Research on the Financial Difficulty Level Recognition of Needy Students Based Upon the Decision Tree C5.0

Xichun Luo1*, Linghui Kong1 and Jiongxian Wang1

1 Student Work Department, City College of Dongguan University of Technology, Dongguan, Guangdong, 518000, China

*Corresponding author’s e-mail: luoxc@ccdgut.edu.cn

Abstract Accurate recognition on the needy students is a core part of the college student subsidy and management work. Supported by the predictive capability of data mining model, this thesis studied the data sample of the recognition on needy students in a college. It selected 33 explanatory variables and one target variable through data pretreatment. Then by means of decision tree C5.0, the data sample was used to build the model. By calculation, the predictive accuracy of decision tree C5.0 was close to 90% on identifying the financial difficulty level of needy students. Therefore, it possessed excellent predictive effect. In addition, Kappa coefficient evaluation method was used to further prove that the model possessed favorable predictive effects. This study aimed to provide decision basis for subsidizing the needy students in colleges and universities, consequently improving the “targeted subsidy” work.

1. Introduction

Along with the smooth operation and rapid development in China’s social economy, it has been a significant mission in our national social structural reform to improve the existing system of higher education. With the in-depth reform on China’s current higher education system, the issues of needy students in colleges and universities show an upward trend. To solve these problems, the college executives are requested to handle the core issues about needy students by improving the student subsidy system in colleges and universities, ensuring the needy college students to accept education in peace, as well as promoting the comprehensive development of needy students in colleges. The ultimate approach to solve this core part of work is accurately identifying the subsidy objectives, then focusing on the needy student group and subsidizing them effectively.[1][2][3]Centered on this issue, the author used the data mining algorithm to build the predictive model so as to improve the accuracy and efficiency of targeted subsidy.

The data mining emerged as a new type of data analyzing technology. It refers to the process of exploring latent rules and extracting the knowledge from a great deal of data. Data mining is also known as Knowledge Discovery in Databases (KDD for short). The most fascinating part is, it can be used to build the predictive model. The decision tree in data mining is a kind of supervised learning method.[4] It constructs the classification trees via the probability of known events. Furthermore, the decision tree is widely used in various fields like finance, medicine, engineering and else. It possesses the intuitive, efficient, predictable and more other advantageous characteristics. In this thesis, the model was built to probe into the needy students through decision tree. Accordingly, it assessed the financial difficulty level of needy students and provided scientific support for solving the recognition issues in the work of subsidizing needy college students. [5]
2. Brief Introduction of Decision Tree C5.0

The decision tree C5.0 is a type of classification tree which was proposed by four American scientists in 1970. Until 1984, the study on this theoretical model had been mature basically. The basic idea of classification trees is that it divides the variable by two subsets in the light of certain classification rule. Then the two subsets are again divided into two parts. Each subset is known as the tree node. The division was undertaken repeatedly until it is appropriate. The ultimate subset is named as the end node. Through the study on the criterion of tree nodes, it can be used to predict a lot of practical issues.

During the process when classification tree developed, J.R. Quinlan proposed the ID3 algorithm in 1979. Then in the year 1983 and 1986, the algorithm was summarized and simplified by him. Consequently, it became a typical classification tree. Again, the C4.5 algorithm was developed in 1993. In recent years the C5.0 algorithm was raised. The basic idea of decision tree C5.0 algorithm is similar as the classification trees. It also used the variables collection to build the tree-shaped model and pruned it according to certain classification rules. [6][7]

(1) Structure of Trees

Faced with a multi-category training data set, the decision tree used the “IF” + “THEN” method to fulfill classification. It mainly referred to two aspects of issues: on the one hand, it aimed to solve how to select an optimal grouping variable for present situation from a collection of variables; on the other, it aimed to solve how to find out an optimal split point from so many values of grouping variables.

Guided by the information theory that was presented by C.E. Shannon in 1948, the information gain rates were the most appropriate grouping variables and the split point which were determined by the criteria.

Before the decision tree was built, the output variable was totally random for information sink. Its average uncertainty was:

\[ Ent(U) = \sum_i P(u_i) \log_2 \frac{1}{P(u_i)} = -\sum_i P(u_i) \log_2 P(u_i) \tag{1} \]

During the process when decision tree was built, it calculated the conditional entropy of each variable. In case the variable was X1, its conditional entropy was:

\[ Ent(U \mid X1) = \sum_j P(x_{1j}) (-\sum_i P(u_i \mid x_{1j}) \log_2 P(u_i \mid x_{1j})) \tag{2} \]

So the information entropy of X1 was:

\[ Gain(U, X1) = Ent(U) - Ent(U \mid X1) \tag{3} \]

The information entropy of each input variable was calculated. The result that had the maximum value was selected as the optimal grouping variable.

