Application of an Ensemble Learning Method in the Risk Prediction of Electricity Tariff Recovery

Huaguang Wu*, Mingxing Zhang*, Baohua Jin

School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, China

*Corresponding author e-mail: 15738893736@163.com, *342027523@qq.com, 450082368@qq.com

Abstract. Electricity tariff recovery has always been an important part of power companies. However, there are often some users who are in arrears in electricity charges. Therefore, how to make effective risk assessment for electricity customers is the key work at present. In this paper, an ensemble learning method named gradient boosting decision tree-random forest-adaboost-logical regression (GRAL) is proposed to construct a risk prediction model of electricity tariff recovery. Firstly, the data provided by the power company was preprocessed with feature engineering. Secondly, the GRAL ensemble learning method and the traditional logistic regression algorithm are used separately to construct the risk prediction model of electricity tariff recovery to predict the probability that the user will default next month. Finally, we make a risk assessment for users according the probability and predict whether the user in different risk levels will be in arrears. Experimental analysis shows that the proposed GRAL ensemble learning method has better experimental results compared with the traditional logistic regression algorithm, this method can effectively evaluate the payment behavior of power customers and make a more accurate prediction about whether users will be in arrears.

1. Introduction
Power resources are the top priority of the national economy and people's livelihood. With the continuous deepening of china's economic system reform, the power industry is also facing many challenges [1]. All along, electricity tariff recovery is the key content of power marketing, and plays a decisive role as the last part of the marketing of power supply companies [2]. However, in actual work, there are often some users who are in arrears of electricity charges. There are many reasons for the arrears of electricity charges, for example, the lack of awareness of timely payment or financial difficulties may lead to the failure of completing the electricity bill in time [3]. To this end, power companies have been committed to improving the rate of electricity tariff recovery through various methods for many years. These methods are often the use of management means, such as strengthening the leadership work of power companies [4], focusing on key enterprises, etc. [5].

In recent years, many scholars have proposed user risk prediction methods based on power big data. This paper [6] proposed the PSO-SVM algorithm to solve the risk warning model. Another paper [7] developed a prediction system for assessing the arrear risks of power customers, but they did not make
a subtilized analysis of the risk of user arrears. Zhanhui Xiao et al. conducted a group analysis of users through probability clustering, and made a risk prediction for each group of users with logistic regression [8]. But the prediction accuracy of each group was not high. In the current study, many literature [9, 10] have done a lot of research on power customer segmentation which mostly focused on single dimension clustering for customers from the business point of view. At present, there is still lack of in-depth research on two aspects of differentiated service strategy and strategy application evaluation, and there is still room for improvement in the field of landing applications [11-13]. In addition, there are many kinds of power big data with complex data structure, however, few scholars have made detailed analysis on the feature engineering part before the model construction.

In our work, we first get the data suitable for model training by processing the data with feature engineering. Secondly, the GRAL ensemble learning method and the traditional logistic regression algorithm are used separately to construct the risk prediction model of electricity tariff recovery to predict the probability that the user will default next month. And then the users are divided into three different risk levels: high risk, medium risk and low risk according to the probability values. Finally, we predict whether the user in three different risk levels will be in arrears. Experimental analysis shows that compared with the traditional logistic regression algorithm, the proposed GRAL ensemble learning method has better experimental results. This method can effectively evaluate the payment behavior of power customers and make a more accurate prediction about whether users will be in arrears.

2. Component learners

2.1. Gradient boosting decision tree

The gradient boosting decision tree (GBDT) model was proposed by Jerome Friedman in 1999[14]. It is an application of the combination of decision tree and boosting method. When we use a decision tree for classification, there is a residual of the real result and the predicted result. This residual is the training data of next decision tree in the GBDT. The first tree is created with the original data. This is also the reflection of boosting in GBDT. The algorithm idea is shown in figure 1.

