Electric Vehicle Marketing Planning Based on Logic Regression and Boosted Tree

YeYu Chai(✉), YiTing Zhao, and BaoRun Li

Academy of Automation, Wuhan University of Technology, Wuhan, Hubei, China
3382825354@qq.com

Abstract. This paper studies the different target requirements of electric vehicle market sales under different conditions. After cleaning the data, the rank-sum ratio comprehensive evaluation model is established to compare the satisfaction of different flat brand electric vehicles. At the same time, Logistic regression model was used to analyze the purchase of vehicles and the visibility of various factors. Establish customer mining model based on Boosted trees; finally, the artificial fish swarm algorithm is used to optimize to establish a marketing strategy selection model.

Keywords: Sales strategy · RSR comprehensive evaluation · Logistic regression · Boosted trees · Artificial fish swarm algorithm

1 Introduction

New energy automobile industry is a strategic emerging industry with broad market prospects. However, compared with traditional vehicles, consumers still have some doubts in some fields, and their market sales need scientific decision-making. In this paper, through the establishment of rank-sum ratio evaluation model, multivariate logistic regression model, Boosted trees model [1] and artificial fish swarm algorithm model, we solve the analysis of customer satisfaction in the sales process, the screening of important indicators affecting sales, the prediction of customer purchase behavior and the planning of marketing strategy, and the cleaning of existing data. The model used in this paper makes the results objective and rigorous, has high credibility, and uses intelligent algorithm to converge quickly and accurately.

2 Materials and Methods

2.1 Data Pre-processing

2.1.1 Sources of Data

To study consumers' purchase intention of electric vehicles and formulate corresponding sales strategies, the data used in this paper are derived from 2021 "Huashu cup" C.

Z. Zeng et al. (Eds.): ECIT 2022, AHE 11, pp. 711–724, 2023.
https://doi.org/10.2991/978-94-6463-005-3_73

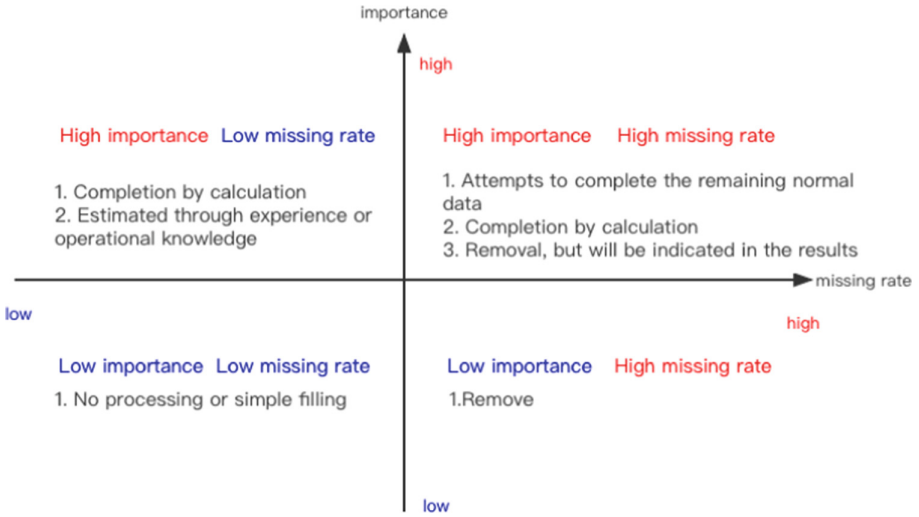

Fig. 1. Removed and completed action standard division chart.

2.1.2 Data Cleaning

2.1.2.1 Data Cleaning Path

For missing data removal in this topic, the standard of action for completion is broken down by data importance and missing rates [5], as shown in Fig. 1.

