Eng Int Syst (2021) 2: 89–96

Engineering
Intelligent Systems

The Prediction Model of College English Performance Based on Data Mining Under the Concept of OBE

Yanfeng Yue*

School of International Education, University of Science and Technology Liaoning, Anshan, Liaoning 114051, People's Republic of China

With the penetration of the concept of Outcome Based Education (OBE) and the rapid development and implementation of information technology in the field of education, how to combine the OBE concept with data mining technology to obtain a prediction of college English performance is a research topic worth exploring. In this paper, firstly, an overview is given of the OBE concept and education data mining, and then the focus will shift to the construction of a college English score prediction model based on data mining. Specifically, the "Modeler+ neural network model" will be used to complete data mining and prediction analysis. Finally, based on the analysis results, several teaching suggestions are offered.

Keywords: OBE concept; Education data mining; College English; "SPSS Modeler" + "neural network model"

1. INTRODUCTION

With the advancement of informatization in the field of education, various information management platforms have been established for the collection and storage of data related to daily teaching activities, student learning and other aspects of formal education. However, in most universities, these platforms are used mainly for simple data collection and analysis, without in-depth mining. In recent years, with the increase in the amount of big data being generated, data mining and data analysis technology have gradually penetrated the field of education, making it possible to mine and evaluate massive amounts of data in university databases.

Many domestic scholars have applied data mining technology to study education-related data, contributing significantly to the research findings in this area. Shu & Qu (2014) used stepwise regression and neural network technology to analyze the examination results of college students and to determine their influencing factors. Shi, Qian & Sun (2016) used statistical analysis, visualization, the association rule algorithm and the clustering algorithm to analyze a large amount of learning data generated by the network learning process. Based on the analysis results, they proposed several ways of monitoring and managing the network learning process. You & Sun (2016) used a multiple linear regression model to predict the exam scores of students in various courses in universities, and conducted teaching intervention according to the predicted results to improve the students' exam results. Chen & Zhu (2017) applied a nested integrated learning method to build a classification prediction model of online learners' exam results. It enables the influencing factors to be determined in order to predict the exam results of online learners. It can also be used to send alerts regarding academic performance, make performance predictions, and assess online learning tasks. Sun L, Zhang K & Ding (2016) applied a clustering algorithm to the analysis of the scores and learning information of online academic education undergraduate English courses, that enabled the segmented

*Corresponding address: No. 185, Qianshan Middle Road, Anshan, Liaoning 114051, People's Republic of China. Email: y6f90o@163.com

Table 1 Statistics for girls and boys.

	Coding	Number	Ratio
Girl	0	1254	41.07%
Boy	1	1799	58.93%

prediction of the scores of English examinations of adult students.

The results obtained by Chinese scholars show that their research was mainly concerned with the "test score", reflecting the traditional Chinese focus on exam-oriented education. This paper argues that while "test scores" are the best way to measure a student's learning achievement, the overall "test scores" do not reflect the other skills that individuals acquire from the learning process. Therefore, educators should consider and pay more attention to the individual's "ability" in terms of other skills. The Outcome-based Education (OBE) concept emphasizes that the purpose of education is to develop "competence". This paper, from the perspective of OBE, studies the neural network model of "listening", "reading" and "writing" skills to be acquired.

2. THEORY

2.1 Concept of OBE

The notion of OBE first appeared in the basic education reforms implemented in America and Australia. In the OBE system, educators must have a clear vision of the skills and levels that students should have upon graduation, and then design appropriate educational programs to ensure that students achieve those goals. Student output, rather than textbooks or teacher experience, drives the functioning of this type of education system, in stark contrast to the traditional content-driven and input-driven education approach. Hence, in China, OBE can be regarded as an innovative educational paradigm.

OBE requires that schools and teachers first clarify the learning outcomes, match a range of diversified and flexible personalized learning approaches with the students' requirements, set challenging tasks that will engage students in self-directed learning, and then use feedback and student outcomes to improve the original curriculum design.

