Intelligent Research on Customer Complaints Based on 95598 Work Orders Track

Xu Xin¹², Fu Jun¹², Wang Li¹², Gong Dongmei¹², Wang Yujun¹², Tang Peilin³
¹Institute of Electric Power Science of State Grid Jibei Electric Power Co., Ltd., Beijing 100045
²Huadian Electric Power Research Institute Co., Ltd., Beijing 100045
³Chengde Power Supply Company of State Grid Jibei Electric Power Co., Ltd., Chengde, Hebei Province 067000

Abstract. With the rapid development of society and economy, the living standard of people is continuously improving, which makes the society's demand for electricity increasing day by day, as well as the demand for power supply capability and service quality. When the power supply service enjoyed by customers does not meet the expected value, it will stimulate various demands of customers and even the customer complaints. How to accurately grasp customer demands and effectively prevent customer complaints is the key to improve customer satisfaction of power supply service and maintain a good image of power enterprises. The purpose of this article is to excavate the internal connection of the existing data through the serialization research based on the national power grid data, so as to predict whether users will make complaints after the call.

1. Data Preprocessing and Feature Construction

1.1 Field selection
This project is based on the historical work order data of the State Grid. The data are used in 2017 and January to June 2018; Nine types of work orders are involved 'report', 'complaint', 'business consultation', 'service application', 'suggestion', 'opinion', 'customer reminder', 'praise' and 'business supervision'; The total amount of data is over 2.4 million. Fields involved: 98. There are 9 fields finally selected. The details are as follows:

| Field name | Field meaning |
|------------|--------------|
| ACCEPTCONTENT | Accepted content |
| HANDLESI TUA TION | Handle situation |
| ORDERNO | State Grid Work Order No. |
| CALLNO | Calling number |
| TEL | Contact number |
| TEL2 | Contact number 2 |
| REQSTARTDATE | Request start time |
| REQFINISHDAT E | Request end time |
| TYPECODE | Business type |
1.2 Missing value processing
Statistics on missing fields are as follows:

| Field name         | Total data | Missing value | Missing rate |
|--------------------|------------|---------------|--------------|
| ORDERNO            | 2400957    | 0             | 0.00%        |
| CALLNO             | 2400957    | 47421         | 1.98%        |
| REQSTARTDATE       | 2400957    | 1010          | 0.04%        |
| REQFINISHDATE      | 2400957    | 1101290       | 45.87%       |
| TEL                | 2400957    | 65            | 0.00%        |
| TEL2               | 2400957    | 1173702       | 48.88%       |
| TYPECODE           | 2400957    | 0             | 0.00%        |
| ACCEPTCONTENT      | 2400957    | 20081         | 0.84%        |
| HANDLESITUATION    | 2400957    | 1923518       | 80.11%       |

(1) CALLNO: This feature is used to distinguish different users. Wouldn't it be better to use the field CUSTNO directly? However, since this field is not required and the missing rate is over 50%, it is very unreasonable to study the sequence if the data is deleted directly. Other filling methods are not applicable, so CALLNO is chosen instead, which cannot guarantee 100% accuracy. But as far as the actual situation is concerned, it should be basically consistent. Similarly, the missing value of CALLNO can also be filled with TEL and TEL2.

(2) REQSTARTDATE and REQFINISHDATE: These two fields are mainly needed for the subsequent construction of "talk duration" feature and cannot be better filled, so the missing value is "None" for unified processing, and only one value can be set for the construction feature in case of this situation.

(3) ACCEPTCONTENT: Because it is text data and the missing rate is very low, the uniform value of "None" is directly used to fill it, which will not have a great impact on the nature of the model.

(4) HANDLESITUATION: It is also text data with a high missing rate. However, the reason for missing this field is that the work order is completed once. These missing data actually indicate that it has such a same attribute. "None" is used for filling to ensure data consistency and no influence on subsequent construction features.

