Majumdar, R. et al. (Eds.) (2025). Proceedings of the 1st International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education

Predicting Academic Performance from Student Behaviors and Content Engagement

Ye YUAN[a*] & Brendan FLANAGAN[b] & Runze HOU[c]

[a] *Graduate School of Informatics, Kyoto University, Japan*

[b] *Center for Innovative Research and Education in Data Science, Kyoto University, Japan*

[c] *School of Optoelectronic Engineering and Instrumentation Science, Dalian University of Technology, China*

*adamputh0412@gmail.com

Abstract: The digitalization of education has produced vast learning data, enabling AI-driven analytics to enhance teaching and learning. A key task is predicting academic performance to identify students needing extra support to pass. Much work has been done to improve the results of this task and their interpretation. However, prior research often isolates student behavior and learning materials, leading to a possible disconnect between learning content and behavior. This study bridges that gap by analyzing how students' interactions with digital content can predict academic performance. Educational log data are preprocessed to extract meaningful features that represent a combination of behavior and content, which are then refined using the Null Importance method. A LightGBM model is trained for prediction, and SHAP analysis reveals key behavioral factors linked to success. Results show that high-performing students engage more strategically and actively with critical materials. By combining behavioral and content-based analytics, this study offers a framework for early detection of learning issues and supports targeted, adaptive interventions to improve student outcomes.

Keywords: Learning analytics, student performance prediction, digital learning behavior, LightGBM, SHAP.

1. Introduction

With the rapid digitalization of education, Learning Management Systems (LMS) and digital materials have become widely adopted, generating vast amounts of educational data. This has sparked growing interest in data-driven education, where learning analytics and artificial intelligence enhance teaching and learning. However, research on effectively analyzing and utilizing such data is still in its early stages and there remains a limited understanding of what learner data can be collected, how to analyze it, and how to apply it to improve outcomes (Romero & Ventura, 2020). This study addresses these gaps by examining how students' interactions with digital learning resources can be used to predict academic performance and uncover behavioral patterns linked to effective learning. We aim to support personalized and adaptive learning systems by developing and evaluating analytical methods.

Related work has investigated factors influencing academic performance based on student interactions with digital platforms. For example, Paas et al. (1994) found that engaging with specific content and key pages correlates with better grades. Other research has linked behaviors like highlighting, reviewing, and study timing to academic achievement (Akçapinar et al., 2020; Flanagan et al., 2022; Oi et al., 2015), while device usage has also been shown to impact outcomes (Sung et al., 2016). These findings highlight the importance of both behavioral and contextual factors.

However, many previous studies examine behavior and content separately (Akçapinar et al., 2020; Flanagan et al., 2022). Our research integrates these aspects, recognizing that not only what students study but also how they engage with materials affects performance. For instance, briefly viewing a key resource may offer little benefit, whereas deep engagement, such as note-taking or repeated review, can lead to better outcomes.

To test this, we preprocess behavioral log data and engineer features such as binary indicators for material access. Due to the large dataset, we apply the Null Importance method to retain only the most relevant features. A LightGBM model is then trained to predict academic performance. Finally, we use SHAP (SHapley Additive Explanations) to interpret the model's predictions. SHAP identifies the behavioral patterns and learning materials most influential to student success, providing actionable insights for early intervention and personalized support. This integrated approach demonstrates the value of combining behavioral and content-based analytics to enhance educational outcomes.

2. **Method**

2.1 *Dataset*

The dataset used in this study originates from the 4th Educational Data Analysis Competition (EDE, 2025) and was collected via the LEAF Platform (Flanagan & Ogata, 2018). It contains 597,832 student operation records and score data for up to 1,000 students. The dataset comprises four files: three hierarchical operation log files, containing data from 5 weeks, 10 weeks and all weeks of a 15 week Japanese junior highschool course. Another file contains the final cumulative student score record. Each operation log records detailed learning activities, including the following attributes:

- *userid*: A unique identifier for each student.
- *contentid*: The ID of the learning material accessed.
- *operationname*: The type of operation, such as "OPEN," "ADD BOOKMARK," "ADD MARKER," and so on.
- *pageno*: The page number within the content where the operation occurred.
- *marker*: Annotations made by students, such as marking content as "difficult" or "important."
- *memo_length*: The length of any notes added by the student.
- *devicecode*: The type of device used (e.g., "pc" or "mobile").
- *eventtime*: The timestamp of the operation.

The final academic performance contains students' quiz scores:

- *userid*: A unique identifier for each student.
- *score*: The quiz score of the corresponding student.

By integrating the operation logs with quiz scores, this dataset enables the analysis of learning behavior patterns and their correlation with academic performance. The data structure allows for investigating factors such as how different engagement strategies influence student outcomes and which operations are most predictive of success.

2.2 *Data Preprocessing and Feature Engineering*

Our method adopts the same model architecture as the previous method (Akçapinar et al., 2020; Flanagan et al., 2022), but the key difference lies in our data processing and feature engineering. The previous method treats content and operations as separate factors, assuming only important content or operations affect student performance. This is problematic, for instance, if a student skips essential content, it cannot be assumed they've mastered it. Our method addresses this by considering both content importance and the type of operations performed.

