Research on Influencing Factors of College Evaluation Results Based on Machine Learning Model

Guorui Zhao[1](✉), Tingting Tian[2], Qiwen Li[3], and Shuosa Zeng[3]

[1] Department of Basic Courses, Guangdong Ocean University, Yangjiang, Guangdong, China
zhaoguoruimath@foxmail.com

[2] College of Material Science and Engineering, Guangdong Ocean University, Yangjiang, Guangdong, China

[3] Business College, Guangdong Ocean University, Yangjiang, Guangdong, China

Abstract. This paper introduces the machine learning model into the research of determining the factors on college evaluations. Taking the "double-high colleges" of the Ministry of Education as a research case, compared with support vector machines, decision trees and other models, Lasso-Logistic can more efficiently compress and select the key explanatory variables. The overall prediction accuracy of the model is nearly 80%. The results show that the landmark achievements of majors, teachers, and students are the core influencing factors, and factors such as school establishment time, number of teachers, student-teacher ratio and other factors have no substantial impact.

Keywords: college evaluation results · influencing factors · determination · Lasso-logistic model

1 Introduction

In the field of higher education, the government has upgraded the construction of "double first-class" universities and "double-high" colleges to the level of strategic development. In 2017 and 2019, the first batch of "double first-class" and "double high" colleges and universities were selected. How to determine the degree of influence of various factors on the evaluation results and clarify its internal mechanism is currently a hot issue in the fields of educational technology and educational evaluation, especially in the field of Education Data Mining (EDM). Its research provides a reference for colleges to implement precise policies.

In this research field, Cui et al. (2017) conducted an in-depth analysis on the evaluation standards and systems of world-class universities [1]; Kong et al. (2019) analyzed the principles and methods of "double first-class" university evaluation [2]; Yu et al. (2020) conducted a research on the construction of a "double first-class" university evaluation data platform [3]; Lin (2020) made a systematic analysis on the design of the evaluation index system for "double high" colleges and universities [4]; Chen et al. (2020)

F. Huang et al. (Eds.): ICAIE 2023, AHCS 15, pp. 4–14, 2024.
https://doi.org/10.2991/978-94-6463-242-2_2

conducted a research on the evaluation index system of the ranking of higher vocational colleges [5]. Judging from the existing literature, in terms of research objects, there are more and more in-depth researches on "Double First-Class", and less research on "Double High". In terms of research methods, there are more qualitative researches, but less quantitative researches. Through preliminary research, we found that there are few articles using machine learning models to study this issue.

Lasso (Least Absolute Shrinkage and Selection Operator), as a model developed in the field of machine learning in recent years, is widely used in bioinformatics, medicine, economics and other fields. Among them, representative studies include: Fang et al. (2014) introduced Lasso into the field of bank personal credit risk assessment [6]; Zhang et al. (2018) introduced Lasso into the field of constructing a national well-being evaluation index system [7]; Sun et al. (2019) introduced Lasso into the field of predicting the graduation destination of college students [8]. However, there are few studies on applying Lasso to education, especially in the field of EDM.

The contribution of this paper is mainly in three aspects: First, in the construction of the "double high" evaluation index system, we traced the source, sorted out the declarations of all colleges, and sorted out all the indicators. Second, most of the Lasso studies use a single algorithm to analyze, and often only have the results of variable selection, but the process is often ignored. We fully display the analysis process to support the rationality of the Lasso model. Third, on the basis of variable selection, we analyzed the factors with significant influence and provided relevant policy recommendations.

2 Mechanism Analysis of Lasso-Logistic Model

As an algorithm that combines variable selection and parameter estimation, Lasso was first proposed by Tibshirani [9] in 1996, and its mechanism is as follows:

Given data $D = \left(X^j, y_j\right), j = 1, 2, \cdots n$, where $X^j = \left(x_{j1}, \cdots x_{jm}\right)$ and y_j are denoted as the explanatory and explained variables, $\hat{\beta} = \left(\hat{\beta}_1, \cdots, \hat{\beta}_p\right)^T$ are the coefficients. Consider the simplest linear regression model, and its optimization goal is

$$\hat{\beta} = \arg\min\{\sum_{j=1}^{n}(y_j - \sum_{i}\hat{\beta}_i x_{ji})^2\} \tag{1}$$

When there are few samples and many variables, the model is prone to over-fitting. In order to alleviate the problem of over-fitting, norm regularization can be introduced.

