

Student-Disjoint Cross-Module Benchmarking for Calibrated Early Warning in Online Learning

Mika Lim

University Innovation Center, Seoul National University, Seoul, Republic of Korea

Abstract—Early-warning models in online learning are commonly evaluated with random data partitions, even though deployment often requires transfer to a course that supplied no training data. This study establishes a leakage-resistant benchmark for cross-module risk prediction using the Open University Learning Analytics Dataset. The analysis covered 32,593 enrollments from 28,785 learners, seven modules, 22 module presentations, and 10,655,280 daily virtual learning environment interaction records. Predictions were produced after 7, 14, and 28 days for learners still active at each cutoff. Logistic regression, random forest, and histogram gradient boosting were evaluated by leaving one entire module out and removing every learner who also appeared in the held-out module from training. Probability calibration was learned only from student-grouped out-of-fold predictions, using the same nested procedure in both validation designs. Random student-disjoint cross-validation yielded an area under the receiver operating characteristic curve (AUROC) of 0.716, whereas held-out-module evaluation yielded an AUROC of 0.661, revealing substantial contextual optimism. Under the fixed hyperparameter configurations used in this benchmark, logistic regression achieved higher cross-module performance than both nonlinear models. A virtual-learning-environment-only specification preserved discrimination while improving precision-recall performance and Brier score relative to adding assessment-submission features. Calibration nevertheless varied markedly across modules, indicating that global post-hoc calibration did not eliminate course-level distribution shift. The benchmark shows that student independence, course separation, feature availability, and probability quality must be evaluated jointly before an early-warning model is considered transferable.

Keywords—*Learning analytics; educational data mining; early warning; student risk prediction; cross-course generalization; probability calibration; data leakage; online learning*

I. INTRODUCTION

Learning management systems record how learners access resources, participate in activities, and submit assessments. These traces support early-warning models that prioritize timely academic support. Research using the Open University Learning Analytics Dataset (OULAD) and other large online-learning corpora has shown that behavioral variables can predict failure or withdrawal before a course ends [1]–[4]. Transfer to a new course, however, is a different problem from predicting held-out learners sampled from the same mixture of courses. Course organization, assessment timing, resource structure, and outcome prevalence can change the relationship between a feature and the target. A model may therefore appear accurate in a random split yet fail when used in an unseen instructional context.

The distinction matters technically and educationally. Repeated learners can enroll in more than one module, allowing student-level regularities to cross a nominal train-test boundary. Random partitions also mix course contexts, so the test distribution resembles the training distribution more closely than a future deployment course. Leakage and distribution mismatch can inflate claims in machine-learning-based science [10]. Recent learning-analytics research has consequently emphasized transfer, external validity, and model stability across courses and cohorts [5]–[9]. Yet many studies still optimize discrimination alone. Early-warning systems use predicted probabilities to allocate scarce support; poor calibration can over- or under-estimate the number of learners requiring intervention even when rank discrimination is acceptable [11], [12].

A second concern is feature legitimacy. Assessment scores can be highly predictive, but OULAD reports submission dates rather than the time at which grades became available. Treating every recorded score as known at the prediction cutoff can therefore introduce timing ambiguity. Demographic variables may improve average performance while creating governance and fairness concerns. Course-specific assessment counts and due-date patterns may also transfer poorly because they encode instructional design rather than general engagement. A deployable benchmark should distinguish registration data, virtual learning environment (VLE) behavior, submission behavior, recorded scores, and protected or sensitive attributes instead of aggregating them into an undifferentiated feature set.

The computational contribution of this study is an evaluation protocol rather than a new classifier: a reproducible student-disjoint, leave-one-module-out benchmark for calibrated early-warning prediction on all seven OULAD modules. It makes four contributions. First, it separates both modules and learner identities between outer training and test sets. Second, it evaluates three prediction horizons while defining a dynamic risk set of learners who remain enrolled at each cutoff. Third, it compares feature families according to operational availability and transferability. Fourth, it evaluates probability quality, module heterogeneity, paired model differences, and model-agnostic feature importance rather than reporting a single accuracy value. The technical novelty lies in coupling these safeguards within a single transfer benchmark rather than evaluating them independently: learner identities and modules are separated at the outer boundary, the risk set is redefined at each prediction cutoff, calibration is learned entirely within training data, and feature families are compared under the same deployment-oriented protocol.

RQ1. How do validation design, prediction horizon, and model class affect cross-module discrimination and probability quality?

