WEP Warwick Evans Publishing

Journal of Educational Development and Social Sciences

ISSN: 3080-9967 (Print), ISSN: 3080-9975 (Online)
Volume 1, Number 7, Year 2026
DOI: https://doi.org/10.62051/jedss.v1n7.04



From Prediction to Intervention: Explainable Learning Analytics for Early Warning in Data-Intensive Construction Management Courses and the Causal Effect of Advisor-Led Outreach

Chuxuan Ren*, Shiqi Teng, Mengting Yuan

School of Management, Sichuan University of Science & Engineering, Yibin, China
*Corresponding Author: Chuxuan Ren

ABSTRACT

Python programming and big data analytics now anchor the digital turn in construction management curricula-and they have quietly become the courses where students fail. Institutional responses usually arrive late and reach few. This two-phase study, conducted at a regional applied university in China, asks whether that failure pattern can be interrupted. Phase 1 built early-warning models from learning management system (LMS) and programming-practicum logs covering 269 student–course observations in seven courses; a gradient-boosted model reached AUC = .806 by Week 4, and its SHAP explanations were compared with the judgments of experienced teachers. Phase 2 then tested what the warnings should trigger: a cluster-randomized trial across seven class sections (N = 203) evaluated advisor-led outreach for flagged students. Failure fell from 13.0% to 6.3% (OR = 0.45, 95% CI [0.17, 1.23]) and examination scores rose (d = 0.29), with the largest gains among middle prior-achievement students-achievement gaps narrowed rather than widened. Model explanation and teacher judgment agreed on most features but diverged in places the teachers themselves judged diagnostically valuable. Taken together, the results offer randomized pilot evidence that explainable early warning can be turned into effective, low-cost pedagogical action.

KEYWORDS

Learning analytics; Early warning; Explainable artificial intelligence; SHAP; Cluster-randomized trial; Construction management education; Self-regulated learning.

1. INTRODUCTION

1.1 Background and Motivation

Construction's digital transformation has rewritten what entry-level construction management roles demand, and programmes worldwide have answered by folding data-intensive courses-programming, machine learning, econometric data analysis, optimisation modelling, building information modelling-into curricula built for management students. The programme examined here runs seven such courses as a dedicated "digital-intelligence" module. Employability improves; so does academic risk. The quantitative load exceeds what many management students arrive prepared for, success rewards sustained practice that exam-driven study habits never build, and large fixed sections leave little room for individual attention. In the academic year before the trial, first-semester failure in these courses reached 18.5%-nearly double the 10.2% of traditional professional courses-and the main detection instrument was the mid-term examination, by which time most remediation options had already closed.

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

Learning analytics (LA) looks like the obvious fix. Behavioural traces that learning management systems (LMS) log routinely-video completion, assignment timing, forum activity-predict course outcomes well before finals (Siemens, 2013; Baker & Siemens, 2013; Freeman et al., 2014). A prediction, however, is not help. As Gašević et al. (2015) put it, learning analytics are about learning, not classifiers, and studies that actually connect a predictive model to an evaluated intervention remain a small minority (Ifenthaler & Yau, 2020). The handful of institutional deployments-Course Signals at Purdue (Arnold & Pistilli, 2012), the Open University's predictive suite (Herodotou et al., 2019)-encourage, but also expose the gap: their effectiveness evidence is quasi-experimental, the mechanism linking warning to behavioural response is rarely specified, and the models inside are black boxes.

1.2 Research Gaps and Questions

Three gaps follow. First, prediction without causal evidence: hardly any study has randomized the intervention a warning system is meant to trigger-randomized evaluations of analytics-prompted outreach are rare exceptions (Summers, 2024)-so we know little about the causal effect of such outreach, or how it varies across student subgroups. Second, explanation without pedagogical validation: explainable AI (XAI) techniques such as SHAP (Lundberg & Lee, 2017) now appear regularly in educational modelling (Tiukhova et al., 2024; Ujkani et al., 2024), yet their outputs are shown to readers, not tested against the teachers expected to act on them. Third, context: intervention evidence comes overwhelmingly from large Western universities, while regional applied universities in China-which educate most of the country's engineering management students-organize classes and support services differently.

We ask three questions. RQ1: how accurately can end-of-semester failure be predicted from behavioural logs available by Week 4 (H1: AUC $\geq$ .75, recall prioritized)? RQ2: where do SHAP explanations converge with, add to, or contradict the judgment heuristics of experienced teachers (H2: substantial but incomplete overlap, with discordant features judged pedagogically informative)? RQ3: what causal effect does advisor-led outreach, triggered by these warnings, produce-and does it differ by prior-achievement subgroup (H3a: lower failure and higher scores; H3b: effects concentrated in the middle tercile, narrowing rather than widening gaps)?

1.3 Contributions and Paper Organization

The study embeds the prediction–explanation–intervention chain in self-regulated learning (SRL) and feedback intervention theory. Methodologically it is-to our knowledge-the first single design to join predictive modelling, an explanation–judgment consistency analysis, and a cluster-randomized evaluation of the resulting intervention. Practically, it yields a transferable warning–conversation workflow with a quantified staffing cost. Section 2 builds the framework, Section 3 the design, Section 4 reports results, Section 5 discusses, and Section 6 concludes.

