

Research on Credit Risk Assessment Model of Enterprises' Unpaid Electricity Charge Based on Machine Learning Method

Jiankun Hu Haiji Wang Xueying Lu

(China Energy Construction Group Guangdong Electric Power Design and Research Institute Co., Ltd.)

Abstract:Based on machine learning methods and employing credit big data, this paper explores the idea of constructing a credit risk assessment model for enterprises' unpaid electricity charges, proposes a system of indicators affecting the risk of unpaid electricity charges, and elaborates on the path of screening key indicators. In this regard, a method for calculating the credit risk score of enterprises' unpaid electricity charges is given, and suggestions are provided for risk response strategies.

1. Introduction

When controlling the power market, it is necessary for them to fully consider the role of market economic benefits in value. With the expansion of China's electricity market, the problem of illegal use of electricity has become one of the factors that cannot be ignored to affect the stability of the electricity trading order. To create a social environment of honest electricity consumption, more and more areas have brought the information of electricity customers stealing electricity, defaulting on electricity consumption and paying electricity charges into the credit information system of the People's Bank of China, which is displayed in personal or enterprise credit reports, thus developing the power credit information business [1]. In particular, power supply companies are now extensively adopting the credit payment mode of "using electricity prior to charging", which further promotes the development of the power credit business. Power supply companies are in dire straits for effective credit products to identify the credit level of power users and avoid power marketing risks.

2. Research on evaluation method

2.1 Connotation of machine learning method

Machine learning, as a new technology that gradually developed and matured with the rise of the Internet and big data, has nearly detached from the empirical distribution assumptions of classical statistics. Under the concept of big data, the sample is the total, so the model is more dependent on data. In recent years, the country,

provinces and cities have accelerated the pace of collecting and sharing credit information of enterprises involved, including information on electricity consumption in violation of regulations and breaches, which makes it possible to build a credit big data analysis model based on machine learning methods. The model is mainly displayed in the form of an enterprise credit portrait, which judges the credit risk of unpaid electricity charges from data details to scores of various dimensions to the final comprehensive scores.

2.2 Construction steps of evaluation model

The construction steps of credit risk assessment model of unpaid electricity charges are as follows: 1) Configure the basic information of the model; 2) Configure the source of model data; 3) Preliminarily construct the index system and preprocess the index data according to the characteristics of power users' unpaid electricity charges; 4) Select feature engineering method, set reasonable feature selection threshold, and mine more valuable information; 5) Divide the data set into a training set and a verification set, wherein the training set is used for training the model and the verification set is used for verifying the model; 6) Choose model algorithms for training according to actual needs, including logistic regression algorithm, decision tree algorithm, clustering algorithm, etc. 7) Adjust and optimize the model according to the training situation of the model, and get the evaluation results of the model. 8) Configure and deploy the model online, and verify the effect and monitor the operation after it goes online. The details are shown (see Figure 1):

Jiankun Hu*: hujiankun@gedi.com.cn
Haiji Wang: wanghaiji@gedi.com.cn
Xueying Lu: luxueying@gedi.com.cn

