Research on the Loss of Catering Service Merchants in Colleges and Universities

Huiping Han*

China University of Geosciences (Beijing), Beijing, 100083, China

Abstract: In recent years, the loss of catering service merchants in colleges and universities has become a major problem restricting the development of catering. Therefore, it is an important task for catering managers to predict and analyze the lost merchants and take measures to improve the retention rate of old merchants and reduce operating costs. Based on the flow data of catering in Beijing universities, this paper uses a machine learning classification algorithm to predict the loss of merchants in two categories, selects the best classification algorithm, and uses this algorithm as a starting point to enable the upgrading of the catering business, consolidate the three-pronged education function of logistics service, management, and scientific research, and ensure the comprehensive development of the school.

Keywords: University catering; Merchant loss; Data algorithm

DOI: 10.12346/fhe.v4i4.8318

1. Introduction

In the literature review, it is found that there are much research on customer churn prediction based on traditional industries such as telecommunications, banking and insurance, but the research on the loss of catering merchants in colleges and universities is blank. The purpose of this paper is to use the method of combining industry experience and theoretical knowledge to define the innovation of merchant churn, combine the attribute data such as flow, evaluation, sales volume conversion, emotional destination generated in the operation of merchants as an effective feature set, and carry out experiments through machine learning classification algorithm, predict and analyze the loss situation, and help analyze and solve the problem of merchant churn.

2. Merchant Loss Prediction Framework

The research framework of merchant churn prediction is mainly divided into data processing, feature engineering and model prediction. Data processing includes data collection, cleaning, modeling, standardization and data dimensionality reduction [1], and the results are stored in the distributed database. Feature engineering includes feature processing, selection and construction. The model prediction is to select the best model through model training and model evaluation.

3. Data Sampling

The data of this study are from the flow of catering centers in 132 universities.

3.1 Definition of Innovation

According to the experience of the loss of catering merchants in colleges and universities, the Vintage analysis method and the migration rate analysis method commonly used in the prediction of user loss in the field of financial risk control are used as the innovative definition of the label period of merchant loss.

The vintage analysis method is used to determine asset quality, analyze account change rules, determine account maturity and analyze influencing factors. This study uses this method for reference to judge the loss of merchants. Make statistics on catering merchants from March to June and September to December 2020 to 2022. The shorter the continuous time of failing to complete the turnover, the smaller the turnover rate. On the contrary, the longer the time, the greater the turnover rate and the higher the possibility of misjudgment.

The mobility analysis method can vividly show the changing trend in the whole merchant life cycle. Assumption: If a merchant in a dangerous state fails to complete the turnover for N consecutive days, it will either complete or deteriorate to N + 1 days without completion after one day, and the calculation formula of migration rate is:

Merchant uncompleted rate = continuous N + 1 uncompleted status/continuous N uncompleted status

It has not been completed for 21 consecutive days, and the merchant migration rate tends to be stable. This time

node takes into account timeliness and accuracy.

Combining vintage analysis method, migration rate analysis method and experience judgment, it is finally determined that the label period of lost merchants is 21 days.

3.2 Characteristic Sampling Period

The common feature of merchant loss is that it is difficult to maintain operating costs due to the sharp decline in turnover. In order to make the characteristics of the loss prediction model cover the characteristic elements of merchant loss comprehensively, it is necessary to determine a reasonable characteristic sampling time period. After the statistics and analysis of the loss time history of all lost merchants in three years, this study found that the merchants with 16 weeks of loss accounted for 50%, and the merchants with 32 weeks of loss accounted for 70%. At present, the feature sampling period is initially selected as 32 weeks, that is, observe the behavior of merchants in the past 32 weeks to construct model features, minimize the sampling period and improve the prediction efficiency on the premise of ensuring the data sampling thickness. The historical data distribution of merchant loss is shown in Table 1.

Table 1. Merchant death history.

Death course (week)	quantity	Proportion of quantity	Cumulative proportion
4	18833	0.2	0.2
8	12265	0.13	0.33
12	9537	0.1	0.43
16	7205	0.08	0.51
20	5639	0.06	0.57
24	4586	0.05	0.61
28	4123	0.04	0.66
32	3946	0.04	0.7
36	3185	0.03	0.73
40	2559	0.03	0.76
44	2436	0.03	0.79
48	2194	0.02	0.81

3.3 Data Acquisition and Processing

This study uses big data technology for data processing operations. Through data extraction, conversion and loading, the data are synchronized to the database, the database model is designed, and the data are standardized.

There are many missing values and outliers in the extracted data. Data cleaning is needed to obtain a series of characteristic indicators based on the summary of merchant granularity and time dimension, which can be provided for data analysis and mining [2].

4. Model Establishment

4.1 Modeling Scheme

As shown in Figure 1, the data are divided by month, and the model is composed of characteristic sampling period, observation point and label period. The characteristic sampling period is 8 months (32 weeks), and the label period is 2 months (8 weeks). The required samples can be obtained by continuously sliding the time window forward.

