Research on Marketing Strategy of Electric Vehicle Based on Bayesian Processing and Natural Language Analysis

Chengyi Cao*

Guangdong University of Technology, Guangzhou, Guangdong, 510006

* Corresponding Author. Email: caochengyi0329@163.com

Abstract. The automobile industry is an important pillar of the national economy, and the new energy automobile industry is a strategic emerging industry. This paper carries on the mathematical modeling analysis of the new energy electric vehicle. First of all, using the naive Bayesian processor in natural language processing, make a judgment through theme analysis and emotion analysis, define the coefficient into the model, and finally compare target customers' satisfaction with different brands of cars. Then use DBSCAN and UMAP algorithm to divide the fluctuating data index on the sequence, analyze its linkage by establishing linkage factor to study its correlation degree, and then quantify the problem, establish abnormal coefficient, analyze the persistence of data run out for a single index, and establish an algorithm for evaluating abnormal coefficient. According to the distribution of different data indicators, the impact index of different factors on the sales of different brands of electric vehicles is determined.

Keywords: Automobile industry, Cluster analysis, Natural language processing, Affective analysis.

1. Introduction

The automobile industry is an important pillar of the national economy, and the new energy automobile industry is a strategic emerging industry [1]. The development of new energy vehicles represented by electric vehicles is an effective way to solve energy and environmental problems, and the market prospect is broad. However, the electric vehicle is a new thing after all. Compared with traditional cars, consumers still have doubts in some areas, such as battery problems, and their market sales need scientific decision-making. Therefore, this paper collects the latest launch of three brand electric vehicles by an automobile company. In order to study consumers' willingness to buy electric vehicles and formulate corresponding sales strategies, the sales department invited 1964 target customers to experience the three brands of electric vehicles [2].

2. Data processing

2.1. Missing value processing

The specific experience data include the satisfaction score of battery technical performance (battery durability and charging convenience) (Full Score: 100, the same below), A1, the overall performance satisfaction score of comfort (environmental protection and space seats), A2, the overall satisfaction score of the economy (energy consumption and maintenance rate) A3, the overall satisfaction score of safety performance (braking and driving vision) A4, the overall satisfaction score of power performance[3-5] (climbing and acceleration) A5 Overall satisfaction score A6 for driving and handling performance (cornering and high-speed stability), a7 for appearance and interior decoration, and A8 for configuration and quality. In addition, there is information about the personal characteristics of the target customer experience.

After analyzing the most missing B7 data in the time domain, it is decided to use the decision tree algorithm to complete the data based on the regression model. In order to determine the possible impact of different indicators on the pricing of shared vehicles, we analyze multiple groups of indicators that may affect financial data. After that, the proportion of missing values[6-8] will be greater than 42 7% of the data of the indicators are deleted; The data of indicators with missing values

accounting for 21% to 46% are filled with the mean value; Fill in the data where the missing value accounts for less than 16% with random forest.

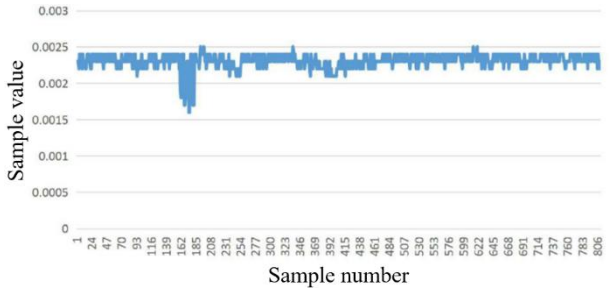

Figure 1 Sample value

As shown in the figure, B7 in the time domain has been given in the data, but it is significantly lower than the mean value in the numerical interval [124201]. After analyzing this part of the data, we find that the value fluctuates to a certain extent, so we screen some data for series comparison.

