Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 11, No. 2, 2023

Study of Bursary Grade Prediction Based on Integrated Learning

Qian Niu, Xiaotong Ye

Sichuan University of Science & Engineering, Yibin, 644000, China

Abstract: With the advancement of digital campus, many scholars have addressed the problems of low efficiency of financial aid work in colleges and universities and insufficient fairness and impartiality of the evaluation process, and proposed the solution of using big data and machine learning algorithms to identify the students who are really in need of help and realise the fairness of bursary awarding. Firstly, multi-dimensional feature vectors are constructed by processing and analysing students' daily behavioural data, then Multi-SMOTE oversampling technology is used to solve the problem of balancing classification samples in the process of bursary evaluation, and then Stacking model fusion algorithms are used to integrate multiple classifiers in order to improve the prediction accuracy. The experimental results prove that the method is significantly better than a single classification model in terms of precision and recall, which can improve the efficiency of college financial aid work, as well as ensure the fairness and accuracy of the financial aid process and provide better financial aid services for students.

Keywords: Bursary evaluation, Multi-SMOTE, Stacking.

1. Introduction

Although the reform of information technology has improved the quality of life of human beings, the phenomenon of exponential growth of data generated by human life cannot be ignored, and these seemingly insignificant data in fact contain enormous value waiting for human beings to explore and explore. Colleges and universities as the birthplace of big data technology, in fact, there are many aspects of the data has not yet been integrated and analyzed, the identification of poor students and financial assistance is one of the important content of the student work department, but the assessment of grants is still mainly carried out offline, with a large number of people involved, the process is complex, subject to the human factor has a greater impact on the identification of the fairness and impartiality of the difficult to ensure that the identification of justice and fairness, so through the behavior of the students at school to predict the economic situation of the students, it is very difficult to ensure that the student's behavior in the school. Therefore, it is more urgent to predict the economic situation of students through their behavior in school and build a student financial aid management system.

With the continuous improvement of campus informatization construction, the function of campus one-card is more and more powerful, and some scholars began to use data mining technology to analyze the potential value and information in one-card. Yunchuan Chen[1] et al. analyzed the campus one-card consumption data to predict students' family difficulties by using decision tree algorithm and plain Bayesian algorithm, and after comparing the results of the two models, they found that the plain Bayesian algorithm is better in terms of accuracy and precision, and the decision tree algorithm is better in terms of recall rate and AUC. Jie Ding et al. used matrix theory and cloud data platform to improve the association rules, and analyzed to get the consumption behavior characteristics of poor student groups by classifying the consumption level of students, which provided the basis for the identification of poverty in colleges and universities. Chunzi Tian[2] et al. take the one-card data as the source of data analysis, use K-Means and DBSCAN algorithm to mine the value information in the students' one-card data, and analyze the characteristics of the students' consumption behaviors during the school period, which provides help for the school's management mode. Ligan Gong[3] et al. In order to analyze the characteristics of consumption behavior, using the Spark framework, using feature engineering to construct the clustering features in the campus one-card data, to provide data support for colleges and universities to carry out accurate financial aid or psychological counseling. Zefeng Han[4] et al. proposed a Consume2Vec model based on deep neural network to mine the consumption data of school students, and established a SACD model for detecting abnormal consumption information, and finally obtained the consumption patterns and characteristics of students.

It can be seen that at present, the identification of poverty grants in colleges and universities is more and more inclined to the level of big data. Scholarship determination work is complicated and complex, manual poverty determination work is often with a certain degree of subjective thinking, and the authenticity of the identification materials submitted by the students is also debatable, does not exclude some students in order to obtain the scholarship quota and exaggerate the degree of poverty in the family, which may make some of the students who should be subsidized do not get the subsidy. Reading the literature, we found that more and more scholars have begun to use the campus card to help carry out the work of scholarship recognition, because the card contains students' daily consumption data, these data can not be faked, from which we can dig out the students' consumption behavior, and we can more objectively and accurately judge whether the students are poor or not, and then according to the students' grades, daily access to libraries and the results of the previous year's scholarships, we can more accurately assess the students' poverty. The data can be mined from these data to find out students' consumption behavior, which can more objectively and accurately judge whether a student is poor or not.

