ISSN: 2688-9323

Research on Logistic Model in Judging Car Satisfaction

Chuan Sun[1,2], Mingxia Wu[1,2], Jing Yang[1,2]

[1]China Automotive Technology&Research Center Co., Ltd. Tianjin China

[2]China Auto Information Technology (Tianjin) Co., Ltd. Tianjin China

Abstract

Firstly, the sample data was analyzed and preprocessed. Second, in order to quantitatively study which factors will affect the sales of different brands of electric cars, we choose to build the logistic model. First, we divide all target customers into three kinds based on the brand of new energy cars, and build the logistic model for each kind. Secondly, in allusion to the first kind of model, battery technical performance, economy, expenditure on annual mortgage, and expenditure on annual car loan have a greater impact on joint venture brands, in allusion to the second kind of model, battery technical performance, safety, household disposable income, expenditure on annual mortgage, and expenditure on annual car loan have a greater impact on self-owned brands. Then for the third kind of model, because the collinearity among different factors is too strong, we build the model after eliminating the effect of collinearity, obtain that comfort, economy, safety, and expenditure on annual mortgage have a greater impact on the new power brands.

Keywords

New Energy Car; Logistic Model; Model Checking; Dynamic Programming.

1. Background and Restatement of Problem

With the development of the automobile industry and the development and utilization of new energy technologies, under the strong support of national policies, China's new energy car industry has achieved rapid development in recent years, and it also has the basis for rapid development. Judging from the market, with the rapid improvement of domestic new energy cars in the fields of intelligent networking and other new technologies, new energy cars have been recognized and accepted by increasing young mainstream consumer groups, new energy cars are gradually becoming one of the main sources of domestic car consumption. Under the dominance of the globalization trend, new energy cars are no longer as simple as an emerging industry, they are developing at an alarming rate and quickly seizing the domestic car market, according to data released by the CPCA, in November 2020, the retail volume of new energy cars reached 119000, increased by 97000 year-on-year, year-on-year increase was 136.5%.

Therefore, different brands of car companies have begun to increase their emphasis on the sales of new energy cars. Since new energy cars are a new industry, the public has eight aspects of their performance such as comfort (environmental protection and space seats), economy (energy consumption and hedging value), and safety performance (brake and driving vision), and the market popularity of new energy cars is not high. A car company launched three brands, in order to study consumers' willingness to buy electric cars and formulate corresponding sales strategies, the sales department invited nearly 2000 target customers to experience the three brands of electric cars and conducted market research.

2. Assumptions of Model

(1) The target customer's information in the sample data is true and effective.

(2) Whether the target customer chooses to buy car is only related to these factors and has nothing to do with other factors.

(3) One third of the data in in column B7 sample data one is ###, so we treat ### as 0 for processing

3. Description of Main Symbols

Table 1. Symbol description

symbol	symbol description
a_ii=1⋯8	the i-th index
y	probability of purchasing car
w_ii=1⋯8	the weight of the i-th index
A	sum of all indexes multiplied by weights
k	direct proportion coefficient, constant

4. Building and Solution of Model

4.1. Data Preprocessing

By comparing sample data two with sample data one, we found that there is the following logical relationship among different data: 1. Every data is not zero except b16 and b17; the number range is from a1 to a7, b16, the number range of b17 is greater than or equal to 0, less than or equal to 100; both b14 and b15 should be less than b13;

When processing missing data, for the messy code phenomenon in the data in b7 (the number of children), we believe that one-third of the missing data will not appear in a set of data, so we change all messy codes to 0. We did not find any vacant data through the screening, therefore, according to the logical relationship 1, we re-screened whether any data was 0, after the second screening, no data was found to be 0, therefore, we believe that there is no missing data in this set of data.

When processing abnormal data, firstly, we found out that the abnormal point is far away from other objects after preliminary searching, so we choose the distance abnormal point detection based on KNN, this method uses the size of k nearest neighbor distances to measure whether an object is far away from most points, namely, the abnormal score of an object is the sum d of the k nearest neighbor distances from the object to it, under the premise of a given threshold t, as long as d>t, the sample is considered to be abnormal point. We screened out 4 abnormal points in this way. Secondly, according to the logical relationship 3, there are 93 abnormal points where the household income is less than the personal income, and finally 97 abnormal points are obtained.

