Analysis of PV Poverty Alleviation Users Based on Big Data

Jian Li[1, a]
State Grid Xinjiang Electric Power Corporation, Wulumuqi, 830000
406839653@qq.com

Abstract. Based on the historical data of power customers, the index system needed to determine the model is determined based on the customer's basic attributes, power usage behavior, payment behavior, customer credit, geographic region, and weather environment. Through the correlation coefficient matrix and information value (IV) of the indicator, the index variables that finally enter the model are selected, the variables are grouped by the optimal grouping method and the weight of Evidence (WOE) is transformed. Based on the processed data, the logistic regression algorithm is used to construct the PV poverty alleviation user analysis model, and the users are classified into high, medium and low grade suspected poor users according to the analysis model, and the analysis results are fed back to the government's poverty alleviation office for confirmation feedback. This paper provides a basis for selecting poor households in the Xinjiang Poverty Alleviation Competition.

Keywords: Photovoltaic poverty alleviation, logistic regression algorithm, user label, index system.

1. Introduction

Eliminating poverty, improving people's livelihood and gradually achieving common prosperity are the essential requirements of socialism and an important mission of our party. During the "Thirteenth Five-Year Plan" period, it was the stage of decisive victory in building a well-off society in an all-round way and achieving the goal of the first hundred years of struggle. It was also the stage of winning the battle to fight poverty. According to the "China Rural Poverty Alleviation and Development Program (2011-2020)",the State Council's "Opinions on Solidly Promoting Rural Poverty Alleviation and Development Work on Innovation Mechanisms", "The Decision of the Central Committee of the Communist Party of China on Winning Poverty Alleviation" and the State Council's Implementation Plan for Precision Poverty Alleviation "The Outline of the Thirteenth Five-Year Plan for National Economic and Social Development of the People's Republic of China" mainly clarifies the overall thinking, basic objectives, major tasks and major measures of the country's poverty alleviation during the 13th Five-Year Plan period. The action guide for work is an important basis for all relevant parties to formulate relevant poverty alleviation plans.

Over the years, by inspiring the comprehensive effects of special poverty alleviation, industry poverty alleviation, social poverty alleviation and poverty alleviation in poverty alleviation, Xinjiang has built a new pattern of “four in one” poverty alleviation with multiple measures and mutual support. Photovoltaic poverty alleviation pilot project to help poor households install distributed photovoltaic power generation system to increase the basic living income of the poor. Xinjiang encourages enterprises to fulfill their social responsibilities, actively participate in photovoltaic poverty alleviation operations through donations or donation of equipment, work together to solve distributed photovoltaic power generation equipment and funding problems. According to the overall goal of poverty alleviation in Xinjiang, by 2020, Xinjiang must ensure that the 2.61 million poor people under the current standards are all out of poverty, all 35 key poverty-stricken counties have been removed, 3029 poverty-stricken villages have all withdrawn, solving the regional overall poverty in the four southern states.

With the development of information technology, China began to use data mining technology to analyze the characteristics of massive users and try to find out the characteristics of users' electricity consumption[1-3].Based on the statistical analysis of the factors of poor users, the use of logistic regression model for the use of poor users[4-6].Modeling and analyzing the key factors of electric

power and its influence degree, deeply analyzing the characteristics of user electricity consumption, user payment, user's family resident population, low-income households, five-guarantee households and designing the key influence variables related to poverty user level identification[7-10].The decision tree algorithm was used to establish the poverty user level recognition model. Based on the theoretical analysis, the improved LR-Bagging algorithm was used to predict the poor user level[11-12].However, multiple algorithm models were not used for presentation analysis to introduce the combined modeling method, and the performance of multiple algorithms was comprehensively compared to get the optimal model[13-14].The model has a good explanation for the identification of poor users, and can be actively applied to the management of PV poverty alleviation users, effectively guiding the healthy operation of the precise poverty alleviation business.

