Research on discrimination algorithm of electric energy substitution user based on logistic regression model

Jihuan Chen 1, a, Yudong Lu 2, b, Qiang Liu 1, c, Lei Li 1, d, Bo Miao 3, e

1 State Grid Zhejiang Electric Power Company, Hangzhou, China
2 Zhejiang Electric Power Research Institute, Hangzhou, China
3 China Electric Power Research Institute, Beijing, China

a chenjihuan@vip.qq.com.cn; b hbyudo@126.com; c liuqiang@zj.sgcc.com.cn;
d li_lei@zj.sgcc.com.cn; e miaobo@epri.sgcc.com.cn

Abstract. The current power grid is found with low visiting efficiency, information integration difficulty and other problems in case of promotion of electric energy substitution. The big data technology is featured by high efficiency, high speed and convenience in terms of information processing, so it can realize the precise localization and differentiated marketing objectives of each enterprise. This paper, based on the analysis of energy consumption status of current power grid user and by means of "industry-field-user" architectural analysis, classifies user as per industry and equipment characteristics and combines with existing data of current power grid to exploit the energy consumption and electricity use characteristics of electric energy substitution user and build discrimination model of electric energy substitution user so as to improve the working efficiency and economic benefit of power grid business.

1. Introduction

Severe smog and other pollution problems are the consequence of unreasonable energy development and structural contradiction in China. To solve environmental problem, it is necessary to change the energy structure dominated by coal, reduce the use of coal, petroleum and other fossil energy as possible, control the energy capacity and adjust the energy consumption structure[1]. In this situation, the electric energy substitution strategy is proposed to actively promote the new energy consumption mode of "replacing coal by electricity, replacing oil by electricity, taking electricity from afar", continuously improve the proportion of electric energy in energy end-use, heavily optimize energy structure, promote energy conservation and emission reduction, relieve air pollution and improve environmental quality and sustainable economic development ability.

Currently, the research of domestic and foreign scholars on electric energy substitution mainly includes as follows: According to the research of Zhao Huiru and other scholars, the main factors affecting the electric energy consumption in our country include economic development level, electricity price, population growth, technological progress, energy policy adjustment and so on[2-4]. Literature [5] expounds the opportunity, challenge and policy selection in electric energy substitution strategy from the macroscopic point of view. Literature [6] combines with the hot water supply system of colleges and universities to comprehensively quantify and evaluate the economic benefit and environmental benefit from electric energy substitution so as to provide some reference for further promoting electric energy
substitution technology. Literature [7] expounds the advantages from overall implementation of electric energy substitution project from the prospective of energy conservation and emission reduction, including the improvement of electric energy proportion in end-use energy, reduction of urban smog and promotion of ecological civilization construction Literature [8] defines the electric energy substitution quantity to quantify the electric energy substitution potential, builds the environmental load model concerning electric energy substitution and uses the decoupling theory model to define many model parameters in many situations so as to realize the valid prediction of end electric energy substitution quantity under different situations.

According to above research literature, most of scholars develop their researches at macroscopic analysis level but less describe how to promote the electric energy substitution project in enterprise and user end. According to the specific business of electric power company, the current electric energy substitution is mainly completed in the form of field visit. However, this traditional method has two problems: one is low visiting efficiency and high blindness of customer manager and the other is failure to integrate and share information. Therefore, it is necessary to further exploit the characteristics of potential user in electric energy substitution project from the perspectives of theoretic research and actual business and use the big data analysis and model to reduce the electric energy substitution cost and improve substitution efficiency.

This paper takes the application of big data mining technology to Zhejiang power grid as an entry point to analyze the power consumption behavior characteristics of electricity user and build a reasonable and efficient potential user discrimination model of electric energy substitution so as to improve the specialization and accuracy of power grid enterprise in developing the electric energy substitution work.

