

Analysis of New Energy Vehicle Sales Based on Variable Selection-xgboost Integrated Model

Quanying LU

College of Economics and Management, Beijing University of Technology, Beijing 100124, China

E-mail: luquanying@bjut.edu.cn

Tengjian MA

College of Economics and Management, Beijing University of Technology, Beijing 100124, China

E-mail: mtj24@emails.bjut.edu.cn

Xiran LIU

College of Economics and Management, Beijing University of Technology, Beijing 100124, China

E-mail: 13121820036@163.com

Jingru GUO

College of Economics and Management, Beijing University of Technology, Beijing 100124, China

E-mail: Guojr@emails.bjut.edu.cn

Qijing YAN*

*School of Mathematics, Statistics and Mechanics, Beijing University of Technology, Beijing 100124,
China*

E-mail: yanqijing@bjut.edu.cn

Abstract The development of the new energy vehicle (NEV) industry is pivotal in addressing China's energy security challenges and restructuring its automotive sector. It is a critical measure for achieving China's carbon neutrality goals. Accurately analyzing and forecasting NEV sales volumes carries significant implications for industry planning and policy formulation. This paper proposes an integrated modeling framework that combines variable selection techniques with the XGBoost algorithm to forecast NEV sales in China. We compare three distinct variable selection methods and evaluate their performance against commonly used benchmark forecasting models. The prediction performance is assessed using two metrics: root mean squared error (RMSE) and mean absolute percentage error (MAPE). Our empirical results demonstrate that the variable selection-XGBoost integrated model outperforms both univariate models and models that do not incorporate core factor extraction, showing superior accuracy in both in-sample and out-of-sample predictions. Among the variable selection-XGBoost models,

Received December 14, 2024, accepted June 18, 2025

Supported by National Natural Science Foundation of China (72573014, 72103185); Natural Science Foundation of Beijing (9254021) and MOE (Ministry of Education in China) Project of Humanities and Social Sciences (23YJCZH259)

*Corresponding author

the SSL-XGBoost model yields the best performance, followed by GLMNET-XGBoost and LARS-XGBoost. The SSL method selects the greatest number of core variables. Cross-validation results indicate that integrating XGBoost with variable selection significantly reduces prediction errors. The variable selection-XGBoost integrated model surpasses traditional models in terms of both accuracy and reliability.

Keywords variable selection-XGBoost; integrated forecasting; new energy vehicle

1 Introduction

Industrialization and urbanization have driven increased travel demand and vehicle ownership¹. This growth has led to oil shortages and greenhouse gas emissions, creating significant challenges for energy security and climate stability^[1]. Amid rising international energy disputes, volatile oil prices^[2,3], and sustained vehicle demand, countries are investing heavily in renewable energy research². The promotion of NEV has become a critical strategy for addressing the global energy crisis and climate change. Accurate sales forecasting plays a crucial role in driving the production and consumption of NEV, while also providing valuable decision-making support for national energy policies. It helps optimize the industrial layout of NEV, thereby effectively promoting green, low-carbon development and facilitating the transformation of the national energy structure. As a major automotive market, advancing the development of the new energy vehicle industry is China's essential path from being a large automobile country to a strong automobile country. It is also a key measure for China to achieve the "carbon neutrality" goal.

China stands as both the primary manufacturer and largest market for NEVs globally. The government has implemented various supportive policies to promote high-quality development of the NEV industry. While NEVs were designated as one of the seven strategic emerging industries in China's 12th Five-Year Plan³, initial growth was modest. Limited core technology development and inadequate supply chain preparation constrained market expansion. Sales increased from 5,000 units in 2009 to 17,600 units in 2013⁴, reflecting an average annual growth rate of 37%^[4]. The NEV market in China experienced a breakthrough growth phase starting in 2014, driven by technological maturation and supportive policy measures. These policies included tax exemptions on NEV purchases, manufacturer subsidies, and the standardization of charging infrastructure. Sales surged from 73,000 units in 2014 to 580,000 units in 2017,

¹National Bureau of Statistics of China (NBSC). China Energy Statistical Yearbook, 1991–2021 National Bureau of Statistics of China & National Energy Administration of China.

<http://www.stats.gov.cn/>.

²IEA. Global EV Data Explorer.

<https://www.iea.org/data-and-statistics/data-tools/global-ev-data-explorer>.

³National Development and Reform Commission. Outline of the Twelfth Five-Year Plan for National Economic and Social Development of the People's Republic of China [EB/OL]. (2011-03-16). https://www.gov.cn/zhuanti/2011-03/16/content_2623428.htm.

⁴China Association of Automobile Manufacturers. <http://www.caam.org.cn/>.

representing an average annual growth rate of 99.5%⁵. However, when China began phasing out financial subsidies for NEV during 2018–2020^[5], the market growth rate decelerated significantly to approximately 2.3% annually¹. By 2023, China’s NEV industry has maintained robust growth, driven by market demand, macroeconomic policies, and industry-wide efforts. China has led global NEV production and sales for nine consecutive years⁶, with a target to reach approximately 20% of total new vehicle sales by 2025⁷. The government’s support mechanism has evolved from manufacturer-focused subsidies to consumer-oriented incentives^[6], effectively stimulating NEV consumption. Furthermore, the tax exemption policy for NEV purchases will be extended and optimized through 2027, with projected tax savings of 520 billion yuan for the 2024–2027 period⁸. With the continuous strengthening of policy support, the sales of NEVs have shown a clear upward trend. However, sales forecasting is influenced by various factors, particularly the combined effects of market demand, policy changes, technological advancements, and consumer preferences. Moreover, NEV sales exhibit significant seasonal fluctuations and complex market volatility, leading to large errors and low accuracy in traditional forecasting models.

To bridge this critical gap, we innovatively develop an integrated modeling framework that systematically combines advanced variable selection methodologies with the powerful machine learning algorithm, XGBoost, specifically tailored for the complex, high-dimensional nature of NEV sales. The combined use of feature extraction and XGBoost allows for an efficient and accurate forecast of NEV sales by first reducing the number of input variables to only the most relevant ones and then using a powerful machine learning model to capture complex patterns in the data. This combination addresses the challenges posed by large and complex datasets, improves the model’s predictive power, and ensures that the forecast accounts for both the most influential features and the underlying nonlinear relationships between those features and NEV sales. Based on the findings, the integrated model significantly enhances the forecasting performance of NEV sales. This provides a theoretical basis for optimizing the structure of China’s NEV industry and for the formulation of related policies. It holds substantial practical

⁵IEA. Energy Statistics Data Browser.