$$\hat{\beta} = \arg\min\left\{\sum_{j=1}^{n}\left[(y_j - \sum_{i}\hat{\beta}_i x_{ji})^2 + \lambda\sum_{i}|\beta_i|\right]\right\} \tag{2}$$

(2) represents the penalty for the coefficient, λ is the adjustment coefficient that controls the degree of compression of each variable. The coefficient of unimportant variable is compressed to 0 through the change of λ to adjust the selection of the variable. The smaller the λ, the smaller the punishment, the more variables will be retained; the larger the λ, the greater the punishment will be, and the fewer variables will be retained. In

terms of model solving, Efron et al. introduced the minimum angle regression algorithm in 2004, and the Lasso model can be solved more efficiently [10].

For the solution of λ, this paper uses the lars package in R language, combined with Mallow C_p criterion and generalized cross-validation. Choose s from p independent variables for regression,

$$C_p = \frac{SSE_p}{s^2} - n + 2p \tag{3}$$

where $SSE_p = \sum_{j=1}^{n} \left(y_j - \hat{y}_j\right)^2$, we get the minimum value λ of C_p through continuous iteration based on Mallows C_p criterion.

Then Logistic regression is performed on the remaining variables of Lasso compression. Suppose the number of compressed variables is m, and the explanatory variable y_j is a binary 0–1 variable. Set $P = P\left(y_j \middle| X^j\right)$, then

$$\log \frac{P}{1-P} = X^T \beta^{Lasso} = \beta_0 + \sum_{j=1}^{m} \beta_j x_j \tag{4}$$

3 Data Description

The data of the paper comes from the 2019 application materials for high-level vocational colleges and professional construction plans with Chinese characteristics (referred to as the “Double-High”). We collected a total of 230 samples of colleges. Among this, fifty-six colleges including Shenzhen Vocational and Technical College were included in the construction of high-level vocational colleges, 141 colleges including Beijing Agricultural Vocational College were included in the construction of high-level groups of majors, and 33 colleges were not selected.

In order to avoid problems of asymmetric data distribution, we use the variable whether to be selected as a high-level higher college as the explained variable Y (0− No, 1− Yes). In addition, a secondary index system was constructed based on the application materials. Among them, the explanatory variables include 2 primary indicators and 50 secondary indicators. Table 1 shows the specific grading index system. To verify the prediction effect of the subsequent mathematical model, the sample data set was randomly divided into training set and test set according to the ratio of 8:2.

4 Empirical Analysis

We use the Lasso-logistic model to analyze the influencing factors of Y. Using the lars package in the R language, the harmonic parameter values λ are selected through the cross validation method (CV). The trend of the estimated value of λ based on CV is shown in Fig. 1, where the saturation that minimizes the mean square error is between 0 and 0.2. Because it is random grouping, the difference of each grouping leads to different results of λ. And the value of λ is different, the degree of model compression

Table 1. Index system of explanatory variable classification

Group	***symbol***	***Description***
Basic status and basic conditions	x_1	School establishment time (years)
	x_2	Area of teaching, research and auxiliary rooms per student (m^2/student)
	x_3	Student dormitory area per student (m^2/student)
	x_4	The total value of school fixed assets (ten thousand yuan)
	x_5	Per student value of teaching and scientific research equipment (yuan/student)
	x_6	Total number of school staff (person)
	x_7	Proportion of full-time teachers with double teacher quality (%)
	x_8	Teaching hours of part-time teachers (class hours)
	x_9	The proportion of part-time teachers' teaching hours to the total number of majors (%)
	x_{10}	Number of students in full-time general higher vocational education (person)
	x_{11}	Number of students at the starting point of secondary vocational school (person)
	x_{12}	Number of overseas students (persons)
	x_{13}	Equivalent to the number of students in school (person)
	x_{14}	Number of cooperative enterprise orders (persons)
	x_{15}	Number of courses jointly developed by cooperative enterprises and schools (courses)
	x_{16}	Number of intern students accepted by cooperative enterprises (persons)
	x_{17}	Proportion of employment of graduates accepted by cooperative enterprises (%)
	x_{18}	The total building area of the school building (m^2)
	x_{19}	Area of laboratories and practice places per student (m^2/student)
	x_{20}	Number of paper books per student (volumes/student)

(*continued*)

Table 1. (*continued*)

