

Article

A CatBoost-Based Prediction Framework for Logistics Industry Prosperity Index to Support Sustainable Decision-Making: An Empirical Study from China

Yule Liu ¹, Qiong Li ¹, Changxi Ma ²  and Xuecai Xu ^{3,*} 

¹ School of International Education, Lanzhou University of Finance and Economics, Lanzhou 730101, China; lyuleleyu@163.com (Y.L.); liq@lzufe.edu.cn (Q.L.)

² School of Traffic and Transportation, Lanzhou Jiaotong University, Lanzhou 730070, China; machangxi@mail.lzjtu.cn

³ School of Civil Engineering and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

* Correspondence: xuecai_xu@hust.edu.cn

Abstract

The logistics industry serves as a vital engine for economic growth, yet its prosperity is influenced by complex and dynamic factors. Accurate forecasting of the Logistics Industry Prosperity Index (LPI) is essential for optimizing resource allocation, enhancing operational efficiency, and mitigating potential risks, thereby supporting sustainable development and digital transformation. However, existing forecasting models often struggle with flexibility, interpretability, and handling complex nonlinear data. To address these challenges, this study proposes an innovative prediction framework based on the CatBoost algorithm and constructs an end-to-end prediction process integrating Bayesian optimization for hyper-parameter tuning and a multidimensional evaluation system. The proposed framework is validated using a unique multidimensional dataset comprising 12 key indicators from Lanzhou City, China, spanning January 2022 to March 2025. Empirical results demonstrate that the CatBoost model significantly outperforms traditional and other machine learning approaches, including ARIMA, SVM, and XGBoost, achieving an R^2 of 0.963 and a MAPE of 0.001%. From a theoretical perspective, this study enriches logistics prosperity forecasting and early-warning methodologies by introducing a highly accurate and robust learning-based framework. From a practical perspective, it provides governments and logistics enterprises with a reliable, data-driven tool for real-time decision support, strategic planning, and proactive risk management.

Keywords: logistics industry prosperity index (LPI); CatBoost; machine learning; sustainable decision-making; forecasting



Academic Editor: Giada La Scalia

Received: 19 January 2026

Revised: 19 February 2026

Accepted: 22 February 2026

Published: 24 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

China's economy now stands at a pivotal turning point, transitioning decisively from rapid expansion towards high-quality development. Driving industrial structure optimization and boosting economic efficiency has consequently surged to the forefront for governments at all levels, enterprises, and the public. Against this profound macroeconomic backdrop, the logistics industry—often hailed as the “goldmine at the feet of enterprises”—has powerfully emerged as a vital new growth engine. By masterfully integrating resources, optimizing allocation, and creating substantial value, it unlocks

significant profits for major corporations and serves as a foundational pillar supporting the national economy's stability.

The Logistics Industry Prosperity Index (LPI), serving as the sector's essential "barometer," meticulously tracks 12 critical sub-indices spanning total business volume, workforce size, and new orders. As the industry becomes increasingly complex and susceptible to global fluctuations, the ability to accurately forecast the LPI is of paramount importance. Reliable forecasting provides a forward-looking perspective that enables stakeholders to anticipate market shifts rather than merely reacting to them. It acts as a crucial informational tool that bridges the gap between current operational data and future strategic planning, ensuring that the industry remains resilient in the face of economic volatility.

Research into LPI forecasting methods carries immense practical significance for steering government macroeconomic policy, optimizing resource allocation, and proactively mitigating potential risks. For government bodies, precise predictions are necessary to formulate supportive regulations and infrastructure plans that align with future demand. For enterprises, forecasting is indispensable for inventory management, capital investment, and maintaining competitive advantages. Without accurate predictive capabilities, the industry risks resource wastage and operational inefficiencies, which are detrimental to the goal of sustainable development. Therefore, developing robust forecasting mechanisms is not just an academic exercise but a fundamental necessity for the long-term health and sustainability of the entire logistics ecosystem.

In response to the problems of poor model flexibility and weak interpretability in current LPI prediction, this paper proposes an innovative method based on the CatBoost algorithm. While existing models often struggle with non-linear shocks and complex time-series data, this study leverages CatBoost's unique advantages—such as ordered boosting and symmetric tree structures—to alleviate gradient bias and overfitting. By constructing an end-to-end prediction process that integrates Bayesian optimization and a multi-dimensional evaluation system, we aim to provide a highly accurate and reliable decision support tool. This study aims to demonstrate that the proposed framework not only outperforms traditional methods but also provides the scientific rigor required to support the high-quality development and digital transformation of the modern logistics industry.

Research on forecasting within the logistics industry has evolved significantly over the past decade. Huang et al. presented a development prediction and dynamic analysis of the modern logistics industry in Henan Province by applying a group of grey predictive models and buffer operators. Their results indicated that the modern logistics industry in Henan Province experienced steady growth [1]. Extending this scope to the national level, Xu et al. focused on demand forecasting for the Chinese logistics industry, introducing a novel adaptive grey model to address time-critical prediction accuracy issues. Comparative analysis concluded that the proposed model outperformed other methods in short-term forecasting [2]. Similarly, Guo and Li forecasted carbon emissions in the regional logistics industry. By comparing the traditional GM(1,N) model, a GM(1,N) model with background value optimization, and a GM(1,N) model with expression optimization, they found that the latter exhibited superior performance [3]. In line with environmental objectives, Chen et al. predicted carbon emission levels in China's logistics industry using a PSO-SVR model to achieve the "dual carbon" goals. They employed grey relational analysis to screen influencing factors and utilized the particle swarm optimization (PSO) algorithm to optimize the penalty coefficient and kernel function parameters of the support vector regression (SVR) model. The study demonstrated that the proposed PSO-SVR model is effective for achieving "dual carbon" targets and upgrading the logistics industry [4]. Furthermore, in the context of the new era, Wen et al. predicted the development trend

of the “Internet Plus” logistics industry under the “Belt and Road” strategy, providing insights that adjusted strategic approaches and development directions [5].

Existing forecasting models in the Logistics Performance Index (LPI) field primarily employ two technical approaches: traditional statistics and machine learning. Regarding traditional statistical models, d’Aleo et al. utilized an explanatory linear regression framework to investigate the mediating role of the LPI in the relationship between the Global Competitiveness Index (GCI) and gross domestic product (GDP) across European countries from 2007 to 2014 [6]. Martí et al. proposed a method based on data envelopment analysis (DEA) to construct a comprehensive composite logistics performance indicator (DEA-LPI), enabling benchmarking and comparative evaluation of logistics performance among countries within the LPI system [7]. Vishal et al. integrated expert opinions with the fuzzy ISM–MICMAC approach to identify logistics performance indicators critical to economic development, thereby establishing a logistics-oriented framework for understanding and forecasting future economic performance [8]. Hanif et al. applied a vector autoregression (VAR) model to examine dynamic time-series relationships between modern logistics development and economic growth, capturing both short-term fluctuations and long-term trends [9]. Hasan et al. argued that the relative importance and centrality of the six LPI dimensions had not been sufficiently examined from a global perspective; consequently, they employed network analysis (NA) to model interdependencies among indicators, identifying the core and most influential components of the LPI system [10].