1.3 Feature construction
There are 10 final modeling selection features, which will be explained one by one:

1.3.1 Emotion Score-Comprehensive Consideration: The calculation of emotion score is based on ACCEPTCONTENT and HANDLESITUATION:

![Figure. 1. Flow chart of emotion score feature construction](image_url)
(1) First of all, a function of emotion score needs to be defined to calculate the emotion function score of the current work order (i.e. time and sequence issues are not considered). The emotional function score of the current text (accepted content or processing situation) is the sum of the scores of each emotional word (positive score for positive words and negative score for negative words).

(2) Calculate the emotion function score of the current text by using the emotion score function.

(3) Calculate the sequence emotion score of the current document by using a preset index lifting formula.

(4) Calculate the final cumulative emotion score of the current work order: Accumulate the sequence emotion score from the first work order in the sequence to the current work order.

(5) The above-mentioned methods calculate the final cumulative emotional scores of the accepted content and the processing situation respectively, and construct this feature according to the positive and negative relationship between the two.

1.3.2 Last Call Interval: Subtract the last incoming call time (days) from the current work order customer call time.

1.3.3 Whether there are complaints in history: search the history work order to find out whether there are complaint work order records of current work order customers, and construct binary features.

1.3.4 Statistical sequence analysis based on accepted processing contents:

Based on logical business thinking, it is believed that customers who show dissatisfaction in accepting content and have a long process or unresolved problems are more likely to complain. Therefore, the content of ACCEPTCONTENT (Accepted Content) HANDLESITUATION is mined. By comparing the ACCEPTCONTENT of positive and negative samples, after the text is cleaned of useless characters, it is found that some expressions appear more frequently in complaint samples than in non-complaint samples. Define rules through statistical analysis: if a rule occurs more than 5 times in a positive sample and the frequency in a positive sample/the frequency in a negative sample >= 3, it is used as a seed expression to traverse the ACCEPTCONTENT of all work orders. If it is expressed in the seed expression, the score label=1, otherwise it is 0. HANDLESITUATIOND is analyzed in the same way, but statistics show that the expressions with more frequency in positive and negative samples are basically indistinguishable, and the missing ratio of such data is 80%. The same analysis cannot be performed. Based on the acceptance situation, if the acceptance content=1 and the HANDLESITUATIOND is not empty, its feature is set to 1, otherwise it is 0. Draw lessons from the thinking of emotional score and increase the thinking of time series. Rules are cumulative form (tested index form has no discrimination at all, give up) and take it as the final feature.
1.3.5 Equals the lifting sequence: According to the results of sequence pattern mining, some sequences with high improvement on complaints (the contents of the sequences are the values of TYPECODE) are found to match the sequences where the current work order is located, and binary features are constructed.

1.3.6 Call Time Period: hours extracted from call time.

1.3.7 Call duration: the difference (in minutes) between REQFINISHDATE and REQSTARTDATE, and the value with missing value is 100.

1.3.8 Label fuzzification:

1.3.9 Month: the month selected from the incoming call time.

1.3.10 Sequence Length: Find the sequence according to the sequence construction method in sequence pattern mining, and look at the length of the sequence where the current work order is located.

1.4 Data Processing for Overlapping Positive and Negative Samples
Because the complaint itself is highly subjective, there are more or less cases where data with the same characteristics appear in both positive and negative samples. In practice, if this part of data is not processed, it will lead to a serious impact on the prediction of complaints: the recall rate is too low or when the recall rate is only slightly increased, the accuracy rate will be greatly reduced. This requires us to make a trade-off when dealing with the classification of overlapping parts. Through statistical analysis and effect comparison, the overlapping data are redistributed in positive and negative samples.

1.5 Sample imbalance treatment
Positive and negative samples in this project (positive samples: the user's next work order is a complaint work order; Negative sample: the user's next work order is not a complaint work order) ratio is 1: 230, which has serious imbalance problem. After the above steps, the data of positive and negative samples are determined, and then the positive samples are expanded by SMOTE supersampling.

