Fuzzy Systems and Data Mining VI
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA200690

An Improved Decision Tree Algorithm for Electricity Theft Prediction and Analysis

Chen Xiaofeng, Zhang Lipeng[1], Xu Zhongping, Zhu Feng, Qi Xiaoming
State Grid Information & Telecommunication Accenture Information Technology Co., Ltd, Beijing, 100032, China

Abstract. At present, there are many kinds of electricity theft and the corresponding approaches to combat this are insufficient. Manual approaches result in a heavy staff workload and are inefficient. In this paper, the data from an electricity information acquisition system is collected and mined using Python. Based on an understanding of the business and an analysis of the information value (IV) measure, important characteristic indexes are selected and an improved decision tree algorithm is used to construct a model of power theft by users. This method effectively narrows the range of users suspected of power theft, improving the pertinence of audit, and providing strong support for reducing the financial losses of power supply enterprises and ensuring the safety of power grid operation.

Keywords. Electricity Information Acquisition System, Python Platform, Improved Decision Tree Algorithm, Analysis Model of Power Theft, Strong Support

1. Introduction

The theft of electric power seriously disturbs the normal operation of power supply, which not only causes huge economic loss to a country, but also affects the economic benefit of power supply enterprises[1-2]. The traditional detection of power theft mainly relies on manual operation, which not only requires a large amount of human resources and increases the operating cost of power grid companies, but also has low detection efficiency, lagging behind the occurrence of power theft, and has evidence is difficult to obtain[3].

With the development of data mining and other technologies, many anti-theft analysis and prediction methods have emerged[4-6]. Cheng[7] proposed a multi-dimensional characteristic factor correlation model based on the k-means clustering algorithm and parameters taken from the electricity information acquisition system, so as to identify suspected power theft users. Based on the LeNet-5 convolutional neural network model, Zheng[8] modeled and analyzed daily electricity consumption data, selected the abnormal electricity consumption mode, and then used a double-layer deep network to comprehensively analyze user information, line loss in the station area, alarm information, and other data, laying a foundation for the realization of accurate power theft detection. Kangningning[9] adopted FCM clustering

[1] Corresponding Author: Zhang Lipeng, State Grid Information & Telecommunication Accenture Information Technology Co., Ltd, Email: Philip_Zhang1@163.com

and an improved SVR model to detect the power consumption behavior of users suspected of stealing electricity. This effectively narrowed the scope of detection, overcame the issue of having only a few samples of theft behavior, and improved the detection efficiency. Based on the Hadoop big data platform, Wu[10] collected data related to the internal customer power consumption behavior patterns of power enterprises, studied the relationship between customer power consumption and power theft, conducted training and modeling on the platform with a neural network, and applied the latest data verification results to prove that it could effectively improve the detection of, and response to, power theft. Through supervised machine learning from core user data such as audit, business expansion, electricity charges, line loss, measurement, and customer service, Cai[11] established a predictive power theft classification model, and assisted in the arrangement of electricity consumption inspection plans, effectively reducing the loss of state-owned assets.

As a new means of data processing, big data and artificial intelligence technologies can perform effective data analysis for a large number of complex scenarios. Based on a variety of user electrical information, this paper makes full use of Python to mine the data, applying logistic regression, random forests, decision trees, and other methods. The decision tree algorithm is made more accurate after business analysis and the application of 50% crossvalidation. The decision tree algorithm will be further improved, so as to more efficiently detect power theft and other abnormal behavior, as well as improving the existing collection system's audit efficiency and power management ability.

2. Principle analysis and algorithm selection

2.1 Analysis of the principle of electricity theft

When considering electricity theft, the user does not steal electrical energy as such, but in some way tampers with the metering device [12], causing a deviation in the reading and thus achieving the goal of paying less money. The expression for a watt-hour meter is as follows:

$$W = Pt = UI\cos\varphi \tag{1}$$

It can be seen from the above equation that the measured electrical energy is mainly related to the power and power consumption time. The metering power is subject to the metering voltage and the current of the watt-hour meter[13], metering voltage, phase-shifting steals electricity[14]. The influence of the three electrical quantities and the change of any factor related to the measurement of electrical energy will interfere with its measurement, thus achieving the purpose of power theft[15]. Various common means of stealing electricity are shown in Figure 1.

