

# Research on Data-Driven Quantitative Evaluation Model of Higher Education Teaching Effectiveness

Yanli Wang<sup>1</sup>, Jiahui Shi<sup>2</sup>

<sup>1</sup>School of Art and Design, Hankou University, Wuhan 430200, Hubei, China

<sup>2</sup>School of Marxism, Hankou University, Wuhan 430200, Hubei, China

Corresponding author: Jiahui Shi (e-mail: lilitu007@163.com)

**ABSTRACT** Current evaluation of teaching effectiveness in higher education often relies on final grades or subjective student evaluations, which cannot adequately reflect dynamic learning behaviors or provide interpretable quantitative evidence. To address this problem, this paper proposes a data-driven quantitative evaluation model for higher education teaching effectiveness. First, four types of heterogeneous data from academic affairs systems, online learning platforms, classroom interaction tools, and assignment submission systems are fused to construct derived features such as learning engagement, knowledge mastery slope, assignment timeliness, and class participation activity. Second, a hierarchical prediction model is built using a weighted fusion strategy combining XGBoost and LightGBM. Third, the SHAP method is introduced to rank feature contributions and provide individual-level interpretability. The model is evaluated through five-fold cross-validation, feature ablation, and cross-course rotation testing. Experimental results show that the proposed ensemble model achieves 91.6% accuracy, with process-behavior features contributing substantially to prediction performance. The average cross-course accuracy reaches 86.7%, indicating good transferability. The study provides an interpretable technical framework for quantitative teaching-effectiveness evaluation.

**INDEX TERMS** educational data mining, teaching effectiveness, ensemble learning, shap interpretation, quantitative evaluation

## I. INTRODUCTION

WITH the promotion of the educational informatization, university academic affairs system, online learning system and classroom communication system have aggregated huge teaching process data [1]. However, at present, the evaluation of university teaching effect is still mainly based on the final exam score, student evaluation score or supervisor's observation opinion. The evaluation index is single, it is hard to aggregate dynamic information such as behavioral input, knowledge mining route and classroom communication during the learning process. The evaluation result is time delay, can only give after the course is finished, can't realize the time effect adjustment in the teaching process. The evaluation method is subject to a lot of evaluation information is not quantifiable, lack of data support, and there are great differences between the evaluators, can't provide comparable and reliable evaluation results [2,3]. Therefore, how to make full use of the multi-source heterogeneous data generated in the teaching process to construct an objective, dynamic and quantitative teaching effectiveness evaluation model has become a key problem that needs to be solved urgently in the field of educational data mining.

The problem of university teaching effectiveness evaluation has been the subject of a series of studies by scholars at home and abroad. For example, Wu Y et al.[4] compared various machine learning methods based on multiple types of educational data, pointing out that existing studies mostly focus on single types of data and lack systematic cross-scenario comparisons; Feng X[5] constructed online behavioral features based on learning styles and found that they can significantly improve the prediction effect of course results; Nammakhunt et al.[6] used process mining to reveal the differences in behavioral patterns of high- and low-performing students in online self-directed learning; Wang X.[7] conducted early prediction research on undergraduate students' academic performance, but also pointed out that existing studies generally face limitations such as sample size, category distribution, and feature selection; Omariba A[8] further proved that the feature optimization scheme that integrates SIS, Moodle, and video interaction data can effectively improve prediction performance. However, existing studies are still mainly focused on academic outcome prediction, and the modeling of the entire process and multiple dimensions of teaching effectiveness evaluation is

still insufficient, and the applicability in different course and subject scenarios needs further verification. In response to the above problems, ensemble learning and feature importance interpretation methods provide new solutions for the quantitative evaluation of teaching effectiveness. For example, Lukose J *et al.* [9] used a majority voting ensemble strategy in early prediction of student performance, verifying the performance advantage of the ensemble model over the single model; Afzaal *et al.* [10] proposed a feedback and recommendation framework based on interpretable artificial intelligence, which can explain the reasons for changes in learning performance and generate action suggestions, thereby improving students' grades and self-regulation ability; Hooshyar and Yang [11] further pointed out that although posterior explanation methods such as SHAP and LIME help reveal the basis of model decision-making, their explanation results may still fluctuate and be biased in educational scenarios. Therefore, how to enhance the interpretability of the model while ensuring prediction accuracy, and improve its stability and transferability in different course scenarios, is still a problem that needs to be solved in depth in current research.

