

Coupling of High-Skilled Talent Cultivation in the Manufacturing Industry and Regional Economic Performance in Vocational Education by Integrating Machine Learning and DEA Models

Yankun Wu, Lei Fu

Sichuan University of Science & Engineering, Zigong 643002, Sichuan, China

Corresponding author: Lei Fu (e-mail: 1244203167@qq.com)

ABSTRACT This study develops a two-stage analytical framework integrating super-efficiency Slacks-Based Measure Data Envelopment Analysis (SBM-DEA), extreme gradient boosting (XGBoost), and SHapley Additive exPlanations (SHAP) to investigate the coupling relationship between high-skilled talent cultivation in vocational education and regional economic performance in manufacturing industries. The super-efficiency SBM-DEA model is employed to quantify regional talent–economy coupling efficiency, while XGBoost captures nonlinear interactions among educational, industrial, and innovation-related factors. SHAP-based interpretability analysis is further conducted to identify the dominant drivers and their marginal effects. Experimental results show that the proposed framework achieves a prediction goodness-of-fit (R^2) of 0.91 and reveals the industrial upgrading index and dual-qualified teacher ratio as the most influential factors, with mean absolute SHAP values of 0.128 and 0.095, respectively. Dependency analysis further identifies a critical industrial upgrading threshold near 2.8, beyond which marginal benefits gradually diminish. The findings highlight the importance of coordinated talent development in supporting intelligent manufacturing, industrial communication systems, and digital production infrastructures, where the integration of skilled technical personnel, information technologies, and industrial innovation is essential for sustainable regional development. This study provides a data-driven engineering methodology for evaluating and optimizing the collaborative evolution of vocational education and manufacturing economies.

INDEX TERMS high-skilled talent cultivation, SBM-DEA, industrial communication systems, SHAP interpretability

I. INTRODUCTION

HIGH-SKILLED individuals are an essential support for the transformation and upgrading of manufacturing industries, serving as the core talent pipeline necessary for modernizing and enhancing global competitiveness. The quality of training for such individuals and the performance of regional economies have become vital factors influencing national industrial competitiveness. As a primary channel for supplying highly skilled personnel, the coupling efficiency between vocational education and the regional economic system directly determines the efficiency of human resource allocation and industrial development potential [1-3]. Accurately revealing the intrinsic mechanism of interactions of the two is urgent practical significance for adjusting optimal use of educational resources and achieving high-quality development of regional economies.

Current studies are confronted with various difficulties in quantifying the coupling relationship between vocational education and the economy. Firstly, it is difficult to characterize the system, as its endogenous mechanisms are relatively complex and include multidimensional variables such as educational input, education and training processes, demand from industry, and economic outputs. These multidimensional variables exhibit strong nonlinear interactions and a linear model often lacks in characterizing the system's dynamic relationship [4,5]. Secondly, the coupling state needs to be assessed in a way that reflects the relative efficiency of resource allocation. However, existing evaluation methods fail to adequately consider measurement biases caused by slack variables when dealing with the efficiency of multiple inputs and outputs, and cannot further distinguish effective decision-making units [6,7]. Most importantly, existing re-

search often stops at calculating efficiency values and lacks the ability to reverse-engineer influencing factors from the results, thus making it difficult to systematically reveal the key paths and structural bottlenecks driving the evolution of the coupling state [8,9].

To address these challenges, the academic community has undertaken a series of explorations. Some studies have used traditional DEA models to assess the efficiency of vocational education resource allocation, but their radial and angular measurement characteristics make it impossible to comprehensively measure the relaxation improvements of input and output, and the efficiency frontier contains a large number of parallel units [10,11]. Other studies have applied multiple regression or structural equation modeling to analyze influencing factors, but these methods have limited ability to capture complex nonlinear relationships and multicollinearity problems among variables [12,13]. In recent years, machine learning algorithms such as random forests have been applied to education economic forecasting, but their "black box" nature hinders the interpretability of model results at the mechanistic analysis level, making it difficult to provide clear decision-making insights [14,15]. Overall, none of the above methods have achieved a complete closed-loop from accurate efficiency measurement to in-depth analysis of coupling mechanisms, nor can they support precise regulation of the regional talent-economic coordinated development path. To address the significant shortcomings of existing methods in deconstructing the coupling mechanism of the vocational education-economic complex system, this study innovatively constructs a two-stage analytical framework. The first stage employs a non-radial, non-angular, super-efficiency SBM-DEA model to accurately measure the coupling efficiency of the vocational education-economic system in each province. This model effectively solves the problems of handling slack variables and ranking effective units in traditional DEA. The second stage uses the DEA efficiency value as the dependent variable, integrates multi-dimensional system features to train an extreme gradient boosting tree model, and further employs a game theory-based SHAP interpretability framework to perform attribution analysis on the prediction results of the optimal machine learning model, quantifying the marginal contribution of each feature to the coupling efficiency and its direction of action. Not only does this method provide precise evaluation of the coupling state of the system, but it also penetrates the causal logic behind the efficiency values, illuminating the "black box" that is driving coupling evolution. It creatively uncovers the key factors and nonlinear action channels that drive coupling evolution, providing a new methodological perspective to understand the deep-rooted laws governing the interaction of talent supply and regional economy.

II. COUPLING EFFICIENCY MEASUREMENT AND DRIVING MECHANISM FRAMEWORK

A. OVERALL FRAMEWORK DESIGN

To thoroughly dismantle the intricate coupling relationship between vocational education in manufacturing talent training and the regional economic performance, this study designs a stepwise two-stage analytical framework. Progressing from "state evaluation" to "mechanism analysis" in a linear pattern, the framework's goal is first to quantify the system's coupling efficiency level and then to further investigate its inherent driving factors and action paths. The first stage acts as its primary approach through Data Envelopment Analysis (DEA), focusing on the static assessment of a multi-dimensional inputoutput system's relative efficiency, to which it outputs quantifiable target variables for the analysis in the next stages. The second stage applies machine learning and interpretability techniques to reveal the implicit nonlinear laws and causal contributions between influencing factors and coupling efficiency. The two stages are closely linked through data flow, forming a complete analytical loop from describing phenomena to revealing mechanisms. Its overall architecture and data flow are shown in Figure 1.