2.1.2.2 Supplement of Missing Data

In this question, the only missing experience data is B7: How many children do you have. Use Boosted tree algorithm for reasonable prediction. This algorithm belongs to one of the decision tree algorithms. Its principle is that different weights are set for the samples based on the classification results of the previous data. If the sample point in the previous data is predicted wrongly, the weight of the sample point will be increased until a predetermined small enough error rate is reached and the final improvement result is obtained.

2.1.2.3 Finding Abnormal Data

The main characteristics of logical errors are reflected in the self-contradiction of the filled experience data, which are shown in the following Table 1.

The main feature of range error is that the deviation between a certain data and other normal data in its data set is too large, which has obvious characteristics.

The evaluation standard is that if the absolute value of the difference between the data and the average value of the total data is greater than 100, the data is abnormal data with range error, which is caused by subjective or objective reasons of the customer.

2.2 Customer Satisfaction Analysis of Different Cars

Comparing customers' satisfaction with different automobiles is a comprehensive evaluation of three brands of automobiles by customers' satisfaction scores of three brands

Table 1. Specific performance of logical errors.

Type of logical error	Embodiment
Direct logic error	• For customers who choose 3: 'Married/cohabitation without children' or 4: 'Married/cohabitation without children (living with their parents)', the number of children should be 0. • For customers whose data in B13: 'your family annual income' is less than B14: 'your personal annual income', their family annual income should be greater than or equal to personal annual income.
Indirect logical errors	• Some B12: 'Position' data are 9: 'individual/small company owners', while B11: 'work unit' data are 3: 'research/education/culture/health/medical institutions', or 4: 'private/private enterprises (with more than 8 employees)', or 6: 'joint venture' customers whose positions do not match the nature of the unit in which they are located

of automobiles, and the rank sum ratio comprehensive evaluation method is used for comprehensive evaluation. This evaluation method has no special requirements for data and can eliminate the influence of abnormal data, which is more accurate than the simple nonparametric method [3].

The basic principle of rank sum ratio comprehensive evaluation method is to obtain the dimensionless statistics RSR in an n row m column matrix by rank transformation. On this basis, the concept and method of parameter statistical analysis are used to study the distribution of RSR, and the RSR value is used to directly rank or classify the evaluation objects. So as to make a comprehensive evaluation of the evaluation object.

The m evaluation indexes of an evaluation objects are arranged in the original data table of n rows and m columns, and the rank of each evaluation object of each index is compiled. Among them, the rank of the benefit index is ranked from small to large, the rank of the cost index is ranked from large to small, and the average rank is compiled for the same index data. The obtained rank matrix is denoted as

$$R = \left(R_{ij}\right) n \times m \tag{1}$$

(1) Calculate rank sum ratio.

$$RSR_i = \frac{1}{mn} \sum_{j=1}^{m} R_{ij} \tag{2}$$

Draw the *RSR* frequency distribution map, list each group frequency f_i, calculate each group cumulative frequency cf_i to calculate cumulative frequency:

$$RSR = a + b \times Probit \tag{3}$$

The p_i quantile of the standard normal distribution is added to 5, which is converted to the frequency unit *Probit*. Taking the probability unit *Probit* corresponding to the

cumulative frequency as the independent variable and RSR_i as the dependent variable, the regression equation is calculated:

$$RSR = a + b \times Probit \tag{4}$$

The significance of calculating the linear regression equation is to obtain the *RSR* fitting value corresponding to the *Probit* critical value, and evaluate the index by comparing the *RSR* fitting value.

2.3 Analysis of the Greatest Impact on Automobile Sales Indicators

The greatest impact on automobile sales is the highest correlation between indicators and purchase and non-purchase behaviour. Because the data belongs to the cross-sectional data of the questionnaire and the dependent variable is the 0–1 variable of purchase and non-purchase, we establish a multivariate Logistic regression model, and use goodness of fit as the quantitative correlation between indicators and dependent variables [2].