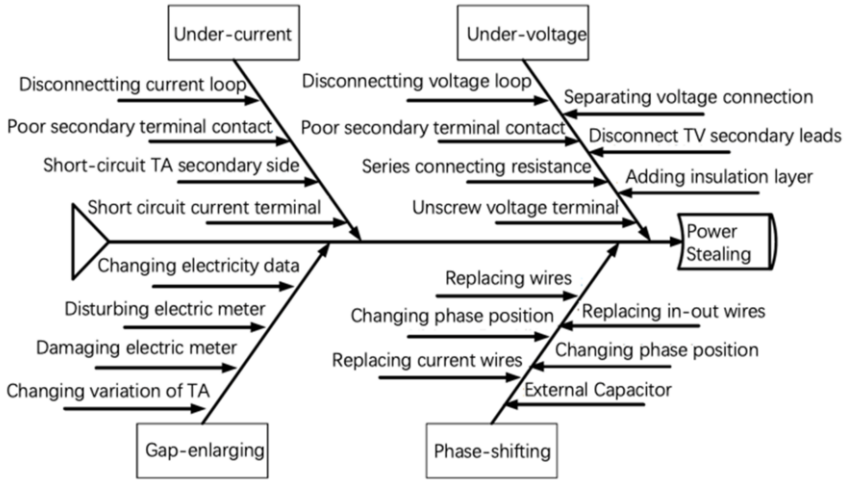

Figure 1. Methods of power theft.

2.2 Algorithm selection

The analysis and automatic recognition of power theft is a classification problem. Logistic regression, CART decision trees, random forests, and other algorithms can be used for this purpose[16]. For each problem, different algorithms have differet respective problems with adaptability. A comparison of the advantages and disadvantages of these main algorithms is shown in Table 1.

Table 1. Comparison of algorithms

Algorithm	Advantages	Disadvantages
Logistic Regression	1. Fast analysis and is suitable for dichotomy problems; 2. Has good robustness and will not be affected by slight multicollinearity.	1. Fit accuracy is low; 2. It cannot process data with missing values.
CART decision tree	1. High speed and accuracy, and can handle both continuous and discrete fields 2. There is no limitation on the uniqueness of data attributes; 3. No sensitivity to missing values; 4. Computing efficiency can be improved without remodelling	1. Processing time series data is a heavy workload; 2. Overfitting may occur; 3. Does not work well when dealing with data that is highly correlated.
Random Forest	1. No sensitivity to missing values and outliers; 2. High dimensional data can be processed without characteristic selection; 3. Both discrete and continuous data can be processed.	1. Overfitting may occur; 2. The data processing effect of different attributes is not good.

In order to further improve the performance of the model in predicting unknown data, this article chooses an area with 100 users for the electricity data (50 homes without power theft, 50 homes with power theft). Fifty percent cross validation is introduced to compare the different machine learning algorithms. The logistic regression, decision tree, and random forest models were trained through 50% cross validation[17]. In this environment, the decision tree has the highest accuracy for this data, so this paper focuses on the decision tree algorithm for modeling. The validation results are shown in Figure 2.

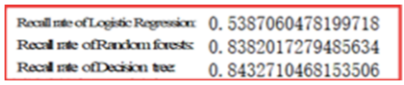

Figure 2. Model verification results

2.3 Decision tree algorithm (CART)

Suppose a training set is given $D = (x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N), x_i = \left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)}\right)^T$, where n is the number of features, $y_i \in 1,2,\ldots,K$ is the number of categories, N is the sample size, $i = 1,2,\ldots,N$. The decision tree model is constructed according to the given training set to achieve the goal of correct classification. The specific method is as follows: the information gain of each feature is calculated. The feature with the maximum value is chosen to be the root node, and the method above is recursively called on each child node to construct the decision tree until the information gain of all features is very small or there is no feature remaining to choose from. Also, in order to prevent overfitting, it is necessary to prune the generated decision tree.

