

PAPER

An Explainable Early Warning System for Student Academic Risk Prediction Based on Time-Windowing and SHAP

Yanyan Zhu¹, Niu Shi¹, Yan Chen², Peiqiang Wang¹(✉)

¹Pingdingshan Polytechnic College, Pingdingshan, Henan, China

²Pingdingshan Tianan Coal Mining Co., Ltd., Pingdingshan, Henan, China

wpq_313@126.com

ABSTRACT

High dropout rates in online learning remain a persistent challenge, while many early warning systems (EWS) provide risk scores without explaining what instructors should do next. This study proposes an explainable EWS that integrates XGBoost with SHAP under a strict 30-day time-windowing protocol on the OULAD dataset, enhanced with five behavior-pattern features. The 30-day window is validated through a 14/30/60-day ablation, with PR-AUC reported to address class imbalance and robustness tested on the KDD Cup 2015 MOOC dataset. To preserve prospective prediction, OULAD records from students who had already withdrawn before the prediction-window endpoint were excluded. On OULAD Day-30, XGBoost achieved the strongest ranking and F1 performance under student-grouped cross-validation, with AUC 0.7865, PR-AUC 0.7652, and F1 0.6745. Day-14 reduced AUC to 0.7353 due to sparse early assignment submissions, whereas Day-60 improved AUC to 0.8154 but reduced intervention time. On KDD Cup 2015, XGBoost achieved AUC 0.8639 and PR-AUC 0.9498 under username-grouped cross-validation. SHAP identified first-assignment score, assignment lead time, active days, and selected academic-background features as key risk drivers, supporting targeted interventions rather than generic risk alerts.

KEYWORDS

educational data mining (EDM), early warning system, explainable artificial intelligence, SHAP, XGBoost, online learning

1 INTRODUCTION

Distance education and massive open online courses (MOOCs) have expanded rapidly in recent years [1], [2], but the flexibility of online learning is accompanied by high attrition and dropout rates [3]. To address this, researchers have applied educational data mining (EDM) and Learning Analytics (LA) to build early warning systems (EWS) for students at academic risk [4], [5].

Zhu, Y., Shi, N., Chen, Y., Wang, P. (2026). An Explainable Early Warning System for Student Academic Risk Prediction Based on Time-Windowing and SHAP. *International Journal of Emerging Technologies in Learning (iJET)*, 21(4), pp. 41–55. <https://doi.org/10.3991/ijet.v21i04.62480>

Article submitted 2026-06-03. Revision uploaded 2026-07-18. Final acceptance 2026-07-19.

© 2026 by the authors of this article. Published under CC-BY.

Although deep-learning and ensemble-tree models keep improving prediction accuracy [6], their “black-box” nature limits frontline adoption: a probability score alone cannot tell an instructor whether a student needs academic remediation, motivational support, platform-access assistance, or a simple follow-up message, whereas surfacing the underlying risk causes can [7], [8]. Many existing studies further rely on full-semester data for retrospective prediction, leaving little time for action [9].

Classical statistical methods—logistic regression and support vector machines—have long served as baselines for student-failure prediction [10], [11]. However, educational data are heterogeneous tables indexed by student–course–time, mixing continuous behavioral logs with high-cardinality categorical attributes [12], which makes ensemble tree models—handling non-linear interactions and missing values natively—a more natural fit [13]. XGBoost [14] in particular has become a mainstream choice in learning analytics due to its regularization and efficiency, and recent benchmarks confirm that XGBoost + SHAP can provide both accurate predictions and interpretable results on academic-risk tasks [15]. Other gradient-boosting variants such as Random Forest, LightGBM [16], and CatBoost reach comparable performance.

Deep sequence models (RNN or LSTM, Transformers) can capture temporal dependencies without manual aggregation, but on mid-sized academic datasets such as OULAD (~30k records) their gains are typically offset by overfitting and weaker interpretability [6], [17]. On larger MOOC datasets such as KDD Cup 2015 (~120k records) the benefits are more apparent, yet additional training cost and opacity limit frontline use. We therefore adopt XGBoost as the primary classifier and retain LR and linear SVM as baselines.

A specialized line of work focuses on early warning itself: unlike retrospective full-semester prediction, an EWS must judge risk from a short post-enrollment signal window [9]. Two strategies dominate—fixed windows (e.g., first two weeks or first month) and rolling/cumulative windows updated daily or weekly [18]. Fixed windows are simpler and cross-comparable, but the chosen window length is rarely justified empirically. Section 3.3 closes this gap with a 14/30/60-day ablation paired with feature-coverage and offline-range diagnostics.

Explainable AI (XAI) is increasingly used in education [7]. SHAP [19], grounded in Shapley values and satisfying additivity, symmetry, and consistency, has become a common method for explaining tree-based models, while LIME [20] remains a complementary post-hoc approach. However, prior XAI work in learning analytics often remains limited to global feature-importance analysis and provides insufficient support for per-student diagnosis or teacher decision-making [21]. Although SHAP has begun to appear in EDM [22], few studies simultaneously address three requirements: strict time-windowing to avoid retrospective attribution, combined global/local explanations grounded in instructionally meaningful cases, and similar findings across datasets. Sections 2.2, 3.3, 3.4, and 3.6 address these requirements, with KDD Cup 2015 providing an additional test setting.

In summary, this study proposes an explainable EWS with four contributions: (i) a strict 30-day time-windowing strategy validated by a 14/30/60-day ablation and protected against both student-level leakage and early-withdrawal outcome leakage; (ii) tree-model evaluation that adds PR-AUC alongside Recall and F1 to remain informative under class imbalance [23]; (iii) SHAP analysis that links risk scores to concrete signals teachers can inspect, distinguishing academic-performance bottlenecks from engagement-disconnection patterns; and (iv) cross-dataset verification on OULAD [24] and KDD Cup 2015 [25], aligning with recent calls for ML+XAI methods that can inform practical teaching interventions [26].