RQ2. Which feature families provide the most transferable balance of discrimination and calibration for an unseen module?

RQ3. How stable are calibrated risks across held-out modules, outcomes, and time, and which observed behaviors drive the predictions?

II. RELATED WORK

A. Early Risk Prediction from Learning Traces

OULAD links demographics, registration, assessments, course structure, and daily VLE interactions for seven anonymized modules [1]. Deep and ensemble models have achieved strong within-dataset prediction, and flexible feature-selection methods can reduce manual engineering [2], [3]. Higher temporal granularity often improves early prediction [4]. These findings establish the predictive value of activity traces, but a model trained and tested within a pooled course mixture answers a narrower question than transfer to a previously unseen module. High apparent accuracy may also depend on final-course information, assessment schedules, or learners represented in both partitions.

B. Cross-Course Generalization

Transfer-learning research has trained models across multiple massive open online courses and used course metadata or prior iterations to warm-start prediction [5]. Other work shows that generalizability depends on course and cohort context and that behavioral models can lose stability after transfer [6], [13]. Recent studies have evaluated digital-trace models across computer-science courses [7] and multiple technical-education courses [8], while modelability has been proposed as a design principle for scalable predictive learning analytics [9]. The present study complements this literature by combining course-disjoint testing with complete learner-identity separation, multiple early cutoffs, feature-availability ablations, and calibration analysis on a widely reused public benchmark.

C. Calibration, Errors, and Transparency

The area under the receiver operating characteristic curve (AUROC) measures ranking, not whether a predicted risk of 0.60 corresponds to an observed event rate near 0.60. Modern classifiers can be miscalibrated even when discrimination is strong [14]. Platt scaling provides a simple post-hoc mapping from model scores to probabilities [15], while Brier score [17] and expected calibration error [14] summarize probability accuracy. Because non-success is common but not balanced uniformly across modules, area under the precision-recall curve is also necessary [16]. Educational consequences further require examination of errors and explanation: confident mistakes can erode adoption, and important contextual causes of non-success may remain absent from platform data [12]. Transparent comparisons between simple and complex models are therefore preferable to assuming that higher-capacity algorithms will generalize better [21].

III. METHODS

A. Data Source and Integrity Checks

The study used OULAD, released under a Creative Commons Attribution 4.0 license [1]. The seven relational tables were joined by module, presentation, learner, assessment, and VLE-site identifiers. Before modeling, record counts and module-presentation coverage were checked against the data descriptor. The verified data contained 32,593 enrollment records, 28,785 unique learners, seven modules, 22 presentations, 206 assessments, 6,364 VLE sites, and 10,655,280 student-VLE daily interaction records. No duplicate enrollment keys or missing outcomes were present.

The binary target was non-success, defined as a final result of Fail or Withdrawn; Pass and Distinction formed the comparison class. At each prediction day, learners whose recorded unregistration date preceded the cutoff were excluded because they were no longer candidates for prospective support. This dynamic risk-set definition produced 29,196 active enrollments at day 7, 28,128 at day 14, and 27,553 at day 28. The positive prevalence declined from 0.473 to 0.442 as already-withdrawn learners left the active set. A fixed-cohort sensitivity analysis restricted all three horizons to the 27,553 enrollments still active at day 28. The active cohorts and corresponding outcome prevalence are summarized in Table I.

TABLE I. ACTIVE COHORTS AND OUTCOME PREVALENCE

Day	Active n	Non-success n	Prevalence
7	29,196	13,811	0.473
14	28,128	12,743	0.453
28	27,553	12,168	0.442

B. Feature Construction

All time-varying features used records dated from course day 0 through the day preceding the cutoff, so no same-day or future event entered a prediction. Registration variables comprised registration timing, previous attempts, and studied credits. VLE features summarized total, recent, and first-week clicks; active days and their ratio to elapsed days; daily mean, standard deviation, and maximum; number of unique resources; recent-to-first-week ratio; and clicks within 14 activity categories. Nonnegative skewed counts were transformed by $\log(1+x)$. Submission-behavior features counted submitted and late assessments, indicated whether any assessment had been submitted, and summarized timing relative to due dates.

Five feature specifications were evaluated. Registration used three static variables. VLE added 23 module-agnostic activity features. Operational added four assessment-submission variables but no scores. Score-augmented added mean, minimum, and maximum recorded score as a timing-sensitivity analysis. Demographic-augmented added gender, region, prior education, deprivation band, age band, and disability to the operational set. The primary transfer interpretation emphasized the VLE and operational specifications; scores and demographics were not treated as required deployment inputs.

