Research on User's Discrimination Algorithm of Power Substitution Based on Logistic Regression Model

1st Qiang Liu
State Grid Zhejiang Electric Power Company
Hangzhou 310007, China
Liu_qiang_yx@zj.sgcc.com.cn

2nd Bo Miao
China Electric Power Research Institute
Beijing 100192, China
miaobo@epri.sgcc.com.cn

3rd Yi Liang *
School of Economics and Management
North China Electric Power University
Beijing 102206, China
lianglouis@126.com
*The Corresponding author

4th Changzu Li
School of Economics and Management
North China Electric Power University
Beijing 102206, China
1047510367@qq.com

5th Dongxiao Niu
School of Economics and Management
North China Electric Power University
Beijing 102206, China
ndx@ncepu.edu.cn

***Abstract*—State Grid exist the problems of low efficiency and integrating information in the promotion of power substitution, big data technology has the characteristics of high efficiency, high speed, convenient in information processing, can realize the purposes of accurate positioning, differentiated marketing. Based on the analysis of the current State Grid users, the paper summarizes the users' characteristics of clustering according to the industry and equipment through the "industry - field - user" three layer structure analysis method, combined with the exploration of the current network of existing data and finishing the digging energy, electricity characteristics, constructing power substitution of user discrimination model, so as to enhance State Grid's business efficiency and economic benefits.**

Keywords—Big data, power substitution, User's discrimination Logistic regression

I. INTRODUCTION

Serious pollution such as smog is the result of unreasonable energy development in China and long-term accumulation and concentration of structural contradictions. To solve environmental problems, we must change the coal-based energy structure, minimize the use of fossil energy such as coal and oil, control the total amount of energy and adjust the energy consumption structure [1]. Under this situation, the state has proposed a power substitution strategy, and actively advocates a new energy consumption model of "returning coal by electricity, replacing oil by electricity, and electricity from afar", continuously increasing the proportion of electric energy in terminal energy consumption, and vigorously optimizing energy structure. Promote energy conservation and emission reduction, reduce air pollution, and improve environmental quality and sustainable economic development.

At present, Zhao Huiru and other scholars, the main domestic and foreign scholars studying electric energy substitution, have studied the main factors which affect China's power consumption, such as economic development level, electricity price, population growth, technological progress, and adjustment of energy policy [2-4]. Literature [5] elaborates on the opportunities, challenges and policy choices in the energy substitution strategy from a macro level. Literature [6] combined with the university's domestic hot water supply system to specifically quantify the economic and environmental benefits brought by the replacement of electric energy, and comprehensively quantify the benefits of electric energy replacement in the domestic hot water supply system, in order to provide a reference for further promotion of electric energy replacement technology. Literature [7] expounded the advantages of comprehensive implementation of electric energy substitution projects from the perspective of energy conservation and emission reduction, not only improving the proportion of electric energy in terminal energy consumption, but also playing an important role in alleviating urban haze and promoting ecological civilization. Literature [8] defines the amount of electric energy substitution to quantify the potential of electric energy substitution, establishes an environmental load model for electric energy substitution, and determines various model parameters under multiple scenarios through decoupling theory model to achieve an effective prediction of the amount of power replacement in the terminal.

It can be seen from the above research literature that most scholars' research directions on electric energy substitution mainly focus on the macro analysis level, but less on how to promote electric energy substitution projects in enterprises and users. It can be seen from the specific business of the power company that the current electric energy substitution reform is mainly completed by means of manual field visits, but there are two major problems in this traditional way: First, the efficiency of field visits is low, and the account manager is blind; It is impossible to integrate and share information. Therefore, it is necessary to carry out in-depth mining of potential user features in electric energy substitution projects from both theoretical research and practical business perspectives, using big data analysis and application models to reduce the cost of electric energy replacement and improve the efficiency of transformation.

  DOI: 10.25236/iccpb.2018.036

This paper is based on the application of big data mining technology in Zhejiang Power Grid. By analyzing the characteristics of electricity consumption behavior of electricity customers, this paper constructs a reasonable and efficient user substitution model for electric energy replacement potential of power grid companies, thus improving the specialization and accuracy of grid enterprises in carrying out electric energy substitution work.

