmap for the coal industry

Model Forecasting Results

Seven candidate models were uniformly measured for error metrics including MSE and RMSE via chronological rolling cross-validation, and the chronological rolling cross-validation MSE comparison results of each model are presented in Figure 8. As reflected by the quantitative values in the bar chart, the Lasso regression yields the lowest error among all algorithms, with an MSE of merely 0.001588. By contrast, ridge regression achieves a considerably higher MSE of 0.0182, which is 11.4 times that of the optimal model. The ensemble models such as random forest and XGBoost obtain MSE values of 0.0024 and 0.0045, respectively, both showing inferior fitting accuracy compared with Lasso regression. The complete quantitative evaluation metrics for Lasso regression are summarized in Table 7. This model can explain 76.42% of the fluctuation in coal-industry output value. Its MAE of 0.029847 indicates that the average prediction deviation for monthly year-on-year output value is only 2.98 percentage points, and such prediction accuracy can support the cycle judgment of the coal industry. The MAE indicates that the average monthly prediction deviation of year-on-year output value growth is 2.98 percentage points, which meets the requirement for industry-cycle monitoring.

Table 7. Evaluation metrics for the coal-industry forecasting model

Metric	Value	Metric	Value
MSE	0.001588	RMSE	0.039851
MAE	0.029847	R^2	0.764184

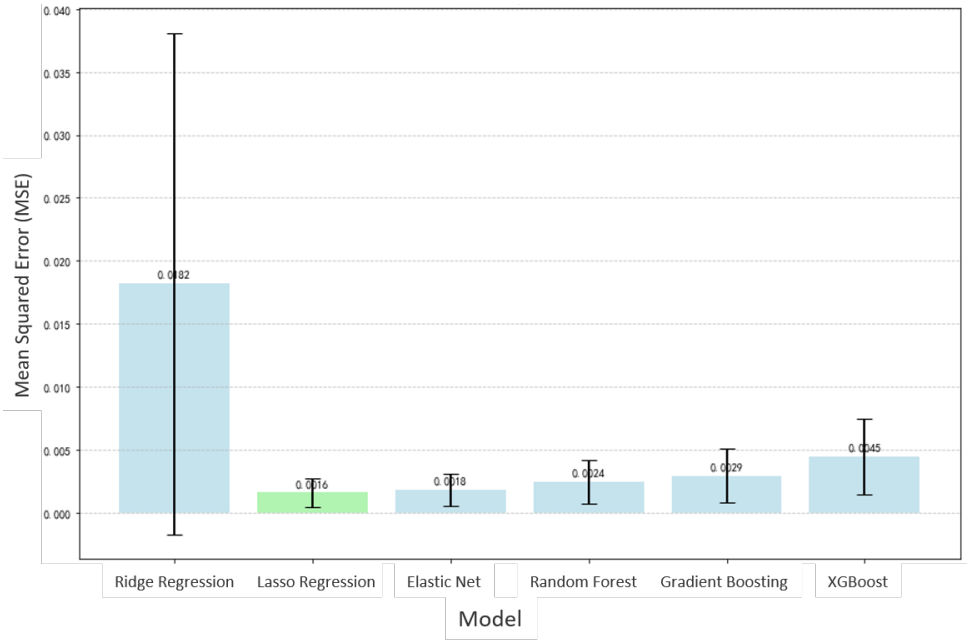


Figure 8. Comparison of five-fold cross-validation results for the coal-industry models (k = 5)

Model Forecasting Results

The prediction curves of the seven models on the test set are overlaid with the time-series of real year-on-year output value of the coal industry, yielding the multi-model prediction comparison curves shown in Figure 9. Visual comparison of curve-fitting performance reveals that ridge regression produces substantial underestimation in the first half of 2024; XGBoost and random forest exhibit obvious deviations at the business-cycle inflection points. The prediction curve of Lasso regression closely tracks the real time-series throughout the whole period, with remarkably superior stability relative to the other algorithms.

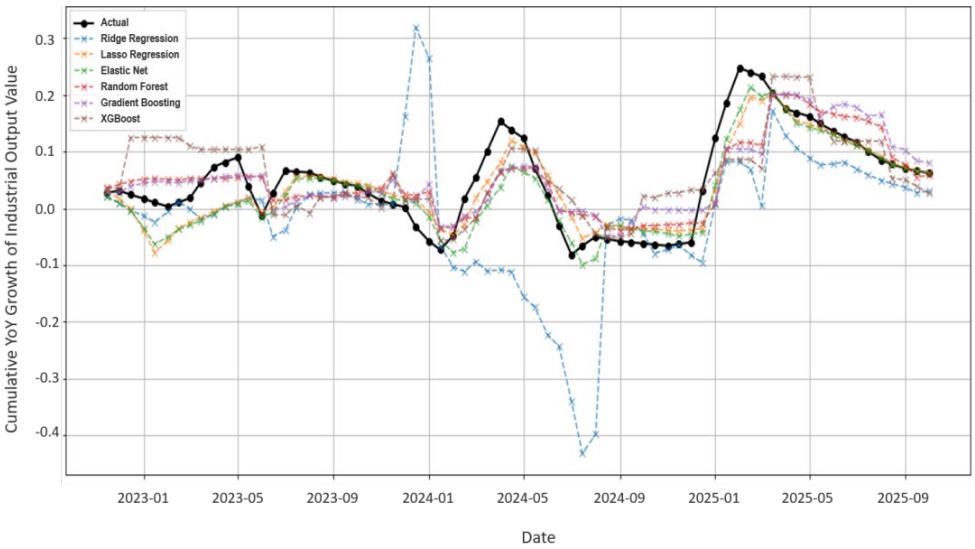


Figure 9. Comparison of test-set predictions across candidate models for the coal industry

The optimal Lasso regression is selected to independently output the predicted sequence. The comparison between predicted values and real year-on-year output values is illustrated in Figure 10. Throughout the whole cycle, the model only presents minor deviations during the troughs in May 2023 and the first quarter of 2024. No systematic overestimation or underestimation is observed overall, indicating stable and reliable time-series fitting performance.

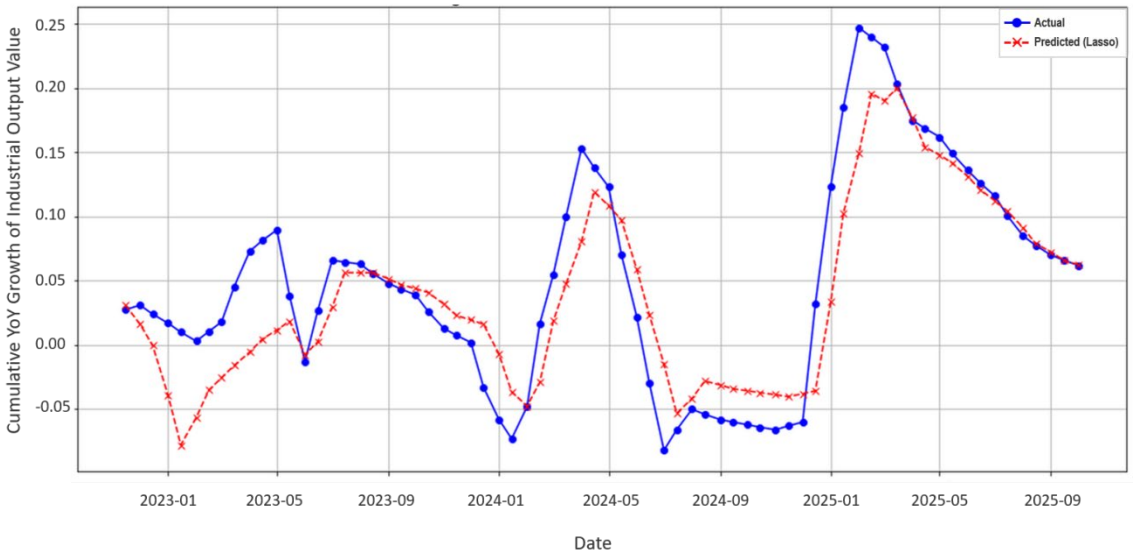


Figure 10. Comparison of Lasso regression predictions and observed values for the coal industry

