Tourism Management and Technology Economy (2021) 4: 40-43
Clausius Scientific Press, Canada

DOI: 10.23977/tmte.2021.040207
ISSN 2616-2199

Sales Strategy Model of Electric Vehicle Target Customers Based on XGBoost Algorithm

Jiahui Wang, Anni Ye

School of Mathematics, Hangzhou Normal University, Hangzhou, Zhejiang, 310000, China

Keywords: Data Cleaning, Customer Mining Model, XGboost Algorithm

Abstract: This paper studies the sales strategy of target customers of electric vehicles, and establishes customer mining models for three brands of electric vehicles, so that the sales department can make scientific decisions and realize the rapid development of the enterprise. Firstly, we deal with the main characteristics of the three brands and group the predicted target customers according to different experience brands; Secondly, based on xgboost algorithm, we establish customer mining models for three brands with customer purchase intention as the objective function. The results show that the prediction accuracy of the model for the purchase behavior of three brand target customers reaches 96.83%, 97.21% and 95.00% respectively.

1. Introduction

With the sustained and rapid development of China's economy, people have a strong demand for cars. In order to improve automobile performance and meet people's needs, automobile enterprises continue to increase automobile R &amp; D investment and adjust marketing strategies. In order to understand the performance indicators and audience characteristics of the newly launched automobile brand, in order to formulate reasonable sales strategies, an automobile company invited 1964 target customers to experience and conduct research on three brand electric vehicles. Based on the basic situation of customers and the score data of electric vehicle experience satisfaction, this paper carries out data analysis and mining modeling, studies the internal factors of automobile sales, establishes the target customer mining model of different brands of electric vehicles combined with the research results, and evaluates the performance of the model, so as to help the company make scientific decision on sales plan.

2. Model Establishment and Solution

2.1 Data Preprocessing

When establishing the customer mining model of independent brand and new power brand electric vehicles, keep the main variables and eliminate the other variables. In particular, among the extracted main variables, one variable is to further refine the customer's personal characteristics. We build a new classification standard to classify the characteristics. Specifically, we use Excel to assign 1 to the target customers who meet the married / cohabiting childless situation of marriage and family, and the rest to 0.

Due to the different brands of electric vehicles experienced by target customers, we need to group the predicted target customers in order to make a reasonable prediction for different vehicle sales. Now we use Excel to group and get the data files corresponding to three brands: predict1.csv, predict2.csv and predict3.csv.

2.2 Establishment of customer mining model based on XGBoost algorithm

In order to further improve the prediction effect of customer purchase intention, this paper adds a regular term function to the original gradient lifting decision tree algorithm based on XGBoost algorithm to avoid over fitting problem. At the same time, the second-order derivative technology is introduced to solve the optimization problem, and a customer mining model with the purchase intention of target customers as the output is established.

Assuming that S can be used to describe the purchase intention of the target customer for a brand of electric vehicle, the purchase intention of customer i can be recorded as S_i

For customers experiencing joint venture brand electric vehicles, the purchase intention can be expressed as:

$$S_i(1) = F(X3, X4, X10\_3, X15, X17, X18)$$

For customers who experience their own brand electric vehicles, their purchase intention can be expressed as:

$$\mathrm{S_i(2) = F(X1, X3, X14, X15, X17, X18)}$$

For customers experiencing new power brand electric vehicles, the purchase intention can be expressed as:

$$S_i(3) = F(X2, X3, X9, X17)$$

The optimization function of target customers' purchase intention is:

$$\mathrm{S_i} = \sum_{\mathrm{i=1}}^{\mathrm{q}} \mathrm{L}(\mathrm{y_i}, \overline{\mathrm{y}_\imath}^{\mathrm{t}}) + \sum_{\mathrm{i=i}}^{\mathrm{t}} \theta(\mathrm{g_i}) = \sum_{\mathrm{i=1}}^{\mathrm{q}} \mathrm{L}\left(\mathrm{y_i}, \overline{\mathrm{y}_\imath}^{\mathrm{t-1}} + \mathrm{g_t}(\mathrm{x_i})\right) + \theta(\mathrm{g_t}) + \mathrm{constant}$$

Where **L** is the loss function, $\boldsymbol{\theta}(\boldsymbol{g_t})$ is the regular term and $\mathbf{constant}$ is a constant.

