

Hybrid Deep Learning Framework for Student Grade Prediction

Fengrui Xiao

*Beijing-Dublin International College, Beijing University of Technology, Beijing, China
23371203@emails.bjut.edu.cn*

Abstract: This paper constructs a hybrid prediction framework combining Multi-Layer Perceptron, Long Short-Term Memory, and Transformer models to predict students' academic performance. The framework takes the Grades dataset, combines advanced feature engineering technologies such as time attenuation weighting and topic correlation matrix, and uses mean square error (MSE) and R^2 index to evaluate the prediction performance. The results show that the performance of the Long Short-Term Memory (LSTM) model is better than other models, and the minimum mean square error of its test set is 4.4821, proving the LSTM model's effectiveness in capturing students' learning time series patterns. The transformer model also performs well, but the mean square error is slightly higher. In addition, the interpretability analysis of SHapley Additive exPlanations (SHAP) reveals the significant contribution of G2 characteristics to the prediction, which provides a basis for targeted education intervention. This study emphasizes the potential of deep learning in educational data mining and emphasizes the importance of interpretability and feature diversity. Future work may explore the integration of more complex feature engineering technologies and further improve the prediction accuracy by combining the advantages of LSTM and transformer models.

Keywords: Learning behavior analysis, academic performance prediction, Long Short-Term Memory, attention mechanism

1. Introduction

Predicting and analyzing students' academic performance is a key step to improving the quality of education. In the research field of student achievement prediction, scholars have made remarkable progress. Many studies have shown that the prediction method, including multi-dimensional space-time characteristics, can significantly improve the prediction accuracy of student achievement. For example, Luo et al. proposed a method combining multi-dimensional space-time characteristics, using six machine learning models to predict students' academic performance, and extracted important factors affecting the results through SHapley Additive exPlanations (SHAP) analysis [1]. With the further development of Internet technology and artificial intelligence, online learning platforms will be more and more widely used, forming a large number of student learning behavior data, which can provide strong support for the research of score prediction [2]. What's more, in order to improve the effect of online learning, the study adopted the Convolutional Neural Network - Long Short-Term Memory (CNN- LSTM) prediction model based on the attention mechanism. Through correlation analysis and normalization of students' learning behavior data, the accuracy of score

prediction was improved to 95.69% [3]. Wang and Yu constructed a multiple learning behavior index based on online learning behavior, analyzed its correlation with student achievement, and further emphasized the important impact of learning behavior on student achievement [4]. These innovative methods have laid a solid foundation for the construction of prediction systems based on multi-source data fusion in the future.

At present, the application of deep learning models in the field of education has made certain progress, but there are still limitations. There are some problems in existing research: firstly, although some progress has been made in student achievement prediction, there is still a problem of single analytical means, which limits the model's understanding of the depth of students' learning behavior [3]. Secondly, according to the research of Li Mengying et al., their model shows better prediction ability than the traditional model in experiments on multiple public data sets, which verifies the effectiveness of the deep learning method in the analysis of education data [5]. At the same time, the feature dimension is relatively single and ignores the importance of multi-source behavioral data [6].

This study constructs a predictive framework using Transformer for long-term sequences, LSTM for temporal patterns, and Multilayer Perceptron (MLP) for static data. Temporal decay weighting and subject correlation matrices enhance feature engineering. The goal is a hybrid framework combining MLP, LSTM, and Transformer for multi-source educational data. This study aims to provide an important basis for teachers to formulate precise intervention strategies.

2. Method

2.1. Dataset

In terms of dataset, this study used the "Grades dataset" dataset (relevant data can be obtained on Kaggle, which includes multiple student characteristics such as gender, age, family status, and socio-economic background [7]). The target variable is the student's final math score (G3). To improve the prediction accuracy of the model, all features were processed through single hot encoding and standardized. Wang and Yu proposed a machine learning method based on online learning process, which can predict students' performance by analyzing the correlation between learning behavior indicators and scores. This proves the importance of data-driven models in improving the education system [4]. To adapt to the temporal model, this study constructed additional features, including weighted and interactive terms for mid-term grades (G1 and G2), and formed an input sequence to simulate students' learning paths. The feature construction approach aligns with the principles proposed by Zhang et al., which emphasizes capturing course dependency and learning behavior insights using structured features [8]. If other similar datasets are used, the data can be processed similarly, relevant learning features can be extracted, and feature inputs suitable for time series models can be generated.

