Prediction of Students' Academic Learning Performance Based on Big Learning Data

Jing Sun

College of Information Science and Engineering, Northeastern University, Shenyang, P.R. China
sunjing@mail.neu.edu.cn

Abstract:
With the development of computer technology and network technology, intelligent teaching platform is gradually applied to the field of education, such as classroom on rainy days. It collects and records a large amount of classroom data, such as student attendance, classroom interaction and classroom tests. How to analyze and make use of these data is very important to understand students' learning status and predict their final achievements in the future. In this paper, 171 students majoring in electrical engineering in a university are selected to evaluate the correlation between their final grades and the three indicators in the classroom data. The results show that there is a strong correlation between the final score and the test score. The students with lower class test scores may have a higher risk of failing the final score, which provides useful experience for the judgment of failing the final score, and actively takes preventive and control measures in the follow-up teaching process to urge students to pay attention to classroom teaching.

Keywords: *Prediction, Learning Data, ROC*

1 INTRODUCTION

Early warning of students' learning disabilities is not only a basic need, but also a difficulty. With the information technology and its application in the field of education, various educational data on learning process and behavior have been gradually collected by colleges and universities recently. For example, such data: Students' gender, self-study time, attendance rate, progress and results of completing homework, class participation in class, class practice and results, etc. There are useful and possibly not obvious links between these information, such as: how classroom learning affects academic performance of the subject, and whether gender has an impact on mastering the subject. Some new modern technologies, such as big data analysis and learning analysis, have become more and more widely used in the field of education. In order to optimize the teaching process and promote students' more scientific and efficient learning, it is necessary to further study the early warning method based on big learning data.

Electrotechnics Course is a professional basic course for non electrical majors in Colleges and universities. The study of this course is beneficial to cultivate students' logical thinking ability, improve students' practical operation ability, and help students form a rigorous and standardized scientific attitude. The final grade of Electrotechnics course is weighted by the usual grade, experimental grade and final examination grade. The evaluation of usual performance is provided by Rain Classroom teaching software, which can record all links of teaching in real time, such as attendance, interaction, classroom test and other data, so as to realize the quantitative assessment of the whole learning process.

Many scholars have carried out relevant research based on various educational data [1] [2] [3] [6] [7] [8] [9] [10] [11] [12] [13] [15]. This research uses correlation analysis, regression analysis and factor analysis to find that the online learning data in different periods are positively correlated with the final examination results. This research used logistic regression method to comprehensively analyze the influencing factors of final grade. It was found that students with different majorsand different feelings of classroom atmosphere will affect students' grade.

The Ra in Classroom is an excellent new intelligent teaching software. Through the mobile network platform that can be used by mobile phones, it can accurately and real-time record all data of teaching links

B. Fox et al. (Eds.): IC-ICAIE 2022, AHCS 9, pp. 1063-1068, 2023.
https://doi.org/10.2991/978-94-6463-040-4_160

in the background, such as attendance, interaction, evaluation and so on. This study uses Sperman correlation analysis to analyze the relationship between the final grade of electrical engineering course and the classroom data recorded in theRain Classroom.Then ROC curve is used to evaluate the sensitivity, specificity and accuracy of classroom data to infer the failure of final grade, and find out the best cut-off value, in order to provide a method forquickly and simply inferring the possibility of a student's failure in final exam, and urge students to pay attention to classroom teaching.

2 DATA AND METHOD

2.1 Research object

There are six classes of Electrotechnics Course at a certain level in a university, with a total of 171 students, which are divided into three large classes for teaching. The teaching contents and teaching progress of each class are consistent. After the classroom study of theory course and electrical engineering experiment, the students will have a unified final closed book examination. The final grade of the course is weighted by the usual grade, experimental grade and final examination grade.

The course data collection is carried out during the teaching process from March 2021 to May 2021. The classroom data are the class attendance times, barrage times and classroom test grades provided by the Rain Classroom. These three parts of data are normalized first. The processing method is to divide each student's grade by the highest grade of the all, multiply by 100 to get each student's three classroom grades, and get attendance grades, interaction grades and test grades respectively as the independent variable of this study. The final grade was completed within about 2 weeks after the end of all teaching activities. The final grade was selected as the dependent variable of this study.