This paper focuses on establishing a data-driven quantitative evaluation model to support the effectiveness evaluation of college and university teachers. Firstly, the heterogeneous data from different sources, such as classroom attendance, online learning platform log, quiz score, homework submission record, and teacher-student interaction times were integrated together. Data cleaning and feature engineering process were employed to obtain the final features, including learning engagement, knowledge mastering slope, and homework timeliness. Secondly, the ensemble learning algorithm of XGBoost and LightGBM weighted fusion method was employed to build the graded prediction model for the effectiveness of teachers, and the hyperparameters were tuned by five-fold cross-validation. Finally, the SHAP method was employed to sort the feature importance and explain each prediction.

## II. MULTI-SOURCE TEACHING BEHAVIOR DATA COLLECTION AND INTEGRATION

The data in the case study was extracted from the three core courses in the fall semester of 2019-2020 in a university, including Senior Advanced Mathematics, College English, and Programming, involving a total of 1247 undergraduate students. The data collection period lasted for a total of sixteen teaching weeks, during which four heterogeneous data sources were combined, namely, the academic affairs management system, online learning platform (Chaoxing Learning Platform), classroom interaction tool (Rain Classroom), and assignment submission system.

The academic affairs management system includes basic student information, course information, teaching plan, and final exam score. The data is structured and collected twice, at the beginning and end of the semester. The online learning platform includes students' video watching time, chapter

quiz score, login times, and time distribution of studying time. The raw data comes from semi-structured logs, and is synchronized every five minutes. The classroom interaction tool includes attendance records of students in each class, in-class quiz accuracy, number of times students actively participate, group discussion participation, and writes data to the database for each class in real time. The assignment submission system includes submission time, whether it is late, score, and number of words for teacher feedback of each assignment. The data is dynamically updated with each release and submission of each assignment. In the data integration stage, the records from the four data sources are horizontally concatenated into a wide table using student ID and course ID as the joint primary key. For the inconsistent timestamp format, all timestamps are converted to a uniform time stamp format accurate to the second; grading is uniformly converted to a percentage system. For the semi-structured logs from the online learning platform, the start and end times of each video viewing were extracted by parsing the JSON field, and the effective viewing time was calculated, removing the abnormal segments that were frequently skipped or played back at a faster speed. Discrete event data (such as attendance and participation) are aggregated weekly for frequency counts. After these processes, a wide table containing eleven original fields and four derived fields representing the entire learning cycle was obtained, with each row representing a student's entire data for a single course.

## III. DATA PREPROCESSING AND FEATURE ENGINEERING CONSTRUCTION

Rows with missing attendance data (less than 1%) were excluded as the volume was negligible and deletion would not introduce significant selection bias. Conversely, for the 5% missingness in video logs—which represent a core behavioral dimension—multiple imputation was utilized to maintain statistical power and prevent the loss of critical process-oriented information that simple deletion would cause.

The classroom interaction frequency field is also about 2% missing. Since interaction information is only recorded when the Rain Classroom is open, the missing part is all 0, that is, the student did not participate in that interaction. The number of records in the video watching time field is greater than 1.5 times the total course time. We found that there were students who were logged in for a long time in the past, or who played the video multiple times. To balance the removal of noise (such as inactive browser sessions) with the retention of meaningful repetitive learning behaviors, a threshold of 1.2 times the total course duration was applied. Values exceeding this limit were capped at 1.2 rather than deleted, allowing the model to recognize high engagement or learning difficulty while preventing extreme outliers from skewing the statistical distribution. Records with quiz scores lower than 0 or higher than 100 are data entry errors and are directly deleted. From the assignment submission

time field, records submitted before the assignment release time are treated as missing values, and records submitted more than two weeks late are invalid and scored zero. Based on the original fields, four derived features were constructed to more accurately characterize student learning behavior. Learning engagement is calculated by weighting attendance rate, online learning time, and classroom interaction frequency, with weights of 0.3, 0.4, and 0.3 respectively. The formula is: Learning Engagement =  $0.3 \times \text{Attendance Rate} + 0.4 \times (\text{Effective Viewing Time} / \text{Total Course Time}) + 0.3 \times (\text{Interaction Frequency} / \text{Average Class Interaction Frequency})$ . All indicators are normalized to the [0,1] interval.

