ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20220628151845
Preliminary communication

LSTM Deep Neural Network Based Power Data Credit Tagging Technology

Ding LI, Jiayi CHEN*, Zhuo WANG, Yuehui SONG

Abstract: The value of power data credit reporting in the social credit system continues to increase, and the government, users and the whole society have deep expectations and support for power data credit reporting. This paper will combine the data labeling theory as the support, define the power data label and explain its labeling implementation. Based on the construction of knowledge graph, the method of labeling power data is introduced in detail: demand analysis method, index selection method, data cleaning method and data desensitization method. Use the sorted data labels to establish a label system for power data, and through its system, visualize the comprehensive situation of enterprise power data credit information to meet the development of power data credit business. This paper takes shell enterprises as the main representatives of credit risk enterprises, analyzes the power data in the three stages before and after loans, and builds a value mining model for power credit data. In the future, the data labeling technology and value mining model of the power data credit business will be comprehensively applied, and the power data label library and credit model library will be established and continuously improved, so as to facilitate the evaluation of the operation of the enterprise at different stages.

Keywords: labeling; logistic regression model; power credit data; value mining;

1 INTRODUCTION

The power data credit service is widely expected from the government, society and users. As of June 2019, my country has established the largest credit investigation system of the world, accumulatively collecting information about 990 million natural persons and 25.91 million enterprises. The average daily inquiries of personal and enterprise credit reports reached 5.5 million and 300 000 times respectively. Electricity credit reporting business has gradually become an important part of the social credit reporting system, and there is an urgent need for power data and related credit reporting products from all walks of life. The use of power data can timely reflect the production and consumption behavior characteristics of microeconomic entities, and analyze and judge business risks. Commercial value is emerging rapidly. Big data technology is accelerating its penetration and integration with traditional industries, promoting the upgrading of manufacturing and service industry technologies, and the rapid rise of the information economy. The use of big data technology can be reflected in various aspects of power system operation, providing strong support for the commercial utilization of power information.

With the vigorous development of the Internet and big data technology in recent years, various data products have emerged. At present, State Grid can directly collect customers' electricity consumption data through the electricity consumption information collection system. By establishing multi-dimensional data analysis models and various types of application scenarios, it provides favorable conditions for serving enterprises and customers. However, all kinds of data are scattered in various application systems, forming "data islands", and the huge value of data cannot be mined from scattered data. The implementation method of data labeling is a set of "middle-tier" components between the big data platform and upper-level business applications, forming the best practice plan for data governance. The core concept is to realize data business and computing intelligence. Data tags are the "bridge" between "data" and "business". Based on value mining technology, the data tag library can effectively solve the problems of building data services based on basic data, poor product flexibility, long construction period, and high cost, effectively supporting data. Taiwan business data is integrated and shared to realize a virtuous circle of business dataization and data businessization. Through the design and optimization of the data labeling system and the construction of label libraries for customers and equipment, the data value of data in various businesses can be fully utilized, the realization of data value-added services can be effectively promoted, and the company can serve the society and government functions. The company's service transformation will promote the development of the company's energy ecology and create a new business growth point for the company.

Modern credit risk management techniques have become more sophisticated, while there is a research gap on the entry of power credit data information into credit risk management. In order to fill this gap and improve the level of credit risk management, through various data mining techniques, this paper optimizes and improves the design of the data label system, establishes a power data label library, and builds a power data label library, so as to effectively promote data sharing and accelerate the use of data labels in various business applications value.

The remainder of this paper is organized as follows. In Section 2, a review of the existing studies was conducted. In Section 3 and 4, an improved method is constructed and developed. In Section 5, the proposed method is applied into the problem and evaluated the validity and superiority of the proposed method. Finally, the paper is concluded in Section 6.

2 LITERATURE REVIEW

In this section, we briefly present the concepts related to data labeling and text-mining.

2.1 Data Labelling

Data labeling is a simple and effective method to improve the efficiency of incremental data clustering, which is the process of assigning each newly added data point to the most similar cluster [1]. The process of establishing data labels is a process in which business

experts and data managers organize data physical tables together and abstract them into label expressions [2]. With the increasing application of big data and data mining, the connotation and extension of data labels are also expanding, and a variety of advanced data applications have been developed, such as customer credit rating and personalized marketing services [3], user behavior perception [4], power grid equipment portraits and power grid specifications [5], material price forecasting [6], operation monitoring analysis in the field of power grid dispatching, etc. [7]. This paper mainly studies the classification and management of data labels themselves. According to the labeling object, it can be divided into data model label, dataset label and unstructured label; according to the definition method, it can be divided into fact data label, statistical data label, model data label and combination definition label; according to label usage, it can be divided into data management label, business class tags, and sample data tags.

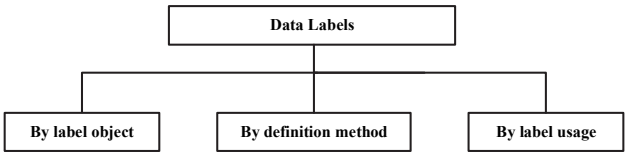

Figure 1 Data label classification framework

Data tags are usually a collection of highly condensed feature items specified by humans, and the information hidden behind the data is visualized through one or more tag applications. The creation, maintenance, and application of data tags constitute a data tag system [8]. From the perspective of system composition, data tags themselves are also a type of data application. The difference is that tags can be applied directly or become an important part of other data applications. The tag library is a unified encapsulation of data source data in the form of tags to realize the precipitation and sharing of business knowledge, to support the rapid construction and release of data application products, and to improve the efficiency of enterprise data application construction and delivery. The data label system is the core of the construction of the enterprise label library, and also the foundation of the enterprise data management and data application system. When planning and designing the system, the integrity of the business logic and the development and changes of the business and the data must be considered. The requirements for ability improvement should be based on the actual needs of enterprise operation and management. After sufficient research and discussion, the main framework of the labeling system should be determined, and then the implementation will be divided into pieces [9].

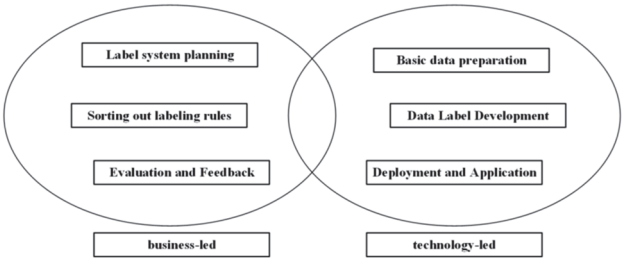

Figure 2 Creation of data label

2.2 Value Mining

Value mining is a process by which users obtain high-quality information from a given text. The need for value mining has grown considerably over the past few years. Coupled with big data analytics, the field of value mining is constantly evolving. Credit is a major sector that can benefit from these technologies; the analysis of large volumes of credit data is both a need and an advantage for businesses, governments and the public. This section discusses some important and widely used techniques for analyzing textual data in the context of credit.