Preprocessing was fit only on each training set. Numeric missing values used the training median, counts used structural zeros where no event occurred, and demographic categories used the training mode plus one-hot encoding.

C. Models and Calibration

Three classifiers represented increasing functional complexity: L2-regularized logistic regression, a random forest with 160 trees and a minimum leaf size of 10 [18], and histogram gradient boosting with 180 iterations, learning rate 0.08, 31 leaf nodes, and L2 regularization. Hyperparameters were fixed across modules and cutoffs to avoid adapting the benchmark to a held-out target. Accordingly, comparisons among model classes refer to these prespecified configurations and should not be interpreted as an exhaustive test of each algorithm after hyperparameter optimization. Model-class comparisons used the operational specification at all three cutoffs. Day-28 feature-family analyses used logistic regression, the strongest cross-module model in that comparison, to isolate feature effects without multiplying the model grid. Logistic numeric inputs were standardized within the training pipeline. Implementations used scikit-learn [19] with a fixed random seed.

For each outer fold, 4-fold stratified group cross-validation generated out-of-fold predictions within the training modules, with learner identifier as the group. A Platt logistic calibrator was fit to those out-of-fold scores and then applied to the untouched held-out module [15]. The F1 threshold was selected from calibrated training predictions only. Uncalibrated and calibrated probabilities were retained; threshold-dependent results were secondary because intervention capacity was not specified.

D. Validation Design and Metrics

The principal evaluation used seven leave-one-module-out folds. In each fold, all enrollments from one module were held out. Every enrollment of a learner appearing in that test module was then removed from the remaining training modules. Depending on the fold, this identity filter removed 2 to 1,772 otherwise eligible training rows. This design estimates transfer to a new module for learners unseen during model fitting. A student-disjoint stratified 5-fold split that ignored module boundaries served as the conventional comparator. The comparator used the same 4-fold, student-grouped inner procedure for Platt calibration, ensuring that the contrast changed the module boundary rather than the calibration protocol.

Discrimination was assessed with AUROC and area under the precision-recall curve (AUPRC). Probability quality was assessed with Brier score, log loss, and expected calibration error computed with ten equal-frequency bins. Accuracy, balanced accuracy, precision, recall, and F1 were reported for training-derived thresholds. Student-cluster bootstrapping preserved repeated enrollments when calculating 95 percent intervals and paired model differences. Permutation importance was computed within each held-out module as the mean decrease in AUROC over three independent permutations of each feature. No normalization was applied to the importance values. Module-specific importance estimates were then aggregated using held-out-module sample sizes as weights. Negative permutation importance values were retained without

truncation or transformation; Fig. 5 displays only the nine highest-ranked features.

IV. RESULTS

A. Model and Horizon Comparison

Under the prespecified hyperparameter configurations, logistic regression achieved the highest pooled cross-module AUROC at every horizon. With operational features, calibrated AUROC increased from 0.654 at day 7 to 0.667 at day 14 and then decreased slightly to 0.661 at day 28 (Fig. 1 and Table II). The corresponding day-28 AUPRC, Brier score, and expected calibration error were 0.594, 0.237, and 0.071; AUPRC exceeded the active-cohort prevalence baseline of 0.442. The student-cluster bootstrap interval for day-28 AUROC was 0.655 to 0.667. Under these fixed configurations, random forest and histogram gradient boosting showed lower transfer performance than logistic regression. At day 28, logistic regression exceeded random forest by 0.0136 AUROC (95 percent interval 0.0104 to 0.0170) and histogram gradient boosting by 0.0082 (0.0050 to 0.0115). Its Brier score was lower by 0.0063 and 0.0060, respectively, with both paired intervals excluding zero.

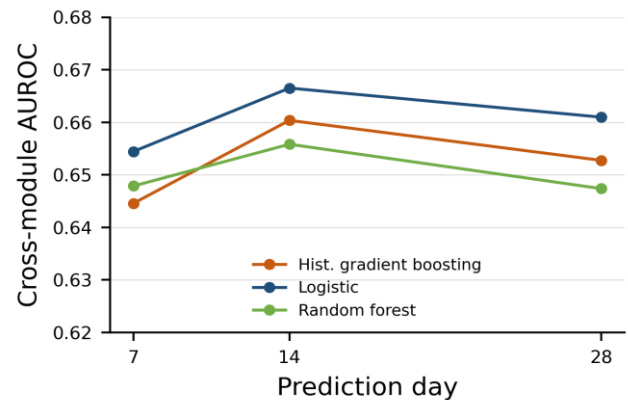


Fig. 1. Student-disjoint held-out-module AUROC by prediction day and model using operational features.