2.2 Educational Data Mining

Educational data mining refers to the comprehensive application of mathematical statistics, machine learning and data mining techniques and methods to the processing and analysis of big data in the education sector. Through data mining, correlations can be found between learners' learning results and learning content, learning resources, teaching behavior and other variables, so as to predict learners' future learning trends.

3. PREDICTION MODEL CONSTRUCTION AND ANALYSIS OF COLLEGE ENGLISH SCORE DATA MINING UNDER THE CONCEPT OF OBE

3.1 Initial Model Construction and Evaluation

3.1.1 Data Mining Platform and Setting Process

In this study, the researchers use IBM's SPSS Modeler 18.0 as a data mining platform. Some scholars such as Ye (2019) use a decision tree model (C5) to predict students' grades; however, in this paper, a neural network model is used. Two neural network models are included in the SPSS Modeler: a neural network model and a Bayesian neural network model. The former is based on the BP algorithm, while the latter is based on the Bayesian method.

3.1.2 Data Extraction and Preprocessing

In this study, cross-section data is used, collated before modeling rather than processing the data in SPSS Modeler. First, the data obtained for 2019 were extracted from the CET-4 database, summarized, and recorded in the same ECXEL file. The statistics for the boys and girls in this sample are shown in Table 1.

This paper focuses mainly on the influence, on their CET-4 scores, of students' individual daily performance and the overall class performance. In addition to gender, this paper obtains data on attendance rate, class mean for assignment scores, final exam scores, and class averages. In traditional teaching and learning approaches, the CET-4 score is used as the total score; on the other hand, OBE focuses more on students' specific abilities, such as listening, reading and writing. The total CET-4 score is derived from scores for four sections: listening, reading, translating and writing. The grade structure for Translating is 248.5:248.5:106.5:106.5. If "translating" and "writing" are changed to "writing", then the fraction structure into 248.5:248.5:213, obtaining the total score for Listening, Reading and Writing data respectively. The descriptive statistics for these data are shown in Table 2.

Then, the scores are assigned grades: 1 (0–60%), 2 (60–90%) or 3 (90–100%) as shown in Table 3.

3.1.3 Establish a Neural Network Model

In this paper, five groups of models are constructed using the neural network module, as shown in Table 4.

Finally, using SPSS Modeler, we constructed the "flow" diagram shown in Figure 1.

Table 2 Descriptive statistics.

Index	n	Min	Max	Mean	Std
Attendance rate	3353	0.10%	99.97%	75.67%	28.80%
Class mean for assignment scores	3353	0	100	89.77	28.616
Final exam score	3353	0	143	85.05	43.538
The average level of the class	3353	0	100	80.51	28.6422
Total score	3353	83	233	148.54	25.057
Listening	3353	83	239	151.01	25.62
Reading	3353	71	202	134.47	26.928
Writing	3353	237	675	144.67	71.746

Table 3 Hierarchy of data.

Grades (coding)	1	2	3
Total score	[0,426)	[426,639)	[639,710]
Listening	[0,149.1)	[149.1,223.7)	[223.7,248.5]
Reading	[0,149.1)	[149.1,223.7)	[223.7,248.5]
Writing	[0,127.8)	[127.8,191.7)	[191.7,213]
Attendance rate	[0,60%)	[60%,90%)	[90%,100%]
Class mean for assignment scores	[0,60)	[60,90)	[90,100]
Final exam score	[0,90)	[90,135)	[135,150]
The average level of the class	[0,90)	[90,135)	[135,150]

Table 4 Variable settings for the model.

Model	Object	Variable
Model 1	Total score	Sex; attendance rate; class mean for assignment scores; final exam score; the average level of the class
Model 2	Listening	
Model 3	Reading	
Model 4	Writing	
Model 5	Listening, Reading & Writing	

3.1.4 Model Evaluation

After setting "Flow", click "Run", the results shown in Table 5 were obtained.

From the constructed model, the scores can be predicted. The comparison between predicted values and observed values of each model is shown in Figure 2 and Figure 3.