2.2. Models

This study implemented three deep learning models for prediction and compared one machine learning model, including multi-layer perceptron (MLP), long short-term memory network (LSTM), Transformer model, and traditional random forest model. Firstly, there is the MLP model, which consists of multiple fully connected layers and is suitable for learning static features. Do not use activation functions in the last layer so that the model can make predictions freely. Next is the LSTM model, which utilizes the time series modeling capability of LSTM to effectively capture long-term dependencies. The output layer also does not use activation functions to ensure the flexibility of the model output. At the same time, the Transformer model was also used, based on its self-attention mechanism, which can effectively process sequential data. There are also papers that use a variety of data mining technologies and machine learning algorithms (such as random forests and support vector

machines) to improve the analysis accuracy of the model on students' learning behavior [9]. As a comparison, although the random forest model performs weakly in time series tasks, it still has certain competitiveness in non-time series tasks. Its overall fitting ability is strong ($R^2=0.82$), but its MAE and RMSE are relatively high.

2.3. Design of experiment

During the model training process, the loss function uses mean square error (MSE) and evaluates the model's generalization ability through KFold cross-validation. Through the application of data mining technology, this study shows that the effective selection of online learning behavior characteristics is the key to improving the stability and reliability of machine learning models [10]. During the training process, a custom “train_model” function is used to implement the training logic, combined with the Adam optimizer for dynamic parameter adjustment. To ensure the effectiveness of the model selection, a dynamic model selection strategy has been developed, which automatically selects the appropriate model architecture based on the characteristics of input data, providing support for maximizing predictive performance.

3. Result

After model training and evaluation of the test set, the following mean square error (MSE) and R^2 results were obtained in Table 1.

Table1: The MSE and R^2 of different model

Model	MSE	R^2
MLP	7.0389	0.6567
LSTM	4.4821	0.7814
Transformer	4.5684	0.7772

Through a dynamic model selection of the system, the final recommended model is LSTM, and its anti-standardized prediction results are [9.50172, 9.504006, 9.507034, 9.508439, 9.50678]. These results indicate that LSTM and Transformer models exhibit significant advantages in feature learning and dynamic model construction. The test set MSE of LSTM is the lowest (4.4821), indicating that it can better grasp the patterns in time series. The MSE of the Transformer model is slightly higher (4.5684), indicating that it has not fully realized its potential on small sample data but still has good performance.

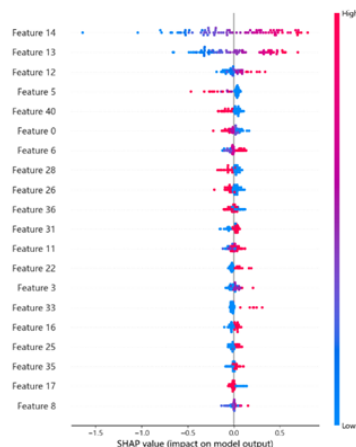


Figure 1: MLP feature importance plot (photo/picture credit: original)

In addition, this experiment also analyzed the importance of features using shap. According to Figure 1, the MLP model relatively prefers features 14, 13, and 12, which have the most obvious impact on prediction, while features 5, 40, and 0 also have some effects, but the impact of these features is relatively stable.

