International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(21s) 2026 pp. 11-18
https://doi.org/10.70917/ijcisim-2026-5269
Article

Quantifying the Horizon-Reliability-Reachability Trade-off in Longitudinal Learning Analytics

Yanying An[1], Ran Wang[2]*

[1] College of Horticulture Science, Qingdao Agricultural University, Qingdao, China
[2] College of Horticulture Science, Qingdao Agricultural University, Qingdao, China
[1]st author email: monkeypinky92@gmail.com; 2nd author email: ranwangqau@gmail.com
[1] https://orcid.org/0000-0001-9796-388X; [2] https://orcid.org/0000-0003-1782-3356

Abstract: Early-warning systems in Learning Analytics are usually evaluated on how accurately a model ranks students by risk at a single chosen point in the course. This paper argues that question is incomplete, and quantifies what it omits using the Open University Learning Analytics Dataset (OULAD, 32,314 enrollments, 7 modules, 22 presentations). At five prediction horizons (weeks 1-4 and 6), we define a landmark-style risk set R_t containing only students not yet withdrawn, so a model is never credited with predicting a withdrawal that already happened. Reliability rises with horizon (best-model AUC 0.741 at week 1 to 0.808 at week 6, all consecutive gains significant by paired DeLong test, $p<1e-5$), but reachability -- the share of the population still in R_t -- falls over the same period, from 89.5% to 82.9%, and the students who exit are disproportionately those who would have been flagged (future-risk prevalence within R_t falls from 47.2% to 43.0%). XGBoost's advantage over logistic regression is present from the earliest horizon (+0.008 AUC, $p<0.001$) but amplifies roughly 2.4x by week 6 (+0.019 AUC, $p<1e-10$), surviving leakage-proof hyperparameter tuning. Under strict walk-forward temporal holdout, the horizon-reliability pattern generalizes in 6 of 7 modules, while conventional random-split evaluation overstates future-cohort generalization by a mean of 2.2 AUC points. These results reframe the central question for early-warning deployment: not only how reliable a prediction is, but how many students remain reachable by the time it is reliable enough to act on.
Keywords: Learning Analytics; early-warning systems; dropout prediction; landmark analysis; calibration; temporal generalization; OULAD

1. Introduction

Early-warning systems (EWS) are now a standard instrument in Learning Analytics: a model consumes behavioural and demographic data and returns a risk score an instructor or advisor can act on. Two design choices sit underneath almost every such system, usually made implicitly rather than evaluated on their own terms: when during a course to predict, and who the prediction is still useful for once it is made.

The timing question has recently received a systematic treatment. A 2026 survival-oriented benchmark on this same dataset compares prediction horizons under a unified protocol that integrates predictive performance, ablation, explainability, and calibration layers, and finds that the dominant predictive signal is temporal and behavioural rather than static or demographic [1]. That paper is explicit about the boundary of its generalization claim: its enrollment-level train/test split shares all 7 modules and all 4 presentations, so its findings hold within a shared curricular context across enrollments, not to an unseen module or presentation [1]. It does not report what happens to the population itself as the prediction horizon lengthens.

A second, independently motivated 2026 study addresses a closely related concern from the opposite direction: probability reliability under cross-module dataset shift [2]. That work already rebuilds an eligibility-aware risk set at every cutoff -- days 14, 28, 42, and 56 -- explicitly excluding students who have already unregistered by the cutoff

from being retroactively treated as active candidates, and restricting every predictor to information available before the cutoff [2]. The risk-set construction there serves as a leakage-control mechanism internal to a different research question; its shrinkage and compositional change are not themselves the object of study.

This paper takes the same landmark-style eligibility logic and asks a different question of it: not "how do we predict correctly at each cutoff," but how much of the addressable population, and which part of it, disappears as the prediction cutoff moves later. We call the resulting empirical quantity reachability, Reachability(t) = |R_t| / |R_0|, and report its trajectory and compositional shift as a finding in its own right, alongside the reliability trajectory. The result is a three-way trade-off rather than the usual two-way one: Earliness, Reliability, and Reachability.