The criterion of feature selection is information gain. Empirical entropy $H(D)$ represents the uncertainty of classifying the data set D, empirical conditional entropy $H(D|A)$ represents the uncertainty of classifying data set D under the conditions given by the feature and $|D|$ represents the sample size. Suppose there are K classes $C_k, k = 1,2,\ldots,K$, and $|C_k|$ is the number of samples belonging to class C_k, $\sum_{k=1}^{K}|C_k| = |D|$. Let feature A have n different values, $\{a_1, a_2, \ldots, a_n\}$; according to the value of feature A, divide D into n subsets $D_1, D_2, \ldots, D_n$, and $|D_i|$ is the number of samples of D_i, $\sum_{i=1}^{n}|D_i| = |D|$. The sample set belonging to class C_k in subset D_{ik} is denoted as D_i, The information gain is calculated as follows:

(1) $H(D) = -\sum_{k=1}^{K}\frac{|C_k|}{|D|}log_2\frac{|C_k|}{|D|}$ (2)

(2) $H(D|A) = \sum_{i=1}^{n}\frac{|D_i|}{|D|}H(D_i) = -\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{D_{ik}}{D_i}log_2\frac{D_{ik}}{D_i}$ (3)

(3) Calculating the information gain, $g(D,A) = H(D) - H(D|A)$ (4)

In general, the greater the information gain, the greater the "purity enhancement" obtained by this feature for the partition of the data set. However, it has the disadvantage of favoring attributes with more values, so the model established based on this does not have good generality. As a result, the information gain ratio is introduced:

$$g\_ratio(D,A) = \frac{g(D,A)}{H_A(D)} \quad (5)$$

$$H_A(D) = -\sum_{i=1}^{n}\frac{|D_i|}{|D|}log_2\frac{|D_i|}{|D|} \quad (6)$$

As the information gain ratio can be less number of desirable properties of preference, we, based on the modified, using the heuristic algorithm, namely from the candidate first find out the information gain above average in attribute, and then choose the highest information gain rate attributes, as the current property root node, in the subsequent also adopt heuristic method of recursive calls to produce the final decision tree.

3. Application scenarios

Based on big data and artificial intelligence, modeling power theft by key users is mainly divided into five parts: data exploration, data cleaning, feature extraction, model construction, and model evaluation. This paper focuses on 4250 power users in a certain area and 112,862 power records for power theft identification.

3.1 Data processing

3.1.1 Data consolidation

For this, the following high voltage power factors were collected: power, current, voltage, and meter code 27 indicators in total. Based on the collected data, using the metering point number and data date as the unique identifier, we will use pandas. This is merged to connect the data in different tables and so form the data table in Table 2.

Table 2. Data after transformation

Category	Name	Description
User	Name of measuring point	The name of the measuring point
	Number of measuring point	The measurement point number of each data item is uniquely identified by the measurement point number
	Date	The date of each data entry
Current	IA	The amplitude of phase A current in one day
	IB	The amplitude of phase B current in one day
	IC	The amplitude of phase C current in one day
	Current	The amplitude of current in one day
	Three-phase current unbalance rate	The magnitude of the three-phase current unbalance rate in one day
Voltage	UA	The amplitude of the phase A voltage in one day
	UB	The amplitude of the phase B voltage in one day
	UC	The amplitude of the phase C voltage in one day
	Voltage	The amplitude of the voltage in one day
	Three-phase voltage unbalance rate	The magnitude of the three-phase voltage unbalance rate
Power	Peak	Peak of total active power in one day
	Valley	Valley of total active power in one day
	Difference	Peak-valley difference for total active power
	Active A	Active power in A direction in three-phase electricity
	Active C	Active power in C direction in three-phase electricity
	Active B	Active power in B direction in three-phase electricity
	Total Active	Active power in three-phase electricity

