WEP Warwick Evans Publishing

International Journal of Global Economics and Management

ISSN: 3005-9690 (Print), ISSN: 3005-8090 (Online) | Volume 7, Number 1, Year 2025
DOI: https://doi.org/10.62051/ijgem.v7n1.04
Journal homepage: https://ijgem.org

Analysis of the Influencing Factors Affecting the Satisfaction of New Energy Vehicles Based on CatBoost Model

Weiguang Jin [1, 2], Shilan Hu [1, 2], Shouxi Wu [1, 2, *]

[1] China Automotive Technology & Research Center Co., Ltd., China
[2] China Auto Information Technology (Tianjin) Co., Ltd., Tianjin, China
*Corresponding Author: wushouxi@catarc.ac.cn

ABSTRACT

In order to solve the problem that automobile enterprises want to be able to keep track of the satisfaction of a large number of consumers with new energy vehicle products at all times, this paper proposes a scheme based on the CatBoost machine learning model to analyze the influencing factors of the satisfaction of new energy vehicle products. This scheme first uses market research methods and SPSS software to obtain the relevant original data; Next, the box plot is used to find outliers, the R language function is used to find missing values, and the missing values are filled with mean and mode respectively according to the characteristics of the dimension data, and then the Min-Max method is used to standardize the data. Finally, the principle of the CatBoost model is introduced; The results show that range, safety and economy are the indicators that consumers are more concerned about, and the results meet the actual requirements. Based on the research conclusions of this paper, it can provide reference and basis for improving product satisfaction and precision marketing of enterprises.

KEYWORDS

CatBoost; SPSS; Box plot; Min-Max; Standardization

1. RESEARCH BACKGROUND

In 2024, production and sales of new energy vehicles reached 12.888 million units and 12.866 million units, respectively,with year-on-year increases of 34.4% and 35.5%. The sales volume of new energy vehicles (NEVs) reached 40.9% of the total sales volume of new vehicles in China, and NEVs will continue to be an important growth point for China's automotive industry. Although the new energy vehicle market has broad development prospects, it also faces some challenges. According to market feedback, user satisfaction has become an important factor influencing the further development of new energy vehicles. User satisfaction not only reflects consumers' recognition of the product, but also directly relates to the brand's reputation and the stability as well as the expansion of market share. Therefore, an in-depth study of the factors influencing the satisfaction of new energy vehicle products is of great practical significance for enterprises to improve product quality and service level and enhance market competitiveness.

Traditional methods of analyzing influencing factors of satisfaction, such as questionnaires often have difficulty dealing with complex data relationships and numerous influencing factors, while machine learning algorithms have significant advantages in dealing with large-scale data and complex relationships. CatBoost, as a decision tree-based gradient boosting algorithm, has been widely used in many fields for its advantages such as handling categorical features, preventing overfitting, and reducing gradient bias.

For example, Reference [1], in order to comprehensively evaluate the brand assets of different new energy vehicles and further promote the development of the new energy vehicle industry, proposed a new energy vehicle brand asset evaluation system constructed based on multi-source heterogeneous data and the CatBoost model. the brand equity of new energy vehicles of China's independent brands is higher than that of joint venture brands, which mainly benefits from the fact that in recent years, China's new energy vehicles have achieved overtaking on curves.It can be seen that applying the CatBoost algorithm to the analysis of factors influencing the satisfaction of new energy vehicle products can more accurately explore the complex relationship between each factor and satisfaction, identify the key influencing factors, and provide strong support for enterprises to formulate targeted improvement strategies. Through the analysis of a large amount of user data, the CatBoost model can identify which factors have the most significant impact on user satisfaction, thereby helping enterprises concentrate resources on optimization and improvement. Therefore, it is of great theoretical and practical value to conduct research on the factors influencing the satisfaction of new energy vehicle products based on CatBoost.

Calculating the satisfaction of new energy vehicles involves not only some numerical parameters but also a large number of text-category parameters. To better calculate the satisfaction and improve the accuracy rate, this paper proposes a method called "Analysis of Influencing Factors of Automotive Product Satisfaction based on CatBoost", where CatBoost is a type of Boosting algorithm. CatBoost is an improved algorithm under the framework of the GBDT algorithm, which is based on the symmetric decision tree algorithm. Meanwhile, CatBoost is a GBDT framework with few parameters, support for categorical variables and high accuracy, mainly addressing the efficient and reasonable processing of categorical features, handling gradient bias and prediction offset problems, and improving the accuracy and generalization ability [2-4] of the algorithm.