The specific construction process of this function is as follows:

Step 1: build a decision tree

Let $g(x_i)$ be a decision tree (the number of leaf nodes is p), select the feature with the largest information gain, split the current node (recorded as o), and the left and right nodes after splitting are recorded as L and R respectively. The splitting gain can be expressed as:

$$\mathrm{Gain} = \mathrm{g_O} - \mathrm{g_L} - \mathrm{g_R}$$

Step 2: integrated decision tree

Based on the addition model, the single decision tree is superimposed layer by layer, and multiple weak classifiers are set into a strong classifier. The predicted value of the model at this time can be expressed as the addition model of N decision trees:

$$\overline{\mathrm{y}_\imath} = \sum_{\mathrm{n=1}}^{\mathrm{N}} \mathrm{g_n}(\mathrm{x_i})$$

In the t-th iteration, add the decision tree g_t that minimizes the objective function, and the predicted value of the model is:

$$\bar{y}_\imath^{\,t} = \sum_{n=1}^{t} g_n(x_i) = \bar{y}_\imath^{\,t-1} + g_t(x_i)$$

The objective function is: $S_i^{\ t} = \sum_{i=1}^{q} L(y_i, \bar{y}_\imath^{\,t})$

Using the second-order expansion of Taylor formula, the objective function can be further reduced to: $S_i^{\ t} = \sum_{i=1}^{q}[L(y_i, \bar{y}_\imath^{\,t}) + u_i g_t(x_i) + \frac{1}{2} v_i g_t^{\,2}(x_i)]$

Step 3: add regular term function

In order to avoid the over fitting problem, this paper adds the regular term function on the basis of the original model:

$$\theta(g_t) = \gamma P + \frac{1}{2}\lambda \sum_{i=1}^{P} \omega_i^{\,2}$$

Then, in the iteration of step t, the objective function can be transformed into:

$$S_i^{\ t} = \sum_{i=1}^{q} L(y_i, \bar{y}_\imath^{\,t}) + \sum_{i=i}^{t} \theta(g_i) = \sum_{i=1}^{q} L(y_i, \bar{y}_\imath^{\,t-1} + g_t(x_i)) + \theta(g_t) + constant$$

Set a is defined as the set of training samples at the leaf node, which can be transformed into the sum of quadratic functions:

$$S_i^{\ t} = \sum_{j=1}^{q} \left[u_i \omega_z(x_i) + \frac{1}{2} v_i \omega_{z(x_i)}^{\ \ 2}\right] + \gamma P + \frac{1}{2}\gamma \sum_{i=1}^{P} \omega_i^{\,2}$$

$$= \sum_{i=1}^{P} [(\sum\nolimits_{i \in A_j} u_i)\, \omega_i + \frac{1}{2}\left(\sum\nolimits_{i \in A_j} v_i + \lambda)\, \omega_i^{\,2}\right] + \gamma P$$

Let $G_j = \sum_{i \in A_j} u_i \, , H_j = \sum_{i \in A_j} v_i$, then the above formula can be transformed into:

$$S_i^{\ t} = \sum_{i=1}^{P} [G_j \omega_i + \frac{1}{2}(\omega_i + \lambda)\omega_i^{\,2}] + \gamma P$$

According to $\frac{\partial s_i^{\ t}}{\partial \omega_i} = 0$, we can get

$$\widehat{\omega_\imath} = -\frac{G}{H_i + \lambda}$$

By substituting $\widehat{\omega_\imath}$ into the objective function: $S_i^{\ t} = -\frac{1}{2}\sum_{i=1}^{P} \frac{G_i^{\,2}}{H_i + \lambda} + \gamma P$

Thus, the gain of each node splitting is:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^{\ 2}}{H_L + \lambda} + \frac{G_R^{\ 2}}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma$$

2.3 Model solution and analysis

On the one hand, the prediction accuracy of the customer mining models corresponding to joint venture brands, independent brands and new power brands reached 96.83%, 97.21% and 95.00% respectively.

