Research on Distributed Photovoltaic Power Station Builders Segmentation Based on Data Mining

Ke Yang*, Hejian Wang, Junsheng Wang and Shaozhen Chen
State Grid Electronic Commerce Co., Ltd., Xicheng District, Beijing, China

*Corresponding author email: keyang@amss.ac.cn

Abstract. Simply according to the investment scale of distributed photovoltaic power station builders, the subdivision of the builder fails to fully consider their potential value, thus underestimating the contribution of the builder to the e-commerce platform. This paper constructs the value index evaluation system of the distributed photovoltaic power station from the current value and potential value of the builders. Based on the objective data of the builders and the clustering algorithm, this paper establishes a two-dimensional subdivision model of the builders based on the current value and potential value of the builders. This paper proposes specific marketing strategies for the value characteristics of each type of builders, which provides a scientific basis for e-commerce platform marketing decisions.

1. Introduction
In recent years, China's photovoltaic power generation industry has made rapid progress with the support of national policies, among which distributed photovoltaic power generation is surging forward. However, the source of funds has become an important factor restricting the development of distributed photovoltaic. Photovoltaic financial products based on the e-commerce platform emerge as the time requires, which not only solves the problems of large initial investment and long payback period faced by the photovoltaic power stations builders, but also realizes the good allocation of social surplus capital and solar energy resources, providing a strong driving force for the development of the photovoltaic industry. For the e-commerce platform, identifying the most valuable operators of distributed photovoltaic power station plays a significant role in promoting the revenue of the platform. Customer segmentation refers to dividing an enterprise's existing customers into different customer groups according to certain standards, which enables the enterprise to adopt differentiated marketing strategies for different customer groups, thus effectively reducing costs and achieving stronger and more profitable market penetration.

1.1. Literature Review
At present, segmentation from the perspective of customer value has been widely recognized by many scholars and researchers. Zhao [1] established a customer segmentation model for the communication industry based on two dimensions of customer loyalty and customer value. Meng et al. [2] proposed a three-dimensional customer segmentation model and measurement method based on the current dominant contribution, potential dominant contribution and potential implicit contribution of customers. Xu et al. [3] conducted clustering analysis on real estate customer value based on SOM neural network algorithm, and analyzed the purchase patterns and value characteristics of different target customer groups. Chu [4] applied the k-mean partitioning algorithm to subdivide the customers based on the current and potential value that mobile customers contribute to the enterprise during the
use of wireless Internet traffic. Wu [5] established a two-dimensional customer value segmentation model for the power customers based on the current and potential value of the customer, by constructing the customer value evaluation index system. Xie et al. [6] improved the traditional RFM model to the TFM customer segmentation model for the customer base of automobile 4S stores, and realized customer segmentation through k-mean clustering algorithm based on customer behavior. But they are all for B2C non-contract environments, and their model only applies to consumables that require frequent repeat purchases. However, once a distributed photovoltaic power station is built, it can usually operate for more than 10 years, and the customer value evaluation index system in the literature is no longer applicable.

1.2. Organization of the Paper
The rest of the paper is organized as follows: in the section 2, we construct the value evaluation index system for the constructors of distributed photovoltaic power stations. Then in the section 3 we establish a segmentation model for the real constructors' data through clustering method, and finally propose differentiated marketing strategies for each segmentation category. We conclude the whole paper in section 4 by summarizing our results and providing several directions worthy of future research.

2. The Construction of Customer Value Index System
Based on the objective data of the builder, this paper chooses to construct a customer value evaluation index system to measure customer value indirectly. This paper holds that the customer value includes two important aspects: the current value of the customer is reflected in the asset scale of the distributed photovoltaic power station, while the potential value of the customer is reflected in the value of the builder.