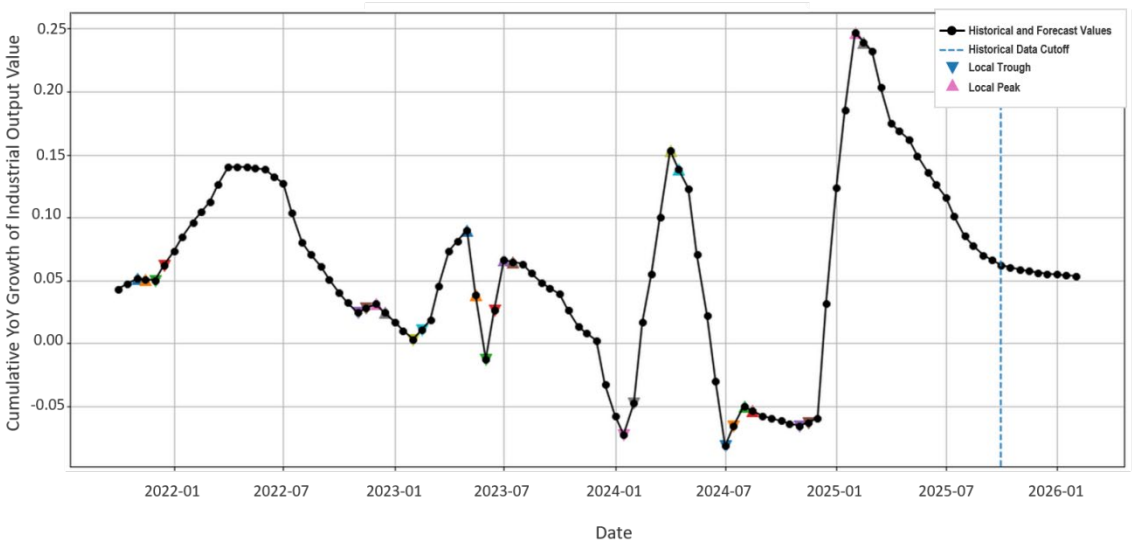


Figure 11. Future multi-step forecast for the coal industry

An eight-step rolling prediction is performed based on the trained Lasso regression. Historical actual data and future prediction intervals are integrated, while business-cycle peak and trough inflection points are marked. The complete time-series trend is shown in Figure 11. The precise predicted values for each future time point are listed in Table 8. The results indicate that the overall development of the coal industry in Luliang City will maintain a continuous upward trend compared with last year, whereas the cumulative year-on-year growth rate will slow down relative to the preceding time point. Overall, the cumulative year-on-year growth will remain at approximately 5 % over the next three months.

Table 8. Future multi-step forecasts for the coal industry

No.	Forecast Date	Predicted Value
1	2025-10-15	0.060052
2	2025-11-01	0.058447
3	2025-11-15	0.057110
4	2025-12-01	0.056021
5	2025-12-15	0.055133
6	2026-01-01	0.054409
7	2026-01-15	0.053819
8	2026-02-01	0.053338

3.3. Liquor Industry Forecasting Analysis

Data for this analysis were obtained from the Lüliang Municipal Bureau of Statistics. The predictor set included electricity consumption in the liquor industry and its cumulative year-on-year (YoY) growth. It also included the end-of-period balances and YoY growth rates of broad money (M2), narrow money (M1), and currency in circulation (M0). Additional variables were the Producer Price Index for Industrial Products (PPI; same month of the previous year = 100), liquor output, and cumulative YoY growth in liquor output. Industry-level electricity indicators covered mining, ferrous-metal mining and dressing, real estate, paper and paper products, cultural, educational, arts and crafts, sports and entertainment products, and leasing and business services.

Variable Correlation Analysis

Figure 12 shows that cumulative YoY growth in industrial output value is most strongly correlated with cumulative YoY growth in electricity consumption in the mining industry (0.84), ferrous-metal mining and dressing (0.80), the PPI (0.70), total electricity consumption (0.69), and the end-of-period M1 balance (0.68). Its correlations with production-volume indicators are relatively weak. This pattern suggests that short-term changes in liquor industrial output value are more closely associated with broader economic conditions than with contemporaneous production volume

.

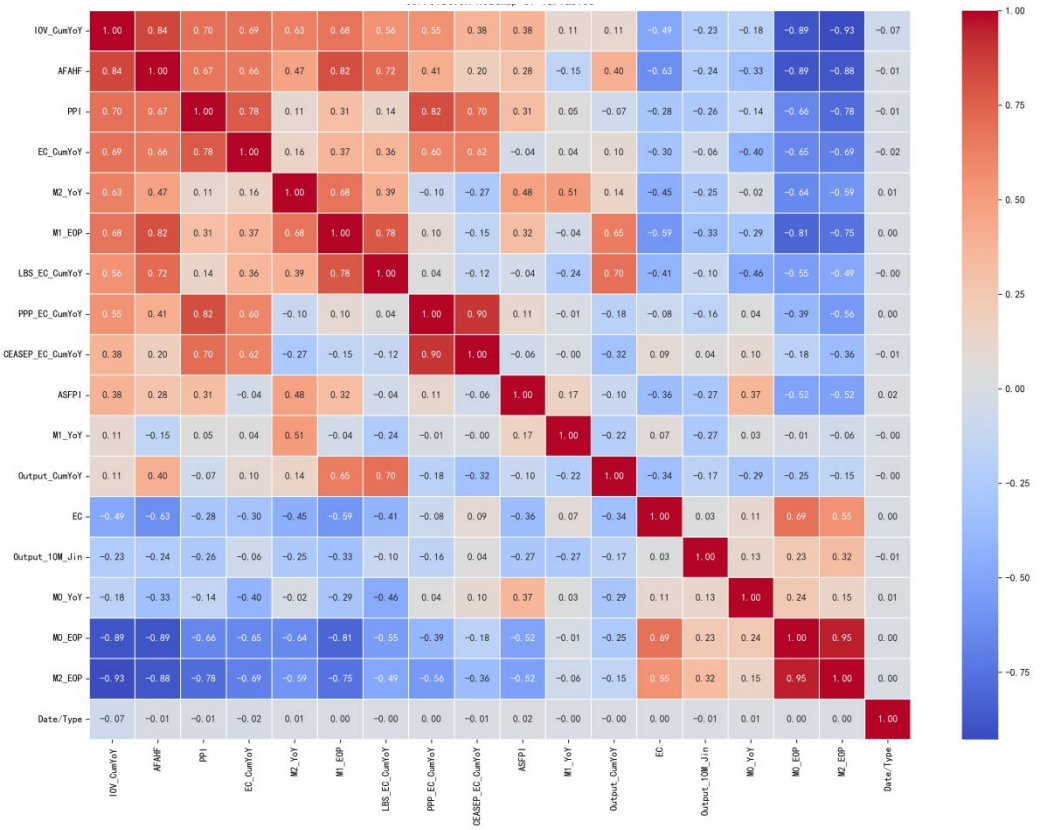


Figure 12. Variable correlation heatmap for the liquor industry

Model Forecasting Results

Uniform performance tests are conducted on seven candidate models via chronological rolling cross-validation, and the MSE error comparison results for each model are presented in Figure 13. As observed from the numerical comparison of bar-chart values, Lasso regression delivers the optimal performance with an MSE of merely 0.000826. The MSE values of ridge regression, random forest, and XGBoost reach 0.0074, 0.0019, and 0.0031, respectively; their errors are 7.96-fold, 1.31-fold, and 2.75-fold higher than those of Lasso regression. The ensemble-learning models fail to demonstrate advantages in prediction accuracy. The full evaluation metrics of the Lasso-regression model for the liquor industry are summarized in Table 9. This model can explain 85.91 % of the output-value fluctuation in the liquor industry. Its MAE equals 0.021436, corresponding to an average prediction deviation of 2.14 percentage points for monthly year-on-year output value. Such high explanatory power verifies that the model is well-suited for prediction tasks in the liquor industry. The MAE indicates that the average monthly prediction deviation of year-on-year output value growth is 2.14 percentage points, which meets the requirement for industry-cycle monitoring.

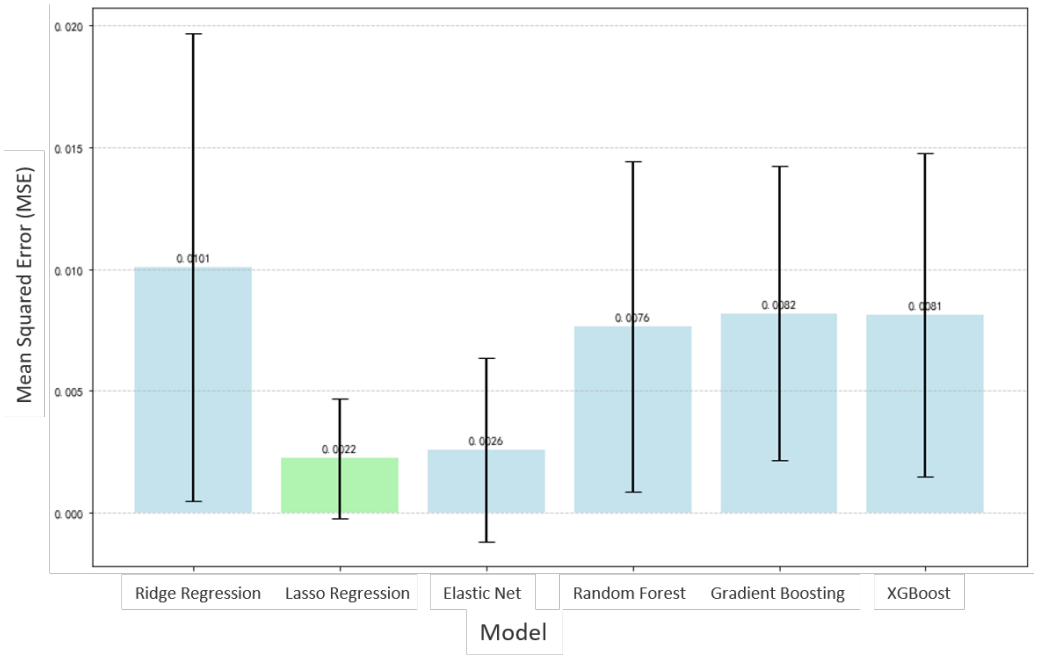


Figure 13. Comparison of five-fold cross-validation results for the liquor-industry models (k = 5)

