Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 14, No. 2, 2024

Research on Potential Customer Mining Models for Electric Vehicles Based on Machine Learning Models and Service Growth Models

Jinxin Wang

GAC TOYOTA MOTOR CO., LTD, Guangzhou 510000, China

Abstract: This study investigates the factors that impact the sales of electric vehicles (EVs) across different brands. Initially, multiple linear regression was performed on the data, and the variance inflation factors were all found to be less than 10. For the first category of indicators, one-way ANOVA was conducted to separate the three types of brands. For the second category of indicators, a logistic regression model was developed to complete the selection. Through multiple rounds of filtering, the following indicators were determined: For the first category, indicators affecting Brand 1 are A1, A2, A3, A4; indicators affecting Brand 2 are A1, A3; indicators affecting Brand 3 are A1, A2, A3, A5. For all brands, the second category of influencing indicators includes B4, B10, B11, B16, B17. A potential customer mining model was then established. Logistic regression models were created based on the sales influencing factors for different brands, and the model formulas for the three types of brands were calculated. The purchase intentions of 15 target customers were predicted, with an accuracy rate of up to 90%. Finally, the study explores whether increasing service efforts could change customer purchase intentions. A stepwise percentage service growth model was established, incorporating an initial service difficulty coefficient to determine how service satisfaction could be altered. Customers 2, 8, and 11 were selected for validation, and their respective service increment percentages were derived, leading to sales strategies tailored to these three customers.

Keywords: Single factor variance; Logistic regression; Service intensity; Service growth model.

1. Introduction

The development of electric vehicles (EVs) has introduced new growth points and development opportunities for the automotive industry, driving technological innovation and enhancing international competitiveness. Additionally, EVs may stimulate the development of upstream and downstream industries, such as battery manufacturing and charging infrastructure. The growth of EVs also contributes to the transformation and upgrading of the economic structure, promotes the development of a green economy, and has the potential to alter lifestyles and urban traffic patterns. To increase the market share of electric vehicles, improve consumer satisfaction, and promote rapid development, this paper primarily studies the key factors influencing electric vehicles and how to identify potential customers and convert them into actual users.

The issue of predicting potential customers has appeared across various industries and has garnered significant attention from experts and scholars. Wang Shengjie [1] and others proposed a customer churn prediction study based on machine learning models. This approach aims to improve the accuracy of existing machine learning models and enhance their interpretability. The method constructs a prediction model through data preprocessing and feature engineering, selecting five key performance indicators to evaluate model performance. Experimental results show that the proposed model outperforms current mainstream machine learning prediction models on the selected evaluation indicators. This paper provides a scientific basis for telecommunications companies to develop targeted customer retention strategies.

Wang Yulin [2] proposed an integrated learning-based customer churn prediction model for the broadcasting and television industry in response to the increasingly competitive internet era. This model, which integrates multiple single prediction models with complementary advantages based on the Stacking ensemble framework, analyzes the training principles of different machine learning models to achieve accurate predictions of customer churn in the broadcasting industry. Research shows that the proposed model effectively improves the accuracy of customer churn prediction, benefiting customer retention efforts for broadcasting operators.

Gao Tianchen [3] addressed the issue of retaining existing clients and preventing potential client churn in the securities industry by proposing a broker client churn early warning model based on high-dimensional feature factors. The study first explores the definition of client churn in the securities industry and then proposes an independence screening method based on high-dimensional feature factors. Finally, customer churn early warning models for daily and weekly predictions are constructed based on the selected factors. The results indicate that the external sample AUC value of the daily churn early warning model can reach an average of over 0.95, demonstrating good predictive accuracy. It is evident that machine learning models and feature engineering have characteristics of low computational cost, high prediction efficiency, and greater model stability in preventing customer churn and identifying potential customers. Therefore, this paper applies machine learning models to mine potential customers in the automotive industry.

The customer service growth model is a comprehensive strategic framework aimed at significantly increasing the number of customers by improving service quality, expanding service coverage, and enhancing customer service experience and awareness. This model not only focuses on meeting current customer needs but also on establishing long-term, stable customer relationships and achieving natural customer

base expansion through word-of-mouth and customer loyalty programs.

In summary, this paper proposes an innovative potential customer mining model for electric vehicles, which cleverly combines machine learning algorithms with service growth strategies. It aims to address the increasingly competitive environment in the electric vehicle industry, accurately identify and mine more potential customers, and provide strong market support for EV manufacturers and their sales channels.