From Figures 2 and 3, it can be seen that model 1 is the best predictor of level 3, while models 2 to 4 are more accurate in predicting levels 1 and 2

3.2 Model Optimizing

In IBM's SPSS Modeler 18.0, the neural network module has a model optimization function that includes a reference model, a naive model and a states model. This study uses the model optimization function for the purpose of "increasing accuracy". The accuracy obtained after model optimization is shown in Table 6.

The results show that the reference model is adopted to improve the accuracy of models 1 and 4. Model 5, on the other hand, does not allow optimization because of multiple targets.

3.3 Model Prediction Analysis

The order of importance of the predictive variables of each model is shown in Figure 4 to Figure 8.

The prediction results show that in the optimized models 1 to 4, the variable ranking first in importance is gender, while

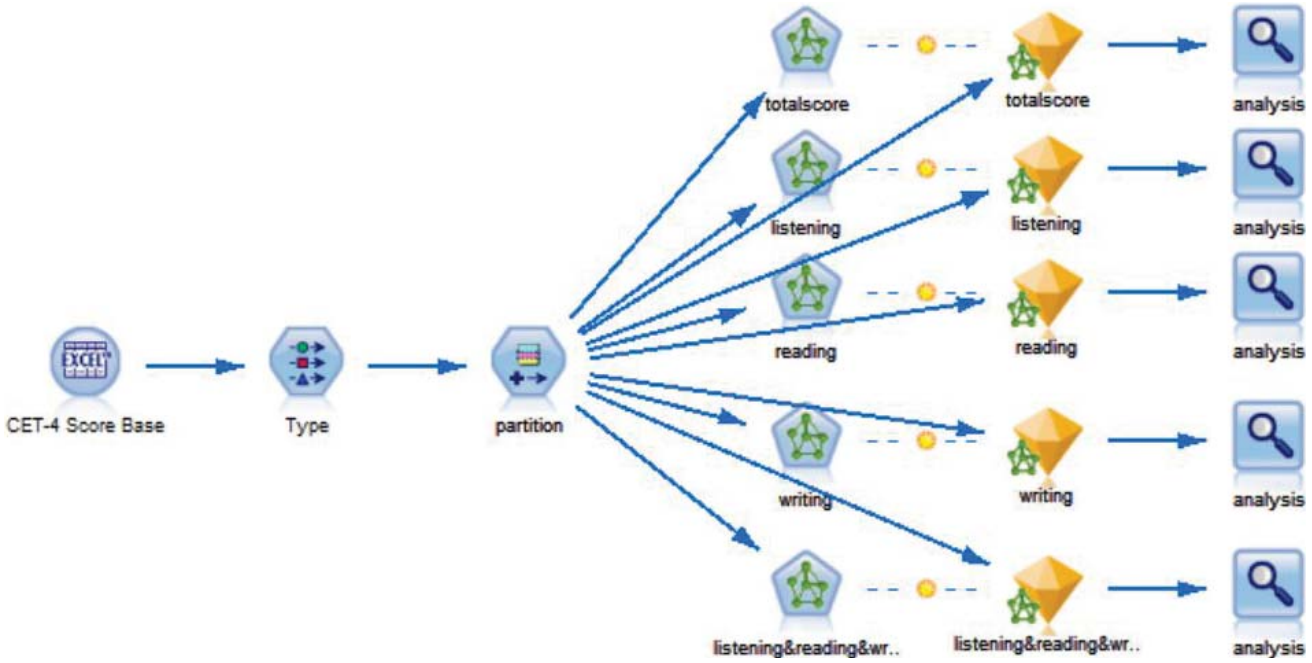

Figure 1 "flow" of SPSS Modeler.

Table 5 Model evaluation.