We implemented a structured preprocessing pipeline to convert raw operation logs into meaningful features. While we retain some common steps, such as deduplication and standardizing operation names, our method differs significantly in how learning behavior is quantified. We filtered content and page-level operations using a quantile-based threshold to retain only frequently accessed items. We then generated features such as binary content access flags, operation counts, estimated time spent per content, and the number of distinct operation types to reflect behavioral diversity. We introduced n-grams (n = 1, 2, 3) from operation sequences within each content to model sequential behavior. We also incorporated user annotations (e.g., marking content as “important” or “difficult,” note frequency/length), device usage, and temporal features like hourly activity, weekday usage, and study intervals (first 5, 10, and 15 weeks). Finally, we applied the Null Importance method for feature selection and ensured all inputs were formatted as floating-point values for model compatibility. This end-to-end pipeline yields a more behaviorally informative dataset, leading to improved performance prediction compared to previous methods.

2.3 *Model Selection and Analysis*

To accurately predict student performance, we employ LightGBM, a gradient-boosting framework known for its efficiency, scalability, and ability to handle large datasets with missing values (Ke et al., 2017). The model is trained on preprocessed student operation data and quiz scores, with multiple iterations to improve robustness. The dataset is split into training and test sets based on study periods (5, 10, and 15 weeks), using a random sampling strategy over 10 trials to ensure diversity. After feature selection, LightGBM is tuned with key hyperparameters such as the number of leaves and minimum child samples to balance model complexity and generalization. Its histogram-based algorithm reduces memory usage and training time, while support for early stopping and feature importance makes it well-suited for educational data mining.

For interpretability, SHAP (SHapley Additive Explanations) is employed to analyze feature contributions (Lundberg & Lee, 2017). SHAP provides global and local explanations, fairly attributing each feature’s impact even when features are correlated. This helps identify key behavioral patterns associated with academic success and offers actionable insights for personalized learning. LightGBM and SHAP form a transparent and practical framework for predicting student performance and analyzing educational data.

2.4 *Evaluation*

Root Mean Squared Error (RMSE) is used to evaluate the model’s predictive accuracy by measuring the difference between predicted and actual quiz scores. Since RMSE penalizes more significant errors more heavily, it provides a reliable indicator of overall prediction quality. To assess model performance, we evaluate RMSE across test sets corresponding to study periods of 5, 10, and 15 weeks. The evaluation is repeated over multiple iterations to ensure result stability, and the RMSE values are averaged. This approach allows us to examine how well the model generalizes across different stages of student learning.

3. Results

3.1 *Prediction Performance*

We conducted comprehensive experiments to evaluate our method (OM) against a previous method (PM), using Root Mean Squared Error (RMSE) as the evaluation metric. As shown in **Table 1**, OM outperforms PM particularly in the shorter periods achieving lower RMSE values, and narrower confidence intervals in others, indicating more accurate and stable predictions of student performance. In the table, LR denotes the learning rate, and Train represents the RMSE on the training set. The evaluation results are reported across four subsets: 5w, 10w, 15w, and All, each corresponding to different student operation data time frames. Precisely, 5w reflects RMSE using the first five weeks of data, 10w uses the first ten weeks, 15w includes the first fifteen weeks, and All represents the overall mean RMSE across the entire evaluation set.

Table 1. Mean RMSE *Prediction Performance Between OM and PM.*

	Train	5w	10w	15w	All
PM (LR=0.05)	0.5426±0.0310	8.5902±1.0052	7.6383±0.6442	7.4984±0.6938	7.9330±0.6997
OM (LR=0.05)	0.4916±0.0416	8.1084±1.0906	7.6939±0.5404	7.5190±0.5864	7.7914±0.6288
PM (LR=0.01)	2.8137±0.0520	8.5941±1.0581	7.5549±0.6346	7.5481±0.7513	7.9265±0.7088
OM (LR=0.01)	2.6372±0.0626	8.1782±1.1648	7.6121±0.4954	7.5514±0.6842	7.8000±0.6829