Table 9. Performance metrics for the selected liquor-industry model

Metric	Value	Metric	Value
MSE	0.002228	RMSE	0.047207
MAE	0.034098	R^2	0.858634

Model Forecasting Results

The predicted sequences of the seven models on the test set are overlaid with the time-series of real year-on-year output value for the liquor industry to plot multi-model prediction comparison curves, as shown in Figure 14. Comparison of curve trends reveals that ridge regression suffers from persistent overestimation, while random forest and XGBoost exhibit obvious fitting deviations at business-cycle inflection points. In contrast, the prediction curve of Lasso regression achieves a higher overall fitting degree with the real-time series, and it can well capture both the upward and downward phases of business cycles.

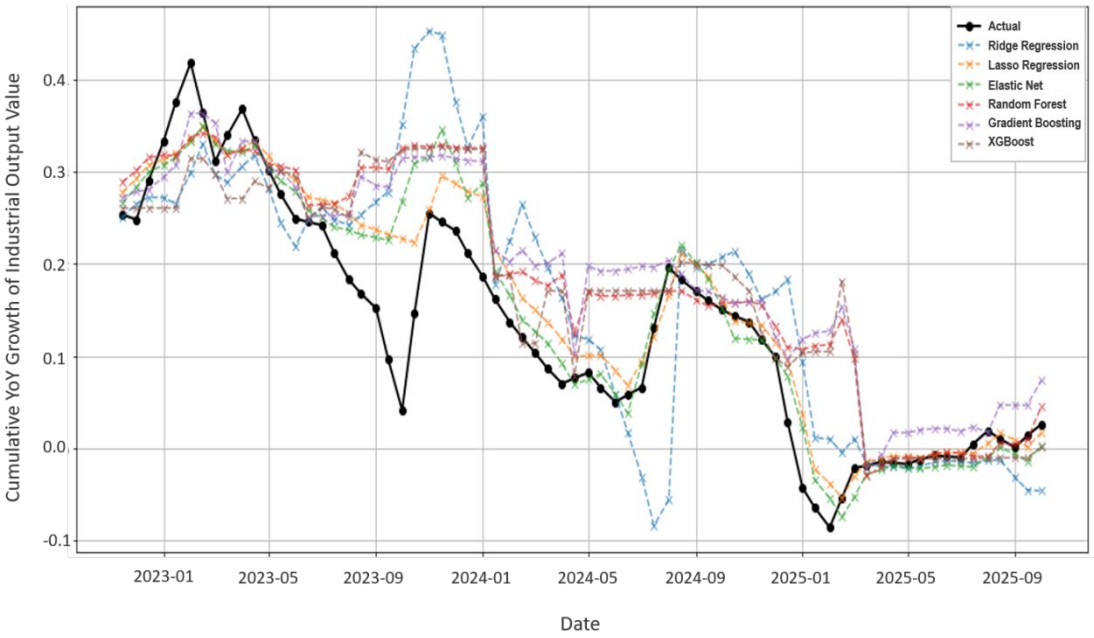


Figure 14. Comparison of cross-validation predictions across liquor-industry models

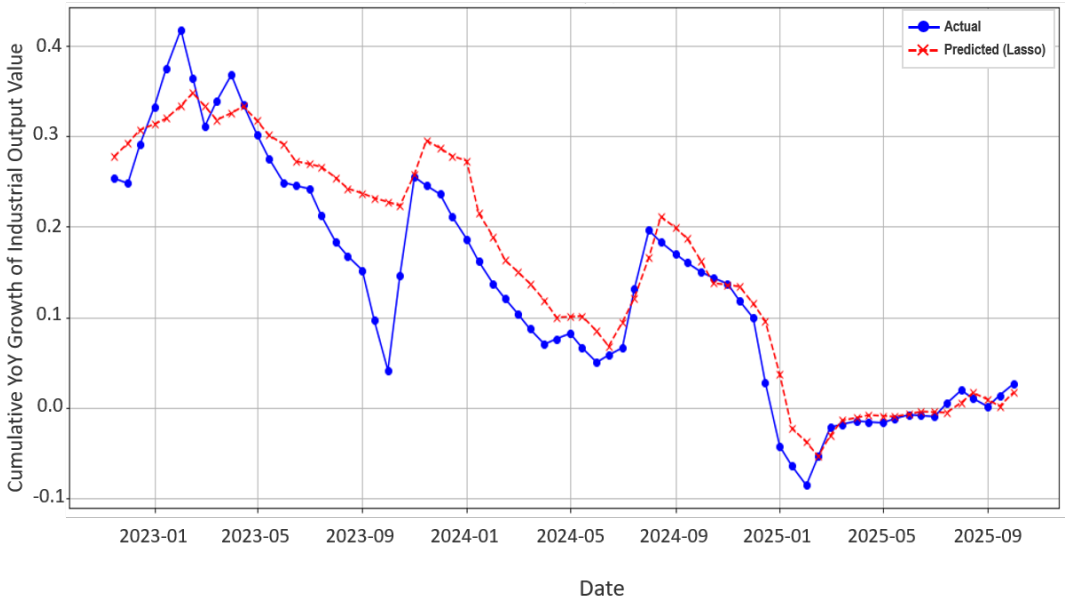


Figure 15. Lasso regression predictions and observed values for the liquor industry

Further, the comparison curve between predicted and actual values of the optimal Lasso regression model is plotted, as illustrated in Figure 15. Across the full time-series interval, the model only produces minor underestimation during the short-term surge in mid-2024. No systematic bias is observed, and its overall stability meets the requirements for industrial monitoring.

An eight-step rolling prediction is implemented using the trained Lasso regression. The historical series and future prediction intervals are output, and key inflection points are marked. The complete time-series variations are presented in Figure 16. The specific predicted values for each future time node are listed in Table 10. According to the quantitative results, the cumulative year-on-year output value of the liquor industry drops slightly from 0.048162 on 2025-10-15 to 0.041273 on 2026-02-01, with a total decline of 0.006889 across the eight periods. It experiences a mild monthly average decline yet maintains positive growth above 4 % throughout, demonstrating strong development resilience.

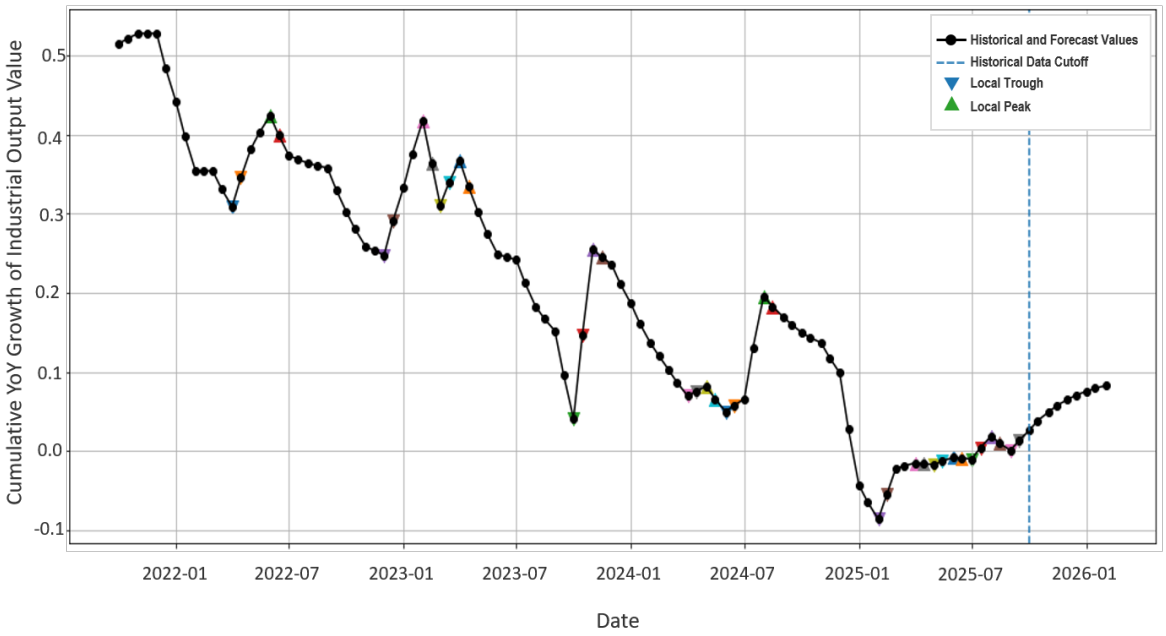


Figure 16. Multi-step forecast and turning points for the liquor industry

Table 10. Future multi-step forecasts for the liquor industry

No.	Forecast Date	Predicted Value
1	2025-10-15	0.038741
2	2025-11-01	0.049225
3	2025-11-15	0.057853
4	2025-12-01	0.064952
5	2025-12-15	0.070795
6	2026-01-01	0.075602
7	2026-01-15	0.079558
8	2026-02-01	0.082814