2. RESEARCH PATHWAYS

The research of this article consists of four steps: data collection, data cleaning, model building, and result discussion. The most important of these is the data cleaning module, and the quality of data cleaning affects the accuracy of the model. The roadmap for this study is shown in Figure 1.

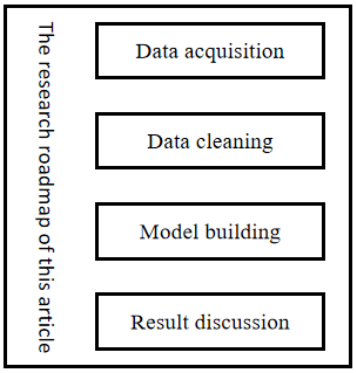

Figure 1. Roadmap for this study

2.1. Data Acquisition Module

The data in this paper are derived from a market research questionnaire. The interviewees are researchers with decades of research experience, and the interviewees are owners of new energy vehicles or actual drivers of new energy vehicles in the past 1-2 years. The data processing software is SPSS professional software. The data dimensions numbered 4-9 are scored 1-10, and the characteristics of other numbers are text-based data. The data feature dimensions of the data collected in this paper are shown in Table 1.

Table 1. Replacement of metric variables

Serial numbers	Indicators	Serial number	Indicators
1	Vehicle model	10	Gender
2	Motivation type	11	Age group
3	Brand	12	Highest degree
4	Economy (Energy consumption and retention rate)	13	Industry
5	Safety performance (braking and driving vision)	14	Career stage
6	Battery life	15	Marriage
7	Dynamic performance (climbing and acceleration)	16	Residence location
8	Comfort (Space seats)	17	Income level
9	drivability		

2.2. Data Cleaning

Data cleaning in this article includes finding and handling outliers and missing values, data standardization, etc [5-7].

(1) Methods for finding outliers

1) Outlier lookup method

As an effective tool for visualizing data distribution, the box plot visually presents data characteristics through five core statistics: the lower edge (minimum), the lower bound of the box (25th percentile /Q1), the middle line of the box (median), the upper bound of the box (75th percentile /Q3), and the upper edge (maximum). The method sets the reasonable range of the data by calculating the interquartile range (IQR=Q3-Q1), extending the upper and lower boundaries to Q3+1.5IQR and Q1-1.5IQR respectively, and discrete points outside this range will be judged as potential outliers. The determination mechanism based on statistical distance gives the box plot a significant advantage [7] in revealing data dispersion and locating outliers. The box plot diagram is shown in Figure 2 below.

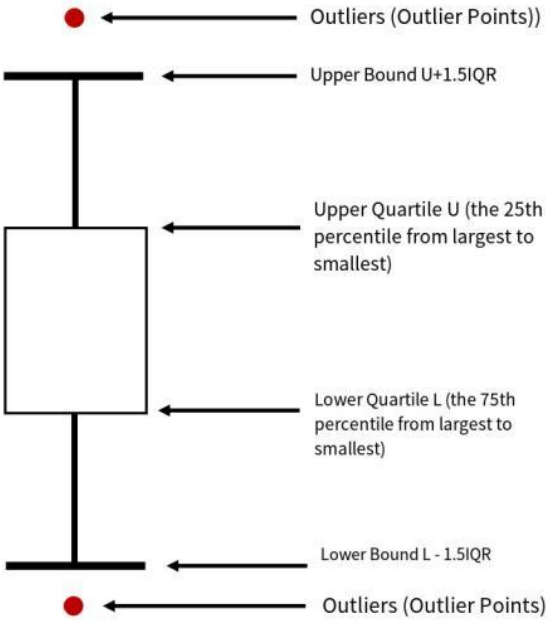

Figure 2. Diagram of the box plot

2) Outlier handling method

The method for handling outliers in this article: Remove outliers and treat them as missing values.

(2) Methods for finding and handling missing values

1) Methods for finding missing values

In the R language environment, missing value identification can be achieved through the is.na() function, which generates Boolean judgment matrices when acting on vectors or data frames, precisely marking whether each data unit is missing or not. For multi-dimensional data structures, the complete. cases() method can systematically screen all observed samples, and the TRUE/FALSE sequence it returns corresponds to the integrity status of the data rows. By combining the sum() function for logical value counting or the table() function for distribution statistics, researchers can quantitatively evaluate data missing patterns, including key quality indicators such as global missing rate and variable dimension missing hotspots, providing a reliable basis [5] for subsequent data cleaning decisions.

