IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON COLLABORATION FOR INTERNET OF THINGS

Received November 18, 2018, accepted December 1, 2018, date of publication December 14, 2018,
date of current version February 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2886551

Analysis and Identification of Power Blackout-Sensitive Users by Using Big Data in the Energy System

CHUNYAN SHUAI[1], (Member, IEEE), HENGCHENG YANG[2], XIN OUYANG[3], MINGWEI HE[1], ZEWEIYI GONG[4], AND WANNENG SHU[5], (Member, IEEE)

[1]Faculty of Transportation Engineering, Kunming University of Science and Technology, Kunming 650093, China
[2]Faculty of Electrical Power Engineering, Kunming University of Science and Technology, Kunming 650093, China
[3]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650093, China
[4]Yunnan Electric Power Research Institute, Yunnan Power Grid Co., Ltd., Kunming 650217, China
[5]College of Computer Science, South-Central University for Nationalities, Wuhan 430074, China

Corresponding author: Xin Ouyang (kmoyx@hotmai.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0306405, and in part by the National Natural Science Foundation of China under Grant 61562056, Grant 61364008, and Grant 61603420.

ABSTRACT With the further liberalization of the electricity market of China, customers' requirements, characteristics, and distribution, as well as the quality, security, and reliability of power supplies without interruption, have received considerable attention from power companies, policymakers, and researchers. How to deeply explore the distribution characteristics of electricity customers and analyze their sensitivities to electricity blackouts has become an especially important problem. This paper takes over 0.1 billion data, collected by various smart devices of the Internet of Things in the power system of China, such as smart meters, intelligent power consumption interactive terminals, data concentrators, and other cross-platform data, for example, 95 598 telephone records, complaint information, user bills, user information, and maintenance records, as study objects, to analyze the consumption characteristics of power users. It has been found that there is a wide range of power users who pay different electricity bills; a long-tail distribution following a power law lies in the number of users versus their paid electricity bills. Meanwhile, there are two Pareto effects (2-8 rule): the number of residents and non-residents versus their electricity bills, and the number of large industrial users and general industry (business users) versus in their electricity consumption and bills. Then, a decision tree algorithm is proposed to capture the characteristics of electricity consumers and to recognize the crowd who is power blackout sensitive. The evaluation indexes and parameters of the decision tree are discussed in detail, and a comparison with other intelligent algorithms shows that the decision tree has a good recognition performance over that of others, and the characteristics used to identify the blackout-sensitive crowd is various. All the results state that except for economic factors, positive social effects should also be considered. Various marketing strategies to satisfy different requirements of power users should be provided to promote long-term relationships between the power companies and power customers.

INDEX TERMS Blackout sensitivity, big data, decision tree, electricity market, Internet of Things, long-tailed, Pareto effect.

NOMENCLATURE

Notations	Description
Y	The dependent variable, or target variable. It can be ordinal categorical, nominal categorical or continuous. If Y is categorical with J classes, its class takes values in $C = \{1, \ldots, \}$.
X_i, i $= 1, \ldots, M$	The set of all predictor variables. A predictor can be ordinal categorical, nominal categorical or continuous.
$D = \{x_i, y_i\}_{i=1}^{N}$	The dataset, which consists of N samples.
W_n	The case weight associated with case n.
f_n	The frequency weight associated with case n. A non-integral positive value is rounded to its nearest integer.

The associate editor coordinating the review of this manuscript and approving it for publication was Chunsheng Zhu.

I. INTRODUCTION

In 2015, the state council of China released a file containing ''some opinions on further deepening the reform of the electric power system''. Its core is to establish an electricity market and implement it in such a way that the market plays a decisive role in resource allocation [1]–[4]. The primary objective is to create a market-oriented electricity price that can accurately reflect the relationship between market supply and demand and a reform of the electricity retail side becomes a key component of power system reform. A lot of social capital has inflowed to the electricity retail business, and active consumers who engage in energy consumption and production and provide ancillary services in a dynamic and interactive manner will be an integral part of the future smart grid. In these contexts, the original Power Grid Corps must improve power quality and security, change the original extensive operation mode, customer service content to satisfy the increasingly personalized and accurate requirements of customers [5], [6]. Aside from the power quality and price [7], the security and reliability of the electricity supply–that is, a high service level and a constant supply of electricity without interruption of service–are another principle consideration for users [8]. However, power failures caused by natural factors [9] and unnatural factors, result in huge losses and threaten the integrity of electric energy systems around the world [8], [10]. The common causes of outages can be categorized into supply side factors, demand side factors and political economy factors. In terms of the different impacts to the residents living in the area and the different damages (economic and others) [8], [11], [12], the power supply reliability in China is divided into three levels [13]: (1) The first level load. The interruption of the power supply at this level can cause personal accidents and large economic losses. (2) The second level load. The power supply interruption of this level will cause greater economic losses. (3) The third level load. This level consists of those who are not in the first or secondary loads, such as small towns, housing estates, auxiliary workshops of a factory, etc. Due to different losses, different power customers have different reflections and ''degrees of sensitivity'' to blackouts and they can be divided into blackout-sensitive and insensitive classes. Previous works on electricity outages have focused primarily on the physical and infrastructural causes of power disruptions and restorations. Such research is essential for understanding how to reduce the number and duration of outages. But, they provide little insight into who is more sensitive to power outages.

Due to the opening power market and fast growing electricity consumption [7], [13], how to deeply explore distribution characteristics of electricity customers and analyze their sensitivities to electricity blackouts has become an especially important problem. Currently, with a number of smart devices and sensors of Internet of things (IOT) utilized in the electric power system, such as different smart meters, various terminals, monitoring equipment, telemetering devices, data concentrators, etc., a large volumes of data have been collected,

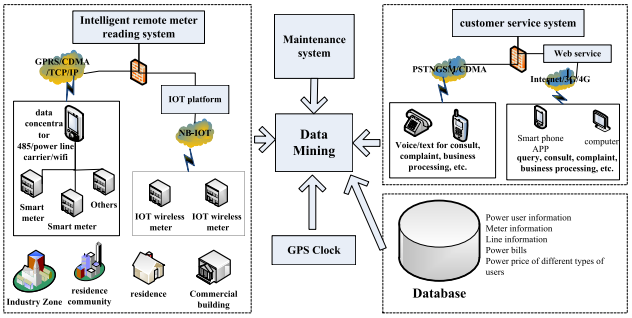


FIGURE 1. Various data of IOT devices and different information platforms used to data mining.

as Fig. 1 shown. Meanwhile, big data mining techniques have been applied to state estimation, forecasting, and control problems of power devices [14]–[16]. In this paper, we firstly fuse various over 100 millions data of multiple platforms and sensors, including smart meters, user information, 95598 telephone records, complaint information, user bills, maintenance records, and so on. And then, we apply statistical analysis and data mining technologies to capture the consumption characteristics of power users and identify their sensitive degrees to electricity blackouts. The main contributions of this paper are as flowing.

- It is found that power users pay a wide range of electricity bill amounts and most users (approximately 90%) pay electricity bills less than RMB 2,000 per year. A long-tailed distribution lies in the number of users against their paid electricity bills and when electricity bills are less than RMB 10,000, the long-tailed distribution follows a power law $y = 4^{*}10^{8}x^{-1531}$ with goodness of fit $R^2 = 0.9808$.
- It is found that there are two Pareto effects (2-8 rules): one is the number of residents and non-residents versus their electricity consumption, and the other is the the numbers of industry users and general industry (business users) versus their electricity consumption and bills.
- The blackout sensitive crowds distribute in all walks of life. The factors, such as industry types, economic losses, interruption duration, interruption time, location (urban or rural area) and so on, will effect the customer's sensitivity degree to power outages.
- A decision tree algorithm can effectively capture characteristics of electricity consumers' behaviors and recognize power blackout sensitive users with a high recognition recall and precision rate over other methods.

These results can provide reasonable support for planning for a power outage or normal maintenance operation and loss assessment in the case of load overload, attaining better social and economic effects and decreasing consumers' payments.

II. LITERATURE REVIEW

Compared with the most developed and well-established European electricity markets [17], the electricity market of

China is in its initial opening stage, and its reform will be in favor of economic developments [6], [18]. This requires an in-depth analysis of customers and the power market. Currently, a vast amount of data in the power field has been collected, and researches studying the power market and users' requirements have been carried out widely. These researches focus on using data mining and machine learning technologies to analyze power market characteristics, mine electricity customers' behaviors, evaluate and forecast power outages, and so on [14]–[16].