Model	Accuracy	Validity	N.1 importance	Neure
Model 1	83.0%	82.97%	Final exam score	8
Model 2	88%	87.95%	Final exam score	7
Model 3	91.9%	91.06%	Final exam score	7
Model 4	77.9%	78.06%	Sex	6
Model 5	85.2%	Listening: 87.26%; Reading: 91.68%; Writing: 75.96%	Final exam score	10

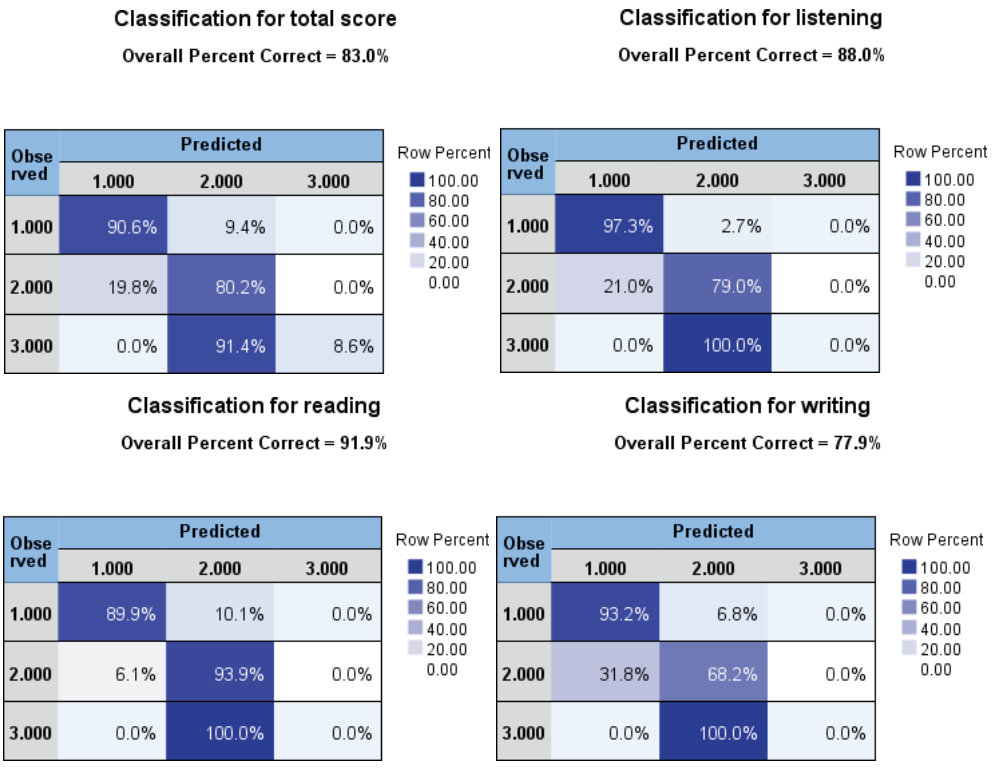

Figure 2 The coincidence ratio between predicted values and observed values from Model 1 to Model 4.