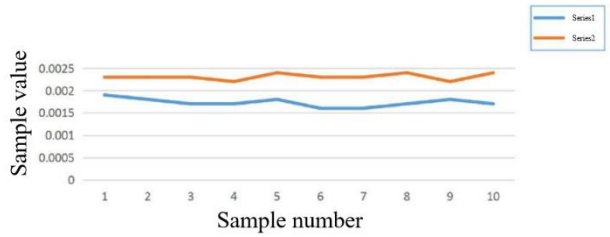

Figure 2 Comparison diagram of sample values of different series

2.2. Exception elimination

In this paper, the laida criterion is used to eliminate the abnormal value, which is 3σ criterion: for some special data, if the deviation is greater than 3σ, the data is the abnormal value and should be eliminated in time. The calculation formula of σ is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}} \tag{1}$$

x_i refers to the original data of the test value, $\bar{x}$ refers to the average value of the test value, and the sample size is n = 7329. When the deviation is greater than 3σ, the data is an abnormal value. The judgment criteria of abnormal value are as follows:

$$\left|x_i - \bar{x}\right| > 3\sigma \tag{2}$$

2.3. Final processing result

After preprocessing the sample data, the known data before processing and the new data obtained after preprocessing are extracted. The data of 20 kinds of indicators have changed to some extent in the data set. It is compared and analyzes the data of each type of index before and after processing. We select the expectation, standard deviation, standard error, and range of the basic statistics as the comparison to a certain extent.

Table 1. Descriptive statistics of sample raw data

	A1	A2	A3	A4	A5	A6
Missing degree	0.35	0.45	0.12	0.45	0.57	0.22
Continuous degree	0.17	0.11	0.32	0.07	0.08	0.31
Standard deviation	0.0035	0.00017	0.00015	0.0037	0.0022	0.0029
Standard evaluation error.	0.0018	0.024	0.00039	0.008	0.006	0.007
Range	0.016	0.018	0.027	0.044	0.031	0.023

Table 2. Descriptive statistics of sample preprocessing data

	A1	A2	A3	A4	A5	A6
Missing degree	0.14	0.13	0.07	0.19	0.21	0.04
Continuous degree	0.38	0.31	0.56	0.49	0.27	0.81
Standard deviation	0.0041	0.00019	0.00016	0.0038	0.0027	0.0033
Standard evaluation error.	0.0015	0.00033	0.00027	0.0021	0.0015	0.0017
Range	0.015	0.021	0.031	0.045	0.039	0.027

Firstly, take some data of the target customer's personal characteristics survey table as an example. First, do Chinese word segmentation on python, and then assume that the data object attributes are independent of each other. The calculation method of P $(x|y_k)$ is as follows:

$$P(\mathbf{x}|\ y_k)=\prod_{d=1}^{D}P(x_d|\ y_k)= \\ P(x_1|\ y_k)P(x_2|\ y_k)\ldots P(x_D|\ y_k) \tag{3}$$

If attribute ad is a discrete attribute or a classification attribute. For data objects belonging to category YK in the training set, there are n different attribute values under attribute ad; There are m data objects in the training set that belong to category YK, and the attribute value under attribute ad is XD. Therefore, P (xdyk) is calculated as follows:

$$P(x_d|\ y_k)=\frac{m}{n} \tag{4}$$

Then we get the distribution of propensity factors of different satisfaction, and then we analyze the emotion and text of multiple car purchase propensity. Finally, we get the traditional target customers' satisfaction with different brands of cars, 69% for car 1, 71% for car 2, and 50% for car 3.

3. Division of satisfaction data

3.1. Clustering effect of satisfaction tendency

In order to study the linkage between satisfaction data, clustering and dimensionality reduction of satisfaction data are carried out. DBSCAN processes these indicators. In order to better obtain their correlation, in the DBSCAN algorithm, starting from the core object, find all samples with the density of the core objective to form a "cluster." Determine all core objects according to the given neighborhood parameters EPS and minpts, select an unprocessed core object for each core object, find the sample with its density, and generate a cluster "cluster." Only some sensors in DBSCAN centers are described.

Table 3. Cluster center

DBSCAN center		
	center	
	1	2
H1	0.202297344	1.591841612
H2	0.13063839	4.554818516
H3	0.018999691	1.622374611
H4	0.564216366	0.058818016
H5	0.008865604	0.079433632
H6	0.001839619	0.037847166
H7	0.00333293	0.277054634
H8	0.008082933	0.523583355
H9	0.010664473	0.726681881
H10	0.202297344	1.591841612

Using the UAMP algorithm to reduce the dimension is convenient for dealing with the linkage relationship of influencing factors. In practice, because the membership degree of the fuzzy set

gradually decays to almost zero, it is only necessary to calculate the membership degree of the nearest neighborhood of each point. Therefore, a fast and efficient method is needed to calculate (approximate) nearest neighbors, which is also true in high-dimensional data space.

