Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 16, No. 3, 2025

Machine Learning for New Energy Vehicle Sales Prediction and Carbon Emission Impact

Yixuan Li, Zhuoying Zhao, Xinxin Su, Yunquan Song *

College of Science, China University of Petroleum (East China), Qingdao, Shandong, 266580, China
* **Corresponding author**: Yunquan Song (Email: syqfly1980@163.com)

Abstract: China's new energy vehicle market is in a booming stage, with the number of new energy vehicle buyers increasing year by year. The promotion of new energy vehicles is of great significance in reducing carbon emissions, environmental protection and promoting the construction of a clean and low-carbon society. In-depth analysis of market demand through the use of data mining and machine learning algorithms can help companies improve the market competitiveness and promotion of new energy vehicles. At the same time, in-depth study of the reduction effect of new energy vehicles on carbon emissions will help to comprehensively assess their environmental benefits and provide a scientific basis for the government to formulate carbon emission reduction policies and promote the development of the new energy vehicle industry. In this paper, we screened the multifaceted factors affecting the sales of new energy vehicles by using the random forest model, and then used the support vector regression machine to predict the sales of different brands of new energy vehicles in the next two months, and further verified the accuracy and reliability of the model after the effect evaluation on the test set. In addition, this paper establishes a panel regression model of the carbon emission reduction effect of new energy vehicles by introducing control variables and using panel data from 31 provinces, autonomous regions and municipalities in China from 2017 to 2022. The panel regression model verifies that the development of the new energy vehicle industry is conducive to promoting carbon emission reduction from the overall national level.

Keywords: Carbon emission reduction; Feature screening; Machine learning; New energy vehicles.

1. Introduction

With the rapid development of the automobile industry, the problems of environmental pollution and shortage of oil resources have become increasingly serious. The national dual-carbon policy of 'peak carbon' and 'carbon neutrality' has become a necessary way to cope with climate change, which greatly promotes the rapid development of the new energy automobile industry. In 2009, the development of new energy vehicles was officially proposed as a strategy, and in 2020, China explicitly proposed that the sales of new clean energy vehicles in 2025 should exceed 20 per cent of the country's total new vehicle sales. The government has introduced a series of policies to support the development of new energy vehicles, including subsidy policies, charging infrastructure construction, and tax breaks, providing strong support for the promotion of new energy vehicles. [1]

Data show that since 2011, China's new energy vehicle sales show rapid growth. [2] China's new energy vehicle production and sales will reach 9.587 million and 9.495 million in 2023, growing by 35.8% and 37.9% respectively, with a market share of 31.6%. This data shows that the popularity of new energy vehicles is increasing in the Chinese market, and consumer recognition and acceptance of new energy vehicles is gradually increasing. The market competitiveness of new energy vehicles is also increasing, and major automobile manufacturers have increased the research and development and promotion of new energy vehicle product lines, pushing the new energy vehicle industry towards a more prosperous stage of development.

As a clean and low-carbon mode of transport, new energy vehicles will become the main development direction of the automotive industry in the future, playing an important role in achieving sustainable development and building a green and low-carbon society. Therefore, it is crucial to study the sales volume of new energy vehicles and explore their impact on carbon emissions. [3]

The purpose of this paper is to explore the impact of new energy vehicles on carbon emissions by data mining the characteristic indicators and historical sales of different brands of new energy vehicles and using provincial panel data. [4] On the one hand, by studying the sales of new energy vehicles, it is possible to gain a deeper understanding of the impact of various factors on sales, provide reference for the formulation of more targeted policies and market strategies. On the other hand, by analysing the reduction effect of new energy vehicles on carbon emissions, the environmental benefits of new energy vehicles can be assessed more comprehensively, providing more in-depth reference and guidance for the government to formulate emission reduction policies and the development of new energy vehicle industry. [4]

2. Method

2.1. Random Forest

When forecasting the sales volume of new energy vehicles, considering that some of the selected features may not have much relevance to the sales volume, in order to ensure model simplicity, remove redundant or irrelevant influences, and improve the model fitting goodness, we assess the feature importance by using the Random Forest method in order to explore the degree of influence of different features on the sales volume of new energy vehicles.