2. Text part

2.1. Modeling idea and architecture design

This project adopts the "industry-field-user" architecture to build the potential user discrimination model: firstly, analyze the basic situation of electric energy substitution industry, classify the users into 19 fields upon different energy-consuming equipment used by power grid users in different industries and different production cycle and production characteristics in each industry; secondly, build the appropriate mathematical model parameters and further exploit users in each field. Taking the metal manufacturing industry as an example in this paper, the specific model building process is as shown in table 1.

| Industry          | Field                        | User mining                  |
|-------------------|------------------------------|------------------------------|
| Metal smelting    | Metal manufacturing field    | Characteristic analysis      |
| Metal finishing   | Textile printing and dyeing field | Index selection            |
| Metal pressing    |                              | Model building               |
| Garment making    |                              | Model verification           |
| Leather manufacturing |                        |                              |
| Fiber manufacturing|                              |                              |
| Education         | Public service and management field |                              |
| Medical treatment |                              |                              |
| Organ/team        |                              |                              |
2.2. Industry analysis and field division

According to National Industries Classification (GB/T4754-2002), this country includes 20 classes and 913 subclasses. When the industry selection is done for electric energy substitution, the Pareto principle shall be followed with a matter of emphasis, namely, focus on the industry with high proportion of coal, petroleum, natural gas and petrochemical energy.

I. According to the energy consumption structure in various industries across the country, the industry accounts for 69.4%. The non-metallic mineral product and manufacturing industry, chemical material product and manufacturing industry and metal smelting and rolling industry are main energy users and their energy are dominated by coal, electricity and petroleum. In five industries with maximum energy consumption, the total coal and petroleum consumption accounts for more than 50%. The promotion of electric energy substitution shall focus on manufacturing industry, processing industry and other high energy-consumption enterprises.

II. According to electricity consumption, in Zhejiang power grid, the high voltage and non-resident users accounted for 12.96% of total user in 2015 but their electricity consumption reached 85.9% of whole province, of which the leather manufacturing, metal smelting, metal finish machining and chemical manufacturing industries are main electricity users.

According to the instructions in Guidance on Promoting Electric Energy Substitution issued by the State Council, main energy-consuming equipment in each industry which may be subject to electric energy substitution concentrate mainly on electric boiler, kiln and heating. In combination with above two industry selection principles and the classification of energy-consuming equipment included in Guidance, select 84 key energy consumption industries from 913 subdivided trade structures and divide them into 19 fields and then build model for each field as follows:

| Field code | Field division                                                                 | Substituted equipment          |
|------------|--------------------------------------------------------------------------------|--------------------------------|
| 1          | Fishery industry                                                               | Water pump                     |
| 2          | Irrigation and drainage                                                        | Water pump                     |
| 3          | Agriculture, forest and animal product industry                                | Water pump                     |
| 4          | Construction industry                                                          | Heating                        |
| 5          | Traffic, storage and post industry                                             | Cool storage                   |
| 6          | Information transmission, computer service and software industry               | Heating, cooker                |
| 7          | Commerce, accommodation and catering industry                                  | Heating, cooker                |
| 8          | Finance, real estate, business and resident service industry                    | Heating, cooker                |
| 9          | Public service and management organization                                     | Heating, cooker                |
| 10         | Mining industry                                                                | Kiln                           |
| 11         | Food, drinking and tobacco industry                                            | Baking equipment               |
| 12         | Textile printing and dyeing industry                                           | Boiler                         |
| 13         | Metal casting industry                                                          | Boiler, kiln                   |
| 14         | Paper product, printing and sporting goods industry                            | Boiler, steamer                |
| 15         | Chemical engineering manufacturing industry                                     | Boiler                         |
| 16         | Rubber and plastic product industry                                            | Boiler, frequency furnace      |
| 17         | Non-metallic mineral product manufacturing industry                             | Boiler, frequency furnace      |
| 18         | Wood processing and product and furniture making industry                      | Boiler, steamer                |
| 19         | Manufacturing industry (transport, optics, processing)                          | Boiler                         |
2.3. Building of model index system

2.3.1. Analysis of model influence factor. The factors which are used to discriminate the enterprise energy consumption can be considered from following four aspects: basic enterprise information, enterprise production and operation information, enterprise electricity use characteristics and enterprise electricity use habit; these information can be acquired from power grid marketing system and electricity consumption acquisition system respectively:

(1) Analysis of basic enterprise information
In different industry, there are different production equipment, product cycle and electricity use characteristics to affect their energy consumption and energy utilization rule; in same industry, the energy-consuming equipment of the enterprises with different age have time-period characteristics. For example, those enterprises which are established after 2016 mainly adopt electricity and natural gas as energy source. As shown below, the age distribution has a certain difference between target enterprise and non-target enterprise due to industry, history and other factors.