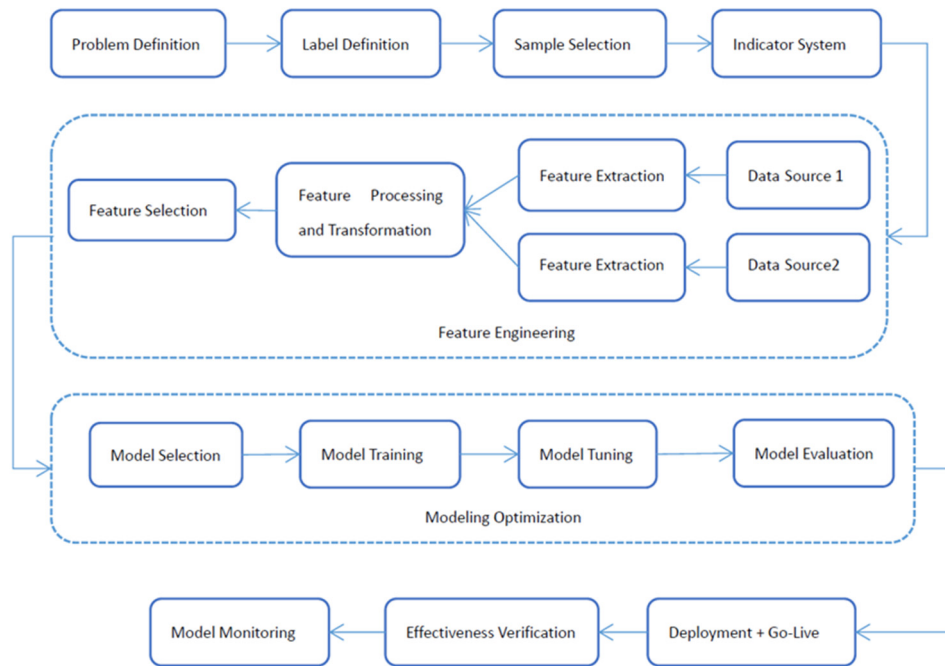


Figure 1. Evaluation model construction process

3. Construction of model indicators

3.1 Label definition

Before training a credit risk assessment model, it is bound to get the sample label of each sample, that is, the so-called y value. The definition of sample label includes the definition of positive sample and negative sample.

Positive sample: Power customers whose electricity charges are more than n_1 RMB or more than n_2 days overdue in the past year.

Negative sample: Power customers whose unpaid electricity charges are less than or equal to m_1 and overdue for less than or equal to m_2 days in the past year.

The determination of n and m can be determined utilizing data analysis. For example, through the statistical analysis of historical data, it is highly probable that ordinary customers will not continue to pay fees after how long they are overdue [2].

3.2 Sample selection

The selection of samples should follow the principles of representativeness, adequacy, timeliness and exclusion. 1) Representativeness: Samples should fully represent the whole population. 2) Adequacy: Different models have certain requirements on the amount of sample data, for which scorecard models generally require no less than 1,500 positive and negative samples and no more than 50,000 overall samples. 3) Exclusiveness: It is necessary to exclude samples with unclear classification and select samples with obvious classification to enter the model training.

The data quantity of model samples, the ratio of positive and negative samples and the data quality are

directly related to the training effect of the algorithm. When selecting sample data, the balance of sample numbers should be achieved as much as possible. Therefore, it is necessary to adopt corresponding sampling methods according to the actual sample proportion, such as random sampling, equal ratio sampling, stratified sampling and so on.

3.3 Preliminary screening of indicators

The construction of the model indicator system should follow the principles of orientation, applicability and objectivity. Drawing lessons from the national standard *Enterprise Credit Evaluation* (GB/T 23794-2015), this paper divides the indicators to measure the risk of a power customer's unpaid electricity charges into two aspects: payment ability and payment willingness. The source data include enterprise basic information, historical electricity consumption information, public credit information, market credit information and so on.

Payment willingness: The judgment of payment willingness can be divided into three modules: "historical performance", "default cost" and "payment mode". Customers with good historical credit records do not mean that they will not default in the future, but customers with poor historical credit records must have a relatively high probability of subsequent default; default cost determines which choice is more likely when the customer is faced with operational difficulties and the choice of "paying" or "not paying" when the credit expires; payment modes include power marketing strategy, pricing strategy and payment methods of power supply companies, etc. Different payment modes may affect the payment willingness of power customers. Among them, for credit granted to small and micro enterprises, payment willingness is strongly related to the credit situation of the business owner himself. Therefore, it is essential to take

the historical electricity consumption records and public credit records of enterprises as the key review dimensions.

Payment ability: Payment ability is strongly related to the profit and debt repayment of the enterprise itself. At present, the data sources used to judge payment ability are tax data, social security data and so on. Financial data is a strong correlation variable of debt repayment, and it is also a variable widely used in the scorecard model. However, in the absence of financial data, it is vital to use historical static weak variables such as operating conditions for risk assessment. Studies have shown that the internal and external conditions of the daily operation of enterprises [3] do affect the payment ability of enterprises for electricity charges, such as the development environment of the industry, the situation of epidemic prevention and control in the region, etc., and the electricity consumption status may also reflect the stability of the operation of enterprises. Therefore, the relevant indicators of the operation status and electricity consumption status of enterprises can be considered in the modeling process.