In fact, the formation of poor user identification is more complicated. This paper analyzes the historical data of users from the basic data, electricity bill information and historical payment records of low-voltage resident customers, deeply explores the key factors affecting customer electricity usage and builds the influence of poor users. The factor indicator system is based on logistic regression algorithm to construct a quantitative analysis model of electricity tariffs, predicting the level of poor users, and providing a theoretical basis for formulating precise poverty alleviation.

2. Main Method

2.1 Optimal Variable Grouping Method

Variable grouping is the process of merging certain categories of categorical variables to reduce their cardinality, or segmenting a numeric variable into a categorical variable. The method is based on the splitting of the decision tree model to find the optimal grouping scheme, and the predictive power index is maximized by combining the categories of variables. That is, the optimal binary segmentation point is first found by the principle of maximizing a certain predictive power index, and then the previous step is repeated in each subcategory, and the segmentation is stopped when the maximum number of packets is reached.

2.2 WOE (Weight of Evidence) Evidence Weight Conversion Method

The categorical variables are transformed into numerical variables to reduce the complexity of the modeling process, and the logistic regression model can be transformed into a standard scorecard format to facilitate the interpretation and application of subsequent model results. For the i-th group of a classification independent variable, the calculation formula of WOE is as follows (1):

$$WOE(x_i) = In\left(\frac{p_i}{q_i}\right) = In\left(\frac{n_{i1}/n_{*1}}{n_{i2}/n_{*2}}\right) \tag{1}$$

The information value represents the ratio of the responding customer to the unresponsive customer in the current group and the difference in this ratio among all samples. The larger the WOE, the greater the difference, the greater the likelihood that the sample in this group will respond.

2.3 Logistic Regression Method

Logistic regression is a further development based on a linear model. Its general form is: $p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$, $P = \frac{1}{1+e^{-p}}$. Logistic regression model has fast calculation speed, obvious results and good fitting effect. It is widely used in big data, machine learning, economics and other fields.

2.4 Score card function method

A user's rating can be expressed as (2):

$$\begin{aligned} score = A + B\beta_0 + \left(B\beta_1\omega_{11}\right)\delta_{11} + \left(B\beta_1\omega_{12}\right)\delta_{12} + ... \\ + \left(B\beta_r\omega_{p1}\right)\delta_{r1} + \left(B\beta_r\omega_{p2}\right)\delta_{r2} + ... \end{aligned} \tag{2}$$

among them $B = PDO / In(2)$, $A = BaseScore - B * In(odds)$, β is the model parameter. ω is the WOE conversion value, δ is a binary variable (1 or 0) indicating whether the variable takes a certain value. r is the number of model variables. The user score is controlled between 0-100 points. The algorithm based on the standard score card converts the logistic regression model result into a score card form, and the user's final score is the sum of the corresponding scores of the variables.

3. Model Construction and Application

3.1 Target Customers

Combine business understanding and current situation analysis to define the definition of poor users in the power business. Different definitions correspond to different business performances and can lead to different business response strategies. The target users selected in this paper are low-voltage resident customers whose monthly electricity consumption is below 50 degrees, and households with 4 or more households living in the household.

3.2 Construction Ideas

First, for the more relevant indicators, the reserved part can be used; at the same time, the derivati ve variables are created to prepare the data for modeling. Second, do a preliminary exploratory analy sis of the data. Through the descriptive statistical analysis of power user characteristics analysis, pow er trend analysis, user power behavior analysis, etc., preliminary construction ideas are formed to pre pare for the selection of indicators. And then select relevant indicators and establish an indicator syst em. Corresponding to the entry requirements of the logistic regression model, cluster analysis, correl ation analysis and principal component analysis are carried out. The purpose is to reduce the weight of each model variable and obtain the weight of each index, transform the data through internal class ification of variables and WOE weight conversion which meet the data requirements of model model ing. Finally, the model is trained and tested, evaluated by model evaluation, and repeated training to obtain the best model.