2.3.1 Multivariate Logistic Regression Model

Multivariate Logistic regression model can determine the role and intensity of explanatory variable X_n in predicting the occurrence probability of classification variable Y. Assuming that X is a reaction variable, P is the response probability of the model, the corresponding regression model equation is as follows:

$$\ln\left(\frac{\mathrm{p_1}}{1-\mathrm{p_1}}\right) = \alpha + \sum_{\mathrm{k=1}}^{\mathrm{k}} \beta_{\mathrm{k}} x_{\mathrm{ki}} \tag{5}$$

Where $p_1 = P(y_i = 1|x_{1i}, x_{2i}, ..., x_{ki})$ denotes the probability of occurrence of events given the values of series independent variables $x_{1i}, x_{2i}, ..., x_{ki}$; α is the intercept, β is the slope. The probability of an event is a nonlinear function composed of explanatory variable X_i, and the expression is as follows:

$$p = \frac{\exp(\alpha + \beta_1 X_1 + \cdots + \beta_n X_n)}{1 + \exp(\alpha + \beta_1 X_1 + \cdots + \beta_n X_n)} \tag{6}$$

2.3.2 Model Verification of Likelihood Ratio Test Method

Multivariate Logistic regression model has its own model evaluation coefficient: goodness of fit. The fitting degree of the observed values is evaluated according to the regression line, and the symbol is R^2. The maximum value of R^2 is 1. The closer the value of R^2 is to 1, the better the fitting degree of the regression line to the observed value is. On the contrary, the smaller the value of R^2 is, the worse the fitting degree of the regression line to the observed value is.

R^2 measures the overall fitting degree of the regression equation. However, the goodness of fit R^2 of multivariate Logistic regression model tends to be dependent variable, so it is necessary to calculate the Austell $R^2_{\mathrm{Cox\&Snell}}$ which tends to be independent variable.

$$R^2_{\mathrm{Cox\&Snell}} = 1 - \exp\left(-\frac{2}{n}[\ln(B) - \ln(0)]\right) \tag{7}$$

ln(0) corresponds to the log-likelihood value of the Logistic model that does not contain the factors to be tested, ln(0) corresponds to the log-likelihood value of the Logistic model that refits the factors to be tested. Finally, the indicators are quantified by the chi-square value. If the chi-square value is less than 0.05, it shows that the variable is related to whether the customer purchases.

2.4 Forecast the Purchase Strategy of Target Customers

We have obtained the influence of several main experience items on whether customers buy the vehicles they experience, which is only a general summary and preliminary understanding of the influencing factors and functions of the experience items. In fact, in addition to the main experience items that affect the purchase of vehicles, there are also secondary experience items, which also affect the purchase of vehicles. Primary and secondary experience items have biased information flow for the purchase or non-purchase of vehicles, which supervises the results. We need to build a model that can close to the specified target, make full use of the given data, and predict the specified target according to the direction and intensity of the information flow. Based on this supervised learning problem, we use the boosted trees algorithm to weight each experience item [4] and use the multi-feature training dataset x_i to predict a target variable y_i.

2.4.1 Boosted Trees Model Establishment

2.4.1.1 Model Preparation

The predicted value y_i of whether the customer buys the car is determined by multiple experience items as a given input value. The predicted value is a weighted linear combination of input eigenvalues, which is expressed as follows:

$$\widehat{y_i} = \sum_j \theta_i x_{ij} \tag{8}$$

Parameter θ is the uncertain part learned from the data. The better the parameter θ is, the better it can fit the training data x_i and the label y_i.

The objective function $obj(\theta)$ is to measure how good the model fits the data, which consists of training loss and regularization:

$$obj(\theta) = L(\theta) + \Omega(\theta) \tag{9}$$

$L(\theta)$ is the training loss function, and Ω is a regular term? The training loss function measures the prediction ability of the model on the training set. For logical regression problems, the loss function $L(\theta)$ is a logical loss, and the formula is as follows:

$$L(\theta) = \sum_i \left[y_i \ln\left(1 + e^{-\hat{y}_i}\right) + (1 - y_i) \ln\left(1 + e^{\hat{y}_i}\right) \right] \tag{10}$$

The function of the regularization term is to control the complexity of the model, which can avoid the occurrence of overfitting, that is, deviation variance equilibrium.