2. Relevant Theory

2.1. Multi-SMOTE based over-sampling approach

Multi-SMOTE is an improved oversampling algorithm based on SMOTE and Borderline-SMOTE, and its main idea is to generate new minority samples by using neighboring minority samples with multiple neighboring majority samples. It introduces a new weighting method to determine the "majority" category of each minority sample by calculating the distance of each minority sample from each majority sample[5] . The minority samples from each category are then oversampled using the SMOTE algorithm to generate new synthetic samples.

Unlike the traditional SMOTE method, Multi-SMOTE selects the appropriate neighborhood size according to the density of the samples in each category when dealing with a few categories of samples, and then generates new samples within each category sample set separately, and finally combines all the generated samples into a new balanced dataset.The advantage of Multi-SMOTE is that it can deal with multiple categories of imbalance problems, avoiding the new problems introduced when dealing with each category's imbalance problem separately. In addition, it takes into account the correlation between the samples and generates more realistic samples.

2.2. Stacking Integrated Learning Approach

The Stacking algorithm is a model fusion algorithm, and its basic idea is to input the prediction results of multiple base models as new features into a meta-model to get more accurate prediction results.

The steps of the Stacking algorithm are:

For the training set data in each base learner, K-fold cross-validation is performed to obtain n feature matrices, and the information from these n matrices is merged to obtain the new training set dataD_1 ;

While performing the cross-validation, the prediction of the test set is performed to get the K kinds of prediction data learned by a base learner, which is averaged as the final prediction result of the base learner on the test set. Similar to step 1, we will get the feature matrices of n training sets, and the information of these n matrices will be merged to get the new test set dataD_2 ;

The training dataset will beD_1 to train the meta-classifier. Then the test dataset is utilizedD_2 and the trained meta-classifier, the test set is predicted to get the final classification result[6] .In Python's machine learning module, each classifier calculates its probability for each category of the sample, and then those categories with the highest probability are used as the output.The flowchart of the Stacking algorithm is shown in Fig. 1.

3. Experimental Design and Performance Analysis

3.1. Data acquisition

The data for this experiment was obtained from DataCastle competition network, which has been fully desensitized because it involves private data. The dataset includes a test set and a training set, both of which contain a one-card consumption data table, a library borrowing data table, a library access data table, a dormitory access data table, a student achievement data table, and a bursary acquisition data table.

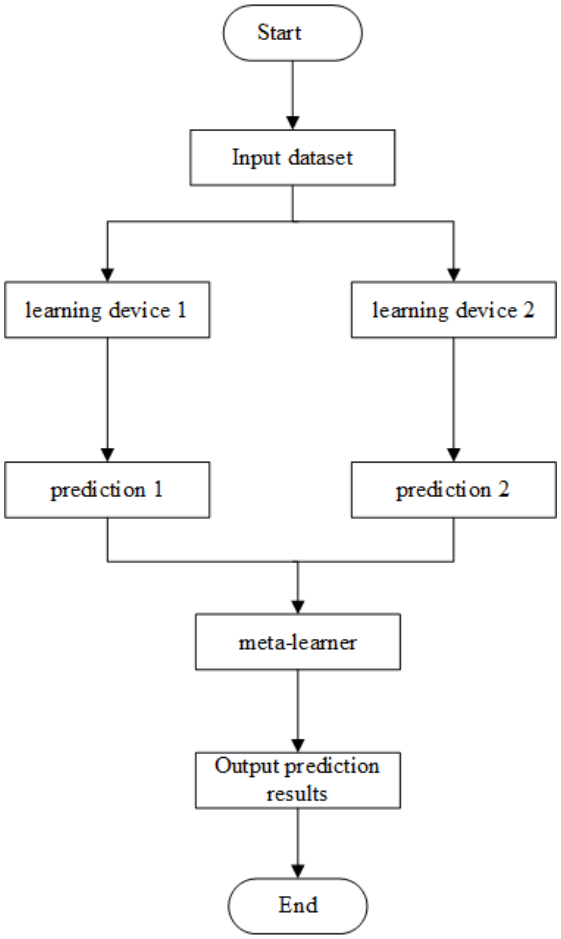

Figure 1. Flowchart of Stacking algorithm

3.2. Feature engineering

3.2.1. Feature construction

First of all, the assessment of student performance needs to follow a unified standard, in the student performance table there is a column of data for student rankings, which is an important reference value for understanding the learning level of students. However, the ranking of students is also associated with the number of students in the college, the ranking of students in two different colleges is directly compared, not comparable, so the construction of a column of "ranking percentage" feature to eliminate the differences in the number of colleges due to the inaccurate assessment of student learning.