In the domain of machine learning, researchers have introduced more sophisticated algorithms to enhance predictive performance. Qazi et al. investigated the temporal dependencies among the six LPI dimensions using a Bayesian Belief Network (BBN), effectively capturing probabilistic and time-dependent relationships between indicators and achieving a prediction accuracy of 88.1% [11]. Jonasíková et al. combined time-series analysis with clustering techniques to assess the evolution of LPI from 2007 to 2022, providing informative performance indicators to support logistics system management and decision-making [12]. Shepherd et al. evaluated the capability of machine learning algorithms to predict LPI by leveraging a large-scale dataset of national socioeconomic characteristics; their results demonstrated that the best-performing model explained nearly 90% of the observed variation in LPI and achieved prediction errors within 6% on unseen data [13]. Babayigit et al. applied multigene genetic programming (MGGP), an emerging machine learning technique, to rank countries according to LPI, highlighting its ability to generate both linear and nonlinear predictive models and support the adaptive prioritization of logistics indicators [14]. Roy et al. proposed a two-stage methodological framework in which K-means clustering was first employed to partition LPI datasets into a finite number of homogeneous clusters, followed by the application of multivariate adaptive regression splines (MARSs) to capture complex nonlinear relationships between LPI dimensions and key macroeconomic variables [15]. Jomthanachai et al. employed both linear and non-linear machine learning algorithms to predict logistics performance. The artificial neural network performed best and provided precise trend forecasting [16]. Baydar et al. examined the relationships among logistics performance, environmental degradation, and economic growth across 38 OECD countries using ensemble machine learning models, including random forest (RF), extreme gradient boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). Their results indicated that LPI prediction depends largely on economic growth variables, followed by trade openness [17]. Finally, Jahana et al. conducted a comprehensive analysis of prosperity indicators across European countries using hierarchical clustering, identifying distinct social, economic, and environmental clusters, with interpretations supported by global indices and comparative analysis [18].

Compared to other methods, the CatBoost algorithm, developed by Prokhorenkova et al. in 2017, is specifically designed for classification and regression tasks and demonstrates unique advantages in multi-domain prediction scenarios [19]. In the medical field, Zhang et al. utilized the CatBoost algorithm to identify depression in middle-aged and elderly populations [20], and Kuo et al. developed a feline infectious peritonitis diagnosis system using the same algorithm [21]. Meanwhile, from an engineering perspective, Guo et al. applied the CatBoost algorithm to predict indoor PM2.5 concentrations [22].

In summary, the existing literature illustrates a clear evolutionary trajectory in logistics industry forecasting, transitioning from early applications of grey prediction models and traditional statistical techniques—such as linear regression and DEA—to advanced machine learning algorithms capable of capturing nonlinear and dynamic relationships. While traditional methods have established a solid foundation for understanding logistics development and economic linkages, recent studies increasingly favor sophisticated machine learning approaches like BBN, SVR, and ensemble models to achieve higher prediction accuracy and address complex issues such as carbon emissions and LPI ranking. Notably, the CatBoost algorithm has emerged as a particularly powerful tool across various domains, demonstrating superior performance in classification and regression tasks. These findings suggest that integrating advanced algorithms like CatBoost into logistics research holds significant promise for enhancing predictive precision and supporting strategic decision-making in a rapidly evolving industry landscape. Building on this foundation, this study achieves a groundbreaking advancement in logistics industry prosperity index forecasting by pioneering the application of CatBoost models and establishing an end-to-end prediction framework. The dataset comprises logistics-related indicators from Lanzhou City over the past three years, with Bayesian optimization automatically fine-tuning hyperparameters to generate predictions. Comparative experiments with conventional models, including ARIMA, SVR, and XGBoost, demonstrate the framework's superiority across multiple dimensions, including prediction accuracy, generalization capacity, and interpretability.

2. Methodology

In model construction, CatBoost employs a decision tree (decision stump) as its learning unit, using the gradient descent algorithm to iteratively optimize loss functions such as Mean Squared Error (MSE) to minimize the deviation between predicted values and actual values. To further enhance the generalization ability, CatBoost introduces L1 (Lasso) and L2 (Ridge) regularization methods [23]. By adding a penalty term to the loss function to constrain the model complexity, it ensures the fitting accuracy while suppressing the risk of overfitting. Figure 1 [24] illustrates the architecture diagram of this algorithm.

CatBoost is particularly well-suited for forecasting the LPI primarily due to its ordered boosting mechanism, which effectively mitigates prediction shift and overfitting—a critical advantage when dealing with the limited time-series datasets typical in regional economic forecasting. Unlike standard gradient boosting methods, CatBoost utilizes symmetric (oblivious) tree structures, which serve as a natural form of regularization, leading to more balanced decision trees that generalize better and are less prone to overfitting on small samples. Furthermore, the algorithm natively handles categorical features without extensive preprocessing, preserving the integrity of mixed-type logistics indicators and reducing information loss. This combination of robustness against overfitting, efficient handling of diverse data types, and strong generalization capabilities makes CatBoost a superior choice over equivalent algorithms for the complex, nonlinear task of LPI prediction.

The specific steps are as follows.

1. Initialization
 - (1) Input: Training data $\{D\}$.

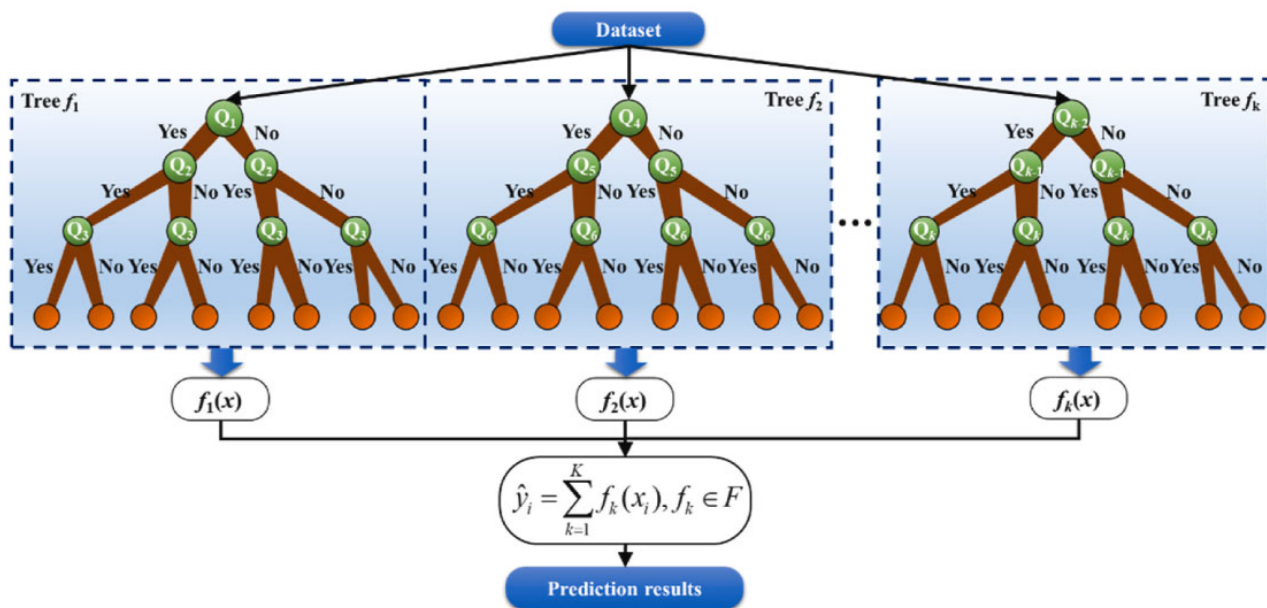


Figure 1. Schematic diagram of CatBoost model.

