

Use of Artificial Intelligence in Open and Distance Learning: Opportunities and Challenges

Dr. Jayanta Doley

Associate Professor, Head, Department of Education, Bihpuria College, Bihpuria, Lakhimpur, Assam

Email l'd. jayantadoley20@gmail.com

HOW TO CITE:

Jayanta Doley (2026). Use of Artificial Intelligence in Open and Distance Learning: Opportunities and Challenges
International Journal of Special Education, 41(23s), 1562 - 1589

ABSTRACT:

Artificial intelligence (AI) can transform digital traces from open and distance learning (ODL) into early-warning information, but practical value derives from leakage control, robust validation, interpretability, and responsible use rather than merely discrimination. This study evaluates a leakage-controlled early-window pipeline on a public learner-level extract from the Open University Learning Analytics Dataset (OULAD), course CCC. The analytic sample comprised 1,043 unique learners with weekly assessment-score sums and virtual learning environment (VLE) click aggregates plus demographic/context variables. Risk was operationalized as $\text{final_score} < 40$; 379 learners (36.3%) met this score-defined criterion, 296 of whom had $\text{final_score} = 0$. Hence, the endpoint is interpreted as low/zero final performance rather than a reconstructed withdrawal label. Only weeks 1-4 of the eight observed weekly aggregates were used as predictors. Non-parametric inference, effect sizes, Spearman correlations, six machine-learning models, five-fold stratified cross-validation, bootstrap confidence intervals, feature-family ablation, SHAP explanations, and subgroup/error analyses were assembled. At-risk learners recorded substantially fewer early clicks (median 246 vs. 703.5; rank-biserial $r = -0.614$), lower positive-score occurrence (0.50 vs. 1.00; $r = -0.773$), and lower early mean score (17.5 vs. 77.75; $r = -0.880$), all Holm-adjusted $p < 0.001$. Regularized logistic regression achieved held-out ROC-AUC 0.936 (95% CI 0.901-0.968), precision-recall performance 0.915 (0.869-0.955), F1 0.789, and recall 0.789. Random Forest produced comparable discrimination and was used for TreeSHAP analysis, which ranked early mean score, positive-score occurrence, non-zero assessment performance, score trend, and recent-click share as the most informative predictors. Assessment-only features nearly matched the full feature set, whereas engagement alone remained informative and demographic/context variables were comparatively weak. The findings support human-in-the-loop risk ranking and targeted support, but formal probability calibration, external validation, sensitivity analysis of the zero-score subgroup, and comparison with very simple assessment rules remain necessary before

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the terms and conditions

of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).	deployment. AI in ODL is most defensible as transparent decision support that augments rather than replaces educational judgment. Keywords: artificial intelligence; open and distance learning; OULAD; learning analytics; early-warning systems; explainable AI; student-risk prediction
--	--

1. Introduction

Open and distance learning (ODL) expands participation by decoupling access to teaching from fixed place and, increasingly, from fixed time. This flexibility also transforms more learning activities into digitally mediated environments in which enrolment, assessment, resource access, and interaction traces can be observed at scale. The Open University Learning Analytics Dataset (OULAD) exemplifies this data-rich setting: it links learner characteristics, registrations, assessments, course metadata, and virtual learning environment (VLE) activity and was explicitly released to support reproducible learning-analytics research [1]. Recent systematic work confirms that learning analytics in distance education is increasingly used for monitoring, prediction, engagement analysis, and support, showing considerable heterogeneity in methods and implementation maturity [31].

AI extends learning analytics from descriptive monitoring toward predictive and adaptive decision support. Contemporary higher-education research includes student-success prediction, recommender systems, automated feedback, intelligent tutoring, and administrative decision support [4]. Yet the educational usefulness of these systems cannot be inferred from technical performance alone. Predictive systems influence who receives attention, when support is triggered, and what explanation is given to tutors and learners. Consequently, the same model can be technically accurate but operationally weak if its features are temporally unavailable, its probabilities are poorly calibrated, its errors burden particular groups, or its output cannot be translated into proportionate intervention.

This distinction is especially important in ODL. Learners participate asynchronously, engagement

can be intermittent without necessarily indicating failure, and life circumstances that affect persistence may be absent from platform traces. Large-scale institutional work at The Open University has demonstrated that predictive learning analytics can be implemented in distance education, but it also shows that tutor engagement and organizational processes mediate whether predictions become useful support [7]. Related qualitative error analysis has shown that algorithmic errors may reflect unobserved circumstances rather than simple model deficiency [10]. Thus, an empirically defensible study of AI in ODL should connect learner behavior, statistical evidence, predictive validity, explanations, errors, and intervention implications rather than treating classification accuracy as the endpoint.

The recent literature also narrows the space for credible novelty. OULAD has been used extensively for performance and risk prediction across multiple machine-learning families [11]. Explainable AI (XAI) is no longer unusual in educational prediction; SHAP, local explanations, and interpretable models have been used to support personalized intervention [13]. Recent OULAD research has further incorporated early-window prediction, calibration, fairness auditing, selective prediction, and capacity-aware alerting [37]. Therefore, novelty cannot defensibly rest on comparing several classifiers or adding a SHAP plot. The unresolved methodological need is a compact, reproducible chain of evidence in which early behavioral signals are statistically validated, model complexity is tested rather than presumed beneficial, explanations are linked to error patterns and feature ablation, and the resulting signals are interpreted as actionable but non-causal evidence for ODL support.

The objective of this study is to evaluate how early assessment and VLE behavior can be transformed into interpretable AI-based risk information using a public OULAD-derived learner cohort. The study addresses five research questions: RQ1, which early engagement, assessment, and learner-context characteristics are associated with final academic risk? RQ2, how accurately can complementary machine-learning models identify at-risk learners using only information available in the first half of the observed learning window? RQ3, which variables contribute most strongly to prediction, and how much predictive information is carried by engagement, assessment, and demographic/context feature families when modeled separately? RQ4, how consistent are conclusions across statistical tests, cross-validation, model comparison, SHAP explanations, and error analysis? RQ5, how can these empirical findings be translated into responsible AI-enabled support in ODL?

The study contributes four elements. First, it implements a temporally restricted weeks-1-4 prediction design and excludes later assessments, later VLE clicks, and final_score from the predictor matrix. Second, it integrates effect-size-aware statistical analysis with six complementary classifiers and cross-validated model comparison. Third, it combines feature-family ablation, global/local SHAP evidence, preliminary subgroup performance analysis, and false-positive/false-negative profiling to distinguish predictive association from educational interpretation. Fourth, it derives an intervention framework in which risk scores are accompanied by observable reason codes, tutor review, proportionate support, and continuous monitoring. This contribution is integrative and deployment-oriented rather than an algorithmic novelty claim.