Credit risk was first studied in western developed countries. For the development of corporate credit risk assessment methods, traditional methods, such as 5C factor analysis, are mainly qualitative analyses. Although widely used, these traditional methods still have some limitations. Later, the Zeta model was built. Subsequently, the credit monitoring model (KMV), credit measurement, credit risk model and credit portfolio view model appeared one after another. Martin first used a logistic model to predict a firm credit risk [10]. The results show that the logistic model has high accuracy and has many advantages over qualitative methods. Using statistical analysis and forecasting techniques, Krichene analyzes and determines the level of risk involved in lending and discusses short-term loan default forecasts for commercial banks in Tunisia [11]. Oreški and Oreški implement a dataset of retail credit in Croatia and Germany Genetic Algorithms and Neural Networks, propose a new classification technique optimized for cost sensitivity and applied to retail credit risk assessment [12]. The results demonstrate the potential of the new technique in terms of misclassification costs.

For the quantitative research of enterprise credit risk assessment, some scholars use traditional and modern methods to study credit risk measurement. Fu and Zhu talked about network supplier credit management models based on Petri net [13]. Hartono et al. examine the utility of Project Risk Management Maturity (PRMM) for project organizations in different domains [14]. Haghighi and Torabi quantified and analyzed uncertain information collected by experts and established a practical fuzzy risk assessment framework to deal with various potential risks faced by hospital information systems [15]. Wang et al. proposed a credit risk assessment strategy combining unsupervised learning and supervised learning, and compared the performance of credit models on different datasets [16]. The results show that this method has great advantages, and can be generalized to credit databases of other financial institutions. Zheng and Zhang constructed a dynamic Bayesian network model of supply chain risk to describe the nature of supply chain risk, and the results showed that supply chain risk changes over time and converges within a certain stable interval, occurrence time and holding time satisfy several Poisson processes [17].

To sum up, based on the above research results, scholars continue to innovate and improve enterprise risk assessment, in particular, the application of neural network algorithms is advantageous [18-20]. Modern credit risk management techniques have become more sophisticated. However, there is relatively little literature on the entry of power credit data information into credit risk management. In order to fill this gap and improve the level of credit risk

management, this paper introduces the power credit data information into the corresponding model. It is a useful supplement to the analysis of power credit data, which is beneficial for enterprises to accurately assess and manage their own credit risks in real time.

3 RESEARCH METHODOLOGY

In order to enhance the application value of business data and effectively support the construction of power grid informatization, the power business data identification, data management, and data processing are carried out based on tag technology. The labeling system has the characteristics of high speed, high flexibility, and strong pertinence. When the entire data structure is more complex and the combination of multiple computing and storage resources is required, the value of the labeling system is more obvious.

3.1 Definition and Implementation of Power Data Label

Power data labeling discovers potential data associations by mining the raw data sets and generates power data labels. Among them, the manual label refers to the label constructed manually by business personnel based on business experience judgment. Non-manual labels are data labels generated by writing business rules or developing and constructing algorithm models. Due to the diversification of data sources, the data traceability technology in big data can be used to record the source of data, its dissemination and calculation process, and provide auxiliary support for later mining and decision-making. Establishing data labels is a process of building a bridge between business and data, and it is also a process of expressing and solidifying the experience of business experts and the physical table structure of data. The implementation of electronic data labels builds a basic knowledge base based on knowledge graphs, and then uses a set of "middle-tier" components between the big data platform and upper-level business applications to form the best practices for data governance. The figure below shows the implementation of electronic data labels.

First build a knowledge graph, based on existing business experience, build a basic knowledge base, which can be dynamically updated later, followed by a heterogeneous computing platform that accesses multiple data sources, and the top layer is the various business applications. Between the computing platform and business applications, there are various big data governance components established with the labeling system as the core, and these components are divided into multiple layers. The lowest layer is data business, which provides basic guarantee for label service; the rule engine is a label production machine, which can continuously produce labels for business data; label factory and label center provide label derivative combination and label management and use audit functions; the upper layer is a series of general-purpose computing applications based on tags, including but not limited to integrated analysis, data relocation, etc.

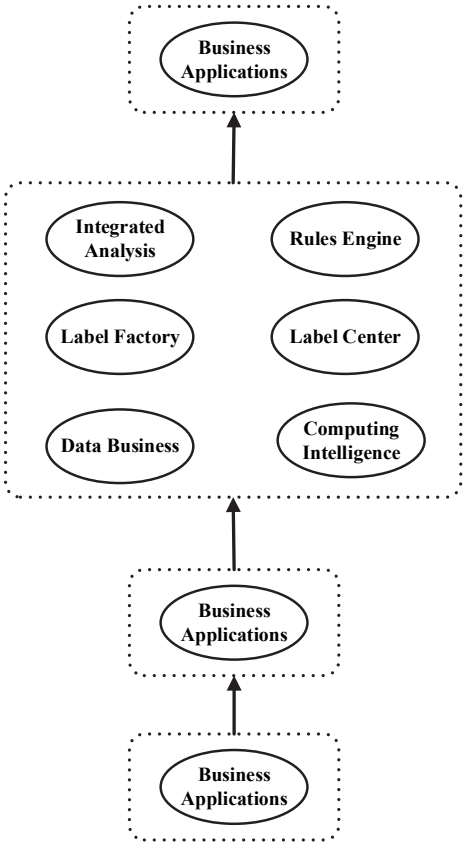

Figure 3 Implementation of electronic data labels

3.2 Overview of Labeling Construction

The overall steps of labeling construction are divided into: building knowledge graph, identifying requirements, selecting indicators, refining labels, building models, and practical business applications.