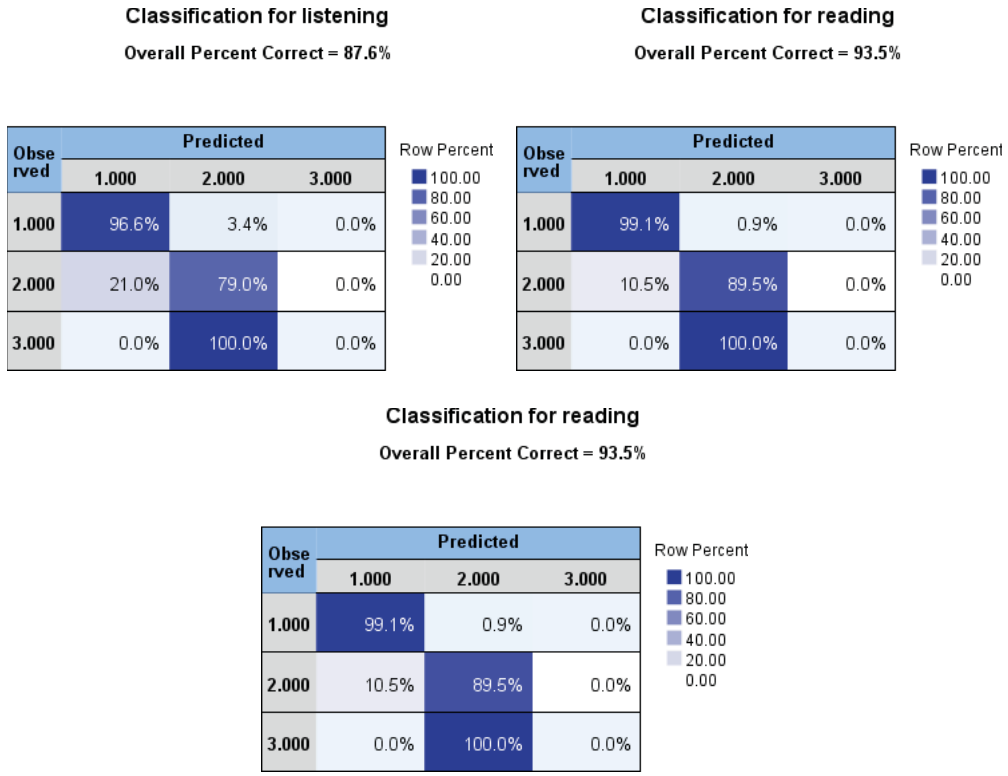

Figure 3 The coincidence ratio between predicted values and observed values from Model 5.

Table 6 Model optimization.

Model	Reference model	Naive model	Entirety model
Model 1	81.5%*	57.7%	69.9%
Model 2	86.8%*	54.8%	73.6%
Model 3	92.3%*	58.0%	88.3%
Model 4	77.3%*	57.6%	64.9%

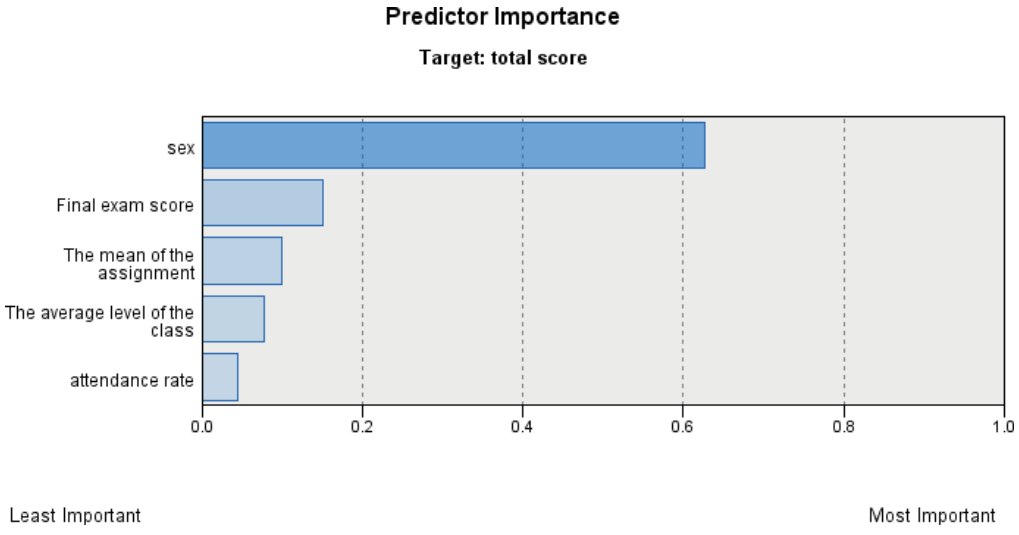

Figure 4 Order of importance of predictive variables in Model 1.

in model 5, Sex ranks fourth. Thus, it can be seen that when predicting a single goal, a different model should be constructed for each gender. However, when targeting listening, reading and writing, gender can be the main variable. Additional information can be obtained from Figures 4∼8.

4. APPLICATION OF PREDICTION MODEL CONCLUSIONS

First, gender appears to be the main factor determining student performance. School administrators should be concerned