II. Modeling Ideas and Architecture Design

This paper uses the "industry-domain-user" three-tier architecture to build a potential user discriminant model: Firstly, the basic situation of the electric energy substitution industry is analyzed. According to the difference of energy-consuming equipment used by grid users in different industries and the differences in production cycle and production characteristics of various industries, users are divided into 19 fields, and then appropriate mathematics is constructed for different fields. And dig deep into the users in each area. This paper takes metal manufacturing as an example. The specific process of model construction is shown in Figure 1.

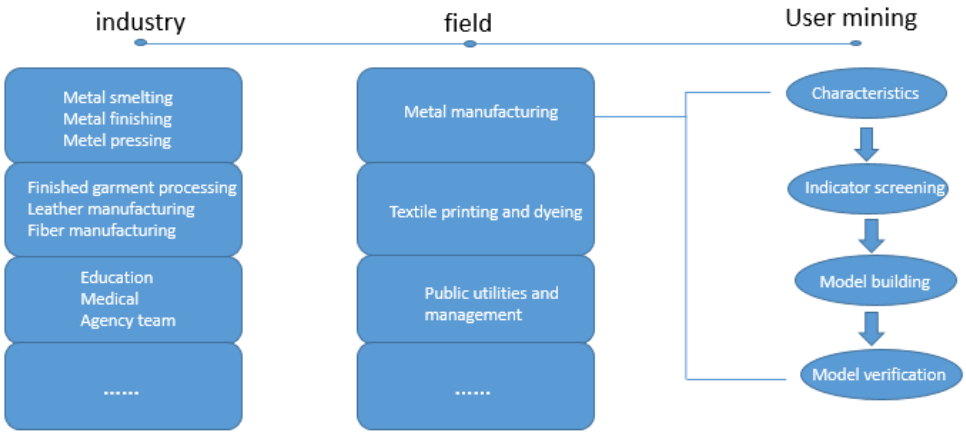

Figure. 1 "Industry - Domain - User" three-tier architecture

According to the "National Economic Industry Classification" (GB/T4754-2002), there are 913 subcategories of 20 categories in China. The replacement of electric energy should follow the "28 principle" in the selection of industries, with some emphasis on selecting petrochemicals such as coal, oil and natural gas. Energy has a relatively high priority in the industry.

(1) From the perspective of energy consumption structure of various industries in the country, industrial consumption accounts for 69.4%. Non-metallic mineral products industry, chemical raw material product manufacturing, metal smelting and rolling processing industry are the main energy-consuming units, and the main energy consumption is mainly coal, electricity and oil. Among the five industries with the highest energy consumption, the total consumption of coal and oil is more than 50% (Figure 2). The replacement of electric energy should focus on manufacturing and Transformation of industrial energy-intensive enterprises.

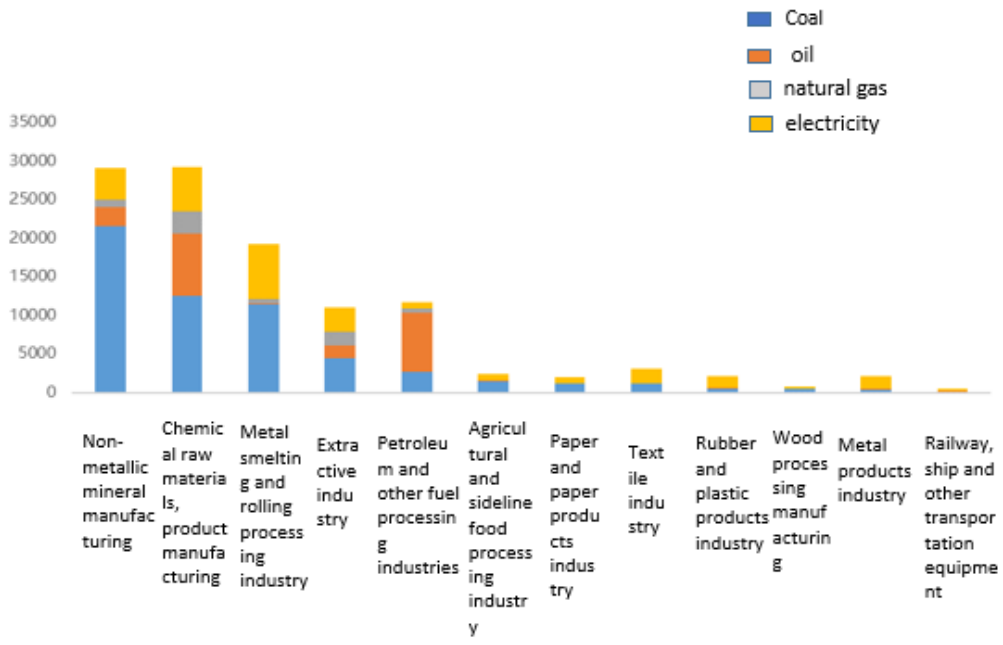

Figure. 2 Total energy consumption in various industries in 2014 (Source: China Energy Statistical Yearbook 2015)