Electricity market and consumption. Zhang *et al.* [18] propose a novel dynamic system model to clarify causal relationships between electricity consumption, electricity supply, electricity price, and their complex interrelationships. By means of Lyapunov exponents, artificial neural networks and bifurcation diagrams, they study the statistical data of the Chinese electricity market from 1997 to 2012, and establish four efficient single regulatory strategies. The results show that excessive regulation of Chinese electricity market should be avoided, and that integrated regulatory strategies and peak-valley price are better than a single regulatory strategy. Shiu *et al.* [19] apply the error-correction model to examine the causal relationship between electricity consumption and the real GDP of China during 1971–2000. The estimation results indicate that the real GDP and electricity consumption of China are cointegrated and that there is a unidirectional Granger causality running from electricity consumption to real GDP but not vice versa. Furthermore, Zhang *et al.* [20] provide an extensive overview of the relationship between Chinese electricity consumption and economic growth in three dimensions, i.e. the time dimension, the regional dimension and the industrial dimension. By means of a multivariate model, Johansen cointegration test and vector error correction model, Al-Bajjali *et al.* [21] find that GDP, urbanization, population, structure of economy and aggregate water consumption are significantly and positively related to electricity consumption, while electricity price is significantly and negatively related to electricity consumption.

Li *et al.* [6] summarize the current power demand response status, feasible demand response market schemes and demand response pilot projects in China. They point out that challenges associated with demand response are the result of lack of a suitable market mechanism for the current Chinese power industry. Wang *et al.* [7] present a systematic review on the research and development status of the residential tiered electricity price policy in China. By using the electricity consumption data of smart meters for 10,000 Australian households for a year, Bedingfield *et al.* [22] present a new adaptable and scalable algorithm to understand electricity usage behaviors and provide customized electricity billing. Rathod *et al.* [23] carry out a K-means clustering algorithm on data from 20,000 consumer meters in the city of Sangli to form different clusters. Then, association analysis is employed to discover electricity consumption patterns at the regional level in a city and extract knowledge concerning electricity consumption with respect to atmospheric temperature and physical distance from geographic features, like rivers, farms, the ground and highways. An pattern characterization framework [24] incrementally explores and extracts actionable knowledge versus the time from meter's stream data. Knowledge discovery in database (KDD) [25] is applied to identify typical load profiles of medium voltage electricity customer characterization. A rule set for automatic classification is also developed to classify new consumers of a real database. Huang *et al.* [26] investigate the application and effectiveness of several data mining approaches for electricity market price classification, and proposes a new data model for forming the initial data set for price classification and forecasting.

Power outages. Societies are highly reliant on power systems for their energy needs. Reliability assessment is performed in both planning and operation of the power system. Although there has been increasing interest in hardening the power system to be resilient against power outages, the risk of power outages cannot be completely diminished. In addition, power blackouts have resulted in various impacts to residents' living spaces, public services and facilities, and huge damages to the economy [11], [12]. By using publicly available information of historical major power outages, socio-economic data, state-level climatological observations, electricity consumption patterns and land-use data, Mukherjee *et al.* [28] have developed a two-stage hybrid risk estimation model to address power outage risks. The results suggest that power outage risk is a function of various severe weather-induced factors, such as the type of natural hazard, expanse of overhead transmission systems, the extent of state-level rural versus urban areas, and, potentially, the levels of investment in operations/maintenance activities.

Data mining and intelligent algorithms are also used to forecast faults to reduce blackouts and locate faults after blackouts. Diao *et al.* [29] present an online voltage security assessment scheme using synchronized phasor measurements and periodically updated decision trees (DTs) to avoid blackouts. By using the past blackout representative events, the DTs are first trained offline to determine detailed voltage security, and then forecast 24-h ahead operating conditions. The DTs are also updated hourly by using newly predicted conditions for robustness improvement. Morales *et al.* [30] present a data mining methodology to perform signal detection, calculate traveling wave times and determine the lightning stroke location along the transmission line after an outage. Since the methodology is immune to flash currents, it is swift and effective in locating the impact point, especially in situations that the factors, such as flash current values, inception angles, distances from the impact point to protection relays and direct and indirect lightning, are considered. Han *et al.* [31] integrate various data mining techniques to analyze power plant faults and introduces a random forest to predict fault types of the power plant and rank important features. Xu *et al.* [32] extend the fuzzy classification algorithm to the E-algorithm to alleviate

the effect of imbalanced data, and then applies this algorithm to identify the distribution of faults on Duke Energy outage data. Dubey *et al.* [33] propose DTs and random forests for enhancing distance relay performance to reduce power outages during power swings in both compensated and uncompensated power transmission networks. Meanwhile, data mining and intelligent algorithms are applied in risk assessment of power failures. Castillo and Anya [27] review various models and algorithms that predict hazards and the losses of energy services to customers. Kamali *et al.* [34] propose a new two-stage scheme to predict the risk of a blackout in electric energy systems. In the first stage, the boundaries of electric islands are determined by using a mixed integer non-linear programming model that minimizes the cost of generation re-dispatch. In the second step, various scenarios, including the island and non-island conditions, are generated and a DT classification technique is utilized to predict the risk of a blackout. The proposed algorithm is simulated over the IEEE 39-bus test system to demonstrate its performance in online applications. To overcome the drawbacks of the high computational cost in classical N-k-induced cascading contingency and outage analysis, a de-correlated neural network ensembles algorithm [35] is proposed, which is comprised of a cascading failure simulation module for post-contingency analysis and a risk evaluation module.

Although most of the aforementioned studies state that the data mining and intelligent algorithms have been extensively used in the areas of power markets, power consumption, power outage faults and other aspects of the power field, seldom does research focus on studying the crowd who is sensitive to power failures.

III. CHARACTERISTICS OF ELECTRICITY CUSTOMERS

A. LONG TAIL AND PARETO EFFECT

Long-tail distribution. For a random variable X and all $t > 0$, there is the following distribution function F:

$$\lim_{x\to\infty} pr[X > x + t|X > x] = 1, \tag{1}$$

or equivalently

$$F(x + t) \sim F(x) \quad as\ x \to \infty. \tag{2}$$

F is called a distribution with a long tail, a kind of heavy-tail distribution [36]. Intuitively, it is interpreted as if the long-tailed quantity exceeds some high level as the probability approaches 1. In statistics and business, the distribution could involve popularities, random numbers of occurrences of events with various probabilities [37]. In business, the term long tail is applied to rank-size or rank-frequency distributions, which often form power laws and are thus long-tailed distributions [36] in the statistical sense.

Pareto effect. The Pareto effect (also known as the 2-8 rule, the law of the vital few) states that, for many events, roughly 80% of the effects come from 20% of the causes [36]. Mathematically, the 2-8 rule is roughly followed by a power law distribution (also known as a Pareto distribution) for a particular set of parameters, and many natural phenomena have been shown empirically to exhibit such a distribution [36]. In economics, the law is 80% of the economic profits derive from 20% of important customers, and 80% of the secondary customers generate only 20% of the total profits.

IV. CLASSIFICATION/REGRESSION TREES AND CHAID ALGORITHM

The classification algorithms are supervised, as the class of each object in the data set is known a priori. The objective of these algorithms consists of learning a function or a set of rules, denoted as a classifier, which allows assigning a new (unobserved) object to the correct class. There are several types of algorithms used to train classifiers, which can be organized by learning strategy:

- **Statistical models** suppose the classes of objects are generated in terms of some probabilistic distribution, such as linear and quadratic discriminant analysis [40];
- **Artificial neural networks (ANNs)** attempt to model the human brain mathematically. An example is the back propagation multilayer perceptron algorithm [41];
- **Support vector machine algorithms (SVMs)** try to seek out hyperplanes in a high-dimensional feature space that separates the data into different categories. A new item is classified according to its location relative to the established hyperplanes [42];
- **Ensembles of classifiers combine multiple classifiers** construct a more robust classifier, generally by applying a voting mechanism, such as boosting algorithms [43].
- **Classification and regression tree algorithms** build prediction models from observed data. The models are achieved by iteratively dividing the data space and matching a prediction model in each partition. As a result, the division can be graphically described as a decision tree. Regression trees are designed for dependent variables and take ordered discrete or continuous values as study objects, with prediction error usually measured in terms of the squared difference between the predicted and observed values. Classification trees study finitely many unordered dependent values and classify them into different classes with misclassification costs, as shown in Fig. 2. The algorithms include C4.5, CART, chi-squared automatic interaction detector (CHAID) [46], random forest [47], and so on. Since the CHAID algorithm employs an χ^2 test, p-value and Bonferroni adjustment to implement multi-branched trees with high classification accuracy, this paper uses it to recognize electricity outage-sensitive users.