3.3 Data Preparation

Based on the data of a county in March 2017-Aug.2017 as the basic data, whether the minimum amount of electricity was generated as the target variable in Sep. and Oct.2017, model training was carried out. The basic data mainly includes:

Basic attributes: user number, power supply unit, meter reading segment number, whether five-guarantee users, whether low-income users, etc.;

Electricity consumption data: user classification, industry classification, power supply voltage, contract capacity, load level, etc.;

Electricity consumption: electricity consumption, electricity bill, amount of liquidated damages, number of arrears, number of billing, etc.;

Payment behavior: payment method, number of payment changes, etc.;

External data: economic data of each county, weather data, information system of poverty alleviation office information system.

3.4 Exploratory analysis

.1 Analysis of comprehensive characteristics of electricity customers

Based on the decision tree algorithm, the classification model of resident users is constructed, and the comprehensive characteristics of suspected poor users are mastered. An example of classification results is shown in Table 1. Users with feature rule one or rule two are suspected to be poor users.

Table 1 Example of suspected poor user rules

	Variable name	**Feature rule**		**Variable name**	**Feature rule**
Rule one	Cash contribution ratio	>0.4 and <0.71	Rule two	Cash contribution ratio	>0.96
	Number of payment channels	>1		Payment times	>2 and <=4
	6th - 10th payment times	>0		Number of payments before the 5th	>0 and <=2
	Billing times in six months	>2		6th - 20th payment times	<=2
	The average billing power for three	<3		21th - 25th payment times	<=2

	months is the reciprocal ranking of the users of the meter reading segment.				
	Resident population	>4		Historical 6-month electricity bill average	<=10
				Power supply unit number	*******06
				Payment channel	cash

.2 Analysis of customer behavior characteristics of electricity users

Based on the suspected poor user classification model we have constructed, according to the IV value, the output importance indicator variable is shown in Fig. 1.

At the same time we can get:

In the importance of the variable, according to the user segment of the meter reading segment whose power is lower than the reference variable, the resident resident population is greater than 4, and analyzes the range of suspected poor users in the meter reading segment

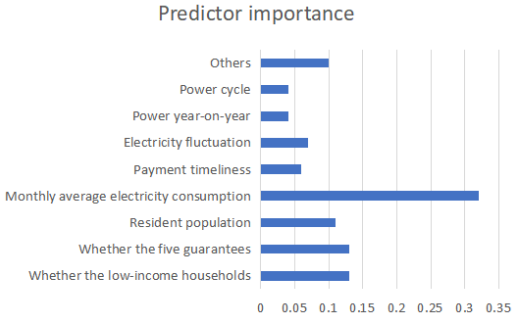

Fig. 1 Importance indicator variable

Table 2 Electricity consumption which less than the baseline electricity distribution

Power supply station	**Copy table name**	**Number of meter reading users**	**Monthly electricity consumption is less than 50 degrees**	**Resident populati on>4**	**Minimum number of users**	**Five guarantee users**
A County Power Supply Office	Low pressure A village 001 meter reading section	227	5	17	0	1
A County Power Supply Office	Low pressure A village 002 meter reading section	300	7	14	0	1
A County Power Supply Office	Low pressure A village 003 meter reading section	210	4	17	0	0
A County Power Supply Office	Low pressure B village 001 meter reading section	320	4	22	0	0

A County Power Supply Office	Low pressure B village 002 meter reading section	310	5	20	0	0
A County Power Supply Office	Low pressure B village 003 meter reading section	272	6	18	0	0

It can be seen from the above table: The monthly electricity consumption of users below 2% is within the reference power. Among such users, according to the density of the resident population, some users with less population are removed.

The monthly user electricity consumption ranking changes have an important relationship with the user's poverty level. Further analyze the characteristics of this variable on the degree of poverty, and at the same time give the suspected that the poor users accounted for the top five users of electricity consumption changes.