(2) From the perspective of electricity consumption, this article takes Zhejiang Province as an example. In 2015, the number of high-voltage and non-residential users in Zhejiang Power Grid accounted for only 12.96% of all users, but electricity consumption accounted for 85.9% of the province, including leather manufacturing, metal smelting, metal finishing, chemical manufacturing and other manufacturing industries which are the main force of electricity.

According to the "Guiding Opinions on Promoting Alternative Energy" issued by the State Council, the main energy-consuming equipment that can be replaced by electric energy in various industries is concentrated in electric boilers, furnaces, and heating. According to the selection principles of the above two industries and the categories of energy-consuming equipment involved in the Opinions, 84 key energy-using industries were selected from 913 sub-sectors, and divided into 19 fields, and models were constructed for each field. As shown in Table 1.

TABLE 1. 19 AREAS AND CORRESPONDING ALTERNATIVE TECHNOLOGIES

Domain coding	Domain division	Alternative equipment
1	Fishery	Water pump
2	irrigation and drainage	Water pump
3	Agriculture, forestry and animal husbandry	Water pump
4	Construction industry	heating
5	Transportation, warehousing and postal services	Cool storage
6	Information transmission, computer services and software industry	Heating, cooking
7	Business, accommodation and catering	Heating, cooking
8	Finance, real estate, business and residential services	Heating, cooking
9	Public utilities and management organizations	Heating, cooking
10	mining industry	Kiln
11	Food, beverage and tobacco manufacturing	Baking equipment
12	Textile printing and dyeing industry	Boiler
13	Metal foundry	Boiler, kiln
14	Paper products, printing and stationery products manufacturing	Boiler, kiln
15	Chemical manufacturing	Boiler
16	Rubber and plastic products industry	Kiln, frequency furnace
17	Non-metallic mineral manufacturing	Kiln, frequency furnace
18	Wood processing and products and household products industry	Boiler, steamer
19	Manufacturing (transportation, optics, processing)	Boiler

III. MODEL INDICATOR SYSTEM CONSTRUCTION

A. Model Influence Factor Analysis.

The factors used to identify the energy consumption of enterprises can be considered from the following four aspects: basic information of enterprises, information on production and operation of enterprises, characteristics of electricity consumption of enterprises and characteristics of electricity consumption habits of enterprises. These four types of data can be separately used from the marketing system and use of the grid. Obtained in the electric acquisition system:

(1) Analysis of basic information of enterprises

Machine production equipment, product production cycle, and electricity consumption characteristics used in different sub-sectors are different, which affects the energy consumption and energy consumption rules; Energy-consuming equipment used by enterprises of different ages in the same industry has certain time-course characteristics. For example, enterprises established after 2015 are mainly affected by electricity and natural gas because of the influence of policies. As shown in Figure 3, due to factors such as industry and history, there is a certain difference in the distribution of households between target and non-target enterprises.