2.2 Research methods

The statistical method was IBM SPSS 18 statistical analysis software for statistical processing and normal distribution test of data. Sperman correlation analysis was used to evaluate the correlation between final performance analysis and attendance performance, interaction performance and test performance.

The main mathematical tool in this paper is receiveroperatingcharacteristiccurve, or ROC curve for short. At first, ROC curve was used in military, but now it is more used in medical field to judge whether certain factors have diagnostic value for the diagnosis of certain diseases[4] [5] [14] [16]. ROC curve is drawn according to a series of different binary classification methods (boundary value or decision threshold). The abscissa x-axis is 1-specificity, also known as false positive rate (false positive rate). The closer the x-axis is to zero, the higher the accuracy is; The Y axis of the vertical axis is called sensitivity, also known as true positive rate (sensitivity). The larger the Y axis, the better the accuracy.

According to the curve position, the whole graph is divided into two parts. The area under the curve is called AUC (area under curve), which indicates the prediction accuracy. The higher the AUC value, that is, the larger the area under the curve, the higher the prediction accuracy. The closer the curve is to the upper left corner (the smaller X and the larger y), the higher the prediction accuracy.

According to the final grade standard, the students are divided into the pass group and the fail group. ROC curve takes the attendance grade, interaction grade and test grade as the test variables, and the final grade (fail / pass) as the state variable to draw the ROC curve. The area under the curve AUC =0.5 is the reference line. The value of AUC is 0~1. According to the evaluation criteria of AUC diagnostic model: AUC is 0.5-0.7, which has low diagnostic value; AUC is 0.7-0.9, which has certain diagnostic value; AUC is above 0.9, which has high diagnostic value. The value corresponding to the maximum value of the Jordan index is the best cut-off value. Set $P < 0.05$ as statistically significant difference.

3 RESULT

3.1 Normality test

The final grade, attendance grade, interaction grade and test grade of this paper were statistically analyzed. The experimental results are shown in Table 1.

Table 1: Summary statistics

	Min	Max	Average	S.D.
Final Grade	36.7	95.1	73.5	12.8
Attendance Grade	81.3	100	98.6	3.2
Interaction Grade	0.8	100	35.3	19.8
Test Grade	13.4	87.5	60.6	14.0

Table 1 shows that the average final grade is 73.5 and the standard deviation is 12.8. The average attendance grade is 98.6 and the standard deviation is 3.2. Students' attendance is good and most of them can attend classes. The minimum interaction grade is only 0.8, with an average of 35.3 and a standard deviation of 19.8. It can be seen that most students' classroom interaction is not optimistic, and a few students speak actively and participate in the classroom. The minimum value of classroom test is 13.4, the maximum value is 87.6, the mean value is 60.6, and the standard deviation reaches 14.0. It can be seen that the overall difficulty of classroom test is difficult.

The final grade, attendance grade, interaction grade and test grade of this paper are continuous variables. Shapiro Wilk test method is used to test the normality of all samples. Significance level α is setat 0.05. The experimental results are shown in Table2. It can be seen from Table 2 that the p value is 0.0001, the significance p value of all data is less than 0.01, and the tested parameter samples do not conform to the normal distribution.

Table 2: Shapiro Wilk test results

	W	df	Sig.
Final Grade	0.957	171	0.000
Attendance Grade	0.488	171	0.000
Interaction Grade	0.963	171	0.000
Test Grade	0.946	171	0.000

3.2 Correlation analysis between final grades and classroom data

In order to analyze the correlation between final grades and classroom data, and further quantify the correlation between final grades and classroom data, the correlation between final grades and attendance grades, interactive grades and test grades is studied by using the correlation coefficient analysis method. According to the normality test results in Section 3.1, the final grade, attendance grade, interaction grade and test grade do not conform to the normal distribution, so Spearman rank correlation is selected for correlation analysis.

The experimental results are shown in Table 3. The results showed that there was a positive correlation between the final grade and attendance grade, interaction grade and test grade. The correlation coefficient r value was 0.289, 0.451, 0.632, and the p value was 0.000, and the difference is statistically significant. According to the correlation coefficient r listed in Table 3, we can get the correlation degree between the indicators of classroom data and the final grade. Among them, the final grade has a low correlation with the attendance grade, a medium correlation between the final grade and the interactive grade, and a strong correlation between the final grade and the test grade.