Regularity $\Omega(\theta)$ is added by complexity $\Omega(f_i)$. The formula of single complexity is as follows:

$$\Omega(f_i) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{11}$$

where, γ and λ are parameters; T is the number of leaves, namely the number of customers; ω is the leaf node score vector, that is, the selection result of the customer for the experience item. The objective function can therefore be expressed as:

$$obj = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega(f_i) \tag{12}$$

2.4.1.2 Model Training

The data used in the training set are the new data obtained after data cleaning in appendix 1, question 1, and the new data are imported for model training.

For the validation set of the model, we adopt k-fold cross-validation to determine the network structure and complexity parameters γ, λ of the model. For a customer experience item dataset D, it is randomly divided into k subsets of similar size but mutually exclusive. Each time, use the union set of $k - 1$ subsets as the training set, and the remaining subset as the test set; in this way, the training set/test set of k groups of this experience item can be obtained, so that k times of training and testing can be carried out, and finally the average result of the k times of testing is returned. Usually, k is 10, that is, 10-fold cross validation.

2.5 Marketing Strategy Planning for Unpurchased Customers

One of the customers who do not purchase electric vehicles from each brand is selected to implement the marketing strategy, so that they purchase electric vehicles. Increasing service intensity can improve the experience satisfaction. The more the percentage is increased, the higher the difficulty is, and the maximum cannot be greater than 5%. This problem can be transformed into an optimization problem, that is, to find out the strategy to improve the minimum experience satisfaction when customers purchase vehicles.

The selection of intelligent algorithms has a good effect on solving the optimization problem. In this paper, the artificial fish swarm algorithm is used, and the algorithm is robust. The requirements for parameter setting are not high and the allowable range is large. Has good global optimization ability, can quickly jump out of local optimum [6].

2.5.1 Determination of Fitness Function

In this problem, the minimum increase in experience satisfaction can be taken as the optimization objective of the model, and at the same time, the maximum increase in experience satisfaction needs to be met by 5%. The established customer mining model is used to predict whether customers purchase electric vehicles after improving satisfaction. When customers purchase vehicles, the sum of the eight percentage increases in satisfaction is taken as the fitness. When satisfaction increases but customers do not

purchase electric vehicles, a higher fitness is given. Assuming that the *j*th percentage increase in satisfaction of customer *i* is a_{ij}, and whether customers i purchase vehicles is buy_i, the fitness function can be established as follows:

$$minFitness = \begin{cases} \sum_{j=1}^{8} a_{ij} \, (buy_i = 1, i = 3, 4, \cdots, 14, 15) \\ 100 \, (buy_i = 0) \end{cases} \tag{13}$$

The constraint condition is:

$$0 \leq a_{ij} \leq 5\% \tag{14}$$

2.5.2 Determination of Crowding Factor

Crowding factor affects the optimization results by controlling whether artificial fish perform tail chasing and clustering behaviour. The introduction of δ avoids the over-crowding of artificial fish and falls into the local optimal solution. At the same time, this parameter will make the artificial fish in the attachment to the extreme point have the effect of mutual exclusion, and it is difficult to accurately approximate to the extreme point. Therefore, it is necessary to determine the appropriate crowding factor. Since this topic is to find the minimum value of the fitness function, the crowding factor can be defined as:

$$\delta = \alpha n_{max}, 0 < \alpha \leq 1 \tag{15}$$

Where α is the extreme close level and n_{max} is the maximum number of artificial fishes expected to aggregate in the neighbourhood.

3 Results and Discussion

3.1 Data Cleaning

3.1.1 Modification of Abnormal Data

1) Scope error
   According to the scope and the screening function of Excel, the abnormal value of each type of experience data set is found. If an experience data set has an abnormal value, the scatter diagram is drawn and the position of the abnormal value is locked. The experience data sets with abnormal data are a1, a3, a5, and B17. The Fig. 2 is the scatter plot of the data of a5.

