High School English Performance Analysis Using Interpretable Machine Learning Approach

Shufang Qu(✉) and Hun Lee Koay

School of Business, Malaysia University of Science and Technology, Petaling Jaya, Malaysia
qu.shufang@phd.must.edu.my, hlkoay@must.edu.my

Abstract. Currently, English learning is being taken to a new level for Chinese students. The transformation of English language teaching and learning is in full swing across the education sector in China. However, few studies have discussed the relationship between high school students' English language learning performance and other courses. To address these issues and gain a practical understanding of the relationship between students' English learning performance and other courses, this paper first collects learning data from 532 students in 10 courses at a high school in China. Second, this paper uses an integrated learning algorithm called XGBoost to predict the students' English learning performance. Specifically, the dataset is divided into a training set and a test set, and we train the model on the training set and test it on the test set. The test results show that the method in this paper can predict students' English performance well (MAE < 0.03, MSE < 0.07, RMSE < 0.04). Moreover, Chinese and mathematics scores were highly correlated with students' English scores. Based on the above findings, this paper further proposes relevant teaching suggestions. The results of this paper provide a practical reference for teaching on each campus.

Keywords: Machine learning · Interpretable · XGBoost algorithm · English teaching

1 Introduction

English as a foreign language has been focused on in modern China for dozens of years. In China these years, great improvement has been made in Chinese education, so the improvement and progress of English teaching have been achieved [1, 2]. Although the improvement and progress are great, there still exist many problems, for example, the English teachers' moral states need improvement, and the moral states determine the teachers' work and achievements. The teachers need to improve because society develops very fast and very thorough. The teachers determine students' performances and achievements in their studies.

To be a teacher, he must be very knowledgeable and contribute his knowledge, abilities, time, and energy to his students [3]. In this way, the students will be well developed. Nowadays, however, English teachers in China are very lacking in knowledge

V. Balakrishnan et al. (Eds.): ICMEIM 2022, AHSSEH 6, pp. 246–254, 2022.
https://doi.org/10.2991/978-94-6463-044-2_33

about the relationship between students' curriculum. This has resulted in teaching blind spots [4]. Students' learning is often correlated; for example, students who do well in Chinese tend to do better in English as well. However, math scores tend to have inhibitory performance on English scores. This may be since Chinese and English are both language courses and students can easily handle both courses once they have mastered a common learning approach [5, 6]. Mathematics courses, on the other hand, tend to require students to provide a more rational and logical perspective on problems, thus creating a difference in thinking.

The phenomenon we mentioned above has been widely recognized in the teaching of English courses, and may even be said to be a consensus among English teachers. Therefore, it is necessary to explore the correlations between courses in order to guide our teachers to better tailor their teaching to students' learning potential and create a more favorable learning atmosphere [7, 8].

In recent years, with the development of information technology, numerous student academic data have been accumulated on campus. And powerful information processing tools such as machine learning algorithms have emerged, for example. Therefore, incorporating machine learning methods to improve Chinese English teachers' teaching ability is a very promising topic.

To address the above-mentioned issues and gain a practical understanding of the relationship between students' English learning performance and other courses, this paper first collects learning data from 532 students in 10 courses at a high school in China. Second, this paper uses an integrated learning algorithm called XGBoost to predict the students' English learning performance. Specifically, the dataset is divided into a training set and a test set, and we train the model on the training set and test it on the test set. The test results show that the method in this paper can predict students' English performance well (MAE < 0.03, MSE < 0.07, RMSE < 0.04). Moreover, Chinese and mathematics scores were highly correlated with students' English scores. Based on the above findings, this paper further proposes relevant teaching suggestions.

The reset of this paper is organized as follows: Sect. 2 provides a detailed discussion of the methodology of this paper. The obtained results are discussed in Sect. 3. Finally, Sect. 4 summarizes the full paper and suggests future research directions.