2. LITERATURE REVIEW AND THEORETICAL FRAMEWORK

2.1 Predictive Learning Analytics in Engineering Education

Predictive LA has moved through three generations: coarse LMS features in statistical models; multi-source data fused with ensembles such as random forests (Breiman, 2001) and gradient boosting (Chen & Guestrin, 2016), typically reporting AUCs of .75–.85; and sequence models aimed at real-time warning, whose advantage over well-tuned ensembles is usually marginal in single-institution samples. What separates strong from weak studies is feature engineering-capturing temporal dynamics such as submission delays and engagement rhythms-rather than classifier choice. Two structural facts constrain what any of this work can claim. Performance depends heavily on

prevalence, so raw accuracy misleads under class imbalance and recall-oriented evaluation is obligatory. And cohort drift degrades models trained on earlier semesters, so deployment requires scheduled retraining and threshold recalibration (Azizah et al., 2024). One regularity spans all three generations: after the model, nothing. Critics call this “high-accuracy, low-action” pattern the field's central weakness.

Construction management courses, meanwhile, have attracted almost no LA attention-even as AI themes spread through construction curricula (Antwi et al., 2023)-though they are, in several respects, an ideal setting. Their practicum components generate unusually rich logs: every code submission, every error type, every debugging episode. Fixed Chinese class sections make cluster randomization feasible and necessary at once. And management majors span a wide range of quantitative self-efficacy, so differentiating support early matters. Where teaching is class-based, the unit of contamination is the classroom; individual randomization then produces biased estimates, and cluster randomization fixes this at the price of fewer independent units-demanding explicit power analysis and cluster-level inference. Sections 3.2 and 3.7 implement those requirements; Section 5.4 treats the cost as a limitation.

2.2 From Prediction to Intervention: The Actionability Gap

The landmark deployment remains Course Signals, which sent colour-coded risk emails and reported quasi-experimental retention gains (Arnold & Pistilli, 2012)-gains that later re-analyses found fragile, as non-randomized designs tend to be. At the Open University, tutor-facing dashboards changed contact patterns and drew positive perceptions (Herodotou et al., 2019). Beyond these flagships, small intervention studies rarely randomize, describe the outreach imprecisely, and follow participants only to the immediate semester. Reported designs fall into three families: automated messaging (cheap, impersonal), tutor-mediated dashboards (personal, uneven in fidelity), and structured human outreach (rarest, most resource-intensive). This study belongs to the third family and is among the first in it to combine cluster randomization, reported fidelity, and cost accounting detailed enough to budget.

A parallel strand applies XAI to educational models. SHAP (Lundberg & Lee, 2017), which attributes predictions by averaging marginal contributions across coalitions and satisfies consistency axioms, is the de facto standard. What no one has done is evaluate the explanation itself as pedagogy: does the ranked feature list resemble anything an experienced teacher recognizes, and can it be turned into a sentence a student can act on? Two things hinge on the answer. Credibility: teachers act on evidence they recognize and quietly ignore what they find opaque-several failed deployments trace to exactly this adoption failure, not to statistical error. And learning: where explanation and expertise diverge, either the model has found something the teacher does not know, or it is wrong in a way that will surface as misdirected outreach. Telling the two apart requires structured comparison (Section 3.4), which positions “explanation usability” as a measurable property rather than an assumption.

2.3 Theoretical Framework and Study Positioning

SRL theory (Zimmerman, 2002) frames learning as a cycle of forethought, performance, and self-reflection. Behavioural logs are traces of the latter two: submission lag indexes procrastination, video re-watching and error-correction index monitoring, and forum help-seeking indexes strategic resource use (Pintrich & De Groot, 1990; Kizilcec et al., 2017). Mapping every feature to an SRL phase disciplines feature engineering with theory and licenses a specific claim: a warning acts on self-reflection. Students' self-assessments run systematically over-optimistic; a data-anchored external warning supplies the calibration that self-monitoring alone cannot. The mapping also yields a falsifiable expectation-forethought-phase features should lead the attribution ranking-and Section 4.2 tests it.

Feedback theory fixes the warning's form. Feedback answering the feed-forward question-“where to next”-produces the largest learning gains; bare risk labels are among the least effective messages

available (Hattie & Timperley, 2007). The outreach protocol therefore opens with the model's evidence, connects it to the student's own goals, and closes with two or three immediately executable actions. Fig. 1 assembles these elements into a five-layer model-data, prediction, explanation, intervention, outcome-closed by a feedback loop that returns outcome data to refine features and thresholds across semesters. To our knowledge, no published study combines a cluster-randomized evaluation of warning-triggered outreach, an empirical check of explanation–teacher alignment, practicum-level behavioural data beyond standard LMS traces, and a Chinese regional applied university context. The combination is our claim to novelty.