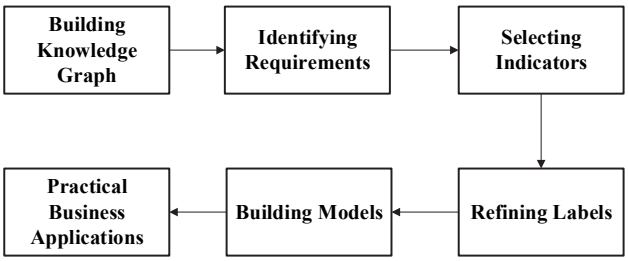

Figure 4 Implementation of electronic data labels

The detailed label construction steps are shown in the following figure. In the stage of building a knowledge graph, a basic knowledge graph, that is, a knowledge base, is constructed based on historical data according to the fields and businesses involved. After that, dynamic optimization is performed according to specific demand scenarios. In the stage of identifying requirements, comprehensively understand the content and scenarios of the requirements in a qualitative and quantitative way, analyze the requirements and extract the requirements. In the index selection stage, data is collected and organized into a data dictionary, and then basic indicators are formed. In the label extraction stage, the data fields are desensitized, and the corresponding data fields are extracted based on the basic indicators. Finally, according to the obtained labels, a model is built, which is then applied to practical problems after training.

Figure 5 Detailed steps for labeling construction

3.3 Detail of Label Construction

A knowledge graph is a semantic network that maps the real world to the data world and consists of nodes and edges, where nodes represent entities or concepts in the physical world, and edges represent attributes of entities or relationships between them. There are top-down and bottom-up methods for knowledge graph construction. The main processes include ontology modeling, knowledge extraction, knowledge fusion, knowledge storage, and knowledge reasoning. The basic knowledge graph is constructed based on three dimensions, namely industrial and commercial data, financial data, and power data. Power data includes payment, power consumption, abnormal behavior, time, etc., as shown in Fig. 6.

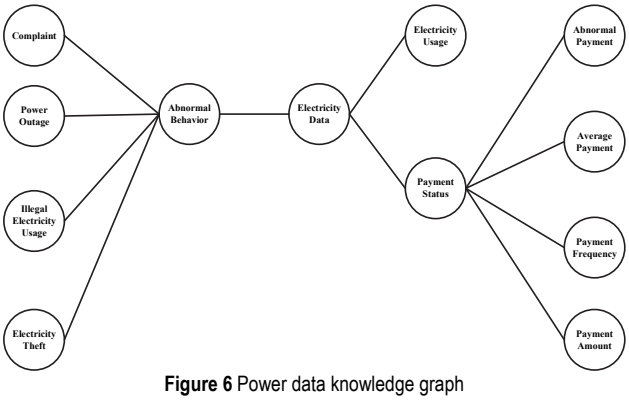

Figure 6 Power data knowledge graph

The existing requirements analysis and modeling methods mainly include structured analysis methods and object-oriented analysis methods. In different scenarios, the demand side will provide multi-dimensional and different data, and it is necessary to select different indicators according to the set demand goals for better follow-up analysis. The data dictionary contains many indicator fields, and indicators need to be selectively selected to prepare for the subsequent generation of key indicators. Based on the initial knowledge graph, it is dynamically updated through some indicator selection methods, and the available indicator data is filtered from the data dictionary, and more dimensions and indicators are added to the knowledge graph to make it more comprehensive and perfect.

Data cleansing can deal with missing values, out-of-bounds values, inconsistent code, duplicate data, etc. After the data cleaning is completed, this paper needs to desensitize some data fields, deform the sensitive information, and transform the data and provide it for use without violating the system rules to ensure that it can be used in access, development, testing and other environments. Safely use desensitized real datasets for reliable protection of sensitive content.

For the power credit data, by processing the data, extract the final label that meets the demand target. The basic attribute labels of electricity include power consumption, load type, etc. According to the data type, this paper roughly classifies the existing data. After the data dictionary is cleaned and desensitized, there are still 8 tables left, which are Enterprise information, electricity bill information, electricity consumption, abnormal conditions, and others, as shown in Tab. 1.

Table 1 Classification of Power Credit Data

Category	Surface
Corporate Information	customer table
Electricity Information	Receivable electricity bills, actual electricity bills
Electricity Usage	Electricity application
Abnormal Situation	Breach of power theft, power outage, inspection results
Other	user table

For descriptive fields, such as the description of the phenomenon of breach of contract and the content of on-site evidence collection records, natural language processing is used to convert the language into machine language. Then, according to the above classification, the data fields of each category are used as the input, and the business status of the enterprise is the output to perform clustering. Seven categories are obtained, which are electricity tariff level, interactive behavior, payment behavior, basic information of the enterprise, electricity usage specification, electricity usage behavior, and electricity usage capacity. In the time dimension, according to the data characteristics of the indicators, the general quarterly, half-yearly, annual and other stages are flexibly used. After that, continue clustering in the same way, and finally get the following label system, as shown in the appendix to Tab. 3. According to the obtained labels, a training model is established, and based on the above indicators, pre-loan, mid-loan, and post-loan models are established to evaluate the business situation of the enterprise at different stages.

4 VALUE MINING MODEL FOR CREDIT

In this paper, we choose to obtain a total of 30 W text data of corporate governance, corporate development quality and corporate development environment of electricity-using enterprises in a well-known platform, covering the period from September 27, 2020 to September 27, 2021(State Grid Xiongan Financial Technology Group Co., Ltd., China), and the obtained processed data are

divided into long and short texts according to the byte size, and different labeling analysis models are used for texts of different lengths.

Electricity credit reporting business has gradually become an important part of the social credit reporting system. All sectors of society have an urgent need for power data and related credit reporting products. The use of power data can timely reflect the production and consumption behavior characteristics of microeconomic entities, and study and judge business risks. In order to better evaluate the loan risk of an enterprise, the power data can be combined for evaluation, and the value mining model of credit data before loan, during loan and after loan can be established respectively.

4.1 Pre-Loan Value Mining Model

In order to better use power data to supervise enterprises before lending, it is necessary to carry out the label construction function of power data credit products based on credit business types, service objects, desensitization rules, product development, business processes and data mining needs. Design and build a basic power data label library. Through the use of user interviews, on-site surveys, questionnaires and other methods to conduct demand analysis, this paper clarifies that the use of power data to evaluate the operating conditions of the enterprise before lending requires the understanding of the overall power consumption of the enterprise, the overall power consumption, whether there are periodic fluctuations, whether the power consumption is stable, and whether there are abnormal conditions such as power outages. Using the structured analysis method, all data fields related to electricity are sorted out and classified into different tables according to the data to which they belong. Then, around the electricity consumption process, a power usage map is drawn, including the customer's own information, normal electricity consumption information, abnormal electricity consumption information, electricity bill information, etc.