<https://www.iea.org/data-and-statistics/data-tools/global-ev-data-explorer>.

⁶National Development and Reform Commission. Responsible comrade of National Development and Reform Commission attends press conference at the State Council Information Office to explain the macroeconomic situation and policies and answer reporters’ questions. (2024-01-19).

https://www.ndrc.gov.cn/fzggw/wld/zyhd/202401/t20240119_1363542.html.

⁷State Council. Comprehensive work program for energy conservation and emission reduction in “The Fourteenth Five-Year Plan”. (2022-01-24)

https://www.gov.cn/zhengce/content/2022-01/24/content_5670202.htm.

⁸The ministry of finance, The state administration of taxation, Ministry of Industry and Information Technology of the People’s Republic of China. Announcement on the Extension and Optimization of Vehicle Purchase Tax Reduction and Exemption Policy for New Energy Vehicles. (2023-06-19). https://www.gov.cn/zhengce/zhengceku/202306/content_6887734.htm.

significance in driving the high-quality development of the NEV industry, alleviating China's energy security pressures, promoting green and low-carbon development, and achieving the "dual carbon" goals. The research framework is shown in Figure 1.

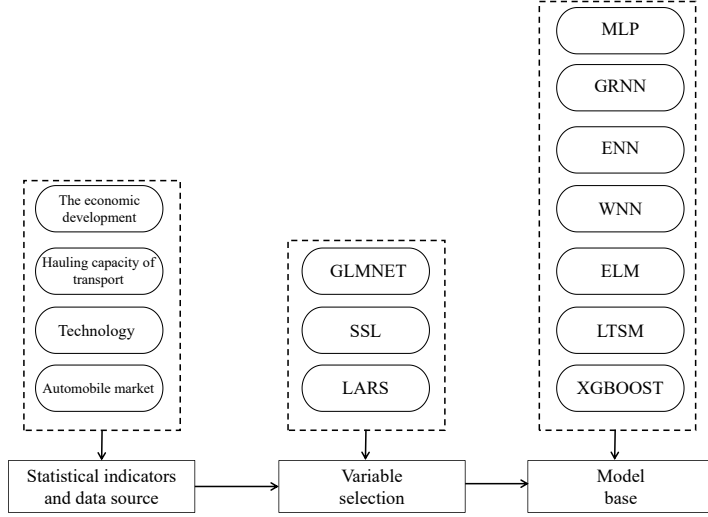


Figure 1 Variable selection-XGBoost integrated model research framework

This paper makes three main marginal contributions distinct:

Firstly, we develop a novel integrated approach that explicitly combines core variable selection methods (GLMNET, SSL, and LARS) with the XGBoost algorithm. This methodological integration not only identifies the most influential factors systematically, but also effectively manages multicollinearity and enhances model interpretability, ultimately yielding superior forecasting accuracy. Unlike previous studies that often overlook the explicit variable selection step or use simpler models, our approach leverages the strengths of regularization and Bayesian-inspired selection techniques to robustly extract the most critical predictors of NEV sales.

Secondly, our research incorporates the Baidu Index as a novel behavioral indicator, reflecting market attention and consumer sentiment towards NEVs. This inclusion innovatively explores the predictive power of internet search patterns, an approach underutilized in current forecasting models for the NEV industry. By quantifying online search behaviors and public attention, our method provides richer behavioral insights and allows for improved responsiveness to emerging market trends, thus significantly enhancing prediction performance.

Thirdly, our study provides an extensive comparative empirical validation, benchmarking the proposed variable selection-XGBoost integrated models against multiple established forecasting methodologies — including autoregressive models (AR, ARMA), neural network variants (MLP, GRNN, ENN, WNN, ELM, LSTM), and conventional XGBoost approaches without variable selection. Empirical results demonstrate conclusively that our proposed integrated model outperforms benchmark models both in-sample and out-of-sample, showcasing statistically significant reductions in forecasting errors as measured by RMSE and MAPE metrics. Furthermore, by

implementing cross-validation-based hyperparameter optimization, we ensure robust model performance and generalizability across forecasting horizons.

The structure of this paper is organized as follows: Section 2 presents a literature review, examining both the factors influencing NEV sales and the forecasting methodologies used in previous studies. Section 3 introduces the variable selection-XGBoost integrated model and provides a detailed description of the model construction. Section 4 applies the model to forecast China's NEV sales, compares its performance with traditional single benchmark models and other integrated models, and presents model optimization to further improve prediction accuracy. Section 5 concludes the paper.

2 Literature Review

We make a simple review from two aspects of influencing factors and forecasting methods as follow.

2.1 Influencing Factor

As the world's leading NEV producer and consumer market, China's NEV sales are influenced by multiple factors. A large number of research has shown that NEV sales are related to China's CPI (consumer price index), transportation and communication, fuel for transportation vehicles; China's PPI (producer price index): automotive manufacturing, vehicle prices, Li Keqiang Index, consumer confidence index, national subsidy policies, highway passenger turnover, search indices, vehicle production, fuel vehicle sales, and charging infrastructure^[7–10]. Ma, et al.^[11] analyzed Chinese consumer online behavior using big data and text mining techniques, identifying electric vehicle pricing, vehicle classification, and powertrain types as the most significant factors affecting consumer response. Similarly, Chen, et al.^[12] employed cognitive behavioral models to confirm that price is a crucial factor in NEV purchasing decisions. Using data mining and deep learning approaches, Wang, et al.^[8] found strong correlations between NEV sales and national policies, charging infrastructure, and safety awareness. These findings align with Qian and Grisolia's mixed logit model results^[13], which also emphasized the importance of government policies and charging infrastructure. Lin and Wu^[13] developed an electric vehicle purchase intention model, demonstrating significant influences from both attitudinal factors (price, government subsidies, vehicle performance, environment) and demographic characteristics (gender, age) on consumer purchasing decisions. Additionally, multiple researches have highlighted charging infrastructure as a critical determinant of NEV market potential^[13,15–17]. Consequently, policy incentives driving charging infrastructure development are crucial for unlocking NEV consumption potential.