Figure 1 illustrates the overall technical roadmap of this study. The process begins with the preprocessing of the raw data and the construction of the indicator system. In the measurement phase, the processed input-output data is fed into the super-efficiency SBM-DEA model to calculate the evaluation value characterizing the system's coupling efficiency. In the analysis phase, this efficiency value, together with the expanded environmental feature variables, constitutes a new dataset used to train the XGBoost (eXtreme Gradient Boosting) machine learning model. Finally, the SHAP interpretability technique is used to analyze the trained optimal model, outputting the marginal contribution of each feature variable to the coupling efficiency, thus completing the transformation from data to insights. This framework ensures synergy between methods, making efficiency evaluation and mechanism mining an organic whole.

B. COUPLING EFFICIENCY MEASUREMENT BASED ON SUPER-EFFICIENCY SBM-DEA

After completing the overall design of the analytical framework, it is necessary to accurately quantify the coupling efficiency of the vocational education-economic system. Traditional DEA models may produce biases when measuring slack variables and cannot further distinguish effective units. To address these limitations, this study adopts a non-radial, non-angular super-efficiency SBM-DEA model to measure coupling efficiency. This model directly handles the slack improvements of inputs and outputs and can obtain efficiency values greater than 1, thereby achieving a sufficient ranking of effective decision-making units [16,17]. It is assumed that there are n decision-making units, each containing m inputs, r_1 expected outputs, and r_2 unexpected outputs. For the decision-making unit DMU0 (Decision Making Unit 0) to be evaluated, its super-efficiency SBM

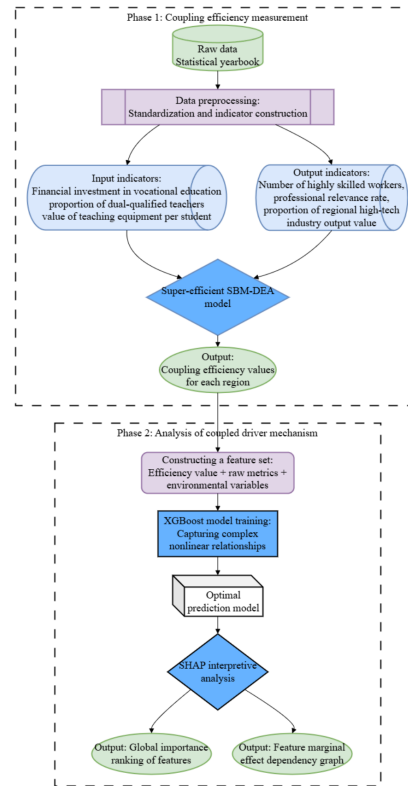


FIGURE 1. Two-stage coupling analysis framework diagram

model can be expressed as the following fractional programming form:

$$\min \rho = \frac{1 + \frac{1}{m} \sum_{i=1}^m s_i^- / x_{i0}}{1 - \frac{1}{r_1 + r_2} \left(\sum_{q=1}^{r_1} s_q^+ / y_{q0}^+ + \sum_{p=1}^{r_2} s_p^- / y_{p0}^- \right)} \quad (1)$$

$$\text{s.t.} \quad \sum_{j=1}^n \lambda_j x_{ij} - s_i^- \leq x_{i0}, \quad i = 1, \dots, m$$

$$\sum_{j=1}^n \lambda_j y_{qj}^+ + s_q^+ \geq y_{q0}^+, \quad q = 1, \dots, r_1 \quad (2)$$

$$\sum_{j=1}^n \lambda_j y_{pj}^- - s_p^- \leq y_{p0}^-, \quad p = 1, \dots, r_2$$

$$1 - \frac{1}{r_1 + r_2} \left(\sum_{q=1}^{r_1} \frac{s_q^+}{y_{q0}^+} + \sum_{p=1}^{r_2} \frac{s_p^-}{y_{p0}^-} \right) > 0$$

$$\lambda_j, s_i^-, s_q^+, s_p^- \geq 0 \quad (3)$$

In Formulas (1), (2), and (3), ρ represents the super-efficiency value of the DMU₀ to be determined, which theoretically can exceed 1. x_{i0} represents the i th input quantity of DMU₀, and y_{q0}^+ and y_{p0}^- represent its expected output quantity of the q th term and its unexpected output quantity of the p th term, respectively. s_i^- , s_q^+ , and s_p^- are slack variables corresponding to insufficient input, insufficient expected output, and excessive unexpected output, respectively. λ_j is the connecting weight variable used to construct the frontier.

ρ comprehensively reflects the average improvement ratio of input redundancy and output insufficiency; the lower the value, the worse the efficiency. This model calculates super-efficiency by excluding DMU₀ from the reference set and

constraining the distance between its linear combination and the frontier [18-19]. By using the Charnes-Cooper transformation, this fractional programming problem can be rewritten as a linear programming problem, and finally the relative efficiency evaluation score of each decision unit with respect to the optimal production frontier can be obtained [20-21]. This evaluation score can be interpreted as a quantitative measure of the coupling strength, representing the target variable for the analytical mechanism driving the second stage of the model.

The input and output indicators in Table 1 are selected based on their theoretical relevance to vocational education–economic coupling and established usage in DEA-based educational efficiency studies. Inputs reflect key educational resources, while expected outputs capture talent supply and economic outcomes. The model does not include separate unexpected outputs, as inefficiencies are captured via slack variables s_q^+ and s_p^- in Eq. (2), consistent with the non-radial SBM approach.

The DEA model incorporates both count-type and ratio-type outputs, as the super-efficiency SBM-DEA model is unit-invariant and operates on standardized data. Prior to modeling, all indicators are normalized to a [0,1] range, which eliminates scale differences and allows variables of different units to be combined without distorting efficiency measurement. This approach is consistent with established SBM-DEA practices where mixed indicator types are used to comprehensively capture multi-dimensional system performance.

