Research on identification of poor students based on the consumption data of Campus Cards

Wei Yang[*], Junli Zhang[a] and Xiangxiu Yao[b]

School of Finance, Xi'an Eurasia University, Xi'an 710065, China

*Corresponding author: yangwei@eurasia.edu, [a]zhangjunli@eurasia.edu, [b]yaoxiangxiu@eurasia.edu

Keywords: Consumption data, Indicator construction, Identification of poor students.

Abstract: The biggest difficulty in targeted poverty alleviation on campus is the difficulty in accurately identifying poor students. This paper makes full use of campus card consumption data, on the basis of building a poverty indicator system, first uses PCA to screen the indicators, and then combines the adaboosting algorithm and the logistic regression algorithm to build a poor student identification model. Finally, the validity of the model is verified by the 2019 student consumption data of a college.

1. Introduction

Precision poverty alleviation on campus is an important part of the national precision poverty alleviation policy. The biggest difficulty in precision poverty alleviation on campus is that it is difficult to accurately identify poor students. The traditional identification of poor students needs to go through multiple processes such as the provision of certification materials by students, the confirmation and verification of each department of the school, and sometimes targeted home visits or investigations. The traditional identification of poor students needs to go through multiple processes such as the provision of certification materials by students, the confirmation and verification of each department of the school, and sometimes targeted home visits or investigations. At the same time, it is also easy to prove that the subsidy is untrue and the student refuses to apply for self-esteem.

The development of big data analysis technology provides a technical opportunity for campuses to accurately identify poor students, make full use of campus card data, especially consumption data, and build a perfect poverty indicator system. Through data mining models and algorithms, we can formulate more scientific data criteria to improve the accuracy of the identification of poor students.

2. Data preprocessing

In order to improve the accuracy of the research, before using the data for analysis and modeling, the collected data needs to be pre-processed such as desensitization, cleaning, and sorting. The processed data are shown in Table 1 and Table 2.

Table.1. Sample Table of Consumption Data

Time spend	student ID	Amount spend	Type use
2019/12/31 23:59:00	16612345088	0.36	Water use

Table.2. Sample Table of Data for Poor Students

Student ID	Poor student
16612345088	1 (True)
15612378649	0 (False)

3. Indicators construction

3.1. Construction of primary indicators

Because the time span of the data is the same, for each student's consumption data, we build indicators from the two dimensions of consumption days and consumption amounts. As student consumption is mainly for catering consumption, we also divide catering consumption into breakfast, lunch, dinner, and non-dinner hours according to dining time; Here, we also refer to other consumptions such as haircuts in the school, bathing, and daily necessities as non-catering consumption. The primary indicators constructed are shown in Fig. 1

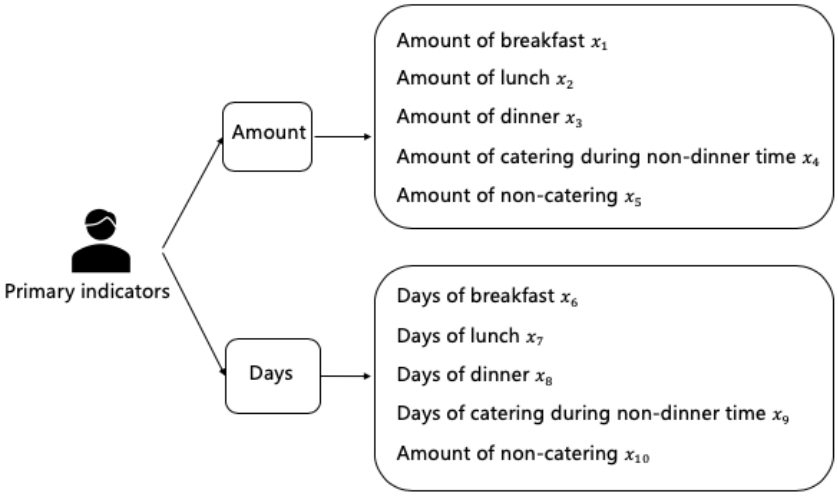

Fig. 1 Primary Indicators

3.2. Construction of intermediate indicators

On the basis of the primary indicators, in order to further explore the hidden poverty-related factors in the consumption data, the interval indicators based on the segmentation of consumption amount and the interval indicators based on the segmentation of transaction time were constructed. The calculation formula is shown in formula (1).

$$\begin{gathered} Rate_h = \sum Rate_{jh} \\ Rate_l = \sum Rate_{jl} \\ Rate_j = \frac{n_{\mathrm{inter}_j}}{n_i} \\ jh \in \{j | Rate_j(notpoor) - Rate_j(poor) \geq \delta_h\} \\ jl \in \{j | Rate_j(notpoor) - Rate_j(poor) \leq -\delta_l\} \end{gathered} \tag{1}$$

In formula (1), $Rate_h$ represents the proportion of high-range amount (or early-range time) of consumption and $Rate_l$ represents the proportion of low-range amount (or late-range time)of consumption.