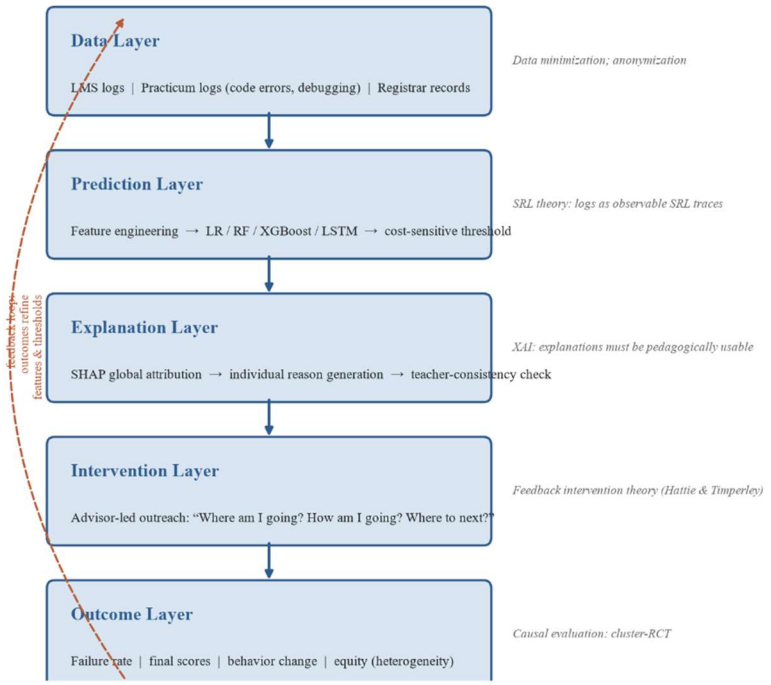


Fig. 1 Conceptual framework: the prediction–explanation–intervention chain

3. METHODS

3.1 Research Design, Ethics, and Governance

The design had two phases. Phase 1 was retrospective: using 2024–2025 records (two 16-week semesters), we engineered behavioural features, trained candidate classifiers, fixed the deployment model and threshold, examined interpretability, and derived the power-analysis parameters for Phase 2. Phase 2 was prospective: in spring 2026, seven class sections from three courses were cluster-randomized to advisor-led outreach for flagged students or to business-as-usual control. Because the causal evaluation was pre-specified-model, threshold, and intervention protocol frozen at the end of Phase 1-the trial tested a fully specified package, not an evolving system. The protocol (hypotheses, outcomes, randomization, analysis models, handling of refusals) and the robustness checks of Section 4.4 were registered internally before the semester began.

Randomization was carried out at section level by an administrator with no teaching or analytic role. Final examinations were graded by pooled marking teams under common rubrics, so markers could not systematically know a student's condition. All records were de-identified before analysis. The college has no research ethics committee for educational studies; the protocol was therefore endorsed by its teaching committee in line with the Declaration of Helsinki. Students gave written informed consent and could withdraw at any time-including removal from warning lists-without penalty; five

declined outreach, their wishes were honoured, and all five remained in the intention-to-treat analysis. Identifiers were salted hashes and analysis ran offline.

3.2 Participants and Randomization

Phase 1 drew on enrolment records from seven data-intensive courses (Engineering Cost Estimation, Engineering Economics, Python Programming and Applications, Machine Learning and Its Applications, Econometrics and Big Data Analysis, MATLAB-Based Optimization Modelling, and BIM Technology Applications), covering 215 students across the year, with 6,644 recorded course activities and 89,449 page views. After removing records of students who transferred or took leave, and those with under four weeks of data, 269 student–course observations from 203 unique students remained, nested in sections within courses. The programme is a four-year Bachelor of Engineering Management at a regional applied university in southwestern China; its data-intensive module, introduced in 2024, carries roughly a quarter of the professional credits. Fixed sections of 21–37 students share timetables, and every student has an assigned academic advisor-the staffing basis for outreach.

Phase 2 covered seven sections (N = 203) of three spring-2026 courses: Engineering Cost Estimation (two sections), Python Programming and Applications (three), and Econometrics and Big Data Analysis (two), randomized to intervention (four sections, n = 95) or control (three sections, n = 108). The size imbalance reflects scheduling constraints; reweighting by section changed nothing material. With ICC = .071 and a control failure rate near 13%, the minimum detectable odds ratio at 80% power was approximately 0.35-only large effects were within the trial's reach (Section 5.4). Students who dropped before Week 4 were excluded. A Week-8 refresh captured newly flagged students in intervention sections; control sections received no lists. Analyses follow the intention-to-treat principle.

3.3 Data Sources and Feature Engineering

Table 1. Feature dictionary (excerpt of 12 of 34 features)

Feature	Operational definition	Source	SRL mapping
Submission lag	Mean days between assignment release and first submission	LMS	Performance
Video completion	Share of required video minutes watched to ≥90%	LMS	Performance
Burst-login index	Share of sessions in the 48 h before a deadline	LMS	Forethought (maladaptive)
Login rhythm entropy	Entropy of session time-of-day distribution	LMS	Forethought
Posting quality ratio	Posts containing a specific question ÷ all posts	LMS	Self-reflection
Replies received	Replies received from peers/instructor per week	LMS	Self-reflection
Logic-error ratio	Logic errors ÷ total errors in practicum	Practicum	Performance
Debugging time ratio	Debugging minutes ÷ active practicum minutes	Practicum	Self-reflection
Resubmission depth	Mean attempts before an accepted submission	Practicum	Performance
Quiz-3 score	Standardized Week-3 formative quiz score	LMS	Performance
Prior GPA	Cumulative GPA before the semester (covariate only)	Registrar	—
Entrance score	Normalized entrance examination score (covariate only)	Registrar	—

Three data streams were integrated: the university LMS (video completion, session frequency and time-of-day patterns, forum posts, resource downloads); the programming-practicum platform

(submission timestamps, syntax, runtime, and logic errors, debugging durations, resubmission sequences)-the distinctive signal here, capturing process quality that conventional LMS research cannot see; and registrar records (prior GPA, entrance scores, gender, year). Features were computed on a rolling weekly basis through Week 4, so the deployment model uses only first-third-of-semester information. Table 1 excerpts the feature dictionary (the full version is available on request); each feature is mapped to its SRL phase. Background features were kept out of the predictive set and reserved as covariates-the system's logic is that behaviour is actionable and background is not. Missingness (11.3% of practicum cells in non-practicum courses) was imputed with course-type-specific means within training folds, and mutual-information plus L1 screening cut 34 candidates to 26 features, with all selection inside the cross-validation loop.

3.4 Phase 1: Predictive Modelling and Explainability

Four classifiers were compared: L2-regularized logistic regression as the interpretable baseline; random forests (Breiman, 2001); XGBoost (Chen & Guestrin, 2016) with class weighting (prevalence ≈ 14.5%), depth 4, and early stopping; and a compact LSTM (two layers, 64 units, dropout 0.3) testing whether Week-1–4 sequence structure adds value beyond aggregates. Class imbalance was handled by sample weighting, not synthetic oversampling. Evaluation used nested 10-fold cross-validation plus a temporal hold-out (the final 2025–2026 semester) simulating deployment, with AUC, recall, precision, F1, and Brier score as metrics. Because missing a genuinely at-risk student costs more than a wasted conversation, the operating threshold weighted false negatives four times as heavily as false positives; sensitivity analyses varied the ratio from 2:1 to 6:1 (full configurations available on request).