	Reactive A	Phase A reactive power
	Reactive B	Phase B reactive power
	Reactive C	Phase C reactive power
	Total Reactive	Total phase reactive power
Electricity consumption	Positive active day freezing	The daily electricity consumption
	Power factor	High voltage total power factor
	Phase Angle	The vector angles of the waveforms of any two phases of three-phase electricity differ by 120 degrees at any time, which is called the phase angle of the two waveforms

3.1.2 Data cleaning and conversion

After exploring the formed data set, it was found that the data set contained 467 missing values; this missing data was then filled using an appropriate method[18]. At the same time, the peak-valley difference, forward active power freezing, active power total, three-phase current unbalance, and three-phase voltage unbalance were converted into five categories via Python to achieve the transformation from continuous to discrete values. The remaining continuous variables are divided into five groups, each of which is based on the quantile discretization function pd.qcut in Python. The variables are divided into five groups according to their rank or sample quantiles[19]. The converted data is shown in Table 3 below.

Table 3. Discrete data result

	Power Factor	Phase Angle	IA	IB	IC	Voltage	Active A	Active B	Active C	Peak	Valley
0	0.3549	1.208	0.051	0.0	-0.054	68.741	0.004	0.0	-0.002	0.0056	0.0017
1	0.3297	1.235	0.048	0.0	-0.045	68.643	0.004	0.0	-0.001	0.0092	0.0016
2	0.2650	1.303	0.044	0.0	-0.048	68.526	0.003	0.0	-0.001	0.0040	0.0017
3	0.2450	1.323	0.050	0.0	-0.055	69.592	0.004	0.0	-0.002	0.0043	0.0018
4	0.3014	1.265	0.048	0.0	-.0.053	68.866	0.004	0.0	-0.002	0.0057	0.0017

3.2 Feature extraction

Data with strong predictive power were selected from the variables with high collinearity correlation between features. Therefore, we use the information value (IV) to screen features with high correlation and strong predictive power[20].

The calculation of IV is based on WOE[21]:

$$WOE_i = \ln\left(\frac{py_i}{pn_i}\right) = \ln\left(\frac{{}^{\#y_i}/_{\#y_T}}{{}^{\#n_i}/_{\#n_T}}\right) = \ln\left(\frac{{}^{\#y_i}/_{\#n_i}}{{}^{\#y_T}/_{\#n_T}}\right) \tag{7}$$

When building a classification model, it is often necessary to filter the independent variables. Specific indicators are needed to measure the predictive power of each independent variable, and according to the size of these quantitative indicators, to determine which variables enter the model. The IV measure can be used to assess the predictive power of an independent variable. The formula for calculating IV is:

$$IV_i = (py_i - pn_i) * WOE_i = (py_i - pn_i) * \ln\left(\frac{py_i}{pn_i}\right)$$
$$= \left({}^{\#y_i}/_{\#y_T} - {}^{\#n_i}/_{\#n_T}\right) * \ln\left(\frac{{}^{\#y_i}/_{\#y_T}}{{}^{\#n_i}/_{\#n_T}}\right) \quad (8)$$

By calculating each characteristic IV quantity, the final result is shown in Figure 3.