Table 3 Monthly user electricity consumption ranking ratio

user name	Electricity fee	energy used	Electricity ranking (ascending order)	Ranking change	Family population
Zhang Yifei	201708	22	1	no	6
Zhang Quangui	201708	24	2	no	5
Zhang Baofeng	201708	30	3	+2	5
Li Quan	201708	31	4	+2	6
Li Fei	201708	32	5	+3	5

3.5 Index System Construction

According to the results of exploratory analysis and data characteristics, a quantitative index model system for poor users is constructed. It is constructed from three dimensions: user basic information, electricity usage behavior and payment behavior. See the picture below for details:

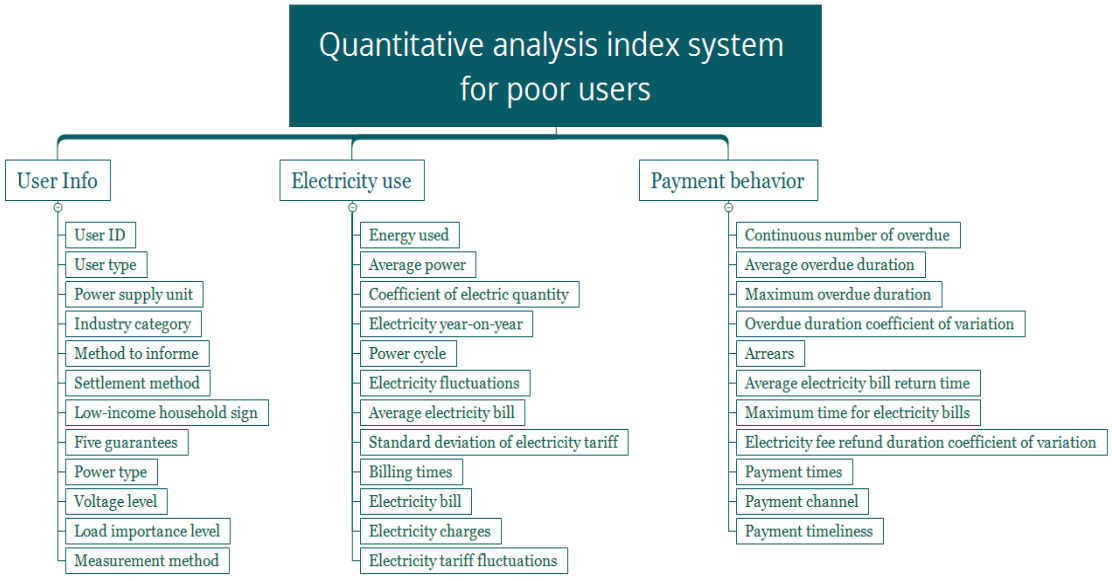

Fig. 2 Quantitative analysis index system for poor users

3.6 Model Construction

Based on R software, logistic regression algorithm is used to realize the predictive analysis model of PV poverty alleviation users.

For the classification index, the variable with high predictive power is selected by the IV value to enter the model; for the continuity index, the correlation coefficient matrix between the variables is

calculated, and the index with the correlation coefficient greater than 0.5 is combined with the IV value for screening. After screening, there are 21 indicators that finally enter the model, as shown in Table 4.

Table 4 Model index

Model index	Indicator name	Model index	Indicator name
Whether the five guarantees	IS_FIVE_INSURED	Billing times in the past 6 months	RELEASE_NUM_6M
Whether the low-income households	IS_LOW_ INSURED	Fluctuation of payment duration	BD_PAY_RATIO
Average battery power in the past 6 months	PQ_HB_6M	Payment times	PAY_NUM
The average electricity consumption in the past 6 months is ranked in this meter reading section.	PQ_TB_6M	Cash contribution ratio	CASH_P
Nearly 6 months payment and timeliness	PAY_IN_TIME_6M	Average electricity bill	T_AMT_AVG
Number of payments after the 25th	PAY_NUM_25AFTER	Power variance	T_PQ_S
21 to 25 payment times	PAY_NUM_25_21	Current electricity consumption in the current period	PQ_TB_LAG
Number of contributions from 11 to 15	PAY_NUM_15_11	Payment channel changes	CHANGE_PAY_NUM
6 to 10 payment times	PAY_NUM_10_6	Payment channel	PAY_MODE
Number of payments before the 6th	PAY_NUM_5_BFORE	Power supply unit	ORG_NO
Payment variance	PAY_TIME_S		

The optimal clustering process and WOE evidence weight conversion were carried out for each index, and a logistic regression model was constructed. The results were quantified and scored. It was verified that the model passed the goodness of fit test and the coefficient of each variable was significant at the 0.05 level. Table 5 shows some of the more influential index coefficients, with an intercept term of -4.13.