This paper responds to three specific gaps. (1) Neither prior study reports a reachability or addressable-population metric as a finding. We report Reachability(t) and its compositional shift explicitly, tied to the operational consequence that a student who has already withdrawn is no longer someone an intervention can reach. (2) Both prior studies show nonlinear models beat simpler baselines, but neither asks how the size of that advantage changes as evidence accumulates. We show the margin is present from the earliest horizon but amplifies roughly 2.4x from week 1 to week 6. (3) [1] explicitly do not attempt unseen-presentation validation; [2] validate a related but distinct cross-module transfer setting. We instead walk forward through every available within-module presentation transition and report the resulting optimism of conventional random-split evaluation relative to this design.

The paper is organized around five research questions. RQ1: how does predictive reliability change as behavioural evidence accumulates across the prediction horizon? RQ2: how does reachability change over the same horizon, and is it compositionally neutral? RQ3: does the advantage of a nonlinear model (XGBoost) over a linear baseline (logistic regression) change in size as the horizon lengthens, and does this survive leakage-proof hyperparameter tuning? RQ4: does the horizon-reliability pattern generalize under strict within-module, walk-forward temporal holdout, and how much does conventional random-split evaluation overstate that generalization? RQ5: how do the behavioural determinants of predicted risk shift across the horizon, and does the reachability trajectory differ systematically by module?

2. Literature Review

2.1 Dropout and Early-Warning Prediction on OULAD

OULAD [3] has become one of the most reused empirical substrates for dropout and at-risk prediction in Learning Analytics and Educational Data Mining. A substantial applied literature compares classifier families on this dataset -- logistic regression, support vector machines, random forests, gradient-boosted trees, and increasingly deep sequential architectures -- generally reporting that tree ensembles and their neural counterparts outperform linear baselines, and that meaningful signal accumulates by the first several weeks of a presentation. This paper does not re-litigate whether nonlinear models outperform linear ones; both [1] and [2] already establish that they do. It asks instead how that margin behaves as a function of the prediction horizon.

2.2 Temporal Risk Prediction and the Timing Question

Reviews of predictive Learning Analytics note that retention and dropout are among the most frequently modelled outcomes, and that comparative studies vary in target definition, temporal scope, feature availability, and evaluation protocol, making cross-study comparison difficult [4] (as cited in [1]). [5] raised a related, earlier methodological concern specifically about how predictive Learning Analytics models are evaluated. Da Silva et al. (2026) respond to this within a single dataset and a single train/test split by treating benchmark architecture itself as the contribution. Their central finding -- that the dominant predictive signal is temporal-behavioural rather than static -- is not contested here; it complements the compositional-shift finding in Section IV.B below.

2.3 Landmark-Style Eligibility and Immortal-Time Bias

The specific mechanism this paper repurposes -- rebuilding the eligible population at every prediction cutoff rather than fixing it once -- echoes landmark analysis in survival and clinical epidemiology, developed to avoid immortal-time bias, the error of implicitly crediting a unit with surviving to a later checkpoint when evaluating an earlier one ([6]). [2] already implement this correctly for OULAD: a registration enters their risk set at cutoff t only if its registration date is observed and earlier than t, and only if its unregistration date is missing or at or after t. The distinction this paper draws is not about who invented the mechanism; it is about what the mechanism is used for. In [2], the risk set exists to make each cutoff's prediction valid, reported as an incidental fact about data volume. Here, the trajectory of |R_t| and the shift in future-at-risk prevalence within it are the reported finding.

2.4 Calibration and Probability Reliability

That a model ranks students correctly by risk is not sufficient for operational use; predicted probabilities must remain numerically coherent with observed outcomes ([7]). [2] address this under cross-module dataset shift: a model calibrated on source modules loses probability reliability when transferred zero-shot to a held-out module, and modest historical target-domain data substantially restores it. Da Silva et al. (2026) find a comparable pattern: their weekly, discrete-time family shows systematically larger calibration gaps than their best static early-window models. This paper's calibration layer (Section III.D) is narrower in scope than either -- post-hoc temperature scaling on a held-out calibration slice of the same evaluation population.