![Figure 1. Schematic diagram of GBDT algorithm](image)

As can be seen from the figure above, GBDT does not train multiple decision trees at the same time. When each decision tree is established, the residual of the previous tree is used to fit the weak learner. The final prediction result of the model are obtained by summing up the results of each decision tree. The formula is as follows:

$$t_{pre} = t_1 + t_2 + \cdots + t_n$$  

(1)

2.2. Random forest

Random forest is an ensemble algorithm. By combining several weak classifiers, the final results are voted or averaged, which makes the results of the whole model have higher accuracy and generalization performance. The idea of random forest classification is:
(1) K samples are selected from original training set d by bootstrap sampling, and each sample size is the same as the original training set;

(2) K sub-training sets are used to randomly select some features from each sub-training set and then select the optimal features for segmentation. K decision tree models \( \{h(X, \theta_k), k = 1, \ldots\} \) are constructed. The parameter set \( \theta_k \) is an independent and identically distributed random vector, and k classification results are obtained under given independent variable x;

(3) According to the voting results, the final classification is voted out of k categories. The final classification decision formula is:

\[
H(x) = \arg \max_{y} \sum_{i=1}^{k} I(h_i(x) = y)
\]  

\( H(x) \) Denotes the combination classification model, \( h_i(x) \) is a single decision tree classification model, \( y \) denotes the output variable. \( I(\varnothing) \) is the presentation function, that is, when the set of classification results contains the classification results of this decision tree model, the function value is 1, otherwise it is 0. Finally, a majority vote is used to determine the final classification.

2.3. Adaboost

Adaboost is an iterative algorithm. The idea of adaboost is to train multiple weak classifiers first, and then combine these weak classifiers to construct a strong classifier to achieve the purpose of learning. In the training of adaboost algorithm, the samples have the same initial weights. First, a weak classifier is trained and the error rate of the classifier is calculated. After each model training, the weight of the sample is adjusted according to the previous learning result. The essence of the adaboost learning process is to constantly change the weight of the samples in the learning until the error is zero or the number of learners reaches the preset value, and then the results of all weak classifier learning are integrated according to the weight, and finally output the result.

2.4. Logistic regression

Logistic regression is essentially a generalized linear regression method based on probability [15], which has been widely used in health evaluation and risk assessment [16, 17].

In this study, we assume that the sample is \( \{X, y\} \) and that y takes 0 or 1 to represent positive and negative classes. Specifically, \( y = 0 \) means "No arrears, that is, users pay electricity bills in time", and \( y = 1 \) means "Arrears, that is, users fail to pay electricity bills in time". \( X \) is the eigenvector of n-dimensional samples, we assume that \( x_1, x_2, \ldots, x_n \) is n eigenvalues. For a sample x, it belongs to a negative class, that is, the probability of arrears can be expressed by the following functions:

\[
p(y = 1|X; \theta) = \frac{1}{1 + \exp(-g(x))}
\]  

\( \theta \) is the model parameter, which is the regression coefficient. For \( g(x) \),

\[
g(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n
\]  

It can be seen from the formula that if the value of the regression coefficient is determined, the logistic regression model can be determined uniquely.
3. The risk prediction model of electricity tariff recovery

3.1. Structure of the model
The structure of risk prediction model for electricity tariff recovery is shown in figure 2. It can be divided into four parts:

1. Getting data: Reading the original data from the big data platform and extracting the basic data of risk users from the data of 240,000 users;

2. Data preparation: Getting the training set and test set by processing the data of risk users with feature engineering, this is the preparation for the next step in building the model;

3. Model training: Getting the trained model by training the user's 12-month historical electricity data with GRAL ensemble method, then the trained model is used to predict the probability of user's arrears next month, and the results are output.

4. Risk assessment and prediction: Users are divided into three different risk levels: high risk, medium risk and low risk according to the probability values, and then the users in the same risk levels are forecasted whether they are arrears or not. After that, the predicted results are fed back to the big data platform.

![Figure 2. The structure of risk model for electricity tariff recovery](image)

3.2. Feature engineering
Data is the basis of model building, and the quality of data processing will affect the effect of the model to a certain extent. Consequently, the data should be processed with feature engineering to obtain data suitable for model training.

3.2.1. Feature selection. Usually, feature selection can reduce the dimension of feature space and improve the accuracy of the model. The work of feature selection in this study was done twice.

1. In this study, there are more than 50 features of the original data. We calculated the Pearson correlation coefficient among features, and deleted the features with low correlation coefficient. The experience removed redundant features and selected 30 features eventually. Due to space limitations, Table 1 here only shows some features.
Table 1. List of features

| Feature                | Description                  | Type       | Remark                      |
|------------------------|------------------------------|------------|-----------------------------|
| Cons_sort_code         | User category                | Categorical| Archival information        |
| Trade_code             | Industry category            | Categorical|                             |
| Sfyq                   | Arrears                      | Categorical|                             |
| Yszb                   | Pre-collection offset ratio  | Number     | Label                       |
| Jfjsl                  | Timely rate of payment       | Number     | Electricity use behavior in the month |
| Cashchk_day_num        | Length of settlement         | Number     |                             |
| T_pq                   | Normal power consumption     | Number     |                             |
| Release_before15       | Is the release date before 15th? | Categorical|                             |

(2) After the completion of the extension of features, the importance analysis of these features is based on the gini index, and then the features with low importance are removed step by step. After many experiments, the most effective feature set for model training is selected.