In the random forest model, the degree of influence of features on the target variable is assessed by calculating the importance value of each feature. The higher the value of importance (VIM) of a feature, it means that the feature plays

a more important role in the prediction of sales. [5] Generally, there are two ways to measure the importance of features: the Gini (Gini) index and the out-of-bag (OOB) error.

Assuming that there are m features, x_1, x_2, ⋯, x_m, in this paper we use the Gini index for evaluation, which is defined as

$$GI_m = \sum_{|K|}^{k=1} \sum_{k' \neq k} P_{mk} P_{mk'}, . \tag{1}$$

where K denotes the number of categories in the random forest, and P_{mk} and $P_{mk'}$ are the proportions of categories k and k', respectively. Let the Gini coefficient of the left node be GI_l and the right node be GI_r, then the score of feature x_j in node m is

$$VIM_{jm}^{(Gini)} = GI_m - GI_l - GI_r, \tag{2}$$

When the nodes appearing in the decision tree are set M, the importance of the feature in the ith tree can be calculated as

$$VIM_{jm}^{(Gini)} = GI_m - GI_l - GI_r, \tag{3}$$

Assuming that there are a total of n trees in the random forest, then

$$VIM_j^{(Gini)} = \sum_{i=1}^{n} VIM_{ij}^{(Gini)}. \tag{4}$$

Finally, normalising the resulting importance scores yields

$$VIM_j = \frac{VIM_j}{\sum_{i=1}^{m} VIM_i}. \tag{5}$$

2.2. Support Vector Regression Machine

Due to the high volatility of the data presented, it is difficult to explore the relationship between the two using conventional methods such as multivariate fitting. Support vector regression machine (SVR), as a machine learning model, is uniquely suited to handle this type of problem. [6] The method is similar to the Support Vector Machine (SVM) in that the goal of SVR is to find a hyperplane such that the interval from the training data points to the hyperplane is as large as possible. This interval is called the 'boundary' and the goal of SVR is to ensure that there are no training data points inside the boundary while minimising the error between the data points outside the boundary and the hyperplane.

For visual representation, the general structure of the SVR is drawn as shown in Fig1

Figure 1. SVR structure diagram

SVR maps the input features into a high-dimensional feature space by means of a kernel trick and builds a linear regression function in that space

$$f(x) = \omega^T \varphi(x) + b, \tag{6}$$

where f(x) is the predicted output value, ω is the feature weight vector, x is the input feature vector, and b is the bias term. SVR usually uses the ε-insensitive loss function, which is defined as follows

$$L(y, f(x)) = \max(0, |y - f(x)| - \varepsilon), \tag{7}$$

where y is the actual value and ε is a preset threshold indicating the model's tolerance to error, i.e., when the gap between the predicted value and the actual value is within ε, the loss is zero; when the gap exceeds ε, the loss is proportional to the gap.

3. Research and analysis

3.1. Data sources

In order to ensure that the data on new energy vehicles is as comprehensive as possible, the data on new energy vehicle models used in this paper comes from China's largest automotive media and automotive service platforms. Our variables include new energy vehicle ownership data and charging pile data from 31 provinces, autonomous regions and municipalities directly under the central government in China from 2013 to 2023. In addition, we screened the top 128 new energy vehicle models including 56 manufacturers or brands in terms of March 2024 sales, and obtained the corresponding 11 vehicle parameters and the monthly sales data of these models from January 2023 to March 2024 from the platform.