2) Methods for handling missing values

The method for handling missing values in this article is as follows:

If the field is numeric data, the method for handling missing values is: Take the average of the feature data of the column to fill in the missing values, as shown in formula (1).

$$x_i^{full} = mean(x_i) \tag{1}$$

Among them, is the missing value filled, is the characteristic data of this column, and mean is the average value function. x_i^{full} x_i

If the field is text-type data, the method for handling missing values is: Take the mode of the column feature data to fill the missing values, as shown in formula (2).

$$x_i^{full} = \mathrm{mod}\, e(x_i) \tag{2}$$

Where is the missing value filled, is the feature data of the column, and mode is the mode function. x_i^{full} x_i

(3) Data normalization

In this paper, maximum-minimum normalization is adopted, and the formula for Min-Max normalization is shown [6] in Formula (3) below.

$$d^{'} = \frac{d - d\_\min}{d\_\max - d\_\min} \tag{3}$$

In formula (1), is the normalized data, $d^{'}$ d is the original data point, d_min is the minimum value of the dataset, and d_max is the maximum value of the dataset. Range normalization (maximum-minimum normalization) maps numerical features to a closed interval [0, 1] through a linear transformation, and its mathematical essence is dimensionless processing to eliminate dimensional differences. While preserving the original data distribution structure, the core advantage of this algorithm lies in maintaining the order relationship between the data, eliminating the interference of extreme values by compressing the numerical scale, and ensuring that the core statistical features such as distribution pattern and skewness characteristics remain unchanged, providing a unified benchmark for the standardization of the input of the CatBoost machine learning model in the following text.

2.3. Model Building

CatBoost is a machine learning algorithm based on Gradient Boosting Decision Trees (GBDT) developed by Yandex, a Russian company, which has demonstrated outstanding performance in many fields. At its core, it is based on the idea of gradient boosting and makes predictions by iteratively building a series of decision trees. Under the framework of gradient boosting, each newly generated decision tree is dedicated to fitting the residuals between the predicted results of all previous trees and the true values, thereby gradually improving the overall performance of the model. The CatBoost model can combine features from feature to feature to form a new set of category features, which greatly enriches the model's feature dimension and improves its accuracy. Since the underlying model of the CatBoost model uses a symmetric tree, it can also prevent overfitting [8-11].

(1) Processing category features

The CatBoost model mainly handles category features as [7] follows:

Step 1: Randomly arrange the set of input feature values to generate multiple random permutations;

Step 2: Calculate the average sample value for samples of the same category;

Step 3: Convert all categorical eigenvalues to numerical values using the following formula:

$$ave\_t\arg et = \frac{CountInClass + \alpha \bullet P}{TotalCount + \alpha} \tag{4}$$

Where ave_target is the transformed feature (numeric type); CountInClass is the number of features in the current category whose eigenvalue is CountInClass; TotalCount is the total number of eigenvalues of this category feature in all samples (including the current sample), that is, the size of the sample without missing values; a Is the weight of the prior probability; P is the prior value for binary classification, and the prior term is the prior probability of the positive example.

(2) Feature recombination

The second aspect is feature recombination, which CatBoost models take a greedy [7] approach only when the current tree considers new splits, as shown in formula (5).

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{G_L + G_R + \lambda}\right] - \gamma \tag{5}$$

Among them, Gain represents the optimal Gini coefficient during the splitting process, $\frac{G_L^2}{H_L + \lambda}$ is the score of the left subtree of the split, $\frac{G_R^2}{H_R + \lambda}$ is the score of the right subtree of the split, $\frac{(G_L + G_R)^2}{G_L + G_R + \lambda}$ is the score when there is no splitting, and γ is the model complexity brought about by adding a new leaf.

2.4. Result Discussion

Through the trained CatBoost model, the influence of each factor on the satisfaction of new energy vehicle products can be further analyzed. The CatBoost model provides feature importance assessment capabilities, which measure the importance of each feature to the prediction result (i.e. satisfaction) by calculating how much each feature contributes to reducing the loss function during model training. The higher the feature importance score, the greater the impact of the factor on satisfaction. The CatBoost model can be used to analyze the importance of each feature, and the data on the importance of each feature in descending order are shown in Table 2 (for space reasons, the first 5 are listed).