3.4. Aluminum Smelting Industry Forecasting Analysis

Data for this analysis were obtained from the Lüliang Municipal Bureau of Statistics. The predictor set included electricity consumption in the aluminum smelting industry and its cumulative YoY growth. It also included the end-of-period balances and YoY growth rates of M2, M1, and M0, together with the PPI. Additional electricity indicators covered nonferrous-metal smelting and rolling, furniture manufacturing, chemical-fiber manufacturing, and railway, shipbuilding, aerospace, and other transportation-equipment manufacturing.

Variable Correlation Analysis

Figure 17 shows that cumulative YoY growth in aluminum-smelting output value is most strongly correlated with cumulative YoY growth in electricity consumption in chemical-fiber manufacturing (0.92). It is also correlated with the PPI (0.69), electricity consumption in furniture manufacturing (0.57), and electricity consumption in railway, shipbuilding, aerospace, and other transportation-equipment manufacturing (0.47).

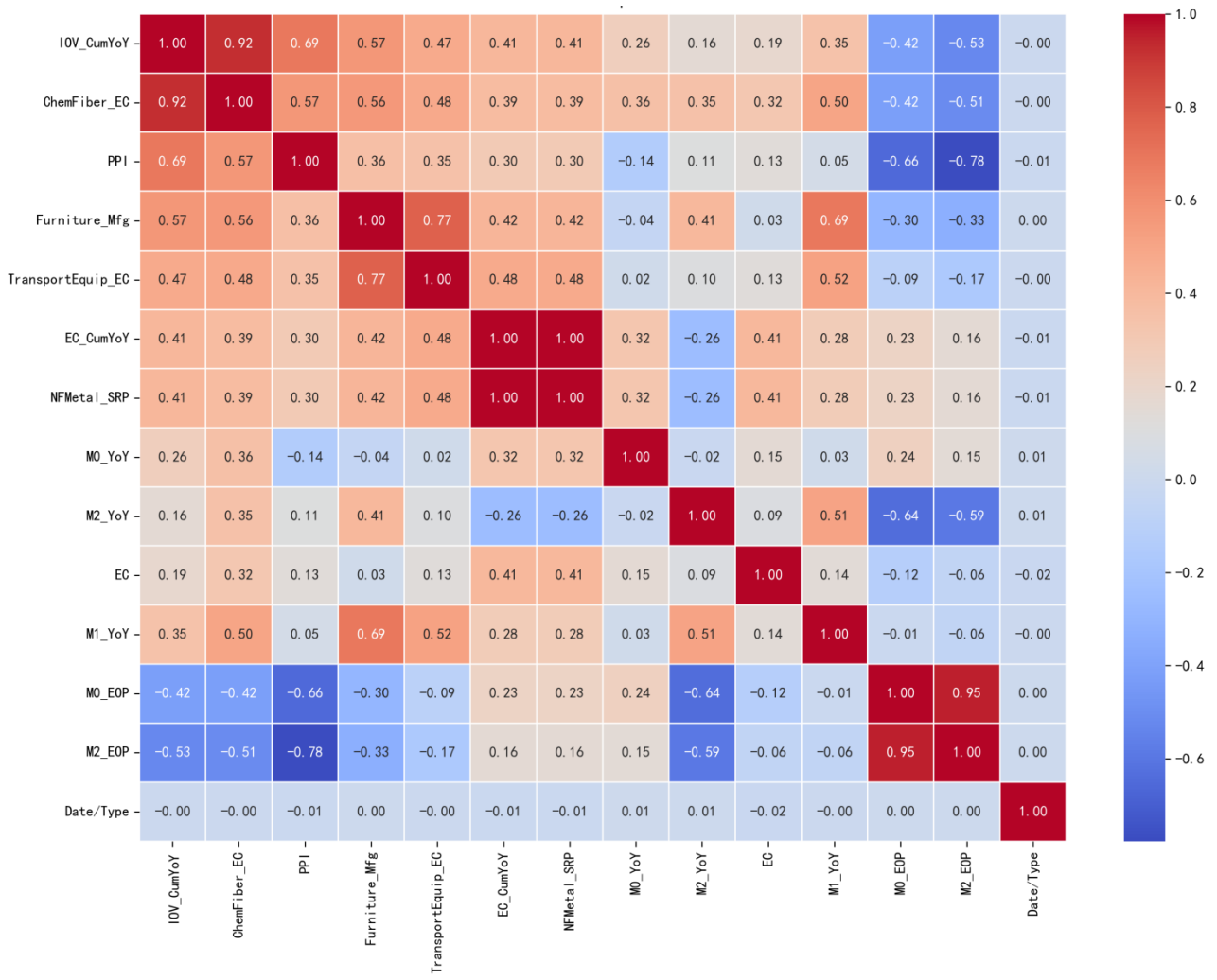


Figure 17. Variable correlation heatmap for the aluminum smelting industry

Model Forecasting Results

Cross-validation is adopted to uniformly calculate error metrics for the seven candidate models, and the quantitative MSE comparison results of each model are shown in Figure 18. As indicated by the bar chart, Lasso regression achieves the lowest error among all models with an MSE of merely 0.001798. Ridge regression yields an MSE of 0.00592, whose error magnitude is approximately 3.3 times that of the Lasso model. The ensemble algorithms, including random forest and XGBoost, both produce MSE values above 0.008, showing a distinct gap in fitting accuracy. The complete evaluation metrics for Lasso regression are summarized in Table 11. With a coefficient of determination, the model can explain 55.96 % of the output-value fluctuation in the aluminum smelting industry. Its MAE equals 0.029023, corresponding to an average prediction deviation of 2.90 percentage points for monthly year-on-year output value, which can support the medium-- and long-term business-cycle judgment for the aluminum industry. The MAE indicates that the average monthly prediction deviation of year-on-year output-value growth is 2.90 percentage points, which meets the requirement for industry-cycle monitoring.

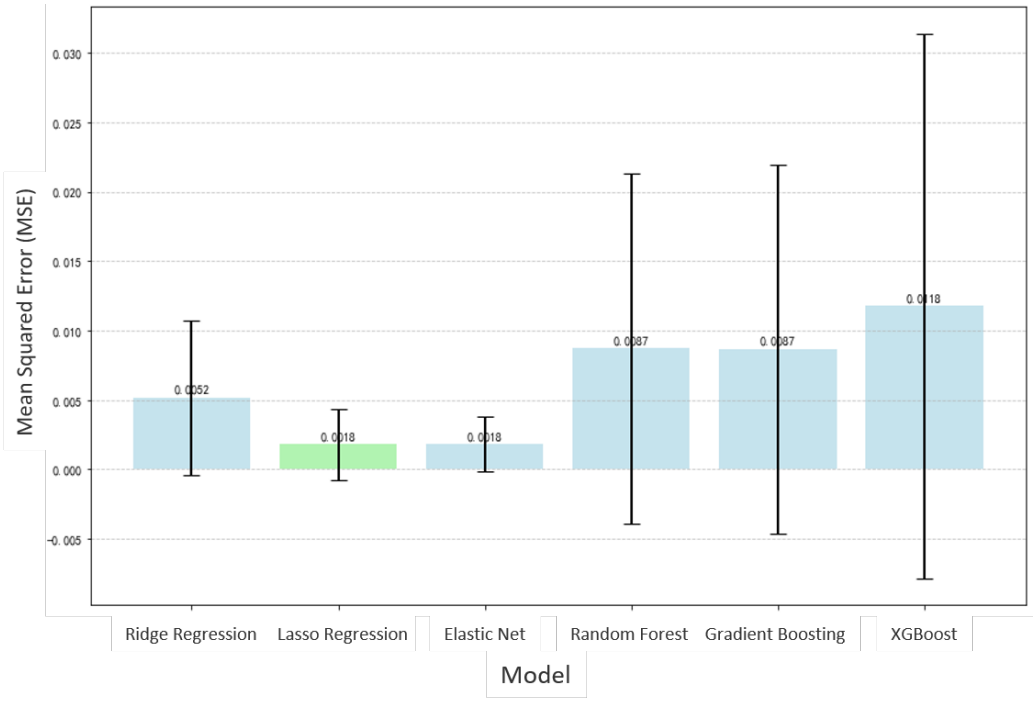


Figure 18. Five-fold cross-validation MSE comparison for the aluminum-smelting models (k = 5)