3.2.2. Extension of features. We get 30 basic features when the first feature selection, which can be divided into two categories, one is used to describe the user's electricity consumption behavior in the current month, and there are altogether 24 features; another category is archival, which describes the user's number, address and other basic information, and there are six features. For the month we want to predict, there is no user's electricity information, only the archival information.

We can verify that these 30 features can not accurately predict whether users will incur arrears in the next month, and lack some historical power characteristics as a reference. We need to join the user's historical data as a supplement, it is shown in table 2 (there are too many features, just take one of them as an example):

Table 2. The method of expanding feature

| Feature | Extended feature | Description                                      | Remark                                      |
|---------|------------------|--------------------------------------------------|---------------------------------------------|
| Jfjsl   | Jfjsl_1          | ‘jfjsl’ in the first month before the forecast month | The 24 features describing the current month's electricity use behavior are expanded in this way. |
|         | Jfjsl_2          | ‘jfjsl’ in the second month before the forecast month |                                            |
|         | Jfjsl_3          | ‘jfjsl’ in the third month before the forecast month |                                            |
|         | Jfjsl_4          | ‘jfjsl’ in the fourth month before the forecast month |                                            |
|         | Jfjsl_5          | ‘jfjsl’ in the fifth month before the forecast month |                                            |
|         | Jfjsl_6          | ‘jfjsl’ in the sixth month before the forecast month |                                            |

After that, 14 of the 24 features are further expanded, which are “yszb”, “jfjsl”, “cashchk_day_num”, “hksc”, “t_pq”, “t_pq_sum”, “release_day”, “charge_day”, “end_day”, “end_day2”, “remain_daynum”“remain_daynum2”, “all_daynum”, “all_daynum2”. In the end, we get a total of 230 features. The extension method for “jfjsl” is shown in table 3 (because there are too many features, just take one of them as an example):
Table 3. The method of further expansion

| Feature   | Extended feature | Description                                                                 | Remark                                                                 |
|-----------|------------------|----------------------------------------------------------------------------|------------------------------------------------------------------------|
| Jfjsl_min | Jfjsl_min        | The minimum value of `jfjsl' in the previous six months                     |                                                                        |
| Jfjsl_max | Jfjsl_max        | The maximum value of `jfjsl' in the previous six months                     |                                                                        |
| Jfjsl_mean| Jfjsl_mean       | The average value of `jfjsl' in the previous six months                     |                                                                        |
| Jfjsl_var | Jfjsl_var        | The variance of `jfjsl' in the previous six months                         |                                                                        |
| Jfjsl_med | Jfjsl_med        | The median of `jfjsl' in the previous six months                           |                                                                        |
|           | Jfjsl_std        | The standard deviation of `jfjsl' in the previous six months               | Fourteen features are extended in this way                             |

3.3. GRAL ensemble learning method

In this paper, we adopt the GRAL ensemble learning method, which trains several different component learners to get multiple output results, and then getting the final result by training a new model which takes the output of the previous component learners as input. In this study, the component learners adopted are GBDT, random forest (RF) and adaboost. The principle of GRAL is showing bellow:

![Figure 3. The principle of GRAL](image)

Step 1: The data is divided into two mutually exclusive sets - data_A and data_B;
Step 2: The data_A is divided into two mutually exclusive parts - data_A1 and data_A2. Three different algorithms (GBDT, RF, adaboost) are used to train data_A1, so as to obtain three trained
component learners which are learner_GBDT, learner_RF and learner_adaboost respectively. Using these three trained component learners to predict the data_A2, three sets of prediction results are obtained, which are pre1_GBDT, pre1_RF and pre1_adaboost.

Step 3: The three trained component learners in step 2 are used to predict the data_B and outputting three sets of prediction results which are pre2_GBDT, pre2_RF and pre2_adaboost;

Step 4: The prediction results of three component learners in step 2 and step 3 are the three newly generated features, and the output result of each component learner correspond to one feature. We merge the features of pre1_GBDT, pre1_RF and pre1_adaboost into a new data named train_new, and merge the features of pre2_GBDT, pre2_RF and pre2_adaboost into a new data named test_new. Next, logistic regression algorithm is used to train the train_new, and the trained model is used to predict the test_new. Finally, the prediction results of the model are output, that is, the possibility of user default.