2.2 Forecast Method

NEV sales forecasting presents a complex nonlinear problem. Traditional research methods can be categorized into three main methodological streams. The Bass diffusion model represents one significant approach to market prediction. Liu, et al.^[18] utilized historical hybrid vehicle

data to construct a Bass model for forecasting China's NEV market performance. Econometric modeling approaches have also been developed to address forecasting challenges. Du, et al.^[19] advanced this methodology by using an extended logistic model to predict Chinese vehicle ownership, while examining future electric vehicle development paths through environmental and energy security constraints. Grey theory models, particularly adaptations of the classical GM(1,1) model with enhanced original sequence smoothness, have emerged as a prevalent forecasting methodology. He, et al.^[20] improved original sequence smoothness by incorporating cumulative and translation transformations with buffer operators into the GM(1,1) model for forecasting China's NEV production and sales. Building on this approach, Ding, et al.^[21] proposed a new information buffer operator, developing a nonlinear grey Bernoulli composite prediction model validated through Norwegian NEV sales data. More recently, Liu, et al.^[22] introduced an improved grey power model with multiple parameters, utilizing the differential form and derived discrete form of the model (denoted as DTGPM) for NEV sales forecasting. Zeng, et al.^[23] constructed a three-parameter grey prediction model with an intelligent buffer operator to enhance forecasting performance under data volatility and limited sample conditions in China's NEV market.

Traditional forecasting models have shown limited effectiveness in predicting NEV sales due to small sample sizes and complex influencing factors. These models struggle to eliminate data fluctuations, resulting in significant prediction errors and low accuracy. Machine learning techniques have addressed these limitations by effectively handling nonlinear data and short-term predictions. Over recent decades, researchers have incorporated an increasing number of factors and models into analyses. Studies have demonstrated that integrating multiple distinct factors can enhance NEV sales forecasting accuracy. However, most existing research directly selects influencing factors and prediction models without systematic identification of core impact variables. While some machine learning methods include variable selection processes, these are typically embedded within prediction models primarily to ensure model robustness^[24,25]. The importance of variable selection has grown with the expansion of data and quantitative indicators. Pre-forecasting variable screening can effectively reduce the impact of redundant variables on model prediction performance^[24,26–29]. However, core variable screening methods remain underutilized in NEV sales forecasting research.

3 Methodology

3.1 The Elastic-net Regularized Generalized Linear Models (GLMNET)

Zou and Hastie^[30] introduced the elastic net method for variable selection, positioning it as the optimal shrinkage method for addressing multicollinearity and variable screening while minimizing accuracy loss. Simulation results demonstrate that the elastic net achieves superior prediction accuracy compared to LASSO. Similar to LASSO, the elastic net performs automatic variable selection and continuous shrinkage while maintaining the ability to select a group of correlated variables. This dual functionality achieves ridge regression's goal of selecting important

features while eliminating features with minimal impact on dependent variables, particularly when the sample size n is smaller than the number of predictors p . The specific formulation can be found in Zou and Hastie's research^[30]. We use the GLMNET method, a package that fits generalized linear models through penalized maximum likelihood, computing regularization paths for LASSO or elastic net penalties across a grid of regularization parameter λ values. GLMNET offers exceptional computational efficiency, effectively utilizing the sparsity of input matrix $\mathbf{X}$. It is applicable to linear, logistic, multinomial, Poisson, and Cox regression models, as well as multi-response linear regression.

$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^N \omega_i l(y_i, \beta_0 + \boldsymbol{\beta}' \mathbf{X}_i) + \lambda[(1 - \alpha)\|\boldsymbol{\beta}\|_2^2/2 + \alpha\|\boldsymbol{\beta}\|_1]. \quad (1)$$

Equation (1) represents a grid of values of λ spanning the entire range, where $l(\mathbf{y}, \boldsymbol{\eta})$ denotes the negative log-likelihood contribution for the i -th observation. For Gaussian distribution, it takes the form $\frac{1}{2}(y - \eta)^2$. The Elastic Net penalty is controlled by the parameter α . It combines LASSO (when $\alpha = 1$) and Ridge regression (when $\alpha = 0$), narrowing the gap between them. The tuning parameter α decides how much weight is given to each type of penalty. Another tuning parameter controls the overall strength of the penalty.

3.2 Least Angle Regression (LARS)

Efron, et al.^[31] introduced the least angle regression selection (LARS) algorithm for high-dimensional linear regression problems. The algorithm's core principle involves constructing a linear regression model by iteratively shrinking coefficients toward the target direction, minimizing the angle between residuals and the most correlated feature variables. The LARS algorithm begins by setting all variable coefficients to zero and identifies the feature variable $\mathbf{X}_{j1}$ most correlated with the dependent variable. It then proceeds along this direction until finding another feature variable $\mathbf{X}_{j2}$ with equal correlation to the current residuals γ . The algorithm continues along the angle bisector between $\mathbf{X}_{j1}$ and $\mathbf{X}_{j2}$ until encountering a third feature variable $\mathbf{X}_{j3}$ with equal correlation to the current residuals. This process continues along the minimum angle direction until feature-residual correlations become insignificant. The coefficient update $\hat{\mathbf{C}}_{kj}$ and update vector $\hat{\mathbf{u}}_k$ of the feature j at each step are as follows:

$$|\hat{\mathbf{C}}_{kj}| = |\mathbf{X}_j'(\hat{\mathbf{y}} - \hat{\boldsymbol{\mu}}_{k-1})|, \quad (2)$$

$$\hat{\boldsymbol{\mu}}_k = \hat{\boldsymbol{\mu}}_{k-1} + \frac{\hat{\gamma}_k}{\bar{\gamma}_k}(\bar{\mathbf{y}}_k - \hat{\boldsymbol{\mu}}_{k-1}). \quad (3)$$

At step k , where $\hat{\mathbf{C}}_k = \max\{|\hat{\mathbf{C}}_{kj}|\}$ represents the maximum correlation, the minimum residual value $\hat{\gamma}_k$ is necessarily smaller than the step size $\bar{\gamma}_k$, ensuring $\hat{\boldsymbol{\mu}}_k$ closer proximity to $\hat{\boldsymbol{\mu}}_{k-1}$ than $\bar{\mathbf{y}}_k$. Compared to traditional forward selection and forward stepwise algorithms, LARS demonstrates superior efficiency while preserving information correlation and effectively addresses the high computational complexity and slow feature selection issues associated with Lasso. The detailed mathematical formulation can be found in Efron, et al.^[31].