Since the asset scale of distributed photovoltaic power station directly determines the profitability of the power station, the asset scale of the power station can be used to describe the current value of the builder. This paper describes the scale of power station assets from two aspects: resource status and power station system. The available area attribute and built up area are selected to analyze the resource status of the builder, and the building cost and integrator attribute are selected to evaluate the power station system. The potential value of the builder can be characterized by two aspects: social capital and willingness to repay, in which the builder's social capital is evaluated by the builder's parents' social hierarchy, while the builder's willingness to repay is characterized by the employer's highest degree.

In the above analysis, the builder's value index is composed of 6 items. The indicator system of customer value measurement is shown in table 1.

| Customer value measurement | The current value | Resources | Available area attribute | Built up area |
|----------------------------|-------------------|-----------|--------------------------|---------------|
|                            |                   | Power plant system | Integrator attribute | Building cost |
|                            |                   | Social capital | Parental social hierarchy | |
|                            |                   | Repayment willingness | Highest degree | |

Table 1. An indicator system for measuring customer value.

3. Customer Segmentation Based on Clustering Model

3.1. Background
Fuzzy c-means (FCM) algorithm is a partitioning based clustering algorithm. Its idea is to maximize the similarity between objects divided into the same cluster and minimize the similarity between different clusters. FCM algorithm is one of the most commonly used clustering algorithms, which has
the characteristics of simple design, wide range of application and easy to realize by computer. It has been applied in many fields and achieved great achievements\cite{7-11}. Jiang et al.\cite{7} summarizes and classifies the existing clustering algorithms for data mining, and analyzes the performance characteristics of various algorithms. Chen et al.\cite{8} verified and analyzed the effectiveness of fuzzy c-means clustering algorithm, and summarized the improved methods. Fan\cite{9} summarized the research results of suppression fuzzy c-means clustering. Jia and Zhang\cite{10} used fuzzy c-means clustering algorithm in the recommendation system to achieve a good recommendation effect.

FCM is an improvement of k-mean algorithm. The main difference from the k-mean algorithm is that k-mean algorithm is rigid in data division, while FCM is a flexible fuzzy division. Membership degree valued at [0,1] is used to determine the degree to which each given data point belongs to each class. FCM divides n samples \(x_i (i=1,2,\cdots,n)\) into c clusters \(F_1,F_2,\cdots,F_c\), and finds the clustering center \(c_i\) of each group, so as to minimize the objective function of the non-similarity index. 

\(u_{ij}\) represents the degree to which the sample \(x_i\) belongs to the cluster \(F_j\), namely, \(0 \leq u_{ij} \leq 1\). \(u_{ij} = 1\) represents \(x_i\) completely belongs to the set \(F_j\), which is equivalent to the concept of a traditional set \(x_i \in F_j\).

\(u_i = (u_{i1},u_{i2},\cdots,u_{ic})\) is the membership vector of \(x_i\), after the normalization operation, the membership of a sample meet: \(\sum_{j=1}^{c} u_{ij} = 1, i=1,2,\cdots,n\). \(U = (u_{11},u_{12},\cdots,u_{nc})\) is the membership matrix for all samples. The objective function of FCM is defined as

\[
E = \sum_{j=1}^{c} \sum_{i=1}^{n} u_{ij}^m \text{dist}(x_i, c_j)^2,
\]

where, \(m\) is the fuzzy weighting coefficient, whose value is greater than 1. Through the Lagrange multiplier method, the clustering center and membership of formula (1) can be minimized. The Lagrange function of the above optimization problem is as follows:

\[
\bar{E} = \sum_{j=1}^{c} \sum_{i=1}^{n} u_{ij}^m \text{dist}(x_i, c_j)^2 + \sum_{j=1}^{c} \lambda_j (\sum_{i=1}^{n} u_{ij} - 1).
\]

By taking the partial derivative of the parameters, we can obtain:

\[
c_j = \frac{\sum_{i=1}^{n} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m},
\]

\[
u_{ij} = \frac{1}{\sum_{j=1}^{c} \left( \frac{\text{dist}(x_i, c_j)}{\text{dist}(x_i, c_k)} \right)}.
\]

The basic steps of FCM algorithm in this paper are as follows:

Input: sample set \(D = \{x_1, x_2,\ldots, x_n\}\); cluster number \(c\); fuzzy weighting coefficient \(m\).