2.5 Generalization Across Cohorts and Presentations

[5] argue that closing the methodological gap in predictive Learning Analytics requires evaluation beyond a single held-out split from the same population a model was trained on. [2] address a cross-module version of this with a complete leave-one-module-out cycle. Da Silva et al. (2026) explicitly do not attempt this: their conclusion is scoped to generalization across enrollments under shared curricular context. This paper's RQ4 targets a third, narrower transfer axis that neither directly measures: within-module generalization across successive presentations over time.

3. Methodology

3.1 Dataset, Entry Population, and Landmark-Style Risk Sets

The empirical basis is OULAD [3], comprising 32,593 enrollments across 7 modules and 22 module-presentation offerings, integrating demographic records, registration timing, assessment submissions and scores, and virtual learning environment (VLE) clickstream interactions.

The base population is restricted to enrollments with an observed registration date on or before the presentation start, giving N = 32,314 (99.1% of raw enrollments). At each prediction horizon t in {1, 2, 3, 4, 6} weeks (cutoff day 7t), the risk set is R_t = { i in base population : date_unregistration_i is missing, or date_unregistration_i > 7t }. This is the landmark-style logic used by [2]; here it is applied at weekly cutoffs and, distinctively, its output is treated as a primary result rather than solely a validity precondition. Because R_t is defined purely by withdrawal timing on a fixed base population, the risk sets nest: R_week6 is a subset of R_week4, which is a subset of R_week3, R_week2, and R_week1. Reachability at horizon t is defined as Reachability(t) = |R_t| / |R_0|, where R_0 is the base population.

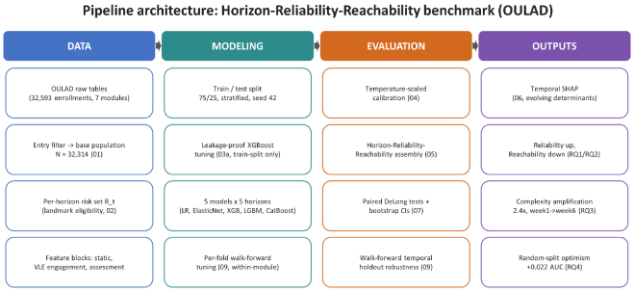

Fig. 1 Pipeline architecture across four phases: data preparation, modeling, evaluation, and outputs.

3.2 Outcome Definition

The prediction target is binary at-risk status: y_i = 1 if enrollment i's final result is Fail or Withdrawn, y_i = 0 if Pass or Distinction. Within R_t, y_i is better read as future at-risk status, since any enrollment that has already withdrawn by t is, by construction of R_t, no longer in the evaluated population.

3.3 Feature Engineering

For each horizon, cumulative behavioural and assessment features are computed using only data with date <= 7t: total clicks, active days, distinct materials accessed, clicks by activity-type bucket, a 7-day recency click share, days since first/last click, and -- restricted to TMA/CMA assessments with a due date on or before the cutoff -- counts of assessments due, submitted, and missed, mean score, and mean submission lateness. Static demographic features are included at every horizon. Missing behavioural or assessment values are encoded with missing-indicator flags rather than imputed as zero.

3.4 Models, Calibration, and Leakage-Proof Hyperparameter Tuning

Five classifiers are compared at every horizon: L2-regularized logistic regression, ElasticNet-penalized logistic regression, XGBoost ([8]), LightGBM, and CatBoost. A single student-level train/test split (75/25, stratified by module x outcome, seed 42) is fixed once on the base population and reused at every horizon. XGBoost is hyperparameter-tuned; the other four models retain fixed settings. The search space covers seven parameters via RandomizedSearchCV (25 candidate draws, inner 5-fold stratified cross-validation, AUC scoring), using only that horizon's training split. Tuning moved held-out test AUC by at most 0.002 relative to untuned defaults at every horizon.

Post-hoc temperature scaling is fit on a calibration slice carved from the test set; expected calibration error (ECE) and Brier score are always reported on the remaining eval half, so a temperature fit never inflates its own reported calibration.