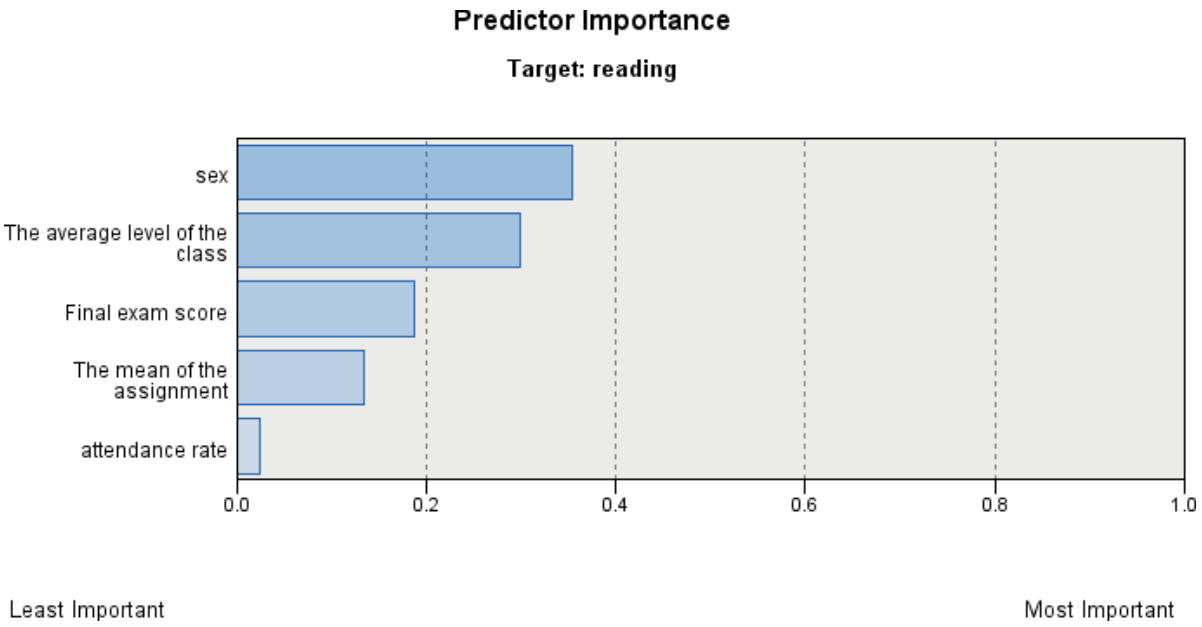

Figure 5 Order of importance of predictive variables in Model 2.

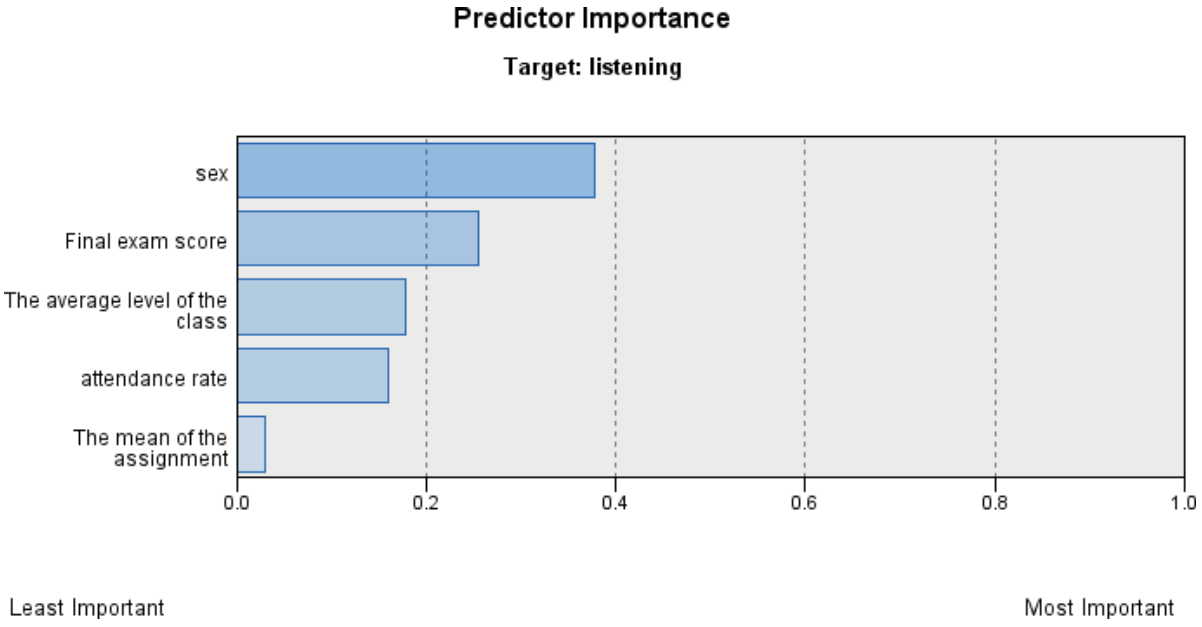

Figure 6 Order of importance of predictive variables in Model 3.