2. Literature Review

2.1 AI, learning analytics, and ODL

AI in higher education has grown rapidly, but applications and evidence remain uneven across disciplines and purposes. Crompton and Burke's review of 138 studies found substantial

expansion of AI research and a broad range of student- and institution-facing applications [3]. Bond et al. subsequently argued that the field requires stronger ethics, collaboration, and methodological rigor, a conclusion directly relevant to predictive ODL systems because apparently successful models can still be difficult to reproduce or deploy responsibly [4].

Learning analytics provides the bridge from digital educational data to AI-supported decisions. Romero and Ventura describe educational data mining and learning analytics as overlapping fields that use behavioral and performance data for prediction, profiling, recommendation, and understanding learning processes [5]. In distance education specifically, Palanci et al. found that prediction and engagement monitoring are prominent applications, while the diversity of platforms, measures, and designs complicates generalization [31]. This makes public datasets useful for methodological reproducibility but also makes transfer to new institutions an empirical question rather than an assumption.

Evidence-to-action remains a recurring weakness. Ifenthaler and Yau's systematic review found that learning analytics can support study success but emphasized the need for stronger evidence linking analytics to effective interventions [6]. Zhao et al. similarly designed a human-centered learning-analytics framework around the step that converts analytic results into an intervention strategy, explicitly addressing the gap between prediction and educational action [33]. Ouyang and Zhang's 2024 review of AI-driven learning analytics found a lack of instructional intervention support and argued for actionable suggestions grounded in educational theory [34].

2.2 Student-risk prediction and early-warning design

Student-risk prediction is one of the most mature AI applications in online and distance learning. Adnan et al. demonstrated that predictive performance changes as different percentages of course activity become available, establishing the importance of explicit prediction windows for

early intervention [38]. Bañeres et al. went beyond retrospective prediction by integrating an early-warning system with intervention in online learning [14]. These studies underscore a central design constraint: a risk model is educationally valuable only if it operates early enough for support to remain feasible.

OULAD has become a benchmark for this work because it combines longitudinal VLE behavior, assessments, and learner context. Jin et al.'s 2024 systematic review of OULAD predictive modeling showed substantial use of traditional machine learning and multiple target definitions, but also highlighted variation in features and evaluation practice [11]. Ujkani et al. applied XAI to OULAD course-success prediction and argued that interpretability is necessary for translating model outputs into support [35]. Wu et al. extended this direction by evaluating four-week withdrawal-risk models with calibration, fairness auditing, and policy-linked alerting [37]. These developments imply that a contemporary empirical contribution should report not only discrimination but also uncertainty, errors, explainability, and operational constraints.

A further concern is data leakage. Educational records often include variables recorded after the practical point at which an intervention decision would be made. Kapoor and Narayanan show more broadly that leakage can produce inflated and irreproducible machine-learning claims when information unavailable at prediction time enters model development [23]. In learning analytics, leakage can arise from using end-of-course outcomes, later assessments, or post-withdrawal activity as predictors. Temporal feature windows should therefore be treated as a first-order methodological requirement rather than an implementation detail.

2.3 Explainability, errors, and responsible educational AI

Explainability in educational AI is not synonymous with feature importance. Khosravi et al. argue that explanations must be evaluated relative to educational stakeholders and decisions [12]. Nagy and Molontay combined interpretable dropout prediction with personalized

intervention, illustrating how feature attributions can be connected to support design rather than presented as model diagnostics alone [13]. Gunasekara and Saarela's OULAD study further shows the continuing need to evaluate both predictive performance and explainability rather than assuming that complex models are preferable [36].

SHAP provides a mathematically consistent additive attribution framework and efficient algorithms for tree ensembles [25]. TreeSHAP supports global ranking and case-level decomposition, but attribution remains predictive: it quantifies how a fitted model uses a feature, not what would happen if the feature were changed by intervention [24]. This distinction is crucial in education because engagement and assessment variables can be consequences of latent motivation, time constraints, prior preparation, or course design.

Responsible AI adds institutional obligations beyond explanation. Empirical work on learning-analytics ethics identifies privacy, informed use, transparency, and consequences for learners as recurring concerns [17]. Baker and Hawn show that algorithmic bias in education can arise from data, labels, measurement, and deployment processes, so removing sensitive attributes does not by itself ensure fairness [18]. Nguyen et al. synthesize ethical principles for AI in education around transparency, accountability, privacy, human agency, and fairness [19]. Human-centered reviews likewise find limited end-user involvement and insufficient attention to safety and trustworthiness in many learning-analytics/AI systems [16].

The resulting research gap is specific. Recent studies demonstrate strong predictive models, OULAD-based XAI, early-window prediction, and even deployment-oriented fairness analysis. However, many empirical papers still privilege predictive scores over a joined interpretation of (i) statistically validated behavior differences, (ii) the incremental value of behavior versus assessment/context features, (iii) whether complex models materially outperform simpler baselines, (iv) why false positives and false

negatives occur in behavioral terms, and (v) what concrete safeguards should govern intervention. The present study addresses this gap through a single evidence chain and explicitly avoids claiming that model attributions are causal effects.

3. Methodology

3.1 Dataset and analytical cohort

The study used a public learner-level extract labeled as OULAD course CCC in the Dataset_MLA repository [39]. The repository describes score_1–score_8 as weekly assessment scores, sum_click_1–sum_click_8 as weekly VLE interaction totals, final_score as the learner's final module result, and the remaining fields as learner-context attributes. OULAD itself is an anonymized public dataset of course presentations, students, assessments, registrations, and VLE interactions [1]. The present analysis used the repository's record-level CCC extract rather than reconstructing the official OULAD relational tables. This choice

makes the weekly representation explicit and easy to audit, but it also means that the exact derivation of the weekly aggregates and final_score from the original OULAD tables is not independently reconstructed in this study.

The raw analytic file contained 1,043 records and 25 variables. Learner identifiers were unique (1,043 unique id_student values), no exact duplicate rows were detected, and no learner outcome was imputed. The target was defined operationally as at-risk when final_score < 40 and not-at-risk when final_score ≥ 40. A 40-point cut point is consistent with the Open University undergraduate scale documented in institutional policy [40]; however, because the third-party extract does not identify the original result rule used to construct final_score, this threshold is a study definition rather than a reconstruction of CCC's official historical final_result label. The endpoint should therefore be interpreted strictly as low final-score risk, not as withdrawal or a reconstructed institutional result category.