A. CHAID

CHAID [45], [46] was originally designed for classification and then extended to regression, for splitting, which resembles stepwise regression. CHAID can recognize three variable types: ordered with missing values (called floating), ordered without missing values (called monotonic) and categorical. If variable X is floating or monotonic, node N_t is divided into 10 sub-nodes, plus one for missing values, or each sub-node

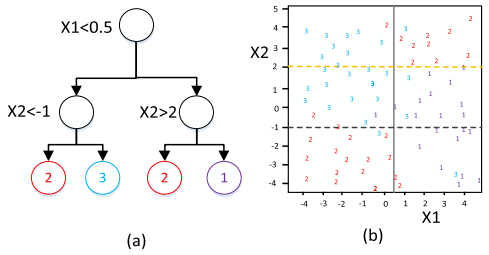


FIGURE 2. Decision tree structure (left) and partition (right) of a classification tree model with three classes, labeled 1, 2 and 3. At each intermediate node, a case goes to the left child node if and only if the condition is satisfied. The perfected class is given beneath each leaf node.

is set to an interval of X values. If X is categorical, N_t is divided into one sub-node for every category of X. Pairs of sub-nodes are then regarded for merging by applying an χ^2 test and Bonferroni-adjusted tests according to the p-value. The merged sub-nodes are then considered for splitting again by virtue of Bonferroni-adjusted tests. Each X variable is evaluated with a Bonferroni adjustment, and the one with the smallest p-value is chosen to be divided. The algorithm can be described as follows.

B. ALGORITHM

Merging. For every predictor variable X, non-significant categories are merged. If X is chosen to be split, each category of X will lead to one sub-node of X. The merging process computes the p-value as well as in the splitting step.

(1) If X only has 1 class, the p-value is set to be 1 and the merging stops.
(2) If X has 2 classes, jump to step (8).
(3) Else, for ordinal variable X, the similar pair being allowed to merge is two adjacent classes. For nominal variable X, the most similar pair is the pair with the largest p-value in relation to the dependent variable Y. The p-value calculation will be presented in subsection C.
(4) If the pair's largest p-value is larger than α_{merge}, a threshold defined in advance, it is merged and a new compound category of X is produced; if it is not, jump to step (7).
(5) (*Optional*) If the new compound category contains more than three categories, then a best binary division, whose p-value is the smallest and not beyond α_{split}, a user defined threshold, is performed within the compound class.
(6) Return to step (2).
(7) (*Optional*) A category with observations below the user-defined minimum size is merged with the most similar class with the largest p-value.
(8) For the merged classes, the p-value is calculated by using Bonferroni adjustments (detailed in subsection D).

Splitting. In the merging process, the adjusted p-value is obtained, and by comparing the p-value against each predictor, the "best" splitting for each variable X is found.

(1) The predictor X with the smallest adjusted p-value is selected.
(2) If the adjusted p-value is less than or equal to α_{split}, the node is split. Else, the node is regarded as an ultimate node.

Stopping. If any of the following stopping rules is satisfied, the growth of the tree should be stopped.

(1) All cases in a node have identical values of the dependent variable.
(2) All cases in a node have identical values for each predictor.
(3) The current tree depth is over the user defined maximum tree depth.
(4) The size of a node is less than the user-specified minimum node size or the number of child nodes is 1.

C. P-VALUE

In the above algorithm, computations of p-values rely on the type of dependent variables. The merging process requires the p-value for a pair category or all categories of X. Suppose that in data D, there are I categories of independent variables X, and J values of dependent variable Y. The p-value calculation is as follows.

1) CONTINUOUS DEPENDENT VARIABLE

For the continuous dependent variable Y, an ANOVA F test [48] is performed to check for different classes of X.Hence, the p-value is derived from the ANOVA F-statistic as

$$F = \frac{\sum_{i=1}^{I}\sum_{n\in D} w_n f_n I(x_n = i)(\bar{y}_i - \bar{y})^2/(I-1)}{\sum_{i=1}^{I}\sum_{n\in D} w_n f_n I(x_n = i)(y_n - \bar{y}_i)^2/(N_f - I)}, \tag{3}$$

$$p = pr(F(I-1, N_f - I) > F) \tag{4}$$

where

$$\bar{y}_i = \frac{\sum_{n\in D} w_n f_n y_n (x_n = i)}{\sum_{n\in D} w_n f_n I(x_n = i)}, \tag{5}$$

$$\bar{y} = \frac{\sum_{n\in D} w_n f_n y_n}{\sum_{n\in D} W_n f_n} \tag{6}$$

and

$$N_f = \sum_{n\in D} f_n. \tag{7}$$

As a random variable, $F(I-1, N_f - 1)$ follows an F-distribution with degrees of freedom I and $N_f - I$.

2) NOMINAL DEPENDENT VARIABLE

For the nominal categorical dependent variable Y, a count table is built by applying classes of X as rows and categories of Y as columns. Under the null hypothesis, the expected cell frequencies are evaluated. The expected and observed cell

frequencies are utilized to compute a likelihood ratio statistic or Pearson chi-squared statistic. The p-value is computed based on any of the frequencies, and they are

$$G^2 = 2\sum_{j=1}^{J}\sum_{i=1}^{I}\ln(n_{ij}/\hat{m}_{ij}) \tag{8}$$

and

$$X^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(n_{ij}/\hat{m}_{ij})}{\hat{m}_{ij}}. \tag{9}$$

For the cell $(x_n = i, y_n = j)$, $n_{ij} = \sum_{n\in D} f_n I(x_n = i \wedge y_n = j)$ is the observed cell frequency, and $\hat{m}_{ij}$ is the expected cell frequency. Without case weights, the expected cell frequencies are

$$\hat{m}_{ij} = \frac{\sum_{j=1}^{J_i} n_{ij} \sum_{i=1}^{I_i} n_{ij}}{\sum_{j=1}^{J_i}\sum_{i=1}^{I_i} n_{ij}}. \tag{10}$$

The p-value is given by

$$p = \begin{cases} Pr(\chi_d^2 > X^2) & Pearson's\ Chi-square\ test \\ pr(\chi_d^2 > G^2) & likelihood\ ratio\ test, \end{cases} \tag{11}$$

where χ_d^2 obeys a chi-squared distribution with degrees of freedom $d = (J-1)(I-1)$.

3) ORDINAL DEPENDENT VARIABLE

For the categorical ordinal dependent variable Y, the null hypothesis of the independence of X and Y is tested against the row effects model, where the rows are the categories of X and the columns are the classes of Y [45]. Under the hypothesis of independence, expected cell frequencies $\hat{m}_{ij}$ are estimated, and under the hypothesis of the data following a row effects model, expected cell frequencies $\hat{\hat{m}}_{ij}$ are estimated. The likelihood ratio statistic and the p-value are

$$H^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{\hat{m}}_{ij}\ln(\hat{\hat{m}}_{ij}/\hat{m}_{ij}) \tag{12}$$

and

$$p = pr(\chi_{I-1}^2 > H^2). \tag{13}$$

D. BONFERRONI ADJUSTMENTS

Suppose that, originally, a variable with I categories is reduced to r categories by the merging step. Bonferroni multiplier B is the number of possible ways in which I categories can be merged into r categories. For $2 \le r \le I$, B can be obtained by using following equation.

$$B = \begin{cases} \binom{I-1}{r-1} & \text{Ordinal predictor} \\ \sum_{v=0}^{r-1}(-1)^v \frac{(r-v)^I}{v!(r-v)!} & \text{Nominal predictor} \\ \binom{I-2}{r-2} + r\binom{I-2}{r-2} & \text{ordinal with a missing category.} \end{cases} \tag{14}$$

TABLE 1. Original tables.