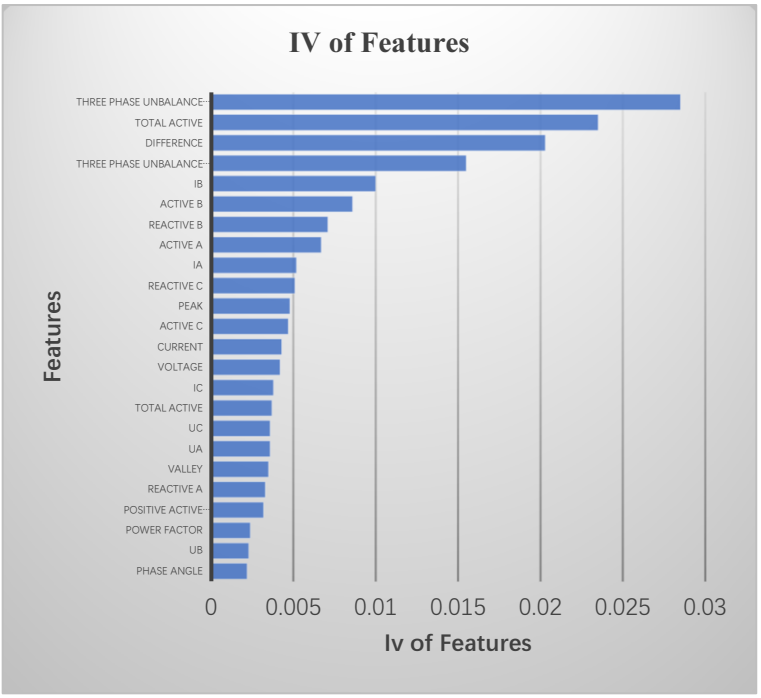

Figure 3. Feature IV.

According to the correlation coefficient and correlation of each index, the P value is used to preliminarily judge the correlation of variables, eliminate highly linear correlated variables, reduce the redundancy of variables, and simplify the input parameters of the model, so as to improve the data quality and improve the accuracy and performance of the model. The correlation between variables is shown in Figure 4.

In order to address the collinearity problem between these features, features with correlation coefficient >0.75 were screened out. Finally, based on the understanding of the business and the judgment of IV measure, 16 important indicators were selected, as shown in Table 4.

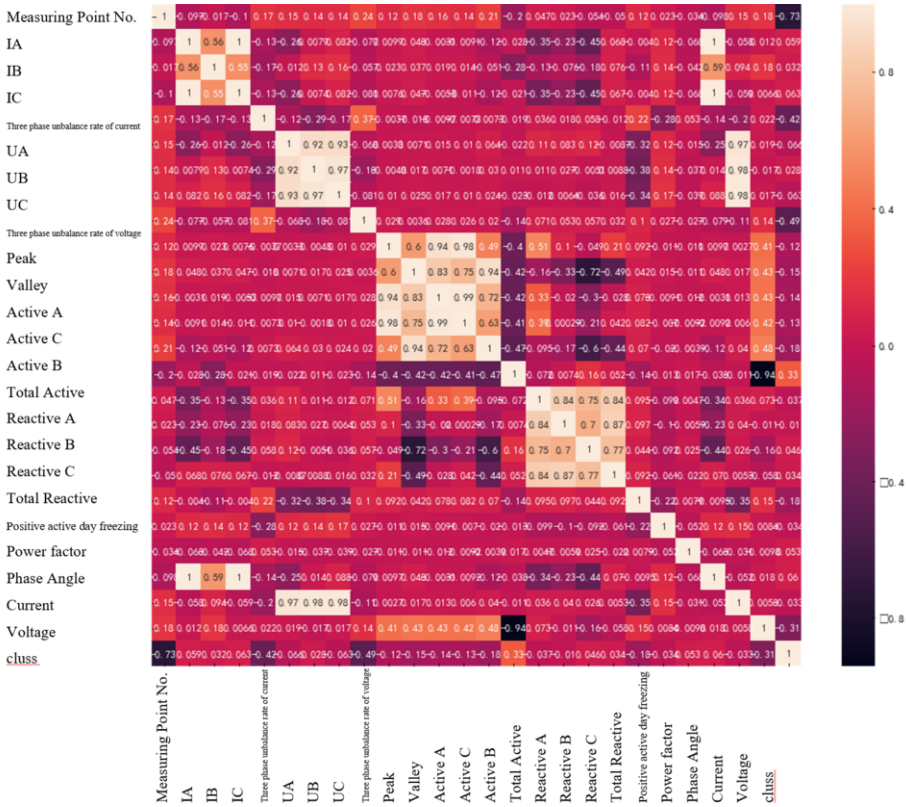

Figure 4. Variable correlation.