Table 1. Dataset and analytical-cohort characteristics

Characteristic	Value
Dataset	Public OULAD-derived CCC learner extract
Learner records	1,043
Unique learners	1,043
Raw variables	25
Weekly assessment variables	8 (score_1–score_8)
Weekly VLE-click variables	8 (sum_click_1–sum_click_8)
Outcome	final_score (0–100)
At-risk definition	final_score < 40
At risk	379 (36.3%)
Not at risk	664 (63.7%)
Early prediction window	Weeks 1–4 only
Exact duplicate rows	0

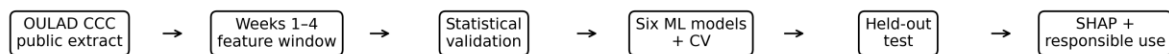


Fig. 1. Empirical workflow from the public OULAD-derived CCC cohort through early-window feature construction, statistical validation, machine learning, held-out evaluation, SHAP explanation, and educational interpretation.

3.2 Data quality, preprocessing, and leakage control

Data quality was audited before feature construction. Seventy-six records (7.29%) had missing IMD band; all other raw variables were complete. Ninety-two imported IMD values had been converted by spreadsheet date formatting into a representation corresponding to the legitimate OULAD 10-20% category; these values were deterministically restored before analysis. Missing IMD values were not filled globally. Within each model-training fold, categorical missingness was imputed with the training-fold mode. Numerical predictors were median-imputed within training and standardized where required. One-hot encoding used unknown-category handling to permit unseen levels in held-out data.

Leakage prevention governed feature construction. Although eight weekly assessment and click measures and `final_score` were present in the source, predictors were restricted to weeks 1-4. `score_5-score_8`, `sum_click_5-sum_click_8`,

and `final_score` were excluded from the predictor matrix. The target was used only for stratification, statistical comparisons, and outcome evaluation. Preprocessing was encapsulated inside scikit-learn pipelines so imputation, scaling, and categorical encoding were learned from training folds only. The held-out test set was untouched by hyperparameter tuning.

The source does not provide event-level timestamps, assessment-release dates, or the transformation used to create each weekly aggregate. The study should therefore be interpreted as an analysis of the first four observed weekly variables in the supplied extract, not as a day-specific real-time deployment simulation. In addition, the extract alone does not establish whether `final_score` is mathematically independent of the component `score_1-score_8` variables. The resulting task is best understood as forecasting a later score-defined outcome from early recorded assessment and engagement information, rather than as evidence that the predictors are independent causal precursors.

Table 2. Analytical variables and engineered early-window features

Feature	Construction (weeks 1-4)	Educational interpretation
<code>early_total_clicks</code>	$\Sigma \text{sum_click_1} \dots \text{sum_click_4}$	Overall early VLE activity volume
<code>early_active_weeks</code>	Count of weeks with clicks >0	Continuity of platform participation
<code>early_click_cv</code>	$\text{SD}(\text{clicks})/\text{mean}(\text{clicks})$	Variability/inconsistency of weekly engagement

click_trend	OLS slope of weekly click totals	Direction of engagement over time
recent_click_share	(week 3 + week 4 clicks)/weeks 1–4 clicks	Recency of engagement within early window
assessment_participation	Fraction of score_1...score_4 >0	Positive-score occurrence proxy; recorded zero is ambiguous
early_mean_score	Mean score_1...score_4 including zeros	Recorded early score level including zero values
early_nonzero_mean_score	Mean of positive early scores	Performance among strictly positive recorded scores
score_trend	OLS slope of score_1...score_4	Direction of early assessment performance
studied_credits	Source variable	Study-load context
Demographic/context	Gender, region, education, IMD, age, disability	Context/audit variables

Table 3. Data preprocessing and quality controls

Issue	Observed	Treatment
Exact duplicates	0	No removal required
Duplicate learner IDs	0	No aggregation required
IMD missingness	76/1,043 (7.29%)	Fold-wise most-frequent imputation
IMD spreadsheet artifact	92 values	Restored to 10-20% after spreadsheet/date-format artifact
Other missing variables	0	No global filling
Outcome imbalance	379/1,043 at risk (36.3%)	Class weighting; held-out distribution preserved
Leakage risk	Weeks 5–8 and final score available in source	Excluded from all predictors
Numeric scaling	Required for LR/SVM	Fit inside training pipeline
Categorical encoding	Nominal variables	Fold-wise one-hot encoding

3.3 Statistical analysis

Exploratory distributions showed zero inflation, bounded assessment variables, ceiling effects in active weeks, and marked right-skew in click counts (early_total_clicks skewness=3.26). Consequently, primary group comparisons used two-sided Mann-Whitney U tests rather than assuming Gaussian residuals. Rank-biserial correlation quantified effect size, with its sign defined as at-risk minus not-at-risk. Median differences were accompanied by non-parametric bootstrap 95% confidence intervals. Spearman correlation assessed monotonic relationships with final_score. Categorical associations used Pearson chi-square tests and Cramér's V; results involving sparse education/age cells were interpreted cautiously. Holm correction controlled family-wise multiplicity within the planned comparisons. The significance threshold was $\alpha=0.05$, but practical interpretation emphasized effect size and confidence intervals rather than significance alone.

3.4 Machine-learning design

Six complementary classifiers were evaluated: regularized logistic regression (transparent linear baseline), decision tree (intrinsically interpretable non-linear baseline), Random Forest (bagged ensemble), support vector machine (kernel classifier), XGBoost, and LightGBM (boosted tree ensembles). Deep learning was not used because the cohort size ($n=1,043$) and tabular feature structure do not provide a clear methodological justification for additional model complexity.

Data were split 80:20 using stratification and random seed 42, producing 834 training and 209

held-out observations. Hyperparameters were selected exclusively within the training set using five-fold StratifiedKFold with shuffling (seed 42) and GridSearchCV. Average precision was the tuning objective because the at-risk group was the minority class; throughout the manuscript, the label PR-AUC refers to this average-precision summary of precision-recall performance rather than trapezoidal integration of the precision-recall curve. Class weighting was applied to logistic regression, decision tree, Random Forest, SVM, and LightGBM; no SMOTE or synthetic observations were used because the 36.3% prevalence represented moderate rather than extreme imbalance. XGBoost used conservative subsampling/column sampling and tuned tree depth/minimum child weight.