Table 5 indicator coefficients

Model index	coefficient	Model index	coefficient
Constant term	-4.1329079	Number of contributions from 11 to 15	-0.3626212
Power supply unit	0.9268939	Number of payments before the 6th	-0.3664164
Average value of payment for the past 6 months	-0.5510513	Average electricity bill	-0.2903592
Power variance	-0.5277196	Current electricity consumption in the current period	-0.279614
Payment variance	-0.329336	Average power consumption in the past 6 months	-0.2885892
Payment times	-0.362532	Payment channel	0.1844638
Number of payments after the 25th	-0.2324154		

3.7 Model Evaluation

1) Model Results

Based on the results of the predictive analysis model of PV poverty alleviation users, the standard scorecards for predictive analysis of poor users are constructed based on the scorecard function. The scores of some important indicators are shown in Table 6.

Table 6 Indicator score example

Variable name	Variable grouping	Score	Variable name	Variable grouping	Score
Average monthly electricity consumption in the past 6 months	(-1,0]	1.71	Power variance	(-1,0]	2.86
	(0,3.33]	1.47		(0,5.33]	2.34
	(3.33,6.67]	2.02		(5.33,32.33]	2.43
	(6.67,10.33]	2.33		(32.33,50]	2.20
	(10.33,13.67]	2.59		(50,544.5]	2.30
	(13.67,17.33]	2.93		(544.5,722]	2.44
	(17.33,22]	3.20		(722,2094.33]	2.34
	(22,57]	3.84		(2094.33,Inf]	2.45
Power supply unit	S County A	2.40	Number of payments after the 6th	(-1,0]	2.44
	S County B	2.46		(0,1]	2.39
	S County C Office	1.96		(1,2]	2.36
	S County D Office	2.50		(2,3]	2.17
	S County E Office	2.62		(3,Inf]	1.82

2) Model evaluation

The model effect evaluation map is drawn according to the model hit rate, coverage rate and lift rate. As shown in Fig. 3, as the number of samples increases, the hit rate and the degree of lift gradually increase, and the coverage rate gradually decreases.

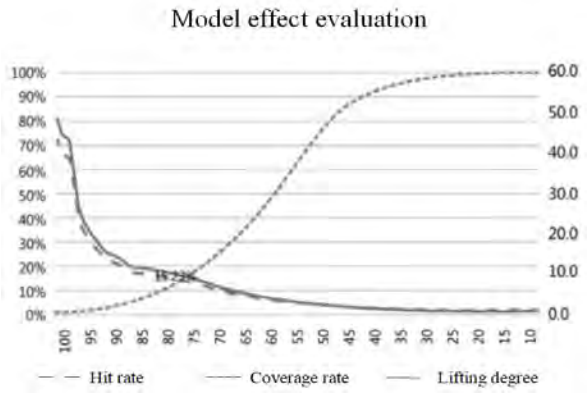

Fig. 3 model renderings

In the figure, the position of the intersection of the coverage rate and the hit rate is the profit balance point. That is, at the intersection, in the case of considering the marketing cost, the balance of payments can be achieved; on the left side of the intersection, the hit rate is higher than the coverage rate. In the case of considering marketing costs, investing less money can get better returns; on the right side of the intersection, the hit rate is lower than the coverage rate, and in the case of considering the marketing cost, you need to invest more cost to get better income.

Therefore, in consideration of marketing costs, it is recommended to define customers with more than 90 points as high-level users. At this time, the hit rate is 22.8%, the coverage rate is 4.1%, and the promotion rate is 15.4. Regardless of marketing costs, it is recommended to define customers with more than 70 points as high-level users. At this time, the hit rate is 11.4%, the coverage rate is 26.3%, and the promotion rate is 7.7.