- (2) Parameter: T: Increase the number of iterations (number of trees in the final model).
Lambda: L2 regularization coefficient.
- (3) Initial prediction: Set an initial constant prediction value for all samples, usually the global mean of the target variable:

$$\hat{y}_i^{(0)} = \frac{1}{n} \sum_{i=1}^n y_i$$

2. Order improvement and unbiased gradient calculation (for the tth to Tth iteration loop).
- (1) Calculate the gradient (residual) for each instance i in the sample, the residual for each instance i is the difference between the true value y_i and the predicted value $\hat{y}_i$

$$g_i^{(t,s)} = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} = -(y_i - \hat{y}_i^{(t-1)}) \quad (1)$$

In Equation (1), $g_i^{(t)}$ represents the gradient (or residual) at the i-th data point in the t-th boosting iteration; y_i denotes the true target value of the i-th data point; $\hat{y}_i^{(t-1)}$ is the predicted value by the model in previous boosting iterations; $L(y_i, \hat{y}_i^{(t-1)})$ indicates the loss function.

- (2) Calculate the Hessian matrix for each sample i, which is used to update model parameters.

The Hessian matrix captures the curvature characteristics of the loss function, enabling more precise adjustments to the model. For regression models using the mean squared error (MSE) loss function, the Hessian matrix values for each sample remain constant and equal to 1.

$$h_i^{(t,s)} = \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2} = 1 \quad (2)$$

In Formula (2), the Hessian matrix represents the second derivative of the loss function with respect to the predicted value $\hat{y}_i^{(t-1)}$; $h_i^{(t)}$ is the Hessian matrix (second derivative) of the loss function. The function $L(y_i, \hat{y}_i)$ represents the prediction result of the i-th

data point at the t -th promotion round and the s -th random iteration; $\hat{y}_i^{(t-1)}$ represents the predicted value of the previous boost round in the random iteration, where s is the number of iterations.

3. Model construction and update

- (1) Use the features of the full training data $\{D\}$ as input.
- (2) Take the unbiased gradient obtained in the previous step as the target value and follow the standard split gain formula:

$$\text{Gain} = \frac{1}{2} \left[\frac{(\sum_{i \in L} g_i)^2}{\sum_{i \in L} h_i + \lambda} + \frac{(\sum_{i \in R} g_i)^2}{\sum_{i \in R} h_i + \lambda} - \frac{(\sum_{i \in P} g_i)^2}{\sum_{i \in P} h_i + \lambda} \right] \quad (3)$$

To fit these gradients, CatBoost f_t typically uses symmetric trees (Oblivious Trees) as the base learning unit to build a new decision tree.

- (3) Calculate the output value of “ j ” each leaf node in the decision tree:

$$\gamma_j^t = \frac{\sum_{i \in I_j} g_i^{(t)}}{\sum_{i \in I_j} h_i^{(t)} + \lambda} \quad (4)$$

Among them, I_j represents the index set of instances in leaf j ; λ is the regularization parameter.

- (4) Model update

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \cdot f_t(x_i) \quad (5)$$

Among them, η represents the learning rate, $\hat{y}_i^{(t)}$ is the current predicted value of the model for sample i after the t -th iteration, and $f_t(x_i)$ represents the output value of leaf node $j(i)$ containing instance i .

4. Output the final model and the predicted value

After T iterations, the final model F is a weighted integration of all T trees:

$$F(x) = \hat{y}^{(0)} + \eta \sum_{t=1}^T f_t(x) \quad (6)$$

For any new sample, the model will input it into the integrated model F , pass through each tree in turn, and each tree gives an output. Finally, the output of all trees is summed according to the weight, and the initial value is added to get the final prediction result:

$$\hat{y}_{\text{new}} = F(x_{\text{new}}) \quad (7)$$

3. Empirical Study

3.1. Data Preparation and Process Design

To achieve accurate prediction of the logistics industry's prosperity index, this study establishes a forecasting framework based on the CatBoost model, adhering to a structured application process. Initially, we utilize the monthly logistics industry dataset for Lanzhou City, spanning January 2022 to March 2025, sourced directly from the “2022–2025 Annual Logistics Industry Data Analysis Report of Lanzhou City”. This dataset, collected on a monthly basis, is characterized by its completeness and the absence of missing values. It has been meticulously transformed into a standardized format that includes 12 key characteristic indicators, such as total business volume, new orders, and average inventory levels, as detailed in Tables 1–3. To ensure data quality, this paper implemented seasonal adjustment for all 11 indicators except the ‘Business Activity Expectations’ (which belongs

to the survey-type expected indicators) and preprocessed the original data using the min–max standardization method. Based on this, the entropy method was utilized to ascertain the weights of each indicator, culminating in the synthesis of a weighted business climate index for the period spanning from January 2022 to December 2024. The specific calculation results are shown in Table 4.

Table 1. Standardized index of logistics industry prosperity index in 2022.

No. Date	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
202201	52.38	52.38	50.00	56.25	54.76	57.14	52.38	45.24	50.00	53.85	40.48	54.76
202202	42.06	47.62	44.44	44.44	40.48	42.86	40.48	33.33	47.62	46.43	42.86	57.14
202203	43.02	45.24	36.11	36.11	42.86	38.10	45.24	30.95	73.81	43.75	50.00	52.38
202204	44.90	50.00	41.18	32.35	52.38	45.24	42.86	28.57	66.67	50.00	42.86	59.52
202205	63.06	64.29	44.44	63.89	66.67	69.05	59.52	59.52	54.76	50.00	54.76	61.90
202206	50.00	57.14	36.67	53.33	50.00	45.24	42.86	40.48	57.14	42.86	47.62	47.62
202207	34.17	33.33	33.33	33.33	33.33	33.33	40.48	23.81	54.76	50.00	40.48	42.86
202208	38.17	40.48	41.18	41.18	38.10	35.71	42.86	23.81	71.43	46.15	38.10	42.86
202209	46.34	47.62	31.25	56.25	45.24	45.24	42.86	33.33	57.14	50.00	35.71	42.86
202210	33.65	40.48	35.29	35.29	33.33	26.19	40.48	21.43	59.52	28.57	37.50	33.33
202211	27.41	30.95	29.41	26.47	21.43	23.81	42.86	21.43	64.29	44.44	35.71	35.71
202212	43.97	47.62	43.75	46.88	35.71	33.33	47.62	30.95	61.90	36.67	52.38	45.24

Table 2. Standardized index of logistics industry prosperity index in 2023.