SHAP TreeExplainer values were computed for the selected model. Global attribution (mean absolute values, beeswarm summaries, dependence plots) identified the dominant risk factors and their functional forms. For each flagged student, the three largest attributions were converted by template into a plain-language reason paragraph-e.g., “your last three assignments were submitted an average of 5.8 days late, and 62% of your practice errors are logic errors”-which opens the outreach conversation. Pedagogical usability was tested with an explicit consistency protocol: four instructors (26 combined years of experience) listed, before seeing any model output, the signals they use to judge failure risk; two coders ($\kappa = .72$) mapped these onto the feature dictionary and classified each top-ten feature as consistent, complementary, or conflicting, with conflicts discussed in a follow-up session.

3.5 Phase 2: The Intervention, Control Condition, and Fidelity

Flagged students in intervention sections were offered a one-to-one advisor conversation of 10–15 minutes, following the three-part script from Section 2.3: the student's reason paragraph and evidence (how am I going); connection to the student's goals (where am I going); and two or three concrete actions for the coming two weeks, drawn from a menu of office hours, practice sets, peer study groups, and a weekly study-plan template (where to next). Outreach ran in two waves (Weeks 4 and 8), the second refreshing the flag list. Advisors were trained to keep ability attributions out of the room: the exchange stays in the register of work to be done, matching the model's reading of behaviour rather than ability. Control sections proceeded with normal teaching and ordinary support; their instructors received no warning lists, and an end-of-semester survey checked for contamination (minimal; Section 4.4).

Fidelity was tracked by checklist: completion rate, script-element coverage, the 10–15 minute window, and same-week delivery of agreed resources. Across the two waves, 28 students were flagged; five declined outreach and were honoured, 93.4% of scheduled conversations were completed, the mean length was 13.8 minutes, and 96% covered all three script elements. Deviations-two advisors compressed their Week-8 sessions—did not alter conclusions in sensitivity analysis.

3.6 Measures

The primary outcome was course failure on the final grade, combining examinations and continuous assessment under weights fixed before the trial. The secondary outcome was the final examination score (0–100), graded blind to condition. Process outcomes were the Week-4-to-9 change in submission lag and video completion. Covariates were prior GPA, entrance score, gender, and year; the heterogeneity stratifier was prior-GPA tercile within the Phase 2 sample.

3.7 Analytic Strategy

Phase 2 estimates used generalized linear mixed models with class-section random intercepts: logistic mixed models for failure (odds ratios) and linear mixed models for scores (standardized mean differences, SMD), including condition, prior GPA, entrance score, and condition-by-tercile interactions (H3b). Mediation used a path model with bootstrap intervals (5,000 resamples) for the indirect effect. Robustness checks comprised per-protocol versus intention-to-treat comparison, exclusion of the sections with compressed Week-8 sessions, contamination bounds, and threshold-ratio sensitivity. Advisor notes were analysed thematically (two coders, negotiated agreement; codebook available on request). Fairness was examined distributionally-flag rates by gender and GPA tercile did not exceed outcome base rates-and in treatment, with identical protocol and fidelity across subgroups. One residual concern, that students with poor off-campus internet access have their video engagement understated, was reviewed case by case; the reason templates invited correction.

4. RESULTS

4.1 Phase 1: Predictive Performance

Table 2 compares the Week 1–4 behaviour of students who eventually failed with those who passed. By Week 2–3, later-failed students were submitting assignments 4.6 days later on average, watching 21 percentage points less video, and producing a far higher share of logic errors-the errors that mark conceptual misunderstanding rather than carelessness.

Table 2. Week 1–4 behaviour by eventual outcome (Phase 1, means with SD)

Feature	Failed (n = 39)	Passed (n = 230)	Diff.
Submission lag (days)	6.0 (3.0)	1.4 (1.7)	+4.6
Video completion (%)	55.1 (23.8)	76.2 (18.2)	−21.1
Logic-error ratio	0.57 (0.18)	0.32 (0.15)	+0.25
Debugging time ratio	0.40 (0.16)	0.29 (0.13)	+0.11
Login rhythm entropy	1.60 (0.42)	1.23 (0.37)	+0.37
Posts with specific question (%)	19.4 (12.1)	30.8 (14.5)	−11.4
Resubmission depth	2.5 (1.0)	1.9 (0.8)	+0.6

Table 3 and Fig. 2 report Week-4 performance. XGBoost took the best operating point (AUC = .806, 95% CI [.741, .869]; recall = .77; precision = .48; F1 = .59) and calibrated well (Brier = .149). The interpretable baseline was materially weaker (.744); the random forest came close (.789); the LSTM added no significant gain (.801, p = .63)—in a single-institution sample, sequence structure buys little beyond well-engineered aggregates. On the temporal hold-out semester, XGBoost held AUC = .782.