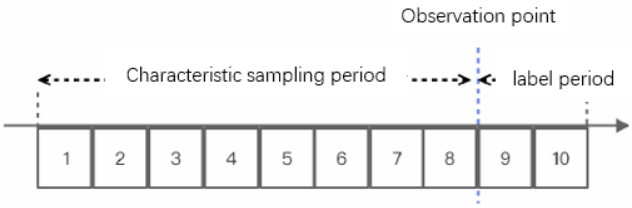

Figure 1. Characteristic sampling period and label period.

The longer the label period is, the more obvious the risk exposure of merchant loss is. The farther away from the observation point is, the older the historical data used to extract the sample characteristics will be, the higher the distortion will be, the greater the difference between the model sample and the future real sample will be, the greater the unknown information in the label period is, and the more difficult the model prediction will become. On the contrary, the shorter the label period is, the risk of merchant loss has not been fully revealed. The screened lost merchants may not be the real lost merchants, and the error is large, and the sample may not be the real sample. At present, the historical characteristic data sampled are the first 8 months, and the label period is the next 2 months. Considering the thickness of data sampling, the first business operation of new merchants is less than 8 months, which is not enough to support the loss prediction research. And because the new merchants are in the cold start stage, there are many uncertain factors in operation, and there are many accidental factors in the loss situation, which is not suitable for loss prediction. Therefore, this study is mainly aimed at the old merchants who have operated for 8 months.

Take lost merchants as positive samples and non-lost merchants as negative samples. Through data exploration, it is found that about 35% of merchants meet the definition of loss.

4.2 Experimental Data Set

The experimental data set covers the effective merchants since March 2020. Considering the sampling thickness, the new merchants with an operating period of less than 8 months are filtered out, and the total number

of old merchants is 800000. The experimental data set is divided into training set and test set according to the ratio of 7:3. After taking the merchants who failed to complete the turnover for 4 consecutive months as the definition of lost merchants, the data exploration found that about 30% of the effective merchants met the definition of lost merchants. If the lost merchants are regarded as positive samples and non-lost merchants as negative samples, the positive and negative ratio is 1:4. The data distribution is shown in Table 2.

Table 2. Division of experimental data set.

	Training Set	Test Set	Total
Positive sample	28000	12000	40000
Negative sample	112000	48000	160000
Total	140000	60000	

4.3 Classifier Experiment

The models used are logical regression, random forest, linear support and gradient upgrading [3].

4.3.1 Logical Regression II Classification

After constant interference of parameters, when the maximum iteration of the model is 100 times, the regularization coefficient is set to 1, and the minimum convergence error is 0.000001, the model effect is optimal. The critical threshold of the positive and negative samples selected in the experiment is 0.5. If it is greater than 0.5, it is positive. It determines that the merchant is in loss status, and vice versa. As shown in Figure 2, AUC value under ROC curve is 0.8726, K-S value is 0.6351, and F1 value is 0.7778.

4.3.2 Random Forest Classifier

After parameter adjustment, when the tree of the tree is set to 100, the minimum number of leaf node data is 2, the maximum depth of a single tree is 1000, and the number of random data input by a single tree is 100000, the prediction effect of the model reaches the best. According to Figure 3, the AUC value under the ROC curve is 0.9245, the K-S value is 0.686, and the F1 value is 0.8074, which is better than the result of the second classification of logistic regression.

4.3.3 Linear Support Vector Machine Classifier

In the prediction experiment of linear support vector machine model, the positive sample label value is set to 1, the positive case penalty factor is set to 1, the negative case penalty factor is set to 2, and the convergence coefficient is set to 0.001. As shown in Figure 4, AUC value under ROC curve is 0.8647, K-S value is 0.6204, and F1 value is 0.7686.

4.3.4 Gradient lifting decision tree classifier

After parameter optimization, when the number of trees is set to 500, the maximum leafy subtree is 32, the maximum depth is 10, the training collection sample ratio is set to 0.7, and the feature ratio is set to 0.7, the effect is optimal. As shown in Figure 5, AUC value under ROC curve is 0.9288, K-S value is 0.6939, and F1 value is 0.8122.

4.3.5 Effect Comparison

Through the comparison of experimental results, it is known that the gradient lifting decision tree algorithm is superior to the other three, and the best effect is used in the research of merchant loss prediction.

5. Ways and Means to Solve Problems

Colleges and universities should use holistic thinking, excavate educational characteristics, improve the organizational structure, analyze the loss of merchants from the material, spiritual, institutional and cultural dimensions, and study the corresponding countermeasures.

5.1 Material Level

Infrastructure construction should be in place and the spatial pattern should be optimized.

5.2 Spiritual Level

People-oriented, highlighting the strategic significance of employees, and enriching the business mission, vision and values. Subdivide merchants and provide targeted relationship maintenance services.

5.3 Institutional Level

Industry competitive salary and welfare should be in place. Personalized talent training programs and employee competency training should be in place. Introduce 360-degree evaluation to improve the performance appraisal cycle and optimize the application of performance appraisal results. Introduce diversified incentives and strengthen employees' spiritual incentives and shareholding incentives.

5.4 Cultural Level

Enrich campus activities, introduce inter-school activities and social activities, and enhance the cohesion and sense of belonging of employees.