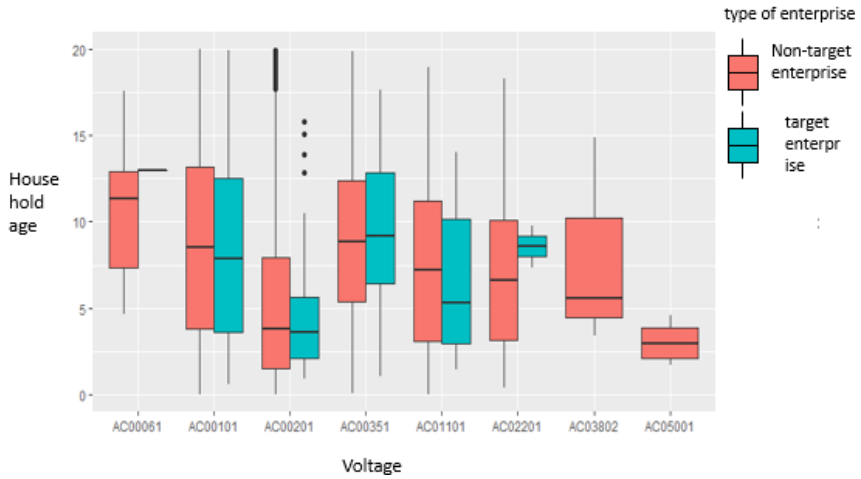

Figure. 3 Target Age vs. Non-Target Age at Different Voltages

(2) Analysis of production and operation information of enterprises

Whether the business operation has a direct impact on the company's subsequent production plan and equipment transformation willingness, enterprises with better business conditions will have a higher willingness to accept electric energy instead; Recently, the company's capacity reduction will also affect the company's alternative possibilities. Recently, there have been reductions, especially those with a higher ratio of capacity reduction to contract capacity. If the production schedule has not been put on hold, the capacity has not been reduced. It is very likely that companies will replace electricity with other energy sources in the near future.

(3) Analysis of power consumption characteristics of enterprises

The power consumption characteristics of enterprises are based on the human-powered division of power grid standards, mainly in four aspects: 1) user category; 2) power supply voltage; 3) power consumption category; 4) operating capacity.

The user categories are divided into high-voltage users, low-voltage non-resident users and low-voltage residential users. Different types of users have large differences in their electrical characteristics.

The power supply voltage is divided according to the voltage level when the user establishes the household. Since the voltage required by the high energy-consuming equipment of various enterprises is different, the indicator can play a good role in model identification. For example, the enterprises with voltage 380V in the textile industry are mostly small. Workshops or sub-processing, generally do not use boilers or setting machines, so the possibility of being replaced by electrical energy is very small.

According to the industry characteristics of the enterprise, the types of electricity can be divided into large industrial electricity, general industrial and commercial electricity, and residential electricity. Enterprises in different power categories have significant differences in power usage cycle, peak usage, and power consumption.

The operating capacity refers to the actual required capacity during the user's electricity use. Combined with the size, capacity and current capacity analysis of the enterprise, it can understand the proportion of the company's electric energy in its production process. The lower the specific gravity of the electric energy, the more it is used. The greater the possibility of other energy sources. As shown in Figure 4, the target company is more distributed in high operating capacity than non-target enterprises.

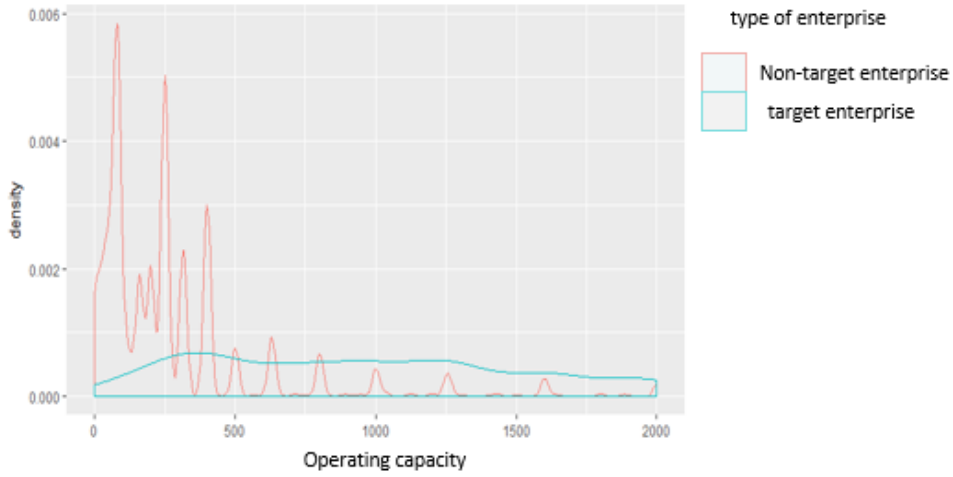

Figure. 4 Operating capacity density distribution of target and non-target enterprises

(4) Analysis of the characteristics of enterprise electricity habits

The enterprise's electricity habits are based on the electricity and load electricity consumption data collected by the grid electricity collection system. Enterprises using coal, oil, natural gas and other energy sources and enterprises using only electric energy are on the peak and valley electricity and electricity load. There are significant differences in daily, monthly and quarterly performances. As shown in Figure 5, after the metal casting industry (top left) of the hardware and electrical appliances has been replaced by electric energy, there is a peak in the electricity load during the early morning hours. In combination with the actual situation analysis of the business, the metal foundry industry, especially the casting industry, generally uses kiln equipment for metal melting, and the equipment is used from 1 to 6 in the morning. After the electric energy is replaced, the enterprise replaces the coal-fired kiln with electric kiln, causing a peak in the electrical load during this period.