TABLE II. CROSS-MODULE PERFORMANCE WITH OPERATIONAL FEATURES

Day	Model	AUROC	AUPRC	Brier
7	Logistic	0.654	0.611	0.232
7	RF	0.648	0.598	0.234
7	HGB	0.645	0.600	0.234
14	Logistic	0.667	0.615	0.229
14	RF	0.656	0.596	0.231
14	HGB	0.660	0.604	0.230
28	Logistic	0.661	0.594	0.237
28	RF	0.647	0.575	0.243
28	HGB	0.653	0.574	0.243

RF: random forest; HGB: histogram gradient boosting. Values use Platt-calibrated probabilities.

B. Validation-Design Gap

With the same nested calibration protocol, the conventional student-disjoint 5-fold design, which mixed modules across

folds, produced AUROC 0.716, AUPRC 0.673, Brier score 0.211, and expected calibration error 0.017 at day 28. The stricter student-disjoint held-out-module design produced 0.661, 0.594, 0.237, and 0.071, respectively (Fig. 2). Thus, merely preventing learner overlap did not simulate deployment to an unseen module. Module separation exposed an AUROC optimism gap of 0.055 and a more than fourfold increase in calibration error. This result answers the validation component of RQ1: context mixing remained a major source of optimistic evaluation after learner identities were separated.

The magnitude of this gap should not be treated as directly equivalent to prior transfer results because the prediction targets, datasets, and validation schemes differ across studies. Kleter et al. [6] evaluated both numerical and binary predictive models across courses and cohorts and reported performance degradation after generalization, with the extent of degradation depending on initial model performance. Schwerter et al. [7] reported in-sample AUROC ranges of 0.73–0.81 for Elastic Net, 0.94–1.00 for random forest, and 0.77–0.99 for XGBoost, while cross-course and cross-institution performance declined and varied substantially across source-target pairings. By contrast, the present 0.716-to-0.661 comparison holds the dataset, outcome definition, feature specification, learner-disjoint rule, and nested calibration procedure constant, so the 0.055 difference more directly isolates the effect of enforcing module separation.

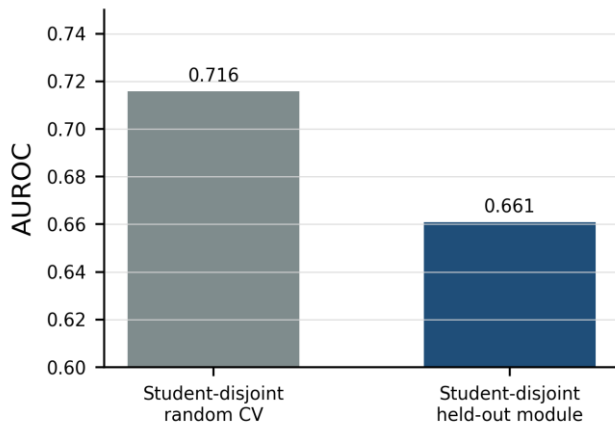


Fig. 2. Validation design changes apparent day-28 logistic performance despite student-disjoint partitions.

C. Feature Transferability

Feature-family results showed that additional information was not uniformly transferable (Table III and Fig. 3). Registration data alone provided weak discrimination. VLE features increased AUROC to 0.660 and AUPRC to 0.604 while achieving the lowest Brier score among the behavior-based models, 0.230. Adding submission behavior changed AUROC by less than 0.001 but reduced AUPRC by 0.010 and increased Brier score by 0.0068. In paired student-cluster bootstrap comparisons, the VLE-minus-operational AUROC difference was -0.0009 (95 percent interval -0.0055 to 0.0036), whereas its AUPRC advantage was 0.0102 (0.0057 to 0.0145) and its Brier difference was -0.0068 (-0.0085 to -0.0050). The latter two improvements were robust.

Recorded scores raised AUROC to 0.672 but reduced AUPRC and worsened Brier score. Demographics yielded the highest AUROC, 0.684, and AUPRC, 0.619, but used six sensitive or governance-relevant variables. These sensitivity results do not establish that such variables should be deployed. Rather, they demonstrate that feature selection changes the balance between ranking, probability quality, timing validity, and responsible use. The VLE-only specification provided the strongest transfer-oriented compromise for RQ2.