Table 11. Performance metrics for the selected aluminum-smelting model

Metric	Value	Metric	Value
MSE	0.001797	RMSE	0.042388
MAE	0.029023	R^2	0.559614

Model Forecasting Results

The prediction curves of all seven types of models on the test set are superimposed with the time-series year-on-year real output value of aluminum smelting to plot a multi-model fitting comparison diagram, and the results are shown in Figure 19. The curves intuitively reveal that ridge regression and ensemble-learning algorithms exhibit substantial prediction deviations at cyclical troughs and abrupt prosperity-shift nodes. In contrast, the prediction curve of Lasso regression generally fits the real variation trend and achieves better control over prediction errors at extreme points. The predicted series from the optimal Lasso regression model is extracted for separate presentation. Figure 20 compares its predicted values against the year-on-year real output value. Slight underestimation is observed only in early 2023 and mid-2024 across the whole time span; no persistent systematic deviation occurs, indicating good overall fitting stability.

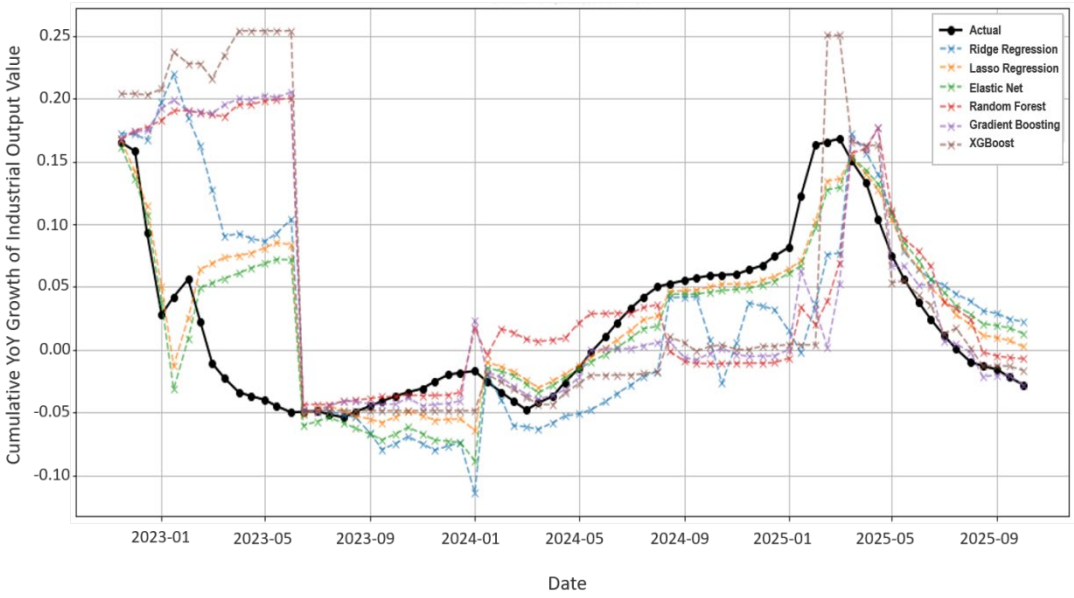


Figure 19. Comparison of cross-validation predictions across aluminum-smelting models

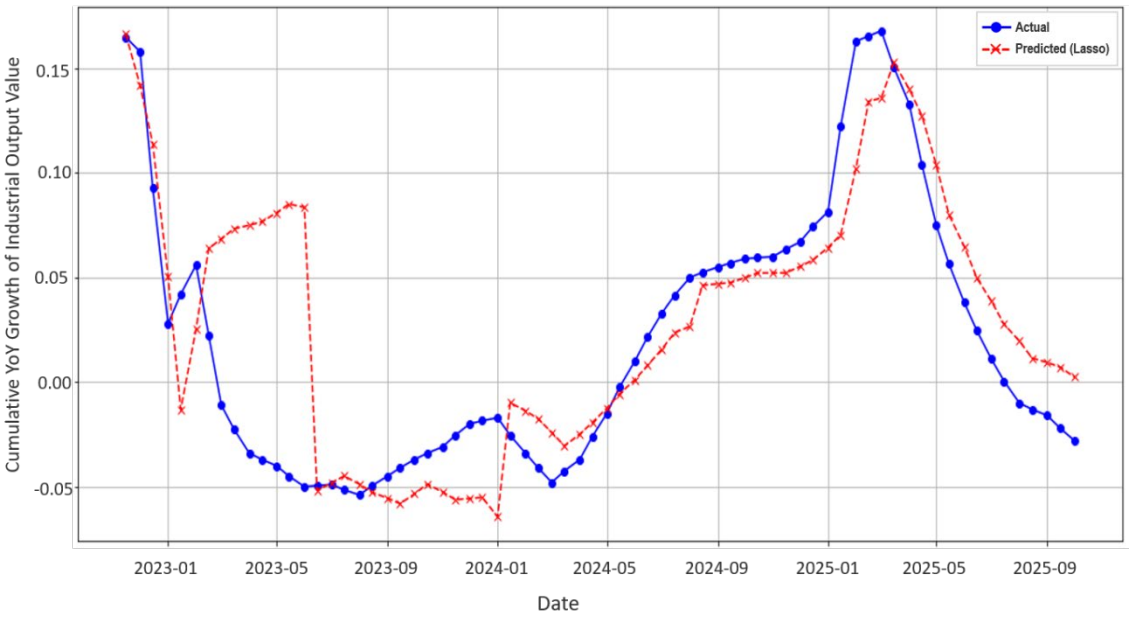


Figure 20. Lasso regression predictions and observed values for the aluminum smelting industry

Based on the trained Lasso regression, an eight-step rolling prediction is carried out. Historical time-series data and future prediction intervals are integrated, and business-cycle peak and trough inflection points are marked. The complete trend is shown in Figure 21. The predicted cumulative year-on-year output values of the industry for the next three months are listed in Table 12. These predicted cumulative year-on-year values for the aluminum smelting industry present a V-shaped recovery pattern. The results indicate that the overall development of Luliang's aluminum smelting industry will first decline compared with last year, yet the cumulative year-on-year growth will turn from negative to positive. It signals that the industry will enter a high-quality development stage featured by "stable volume and improved quality". This is particularly beneficial for leading enterprises that have deployed new-energy tracks and possess green and low-carbon technologies.

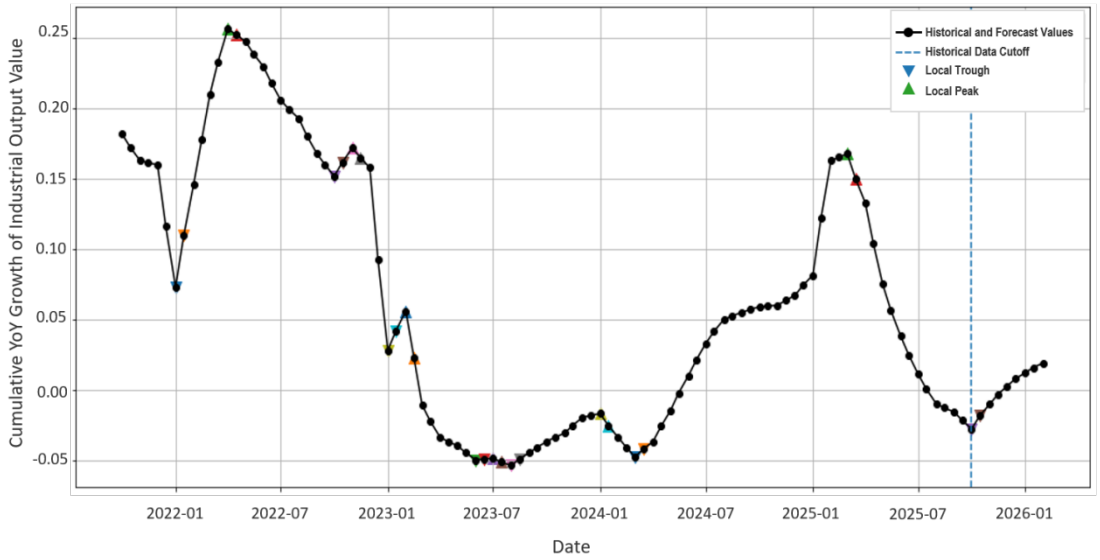


Figure 21. Multi-step forecast and turning points for the aluminum smelting industry

Table 12. Multi-step forecasts for the aluminum smelting industry

No.	Forecast Date	Predicted Value
1	2025-10-15	-0.018529
2	2025-11-01	-0.010413
3	2025-11-15	-0.003459
4	2025-12-01	0.002501
5	2025-12-15	0.007608
6	2026-01-01	0.011984
7	2026-01-15	0.015734
8	2026-02-01	0.018947

3.5. Steel Industry Forecasting Analysis

Data for this analysis were obtained from the Lüliang Municipal Bureau of Statistics. The predictor set included electricity consumption in the steel industry and its cumulative YoY growth. It also included the end-of-period balances and YoY growth rates of M2, M1, and M0, together with cumulative YoY growth in steel output.