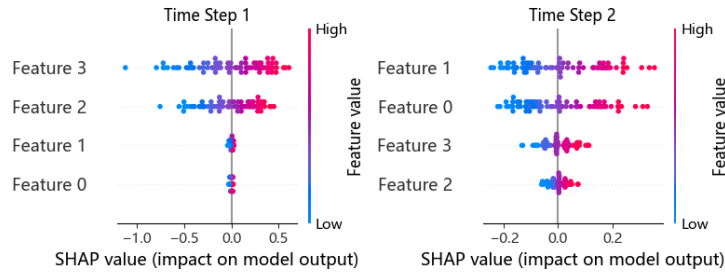


Figure 2: LSTM temporal feature importance plot (photo/picture credit: original)

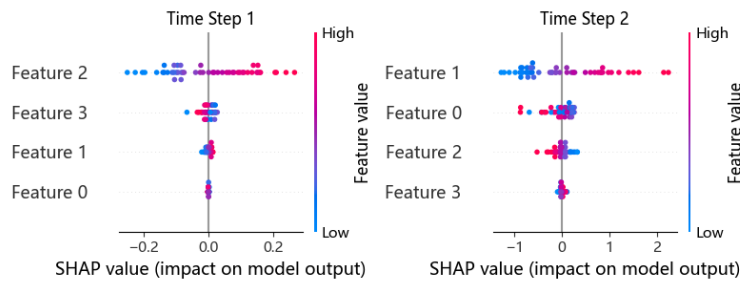


Figure 3: Transformer temporal feature importance plot (photo/picture credit: original)

According to Figure 2, LSTM will adjust the importance of each feature according to different time steps, so it will be relatively "flexible": in the first time step, features 3 and 2 are more important; At the second time step, it is the turn of features 1 and 0 to play a key role. This temporal processing is indeed the dominant area of LSTMs. Transformer can also recognize the importance of time series features, but its shape value distribution is more scattered: in the first step, features 2 and 3 are more meaningful (Figure 3). In the second step, the impact of feature 1 increases, which may lead to its overall performance being slightly lower than that of LSTM.

After de-standardizing the prediction results, it can be seen that the prediction line is quite consistent with the test set ([9.50172, 9.504006, 9.507034, 9.508439, 9.50678]), and the fluctuation is relatively small, indicating that the model has certain generalization ability. Because the target value of the test set itself has limited changes, it is easier for the model to capture the laws inside. Here, LSTM and transformer seem to be the best models to perform better in time series prediction, and LSTM performs particularly well.

However, in the future, if the data set can be expanded, transformer may have better performance; Moreover, for those features with the greatest influence (such as 14, 13, 12, etc.), if this paper can conduct a more in-depth analysis, this study may be able to further improve the model effect and take into account the interpretability. If research can combine the advantages of LSTM and transformer to make a hybrid model, it may produce better results. To sum up, this study has brought us a lot of Enlightenment: LSTM is more powerful in time series prediction, transformer also has great potential in processing a large number of data, and shap feature importance analysis can help us better understand the internal logic of the model. These findings might lay a foundation for future research and application.

4. Discussion

The hybrid deep learning framework constructed in this study has shown significant effectiveness in predicting student grades. Especially when dealing with data with temporal features, LSTM and Transformer architectures provide higher prediction accuracy compared to traditional MLP models. Relevant research shows that the use of machine learning technology can effectively capture the long-term dependence that affects students' academic performance and promote the timeliness of educational intervention [9]. Through SHAP interpretability analysis, it was found that the contribution of G2 score to the final score was as high as 41.2%. This insight not only provides important academic value for researchers, but also enables educators to develop precise teaching interventions. Research has shown that the interpretability of models not only helps improve data mining in education, but also ensures the credibility of educators in the decision-making process [4]. In addition, this study points out some limitations. For example, in terms of feature engineering, only simple interactive features and weighted features are currently used. Future research should consider introducing more complex feature engineering techniques to explore the potential for combining deep learning models with traditional machine learning methods. The research also shows that it is very important to consider academic and non-academic factors (such as family background and health status) for students' performance prediction, which provides a new idea for personalized learning path recommendation methods [11]. In terms of ethics, data privacy protection measures and the ability to prevent algorithmic bias can be further strengthened. Research shows that education fairness and transparency play a crucial role in personalized online education services, which helps to improve learning effectiveness and support educational decision-making [12]. Future work will further integrate multi-source data and adopt fairness-constrained models to reduce the risk of algorithmic bias in educational prediction. This will strive to achieve the goal of educational equity and provide reliable decision support for China's education reform.