Output: clustering center point vectors \(c\); fuzzy partition matrix \(U\).

1) Initialize the membership matrix \(U\) with a random number between [0,1] and make it meet the normalization conditions.

2) Calculate c clustering centers \(c_i, i=1,2,\cdots,c\), with formula (2).

3) Calculate the objective function according to formula (1). If it is less than a certain threshold, or if its change from the value of the last value function is less than a threshold, the algorithm stops.

4) Calculate the new \(U\) matrix with (3) and return 2).

The algorithm outputs c clustering center point vectors and a fuzzy partition matrix, which represents the membership of each sample point to each cluster. According to this partition matrix, the class of
each sample point is determined according to the maximum membership principle of fuzzy set, that is, the class labels of $x_i (i=1,2,\ldots,n)$ shall be determined by the following formula:

$$k \in \arg \max_{j=1,2,\ldots,c} \{u_{ij}\}.$$  \hspace{1cm} (4)

If more than one component is maximized, one class is selected at random.

3.2. Construction of Customer Segmentation Model

The experimental data in this paper are taken from the real data of 809 distributed photovoltaic power stations and normalized. The normalization method adopts the minimum and maximum to normalize, namely,

$$x_{i,j} = \frac{x_{i,j} - \min_{j} x_{i,j}}{\max_{j} x_{i,j} - \min_{j} x_{i,j}}.$$  

Since the current value and potential value of the builder are two independent dimensions, this paper adopts clustering on the two dimensions respectively, and then makes combination to realize the two-dimensional customer segmentation model. In this paper, fuzzy c-means clustering is adopted, and MATLAB 2015b is used as the programming environment (using its own fuzzy c-means function), and the algorithm results are analyzed.

The implementation results of the clustering algorithm have multiple measurement criteria. In this paper, four measurement criteria are adopted: Compactness (CP), Separation (SP), Davies-Bouldin Index (DBI) and Dunn Validity Index (DVI). The definition of Compactness is

$$CP = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{x \in C_i} \text{dist}(x, e_i),$$

which $CP = \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{x \in C_i} \text{dist}(x, e_i)$, $k$ represents the number of clusters. The smaller CP is, the closer the clustering distance within classes is. The definition of Separation is

$$SP = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j \neq i} \text{dist}(e_i, e_j),$$

the larger the SP is, the farther the clustering distance between classes is. The definition of Davies-Bouldin Index is

$$DBI = \frac{1}{k} \max_{i \neq j} \left( \frac{CP_i + CP_j}{\text{dist}(e_i, e_j)} \right),$$

smaller DB means less distance between classes and more distance within classes. The definition of Dunn Validity Index is

$$DVI = \frac{\min_{i \neq j} \min_{x \in C_i, y \in C_j} \text{dist}(x, y)}{\max_{i \neq j} \max_{x \in C_i, y \in C_j} \text{dist}(x, y)},$$

larger DVI means more distance between classes and less distance within classes.

3.3. Experiment Results

| Table 2. Performance of FCM clustering algorithm in current value dimension. |
|-----------------------------|-------------|-------------|-------------|-------------|
|                             | CP          | SP          | DBI         | DVI         |
| Own                         | 0.2377      | 0.2033      | 2.8528      | 0.0053      |
| Lease                       | 0.2412      | 0.2242      | 2.5205      | 0.0087      |