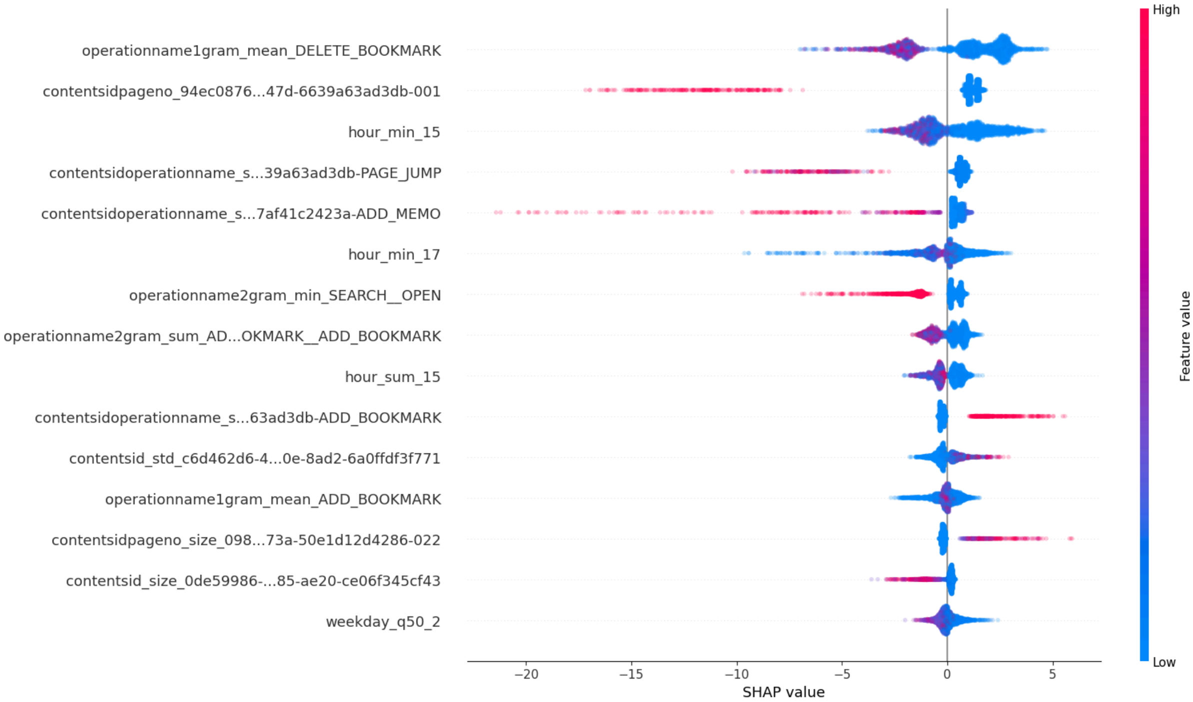

Figure 1. SHAP value(Learning Rate=0.05)

3.2 *Feature Importance*

After completing all experiments, we conducted a SHAP analysis to evaluate the contribution of each feature to the model's predictions. The results are shown in Figure 1, where the vertical axis lists the features, the color bar indicates feature values, and the horizontal axis shows their positive or negative impact on the predicted outcomes. Due to space constraints, only the top 15 features are displayed. This analysis offers practical insights into which student behaviors are most influential. For instance, the feature *contentidoperationname_s…63ad3d3db-ADD_BOOKMARK* reflects how often a student bookmarked a specific learning material—an action associated with better academic performance, likely indicating perceived value or intent to revisit. In contrast, *contentidoperationname_s…39a63ad3db-PAGE_JUMP* captures how often students skipped that content, which is linked to lower performance, possibly due to disengagement.

These findings highlight the importance of learning content and student interaction patterns, offering valuable guidance for instructional strategies and validating the effectiveness of our approach.

4. Discussion and Conclusion

In summary, we propose a feature processing method that outperforms previous approaches and provides practical value for identifying effective learning materials and improving teaching strategies. Our results show that not all frequently used materials equally benefit student performance; more important than access frequency is how students engage with the content—deep, intentional operations such as note-taking and repeated review lead to better outcomes, while passive use has limited impact thus reaffirming results in previous research (Akçapinar et al., 2020). Unlike prior work on usage counts or simple metrics (Paas et al., 1994; Akçapinar et al., 2020), our method incorporates engagement quality and context. By highlighting the importance of meaningful operation, this study contributes to learning analytics. It supports the development of adaptive systems that recommend key materials and effective ways to use them. Some possible limitations to this method are that it will be dependent on specific learning materials that were present in the training data, and this could hinder the generalizability of the method for predicting performance on reading behavior. However it highlights that the analysis of learning contents in addition to reading behavior could lead to better performance prediction and warrants further investigation in future research.

Acknowledgments

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) JP20H01722 and JP23H01001 (Exploratory) JP21K19824.

References

Akçapinar, G., Chen, M. R. A., Majumdar, R., Flanagan, B., & Ogata, H. (2020, March). Exploring student approaches to learning through sequence analysis of reading logs. In Proceedings of the tenth international conference on learning analytics & knowledge (pp. 106-111).

EDE (2025). 4th Educational Data Analysis Competition. https://sites.google.com/view/ede-datachallenge-4th

Flanagan, B., & Ogata, H. (2018). Learning analytics platform in higher education in Japan. Knowledge Management & E-Learning: An International Journal, 10(4), 469-484.

Flanagan, B., Majumdar, R., & Ogata, H. (2022). Early-warning prediction of student performance and engagement in open book assessment by reading behavior analysis. International Journal of Educational Technology in Higher Education, 19(1), 41.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Misato, O. I., Okubo, F., Shimada, A., Chengjiu, Y. I. N., & Ogata, H. (2015, November). Analysis of Preview and Review Patterns in Undergraduates' E-Book Logs. In the International Conference on Computers in Education (pp. 166-171).

Paas, F. G., & Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. Educational psychology review, 6, 351-371.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley interdisciplinary reviews: Data mining and knowledge discovery, 10(3), e1355.

Sung, Y. T., Chang, K. E., & Liu, T. C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. Computers & Education, 94, 252-275.