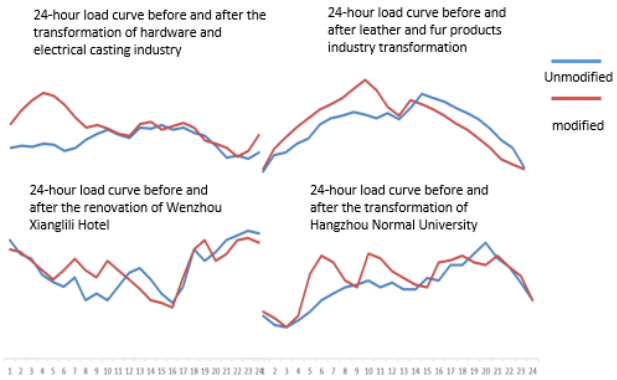

Figure. 5 24 hours load curve before and after transformation of enterprises in different types of industries

B. Model variable screening and establishment.

Through the analysis and sorting of the influencing factors of the above variables, 13 model variables were initially identified, as shown in Table 2.

TABLE 2 MODEL VARIABLES SCREENING RESULTS

Attributes	Analysis index
Basic Information	industry sector
	Household age
Operating conditions	Industry boom
	Volume reduction rate
Electrical characteristics	Supply voltage
	User classification
	Power category
	Operating capacity
Electricity habit	Peak electricity index
	Daily load index
	Quarterly load index
	Moon peak electricity index

In order to analyze the degree of influence of each variable x_i on the substitution result, the above influential factors can be used as independent variables, and the substitution result y_i is used as the dependent variable. Correlation analysis is used to investigate the correlation between the independent variable and the dependent variable.

According to the correlation analysis between the dependent variable (alternative result) and the independent variable (Table 3), the correlation coefficient between voltage type, load type, volume reduction rate, peak charge index and target variable is small, and these indicators are weakly linear with the model results. It can be discarded in the later construction of the model. According to the significance level of the input variables and the target variables, except for the load type P value of 0.77>0.05, the significance test was not passed, and other indicators passed the Pearson test.

In addition, the multicollinearity between variables can be determined by the kappa condition number. The kappa condition number is 12.776<100 by multicollinearity analysis between variables. Therefore, the model input variables have a low degree of collinearity and can be used for all. The construction of the model.

TABLE 3 VARIOUS TYPES OF INDICATORS VARIABLES AND THE TARGET VARIABLE CORRELATION COEFFICIENT AND SIGNIFICANCE LEVEL

Index	Pearson correlation coefficient	P-value:pearson	Spearman correlation coefficient	P-value:spearman
Operating capacity	0.28	0	0.00	0
Household age	-0.11	0	0.00	0
Power category	0.62	0	0.00	0
Supply voltage	0.06	0.06	0.00	0
Load type	0.01	0.77	1.98	0.9
Volume reduction rate	-0.08	0.03	0.00	0
Daily load index	0.38	0	0.00	0
Quarterly load index	0.33	0	0.00	0
Industry boom	0.23	0	0.00	0
Peak electricity index	-0.14	0	0.42	0.19
Monthly load index	0.57	0	0.00	0
Moon peak electricity index	-0.48	0	0.00	0

IV. Construction of Logistic Regression Model Based on Data Mining

A. Model technical principle description.

The specific formula of the Logistic function (Sigmoid function) is as follows:

$$g(z)=\frac{1}{1+e^{-z}} \tag{1}$$

Where e is the natural logarithm and z is the steepness of the curve,

For the case of linear classification, the boundary form is as follows:

$$\theta_0+\theta_1 x_1+\cdots+\theta_n x_n=\sum_{i=1}^{n}\theta_i x_i=\theta^T x \tag{2}$$

Where θ_0 is a constant, $\theta_1,\theta_2,\cdots,\theta_n$ are the coefficients of the variable, $x_1,x_2,\cdots,x_n$ is a specific variable. In this paper, the variable is a specific electricity consumption indicator that can influence the result of the user's electric energy substitution potential, such as voltage, load, electricity, electricity consumption, Enterprise size, etc.