Held-out metrics were accuracy, precision, recall/sensitivity, specificity, F1, ROC-AUC, PR-AUC (average precision), Brier score, and the confusion matrix. Five-fold cross-validation summarized mean and standard deviation of discrimination and threshold metrics. Logistic-regression held-out metrics were accompanied by 1,200 bootstrap confidence intervals. Because only one five-fold cross-validation run was available and folds are not statistically independent, no formal claim of pairwise algorithmic superiority is made. The held-out set is used for final descriptive evaluation of all fitted model families, not to establish a statistically selected winner. Brier score is reported as an overall probabilistic accuracy measure, but calibration curves, calibration intercept/slope, and post-hoc recalibration were not evaluated; consequently, model scores should not be interpreted as deployment-ready probabilities.

Table 4. Machine-learning models and optimized hyperparameters

Model	Selected hyperparameters
Logistic Regression	{"model__C": 0.1}
Decision Tree	{"model__max_depth": 5, "model__min_samples_leaf": 15}
Random Forest	{"model__max_depth": null, "model__max_features": "sqrt", "model__min_samples_leaf": 2}

SVM	{"model__C": 0.5, "model__gamma": "scale"}
XGBoost	{"model__max_depth": 2, "model__min_child_weight": 1}
LightGBM	{"model__min_child_samples": 25, "model__num_leaves": 7}

3.5 Explainability, ablation, error, and subgroup analyses

TreeSHAP was applied to Random Forest because it was the leading tree-based model in training-set cross-validation and closely matched the best held-out discrimination. Mean absolute SHAP values provided global importance, while the SHAP summary distribution showed the direction and heterogeneity of feature contributions to the Random-Forest model output. One-hot encoded features were aggregated back to their source variable for global ranking. SHAP values are treated as model-specific attributions, not causal estimates and not explanations of the logistic-regression model.

Feature-family ablation evaluated four prespecified sets under the logistic-regression pipeline: demographic/context only, engagement only, assessment only, and the full early-window set. These analyses compare how much predictive information each family carries when modeled in the reported configurations; they do not by themselves estimate a paired marginal gain from adding engagement to an assessment model. Error analysis compared early behavioral profiles for correct predictions, false negatives, and false positives. Preliminary subgroup performance was reported by gender and age when subgroup sample size was at least 20; smaller disability-positive and oldest-age groups were not treated as stable fairness estimates.

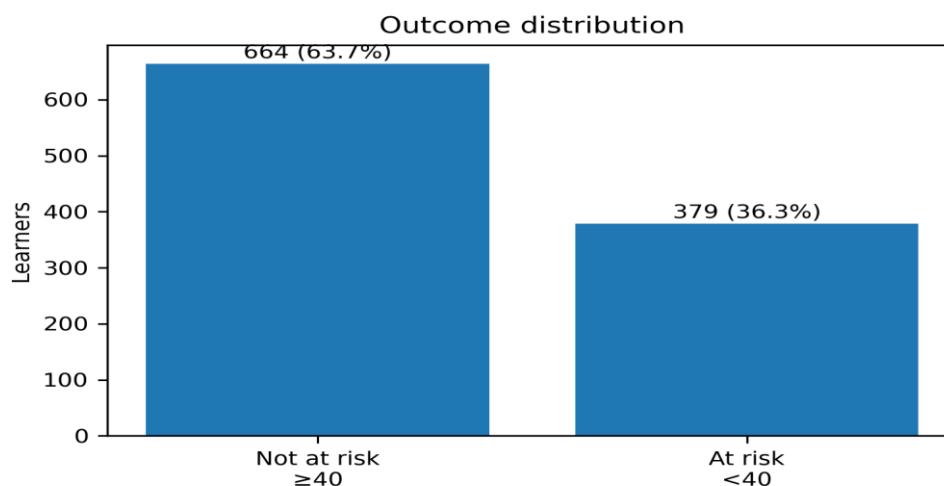


Fig. 2. Outcome-class distribution for the 1,043 learners. The at-risk class (<40 final score) comprises 36.3% of the cohort.



Fig. 3. Data-quality audit. IMD band was the only raw variable with missingness (7.29%); other variables were complete.

4. AI-BASED ANALYSIS

4.1 Prediction task and rationale

The modeling task asks whether a learner whose first four observed weekly variables are known will later have `final_score` < 40. This is an early-window, score-threshold classification problem rather than a withdrawal model. The output is intended as a prioritization signal for additional human review and support. The first-half cut-off balances two competing objectives: earlier decisions preserve intervention time, whereas later windows contain more direct evidence of performance. Because 296 learners had `final_score` = 0 and the extract does not distinguish the reasons for those zero outcomes, the target combines zero and low final performance. No result in this paper implies that changing a feature would causally change a learner's final score.

4.2 Evaluation priorities and imbalance handling

The primary ranking metric was PR-AUC (operationalized as average precision rather than accuracy), which reduces the influence of the majority class and aligns evaluation with the task of retrieving at-risk learners from an imbalanced cohort [28]. Model interpretation also considered ROC-AUC, recall, F1, Brier score, cross-validation stability, explanation burden, and governance. The 0.50 threshold used for confusion-matrix reporting is an illustrative analysis threshold, not an optimized institutional

decision boundary. Because formal calibration was not assessed and class weighting can affect probability interpretation, deployment should use locally validated ranking/threshold procedures and calibration before treating model outputs as probabilities. Synthetic resampling was intentionally avoided to preserve the observed training distribution.

4.3 Reproducibility

The empirical pipeline used Python 3.13.5, pandas 2.2.3, NumPy 2.3.5, SciPy 1.17.0, scikit-learn 1.8.0, XGBoost 3.1.3, LightGBM 4.6.0, SHAP 0.50.0, and Matplotlib 3.10.8. All stochastic model and split operations used `random_state`=42 where supported. Feature definitions, target construction, preprocessing, search spaces, and evaluation criteria were fixed before held-out evaluation. The analyzed learner extract is publicly identified in [39]. Reproducibility is nevertheless constrained by the use of a third-party derived file rather than a reconstruction from the official OULAD relational tables, and by the absence in the manuscript of a versioned derivation script showing how the weekly aggregates and `final_score` were produced from OULAD. These provenance constraints are treated as limitations rather than filled by assumption.