TABLE III. DAY-28 LOGISTIC FEATURE-SET ABLATION

Feature set	p	AUROC	AUPRC	Brier	ECE
Registration	3	0.551	0.507	0.243	0.015
VLE	26	0.660	0.604	0.230	0.044
Operational	30	0.661	0.594	0.237	0.071
+ Scores	33	0.672	0.574	0.238	0.084
+ Demographics	36*	0.684	0.619	0.230	0.067

p counts six demographic fields before one-hot expansion. ECE: expected calibration error.

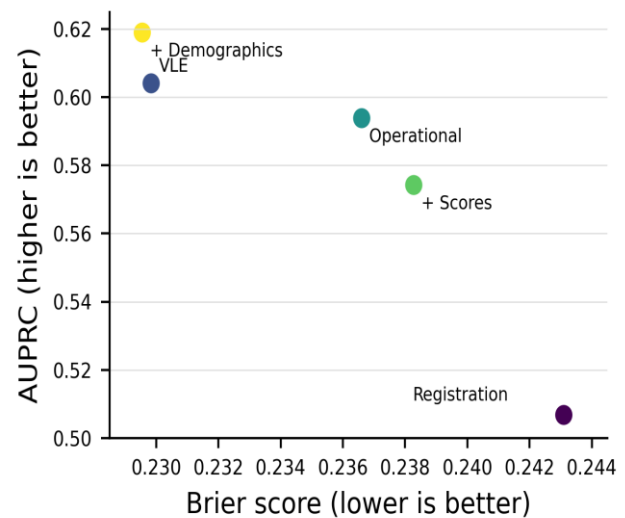


Fig. 3. Trade-off between precision-recall performance and probability accuracy across feature families.

D. Module Heterogeneity and Calibration

Held-out-module results varied substantially. For the day-28 operational logistic model, AUROC ranged from 0.635 for GGG to 0.736 for EEE. Expected calibration error ranged from 0.027 for CCC to 0.357 for GGG. The VLE-only reliability curves showed the same broader pattern: some modules lay above the identity line and others below it (Fig. 4). Global Platt scaling slightly improved pooled expected calibration error for logistic regression, from 0.073 to 0.071, but it did not correct target-specific base-rate and feature-distribution shifts. For random forest and histogram gradient boosting at day 28, Platt scaling worsened pooled Brier score and expected calibration error. Calibration therefore must be checked on each destination context rather than assumed to transfer with discrimination.

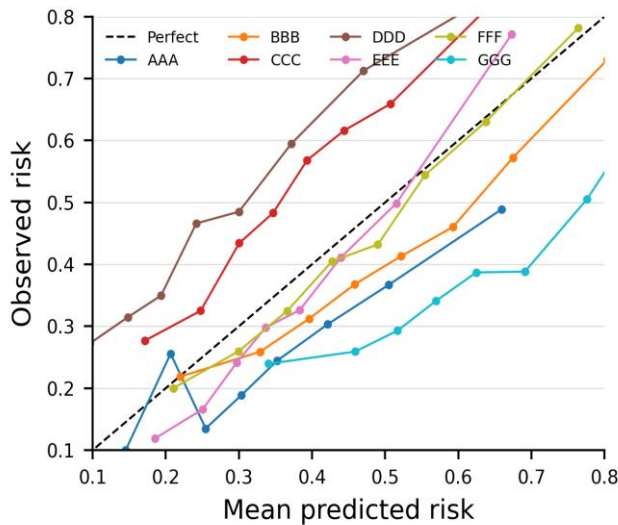


Fig. 4. Reliability of the day-28 VLE-only logistic model in each held-out module.

E. Time, Risk-Set Composition, and Errors

The fixed day-28 cohort analysis separated increasing observation length from attrition-induced changes in the active population. For operational logistic regression on the same 27,553 enrollments, AUROC was 0.644 at day 7, 0.665 at day 14, and 0.661 at day 28. The day-28 decline therefore was not explained solely by removal of already-withdrawn learners. For the VLE-only model, AUROC changed from 0.643 to 0.658 to 0.660, while AUPRC increased from 0.569 to 0.594 to 0.604. Submission behavior helped at day 14 but harmed day-28 probability quality, consistent with growing cross-module differences in assessment schedules. More elapsed time did not guarantee better transferable operational features.