3.5 Statistical Testing

Paired DeLong tests ([9] formulation) compare AUCs that share the same evaluated students: (i) XGBoost versus logistic regression within each horizon (RQ3), and (ii) horizon t versus horizon t+1 for XGBoost (RQ1), where week-t predictions are first restricted to the intersection with R_{t+1}. Bootstrap 95% confidence intervals (2000 resamples) are reported for every (horizon, model) AUC/PR-AUC/Brier combination as a model-agnostic complement.

3.6 Cross-Presentation Robustness: Within-Module Walk-Forward Temporal Holdout

A walk-forward design is used per module. For a module with chronologically sorted presentations [P1, ..., Pn], fold k trains on [P1..Pk] and evaluates on P_{k+1}, always strictly later than everything used to train it. XGBoost hyperparameters are re-tuned per fold rather than reused from Section III.D, because that search's training population pools students from every presentation, including whichever presentation a given fold treats as unseen future data.

3.7 Reproducibility

The analysis pipeline is organized as a sequence of versioned scripts, each consuming the frozen output of the previous stage. A single random seed (42) governs the train/test split, all model random states, the hyperparameter search, and bootstrap resampling.

4. Results

4.1 RQ1 – Temporal Reliability

AUC rises monotonically for every model as the evidence window lengthens. For XGBoost -- the frontier winner at every horizon -- the increase from week 1 to week 6 is 0.7408 to 0.8076 (+0.067 AUC). Every consecutive-week increase is statistically significant via paired DeLong test on the shrinking-population intersection (week 1 to 2 p = 5.5e-6; week 2 to 3 p < 1e-15; week 3 to 4 p = 3.6e-13; week 4 to 6 p < 1e-15).

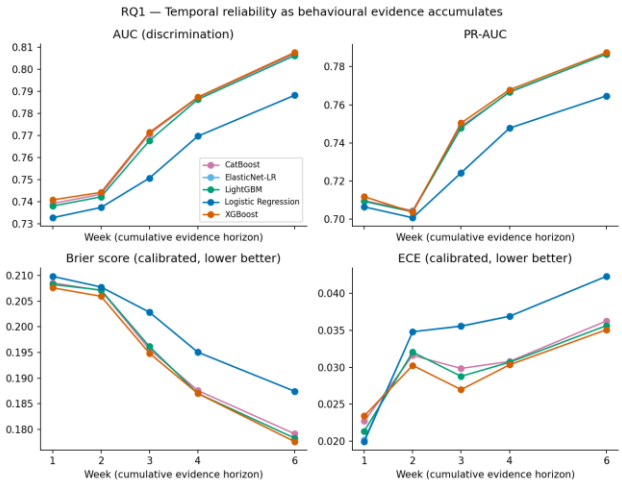

Fig. 2 AUC, PR-AUC, calibrated Brier score, and calibrated ECE by model across the five prediction horizons.

4.2 RQ2 – Reachability: The Third Axis

Reachability, not just reliability, changes with horizon (Table I). Between week 1 and week 6, reachability falls from 89.5% to 82.9% -- 2,141 students who were addressable at week 1 have already withdrawn by week 6. Future-at-risk prevalence within R_t also falls over the same window (0.472 to 0.430): early withdrawal progressively removes a disproportionately high-risk subgroup from the addressable population.

TABLE I
Earliness, Reachability, and Reliability by Horizon

Week	Reachable |R_t|	Reachability	Future-Risk Prev.	Best AUC
1	28,936	89.5%	0.472	0.741
2	27,881	86.3%	0.452	0.744
3	27,671	85.6%	0.448	0.771
4	27,307	84.5%	0.441	0.787
6	26,795	82.9%	0.430	0.808

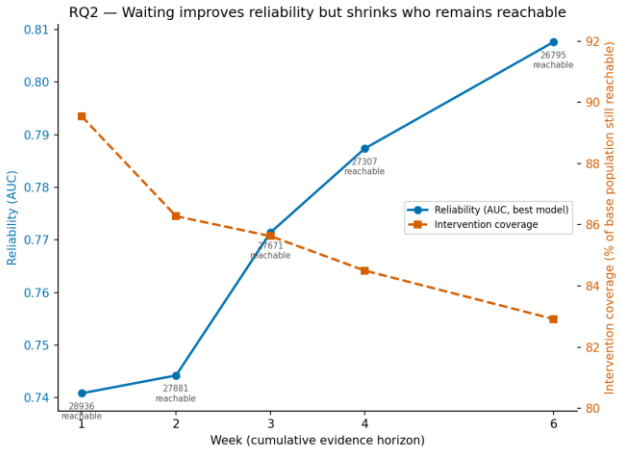

Fig. 3 Reliability (AUC) rising against reachability falling, by horizon.