3.3 Spike-slab Lasso (SSL)

While model averaging is recognized as an effective approach for handling variable selection and hypothesis testing, this perspective holds true only in Bayesian settings where model probabilities are a prerequisite. Ročková and George^[32] advanced beyond the separable penalty framework based on full Bayesian spike and slab functions by introducing a novel adaptive penalty function. The SSL model achieves multiple adjustments through an inseparable Bayesian penalty function that leverages inter-coordinate strength and adapts to ensemble sparsity information. It employs a Laplace mixture sequence with increasing spike penalty function λ_0 while maintaining a small fixed constant. This distinguishes it from LASSO, which uses a single Laplace prior sequence with increasing penalty function λ . Furthermore, the methodology re-deploys the expectation-maximization approach to Bayesian variable selection (EMVS, an efficient EM algorithm for extending Bayesian models with Gaussian spike-slab mixtures) for SSL priors^[30]. This approach reduces biased estimates through automatic threshold variable selection while achieving faster convergence. The detailed mathematical formulation can be found in Rocková and George's research^[32].

3.4 XGBoost

Chen and Guestrin^[33] introduced the XGBoost algorithm, which represents an efficient implementation of the gradient boosting concept. The gradient boosting method iteratively constructs classifiers that minimize loss along the steepest gradient direction. XGBoost builds upon decision trees by using an ensemble strategy, utilizing gradient boosting to progressively reduce the loss of previously generated decision trees while generating new trees for the model. This approach ensures robust final decision-making, as illustrated in Figure 2.

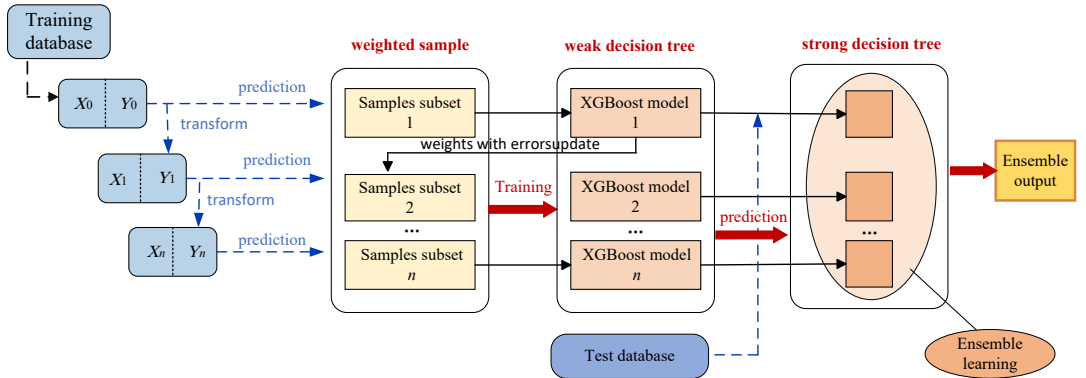


Figure 2 Structure of XGBoost algorithm

XGBoost adds one tree in each iteration, forming a linear combination of K trees:

$$\hat{y}_i = \sum_{k=1}^K f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F}, \quad (4)$$

where $\mathcal{F}$ represents the function space containing all trees, and $f_k(\mathbf{x}_i)$ denotes the weight of

the leaf node where the i -th sample is classified in the k -th tree. The objective function of the XGBoost model is defined as:

$$\text{Obj}_j = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k), \quad (5)$$

where n represents the number of samples, $l(y_i, \hat{y}_i)$ denotes the loss function measuring the difference between prediction y_i and target $\hat{y}_i$, K indicates the number of trees, $\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2$ is the regularization term $\Omega(f_k)$ for the k -th tree controlling model complexity, γ denotes the penalty for increased tree complexity, T represents the number of leaf nodes, λ is the weight of the Ridge regularization term controlling leaf node weight magnitude, ω represents the leaf node weight vector.

Feature importance metrics evaluate the significance of each feature within its feature set. XGBoost bases its decision tree construction on three key metrics: FScore (feature split frequency), AverageGain (average feature gain), and AverageCover (average feature coverage). The metrics are defined as:

$$\text{FScore} = |\mathbf{X}|, \quad (6)$$

$$\text{AverageGain} = \frac{\sum \text{Gain}_x}{\text{FScore}}, \quad (7)$$

$$\text{AverageCover} = \frac{\sum \text{Cover}_x}{\text{FScore}}, \quad (8)$$

where $\mathbf{X}$ represents the set of leaf nodes where the target feature is classified, Gain (Gain = $\sum_{\text{left}} \omega + \sum_{\text{right}} \omega - \sum_{\text{nosplit}} \omega$) denotes the node gain value during splitting at each leaf node in X , Cover indicates the number of samples falling into each node in X . Detailed mathematical formulations can be found in Chen and Guestrin^[33].

4 Empirical Analysis

4.1 Dataset

Based on existing domestic and international research literature, this study takes the sales of NEV in China as the dependent variable and selects 12 initial feature variables from the perspectives of economic development, automotive market, transportation capacity, and technology. Due to the limited sample size before 2016 and to ensure data completeness, the sample period is selected from January 2016 to December 2023. The data is logarithmically transformed to ensure stationarity and reduce heteroscedasticity.

In terms of economic development, we select economic indicators highly related to NEV sales, including China's CPI for transportation and communication (fuel for vehicles), China's PPI for automotive manufacturing, gasoline prices, and vehicle purchase tax. For transportation capacity, we use China's highway passenger turnover as the indicator. Technology, which is also a key factor influencing NEV purchasing decisions^[9], is represented by the Baidu Index. The automotive market is represented by indicators such as the sales and production volumes of

conventional fuel vehicles, the number of public charging stations, and the production volume of NEV in China^[8]. Specific descriptions of the variables and their data sources are provided in Table 1.

The Baidu Index⁹ measures weighted search frequencies of keywords in Baidu web searches^[34], serving as an information search platform that reflects user attention to different keywords over time. Research has shown a positive correlation between online search behavior and offline purchasing behavior, making it valuable for purchase behavior prediction^[35].