Table 3: Correlation analysis between final grade and attendance grade

	Final Grade	
	r	P
Attendance Grade	0.289**	0.000
Interaction Grade	0.451**	0.000
Test Grade	0.632**	0.000

3.3 Using ROC curve to evaluate final grades

According to the final grade standard, the students are divided into pass group and fail group. The ROC curve is evaluated with attendance grade, interaction grade and test grade as test variables and final grade (fail / pass) as state variables.

In the SPSS test variables, the final grade passing is represented by "1", the unqualified is represented by "0", and the test variables are represented by the corresponding scores of each part. The ROC curves of the diagnostic accuracy of attendance, interaction and test scores to the failure of the final grade are made, and the ROC curves of the passing accuracy of the final grade are drawn respectively. The AUC value under the 95% confidence interval (CI) is calculated to evaluate the value of attendance, interaction and test scores to the failure of the final grade, and the value of passing the final grade is compared.

The results are shown in Figure 1. It can be seen from Table 4 that the area under the ROC curve of attendance performance is 0.647, the cut-off value is 96.9, the sensitivity is 41.2%, the specificity is 86.9%, and the accuracy is 77.8%; The area under the ROC curve of interactive performance was 0.684, the cut-off value was 35.0, the sensitivity was 79.4%, the specificity was 53.3%, and the accuracy was 58.5%; The area under the ROC curve of the test grade is 0.805, the cut-off value is 54.7, the sensitivity is 76.5%, the specificity is 80.3%, and the accuracy is 79.5%.

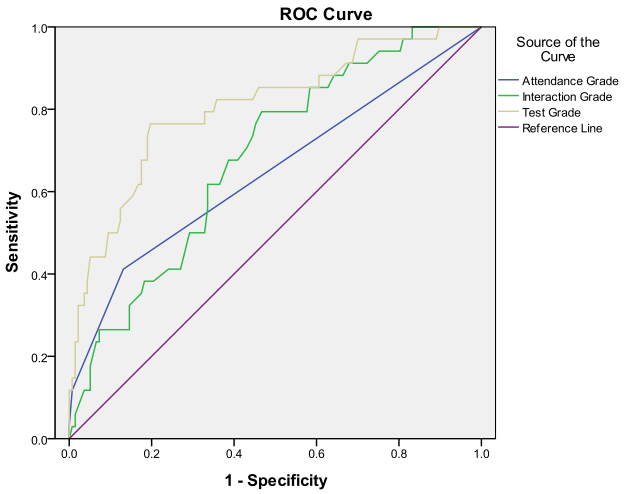

Figure1: ROC curve

Table 4: Predictive value of classroom data for failing final grades

Test Result Variable(s)	AUC	SEM	95% Confidence interval	cut-off value	Sensitivity (%)	Specificity (%)	P
Attendance Grade	0.647	0.058	0.533-0.760	96.9	41.2	86.9	0.008
Interaction Grade	0.684	0.048	0.590-0.777	35.0	79.4	53.3	0.001
Test Grade	0.805	0.045	0.718-0.893	54.7	76.5	80.3	0.000

In order to further verify the diagnostic performance of ROC curve evaluation, the diagnostic accuracy of the best critical value obtained from the experimental results for unqualified final scores is verified by using the final scores and classroom data of another group of 32 people. The results are shown in Table 5.

Taking the diagnostic accuracy of Y value as 100%, the diagnostic accuracy of attendance grade, interaction grade and test grade were 78.13%, 46.88% and 75.00% respectively. Through Y value, attendance grade, interaction grade and test grade, 32 students were diagnosed as having failed in the final exam. 6,3, 17 and 6 students were diagnosed as having failed in the final exam.

The failure rates of the final exam were 18.75%, 13.64%, 53.13% and 18.72% respectively.

The failure rate of final exam diagnosed by attendance grade alone is 4.11% lower than Y value, the failure rate of final exam diagnosed by interaction grade alone is 34.38% higer than Y value, and the failure rate of final exam diagnosed by test grade alone is 0.03% lower than Y value.