C. COUPLED DRIVER MECHANISM BASED ON XGBOOST-SHAP

After obtaining the coupling efficiency measurement results, it is necessary to further reveal the key driving factors affecting efficiency and their complex mechanisms. This study adopts a mechanistic analysis model that combines the extreme gradient boosting tree XGBoost with the SHAP interpretation framework. XGBoost is an ensemble learning algorithm based on gradient boosting decision trees. It effectively controls the model complexity through an additive training strategy and a regularization term, and has a strong ability to handle nonlinear relationships and feature interactions [22,23]. Given a dataset containing n samples, its learning objective function $L(\phi)$ is composed of the loss function l and the regularization term Ω , as shown in Formula (4):

$$L(\phi) = \sum_{i=1}^n l(\hat{y}_i, y_i) + \sum_{k=1}^K \Omega(f_k) \quad (4)$$

In Formula 4, $\hat{y}_i$ and y_i represent the predicted value and the true value of the i th sample, respectively; K is the total number of trees; f_k represents the structure of the k th tree. The regularization term is used to penalize the complexity of the tree [24-25]:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (5)$$

T represents the number of leaf nodes; w represents the leaf weight score; γ and λ represent the hyperparameters controlling the intensity of the penalty. This objective function keeps the model concise while fitting the data and improves its generalization ability. The model uses the coupling efficiency values of each province calculated in the previous stage as the prediction target, and uses the feature set constructed by the input, output, and environmental variables listed in Table 1 as input for training.

Table 1 systematically presents all the variables used in the driving factor analysis phase. The indicator system is divided into three dimensions: input, output, and environment. The input dimension focuses on the allocation of human, financial, and material resources in the vocational education system; the output dimension measures the direct results of talent supply and the ultimate manifestation of economic benefits; the environmental dimension covers external conditions such as regional macroeconomics and innovation. Table 1 clarifies the operational definitions and data sources of each indicator, providing a complete feature space definition for model construction and ensuring the transparency and reproducibility of the analytical basis.

The classification of "High-Tech Industry Output Value Proportion" as an environmental variable rather than a system output in the DEA stage is intentional. While it reflects economic outcomes, it is treated as an exogenous contextual factor influencing the vocational education system's coupling performance, not as a direct output of

educational resource conversion. This distinction aligns with the study's focus on measuring how effectively educational inputs translate into talent outputs (e.g., skilled graduates and job-matching), with broader economic outcomes positioned as moderating factors. In the subsequent XGBoost-SHAP stage, all variables—including environmental ones—are treated as features to analyze their collective influence on coupling efficiency.

After obtaining the optimal XGBoost prediction model through training, the SHAP framework is used for post-hoc interpretation. SHAP is based on the Shapley value in cooperative game theory, providing a consistent and fair allocation of the contribution of each feature in each predicted sample [26,27]. For the predicted value of the i -th sample, its SHAP additive interpretation model is shown in Formula (6):

$$\hat{f}(x_i) = \phi_0 + \sum_{j=1}^M \phi_j^i \quad (6)$$

In Formula 6, $\hat{f}(x_i)$ represents the model's predicted value for the sample i ; ϕ_0 represents the baseline value of all sample predictions; M is the total number of features; ϕ_j^i is the SHAP value of the j th feature on that sample. Its absolute value characterizes the magnitude of the feature's influence, and the sign indicates the direction of the influence. By calculating the SHAP values of all samples, the global importance ranking of features can be obtained, and a SHAP dependency graph can be further plotted to reveal potential non-monotonic or interactive relationships between individual features and the predicted output. This method transforms XGBoost's high predictive performance into interpretable decision insights, precisely quantifying the marginal contribution of each driving factor to coupling efficiency.

III. EXPERIMENTAL DESIGN AND DATA DESCRIPTION

A. DATA SOURCES AND PREPROCESSING

This study uses 30 provincial-level administrative regions in China as decision-making units and selects panel data from 2015 to 2022 as the observation sample. The original data is all from publicly published authoritative statistical yearbooks and official database platforms. To ensure the completeness and consistency of the data, linear interpolation is used to fill in the missing data in the initial collection. For indicators whose statistical definitions change, retrospective adjustments are made based on the latest definitions [28,29]. Considering the significant differences in the dimensions and orders of magnitude of each indicator, the Min-Max standardization method is used to linearly transform all original data to the $[0,1]$ interval to eliminate the influence of dimensions and improve the convergence speed and stability of the subsequent machine learning model [30,31]. After the above preprocessing, the descriptive statistics of all variables are shown in Table 2.

TABLE 1. Evaluation index system for coupled systems

Dimension	Indicator Name	Calculation Method	Unit	Data Source
Input Indicators	Vocational Education Fiscal Investment	Public financial budget expenditure in the vocational education sector	10 thousand CNY	China Educational Finance Statistical Yearbook
Input Indicators	Dual-Qualified Teacher Ratio	Number of dual-qualified full-time teachers / Total number of full-time teachers	%	China Education Statistical Yearbook
Input Indicators	Per Student Teaching Equipment Value	Total value of teaching and scientific research instrument equipment assets / Number of enrolled students	CNY/student	China Education Statistical Yearbook
Output Indicators	High-Skilled Talent Quantity	Number of individuals obtaining professional qualifications for Senior Worker, Technician, or Senior Technician	person	Statistical Yearbook of Human Resources and Social Security Affairs
Output Indicators	Graduate Job-Major Match Rate	Proportion of graduates employed in jobs related to their field of study	%	National Higher Vocational Colleges Talent Cultivation Work Status Data Collection and Management Platform
Output Indicators	High-Tech Industry Output Value Proportion	Output value of high-tech industries / Total industrial output value	%	China Statistical Yearbook on High Technology Industry
Output Indicators	Regional R&D Investment Intensity	R&D internal expenditure / Regional Gross Domestic Product	%	China Statistical Yearbook on Science and Technology
Environmental Variables	Industrial Structure Service-Oriented Index	Value-added of the tertiary sector / Value-added of the secondary sector (Note: This ratio reflects the degree of service-oriented structural shift, which is one dimension of industrial upgrading in the context of regional economic transformation.)	-	China Statistical Yearbook
Environmental Variables	Innovation Policy Support	Annual number of government documents issued related to technological innovation	document	Public data from provincial and municipal government gazettes