To find out the most appropriate split point from so many values in variables, if the variable was numerical variable, it used the ChiMerge method. Its basic idea was sorting the variables as per the values in ascending order; then defining the initial interval and separating the input variable values into some groups. Next it worked out the contingency table of the two adjacent groups of input variable and the output variable. On this basis, the Chi-square observed value was calculated. If the observed value was less than the critical value, the two intervals were supposed to emerge into one. Repeat the above procedure until all the Chi-square observed values were equal or larger than the critical value.

(2) Pruning of Trees

The key to tree pruning was mainly about the error estimation and the pruning criteria. The error evaluation used the evaluating method of confidence interval in statistics. It directly estimated the error in the training set:
\[
P\left(\frac{f_i - e_i}{\sqrt{\frac{f_i(1-f_i)}{N_i}}} < \frac{z_\alpha}{2}\right) = 1 - \alpha
\]

(4)

\[
e_j = f_i + z_\alpha \sqrt{\frac{f_i(1-f_i)}{N_i}}
\]

(5)

In it, \( z_\alpha \) was the critical value.

The default confidence degree of C5.0 was \( 1-0.25=75\% \). In case \( \alpha \) was 0.25, \( z_{0.25} = 1.15 \).

The criterion for pruning was established on basis of error estimation. C5.0 model judged whether to prune in accordance with the “reduction-error” method. It calculated the weighted error of the node in the sub-tree; then compared this error with that of the parent node. If the former was larger, the sub-tree could be cut off; or else, it couldn’t be cut off. The formula of weighted error was:

\[
\sum_{i=1}^{k} p_i e_i > e, i = 1, 2, \cdots, k
\]

(6)

3. Data Pretreatment

The data sources in this thesis stem from the collection of student family background information in a college. Thirty three explanatory variables and one target variable were filtered from these data. In them, the explanatory variables consisted of whether it was recorded financially difficult family on file, whether he or she was the destitute dependant, whether the household lived on subsistence allowances, whether he or she was the offspring of destitute worker, whether it was a low-income family with difficulties in the urban area and else. The target variable was difficulty level. It was a categorical variable which was classified into four groups: “no difficulty”, “general difficulty”, “much difficulty” and “particular difficulty”. 1888 pieces of data were retained after data pretreatment. The details of each interfering factor and assigned value were shown in Table 1.

| Serial Number | Type                  | Value Assignment Conditions |
|---------------|-----------------------|----------------------------|
| C1            | Is it a recorded financially difficult family on file? classification | A11: Yes A12: No |
| C2            | Is he or she the destitute dependant? classification | A21: Yes A22: No |
| C3            | Is it an urban or rural household that lives on subsistence allowances? classification | A31: Yes A32: No |
| C4            | Is he or she the offspring of destitute worker? classification | A41: Yes A42: No |
| C5            | Is it a low-income family with difficulties in the urban area? classification | A51: Yes A52: No |
| C6            | Is he or she an orphan? classification | A61: Yes A62: No |
| C7            | Is it impossible for the parents to fulfill their obligations of supporting their offspring? classification | A71: Yes A72: No |
| C8            | Is he or she raised by single parent? classification | A81: Yes A82: No |
| C9  | Does he or she enjoy the preferential treatment? Is he or she the offspring of a police officer who died while on duty? Is the student disabled individual? Does the student have any major illness? The number of supported family members The number of supported elderly people The number of family members who lost their jobs The employment status of parents The educational background of parents The age of parents The number of labor force in family The family size The number of students in family Is there any patient with major diseases in family? Is there any parent who lost the job or lost the ability to work? Is he or she the offspring of the disabled The disability grade of the student’s father The disability grade of the student’s mother The debt amount of family income The annual per capita household income The type of major household income source | classification | A91: Yes A92: No A101: Yes A102: No A111: Yes A112: No A121: Yes A122: No | }
4. The Structure and Analysis of the Financial Difficulty Level Recognition Model on Needy Students

4.1 Structure and Analysis of Decision Tree C5.0 Model
Firstly the author drew 80% samples at random from the data sample set. The selected samples were used as training set to build the model. The balance 20% samples were used as the testing set to validate the model. Stratified sampling method was used to ensure the samples to be more representative. The training set and the testing set were drawn out as per the proportion of 4:1 from the four difficulty levels of data sample sets respectively; then they were merged separately as the training set and the testing set. By this way, it could prevent too many samples from belonging to single classification in the training set that would occur by direct sampling.

Use the training set to build the model and then use the testing set to validate the model. Accordingly, the classification results of the built decision tree model versus the training set and the testing set were shown in Table 2 and Table 3.