$Rate_j$represents the proportion of each interval; n_{inter_j} represents the number of high or low consumption in the j interval; n_j indicates the total number in the j interval; jh means that it belongs to the high-range, jl means that it belongs to the low-range.

The intermediate indicators constructed are shown in Fig. 2.

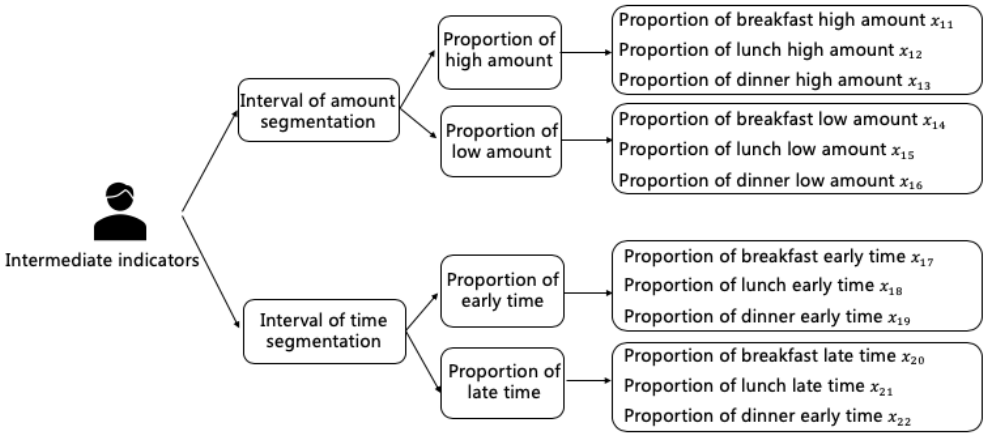

Fig. 2 Intermediate indicators

3.3. Construction of advanced indicators

Information entropy is a concept used to measure the amount of information in information theory. The more orderly a system is, the lower the entropy of information; conversely, the more chaotic a system is, the higher the entropy of information. The entropy of information reflects a measure of the degree of ordering of the system. Here, the theory of information entropy is introduced to reflect the regularity of student consumption amount and consumption time. The calculation formula is shown in equation (2).

$$Entropy_i = -\sum_j \frac{n_{inter_j}}{n_i} \log_2 \frac{n_{inter_j}}{n_i} \tag{2}$$

$$\begin{cases} n_i = \sum_j n_{inter_j} \\ interval \cdot j \le inter_j \le interval \cdot (j+1) \\ interval = (\max(expand) - \min(expand)) / k \end{cases}$$

$Entropy_i$ is the information entropy of the consumption amount of the i student, and $expand$ is the consumption amount. The consumption time entropy is as same as the consumption amount entropy. The advanced indicators constructed are shown in Fig. 3

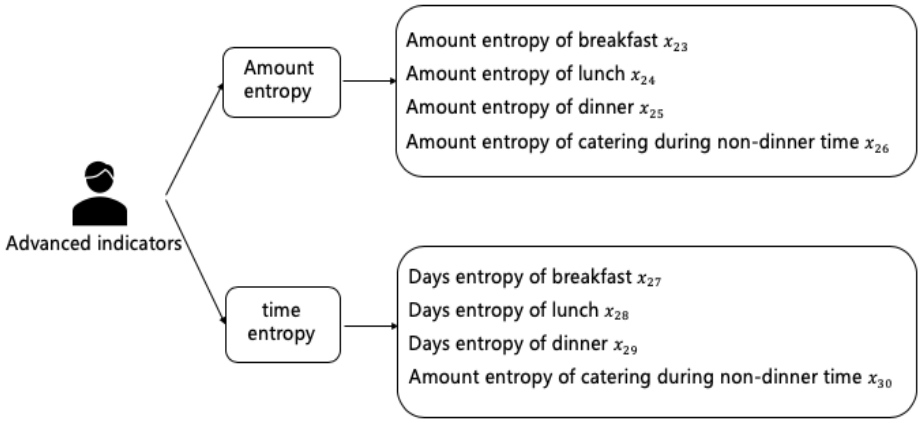

Fig. 3 Advanced Indicators

4. Poverty Indicators screening based on PCA

Since there are many indicators constructed, there may be problems of high information repetition rate among the indicators and large resource occupation, so the existing indicators need to

be screened. Here, the principal component analysis method is used to eliminate redundant indicators, and the correlation analysis is used for further indicators screening. poverty index screening.