	Dataset	Description	Records	Attributes
1	arc_s_95598_wkst	Records when users call 95598 (China power grid customer service telephone number)	1573046	13
2	comm_rec	Power users' call information	1593088	7
3	s_region_outage	Power outage lines	238072	57
4	s_info_overse	telephone business	15898	10
5	c_cons	Information of power users	1968846	12
6	c_cons_prcc	Power price of different types of users	16574319	3
7	cont_info	Electricity bills of low-income residents	7386	4
8	c_rca_cons	Information about low-income residents	22192	4
9	a_rcved_flow	Paid electricity bills of power users	3249742	9
10	arc_a_rcvbl	Electricity bill receivable	6466654	9
11	c_meter_read	Information about users' electricity meter	40821270	2
12	c_meter	Running power meter	25807825	6
13	a_pay_flow	Tables about power bills	5483788	5
14	c_tags	Power users with labels	430326	2
	Total		104249759	143

V. DATASET, PRE-PROCESS AND EVALUATION INDEXES

A. DATASET AND PRE-PROCESS

The desensitization datasets come from a developed province of China during the entirety of 2015, includes 14 tables, 104,249,759 records, a total of 1,968,846 users (where 430,326 users have labels) and 143 attributes, as shown in Table 1.

The whole data process is shown in Fig. 3, including data cleaning and filtering, association analysis, data fusion, statistical analysis, classification and comparison. Datasets 1, 2 and 3 are utilized to perform statistical analysis of the distribution of electricity customers. By applying union and intersection operations and association analysis on 14 tables, some complex text features, repeated data and features that are irrelevant to power outage sensitivity are filtered and deleted. By combing and intersecting multiple tables, the dataset with labels that are sensitive or insensitive to

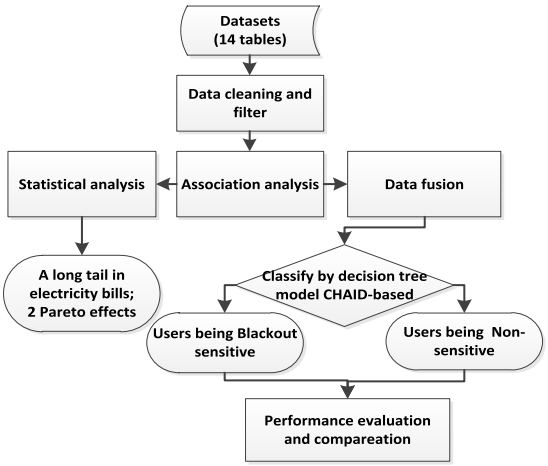


FIGURE 3. Data process.

TABLE 2. Features.

	Variables	Description
1	cons_no	Coding of the users when they pay the electricity bills (used to associate the tables; comes from table a_rcved_flow)
2	cust_no	Coding of the power users, generated automatically come from table c_cons
3	cons_id	User ID. Comes from table c_cons (used to associate the tables)
4	app_no	Coding of a work order generated automatically by calling 95598. Comes from the arc_s_95598_wkst table
5	Id	ID of a work order accepted by the 95598 service. Comes from table arc_s_95598_wkst
6	calling_no	Phone number of a calling user
7	trade_code	Type of business
8	cons_sort_code	Type of consumer
9	elec_type	Electricity type
10	contract_cap	Capacity of the contract
11	status_code	Status of the user (0- used, 1-unused)
12	lode_attr_code	Code of power load attribute
13	urban_rural_flag	Flags indicating urban or rural users
14	busi_type_code	Type of business accepted by calling 95598, including: request repair, apply for new services, consulting and so on
15	wkst_busi_type_code	Business type of work order after issued
16	org_no	Coding of the local power station
17	prov_org_no	Coding of the provincial power supply company organization in the province
18	city_org_no	Coding of the power supply company of a city
19	req_begin_date	Begin time of a call requesting for a service (including repair, applying for a new business, restoring electricity and so on)
20	req_finish_date	Time the service is completed at
21	handle_time	The whole processing time of a work order
22	Tag	0-sensitive，1-insensitive
23	accept_content	A detailed description of a work order accepted by 95598

a power outage becomes 430,326 records with 23 features, as shown in Table 2. 70% of the data with label 0 is randomly selected as the training set, and 30% as the test set, in which 3.02% of the users of the data are blackout sensitive, as Table 3 shows.

Meanwhile, different attributes and orders of magnitude of the features will result in inaccuracy. The traditional standardizations consist of the maximum-minimum value

TABLE 3. Dataset sensitivity and insensitivity to power outages.

	Percentage	Number of users	Tags and percentages (%) 0-sensitive	1-non-sensitive
Train dataset	70%	409490	0(1)	96.97(3.02)
Test dataset	30%	175497	0(1)	96.97(3.02)

standardization and Z-score standardization. This paper applies a Z-score to standardize the dataset:

$$x_i^o = \frac{x_i - \mu}{\delta}(i = 1, \ldots, k), \tag{15}$$

where $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$ and $\delta^2 = \frac{\sum(x-\mu)^2}{N}$.

B. EVALUATION INDEX

To accurately evaluate the performance of a decision tree, the usual evaluation indexes, described below, are used.

True positive (TP). The users are sensitive to power failure.

True negative (TN). The users are judged to be insensitive to a power failure.

False positive (FP). The users are judged to be sensitive to power failure but in fact are insensitive. The false positive rate **(FPR)**is the probability of an FP.

$$FPR = \frac{FP}{FP + TN}. \tag{16}$$

False negative (FN): The users are judged to be insensitive to a power failure but the actual are sensitive. The false negative rate **(FNR)** is the probability of the FN.

Precision (P): The detected proportion of customers who really are sensitive to power outages.

$$P = \frac{TP}{TP + FP}. \tag{17}$$

Recall (R, probability of detection): The probability of correctly detecting the proportion of customers who are insensitive to power failure.

$$R = TPR = \frac{TP}{TP + FN} \tag{18}$$

and

$$FNR = 1 - R. \tag{19}$$

F1 is the harmonic mean of precision and recall.

$$F1 = \frac{2^*P^*R}{P + R} \tag{20}$$

Accuracy (A): The proportion of customers who are sensitive to power blackouts and are recognized as sensitive users and the customers who are insensitive to power blackouts and are recognized as insensitive users.

$$A = \frac{TP + TN}{TP + FN + FP + TN}. \tag{21}$$

Fundamentally, P and FPR reflect the recognition precision and detection errors. R and FNR reflect the recognition

TABLE 4. Electricity bill grade, Number of users and electricity bills.

	Electricity Bill Grade (RMB)	Number of Users	Electricity Bill (RMB)	Percentage of Users（%）	Cumulative Percentage of Users (%)	Percentage of the Electricity Bill（%）	Cumulative Percent of the Electricity Bills (%)
0	-7799273-0(Arrearage)	2631	NULL	NULL	0.17	NULL	
1	0-50(25)	429802	8596461	21.8	21.80	0.04	0.04
2	50-100(75)	484958	29098916	24.6	46.40	0.14	0.18
3	100-150(125)	282627	28264125	14.4	60.80	0.13	0.31
4	150-200(175)	164949	23094037	8.37	69.17	0.11	0.42
5	200-250(225)	99049	17829619	5.03	74.20	0.08	0.5
6	250-300(275)	59377	13063485	3.01	77.21	0.06	0.56
7	300-400(325)	70598	19768332	3.59	**80.80**	0.10	0.66
8	400-500(450)	47915	17250141	2.43	83.23	0.08	0.74
9	500-1K(750)	94685	56813986	4.81	88.04	0.28	1.02
10	1K-2K(1.5K)	52085	62504677	2.64	**90.68**	0.30	1.32
11	2K-5K(3.5K)	49263	137942830	2.51	93.19	0.67	1.99
12	5K-10K(7.5K)	29979	179883463	1.52	94.71	0.87	2.86
13	10K-20K(15K)	28558	342716462	1.45	96.16	1.66	4.52
14	20K-50K(35K)	34621	969432021	1.76	97.92	4.70	9.22
15	50K-100K(75K)	16432	985966240	0.84	98.76	4.77	13.99
16	100K-1M(500K)	19390	2326916344	0.98	99.74	11.3	25.29
17	1M-37.7M(above 1M)	1927	15415289525	0.09	99.83	**74.7**	100
	Total (without Arrearage)	1968846	20634430665	1	100 (arrearage added to payments)	1	

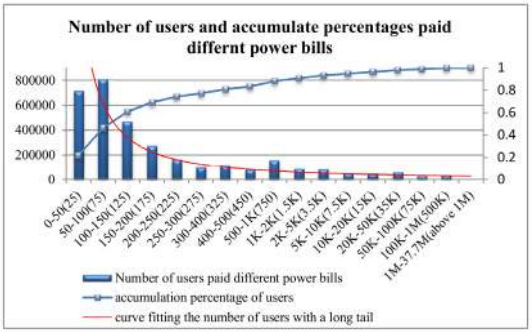


FIGURE 4. User distribution vs. their electricity bills annually.

coverage and missing rate. F1 is the balance of the P and R values. Usually, a big TN will lead to the accuracy (A) being far greater than precision (P). Therefore, P, R and FPR can response to the overall performance of a model.

