Analysis of Influencing Factors of New Energy Vehicle Satisfaction Based On Scenario Thinking and Catboost Model

Fan Zhang¹,² *, Jing Yang¹,², BenShuang Liang¹,², Hongjing He¹,²

¹China Automotive Research (Tianjin) Automotive Information Consulting Co., Ltd., Tianjin, China
²China Automotive Technology and Research Center Co., Ltd. (CATARC)

*Corresponding author: zhangfan2019@catarc.ac.cn

Abstract. In order to accurately grasp the influencing factors of consumer product satisfaction with new energy vehicles, this paper uses scenario-based thinking as the framework and a large amount of multi-source heterogeneous data as the basis. Dimensional matching is performed on the structured data, and the random forest machine learning method is used to obtain complete sample data to reflect the overall market characteristics; thus, the CatBoost model is trained based on the complete sample data, and the degree of influence of each feature on satisfaction is analyzed. Through research and analysis, it is found that not only the performance of the product itself has a greater impact on product satisfaction, but the characteristics of consumers' own attributes are also an important factor affecting product satisfaction. Based on the research conclusions of this article, it can provide reference and basis for the improvement of enterprise product satisfaction and precision marketing.

Keywords: Scenario, Random Forest, CatBoost, Satisfaction, Features, Precision Marketing.

1. Introduction

Research background With the gradual decline of China’s new energy vehicle subsidies, and new Internet car-making forces and joint venture brands are accelerating their entry into the new energy vehicle market, the new energy market is highly competitive and products are uneven, which is not conducive for consumers to buy ideal models. It is also not conducive to the upgrading of new energy vehicles by car companies. Therefore, it is necessary for us to study the satisfaction of new energy vehicles. At the same time, we must accurately grasp the demand trends and satisfaction evaluations of new energy vehicle consumers, so as to move forward and help China's new energy. The car will set sail in the future.

The study of satisfaction with new energy vehicles can help consumers to understand the current performance of products and services of various new energy models, and provide a reference for consumers to choose and buy cars. At the same time, related enterprises can obtain the following points according to the satisfaction report of new energy vehicles: first, they can dig the pain points of new energy vehicle products and services, and determine the direction of improvement in the future; Second, we can understand the current consumer characteristics and ideology of the new energy
vehicle market, and find target users; Thirdly, it can study the demand of consumers in the new energy vehicle market and the factors affecting decision-making; Fourth, consumers' views on the future development direction of new energy vehicles and demand trends can be explored. Fifth, we can understand consumers' awareness and acceptance of new energy vehicle brands and policies. To sum up, it can be seen that the study of satisfaction with new energy vehicles is a historical burden, so a more efficient, more accurate and more convenient method is used to study new energy. Car satisfaction is also an urgent priority.

The calculation of new energy vehicle satisfaction includes not only some numerical parameters, but also a large number of text category parameters. In order to better calculate satisfaction and improve accuracy, this paper proposes a method "CatBoost-based automotive product satisfaction factors "Analysis", CatBoost is a kind of Boosting family of algorithms. It is an improved algorithm under the framework of the GBDT algorithm. It is a symmetric decision tree algorithm. It is a GBDT with few parameters, support for categorical variables and high accuracy. The framework mainly solves the problem of efficiently and reasonably processing categorical features, dealing with gradient deviation and prediction offset, and improving the accuracy and generalization ability of the algorithm [1-3].

2. Research path
This article is based on the raw data of new energy vehicle satisfaction in recent years. The structure of this method mainly includes collecting multi-source heterogeneous data, defining scenario-based thinking dimensions, multi-source heterogeneous data dimension matching, data preprocessing, CatBoost model calculation and obtaining satisfaction data report. The structure diagram of this paper is shown in Figure 1 below.

![Figure 1 Structure diagram of the method in this paper](image)

2.1. Collect multi-source heterogeneous data
The multi-source heterogeneous data in this article comes from various automobile websites and survey questionnaires, as shown in Figure 2 below.