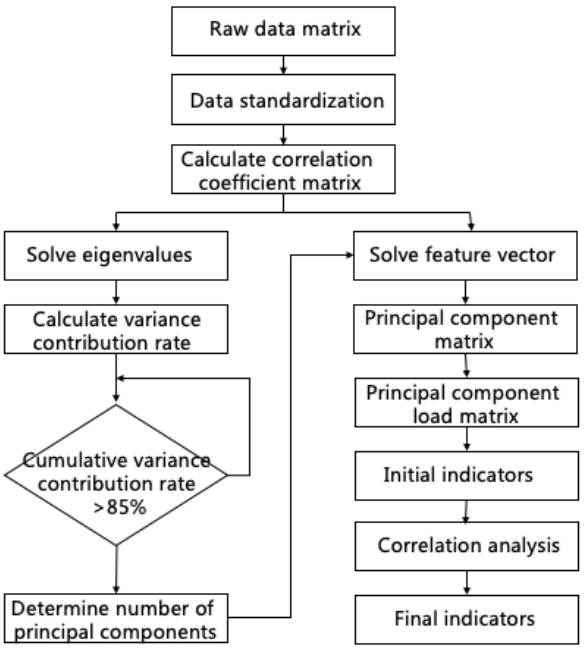

Fig. 4 Flow chart of poverty index screening

5. Accurate identification model of poor students

The flow of poor students identification model is as follows:

(1) The sample data is divided into training data and test data.

(2) The N decision trees are constructed as weak classifiers. The training data is used to classify the weak classifiers. When the classification error rate is less than the threshold ε, Strong classifier is obtained which is the linear combination of weak classifiers, and the test data is used to verify the classification effect of the adaboosting algorithm.

(3) The classifier is constructed based on the Logistic Regression algorithm to find the model that minimizes the loss function of the classifier, and the test data is also used to evaluate the classification effect of the model.

(4) In the adaboosting algorithm, let each tree calculate the vote rate for each student's poverty vote, and then get the poverty support rate of each student n/N,

(5) The poverty probability p of each student is obtained by the Logistic Regression algorithm;

(6) from each student's poverty support rate and poverty probability, the student's poverty index is calculated, the student's poverty level is graded, and the truly poor students are identified.