2 MATERIALS AND METHODS

The explainable early warning system framework proposed in this study comprises four primary phases: data input and preprocessing, time-windowing and feature engineering, model training and evaluation, and SHAP-based explainability and intervention guidance. The intended output is not merely a binary warning flag, but a summary instructors can use to connect each high-risk prediction with observable evidence such as weak early assessment performance, delayed submission, inactivity, or loss of engagement rhythm. The methodological framework is illustrated in Figure 1.

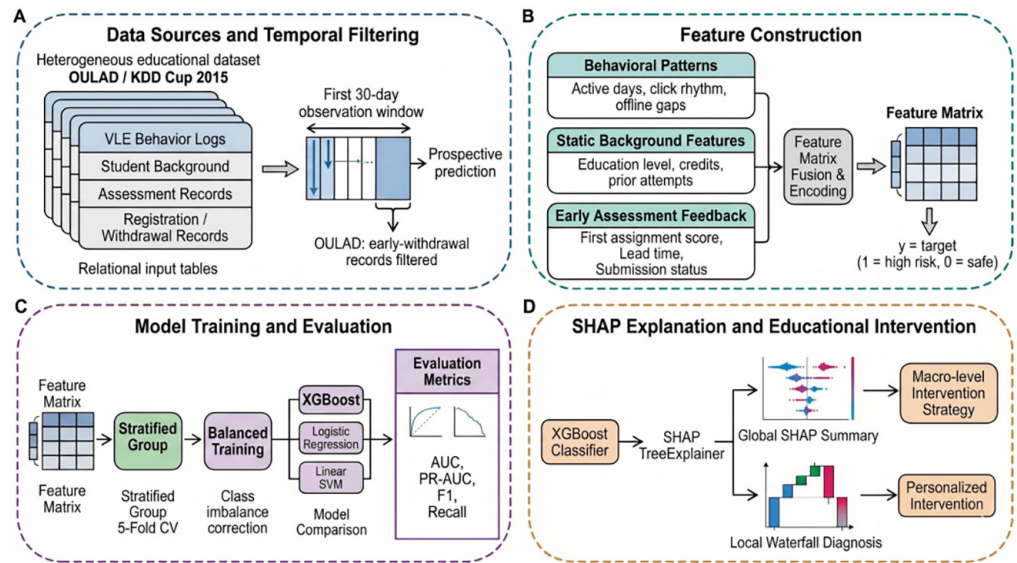


Fig. 1. System framework

2.1 Dataset description

This study utilizes the open university learning analytics dataset (OULAD) [24], a relational database containing student demographics, learning behavior logs, academic assessment results, and registration records. The study defines a binary target variable: students who pass or achieve distinction are labeled as 0 (safe), while students who fail or withdraw after the prediction-window endpoint are labeled as 1 (high risk). To preserve the prospective nature of the early-warning task, withdrawal timing is controlled using `date_unregistration` from `studentRegistration.csv`: records with `date_unregistration` less than or equal to the window endpoint are excluded from the corresponding evaluation. The OULAD raw table contains 32,593 student-course registration records; after excluding withdrawals already observed by Day-30, the main Day-30 modeling sample contains 27,474 records.

2.2 Time-windowing and feature engineering

To ensure the forward-looking nature of the early warning system, this study sets the time window as the closed interval from the first day of the course to the 30th day, excluding pre-heating behavioral data before the course begins. Feature extraction primarily encompasses two dimensions: dynamic learning behavior features and static background features.

Dynamic behavior features are aggregated from students' VLE logs within the 30-day window. Basic engagement is measured by `forum_interactions` and `courseware_clicks`, while five derived behavioral-pattern features capture deeper learning regularities: `max_offline_days` for disengagement risk, `active_days` for learning continuity, `daily_click_std` for distinguishing steady learning from cramming, `activity_type_diversity` for resource-use breadth, and `click_slope_15` for changes in engagement between the first and second halves of the window.

Academic feedback features are organized around the first assignment: the first assignment score (`first_assignment_score`, reflecting early knowledge mastery, strictly adopting submission records where `date_submitted` $\leq$ 30); submission status (`assessment_submitted`, derived from whether `first_assignment_score` is non-empty); and days submitted ahead of the deadline (`assignment_lead_days` = `deadline` – `submission_date`, where positive values indicate early submission and negative values indicate late submission, serving as a proxy indicator for procrastination behavior).

Regarding static background and historical features, in addition to demographic characteristics such as gender, highest education level, age band, geographic region, poverty index, and disability status, dimensions such as course code, credit burden, and historical retake counts are also included. All the aforementioned features are processed via one-hot encoding or categorical mapping to form the final input matrix.

2.3 Model selection

This study adopts XGBoost [14] as the primary classifier. This model has advantages in handling heterogeneous tabular data and capturing complex feature interactions. Meanwhile, Logistic Regression is used as the first set of baseline models; Linear Support Vector Machine (LinearSVC combined with Platt scaling for calibrated output probabilities) is introduced as the second set of baseline models to address the excessively high computational cost of the full OULAD dataset under the traditional RBF kernel.

To ensure comparability between the OULAD and KDD Cup 2015 experiments, XGBoost employs the same set of hyperparameters across both datasets: `n_estimators` = 100, `max_depth` = 4, `learning_rate` = 0.1, and corrects for class imbalance using `scale_pos_weight`. Compared to the default settings (`max_depth` = 6, `learning_rate` = 0.3), this configuration features shallower trees and a lower learning rate, which helps suppress overfitting on medium-scale educational data. Logistic regression and linear SVM are balanced using `class_weight` = 'balanced'. For OULAD, all models are evaluated under 5-fold stratified group cross-validation using `id_student` as the grouping variable, preventing records from the same student from appearing in both training and test folds. KDD Cup 2015 is evaluated under the same grouped-validation principle using `username` as the grouping variable, preventing enrollments from the same learner from appearing in both training and test folds.