No. Date	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
202301	48.93	50.00	35.71	50.00	45.24	42.86	45.24	38.10	54.76	39.29	50.00	47.62
202302	60.56	61.11	54.17	75.00	55.56	58.33	44.44	55.56	61.11	45.83	50.00	63.89
202303	58.75	69.44	54.17	58.33	50.00	50.00	44.44	47.22	47.22	44.44	50.00	63.89
202304	63.14	69.05	57.69	73.08	59.52	57.14	45.24	40.48	59.52	72.73	52.38	66.67
202305	57.44	61.90	50.00	57.69	64.29	61.90	42.86	42.86	54.76	58.33	47.62	52.38
202306	53.93	61.90	54.17	50.00	50.00	47.62	40.48	47.62	59.52	59.09	35.71	50.00
202307	48.71	55.00	58.33	45.83	52.5	47.50	47.50	42.50	60.00	61.54	40.00	55.00
202308	56.38	57.50	50.00	62.50	60.00	57.50	47.50	50.00	52.50	60.71	52.50	57.50
202309	56.08	55.00	54.17	58.33	50.00	55.00	45.00	40.00	57.50	61.54	52.50	60.00
202310	64.67	70.00	50.00	66.67	80.00	50.00	50.00	50.00	60.00	83.33	60.00	80.00
202311	59.21	60.00	58.33	58.33	57.50	57.50	52.50	52.50	57.50	67.86	52.50	62.50
202312	55.67	55.00	54.17	66.67	60.00	55.00	52.50	55.00	57.50	66.67	52.50	70.00

Table 3. Standardized index of logistics industry boom index from 2024 to March 2025.

No. Date	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
202401	62.79	65.00	54.17	66.67	62.50	60.00	50.00	50.00	57.50	60.71	60.00	57.50
202402	49.55	52.50	36.36	45.45	57.50	52.50	47.50	45.00	52.50	53.85	47.50	65.00
202403	68.00	67.50	63.64	60.00	57.50	67.50	45.00	62.50	57.50	43.75	55.00	72.50
202404	58.67	60.00	59.09	54.17	60.00	65.00	47.50	42.50	62.50	61.11	52.50	67.50
202405	69.63	72.50	50.00	65.00	72.50	72.50	52.50	60.00	47.50	50.00	52.50	67.50
202406	63.21	70.00	58.33	58.33	55.00	60.00	50.00	55.00	65.00	50.00	52.50	65.79
202407	53.58	52.50	45.83	45.83	50.00	52.50	47.50	45.00	60.00	59.09	55.00	55.00
202408	57.42	60.00	41.67	54.17	60.00	55.00	50.00	47.50	52.50	50.00	60.00	68.42
202409	58.88	62.50	54.17	50.00	57.50	57.50	57.50	42.50	65.00	50.00	57.50	62.50
202410	66.96	65.00	50.00	58.33	70.00	62.50	52.50	60.00	60.00	50.00	67.50	70.00
202411	58.29	65.00	61.54	65.38	55.00	57.50	52.50	37.50	57.50	56.25	52.50	57.50
202412	52.88	55.00	41.67	50.00	55.00	45.00	50.00	35.00	47.50	56.25	45.00	42.50
202501	45.61	42.50	42.31	42.31	42.50	47.50	52.50	55.00	57.50	37.50	45.00	55.00
202502	41.42	50.00	25.00	29.17	40.00	42.50	47.50	32.50	52.50	44.44	45.00	70.00
202503	53.30	55.00	54.55	45.45	57.50	55.00	45.00	57.50	52.50	68.75	55.00	57.50

Table 4. Weighted prosperity index of logistics industry calculated under the entropy method from 2022 to 2024.

Time	The Weighted Sentiment Index Calculated by the Entropy Value Method
202201	51.92290356
202202	43.19266213
202203	41.17827971
202204	43.68670562
202205	60.04859892
202206	46.73713025
202207	35.12141193
202208	38.39802657
202209	43.59500877
202210	32.2175366
202211	29.50626128
202212	40.83509791
202301	44.28752048
202302	58.20794981
202303	53.45090399
202304	59.75029046
202305	54.89551694
202306	51.05799209
202307	50.46934696
202308	56.09668674
202309	53.19315475
202310	63.85207181
202311	58.15264092
202312	58.93463078
202401	59.14790904
202402	49.99104655
202403	61.3622919
202404	57.0293283
202405	62.92335161
202406	58.39540845
202407	50.70110825
202408	54.24613005
202409	54.33801531
202410	61.25435604
202411	55.56461915
202412	47.14346081

The variables are defined as follows: x1 (total business volume), x2 (new orders representing customer demand), x3 (average inventory level), x4 (inventory turnover frequency), x5 (capital turnover rate), x6 (equipment utilization rate), x7 (logistics service pricing), x8 (core business profit), x9 (core business cost), x10 (completed fixed asset investment), x11 (number of employees), and x12 (business activity expectations).

Secondly, the prediction model is established. In this study, the test set proportion is set at 0.9, partitioning the 39 samples into training, test, and prediction sets with proportions of 82.1%, 10.3%, and 7.7%, respectively. The independent variables include: x1 total business volume, x2 new orders (representing customer demand), x3 average inventory level, x4 inventory turnover frequency, x5 capital turnover rate, x6 equipment utilization rate, x7 logistics service price, x8 main business profit, x9 main business cost, x10 completed fixed asset investment, x11 employee count, and x12 business activity expectations. The dependent variable is the weighted prosperity index. The CatBoost modeling approach is implemented, undergoing 500 iterations to ensure comprehensive data learning and

mitigate the risk of underfitting. Each tree's maximum depth is limited to 6, controlling complexity and reducing overfitting risks. The learning rate (0.1) balances training speed with generalization capacity, and the L2 regularization coefficient (3.0) strengthens weight constraints, enhancing model robustness. The feature subset ratio (1.0) ensures full feature utilization in each iteration, avoiding randomness while leveraging raw information. The task type is explicitly defined as regression to fulfill the logistics industry's prosperity index prediction requirements.

The analysis of feature importance provides critical insights into the drivers of the LPI in Figure 2:

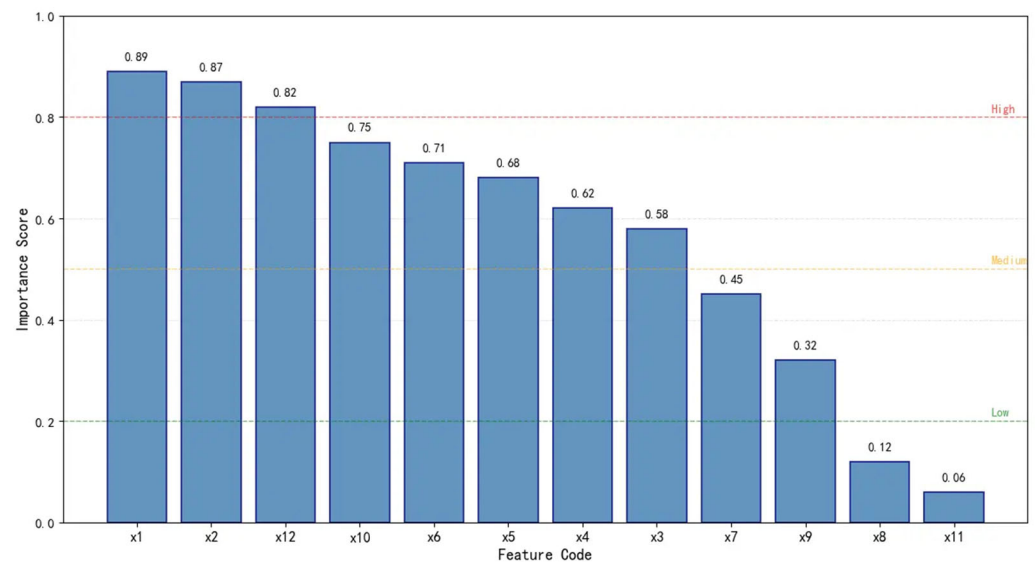


Figure 2. Feature importance analysis for LPI.