2. Research Roadmap

This research roadmap consists of four aspects. The first step involves data collection, including market research data and consumer data. The second step is the establishment of a customer mining model, with the primary focus on using logistic regression models for prediction. The third part involves developing a customer service growth model based on the results of the second step to identify more automotive customers. The fourth step is analyzing the results of the models. The research roadmap is illustrated in Figure 1 below.

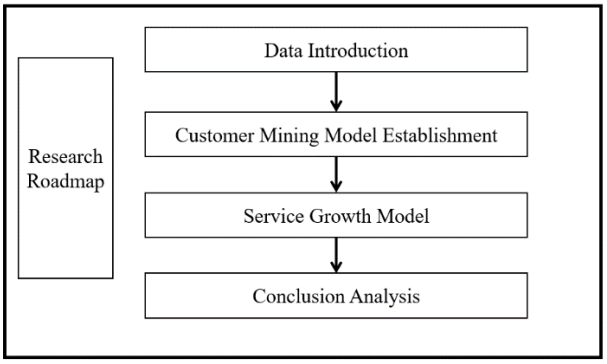

Figure 1. Research roadmap of this paper

2.1. Data Import

The data fields and their values used in this study are shown in Table 1.

Table 1. Data fields and their values

User Attribute	Variable Symbol	User Attribute	Variable Symbol
Purchase or Not	Y	Total Family Members	B5
Power Energy Type	P	Marital Status	B6
Battery Technology Performance	A1	Number of Children	B7
Comfort	A2	Age	B8
Economic Efficiency	A3	Education Level	B9
Safety	A4	Years of Work	B10
Power Performance	A5	Nature of Employer	B11
Driving Control	A6	Position	B12
Overall Appearance & Interior	A7	Annual Family Income	B13
Quality	A8	Personal Annual Income	B14
Household Registration	B1	Family Disposable Income	B15
Years of Residence	B2	Mortgage Ratio	B16
Living Area	B3	Car Loan Ratio	B17
Years of Driving	B4		

For Table 1, the goal is to identify which factors among many may influence the sales of different brands of electric vehicles. Since this involves both the attributes of the electric vehicles themselves and the personal characteristics of the users, the indicators are divided into two categories. The first category (a) consists of factors related to the electric vehicles themselves, and the second category (B) consists of the users' personal characteristics. Given the complexity of the indicators, different selection methods were applied to the two categories. For the first category of indicators, a single-factor analysis of variance and a probit regression model were used to select indicators for the three brands. For the second category of indicators, a logistic regression model was established to select the indicators related to users' personal characteristics.

2.1.1. The first type of index selection

The first type of indicator selection model was established. A multiple linear regression analysis was carried out on a total of 8 indicators A1-A8 of each brand, and the variance inflation factor was obtained. Table 2 shows the variance inflation factor of indicator A of the three brands.

Table 2. Variance inflation factor of A index of the three brands

Indicators	Brand 1	Brand 2	Brand 3
A1	3.5	3.2	3.7
A2	4.7	5.1	7.0
A3	2.4	2.6	3.4
A4	4.6	4.2	4.4
A5	3.9	4.4	4.4
A6	4.4	4.4	4.6
A7	4.0	4.0	4.9
A8	3.4	4.1	4.7

The variance inflation factors (VIFs) for the a indicators of the three brands are all less than 10, indicating no multicollinearity among the a indicators. Therefore, methods like principal component analysis, ridge regression, or partial least squares regression cannot be used [4-5]. Subsequently, a probit regression model was applied to the a indicators of the three brands. After conducting significance tests on the overall regression coefficients, goodness-of-fit tests, and

normality tests, it was found that the model fit for the second brand was very good, while the model fit for the first and third brands was poor. Since the data passed the tests for variance analysis, a one-way analysis of variance (ANOVA) model was established for the first and third brands. Table 3 shows the ANOVA results for Brand 1 and Brand 3.

Table 3. Results of ANOVA for Brand One and brand three

Indicators	Brand One	Brand Three
A1	0	0
A2	0	0
A3	0	0
A4	0	0.12
A5	0.0031	0
A6	0.0016	0.0033
A7	0.0046	0.0332
A8	0.0028	0.0094

Using the one-way analysis of Variance model, we select 4 indicators A1, A2, A3 and A4 for brand 1, and A1, A2, A3 and A5 for brand 3.