2.4 SHAP explainability framework

To overcome the "black-box" limitations of ensemble tree models, this system introduces the SHAP (SHapley Additive exPlanations) framework [19]. Based on Shapley values from cooperative game theory, SHAP is an additive feature attribution method backed by mathematical axioms (such as local accuracy, missingness, and consistency). This study integrates the TreeExplainer—which is specifically optimized for tree-based models—from the SHAP library into the trained XGBoost model [27]. Compared to traditional heuristic feature importance metrics based on Gain or Weight, TreeExplainer can precisely calculate the marginal contribution of

features under complex interactions in polynomial time. Specifically, SHAP values decompose the model output on the raw prediction scale into a linear combination of contributions from all input features. For tree-based binary classifiers, this output can be interpreted as the model margin rather than a direct probability unless probability-based SHAP output is explicitly specified. Under this framework, the system provides interpretability support across two dimensions: at the global level, by aggregating the SHAP values of all test samples to construct directional feature importance plots, revealing the dominant factors behind high-risk academic outcomes; and at the local individual level, by utilizing waterfall plots to quantify the independent “push-and-pull” effects of a specific student’s micro-behaviors (e.g., offline days, assignment scores) on their final high-risk model output, thereby providing educators with risk attribution that can guide follow-up.

3 RESULTS

3.1 Experimental setup

The evaluation protocol covers model performance and explainability analysis. For OULAD, performance is assessed using 5-fold stratified group cross-validation by student identifier, reporting AUC, PR-AUC, Accuracy, Precision, Recall, and F1, with grouped out-of-fold probabilities used for ROC plotting. PR-AUC is included to better evaluate imbalanced settings, especially the KDD Cup 2015 dataset with a 79% dropout rate. For SHAP analysis, diagnostic cases are drawn from a held-out student-grouped fold to avoid overlap between training and explanation samples.

3.2 Model performance comparison

After feature engineering and cross-validation, XGBoost achieves the best AUC, PR-AUC, Accuracy, F1, and Recall, while Linear SVM obtains the highest Precision, as shown in Figure 2 and Table 1.

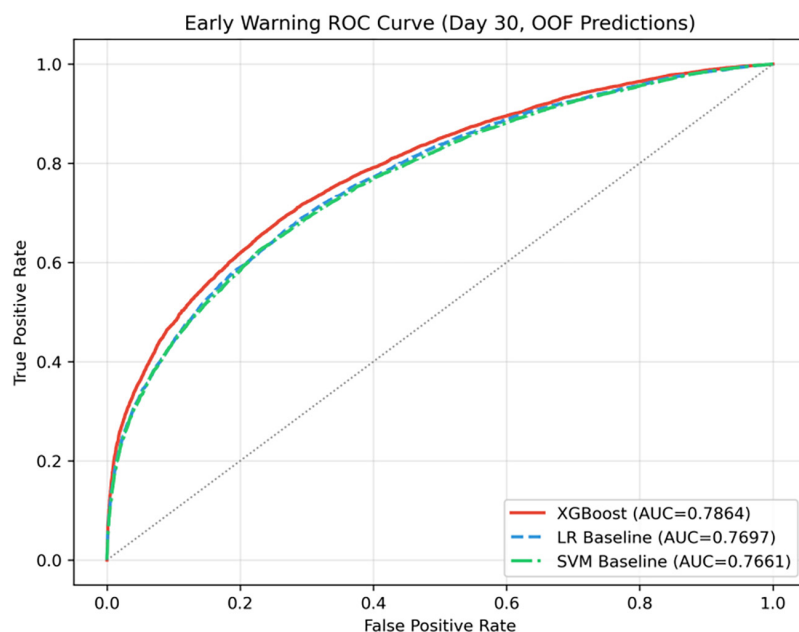


Fig. 2. ROC curve comparison

Table 1. Prediction performance comparison of early warning models (OULAD, Day-30)

Model	AUC	PR-AUC	Accuracy	F1-Score	Precision	Recall
Linear SVM	0.7660	0.7408	0.7053	0.6263	0.7077	0.5619
Logistic Regression	0.7697	0.7449	0.7027	0.6624	0.6618	0.6631
XGBoost	0.7865	0.7652	0.7178	0.6745	0.6846	0.6648

The AUC reaches 0.7865, leading LR (0.7697) and SVM (0.7660) by approximately 1.7 and 2.1 percentage points, respectively; the PR-AUC is 0.7652, which is also the highest among the three models. After excluding students who had already withdrawn by Day-30, the OULAD positive class accounts for approximately 44.0%, and the task becomes a stricter prospective prediction problem rather than retrospective recognition of already observed withdrawals. It is particularly worth noting that the features for the first assignment score and the days submitted ahead of the deadline are strictly limited to submission records where $\text{date_submitted} \leq 30$, avoiding the use of future assignment signals outside the window.

Class balancing was applied during model construction: Logistic Regression and Linear SVM used class-weight adjustment, while XGBoost used positive-negative sample ratio scaling. For early warning, AUC reflects overall ranking ability, F1 balances false negatives and false positives, and Recall measures missed-risk detection. XGBoost achieved the best performance on these key metrics, with an AUC of 0.7865, F1 of 0.6745, and Recall of 0.6648, outperforming both linear baselines. Under the same student-grouped and early-withdrawal-filtered protocol, the added behavioral-pattern features further improved XGBoost's AUC from 0.7797 to 0.7865, indicating that continuous engagement patterns provide additional predictive value. This advantage is attributable to XGBoost's ability to capture non-linear interactions, together with shallow-tree and low-learning-rate settings that help reduce overfitting on medium-scale educational data.