![Figure 2 Schematic diagram of multi-source heterogeneous data in this article](image)

2.2. Define the contextual thinking dimension
Scenarioization refers to a specific and real life scenario, which is often related to consumption scenarios. It includes basic user attributes (gender, age, etc.), various indicators, engineering indicators, and car usage scenarios (car usage time, car usage location), Car usage scenarios, driving habits, perception evaluation, etc.). Researching scenario-based thinking helps to discover consumers’
“subjective preferences” in time, and acts as a catalyst for the company’s precision marketing and product upgrades. Figure 3 below shows pure engineering indicators and consumption based on scenario-based experience. The perception and attitude of the people are different. Based on this, the scene-oriented dimensions of the automotive field are defined, as shown in Table 1 below.

![Figure 3](image)

**Figure 3** The difference between pure engineering indicators and consumer perceptions based on scenario-based experience.

**Table 1** Scenario-based dimensions defined

| Mode | Power type | Brand type | Economic type | Safety | Endurance | Recharge | Dynamic | Comfort | Drivability |
|------|------------|------------|---------------|--------|-----------|----------|---------|---------|-------------|
| Gender generation | Highest education | Industry | Career stage | Marriage | Living location | Income level |

2.3. Multi-source heterogeneous data dimension matching

Websites such as Autohome, Bitauto.com, Owner’s Home, and AutoQuality.com mainly use text data. Here, the phrase mapping method is used to split and match the text data. The mapping dictionary comes from the CATARC keyword group mapping dictionary, such as After a sentence is split, the words include: saving money, economy, cost-effective and other words, then the sample data at this time is defined as an economic index, and the LSTM model is used to calculate the satisfaction degree of this index, and so on. The questionnaire data comes from the survey questionnaire.

2.3.1. Merger integration features

This module is mainly to integrate and merge feature parameters that have similar functions and are meaningless to model calculations. For example, the feature "power type" in Table 1 has only one feature value "pure electric". This feature is calculated according to the following formula (1) for machine learning Contribution rate of model establishment [4].

\[
g = 1 - \frac{1}{n}
\]

Among them, n represents the sum of all possible values of this feature, and g represents the contribution rate. From this formula, the contribution rate of the parameter "power type" can be calculated to be 0 (power types are pure electric), which has no practical meaning. Therefore, this feature parameter can be deleted from the data in this article. In other practical problems, the contribution rate threshold of the model can be set according to the number and accuracy of feature parameters. The feature list after merging and integrating the parameters is shown in Table 2. Table 1 reduces one parameter.
Table 2 List of features after merging and integrating parameters

| Model Type | Brand | Economical | Safety | Endurance | Recharge | Dynamic | Comfort | Drivability | Gender |
|------------|-------|------------|--------|-----------|----------|---------|---------|-------------|--------|
| Generati  | Highest education | Industry | Career stage | Marriage | Living location | Income level |

2.4. Data preprocessing

Data preprocessing includes eliminating outliers (also called outliers, hereinafter referred to as outliers) and filling in missing values. To ensure the accuracy and efficiency of the model, it is necessary to fill in missing values and handle outliers before building the model.

2.4.1. Finding missing values

Missing values may appear due to human accidental negligence. Missing values may appear in any feature of the data set. Missing values can be obtained by mathematical statistics or observed by human eyes. In order to save human resources, in this article, use mathematical statistics to find missing values. Assuming that the white areas in Figure 4 are missing values, the judgment method is as follows (assuming the data dimension is m rows and n columns) [5-7]:

(1) First locate the sample location where the missing value is located. If the number of feature values for each feature is inconsistent with the number of samples, you can judge that there are missing values under the feature; the specific method is shown in Figure 4 below, first locate f1, and then scan D1 down to Dm. If the number of samples under the f1 feature (the number of data after statistics) is inconsistent with the number of samples from D1 to Dm (m) at this time, it means there are missing values under the f1 feature; Position f2 by analogy until fn is judged. Figure 4 below shows that the sample position where the missing value is located is fn-1;

(2) After locating the sample location where the missing value is located, locate the feature point where the missing value is located. The specific method is shown in Figure 4 below. First locate D1, then scan from f1 to the right to fn. The number of features of D1 (n) is inconsistent with the number of features from f1 to fn (the number of data after statistics), which means that there are missing values in the D1 sample; and so on, until Dm is judged. From Figure 4 below, it can be seen that the feature position of the missing value is Dm-1;