4.3 RQ3 – Model Complexity: Amplification, Not a Threshold Effect

XGBoost's AUC advantage over logistic regression is statistically significant from the first horizon, growing roughly 2.4-fold from week 1 (+0.008 AUC, p = 2.6e-4) to week 6 (+0.019 AUC, p = 6.0e-11). This is read as progressive amplification of nonlinear modelling gains as evidence accumulates and diversifies, rather than a threshold effect.

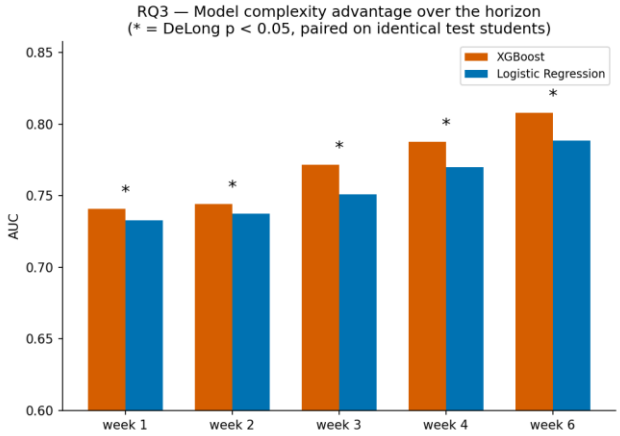

Fig. 4 XGBoost versus logistic-regression AUC by horizon, with DeLong significance markers.

4.4 RQ4 – Within-Module Temporal Generalization and Random-Split Optimism

Under strict walk-forward temporal holdout, the rising-with-horizon pattern survives in 6 of 7 modules. Comparing each fold's temporal-holdout AUC to the same module/week's random-split AUC gives a mean gap of +0.0223 AUC (median +0.0282) after per-fold leakage-proof tuning -- a smaller gap than the +0.035 obtained under fixed, untuned hyperparameters, indicating part of the original estimate was inflated by suboptimal defaults rather than genuine distribution shift, though a real gap persists.

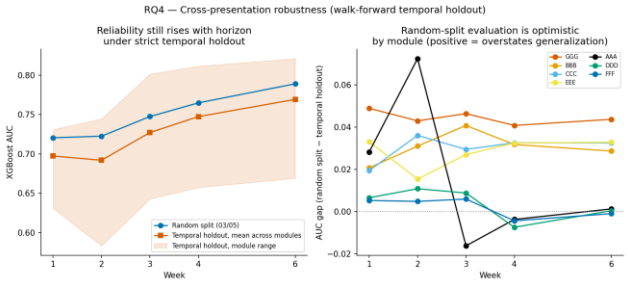

Fig. 5 Random-split vs. temporal-holdout AUC by horizon (left) and the resulting optimism gap by module (right).

4.5 RQ5 – Evolving Determinants of Risk and Module Heterogeneity

XGBoost SHAP ([10]) importances show a clear compositional shift: demographic and structural features dominate at weeks 1-2, while assessment-participation features (active days, average score) climb sharply from week 3 onward. Module-stratified reachability shows two distinct attrition patterns: front-loaded (module BBB, already low reachability by week 1, then flat) and persistent (module CCC, declining continuously through week 6).