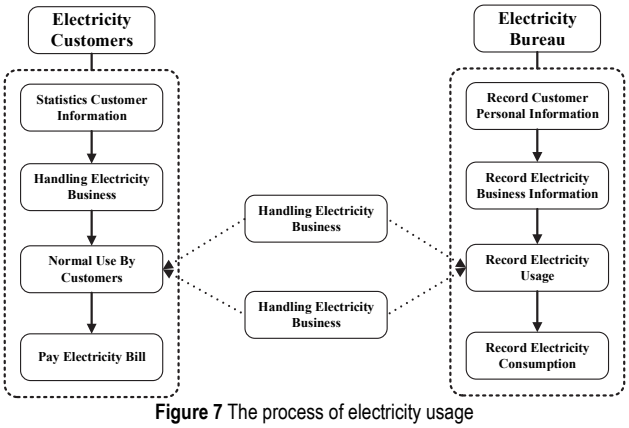

Figure 7 The process of electricity usage

Based on the above figure, the demand points of the pre-loan model are as follows: (1) Power consumption describes the user's power consumption, whether it is normal and stable, and whether it conforms to its industry and business characteristics, including power consumption and average power consumption in different time dimensions. Electricity consumption level, etc. (2) Electricity bills describe the user's electricity bill payment, including electricity bills in different time dimensions, whether to pay bills on time, etc. (3) Abnormal situation describes whether the user uses electricity illegally, including default electricity consumption and power outages information, etc. In this paper, the data cleaning tool Integrity is used for data cleaning. For descriptive fields, it is mainly to fill in missing data. For numeric fields, it is mainly to determine the type of the field and determine the data value range of the field. According to the key points of the pre-loan demand analysis, use the bidirectional maximum matching (Bi-MM) to match the key points with the key points of the obtained knowledge map, find the corresponding indicators through the connection of the key points, and convert the forward and reverse directions. Match results are compared to get the most accurate result. The pre-loan demand includes three aspects: electricity consumption, abnormal conditions, and electricity charges. It is further decomposed to include electricity consumption, time dimension, electricity consumption level, number of power outages, and electricity charges. After matching with the knowledge graph, a corresponding label is formed, which is in the appendix Tab. 4.

On the one hand, it compares and analyzes the distribution of electricity consumption and electricity costs of enterprises in the past 12 months with the distribution of electricity consumption and electricity costs of industries in the past 12 months. If the distribution and industry correlation are high, the company has a higher score and is less likely to be a shell company. On the other hand, analyze the change trend of the company's own electricity consumption, such as the number of power outages and the growth of electricity consumption. If the number of power outages is high and the growth of electricity consumption is poor, the score is low, and the company is more likely to be a shell company. The pre-loan business index system model is used to analyze the business situation of the pre-loan enterprise, as shown in the appendix Tab. 5.

4.2 In-Loan Value Mining Model

Based on the analysis of the historical power consumption data and financial data of the enterprise, mine the relationship between historical power consumption data and financial data, establish an association relationship, and then predict the latest financial status of the enterprise according to the latest power consumption data. The industry and the corresponding scale, objectively evaluate other productivity conditions; include the enterprise's electricity stealing records and default electricity consumption records into the dimension of evaluating enterprise credit. Finally, the credit indicator system model in the loan is obtained, as shown in the appendix Tab. 6.

4.3 Post-Loan Value Mining Model

The operation rule of the post-loan early warning indicator system model is to obtain the corresponding rating results such as "expansion, stability, contraction, warning" by checking the pre-set data of the core indicators of the enterprise post-loan early warning model one by one.

In the post-loan early warning scenario, the rating result of the enterprise is an alarm if any indicator is an alarm; when no indicator is an alarm, the rating result of the enterprise is a contraction if there is any indicator; Expansion means output expansion, and no expansion means the final output rating is stable. The post-loan early warning indicator system model is shown in the appendix Tab. 7.

5 MODEL SOLVING AND RESULT

Logistic regression is one of the most commonly used statistical learning algorithms for binary classification tasks and performs well in many tasks. As a nonlinear probability model, the data required by the logistic regression model is relatively easy to obtain and more flexible to process than other credit data evaluation models. According to the domestic and foreign research results summarized in this paper, it is not difficult to find that the logistic regression model has a very high utilization rate and recognition rate. The logistic regression model constructed in this paper is shown below.

$$P = \frac{1}{1+\mathrm{e}^{-s}} \tag{1}$$

$$S = \alpha + \sum_{i-1}^{n} \beta_i \chi_i \tag{2}$$

where $0 \le P \le 1$ and P is the probability value of credit risk for enterprises, α is the constant term, β_i is the estimated coefficient, χ_i is the factors affecting credit risk prediction. The probability function of logistic regression model is formed as a curve, $S \in (-\infty, +\infty)$.

$$\lim_{x\to+\infty} P = \lim_{x\to+\infty} \frac{1}{1+\mathrm{e}^{-s}} = 1 \tag{3}$$

$$\lim_{x\to-\infty} P = \lim_{x\to-\infty} \frac{1}{1+\mathrm{e}^{-s}} = 0 \tag{4}$$

As we can see, the closer the default probability is to 1, the greater the default risk; the closer the default probability is to 0, the smaller the default probability. Assuming Y represents the default event of a corporate loan, and its value is 0 or 1. Then, we can get the logarithm of Y.

$$\mathrm{Prob}(Y=1) = \ln \frac{P}{1-P} = \alpha + \beta_1 \chi_1 + \cdots + \beta_n \chi_n \tag{5}$$

Further simplification can be obtained as follows.

$$P = \left(\frac{\mathrm{e}^{Z}}{1-\mathrm{e}^{Z}} \right) \left(Z + \beta_1 \chi_1 + \cdots + \beta_n \chi_n \right) \tag{6}$$

Logistic model does not strictly stipulate the critical value. This paper sets the default threshold $P = 0.5$. If P is less than 0.5, we can say that the default probability is low, the enterprise credit level is high, and the risk of the credit loan is small. If P is more than 0.5, we can say that the default probability is high, the enterprise credit level is low, and the risk of the credit loan is high. Therefore, P value can be used as a reference value for enterprise credit risk prediction. The Adam optimization algorithm was used to train the neural network and the experimental results show that the Adam algorithm works excellently in Python tool.

Since the index is in the form of a percentage, this paper uses the annual power credit data to construct a logistic regression model with pre-loan, in-loan and post-loan indicators respectively. The regression results are shown in Tab. 2, and by examining the Hosmer & Lemeshow test values, we can see that the model has a very good fit. From the significance of the indicators, the first and fourth principal components of the pre-loan model are positively correlated with the credit level of technology companies. For the credit model, the second and fifth principal components are positively correlated with the credit level of the enterprise. For the post-loan model, the second, third, fourth, seventh, and eighth principal components are positively correlated with the firm credit level.