TABLE 2. Descriptive statistics

Variable categories	Variable name	Sample size	Mean	Standard deviation	Minimum value	Maximum value
Input	Vocational education fiscal investment	240	125.43	85.67	15.20	450.15
Input	Dual-Qualified teacher ratio	240	0.28	0.09	0.11	0.52
Input	Per student teaching equipment value	240	1.15	0.58	0.32	3.05
Output	High-Skilled talent quantity	240	12.54	9.88	1.05	55.10
Output	Graduate Job-Major match rate	240	0.71	0.12	0.45	0.93
Environment	High-Tech industry output value proportion	240	0.25	0.11	0.08	0.59
Environment	Regional R&D investment intensity	240	0.02	0.01	0.004	0.06
Environment	Industrial upgrading index	240	1.18	0.75	0.52	4.85
Environment	Government innovation policy support	240	8.35	5.22	1.00	25.00

Table 2 presents the distribution characteristics of each variable after preprocessing. The total number of observed samples is 240, covering the entire spatiotemporal range defined in the study. The data shows that the mean and standard deviation of each indicator reflect significant differences among provinces in vocational education investment, economic output, and innovation environment. For example, the standard deviation of the number of highly skilled personnel is close to the mean, indicating a highly uneven distribution of talent reserves across regions; the difference between the maximum and minimum values of R&D investment intensity exceeds tenfold, highlighting a huge gap in regional innovation capabilities. These descriptive statistical results provide a data foundation for subsequent model analysis, and the regional heterogeneity they reveal also confirms the necessity of analyzing the driving mechanisms.

B. COMPARISON MODEL AND EVALUATION INDICATORS

To verify the effectiveness of the framework constructed in this paper, a systematic comparative experiment is set up. In the coupling efficiency measurement stage, the classic radial CCR (Charnes-Cooper-Rhodes) model and the BCC (Banker-Charnes-Cooper) model considering variable scale returns are selected as baseline comparison models to study the performance of the models in handling slack variables and distinguishing effective units [32,33]. In the coupling-driven mechanism analysis stage, random forest and support vector machine regression are selected as the comparative baselines of the XGBoost model [34,35]. As an ensemble learning method, random forest has good anti-

overfitting ability by constructing multiple decision trees and integrating their prediction results. Support vector machine regression maps data to a high-dimensional feature space through kernel functions to find the optimal linear splitting hyperplane, which is suitable for small sample nonlinear problems. The above comparative models can test the performance of XGBoost in capturing the complex relationship of vocational education-economic system from the perspective of different algorithm principles.

Model performance is evaluated using two metrics: the coefficient of determination and the mean square error. The coefficient of determination R^2 measures the model's explanatory power for the variance of the target variable; a value closer to 1 indicates a higher goodness of fit. The MSE (mean square error) calculates the expected value of the squared difference between the predicted and actual values, directly reflecting prediction accuracy; a smaller MSE value indicates lower prediction error. These two metrics work together to create a holistic system for evaluating model performance. All models are built on the same training and test sets—with 30% of the total set used as the test set. The final hyperparameter specifications are shown in Table 3.

TABLE 3. Hyperparameter configuration of machine learning models

Model name	Hyperparameter name	Parameter value
XGBoost	Learning rate	0.1
XGBoost	Max depth	6
XGBoost	n_estimators	200
XGBoost	Subsample	0.8
Random Forest	n_estimators	300
Random Forest	Max features	'sqrt'
Random Forest	min_samples_split	5
SVR (Support Vector Regression)	kernel	'rbf'
SVR (Support Vector Regression)	C	1.0
SVR (Support Vector Regression)	gamma	'scale'

The primary machine learning models utilized in this study and their optimal hyperparameter values are described in Table 3. The optimized learning rate for XGBoost is 0.1 with 200 estimators to ensure effective trainings, while Random Forest employs a parameter of 300 trees to produce stable predictions. The Support Vector Machine regression is optimized with the radial basis function kernel to accommodate any nonlinear relationships. These combinations of hyperparameter values provide the optimal trade-off between model performance and computational intensity, and ensure there is appropriate assurance for replicating the experiments. The unified hyperparameter optimization process ensures fair comparison of all comparative models under optimal configurations, thereby guaranteeing the scientific validity and reliability of subsequent performance comparison results.

IV. RESULTS

A. SPATIOTEMPORAL DIFFERENTIATION CHARACTERISTICS OF COUPLING EFFICIENCY

To reveal the spatial distribution pattern of the coupling efficiency of the vocational education-economic system, this study first calculates the coupling efficiency value of each provincial-level administrative region based on the super-efficiency SBM-DEA model, and analyzes its geographical distribution pattern. The system identifies high-efficiency clusters and low-efficiency weak areas, and the spatial distribution characteristics of coupling efficiency in different regions are shown in Figure 2.

Figure 2 uses geographic coordinates and employs a dual encoding method of color gradient and point size to display the coupling efficiency values of each province. The horizontal axis represents the longitude range of 70°E to 135°E, and the vertical axis represents the latitude range of 15°N to 55°N. The point size is proportional to the efficiency value. Data shows a significant stepped distribution of efficiency values, with higher values in the eastern coastal areas and lower values in the western inland areas. Specifically, the efficiency values of Shanghai, Jiangsu, Beijing, and Guangdong are 1.30, 1.20, 1.25, and 1.22, respectively, all exceeding the production frontier. These regions have developed

industrial systems and well-established vocational education systems that create a positive interaction. In contrast, the efficiency ratings of Tibet, Qinghai, and Ningxia are merely 0.50, 0.55, and 0.61, respectively. Their relatively lagging industrial development and insufficient investment levels in vocational education resources profoundly restrict their system coupling efficiency. The efficiency values of central province regions, such as Henan and Hubei, are situated in the middle range of 0.89-1.05, which reflects the transitional characteristics of their transformation and development process. This spatial differentiation is essentially attributed to the disparities in their economic development levels, levels of industrial structure, and allocation of vocational education resources among different regions, ultimately forming a distribution pattern that degrades from eastern to western regions.