Table 4. Effective IV of features

Order Number	Features	IV of Features	Order Number	Features	IV of Features
1	Phase Angle	0.0022	13	Active C	0.0047
2	UB	0.0023	14	Peak	0.0048
3	Power factor	0.0024	15	Reactive C	0.0051
4	Positive active day freezing	0.0032	16	IA	0.0052
5	Reactive A	0.0033	17	Active A	0.0067
6	Valley	0.0035	18	Reactive B	0.0071
7	UA	0.0036	19	Active B	0.0086
8	UC	0.0036	20	IB	0.01
9	Total Active	0.0037	21	Three phase unbalance rate of current	0.0155
10	IC	0.0038	22	Difference	0.0203
11	Voltage	0.0042	23	Total Active	0.0235
12	Current	0.0043	24	Three phase unbalance rate of voltage	0.0285

3.3 Modeling and output

3.3.1 Model building

After processing, the data is divided into a training set and test set, and the optimal partition ratio is shown in Table 5.

Table 5. The accuracy of the model when dividing the test set and training set

Order Number	Split Scale (Test Set : Training Set)	Accuracy
1	4 (test) : 6 (train)	93.5
2	5 (test) : 5 (train)	93.6
3	2 (test) : 8 (train)	94.1
4	3 (test) : 7 (train)	94.5

According to these results, the final partition was achieved using a 7:3 training:test split. The improved decision tree method is used for training using the segmented training set, and the anti-theft model is obtained.

3.3.2 Model output

Python is used to calculate the feature importance for the power theft analysis model. In the model trained by the decision tree, the value of clf.feature_importances_ (key variable) is extracted as the coefficient of indicator feature importance. The calculation principle is to determine the weight of each index according to the information gain ratio of the dependent variable. The larger the information gain ratio, the higher the index weight. The resulting weights are shown in Figure 5.

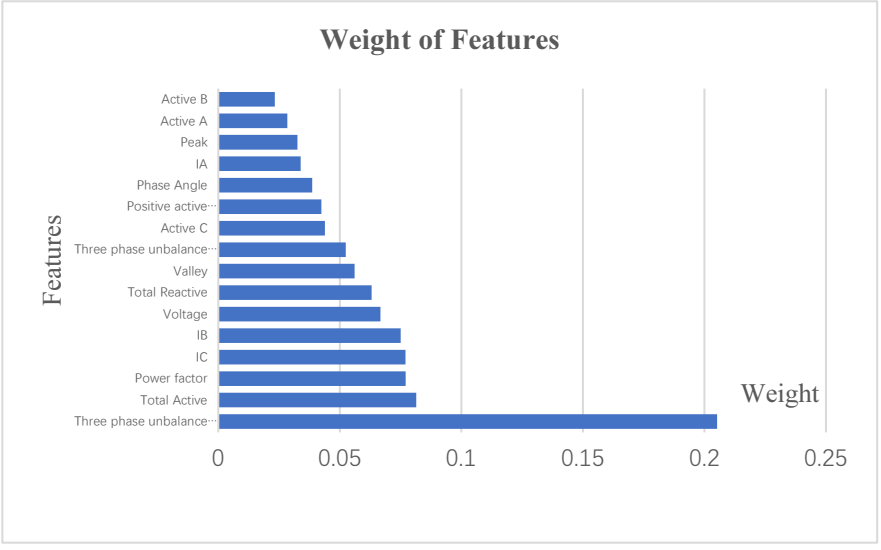

Figure 5. Weight of features.

According to the learned decision tree model, the three-phase voltage unbalance, total active power, power factor, and other factors have a high influence in determining power theft.

After the above data processing and model training, the classification system is established. Whether users steal electricity is the model prediction category, which is divided into users who steal electricity and normal users (1 represents users who steal electricity, 2 represents normal users). The prediction results are shown in Table 6 below.

Table 6. Prediction results

	0	1	2	3	4	5	6	7	8	9	…	3715	3716	3717	3718	3719	3720	3721	3722	3723	3724
test_est	2	2	2	2	2	2	2	2	2	2	…	2	2	2	1	2	2	1	2	2	2

The calculated output of the model includes the number of metering points, data date, suspicion coefficient of power theft, etc., as shown in Table 7.