The keyword selection process began with an extensive literature review, combining direct selection and range selection methods^[36]. This process yielded ten NEV-related keywords: ‘BYD

Table 1 Initially selected feature variables and their explanations

First-level indicator	Second-level indicator	Variable	Unit	Symbol	Data sources
Implicit variable		China’s new energy vehicle sales	vehicles/month	Y	China Association of Automobile Manufacturers (CAAM)
Economic development	Economic indicator	China’s CPI for vehicle fuel: transportation and communication category	same month last year = 100	X_1	Wind
		China’s PPI: automobile	%	X_2	Wind
		Gasoline prices in China	yuan/liter	X_3	East Money
		Li Keqiang Index	%	X_4	Wind
		Consumer confidence: National Index	-	X_5	National Bureau of Statistics of China (NBS)
	Economic policy	China’s tax revenue: Vehicle purchase tax	billions of dollars/month	X_6	Wind
Transportation capacity		China’s road passenger turnover	person/kilometer	X_7	National Bureau of Statistics of China (NBS)
Technology		Baidu Index	-	X_8	Baidu Index official website
Automotive market	Fuel vehicle market	China fuel vehicle sales	vehicles/month	X_9	China Association of Automobile Manufacturers (CAAM)
		China’s fuel vehicle production	vehicles/month	X_{10}	China Association of Automobile Manufacturers (CAAM)
	New energy vehicle market	China’s public charging station inventory	units/month	X_{11}	Wind
		China’s NEV production	million vehicles/month	X_{12}	China Association of Automobile Manufacturers (CAAM)

⁹<https://index.baidu.com/v2/index.html#/>.

Electric Car Price’ ‘BYD Company’ ‘BYD Han’ ‘BYD Qin Price’ ‘Electric Vehicle’ ‘Electric Vehicle Network’ ‘Chery QQ’ ‘Tesla’ ‘New Energy Technology’, and ‘NEV Purchase Tax’. To ensure objectivity, we validated our keyword selection using two mainstream long-tail keyword mining tools: Aizhan¹⁰ and Ciku¹¹. This validation confirmed that our selected keywords had significantly higher search volumes than related alternatives, effectively reflecting public attention and demand. We collected daily aggregated mobile and PC search data for all selected keywords nationwide from January 2016 to April 2024, which we then converted to monthly frequency data. Finally, we synthesized a NEV keyword Baidu Index using principal component analysis (PCA). SPSS testing revealed a KMO test value of 0.818 and Bartlett’s sphericity test approximate chi-square value of 590.975, with p -value < 0.05 (Table 2), indicating statistical significance. Table 3 shows two extracted principal components, F_1 and F_2 , with a cumulative variance contribution of 64.097%, capturing the majority of information. The ten keyword indicators are recorded sequentially as $W_1, W_2, \dots, W_m$ ($m = 10$); eigenvalues are denoted by λ_i ($i = 1, 2, \dots, m$), as shown in Table 3 and eigenvectors are denoted by $a_i = (a_{1i}, a_{2i}, \dots, a_{mi})$, as shown in Table 4. Each principal component’s corresponding contribution rate is $\lambda_i / \sum_{i=1}^2 \lambda_i$. Using Equation (9), we calculated values for both principal components, then weighted them to compute the final composite principal component value F_x , representing the synthesized Baidu

Table 2 KMO and bartlett’s test of sphericity

KMO measure of sampling adequacy		0.818
Bartlett’s test of sphericity	approximate chi-square	590.975
	degrees of freedom	45
	statistical significance	<0.001

Table 3 Total variance explained

Component	Initial eigenvalue			Extracted sum of squared loadings		
	Total	Percentage of variance	Cumulative %	Total	Percentage of variance	Cumulative %
1	5.057	50.568	50.568	5.057	50.568	50.568
2	1.353	13.529	64.097	1.353	13.529	64.097
3	0.902	9.025	73.122			
4	0.703	7.026	80.148			
5	0.587	5.866	86.014			
6	0.486	4.862	90.876			
7	0.435	4.351	95.227			
8	0.225	2.247	97.474			
9	0.168	1.679	99.154			
10	0.085	0.846	100.000			

¹⁰<https://aizhan.com/>.

¹¹<http://www.ciku5.com/>.

Table 4 Component matrix

	Component	
	1	2
W_1	-0.906	0.288
W_2	0.659	0.567
W_3	0.759	0.102
W_4	-0.708	0.354
W_5	0.759	0.009
W_6	-0.860	0.266
W_7	0.552	0.637
W_8	0.214	0.540
W_9	0.642	-0.026
W_{10}	-0.804	0.209

Index for the keywords, as shown in Equation (10).

$$F_n = a_{1n} \cdot W_1 + a_{2n} \cdot W_2 + \cdots + a_{mn} \cdot W_m \quad (n = 1, 2), \quad (9)$$

$$F_W = \sum_{i=1}^2 \frac{\lambda_i}{\sum_{i=1}^2 \lambda_i} F_i. \quad (10)$$

4.2 Analysis of Core Factors

To ensure data stationarity, we applied logarithmic transformation to all 12 statistical indicators. Table 5 presents the core factor extraction results. The analysis shows that the SSL method identified the most variables (6), including China's CPI for vehicle fuel: transportation and communication category (X_1), Li Keqiang Index (X_4), China's tax revenue: Vehicle purchase tax (X_6), China's fuel vehicle production (X_{10}), China's public charging station inventory (X_{11}), and NEV production volume (X_{12}). The LARS method selected five variables: China's gasoline prices (X_3), Li Keqiang Index (X_4), China's tax revenue: Vehicle purchase tax (X_6), Baidu Index (X_8), and China's NEV production (X_{12}). The GLMNET method selected the fewest variables (3), which form a subset of the variables identified by the SSL method.

Table 5 Selected key features by GLMNET, SSL and LARS method

Methods	Selected feature ID
GLMNET	X_{10}, X_{11}, X_{12}
SSL	$X_1, X_4, X_6, X_{10}, X_{11}, X_{12}$
LARS	$X_3, X_4, X_6, X_8, X_{12}$

An integrated analysis of the variable selection results from the three methods reveals several key insights:

1) Economic development

Our analysis of economic development factors reveals that, collectively, the SSL and LARS methods selected significant variables including the China's CPI for vehicle fuel: transportation and communication category (X_1), China's gasoline prices (X_3), Li Keqiang Index (X_4), and vehicle purchase tax revenue (X_6).

The transportation and communication component of China's CPI encompasses various transportation-related costs, including fuel price fluctuations. From an energy security perspective, large-scale NEV adoption effectively reduces dependence on foreign oil supplies, with rising gasoline prices driving NEV development^[37]. From an economic perspective, automobiles and gasoline are complementary goods. Higher oil prices reduce fuel demand, subsequently decreasing traditional vehicle sales, suggesting that gasoline price regulation can influence NEV sales growth.

The Li Keqiang Index serves as an economic indicator, incorporating multiple macroeconomic metrics closely tied to economic activity^[38]. Its assessment of overall economic conditions significantly influences consumer confidence, purchasing power, and NEV market consumption patterns.