The prediction function constructed by equations (1) and (2) is:

$$h_\theta(x)=g\left(\theta^T x\right)=\frac{1}{1+e^{-\theta^T x}} \tag{3}$$

Where $g\left(\theta^T x\right)=g(z)$, indicating whether the two results are replaced, is the probability of taking 1, otherwise taking the probability of 0, so the above formula can be converted into:

$$P\left(y=1\middle|x;\theta\right)=h_\theta(x) \tag{4}$$

$$P\left(y=0\middle|x;\theta\right)=1-h_\theta(x) \tag{5}$$

The combination of (4) and (5) can be written as:

$$P\left(y\middle|x;\theta\right)=\left(h_\theta(x)\right)^y\left(1-h_\theta(x)\right)^{1-y} \tag{6}$$

In the processing of variables, we should pay attention to the following two points: 1. The linearization characteristics of the logistic regression model make the direction of influence of continuous variables singular, but according to the actual business experience, some variables do not exhibit such characteristics, such as voltage at 10KV. The possibility of enterprise substitution is the highest, and the possibility of above or below 10KV is low, showing the peak-like feature of the middle and high sides. The above-mentioned problem can be avoided by artificially converting or slicing the partial variables, but this will inevitably result in The distortion of some data, so the accuracy and quantity of the slice is an important factor to ensure the effect of the model; 2. The absolute value of each variable varies greatly, and it is easy to cause a single variable to have too much influence on the function. Therefore, the continuous variable is normalized before the model variable is input.

For the logarithm of (6), the maximum likelihood estimation Cost function can be derived as:

$$Cost\left(h_\theta(x),y\right)=\begin{cases}-\log\left(h_\theta(x)\right) & \text{if } y=1\\ -\log\left(1-h_\theta(x)\right) & \text{if } y=0\end{cases} \tag{7}$$

The Cost Loss function here uses the maximum likelihood estimate, so the smaller the Cost value, the more the function converges and the better the estimation of the model. To find the minimum value of the Cost function, you can use the gradient descent method, which is available according to the gradient descent method:

$$\theta_j:=\theta_j-\alpha\frac{\partial}{\partial\theta_j}J(\theta),\quad(j=0,1,\cdots,n) \tag{8}$$

By the gradient descent method, (8) can be written as:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) x_j^{(i)}, \quad (j = 0,1,\cdots,n) \tag{9}$$

It can be seen from equation (9) that when Cost is the smallest, that is, the sum of residuals $\left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right)$ is the smallest, the whole model function and the actual result are best fitted. In this paper, the Cost Loss value is verified using the logistic regression of the R language and the comparison of the residual sums.

V. Model output and Application.

As shown in Table 4, for the electric energy substitution potential user discriminant model coefficient output results (metal manufacturing), in addition to the industry climate index, other variables have a significant impact on the model results (P <= 0.05), of which the electricity type is large industry When electricity and voltage are 10KV, the possibility of enterprise electric energy substitution is greater, while other variables, especially the higher the load index and peak electric energy index, the less likely the enterprise to replace energy, because the higher the load index, the enterprise The greater the fluctuation of the electrical load, the greater the proportion of energy consumed in its main production equipment, and the less likely it is to replace the electrical energy.

Table 4 Model results output - metal manufacturing (R Language Analysis)

Index	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.33E+00	1.67E+00	-1.998	0.045675*
Operating capacity	-1.04E-06	1.38E-04	-7.534	4.94e-14***
Household age	-1.52E-02	1.18E-02	-1.29	0.037219*
User Type - Big Industry	5.47E+00	7.70E-01	7.113	1.13e-12***
User Type - General Business	1.48E+00	9.00E-02	8.325	5.3901e-12*
Daily load indicator	-4.63E+00	2.44E+00	-1.901	0.047333*
Quarterly load index	-2.39E+00	5.62E-01	-4.254	2.10e-05***
Industry boom	1.64E-03	4.42E-03	0.371	0.710298.
Monthly load index	-1.14E-03	9.34E-04	-1.225	0.020706*
Moon peak electricity index	-4.54E+00	7.20E-01	-6.308	2.83e-10***
Voltage -10kv	7.77E-02	2.73E-02.	2.847	0.004411**
Voltage -380v	1.49E-02	5.13E-03	-1.412	3.07e-04***
……	……	……	……	……
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

The model reserves 30% of the data as a verification set, which is used to verify the applicability of the model. It can be seen from Figure 6 that the overall coverage and hit rate of the model are better, and the hit rate has a small fluctuation in the latter half. It is found that this is due to the fact that some of the industry's target variables are too small, resulting in the model's hit rate being over-fitting in the first half, which has little effect on the actual application of the model.