Group	***symbol***	***Description***
	x_{21}	Total value of teaching and scientific research equipment (ten thousand yuan)
	x_{22}	Internet access bandwidth (Mbps)
	x_{23}	Number of full-time teachers in the school (person)
	x_{24}	The total number of part-time teachers in the 2017–2018 school year (person)
	x_{25}	2017–2018 academic year total professional class hours (class hours)
Basic status and basic conditions Landmark achievement	x_{26}	Non-academic training scale (person)
	x_{27}	Number of students at the starting point of ordinary high school (person)
	x_{28}	Number of students enrolled in the two years after the five-year higher vocational school (person)
	x_{29}	Number of other students in school (person)
	x_{30}	Student-teacher ratio (x:1)
	x_{31}	Number of part-time teachers in schools supported by cooperative enterprises (person)
	x_{32}	Number of textbooks jointly developed by cooperative enterprises and schools (types)
	x_{33}	Cooperative companies accept employment of 2018 graduates (persons)
	x_{34}	The total value of equipment donated by the cooperative enterprise to the school (ten thousand yuan)
	x_{35}	Number of national-level vocational education professional teaching resources database
	x_{36}	The number of national education and teaching reform pilots
	x_{37}	Number of national key majors
	x_{38}	Typical number of employment and entrepreneurship in the country
	x_{39}	Number of national honors for teachers
	x_{40}	Number of National Vocational College Skills Competitions Organized

(*continued*)

Table 1. (*continued*)

Group	***symbol***	***Description***
	x_{41}	Number of national teaching achievement awards
	x_{42}	Number of awards won in the National Vocational College Teaching Ability Competition
	x_{43}	The number of students' national and above competition awards
	x_{44}	The number of national teaching achievement awards won in the past two sessions
	x_{45}	Presided over the number of national professional teaching resource database projects
	x_{46}	The number of national education and teaching reform pilot projects undertaken
	x_{47}	Number of key majors in national demonstration and backbone higher vocational schools
	x_{48}	School employment work was rated as a national employment and entrepreneurship model
	x_{49}	Teachers have won national awards
	x_{50}	Hosted the National Vocational College Skills Competition in the past five years

Note: The indicator time period is 2017–2018. In the qualitative variable, "1" represents "yes" and "0" represents "no"

will also change, and the number of variables selected by the model each time will also be affected. Tibshirani believes that when the mean square error of the model is small, we generally choose λ that makes the model relatively concise. In addition, in order to ensure the stability of the model, we repeatedly set different random numbers, perform CV 10 times, and take the mean value of parameter λ. The average value of λ is calculated to be about 0.101.

Figure 2 shows the path of the variable coefficient along with the selection of the harmonic parameter λ. When λ takes the minimum value, only x_{37} is selected. As λ increases, x_{41}, x_{35}, etc. are selected into the model accordingly. When λ is close to 1, all 50 variables are selected into the model. x_{21}, x_{35}, x_{37}, x_{38}, x_{41}, x_{41}, x_{42}, x_{43}, x_{46} and x_{47} are selected into the model based on the ideal values $\lambda=0.101$. At this time, logistic regression is performed on the variables after Lasso compression, and the parameter estimates are shown in Table 2.