2 Methods and Meterials

2.1 XGBoost Algorithm

XGBoost was first proposed by Chen and Guestrin as an improvement over the gradient boosted decision tree GBDT [9, 10]. Conventional trees only use first-order derivatives, but XGBoost regression (XBR) innovatively introduced second-order derivatives and regular terms, making the algorithm good in training and rapid in computing. The XBR learning process is as follows.

Assume that there are K trees in the model, first we define the basic process as:

$$\begin{aligned} &\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \\ &with\, F = \{f(x) = w_{q(x)}\}(q : R^m \rightarrow \{1, 2, .., T\}, w \in R^T) \end{aligned} \tag{1}$$

where $\hat{y}_i$ is the prediction value of the model for the data. F denotes the set of all trees, $f(x)$ is the function of one tree. T is the number of leaf nodes of the tree, $q(x)$ is the mapping function of the sample data corresponding to a leaf node on the tree, and $w_{q(x)}$ is the score of the leaf node.

Then comes the objective function, which is defined as:

$$Obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f(k)) \tag{2}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

where $\sum_{i}^{n} l(y_i, \widehat{y_i})$ is the model loss function, $\Omega(f_k)$ is the regular term of tree k, and γ and λ are XGBoost customizing parameters that, respectively, limit the number of leaf nodes and control the size of the node score; other variables are as for the previous equations.

The XGBoost algorithm is an ensemble technique that trains cumulatively and successively to iteratively optimize the objective function until the objective function reaches a minimum value, at which time training is complete. The training process starts with the optimization of the first tree, and when the model iterates to tree t, it is given by:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-)} + f_t(x_i) \tag{4}$$

If the loss function is squared error, the objective function can be changed to:

$$Obj^{(t)} = \sum_{i}^{n} (y_i - \hat{y}_i^{(t)})^2 + \sum_{i=1}^{t} \Omega(f_t) \tag{5}$$

And it's easy to change to:

$$Obj^{(t)} = \sum_{i}^{n} [2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2] + \Omega(f_t) + cons \tag{6}$$

where $\hat{y}_i^{(t)}$ is the prediction value of the model that has iterated to t trees, $\hat{y}_i^{(t-1)}$ is the prediction of the model after optimization of the previous $t-1$ trees, $f_t(x_i)$ is the score of the newly added t trees, and constant is the sum of the regularization terms of the previous $t-1$ trees.

2.2 Data Description

In this paper, we collected student academic data from a high school in China. Specifically, this data includes the examination data of 532 students in a particular grade, as shown in Table 1.

Table 1. Description of the obtained data-set

Academic course	Symbols
Chinese	X1
Mathematics	X2
Physics	X3
Chemistry	X4
Biology	X5
Politics	X6
History	X7
Geography	X8
Technology	X9
English	Y

Table 2. Statistical description of the research data

	MEAN	STD	MIN	MEDIAN	MAX
X1	90.00	10.71	43.00	90.50	116.00
X2	90.70	23.32	25.00	93.25	142.00
X3	47.32	21.75	0.00	43.00	95.00
X4	59.18	14.78	15.00	60.00	94.00
X5	59.79	13.90	18.00	60.00	95.00
X6	63.84	11.90	22.00	64.50	90.00
X7	57.77	11.70	23.00	58.00	91.00
X8	68.77	10.57	27.00	69.00	94.00
X9	48.54	10.48	17.00	48.00	88.00
Y	83.19	16.15	25.00	83.50	126.00

A statistical description of the data used in this paper is presented in Table 2. In order to obtain more reasonable results, later we will first pre-process the data in order to normalize them to [0,1], as shown in Eq. (7).

$$x_{ij}^{'} = \frac{\max\{x_{ij}, ..., x_{nj}\}}{\max\{x_{ij}, ..., x_{nj}\} - \min\{x_{ij}, ..., x_{nj}\}}, \tag{7}$$

where x_{ij} represents the value of the j-th indictors of the i-th sample.