Secondly, the data in the book borrowing table and library access table can also be used to corroborate the students' usual student status, in which the number of times each student borrows books is mainly extracted from the book borrowing table, and the number of times each student enters the library per year is mainly extracted from the library access table.

In addition, the most important criterion for judging whether a student can get a scholarship is the economic situation, and this aspect is mainly determined by the student's consumption in school. The consumption characteristics of students are constructed by counting the total number of times each student consumes, the total amount of consumption, the maximum amount of single consumption, the minimum amount of single consumption, and the average amount of single consumption in the data table of the one-card.

3.2.2. Missing value processing

The presence of null values in the data set will have an impact on the accuracy of the experimental results, and it is necessary to go to fill in the null values, so as to make the data more complete, usually based on the principle of statistics, based on the distribution of other data in the data set on the missing values to fill in. For some students' data in the data set have too high missing rate, or some samples have less relevance to the research results, the samples in this category can be deleted directly.

3.2.3. Category feature coding

For non-numerical features are numerical using one-hot coding, and then the feature data are normalized to eliminate the magnitude differences between features, the normalization process is shown in Eqs. (1)-(3).

$$t'_{ij} = \frac{t_{ij} - AVG_j}{\sigma_j} \tag{1}$$

$$AVG_j = \frac{1}{n}(t_{1j} + t_{2j} + \cdots + t_{nj}) \tag{2}$$

$$\sigma_j = \sqrt{\frac{1}{n}\left[(t_{1j} - AVG_j)^2 + (t_{2j} - AVG_j)^2 + \cdots + (t_{nj} - AVG_j)^2\right]} \tag{3}$$

where, t_{ij} is the eigenvalue, normalized to t_{ij}', the eigenmean is AVG_j, and the eigenstandard deviation is σ_j.

3.2.4. Data imbalance treatment

In this experiment, due to the nature of the bursary assessment question, the majority of the population was not awarded a bursary, and therefore the results in the bursary level classification were bound to be skewed, resulting in an inflated precision in the classification result of 0 for the amount of bursary received. Where the distribution of bursaries is shown in Figure 3, the distribution of sample sizes is very unbalanced, with only about 17% of the total sample receiving financial assistance. Due to the limited number of total samples, in order to solve this sample imbalance problem, the experiment used the oversampling method to increase the sample size of a few categories.

The oversampling process resulted in a significant increase in the sample size for each grant amount category. There were originally 5,272 in the training set who did not receive a grant, 504 who received $1,000, 325 who received $1,500, and 267 who received $2,000, and the oversampling process increased the samples of those who did not receive a grant, those who received $1,000, those who received $1,500, and those who received $2,000 by a factor of 5, 8, and 10, respectively.

In order to maintain a balanced distribution of data and to ensure the accuracy and reliability of the model evaluation, the training set was divided into five stratified subsets using a stratified sampling technique to divide the cross-validation set. In the testing phase of each fold, one of the subsets will be used as the test set while the other subsets will be used as the training set. In this way, all the samples can be used for training and evaluation of the model, effectively assessing the generalization ability of the model and reducing the bias introduced by the uncertainty of sample selection.

