Study on the Individual and Environmental Factors of College Adaptation of Freshmen: Research Based on Machine Learning

Peng WANG, Yan-lu WANG, Xiang-ping ZHAN*, Jun WANG, Yun YAN, Chen-hui ZHANG and Wei-xuan MENG

No. 1500, the University Road, The university technology garden in Changqing District, Jinan, School of Psychology, Shandong Normal University, 250358 China

*Corresponding author

Keywords: Freshmen, School adaptation, Predictive factors, Machine learning.

Abstract. In this study, we used the method of multiple regression and machine learning (Ridge regression and Lasso regression) to investigate the influence of individual and environmental factors on school adaptation at different time. The results shows that: (1) College students’ adaptation is negatively correlated with shyness, loneliness, Internet addiction and non-adaptive cognition; meanwhile positively correlated with self-awareness, active coping style, self-esteem, interpersonal trust and social support. (2) Lasso regression of machine learning can effectively reduce dimensions and the number of variables, further simplify the model to strengthen the explanation. (3) The key factors of college students’ adaptation are shyness, coping style, self-esteem, loneliness, Internet addiction and social support.

Introduction

School adaptation is a multidimensional concept with a wide range of meanings. At present, there is no unified definition in the academic field, but most scholars agree that school adaptation is an interaction state between students, school environment and school activities [1]. When enter the university, due to sense of powerlessness and confusion without goals, freshmen may have various problems in thinking, life, psychology, study and other aspects; at the same time their physical and mental health may be affected to different degrees [2]. In the present study, the predictors of school adaptation can be roughly classified into three categories. The first is subjective factors, namely individual factors. The Second is objective factors, namely environmental factors. The third is the way of interaction between individual and environment, such as coping style [3].

In our study, high-dimensional data containing 10 variables are involved. It is found that high-dimensional data often have more features, most of which are irrelevant with variables to be explained. At the same time, there are strong correlations between some features, resulting in collinearity [4], which often affects the interpretation of dependent variables. Therefore, dimensionality reduction is generally required for high-dimensional feature data. Machine learning aims to use computers to simulate human learning activities, which is a remedy to identify existing knowledge, acquire new knowledge, continuously improve performance and realize self-perfection [5]. In recent years, with the development of machine learning, many researchers of psychology have tried to use machine learning algorithms. Li et al. (2015) found that it was completely possible to calculate the psychological characteristics of netizens, such as personality characteristics and mental health, with the data of microblog users [6]. Zhu et al. (2015) pointed out that for many
aspects of current psychological research, big data and machine learning technology could directly improve efficiency and enhance validity of the research [7].

Therefore, we adopted machine learning method to effectively process high-dimensional data to investigate the influences of numerous individual and environmental factors on school adaptation at different time. College adaptation is taken as the dependent variable, and other factors involved are taken as independent variables. Meanwhile, college students' test scores at the end of semester are added as the reference variable.

**Method**

**Participants**

In the first test, 800 freshmen were selected by cluster sampling method, and 750 valid questionnaires were finally collected, with an effective rate of 93.75%. Participants are from two universities in Shandong province of China. There are 426 participants from A university, 324 participants from B university, 262 males and 488 females (\(Mage= 18.23, SD= 1.06\)). One year after the interval, 250 college students were selected for the second follow-up test, with 229 valid questionnaires, the effective rate of which was 91.6%, including 105 males and 124 females (\(Mage=19.21, SD= 0.74\)).

**Measurement**

a) Chinese College Student Adjustment Scale

b) Scales of individual factor

    College Students' Shyness Scale, Self-consciousness Scale, Self-esteem Scale, and Loneliness Scale

c) Scales of environmental factor

    Interpersonal Trust Scale and Social Support Scale

d) Scales of individual and environment interaction

    Simple Coping Style Scale, Chinese Internet Addiction Scale, and Non-adaptive Cognitive Questionnaire

**Data Analysis**

SPSS 20.0 and Amos17.0 were used for data statistical analysis, and machine learning python program was used for data dimension reduction, Ridge regression and Lasso regression.