The initial indicator system of the model proposed in this paper is shown (see Table 1):

Table 1. Indicator system of credit risk assessment model for unpaid electricity bills

Primary indicator	Secondary indicator	Tertiary indicator
Payment willingness	Historical performance	Continuous normal electricity charge payment record
		Historical credit records
		Administrative penalty record
		Administrative compulsory record
		List of business anomalies
		Supervision and inspection results
		Credit evaluation results
		Judicial record
		Red and black list records
	Records of unpaid expenses of other public utilities	
	Default cost	Registered capital
		Intangible assets
		Credit status of enterprise shareholders and executives
Payment mode	Corresponding power supply company	
	Electricity price	
	Payment method	
Payment ability	Financial position	Solvency: Current ratio, quick ratio and asset-liability ratio
		Profitability: Gross profit margin, net interest rate, return on net assets and return on assets
		Operating ability: Average turnover days of accounts receivable, inventory turnover days, current assets turnover rate and total assets turnover rate

Operation status	Operating level: Total operating income, total tax payment, total social security payment, total liabilities and total assets
	Business environment: Industry category and region
	Operating status, such as in operation, out of business, revocation, etc.
	Operating years
Electricity consumption status	Electricity consumption trend
	Abnormal electricity consumption
	Average peak power consumption
	Power consumption grade
	Electricity quantity analysis
	Electricity charge analysis

3.4 Data preprocessing

Data preprocessing is a process of detecting, correcting or deleting damaged, inaccurate or unsuitable samples from index data. The purpose is to make data adapt to the model and match the needs of the model. The treatment method is described as follows.

(1) Dimensionless.

It is an operation of converting data from different specifications to the same specification, or data from different distributions to a specific distribution.

(2) Normalization of maximum and minimum values.

It is a process of centralizing data according to the minimum value, then scaling it according to the range and converging to between 0 and 1. The formula is as follows:

$$x^* = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

(3) Data standardization.

It is to centralize the data according to the mean value, and then scale the data according to the standard deviation. The scaled data shall obey the normal distribution with a mean value of 0 and a variance of 1. The formula is as follows:

$$x^* = \frac{x - \mu}{\sigma} \quad (2)$$

(4) Processing of missing values.

Retention: If logically missing values have a clear business meaning, the missing values should be retained or marked with special values.

Filling: If the filling logic is clear, the value should be filled according to the logic. In the case of random deletion, the mean, mode or median are used for filling.

Deletion: If the missing rate of a feature variable is too high and the correlation with the target variable is low, it can be considered to delete the variable.

(5) Outlier handling

Deletion: In the case of ensuring sufficient sample size and not affecting the original distribution of variables, the variable can be considered for deletion.

Box method: Through box value conversion, the adaptability of the model to the extreme value is realized,

and the leverage effect of the extreme value is reduced.

Average value correction: If the sample size of the data is very small, the average value of the two observed values before and after can be used to correct the outlier.

3.5 Feature engineering

Feature engineering refers to the process of extracting key information from source data for combination and mining strong correlation indicators through data processing methods. Actually, not all features meet the requirements of the model. Features not in compliance with the requirements in the event of entering the model will affect the stability and accuracy of the model. Under normal circumstances, firstly, screening is based on feature quality and distribution, such as screening the indicators with too large missing value, too small feature variance, single value and too large distribution difference between the training set and prediction set. Secondly, filtering based on feature importance and deleting features with small contributions can reduce the complexity of the model and improve its interpretability of the model. In the scoring model, feature IV value and other methods can be generally used for feature screening. The formula for measuring IV value is as follows:

$$IV = \sum_{i=1}^r \left(\frac{n_{i1}}{n_1} - \frac{n_{i2}}{n_2} \right) \ln \left(\frac{n_{i1}/n_1}{n_{i2}/n_2} \right)$$

Wherein, n_{i1} represents the number of normal customers corresponding to the i -th attribute of the characteristic variable, n_1 refers to the total number of low-risk customers in the sample, n_{i2} denotes the number of abnormal customers corresponding to the i -th attribute of the characteristic variable, and n_2 indicates the total number of abnormal customers in the sample. When IV is greater than 0.3, it shows that the feature has a good predictive ability and should be selected into the model.