Variable Correlation Analysis

Figure 22 shows that cumulative YoY growth in steel-industry output value has the strongest positive correlations with YoY growth in M0 (0.51) and YoY growth in M1 (0.39)

.

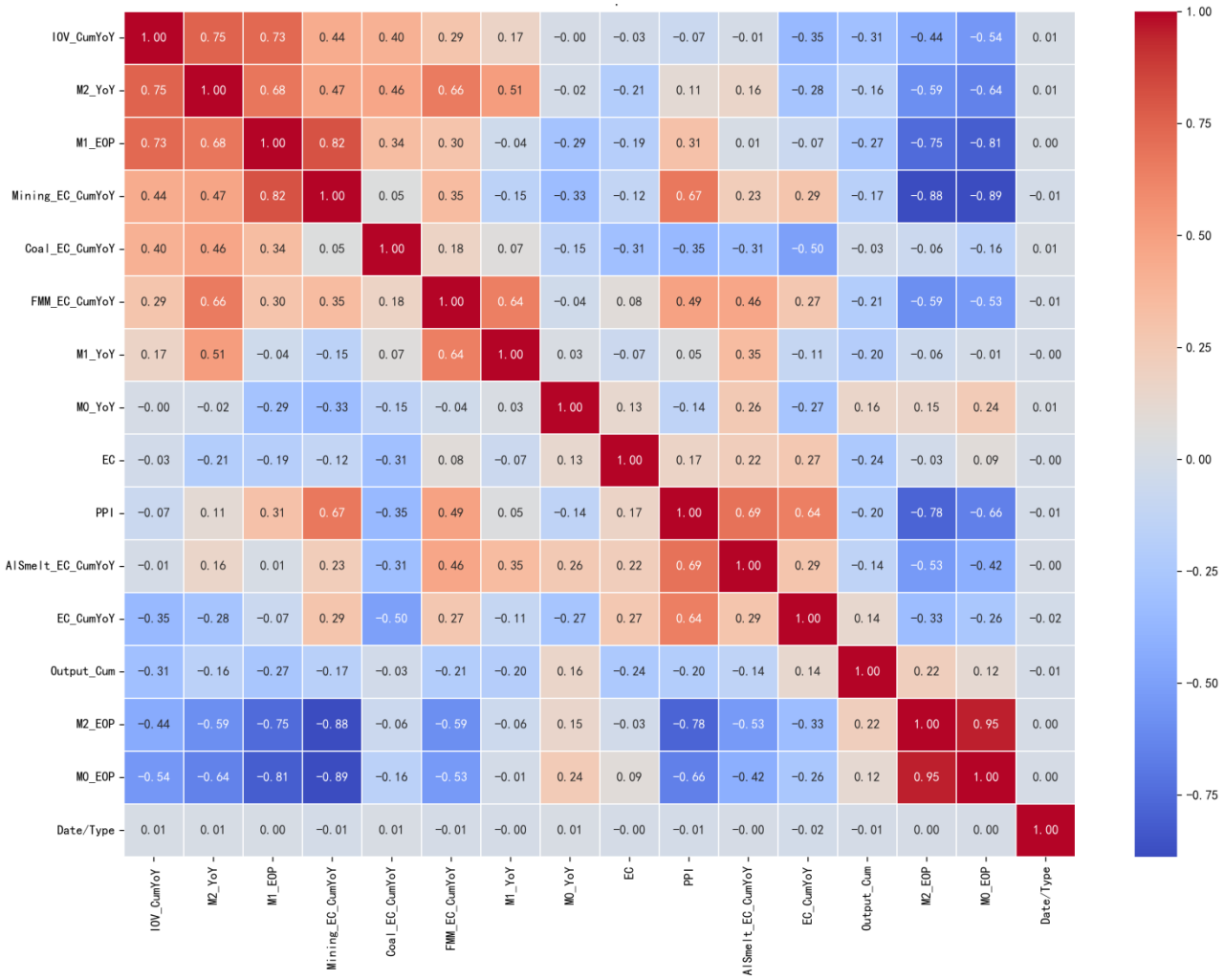


Figure 22. Variable correlation heatmap for the steel industry

Model Forecasting Results

Cross-validation is employed to uniformly calculate the mean squared error for the seven candidate models, and the bar chart for MSE comparison of each model is illustrated in Figure 23. Quantitative comparison shows that the Elastic-Net model yields the minimum error among all models, with an MSE of merely 0.001249. The MSE of Lasso regression is 0.0014, while ridge regression achieves a high MSE of 0.0042. The errors of ensemble algorithms such as random forest and XGBoost are significantly higher than those of the Elastic-Net model. The full quantitative evaluation metrics of the optimal Elastic-Net model for the steel industry are summarized in Table 13. This model can explain 69.40 % of the output-value fluctuation in the steel industry. Its MAE equals 0.025683, indicating an average prediction deviation of only 2.57 percentage points for monthly year-on-year output value. Capable of both feature selection and multicollinearity suppression, the model adapts well to the data characteristics of the steel industry mixed with multi-monetary and multi-electricity-consumption variables. The MAE indicates that the average monthly prediction deviation of year-on-year output-value growth is 2.57 percentage points, which meets the requirement for industry-cycle monitoring.

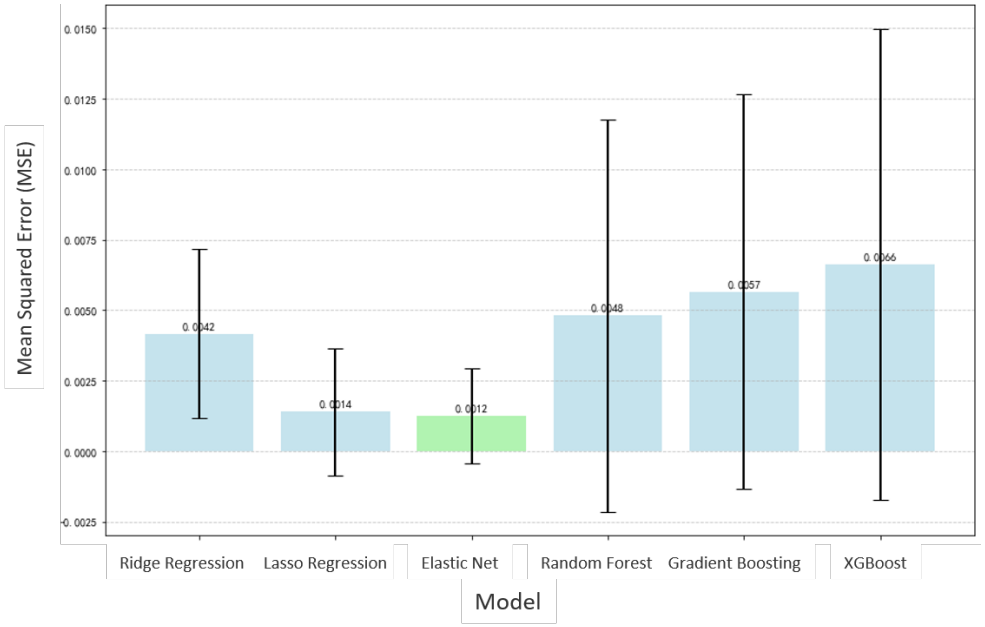


Figure 23. Five-fold cross-validation MSE comparison for the steel-industry models (k = 5)

Table 13. Performance metrics for the selected steel-industry model

Metric	Value	Metric	Value
MSE	0.001249	RMSE	0.035335
MAE	0.025683	R^2	0.693983

Model Forecasting Results

The prediction curves of the seven models on the test set are overlaid with the time-series of real year-on-year output value for the steel industry to generate a comparison plot, as shown in Figure 24. An intuitive observation of curve-fitting performance reveals that ridge regression and XGBoost suffer from severe underestimation during the trough period in mid-2024. Lasso regression presents minor deviations in fitting the long-term downward cycle. By contrast, the prediction curve of the Elastic-Net closely tracks the real time-series throughout the whole period, and it outperforms other algorithms in capturing peak and trough inflection points.