2.1.2. Selection of the second type of indicators

Because the dependent variable here is a categorical variable rather than a continuous variable, it violates the assumptions of linear regression analysis on data, and the methods of linear regression analysis are no longer applicable at this time. The logistic regression model is suitable for dealing with the case where the dependent variable is a categorical variable. The logistic regression model does not directly analyze the relationship between y and x, but analyzes the relationship between the probability p of y taking a certain value and the x value. Therefore, we consider using logistic regression model to screen the second type of indicators here.

Logistic model principle:

Probabilistic nonlinear regression model is used to study binary observations. logistic regression is essentially linear regression with a layer of functional mapping added to the feature-to-outcome mapping. Then the Logistic regression model can be represented as, here called the Logistic function.

$$P(y=1|x)=\pi(x)=\frac{1}{1+e^{-g(x)}} \quad f(x)=\frac{1}{1+e^{-x}}$$

Where. $g(x)=w_0+w_1x_1+\cdots+w_nx_n$ LR uses maximum likelihood estimation method to estimate the model, the partial derivative of the log-likelihood function, the LR loss function is the log-loss function, and the gradient descent method is used to minimize the loss function, so as to obtain the LR regression model.

According to the above formula combined with the probability density function of the discrete distribution, the following formula (1) is obtained.

$$p(y|x;\theta)=(h_\theta(x))^y(1-h_\theta(x))^{1-y} \quad (1)$$

According to the likelihood function, as in formula (2).

$$L(\theta)=p(\vec{y}|X;\theta)$$

$$L(\theta)=\prod_{i=1}^{m}p(y^{(i)}|x^{(i)};\theta) \quad (2)$$

$$L(\theta)=\prod_{i=1}^{m}(h_\theta(x^{(i)}))^{y^{(i)}}(1-h_\theta(x^{(i)}))^{1-y^{(i)}}$$

Therefore, the loss function of LR model is shown in formula (3).

$$l(\theta)=-\frac{1}{m}[\sum_{i=1}^{m}(y^{(i)}\log h_\theta(x^{(i)})+(1-y^{(i)})\log(1-h_\theta(x^{(i)})))] \quad (3)$$

After sorting out the two types of index selection models, the index selection model of the second question is obtained. The attributes of the three brands of electric vehicles are different, so the selection of the first type of indicators is different, respectively A1, A2, A3, A4 and A1, A3 and A1, A2, A3, A5. For the second type of user characteristics, a total of five indicators, B4, B10, B11, B16 and B17, are selected.

2.2. The establishment of customer mining model

By bringing the indicators of the three brands into the logistic regression model respectively, the constant value and value of each indicator can be obtained, so as to establish the logistic regression model of each brand, from which the probability of the user buying the brand car can be judged. The cleaned customer data is used as a training set to solve the optimal threshold value. If the probability of solving in the logistic regression model is greater than the threshold value, it is considered that the user will choose to buy the car; otherwise, it is considered that the user will not buy the car.

2.2.1. Customer mining model solution

First, the significance test of the overall regression coefficient of the three brand indicators and the goodness of fit test of the model are carried out. The results show that the fitting effect of the three brand indicators is very good (see the appendix for the results), which is suitable for the establishment of logistic regression model. Therefore, the indicators of the three brands were respectively brought into the logistic regression, and the constant values and β values were calculated respectively as shown in Table 4.

Table 4. Logistic regression parameter estimation table for the three brands

Brand I		Brand Two		Brand Three	
Constant value	-60	Constant value	-33.7	Constant value	-44.3
A1	0.3	A1	0.26	A1	-0.01
A2	0.2	A3	0.01	A2	0.23
A3	0.3	B4	0.07	A3	0.33
A4	-0.1	B10	-0.08	A5	0.01
B4	0.2	B11	0.34	B4	0.24
B10	-0.03	B16	-0.17	B10	-0.5
B11	0.6	B17	-0.19	B11	0.3
B16	-0.17			B16	-0.2
B17	-12.5			B17	-0.1

2.2.2. Predict the purchase intention of target customers

Based on the established customer mining model, 15 target customers will be judged whether they will buy the corresponding brand of electric vehicles. The prediction table is shown in Table 5 below.