about the huge discrepancy in CET-4 scores between boys and girls. Although there may be various reason for this phenomenon, previous research findings indicate that women have a language advantage over their male counterparts. Female college students tend to have good performance in various language tests, especially CET-4. This researcher believes that in addition to gender differences, there are other factors that could explain why there is a huge difference between boys' and girls' CET-4 scores for university English courses. This is an issue that merits further study.

Second, "final examination score" and "average homework score" are still the decisive factors determining the CET-4 score. In addition, it has been found that "attendance" is directly related to learning outcomes and test scores. This is not surprising given that attendance is an indicator of a student's attitude towards learning. Course administrators should pay attention to the teaching and learning modes being adopted, encourage students to change from passive learning to active learning in senior high school, help students to establish a positive learning attitude, and ensure the school's environment is positive and conducive to learning.

Finally, the average level of the class is also an important factor affecting the CET-4 score, and reflects the quality of teaching. Therefore, the methods used by teachers to teach college-level English should evolve and keep pace with modern learning theories. Moreover, more attention should be given to the practical application of English, as this is an important factor that influences skills development and retention. During this time, students are still in the process of changing from passive to active learners and should be given

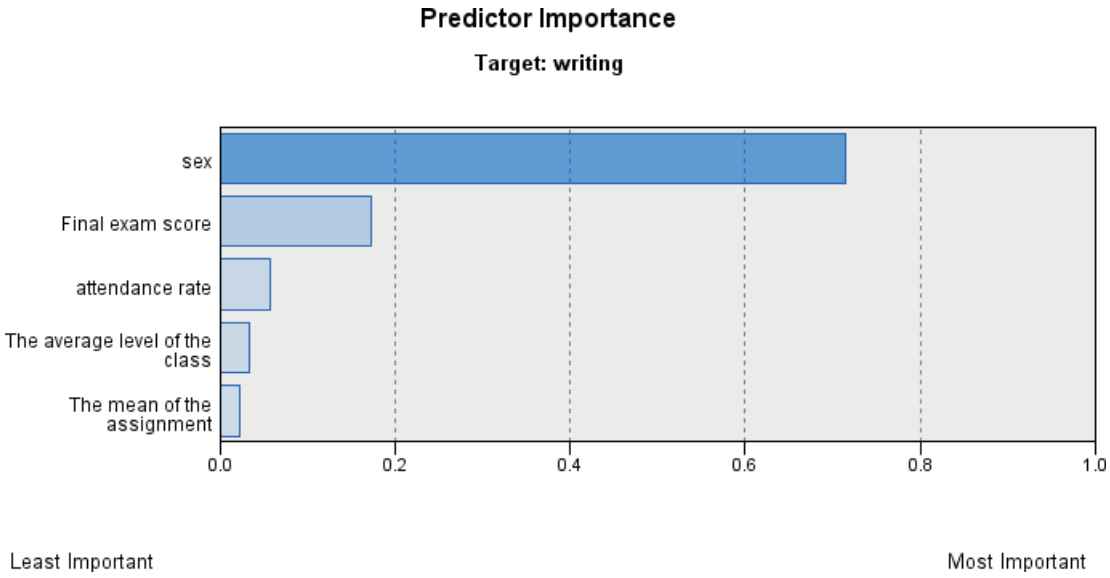

Figure 7 Order of importance of predictive variables in Model 4.