4. Research on the model algorithm

4.1 Selection of model algorithm

The logistic regression model is a generalized linear regression analysis model, which has a strong explanatory ability, simple thinking and strong generalization ability, and its stability is only directly related to the stability of the variables entering the model. However, logistic regression generally only considers the linear relationship between variables and dependent variables, but in practice, the relationship between variables and dependent variables is more complex.

The tree model algorithm adopts a tree structure and uses layer-by-layer reasoning to realize the final classification. The decision tree can learn the deep relationship between features and express the complex relationship between variables and dependent variables, so as to achieve a good prediction effect. Compared with the logistic regression model, the decision tree has better generalization ability and interpretability on large data sets. Given the complexity of the payment behavior of power

users, the tree model algorithm is more suitable for default risk assessment scenarios. The calculation method is as follows:

(1) The generation of a classification tree

The classification tree uses the Gini index to select the optimal feature, and at the same time determines the optimal binary segmentation point of the feature. The Gini coefficient represents the impurity of the model. The smaller the Gini coefficient, the lower the purity and the better the characteristics.

For sample D , if D is divided into D_1 and D_2 according to a certain value A of feature A , the Gini coefficient expression of D under the condition of feature A is:

$$\text{Gini}(D) = \frac{|D_1|}{D} \text{Gini}(D_1) + \frac{|D_2|}{D} \text{Gini}(D_2) \quad (3)$$

(2) Score mapping

The forecast output is mapped to a score:

$$\text{score} = A + B * \log_2 \left(\frac{1 - \text{pred}}{\text{pred}} - \text{lag} \right) \quad (4)$$

Where A is the basic score and B is the step size. pred is the output of the decision tree model, and lag is the threshold of basic likelihood probability.

4.2 Model validation and optimization

AUC index is a common “weapon” for model verification. The AUC index is not affected by category imbalance, and different sample ratios will not affect AUC evaluation results. AUC is calculated as follows: M is the positive sample number, N is the negative sample number, and rank_i is the position of the i -th positive sample.

$$\text{AUC} = \frac{\sum_{i \in \text{positiveClass}} \text{rank}_i - \frac{M \times (M+1)}{2}}{M \times N} \quad (5)$$

When the prediction results of machine learning have an impact on users' finance, privacy and security, an interpretable and convincing risk control conclusion becomes very important. In this paper, it is suggested to use an interpretable model as the underlying algorithm. Relying on an intelligent machine learning platform, the principle and effect of the model are monitored in real time, mainly from the dimensions of tree model decision structure, model accuracy, feature importance, model report, etc., and timely adjustments and optimizations are made when anomalies are found.

For the sake of enhancing the interpretability, scientificity and practicability of the model, the expert experience method can also be used to modify the model [4]. By organizing experts in the industry, we discuss the scoring model and adjust the weight and calculation method of each indicator according to the characteristics of the power supply and consumption business and the needs of regulatory authorities or power supply companies. The expert team includes but is not limited to experts in the credit field, power field and finance field.

4.3 Examples of key indicator analysis

This paper selects the data of power customers in a certain area as a sample to model and classifies the risk level of users according to their credit scores. Because it is difficult to obtain some data such as bank credit, this paper mainly

analyzes the basic information of enterprises, public credit information and electricity consumption behavior information.

Through example calculation, it is found that the key indicators affecting the risk of unpaid electricity charges of enterprises include the number of times they have been implemented in recent three years, the number of overdue electricity charges paid in the previous year, and whether they are currently included in the business exception list. This paper takes the number of executions in the recent three years as an example to illustrate. Considering the distribution of the execution times of enterprises in the recent three years, the execution times are divided into three situations: never, once, twice and more than 3 times. The information values of this indicator are shown (see Table 2):

Table 2. Sample data of the key index

Number of executions in recent three years	Normal user frequency	Abnormal user frequency	IV value
Never	1649	1262	0.0575
Once	83	376	0.2459
Twice	60	145	0.0417
More than 3 times	8	17	0.0038
Total	1800	1800	0.3489

Because the IV value of this indicator is greater than 0.3, it shows that this indicator has a good predictive ability and should be included in the indicator system. It can also be seen from the above table that the proportion of abnormal customers increases in the group of companies that have been executed more frequently. Therefore, by monitoring the executed records of power customers in the recent three years, the risk of customers' unpaid electricity charges can be evaluated, and corresponding risk response strategies can be taken.