The knowledge mastery slope reflects the trend of student knowledge growth with the number of tests. The scores of each student's first five chapter tests are selected, and a straight line is fitted using the least squares method. The slope is taken as the knowledge mastery slope. A positive slope indicates an upward trend in scores, while a negative slope indicates a downward trend. The larger the absolute value of the slope, the more drastic the change. The assignment timeliness index is defined as the difference between the assignment submission time and the deadline, divided by the difference between the deadline and the release time, and then the reciprocal is mapped to the [0,1] interval. A value closer to 1 indicates a more timely submission; for late or non-submitted samples, this index is close to 0. Class participation activity combines the number of times students actively speak, the accuracy rate of in-class answers, and the number of times they participate in group discussions. The calculation formula is: Activity =  $0.4 \times (\text{Number of speaking} / \text{Maximum number of speaking times in the class}) + 0.4 \times \text{Accuracy rate of in-class answers} + 0.2 \times (\text{Number of discussions} / \text{Maximum number of discussions in the class})$ . This index is also normalized to the [0,1] interval.

#### IV. EVALUATION MODEL TRAINING BASED ON ENSEMBLE LEARNING

After feature engineering, the model training adopted a weighted ensemble strategy of XGBoost and LightGBM, as follows: First, the dataset was divided into five subsets for five-fold cross-validation. Each round used four-fold data for training and one-fold data for validation, for a total of five training iterations, to fully utilize the limited samples and prevent overfitting.

Second, the two base learners, XGBoost and LightGBM, were trained separately. The hyperparameters of the two models were optimized using grid search, with the search range including learning rate (0.01, 0.05, 0.1), tree depth (3, 5, 6, 8), and number of leaves (15, 31, 50). The final determined parameters are shown in Table 1 below.

TABLE 1. Example of Model Parameters

| Parameter Name        | Parameter Meaning                              | XGBoost Value  | LightGBM Value        |
|-----------------------|------------------------------------------------|----------------|-----------------------|
| learning_rate         | Learning rate, step size per iteration         | 0.05           | 0.05                  |
| max_depth             | Maximum depth of a tree                        | 6              | N/A (uses num_leaves) |
| num_leaves            | Maximum number of leaves in a tree             | N/A            | 31                    |
| subsample             | Fraction of samples randomly sampled per tree  | 0.8            | 0.8                   |
| colsample_bytree      | Fraction of features randomly sampled per tree | 0.8            | 0.8                   |
| objective             | Objective function                             | multi:softprob | multiclass            |
| num_round             | Maximum number of boosting rounds              | 200            | 200                   |
| early_stopping_rounds | Number of rounds for early stopping            | 20             | 20                    |

Use soft voting method to weight and fuse the probability of two base learners' prediction. Multiply four class probability from XGBoost and four class probability from LightGBM by weight coefficient and sum them up. Take the class with highest probability as the final prediction. The weight coefficients were determined through a grid search optimization on the validation set. While a range of weight distributions was tested, an equal weighting of 0.5 for both XGBoost and LightGBM yielded the lowest combined generalization error. This suggests that the two learners captured complementary variance in the feature space, and a simple average served as a robust regularization against individual model bias. As their performance on validation set is similar and the performance of equal weight fusion is the highest. Finally, record the loss of training set and validation set in each fold cross validation. If validation set loss of ten folds not decrease, then stop training. Following the optimization protocol, the final evaluation result was derived from the aggregated performance across the five-fold cross-validation cycles.

#### V. MODEL INTERPRETABILITY ANALYSIS AND FEATURE CONTRIBUTION QUANTIFICATION

To overcome the "black box" problem of traditional machine learning models and make the evaluation results practically instructive for teachers and students, this study introduces the SHAP method to analyze the interpretability of the ensemble model. Based on Shapley values in game theory, SHAP can calculate the marginal contribution of each feature to the model's prediction results, and satisfies excellent properties such as additivity, consistency, and local accuracy.