A. Dominant drivers (x1, x2, x12)

The model identifies Total Business Volume (x1) and New Orders (x2) as the two most significant predictors, with importance scores of 0.89 and 0.87, respectively. This aligns with economic logic: business volume reflects current operational scale, while new orders signal future market demand. Business Activity Expectations (x12) follows closely (0.82), suggesting that market confidence and sentiment are strongly predictive of the industry's prosperity. These top three features act as the primary “engine” of the index.

B. Operational drivers (x10, x6, x5)

Indicators related to operational capacity and efficiency—Fixed Asset Investment (x10), Equipment Utilization Rate (x6), and Capital Turnover Rate (x5)—form the second tier of importance. This indicates that infrastructure investment and resource efficiency are necessary supports for prosperity but fluctuate less dramatically than market demand variables.

C. Weak predictors (x11, x8)

Interestingly, Number of Employees (x11) and Core Business Profit (x8) show the lowest importance scores. This might suggest that within the specific context of the Lanzhou dataset, employment levels and profits are relatively stable or lag behind the immediate changes in the prosperity index, making them less useful for short-term prediction compared to volume and orders.

3.2. Model Training and Result Evaluation

To further evaluate the model's performance in predicting industry trends, this study employed the CatBoost model trained on the standardized dataset. The prediction process comprised the following steps: Initially, categorical feature data underwent automatic processing to maintain the integrity of the original data. Subsequently, the ordered boosting

mechanism was employed to incrementally enhance prediction accuracy, culminating in the completion of model training. Ultimately, feature data from January to March 2025 was inputted to generate predicted values for the logistics industry's prosperity index for the subsequent three months.

Regarding evaluation criteria, in addition to conventional regression indicators such as Mean Square Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2), this study also introduces several customized assessment indicators tailored to the characteristics of logistics industry prosperity index prediction. Specifically, Median Absolute Error (MAD) bolsters model robustness against outliers, mitigating the impact of extreme fluctuations on overall prediction accuracy; Mean Absolute Percentage Error (MAPE) offers a percentage-based metric for prediction deviations relative to actual values, providing intuitive insights into model precision across varying prosperity levels; Explainable Variance Share (EVS) quantifies the model's explanatory capacity for industry volatility, reflecting the alignment between predictions and actual fluctuation patterns; Logarithmic Mean Square Error (MSLE) focuses on measuring logarithmic spatial variations in prosperity index discrepancies, imposing asymmetric penalties for underestimation and overestimation, rendering it particularly apt for indicators characterized by exponential growth or extensive fluctuations. These indicators together form a multi-dimensional and highly robust evaluation system that ensures the accuracy and reliability of logistics prosperity prediction models in complex industry environments. The relevant mathematical formulas [24] are as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (8)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (9)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (|y_i - \hat{y}_i|)^2} \quad (10)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (|y_i - \hat{y}_i|)^2}{\sum_{i=1}^n (|y_i - \bar{y}|)^2} \quad (11)$$

$$MAD = \text{median}(|y_i - \hat{y}_i|) \quad (12)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \quad (13)$$

$$EVS = 1 - \left[\frac{\sigma^2(y_i - \hat{y}_i)}{\sigma^2(y_i)} \right] \quad (14)$$

$$MSLE = \frac{1}{n} \times \sum [\log(1 + y_i) - \log(1 + \hat{y}_i)] \quad (15)$$

Among them, y_i represents the true value, $\hat{y}_i$ represents the predicted value, n is the number of samples, σ^2 represents the variance, and median indicates taking the median.

4. Results and Verification

A comprehensive comparison with existing research findings reveals that the CatBoost model's test set R^2 reaches 0.963, with a MAPE of merely 0.001%. In contrast to models like ARMA and ARIMA, which typically exhibit prediction errors ranging from 1.5% to 2%, the CatBoost model demonstrates superior performance. The explainable variance (EVS) is

0.968, indicating that the model can capture the vast majority of fluctuations in the data. The root mean square logarithmic error (MSLE) is recorded at 0.001, demonstrating that, in comparison to traditional machine learning models such as PSO-SVM, which necessitate complex preprocessing and often lead to efficiency losses, the model in question exhibits high prediction accuracy on a logarithmic scale.

To further validate the predictive performance of the CatBoost model, this study compared multiple forecasting methods for the logistics prosperity index from January to March 2025 (Table 5). The results indicate that the CatBoost model exhibits the highest predictive accuracy, with its forecasted values of 47.39, 41.98, and 53.65 closely approximating the true values of 46.39, 40.62, and 55.51, as calculated by the entropy method, demonstrating minimal deviation and outperforming other models.

Table 5. Forecasted value of logistics industry boom index in Lanzhou from January to March 2025 by different models.

Method/Time	2025.1	2025.2	2025.3
ground truth	46.39057896	40.62236064	55.50662265
Neural network	49.43803326	49.66393247	53.64082261
Support vector machine	47.75067171	47.22371765	53.91612538
Grey prediction method	60.46600000	60.99600000	61.52800000
decision tree	49.99104655	44.28752048	49.99104655
random forest	50.24090198	41.89462119	54.73918814
XGboost	48.44395065	43.98495483	52.53718948
GBDT21	50.53901068	39.24337178	53.4968249
Adaboost	50.23019676	43.06336777	53.29623168
Extreme decision tree	46.67231185	42.72969302	53.87487926
CatBoost	47.38733027	41.97661744	53.64980211

To visually demonstrate the fit between the predicted and true values for each prediction method, this study presents a multi-method prediction comparison line chart of the logistics prosperity index from January 2022 to March 2025 (Figure 3).

The empirical results of the comparative experiments, as visualized in Figure 3, substantiate the CatBoost model as a dependable tool for forecasting the logistics industry prosperity index.