3.3 Time window ablation experiment

To demonstrate the rationality of using the first 30 days of the course as the prediction window, this study compared the model performance under three time windows: 14 days, 30 days, and 60 days, while keeping all other experimental settings (student-grouped 5-fold stratified cross-validation, XGBoost model, identical feature construction process) consistent. The boundary values of the offline days feature and the date_submitted threshold for the first assignment score were adjusted synchronously with the window length to ensure that the signals under different windows maintained consistent criteria and did not incorporate future information from outside the window.

Table 2. Results of time window ablation experiment (OULAD, XGBoost)

Time Window	AUC	PR-AUC	Recall	F1	Intervention Timeliness
Day-14	0.7353	0.6980	0.6470	0.6426	Earliest
Day-30	0.7865	0.7652	0.6648	0.6745	Recommended
Day-60	0.8154	0.7890	0.6929	0.6917	Late

As shown in Table 2, Day-30 clearly outperforms Day-14, with gains of 5.1, 6.7, and 1.8 percentage points in AUC, PR-AUC, and Recall, respectively. This gap is mainly

due to sparse early signals: after excluding students who had already withdrawn, only 13.6% of eligible Day-14 records had submitted the first assignment, compared with 75.9% by Day-30. In addition, the Day-14 window limits max_offline_days to 15 days, reducing its ability to distinguish mild from severe disengagement, whereas Day-30 provides a wider offline-behavior range.

Although Day-60 further improves AUC, PR-AUC, and Recall by 2.9, 2.4, and 2.8 percentage points over Day-30, it reduces the remaining intervention time from approximately 89% to 78% of a typical 9-month OULAD course. Therefore, Day-30 is adopted as the default window because it balances predictive signal coverage with sufficient time for meaningful intervention.

3.4 Cross-dataset generalization framework analysis

To verify the structural robustness of the proposed time-windowing and explainability framework, this study conducted cross-validation on the KDD Cup 2015 large-scale online course dataset. There are substantial differences in task definitions between the two datasets: OULAD predicts failure or withdrawal after the Day-30 prediction point in a full-time distance education scenario (where the positive class accounts for approximately 44% after early-withdrawal filtering), whereas KDD Cup 2015 predicts dropout within the subsequent 10 days after the first 30 days of the course in a MOOC scenario (where the dropout class accounts for as high as 79%).

Table 3. Cross-dataset generalization comparison

Dataset	Task Definition (Positive Ratio)	Model	AUC	PR-AUC
OULAD	Fail/withdrawal after Day-30 (~44%)	XGBoost	0.7865	0.7652
		Logistic Regression	0.7697	0.7449
		Linear SVM	0.7660	0.7408
KDD Cup 2015	Dropout in next 10 days (~79%)	XGBoost	0.8639	0.9498
		Logistic Regression	0.8545	0.9445
		Linear SVM	0.8519	0.9434

As shown in Table 3, 120,542 valid KDD Cup 2015 records with 48 features were evaluated using 5-fold stratified group cross-validation by username. XGBoost achieved the best threshold-independent performance, with AUC 0.8639 and PR-AUC 0.9498, slightly outperforming LR and SVM. Because the dropout class accounts for about 79% of the dataset, PR-AUC is reported alongside AUC to better reflect precision-recall performance under class imbalance. Although Linear SVM achieved high recall at the default threshold, its intervention value is limited because dropouts are already the majority class. Overall, XGBoost shows stronger feature discrimination across both ranking metrics.

The SHAP results show similar patterns across the two datasets: dynamic behavioral and assessment features rank highly, including assignment-related signals, active days, and offline behavior in OULAD, and event counts and offline behavior in KDD Cup 2015. This suggests that observable learning-process features are relatively stable indicators of academic risk, while demographic and static background variables play a secondary role.

3.5 Global explanation: the dominance of learning behaviors and historical backgrounds

To understand the overall drivers of the model's predictions, we analyzed the global feature contributions, as visualized in Figure 3.