   Customer No. 1 of data a1, customer No. 1964 of data a3, and customer No. 480 of data a5 have filled in abnormal data greater than 100. We use the average number of the sum of corresponding normal data to replace them.

   Customer 223 of data B17 filled in an abnormal data and deleted the data.
2) Logic error

- When B6 = 3 or 4, the number of B7 children is not 0, directly corrected to 0.
- When B11 = 3 or 4 or 6, B12 = 9, the B11 data is corrected to 7.

If there is data less than B14 in B13, the data of B13 are deleted.

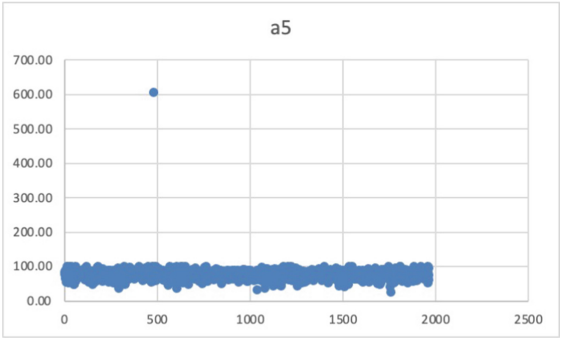

Fig. 2. a5 data scatter diagram.

Table 2. RSR and Probit distribution table.

brand	Value of RSR	Average of rank/n ∗ 100%	Value of Probit
1	0.375	33.3	4.569
2	0.625	66.7	5.431
3	1.000	91.7	6.383

Table 3. Sorting results.

brand	Value of RSR	RSR Ranking	RSR fitting value	Ranking level
1	0.375	3	0.359	1
2	0.625	2	0.656	2
3	1.000	1	0.985	3

3.2 Satisfaction Analysis

For the data of a1, a2, a3…, a8, the ranks of 1, 2, 3 cars are calculated, and the *Probit* value is calculated in Table 2.

The RSR critical value is calculated, and the interval comparison is carried out through the RSR fitting value and the RSR critical value (fitting value) in the previous table, so as to obtain the grading level. The larger the grading level number is, the higher the grading level is, that is, the better the effect is, and the higher the customer satisfaction is (Table 3).

Table 4. Summary of influencing factors.

brand	Customers are willing to buy the car's remarkable impact experience project
1	a1, a3, B16, B17
2	a1, a3, a5, B11, B15, B16, B17
3	a2, a3, B6, B16

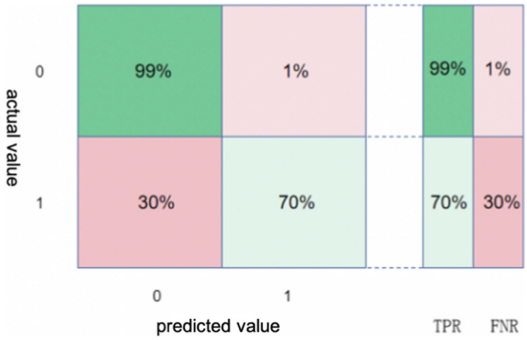

Fig. 3. TPR and FNR

3.3 Analysis of the Most Influential Indicators on Electric Vehicle Sales

Spss software was used to solve the results of multivariate logistic regression model, and the chi-square value was used to quantify the results. The indicators with chi-square value less than 0.05 were selected as indicators with significant influence on the purchase of the car. The influencing factors are summarized in the Table 4.

Through the intersection of experience data, we can know that a1, a3, B16, B17 are the general factors that affect customers' willingness to buy electric vehicles. That is, the customer's own economic burden strength, the customer's evaluation of vehicle comfort and safety will affect the sales of different brands of electric vehicles.