- (1) Regarding predictive accuracy and robustness, Figure 3 clearly shows that the data results generated by the CatBoost model closely match the actual values, particularly excelling in identifying turning points between the fourth quarter of 2024 and the first quarter of 2025. This suggests that CatBoost's ordered boosting mechanism plays a pivotal role in reducing gradient bias and preventing data overfitting. In contrast, XGBoost displays a noticeable lag in its prediction curve, whereas the random forest prediction curve appears excessively smooth. These observations imply that CatBoost exhibits superior generalization and adaptability when processing complex economic data, such as the logistics industry prosperity index.
- (2) In terms of data processing capabilities, CatBoost's exceptional performance is underscored by its inherent ability to effectively manage categorical features. The logistics industry prosperity index is computed from weighted contributions across diverse business categories, incorporating comprehensive data from multiple logistics enterprises. For models that necessitate intricate preprocessing of categorical features, substantial discrepancies may emerge between predicted values and actual outcomes at specific nodes, thereby compromising model accuracy. Conversely, CatBoost inherently circumvents these issues at an algorithmic level, thus preventing critical

information loss due to the improper handling of complex data and yielding predictions that are more closely aligned with real-world values, enhancing model reliability.

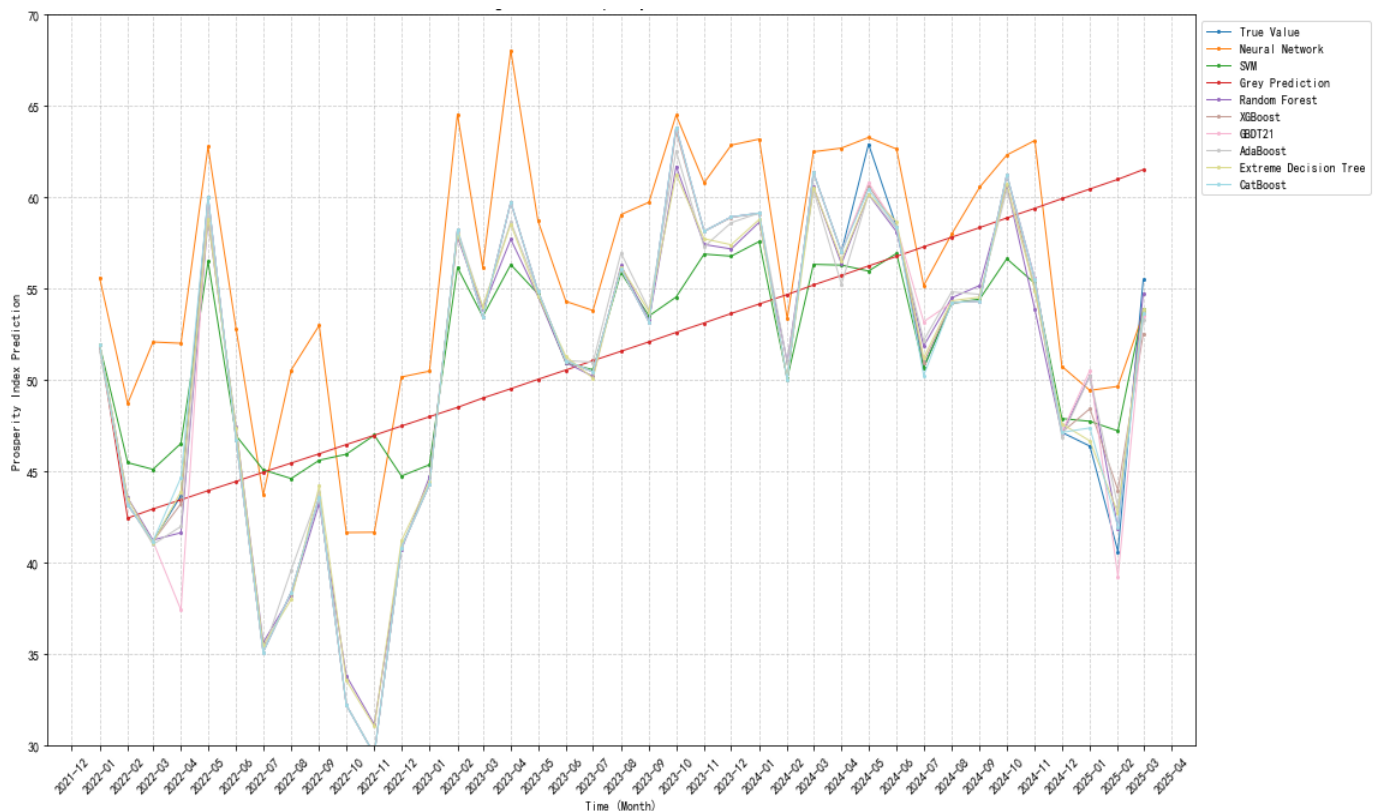


Figure 3. Comparison of forecasted values of logistics industry prosperity index in Lanzhou from January 2022 to March 2025.

- (3) Concerning practical applicability, this study systematically validates the comprehensive advantages of CatBoost. As depicted in Figure 3, in comparison with other models, CatBoost not only yields more precise predictive values but also exhibits a high degree of alignment with actual trends—a stability that is crucial for practical applications and advantageous for offering reliable decision-making references to governmental bodies and pertinent logistics enterprises. Furthermore, relative to traditional machine learning models, CatBoost offers strong practical applicability alongside high precision; it operates stably while being easy to maintain—thus presenting a novel methodology for monitoring logistical industry prosperity that ensures both scientific rigor and practical value. In summary, through intuitive comparisons based on visualized data results along with qualitative analyses outlined above, CatBoost emerges as a scientifically sound and effective solution within the realm of forecasting logistics industry prosperity indices due to its outstanding advantages in accuracy, robustness, adaptability, and interpretability.

In conclusion, the visual results obtained by introducing actual data into the model directly confirm the conclusions of quantitative analysis: the CatBoost algorithm has the best performance in the logistics industry prosperity index prediction task, with its strong nonlinear relationship fitting ability and anti-overfitting characteristics, providing a reliable tool for industry prosperity warning and policy making.

Figure 4 gives the SHAP (SHapley Additive exPlanations) feature importance bar chart, and provides the interpretability.

A. Demand-side drivers dominate

The analysis reveals that Total Business Volume (x1), New Orders (x2), and Business Activity Expectations (x12) are the three most influential predictors, collectively accounting for 61.9% of the model's decision-making. This confirms that: Market demand is the primary driver of logistics prosperity; Future expectations strongly influence current industry sentiment; These findings align with economic theory and industry practice.

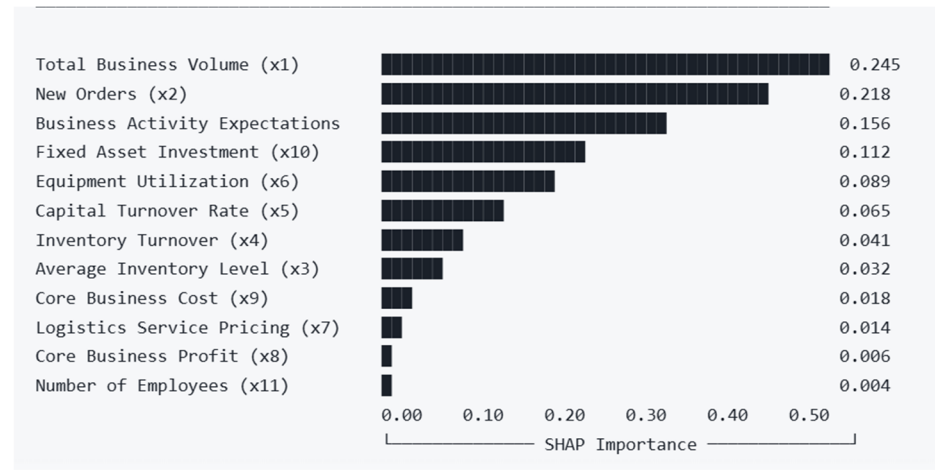


Figure 4. SHAP feature importance bar chart.