2.3 Evaluation Metrics

In order to better evaluate the prediction results made by the adopted method, the following regression evaluation indicators are used in this paper.

$$MAE = \frac{\sum_{i=1}^{n} \hat{y}_i - y_i}{n} \tag{8}$$

$$MSE = \sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n} \tag{9}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{10}$$

where n denotes the number of samples, $\hat{y}_i$ is the predicted value of the model, y_i means the true value of the response.

3 Results and Discussions

3.1 Pre-processing

As shown in the previous section, we first normalize the collected data to better exploit the performance of the machine learning algorithm. Table 3 shows the results of the statistical description of the sample after normalization.

From Table 3, we can find that the normalized data are all in the same interval range, which is favorable to the prediction results in the subsequent section. It is worth mentioning that we further divide the data by dividing 80% of the data into a training set to train the XGBoost model and 20% of the data into a test set to test the prediction performance of the XGBoost model.

Table 3. Statistical description of the pre-processed data

	MEAN	STD	MIN	MEDIAN	MAX
X1	0.64	0.15	0.00	0.65	1.00
X2	0.56	0.20	0.00	0.58	1.00
X3	0.50	0.23	0.00	0.45	1.00
X4	0.56	0.19	0.00	0.57	1.00
X5	0.54	0.18	0.00	0.55	1.00
X6	0.62	0.18	0.00	0.63	1.00
X7	0.51	0.17	0.00	0.51	1.00
X8	0.62	0.16	0.00	0.63	1.00
X9	0.44	0.15	0.00	0.44	1.00
Y	0.58	0.16	0.00	0.58	1.00

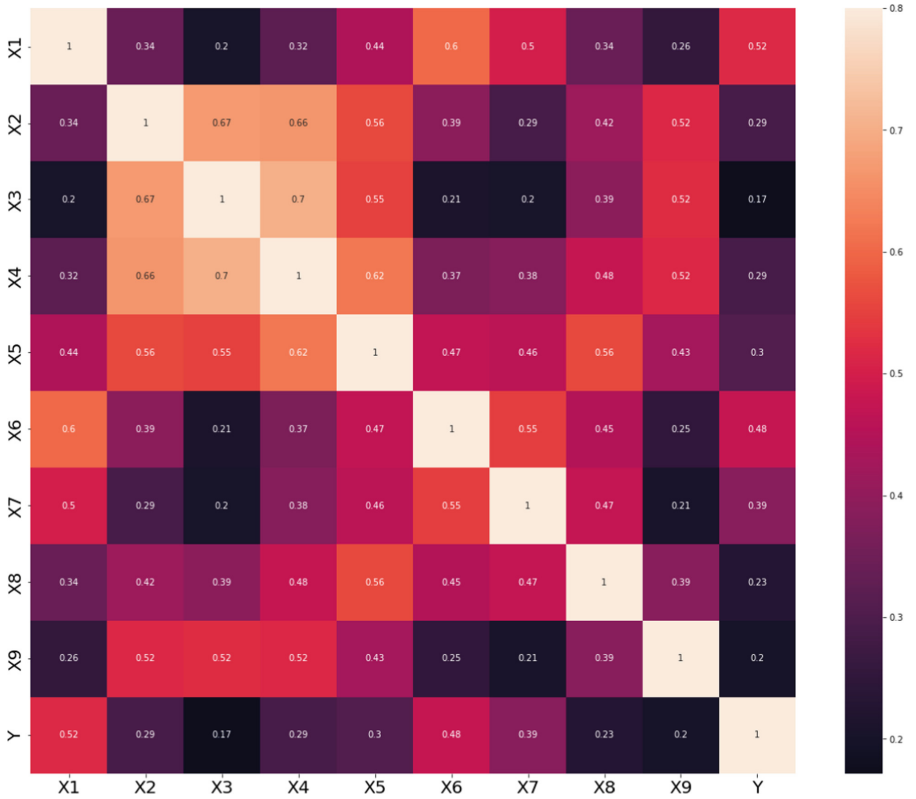

Fig. 1. Correlation heat map

3.2 Correlation Analysis

Before making predictions, we first analyze the correlations between features to find out if there are significant correlated features present [11]. The correlation coefficient is calculated as in Eq. (11). We present our results in Fig. 1.

$$\rho_{x_1 x_2} = \frac{Cov(X_1, X_2)}{\sqrt{DX_1, DX_2}} \tag{11}$$

As can be seen from the figure, all X did not show a significant correlation with Y (basically less than 0.5). Therefore, we decided to keep all X variables for input into the XGBoost algorithm.