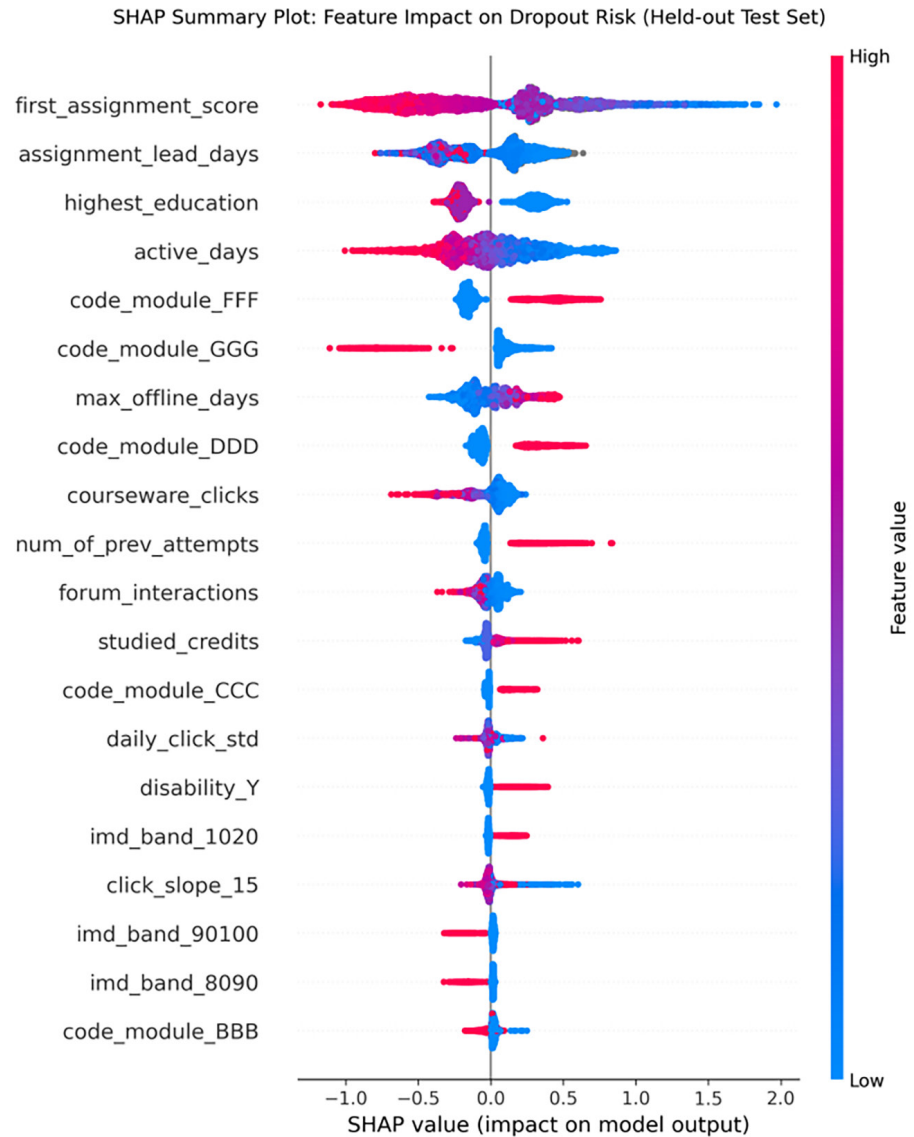


Fig. 3. Global SHAP summary plot

The global SHAP analysis reveals that the strongest contributor is `first_assignment_score`, followed by `assignment_lead_days`, `highest_education`, `active_days`, and several course-module indicators. The maximum consecutive offline days feature remains an important behavioral signal, though after excluding students who had already withdrawn by Day-30, its global rank is lower than the direct assignment-feedback variables. Secondary behavioral indicators, such as the slope between the first and second halves of the window (`click_slope_15`) and forum interactions, rank in the middle. Purely demographic background features (such as gender, poverty index, region, and age band) are relegated to the lowest positions.

In terms of directionality, long periods of non-login, low active days, and a late-stage decline in click rhythm significantly increase the model's high-risk output. Conversely, better initial assignment scores, more days submitted ahead of the deadline, and higher active days exert a protective effect. At an empirical level, this profile supports the view that acquired learning engagement and past academic experience are the core elements determining academic performance: merely "how much was studied" (total clicks) is only a surface signal; "whether the study was continuous, whether assignments were completed early, and whether engagement was maintained under distributed cognitive load" serve as deeper criteria for risk identification.

3.6 Local explanation: individualized diagnostic case studies

By utilizing SHAP waterfall plots, specific attribution can be performed for individual high-risk students, turning generic risk scores into information that can guide student-specific support. Because the waterfall plots are reported on the raw model-output scale, the displayed output values should be read as model margins rather than probabilities; the SHAP signs indicate whether each feature increases or decreases the high-risk output. The two cases below illustrate how the same warning status can imply different instructional responses: one requires academic remediation despite visible engagement, whereas the other requires re-engagement despite acceptable early performance.

Case A:

The risk attribution for a student presenting a typical "high input, low output" profile is detailed in Figure 4.

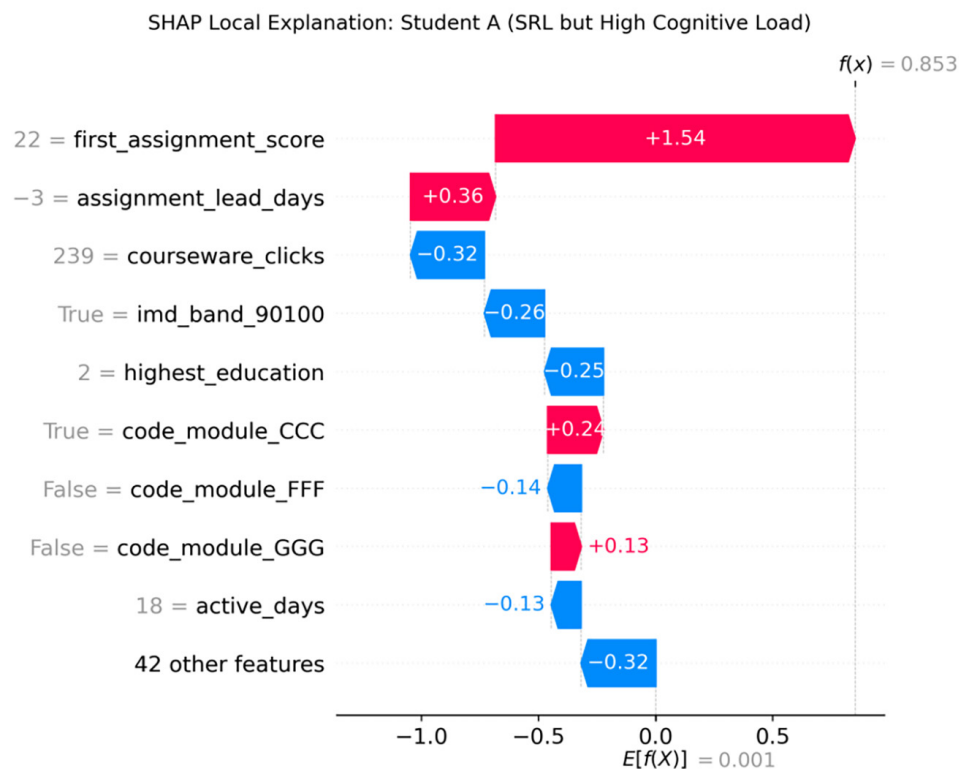


Fig. 4. Local SHAP waterfall plot for Case A ("High Input, Low Output")

This student displayed substantial engagement within the 30-day window, with 239 courseware clicks and 18 active days, both of which reduced the high-risk output. However, the first-assignment score was only 22 (SHAP +1.54), becoming the dominant risk driver. A late submission, reflected by `assignment_lead_days` = -3 (SHAP +0.36), further increased the risk. This profile suggests that the student was behaviorally active but academically unsuccessful. The recommended intervention is therefore not another generic participation reminder, but rapid academic tutoring, diagnosis of prerequisite knowledge gaps, worked-example feedback on the first assignment, and follow-up checking before the next assessment deadline.