Using the curve family of $\frac{1}{1+ay^{2b}}$, since the Laplace operator represented by topology is the approximation of the Laplace Beltrami operator of manifold,

$$\rho_i = \min\{d\left(x_i, x_{i_j}\right) | 1 \le j \le k, d\left(x_i, x_{i_j}\right) > 0\} \tag{5}$$

$$\sum_{j=1}^{k} \exp\left(\frac{-\max\left(0, d\left(x_i, x_{i_j}\right) - \rho_i\right)}{\sigma_i}\right) = \log_2(k) \tag{6}$$

Calculation of distribution:

$$p_{i|j} = \exp\left(\frac{-\max\left(0, d\left(x_i, x_{i_j}\right) - \rho_i\right)}{\sigma_i}\right) \tag{7}$$

$$p_{ij} = p_{i|j} + p_{j|i} - p_{i|j} p_{j|i} \tag{8}$$

Get the weight value of each factor, and then adjust the new sensor relationship. Through the total variance explained, we can see the contribution degree of each component and the percentage of variance. The required components are collected at a cumulative rate of more than 83%, and then the projection of the digital data set is calculated and obtained.

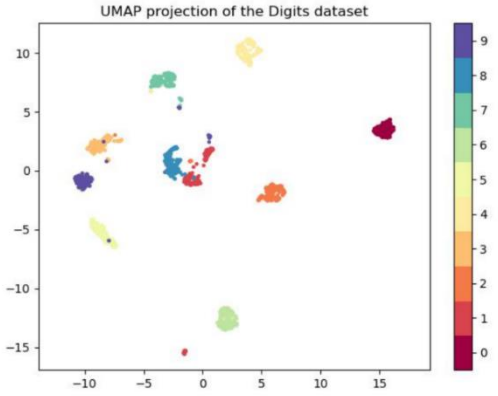

Figure 3. Digital dataset projection

Then, according to the coefficient factor and the projection map, the linkage factors a, B, C, and D of the sensor are defined, and the proportion of the factor is 0 34、0. 20、0. 11、0. 35。

Table 4. Influencing factors and linkage factors

Linkage factor	Proportion coefficient
A	0.34
B	0.20
C	0.11
D	0.35

4. Analysis of influencing factors and sensor linkage

After analyzing the sensor association degree above, the association rules are processed. First, decompose the eigenvalues:

$$J(a,b) = a^T S_{XY} b - \frac{\lambda}{2}\left(a^T S_{XX} a - 1\right) - \frac{\theta}{2}\left(b^T S_{YY} b - 1\right) \tag{9}$$

Then the correlation target is optimized by the Lagrange coefficient:

$$S_{XX}^{-1} S_{XY} b = \lambda a \quad S_{YY}^{-1} S_{YX} a = \lambda b \tag{10}$$

Finally, combined with the fluctuation characteristics of sensors in the same reactor:

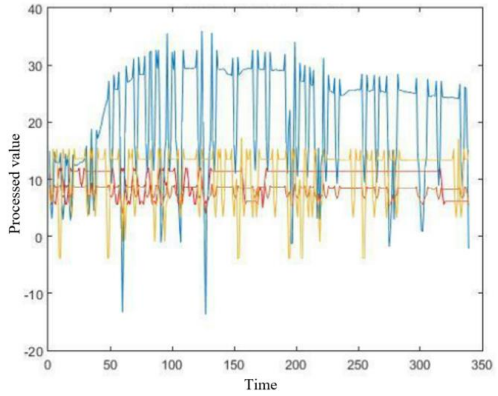

Figure 4. Sensor fluctuation of influencing factors in the same reactor

By simultaneous interpreting, the optimization process of the correlation factor and the fluctuation characteristics of the actual sensor data, the comparison results of the linkage factors of different sensors are finally determined.