Table 2 Indicator of Post-loan value mining model

Indicator	Pre-loan	In-loan	Post-loan
F_1	–0.886	–0.128	0.519
F_2	–0.759	–2.028	–6.401
F_3	0.181	0.151	–2.822
F_4	–0.307	0.132	–14.546
F_5	–1.542	–1.812	–3.017
F_6	–1.239	0.234	2.711
F_7	—	–0.374	–2.477
F_8	—	0.058	–12.647
F_9	—	–0.912	–9.885
F_{10}	—	–0.431	–1.748
F_{11}	—	–1.595	2.027
F_{12}	—	0.179	–1.843
F_{13}	—	—	–9.304
Constant	–2.315	–1.665	–12.14
Chi-Square	24.346	26.811	76.92
Hosmer & Lemeshow test sig	0.093	0.496	0.935
–2 log likelihood	82.851	84.524	28.186
Cox-Snell *R* Square	0.202	0.194	0.497
Negelkerke *R* Square	0.318	0.354	0.726

The results of empirical test show that, from the regression results for Post-loan model, factors such as monthly electricity consumption, year-on-year electricity consumption, closest 12 Monthly electricity consumption, electricity consumption in the past 3 months, close12 month-to-month electricity consumption, and close12 cumulative length of monthly late payment have a greater impact on enterprise credit risk. The closer to the time of default, the higher the prediction accuracy of the model, and the more obvious the impact of power credit data information on default risk. The prediction accuracy rate of the constructed post-loan model is 95.5%, indicating that the model has high effectiveness and accuracy. The model can be used as one of the reliable scientific tools for credit risk assessment.

In addition, the findings of this paper have the following implications for enterprise risk. First, a successful business should focus on improving its profitability, operating capabilities and growth capabilities. The market competition among enterprises is becoming more and more intense. The improvement of profitability and management level promotes the development of

enterprises from the inside out. While accelerating their own development and improving their competitiveness, enterprises should strengthen the management of key financial indicators, increase inventory turnover, reduce inventory backlog, increase working capital, and speed up capital turnover. Through effective internal management, companies can enhance their external operating capabilities, reduce credit risk, and stimulate growth. Second, make full use of side information such as power credit data and apply it to enterprise credit risk assessment. Unstructured textual information contains a lot of incremental information. The external media, the public's evaluation of the company and the attitude of the company's own management will significantly affect the company's credit risk assessment. Enterprises should do a good job in managing external public opinion while doing a good job in their own risk management. It is important not only to pay attention to fluctuations in financial indicators, but also to consider the impact and shadow of non-financial factors on the business. Third, improve the enterprise's awareness of credit risk crisis. Credit risk is a gradual accumulation process. Enterprise managers can predict the probability of financial distress in advance through various indicators, and adjust business strategies in time before financial crisis occurs. Managers can then change financial policies to avoid a recession.

6 CONCLUSION

According to the credit reporting business types, service objects, desensitization rules, product development, business process and data mining requirements, this research has carried out a functional design for the label construction of power data credit reporting products, and constructed a power data basic label library with 180 labels. Combined with the label construction of power data, an analysis method of credit data indicators is proposed, and a credit data value mining model based on detailed data is constructed, including pre-loan, mid-loan, and post-loan models, and finally a credit data model library based on power data is formed, which is convenient for evaluating the operating conditions of the enterprise at different stages. When only financial indicators are used to predict corporate credit risk, there are many problems, such as unreliable data. And power credit data has a lot of incremental information in predicting corporate credit risk. Therefore, this paper uses the unstructured textual information of power credit data to evaluate corporate credit risk. This paper uses web crawler and value mining technology to quantify the power credit data information of enterprises. A Logistic regression model combined with power credit data information indicators is constructed. We empirically test the validity of the logistic regression model with data samples before, during and after lending. There are still two deficiencies in this paper in theory and method. First, due to space limitations, the introduction of the method in Section III is not detailed enough, and the explanation is not comprehensive enough. Secondly, the sample size selected in this paper is insufficient, which may lead to a certain degree of deviation between the results and the actual situation. Perhaps the results of the model need further tuning. Third, text mining methods should be further improved and optimized to mine more valuable variables and meet the needs of more information. At the same time, the empirical sample size can be expanded, the quality of collected data can be improved, and more comprehensive and reliable conclusions can be drawn.

Acknowledgements

This work was funded by Science and Technology Project of State Grid Corporation of China (1400-202057219 A-0-0-00). We appreciate their support very much.

7 REFERENCES

[1] Kielesińska, A. (2020). Safety of Imported Machines - Selected Issues in the Context of Polish (UE) Regulation. *System Safety: Human - Technical Facility - Environment*, *2*(1), 174-182. https://doi.org/10.2478/czoto-2020-0021

[2] Klimecka-Tatar, D. (2017). Value Stream Mapping as Lean Production Tool to Improve the Production Process Organization-Case Study in Packaging Manufacturing. *Production Engineering Archives*, *17*(17), 40-44. https://doi.org/10.30657/pea.2017.17.09

[3] Knop, K., Borkowski, S., & Stachurski, S. (2008). Modernity of Filling Machine in theAspect of the Exact Packing. *Quality Improvement and Machines Exploitation*, 63-68.

[4] Knop, K. (2016). Using a QFD Method and CTQ Tree to Identify the Areas Needing Improvement in the Product - Farm Truck Trailer, Scientific Papers of Silesian University of Technology. *Organization and Management Series*, *1947*(87), 219-236.