5. Results And Analysis

5.1 Dataset, outcome, and data quality

The final cohort included 1,043 unique learners. The mean final score was 50.11 (SD 36.95,

median 56, interquartile 0-82; range 0-100). A substantial mass occurred at zero: 296 learners (28.4%). Of the 379 learners qualifying for the $\text{final_score} < 40$ risk definition, 296 (78.1%) had $\text{final_score} = 0$. The remaining 664 (63.7%) were not at risk by the study definition. The importance of this concentration is that the extract does not distinguish whether a zero final score represents non-completion, non-submission, or very low assessed performance; the target thus represents low/zero final performance rather than a single homogeneous educational state.

The quality audit found no duplicate learner identifiers and no exact duplicate records. IMD band was missing for 76 learners (7.29%). Because IMD was a contextual categorical variable rather than the primary behavioral signal, fold-wise imputation prevented deletion of otherwise complete learner histories. Figure 3 shows the missingness profile. Deterministic repair of 92 spreadsheet-converted 10-20% IMD labels restored a source category but did not involve statistical imputation.

Table 5. Descriptive statistics for major analytical variables

Variable	Mean	SD	Median	IQR	Range
final_score	50.11	36.95	56.00	0.00–82.00	0.00–100.00
early_total_clicks	721.55	780.35	523.00	266.50–892.00	0.00–7579.00
early_active_weeks	3.64	0.92	4.00	4.00–4.00	0.00–4.00
early_click_cv	0.68	0.38	0.61	0.45–0.83	0.00–1.73
recent_click_share	0.45	0.22	0.49	0.36–0.59	0.00–1.00
assessment_participation	0.77	0.34	1.00	0.50–1.00	0.00–1.00
early_mean_score	56.90	31.38	63.50	34.00–83.75	0.00–100.00
early_nonzero_mean_score	64.29	27.90	71.25	50.96–85.42	0.00–100.00
score_trend	-4.45	10.62	-1.50	-10.05–1.70	-39.40–30.40
studied_credits	69.88	37.59	60.00	60.00–90.00	30.00–630.00

5.2 Learner engagement and assessment patterns

Early engagement differed sharply between outcome groups. At-risk learners produced a median of 246 clicks in weeks 1-4 compared with 703.5 for not-at-risk learners. Engagement was also less sustained within the window: although both groups had a median of four active weeks due to a ceiling effect, at-risk learners averaged 3.02 active weeks versus 3.99 for not-at-risk learners. This shows the click distribution to be

strongly right-skewed and for the at-risk distribution to be shifted toward lower activity. Figure 4.

Temporal structure provided additional information beyond the volume of clicks. At-risk learners had a median click coefficient of variation of 0.802 versus 0.558 for not-at-risk learners, indicating more uneven early engagement. Recent-click share was 0.310 for at-risk learners versus 0.529 for not-at-risk learners,

indicating a smaller proportion of first-four-week activity in weeks 3-4. In contrast, the raw linear click trend produced a small unadjusted difference that failed a Holm correction, showing that a simple slope was less informative than relative recency and consistency.

Assessment behavior diverged more strongly between groups. The median positive-score occurrence across score_1-score_4 was 0.50 in the at-risk group and 1.00 in the not-at-risk group. The median early mean score with

recorded zeros retained was 17.5 versus 77.75. Because the extract does not distinguish an unsubmitted assessment from an attempted assessment recording zero, `assessment_participation` is a positive-score occurrence proxy rather than a verified submission indicator. Even with zeros excluded, the median positive-score performance gap was large: 45.0 versus 80.0, showing that conditional performance differed substantially while leaving the meaning of recorded zeros unresolved.

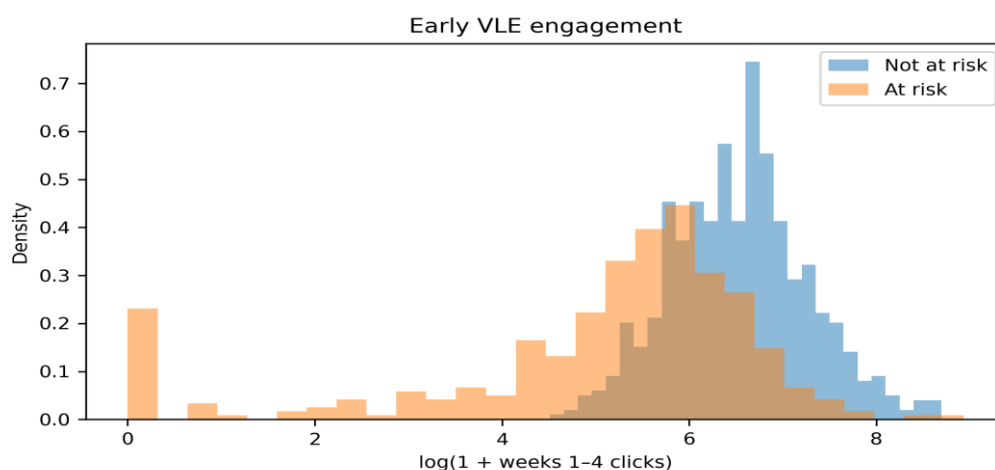


Fig. 4. Distribution of log-transformed weeks-1-4 VLE clicks by outcome group. At-risk learners are concentrated at lower engagement levels.

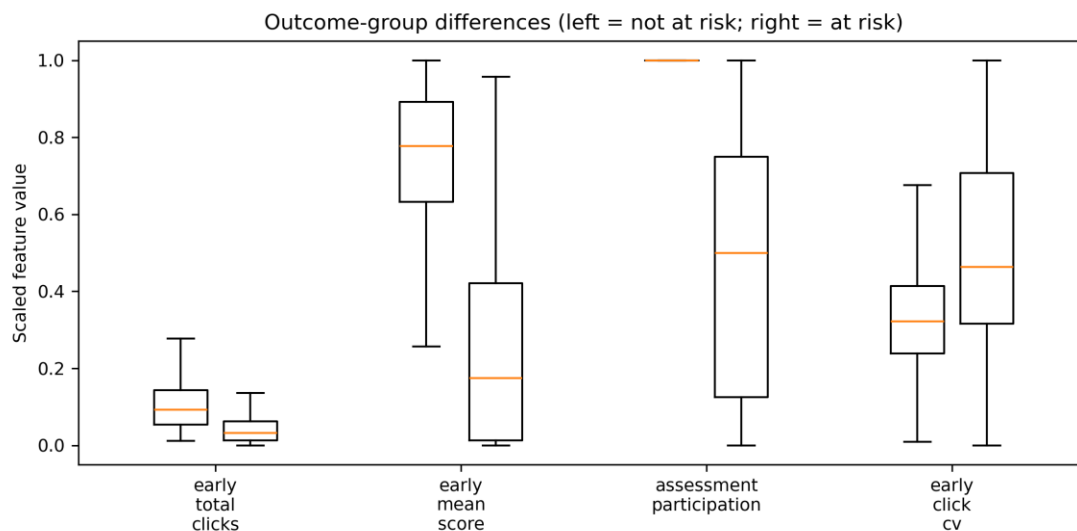


Fig. 5. Scaled behavioral characteristics across outcome groups. The plot highlights lower clicks, lower assessment participation and scores, and higher engagement variability among at-risk learners.