ROC illustrates the diagnostic ability of a classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against FPR at various threshold settings.

AUC is the area under the ROC curve. In general, AUC is between 0.5 and 1, and the higher AUC is, the better the differentiation ability of the model is.

VI. CASE ANALYSIS

A. LONG TAIL

According to the amount of electricity consumed in 2015, a more in-depth hierarchical discussion of electricity customers is conducted. Since there was a total of 3,275,635 users (without 4383 arrearages) contribute RMB 42,968,069,025 (from RMB 0-32,100,000,000), their ranges are too big to be displayed clearly. Consequently, we divide them into different grades to generate statistics and exhibit them in Table 6, in which, 3,275,635 users pay an electricity fee between RMB 0 and 50 yearly, 807,921 users are in the range of RMB 50-100, and the amount of electricity paid is only 0.04% and 0.14%. An accumulated 80% of users pay less than 1% of the electricity bills. On the contrary, users who paid RMB 100K-37.7M per year account for 0.98% and 0.09%, but the accumulated amount paid is beyond 85% power bills.

According to Table 4, Fig. 4 displays the users' distribution. Overall, if the interval inequality of the electricity bill is ignored, a long tail of the number of customers versus electricity bills can be sketched out (the red line in Fig. 3). It can be seen that the power users pay a wide range of electricity bill amounts, and most users (approximately 90%) pay electricity bills less than RMB 2,000 per year, in which 80% of users pay below RMB 400 per-year. Since a too-long tail makes it difficult to observe the distribution of users against their electricity bills, to clearly observe the distribution, we extract 1,865,287 users, of which the electricity bills are in the range of RMB 0-10,000, to fit a curve, as shown in Fig. 5. It is found that there is a long tail following a power law distribution with the probability density function $y = 4^{*}10^{8}x^{-1531}$. Due to goodness of fit $R^2 = 0.9808$ and $R^2 \rightarrow 1$, the fitting is reliable.

Furthermore, to make clear the contributions of different industries to electricity bills, according to the voltage levels and types of users [7], power users are divided into four classes, as Table 5 shows. The transverse one is among similar

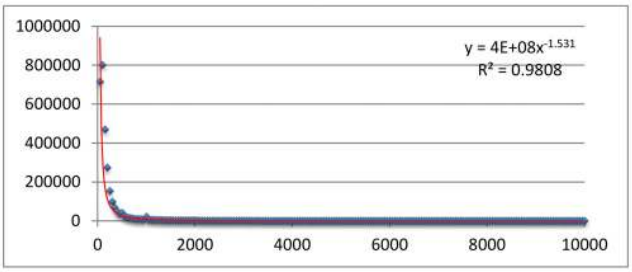


FIGURE 5. A long tail following a power law existing in the number of power users who pay electricity bills in the range of RMB 0-10,000.

TABLE 5. Original types of electricity users.

Type	Number of users	Percentage (%)
high-voltage	100518	5.10
general industry and commerce	216111	10.90
low-voltage resident	1652172	83.90
Others	45	0.002

TABLE 6. Payments of residents and non-residents.

Types	Number of users	Percentage (%)	Total electricity bill (RMB)	Percentage of electricity (%)
Residents	1999224	82.36%	501332316	2.43
non-residents	428020	17.63%	20133098336	97.57

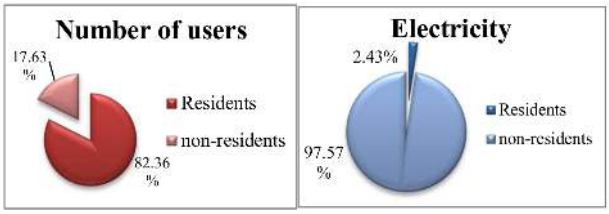


FIGURE 6. Electricity bill payments of residents and non-residents.

users of different voltage levels and the longitudinal one is between different industrial users under the same voltage level.

Low-voltage residential electricity accounts for up to 83.90% of electricity usage, while the percentage of high voltage electricity usage is only 5.1%. In addition, Table 6 shows the electricity consumption and payment of electricity, divided by residents and non-residents.

According to the statistics of Table 6 and Fig. 6, the total number of users is 1,968,846, and the number of residents accounts for 82.36% of the total users, contributing over 2.43% of electricity fees. On the contrary, non-residents account for about 17.63% of the total, paying 97.57% of electricity bills. This observation is partly in line with the extended Pareto effects, namely 20% of important customers make more than 80% of the contributions.

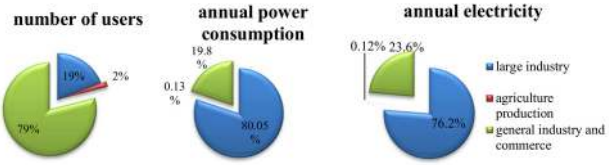


FIGURE 7. Proportion of non-resident electricity customers, electricity consumption and electricity bills.

The electricity consumption of non-residents mainly includes large industrial electricity, the farming industry, and general industrial and commercial electricity, as shown in Fig. 7. In the non-residents area, the power consumption of farming only accounts for 1.67% of electricity customers, 0.13% of the electricity consumption, and 0.12% of the total paid annually. From Fig. 7, it can be seen that the Pareto effect also exists in general industrial, commercial electricity consumption and big industrial electricity consumers. 20% of non-residential electricity customers in large industrial areas contribute 80% of the profiles.

The long tail theory is an improvement on and perfection of the Pareto effect [36]. Both the Pareto effect and the long-tail distribution reflect that, in developed areas of China, low-voltage residential electricity accounts for 80% of electricity usage, which results in a higher electrical loss for the power supply company; the less non-residential users and industrial electricity have become the dominant electricity consumers. Electricity sale companies should make different sales strategies for residents and non-residents to gain higher economic benefits, and ensure power supply reliability and quality for non-residents. Aside from the economic considerations, additional factors should also be taken into account. The next section will continue to analyze other factors and recognize users who are sensitive to power blackouts.

B. BLACKOUT SENSITIVE

1) PERFORMANCE

After pre-processing, the power outage-sensitive dataset includes 23 properties (Section V, Table 3) and a total of 430,326 users with labels. Let the $p = 0.02$. This paper trains and tests the decision tree model using CHAID. As discussed in section IV, we firstly use the precision (P), recall (R) and FPR to evaluate the performance of this model. 70% of the data is randomly selected as the training set, and 30% as the test set. Too many branches and layers may result in a long time delay and over-fitting, while less branches and layers will lead to a low performance. Therefore, proper parameters should be taken into account. The model training and test are carried out under different branches and layers, and the results of the test are shown in Table 7, as well as in Figs. 8, 9,10 and 11.

As shown in Figs. 8 and 9, overall, for a fixed branch (or layer), P and R grow with the number of layer (or branch). P quickly reaches 90%, while R increases slowly. In light of R, we divide Table 7 into three parts: R <= 50% in the blue dashed frame, R > 80% in the red dashed frame and

TABLE 7. P, R and FPR changes with layers and branches.

Layer	Branch	2(%)	3(%)	4(%)	5(%)	6(%)	7(%)	8(%)	9(%)
2	P	0.00	97.79	99.47	89.38	91.11	93.04	96.43	96.07
	R	0.00	17.77	19.29	32.40	34.74	36.98	42.90	45.83
	FPR	0.00	0.01	0.00	0.06	0.05	0.04	0.03	0.03
3	P	0.00	95.34	95.93	92.16	95.10	95.50	96.05	96.32
	R	14.6	22.17	35.94	52.40	56.39	62.00	67.64	72.35
	FPR	0.13	0.02	0.02	0.07	0.05	0.05	0.05	0.04
4	P	0.00	96.57	93.13	95.16	91.87	94.08	94.53	94.45
	R	15.3	24.78	49.42	59.17	70.43	78.05	84.15	88.85
	FPR	0.11	0.01	0.06	0.05	0.10	0.08	0.08	0.08
5	P	0.00	90.51	94.36	95.07	95.72	96.26	94.83	**95.15**
	R	28.7	43.66	55.46	67.46	75.03	81.71	87.96	**90.81**
	FPR	0.06	0.07	0.05	0.05	0.05	0.05	0.07	**0.07**
6	P	0.00	91.30	93.57	95.10	95.46	**94.79**	**94.46**	94.21
	R	36.1	54.64	64.13	75.76	83.68	**90.25**	**92.68**	93.30
	FPR	0.07	0.08	0.07	0.06	0.06	**0.08**	**0.09**	0.09
7	P	0.00	90.20	94.01	95.41	**94.39**	94.93	94.49	94.35
	R	36.4	57.68	67.88	80.11	**87.60**	92.01	93.11	93.37
	FPR	0.07	0.09	0.07	0.06	**0.09**	0.08	0.09	0.09
8	P	0.04	90.37	94.28	94.81	**94.73**	94.34	94.02	94.03
	R	42.7	61.94	74.04	85.42	**91.42**	93.06	93.28	93.30
	FPR	0.09	0.11	0.07	0.08	**0.08**	0.09	0.10	0.10
9	P	0.04	90.09	93.93	**94.88**	94.80	94.40	94.17	94.19
	R	40.0	64.42	93.12	**87.49**	91.52	93.03	93.09	93.27
	FPR	0.10	0.11	0.09	**0.08**	0.08	0.09	0.09	0.10