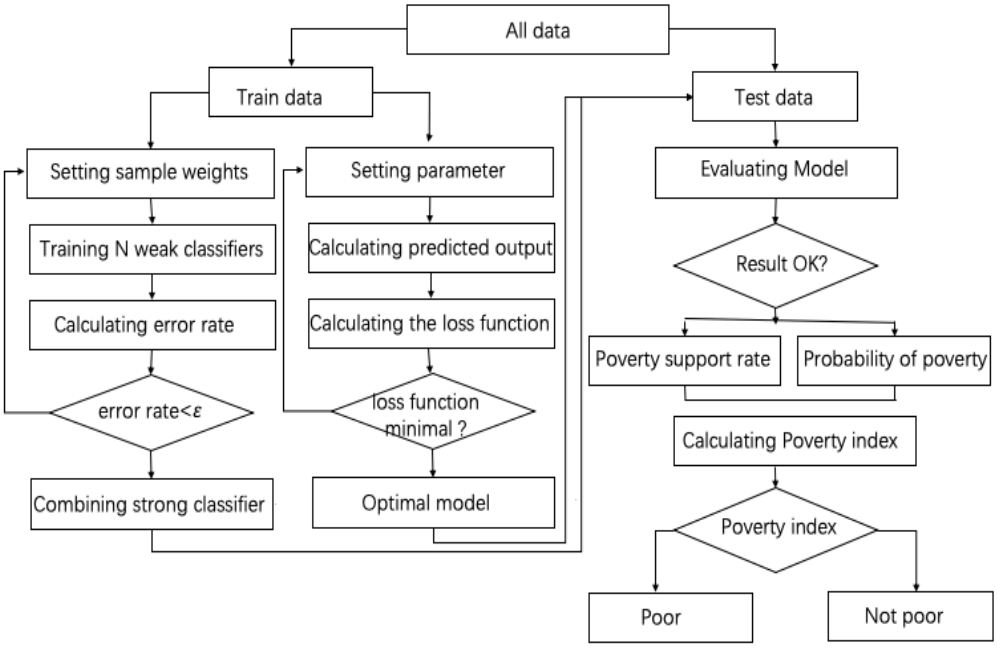

Fig. 5 Flow of poor students identification model

6. Example

This article selected the consumption data of a college student's campus card in 2019, with a total of 10411794 items, involving 20027 students. We also extracted data from those identified as poor students. There are 4655 poor students. From Table 3, the accuracy of the final identification of poor students is 86.83%. The research results are as follows.

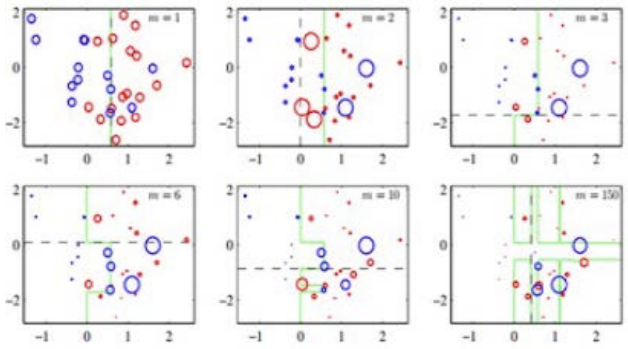

Fig. 6 Poverty support rate graph of some students

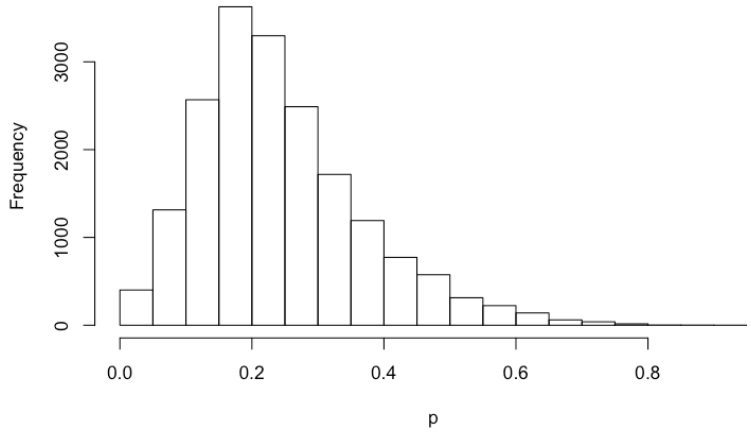

Fig. 7 Poverty probability histogram

Table.3. The final result of the accurate identification model

real / predict	poor	non-poor
poor	4042	3071
non-poor	613	12301

7. Conclusions

The research results can be used as a reference for schools to assess poverty and effectively improve the efficiency of identifying poor students. You can also optimize the model in a wider data dimension by adding other data such as basic information, access control, library, learning, family situation, etc. to improve its accuracy.

Acknowledgement

Research Fund Project of Xi'an Eurasian College (2018XJSK16): Student Behavior Analysis and Research Based on Campus Card Data.

References

[1] Luo Lilin. Research on the construction of college precise funding model in the perspective of big data [J]. Journal of Chongqing University (Social Science Edition), 2018, 24 (2): 197-204.

[2] Mu Yang, Zhang Yongfu. Reconstruction of the identification system for poor students in universities [J]. Journal of Northwestern Polytechnical University (Social Science Edition), 2017, 37 (1): 70-73.

[3] Song Dechang. Research on the evaluation method of student economic status based on campus card [J]. Journal of Sun Yat-sen University (Natural Science Edition), 2009, 48 (S1): 9-11.

[4] Duan Xumei, Hu Mengying, Zhong Junnan. Constructing a mathematical model for the identification of poor students in universities and its applications [J]. Journal of Jilin Education College, 2015, 31 (5): 150-151.

[5] Fan Bo, Jiang Yuguo. Design of auxiliary system for identification of poor students based on data mining [J]. Software Guide, 2015, 14 (12): 134-135.