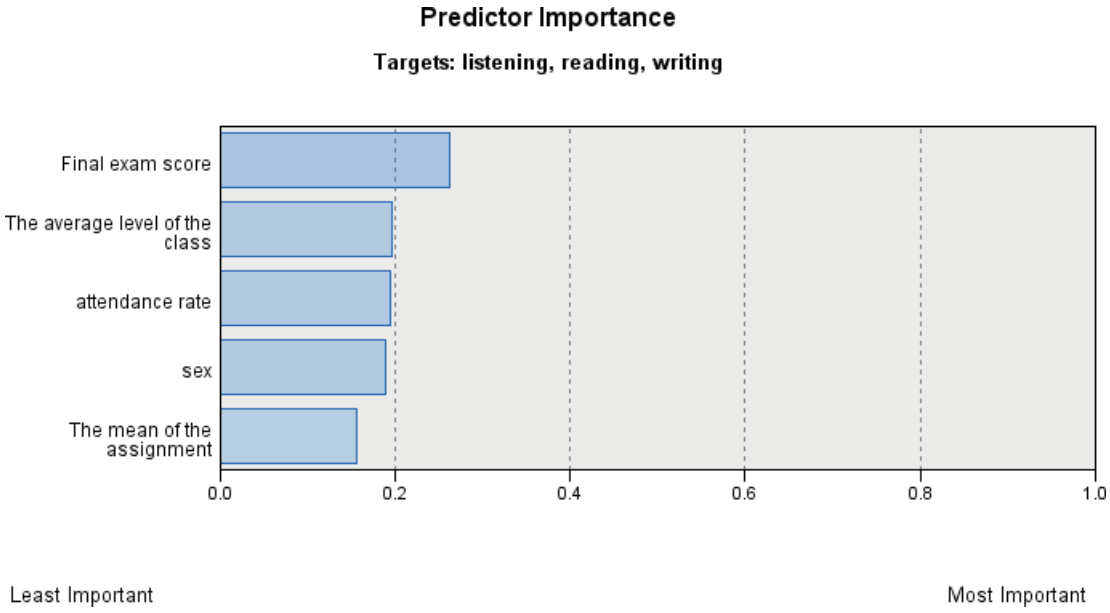

Figure 8 Order of importance of predictive variables in Model 5.

every encouragement to do so in the classroom. High-quality courses can stimulate students' interest in learning which in turn will lead to better CET-4 scores and pass rates.

5. CONCLUSION

At present, with the rapid development of various online education platforms, a large amount of new data is constantly being integrated into university databases. However, university education is still being delivered mainly through face-to-face classroom teaching. This means that while new types of data continue to be amassed, large amounts of traditional data are still being collected and stored.

With the continuous development of educational data mining research, there is now an abundance of educational data mining and analysis methods. Nevertheless, the value of traditional educational data will be rediscovered. The biggest advantage of such data over new data is that the data are huge and have been accumulated over a longer period of time. Therefore, the mining of massive traditional data is more likely to find the deep-rooted laws in the development process of colleges and universities, teachers' teaching and students' learning, etc., which will help to improve the management of teaching and teacher quality, and enhance students' learning.

REFERENCES

1. Chen, Z.J. & Zhu, X.L. (2017). Research on Online Learner Academic Performance Prediction Modeling Based on Educational Data Mining. *China Audio-Visual Education*, (12), 75–81.

2. Shi, S., Qian, Y. & Sun, L. (2016). Research on Network Learning Process Supervision based on Education Data Mining. *Modern Education Technology*, (6), 87–93.
3. Shu, Z.M. & Qu, Q.F. (2014). Analysis of College Students' Learning Outcomes Based on Educational Data Mining. *Journal of Northeastern University (Social Sciences)*, (3), 309–314.
4. Sun, L., Zhang, K. & Ding, B. (2016). Research and Implementation of Subdivision Prediction of Online Education Based on Data Mining. *China Distance Education*, (12), 22–29.
5. Ye, Z.J. (2019). Modeling of College English Test Pass Rate prediction based on Data Mining. *Journal of Changchun Normal University*, 38(12), 55–62.
6. You, J.X. & Sun, Z. (2016). Research on the Prediction and Intervention of College Students' Academic Performance on cloud Learning Platform. *China Distance Education*, (9), 14–20.