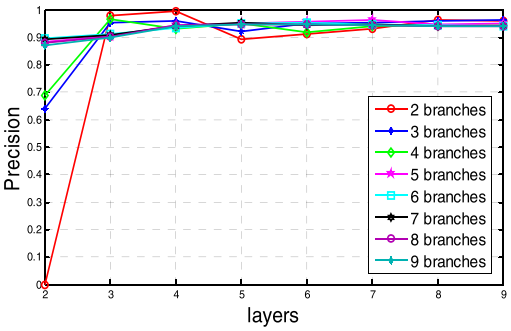


FIGURE 8. Precision versus layers under different branches.

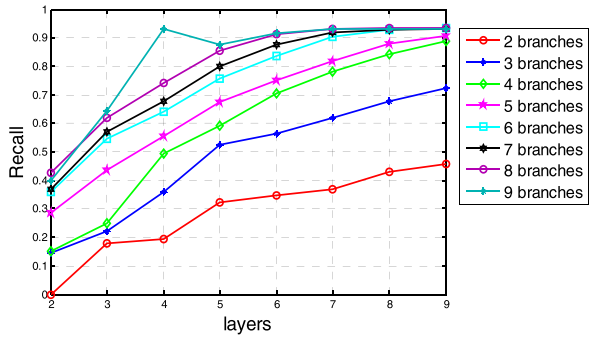


FIGURE 9. Recall vs. layers under different branches.

between them is where R is equal to other values. The blue dashed frame part in Table 7 and Figs. 9 and 10 shows that the recognition precisions and recalls of the model are more sensitive to branches than layers, and a binary tree is unsuitable for recognizing blackout-sensitive users. When branch $= 2$, the DT evolves into a binary tree (the yellow part in Table 7), whose recall rates are low. That means identification of power

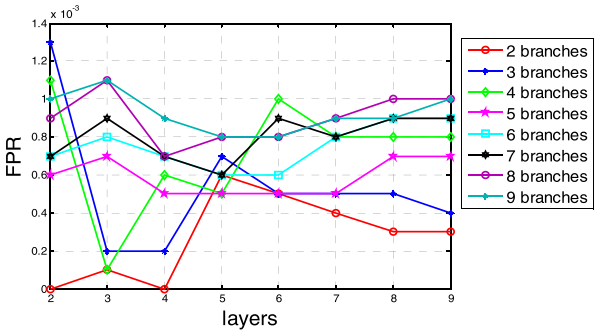


FIGURE 10. FPR under different layers and branches.

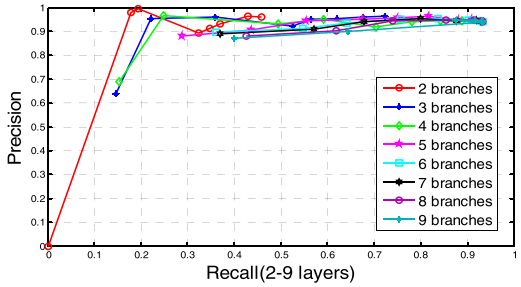


FIGURE 11. Precision vs. recall under different branches.

outage-sensitive users is a multi-class problem rather than binary classification.

Fig. 10 demonstrates that the FPRs change with the layers under different branches. Although the FPRs fluctuate under all parameter settings, the FPRs are very low (10^{-3}). Fig. 11 illustrates that recall increases with the branches and layers and will not result in the decline of P. From the different colored parts in Table 7 and Figs. 8, 9,10 and 11, it can be seen that only when there is an increase in both the number of branches and layers will a good performance be achieved. Moreover, to more clearly observe the changes in performance, we calculate the left three indexes and demonstrate all of them in three dimensional graphs, as shown in Fig. 12.

The recognition P of the model increases rapidly with the braches and layers, and is always relatively higher than R. From the different colors in Table 7 and R in Fig. 12, it can be seen, in the case of branch $>= 7$ and layer $>= 7$, that R is over 90% and FNR decreases to less than 10%. Because of the high detection P, the accuracy A is also very high. In terms of formula (20), the value and change trend of F1 is more similar to that of R.

2) DISCUSSION

Since power failures will cause various influences in the economy and society. It will affect many aspects of people, such as work, life, study, travel and so on, and result in different losses for the enterprise or institution. Due to various factors, people's sensitivity degrees to power failures are different. As Figs 8 to 11 shown, with the branches and layers growing, these factors are more taken into consideration,

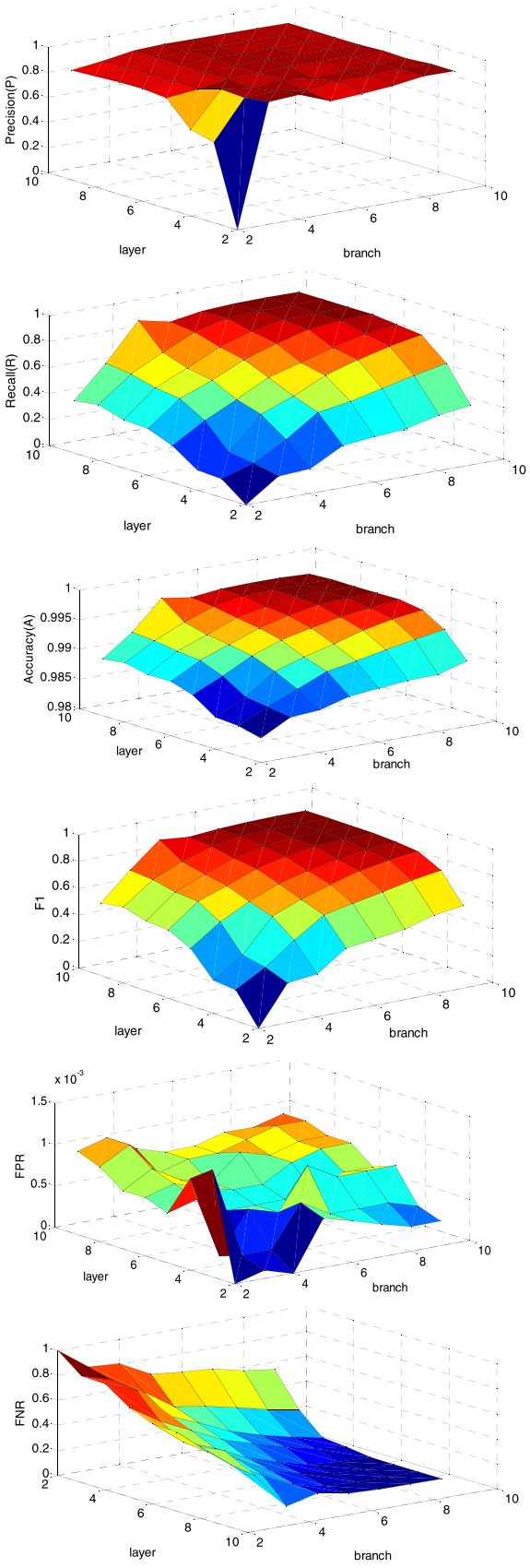


FIGURE 12. P, R, A, F1, FPR and FNR changing with layers and branches on the test dataset.

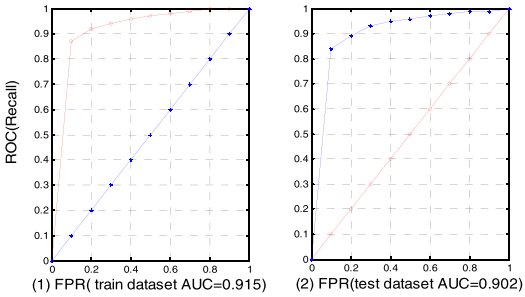


FIGURE 13. The ROC curve and AUC with branch = 7 and layer = 6.