![Figure 1. Age performance of target and non-target enterprises under different voltage](image)

(2) Analysis of enterprise production and operation information
Whether the business operation of an enterprise is brisk can directly its intention on subsequent production plan and equipment modification. The enterprise with good business operation has higher intention on electric energy substitution. In addition, the recent capacity reduction of an enterprise can also affect its substitution possibility. For the enterprise has capacity reduction recently, especially the reduced capacity accounts for high proportion in contract capacity, if its production progress is not delayed and the productivity is not reduced, then the enterprise is very likely to use other energy instead of electric energy.

(3) Analysis of enterprise electricity use characteristics
The enterprise electricity use characteristics are artificial division of users based on power grid use standard, mainly including four aspects: 1. user type; 2. supply voltage; 3. electricity use type; 4. operating capacity.

The user type includes high-voltage user, low-voltage non-resident user and low-voltage resident. Different types of users are very different in electricity use characteristics.

The supply voltage is classified as per the voltage when a user is registered. Because the high energy-consuming equipment of various enterprises have different voltage requirements, the index can play a good role in model identification. For example, most of 380V enterprises in textile industry are tiny workshops or OEMs and they generally don't use boiler or setting machine, so their possibility of electric energy substitution is very small.

The electricity use type is classified as per industry characteristics of enterprise, including great industry electricity, general industry and commerce electricity and resident electricity. The enterprises
with different electricity use type are obviously different in terms of electricity use cycle, on-peak demand and electricity consumption.

The operating capacity means the actual required capacity of user. Through analysis of enterprise scale, productivity and current capacity, it is allowed to know the proportion of electric energy of an enterprise in its production process. The heavier the proportion is, the larger possibility it uses other energy sources. As shown below, compared with non-target enterprise, the target enterprise has more distribution in terms of high operating capacity.

![Operating capacity density distribution map of target enterprise and non-target enterprise](image)

**Figure 2.** Operating capacity density distribution map of target enterprise and non-target enterprise

(4) Analysis of enterprise electricity use habit

The enterprise electricity use habit means the peak/valley electricity consumption and load electricity consumption data acquired on the basis of power grid electricity consumption acquisition system. There is significant difference in terms of peak/valley electricity and daily, monthly and quarterly electric load between the enterprise using coal, petroleum and natural gas and the enterprise using electric energy only. As shown in Fig.5, after the hardware, electric appliance and metal casting industries (upper left drawing) complete electric energy substitution, there are peak hours of electric load in wee hours. As we well known, for the metal casting industry, especially the casting industry, the metal smelting often adopts kiln equipment and the equipment is often used in wee hours (1:00am~6:00am). After the modification of electric energy substitution, the electric kiln replaces the coal-fired kiln, so the peak load occurs in these hours.

![Load curve of enterprises in different industries within 24h before and after modification of electric energy](image)

**Figure 3.** Load curve of enterprises in different industries within 24h before and after modification of electric energy
2.3.2. Selection and determination of model variables. Through analysis of above variable influence factors, the following 13 model variables are initially determined.

Table 3. Classification of variable model

|   | Basic information                  | industry field       |
|---|------------------------------------|----------------------|
| 2 | age                                | industry cycle       |
| 4 | industrial cycle                   | capacity reduction ratio |
| 6 | supply voltage                     | user classification  |
| 8 | Electricity use type               | operating capacity   |
| 10| peak electricity index             | daily load index     |
| 12| quarterly load index               | monthly peak electricity index |

In order to know the influence of each variable $x_i$ on substitution result, it is allowed to take above influence factor as independent variable and the substitution result $y_i$ as dependent variable and study the correlation of independent variable and dependent variable by means of correlation analysis.

Through correlation analysis of dependent variable (substitution result) and independent variable (as shown in Table 4), the correlation coefficient of voltage type, load type, capacity reduction ratio and peak electricity index with target variable is small, and these indexes have poor linear relation with model result and can be omitted in building model. According to the significance level of input variable and target variable, all indexes pass Pearson test, except that load type $P \ (0.77>0.05)$ fails to pass significance test.