Table 1. New energy vehicle model variable description table

Type	Variable Name	Variable Meaning	Unit
Vehicle Parameters	Length	Length of the vehicle	mm
	Width	Width of the vehicle	mm
	Height	Height of the vehicle	mm
	Wheelbase	Length of the vehicle's wheelbase	mm
	Front Track	Distance between the front wheels	mm
	Rear Track	Distance between the rear wheels	mm
Technical Level	Max Speed	Maximum speed of the vehicle	km/h
	Fuel Type	1-Pure Electric, 0-Other	/
	Level	Seven levels including SUV	/
	Max Price	The highest price of the model	10,000 Rmb
	Min Price	The lowest price of the model	10,000 Rmb
Brand	Manufacturer	56 manufacturers	/
Historical Sales	Monthly Sales	Total monthly sales of the model	Units

3.2. Model study

3.2.1. Feature selection

Based on the importance of the features calculated by the Random Forest method, we ranked the 14 features above and displayed the results in Fig2.

After comprehensive analysis, we decided to eliminate some features and finally retained a total of eight indicators to forecast new energy vehicle sales: class, minimum selling price, maximum speed, maximum selling price, manufacturer, month, high, and number of charging piles nationwide.

In addition, we have performed unique thermal coding for the category indicators corresponding to 'manufacturer', 'level' and 'month', i.e., the values of each categorical variable are converted into a binary vector. However, due to the excessive number of categories of these variables, we use Embedding to reduce the dimensionality of these variables by calculating the weight matrix of the Embedding layer, considering that the sparsity of the data may be detrimental to the computation and model training. Assuming that there are m different samples and n text types, the dimensionality reduction technique reduces the matrix of m$\times n$ dimensions to m×k dimensions

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \rightarrow \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mk} \end{bmatrix}.$$

we use Embedding technique to downscale the data after solo thermal coding, setting the vendor down to 3 dimensions, the level down to 1 dimension, and the month down to 2 dimensions to get the final data.

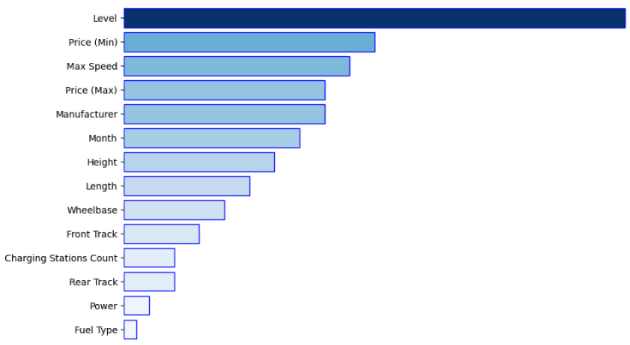

Figure 2. Random forest feature importance score

3.2.2. Model building and optimization

For the objective function

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} w^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right). \tag{8}$$

The constraints are $\begin{cases} y_i - \left(\omega^T \phi\left(x_i\right) - b\right) \le \varepsilon + \xi_i \\ \left(\omega^T \phi\left(x_i\right) + b\right) - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \quad \left(i = 1, 2, \ldots, N\right) \end{cases}$,

where C is the penalty coefficient to balance model complexity and error tolerance, and ξ_i and ξ_i^* are slack variables that allow data points to fall outside the ε interval band.

The Lagrange multipliers α_i, $\alpha_i^* \ge 0$ and μ_i, $\mu_i^* \ge 0$ are introduced to construct the Lagrange function

$$\begin{aligned} \mathcal{L} = {} & \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right) \\ & - \sum_{i=1}^{N}\alpha_i\left[\varepsilon + \xi_i - y_i + \left(\omega^T\varphi\left(x_i\right) + b\right)\right] \\ & - \sum_{i=1}^{N}\alpha_i^*\left[\varepsilon + \xi_i^* + y_i - \left(\omega^T\varphi\left(x_i\right) + b\right)\right] \\ & - \sum_{i=1}^{N}\left(\mu_i\xi_i + \mu_i^*\xi_i^*\right). \end{aligned} \tag{9}$$