B. PERFORMANCE COMPARISON OF DRIVING MECHANISM AND COUPLING EFFICIENCY MEASUREMENT MODEL

After determining the spatial distribution of coupling efficiency, it is necessary to verify the reliability of the model selected in the driving factor analysis stage. This study selects three algorithms: extreme gradient boosting tree, random forest, and support vector machine regression. Based on the same training and test sets, the prediction performance is compared. The optimal mechanism analysis model is determined through systematic analysis of multi-dimensional evaluation indicators. The comprehensive performance comparison of each model under the standardized evaluation framework is shown in Figure 3.

Figure 3 presents the standardized performance scores of the three models on five metrics. Each axis represents the normalized values of goodness-of-fit R^2 , mean square error (MSE), mean absolute error (MAE), maximum error, and training time, respectively. All original metrics are normalized using a minimum-maximum standard, uniformly mapped to the 0-1 range for cross-dimensional comparison; values closer to 1 indicate better performance. The XGBoost model excels in prediction accuracy, achieving an R^2 value of 0.91, significantly higher than Random Forest's 0.82 and Support Vector Machine's 0.68. This is attributed to its gradient boosting mechanism, which uses sequential construction of weak learners to gradually correct residuals, thus more accurately capturing complex nonlinear relationships between variables. In terms of error control, XGBoost's standardized MSE score is 0.933, and its MAE is 0.9, both superior to the comparison models. This is due to its regularization term effectively controlling model complexity and mitigating overfitting. While random forests achieve faster training speeds through parallel ensemble strategies, their averaging mechanism based on independent tree structures limits further improvements in prediction accuracy. Support vector machines (SVMs) are robust in small-sample, high-dimensional spaces; however, their kernel function parameters are sensitive to data distribution,

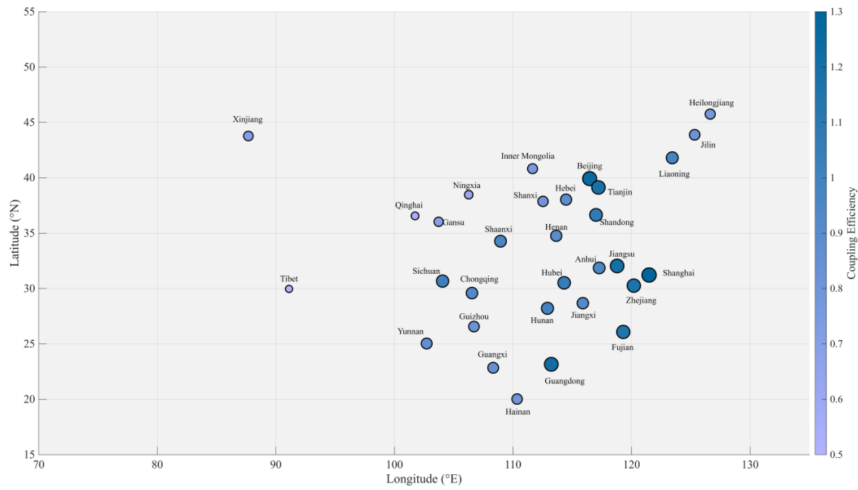


FIGURE 2. Spatial distribution of coupling efficiency

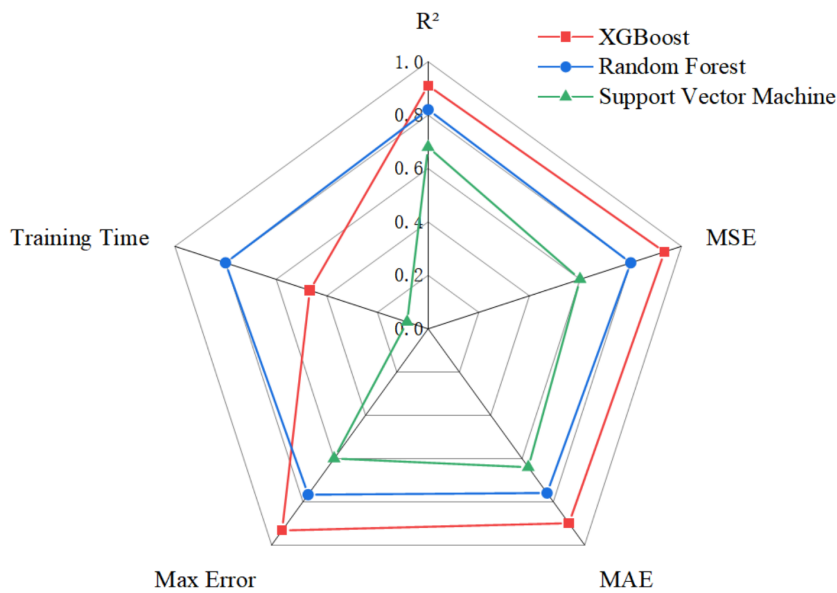


FIGURE 3. Performance comparison of drive mechanism models

resulting in relatively limited generalization ability in large-scale panel data. Overall, XGBoost’s advantages in prediction accuracy and stability provide a reliable guarantee for its subsequent role in driving mechanism analysis. To verify the methodological advantages of the super-efficiency SBM-DEA model in the coupling efficiency measurement stage, this study systematically compares its evaluation results with those of the traditional CCR and BCC models, and analyzes its technical improvements from multiple dimensions such as efficiency distribution, effective unit discrimination, and model discriminative power. The comparison of the quantitative measurement results of different models is shown in Table 4.

TABLE 4. Comparison of efficiency measurement results for different DEA models

Model name	Average efficiency value	Number of effective cells (efficiency value = 1)	Number of effective cells (efficiency value > 1)	Efficiency value standard deviation
CCR	0.824	6	0	0.185
BCC	0.881	9	0	0.162
SBM-DEA	0.895	0	11	0.201

As shown in Table 4, the average efficiency of the super-efficiency SBM-DEA model is 0.895, slightly higher than the 0.881 of the BCC model and the 0.824 of the CCR model. Its standard deviation of efficiency increases to 0.201, indicating that this model has higher discriminative sensitivity to differences in the efficiency of decision-

making units. Limited by their radial measurement principle and inability to handle slack variables, the CCR and BCC models identify 6 and 9 efficient units with an efficiency value of 1, respectively. These units are considered equally optimal under the traditional framework and cannot be further distinguished. The super-efficiency SBM-DEA model used in this paper, by applying non-radial and non-angular relaxation measurements and excluding the units to be evaluated from the reference set, successfully resolves all the aforementioned efficient units into super-efficiency units with an efficiency value greater than 1. This result verifies the methodological advantage of the super-efficiency SBM-DEA model in avoiding the "failure of efficient unit ranking" problem in traditional models and provides more discriminative efficiency input data for subsequent analysis of driving mechanisms.