Table 5: The diagnostic accuracy of the cut-off value of classroom data for failure after ROC curve evaluation

Index	Cut-off value	Number of qualified persons	Number of failed students	Failure rate %	Accuracy
Y	1	26	6	18.75	100
Attendance Grade	96.9	29	3	13.64	78.13
Interaction Grade	35.0	15	17	53.13	46.88
Test Grade	54.7	26	6	18.72	75.00

4 Conclusions

In this paper, we use Sperman correlation to analyze the correlation between final grades and classroom data. The results showed that there was a low correlation between the final grades and attendance scores, a moderate correlation with interaction scores, and a strong correlation between the final grades and usual test scores. It shows that the usual classroom tests can better reflect students' classroom learning level, and students with low attendance, less interaction and low classroom test scores may have a higher risk of failing.

In addition, in order to further analyze the relationship between classroom test scores and final grades, the ROC curve is used to evaluate the usual classroom test scores and infer the risk of students' failure in the final exam. The results show that it has high sensitivity and specificity, and has diagnostic value for predicting the failure in the final exam. By passing the class test results, teachers can find the risk of failing the final exam as soon as possible in the process of students' learning, and formulate preventive measures in time to prevent students from failing.

REFERENCES

[1] Jay B, James M, Anne Z, et al. (2015). Using Learning Analytics to Predict At-Risk Students in Online Graduate Public Affairs and Administration Education. J. Journal of Public Affairs Education. 21(2),247-262.

[2] Kotsiantis, Sotiris B. (2012). Use of machine learning techniques foreducational proposes: a decision support system forforecasting students' grades.J. Artificial Intelligence Review37,331-344.

[3] Li X, Huang H, Chen X & Du P. (2020). Student Achievement Analysis System Based on Data Mining. J. Modern Information Technology. 4, 82-84.

[4] Liu X. 2020.The Diagnostic Value of Fine Needle Aspiration Cytology in Parotid Gland Masses Evaluated by ROC. Jiangxi: Nanchang.

[5] Ma S. 2018. Evaluating on Correlation Factors of Prognosis of Breast Cancer and Efficacy of Neoadjuvant Chemotherapy, Guangdong: Guangzhou

[6] Mi C. Yu N&Peng X. (2019). Method and System Constructing for Learning Situation Early Warning based on Data Mining Techniques. C. The 14th International Conference on Computer Science & Education (ICCSE 2019). 965-969.

[7] Pan X, Guo Q & Lin N. (2021). Investigation on Influencing Factors of Undergraduates' Higher Mathematics Achievement---Analysis Based on Logistic Regression Model. J. College Mathematics. 37, 60-69.

[8] Rangel V S, Bell E R, Monroy C & Whitaker J R. (2015). Toward a New Approach to the Evaluation of a Digital Curriculum Using Learning Analytics. J. Journal of Research on Technology in Education. 47(2),89-104.

[9] Song D, Liu D, Feng X. (2020). Research on Curriculum Score Prediction and Curriculum Early Warning Based on Multi-source Data Analysis Research on Higher Engineering Education. J. Research on Higher Engineering Education. 1, 189-194.

[10] Wang C, Hou Y, Xu M, Zhao G, Xiao C, Yan C & YAN Q. (2021). Logistic Regression Analysis of Academic Performance of Arts and Science Students Majoring in Stomatology. J. Journal of Higher Education. 24, 65-68.

[11] Wei S. (2013). Learning Analytics: Mining the Value of Education Data under the Big Data Era. J. Modern Educational Technology. 23(2),5-11.

[12] Wei S, Han Y, Wang L. (2015). Reflections on Online Teaching and Learning Based on Data Mining and Analysis of Learning Process Data. J. Modern Educational Technology. 24(6),89-95.

[13] Xiong S, Nong Y. (2021). On Correlation Analysis Between Online Learning Data and Students ' Academic Performance---With the University Probability Theory Course as an Example. J. Journal of Southwest China Normal University (Natural Science Edition). 46(11), 84-89.

[14] Zhang Y., 2021. The Application of CT Radiomics and Machine Learning in the Diagnosis of Bosniak III Renal Masses. Jilin: Changchun.

[15] Zeng X, Peng Y, Du D, Liu X, Yang G, Sun Z & Wang Y. (2021). Ordinal Logistic Regression Analysis of Influence Factorson Test Scores of Medical Statistics. J. China Journal of Modern Medicine. 24, 106-112.

[16] Zhu L., 2021.Evaluation of the severity of pulmonary hypertension by electrocardiogram combined with echocardiography, Shandong: Qingdao.