Table 2 Predictive results of training set by the decision tree C5.0

| Actual classification | Predictive Classification | Total | Accuracy | Error Rate |
|-----------------------|--------------------------|-------|----------|-----------|
| No difficulty         | Predictive                |       |          |           |
| General difficulty    | Much difficulty          | Particular difficulty |       |           |
| No difficulty         | 82                        | 2     | 1        | 0         | 85        | 96.47% | 3.53% |
| General difficulty    | 2                         | 609   | 25       | 5         | 641       | 95.01% | 4.99% |
| Much difficulty       | 0                         | 22    | 346      | 12        | 380       | 91.05% | 8.95% |
| Particular difficulty | 0                         | 0     | 5        | 383       | 388       | 98.71% | 1.29% |
| Total                 | 84                        | 633   | 377      | 400       | 1494      | 95.04% | 4.96% |

Table 3 Predictive results of test set by the decision tree C5.0

| Actual classification | Predictive Classification | Total | Accuracy | Error Rate |
|-----------------------|--------------------------|-------|----------|-----------|
| No difficulty         | Predictive                |       |          |           |
| General difficulty    | Much difficulty          | Particular difficulty |       |           |
| No difficulty         | 36                        | 1     | 0        | 0         | 37        | 97.30% | 2.70% |
| General difficulty    | 0                         | 129   | 10       | 3         | 142       | 90.85% | 9.15% |
| Much difficulty       | 1                         | 7     | 87       | 8         | 103       | 84.47% | 15.53% |
| Particular difficulty | 0                         | 0     | 1        | 111       | 112       | 99.11% | 0.89% |
| Total                 | 37                        | 137   | 98       | 122       | 394       | 92.13% | 7.87% |
From Table 2 and Table 3, it indicated that,

(1). Regardless of training set or testing set, the general predictive accuracy of decision tree C5.0 both achieved 90% and above. Therefore, it showed that the model had great predictive effect. It can perfectly be used to provide decision-making support in college student subsidy work.

(2). In terms of the predictive accuracy of particularly difficult classification, both of the training set and the testing set possessed more than 98% accuracy. What’s more, even the misjudgment cases were classified into the much difficult category. The effect of this error on needy students was almost negligible.

(3). The statistics tables showed that, the number of needy students who were misjudged as having no financial difficulty was only two out of the total 1409 needy students. In the test set of 357 needy student samples, only one case was misjudged. It indicated that, by using the data mining algorithm, the probability of missing the subsided needy students was very low. Despite all this, the ancillary mechanism is also supposed to be set up so as to improve the model.

4.2 Kappa Coefficient Evaluation

Kappa coefficient evaluation method was normally used to assess the predictive effect of such multi-classification model. In statistics, Kappa coefficient is a kind of method which evaluates the consistence. It is also frequently used to assess the predictive effect of multi-classification models. The formula of Kappa coefficient is:

$$Kappa = \frac{P_o - P_e}{1 - P_e}$$

In this formula, $P_o$ represented the quotient from the sum of sample that had coincident real categories and predictive categories dividing the number of total samples; $P_e$ was calculated by multiplying the number of real samples and the sum of predictive samples in each category respectively, then dividing the square value of the total number of samples. 

The Kappa coefficient evaluation method was used to assess the confusion matrix of the testing set. It obtained the kappa=0.890 and again proved the model to be an excellent tool in prediction.

5. Conclusions

This thesis built the model of decision tree C5.0 in accordance with the recognition on the needy students from a selected university in Guangdong province. Based upon the mature decision tree C5.0 model, in case a latent student who has financial difficulties needed to be identified, the model can evaluate his or her difficulty level by the family information that was provided by the student. It proved that the predictive results possessed relatively higher accuracy. Therefore, this model is able to provide decision support on college students’ financing work, and thus improving the “targeted subsidy work”.

References

[1] Xiao, F., X., Feng, X., L. (2020). The Purpose, Principle and Strategy of the Targeted Financial Aid for the Poor College Students. Modern Education Management. (3), 117-122

[2] Hong, L. (2018). The History, Problems and Precise Path of Subsidization System for Poor Students in Colleges and Universities. Journal of Educational Science of Hunan Normal University. 017(005), 103-109.

[3] Huang, W., Li, F., Liao, X., & Hu, P. (2017). More money, better performance? the effects of student loans and need-based grants in china’s higher education. China Economic Review. 208-227

[4] Davis, R. H., Edelman, D. B., & Gammerman, A. J. (1992). Machine-learning algorithms for credit-card applications. IMA Journal of Mathematics Applied in Business & Industry. 129-137

[5] Chen, X., Wang, S., L., Li, J., J., Zhang, Z. (2014). Applying Weighted Constraints-Based Decision Tree Method to Impoverished Students Identification. Computer Applications and Software. (12), 136-139
[6] Quinlan, J. R. (1993). C4.5: Programs for machine learning. Morgan Kaufmann. 81-106
[7] Pang, S., L., Gong, J., Z. (2009). C5.0 classification algorithm and its application on individual credit score for banks. Systems Engineering-theory & Practice. (12), 94-104
[8] Zhang, J. B., Jiang, Z., Y., Fang, J., Wang, Z., Y., Cao, M., & Ma, X., Y. (2019). Evaluation of Consistency for HbA1c Kits by Using Kappa Statistic and Bland-Altman Method. Mathematics in Practice and Theory. 49(20):167-175