(3) According to the conclusions of (1) and (2), the position of the missing value is (Dm-1, fn-1);

![Figure 4 Schematic diagram of finding missing values](image)

2.4.2. Missing value processing

After finding the location of the missing value in subsection 2.4.1, the missing value needs to be processed. There are two processing methods [5-7]:

(1) When the proportion of missing value data in the sample number is very small or the sample size is not required, the sample data with missing records can be directly discarded or the mode or mean value can be manually processed (when the characteristics of this column are When the value is
text data, the mode of the characteristic value of this column is used to replace the missing value. When the characteristic value of this column is floating-point data, it is represented by the mean number of the characteristic value of this column).

(2) In actual data, missing data often occupies a considerable proportion or when the number of samples is required in the project. If the sample data with missing values is discarded, a lot of important information will be lost, causing systematic differences between incomplete sample data and complete sample data. Analyzing such data may lead to wrong conclusions.

Because the data in this article requires the number of samples, it is impossible to discard sample data with missing values. The text data and floating-point data in this article account for a large proportion, so this article uses random forest to fill in the missing values.

Filling method: Assuming a data with n features, feature T has missing values (a large number of missing values are more suitable), use T as a label, and other n-1 features and the original data as a new feature matrix, if Other features also have missing values, traverse all the features, starting with the least missing values; the fewer missing values, the less accurate information is needed; to fill a feature, first replace the missing values of other feature values with 0, so Each time through the loop, the features with missing values will be reduced by one. The specific data explanation is shown in Table 3 below, and the specific graphic explanation is shown in Figure 5 below:

| data                          | Description                                                                 |
|-------------------------------|-----------------------------------------------------------------------------|
| Xtrain（Practice characteristics） | Corresponding to the row where the feature ft is not missing N-1 features + labels |
| Ytrain（Training label）       | The value of feature ft is not missing                                       |
| Xtest（Predictive features）   | Corresponding to the row of the missing value of feature ft N-1 features + original label |
| Ytest（Filled missing values） | Feature ft missing value (unknown)                                          |
Figure 5 Detailed graphical explanation

Figure 6 Box plot to find outliers
2.4.3. Judgment and treatment of abnormal value

The source of data outliers in this article may be those caused by man-made recordings or abnormal values caused by machine abnormalities. The outliers are mainly reflected in "economy", "safety", "endurance", "charging", and above these floating-point characteristic parameters such as power, comfort and drivability.

The selection of outliers in the box plot is relatively objective and has certain advantages in identifying outliers. The principle of finding outliers in a box chart, that is, a value greater than or less than the upper and lower bounds set by the box chart is recognized as an outlier, and finding outliers in the box chart is shown in Figure 6 [8-11].

The following defines the lower upper quartile, lower quartile, upper bound, lower bound and IQR.

1. Upper quartile: The upper quartile is assumed to be U, which means that only 1/4 of the values in all samples are greater than U, that is, U is at 25% when sorting from largest to smallest;
2. Lower quartile: The lower quartile is assumed to be L, which means that only 1/4 of the values in all samples are less than L, that is, when sorting from large to small, L is at 75%; Assuming that the interpolation between the upper quartile and the lower quartile is IQR, that is: IQR=U-L;
3. Upper bound: The upper bound is assumed to be: U+1.5IQR, where IQR represents the interpolation between the upper quartile and the lower quartile, that is, IQR=U-L;
4. Lower bound: The lower bound is assumed to be: L-1.5IQR, IQR is the same as (3).

If the data is greater than the upper bound (U+1.5IQR) or the data is smaller than the lower bound (L-1.5IQR), then the data can be judged as an outlier (outlier)

For outliers, you can eliminate the sample data with missing values, or treat the outliers as missing values, and then use the method of random forest to fill in the missing values in subsection 2.4.2 to replace the outliers.