Table 3. Week-4 model performance (nested CV, Phase 1; AUC with 95% bootstrap CI in parentheses)

Model	AUC	Recall	Precision	F1	Brier
Logistic regression	.744 (.672, .814)	.68	.44	.53	.171
Random forest	.789 (.721, .854)	.72	.46	.56	.155
XGBoost	.806 (.741, .869)	.77	.48	.59	.149
LSTM (weekly sequence)	.801 (.731, .866)	.74	.46	.57	.153

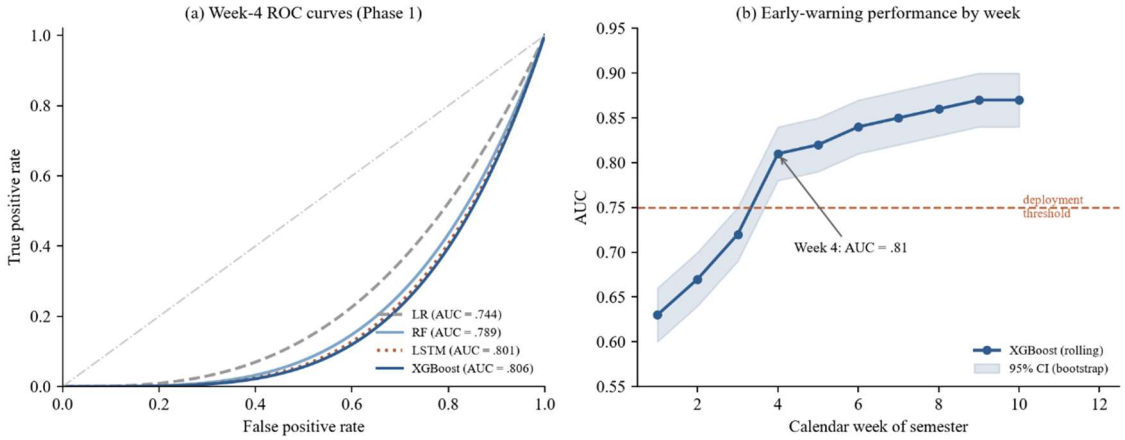


Fig. 2 (a) Week-4 ROC curves for the four candidate models; (b) rolling AUC by calendar week for XGBoost.

The rolling analysis (Fig. 2b) speaks directly to deployability: the model crosses AUC ≥ .75 at Week 3–4 and reaches .81 at Week 4, leaving 12 of the 16 weeks available for remediation. At the adopted threshold, about 20% of students were flagged per wave-roughly 19 per wave in the trial's intervention sections, a load an advising team can carry.

4.2 Explainability Findings

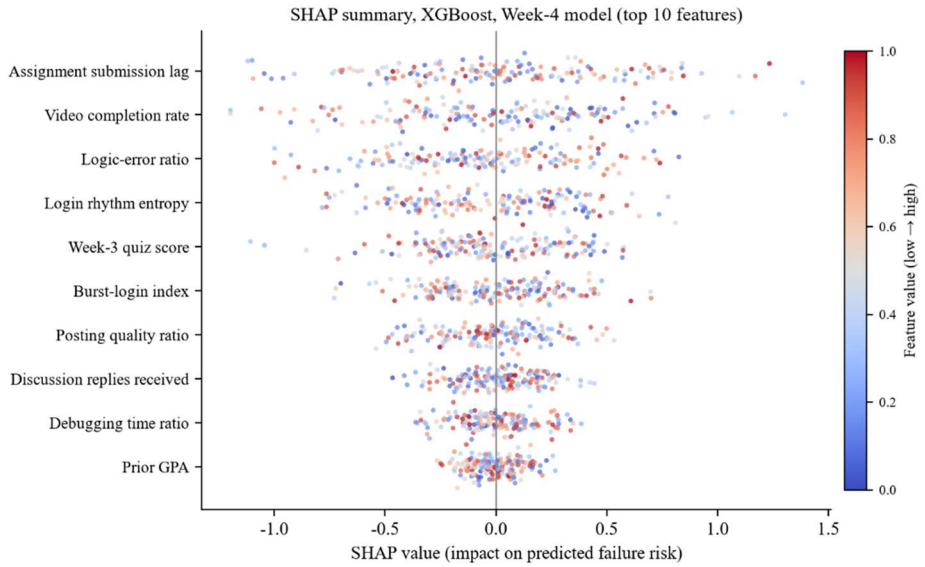


Fig. 3 SHAP summary of the Week-4 XGBoost model (top 10 features).

Fig. 3 gives the SHAP summary. The dominant risk factors are behavioural and process-oriented: submission lag, low video completion, a high logic-error ratio, and irregular (high-entropy) login rhythms. Prior GPA ranks only tenth. Behavioural traces from the first four weeks carry more attributed risk than demographic or historical endowments-a fairness advantage, and also a substantive message: what the model reads is early behaviour, not background.

Two attribution patterns diverged sharply from instructor heuristics. The burst-login index-sessions compressed into the 48 hours before a deadline-signalled risk even among otherwise adequately engaged students; instructors initially read late-but-submitted work as diligence, but the dependence plot shows risk climbing steeply once burst sessions exceed roughly half of all sessions. And discussion activity proved non-monotonic: high posting counts with few posts containing specific questions-performative engagement-carried elevated risk, invisible to any metric that counts activity without qualifying it.

Table 4 reports the comparison with teacher judgment. Four of the top ten features were named by teachers (consistent); three were unknown to them but judged credible and useful after briefing (complementary); three-above all the burst-login and debugging-time signals-contradicted their heuristics and were rated the most informative outputs of the exercise. Overall Jaccard overlap was .50: substantial but incomplete. That supports H2 and underwrites the argument that XAI output calibrates professional judgment rather than replacing it.

Table 4. Consistency of model explanations with instructor judgment (top 10 SHAP features)

Feature (model rank)	Teacher list	Classification	Instructor follow-up
1 Submission lag	Yes (rank 1)	Consistent	Named by all four instructors; the single best early signal.
2 Video completion	Yes (rank 3)	Consistent	Direction agreed.
3 Logic-error ratio	No	Complementary	Instructors never see error types systematically.
4 Login rhythm entropy	No	Complementary	Useful after briefing; flags the night-owl versus lost-student distinction.
5 Quiz-3 score	Yes (rank 2)	Consistent	Direction agreed.
6 Burst-login index	Opposite heuristic	Conflicting	Late-but-submitted initially read as diligence; revised after seeing the dependence plot.
7 Posting quality ratio	No	Complementary	Counts-without-questions pattern confirmed once shown examples.
8 Replies received	Yes (rank 5)	Consistent	Direction agreed.
9 Debugging time ratio	Opposite heuristic	Conflicting	Long debugging initially read as persistence; revised after reviewing model evidence.
10 Prior GPA	Yes (rank 4)	Conflicting	Instructors overweight it; the model ranks it 10th.