3.4 Forecast the Purchase Strategy of Target Customers

3.4.1 Post-training Model Evaluation

1) accuracy
   Import the cleaned data into the model, use ten-fold cross validation for training, and evaluate the trained model. The confusion matrix after training is shown in Figs. 3 and 4.

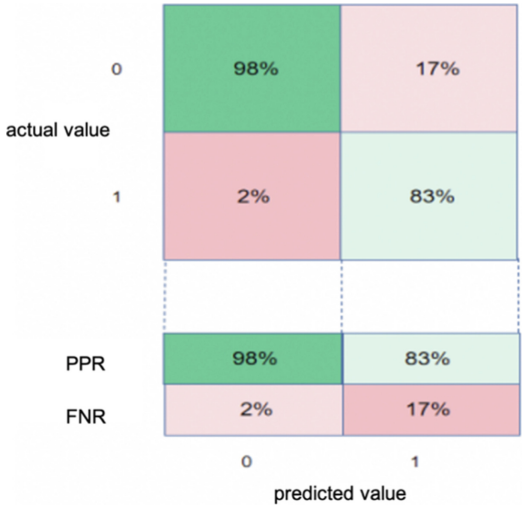

Fig. 4. PPR and FNR

Under the treatment of confusion matrix, the TP part was as high as 98−99%, and the TN and FP parts were low, indicating that the model had high accuracy and strong prediction ability. Accuracy is the ratio of the number of samples to the total number of samples, which can be represented by four numerical values of the confusion matrix:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{16}$$

The accuracy of this training model is 96.9%, and the accuracy is high and credible.

2) ROC and AUC

The ROC curve is ordinated by the true class rate (TPR), and the false positive class rate (FPR) is abscissa. The closer the curve is to 1, the better the effect of the model is. The Fig. 5 are ROC and AUC curves of the model.

It can be intuitively seen that the area surrounded by the ROC curve of the model is about 0.93, indicating that the overall curve is close to 1, and the generalization ability of the model is strong.

3.4.2 Forecasting Using Models

Using the trained model to predict the data, the prediction results are shown in Table 5.

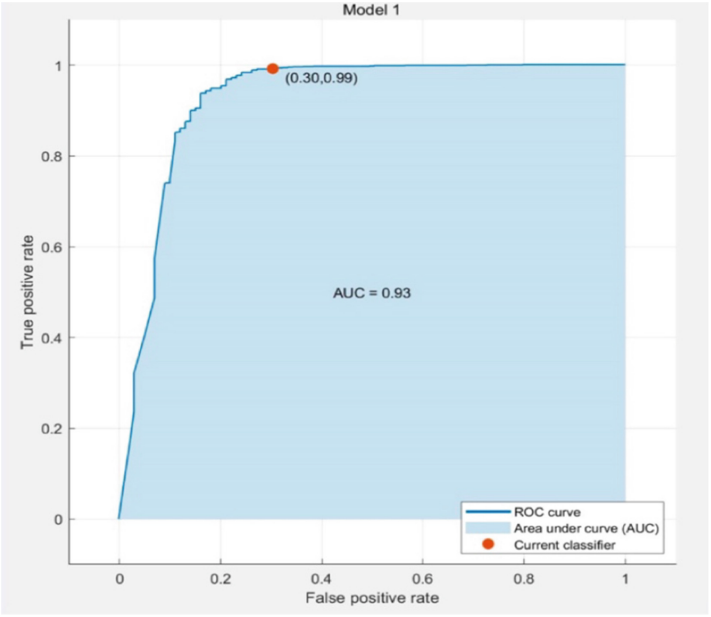

Fig. 5. ROC and AUC curve

Table 5. Prediction results of customer purchase behavior

client	Buy or not (1 Buy)	client	Buy or not (1 Buy)
1	1	8	0
2	1	9	0
3	0	10	0
4	0	11	1
5	1	12	1
6	1	13	0
7	1	14	0

3.5 Solving Marketing Strategy Planning

3.5.1 Algorithm Parameter Determination

In most cases, customers do not purchase electric vehicles, so in most cases, the fitness is given a higher fitness when they do not purchase electric vehicles, which is easy to fall into local optimum. Therefore, attention should be paid to getting rid of the local optimum value in the setting of parameters. At the same time, the setting of parameters should also make the calculation speed not too slow. Therefore, the parameters are set as shown in Table 6.