Tax incentives, particularly purchase tax exemptions, directly impact consumer vehicle costs, positively influencing NEV adoption and application^[39].

2) Technology

The LARS method identified the Baidu Index as significant, with Baidu being the world's largest Chinese search engine. Both GLMNET and SSL methods selected (X_{10}) and China's public charging station inventory (X_{11}) as key variables. Notably, China's NEV production (X_{12}) emerged as a significant feature across all three methods.

The development of charging infrastructure represents a crucial factor influencing consumer acceptance of NEV. Supporting facilities have become a primary source of consumer confidence and motivation for NEV adoption^[40,41].

3) Automotive market

From an economic perspective, NEV and fuel vehicles function as substitutes to some degree, a relationship confirmed by Li, et al.^[15] in their study of optimal investment models for NEV charging infrastructure under public-private partnerships. Consequently, fuel vehicle production shows a negative correlation with NEV sales. Following market supply-demand principles, increased NEV production promotes economies of scale, reducing production costs and incentivizing supply chain and marketing network expansion. This enhanced capacity better meets consumer demand, driving sales growth.

4.3 Forecasting Performance Evaluation

We selected 9 benchmark models to validate the predictive capability of our proposed variable selection-XGBoost integrated model. These include seven multivariate models (MLP, GRNN, ENN, WNN, ELM, LSTM, and XGBoost) and two univariate models (AR and ARMA). A horizontal comparison was conducted to evaluate the prediction performance of our variable

selection-XGBoost integrated model against these benchmark models.

Prediction accuracy was assessed using two evaluation criteria (RMSE and MAPE) for both in-sample and out-of-sample prediction performance, as shown in Table 6. The formulas are as follows:

$$\text{RMSE} = \sqrt{\frac{1}{T_f} \sum_{t=T_e+1}^T (y_t - \hat{y}_t)^2}, \quad (11)$$

Table 6 Forecasting performance

Types	Model	In-sample		Out-of-sample	
		RMSE	MAPE	RMSE	MAPE
Univariate selection	AR	20.19%	44.48%	28.65%	29.80%
	ARMA	20.17%	44.20%	17.18%	17.27%
Without variables selection	MLP	9.79%	19.58%	23.14%	23.40%
	GRNN	8.69%	18.51%	18.66%	15.92%
	ENN	11.11%	20.60%	15.62%	16.17%
	WNN	7.44%	13.69%	18.04%	16.85%
	ELM	9.75%	15.06%	22.23%	20.76%
	LTSM	10.76%	1.20%	41.67%	4.85%
	XGBoost	6.34%	1.06%	10.48%	1.43%
GLMNET	MLP	7.78%	10.24%	16.86%	16.54%
	GRNN	8.07%	15.83%	9.03%	8.96%
	ENN	8.11%	16.45%	9.54%	10.29%
	WNN	6.18%	13.32%	7.60%	7.91%
	ELM	7.06%	13.56%	17.58%	18.05%
	LTSM	5.55%	0.81%	33.60%	3.55%
	XGBoost	4.96%	0.77%	7.14%	0.90%
SSL	MLP	5.29%	11.43%	17.41%	18.40%
	GRNN	6.96%	15.32%	12.41%	11.63%
	ENN	7.97%	16.60%	7.55%	8.04%
	WNN	5.83%	10.90%	7.88%	7.01%
	ELM	7.25%	14.20%	17.91%	15.97%
	LTSM	5.09%	0.70%	40.91%	4.53%
	XGBoost	4.93%	0.77%	7.04%	0.89%
LARS	MLP	9.24%	17.50%	9.87%	8.69%
	GRNN	8.57%	18.44%	10.13%	9.83%
	ENN	8.88%	17.62%	8.33%	8.77%
	WNN	5.56%	10.31%	11.91%	11.20%
	ELM	5.52%	3.96%	18.92%	17.27%
	LTSM	8.18%	1.17%	29.87%	4.31%
	XGBoost	5.41%	0.93%	7.79%	1.02%

$$\text{MAPE} = \sqrt{\frac{1}{T_f} \sum_{t=T_e+1}^T \left| \frac{y_t - \hat{y}_t}{y_t} \right|}, \quad (12)$$

where T_e is the training sample dataset; T_f is the prediction sample dataset; the full sample dataset is $T = T_e + T_f$.

Analyzing the prediction results in Table 6 reveals several key findings.

Our proposed variable selection-XGBoost integrated model demonstrates superior performance in both in-sample and out-of-sample predictions compared to univariate models and models without core factor extraction, as measured by RMSE and MAPE metrics. Among the variable selection-XGBoost integrated models, the SSL-XGBoost integration shows the best performance, followed by GLMNET-XGBoost and LARS-XGBoost. For one-step-ahead in-sample predictions, the SSL-XGBoost integrated model achieved RMSE and MAPE values of 4.93% and 0.77% respectively, showing improvements of 15.26 and 43.71 percentage points compared to the AR model, and 1.41 and 0.29 percentage points compared to the XGBoost model without variable selection. For one-step-ahead out-of-sample predictions, the SSL-XGBoost integrated model achieved RMSE and MAPE values of 7.44% and 0.94%, demonstrating improvements of 21.21% and 28.86% compared to the AR model, and 3.04 and 0.49 percentage points compared to the XGBoost model without variable selection.

These results highlight two significant findings: Firstly, the variable selection-XGBoost integrated models significantly improve prediction accuracy compared to both univariate models and machine learning models without variable selection; second, the SSL method's selection of the largest number of core variables among the three variable selection models suggests that the number of core variables influences prediction outcomes.

4.4 Model Optimization

In machine learning, selecting appropriate hyperparameter combinations is crucial for model performance. While traditional methods involve manual hyperparameter adjustment and performance evaluation, this approach is time-consuming and error-prone. A more efficient approach employs k -fold cross-validation for automatic selection of optimal hyperparameter combinations. The k -fold cross-validation process involves the following steps:

- 1) Randomly dividing the dataset into k non-overlapping subsets of equal size.
- 2) Using $k - 1$ subsets as the training set and the remaining (held out) subset as the test set.
- 3) Repeating this process k times, using a different subset as the test set in each iteration.
- 4) Training k models, calculating test errors for each model on its respective test set, and averaging these k test errors to obtain the cross-validation error.

Before performing cross-validation, a grid was defined to specify the hyperparameters to be tuned and their respective search ranges, as shown in Table 7.