4. Conclusion

Using big data for data mining and modeling analysis can predict future business in a timely manner, so as to effectively manage risks and take targeted measures. In this paper, the application of the logistic regression model is improved, and the indicators entering the model are refined. It has

strong applicability and scalability. It can design labels and application scenarios for different job functions such as meter readers and managers. For example, the meter reader can use the poverty-stricken household grade label to screen the customer group in the meter reading stage, and check the suspected households of the poor households; the management can provide the power marketing information record to the photovoltaic poverty alleviation organization for the high-level suspected poor users. Through the feedback from high-risk users by poverty alleviation institutions, the dynamic assessment of poor households will be realized. Through practical application, it demonstrates the model's ability to integrate data associations. It can be widely used in the design and development of power poverty alleviation households analysis programs to improve the precise positioning of poverty alleviation users and promote accurate poverty alleviation work.

References

[1]. ROBU V, VINYALS M, ROGERS A, et al. Efficient buyer groups for prediction-of-use electricity tariffs[C]. AAAI Conference on Artificial Intelligence, 2014: 451-457.

[2]. PEPERMANS G,WILLEMS B. Cost recovery in congested electricity networks[J]. Zeitschrift Für Energiewirtschaft, 2010, 34(3):195-208.

[3]. Zhang Xiaofeng. Construction of Electric Power Recycling Risk Prevention System for Large Power Customers[J]. Inner Mongolia Science and Technology and Economy, 2013(24): 121-123.

[4]. Huang Y, Yan Y. Research of evaluating credit-risk in power enterprise based on SVM and VIKOR method[C]. IEEE International Conference on Industrial Engineering and Engineering Management. IEEE, 2009:1596-1599.

[5]. Liu Jin. Risk Assessment of Power Customers' Arrears Based on Improved Analytic Hierarchy Process[J]. Electronic World, 2014(19): 182-183.

[6]. WIGAN M R, CLARKE R. Big data’s big unintended consequences[C]. 2013 IEEE 6th International Conference on Computer Society, 2013: 47-53.

[7]. RAHMAN M N, ESMAILPOUR A, Zhao J. Machine Learning with Big Data An Efficient Electricity Generation Forecasting System[J]. Big Data Research, 2016, 5:9-15.

[8]. Yang Huafei, Li Donghua, Cheng Ming. Analysis and Research on Key Technologies and Construction Ideas of Power Big Data[J]. Electric Power Information and Communication Technology, 2015, 13(1): 7-10.

[9]. Zhao Yongliang, Qin Wei, Wu Shangyuan, et al. Risk Prediction of High-Voltage User Electricity Charge Recovery Based on Data Mining[J]. Electric Power Information and Communication Technology, 2015, 13(9): 57-61.

[10]. HUANG Wensi, HAO Yuyong, LI Jinhu, et al. Risk Decision of Power Customers' Arrears Based on Decision Tree Algorithm[J]. Electric Power Information and Communication Technology, 2016(1):19-22.

[11]. Wu Wei, Zhu Zhou. Resident Customer Prediction of Power Arrearage Risk Based on Feature Selection and Improved LR-Bagging Algorithm[J]. Electronic Products World, 2017(4): 70-75.

[12]. Zhang Lu, Pan Mingyu, Tian Heping, et al. Early warning research on the risk of arrears of power customers based on data mining technology [C]. Proceedings of 2017 Smart Grid Development Symposium, 2017, 11: 583-589.

[13]. GRANELL R, AXON C J, WALLOM D C H. Predicting winning and losing businesses when changing electricity tariffs[J]. Applied Energy, 2014, 133(10):298-307.

[14]. KAMILARIS A, KALLURI B, KONDEPUDI S, et al. A literature survey on measuring energy usage for miscellaneous electric loads in offices and commercial buildings[J]. Renewable & Sustainable Energy Reviews, 2014, 34:536-550.

[15]. GRANELL R, AXON C J, WALLOM D C H, et al. Power-use profile analysis of non-domestic consumers for electricity tariff switching[J]. Energy Efficiency, 2016, 9(3):825-841.

[16]. FELDMAN R, SANGER J. Text mining handbook: advanced approaches in analyzing unstructured data[M]. Cambridge University Press, 2006.