C. CONTRIBUTION OF KEY DRIVING FACTORS

Based on the confirmation that the XGBoost model has the best prediction performance, this study further adopts the SHAP interpretability framework to quantify the marginal contribution of each feature to the coupling efficiency, aiming to identify the core elements driving the coordinated development of the system and clarify their relative importance. The global importance ranking of the key driving factors based on the mean absolute SHAP value is shown in Figure 4.

Figure 4 demonstrates the standardized contribution of each driving factor. The horizontal axis represents the mean absolute SHAP value, which is a direct representation of the strength of the influence of a feature on the model output. The vertical axis lists all feature variables in descending order of contribution. Data analysis reveals that industrial upgrading index receives a SHAP value of 0.128, contributing more than any of the other factors. This is a result of the significant additional reliance on high skilled personnel during the industrial structure transformation towards the high-tech service industries, which in turn drives the coupling demand between vocational education and economic development. The proportion of dual-qualified teachers ranks second with a contribution value of 0.095. This confirms the valuable bridging role of a teaching staff with competencies spanning theoretical teaching and practical guidance for connecting industrial demand and educational supply. Research & Development (R&D) investment intensity and the proportion of high-tech industry output value follows with contributions values of 0.087 and 0.078, respectively. This suggests that the innovation ecosystem of a given region indirectly influences through some level of technology-spillover and subsequent industrial upgrading with regards to the scale and quality of skilled personnel training. The SHAP value of government innovation policy support is 0.065, reflecting the coordinating function of policy guidance in integrating educational resources and industrial planning. In contrast, the contributions of purely resource-based inputs, such as per-student teaching equip-

ment resources and fiscal investment, are 0.054 and 0.048, respectively, indicating that improvements in hardware conditions need to be coordinated with innovations in talent cultivation models to achieve maximum effectiveness. The differences in the contribution of various driving factors reveal that system coupling is constrained by both macro-industrial structure and micro-educational quality, and the coordinated advancement of industry demand-driven and education supply-side reforms constitutes the key path to system optimization.

D. NONLINEAR EFFECTS OF DRIVING FACTORS

Based on the identification of key driving factors, it is necessary to further reveal the deep-seated mechanism of their interaction with coupling efficiency. This study focuses on the industrial upgrading index, which has the highest contribution, and explores the changing law of its marginal effect through SHAP dependency analysis. It analyzes the differentiated impact path of this factor on system coupling at different development stages, and thus obtains the nonlinear relationship between the industrial upgrading index and coupling efficiency, as shown in Figure 5.

Figure 5 uses the industrial upgrading index as the horizontal axis and the SHAP value as the vertical axis, illustrating the nonlinear relationship between the two through scatter plots and multinomial fitting curves. Data analysis shows that when the industrial upgrading index is below 1.5, the SHAP value rapidly increases from -0.02 to 0.14, indicating that the increased proportion of the tertiary sector relative to the secondary sector significantly improves human resource allocation efficiency during this stage, and the diversified demand for skilled personnel from the service industry effectively activates the vocational education supply system. When the index is between 1.5 and 2.8, the SHAP value slowly rises from 0.14 to a peak of 0.16, reflecting the diminishing marginal benefits of industrial structure optimization. At this point, industrial upgrading gradually prioritizes the quality of talent over quantity. When the index exceeds 2.8, the SHAP value drops to 0.12, indicating that excessive service-oriented development may weaken the foundation of the real economy, and the structural imbalance between high-tech manufacturing and producer services weakens the industrial foundation for industry-education integration. An industrial upgrading index of 2.8 constitutes a critical threshold for system coupling. Beyond this value, the advancement of the industrial structure no longer necessarily leads to a continuous improvement in coupling efficiency. Maintaining a reasonable industrial structure becomes a prerequisite for ensuring the coordinated development of the system.

V. CONCLUSION

This study constructs a framework for coupling measurement and mechanistic analysis of complex vocational education-economic systems by integrating the super-efficiency SBM-DEA and XGBoost-SHAP models. The