**Result**

**Correlation Between School Adaptation and Its Predictors of Freshmen**

Pearson product-moment correlation analysis of college students’ adaptation and other variables showed that college students' adaptation is significantly negatively correlated with shyness (\(r= -0.311, p< 0.001\)), loneliness (\(r= -0.454, p< 0.001\)), Internet addiction (\(r= -0.313, p< 0.001\)) and non-adaptive cognition (\(r= -0.228, p< 0.001\)); while significantly positively correlated with self-consciousness (\(r= 0.170, p< 0.001\)), simple coping style (\(r= 0.070, p< 0.01\)), self-esteem (\(r= 0.085, p< 0.05\)), interpersonal trust (\(r= 0.081, p< 0.05\)) and social support (\(r= 0.247, p< 0.001\)).
Analysis of the Predictors of Freshmen’s College Adaptation

Regression Analysis. The total score of college student adaptation scale is the dependent variable, and its 9 influencing factors are independent variables. According to the scatter diagram, it is suitable for linear regression, and multiple linear regression equation can be obtained. We named shyness X1, self-awareness X2, simple coping X3, self-esteem X4, interpersonal trust X5, loneliness X6, Internet addiction X7, network adaptive cognition X8, social support X9, and total adaptation score Y. The regression equation is:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 \] (1)

Standardized regression coefficients can be obtained (see Table 1):

<table>
<thead>
<tr>
<th>Variable</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
</tr>
</thead>
<tbody>
<tr>
<td>coefficient</td>
<td>-0.22*</td>
<td>0.063*</td>
<td>0.202*</td>
<td>0.274**</td>
<td>0.056*</td>
<td>-0.154*</td>
<td>-0.064*</td>
<td>-0.171*</td>
<td>0.322**</td>
</tr>
</tbody>
</table>

Adjusted \( R^2 \) is 0.518, and the constant is 118.7. All the standardized coefficients are significant.

Machine Learning Method. Through the above methods, it can be known that when using multiple linear regression model, the regression equation with good results is obtained, but the key factors affecting college students' adaptation cannot be mined. We took a completely new approach to solving this problem using machine learning as the statistical and processing methods. According to the previous studies, we adopted the following two methods to realize dimensionality reduction of high-dimensional data in this study: Ridge regression and Lasso regression.

Ridge Regression. By using python (GUI), inputting the code, selecting the file path, outputting the result, and selecting the best threshold value K, we can get the standardized regression coefficient value of the regression equation (see Table 2). It should be pointed out that Ridge regression does not contain constant.

<table>
<thead>
<tr>
<th>Variable</th>
<th>X1</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
</tr>
</thead>
<tbody>
<tr>
<td>coefficient</td>
<td>-0.259*</td>
<td>0.09**</td>
<td>0.248*</td>
<td>0.051*</td>
<td>-0.21**</td>
<td>-0.175*</td>
<td>-0.188*</td>
<td>0.264**</td>
</tr>
</tbody>
</table>

Adjusted \( R^2 \) is 0.522, and all the standardization coefficients are significant. According to the equation, the independent variable X2 self-consciousness has disappeared in the process of dimension reduction, indicating that in the overall model of Ridge regression modeling, X2 self-consciousness is not an explanatory variable. We realize the original explanation of variable dimension reduction, but as a result of the determination of optimal threshold K has a great deal of thought conceptions, some unimportant coefficient is not directly compressed to 0, so the model of explanatory is bad [8], so it is necessary to adopt a more optimized method.

Lasso Regression. In this method, the absolute value function of the model coefficients is used as a penalty to compress the model coefficients and make some regression coefficients become smaller, or even make some coefficients with small absolute values become 0 directly. The results of our analysis are shown in table 3.
Table 3. The standardized regression coefficients of Lasso regression.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$X_1$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_6$</th>
<th>$X_7$</th>
<th>$X_9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>coefficient</td>
<td>-0.219*</td>
<td>0.239**</td>
<td>0.193*</td>
<td>-0.160*</td>
<td>-0.137*</td>
<td>0.133**</td>
</tr>
</tbody>
</table>

Adjust $R^2$ is 0.593, all the standardized coefficients are significant, and the constant is 106.4. In this regression equation, it is concluded that variables $X_2$ self-awareness, $X_5$ interpersonal trust, and $X_8$ non-adaptive cognition disappear in the process of feature selection and dimension reduction. Lasso regression directly reduces the coefficients of some independent variables that contribute little to the dependent variables to 0 in the operation process, realizing the purpose of dimension reduction and avoiding the subjective arbitrariness of human selection. It is an ideal way of dimension reduction and a good method to realize the key features of data mining.

Thus, we obtain the variables, which contribute the most to the dependent variable of adaptation in the final method. In order: shyness, coping style, self-awareness, loneliness, Internet addiction, social support.