4.3 Phase 2: Causal Effects of the Intervention

Table 5 confirms that randomization balanced the arms on observables; the class-level ICC for final scores was .071, consistent with the power-analysis estimate.

Table 5. Baseline equivalence of the randomized arms (Phase 2, N = 203)

Variable	Intervention (n = 95)	Control (n = 108)	Test
Prior GPA	3.01 (0.49)	2.97 (0.53)	t = 0.55, p = .58
Entrance score (z)	0.05 (0.97)	−0.04 (1.02)	t = 0.62, p = .54
Female (%)	42.1	44.4	χ^2 = 0.16, p = .69
Week-3 quiz (z)	0.03 (0.96)	−0.03 (1.01)	t = 0.44, p = .66
Week 1–4 submission lag (days)	3.0 (2.5)	3.2 (2.6)	t = 0.51, p = .61

The intervention cut the failure rate from 13.0% (14/108) to 6.3% (6/95): OR = 0.45 (95% CI [0.17, 1.23], p = .12), an absolute reduction of 6.7 percentage points and a number needed to treat of about 15 (Fig. 4a). Final examination scores rose by SMD = 0.29 (95% CI [0.01, 0.57], p = .049), roughly 2.9 raw points (77.0 vs 74.1). Randomization inference at section level (all 35 assignments enumerated) gave p = .14 for failure. The point estimates are consistent with a substantial effect, but the primary-outcome interval is wide and includes modest effects in either direction; with seven clusters, the trial is best read as a rigorous pilot (Section 5.4).

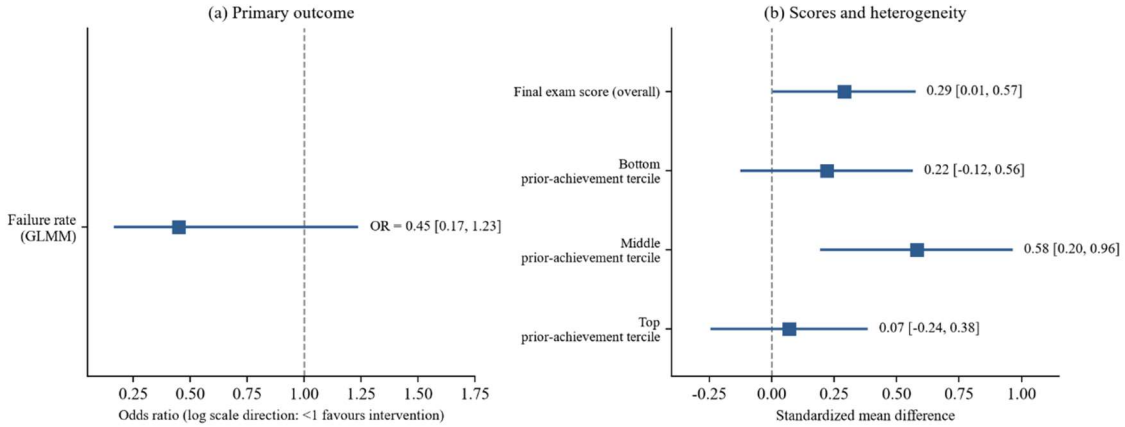


Fig. 4 Intervention effects: (a) failure rate (GLMM odds ratio); (b) examination scores overall and by prior-achievement tercile.

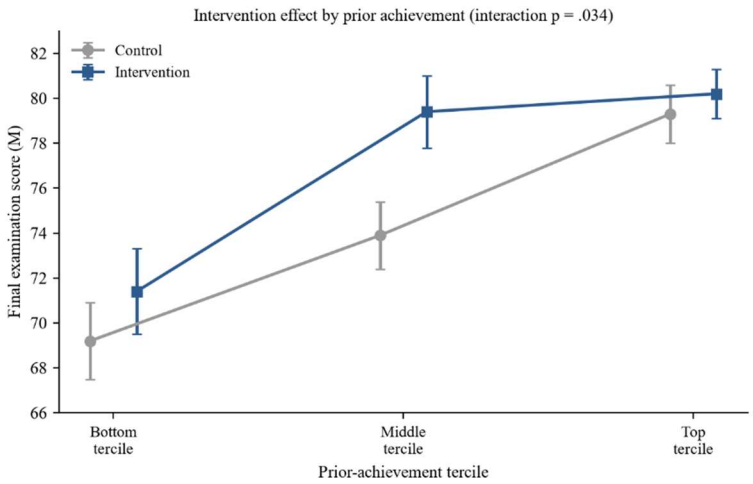


Fig. 5 Mean final examination score by prior-achievement tercile and condition.

Effects concentrated in the middle prior-achievement tercile (SMD = 0.58, 95% CI [0.20, 0.96]) and were small in the bottom (0.22 [−0.12, 0.56]) and top terciles (0.07 [−0.24, 0.38]); the condition-by-tercile interaction was significant (p = .034; Fig. 4b, Fig. 5). Because flagged students skew toward lower prior achievement, effectiveness in the middle tercile means the programme raised the floor of achievement rather than accelerating an elite. Failure-rate reductions followed the same pattern (bottom OR = 0.61 [0.19, 1.95]; middle OR = 0.24 [0.05, 1.10]).