3.4. Risk assessment and prediction

From the process of building the risk model mentioned above, we can see that the logistic regression algorithm will output the user's arrears probability \( P \), and then we set an appropriate threshold \( T_1 \) to divide the user into three levels: high risk, medium risk and low risk according to the model output. Specific principles of division are shown in table 4.

| Threshold \( T_1 \) | The level of risk   |
|---------------------|---------------------|
| \( P \geq 55\% \)   | high-risk           |
| \( 20\% \leq P < 55\% \) | medium-risk        |
| \( P < 20\% \)     | low-risk            |

We define users whose arrears probability is more than 55\% as high risk users, those whose arrears probability is between 20\% and 55\% as medium risk users, and those whose arrears probability is less than 20\% as low risk users. After that, the threshold \( T_2 \) is set according to probability \( P \), users in the same risk level are divided into arrears and non-arrears.

| The level of risk | Threshold \( T_2 \) |
|-------------------|---------------------|
| high-risk         | 55\%                |
| medium-risk       | 22\%                |
| low-risk          | 10\%                |

4. Experiment and analysis

4.1. Dataset

The data source of this study is the payment records and power consumption information of 240,000 high voltage users in a province. It is easy to find that only about 0.6\% of all users are in arrears in a month. It can be seen that the distribution of positive class and negative class is unbalanced extremely. Therefore, we only need to pay attention to these risk users who have a record of arrears.

4.2. Result analysis

In the experimental part, we trained the data of users for 12 months (August 2017 - July 2018) with GRAL ensemble method and traditional logistic regression algorithm respectively, and then we used the two trained models to predict whether users will be in arrears in August 2018, and compared the predicted results with the actual results. The experimental results are shown in table 6.
Table 6. The results of two models

| Risk level | Model            | Recall  | Precision | F1     |
|------------|------------------|---------|-----------|--------|
| high risk  | GRAL             | 100.00% | 80.57%    | 89.24% |
|            | logistic regression | 100.00% | 58.05%    | 73.45% |
| medium risk| GRAL             | 92.08%  | 63.60%    | 75.23% |
|            | logistic regression | 90.58%  | 54.82%    | 68.31% |
| low risk   | GRAL             | 40.89%  | 23.51%    | 29.86% |
|            | logistic regression | 84.69%  | 12.45%    | 21.70% |

The recall, precision, and F1 of GRAL model are higher than those of traditional logistic regression model in high risk users and medium risk users. Especially in precision of high and medium risks, GRAL model has obvious advantages.

Figure 4. Recall, precision and F1 in different risk levels

Figure 4 compares recall, precision, and F1 of the two models in three different risk levels. It can be seen that GRAL model is superior to traditional logistic regression model in recall, precision and F1 except recall of low risk users. In low risk users, although the recall of logistic regression model is higher than that of GRAL model, reaching 84.69%, the precision of logistic regression model is only 12.45%, which is about 1/2 of GRAL model. As mentioned above, when both recall and precision cannot be taken into account at the same time, the harmonic average F1 value of the two models is used to compare the two models. For low risk users, the F1 value of the GRAL model reaches 29.86%, while that of the logistic regression model is only 21.70%. Therefore, it can be considered that the GRAL model is still better than the traditional logistic regression model in low risk users.

In addition, we can also find that neither the GRAL model proposed nor the traditional logistic regression model has a good prediction result for low risk users. Through sampling analysis, it is found that the low-risk users have higher uncertainty in the payment behavior, and most of the low-risk users are within two times of the arrears in the past two years, and the arrears time is highly volatile. Therefore, if we want to predict more accurately whether these users are in arrears, we need to add more external features. In addition, we can consider referencing other algorithms to optimize debugging according to the characteristics of low-risk user data imbalance.

In summary, compared with the traditional logistic regression model, the GRAL ensemble learning method designed in this paper is more suitable for building a risk prediction model of electricity tariff.
recovery. The experimental results also show that the GRAL model can effectively realize the risk early-warning of electricity tariff recovery for power users.

5. Conclusion and future work
In this paper, we propose an ensemble learning method (GRAL) to build a risk early-warning model of electricity tariff recovery. Experimental analysis shows that compared with the traditional logistic regression algorithm, the proposed GRAL ensemble learning method has better experimental results in the prediction of three different risk levels.