At the training-derived decision thresholds, the operational logistic model correctly flagged 76.2 percent of eventual failures and 72.6 percent of withdrawals, but it also flagged 53.8 percent of eventual passes and 47.4 percent of distinctions. These errors caution against treating predicted risk as a deterministic judgment. The output is more appropriate for low-stakes outreach or instructor review than for exclusionary decisions.

F. Feature Importance

Permutation importance for the day-28 operational logistic model identified the number of assessment submissions as the strongest predictor, followed by active-day ratio, total clicks, submission timing, unique resources, and late submissions (Fig. 5). This result should be interpreted as within-held-out-module sensitivity rather than as evidence that feature importance is invariant across modules. The present analysis does not decompose within-module importance from cross-module stability; therefore, the weaker transfer performance observed after adding submission features is compatible with, but does not by itself establish, context-specific importance. A direct within-versus-cross-module importance decomposition would be required to test that mechanism.

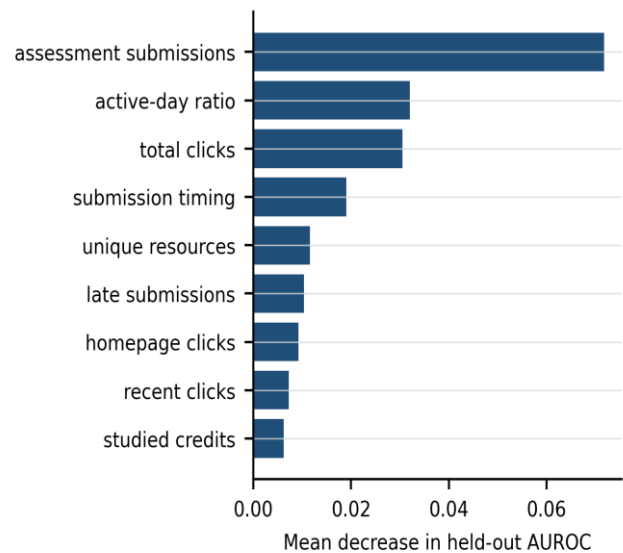


Fig. 5. Weighted permutation importance across day-28 held-out modules for operational logistic regression.

V. DISCUSSION

A. Principal Findings

The benchmark provides three answers. For RQ1, both learner identity and course context affected evaluation. Student-disjoint random cross-validation remained optimistic because modules were mixed across folds. The held-out-module design reduced AUROC by 0.055 and exposed materially worse calibration. Under the prespecified hyperparameter settings, logistic regression showed more reliable transfer than the two nonlinear models. This pattern is consistent with prior evidence that strong within-context performance does not guarantee cross-context generalization [6], [7], [10], but the present comparison cannot separate model-class effects from possible undertuning of the nonlinear models. The result does not imply that logistic regression is universally superior; it shows that model ranking depends on the deployment distribution represented by the validation design.

For RQ2, the VLE-only model offered a more defensible transfer point than the broader operational model. Submission variables were individually important, yet adding them reduced cross-module AUPRC and probability accuracy at day 28. Course-specific deadlines, assessment counts, and marking designs can change what a submission means. This is a negative-result contribution: predictive information can be locally useful while degrading transportability. Recorded scores and demographics altered performance again, but their timing and governance costs make them sensitivity conditions rather than default inputs. A feature's availability, stability, and policy acceptability should therefore be evaluated alongside its importance.

For RQ3, calibration and error rates varied sharply across modules. Platt scaling improved pooled logistic metrics only modestly and could not correct module-specific distribution shifts. The GGG fold was especially miscalibrated even though learner overlap was minimal, indicating that the problem was

contextual rather than identity leakage. Furthermore, the fixed-cohort analysis showed that additional weeks did not monotonically improve every specification. The predictive system changes as both the evidence window and the set of still-active learners evolve. Early-warning evaluation should therefore be framed as repeated risk-set prediction rather than a single static classification task.

B. Implications for Learning Analytics Systems

A practical implementation should treat transfer as a gated software lifecycle rather than a one-time model export. A source-trained model can produce provisional risk rankings for a new module, but its probability scale should be monitored against emerging local outcomes. Until enough labeled target data are available, institutions should communicate risk bands or prioritized outreach rather than precise probabilities. When a previous presentation exists, recalibration and threshold selection should use that presentation without contaminating the target evaluation. The benchmark also supports a parsimonious default: engagement variables that have comparable definitions across courses are preferable to assessment and demographic variables whose semantics or acceptability change by context. This sequence—identity-safe validation, destination monitoring, local recalibration, and threshold governance—translates the benchmark into testable requirements for learning-analytics software.