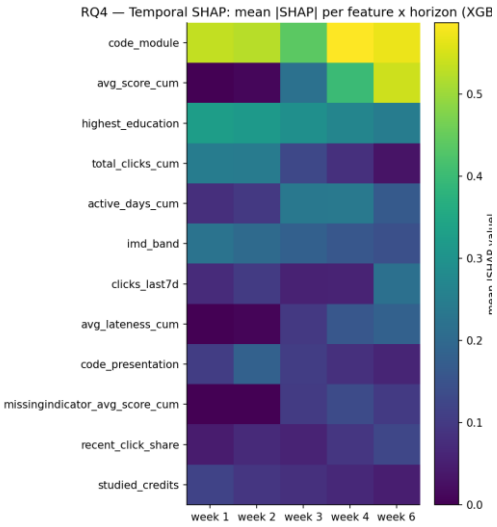

Fig. 6 Mean absolute SHAP value by feature and horizon.

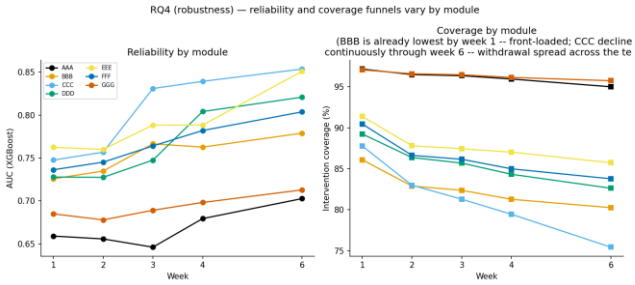

Fig. 7 XGBoost AUC (left) and reachability (right) by module and horizon.

5. Discussion

Several boundaries limit how these results should be read. The train/test split underlying Sections IV.A-C and E is a single random enrollment-level split shared across all modules and presentations, bounded to enrollments under shared curricular context -- which is exactly why Section III.F's walk-forward design exists as a separate analysis.

This paper does not attempt a cross-module leave-one-out transfer study; whether its reachability and amplification findings hold under cross-module transfer is an open question. The central operational implication is that “the model is more reliable if you wait” and “the model is still useful if you wait” are different claims: a deployment that waits until week 6 to issue its first warning is trading against a shrinking, non-randomly-selected population, not simply a fixed population's prediction quality for a better one.

6. Conclusions

Waiting improves prediction: reliability rises with every additional week of behavioural evidence. But waiting also removes students from the population that can still be helped, and does so in a way that is not risk-neutral. Meanwhile, the advantage of model complexity is an amplification present from the start rather than a switch that turns on partway through the course, and conventional random-split evaluation systematically understates the difficulty of predicting a genuinely future cohort, even after controlling for hyperparameter quality. Together, these results reframe the central question for early-warning deployment away from “how accurate can this model become” and toward: when does a prediction become reliable enough to act on, and how many students remain reachable by the time it does?

REFERENCES

1. R. da Silva, J. Eicher, and G. Longo, "A unified survival benchmark for temporal dropout risk prediction in Learning Analytics," *arXiv:2604.08870*, Jul. 2026.
2. Q. Li and J. Tan, "Target-adaptive probability calibration for reliable student risk prediction under cross-module dataset shift," *Research Square*, rs-10648448, Aug. 2026, doi: 10.21203/rs.3.rs-10648448/v1.
3. J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics dataset," *Sci. Data*, vol. 4, p. 170171, 2017, doi: 10.1038/sdata.2017.171.
4. N. Sghir, A. Adadi, and M. Lahmer, "Recent advances in predictive learning analytics: A decade systematic review (2012-2022)," *Educ. Inf. Technol.*, vol. 28, no. 7, pp. 8299-8333, 2023, doi: 10.1007/s10639-022-11536-0.
5. J. Gardner and C. Brooks, "Evaluating predictive models of student success: Closing the methodological gap," *arXiv:1801.08494*, 2018.
6. A. Gleiss, R. Oberbauer, and G. Heinze, "An unjustified benefit: immortal time bias in the analysis of time-dependent events," *Transplant Int.*, vol. 31, no. 2, pp. 125-130, 2018, doi: 10.1111/tri.13081.
7. B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: the Achilles heel of predictive analytics," *BMC Med.*, vol. 17, p. 230, 2019, doi: 10.1186/s12916-019-1466-7.
8. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.
9. X. Sun and W. Xu, "Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves," *IEEE Signal Process. Lett.*, vol. 21, no. 11, pp. 1389-1393, 2014.
10. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765-4774.