On the other hand, the number of negative samples in the sales data of electric vehicles of various brands is much larger than the number of positive samples, and there is a class imbalance. Severely unbalanced class distribution will reduce the applicability of some evaluation criteria, while ROC curve is not affected by class distribution and is suitable for evaluating unbalanced data sets. Therefore, we use ROC curve to reasonably evaluate the prediction performance of three brand customer mining models.

The ROC curve of the customer mining model established for the sales of electric vehicles of joint venture brands, independent brands and new power brands is shown in Figure 1.

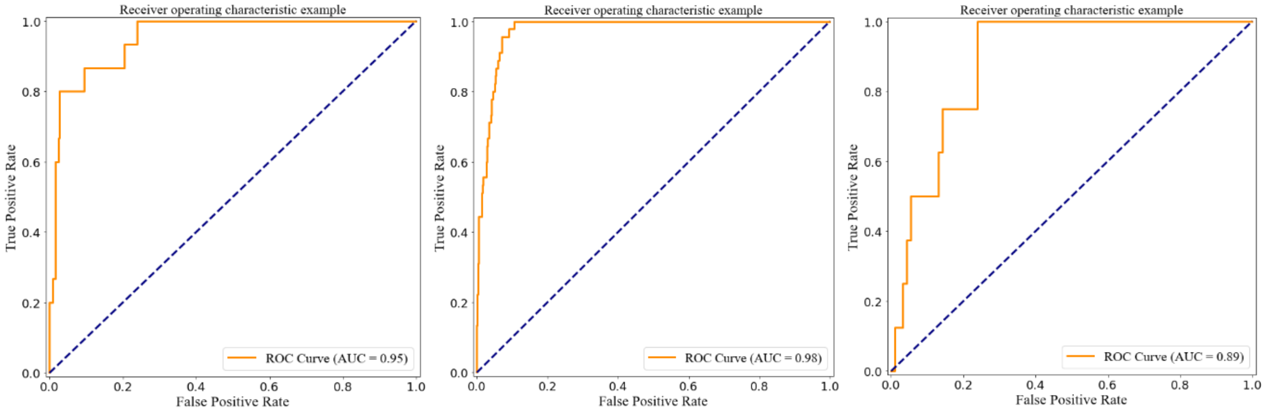

Figure 1: ROC curves of three brand customer mining models

3. Model Evaluation

The three brand customer mining models we established are based on the improved gradient lifting tree algorithm, which has the advantages of flexible objective function and difficult over fitting, and the prediction effect of the model is good. However, the customer mining model based on the improved gradient lifting decision tree algorithm contains more parameters, the process of adjusting parameters is complex, and whether the parameters are adjusted properly affects the prediction performance of the model.

References

[1] Xiong Luo, Li Xingguo. Research on Influencing Factors of family car purchase behavior [J]. Market weekly (Theoretical Research), 2014 (09): 44-47

[2] Li Qianyou. Logist model analysis of family income and car purchase intention [J]. Oriental corporate culture, 2012 (11): 233

[3] Li Bei, Hu Yi, min Shuhui, Guo Ruiqi. Construction of elderly care model selection model for urban elderly based on data mining technology [J]. China health management, 2021, 38 (07): 551-555

[4] Yang Guijun, Xu Xue, Zhao Fuqiang. User score prediction model based on xgboost algorithm and its application [J]. Data analysis and knowledge discovery, 2019, 3 (01): 118-1