From the results of Table 5, it can be seen that customers numbered 1 will buy the first brand electric vehicle, customers numbered 6 will buy the second brand electric vehicle, and customers numbered 12 and 13 will buy the third brand electric vehicle.

2.3. Establishment of service growth model

This section will discuss whether increasing the level of service in a short period can improve satisfaction with A1-A8 and thus change the customer's purchase intention from not buying to buying. Considering that the percentage growth is

relatively small, we established a stepwise percentage service growth model for simplicity. Based on this model, we use the prediction model discussed earlier to determine the purchase intention after different percentage increases. Generally, the difficulty of service is proportional to the percentage of satisfaction improvement. To reflect the objective reality that it becomes more difficult to improve satisfaction by the same percentage for users with higher initial scores, this study introduces the concept of α (service difficulty coefficient). The initial service difficulty coefficient is equivalent to the corresponding satisfaction score for the indicator. For example, if a customer's satisfaction score for A1 is 89.84, then α is set to 89.84. The optimal solution is obtained by using the minimum comprehensive service difficulty as the dependent variable.

Table 5. Prediction table of target customers' purchase intention

Customer Number	Brand number	Whether to buy (0 not enough to buy, 1 to buy)
1	1	1
2	1	0
3	1	0
4	1	0
5	1	0
6	2	1
7	2	0
8	2	0
9	2	0
10	2	0
11	3	0
12	3	1
13	3	1
14	3	0
15	3	0

2.3.1. Establishment of service growth model

(1) Data selection

Selection of target customers: According to the solution of the third question, P of customers 2,8,11 are all 0, and they have each experienced the first car, the second car and the third car are representative. We can try to solve and implement corresponding sales strategies for customers numbered 2,8,11 through the service growth model.

(2) Selection of variables: Since the difficulty of improving various indicators is proportional to the improvement of experience satisfaction, in order to ensure the minimum difficulty of comprehensive service, the index that has the greatest impact on the purchase result should be selected. According to the model of Question 2: On behalf of the users to take the first car data1 = A1, A2, A3, A4, B4 and B10, B11, bl3, B16, B17; On behalf of the user to select the second car data2 = a1, a3, a7, B4, B10, B11, bl3, B16, B17; On behalf of the user to select the third car data3 = A1, A2, A3, A5, B4, B10, B11, bl3, B16, B17.

(3) Model design principle: Taking the first car as an example, in order to simulate all the percentage growth conditions, four layers of cycles are used for the first four indicators of the first car, and six layers of steps are carried out in each cycle, totaling 6*6*6*6=1296 situations. Reset the data back to the original data after the end of each layer cycle, wait for the next cycle, and use i4+(i3-1)*6+(i2-1)*36+(i1-1)*216 to design the matrix with the change size of the cycle, deposit the increased data data1 into it, and get a total of 1296 groups of predictable data. At the same time, the corresponding increase percentage of each variable is stored in it, and the difficulty coefficient is stored in it. At this point, the matrix to be detected is successfully constructed, and the model of the third question is used to check and solve it.

2.3.2. Solution of service growth model

Using the obtained predictable data, the prediction model of the third question was used to conduct a prediction test. Customers with purchase intention greater than the threshold value were assigned a value of 1 and stored in the matrix. After that, the data with purchase intention were extracted, the comprehensive difficulty coefficient was compared, the smallest difficulty coefficient value was taken as the optimal solution, and the corresponding service growth amount of each index was output. The service increment of three different customer experience indicators is shown in Table 6.

Table 6. Shows the service increment of three different customer experience indicators

Vehicle type \ increase percentage Ratio (%) \ variable name	Client 2	Client 8	Customer 11
A1	0	2	0
A2	0	0	3
A3	3	5	5
A4	0	0	0
A5	0	0	0
A6	0	0	0
A7	0	0	0
A8	0	0	0

2.3.3. Analysis of results

(1) For customer 2: If A3 satisfaction is only increased by 3 percentage points, it can happen that the user will change from no purchase to purchase, and at this time, the service difficulty coefficient is the lowest in the case of purchase intention change.

(2) For customer 8: When A1 satisfaction is increased by 2 percentage points and A3 satisfaction is increased by 5 percentage points, the user's intention can happen to change from no purchase to purchase, and at this time the service difficulty coefficient is the lowest in the case of purchase intention change.