4.4.1 Model Algorithm Optimization Steps

We conducted optimal parameter searches for data from 2016–2023 using four approaches: XGBoost without variable selection, GLM-XGBoost, SSL-XGBoost, and LAR-XGBoost. Using

the Grid Search CV function for grid search, we selected a 3-fold cross-validation based on the dataset size. The optimal parameter combinations are presented in Table 8.

Table 7 Hyperparameter grid for cross-validation

Hyperparameter	Scope of adjustment
n_estimators	[100, 200, 300]
max_depth	[3, 5, 7, 9]
learning_rate	[0.01, 0.1, 0.2]
subsample	[0.8, 0.9, 1.0]
colsample_bytree	[0.8, 0.9, 1.0]

Table 8 Optimal parameters

Model	Hyperparameter	In-sample	Out-of-sample
No variable selection-XGBoost	colsample_bytree	1.0	1.0
	learning_rate	0.1	0.2
	max_depth	5.0	7.0
	n_estimators	300	300
	subsample	0.8	0.8
GLM-XGBoost	colsample_bytree	1.0	1.0
	learning_rate	0.01	0.1
	max_depth	5.0	7.0
	n_estimators	300	100
	subsample	0.8	1.0
SSL-XGBoost	colsample_bytree	1.0	1.0
	learning_rate	0.1	0.2
	max_depth	3.0	5.0
	n_estimators	100	100
	subsample	0.9	1.0
LARS-XGBoost	colsample_bytree	1.0	1.0
	learning_rate	0.1	0.1
	max_depth	3.0	7.0
	n_estimators	100	100
	subsample	0.9	1.0

4.4.2 Model Performance Superiority

We evaluated four datasets (XGBoost without variable selection and three variable-screening approaches) using the XGBoost prediction model, yielding the following results:

For in-sample predictions, the optimized models achieved RMSE values of 5.00%, 4.79%, 3.91%, and 4.65%. Compared to the non-optimized models' RMSE values of 6.34%, 4.96%, 4.93%, and 5.31%, the optimization improved accuracy by 0.17% to 1.34%. The optimized

models' MAPE values were 0.78%, 0.75%, 0.52%, and 0.74%, compared to the non-optimized models' values of 1.06%, 0.77%, 0.77%, and 0.93%, showing improvements ranging from 0.02% to 0.28%.

For out-of-sample predictions, the optimized models achieved RMSE values of 7.47%, 7.11%, 7.05%, and 7.37%. Compared to the non-optimized models' RMSE values of 10.48%, 7.14%, 7.44%, and 7.79%, the optimization improved accuracy by 0.03% to 3.01%. The optimized models' MAPE values were 0.99%, 0.89%, 0.87%, and 0.97%, compared to the non-optimized models' values of 1.43%, 0.90%, 0.94%, and 1.02%, showing improvements ranging from 0.01% to 0.44%.

5 Conclusion

In this paper, we propose an integrated modeling framework combining variable selection and machine learning methods to forecast China's NEV sales. Firstly, we employed three representative feature extraction methods, GLMNET, SSL, and LARS, that are suitable for China's NEV market data characteristics to evaluate and screen core influencing factors. Secondly, we compared the integrated model's prediction performance against widely used benchmark methods (MLP, GRNN, ENN, WNN, ELM, LSTM, and XGBoost). Additionally, we incorporated the Baidu Index as a behavioral indicator reflecting market attention and consumer sentiment to explore the predictive potential of internet search patterns for NEV sales trends. Finally, we optimized the model through cross-validation to enhance prediction accuracy.

Variable selection results show that the SSL method identified the most variables (6), followed by LARS (5), and GLMNET (3). Variables selected multiple times include the Li Keqiang Index (X_4), China's tax revenue: Vehicle purchase tax (X_6), China's fuel vehicle production (X_{10}), China's public charging station inventory (X_{11}), China's NEV production (X_{12}). These findings suggest that economic development drives consumption and automotive sales, with NEV sales influenced by market acceptance and economic vitality. With maturing NEV technology, decreasing costs, and continued government tax incentives, NEV market share is expected to grow. Furthermore, NEVs are becoming an increasingly significant segment of the automotive market. Following market operation principles and supply-demand relationships, conventional vehicle production shows a negative correlation with NEV sales, while increased NEV production promotes economies of scale.

The prediction performance results show that variable selection-XGBoost integrated model achieves superior accuracy (RMSE, MAPE) in both in-sample and out-of-sample predictions compared to univariate models and models without core factor extraction. This underscores the importance of selecting core influencing factors before forecasting NEV sales.

Variable selection-XGBoost integrated model demonstrates strong scalability and can be applied to other industries characterized by dynamic consumer behavior, technological change, and evolving policy environments. For example, in the electronics industry — particularly in mobile phones, computers, and other consumer electronics — consumer demand is significantly

influenced by technological innovations and seasonal fluctuations. Additionally, government policies such as subsidies and “old-for-new” programs incentivize consumers to upgrade their devices in exchange for discounts or other incentives. Identifying appropriate variables and applying variable selection-XGBoost integrated model to predict outcomes in other industries presents an intriguing area for further research.