Figure 2. Logical regression II classification.

Figure 3. Random forest.

Figure 4. Linear support vector machine.

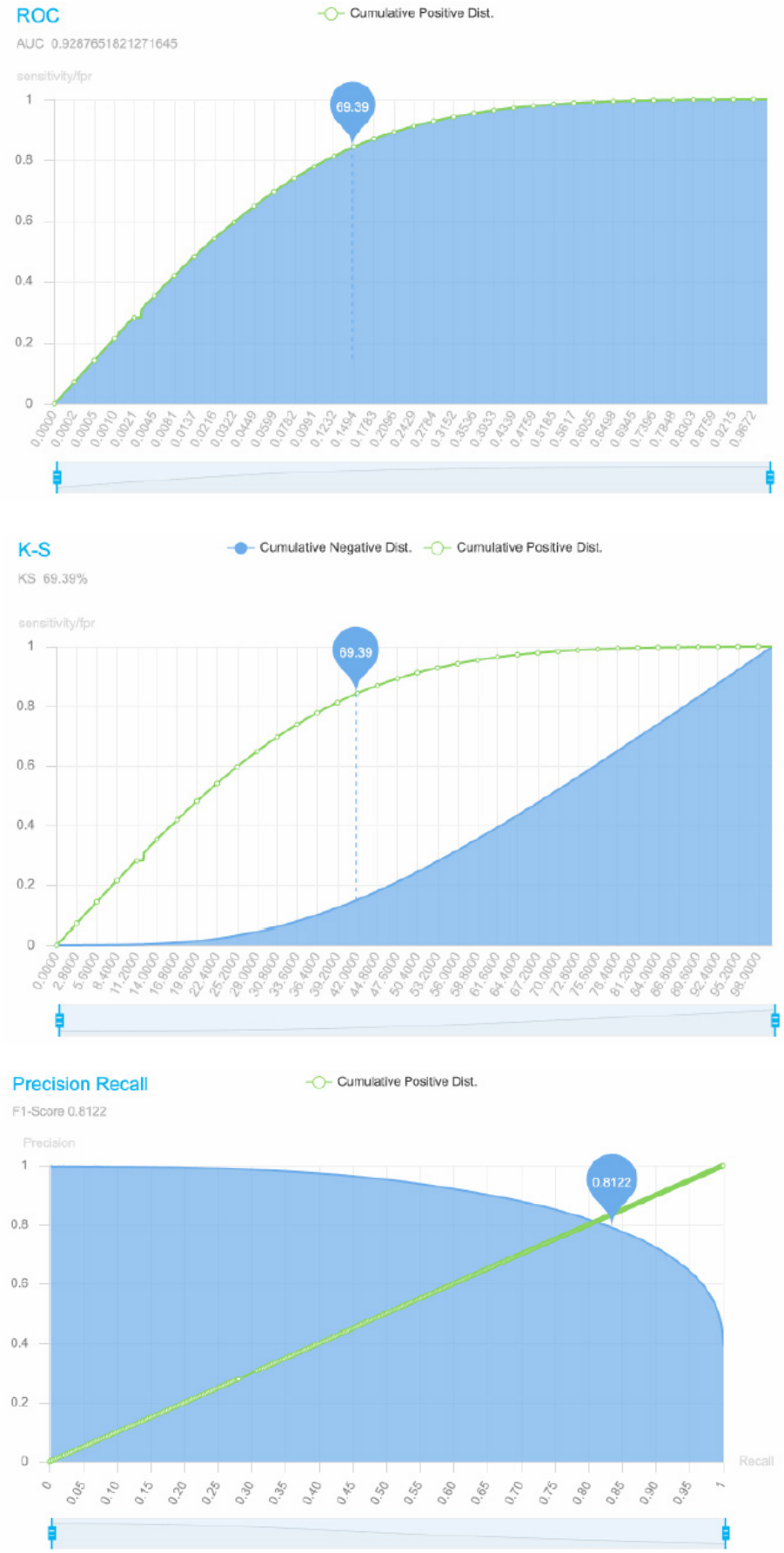

Figure 5. Gradient lifting decision tree.

6. Conclusions

In this paper, based on the data collected from the transaction flow, four models are used for experiments, and the gradient lifting decision tree algorithm is used to obtain the optimal prediction results, which improves the accuracy of the prediction results. In future research, consider introducing industry information into the gradient upgrading decision tree algorithm model. On the basis of prediction, build a loss attribution model, and combine prediction and attribution to form a closed-loop solution of prediction, attribution, intervention and return.

References

[1] Yu, L.P., Li, Y.F., 2018. World Bank Anomaly detection algorithm based on high-dimensional data flow. Computer Engineering. 44(1), 51-55.

[2] Lu, W.W., 2022. Internet service platform merchant loss prediction [Master's thesis]. Shanghai: East China Normal University.

[3] Lee, E., Jang, Y., Yoon, D.M., et al., 2018. Game data mining competition on churn prediction and survival analysis using commercial game log data. IEEE Transactions on Games. 11(3), 215-226.