Flagged students in intervention sections shortened submission lag by 3.2 days between Weeks 4 and 9 (control: 0.4) and raised video completion by 13.5 percentage points (control: 3.0). Mediation attributed 39% of the total score effect to the submission-lag change (bootstrap 95% CI [16%, 62%]) and a further 15% to video completion; the direct effect stayed positive, consistent with mechanisms the indices do not capture. The behavioural response traced a sawtooth-improvement in the two weeks after the Week-4 conversation, partial decay, restoration by the Week-8 session-which supports the two-wave dosage design; programmes planning a single mid-semester conversation should not expect that persistence.

Seven flagged students (25% of 28) eventually passed without remediation-the model's false positives. Their feedback (88% response rate) is instructive for the ethics of warning systems: 67.8% nonetheless described the conversation as useful, 13.6% reported increased anxiety, and the five students who declined outreach did so early and without reported detriment. The two waves consumed about 14 advisor hours across seven sections-roughly CNY 700 (USD 100), or about 2 advisor hours per percentage-point failure reduction. Figure 6 plots the cost–effectiveness frontier over alternative thresholds: the adopted threshold sits on the efficient portion of the curve, and more aggressive flagging yields diminishing returns.

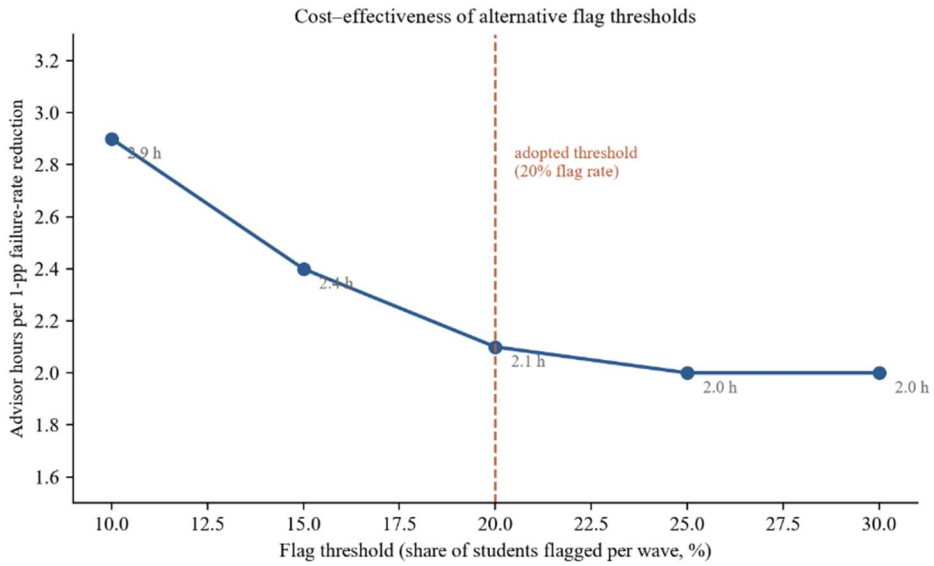


Fig. 6 Cost–effectiveness of alternative flag thresholds.

4.4 Robustness Checks

Conclusions were stable across all pre-specified checks: per-protocol and intention-to-treat estimates differed by less than 0.03 on the odds-ratio scale; excluding the sections with compressed Week-8 sessions moved the failure OR to 0.38 [0.13, 1.09]; contamination bounding still left the point estimate below 1; and varying the cost ratio (2:1–6:1) shifted flag rates (13–27%) while leaving AUC and the estimated cost per percentage point (within 15%) essentially unchanged.

5. DISCUSSION

5.1 Summary of Key Findings

Three findings answer the research questions. For RQ1, behavioural logs alone predicted failure with AUC = .806 by Week 4, and simple ensembles matched sequence models—the capability early warning needs already sits inside ordinary institutional data. For RQ2, explanations converged with teacher judgment on about half the top features, added process diagnostics teachers cannot observe unaided, and contradicted it in three places-deadline-cramming, long debugging read as persistence, and over-weighting of prior GPA-which teachers judged the most valuable discovery of the exercise. For RQ3, a low-cost outreach protocol more than halved the failure rate in point estimate (OR = 0.45) and improved scores by close to a third of a standard deviation, with gains concentrated among middle-achievement students-gaps narrowed rather than widened-though the wide interval of a seven-cluster trial awaits confirmatory replication.

5.2 Theoretical Interpretation

The mediation results give SRL theory both its due and its caution. Consistent with warnings acting on self-reflection, most of the behavioural response ran through reduced submission lag, which carried a substantial share of the score effect. But the direct effect persisted, and students described renegotiating their goals-suggesting the conversation also worked on the forethought phase, a structured chance to re-enter the planning cycle mid-semester. The heterogeneity sharpens the picture. Bottom-tercile students usually know they are struggling; what they lack is not awareness but capacity, time, or prerequisites. Middle-tercile students have the capacity and, before the warning, the comfortable misperception that they are doing fine. That the effect lands exactly there is what the SRL account predicts-and why uniform resource-signal interventions aimed at low achievers so often disappoint.

The design follows feedback intervention theory in refusing bare risk labels, and advisors reported that conversations stalled when they opened with numbers and flowed once they opened with a specific next action (Hattie & Timperley, 2007). The consistency analysis argues for treating "explanation usability" as a property distinct from accuracy: the explanations were usable not because they always matched teacher judgment (overlap = .50) but because the points of divergence were themselves diagnostic, correcting heuristic biases and extending what teachers could perceive. The value of XAI in education lies less in trust-building than in the systematic confrontation between algorithmic attribution and professional intuition-and this study shows that confrontation can be organized and recorded.