Case B:

Conversely, Figure 5 visualizes the risk profile of a student who exhibits a dramatic “disconnected” pattern.

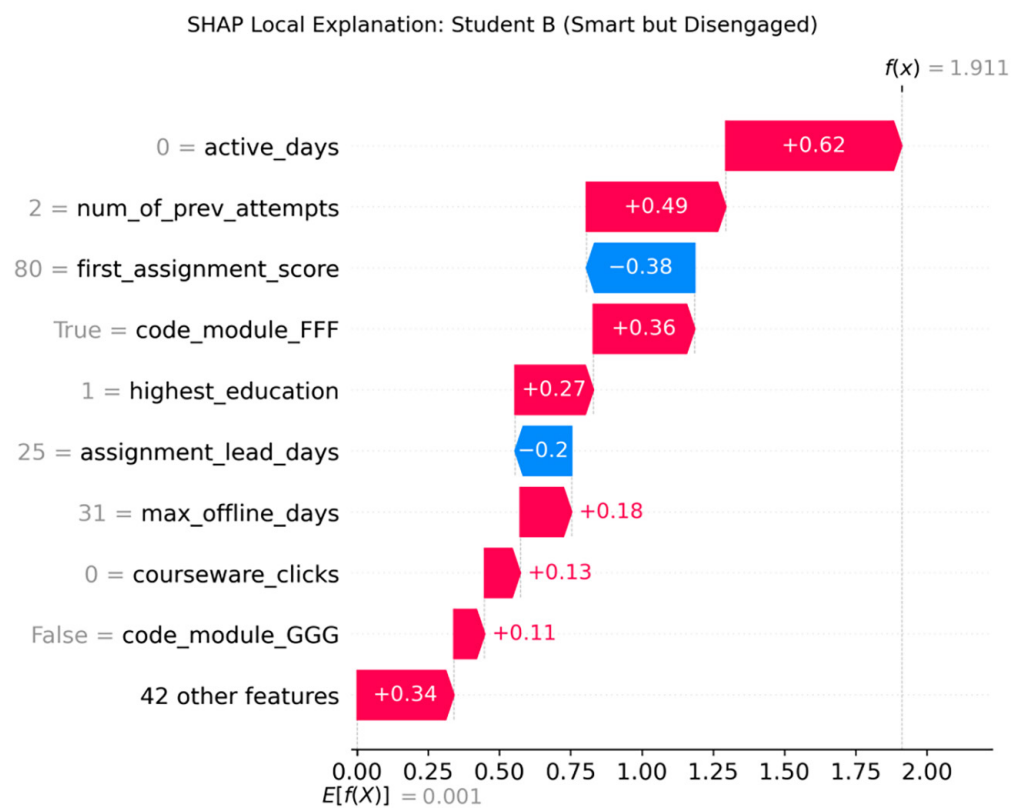


Fig. 5. Local SHAP waterfall plot for Case B (“Disconnected”)

Within the 30-day VLE window, this student showed complete online disengagement, with zero active days (SHAP +0.62), 31 consecutive offline days (SHAP +0.18), and zero courseware clicks (+0.13), all increasing the high-risk output. However, the student achieved a good first-assignment score of 80 (SHAP -0.38) and submitted it 25 days early (SHAP -0.20), indicating a solid academic foundation despite no VLE activity. Therefore, the appropriate intervention is not academic remediation but engagement recovery: instructors should check whether the student is facing platform-access barriers, competing obligations, isolation, or burnout, and then use personalized outreach, study-plan renegotiation, and low-friction re-entry activities to restore participation in the course rhythm.

4 DISCUSSION

4.1 Model performance advantage

The experimental findings converge on three substantive points. First, on the OULAD Day-30 task, XGBoost outperformed both logistic regression and linear SVM on the ranking metrics (AUC and PR-AUC) and on F1, with leads of approximately 1.7 and 2.1 percentage points on AUC over LR and SVM, respectively. This advantage is consistent with recent comparative studies of academic-risk prediction on heterogeneous tabular educational data [10], [15], and is plausibly explained by the tree ensemble's native capacity to model non-linear interactions among assignment scores, offline-day distributions, and engagement-pattern features that linear models can only capture via explicit feature crosses. The shallow-tree, low-learning-rate configuration adopted here (`max_depth` = 4, `learning_rate` = 0.1) further suppresses the overfitting risk typical for medium-scale educational datasets [17]; the five newly introduced behavior-pattern features alone account for an approximate 0.7 percentage-point AUC gain over a version using only basic aggregates.

4.2 Justification of the 30-day window

The 14/30/60-day ablation provides empirical grounding for a design choice that is rarely justified in prior fixed-window EWS work [18], [28]. The 5.1 percentage-point AUC drop from Day-30 to Day-14 is not arbitrary: only 13.6% of eligible Day-14 records (3,832 of 28,127) have a first assignment submitted by Day-14, and the `max_offline_days` feature is structurally capped at 15 days under the inclusive day-0-to-day-14 window, removing the model's ability to distinguish moderately from severely disengaged students. The Day-60 setting reaches a higher AUC (0.8154) but reduces the remaining intervention window from approximately 89% to 78% of a typical 9-month OULAD course. We interpret this trade-off as unfavorable for an EWS, whose practical purpose is to enable timely teacher action rather than to maximize a benchmark score in isolation.