First of all, we cluster the samples in the current value dimension. Since the available area attribute is a binary variable, whose value is owned or leased, in which 599 samples are owned and 210 samples are leased, it is considered to cluster the current value respectively under each attribute. In this experiment, the algorithm was run several times by specifying input parameters, and it was found that the performance of the algorithm was best when divided into three categories. The algorithm results are shown in Table 2. Experiments show that under the two attributes of ownership and lease, the performance indexes of fuzzy c-means clustering are good, and the obtained clustering center is shown in Table 3 below.
Table 3. The clustering center of the current value dimension

| Feature       | \(c_1\)  | \(c_2\)  | \(c_3\)  |
|---------------|---------|---------|---------|
| Own Built up area | 0.2730  | 0.4582  | 0.2182  |
| Own Building cost | 0.2091  | 0.3814  | 0.1596  |
| Own Integrator attribute | 0.8437  | 0.2958  | 0.0715  |
| Lease Built up area | 0.2787  | 0.4996  | 0.2287  |
| Lease Building cost | 0.2187  | 0.4256  | 0.1741  |
| Lease Integrator attribute | 0.8072  | 0.3084  | 0.0973  |

As can be seen from table 3, under the two attributes of ownership and lease, the current value of the builder can be divided into two categories: high and low. The feature of the class center \(c_1\) is that the built up area and building cost of the power station are in the middle level, and the best integrators are used. The center of the class \(c_2\) is characterized by power station built up area and building cost are high, and using a medium level of integration. Overall the current value of the two classes is very high, so they all can be classified as high value. The feature of the class center \(c_3\) is that the built up area and building cost of the power station are low, and the worst integrators are used, so it is classified as low value.

Secondly, we cluster the samples in the potential value dimension. In this experiment, the performance of the algorithm is best when divided into three categories. The algorithm results are shown in table 4. Experiments show that the performance of fuzzy c-means clustering is very good, and the clustering center obtained is shown in table 5 below.

Table 4. Performance of FCM clustering algorithm in potential value dimension.

|     | CP     | SP     | DBI    | DVI    |
|-----|--------|--------|--------|--------|
| FCM | 0.1222 | 0.3806 | 0.4868 | 0.7071 |

Table 5. The clustering center of the potential value dimension.

| Feature                  | \(c_1\)  | \(c_2\)  | \(c_3\)  |
|--------------------------|---------|---------|---------|
| Parental social hierarchy| 0.7501  | 0.5006  | 0.1527  |
| Highest degree           | 0.9995  | 0.5056  | 0.0005  |

It can be clearly observed from table 5 that builders in potential value dimension can be divided into three categories: high, medium and low. High value builders are reflected in the high parental social hierarchy and the highest degree. Middle value builders’ parental social hierarchy and the highest degree are in the medium, and low value builders’ parental social hierarchy and the highest degree are low.

To sum up, the builders are subdivided from two dimensions, and the combination categories are shown in table 6 below.

Table 6. Customer value segmentation based on current and potential value.

| Value | Current | Potient value |
|-------|---------|---------------|
| Own   | (High, High) | (High, High) |
|       | (Low, High) | (High, Middle) |
|       | (High, Low) | (Low, Low)    |
| Lease | (High, High) | (High, High) |
|       | (Low, High) | (High, Middle) |
|       | (High, Low) | (Low, Low)    |
3.4. Marketing Strategy Based on Customer Value Segmentation Model

According to the classification results in table 6, under the condition that the available area attribute is owned and leased, the builders can be divided into 6 categories. Based on customer value, these six categories can be divided into three categories: high value, middle value and low value. High value includes (High, High) and (High, Middle); middle value includes (High, Low) and (Low, High); Low value includes (Low, Middle) and (Low, Low).

(1) High value builders. First of all, the quality of the power stations they build is good. If properly managed in the later stage, the power stations can achieve good profits. Second, their parents are of high social hierarchy and have easy access to social capital. Their own level of education is high, usually able to think rationally, repayment willingness will be strong. To sum up, for such customers, the e-commerce platform of State Grid should actively cooperate with them to achieve mutual benefit and win-win situation.