The Verification and Comparison of Three Methods. We put the original data of the two participants into the regression equation, and conducted Pearson product-moment correlation analysis between the predicted value and the second result. The size of the correlation coefficient reflected the quality of the results. It can be clearly seen from the results in Table 4 that the number of explanatory variables is gradually decreasing, and the error of the model keeps decreasing with the reduction of variables. From multiple regression to Ridge regression to Lasso regression, after removing the irrelevant variables, the predicted value and the actual second correlation coefficient are gradually increased according to the regression equation.

Table 4. Comparison of three methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Dimensionality Reduction</th>
<th>Reserved Variables</th>
<th>$SE$</th>
<th>$R^2$</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLR</td>
<td>NO</td>
<td>9</td>
<td>25.527</td>
<td>0.518</td>
<td>0.474</td>
</tr>
<tr>
<td>RR</td>
<td>YES</td>
<td>8</td>
<td>24.736</td>
<td>0.522</td>
<td>0.533</td>
</tr>
<tr>
<td>Lasso</td>
<td>YES</td>
<td>6</td>
<td>24.158</td>
<td>0.593</td>
<td>0.617</td>
</tr>
</tbody>
</table>

Further Exploration of Longitudinal Study. As a result, we can add the time factor in the independent variables. We selected college students who took the test twice as participants. Time of the first test is T1 and the second is T2. Half of the participants were selected as the training set and the other as the verification set, and the regression equation was established by Lasso regression method. Table 5 shows the results of the standardized regression coefficient of Lasso regression.

Table 5. The standardized regression coefficient of Lasso regression for longitudinal data.

<table>
<thead>
<tr>
<th>variable</th>
<th>Simple coping style</th>
<th>Social support</th>
<th>Self-esteem</th>
<th>loneliness</th>
<th>shyness</th>
<th>Internet addiction</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1 coefficient</td>
<td>0.206**</td>
<td>0.176*</td>
<td>0.197*</td>
<td>-0.237*</td>
<td>-0.134*</td>
<td>-0.132**</td>
</tr>
<tr>
<td>T2 coefficient</td>
<td>0.267*</td>
<td>0.247**</td>
<td>0.201*</td>
<td>-0.148**</td>
<td>-0.102**</td>
<td>-0.083*</td>
</tr>
</tbody>
</table>

The constant term is 71.554, the standardized regression coefficient of adaptation (T1) is 0.455 ($p< 0.05$), the error of standard estimation is 30.083, and the adjusted $R^2$ is 0.624. All the
coefficients are significant. Through the change of two regression coefficients, we found that the change of each independent variable, and rank the importance of each variable to college students' adaptation. The correlation analysis of the predicted results and the verification set shows that the correlation coefficient is 0.587, and the predicted results are good.

**Discussion**

In this study, 750 freshmen from two universities are investigated through tracking design, and the influencing factors of college students' adaptation were reasonably analyzed by three methods. In the end, the key factors of college students' adaptation are identified by the results of Lasso regression. This study further enriches current research on college students' adaptation and relevant factors. The machine learning, a new data processing method, also provides a broader research idea for psychological research.

Through Lasso regression combined with longitudinal data analysis, the key factors of college students' adaptation are obtained: shyness, coping style, self-esteem, loneliness, Internet addiction and social support. In longitudinal study, we find that for individual factors, the shyness level is higher than one year later, self-esteem is relatively stable, and the loneliness level is lower; for environmental Factors, social support is higher than one year later; for Individuals and Environment Interaction factors, coping styles become more positive, and the level of Internet addiction is higher than one year later.

Through the analysis results, it can be clearly seen that from multiple linear regression to Ridge regression and then to Lasso regression, the goal of dimensionality reduction is gradually realized, the number of explanatory variables is gradually reduced, the explanatory power is constantly strengthened, and the model is more concise and clear. According to the research of longitudinal data, we also rank the predictors of college students' school adaptation in order of importance: positive coping style, social support, self-esteem, loneliness, shyness and Internet addiction. This also reveals that not only individual factors, but also environmental factors, the interaction between individuals and environment on college students' adaptation are increasingly important.

**Acknowledgement**

This research was financially supported by 2018 Shandong graduate education quality improvement plan project “high school students of fine arts core quality and career planning graduate education joint training base” (SDYJD18020); and Shandong province social science planning project “development track and intervention of adolescent pathological Internet use” (18CQXJ06). Peng Wang, Jun Wang, Yun Yan, Wei-xuan Meng are co-first authors.

**References**