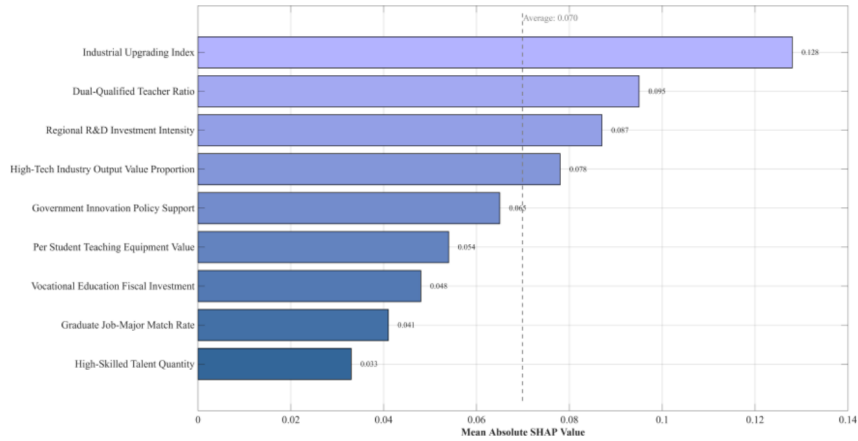


FIGURE 4. Global importance of driving factors based on SHAP values

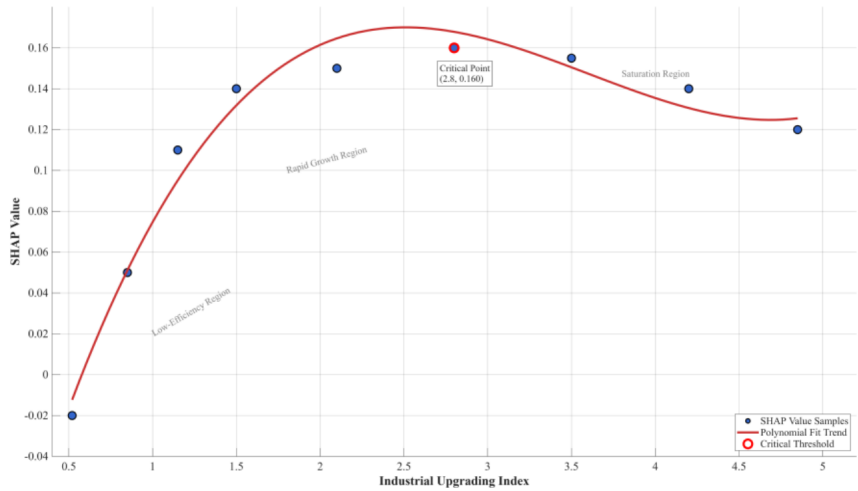


FIGURE 5. Nonlinear impact of industrial upgrading index on coupling efficiency

method first precisely quantifies the coupling efficiency of each province, revealing a spatial distribution with high efficiency in the eastern coastal areas and low efficiency in the western inland areas. Shanghai and Tibet have efficiency values of 1.30 and 0.50, respectively. Then, a machine learning model achieves a standardized R^2 prediction goodness of fit of 0.91, and based on the SHAP value, the industrial structure service-orientation index is identified as having the highest contribution, reflecting the significant role of service sector expansion in shaping the vocational education-economic coupling within the current stage of regional development, with a SHAP value of 0.128. Dependency analysis further reveals that this factor has a critical effect point of 2.8, beyond which marginal benefits diminish. The study’s results utilize interpretable AI at the methodological level to facilitate regional complex systems analysis, offering practical empirical evidence and precise intervention targets to optimize vocational educational resources and industrial synergy.

AUTHOR CONTRIBUTIONS

Conceptualization –Yankun Wu and Lei Fu; methodology – Yankun Wu and Lei Fu; investigation – Yankun Wu and Lei Fu; writing–original draft preparation – Yankun Wu and Lei Fu. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] B. Huang and Z. Z. Zhang, “Synergistic development of vocational education and regional economy in the era of high-quality growth,” *International Educational Research*, vol. 8, no. 3, pp. 66–p66, 2025, doi: 10.30560/ier.v8n3p66.

[2] H. Zhao, N. Stoyanets, L. Cui, and G. Li, “Strategy of vocational education adapting to social and economic development,” *Journal of*