5. Prospects and suggestions

5.1 Recommendations for model construction

Combined with rich data dimensions, the machine learning model is used to comprehensively evaluate enterprise credit risk from point to area, which is beneficial to enhance the usability and high flexibility of the model, and ensure the scalability of the model. If the specific application subjects and application scenarios are different, the model can modify the label definition and adjust the indicator system according to the actual situation. This is because power customers from different channels and even customers from different regions have different risk performances, and different models should be configured based on their respective data.

5.2 Recommendations for risk response

The regulatory authorities can appropriately increase the frequency of daily supervision and inspection for high-risk customers, and track and deal with illegal clues in time. Power supply companies can also adopt differentiated risk

response strategies for power customers with different risk levels:

(1) There is a necessity to sign the *Electricity Charge Settlement Agreement* with new customers to clarify the period and time of electricity charge settlement, which can shorten the settlement period and time for high-risk customers.

(2) The power supply company shall implement the bank guarantee system or handle the asset mortgage electricity charge contract for high-risk production users.

(3) The power supply company provides a "bundled guarantee" between the customers of leasing operation and contracting operation, and the unsecured high-risk customers shall implement advance payment of electricity charges.

(4) Power supply companies should strengthen the management and control of electricity charges, take strict measures such as separate reminders for high risks, and directly link the payment of electricity charges with the order of peak power limit.

(5) The power supply company shall establish the credit rating of users, formulate the warning line of electricity charges by grades, and strictly implement the "Early Warning System of Electricity Charges" and the "Credit Rating System of Customers".

(6) Power supply companies are recommended to improve the level of power supply services, innovate more convenient and efficient payment methods, and improve the satisfaction of power users.

6. Conclusions

This paper explores a method for calculating the credit risk score of enterprises' unpaid electricity charges, and provides suggestions for risk response strategies. It should be noted that when adopting risk response strategies, attention should be paid to avoiding legal risks and ethical risks^[5]. The regulatory authorities should strictly perform their duties of credit supervision and implement disciplinary measures against dishonesty in accordance with laws and regulations. Power supply companies should not take advantage of their dominant position in market transactions, abuse credit data, formulate discriminatory power marketing strategies, and damage the legitimate interests of power users.

About the author

Jiankun Hu*, embarks on Energy Informatization Research and design of Electric Power Information System, has postgraduate qualification, as an senior engineer of Guangdong Electric Power Design and Research Institute Co., Ltd., China Energy Construction Group, hujiankun@gedi.com.cn

Haiji Wang, embarks on Electric Power Information Research and digitalization, has postgraduate qualification, as an senior engineer of Guangdong Electric Power Design and Research Institute Co., Ltd., China Energy Construction Group, wanghaiji@gedi.com.cn

Xueying Lu, embarks on Electric Power Information Research and project management, has postgraduate

qualification, as an senior engineer of Guangdong Electric Power Design and Research Institute Co., Ltd., China Energy Construction Group, luxueying@gedi.com.cn

References

1. Tian Y. J., Zhao Y. Y., Su Y., Xu B. Y. Applicability of power big data for enterprise credit reporting [J]. Electricity and Energy, 2021, 42 (4): 381-385.
2. Xie X H. Risk prediction of power customer arrears based on multi-scale feature extraction [D]. Southeast University (Nanjing). 2020.
3. Xie L. J., Jiang R., Pi Y Q., et al. Hybrid risk early warning algorithm of electricity customer's electricity charge default combined with RF-BP-LR [J]. Journal of Chongqing University of Technology (Natural Science), 2022, 36 (5): 250-258.
4. Li G., Li J. P., Sun X. L., Zhao M. Research on the combination of subjective and objective weights and its rationality [J]. Management Review, 2017, 12:17-26.
5. Zhang T., Algorithm Embedded Public Credit Comprehensive Evaluation: Internal Logic, Risk Review and Regulation Path [J]. Journal of Shenzhen University (Humanities and Social Sciences Edition), 2022, 39 (3): 96-105