The educational interpretation must remain limited. Clicks and submissions are traces of platform interaction, not direct measures of knowledge, motivation, or ability. False positives may reflect strategic low-click study, prior knowledge, offline work, or inaccessible contextual information; false negatives may reflect sudden work, family, health, or technical events [12]. Explanations should therefore support conversation and resource allocation, not label learners as inherently likely to fail. This human-in-the-loop framing is consistent with calls for transparent, calibrated, and ethically governed predictive analytics [11], [20]–[22].

C. Methodological Contribution

The principal contribution is not a new classifier but a stricter, reusable benchmark specification. It defines the estimand as performance on enrollments from a module absent during training; removes all cross-boundary learner identities; fits preprocessing, calibration, and thresholds inside training data; applies the same nested calibration protocol to the conventional comparator; evaluates active learners at explicit temporal cutoffs; separates feature families by operational legitimacy; and reports discrimination, probability quality, heterogeneity, uncertainty, and explanation. This combination is necessary because a standard dataset does not by itself standardize a valid evaluation [10]. The accompanying scripts produce the cohort, features, folds, predictions, tables, and figures from the verified OULAD tables. The contribution therefore lies in the joint specification of these components: none is claimed as novel in isolation, but their integration makes the deployment estimand, leakage controls, calibration boundary, temporal risk set, and feature-availability constraints explicit within one reproducible benchmark.

D. Limitations

OULAD comes from one distance-learning institution and historical course offerings. Module labels and subject content are anonymized, limiting explanations of why particular modules shifted. VLE records are daily aggregates and omit offline study, motivation, communication outside the platform, and unexpected life events. The final result is a useful operational label but does not measure learning quality or the causal effect of intervention. The evaluated algorithms and fixed hyperparameters do not represent an exhaustive search for maximum predictive performance. Consequently, the lower transfer performance of random forest and histogram gradient boosting cannot be separated from the possibility that those prespecified configurations were undertuned for this setting. The study did not tune models for each target module, use target labels for recalibration, or evaluate interventions. These restrictions make the benchmark conservative but do not establish performance at another institution. Moreover, seven held-out modules are sufficient for a benchmark but not for estimating uncertainty over the wider population of courses; the learner-cluster intervals quantify sampling variation among observed learners, not between-course uncertainty. Expected calibration error is also bin-dependent. Finally, the score-augmented analysis cannot determine exactly when a grade became visible; it should not be interpreted as a prospective score-based deployment result. The feature-importance analysis is also descriptive: it summarizes predictive sensitivity within held-out modules but does not estimate a formal within-module versus cross-module importance decomposition.

VI. CONCLUSION

Within the OULAD setting examined here, evaluation on an unseen module produced materially different conclusions from student-disjoint cross-validation that mixed module contexts. Student-disjoint random cross-validation substantially overstated performance relative to a student-disjoint held-out-module benchmark even when both designs used the same nested calibration procedure. Under the prespecified hyperparameter configurations, logistic regression transferred better than the evaluated nonlinear models, and VLE-only engagement features achieved nearly identical discrimination with better precision-recall and probability accuracy than a broader assessment-aware specification. Calibration remained module-dependent, so a global correction was insufficient. The resulting evaluation architecture shifts attention from maximizing a pooled score to establishing whether a model is independent, transferable, calibrated, interpretable, and based on legitimate information. These conclusions are limited to seven anonymized modules from a single distance-learning institution and should not be interpreted as evidence of performance across institutions or course populations more broadly. Future work should replicate the benchmark on at least one independent institutional dataset, incorporate systematic hyperparameter tuning within the training boundary, and test whether feature-importance patterns remain stable across source and destination modules.

DATA AND CODE AVAILABILITY

OULAD is publicly available under the Creative Commons Attribution 4.0 license through the dataset DOI <https://doi.org/10.24432/C5KK69> and is described in [1]. The verified analysis used 10,655,280 student-VLE records spanning all 22 module presentations. Reproducible scripts and exact aggregate outputs are supplied with this manuscript; the source data are not redistributed.

ETHICS STATEMENT

This secondary analysis used an anonymized public dataset and involved no participant contact, intervention, or attempt at re-identification. Predictions were evaluated as research outputs and were not used to make decisions about individual learners.