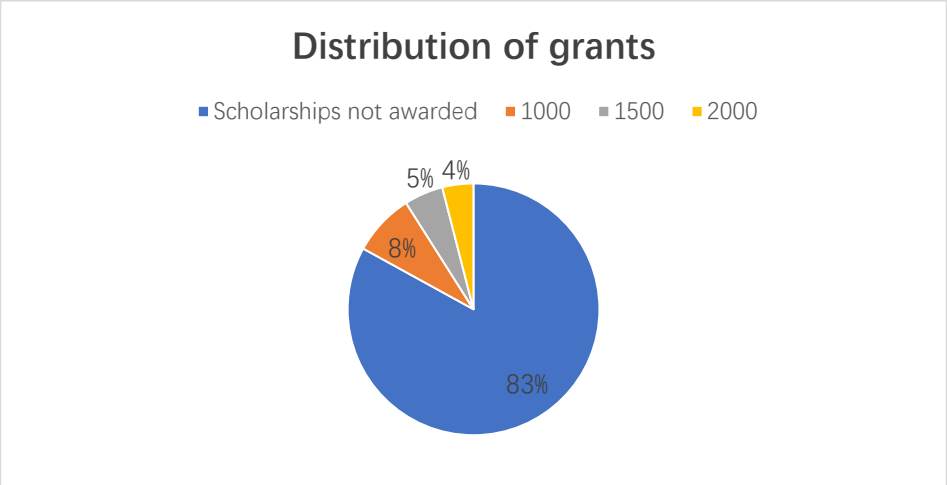

Figure 2. Distribution of grants

3.3. Evaluation criteria

Confusion matrix is a tool used in machine learning to compare the prediction results of classification models[7], each row in the matrix represents the predicted value and each column represents the true value, taking the dichotomous classification as an example, the elements of the confusion matrix are shown in Table 1.

Table 1. Confusion matrix elements

	positive	negative
positive	TP	FP
negative	TN	FN

The precision and recall in the classification metrics are based on the confusion matrix with the following expressions:

$$Precision = \frac{TP}{(TP+FP)} \tag{4}$$

$$Recall = \frac{TP}{(TN+FN)} \tag{5}$$

For multiclassification problems, the F1 of a subclass can be calculated based on the precision and recall of each subclass:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

The F1-score ranges from 0 to 1, and its value depends on the precision and recall, and only if both perform well can the imbalance data be evaluated in a more reasonable way.

4. Experimental Results and Analysis

The results are subjected to five-fold cross-validation, and the results are shown in table 2 below, which shows that integrating Random Forest, KNN, XGBoost, CatBoost, which are weak classifiers, as the base learners of the Stacking framework for integrated learning has better results than the individual models in the model evaluation criteria Precision, Recall, and F1 -score.

Table 2. Results of different classification models

Model	Precision	Recall	F1 -score
KNN	0.7096	0.8424	0.7703
RandomForest	0.7239	0.8062	0.7596
XGBoost	0.8909	0.8512	0.8729
CatBoost	0.9291	0.7532	0.8219
Stacking	0.8846	0.8802	0.8823

5. Summary

This paper centers on the grade prediction problem of college student grants, firstly, the dataset is imbalanced, then the original dataset is feature engineered to screen out the features that have an important influence on the prediction of grant grades, and finally, the Stacking integrated model is constructed with Random Forest, KNN, XGBoost, and CatBoost as the base learners, and the results show that the Stacking integrated model has a significant improvement in precision, recall , and F1-score in these evaluation criteria by comparing it with the other single models. Comparison, the results show that, in a comprehensive view, the Stacking integrated model has a significant improvement in the evaluation criteria of precision, recall , F1 -score.

References

[1] Yunchuan Chen,Hao Song,Ye Zhao et al.Data Mining and Application of Campus card Based on Logistic Regression Algorithm [J]. Journal of Kunming Metallurgy Higher College, 2020, 36(03):57-61.

[2] Chunzi Tian,Wan Yang,Dehui Yang et al. Based on K-Means and DBSCAN clustering algorithms according to the context of student behaviour analysis and research based on comprehensive university data[J]. Science and Technology Innovation,2020(32):86-88.

[3] Lixing Gong,Kun Gu, Xinming Ming et al. Analysis of college students' consumption behavior based on campus card data[J]. Journal of Shenzhen University (Science and Technology Edition),2020,37(S1):150- 154.

[4] Zefeng Han, Tao Yang, Linlin Hou et al. Consume2Vec model-based analysis of campus card big data [J]. Computer Application,2020,40(S1):85-91.

[5] Zhongqiang Liu, Weiwei Zou. Anomaly Detection Model for Consumer Electricity Usage Based on Sampling Technique and LightGBM[J].Computer System Applications, 2021, 30(09):232-236. DOI: 10.15888/j.cnki.csa.008157.

[6] Miaofang Lin. Spectral classification algorithm based on BP neural network with Stacking model fusion[J]. Modern Information Technology, 2022, 6(04):91-94+99. DOI:10.19850/j.cnki.2096-4706.2022.04.024.

[7] Qinli Yu, Haizheng Yu. Credit risk assessment model based on improved SMOTE adaptive integration [J]. Journal of Chongqing University of Technology(Natural Science), 2022, 36(07): 293-302.