5.3 Statistical associations and practical effect sizes

The largest group difference was early mean assessment score ($U=15,099$; Holm-adjusted $p<0.001$; rank-biserial $r=-0.880$). The median early mean difference was -60.25 points with a 95% bootstrap CI from -63.44 to -56.00. The assessment_participation positive-score proxy was similarly strong ($U=28,602.5$; adjusted $p<0.001$; $r=-0.773$) with a median difference of -0.50 (CI -0.75 to -0.50). Early non-zero mean score produced $r=-0.735$, confirming that performance on positive recorded assessment scores also discriminated risk.

VLE behavior showed moderate-to-large effects. Early total clicks produced $U=48,618$, adjusted $p<0.001$, $r=-0.614$, with a median difference of -457.5 clicks (CI -513.76 to -383.48). Recent-click share yielded $r=-0.527$ and click variability

$r=+0.414$ (higher variability was associated with at-risk status) and score trend $r=-0.400$. Studied credits, click trend did not survive Holm correction, preventing their elevation to substantive predictors on statistical evidence alone.

Spearman analysis with continuous final score converged with the group comparisons. Early mean score had $\rho=0.861$, early non-zero mean score $\rho=0.766$, the assessment_participation positive-score proxy $\rho=0.733$, active weeks $\rho=0.523$, total clicks $p<0.0010$, and recent-click share $p<0.0011$ (all adjusted $p<0.0012$). Early click variability was negatively associated with final score ($\rho=-0.336$). These are statistical associations within an observational dataset, not estimates of causal effect.

Table 6. Statistical association and group-comparison results (Holm-adjusted)

Variable	Test	Statistic	Adj. p	Effect	Risk vs non-risk median
early_total_clicks	Mann-Whitney U	48618.00	<0.001	$r=-0.614$	246.00 vs 703.50
early_click_cv	Mann-Whitney U	177905.00	<0.001	$r=0.414$	0.80 vs 0.56
recent_click_share	Mann-Whitney U	59579.00	<0.001	$r=-0.527$	0.31 vs 0.53
assessment_participation	Mann-Whitney U	28602.50	<0.001	$r=-0.773$	0.50 vs 1.00
early_mean_score	Mann-Whitney U	15099.00	<0.001	$r=-0.880$	17.50 vs 77.75
early_nonzero_mean_score	Mann-Whitney U	33329.50	<0.001	$r=-0.735$	45.00 vs 80.00
score_trend	Mann-Whitney U	75537.50	<0.001	$r=-0.400$	-7.00 vs -0.50
studied_credits	Mann-Whitney U	135800.00	0.071	$r=0.079$	60.00 vs 60.00

gender	Chi-square	9.09	0.013	V=0.093	—
region	Chi-square	15.78	0.403	V=0.123	—
imd_band	Chi-square	31.86	0.003	V=0.175	—
disability	Chi-square	0.62	0.430	V=0.024	—

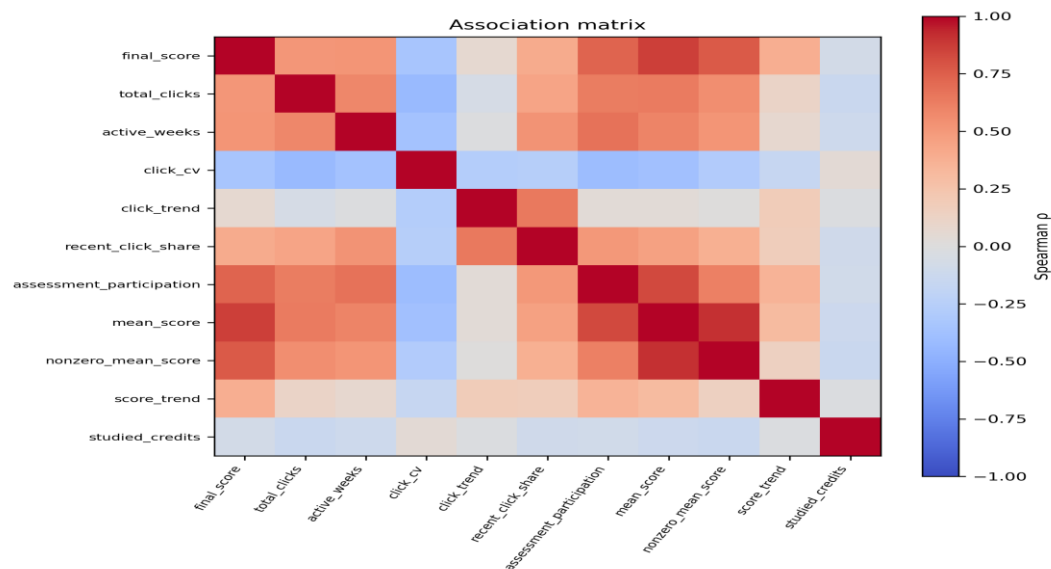


Fig. 6. Spearman association matrix for early behavioral, assessment, workload, and final-score variables.

5.4 Predictive modeling and held-out performance

All six models achieved useful discrimination, but ordering varied by metric. Logistic regression had the highest held-out PR-AUC (0.915) and ROC-AUC (0.936), while SVM had the highest accuracy (0.861) and F1 (0.805). Random Forest closely matched logistic regression in PR-AUC (0.914) and ROC-AUC (0.934) and had the lowest Brier score (0.0969). Decision Tree had the lowest discrimination (PR-AUC 0.851; ROC-AUC 0.904). Because these are single held-out estimates from model families tuned on the training cohort, small differences should be interpreted descriptively rather than as proof that one algorithm is objectively superior.

No model family is declared the statistically selected winner from the held-out comparison. The logistic regression is used in subsequent threshold and error analyses as a transparent

reference model because it combines competitive discrimination with a simpler decision structure. Random Forest and SVM are similarly plausible alternatives on the reported metrics. This framing avoids using the held-out set to make a strong post-hoc model-selection claim and focuses the comparison on practical equivalence, uncertainty, and interpretability.