- Innovations and Sustainability, vol. 6, no. 1, pp. 03-03, 2022, doi: 10.51599/is.2022.06.01.03.
- [3] L. Liu, S. Si, and J. Li, "Research on the effect of regional talent allocation on high-quality economic development-Based on the perspective of innovation-driven growth," Sustainability, vol. 15, no. 7, pp. 6315, 2023, doi: 10.3390/su15076315.
- [4] X. Gong, "Performance evaluation of industry-education integration in higher education from the perspective of coupling coordination-an empirical study based on Chongqing," PLOS ONE, vol. 19, no. 9, Art. no. e0308572, 2024, doi: 10.1371/journal.pone.0308572.
- [5] F. Gao, F. Liu, and S. H. Aldulaimi, "Scientific Model of Vocational Education Teaching Method in Differential Nonlinearity," Applied Mathematics and Nonlinear Sciences, vol. 8, no. 1, pp. 1397-1404, 2023, doi: 10.2478/amns.2022.2.0128.
- [6] V. J. Bolós, R. Benítez, and V. Coll-Serrano, "Continuous models combining slacks-based measures of efficiency and super-efficiency," Central European Journal of Operations Research, vol. 31, no. 2, pp. 363-391, 2023, doi: 10.1007/s10100-022-00813-5.
- [7] S. Zhou and Z. Xu, "Energy efficiency assessment of RCEP member states: A three-stage slack based measurement DEA with undesirable outputs," Energy, vol. 253, Art. no. 124170, 2022, doi: 10.1016/j.energy.2022.124170.
- [8] X. Q. Li, "A Review of Research on the Coordinated Development of New Quality Productivity and Vocational Education: Connotation, Logic, and Path," Journal of Exploration of Vocational Education, vol. 4, no. 3, pp. 58-83, 2025, doi: 10.63650/jeve.v4i3.62.
- [9] Y. Ma and Z. Cui, "Influencing factors and improvement paths of manufacturing innovation performance: Configuration analysis based on TOE framework," Plos one, vol. 18, no. 11, Art. no. e0294630, 2023, doi: 10.1371/journal.pone.0294630.
- [10] L. Long, "Eco-efficiency and effectiveness evaluation toward sustainable urban development in China: A super-efficiency SBM-DEA with undesirable outputs," Environment, Development and Sustainability, vol. 23, no. 10, pp. 14982-14997, 2021, doi: 10.1007/s10668-021-01282-7.
- [11] Y. Sun, D. Wang, F. Yang, and S. Ang, "Efficiency evaluation of higher education systems in China: A double frontier parallel DEA model," Computers & Industrial Engineering, vol. 176, Art. no. 108979, 2023, doi: 10.1016/j.cie.2023.108979.
- [12] S. Grønneberg and J. P. Irmer, "Non-parametric regression among factor scores: Motivation and diagnostics for nonlinear structural equation models," Psychometrika, vol. 89, no. 3, pp. 822-850, 2024, doi: 10.1007/s11336-024-09959-4.
- [13] J. Y. L. Chan, S. M. H. Leow, K. T. Bea, W. K. Cheng, S. W. Phoong, Z. W. Hong, et al., "Mitigating the multicollinearity problem and its machine learning approach: a review," Mathematics, vol. 10, no. 8, pp. 1283, 2022, doi: 10.3390/math10081283.
- [14] M. Atzmueller, J. Fürnkranz, T. Kliegr, and U. Schmid, "Explainable and interpretable machine learning and data mining," Data Mining and Knowledge Discovery, vol. 38, no. 5, pp. 2571-2595, 2024, doi: 10.1007/s10618-024-01041-y.
- [15] U. Tariq, I. Ahmed, M. A. Khan, and A. K. Bashir, "Deep learning for economic transformation: a parametric review," Indonesian journal of electrical engineering and computer science, vol. 35, no. 1, pp. 520-541, 2024, doi: 10.11591/ijeecs.v35.i1.pp520-541.
- [16] H. S. Lee, "An integrated model for SBM and Super-SBM DEA models," Journal of the Operational Research Society, vol. 72, no. 5, pp. 1174-1182, 2021, doi: 10.1080/01605682.2020.1755900.
- [17] Q. Chen, T. Wu, and Y. Yin, "Evaluation and Analysis of Sustainable Development Efficiency of Port Cities in China Using the Super-Efficiency SBM-DEA Model," Polish Journal of Environmental Studies, vol. 33, no. 5, pp. 5037-5050, 2024, doi: 10.15244/pjoes/183085.
- [18] S. Yang, G. Zhao, and F. Li, "The Efficiency Evaluation of DEA Model Incorporating Improved Possibility Theory," Mathematics, vol. 12, no. 19, pp. 3116, 2024, doi: 10.3390/math12193116.
- [19] H. Liao, G. Hao, R. Yasmeen, and W. U. H. Shah, "Evaluation of educational resource utilization efficiency, regional technological heterogeneity, and total factor productivity change in 35 European countries," PLOS ONE, vol. 19, no. 1, Art. no. e0295979, 2024, doi: 10.1371/journal.pone.0295979.
- [20] Y. Ning, Y. Zhang, and G. Wang, "An improved DEA prospect cross-efficiency evaluation method and its application in fund performance analysis," Mathematics, vol. 11, no. 3, pp. 585, 2023, doi: 10.3390/math11030585.
- [21] M. Pokushko, A. Stupina, I. Medina-Bulo, A. Stupin, V. Stasiuk, and R. Pokushko, "THE CHARNES, COOPER AND RHODES MODEL AND ITS APPLICATION," Facta Universitatis, Series: Mathematics and Informatics, vol. 40, no. 1, pp. 197-207, 2025, doi: 10.22190/FUMI240306015P.
- [22] Y. Xia, S. Jiang, L. Meng, and X. Ju, "XGBoost-B-GHM: An Ensemble Model with Feature Selection and GHM Loss Function Optimization for Credit Scoring," Systems, vol. 12, no. 7, pp. 254, 2024, doi: 10.3390/systems12070254.
- [23] J. Dong, Y. Chen, B. Yao, X. Zhang, and N. Zeng, "A neural network boosting regression model based on XGBoost," Applied Soft Computing, vol. 125, Art. no. 109067, 2022, doi: 10.1016/j.asoc.2022.109067.
- [24] G. Zhang and A. Gionis, "Regularized impurity reduction: accurate decision trees with complexity guarantees," Data mining and knowledge discovery, vol. 37, no. 1, pp. 434-475, 2023, doi: 10.1007/s10618-022-00884-7.
- [25] M. Wu, S. Parbhoo, M. C. Hughes, V. Roth, and F. Doshi-Velez, "Optimizing for interpretability in deep neural networks with tree regularization," Journal of Artificial Intelligence Research, vol. 72, pp. 1-37, 2021, doi: 10.1613/jair.1.12558.
- [26] L. H. B. Olsen, I. K. Glad, M. Jullum, and K. Aas, "A comparative study of methods for estimating model-agnostic Shapley value explanations," Data Mining and Knowledge Discovery, vol. 38, no. 4, pp. 1782-1829, 2024, doi: 10.1007/s10618-024-01016-z.
- [27] H. Wang, Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, "Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods," Journal of Big Data, vol. 11, no. 1, pp. 44, 2024, doi: 10.1186/s40537-024-00905-w.
- [28] T. Niedzielski and M. Halicki, "Improving linear interpolation of missing hydrological data by applying integrated autoregressive models," Water Resources Management, vol. 37, no. 14, pp. 5707-5724, 2023, doi: 10.1007/s11269-023-03625-7.
- [29] S. M. Ribeiro and C. L. de Castro, "Missing data in time series: A review of imputation methods and case study," Learning and Nonlinear Models, vol. 20, no. 1, pp. 31-46, 2022, doi: 10.21528/lnlm-vol20-no1-art3.
- [30] P. J. M. Ali, "Investigating the Impact of min-max data normalization on the regression performance of Knearest neighbor with different similarity measurements," ARO-The Scientific Journal of Koya University, vol. 10, no. 1, pp. 85-91, 2022, doi: 10.14500/aro.10955.
- [31] H. A. Ahmed, P. J. M. Ali, A. K. Faeq, and S. M. Abdullah, "An investigation on disparity responds of machine learning algorithms to data normalization method," aro-the scientific journal of koya university, vol. 10, no. 2, pp. 29-37, 2022, doi: 10.14500/aro.10970.
- [32] C. Yongsheng, Z. Deyuan, and W. Xia, "Agricultural Total Factor Productivity Based on Farmers' Perspective: An Example of CCR, BCC, SBM and Technology Optimization Malmquist-Luenberger Index," Journal of Resources and Ecology, vol. 15, no. 2, pp. 267-279, 2024, doi: 10.5814/j.issn.1674-764x.2024.02.003.
- [33] D. Maradin, B. Olgić Draženović, and S. Čegar, "The efficiency of offshore wind energy companies in the European countries: A DEA approach," Energies, vol. 16, no. 9, pp. 3709, 2023, doi: 10.3390/en16093709.
- [34] S. Fatima, A. Hussain, S. B. Amir, S. H. Ahmed, and S. Aslam, "Xgboost and random forest algorithms: an in depth analysis," Pakistan journal of scientific research, vol. 3, no. 1, pp. 26-31, 2023, doi: 10.57041/pjosr.v3i1.946.
- [35] J. Sun, W. Dang, F. Wang, H. Nie, X. Wei, P. Li, et al., "Prediction of TOC content in organic-rich shale using machine learning algorithms: Comparative study of random forest, support vector machine, and XGBoost," Energies, vol. 16, no. 10, pp. 4159, 2023, doi: 10.3390/en16104159.