Table 4. Correlation coefficient and significance level of index variable and target variable

| Index                        | P-value: pearson | P-value: spearman |
|------------------------------|------------------|-------------------|
| Operating capacity           | 0                | 0                 |
| Household age                | 0                | 0                 |
| Electricity use type         | 0                | 0                 |
| Voltage                      | 0.06             | 0                 |
| Load type                    | 0.77             | 0.9               |
| Capacity reduction ratio     | 0.03             | 0                 |
| Daily load index             | 0                | 0                 |
| Quarterly load index         | 0                | 0                 |
| Industrial cycle             | 0                | 0                 |
| Peak electricity index       | 0                | 0.19              |
| Monthly load index           | 0                | 0                 |
| Monthly peak electricity index | 0            | 0                 |

In addition, the inter-variable multicollinearity can be determined by kappa condition number. Through analysis of inter-variable multicollinearity, the kappa condition number is obtained as 12.776<100. Therefore, the model input variables are of low collinearity and can be used to build model in full.
2.4. Building of logistic regression model based on data mining

2.4.1. Technical principle of model. The specific formula of Logistic function (or known as Sigmoid function) is as follows:

\[ g(z) = \frac{1}{1+e^{-z}} \]  

(1)

\( e \) shows natural logarithm, \( z \) shows curve gradient;

For linear classification, the boundary type is as follows:

\[ \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n = \sum_{i=1}^{n} \theta_i x_i = \theta^T x \]  

(2)

Here \( \theta_0 \) shows constant, \( \theta_1, \theta_2, \ldots, \theta_n \) shows variation coefficient, \( x_1, x_2, \ldots, x_n \) shows specific variable; in this research, the variable means the specific electricity use index which can affect the electric energy substitution result of user, such as voltage, load, electricity quantity, electricity use type, enterprise scale and so on.

Based on the equations (1) and (2), the prediction function is constructed as follows:

\[ h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \]  

(3)

where \( g(\theta^T x) = g(z) \), \( g(\theta^T x) \) shows substitution result in this research, of which 1 shows substitution and 0 shows no substitution; therefore, the above equation can also be converted to:

\[ P(y = 1|x; \theta) = h_\theta(x) \]  

(4)

\[ P(y = 0|x; \theta) = 1 - h_\theta(x) \]  

(5)

The equations (4) and (5) can be summarized as follows:

\[ P(y|x; \theta) = (h_\theta(x))^y (1 - h_\theta(x))^{1-y} \]  

(6)

In case of variable processing, it is necessary to attach more importance to following two points: 1. The linearity of logistic regression model enables that there is unicity of influence direction in continuous variable. According to the actual business experience, however, some variables don't present this characteristics. For example, in case of voltage of 10KV, the enterprise substitution possibility is the highest; in case of voltage above or below 10KV, the possibility is relatively low. This is characterized by peak shape. The above problem can be solved by artificial factor transformation or sectioning for some variables, but this inevitably results in some data distortion. Therefore, the sectioning accuracy and quantity are important factor to ensure model effect; 2. Each variable has very large difference of absolute value and this is easy to cause excessive influence weight of single variable on \( \theta^T x \) function. Therefore, the continuous variable shall be subject to normalization processing before entering model variable.

By taking the logarithm of equation (6), it is allowed to infer the maximum likelihood estimation Cost function as:

\[ \text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases} \]  

(7)

Here Cost loss function adopts the maximum likelihood estimation. Therefore, the less Cost value is, the more convergent the function is, the better the model estimation effect is.
It is allowed to obtain the minimum value of Cost function by means of gradient descent which can give the renewal process of $\theta$:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta), \quad (j = 0, 1, \ldots, n)$$

(8)

By means of gradient descent, the equation (7) can be taken as:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left( h_{\theta}(x^{(i)}) - y^{(i)} \right) x_j^{(i)}, \quad (j = 0, 1, \ldots, n)$$

(9)

According to the equation (9), when Cost is the least, namely, the minimum residual \(h_{\theta}(x^{(i)}) - y^{(i)}\) sum, the best imitative effect of model function and actual result will be reached. In this research, Cost loss value is verified by comparison of logistic regression Akaike inspection and residual sum of R language.

2.4.2. Model output and application. The figure below shows the discrimination model output result of electric energy substitution potential users (metal manufacturing industry). In addition to the industrial cycle index, other variables will make obvious effect on model result (P<=0.05). When the electricity use type is great industry electricity and the voltage is 10KV, the electric energy substitution possibility of the enterprise will be large. For other variables, especially the load index and peak electricity index, the higher they are, the less the electric energy substitution possibility of the enterprise is. This is because that the higher load index shows the larger electric load fluctuation, the more proportion of electric energy in energy consumption of main production equipment and the less electric energy substitution possibility.