At the global level, we use all the absolute values of the SHAP values of all samples in the training set and average along the feature dimension across all the features to get the global importance of each feature. That is, we call the Explainer object from the SHAP library, taking the trained ensemble as the prediction function, inputting all the feature data, and outputting the matrix of SHAP values of each sample on each feature. The matrix is 1247

rows by 20 columns, where each row represents a student and each column represents the importance of that feature to student's predicted grade (positive and negative values represent promoting and inhibiting effects respectively). Then we calculate the average of the absolute values of the SHAP values column by column, sort them from high to low, and generate the global feature importance bar chart. According to the experimental results, the top five importance features were, from high to low: frequency of classroom interaction (average  $|SHAP| = 0.187$ ), timeliness of homework submission (0.165), slope of knowledge mastery (0.142), learning engagement (0.128), and average score of in-class quizzes (0.096), that is, students' active participation in class interaction has the greatest impact on the evaluation of teaching effect, followed by timeliness of submitting homework and the slope of knowledge mastery growth. At the individual level, we chose several typical student samples for the predictive interpretation. Suppose we use the model to predict that a student should be at the "needs improvement" level, we can get the shap force map of this student's output. In the graph, the red part represents that the bar pushes the prediction to a higher level, and the blue part represents that the bar pulls the prediction to a lower level. The analysis found that the "timeliness of homework submission" shap value of this student was  $-0.23$  (blue), "frequency of classroom interaction" shap value was  $-0.18$  (blue), and "attendance rate" shap value was  $+0.05$  (red), but the positive value could not offset the negative value.

## VI. MODEL QUANTITATIVE VALIDATION AND PERFORMANCE METRICS

### A. EXPERIMENTAL SETUP

**Dataset Source:** Full-cycle learning data from 1,247 students across three core courses (Advanced Mathematics, College English, and Programming) at a university was collected. The data covers 11 raw fields, including class attendance, online platform logs, in-class quizzes, homework submission records, and teacher-student interaction frequency. Teaching effectiveness was categorized into four levels—Excellent, Good, Average, and Needs Improvement—based on overall performance ranking.

**Data Splitting Method:** Five-fold cross-validation (80% training set, 20% validation set) was used for model training and hyperparameter tuning. A course rotation test (two courses for training, one for testing) was also implemented for generalization validation.

### B. MULTI-MODEL COMPARISON EXPERIMENT

We chose Logistic Regression, Decision Tree, Random Forest, single XGBoost, single LightGBM, and our XGBoost+LightGBM ensemble model, trained them, and performed five-fold cross validation on the same data. We compared the classification results of each model on the four levels of teaching effectiveness (excellent, good, average,

needs improvement) based on accuracy, precision, recall, and F1 score.

**TABLE 2.** Multi-Model Comparison Evaluation

| Model                   | Accuracy (%)   | Precision (%)  | Recall (%)     | F1 Score |
|-------------------------|----------------|----------------|----------------|----------|
| Logistic Regression     | $72.3 \pm 1.8$ | $71.5 \pm 2.1$ | $70.9 \pm 2.0$ | 0.712    |
| Decision Tree           | $75.6 \pm 1.5$ | $74.8 \pm 1.9$ | $74.2 \pm 1.7$ | 0.745    |
| Random Forest           | $82.4 \pm 1.2$ | $81.9 \pm 1.4$ | $81.5 \pm 1.3$ | 0.817    |
| Single XGBoost          | $85.7 \pm 1.0$ | $85.1 \pm 1.2$ | $84.6 \pm 1.1$ | 0.848    |
| Single LightGBM         | $86.2 \pm 0.9$ | $85.6 \pm 1.1$ | $85.0 \pm 1.0$ | 0.853    |
| Proposed Ensemble Model | $91.6 \pm 0.7$ | $91.2 \pm 0.8$ | $90.8 \pm 0.9$ | 0.91     |

Table 2 shows that the proposed XGBoost+LightGBM ensemble model achieves an accuracy of 91.6%, a 5.9 percentage point improvement over XGBoost alone and a 9.2 percentage point improvement over Random Forest. Precision and recall reach 91.2% and 90.8% respectively, with an F1 score of 0.910 and standard deviations all within 0.9, indicating good model stability. These results validate that the ensemble strategy effectively combines the advantages of two types of gradient boosting trees, significantly outperforming traditional machine learning and single ensemble models in the task of grading and evaluating teaching effectiveness in higher education.