[5] Mumcu, Y. & Kimzan, H. S. (2015). The Effect of Visual Product Aesthetics on Consumers' Price Sensitivity. *Procedia Economics and Finance*, *26*, 528-534. https://doi.org/10.1016/S2212-5671(15)00883-7

[6] Pacana, A. & Czerwińska, K. (2019). Analysis of the Causes of Control Panel Inconsistencies in the Gravitational Casting Process by Means of Quality Management Instruments. *Production Engineering Archives*, *25*, 12-16. https://doi.org/10.30657/pea.2019.25.03

[7] Rosak-Szyrocka, J. & Knop, K. (2018). Quality Improvement in the Production Company. *Multidisciplinary Aspects of Production Engineering*, *1*. https://doi.org/10.2478/mape-2018-0066

[8] Siwiec, D. & Pacana, A. (2019). The Use of Quality Management Techniques to Analyse the Cluster of Porosities on the Turbine Outlet Nozzle. *Production Engineering Archives*, *24*, 33-36. https://doi.org/10.30657/pea.2019.24.08

[9] Siwiec, D. & Pacana, A. (2021). Method of Improve the Level of Product Quality. *Production Engineering Archives*, *27*(1), 1-7. https://doi.org/10.30657/pea.2021.27.1

[10] Martin, D. (1977). Early Warning of Bank Failure: A Logit Regression Approach. *Journal of banking & finance*, *1*(3), 249-276. https://doi.org/10.1016/0378-4266(77)90022-X

[11] Krichene, A. (2017). Using a Naive Bayesian Classifier Methodology for Loan Risk Assessment. *Journal of Economics, Finance and Administrative Science*, *22*(42), 3-24. https://doi.org/10.1108/JEFAS-02-2017-0039

[12] Oreški, S. & Oreški, G. (2018). Cost-sensitive Learning from Imbalanced Datasets for Retail Credit Risk Assessment. *TEM Journal*, *7*(1), 59. https://dx.doi.org/10.18421/TEM71-08

[13] Fu, Y. & Zhu, J. (2019). Network Supplier Credit Management: Models Based on Petri Net. *Tehnicki vjesnik - Technical Gazette*, *26*(5), 1434-1443. https://doi.org/10.17559/TV-20190722013047

[14] Hartono, B., Wijaya, D. F., & Arini, H. M. (2019). The Impact of Project Risk Management Maturity on

Performance: Complexity as a Moderating Variable. *International Journal of Engineering Business Management*, *11*, 1847979019855504. https://doi.org/10.1177/1847979019855504

[15] Motevali Haghighi, S. & Torabi, S. A. (2020). Business Continuity-inspired Fuzzy Risk Assessment Framework for Hospital Information Systems. *Enterprise Information Systems*, *14*(7), 1027-1060. https://doi.org/10.1080/17517575.2019.1686657

[16] Bao, W., Lianju, N., & Yue, K. (2019). Integration of Unsupervised and Supervised Machine Learning Algorithms for Credit Risk Assessment. *Expert Systems with Applications*, *128*, 301-315. https://doi.org/10.1016/j.eswa.2019.02.033

[17] Zheng, X. & Zhang, L. (2020). Risk Assessment of Supply-chain Systems: A Probabilistic Inference Method. *Enterprise Information Systems*, *14*(6), 858-877. https://doi.org/10.1080/17517575.2020. 1762004

[18] Kilvisharam Oziuddeen, M. A., Poruran, S., & Caffiyar, M. Y. (2020). A Novel Deep Convolutional Neural Network Architecture Based on Transfer Learning for Handwritten Urdu Character Recognition. *Tehnicki vjesnik - Technical Gazette*, *27*(4), 1160-1165. https://doi.org/10.17559/TV-20190319095323

[19] Qian, Y., Zeng, J., Zhang, S., Xu, D., & Wei, X. (2020). Short-Term Traffic Prediction Based on Genetic Algorithm Improved Neural Network. *Tehnicki vjesnik - Technical Gazette*, *27*(4), 1270-1276. https://doi.org/10.17559/TV-20180402112949

[20] Aydemir, E. & Gulsecen, S. (2019). Arranging Bus Behaviour by Finding the Best Prediction Model with Artificial Neural Networks. *Tehnicki vjesnik - Technical Gazette, 26*(4), 885-892. https://doi.org/10.17559/TV-20170629201111

Contact information:

Ding LI
State Grid Xiongan Financial Technology Group Co., Ltd., China
E-mail: dingli@sgec.sgcc.com.cn

Jiayi CHEN
(Corresponding author)
State Grid Xiongan Financial Technology Group Co., Ltd., China
E-mail: chenjiayi@sgec.sgcc.com.cn

Zhuo WANG
State Grid Xiongan Financial Technology Group Co., Ltd., China
E-mail: zhuowang@sgec.sgcc.com.cn

Yuehui SONG
State Grid Xiongan Financial Technology Group Co., Ltd., China
E-mail: yuehuisong@sgec.sgcc.com.cn

Appendix

Table 3 Business performance indicators

Classification	Indicator	Definition
Power consumption	Number of power outages in the last 12 months	The number of power outages in the last 12 months, the more power outages, the greater the possibility of empty shells
	Enterprise's largest electricity consumption in the last 12 months	The maximum value used to analyze the distribution of electricity consumption in enterprises
	The minimum electricity consumption of the enterprise in the last 12 months	Minima used to analyze the distribution of electricity consumption in enterprises
	The average electricity consumption of the enterprise in the last 12 months	Central trends for analyzing the distribution of electricity usage in the enterprise
	The standard deviation of the electricity consumption of the enterprise in the last twelve months	Used to analyze the discrete degree of electricity consumption distribution in enterprises
	Industry standard deviation of electricity consumption in the last twelve months	The degree of dispersion used to analyze the distribution of electricity consumption in the industry
	The average growth rate of total electricity consumption of enterprises in the last 12 months	Used to analyze the change trend of the company's total monthly electricity consumption
Electricity cost	The average electricity bill of the enterprise in the last 12 months	Central trends used to analyze the distribution of electricity bills in the industry
	Industry average electricity bill in the last twelve months	Central trends for analyzing the distribution of electricity bills for businesses

Table 4 Label system of power credit data

First-level Indicators	Secondary-level Indicator	Definition
Receivable tariff level	The average score of electricity bills receivable by enterprises in the past 12 months	The indicator period is 12 months, which is used to analyze the electricity consumption level of enterprises in the past 12 months
	Industry level of electricity bills receivable by enterprises in the last 12 months	The indicator period is 12 months, which is used to analyze the degree of correlation between enterprises and the average electricity cost of the industry
Actual electricity bill level	The average score of electricity bills received by enterprises in the past 12 months	The indicator period is 12 months, which is used to analyze the long-term electricity bill payment ability of enterprises
	Industry level of electricity bills actually collected by enterprises in the last 12 months	The indicator period is 12 months, which is used to analyze the correlation between enterprises and the average value of electricity bills received by the industry.
Receivable electricity bill fluctuations	Fluctuation of electricity bills receivable in the past 12 months	The fluctuation score of monthly electricity bills receivable in the past 12 months will not be counted for less than 12 months of network access or no previous data.
The actual electricity bill fluctuation	Fluctuation of electricity bills received in the past 12 months	The fluctuation score of the monthly actual electricity bills in the past 12 months will not be counted for less than 12 months or no previous data.
Overdue electricity bill	Cumulative overdue score in the past 12 months	The number of overdue times in the past 12 months is a measure of the quality of the company's payment behavior. The higher the value, the more serious it is.