However, the prediction effect for low risk users is not very good. More external features need to be added if we want to predict whether low risk users will be in arrears more accurately. In addition, because the distribution of positive and negative classes of low risk user data is unbalanced extremely, we can refer to other algorithms to optimize debugging, which will be a great challenge. At the same time, this is also the direction we need to focus on next.

Acknowledgements
This work was supported by “National Natural Science Foundation of China (Grant No. 61672470), the National Key Research and Development Plant (Grant No.2016YFE0100600 and 2016YFE0100300), the Second Education Fund for Industry and Education project "Digital Science and Technology, Wisdom for the Future" (NO. 2018A01094), the Doctoral Research Fund of Zhengzhou University of Light Industry (13501050045), and the Science and Technology Project of Henan Province (182102210617), the Research Fund for the Doctoral Program of Zhengzhou University of Light Industry (NO. 2017BSJJ046), the Second Education Fund for Industry and Education project "Digital Science and Technology, Wisdom for the Future" (NO. 2018A01094), Foundation of Henan Province Educational Committee (No.17A520064), Foundation of the doctoral research project of ZZULI(No.13501050066)”.

References
[1] Anderson J A . Electricity Restructuring: A Review of Efforts around the World and the Consumer Response [J]. Electricity Journal, 2009, 22(3): 70-86.
[2] Lave L B , Apt J , Blumsack S . Rethinking Electricity Deregulation [J]. The Electricity Journal, 2004, 17(8): 11-26.
[3] Hong-Tao X , Yong-Yi S , Wei-Jia L , et al. Research and Practice on Risk Prediction Model of Electricity Tariff Recovery [J]. Electric Power Information and Communication Technology, 2016.
[4] Wei-Dong Su . How to Scientifically Prevent the Power Supplying Section from the Electric Charge Risk [J]. Qinghai Electric Power, 2005.
[5] Ping P , Duo L , Wen-Hong Z . Exploration on an Early Warning Made by County Power Enterprises to Resolve the Risk of Electric Charge [J]. Journal of Chongqing Electric Power College, 2009.
[6] Xiaomu Z , Tao W , Jianjun H , et al. Study on Tariff Risk Early Warning of Electric Power Users Based on PSO-SVM Algorithm[C]// 2018 International Conference on Big Data and Artificial Intelligence (BDAI). IEEE, 2018.
[7] Guo W , Hong W , Li W , et al. Design and Implementation of Electric Charge Arrears Prediction System[C]// Web Information System & Application Conference. IEEE, 2016.
[8] Xiao Z , Feng X , Dang L , et al. Research and application of subtitled customer clustering algorithm in power marketing[C]// 2017 IEEE Conference on Energy Internet and Energy System Integration (EI2). IEEE, 2017.
[9] Peng X , Deng D , Cheng S , et al. Key technologies of electric power big data and its application prospects in smart grid[J]. Zhongguo Dianji Gongcheng Xuebao/Proceedings of the Chinese Society of Electrical Engineering, 2015, 35(3): 503-511.
[10] Cai-Hua S , Yong-Cai W , Yuan-Juan L , et al. Analysis of data mining based customer
classification model for electric power industry [J]. Modern Electronics Technique, 2014.

[11] Li Z, Xingzhe H, Jun H U, et al. Improved K-Means Algorithm Based Analysis on Massive Data of Intelligent Power Utilization [J]. Power System Technology, 2014, 38(10): 2715-2720.

[12] Suxiang Zhang, Jianming Liu, Bingzhen Zhao, et al. Cloud Computing-based Analysis on Residential Electricity Consumption Behavior [J]. Power System Technology, 2013, 37(06): 1542-1546.

[13] Si Liu, Xuhua Fu, Chengjin Ye, et al. Spatial Load Clustering and Integrated Forecasting Method of Distribution Network Considering Regional Difference [J]. Automation of Electric Power System, 2017, 41(03): 70-75+82.

[14] Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine [J]. Annals of Statistics, 2001, 29(5): 1189-1232.

[15] Caesarendra W, Widodo A, Yang B S. Application of relevance vector machine and logistic regression for machine degradation assessment [J]. Mechanical Systems and Signal Processing, 2010, 24(4): 1161-1171.

[16] Hongrui Wang, Longxia Qian, Xinyi Xu. Risk Assessment Model and Application of Water Shortage Based on Fuzzy Probability [J]. Journal of Hydraulic Engineering, 2009, 40(07): 813-821.

[17] Mazzocco T, Hussain A. Novel logistic regression models to aid the diagnosis of dementia [J]. Expert Systems with Applications, 2012, 39(3): 3356-3361.