Table 6. Algorithm parameters.

index	parameter
Population size	20
Visual	5
step	3
crowding factor	17
number of replications	30

Table 7. Fitness results.

client	fitness	client	fitness
1	100	8	100
2	100	9	100
3	100	10	100
4	100	11	100
5	100	12	100
6	100	13	100
7	100	14	3.804

3.5.2 Problem Solving

The unpurchased electric vehicle customer information is substituted into the model to solve the fitness results are shown in Table 7.

It can be seen from the results that the optimal fitness of customers of customer 14 is 100, and the eight satisfactions of customers of customer 14 are increased by 5% and then predicted again. The results are still not without purchasing vehicles, that is, it is impossible to implement marketing strategies for these customers. The analysis of these customers shows that the car loans and housing loans of customers 3, 4, 5, 8, 9 and 13 are high, so they cannot purchase vehicles by improving services, while customers 10 and 15 are too low to increase their satisfaction by only 5%. The artificial fish swarm algorithm is used to solve the sales strategy of customer 14. The results are shown in Fig. 6 and Table 8.

So, in the battery technology performance satisfaction to increase service will increase satisfaction 3.804% can make customers 14 car purchase.

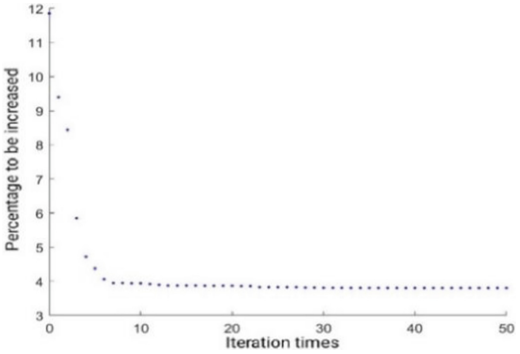

Fig. 6. Iterative graph of artificial fish swarm algorithm

Table 8. Optimal fitness results.

index	Optimal fitness	index	Optimal fitness
a1	3.804	a5	0
a2	0	a6	0
a3	0	a7	0
a4	0	a8	0

4 Conclusions

In this paper, the rank sum ratio evaluation model, multiple logistic regression model, Boosted tree model and artificial fish swarm algorithm are established to analyse the customer satisfaction of different automobile brands in the sales process of electric vehicles. At the same time, the important indicators affecting the sales of automobiles are obtained by the model, and the purchase behaviour of target customers is predicted. The intelligent algorithm is used to plan the sales strategy. The results are reliable and have a certain role in solving the problems in the sales process of electric vehicles.

References

1. Yang G, Xue X, Zhao F (2019) User rating prediction model and application based on XGBoost algorithm. Modern Libr Inf Technol 003(001):118–126
2. Guo J, Zhou Q, Weng H et al (2015) Multilevel logistic regression model analysis of health service utilization and influencing factors of floating population. China Health Econ (3):3
3. Wu Q, Dan Z (2003) Comparison of rank sum ratio method and several common evaluation methods in medical quality evaluation. Chinese Hosp Stat 10(1):3–5
4. Siebert LC, Filho J, Júnior E et al (2019) Predicting customer satisfaction for distribution companies using machine learning. Int J Energy Sector Manage. ahead-of-print(ahead-of-print)

5. Wu T, Zhang B, Wang Y et al (2007) Summary of data cleaning research. Modern Libr Inf Technol 2(12)
6. Li X, Lu F, Tian C et al (2004) Application of artificial fish swarm algorithm for combinatorial optimization problems. J Shandong Univ Eng Edn 34(5)