3.3 Prediction Results

We first train the XGBoost algorithm on the training set and the obtained training results predict the performance on the test set. The obtained prediction results are shown in Table 4. As can be found in Table 4, our developed method exhibits excellent predictive performance (MAE $<$ 0.03, MSE $<$ 0.07, RMSE $<$ 0.04). This indicates that XGBoost

Table 4. Prediction results

Metrics	MAE	MSE	RMSE
Results	0.0248	0.0674	0.0337

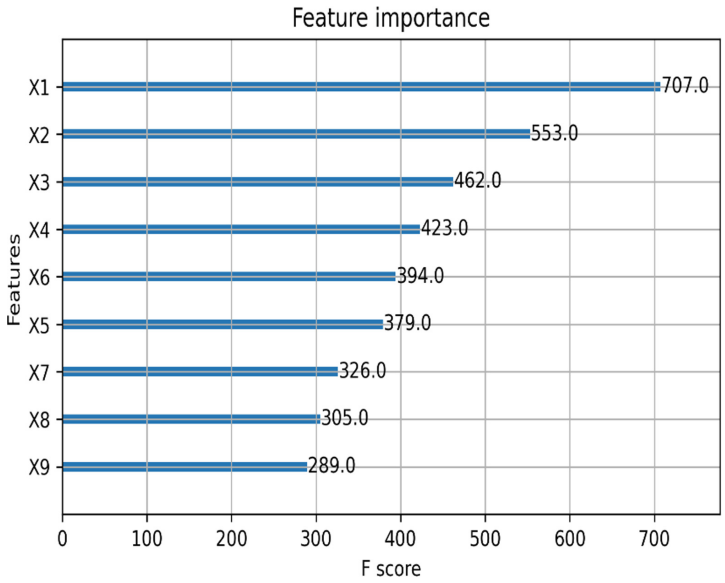

Fig. 2. Feature importance analysis

can accurately predict students' English performance based on their other coursework scores.

However, it is not scientific to simply know the outcome of that prediction. What we are more interested in knowing is which courses play a significant role in predicting students' academic performance in English.

Therefore, we further performed a feature visualization analysis, based on the XGBoost algorithm for feature importance analysis. This method has been extensively explored in existing studies [12–14]. Here, we follow the ideas of previous studies to explain the prediction results of the XGBoost model. The results of the feature importance analysis we performed are presented in Fig. 2.

As can be seen from Fig. 2, the main courses (Chinese X1, Mathematics X2) show the strongest explanatory performance. And according to the importance of the courses, their influence on English performance gradually decreases. Therefore, it can be said that the major course curriculum is crucial in predicting the English performance of this student.

4 Conclusions

Currently, English learning is being taken to a new level for Chinese students. The transformation of English language teaching and learning is in full swing across the education sector in China. Nowadays, however, English teachers in China are very lacking in knowledge about the relationship between students' curriculum.

The phenomenon we mentioned above has been widely recognized in the teaching of English courses, and may even be said to be a consensus among English teachers. Therefore, it is necessary to explore the correlations between courses in order to guide our teachers to better tailor their teaching to students' learning potential and create a more favorable learning atmosphere. However, few studies have discussed the relationship between high school students' English language learning performance and other courses. To address these issues and gain a practical understanding of the relationship between students' English learning performance and other courses, this paper first collects learning data from 532 students in 10 courses at a high school in China. And before we put them into the XGBoost algorithm, we first pre-processed them to be in [0,1]. Second, this paper uses an integrated learning algorithm called XGBoost to predict the students' English learning performance.