(2) Middle value builders. Their problem is that to access social capital is not easy, but the quality of power stations is good, or to access social capital is easy, but the quality of power stations is mediocre. For customers with low current value, the e-commerce platform of State Grid can cooperate with them, but it must strictly supervise the operation and maintenance of the power station it has put into construction to ensure the long-term safe and stable operation of the power station. For customers with low potential value, the e-commerce platform shall strictly review the credit records of the constructors, and actively cooperate with the builders with good credit. In addition, it can also be recommended to improve the scale of power station assets and turn them into high-quality contractors.

(3) Low value builders. For those with low current and potential value, the power station quality is relatively poor, and it is difficult for them to obtain social capital. Besides, they have a weak willingness to repay debts. Therefore, it is suggested that the e-commerce platform of State Grid should not cooperate with them. For the builders which current value is low, but the potential value is medium, it is better to strict review of their credit records and cooperate cautiously. Due to the extra cost of leasing the roof, and the inconvenient maintenance and other problems may be encountered during the operation of the power station, there is a certain risk of long-term stable operation. So the customer value of the builder whose available area attribute is his own is higher than that of the builder whose available area attribute is lease under the same value classification. As for the contractor, the leasing contract should also be strictly reviewed to ensure the long-term and stable operation of the power station.

4. Conclusion

Based on the objective and real data of the photovoltaic power station builter, this paper firstly constructs the builder value evaluation index system. Then, a two-dimensional subdivision model is established by cluster analysis. In our model, the value of constructors can be divided into three levels: high, middle and low, and the composition of each level can be reflected through the clustering center. Finally, a one-to-one marketing strategy is proposed for the segmentation model. In addition, our model can also observe the distribution of the number of customer groups, and identify the most important customer group and the largest customer group, which is of great help for precision marketing.

The marketing strategy presented in this paper provides a scientific basis for the marketing decision of photovoltaic financial products on the e-commerce platform. The model in this paper is also applicable to other similar new energy financial products. However, this paper only subdivides the qualifications of the builders based on objective data, and does not solve the problems of power station asset evaluation, which is an important topic worth studying.

Acknowledgment

This paper was financially supported by National Key Research and Development Program, under the Grants No.2018YFB1500800.
References

[1] Zhao B G. Customer segmentation model and empirical research [J]. Research on financial and economic issues, 2006, 272(7): 85-90.

[2] Meng Q L, Zou N J, Chen X J. Three-dimensional customer segmentation model for customer assets and its application [J]. Mathematical statistics and management, 2008, 27(5): 785-794.

[3] Xu X F, Xing D E, Zong X J. Research on customer value segmentation model based on PCA and SOM neural network algorithm [J]. Electrical automation, 2017, 39(3): 49-52.

[4] Chu G L. Research on telecom customer segmentation based on clustering model [J]. Statistics and decision making, 2014(8).

[5] Wu L, Li C C. Power customer segmentation research based on customer lifetime value [J]. Technology and industry, 2012, 12(11).

[6] Xie P S, Zhang K, Fan H J, Gui X Q, Zhang E Z. Research on TFM Customer Segmentation Model and Method of Automobile 4S Stores. Journal of Chinese Computer Systems, 2019, 40(10): 2165-2169.

[7] Jiang Y, Zhang C Y, Qiu P L, Zhou D F. Clustering algorithm for data mining [J]. Journal of electronics and information technology, 2005, 27(4): 655-662.

[8] Chen X, Fang F, Hu Z L. Overview of fuzzy clustering algorithm [J]. Life science instrument, 2013, 11(06): 33-37.

[9] Fan J L. A brief overview on suppressed fuzzy C-means clustering [J]. Journal of Xi'an University of Posts and telecommunications, 2014, 19(03): 1-5.

[10] Jia J J, Zhang Y C. Comprehensive trust recommendation algorithm based on user fuzzy clustering [J / OL]. Computer Engineering: 1-10 [2020-06-05] https://doi.org/10.19678/j.issn.1000-3428.0057930.

[11] Zhou Z H. Machine learning [M]. Tsinghua university press, 2016.