Table 7. Model calculation result

	Measuring Point No.	Date of data	Three phase unbalance rate of current	Three phase unbalance rate of voltage	Positive active day freezing	Phase Angle	Total Reactive	Total Active	Suspected index of stealing electricity
0	20001294867	2015/10/8	0.005548	0.008772	0.00309	-0.009781	0.000329	-0.214651	0.600000
1	24000119881	2017/3/4	-0.149724	-0.062338	0.00309	0.011057	0.013128	0.003330	1.000000
2	68222	2015/11/10	0.005548	0.008772	0.00309	0.003261	0.000329	0.003330	0.250000
3	40606	2018/5/7	0.005548	0.008772	0.00309	-0.009781	0.013128	0.003330	0.714286
4	3158128	2017/7/27	-0.149724	-0.062338	0.00309	-0.009781	-0.017691	0.003330	1.000000

The model is a dichotomy, with a probability for the categories YES and NO. According to the calculated results, if the probability of YES is greater than or equal to 50% and the probability of NO is less than 50% via the suspicion coefficient of electric theft, then the user is classified as stealing power. Through the suspicion coefficient of electric power theft, the users who are stealing power are accurately identified. If the value falls within the range [0.5,0.8] this is determined to be a suspicion of general electric power theft; if it falls within the range [0.8,1] this is considered to be major electric power theft, as shown in Figure 6.

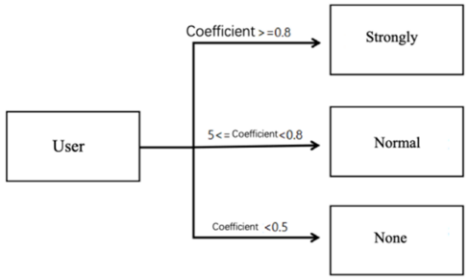

Figure 6. Power theft user division.

The final classification prediction results are shown in Table 8.

Table 8. Anti-theft classification prediction

Measuring Point No.	Date of data	Three-phase unbalance rate of current	Three-phase unbalance rate of voltage	...	Is it stealing electricity	Suspicion of power theft
68222	2015/11/27	0.2215	0.0027	...	Yes	Normal
39826605	2017/8/13	0.9661	1	...	Yes	Strong
107833	2019/3/2	0.0455	0.0043	...	No	None
...	...	...	...	...	...	...

3.4 Evaluation

A true positive TP is where the actual classification in the test set and the predicted result from the model are both that the user is stealing power (i.e. power-theft user); a true negative TN is where both are the normal user category. A false positive FP where the actual classification is a normal user but the predicted result is power theft. A false negative FN is the opposite of this. The formula for precision is: $Precision = \frac{TP}{TP+FP}$, and the formula for recall is: $Recall = \frac{TP}{TP+FP}$. The F_1 score is a harmonic average of the precision and recall, so the larger the F_1 score, the better the model.

$$F_1 = 2 * \frac{precision*recall}{precision+recall} \tag{9}$$

According to the above results, the evaluation of each category is shown in Table 9.

Table 9. Model evaluation

Category	Precision	Recall	F1 score
Power-theft user	88%	83%	85%
Normal user	96%	97%	97%
avg/total	95%	95%	95%

Next, the accuracy of the model is calculated, and the formula is:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{10}$$

Finally, the model accuracy is 94.1%.

4. Conclusion

In this paper, various behaviors of stealing electricity are analyzed and the fishbone diagram representation is given. Python is used to transform and populate data with missing values, apply IV assessment and an understanding of the business, and filter the important metrics. Based on this, an improved decision tree algorithm is used to construct a power theft prediction model. The validity and correctness of the model are verified by experimentation. The results of this paper show that this is an effective method to improve the efficacy of anti-theft audit.

Reference

[1] LAI Zhe. Analysis of Anti-stealing Electricity with Big Data Technology[J]. Electronic Test, 2019(16):50-51+38.

[2] GAO Wei, WEN Tongyang, XU Wei, et al. Analysis of Data Collection Technology and Application of Intelligent Anti Stealing[J]. Science and Technology Innovation Herald, 2019,16(18):32+34.