4.3 Behavioral drivers and individualized diagnosis

The SHAP analysis confirms that risk attribution is dominated by early academic feedback and dynamic learning behavior, with static background variables playing a contextual role. The leading features—`first_assignment_score`, `assignment_lead_days`, `active_days`, and `max_offline_days`—are all observable from VLE activity and assessment submissions within the first 30 days, supporting the view that “what students do” strongly informs early academic-risk prediction [21], [29]. The two waterfall cases further demonstrate that students with comparable predicted risk can require entirely different interventions: the “high-input, low-output” archetype (Case A) calls for academic tutoring to address the bottleneck of low assignment performance despite consistently high engagement, whereas the “disconnected” archetype (Case B) requires re-engagement actions to restore VLE participation despite a strong, early-submitted first assignment. In operational terms, the framework helps connect model prediction with teacher action: it helps instructors decide whether to prioritize remediation, engagement recovery, deadline-management support, or

monitoring. A generic warning system that emits only a risk score cannot disambiguate these archetypes; the SHAP-augmented design addresses this gap in existing XAI-in-education work [21].

4.4 Cross-dataset robustness

Cross-dataset validation on KDD Cup 2015—a MOOC dropout task with a positive-class proportion of approximately 79%, in contrast to the Day-30 OULAD task’s 44% positive-class proportion after early-withdrawal filtering—further establishes that the framework is not dataset-specific. XGBoost retained its lead on both AUC (0.8639) and PR-AUC (0.9498) under username-grouped cross-validation, with PR-AUC being particularly informative under positive-skewed distributions where ROC metrics tend to overstate ranking ability [30]. Across both datasets, observable learning-process features such as assignment feedback, activity intensity, and offline behavior remain central to SHAP explanations, suggesting that disengagement and early academic feedback are transferable risk signals rather than dataset-specific artifacts.

4.5 Limitations and future work

Despite the promising results, this study has several limitations that point toward future research. First, regarding the time-windowing strategy, the ablation study only evaluated three discrete settings (14, 30, and 60 days). Future research could explore more fine-grained, sliding, or dynamic prediction windows (e.g., weekly increments) to pinpoint the exact optimal threshold for intervention. Second, the current target variable simplifies the problem by merging “dropout” (withdrawal) and “failure” into a single “high-risk” binary class. However, the teaching responses for a student who formally withdraws differ significantly from those for a student who persists but ultimately fails. Future work should refine this into a multi-class classification task to provide more nuanced early warnings. Third, the contribution thresholds for extracting the SHAP cases were primarily based on empirical settings, lacking a rigorous sensitivity analysis. Finally, while the framework’s robustness under extreme class imbalance was successfully validated using the KDD Cup 2015 dataset, translating this model into a practical teaching tool requires rigorous Randomized Controlled Trials (RCTs) in real-world educational environments to evaluate its actual impact on teacher decision-making and student retention [31], [32].

5 CONCLUSIONS

This study developed an explainable early warning system (EWS) that balances prediction performance, methodological validity, and intervention timeliness through a validated 30-day time-windowing strategy. This window preserves sufficient early behavioral and assessment signals while leaving approximately 89% of the semester for intervention. By integrating SHAP with XGBoost, the system moves beyond black-box risk prediction and provides interpretable risk attribution that can inform instructor follow-up.

The results show that dynamic learning behaviors and early academic feedback are stronger predictors of academic risk than static demographic features. Global and local SHAP analyses further demonstrate that students with similar risk scores

may require different interventions, such as academic tutoring for low-performing but engaged students or re-engagement support for disconnected students. Overall, the proposed framework links machine learning prediction with teaching action by converting risk detection into early, differentiated support. Future work will incorporate multimodal data, such as forum-text semantics, and classroom deployment studies to improve predictive depth and evaluate how explanation-supported alerts influence teacher decision-making and student persistence.

6 ACKNOWLEDGMENT

This work was supported by the Key Scientific Research Project of Higher Education Institutions in Henan Province, China (Grant No. 25A520059).

7 REFERENCES

- [1] T. R. Liyanagunawardena, A. A. Adams, and S. A. Williams, "MOOCs: A systematic study of the published literature 2008–2012," *International Review of Research in Open and Distance Learning*, vol. 14, no. 3, pp. 202–227, 2013. <https://doi.org/10.19173/irrodl.v14i3.1455>
- [2] K. F. Hew and W. S. Cheung, "Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges," *Educational Research Review*, vol. 12, pp. 45–58, 2014. <https://doi.org/10.1016/j.edurev.2014.05.001>
- [3] Y. Wang and R. Baker, "Content or platform: Why do students complete MOOCs?" *J. Online Learn. Teach.*, vol. 11, no. 1, 2015.
- [4] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 3, 2020. <https://doi.org/10.1002/widm.1355>
- [5] C. Lang, G. Siemens, A. F. Wise, D. Gašević, and A. Merceron, Eds., *Handbook of Learning Analytics*, 2nd ed. Vancouver, BC: Society for Learning Analytics Research (SoLAR), 2022. <https://doi.org/10.18608/hla22>
- [6] J. Gardner and C. Brooks, "Student success prediction in MOOCs," *User Model. User-adapt. Interact.*, vol. 28, no. 2, pp. 127–203, 2018. <https://doi.org/10.1007/s11257-018-9203-z>
- [7] Z. Swiecki et al., "Assessment in the age of artificial intelligence," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100075, 2022. <https://doi.org/10.1016/j.caeai.2022.100075>
- [8] C. Brooks and C. Thompson, "Predictive modelling in teaching and learning," in *The Handbook of Learning Analytics*, Vancouver, BC: Society for Learning Analytics Research (SoLAR), 2022. <https://doi.org/10.18608/hla22.003>
- [9] E. Howard, M. Meehan, and A. Parnell, "Contrasting prediction methods for early warning systems at undergraduate level," *Internet and Higher Education*, vol. 37, pp. 66–75, 2018. <https://doi.org/10.1016/j.iheduc.2018.02.001>
- [10] A. Villar and C. R. V. de Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study," *Discover Artificial Intelligence*, vol. 4, no. 1, 2024. <https://doi.org/10.1007/s44163-023-00079-z>
- [11] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in *15th European Concurrent Engineering Conference 2008, ECEC 2008 – 5th Future Business Technology Conference, FUBUTEC 2008*, 2008.
- [12] A. A. Mubarak, H. Cao, and S. A. M. Ahmed, "Predictive learning analytics using deep learning model in MOOCs' courses videos," *Educ. Inf. Technol. (Dordr.)*, vol. 26, no. 1, pp. 371–392, 2021. <https://doi.org/10.1007/s10639-020-10273-6>