DECLARATION ON GENERATIVE AI

An OpenAI generative AI tool was used for programming assistance, manuscript drafting, language refinement, and document formatting. The human author specified the research problem and retains responsibility for independently verifying the data, code, analyses, citations, interpretation, and all AI-assisted text before submission.

REFERENCES

- [1] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics dataset," *Scientific Data*, vol. 4, art. 170171, 2017, doi: 10.1038/sdata.2017.171.
- [2] H. Waheed, S. U. Hassan, N. R. Aljohani, B. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Computers in Human Behavior*, vol. 104, art. 106189, 2020, doi: 10.1016/j.chb.2019.106189.
- [3] A. Al-Zawqari et al., "A flexible feature selection approach for predicting students' academic performance in online courses," *Computers and Education: Artificial Intelligence*, vol. 3, art. 100103, 2022, doi: 10.1016/j.caeai.2022.100103.
- [4] C. Ye and G. Biswas, "Early prediction of student dropout and performance in MOOCs using higher granularity temporal information," *Journal of Learning Analytics*, vol. 1, no. 3, pp. 169–172, 2014, doi: 10.18608/jla.2014.13.14.
- [5] V. Swamy, M. Marras, and T. Käser, "Meta transfer learning for early success prediction in MOOCs," in *Proc. 9th ACM Conf. Learning at Scale*, 2022, pp. 121–132, doi: 10.1145/3491140.3528273.
- [6] S. Kleter, U. Matzat, and R. Conijn, "Boundary conditions of generalizing predictive models for academic performance: Within cohort versus within course," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 2183–2194, 2024, doi: 10.1109/TLT.2024.3443079.
- [7] J. Schwerter et al., "Cross-course generalizability of SRL-aligned predictive models using digital learning traces," *Computers & Education*, vol. 253, art. 105670, 2026, doi: 10.1016/j.compedu.2026.105670.
- [8] O. Ovtšarenko, "Using AI to forecast student dropout risk in technical education using a learning analytics approach," *Scientific Reports*, vol. 16, art. 14616, 2026, doi: 10.1038/s41598-026-44919-1.
- [9] A. Xu, Y. Zhang, A. B. Blake, and J. W. Stigler, "Modelability as a strategy for improving the generalizability and scalability of predictive models," *Journal of Learning Analytics*, vol. 13, no. 1, pp. 89–109, 2026, doi: 10.18608/jla.2026.9099.
- [10] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, art. 100804, 2023, doi: 10.1016/j.patter.2023.100804.
- [11] A. Mathrani, S. Sarvesh, and T. Umer, "Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics," *Computers and Education Open*, vol. 2, art. 100060, 2021, doi: 10.1016/j.caeo.2021.100060.
- [12] M. Hlosta, C. Herodotou, T. Papatoma, A. Gillespie, and P. Bergamin, "Predictive learning analytics in online education: A deeper understanding through explaining algorithmic errors," *Computers and Education: Artificial Intelligence*, vol. 3, art. 100108, 2022, doi: 10.1016/j.caeai.2022.100108.
- [13] R. Conijn, C. C. P. Snijders, P. A. M. Kleingeld, and U. Matzat, "Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS," *IEEE Transactions on Learning Technologies*, vol. 10, no. 1, pp. 17–29, 2017, doi: 10.1109/TLT.2016.2616312.
- [14] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning*, vol. 70, 2017, pp. 1321–1330.
- [15] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, A. J. Smola et al., Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.
- [16] 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, art. e0118432, 2015, doi: 10.1371/journal.pone.0118432.
- [17] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
- [18] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [19] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [20] H. Khosravi et al., "Explainable artificial intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 3, art. 100074, 2022, doi: 10.1016/j.caeai.2022.100074.
- [21] K. A. Bird, B. L. Castleman, Z. Mabel, and Y. Song, "Bringing transparency to predictive analytics: A systematic comparison of predictive modeling methods in higher education," *AERA Open*, vol. 7, no. 1, pp. 1–19, 2021, doi: 10.1177/23328584211037630.
- [22] M. Karimi-Haghighi, C. Castillo, D. Hernández-Leo, and V. Moreno Oliver, "Predicting early dropout: Calibration and algorithmic fairness considerations," in *Companion Proc. 11th Int. Conf. Learning Analytics and Knowledge (LAK21)*, 2021, pp. 294–304.