Specifically, the dataset is divided into a training set and a test set, and we train the model on the training set and test it on the test set. The test results show that the method in this paper can predict students' English performance well ($MAE < 0.03$, $MSE < 0.07$, $RMSE < 0.04$). Moreover, Chinese and mathematics scores were highly correlated with students' English scores.

Based on the above findings, this paper further proposes relevant teaching suggestions:

1. Teachers should change poor English teaching methods to good ones, so every teacher should keep up with the times. The society is developing rapidly, so every person should also develop or progress rapidly. Bad principles and methods should be discarded and removed to meet the fast requirements of the society. Teachers should have methods of self-improvement;
2. The examination system is necessary in Chinese society because the Chinese believe that equality of people is important. An equal education requires an examination system. English teachers should pay attention to those students who do well in other courses but do not do well in English. This is because perhaps their potential for English learning has not been fully explored.

The results of this paper provide a practical reference for teaching on each campus.

References

1. Liu, H., Zhang, X., & Fang, F. (2021). Young English learners' attitudes towards China English: unpacking their identity construction with implications for secondary level language education in China. Asia Pacific Journal of Education, 1-16.

2. Cheng, J., & Wei, L. (2021). Individual agency and changing language education policy in China: Reactions to the new 'Guidelines on College English Teaching'. Current issues in language planning, 22(1-2), 117-135.
3. Lei, W. (2021). A Survey on Preservice English Teachers' Intercultural Communicative Competence in China. English Language Teaching, 14(1), 37-47.
4. Huang, M., Shi, Y., & Yang, X. (2021). Emergency remote teaching of English as a foreign language during COVID-19: Perspectives from a university in China. IJERI: International Journal of Educational Research and Innovation, (15), 400–418.
5. Xiu, X., & Ibrahim, N. M. B. (2021). Role of Learner Autonomy and Students' Perception in Legitimizing China English as A Variety of English. Eurasian Journal of Applied Linguistics, 7(2), 31-45.
6. Hui, Z. (2021). A Review of Information and Communication Technologies Implementation in English Teaching and Learning in China. Journal of Research in Educational Sciences, 12(14), 32-37.
7. Yu, X. (2021). Foreign language learning anxiety in China: theories and applications in English language teaching: by Deyuan He, Singapore, Springer Nature, 2018, xx+ 221 pp., 74, 96€(eBook), ISBN 978–981–10–7662–6 (Vol. 24, No. 8, pp. 1247-1249). Routledge.
8. Xiong, T., Li, Q., & Hu, G. (2022). Teaching English in the shadow: identity construction of private English language tutors in China. Discourse: Studies in the Cultural Politics of Education, 43(1), 73-85.
9. Yang, L., Zhao, Y., Niu, X., Song, Z., Gao, Q., & Wu, J. (2021). Municipal Solid Waste Forecasting in China Based on Machine Learning Models. Front. Energy Res, 9, 1-13.
10. Wang, Y., Yang, L., Wu, J., Song, Z., & Shi, L. (2022). Mining Campus Big Data: Prediction of Career Choice Using Interpretable Machine Learning Method. Mathematics, 10(8), 1289.
11. Gao, X.; Wang, J.; Yang, L. An Explainable Machine Learning Framework for Forecasting Crude Oil Price during the COVID-19 Pandemic. Axioms 2022, 11, 374.
12. Rajbahadur, G. K., Wang, S., Ansaldi, G., Kamei, Y., & Hassan, A. E. (2021). The impact of feature importance methods on the interpretation of defect classifiers. IEEE Transactions on Software Engineering.
13. Feng, D. C., Wang, W. J., Mangalathu, S., & Taciroglu, E. (2021). Interpretable XGBoost-SHAP machine-learning model for shear strength prediction of squat RC walls. Journal of Structural Engineering, 147(11), 04021173.
14. Ben Jabeur, S., Stef, N., & Carmona, P. (2022). Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering. Computational Economics, 1-27.