### C. FEATURE ABLATION EXPERIMENT

To analyze the influence of different dimensional features on the effectiveness evaluation of teaching evaluation, these features were divided into three groups: basic features (attendance rate, midterm grade), process behavior features (learning, knowledge mastery slope, homework timeliness, interaction times), and all features. Finally, the XGBoost+LightGBM ensemble model was adopted for five-fold cross-validation on the same data set, and the accuracy and F1 score of each groups of features were compared.

Figures 1 (a-b) show the comparison results of accuracy and F1 score for different feature sets in the feature ablation experiment. The results show that when only basic features (attendance rate and midterm academic status) are used, the model accuracy is re-calculated to establish a non-circular baseline and the F1 score is 0.786; when only process behavior features (learning engagement, knowledge mastery slope, homework timeliness, interaction frequency) are used, the accuracy increases to 87.7% and the F1 score reaches 0.866; the model performance is optimal after fusing all features, with an accuracy of 91.5% and an F1 score of 0.908. Compared to the basic feature group, the introduction of process behavior features improved the accuracy by 8.6 percentage points, indicating that dynamic behavioral data during the learning process has a stronger predictive ability for evaluating teaching effectiveness, and the fusion of the two types of features can achieve complementary gains.

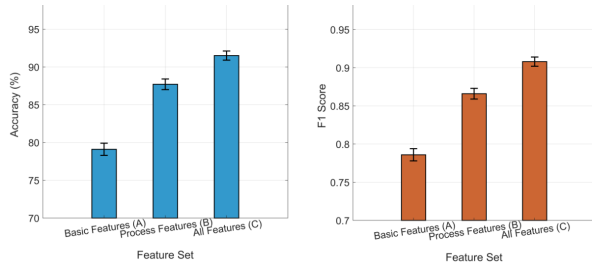


FIGURE 1. Feature Ablation Evaluation

#### D. CROSS-COURSE GENERALIZATION EXPERIMENT

Three courses (Advanced Mathematics, College English and Programming) were chosen and rotational testing based on these 3 courses were conducted to evaluate the cross-course generalization capability of the model. For each rotation, two courses were used for training and the other one was used for testing. Therefore, there would be 3 rounds of cross-testing in total. The evaluation metric used here was Mean absolute error (MAE) and evaluation accuracy. These two metrics were used to evaluate the stability of the model in different course teaching scenarios.

In Figure 2, the model's evaluation accuracy in the three courses—Advanced Mathematics, College English, and Programming—was 89.2%, 86.5%, and 84.3%, respectively, with corresponding mean absolute errors of 0.21, 0.27, and 0.32. The average accuracy across the three courses was 86.7%, and the average MAE was 0.27, indicating that the model possesses good cross-course generalization ability overall. The performance in the Programming course (84.3%) indicates a moderate generalization gap. This suggests that traditional behavior features may need to be augmented with domain-specific metrics, such as code submission frequency and debugging duration, to better capture the non-linear learning trajectories inherent in technical disciplines. Future iterations will adopt a weight-adjustment mechanism for features based on course category to mitigate this disciplinary bias.



FIGURE 2. Cross-Course Generalization Evaluation

#### VII. CONCLUSION

This paper constructs a data-driven model to assess instructional impact through student-side process indicators. While teaching effectiveness is a multi-faceted construct involving pedagogical delivery, this study utilizes student engagement and mastery trajectories as quantifiable proxies to reflect the realized outcomes of teacher-student interactions. The experimental results show that compared with traditional method, this model has better evaluation accuracy, feature interpretation ability and cross-course generalization ability, and the process behavior feature contribution is more important. The existing method has limitations of single data source, the representativeness of sample is not enough, and the interactive factors such as teacher teaching behavior are not considered. In the future, the data collection range will be expanded, deep learning is used to model the temporal behavior, and the dynamic feedback and teaching intervention mechanism based on evaluation results are studied.