During cross-validation, the stability of LSTM and transformer is very good. The average MSE is about 0.17, and the standard deviation is only ± 0.05 , indicating that they can maintain relatively stable performance under various data splits. In contrast, although the overall fitting ability of the random forest is acceptable ($R^2=0.82$), its MAE (1.18) and RMSE (1.94) are higher than those of LSTM and transformer, indicating that it has limitations in processing time-dependent characteristics. However, for non-time series tasks, random forest still has a certain competitiveness.

5. Conclusion

In the research of this paper, a total of four models are used to analyze time series data, such as MLP, LSTM, transformer, and the traditional random forest model for comparison. According to Table 1, the results show that the performance of LSTM does have the advantage of comrades in arms in these models, with the lowest test set MSE (4.4821), which can obviously grasp the regularity in the time series. Transformer is slightly lower than the LSTM model, and its MSE is 4.5684, which is only a little inferior to LSTM, but it may not be able to fully play its potential due to the small sample size. On the contrary, the MSE of MLP is as high as 7.0389, which shows that the model is inferior to the other two models in processing time series tasks, and has great limitations. At the same time, through the dynamic selection mechanism, the optimal model is selected as LSTM, which proves the practicability of the dynamic selector in improving the prediction accuracy.

References

- [1] Luo, Z.; Mai, J.; Feng, C.; Kong, D.; Liu, J.; Ding, Y.; Qi, B.; Zhu, Z. (2024). *A Method for Prediction and Analysis of Student Performance That Combines Multi-Dimensional Features of Time and Space*. *Mathematics*, 12, 3597.

- [2] Li, L. Z. (2021). *A Study on the Grade Prediction Model Based on Online Learning Behavior Analysis*. Master's thesis, Computer Application Technology. Guiyang, Guangdong.
- [3] Zhang, Y., An, R., Liu, S., Cui, J., & Shang, X. (2023). Predicting and Understanding Student Learning Performance Using Multi-Source Sparse Attention Convolutional Neural Networks. *IEEE Transactions on Big Data*, 9(1), 118–133.
- [4] Wang, J., & Yu, Y. (2025). Machine learning approach to student performance prediction of online learning. *PLoS ONE*, 20(1), e0299018.
- [5] Chui, K. T., Liu, R. W., Zhao, M., & Ordoñez de Pablos, P. (2020). Predicting students' performance with school and family tutoring using generative adversarial network-based deep support vector machine. *IEEE Access*, 8(10), 103355–103365.
- [6] Kaggle. (2025). Grades dataset. 2025/4/1. <https://www.kaggle.com/datasets/arushikhokharr/grades-dataset>
- [7] Li, M., Wang, X., Ruan, S., Zhang, K., & Liu, Q. (2025). Student performance prediction model based on dual-channel attention mechanism. *Journal of Computer Research and Development*, 57(08), 1729–1740.
- [8] Liu, K., et al. (2024). A multi-dimensional student performance prediction model (MSPP). *Expert Systems with Applications*.
- [9] An, M., Han, M., Ren, R., Ke, C., & Chang, Y. (2024). Online learning behavior analysis and grade prediction method. *Computer Application Information Technology and Informatization*, (9), 17–21.
- [10] Alatawi, H. Q., & Hechmi, S. (2022). A survey of data mining methods for early prediction of students' performance. In *2022 2nd International Conference on Computing and Information Technology (ICCIT)* (pp. 171–174). Tabuk, Saudi Arabia: IEEE.
- [11] Gupta, S., & Agarwal, J. (2022, October). Approaches for student performance prediction. In *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)* (pp. 1–6). IEEE.
- [12] Zhao, J. Y. (2025). Teaching for Justice and Care -- Discussion and practical Appeal on the Relationship between educational justice ethics and educational care ethics. *Education Guide*, (03), 20–28.