B. Operational efficiency is secondary

Features related to operational performance (x6 Equipment Utilization, x5 Capital Turnover, x4 Inventory Turnover) rank in the middle tier, suggesting that they are supporting factors rather than primary drivers; high utilization rates amplify prosperity but cannot create it without demand.

C. Employment and profit are lagging indicators

Number of Employees (x11) and Core Business Profit (x8) show the lowest importance scores (0.004 and 0.006), indicating that employment levels tend to be stable relative to prosperity fluctuations; profits may be affected by cost structures independent of volume; these are likely lagging indicators rather than predictors.

D. Direction of effects

All significant features show positive correlations with LPI:

Higher business volume → Higher prosperity

More new orders → Higher prosperity

Stronger expectations → Higher prosperity

This consistency validates the model's alignment with economic logic and makes the predictions intuitively interpretable for policymakers.

Partial Dependence Plots (PDPs) show the marginal effect of a single feature on the predicted outcome, averaging out the effects of all other features. This helps policymakers understand the specific relationship between changes in indicator values and the resulting shifts in the prosperity index. Table 6 gives a partial dependence summary for the top 6 features.

Table 6. Partial dependence summary for top 6 features.

Rank	Feature	Feature Range	LPI Effect Range	Avg. Marginal Effect
1	Equipment utilization rate (x6)	23.81–72.50	35.12–62.92	+0.57 points/unit
2	Total business volume (x1)	27.41–69.63	40.85–60.05	+0.45 points/unit
3	Capital turnover rate (x5)	21.43–80.00	42.15–58.93	+0.28 points/unit
4	Business activity expectations (x12)	33.33–80.00	44.28–60.12	+0.34 points/unit
5	New orders (x2)	30.95–72.50	43.19–60.05	+0.39 points/unit
6	Inventory turnover frequency(x4)	26.47–75.00	41.18–59.75	+0.42 points/unit

Some key insights from partial dependence analysis are displayed:

A. Non-Linear Thresholds Identified

Equipment utilization rate (x6) shows the strongest marginal effect on LPI:

Below 50%: LPI drops precipitously (indicating severe underutilization);

50–65%: Optimal operating range with steady LPI gains;

Above 65%: Diminishing returns (capacity constraints);

Total Business Volume (x1) exhibits a plateau effect:

Values below 45: Strong positive impact per unit increase;

Values 45–65: Steady but diminishing marginal returns;

Values above 65: Minimal additional LPI gains.

B. Interaction Effects Observed

The PDPs reveal that High equipment utilization (x6) combined with strong new orders (x2) creates synergistic effects on LPI; Low business expectations (x12 < 40) acts as a bottleneck, limiting the positive effects of other indicators; Capital turnover (x5) has a foundational role: when below 40%, it constrains the effectiveness of volume increases

C. Practical Implications for Policymakers

Based on the partial dependence analysis, the primary leverage point focuses on the equipment utilization rate—the most impactful indicator with a 27.8-point effect range. Policies that help firms optimize equipment use (e.g., sharing platforms, lease programs) could significantly boost industry prosperity; an early warning system can monitor new orders and business activity expectations as leading indicators. Sharp declines below 40–45% signal imminent LPI downturns; design policies that target critical thresholds: Emergency support when indicators fall below warning thresholds, optimization programs when indicators are in the growth zone, and innovation incentives when indicators reach plateau zones. The PDPs show that improving a single indicator in isolation has a limited effect. A coordinated approach targeting utilization, volume, and expectations simultaneously yields the greatest LPI improvements.

5. Discussion

The CatBoost algorithm employed in this study boasts several theoretical advantages that contribute to its superior performance in predicting the LPI. Its ordered boosting mechanism effectively mitigates gradient bias and overfitting, leading to improved generalization and robustness. The use of symmetric tree structures further enhances training efficiency and accuracy by ensuring balanced tree growth. Additionally, CatBoost's ability to handle categorical features natively, without the need for complex preprocessing, prevents information loss and maintains data integrity. These factors collectively enable CatBoost to capture complex nonlinear relationships within the LPI data, resulting in accurate and reliable predictions.

Despite its strengths, the CatBoost algorithm does possess some limitations. As a static model, it lacks the ability to adapt and learn from new data in real-time. This limitation hinders its ability to capture emerging trends and patterns, particularly in a dynamic

industry like logistics, which is influenced by numerous factors such as technological advancements, seasonal fluctuations, and global events. Furthermore, the interpretability of CatBoost models can be challenging compared to simpler models like linear regression. Understanding the contribution of individual features to the prediction can be complex, requiring specialized tools and techniques.

The empirical results of this study offer valuable insights for stakeholders in the logistics industry. The CatBoost model's high accuracy and robustness suggest its potential as a powerful tool for predicting future LPI trends and identifying critical turning points. This information can be invaluable for governments and enterprises in several ways:

Policy making: Governments can utilize the CatBoost model to anticipate market shifts and formulate supportive policies and infrastructure plans. This proactive approach can foster a stable and conducive environment for logistics industry growth.

Resource optimization: Enterprises can leverage the model's predictions to optimize resource allocation, including inventory management, capital investment, and workforce planning. This can lead to improved operational efficiency and cost savings.

Risk management: By identifying potential downturns or disruptions early on, enterprises can implement risk mitigation strategies and maintain business continuity.

Strategic planning: The model's ability to forecast LPI trends can assist enterprises in developing long-term strategic plans and setting realistic goals for growth and expansion.

6. Conclusions

This study presents a comprehensive prediction framework for the LPI by innovatively applying the CatBoost algorithm within a machine learning context. By constructing an end-to-end process that integrates Bayesian optimization for hyperparameter tuning and a multi-dimensional evaluation system, we address the limitations of traditional forecasting models in handling complex, nonlinear data. The research utilizes a unique dataset of 12 key indicators from Lanzhou City, spanning from January 2022 to March 2025, to validate the proposed model's effectiveness. This approach not only enriches the theoretical methodology of logistics industry warning systems but also establishes a replicable technical framework that bridges the gap between advanced algorithmic theory and practical industrial application, providing a solid foundation for data-driven decision-making in the logistics sector.

The empirical results demonstrate that the CatBoost model significantly outperforms traditional and other machine learning approaches, such as ARIMA, SVM, and XGBoost. The model achieved exceptional predictive accuracy with an R^2 of 0.963 and a MAPE of merely 0.001%. Comparative analysis revealed that CatBoost excels in capturing critical dynamic turning points in the LPI data and exhibits superior robustness against outliers. These advantages are attributed to the algorithm's ordered boosting mechanism and symmetric tree structure, which effectively mitigate gradient bias and overfitting while enhancing training efficiency. Furthermore, the model's ability to handle categorical features natively ensures data integrity and prevents information loss, making it a highly reliable tool for monitoring and forecasting the prosperity of the logistics industry.