Since all the data are filled in by individuals, there is no connection among the data, it is impossible to correct or complete the data through interpolation, therefore, we decided to eliminate all abnormal data and carried out statistical description for some data.

4.2. Statistical Description of Data

The averages of satisfaction scores of all the surveyed people for from a1 to a8 are shown in Table 2 below:

Table 2. Averages of a1-a8

a1	a2	a3	a4	a5	a6	a7	a8
77.9447	78.1624	75.9437	78.8826	77.2605	77.91106	78.0572	77.6362

The variance of satisfaction scores of all the surveyed people for a1 to a8 is shown in Table 3 below:

Table 3. Variances of a1-a8

a1	a2	a3	a4	a5	a6	a7	a8
78.2874	80.5761	109.6833	81.8328	87.7275	86.6205	84.1745	90.1010

The averages and variances of everyone's scores for each performance are shown in Fig 1 below:

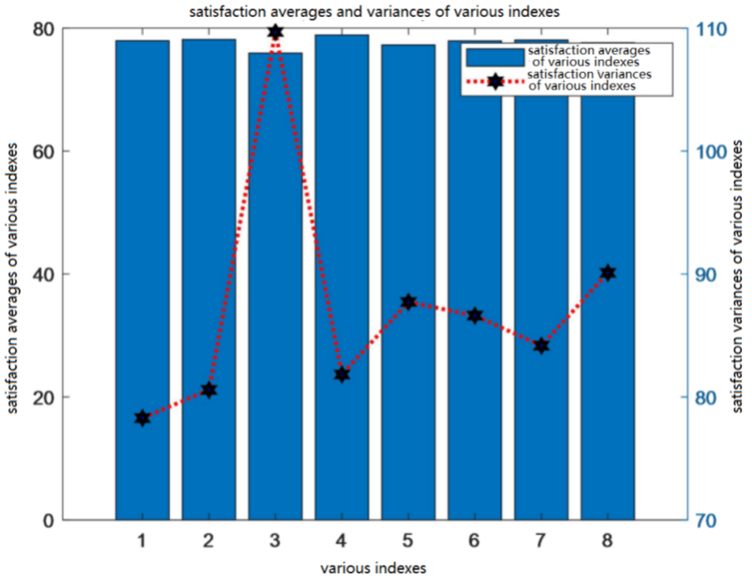

Fig 1. Satisfaction average and variance of various indexes

It can be seen that everyone's scores for 8 performance indexes of car are at about 76 points, they are basically satisfied with the car's performance, but the satisfaction variance is too large, it shows that everyone's views on the car's performance are extreme, people who are unwilling to buy car will give a low score to the performance of these indexes, and people who are unwilling to buy car will give a lower score to the performance of these indexes.

Among all the people who bought car, 1 is the number of people whose family income is less than one hundred thousand; 2 is the number of people whose family income is less than half a million; 3 is the number of people whose family income is less than one million; 4 is the number of people whose family income is more than one million, as shown in Table 4 below:

Table 4. The number of people with household income who buy cars

	t1	t2	t3	t4
the number of people	0	89	6	4

The ratio of people who buy cars with different household incomes is shown in Fig 2:

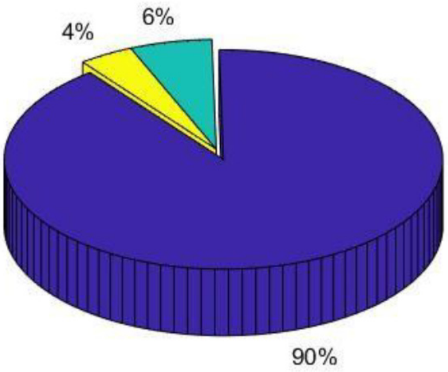

Fig 2. The ratio of people with different household incomes in the people who buy car

We can see that people with household incomes between 100000 and 500000 are the main car buyers from the above figure.

4.3. Analysis

The final intention of target customers is only divided into two kinds, buy or not, namely yes or no, the many factors given in question one may affect the target customer's willingness to buy car, but the influence is different, the impact of different brands is not necessarily the same, so in order to quantitatively analyze the influence, we choose to build the logistic model.