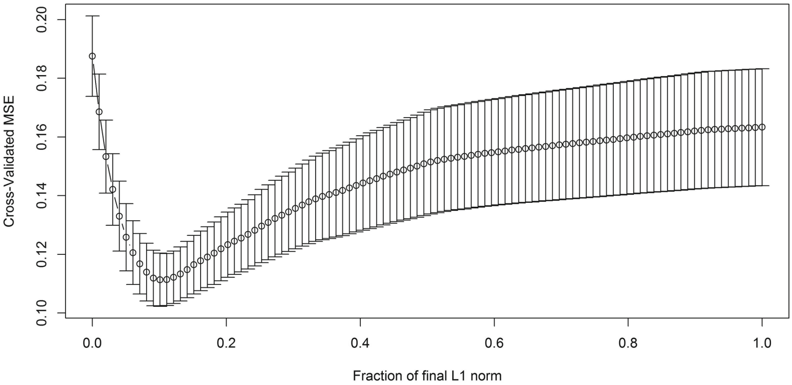

Fig. 1. Reconciliation parameters and corresponding trend graph

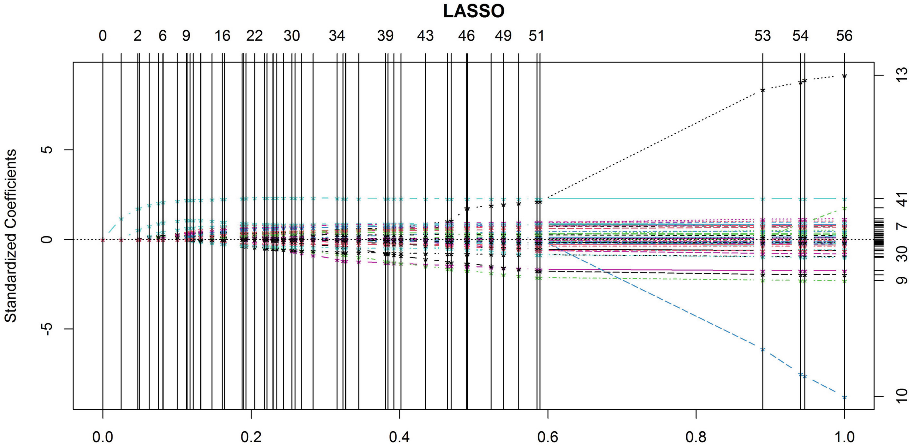

Fig. 2. Path of the Lasso coefficient

Due to too many explanatory variables, logistic regression and stepwise logistic regression models are too complex. The two algorithms does not converge and overfitting. Therefore, we only show the parameter estimation results of the Lasso-Logistic model, as shown in Table 2, from which the regression Eq. (3) can be obtained. Among them, P is the probability of being selected. Then

$$\begin{aligned}\log it(P) = \ln \frac{P}{1-P} = 2.99 + 0.021x_{21} + 1.738x_{35} + 3.831x_{37} \\ +1.44x_{38} + 1.293x_{41} + 0.271x_{42} + 0.074x_{43} + 1.471x_{46} + 2.141x_{47}\end{aligned} \tag{5}$$

We introduce support vector machines, decision trees, random forests and other models to compare the prediction accuracy of selected and unselected double-high colleges on the training set and test set. The details are shown in Tables 3 and 4.

Table 2. Lasso-Logistic model parameter estimates

Explanatory variables	***Description***	***Regression coefficient***	***Standard error of the regression coefficient estimates***
x_{21}	Presided over the number of national professional teaching resource database projects	0.021*	0.011
x_{35}	Number of national-level professional teaching resource libraries for vocational education	1.738**	0.002
X_{37}	Number of National Key Programs	0.383*	0.237
x_{38}	Number of national employment and entrepreneurship models	1.44*	0.608
x_{41}	Number of national-level teaching achievement awards	1.293**	0.425
x_{42}	Number of awards in the National Vocational College Teaching Ability Competition	0.271**	0.116
x_{43}	Number of student awards at national level and above	0.074**	0.027
x_{46}	Number of national education teaching reform pilots undertaken	1.471*	0.766
x_{47}	There are key majors in national model and backbone higher vocational schools	2.141**	1.228

Note: ** and * indicate significant regression coefficients at the 5% and 10% levels, respectively

From Tables 3 and 4, it can be seen that on the training set, the overall accuracy of each model exceeds 90%, but for the "selected group", the accuracy of the Lasso-logistic model is much higher than that of support vector machines and decision trees, second only to Random forest. On the test set, the overall accuracy of the Lasso-logistic model is the highest, reaching 79%. Especially in the "selected group", the accuracy is much higher than that of support vector machines and random forests, second only to decision trees. In addition, the Lasso-logistic model compresses most of the variables and is less complex than other models. Secondly, the Lasso-logistic model is more interpretable.

Table 3. Comparison of prediction accuracy rates of various models on training set (unit: %)

Models	***Training set accuracy***		
	Selected	***Not selected***	***Overall***
Lasso-logistic	0.823	0.946	0.919
Support vector machines	0.734	1	0.941
Decision Trees	0.7	0.852	0.9
Random Forest	1	1	1

Table 4. Comparison of prediction accuracy rates of various models on test set (unit: %)

Models	***Test Set Accuracy***		
	Selected	***Not selected***	***Overall***
Lasso-logistic	0.543	0.857	0.798
Support vector machines	0.224	0.673	0.752
Decision Trees	0.562	0.812	0.783
Random Forest	0.24	0.952	0.741

Based on the above Lasso-Logistic model analysis, we get the following results:

First, there are a total of 34 explanatory variables in the basic state and basic condition group. Lasso selected variable x_{21} for the total value of teaching and scientific research equipment, and it was significant at a significance level of 10%. It shows that x_{21} plays an important role in the selection of “double-high”, and has a statistically significant impact. The variable represents an important manifestation of the college’s ability.