The dyadic problem is obtained by taking the partial derivative of ω, b, ξ_i, ξ_i^* and making it zero

$$\min_{a,a^*}\begin{bmatrix} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(a_i - a_i^*\right)\left(a_j - a_j^*\right)K\left(x_i, x_j\right) \\ +\sum_{i=1}^{N}\left(a_i - a_i^*\right)\varepsilon - \sum_{i=1}^{N}\left(a_i - a_i^*\right)y_i \end{bmatrix}. \tag{10}$$

$$st.\sum_{i=1}^{N}\left(a_i - a_i^*\right) = 0, \quad 0 \le a_i \le C, \quad 0 \le a_i^* \le C, \quad i = 1, 2, \cdots, N$$

Here $K(x_i, x_j)$ is the kernel function used to compute the similarity between the input vectors x_i and x_j. The kernel function is a key component of SVR because it allows the model to find linear solutions in high-dimensional spaces, even if the data in the original feature space is nonlinearly differentiable. After solving the dyadic problem, the regression function is

$$f\left(x\right) = \sum_{i=1}^{N}\left(\alpha_i - \alpha_i^*\right)K\left(x_i, x\right) + b. \tag{11}$$

Considering that the Gaussian kernel function (RBF) has a strong nonlinear mapping ability and fewer parameters, which helps to simplify the model, the Gaussian kernel is chosen as the kernel function in this paper. With a step size of 0.5, the optimal parameter C and Gaussian kernel parameter γ are searched by the GridSearchCV function in python.

In order to assess the performance of the SVR model in regression fitting and to measure its fit to the original sample, the mean square error MSE and the coefficient of determination R^2 are used as evaluation metrics in this paper.

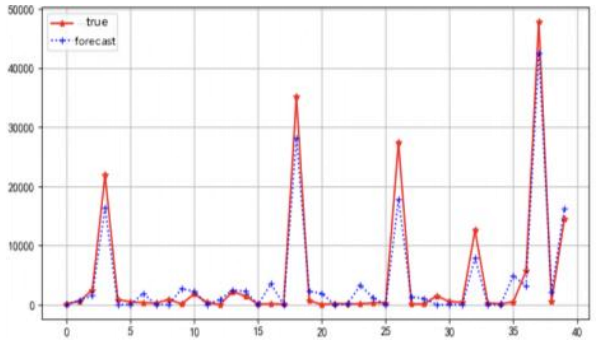

Figure 3. Comparison of predicted and real values

As can be seen from the figure, the coefficient of determination is 0.93, which indicates that the model explains the sample variance to a high degree; at the same time, the mean square error is relatively small and the difference between the predicted and true values is acceptable. Therefore, it can be considered that the results obtained by using the model for prediction are reliable to a certain extent.

3.3. Analysis of carbon reduction effects

(a)

(b)

Figure 4. When people search for new energy vehicles and carbon emissions related.

The carbon emission reduction effect of new energy vehicles is an important topic that has attracted much attention in today's automotive industry and environmental protection. This paper draws the following word cloud diagram by analysing the text of hot words related to new energy vehicles and carbon emission. Words that appear more frequently have larger font sizes in the graph, which also means that the public is paying more attention to them.

This paper adopts a panel regression model to explore the impact of new energy vehicles on carbon emissions with new energy vehicle ownership as the core explanatory variable. The model takes the total carbon emissions as an explanatory variable, and by referring to relevant literature, demographic factors, industrial structure, economic development level, and the development level of traditional fuel vehicles are selected as control variables.

In this paper, the panel data of 31 provinces in China from 2017 to 2022 are selected as the research samples, and the panel regression model variables used are shown in the table below

Table 2. Variable Names and Definitions

Type	Name	Symbol	Unit
Dependent	Carbon Emission Level	CE	10k tons
Core Explanatory	New Energy Vehicle Level	NEV	10k units
Control Variables	Population Factor	PD	People/km²
	Industrial Structure	IS	Billion
	Economic Development Level	PGDP	Yuan
	Traditional Fuel Vehicle Level	EV	10k units