The model includes 30% of data as validation set to validate the applicability of model. According to Fig.4, the model has good coverage rate and hit rate and the hit rate is found with fluctuate in a small range in later half. It is proved through inspection that this is caused by over-fitting of model hit rate in first half because the target variables of some industries are too less and this has little influence on actual application of model.

![Figure 4. Inspection result of model validation set (R language analysis result)](image)

According to the model output result, adopt the query optimization technology in big data storage technology to solidify the model into power grid business system. In order to reach the quick label combination operation, the system will store the data relation of customer and model result in the form of Boolean value so as to ensure the update efficiency of business application and model result.

3. Conclusion

Since the electric energy substitution model was put into use in October 2017, we have completed the potential analysis of the provincial 4,500,000 enterprises other than resident user and made two field visits in Shaoxing and Jiaxing, involves with 4612 enterprises in total. The average hit rate reaches 23.2% and 30.7% respectively, increasing by 5~6 times compared with the visit efficiency before the model is built.
According to industry analysis and two visit results, the potential enterprises mainly distribute in textile, metal, rubber and other manufacturing industries but are few in traffic, information transmission and commercial fields. The visit result also verifies the energy-consuming analysis of the industry.

The research shows that the discrimination model of electric energy substitution potential user can quickly respond to business requirement, help business personnel to select electric energy substitution potential user from mass data and establish the relevant promotion strategy so as to develop the electric energy substitution work. Test results of the model in Shaoxing and Jiaxing show that the model is good to accurately locate user for business personnel and provide science tool for accuracy and differentiation of future promotion strategy and service.

Acknowledgments
This work was financially supported by Zhejiang Science and technology project Research on electric energy substitution potential mining and implementation effect evaluation based on big data analysis(5211DS17002T).

References
[1] Zhao Yinhui. Energy Substitution Analysis and Model Optimization Research under Beijing Objective of Energy Conservation And Emission Reduction[D]. North China Electric Power University Economic and Management School, 2014:1-171
[2] Zheng Jincheng. Analysis of Rural Electric Energy Substitution Potential and Environmental Benefit [D]. North China Electric Power University Economic and Management School, 2015:1-56.
[3] Zhang Lili. Research on Energy Structure Optimization of Zhejiang Province [D]. Zhejiang University of Technology, 2013:1-65.
[4] Maupi&Eric&Lentsolou&Maseka&Lesaoana.Determination of Causal Effect in Observational Studies: Analysis of Correlated Data with Binary End-Points [J]. Mathematics and System Science: English edition, 2012(2):119-125
[5] Li Zuofeng. Research and Practice of Jiangsu Energy Substitution [J]. Power Demand Side Management, 2016, 18(5):1-3
[6] National Bureau of Statistics Energy Statistics Office. China Energy Statistical Yearbook 2015[M]. Beijing: China Statistics Press, 2016
[7] Qu Ziqing, Xin Jieqing, Wu Liang, Zheng Siyuan. Research on Boundary Condition of Electric Energy Substitution Technical Selection for Commercial User [C]. Fifth Meeting of Third Power System Automation Committee of Chinese Society of Electrical Engineering--Symposium, 2015:1-17.
[8] Zhou Youxue. Promotion and Implementation of Electric Energy Substitution Technology [M]. Annual Meeting of Gansu Society of Electrical Engineering, 2014:1-5
[9] Zhang Yun. Starting Energy Consumption Mode--Electric Energy Substitution: Persisting in the Path of Clean, Environmental and Sustainable Development [J]. State Grid, 2013(10):34-40
[10] Xue Yusheng, Lai Yening. Integration of Big Energy Thinking and Big Data Thinking (I) Big Data and Electric Power Big Data [J]. Power System Automation, 2016, 40(1):1-8
[11] Duan Junhong, Zhang Naidan, Zhao Bo, Yan Xiaobin. Research on Foundation System Framework and Application of Electric Power Big Data [J]. Electric Power Information and Communication Technology, 2015, 13(2):92-95
[12] Heng Xingchen, Zhou Li. Application of Distributed Technology in Electric Power Big Data High-performance Processing [J]. Electric Power Information and Communication Technology, 2013, 11(9):40-43.
[13] Zhou Youxue, Wang Qi. State Grid Gansu Branch: Promotion of Electric Energy Substitution according to Local Conditions [J]. China Power Enterprise Management, 2015(2):22-29