	Liquidated damages scores for the past 12 months	The accumulated default amount of the enterprise in the past 12 months
	Last time in arrears	The number of days in arrears for the most recent default
	The company's arrears level in the past 12 months	Whether there is arrears
	The company's arrears level in the past 24 months	Whether there is arrears
	Enterprise arrears level in the past 12 months	Cumulative arrears in the past 12 months
	Enterprise arrears level in the past 24 months	Cumulative arrears in the past 24 months
Electricity account	Enterprise electricity bill balance level	Explain the degree of surplus in the balance of the business
Electricity recovery level	Electricity bill recovery in the last 12 months	Calculate the electricity bill recovery of the company in the past year through arrears and liquidated damages
	Enterprise electricity bill payment industry level	The electricity bill payment of enterprises in the past 12 months at the industry level
Electricity Stealing	Electricity theft level in the past 6 months	Refers to the number of times that users have stolen electricity in the past six months. Electricity theft is illegal, and such behavior is bad.
	Electricity theft level in the past 6 months	Refers to the amount of electricity stolen by customers in the past six months, and measures the severity of the customer's default from the amount
	Default electricity consumption level in the past 6 months	Refers to the number of times the customer has defaulted on electricity use in the past six months, indicating the customer's default behavior, and measuring the frequency of customer default according to the number of defaults
	The level of electricity theft in the past 12 months	Refers to the number of times that users have stolen electricity in the past year. Electricity theft is illegal, and such behavior is bad.
	Electricity theft level in the past 12 months	Refers to the amount of electricity stolen by customers in the past year. The amount of electricity stolen is huge, and the worse the behavior is
Breach of electricity	Default power consumption level in the past 6 months	Refers to the amount of default by the customer in the past six months, and measures the severity of the customer's default from the size of the amount
	Default power consumption level in the past 6 months	Refers to the number of times that the user has breached electricity usage in the past 6 months. The breach of electricity usage is illegal, and such behavior is bad.
	Default electricity consumption level in the past 12 months	Refers to the number of times the customer has defaulted on electricity usage in the past year, indicating the customer's default behavior, and measuring the frequency of customer default according to the number of defaults
	Default power consumption level in the past 12 months	Refers to the amount of default by the customer in the past year, and measures the severity of the customer's default from the size of the amount
Power outage analysis	Business outage levels in the last 12 months	The indicator period is 12 months, which mainly reflects the number of power outages of the enterprise.
Electricity level	The level of enterprise electricity consumption in the industry	Used to analyze the position of enterprise electricity consumption in the industry
	Enterprise Monthly Electricity Index	The comparison between the last month's electricity and the monthly average (12 months) reflects the level of electricity used in the previous month
	The company's monthly electricity consumption	The indicator period is 1 month, which mainly reflects the total electricity of the enterprise.
	Electricity consumption peer ranking	By comparing the accumulated electricity consumption of the enterprise in 12 months with the electricity consumption of enterprises in the same industry (national standard level 3) in the province, it is judged that the enterprise is in: head position, middle position, and tail position
	electricity account	This indicator is used to verify the normal operation of the enterprise
Electricity characteristics	Electricity customer sales status	The user judges whether the company has sold its account in the State Grid Corporation of China
	Power delivery status of electricity customers	The user judges whether the enterprise has the behavior of reporting installation in the State Grid Corporation
	Enterprise electricity consumption scale index	The long-term electricity consumption of the enterprise is less than the industry average electricity consumption, and its production status may be much lower than the industry average level
Electricity Trend	Average electricity growth trend of enterprises in the last 12 months	The indicator period is 12 months and is used to analyze the growth trend of the total electricity consumption of the enterprise
	Monthly electricity consumption year-on-year index of the company in the last 12 months	The monthly electricity consumption in the past 12 months is compared with the previous period, and no statistics are made for less than 12 months of access to the network.
	The average growth level of electricity consumption of enterprises in the past 3 months	The level of the user's three-month power consumption fluctuation in the industry, reflecting whether the production status of the enterprise conforms to the normal law of the industry

	The average growth level of electricity consumption of enterprises in the past 12 months	The indicator period is 12 months, and it is used to analyze the average growth trend of electricity consumption in the past 12 months.
Electricity fluctuation	The fluctuation of electricity consumption in the past 3 months	The monthly electricity consumption of the past 3 months is compared with the previous period, and no statistics will be made if the access to the network is less than 3 months or there is no data from the previous period.
	Fluctuation of electricity consumption in the past 12 months	The percentage change of the accumulated electricity consumption in the past 12 months compared with the previous period, and no statistics will be made if the network access is less than 12 months and there is no data in the previous period.
	The fluctuation of electricity consumption of enterprises in the industry in the past 12 months	The indicator period is 12 months, which mainly reflects the degree of correlation between the power consumption of enterprises and the industry
Electronic channel information	Information quality situation	The authenticity and comprehensiveness of enterprise information can measure the quality of enterprise information
Basic conditions	Risk level	Risk level reflects
		Enterprise risk, the smaller the value, the higher the credit rating
	Credit rating	Customer's credit rating as assessed by credit evaluation
	Legal person ID number	ID number
	Category	It is used to analyze the power consumption of enterprises in different industries and is suitable for various scenarios.
	Enterprise nature	Used to analyze business size
	Company name	This indicator is used to verify the identity information of the enterprise
	Household age	old users, new users

Table 5 Indicator of pre-loan value mining model

Indicator	Definition	Calculation
SCORE_OUTAGE	The more actual power outages, the greater the possibility of empty shells	Number of power outages in the last 12 months
SCORE_DIFF	The smaller the difference in electricity consumption, the greater the possibility of an empty shell	(The company's maximum electricity consumption in the last 12 months - the company's minimum electricity consumption) / The company's average electricity consumption in the last 12 months
SORE_VOLA	Compared with the smaller fluctuations in the industry, the possibility of empty shells is greater	The standard deviation of the average electricity consumption of the enterprise in the last 12 months / the standard deviation of the average electricity consumption of the industry in the last 12 months
SCORE_LEVEL	It is positively correlated with the industry average electricity cost, the bigger the better	The average electricity bill of the enterprise in the last 12 months/the average electricity bill of the industry in the last 12 months
SCORE_INREA	The greater the average electricity consumption growth rate, the less likely the empty shell is	The average growth rate of total electricity consumption of enterprises in the last 12 months
ENT_SHELL_SCORE	The lower the value, the higher the risk of empty shells.	Based on the scores of the above 5 items, based on the machine learning algorithm, calculate the empty shell law of the enterprise