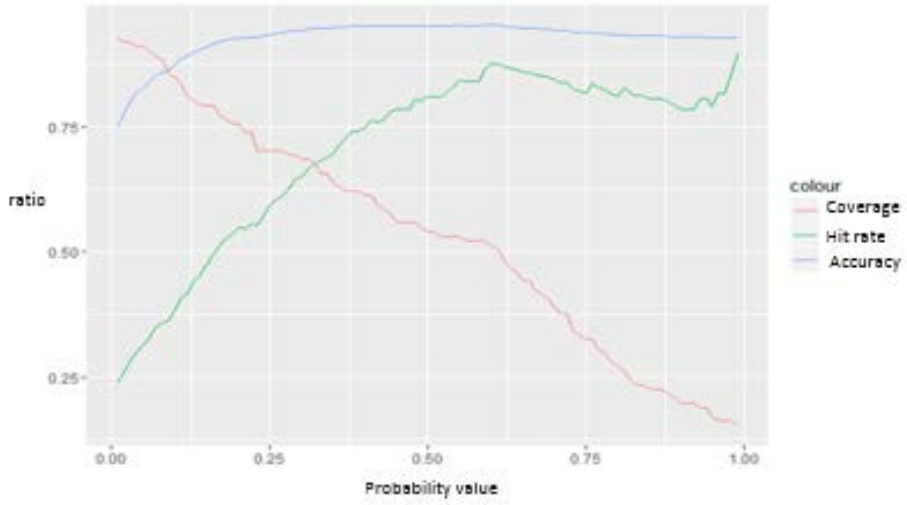

Figure. 6 Model validation set test results (R language analysis)

According to the output of the model, the query optimization technology in the big data storage technology is used to solidify the model into the grid business system. In order to achieve the purpose of fast label combination calculation, the system compares the relationship between the customer and the model result data in a Boolean manner. Store to ensure the efficiency of business applications and model updates.

VI. Application and effectiveness of electric energy substitution model

After the electric energy substitution model was put into use in October 2016, the potential analysis of 4.5 million enterprises other than resident users in the province has been completed, and two on-site field visits have been completed in

Shaoxing and Jiaxing, involving a total of 4,612 enterprises. The average hit rate reached 23.2% and 30.7%, respectively, which was 5-6 times higher than the visit efficiency when the model was not built.

From the perspective of industry, the results of two visits show that there are more potential enterprises in the textile, metal, rubber and other manufacturing industries, and there are almost no potential enterprises in the fields of transportation, information transmission and commerce (Figure 7).The interview results also validated the second part of the industry energy analysis.