TABLE 8. Percentage of users who are blackout-sensitive in different industry types.

ELEC_TYPE	Percentage of users who are blackout-sensitive (%)
high-voltage	2.34
general industry and commerce	3.53
low-voltage resident	3.78
Others	14.29

which improves the performance rapidly. Meanwhile, once these factors have been considered by the decision tree, the performance becomes stable, even, more branches and layers will cause over-fitting and reduce the robustness. Therefore, branches and layers should be selected carefully.

As discussed in subsections A, since the industry type determines the economic benefits, as well as economic losses, caused by power failures, economic factors are important for recognizing blackout-sensitive users. Further analysis finds that the users who are power failure sensitive are distributed in different industries (elec_type), and the proportion in each industry has no significant differences, as Table 8 shows. Aside from economic factors (reflected by elect_type, lode_attr_code and contract_cap), interruption duration and time, the type of business accepted by 95598 (busi_type_code), calling (calling numbers and time), the flag used to indicate urban vs rural users (urban_rural_flag) and the handle time (handle_time) are also key features. During a blackout, the power sensitive crowd consisting of both residents and non-residents will repeatedly dial 95598 to urge them to restore power. The main reasons may be that blackouts result in living or work inconveniences, economic losses, and more.

3) PARAMETERS AND ROC

Considering the trade-off between performance and cost, reasonable parameters are selected, such as the gray part in Table 7 (branch = 9 and layer = 5, branch = 8 and layer = 6, branch = 7 and layer = 6 and branch = 6 and layer = 7). Given branch = 7 and layer = 6, Fig. 13 illustrates the ROC

TABLE 9. Comparison with other methods.

	Decision Tree (6,7)	SVM 40 Iterations	Logical Regression
FN	853	2790	6799
TN	421141	3301	423413
FP	434	418501	47
TP	7898	5734	57
A	99.77%	**2.10%**	98.40%
P	94.79%	**1.35%**	54.62%
R	90.25%	67.27%	**0.83**%
F1	92.47%	26.51%	**1.63%**
FPR	0.08%	**99.20**%	0.01%
FNR	9.74%	31.73%	**99.17**%

curve and the AUC. The smooth curves of the ROC ensure that there is no over-fitting. Meanwhile, the AUC is equal to 0.915 on the training dataset and 0.902 on the test dataset, which means the model has a good ability to discriminate between the data.

4) COMPARISON

With reasonable parameter settings, we compare the results of the decision tree with those of the SVM and logistic regression model. The results are shown in Table 9. It can be seen that under reasonable parameters, the decision tree model has a good identification precision and recall rate over the SVM and logic regression model, plus a lower FPR and FNR, which is more suitable for detecting power outage sensitive users.

VII. CONCLUSION AND FUTURE WORK

With the further liberalization of the electricity market of China, the customers' requirements, characteristics and distribution, as well as the quality, security and reliability of power supplies without interruption, have received considerable attention from power companies, policymakers and researchers. Meanwhile, a large number of smart devices of Internet of things are used in smart power grids, and they provide a huge data support for further understanding the characteristics of Chinese power users. This article adopts statistical analysis and data mining to analyze the users distribution and their attitudes toward power security and reliability–that is, their sensitivity to power outages. The study found that in Chinese power market, The number of users who paid different electricity fees from arrears to tens of millions, and there are tow Pareto effects and a long tail distribution in number of users against power bills. In addition, a decision tree model is used to classify customers and find customers who are sensitive to power failures. The experiments show that the decision tree can accurately capture the characteristics of the crowd who is sensitive to power outages, and makes reasonable classifications with good identification performances over the SVM and logistic regression model.

These findings suggest that electricity fees power users paid annually distribute widely, and users in the large industry field are very import customers of the electricity companies. The power company should ensure the reliability and safety of the power supply to maximize its own benefits. However, since the crowd who is sensitive to power failures distributes in all walks of life, and the proportion in each industry has no significant difference. The factors, such as industry types, economic losses, interruption duration, time, location (urban or rural area) and so on, will effect the customer's sensitivity degree to power outages. For power suppliers and policy makers, in addition to economic factors, positive social effects should also be considered. Companies that sell electricity may construct more reasonable dispatch and repair schemes and power blackout plans during the power consumption peak. Meanwhile, various marketing strategies to satisfy different requirements of power users should be provided to promote suppliers' long-term attractiveness to power customers, such as time-sharing electricity price and peak-valley electricity price.

In the next step, we will further quantitatively analyze the influences of various factors on power failure sensitivity, and further deepen the algorithm based on decision tree, such as random forest and bagged tree.

ACKNOWLEDGMENT

The authors would like to thank State Grid Corporation of China provides them the data.

REFERENCES

[1] *Several Opinions of the Central Committee of the Communist Party of China and the State Council on Further Deepening the Reform of the Power System*, Tech. Communist Party China (CPC) Central Committee and State Council, Beijing, China, 2015.

[2] *The Analysis of the National Power Supply and Demand Situation in the First Quarter of 2015 and the Forecast Report for the Last Three Quarters*, CE Council, Beijing, China, 2015.

[3] H. W. Ngan, "Electricity regulation and electricity market reforms in China," *Energy Policy*, vol. 38, no. 5, pp. 2142–2148, May 2010.

[4] Q. Wang and X. Chen, "China's electricity market-oriented reform: From an absolute to a relative monopoly," *Energy Policy*, vol. 51, pp. 143–148, Dec. 2012.

[5] C. Hu *et al.*, "In the context of the new electricity reform, the purchase and sale approaches and the business model of electric companies in China," *Power Grid Technol.*, vol. 40, no. 11, pp. 3293–3299, 2016.

[6] W. Li, P. Xu, X. Lu, H. Wang, and Z. Pang, "Electricity demand response in China: Status, feasible market schemes and pilots," *Energy*, vol. 114, pp. 981–994, Nov. 2016.

[7] C. Wang, K. Zhou, and S. Yang, "A review of residential tiered electricity pricing in China," *Renew. Sustain. Energy Rev.*, vol. 79, pp. 533–543, Nov. 2017.

[8] P. J. Gertler, A. Kenneth, and L. M. Mobarak, "Electricity reliability and economic development in cities: A microeconomic perspective," eScholarship Univ. California, Berkeley, CA, USA, Tech. Rep., 2017.

[9] Y. Wang, C. Chen, J. Wang, and R. Baldick, "Research on resilience of power systems under natural disasters—A review," *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 1604–1613, Mar. 2016.

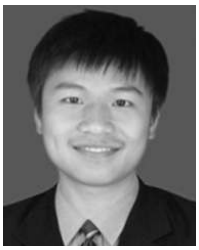
[10] G. Gill, "Major blackout events on each side of the Atlantic," *Loss Prevention Bull.*, vol. 242, pp. 3–8, Apr. 2015.

[11] M. Shuai, W. Chengzhi, Y. Shiwen, G. Hao, Y. Jufang, and H. Hui, "Review on economic loss assessment of power outages," *Procedia Comput. Sci.*, vol. 130, pp. 1158–1163, May 2018.

[12] T. Zachariadis and A. Poullikkas, "The costs of power outages: A case study from cyprus," *Energy Policy*, vol. 51, pp. 630–641, Dec. 2012.