In this group, other explanatory variables didn’t have substantial effect. For example, the time when the school was established (x_1), the number of full-time teachers in the school (x_{23}), the student-teacher ratio (x_{30}), etc. were not selected.

Second, there are a total of 16 explanatory variables in the iconic achievement group. Lasso selected 8 variables. They are the number of national teaching resource banks (x_{35}), the number of national key majors (x_{37}), the number of national employment and entrepreneurship models (x_{38}), the number of national teaching achievement awards (x_{41}), the number of awards won in the national vocational college teaching ability competition (x_{42}), the number of students at national level and above competition awards (x_{43}), the number of national education reform pilots (x_{46}), the number of national demonstrations, and the number of key majors in key higher vocational colleges (x_{47}). Among them, $x_{35}, x_{41}, x_{42}, x_{43}$ and x_{47} are significant at the 5% significance level, which shows that the teaching results of national majors, teacher competitions, and student competitions are the key influencing factors for the selection of “double high”colleges.

5 Conclusion

This paper introduces the Lasso-Logistic model into the research on the factors affecting the evaluation results of colleges in the field of educational data mining, and explores the factors and mechanisms behind the evaluation results. The main research conclusions of this paper are as follows:

First, in terms of specific suggestions, higher vocational colleges should not excessively pursue the number of students and teachers and the size of the school. National-level majors, teacher competitions, and student competitions are the core factors for the selection of double-high colleges. Colleges should lay a solid foundation and do a good job in the construction and accumulation of such landmark achievements. Second, from the perspective of research methods, Lasso can more effectively compress and select key variables, reduce the complexity of the model, and the overall accuracy is better than other algorithms. Therefore, it is more reasonable and scientific to use Lasso to study the determination of multi-class influencing factors.

In short, it is an extremely beneficial attempt to introduce machine learning models, especially sparsity algorithms such as Lasso, into the field of educational data mining.

Acknowledgment. This work is supported by Guangdong Province Philosophy and Social Sciences 2020 Disciplinary Co-construction Project, the Guangdong Provincial College Student Innovation and Entrepreneurship Training Program Project.

References

1. Cui Y, Li J, Pei X, Wan M.(2017) On the Construction and Perfection of the Evaluation System for the Construction of World-class Universities in my country [J]. Academic Degrees & Graduate Education,(11):23–29.
2. Kong X, Zhou C.(2019)Developmental Principles and Methods of “Double First Class” Construction Evaluation [J]. Jiangsu Higher Education, (12):55–61.
3. Yu B, Zhao R, Wang X, Li D. (2021) Research on the construction of a tracking evaluation platform for “double first-class” universities driven by big data [J]. Journal of Chongqing University, 27(02): 122–132.
4. Lin C. (2020) Research on the Performance Evaluation Index System of Vocational Education “Double High Plan” Project [J]. Journal of Ningde Normal University (Philosophy and Social Sciences), (02):92–109+118.
5. Chen B, Gao W. (2020) Research on the Evaluation Index System of Higher Vocational College Ranking [J]. Chinese Vocational and Technical Education, (13): 89–96.
6. Fang K, Zhang G, Zhang. (2014) Individual Credit Risk Prediction Method: Application of a Lasso-logistic Model [J]. The Journal of Quantitative & Technical Economics, 31(02): 125–136.
7. Zhang X, Zhong W, Hong Y.(2018)Analysis of Influential Factors of Civil Happiness in China: Based on LASSO Screening Method [J]. Statistical Research,35(11):3–13.
8. Sun Y, Pan K, Sun Z, He Z.(2019)Ideas and Methods for Predicting the Whereabouts of Graduates of College Students [J].Education Research Monthly,(01):25–35.

9. Tibshirani R.(1996), Regression Shrinkage and Selection via the Lasso [J],Journal of the Royal Statistical Society, B,1 (58),267–288.
10. Efron B, Hastie T, Tibshirani J R .(2004) Least Angle Regression[J]. Annals of Statistics, 32(2):407–451.