Table 6 Indicator of in-loan value mining model

Indicator	Definition	Calculation
Account opening time	How long the account has existed as of the current time	counting
Industry electricity consumption analysis in this province (last 12 months)	The cumulative electricity consumption in the last 12 months is ranked in the industry.	The current company's cumulative electricity consumption in the past 12 months is ranked in the same industry in this province/Number of companies in the same industry in this province * 100%
Monthly electricity consumption year-on-year index in the past 12 months	The monthly electricity consumption in the past 12 months is compared with the previous period, and no statistics are made for less than 12 months of access to the network.	((Number of months with a year-on-year growth of 0~10%) * 0.1 + (Number of months with a year-on-year growth of 10~20%) * 0.2 + (Number of months with a year-on-year growth of 20~50%) * 0.5 + (Year-on-year growth of 50~100%) Number of months) * 0.8 + (Number of months increased by more than 100% year-on-year) * 2) – ((Number of months decreased by 0~10% year-on-year) * 0.1 + (Number of months decreased by 10~20% year-on-year) * 0.2 + (Number of months decreased by 20~50% year-on-year) * 0.5 + (Number of months decreased by 50~80% year-on-year) * 0.8 + (Number of months decreased by 80~100% year-on-year) * 1)
Year-on-year electricity consumption in the past 3 months	The percentage change of the accumulated electricity consumption in the past 3 months compared with the previous period, no statistics will be made if the network access is less than 3 months and there is no data from the previous period.	(Accumulated electricity consumption in the past three months - accumulated electricity consumption in the same period of the previous year for the same period of three months) / Cumulative electricity consumption in

		the same period of the previous year for the same period of three months * 100%
Year-on-year electricity consumption in the past 12 months	The percentage change of the accumulated electricity consumption in the past 12 months compared with the previous period, and no statistics will be made for those who have been connected to the network for less than 12 months and have no data from the previous period.	(Accumulated electricity consumption in the past 12 months - 12 months accumulated electricity consumption in the same period of the previous year) / 12 months accumulated electricity consumption in the same period of the previous year * 100%
Electricity consumption in the past 3 months	The percentage change of the cumulative electricity consumption in the past 3 months compared with the previous 3 months, no statistics will be made if the network access is less than 3 months and there is no data from the previous period.	(Accumulated electricity consumption in the past 3 months - accumulated electricity consumption in the previous 3 months) / accumulated electricity consumption in the previous 3 months * 100%
Electricity consumption in the past 12 months	The percentage change of the cumulative electricity consumption in the past 12 months compared with the previous 6 months, and no statistics will be made if the network access is less than 6 months and there is no data from the previous period.	(Accumulated electricity consumption in the past 12 months - accumulated electricity consumption in the previous 12 months) / accumulated electricity consumption in the previous 12 months * 100%
The number of default electricity usage	The number of default electricity usage in the past 12 months	counting
Current accumulated late payment	Current accumulated late payment	counting
Corporate financial statements for the past 12 months	Mining of the relationship between corporate financial statements and corporate productivity	Create relational model through logistic regression algorithm and correlation algorithm
Industry analysis of the company in the past 12 months	Calculate the industry average using the electricity consumption data of enterprises of the same scale in the same industry	average value
Industry average	Comparison of the company with the industry average	The current company's cumulative output in the past 12 months ranks among the same industry and the same scale in this province/Number of companies in the same industry in this province * 100%

Table 7 Indicator of post-loan value mining model

Indicator	Definition	Thresholds			
		expansion	Stablize	shrink	alert
Industry electricity consumption analysis in this province (last 12 months)	The cumulative electricity consumption in the last 12 months is ranked in the industry.	0 to 30	31 to 60	61 to 90	91 to 100
Monthly electricity consumption year-on-year index in the past 12 months	The monthly electricity consumption in the past 12 months is compared with the previous period, and no statistics are made for less than 12 months of access to the network.	2.4 and above	1.2 to 2.4 the following	–1.2 to below 1.2	–1.2 the following
Year-on-year electricity consumption in the past 3 months	The percentage change of the accumulated electricity consumption in the past 3 months compared with the previous period, no statistics will be made if the network access is less than 3 months and there is no data from the previous period.	20 and above	–10 to below 20	–20 to below –10	below –20
Close 12 Monthly electricity consumption year-on-year	The percentage change of the accumulated electricity consumption in the past 12 months compared with the previous period, and no statistics will be made if the network access is less than 12 months and there is no data in the previous period.	20 and above	–10 to below 20	–20 to below –10	–20 or less
Electricity consumption in the past 3 months	The percentage change of the cumulative electricity consumption in the past 3 months compared with the previous 3 months, no statistics will be done if the network access is less than 3 months and there is no previous data.	20 and above	–10 to below 20	–20 to below –10	–20 or less
Close 12 month-to-month electricity consumption	The percentage change of the accumulated electricity consumption in the past 12 months compared with the previous 6 months, and no statistics will be made if the network access is less than 6 months and there is no data in the previous period.	20 and above	–10 to below 20	–20 to below –10	–20 or less
Close 12 Cumulative length of monthly late payment	The number of days that electricity bills are not paid within the specified time (the number of days for which liquidated damages are incurred) can be accumulated according to the number of transactions. For example, the electricity bill has been overdue for 120 days in January, and the overdue days in February are 90 days, so the total is 210 days.	-	-	30 to 90	90 and above
Close 12 Cumulative monthly late payment amount	The amount of electricity bill payment that is not completed within the specified time (the principal of electricity bill that generates liquidated damages)	-	-	10 000 to 500 000	More than 500 000
Close 12 Number of monthly late payment	The number of times the electricity bill was not paid within the specified time (the number of electricity bills that resulted in liquidated damages)	-	-	2	2 or more
Current accumulated late payment	Current accumulated late payment	-	-	10 000 to 100 000	More than 100 000
The number of default electricity usage in the past 6 months	Close 6 Monthly default electricity usage	-	-	-	1 or more
The number of default electricity usage in the past 12 months	The number of default electricity usage in the past 12 months	-	-	1	2 or more
Electricity theft in the past 6 months	Close 6 Electricity theft per month	-	-	-	1 or more