- [13] N. Saini, C. Joseph, J. Rupani, J. Sridevi, A. Pankajam, and A. Jain, “Enhancing student retention and performance prediction in E-learning environments using supervised machine learning models,” *Journal of Informatics Education and Research*, vol. 5, no. 4, pp. 2137–2144, 2026. <https://doi.org/10.52783/jier.v5i4.4055>
- [14] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>
- [15] A. Bettahi, F. Z. Belouadha, and H. Harroud, “A modular and explainable machine learning pipeline for student dropout prediction in higher education,” *Algorithms*, vol. 18, no. 10, p. 662, 2025. <https://doi.org/10.3390/a18100662>
- [16] G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems*, 2017.
- [17] L. Grinsztajn, E. Oyallon, and G. Varoquaux, “Why do tree-based models still outperform deep learning on typical tabular data?” in *Advances in Neural Information Processing Systems*, 2022, pp. 507–520. <https://doi.org/10.52202/068431-0037>
- [18] M. Hlosta, Z. Zdrahal, and J. Zendulka, “Ouroboros: Early identification of at-risk students without models based on legacy data,” in *ACM International Conference Proceeding Series*, 2017, pp. 6–15. <https://doi.org/10.1145/3027385.3027449>
- [19] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, 2017.
- [20] M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why should I trust you?” explaining the predictions of any classifier,” in *NAACL-HLT 2016–2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, 2016, pp. 97–101. <https://doi.org/10.18653/v1/N16-3020>
- [21] H. Khosravi et al., “Explainable artificial intelligence in education,” *Computers and Education: Artificial Intelligence*, vol. 3, p. 100074, 2022. <https://doi.org/10.1016/j.caeai.2022.100074>
- [22] Z. Liu, X. Zhou, and Y. Liu, “Student dropout prediction using ensemble learning with SHAP-based explainable AI analysis,” *Journal of Social Systems and Policy Analysis*, vol. 2, no. 3, pp. 111–132, 2025. <https://doi.org/10.62762/JSSPA.2025.321501>
- [23] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” *PLoS One*, vol. 10, no. 3, p. e0118432, 2015. <https://doi.org/10.1371/journal.pone.0118432>
- [24] J. Kuzilek, M. Hlosta, and Z. Zdrahal, “Data descriptor: Open university learning analytics dataset,” *Sci. Data*, vol. 4, p. 170171, 2017. <https://doi.org/10.1038/sdata.2017.171>
- [25] W. Feng, J. Tang, and T. X. Liu, “Understanding dropouts in MOOCs,” in *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 2019. <https://doi.org/10.1609/aaai.v33i01.3301517>
- [26] M. R. Khatun, M. A. Mim, M. M. Tasin, and M. M. Hossain, “A hybrid framework of statistical, machine learning, and explainable AI methods for school dropout prediction,” *PLoS One*, vol. 20, no. 9, p. e0331917, 2025. <https://doi.org/10.1371/journal.pone.0331917>
- [27] S. M. Lundberg et al., “From local explanations to global understanding with explainable AI for trees,” *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020. <https://doi.org/10.1038/s42256-019-0138-9>
- [28] M. Adnan et al., “Predicting at-risk students at different percentages of course length for early intervention using machine learning models,” *IEEE Access*, vol. 9, pp. 7519–7539, 2021. <https://doi.org/10.1109/ACCESS.2021.3049446>

- [29] M. Nagy and R. Molontay, “Interpretable dropout prediction: Towards XAI-based personalized intervention,” *Int. J. Artif. Intell. Educ.*, vol. 34, no. 2, pp. 274–300, 2024. <https://doi.org/10.1007/s40593-023-00331-8>
- [30] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, “Learning from class-imbalanced data: Review of methods and applications,” *Science Direct*, vol. 73, pp. 220–239, 2017. <https://doi.org/10.1016/j.eswa.2016.12.035>
- [31] C. Herodotou, G. Naydenova, A. Boroowa, A. Gilmour, and B. Rienties, “How can predictive learning analytics and motivational interventions increase student retention and enhance administrative support in distance education?” *Journal of Learning Analytics*, vol. 7, no. 2, 2020. <https://doi.org/10.18608/jla.2020.72.4>
- [32] I. Borrella and E. Ponce-Cueto, “Design principles and impact of a learning analytics dashboard: Evidence from a randomized MOOC experiment,” *Applied Sciences (Switzerland)*, vol. 15, no. 21, p. 11493, 2025. <https://doi.org/10.3390/app152111493>

8 AUTHORS

Yanyan Zhu received the master’s degree and is currently with Pingdingshan Polytechnic College, Pingdingshan 467001, Henan, China, where she is engaged in teaching and research in safety engineering. Her current research interests include educational data mining, learning analytics, and explainable artificial intelligence in education.

Niu Shi is with Pingdingshan Polytechnic College, Pingdingshan 467001, Henan, China. Current research interests include machine learning applications in education and student academic-risk prediction.

Yan Chen is with the Research Institute of Coal Mining and Utilization, Pingdingshan Tianan Coal Mining Co., Ltd., Pingdingshan 467000, Henan, China. Current research interests include data mining and applied machine learning.

Peiqiang Wang is with Pingdingshan Polytechnic College, Pingdingshan 467001, Henan, China. His current research interests include educational data mining, early warning systems, and explainable artificial intelligence (E-mail: wpq_313@126.com).