4.4. Building of Logistic Model

Logistic regression is mainly used in pathology, which is usually used to study the risk factors of a certain disease, and predict the occurrence of the disease based on the risk factors. Since it is impossible to determine the influence of each factor on the buy intention of the target customers, we select all the above factors as the influencing factors, and build the Logistic regression model to study the influencing factors of customers' buying.

We divide all target customers into three kinds based on the brand of new energy cars, and build the logistic model for each kind.

The logistic model built is as follows:

$$\ln\frac{p}{1-p}=\beta_0+\sum_{k=1}^{n}\beta_k x_k$$

p is the customers' buying probability, the formula is as follows:

$$p=\frac{\exp(\beta_k+\sum_{k=1}^{n}\beta_k x_k)}{1+\exp(\beta_k+\sum_{k=1}^{n}\beta_k x_k)}$$

(1) The first kind:

The programming solution is carried out through MATLAB, the results are as follows:

After screenings many times, it can be concluded that only the test p-value of x1, x3, x24, x25 (corresponding factors: x1-a1 battery performance, x3-a3 comfort, x24-b16 the ratio of the annual mortgage expenditure in the total family income, x25-b17 the ratio of annual car loan expenditure in the family total income) is less than 0.05, namely, these four factors affect the buying of the first kind of electric car. All other values are discarded, and Logistic regression is carried out between the above independent variables and dependent variables, the results are as follows:

Table 5. Pre-estimation results of the first kind of parameters

parameter	freedom	estimated value	standard error	t value	p value
constant term	1	-593.0447	684.6562	-0.8662	0.3864
X1	1	0.3841	0.1432	2.6824	0.0073
X2	1	0.1221	0.1257	0.9714	0.3314
X3	1	0.2267	0.0845	2.6831	0.0073
X4	1	-0.1186	0.1063	-1.1160	0.2644
...	...	...	...	...	...

Table 6. Estimated results of the first kind of parameters

parameter	freedom	estimated value	standard error	t value	p value
constant term	1	-29.3338	5.5197	-5.3144	0.0000
X1	1	0.1490	0.0612	2.4353	0.0149
X2	1	0.1971	0.0557	3.5392	0.0004
X3	1	-0.2136	0.0550	-3.8850	0.0001
X4	1	-0.1839	0.0483	-3.8078	0.0001

Corresponding factors, x1-a1 battery performance, x2-a3 comfort, x3-b16 the ratio of annual mortgage expenditure in the family total annual income, x4-b17 the ratio of car loan expenditure in the family total annual income.

As can be seen from the above table, the p-value test for the constant term and the ratio of x1, x2, x3, and x4 are all less than 0.05, it suggests that comfort, safety, the ratio of annual mortgage expenditure in the family total income, the ratio of annual car loan expenditure in the family total income all pass the test.

On the basis of the above analysis of the results, we get the logistic regression equation as:

$$ln\frac{p}{1-p}=-29.3338+0.149X_1+0.1971X_2-0.2136X_3-0.1839X_4$$

(2) The second type:

Table 7. Pre-estimation results of the second kind of parameters

parameter	freedom	estimated value	standard error	t value	p value
constant term	1	446.9896	305.1177	1.4650	0.1429
X1	1	0.3317	0.0725	4.5748	0.0000
X2	1	0.0295	0.0692	0.4272	0.6693
X3	1	0.1437	0.0478	3.0039	0.0027
X4	1	-0.0020	0.0655	-0.0302	0.9759
...	...	...	...	...	...

After screenings many times, the test p-value of x1, x3, x23, x24, x25 (corresponding factors, x1-a1 battery performance, x3-a3 economy, x23-b15 household disposable income, x24-b16 the ratio of annual mortgage expenditure in the family total income, x25-b17 the ratio of annual car loan expenditure in the family total income) is less than 0.05, namely, these four factors affect the buying of the second kind of electric car, All other values are discarded, and Logistic regression is carried out between the above independent variables and dependent variables, the results are as follows:

Table 8. Estimated results of the second kind of parameters

parameter	freedom	estimated value	standard error	t value	p value
constant term	1	-38.1456	4.1774	-9.1314	0.0000
X1	1	0.3096	0.0493	6.2794	0.0000
X2	1	0.1215	0.0368	3.2794	0.0010
X3	1	0.0923	0.0188	4.9138	0.0000
X4	1	-0.2056	0.0327	-6.2914	0.0000
X5	1	-0.2175	0.0338	-6.4282	0.0000

Corresponding factors, x1-a1 battery performance, x3-a3 economy, x23-b15 household disposable annual income, x24-b16 the ratio of annual mortgage expenditure in the family total income, x25-b17 the ratio of annual car loan expenditure in the family total income.