References

- [1] He X, Ou S, Gan Y, et al. Greenhouse gas consequences of the China dual credit policy. *Nature Communications*, 2020, 11(1): 5212.
- [2] Gong X, Lin B Q. Jump risk, structural breaks and forecasting crude oil futures volatility. *Chinese Journal of Management Science*, 2018, 26(11): 11–21.
- [3] Guo J F, Zhang X M, Cao Q, et al. Electric vehicles contribute to China’s energy security and carbon peaking and carbon neutrality. *Proceedings of the Chinese Academy of Sciences*, 2024, 39(2): 397–407.
- [4] Yuan X L, Liu X, Zuo J. The development of new energy vehicles for a sustainable future: A review. *Renewable Sustainable Energy Rev.*, 2015, 42: 298–305.
- [5] Liu B C, Song C Y, Wang Q S, et al. Research on regional differences of China’s new energy vehicles promotion policies: A perspective of sales volume forecasting. *Energy*, 2022, 248: 123541.
- [6] Xu J, Huang X Q, Luo S D. Production decisions of automobile enterprises and the impact of government subsidies based on heterogeneous consumer preferences. *Journal of Industrial Engineering and Engineering Management*, 2024, 38(5): 190–200.
- [7] Rietmann N, Hügler B, Lieven T. Forecasting the trajectory of electric vehicle sales and the consequences for worldwide CO₂ emissions. *Journal of Cleaner Production*, 2020, 261: 121038.
- [8] Wang L, Fu Z L, Guo W, et al. What influences sales market of new energy vehicles in China? Empirical study based on survey of consumers’ purchase reasons. *Energy Policy*, 2020, 142: 111484.
- [9] Featherman M, Jia S J, Califf C B, et al. The impact of new technologies on consumers beliefs: reducing the perceived risks of electric vehicle adoption. *Technological Forecasting and Social Change*, 2021, 169: 120847.
- [10] Hao Y, Peng B, Zou H, et al. How does differentiated subsidy adjustment influence new energy vehicle sales? *Transportation Research Part A: Policy and Practice*, 2025, 195: 104432.
- [11] Ma S C, Fan Y, Guo J F, et al. Analysing online behaviour to determine Chinese consumers’ preferences for electric vehicles. *Journal of Cleaner Production*, 2019, 229: 244–255.
- [12] Chen K, Ren C R, Gu R, et al. Exploring purchase intentions of new energy vehicles: From the perspective of frugality and the concept of “mianzi”. *Journal of Cleaner Production*, 2019, 230: 700–708.
- [13] Qian L X, Grisolia J M, Soopramanien D. The impact of service and government-policy attributes on consumer preferences for electric vehicles in China. *Transportation Research Part A: Policy and Practice*, 2019, 122: 70–84.
- [14] Lin B Q, Wu W. Why people want to buy electric vehicle: An empirical study in first-tier cities of China. *Energy Policy*, 2018, 112: 233–241.
- [15] Li D D, Lin W M, Gao W W, et al. Optimal investment mode of new energy vehicle charging

- infrastructure under Public-Private-Partnership. Chinese Journal of Management Science. doi: 10.16381/j.cnki.issn1003-207x.2023.1477.
- [16] Wang Z H, Zhao C Y, Yin J H, et al. Purchasing intentions of Chinese citizens on new energy vehicles: How should one respond to current preferential policy? Journal of Cleaner Production, 2017, 161: 1000–1010.
- [17] Xiong Y, Wang L. Policy cognition of potential consumers of new energy vehicles and its sensitivity to purchase willingness. Journal of Cleaner Production, 2020, 261: 121032.
- [18] Liu Y Q, Wang M, Wang J Y. Research on China's new energy vehicle market forecast. Research on Economics and Management, 2016, 37(4): 86–91.
- [19] Du Z L, Lin B Q, Guan C X. Development path of electric vehicles in China under environmental and energy security constraints. Resources Conservation and Recycling, 2019, 143: 17–26.
- [20] He L Y, Pei L L, Yang Y H. An optimised grey buffer operator for forecasting the production and sales of new energy vehicles in China. Science of The Total Environment, 2020, 704: 135321.
- [21] Ding S, Li R, Wu S. A novel composite forecasting framework by adaptive data preprocessing and optimized nonlinear grey Bernoulli model for new energy vehicles sales. Communications in Nonlinear Science and Numerical Simulation, 2021, 99: 105847.
- [22] Liu L Y, Liu S F, Wu L F. New energy vehicle sales forecast based on discrete time grey power model. Chinese Journal of Management Science, 2024, 32(1): 106–114.
- [23] Zeng B, Yin F, Wang J, et al. Prediction and rationality analysis of new energy vehicle sales in China with a novel intelligent buffer operator. Engineering Applications of Artificial Intelligence, 2025, 143: 110030.
- [24] Drezga I, Rahman S. Input variable selection for ANN-based short-term load forecasting. IEEE Transactions on Power Systems, 1998, 13(4): 1238–1244.
- [25] Huang T, Fildes R, Soopramanien D. The value of competitive information in forecasting FMCG retail product sales and the variable selection problem. European Journal of Operational Research, 2014, 237(2): 738–748.
- [26] Chen H Y, Ouyang L W, Ren B M. A time series forecasting model with frequency-domain attention — Based variable selection. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2025, 53(5): 150–156.
- [27] Chai J, Lu Q Y, Hu Y, et al. Analysis and Bayes statistical probability inference of crude oil price change point. Technological Forecasting and Social Change, 2018, 126: 271–283.
- [28] Korobilis D. VAR forecasting using Bayesian variable selection. Journal of Applied Econometrics, 2013, 28(2): 204–230.
- [29] Zhang Y J, Ma F, Wang Y D. Forecasting crude oil prices with a large set of predictors: Can LASSO select powerful predictors? Journal of Empirical Finance, 2019, 54: 97–117.
- [30] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, 2005, 67(2): 301–320.
- [31] Efron B, Hastie T, Johnstone I, et al. Least angle regression. The Annals of Statistics, 2004, 32(2): 407–499.
- [32] Rocková V, George E I. The Spike-and-Slab LASSO. Journal of the American Statistical Association, 2018, 112(521): 431–444.
- [33] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. ACM, 2016: 785–794.

- [34] Liu J S, Qiu Z P, Yu C R. Spatiotemporal differences and influencing factors of public attention to doctor–patient disputes in China. *Chinese Journal of Management Science*, 2024, 32(8): 322–334.
- [35] Wang L, Ning Y J, Jia J M. Predicting sales and market share with online search: Evidence from Chinese automobile market. *Journal of Industrial Engineering and Engineering Management*, 2015, 29(4): 56–64.
- [36] Sun Y, Lu B F. A review of researches on the correlation between Internet search and economic behavior. *Management Review*, 2011, 23(7): 72–77.
- [37] Saint Akadiri S, Bekun F V, Sarkodie S A. Contemporaneous interaction between energy consumption, economic growth and environmental sustainability in South Africa: What drives what? *Science of The Total Environment*, 2019, 686: 468–475.
- [38] Zhang X F, Zhang Y Y. A study on the superiority of the Keqiang index in reflecting China's economic reality. *Statistics & Decision*, 2014(22): 30–32.
- [39] Zhou Y, Pan Y. Subsidization versus tax reduction: The policy choice for new energy vehicles under the constraints of transaction costs. *Journal of Management World*, 2019, 35(10): 133–149.
- [40] Xiong Y Q, Li X L. Impacts of supply and demand side policies of new energy vehicle industry on potential consumers. *China Population, Resources and Environment*, 2018, 28(6): 52–62.
- [41] Zhao S J, Zhao F Q, Hao H, et al. The Current situation and countermeasures in Chinese charging infrastructure of new energy vehicles. *Forum on Science and Technology in China*, 2017(10): 97–104.