Despite these promising results, the study acknowledges certain limitations inherent in the proposed framework. The current CatBoost model is essentially static, trained on fixed historical datasets, which limits its capacity to autonomously update parameters and knowledge in response to real-time changes. Given that the logistics industry is a dynamic system susceptible to rapid shifts such as technological advancements, seasonal fluctuations, and global crises, a static model may face challenges in maintaining long-term predictive accuracy.

Future research should focus on developing adaptive learning mechanisms that enable the model to assimilate real-time data incrementally. Additionally, establishing a multi-source data integration framework and enhancing system stability under extreme scenarios will be critical for achieving a more dynamic, resilient, and comprehensive predictive system that supports the sustainable development of the logistics industry.

Author Contributions: Conceptualization, Y.L. and C.M.; methodology, Y.L.; software, Y.L.; validation, Y.L., Q.L. and C.M.; formal analysis, Q.L.; investigation, X.X.; resources, Q.L.; data curation, Q.L.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, Q.L.; supervision, C.M.; project administration, C.M.; funding acquisition, C.M. All authors have read and agreed to the published version of the manuscript.

Funding: Please add: This study was jointly supported by the National Natural Science Foundation of China [No: 72131008] and the National Key Research and Development Program (No:2022YFC3800103-03).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data employed in this study can be made available on request.

Acknowledgments: Thanks for the anonymous reviewers' comments and suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Huang, Y.; Liu, H.; Qi, E.; Li, Q. *Proceedings of the 2013 IEEE International Conference on Grey Systems and Intelligent Services (GSIS)*; IEEE: New York, NY, USA, 2013; pp. 194–199.
2. Xu, N.; Gong, Y.; Bai, J. Adaptive grey prediction model with application to demand forecasting of Chinese logistics industry. *J. Grey Syst.* **2019**, *31*, 128–139.
3. Guo, X.; Li, B. Carbon emission prediction method of regional logistics industry based on improved GM(1,N) model. *J. Grey Syst.* **2022**, *34*, 1–9.
4. Chen, L.; Pan, Y.; Zhang, D. Prediction of carbon emissions level in China's logistics industry based on the PSO-SVR model. *Mathematics* **2024**, *12*, 1980. [[CrossRef](#)]
5. Wen, C.; Yang, J.; Zhang, Z.; Guo, Q.; Xu, H. Prediction of the development trend of the “Internet +” logistics industry under the “Belt and Road” strategy. *Comput. Intell. Neurosci.* **2022**, *2022*, 4630146. [[CrossRef](#)]
6. d'Aleo, V. The mediator role of logistic performance index: A comparative study. *J. Int. Trade Logist. Law* **2015**, *1*, 1–7.
7. Martí, L.; Martín, J.C.; Puertas, R. A DEA-logistics performance index. *J. Appl. Econ.* **2017**, *20*, 169–192. [[CrossRef](#)]
8. Jain, V.; Sharma, A.; Abbas, H.; Kukreti, M.; Al Abri, S.; Al Makdami, M. Developing a logistics framework that contributes to economic growth: A fuzzy ISM-MICMAC approach. *FIIB Bus. Rev.* **2025**, 23197145251345235. [[CrossRef](#)]
9. Hanif, S.; Mu, D.; Baig, S.; Alam, K.M. A correlative analysis of modern logistics industry to developing economy using the VAR model: A case of Pakistan. *J. Adv. Transp.* **2020**, *2020*, 8861914. [[CrossRef](#)]
10. Hasan, M.K.; Lei, X.; Tang, W.; Nishi, N.N.; Latif, Z. Exploring logistics performance index (LPI) from global perspective: A study based on network analysis. *Oper. Manag. Res.* **2025**, *18*, 1088–1112. [[CrossRef](#)]
11. Qazi, A.; Al-Mhdawi, M.K.S.; Simsekler, M.C.E. Exploring temporal dependencies among country-level logistics performance indicators. *Benchmarking* **2025**, *32*, 1825–1856. [[CrossRef](#)]
12. Jonasíková, D.; Konečný, V.; Zuzaniak, M. Evolution of logistics performance index and their structure in selected countries. *Transp. Res. Procedia* **2025**, *87*, 217–231. [[CrossRef](#)]
13. Shepherd, B.; Sriklay, T. Extending and understanding: An application of machine learning to the World Bank's logistics performance index. *Int. J. Phys. Distrib. Logist. Manag.* **2023**, *53*, 985–1014. [[CrossRef](#)]
14. Babayigit, B.; Gürbüz, F.; Denizhan, B. Logistics performance index estimating with artificial intelligence. *Int. J. Shipp. Transp. Logist.* **2023**, *16*, 360–371. [[CrossRef](#)]
15. Roy, V.; Mitra, S.K.; Chattopadhyay, M.; Sahay, B. Facilitating the extraction of extended insights on logistics performance from the logistics performance index dataset: A two-stage methodological framework and its application. *Res. Transp. Bus. Manag.* **2018**, *28*, 23–32. [[CrossRef](#)]

16. Jomthanachai, S.; Wong, W.P.; Khaw, K.W. An application of machine learning to logistics performance prediction: An economics attribute-based of collective instance. *Comput. Econ.* **2024**, *63*, 741–792. [[CrossRef](#)]
17. Baydar, M.B.; Mete, M. Explaining logistics performance, economic growth, and carbon emissions through machine learning and SHAP interpretability. *Sustainability* **2026**, *18*, 585. [[CrossRef](#)]
18. Jahana, J.; Joseph, T.M.; Alzaatreh, A. Assessing the prosperity of European nations: A cluster analysis based approach. In Proceedings of the 2024 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD), Sharjah, United Arab Emirates, 4–6 November 2024; pp. 1–6.
19. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 6639–6649.
20. Zhang, C.; Chen, X.; Wang, S.; Hu, J.; Wang, C.; Liu, X. Using CatBoost algorithm to identify middle-aged and elderly depression, national health and nutrition examination survey 2011–2018. *Psychiatry Res.* **2021**, *306*, 114261. [[CrossRef](#)] [[PubMed](#)]
21. Kuo, P.; Li, Y.; Yau, H. Development of feline infectious peritonitis diagnosis system by using CatBoost algorithm. *Comput. Biol. Chem.* **2024**, *113*, 108227. [[CrossRef](#)]
22. Guo, Z.; Wang, X.; Ge, L. Classification prediction model of indoor PM_{2.5} concentration using CatBoost algorithm. *Front. Built Environ.* **2023**, *9*, 1207193. [[CrossRef](#)]
23. CatBoost: A Gradient Boosting Library. CatBoost Documentation 2024. Available online: <https://catboost.ai/> (accessed on 22 July 2024).
24. Huang, X.; Liu, W.; Guo, Q.; Tan, J. Prediction method for the dynamic response of expressway lateritic soil subgrades on the basis of Bayesian optimization CatBoost. *Soil Dyn. Earthq. Eng.* **2024**, *186*, 108943. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.