Table 2. Sort the data of each feature's importance from high to low

Feature ranking	Features	Importance
1	Battery life	20.75%
2	Safety	11.32%
3	Economy	11.07%
4	Charging	10.30%
5	Comfort	9.86%

As shown in Table 2, among many influencing factors, range is identified by the model as one of the factors that have a significant impact on satisfaction with new energy vehicle products. With the development of the new energy vehicle market, consumers' expectations for range are constantly increasing. Range is directly related to the convenience and practicality of new energy vehicles. In actual use, insufficient range can cause "range anxiety", limit travel radius and increase time cost, especially during long trips, frequent charging needs can significantly reduce user satisfaction. Long-range models, on the other hand, can cover both daily and long-distance travel scenarios, enhancing the sense of security and convenience of use, thereby increasing user satisfaction.

3. CONCLUSIONS AND ANALYSIS

In this study, a new energy vehicle satisfaction prediction model was constructed based on the CatBoost algorithm, and the correlation between various influencing factors and satisfaction was analyzed in depth. The results show that range and economy are the core factors influencing consumer satisfaction, which are highly consistent with the industry status quo and user perception. Compared with the traditional method of combining questionnaires with linear analysis, this study uses the CatBoost algorithm to effectively handle the nonlinear relationships among multiple factors, breaking through the limitations of simple statistical models and providing a more accurate data mining perspective for the study of satisfaction with new energy vehicles. The research findings have direct guiding significance for optimizing product performance and enhancing user experience.

This study has limitations in terms of data coverage and depth, and the data may not fully cover the characteristics and needs of all regions, all vehicle models, and all user groups, which may affect the generalization ability of the model and the universality of the research results. The representativeness of the data may be insufficient in some niche models or the use of new energy vehicles in specific regions, resulting in limited predictive ability of the model for these special circumstances. In the future, it can be optimized in three aspects: 1) Expand the breadth of the data, increase the sample size and cover more regions, vehicle models and user groups, and supplement data in dimensions such as user scenarios and driving habits; 2) Deepen data cooperation, jointly obtain first-hand data with automakers and industry associations, and integrate multi-platform user feedback through big data technology; 3) Introduce time series analysis to integrate market dynamics and technological trends to enhance the model's ability to predict changes in satisfaction.

REFERENCES

[1] H. Huiying, Y. Jing, Z. Fan and Z. Yishu, "Brand evaluation system for vehicles based on multi-source heterogeneous data and CatBoost model," 2024 IEEE 2nd International Conference on Electrical, Automation and Computer Engineering (ICEACE), Changchun, China, 2024, Pp. 670-676, doi: 10.1109 / ICEACE63551.2024.10898605.

[2] Li Yuhan, Shi Zhen. Evaluation of riding satisfaction of electric vehicles based on Principal component analysis [J]. Automobile Practical Technology, 2020(11):8-10.

[3] Dongfang Yuxiao. The prospect of new energy vehicles is broad, but consumer satisfaction needs to be improved [J]. China World, 2019(07):58-59.

[4] Du Jiarui, Gao Lequan, Liu Yan, Li Bin, Yang Zhongmin. Research on Influencing Factors of Purchase Intention of New Energy Vehicle Consumers in Beijing-Tianjin-Hebei [J]. Cheng, 2019, 38(19):220-223.

[5] Yu Jicheng, Zhou Feng, Wang Jiangchu, et al. Prediction of transmission line loss rate based on multi-dimensional features and GBDT model [J]. Computer Applications and Software, 2022, 39 (06): 82-86+126.

[6] Zhang Fan, Guo Yaxin, Yang Jing, et al. Research on prediction of electric quantity based on GBDT+ feature engineering method [J]. Electron Quality, 2020, (01): 1-4.

[7] Fan Zhang et al 2021 IOP Conf. Ser.: Earth Environ. Sci. 7669 042023DOI 10.1088/1755-1315/769/4/042023.

[8] Zhang Yixiao, Zhao Zhongguo, Zheng Jianghua. CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China [J]. Journal of Hydrology, 2020(prepublish).

[9] Ding Qi. Research on Employee Departure Prediction Based on Catboost algorithm [D]. Shanghai Normal University, 2020.

[10] Dang Cunlu, Wu Wencheng, Li Chaofeng, Li Yongqiang. Research on short-term power load forecasting based on CatBoost algorithm [J]. Journal of Electrical Engineering, 2020, 15 (01):76-82.

[11] Jiang Qigang, Yang Xiuyan, Yang Changbao, Zhao Zhenhe. Object-oriented land use classification based on CatBoost algorithm [J]. Journal of Jilin University (Information Science Edition), 2020, 38(02):185-191.