2.5. Bringing into the CatBoost model

The CatBoost algorithm is a deeply improved version based on the GBDT framework. Its main feature is to use a special way to deal with categorical features. First, do some mathematical statistics on the category features, calculate the frequency of a certain category feature, and generate new floating-point features. In addition, to prevent model overfitting, you can add L2 regular hyperparameters at the end of the model. The CatBoost model can use the feature-to-feature link to combine into a set of new category features, which greatly enriches the feature dimension of the model and improves the accuracy of the model. The basic model of the CatBoost model uses a symmetric tree, which can also prevent the model from overfitting [12-15].

2.6. Analysis of satisfaction data

Through the CatBoost model pair, the importance of each feature can be analyzed, and the importance of each feature to the model is shown in Figure 7.
Figure 7 The importance of each feature to the mode

It can be seen from Table 9 that the factor that has the greatest impact on new energy user product satisfaction is safety, followed by battery life, while the relatively small factors are the user's business stage and the type of brand purchased. The degree of influence of each feature in the model output on product satisfaction meets the needs of people in real life, so this model has a fairly high reliability.

3. Conclusion
For consumers of new energy vehicles, the factors that affect their product satisfaction can be divided into two aspects: product performance and consumers' own attributes. In terms of product performance, the degree of impact on product satisfaction from large to small is: safety, battery life, economy, charging, comfort, drivability, and power. It can be seen that safety and battery life are the most critical factors affecting consumer product satisfaction. Product factors. It can be seen from this that if new energy automobile companies want to improve consumer product satisfaction, they can mainly start with product safety and battery life, and further improve and perfect the performance of vehicles in terms of safety and battery life. At the same time, focus on the needs and differences of consumers of different marriages and families and different income levels for vehicles, and target and meet the needs of consumers with different characteristics, so as to further enhance consumer satisfaction with products

References
[1] Li Yuhan, Shi Zhen. Evaluation of Electric Vehicle Driving Satisfaction Based on Principal Component Analysis [J]. Automotive Practical Technology, 2020(11): 8-10.
[2] Dongfang Yuxiao. New energy vehicles have a broad prospect and consumer satisfaction needs to be improved [J]. China Circle, 2019(07): 58-59.
[3] Du Jiarui, Gao Lequan, Liu Yan, Li Bin, Yang Zhongmin. Research on the influence factors of consumer purchase intention of new energy vehicles in Beijing, Tianjin and Hebei [J]. Cheng, 2019, 38(19): 220-223.
[4] Tang Yunjun, Sun Shuchang. Feature engineering methods in machine learning [J]. Automotive Practical Technology, 2020(12): 70-72.
[5] Yang Hong, Tian Jing, Wang Ke, Zhang Qing, Han Qinghua, Zhang Yanbo. Comparison and application of mixed missing data filling methods [J]. China Health Meter, 2020, 37(03): 395-399.
[6] Luo Changyin, Chen Xuebin, Song Shangwen, Liu Yang. Application of data preprocessing technology in heterogeneous data [J]. Pieces, 2020, 41(05): 6-13.
[7] Su Sifan, Zhu Cui, Zhu Wenjun, Zhao Fengchao. Missing value filling algorithm based on space-time fusion [J]. Computer Engineering and Design, 2020, 41(01): 78-84.
[8] Engineering; Studies from Macau University of Science & Technology Further Understanding of Engineering (Outlier Processing In Multimodal EmotionRecognition) [J]. Journal of Technology & Science, 2020.

[9] Qiao Yilin. Wind power operation data preprocessing technology and its application research [D]. North China Electric Power University (Beijing), 2019.

[10] Yang Kai, Zhao Mingzhou, Mao Wenrui, Zhang Yingnan. Research on data preprocessing based on power grid production projects [J]. Electronic Testing, 2019(01): 80-81.

[11] Wang Xin, Gao Tianci, Fang Jiasheng, Wang Ping. High-speed railway track irregularity abnormal value processing algorithm based on time history [J]. Railway Science and Engineering News, 2018, 15(12): 3029-3036.

[12] 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).

[13] Ding Qi. Research on employee turnover prediction based on Catboost algorithm [D]. Shanghai Normal University, 2020.

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

[15] 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.