[3] XIONG Xia, TAO Xiaofeng, YE Fangbin, et al. Electricity Stealing Detection Method Based on Weighted Algorithm of Station Identification and Correlation Monitoring[J]. Journal of Computer Applications, 2019,39(S2):289-292.

[4] TANG Liangyi, WANG Kun, LI Hongliu, et al. Analysis on the Application of Big Data Technology in Intelligent Anti Stealing and Line Loss Monitoring[J]. Electrotechnical Application, 2017,36(17):48-53.

[5] WEI Yao, ZHU Weiyi, GONG Taorong, et al. Abnormal Electric Consumption Analysis System Using Data Mining[J]. Electric Power Information and Communication Technology, 2014,12(05):70-73.

[6] HOU Dan, LI Gang, ZHAO Wenqing, et al. Application of Big Data Analysis and Visualization Technology in Power Grid Company[J]. Smart Grid, 2015,3(12):1186-1191.

[7] CHENG Junwen, LI Huijuan, CHAO Zhiqiang. Research on Anti-stealing Based on K-means Clustering Algorithm and Electricity Information Collection System［J］Distribution & Utilization, 2018,36(1):75-80.

[8] ZHENG Jianning. Detection Method of Electricity Stealing Behavior Based on Deep Learning[J]. Information Technology, 2019(02):156-159.

[9] Kang Ningning, Li Chuan, Zeng Hu, et al. Electric larceny detectionusing FCM clustering and improved SV R model[J]. Journal of Electronic Measurment and Instrumentation, 2017,31(12):2023-2029.

[10] WU Yiliang. Research on Prediction Platform of Power Stealing Based on Hadoop[J]. Mechanical and Electrical Information, 2017(06):20-21.

[11] Cai Jiarong, Wang Shunyi, Wu Guangcai. The User's Electric Power Prediction and Electricity Inspection Plan Based on Machine Learning Research on Auxiliary Arrangement[J]. Electronic Test, 2018(02):108-109.

[12] Zhang Fangfang. Analysis of Anti-Stealing of Intelligent Electric Energy Meter[J]. Electronic Technology & Software Engineering, 2018(15):206.

[13] LUO Kewei. Analysis of Electric Power Stealing and Anti-Electric Power Stealing Technology in Electric Power Monitoring[J]. New Technology & New Products of China, 2019(24):116-117.

[14] CHEN Jianliang. Technical Analysis of Anti-Stealing in Power Supply Enterprises[J]. China High-Tech Enterprises, 2013(22):119-120.

[15] FU Weizhu, ZHANG Pei, GAO Wei, et al. Case Analysis of The Over Differential Stealing of the Telephone Energy Meter by Maliciously Putting Capacitor into Operation[J]. Science and Technology Innovation Herald, 2016,13(08):7+9.

[16] CUI Hongyan, XU Shuai, Zhang Lifeng, et al. The Key Techniques and Future Vision of Feature Selection in Machine Learning[J]. Journal of Beijing University of Posts and Telecommunications, 2018,41(01):1-12.

[17] XU Chang, CHEN Jinqiong, ZHOU Wen. Improving unbalanced data classification of unsupervised limit learning machine[J]. Journal of Anhui Normal University (Natural Science), 2018,41(06):544-551.

[18] FAN Qiao, GUO Aijun. An Improved Solow Residual Method for TFP Calculating under the Framework of Spatial Econometrical Analysis[J]. The Journal of Quantitative & Technical Economics, 2019,36(08):165-181.

[19] LIU Yun, YUAN Haoheng. Parallel Discretization of Data Preparation Optimization in Data Mining[J]. Journal of Sichuan University(Natural Science Edition), 2018,55(05):993-999.

[20] Yang Jianfeng, Qiao Peirui, Li Yongmei, A Review of Machine-learning Classification and Algorithms[J]. Statistics & Decision, 2019,35(06):36-40.

[21] GENG Juncheng, ZHANG Xiaofei, YUAN Shaoguang, et al. Research and Implementation of Users'outage Sensitivity Score Card Basedon Logistic Regression Model[J]. Power Demand Side Management, 2018,20(03):46-50.