(3) For customer 11: When A2 satisfaction is increased by 3 percentage points and A3 satisfaction is increased by 5 percentage points, the user's intention can happen to change from no purchase to purchase, and at this time, the service difficulty coefficient is the lowest in the case of purchase intention change.

2.3.4. Sales strategy

Sales Strategy for the First Car (Joint Venture Brand): As shown in the table, when A3 increases by 3 percentage points, Customer 2's purchase intention changes from not buying to buying. In actual sales promotion, emphasizing the economic efficiency of joint venture brand vehicles can lead to higher customer satisfaction with the vehicle's economy, thereby boosting the sales of joint venture vehicles.

Sales Strategy for the Second Car (Domestic Brand): As shown in the table, when a increases by 2 percentage points and a3 increases by 5 percentage points, Customer 8's purchase intention changes from not buying to buying. In actual sales promotion, emphasizing the economic efficiency of domestic brand vehicles, with an increase of up to 105% in economic efficiency, and further promoting the battery technology performance of domestic brand vehicles can drive the sales of domestic vehicles.

Sales Strategy for the Third Car (New Energy Brand): As shown in the table, when A2 increases by 3 percentage points and A3 increases by 5 percentage points, Customer 11's purchase intention changes from not buying to buying. In actual sales promotion, emphasizing the economic efficiency of new energy brand vehicles, with an increase of up to 105% in economic efficiency, and further promoting the comfort of new energy brand vehicles can drive the sales of new energy vehicles.

3. Conclusion Analysis

3.1. Model Advantages

During data cleaning, logical judgment was used to filter customer personal characteristics, making the cleaned data more accurate and reflective of actual conditions. Due to the complexity of the indicators, which include both quantitative and qualitative factors with varying dimensions, multiple methods were used to select the indicators. This ensured more accurate selection of indicators related to both brand-specific factors and customer personal characteristics. Based on the data, a logistic regression model was established to calculate the impact of each indicator on whether a customer would purchase a car. This quantified the importance of each indicator, enabling the car company to have a clear understanding of the significance of each factor, leading to more efficient improvement of relevant automotive factors. It also provided a better understanding of the target audience for the car, allowing for more targeted promotion of different car brands.

3.2. Model Limitations

During variance analysis, kurtosis and skewness tests, along with distribution graphs, were used for normality testing. However, the accuracy may not be as high as absolute normality testing, and the data may not be perfectly normally distributed. When using the logistic regression model to determine whether customers would purchase a particular brand of car, the accuracy was higher for the first and third brands, but not as high for the second brand. This could be due to the lower accuracy when selecting indicators using the probit regression model, suggesting that the model could be further optimized.

3.3. Model Expansion and Improvement

The indicator selection model and target customer mining model have broad applications. They can be used in various commercial scenarios, such as insurance promotion, for improving products and identifying target customers. Additionally, a satisfaction improvement model was established, which can be applied to many service industries to find the most efficient ways to enhance customer satisfaction.

References

[1] Wang Shengjie, Zhang Qinghong. Research on customer churn prediction in telecommunication industry based on explainable machine learning model [J]. Telecommunication Science, 2024, 40 (07): 121-133.

[2] WANG Yu-lin. Research on Customer Churn Prediction Model of Broadcasting and Television based on Ensemble Learning [J]. TV technology, 2024, 48 (07) : 59-66. DOI: 10.16280 / j.v ideoe. 2024.07.014.

[3] GAO Tian-Chen, Qu Hao, Wang Feifei, et al. Research on customer churn warning model of securities companies based on high-dimensional feature factors [J]. Journal of Economic Management, 2023, 2 (04): 143-168.

[4] Li Ziyi, Wu Xinyang. Research on expert Missing information processing method based on multiple interpolation and Logistic regression [J]. Information Technology and Informatization, 2024, (08): 44-48.

[5] Miao Weijie, Wu Wenyuan. Privacy protection logic regression model based on homomorphic encryption training program [J/OL]. Computer engineering, 1-12 [2024-09-09]. https://doi.org/10.19678/j.issn.1000-3428.0069639.

[6] WANG Jinting, WANG Shouyang, Zhu Sheng. Risk analysis of queuing Game in Service system under mean-variance criterion [J/OL]. Science in China: Mathematics, 1-14 [2024-09-09]. http://kns.cnki.net/kcms/detail/11.5836.O1.20240807.1541.002.html.