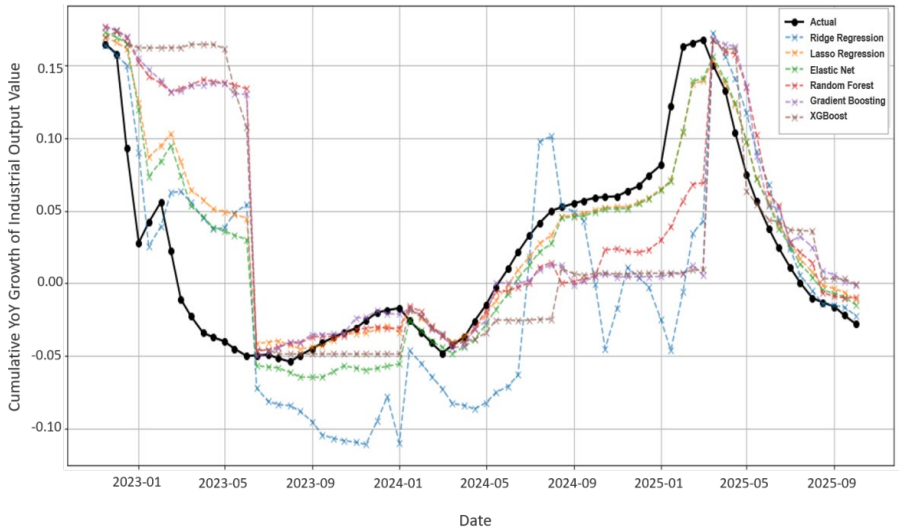


Figure 24. Comparison of cross-validation predictions across steel-industry models

The predicted sequence of the optimal Elastic-Net model is output independently, and the comparison between predicted values and real year-on-year output values is presented in Figure 25. Within the full time-series cycle, the model exhibits no persistent overestimation or underestimation. Only tiny errors occur during the short-term surge in early-2023, and its overall fitting stability satisfies the requirements for steel-industry cycle monitoring.

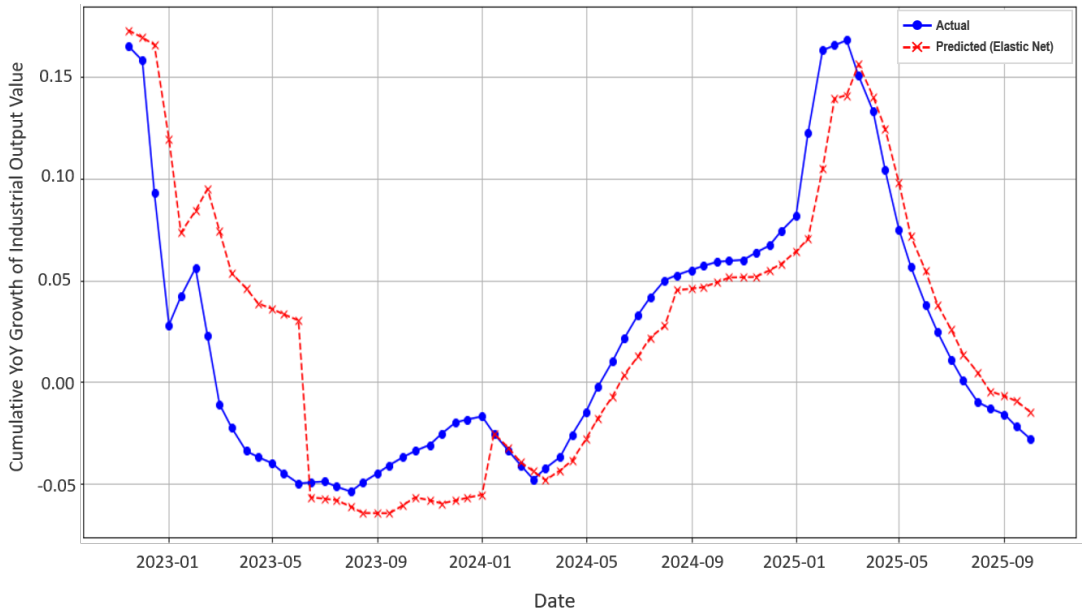


Figure 25. Elastic net regression predictions and observed values for the steel industry

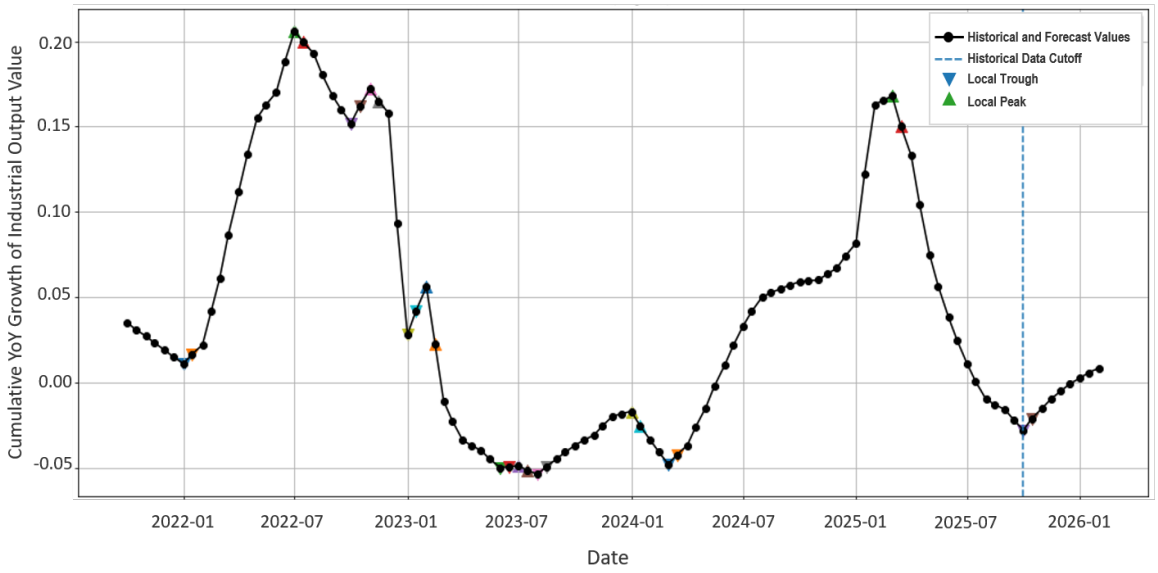


Figure 26. Multi-step forecast and turning points for the steel industry

Based on the trained Elastic-Net model, an eight-step rolling prediction is implemented. Historical data and future prediction intervals are integrated, and business-cycle peak-and-trough inflection points are marked. The complete time-series variation trend is illustrated in Figure 26. The predicted cumulative year-on-year output values of the industry for the next three months are listed in Table 14. The cumulative year-on-year output value of the steel industry will gradually rebound from its declining status in October 2025, achieve positive growth starting from January 2026, and reach a predicted value of 0.84 % in February 2026, indicating an overall recovery trend for the industry.

Table 14. Multi-step forecasts for the steel industry

No.	Forecast Date	Predicted Value
1	2025-10-15	-0.021107
2	2025-11-01	-0.015065
3	2025-11-15	-0.009770
4	2025-12-01	-0.005129
5	2025-12-15	-0.001061
6	2026-01-01	0.002504
7	2026-01-15	0.005628
8	2026-02-01	0.008367

4. Conclusions

Lüliang is an energy-industrial base undergoing economic transformation. This study addresses two key drawbacks of conventional macroeconomic data: low timeliness and insufficient capacity for dynamic regional-economic monitoring. Leveraging the high-frequency, fine-grained, and full-coverage features of power big data, we integrate official economic datasets and industrial electricity consumption records to break regional data silos and build a multi-algorithm-fusion model for industrial-economic forecasting. Cross-validation is used for adaptive model selection. Lasso regression is applied to macroeconomic and most-industry forecasting, whereas the Elastic Net is specially tailored for the steel industry. Consequently, the precision and stability of trend prediction for the regional economy and key industries are substantially enhanced. Empirical results demonstrate obvious heterogeneous coupling between electricity big-data features and industrial prosperity. Lasso regression is selected for the macroeconomy, coal, liquor, and aluminum-smelting industries, while Elastic Net is adopted for the steel industry to simultaneously handle feature selection and multicollinearity. Core driving factors are sector-specific, and the optimal models vary for individual industries. Lasso regression delivers superior comprehensive forecasting performance for macroeconomic, coal, liquor, and aluminum-smelting cases. By contrast, the Elastic Net, which simultaneously implements feature selection and multicollinearity mitigation, is better suited to the steel sector. According to 3–6-month rolling forecasts, Lvliang's aggregate industrial growth faces a moderate short-term slowdown accompanied by prominent industrial divergence. The coal and liquor industries possess solid growth resilience; aluminum smelting exhibits a V-shaped recovery; and the steel sector gradually bottoms out and rebounds. The complete modeling procedure balances prediction accuracy and economic-mechanism interpretability, offering quantitative evidence for government anticipatory regulation, corporate operational planning, and optimized regional power allocation.