The following panel regression model was developed based on the description above:

$$\begin{aligned}\ln CE_{it} &= \alpha_0 + \alpha_1 \ln NEV_{it} \\ &+\alpha_2 \ln PD_{it} + \alpha_3 \ln IS_{it} + \alpha_4 \ln UR_{it} \\ &+\alpha_5 \ln PGDP_{it} + \nu_i + \nu_t + \varepsilon_{it},\end{aligned} \tag{12}$$

where i and t denote different regions and time, respectively. i=1,2,⋯,31; t=1,2,⋯,6. e.g., CE_{ie} denotes the carbon emission intensity of the i_{th} province in the t_{th} year. v_i denotes area fixed effects, v_t denotes time fixed effects, and ε_{it} denotes a random perturbation term.

We obtained the regression results of the carbon reduction effect as shown in the table 3.Based on the results of the panel regression model, the following conclusions can be drawn: analysed from the perspective of the core explanatory variables, the coefficient of the level of development of the new energy automobile industry is significantly positive at the 5% level, which indicates that the development of the new energy automobile industry at the national level still brings about an increase in carbon emissions. This positive relationship may stem from the electricity consumption involved in the production and use of new energy vehicles as well as the carbon emissions resulting from the manufacturing process, which will therefore have a negative impact on the environment.

Analysed from the perspective of control variables, the coefficient of the level of development of traditional fuel vehicles is significantly positive at the 10% level, which further proves that the use of traditional fuel vehicles causes a large amount of carbon emissions. The coefficient of the development level of the new energy automobile industry is small compared with the coefficient of the development level of traditional fuel automobiles, which indicates that the use of new energy automobiles has a significant carbon emission reduction effect relative to traditional fuel automobiles.

Table 3. Variable Names and Definitions

Type	Name	Symbol	Fixed Effect
Core Explanatory	New Energy Vehicle Level	NEV	0.080541** (2.41**)
Control	Population Factor	PD	-0.607040 (0.97)
	Traditional Fuel Vehicle Development Level	EV	0.168806* (1.67*)
	Industrial Structure	IS	-0.203389 (-1.08)
	Economic Development Level	PGDP	0.0005724 (0.00)
	Const		7.202726* (1.92*)
	N		186
	R^2		0.2446

Note: *, **, *** respectively represent significance at the 10%, 5%, 1% levels; figures in parentheses are t-statistics.

Therefore, the nationwide carbon reduction effect of the promotion of new energy vehicles is highlighted by the comparison with traditional fuel vehicles.

In addition, this section divides the overall national panel data into panels for the three major regions of East, Central and West based on the regional division principle in 2005 in

order to further assess the carbon emission reduction effect of the new energy vehicle industry in each region.

4. Conclusions and recommendations

Against the background of the booming development of the new energy vehicle industry, this paper comprehensively analyses the data on parameter configurations, sales price, performance, brand, class, number of charging piles in the country, and historical sales of different models, and the research object includes 128 models from 56 manufacturers. Using exploratory data analysis, random forest feature screening, embedded degradation and time series methods, a sales volume prediction model based on support vector regression machine is successfully constructed. The study draws the following conclusions:

Table 4. Regional Heterogeneity Panel Regression Results

Name	Eastern Region	Central Region	Western Region
NEV	0.0661716 (2.23**)	0.0960427 (0.89)	0.1455851 (1.75*)
PD	-0.3244151 (-0.77)	-0.560743 (-0.35)	3.621133 (1.95*)
EV	0.1378387 (1.61)	0.4987154 (1.10)	0.0789507 (0.73)
IS	-0.1882588 (-0.93)	-0.5044846 (-1.05)	0.4268497 (0.92)
PGDP	-0.0221029 (-0.07)	0.1332108 (0.13)	-1.176645 (-1.41)
const	5.584 (4.389***)	12.05962 (1.23)	2.134163 (0.26)
N	186	186	186
R^2	0.0417	0.0642	0.1023

Note: *, **, *** respectively represent significance at the 10%, 5%, 1% levels; figures in parentheses are t-statistics.