For the logistic-regression reference model, held-out PR-AUC was 0.915 (95% bootstrap CI 0.869-0.955) and ROC-AUC 0.936 (0.901-0.968). Accuracy was 0.847 (0.794-0.895), recall 0.789 (0.693-0.877), specificity 0.880 (0.820-0.935), precision 0.789 (0.694-0.880), and F1 0.789 (0.712-0.857). These intervals quantify the sampling uncertainty in this held-out cohort. They do not account for uncertainty from choosing among model families nor do they establish probability calibration.

Table 7. Held-out model performance

Model	Acc.	Prec.	Recall	Spec.	F1	ROC-AUC	PR-AUC	Brier
Logistic Regression	0.847	0.789	0.789	0.880	0.789	0.936	0.915	0.098
Random Forest	0.842	0.787	0.776	0.880	0.781	0.934	0.914	0.097
XGBoost	0.847	0.824	0.737	0.910	0.778	0.928	0.906	0.108
LightGBM	0.842	0.803	0.750	0.895	0.776	0.923	0.899	0.115
SVM	0.861	0.822	0.789	0.902	0.805	0.924	0.889	0.105
Decision Tree	0.842	0.787	0.776	0.880	0.781	0.904	0.851	0.125

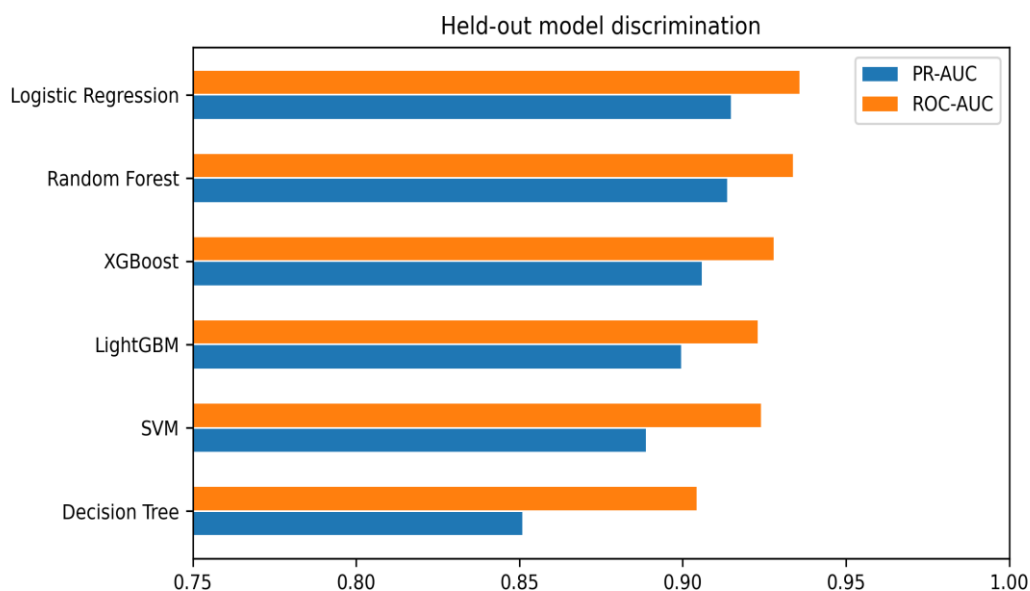


Fig. 7. Held-out discrimination. Logistic regression had the highest PR-AUC and ROC-AUC; SVM had the highest thresholded accuracy/F1.

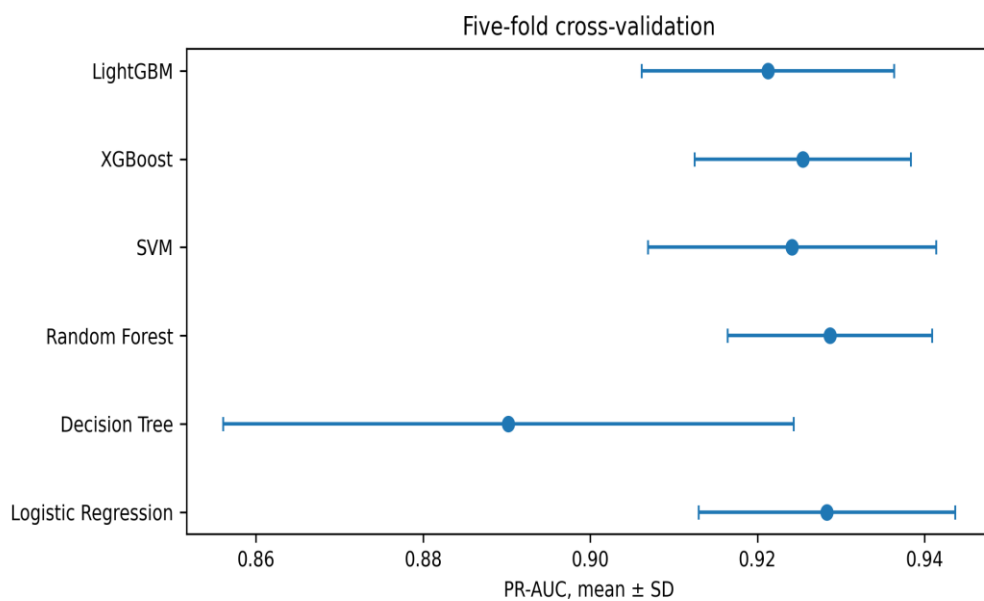
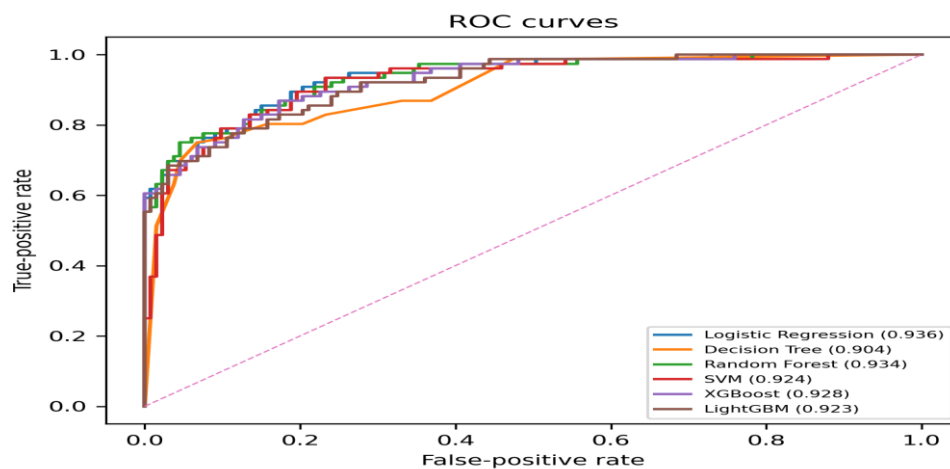
5.5 Cross-validation robustness