The closed-loop "one-model-per-industry" adaptive matching forecasting workflow proposed in this study can provide decision support for anticipatory economic regulation of local governments and risk early warning in the transformation of resource-based cities. Several applicable prerequisites should be noted when deploying this forecasting tool. First, it requires adequate multi-source monthly time-series datasets, including fine-grained industry-specific electricity consumption records, macro-monetary indicators, industrial product price indices, and official industrial statistical indicators. Second, the target city should show prominent pillar-industry-driven economic features, in which a small number of dominant industries dominate local economic fluctuations. Third, continuous time-series samples are necessary to support chronological rolling cross-validation and dynamic-correction-oriented parameter iteration.

It should be emphasized that the model parameters trained with Lüliang-city datasets cannot be directly applied to other resource-based cities. Differences in local industrial composition, feature correlation patterns, and multicollinearity levels will make simple parameter transplantation produce biased and unreliable prediction outputs. Nevertheless, the whole closed-loop "one-model-per-industry" technical framework possesses favorable transferability for other prefecture-level resource-based cities dominated by energy, raw-material and heavy-chemical industries. For new application scenarios, data preprocessing, feature screening and adaptive model re-selection should be re-performed according to local data quality and industrial features. For resource-based cities suffering from severe data missing or frequent industrial-structure changes, additional data reconstruction and cleaning are required before practical deployment.

This research bears the following limitations. First, this study is built upon monthly structured time-series datasets with a limited sample time span. Restricted by data availability, the model cannot well capture the long-run evolutionary patterns of industrial transformation. Consistent with our experimental observation, prediction performance will degrade when facing ultra-long-horizon forecasting tasks beyond the 3-6-month effective

prediction window. Second, the candidate model library only contains regularized statistical models and basic tree-ensemble algorithms, without introducing time-series deep-learning architectures such as LSTM and TCN. As reflected in our experimental results, the existing models exhibit inadequate detection capability for extreme business-cycle turning points, and local prediction bias can be observed at cyclical peaks and troughs. Third, the current dynamic correction module only implements coefficient-level parameter iteration updating. According to the empirical results across different industries, the coupling strength between electricity consumption and economic prosperity may undergo structural shifts in different development phases. However, the present model cannot automatically identify and adapt to such structural changes.

Further research can be advanced from three perspectives. Firstly, multi-source heterogeneous data will be fused, and text mining will be applied to quantify the disturbance effects brought by policies and external market shocks to enrich model input features. Secondly, a hybrid prediction framework integrating linear benchmarks and time-series deep learning will be established, together with time-varying regularization models, to enhance forecasting accuracy at cycle-switching and abrupt-change positions.

Acknowledgements

The authors gratefully acknowledge the financial support from the project “Research on the Key Technologies for Regional Economic Development Index Construction Based on Power Big Data” (Grant No. 5205J023000B), which also covered the article processing charge (APC) of this paper.

References

[1] Bale AS, Tiwari S, Raju KL, Pravesh P, Kishore P, Vinayak N. Environmental surveillance monitoring system in industries using Industrial Internet of Things. Design Eng. 2021;2021(5):1783-1790.

[2] Li YF, Wang BB, Luo H, Fang X. Multi-dimensional assessment of the developing situation of provincial electricity market considering the external economic factors. In: Proceedings of the 2016 China International Conference on Electricity Distribution (CICED); 2016 Aug 10-13; Xi'an, China. Piscataway (NJ): IEEE; 2016. p. 1-5.

[3] Xu S, Chen S, Sailike Y, et al. Industrial added value prediction based on multi-source data from the perspective of electricity consumption. Procedia Comput Sci. 2024; 242:1287-1293.

[4] Liu X, Cifuentes-Faura J, Yang X, Pan J. The green innovation effect of industrial robot applications: Evidence from Chinese manufacturing companies. Technol Forecast Soc Change. 2025; 210:123904.

[5] Haggag M, Abdelhady K, Guirguis M, Saudy M, El Dakhakhni W. Machine learning long-term electricity demand forecasting system for strategic energy investments. Sci Rep. 2026; 16:12471.

[6] Wang J, Tian Y, Jia F, Zhang S. Constructing the industrial prosperity index based on big data of enterprise electricity. IOP Conf Ser Earth Environ Sci. 2020; 571(1):012002.

[7] Wang T, Jia J, Xiao J, et al. Research on a new electric power market trading strategy based on power big data. In: Proceedings of the International Conference on Internet of Things and Machine Learning (IoTML 2022). Proc SPIE. Bellingham (WA): SPIE; 2023; 12640:95-99.

[8] Raj JD, Immanuel JS, Karthik P, Jayanthi M. Prediction of electric vehicle price in India using simple linear regression. In: Proceedings of the International Conference on Trends in Automotive Parts Systems and Applications; 2023 Dec 21; Coimbatore, India. Warrendale (PA): SAE International; 2024. Paper No. 2023-01-5162.

[9] Bayu K, Henidar DR, Hermastiandi F, Prasetya G, Nugraha A. Forecasting electrical energy loads at PT Krakatau Daya Electric using the linear regression method. J Mechatron Artif Intell. 2024; 1(1):45-54.

[10] Nieveen J, Zhang Y, Wendelken S, et al. Polynomial Kalman filter for myoelectric prosthetics using efficient kernel ridge regression. In: Proceedings of the 8th International IEEE/EMBS Conference on Neural Engineering (NER); 2017 May 25-28; Shanghai, China. Piscataway (NJ): IEEE; 2017. p. 432-435.

[11] Zhang Q, Drissi-Habti M. Electric cable insulator damage monitoring by Lasso regression. Machines. 2024;12(1):50.

[12] Sadou H, Hacib T, Le Bihan Y, Meyer O, Acikgoz H. Multiple non-linear regression for microwaves characterization of dielectric materials. Appl Phys A Mater Sci Process. 2025; 131(5):387.

[13] Ullah I, Liu K, Yamamoto T, Zahid M, Jamal A. Prediction of electric vehicle charging duration time using ensemble machine learning algorithm and Shapley additive explanations. Int J Energy Res. 2022; 46(11):15211-15230.

[14] Chauhan D, Yadav A, Mallipeddi R. EAEFA-R: Multiple learning-based ensemble artificial electric field algorithm for global optimization. Knowl Based Syst. 2025; 318:113453.

[15] Kim KH, Cho SM, Song KB. Monthly electric power sales forecasting algorithm using LSTM-XGBoost ensemble model. Trans Korean Inst Electr Eng. 2024; 73(5):766-772.

[16] Sim T, Ryu S, Lee D, Lee S, Chun CJ, Moon H. A machine learning ensemble framework based on a clustering algorithm for improving electric power consumption performance. Sci Rep. 2025; 15:40172.

[17] Melon MMH, Raihan MZU, Maruf SMA, Hasan MN, Biswas MH, Alif HA, Mashrafi MJ, Masum M. Risk-informed flood management under climate variability: A hybrid temporal-ensemble artificial intelligence framework. Bull Atmos Sci Technol. 2026; 7:9.

[18] Li H, Long R, Chen H. Economic transition policies in Chinese resource-based cities: An overview of government efforts. Energy Policy. 2013; 55:251-260.

[19] Xie X, Li K, Liu Z, Ai H. Curse or blessing: How does natural resource dependence affect city-level economic development in China? Aust J Agric Resour Econ. 2021; 65(2):413-448.

[20] Li Q, Zeng FE, Liu S, Yang M, Xu F. The effects of China's sustainable development policy for resource-based cities on local industrial transformation. Resour Policy. 2021; 71:101940.

[21] Ni, Z., Yu, H., Liu, S., Li, J., Lin, W. Basisformer: Attention-based time series forecasting with learnable and interpretable basis. Advances in neural information processing systems, 2023; 36: 71222-71241.

[22] Lim B, Arık SÖ, Loeff N, Pfister T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. Int J Forecast. 2021; 37(4):1748-1764.

[23] Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, Zhang W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2021; 35(12):11106-11115.
[24] Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Series B Stat Methodol. 1996; 58(1):267-288.
[25] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: Data mining, inference, and prediction. 2nd ed. New York: Springer; 2009.
[26] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13-17; San Francisco, CA. New York: Association for Computing Machinery; 2016. p. 785-794.
[27] Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970; 12(1):55-67.
[28] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol. 2005; 67(2):301-320.
[29] Breiman L. Random forests. Mach Learn. 2001; 45(1):5-32.
[30] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001; 29(5):1189-1232.