2. Model building
This project finally uses random forest as classifier to train the model.

2.1 Algorithm——Overview of Random Forests
Random forest is a supervised learning algorithm. Like its name, it creates a forest and makes it random in some way. The "forest" constructed is the integration of decision trees, and most of the time it is trained by the "bagging" method. Bagging method, that is the bootstrap aggregating. Uses randomly placed selection training data and then constructs a classifier, and finally combines the learned models to increase the overall effect.

In short: random forests have built multiple decision trees and combined them together to obtain more accurate and stable predictions. One of the great advantages of random forest is that it can be used for both classification and regression problems, which are currently faced by most machine learning systems.

The specific implementation steps are as follows:

2.1.1 Input as Sample Set:
D={(x,y1),(x2,y2),...(xm,ym)}

The number of iterations of the weak classifier is T, and the output is the final strong classifier: f(x).

2.1.2 For t = 1, 2 ..., t:
(1) Randomly sampling the training set for the t time and collecting m times in total to obtain a sampling set Dt containing m samples.
(2) The t decision tree model G_t(x) is trained with the sample set Dt. When training the nodes of the decision tree model, a part of sample features are selected from all sample features on the nodes, and an optimal feature is selected from these randomly selected part of sample features to divide the left and right subtrees of the decision tree.

In the classification algorithm prediction, the category or one of the categories in which the T weak learners cast the most votes is the final category.

2.2 Model Application
After determining the model algorithm, the processed data set can be trained. First, we randomly disturb the processed data set and divide it into training set and test set according to the ratio of 7:3. Secondly, the model is trained by using the data of training set.

3. Model Optimization and Model Effect

3.1 Model Optimization
There are many ways to optimize the model, and the specific choice depends on the analysis of actual problems. This section will mainly explain the ideas in the optimization process:
(1) New effective feature mining: after the initial features are constructed according to project understanding and business experience, three rounds of feature mining and construction (extracting potentially useful information from existing non-text fields, combining objective factors such as weather from text information, highlighting the role of sequence factors in complaint prediction) are conducted, and the features used in the final modeling are selected in combination with feature distribution map and correlation coefficient.
(2) Model parameter adjustment;
(3) Selection of algorithm and model: From traditional algorithm to integrated learning and then to neural network, compare the results, and select the model with the best prediction effect for the positive samples (current work order users will complain next time they call).

3.2 Model Effect
The effect of the final selection model is as follows (including accuracy rate, recall rate and confusion matrix display):
Figure 4. Presentation of Forecast Results

4. Conclusion
Judging from the prediction results, the effect of positive samples is not perfect especially the recall rate. This is because the difference between positive and negative samples is too large, and there will be data with exactly the same characteristics in the positive and negative samples. According to actual business requirements, we will improve the accuracy of positive samples as high as possible while ensuring that the recall rate cannot be too low. After many attempts and comparisons, the final model was selected.

References
[1] Wang Yizhe, Guo Quanhui, Liu Youfei, et al. 95598 Analysis and Suggestions on Complaint Work Order Data [J]. Jiangxi Electric Power, 2017 (10): 4 - 6.
[2] Luan Yuanyuan, Wang Zhongren, Xi Adan, et al. Research on Customer Complaint Early Warning Model Based on Improved BP Neural Network [C]// China institute of communications Academic Annual Meeting. 2010.
[3] Zhou Wenjie, Yang Lu, Yan Jianfeng. Big Data Driven Complaint Forecast Model [J]. Computer Science, 2016, 6 (7): 53 - 59.
[4] Wei Hongming. Research on the Prediction Model of Mobile Communication Customer Complaints Based on Data Mining [D]. Hengyang: Nanhua University, 2009.
[5] Liu Bing. Emotional Analysis: Mining Views, Emotions and Emotions [M]. Liu Kang, Zhao Jun, Beijing: Mechanical Industry Press, 2017.
[6] Yang Xujun, Hu Zhiyong, Zhang Chongqi, et al. Analysis of Complaint Early Warning and Early Intervention Based on Big Data Model [J]. Information Communications. 2016 (1): 266 - 267.