(1). The sales of new energy vehicles are closely related to model parameters and technology level. Through descriptive analysis, independent t-test and Pearson correlation coefficient test, it is found that body size, selling price and performance are important considerations for consumers to choose new energy vehicles. The sales volume shows seasonal fluctuations, with sales decreasing in January and February, while sales are relatively high in November and December. The impact of the number of charging piles on new energy vehicle sales highlights the importance of infrastructure development in market development.

(2). Random Forest was used for feature screening and combined with the SVR model for prediction. In the effect evaluation on the test set, we obtained a coefficient of determination of 0.93 and a relatively small mean square error, indicating that the model fits well and can better characterise the deep connection between the data. By predicting the sales of different models of new energy vehicles in the next two months, it can provide an important reference for enterprises to formulate production plans.

In addition, the carbon emission reduction effect of the high-quality development of the new energy automobile industry is investigated through panel regression based on the panel data of 31 provinces across China from 2017 to 2022. The specific research findings are as follows:

(1).The results of the national panel regression show that the coefficients of the development level of new energy vehicles and traditional fuel vehicles are significantly positive at the levels of 5% and 10%, respectively, while the coefficient of the development level of new energy vehicles is much smaller than that of traditional fuel vehicles, so it can be shown that new energy vehicles can help to promote China's energy-saving and emission reduction policies compared with traditional fuel vehicles.

(2). The regional heterogeneity analysis of the carbon emission reduction effect of new energy vehicles shows that the development of new energy vehicles has carbon emission reduction benefits compared with traditional fuel vehicles in the eastern and central regions; however, in the western region, the promotion of new energy vehicles does not have significant carbon emission reduction effects.

In conclusion, we believe that, from the manufacturers' perspective, they should continue to improve the technical level and performance of new energy vehicles and launch competitive products to meet the growing demand of consumers; they should formulate flexible and diverse pricing strategies to meet the needs and budgets of different consumers; and they should also actively participate in the technological research and development of new energy vehicles to continuously improve the energy efficiency and environmental performance of their products and to promote sustainable development of the industry as a whole. Sustainable development of the industry as a whole. Consumers should enhance their awareness of new energy vehicles, gain a deeper understanding of their advantages such as environmental protection and energy saving, and pay attention to key factors such as price, performance, and ease of charging to choose a model that meets their individual needs. The government and society should actively support and encourage the promotion and popularisation of new energy vehicles, and reduce the environmental impact of traditional fuel vehicles by formulating policies, providing financial support, and promoting technological innovation.

Through the joint efforts of manufacturers, consumers and the government, the new energy vehicle market will usher in a healthier development, realise the carbon emission reduction effect and make positive contributions to environmental protection and sustainable development.

References

[1] Duan Shengli. Research on the impact of consumer perception on purchasing intention of new energy vehicles: considering the moderating effect of self - efficacy and government support [D]. Jiangsu University, 2022. DOI: 10.27170 / d.cnki. gjsuu. 2022. 002020.

[2] Xu Yan. Development status and trend of new energy vehicles[J]. Automobile Applied Technology,2020,45(24):13-15.DOI:10.16638/j.cnki.1671-7988.2020.24.005.

[3] Li Xinhai. Using “random forest”for classification and regression[J]. Chinese Journal of Applied Entomology, 2013, 50(04):1190-1197.

[4] Zeng Shao-Hua. The Theory Research of Algorithm on Support Vector Regression and Application [D]. Chongqing University, 2006.

[5] WANG Hua,YANG Weihua,YANG Guang. Analysis of Carbon Emissions and Influencing Factors of BuildingOperations in Hebei Province [J]. Journal of Hebei University of Engineering (Natural Science Edition), 2024, 41(02):78-85.

[6] Sixu Mu, Guangdong Huang, Dynamic Time-Frequency Spillover Effects between Carbon Trading Markets, Fossil Fuels, and New Energy Vehicles: Evidence from China, Procedia Computer Science, 221(2023): 885-892.