[13] J. E. Payne, "A survey of the electricity consumption-growth literature," *Appl. Energy*, vol. 87, no. 3, pp. 723–731, Mar. 2010.
[14] C. Tu, X. He, Z. Shuai, and F. Jiang, "Big data issues in smart grid—A review," *Renew. Sustain. Energy Rev.*, vol. 79, pp. 1099–1107, Nov. 2017.
[15] R. J. Bessa, "Chapter 10—Future trends for big data application in power systems," in *Big Data Application in Power Systems*. New York, NY, USA: Elsevier, 2018, pp. 223–242.
[16] J. Morais, Y. Pires, C. Cardoso, and A. Klautau, "An overview of data mining techniques applied to power systems," in *Data Mining and Knowledge Discovery in Real Life Applications*. London, U.K.: InTech, 2009.
[17] L. Morales and J. Hanly, "European power markets—A journey towards efficiency," *Energy Policy*, vol. 116, pp. 78–85, May 2018.
[18] W. Zhang, L. Tian, M. Wang, Z. Zhen, and G. Fang, "The evolution model of electricity market on the stable development in China and its dynamic analysis," *Energy*, vol. 114, pp. 344–359, Nov. 2016.
[19] A. Shiu and P.-L. Lam, "Electricity consumption and economic growth in China," *Energy Policy*, vol. 32, no. 1, pp. 47–54, Jan. 2004.
[20] C. Zhang, K. Zhou, S. Yang, and Z. Shao, "On electricity consumption and economic growth in China," *Renew. Sustain. Energy Rev.*, vol. 76, pp. 353–368, Sep. 2017.
[21] S. K. Al-Bajjali and A. Y. Shamayleh, "Estimating the determinants of electricity consumption in Jordan," *Energy*, vol. 147, pp. 1311–1320, Mar. 2018.
[22] S. Bedingfield, D. Alahakoon, H. Genegedera, and N. Chilamkurti, "Multi-granular electricity consumer load profiling for smart homes using a scalable big data algorithm," *Sustain. Cities Soc.*, vol. 40, pp. 611–624, Jul. 2018.
[23] R. R. Rathod and R. D. Garg, "Regional electricity consumption analysis for consumers using data mining techniques and consumer meter reading data," *Int. J. Elect. Power Energy Syst.*, vol. 78, pp. 368–374, Jun. 2016.
[24] D. De Silva, X. Yu, D. Alahakoon, and G. Holmes, "A data mining framework for electricity consumption analysis from meter data," *IEEE Trans. Ind. Informat.*, vol. 7, no. 3, pp. 399–407, Aug. 2011.
[25] S. Ramos, J. M. Duarte, F. J. Duarte, and Z. Vale, "A data-mining-based methodology to support MV electricity customers' characterization," *Energy Buildings*, vol. 91, pp. 16–25, Mar. 2015.
[26] D. Huang, H. Zareipour, W. D. Rosehart, and N. Amjady, "Data mining for electricity price classification and the application to demand-side management," *IEEE Trans. Smart Grid*, vol. 3, no. 2, pp. 808–817, Jun. 2012.
[27] A. Castillo, "Risk analysis and management in power outage and restoration: A literature survey," *Electr. Power Syst. Res.*, vol. 107, pp. 9–15, Feb. 2014.
[28] S. Mukherjee, R. Nateghi, and M. Hastak, "A multi-hazard approach to assess severe weather-induced major power outage risks in the U.S.," *Rel. Eng. Syst. Saf.*, vol. 175, pp. 283–305, Jul. 2018.
[29] R.Diao *et al.*, "Decision tree-based online voltage security assessment using PMU measurements," *IEEE Trans. Power Syst.*, vol. 24, no. 2, pp. 832–839, May 2009.
[30] J. A. Morales, Z. Anane, and R. J. Cabral, "Automatic lightning stroke location on transmission lines using data mining and synchronized initial travelling," *Electr. Power Syst. Res.*, vol. 163, pp. 547–558, Oct. 2018.
[31] R. Han and Q. Zhou, "Data-driven solutions for power system fault analysis and novelty detection," in *Proc. 11th Int. Conf. Comput. Sci. Educ. (ICCSE)*, Nagoya, Japan, Aug. 2016, pp. 86–91.
[32] L. Xu, M.-Y. Chow, and L. S. Taylor, "Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification *E*-algorithm," *IEEE Trans. Power Syst.*, vol. 22, no. 1, pp. 164–171, Feb. 2007.
[33] R. Dubey, S. R. Samantaray, B. K. Panigrahi, and V. G. Venkoparao, "Data-mining model based adaptive protection scheme to enhance distance relay performance during power swing," *Int. J. Elect. Power Energy Syst.*, vol. 81, pp. 361–370, Oct. 2016.
[34] S. Kamali and T. Amraee, "Blackout prediction in interconnected electric energy systems considering generation re-dispatch and energy curtailment," *Appl. Energy*, vol. 187, pp. 50–61, Feb. 2017.
[35] Y. Jia, Z. Xu, L. L. Lai, and K. P. Wong, "Risk-based power system security analysis considering cascading outages," *IEEE Trans. Ind. Inform.*, vol. 12, no. 2, pp. 872–882, Apr. 2016.
[36] M. E. J. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemp. Phys.*, vol. 46, no. 5, pp. 323–351, Dec. 2005.
[37] M. A. Hart, "The long tail: Why the future of business is selling less of more by Chris Anderson," *J. Product Innov. Manage.*, vol. 24, no. 3, pp. 274–276, Apr. 2007.
[38] G. E. P. Box and R. D. Meyer, "An analysis for unreplicated fractional factorials," *Technometrics*, vol. 28, no. 1, pp. 11–18, Feb. 1986.
[39] P. Marshall, *80/20 Sales and Marketing: The Definitive Guide to Working Less and Making More*. Irvine, CA, USA: Entrepreneur Press, 2013.
[40] C. Anagnostopoulos, D. K. Tasoulis, N. M. Adams, N. G. Pavlidis, and D. J. Hand, "Online linear and quadratic discriminant analysis with adaptive forgetting for streaming classification," *Stat. Anal. Data Mining*, vol. 5, no. 2, pp. 139–166, Apr. 2012.
[41] S. A. Kalogirou, "Applications of artificial neural-networks for energy systems," *Appl. Energy*, vol. 67. nos. 1–2, pp. 17–35, Sep. 2000.
[42] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and A. M. Mohammad, "Detection of abnormalities and electricity theft using genetic support vector machines," in *Proc. EEE Region 10 Conf.*, Nov. 2008, pp. 1–6.
[43] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Found. Trends Comput. Graph. Vis.*, vol. 7, nos. 2–3, pp. 81–227, Feb. 2012.
[44] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," *Energy*, vol. 32, no. 9, pp. 1761–1768, Sep. 2007.
[45] M. van Diepen and P. H. Franses, "Evaluating chi-squared automatic interaction detection," *Inf. Syst.*, vol. 31, no. 8, pp. 814–831, Dec. 2006.
[46] *CHAID Algorithm*. Accessed: Feb. 2019. [Online]. Available: http://www.statsoft.com/Textbook/CHAID-Analysis
[47] S. Xuan, G. Liu, and Z. Li, "Refined weighted random forest and its application to credit card fraud detection," in *Proc. Int. Conf. Comput. Social Netw.*, Nov. 2018, pp. 343–355.
[48] R. G. Lomax and D. L. Hahs-Vaughn, "Statistical concepts: A second course," in *Statistical Concepts: A Second Course*, R. G. Lomax, Ed. London, U.K.: Routledge, 2007.

CHUNYAN SHUAI received the Ph.D. degree in computer science from Tongji University, China, in 2013. She is currently an Associate Professor with the Faculty of Transportation Engineering, Kunming University of Science and Technology, Kunming, China. Her research interests include data retrieval, machine learning, streaming data processing, multidimensional indexing, query optimization, and network security.



HENGCHENG YANG is currently pursuing the master's degree with the Faculty of Electric Power Engineering, Kunming University of Science and Technology, Kunming, China. His current research interests include data retrieval and optimization in high-dimensional spaces.

XIN OUYANG received the Ph.D. degree in computer science from the Kunming University of Science and Technology, China, in 2013, where he is currently a Lecturer with the Faculty of Information Engineering and Automation. His research interests include big data retrieval, streaming data processing, multi-dimensional indexing, and pattern recognition.

MINGWEI HE received the B.S. degree in traffic engineering and the M.S. degree in vehicle operation engineering from the Kunming University of Science and Technology, in 2004 and 2007, respectively, and the Ph.D. degree in traffic system engineering from the Dalian University of Technology, in 2017. He is currently a Lecturer with the Faculty of Transportation Engineering, Kunming University of Science and Technology. His research interests include travel decision behavior, transportation geography, and transportation information.

ZEWEIYI GONG is currently an Engineer with the Yunnan Electric Power Research Institute, Yunnan Power Grid Co., Ltd., Kunming, China. His current research interests include data retrieval and optimization in high-dimensional spaces.

WANNENG SHU (M'81) received the B.E. degree in computer science and technology from Jianghan University, Wuhan, China, in 2003, the M.E. degree in computer application technology from Central China Normal University, Wuhan, in 2007, and the Ph.D. degree in computer software and theory from Wuhan University, Wuhan, in 2013.

He is currently an Associate Professor with the South-Central University for Nationalities, Wuhan. His research interests include the wireless communication, smart city, green cloud computing, machine learning, and intelligent computing.