It can be seen from the above table that the p-value test for the constant term and the ratio of x1, x2, x3, x4, and x5 are all less than 0.05, it suggests the comfort, safety, and household disposable annual income, the ratio of annual mortgage expenditure in the family total income, the ratio of annual car loan expenditure in the family total income all have pass the test, and the model has been well built.

On the basis of the above analysis of the results, we get the logistic regression equation as:

$$ln\frac{p}{1-p} = -38.1456 + 0.3096X_1 + 0.1215X_2 + 0.0923X_3 - 0.2056X_4 - 0.2175X_5$$

(3) The third kind:

Because the dimensions of the data are different, in order to eliminate the influence of the dimensions on the data, we standardize the data, and the results are as follows:

Table 9. Pre-estimation results of the third type of parameters

parameter	freedom	estimated value	standard error	t value	p value
constant term	1	-458.2465	135136936	0.0000	1.0000
X1	1	153.2113	58019964	0.0000	1.0000
X2	1	839.6071	96081553	0.0000	1.0000
X3	1	434.1867	61888618	0.0000	1.0000
X4	1	-58.3388	77070364	0.0000	1.0000
...	...	...	...	...	...

It can be seen from the above table that the p value of all parameters is 1; the above result is due to the strong collinearity among various factors, so we decide to eliminate the collinearity among the factors, and then carry out logistic regression. We adopt the method of removing one or several factors with the largest collinearity to eliminate the influence of collinearity. The collinearity analysis is carried out on factors, and the results are as follows:

Table 10. Collinearity analysis results

parameter	estimated value	standard error	t value	p value	variance inflation factor
constant term	16.5616	14.2458	1.1626	0.2455	0.0000
X1	0.0070	0.0023	3.0107	0.0027	3.8979
X2	-0.0034	0.0024	-1.4335	0.1523	5.1084
X3	0.0021	0.0015	1.3791	0.1685	2.7913
X4	0.0015	0.0023	0.6553	0.5126	4.7217
...	...	...	...	...	...

We found that some variables have values greater than 10 by observing the variance inflation factor, it suggests that there is collinearity among the factors

Table 11. Index collinearity diagnosis

order number	eigenvalue	condition index
1	6.5315	1.0000
2	3.2751	1.4122
3	3.0651	1.4598
4	1.6099	2.0142
...	...	...

Judging by the collinearity index, we found that the collinearity between factor 14 and factor 25 is very strong, and then principal component analysis is carried out to eliminate the effect of collinearity, the results obtained are as follows:

Table 12. Eigenvalues of factors

order number	eigenvalue	difference	contribution	accumulative contribution
1	6.5315	3.2565	0.2613	0.2613
2	3.2751	0.2099	0.1310	0.3923
3	3.0651	1.4553	0.1226	0.5149
4	1.6090	0.0278	0.0644	0.5793
...	...	...	...	...

It can be seen that the cumulative contribution of the first 8 factors is nearly 80%, according to the results of the first two brands, we found that the factor b13-17 is more important and should not be discarded, so we only do logistic regression on the first 8 and last 4 factors and get the following results:

Table 13. Parameter pre-estimation results of the third type after eliminating collinearity

parameter	freedom	estimated value	standard error	t value	p value
constant term	1	-33.3827	10.7936	-3.0928	0.0020
X1	1	0.0668	0.1609	0.4152	0.6780
X2	1	0.4201	0.2283	1.8402	0.0657
X3	1	0.1876	0.1212	1.5469	0.1219
X4	1	-0.2869	0.2166	-1.3251	0.1851
...	...	...	...	...	...