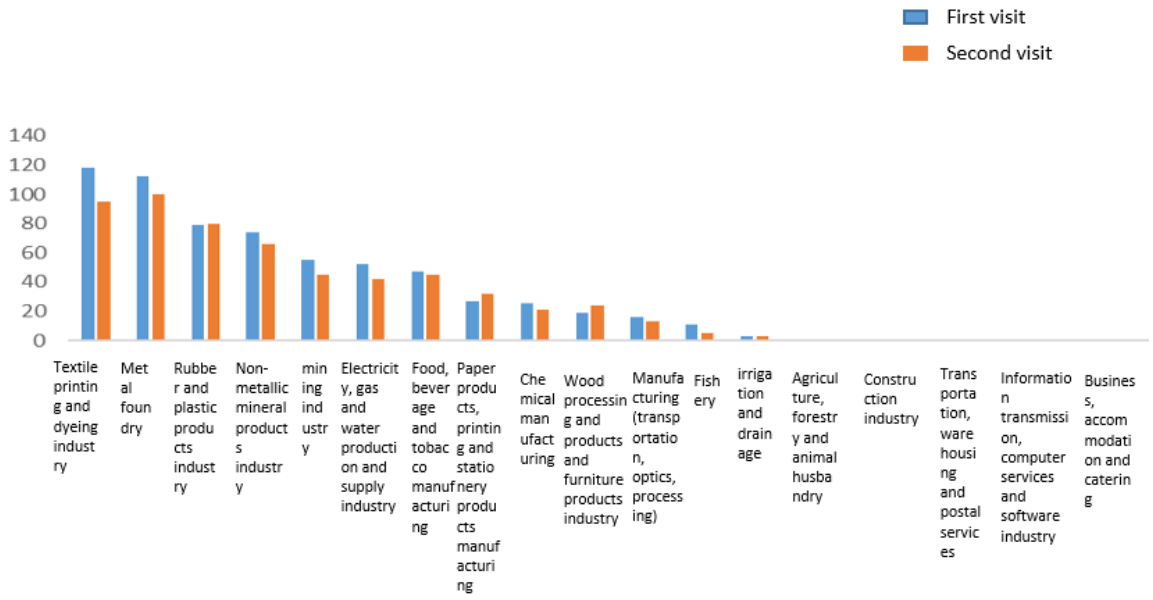

Figure. 7 Distribution of actual potential enterprises in various industries (two visits)

VII. Conclusion

The research shows that the user identification model of electric energy substitution potential can quickly respond to business needs, help business personnel to screen out electric energy to replace potential customers from massive data and formulate corresponding promotion strategies, and carry out daily electric energy substitution work in a targeted manner. The model has achieved good results through pilot trials in Shaoxing and Jiaxing. It provides a scientific tool for the precise positioning of users by business personnel and the advancement and differentiation of strategies and services in the later stage.

Acknowledgements

This work is supported by the State Grid Science and Technology Project (Project name: "Research on Key Technology of Electric Energy Substitution Effect Improvement and Its System Development Application Based on Electric Energy Service Management Platform").

References

[1] Y.H. Zhao. Research on Energy Substitution Analysis and Model Optimization under Beijing Environmental Energy Saving and Emission Reduction Targets[D].North China Electric Power University, 2014, p.1-171.

[2] J.C. Zheng. Rural Electric Energy Substitution Potential and Environmental Benefit Analysis[D]. North China Electric Power University, 2015, p.1-56.

[3] L.L. Zhang. Zhejiang Province Energy Structure Optimization Research[D]. Zhejiang University of Technology, 2013, p.1-65.

[4] Maupi, Eric, Letsoalo, et al. Determination of Causal Effect in Observational Studies: Analysis of Correlated Data with Binary End-Points[J]. Journal of Systems Science and Complexity, 2012(2):119-125.

[5] Z.F. Li: Research and Practice of Jiangsu Energy Substitution[J]. Power Demand Side Management, 2016, 16(5), p.1-3.

[6] Department of Energy Statistics, National Bureau of Statistics. China Energy Statistics Yearbook 2015[M]. China Statistics Press, 2016.

[7] Z.Q. Qu, J.J. Xin, L. Wu and S.Y. Zheng. Study on the boundary conditions of commercial users' alternative technology selection[C]. China Electrical Engineering Society Power System Automation Committee 3rd Meeting and Academic Exchange Conference, 2015, p.1-17.

[8] Y.X. Zhou. Youxue. Promotion and Implementation of Electric Energy Substitution Technology[M]. Gansu Electrical Engineering Society Academic Annual Meeting, 2014, p.1-5.

[9] Y. Zhang. Opening Energy Consumption Mode——Electric Energy Substitution: Taking the Road of Clean, Environmental Protection and Sustainable Development[J]. State Grid, 2013(10), p. 34-40.

[10] Y.S. Xue and Y.N. Lai. The Integration of Big Energy Thinking and Big Data Thinking (I) Big Data and Power Big Data[J]. Automation of Electric Power Systems, 2016, 40(1), p. 1-8.

[11] J.H. Duan, N.D. Zhang, B. Zhao and X.B. Yan. Research on the Architecture and Application of Power Big Data Infrastructure[J]. Electric Power Information and Communication Technology, 2015, 13(2), p. 92-95.

[12] X.C. Heng and L. Zhou. Application of Distributed Technology in High-performance Processing of Power Big Data[J]. Electric Power Information and Communication Technology, 2013, 11(9), p. 40-43.

[13] Y.X. Zhou and Q. Wang. State Grid Gansu Company: Promoting Electric Energy Replacement Based on Local Conditions[J]. China Power Enterprise Management, 2015(2), p. 22-29.