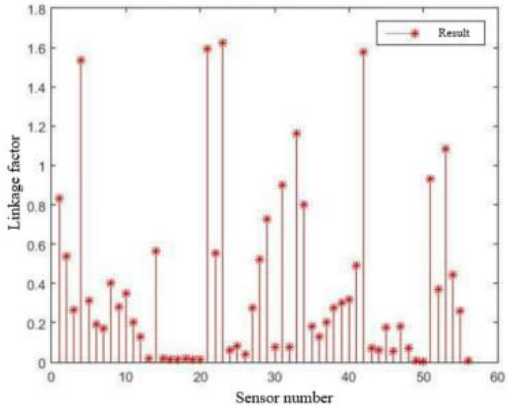

Figure 5. Comparison of simultaneous interpreting factors of different sensors

After the linkage analysis of the influencing factor sensor data above, the abnormality of the sensor data of a single influencing factor is analyzed here. Firstly, through the k-distance neighborhood idea:

$$\begin{aligned} D_k(x^{(i)}) &= \|x^{(i)} - x^{(k=t)}\| \\ &= \sqrt{(x_1^{(i)} - x_1^{(k=t)})^2 + (x_2^{(i)} - x_2^{(k=t)})^2 + \cdots + (x_n^{(i)} - x_n^{(k=t)})^2} \end{aligned} \tag{11}$$

The following abnormal relationship can be expressed as:

$$\begin{aligned} RD_t(x^{(i)}, x^{(j)}) &= \max(D_k(x^{(i)}), \|x^{(i)} - x^{(j)}\|) \\ &= \max(\|x^{(i)} - x^{(k=k)}\|, \|x^{(i)} - x^{(j)}\|) \end{aligned} \tag{12}$$

Then, the characteristics of data regularity, independence, and contingency after removing outliers are analyzed through the expression of N, combined with the jump of actual data. By combining the establishment law of N and the jump of data, the change law of the sensor in the time field is finally obtained.

Table 5. Influencing factor number and its influencing index

A1	A2	B3	B6	A7
32	34	39	54	63
A5	A6	B7	B15	B11
91	92	97	99	

5. Conclusions

The new energy vehicle industry is a strategic emerging industry. Compared with traditional cars, consumers still have doubts in some areas, such as battery problems, and their market sales need scientific decision-making. Therefore, based on the collected data, this paper first analyzes the data and makes visualization processing, classifies the continuous and discrete data, deals with the missing data and abnormal data, uses the naive Bayesian processor in natural language processing, makes the judgment through topic analysis and emotion analysis, and defines the coefficient into the later model. Finally, the data are sorted out to prepare data for further exploration. Finally, the comparison of target customers' satisfaction with different brands of cars is given. Then, it needs to analyze the influence of different factors. This paper uses DBSCAN and UMAP algorithm to divide the fluctuating data indicators on the series, analyzes the linkage by establishing linkage factors to study their correlation degree, and then quantifies the problem, establish abnormal coefficients, analyze the persistence of data run out for a single index, and establish an algorithm for evaluating abnormal coefficients.

References

[1] Fu Tong. Research on the method of model estimation in mathematical modeling-- taking graph model learning as an example [J]. Journal of Jilin normal University of Engineering and Technology, 2020. 36 (12): 117-119.

[2] Yu Fang, Wang Tiansong, Huang Yongfeng. Discussion on the integration of mathematical software into the teaching of mathematical modeling course [J]. Journal of Changji University, 2019 (06): 107111.

[3] Zhao Xiaoyong, Wang Ningning, Wang Lei. Research on outlier ensemble mining based on active learning [J]. Computer Engineering and applications, 2020556 (12): 112-117.

[4] Chen Xiaoyun, Ma Liangzhai. Research on local outlier mining algorithm based on attribute weight [J]. Microcomputer Information, 2010 Ji 26 (33): 9-11.

[5] Zhou Ning, Fu Helin, Yuan Yong. Slope stability evaluation method based on fuzzy neural network [J]. Journal of Underground Space and Engineering, 2009 Journal 5 (S2): 82-88.

[6] Wang Lei. BP Network Implementation Based on Computer MATLAB Neural Network Toolbox [J] Journal of Physics, Conference Series,2020 (2): 23-34.

[7] Zhao Yi. A Survey of Machine Learning algorithms under big data-taking AlphaGO as an example [J]. Information recording material, 2019, 20 (1): 10-12.

[8] Nuha Zamzami, Nizar Bougila. High-dimensional count data clustering based on an exponential approximation to the multinomial Beta-Liouville distribution [J] Information Sciences,2020 (34): 24-27.