Then we gradually eliminate the factors with p-value greater than 0.05 from the largest p-value, finally eliminate until the remaining dependent variable p-values are all less than 0.05, and the result is the 4th, 5th, 6th, and 26th columns (corresponding factors, a2-comfort, a3-economy, a4-security, b16-the ratio of annual mortgage expenditure in the total annual household income) and has an impact on whether to buy car.

Logistic regression is carried out on the above two columns; and the results are as follows:

Table 14. Parameter estimation results of the third type after eliminating collinearity

parameter	freedom	estimated value	standard error	t value	p value
constant term	1	-30.5928	8.2061	-3.7281	0.0002
X1	1	0.3897	0.1342	2.9048	0.0037
X2	1	0.2399	0.0999	2.4005	0.0164
X3	1	-0.2734	0.1186	-2.3053	0.0212
X4	1	-0.1513	0.0603	-2.5092	0.0121

Corresponding factors: x1-comfort, x2-economy, x3-safety, x4- the ratio of annual mortgage expenditure in the family total income.

It can be seen from the above table that the p-value test for the constant term and the ratio of x1, x2, x3, and x4 are all less than 0.05, it suggests comfort, economy, safety, a the ratio of annual mortgage expenditure in the total annual household income, the ratio of annual car loans expenditure in the total annual household income all pass the test, and the model has been well built.

List the logistic regression function:

$$ln\frac{p}{1-p}=-30.5928+0.3897X_1+0.2399X_2-0.2734X_3-0.1513X_4$$

It can be seen from the model that the probability of each customer buying electric car is not 100%, so we set a standard, when the buying probability is greater than 0.5, we regard this customer as choosing to buy car.

4.5. The First Kind of Model

Fit test of the first type of logistic model of the previous question on the model.

Table 15. Fit test of the first type of model table

norm	only contain constant terms	contain linear term
AIC	192.4460	100.6041
BIC	196.7431	122.0897
-2logL	190.4460	90.6041

It can be seen from the above table that for the AIC only containing linear term and constant term, the value of BIC is smaller, so the model with linear term fits better, and it can be concluded that the model is well built.

Test of regression coefficient:

Table 16. Likelihood ratio test of the first type of model

test	chi-square	freedom	p value
likelihood ratio	99.8419	4	0.0000
Wald	48.0496	4	0.0000

It can be seen from the above table that the p-values of the likelihood ratio test and the wald test are both less than 0.05, and the model fits well, according to the built model, we analyze whether users of the first type of electric car buy or not in Table 4, and the results are as follows (Four decimal are retained, and values that are too small are treated as 0)

Table 17. Buying results of users of the first type model

user number	whether buy or not	buying probability
1	yes	0.8286
2	no	0.0103
3	no	0
4	no	0
5	no	0.0167

4.6. The Second Type of Model

According to the first test of previous question on the logistics model:

It can be seen from the above table that for the AIC only containing linear term than and constant term, the value of BIC is smaller, so the model with linear term fits better, and it can be concluded that the model is well built.

Table 18. Fit test of the second type of model

norm	only contain constant terms	contain linear term
AIC	506.2225	220.7297
BIC	511.3034	251.2149
-2logL	504.2225	208.7297

Test of regression coefficient:

Table 19. Likelihood ratio test of the second type of model

test	chi-square	freedom	p value
likelihood ratio	295.4928	5	0.0000
Wald	155.3671	5	0.0000

It can be seen from the above table that the p-values of the wald test of the likelihood ratio test are all less than 0.05, and the model fits well, according to the built model, analyze whether users of the second type of electric car buy or not in Table 4, the results are as follows (the probability keep four decimals, too small value is treated as 0)

Table 20. Buying results of users of the second type model

user number	whether buy or not	buying probability
6	yes	0.8676
7	no	0.1582
8	no	0
9	no	0
10	no	0.0015

4.7. The Third Type of Model

According to the first test of previous question on the logistics model:

Table 21. Fit test of the third type of model

norm	only contain constant terms	contain linear term
AIC	78.2407	43.7258
BIC	81.1459	58.2522
-2logL	76.2407	33.7258

It can be seen from the above table that for the AIC with only linear term and constant term, the value of BIC is smaller, so the model with linear term fits better, and it can be concluded that the model is well built.