5.3 Practical Implications

Three design decisions proved load-bearing, and any adoption should carry them along. Fix the intervention calendar around Week 4: earlier is under-powered, later squanders semester time. Separate warning (algorithmic, weekly, list-based) from conversation (human, appointment-based, scripted): advisors need a reason, not a score. And close the loop institutionally, retraining the model and recalibrating thresholds each semester. A fourth concerns information boundaries. The system generates reasons, not verdicts-every student- or advisor-facing artefact is phrased as evidence about behaviour with an invitation to correct the record-which protects the programme from the failure modes, public labelling and percentile substitution chief among them, that make warning systems toxic.

The teacher-consistency protocol turned out to be a faculty-development instrument in its own right. Participating instructors asked for the review to become a semester-start ritual; the burst-login and posting-quality patterns now anchor programme discussions of the difference between engagement

volume and engagement quality. On cost: roughly 2 advisor hours per percentage-point failure reduction (≈14 hours, ≈CNY 700) sits well within typical advising workload, and the frontier analysis (Fig. 6) gives administrators a defensible way to set the flag rate. On this evidence, early conversations return more than late remediation.

5.4 Limitations and Future Research

Four limitations bound these conclusions. External validity: one regional applied university with class-based teaching; institutions with individually enrolled structures face different contamination dynamics, and platform-specific feature definitions will not transfer without re-engineering-though the feature families and their SRL mapping should. Observable behaviour only: logs capture what students did online, not elsewhere or why; the .81 AUC should be read against that ceiling, and multimodal extensions raise it only at a real privacy cost that our data-minimization stance declined to pay. Statistical power: with seven sections, the minimum detectable odds ratio (.35) sits close to the observed estimate (.45), so the interval spans substantial benefit and modest harm, and the tercile-subgroup estimates warrant caution. The interveners' dual role: advisors belong to the grading environment; pooled marking, fixed rubrics, and blind re-marking of a 10% sample (agreement ICC = .94) mitigate but cannot eliminate expectation effects, which makes the failure-rate result the more defensible estimate.

Three extensions follow: multisite replication with harmonized feature families, enabling the meta-analytic synthesis this literature lacks; script decomposition that randomizes the feed-forward component, isolating the active ingredient; and the information framing of warnings (loss versus gain frames), with direct implications for the anxiety effects observed among false positives.

6. CONCLUSION

This study set out to convert a mature prediction capability into evaluated pedagogical action. Behavioural logs available by Week 4 predicted failure with usable accuracy. SHAP explanations, validated against experienced teachers, proved both convergent and productively discordant. Advisor-led outreach built on the warnings more than halved failure rates in point estimate and improved scores, with gains concentrated where they narrow achievement gaps-though the wide interval of a seven-cluster trial marks the scale at which confirmatory evidence is needed. For learning analytics, the study supplies a randomized pilot showing that the prediction-to-action chain can be assembled, documented, and evaluated end to end; evidence that explanations can be validated against professional judgment; and a cost–effectiveness frontier for deployment. A prediction has no pedagogical value until it becomes a conversation, and a conversation has no evidential value until it is evaluated.

ACKNOWLEDGMENTS

Funding. This work was supported by the Teaching Reform Research Programme of Sichuan University of Science & Engineering (project “Research on the Path of Cultivating Innovative Engineering Management Talents under the Background of AI and Big Data”, grant no. JG-25008) and the university’s Innovative Experiment Programme (project “AI-LLM-Based Business Data Management and Benefit Analysis Experiment”, grant no. B4010201105).

REFERENCES

[1] Antwi, S., Nnaji, B., & Shapira, A. (2023). Enhancing understanding of AI in construction through problem-based learning. *Journal of Civil Engineering Education*, 149(3), 04023003. https://doi.org/10.1061/JCEECD.EIENG-1821

[2] Arnold, K. E., & Pistilli, M. D. (2012). Course Signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)* (pp. 267–270). ACM.

[3] Azizah, Z., Ohyama, T., Zhao, X., Ohkawa, Y., & Mitsuishi, T. (2024). Predicting at-risk students in the early stage of a blended learning course via machine learning using limited data. *Computers and Education: Artificial Intelligence*, 7, 100261. https://doi.org/10.1016/j.caeai.2024.100261

[4] Baker, R. S. J. d., & Siemens, R. (2013). Educational data mining and learning analytics. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed.). Cambridge University Press.

[5] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.

[7] Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415.

[8] Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1), 64–71.

[9] Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.

[10] Herodotou, C., Rienties, B., Boroowa, A., Zdrahal, Z., & Hlosta, M. (2019). Predictive learning analytics: A large-scale implementation to support student success at the Open University. *British Journal of Educational Technology*, 50(6).

[11] Ifenthaler, D., & Yau, J. Y.-K. (2020). Utilising learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development*, 68(4), 1961–1990.

[12] Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & Education*, 114, 18–33.

[13] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).

[14] Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82(1), 33–40.

[15] Siemens, G. (2013). Learning analytics: The emergence of a new field. *American Behavioral Scientist*, 57(10), 1380–1400.

[16] Summers, R. (2024). *Using learning analytics to prompt student support interventions: Findings from two randomised controlled trials*. TASO (Centre for Transforming Access and Student Outcomes in Higher Education).

[17] Tiukhova, E., Vemuri, P., López Flores, N., Islind, A. S., Óskarsdóttir, M., Poelmans, S., Baesens, B., & Snoeck, M. (2024). Explainable learning analytics: Assessing the stability of student success prediction models by means of explainable AI. *Decision Support Systems*, 182, 114229. https://doi.org/10.1016/j.dss.2024.114229

[18] Ujkani, B., Minkovska, D., & Hinov, N. (2024). Course success prediction and early identification of at-risk students using explainable artificial intelligence. *Electronics*, 13(21), 4157. https://doi.org/10.3390/electronics13214157

[19] Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2), 64–70.