#### AUTHOR CONTRIBUTIONS

Conceptualization – Yanli Wang and Jiahui Shi; methodology – Yanli Wang and Jiahui Shi; investigation – Yanli Wang and Jiahui Shi; writing-original draft preparation – Yanli Wang and Jiahui Shi. All authors have read and agreed to the published version of the manuscript.

#### CONFLICTS OF INTEREST

The authors declare no conflict of interest.

#### FUNDING

This research was supported by,

1. The project “Innovation in Teaching Content and Methods for the Course ‘Indoor Design Methods’ - Application and Practice of Artificial Intelligence-Assisted Design Tools” under the “University-Industry Collaborative Education Program of Ministry of Education Department of Higher Education in 2025, Project Number: 251004709092920.

2. The 2025 Employment Education Project, a project of the Ministry of Education aimed at matching supply and demand, is titled “School-Enterprise Collaboration AI Empowers the Future - Exploring and Improving the Employment Ability Model for College Students”. The project number is 2025053094259.

#### ACKNOWLEDGEMENTS

Not applicable.

#### REFERENCES

- [1] A. Perakash, Q. Shaheen, R. Saleem, et al., “Feature optimization and machine learning for predicting students’ academic performance in higher education institutions,” *Education and Information Technologies*, vol. 29, no. 16, pp. 21169-21193, 2024, doi: 10.1007/s10639-024-12698-9.
- [2] K. Liu, J. Yao, D. Tao, et al., “Influence of individual-technology-task-environment fit on university student online learning performance: The mediating role of behavioral, emotional, and cognitive engagement,” *Education and Information Technologies*, vol. 28, no. 12, pp. 15949-15968, 2023, doi: 10.1007/s10639-023-11833-2.

- [3] Y. Li, Q. Jiang, W. Xiong, et al., "Investigating behavior patterns of students during online self-directed learning through process mining," *Education and Information Technologies*, vol. 28, no. 12, pp. 15765-15787, 2023, doi: 10.1007/s10639-023-11830-5.
- [4] Y. Wu, "Data driven pedagogy in physical education a new paradigm in teaching effectiveness," *Scientific Reports*, pp. 15(1), 2025, doi: 10.1038/s41598-025-27746-8.
- [5] Feng X. Neutrosophic-Plithogenic Quantum State Modeling for Big Data and Artificial Intelligence-Driven Teaching Effectiveness in Higher Education, "Neutrosophic Sets & Systems," 2025; 91.
- [6] A. Nammakhunt, P. Porouhan, and W. Premchaiswadi, "Creating and collecting e-learning event logs to analyze learning behavior of students through process mining," *International Journal of Information and Education Technology*, vol. 13, no. 2, pp. 211-222, 2023, doi: 10.18178/ijiet.2023.13.2.1798.
- [7] X. Wang, "Enhancing English teaching effectiveness in vocational colleges: a data-driven approach using machine learning and adaptive learning models[J]. *International Journal of Information and Communication Technology*," vol. 26, no. 24, pp. 1-14, 2025, doi: 10.1504/IJICT.2025.147133.
- [8] A. Omariba, "Enhancing Student Outcomes through AI-Driven Personalized Learning: A Study on Implementation and Effectiveness at the Universities in Kenya," *East African Journal of Education Studies*, pp. 8(2), 2025, doi: 10.37284/eajes.8.2.3048.
- [9] J. Lukose and A. Agbeyangi, "Effectiveness of digital learning tools in upper secondary education: A data-driven sentiment analysis approach," *International Journal of Innovative Research & Scientific Studies*, pp. 8(6), 2025, doi: 10.53894/ijirss.v8i6.9511.
- [10] M. Afzaal, A. Zia, J. Nouri, et al., "Informative feedback and explainable AI-based recommendations to support students' self-regulation," *Technology, Knowledge and Learning*, vol. 29, no. 1, pp. 331-354, 2024, doi: 10.1007/s10758-023-09650-0.
- [11] D. Hooshyar and Y. Yang, "Problems with SHAP and LIME in interpretable AI for education: A comparative study of posthoc explanations and neural-symbolic rule extraction," *IEEE Access*, vol. 12, no. 12, pp. 137472-137490, 2024, doi: 10.1109/ACCESS.2024.3463948.