Test of regression coefficient:

Table 22. Likelihood ratio test of the third type of model

test	chi-square	freedom	p value
likelihood ratio	42.5149	4	0.0000
Wald	25.8108	4	0.0000

It can be seen from the above table that the p-values of the wald test of the likelihood ratio test all are less than 0.05, the model fits well, according to the model built, analyze whether users of the second type of electric car buy or not in Table 4, the results are as follows:

Table 23. Buying results of user of the third type model

user number	whether buy or not	buying probability
11	no	0.2728
12	yes	0.7192
13	no	0.4704
14	no	0.0226
15	no	0.0126

5. Promotion and Evaluation of the Model

Advantages of the model

We analyzed the factors that may be abnormal for problem one, then took out these factors separately to analyze the abnormal points, so that the normal data would not be judged abnormal due to the large difference of other indexes. Moreover, when the amount of data is sufficient, we directly delete the wrong data, which saves time.

For the question two, we used the method of principal component analysis to solve the multicollinearity problem in logistic regression, and then used the method of gradually eliminating related factors from the logistic regression solution, and finally got the factors related to car buying.

Disadvantages of the model

The interpolation and fitting methods can be used to change the wrong data for question one, instead of direct deleting, so that the data has a higher credibility.

Only using principal component analysis find non-collinear factors to do logistic regression for the question two, we did not fundamentally build the new model to solve the complex relationship among various indexes.

Promotion of the model

The model built is easy to understand, and solves the interference of index collinearity on logistic regression, it has good applicability and is closer to real life; it can be used to solve similar problems and has good promotion.

6. Suggestions of Sales Strategies

On the basis of the above research and analysis, the important indexes that affect the buying rate of the company's three brands of electric cars are:

Table 24. Important factors affecting purchase rate

	influencing factor	mainly affect brand
a1	battery technical performance	1,2,3
a2	comfort (environmental protection and space seats)	3
a3	economy (energy consumption and value hedging rate)	1,2,3
a4	safety performance (brake and driving vision)	3
b15	household disposable income	2
b16	the ratio of annual mortgage expenditure in the family total income	1,2,3
b17	the ratio of annual car loans expenditure in the family total income	1,2

It can be seen from the table that the impact on the purchase rate of electric cars can be roughly divided into two aspects, namely car technology and target customer personal aspects. Therefore, we have the following suggestions for the sales department:

1. For the three brands, it is necessary to focus on explaining the technical performance and economy of the battery. The most critical part of electric car is battery technical performance, and it is beyond the characteristics of other types of cars, in the eyes of customers, the better the technical performance of electric cars, the better the performance of electric cars. Similarly, customers will have the psychology of shopping around, under the same brand, which electric vehicle is more economical, the more it will be favored by customers.

2. It is necessary to learn "teach students in accordance with their aptitude." The disposable annual income of different families is not the same, low-income families buy cars mainly for practicality, in allusion to families with lower disposable incomes, electric cars with relatively low prices and complete functionality should be promoted; high-income families buy cars mainly for increasing comfort in practicality, for families with higher annual disposable income, electric cars with powerful functions and very good experience should be promoted, which can reduce the mention of prices accordingly.

3. The new power brand is the latest product, so all aspects should be paid attention to in the sales process. Because the target customers' ability to accept new products is always relatively slow, it is necessary to have a detailed introduction to each aspect during the sales promotion process, accelerate the recognition of new products for customers, and facilitate the sales of cars of new power brands.

References

[1] Sun Lin, Wang Jinjing. China's New Energy Vehicles Enter into the "Fast Lane"[N]. CPPCC Daily, 2021-08-06 (005).

[2] Wang Xiaoyin, Li Zhi, Zhou Baoping. Mathematical Modeling and Mathematical Experiments (Third Edition), Beijing: Science Press, 2019.

[3] Sun Jianyang, Sun Tong, Luo Junlin. Investment Value Analysis and Development Strategy of China's New Energy Vehicle Industry: A Case Study of BYD[J]. New Silk Road Horizon, 2021(05): 82-83+86.

[4] Tian Xiaohui. Research on New Energy Vehicle Marketing Strategy[J]. Auto Time, 2020(20): 98-99.

[5] Li Fan. Research on Marketing Strategies of ZH New Energy Vehicles[D]. Xidian University, 2019.