Cross-validation confirmed that the leading models were closely grouped. Logistic regression PR-AUC was 0.928 ± 0.015 , Random Forest 0.929 ± 0.012 , XGBoost 0.925 ± 0.013 , SVM 0.924 ± 0.017 , and LightGBM 0.921 ± 0.015 . Decision Tree was lower at 0.890 ± 0.034 . XGBoost achieved the highest cross-validated

accuracy (0.894 ± 0.009) and F1 (0.845 ± 0.019), while Random Forest had the numerically highest cross-validated PR-AUC by only 0.001 over logistic regression. The mean and standard deviation overlap supports a restrained conclusion: for this engineered early-window feature set, added algorithmic complexity did not display a clear practical discrimination advantage over simpler regularized logistic regression.

Table 8. Five-fold cross-validation performance on the training cohort

Model	PR-AUC	ROC-AUC	F1	Accuracy
Logistic Regression	0.928±0.015	0.942±0.008	0.827±0.012	0.874±0.009
Random Forest	0.929±0.012	0.942±0.015	0.839±0.022	0.889±0.012
XGBoost	0.925±0.013	0.940±0.013	0.845±0.019	0.894±0.009
LightGBM	0.921±0.015	0.935±0.016	0.830±0.027	0.879±0.020
SVM	0.924±0.017	0.939±0.010	0.834±0.012	0.881±0.009
Decision Tree	0.890±0.034	0.924±0.021	0.816±0.035	0.867±0.025

**Fig. 8. Five-fold PR-AUC mean $\pm$ SD. The leading models overlap strongly, indicating limited practical separation in discrimination.****Fig. 9. Held-out ROC curves for all six classifiers.**

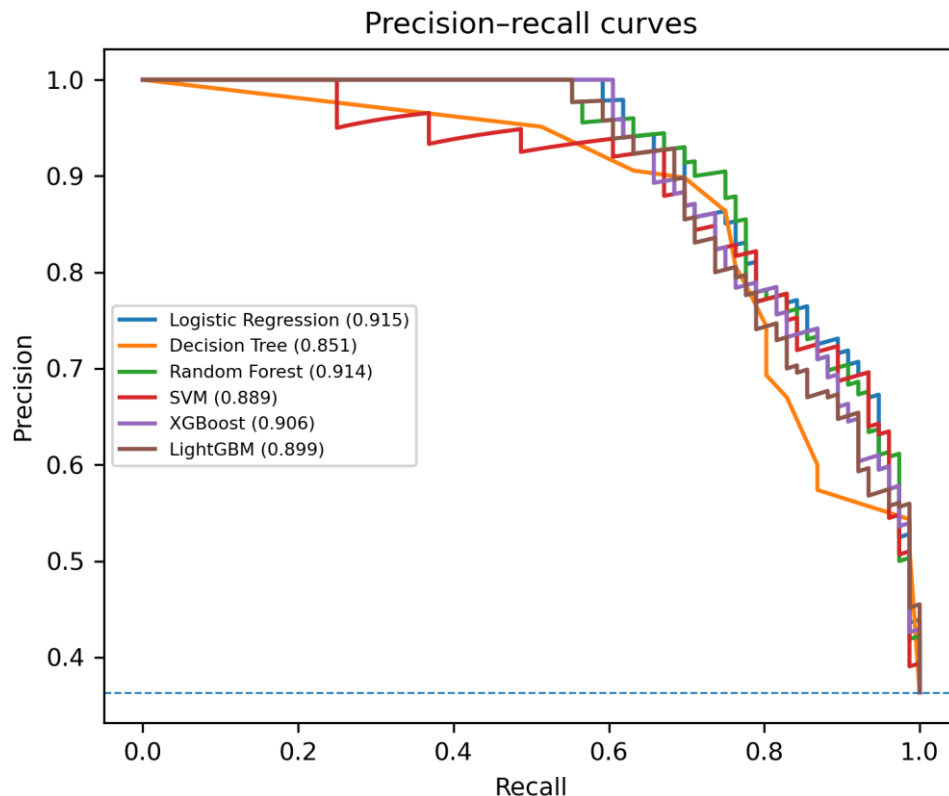


Fig. 10. Held-out precision–recall curves. PR-AUC was the primary ranking metric because the at-risk group was the minority class.

5.6 Confusion-matrix and error analysis

At the illustrative 0.50 threshold, the logistic-regression reference model correctly identified 60 of 76 at-risk learners and 117 of 133 not-at-risk learners. It produced 16 false negatives and 16 false positives (Fig. 11). The symmetry in counts does not imply symmetry in educational cost. A false negative represents a learner who may not receive proactive support, while a false positive can consume tutor capacity and may create unnecessary concern if communicated poorly. Because the model was class-weighted and formal calibration was not assessed, the 0.50 cutoff should not be interpreted as a universally meaningful 50% risk threshold.

Error profiles clarify why purely automated decisions would be inappropriate. False negatives had a median early mean score of 59.5, assessment participation of 1.00, and recent-click share of 0.540. In other words, many missed at-risk learners looked comparatively engaged and academically viable early on. Their median

studied credits were 75, compared with 60 among correct predictions. By contrast, false positives had median early mean score 44.25, assessment participation 0.75, and a strongly negative score trend of -13.85. These learners displayed plausible early-warning signals but ultimately reached the non-risk outcome. The pattern indicates that a warning system should trigger review and supportive contact, not deterministic classification of learner ability.

The bootstrap intervals reinforce this interpretation: even with strong discrimination, recall could range plausibly from 0.693 to 0.877 in samples of this size and thus institutions should design capacity and follow-up processes around uncertainty and periodically reassess thresholds as cohorts change rather than hard-code a universal decision boundary. If decisions require calibrated probabilities rather than ranking scores, calibration should be assessed and, where necessary, performed using data separate from final evaluation.

Table 9. Confusion-matrix and uncertainty summary for logistic regression

Measure	Estimate	95% bootstrap CI
accuracy	0.847	0.794–0.895
precision	0.789	0.694–0.880
recall	0.789	0.693–0.877
specificity	0.880	0.820–0.935
f1	0.789	0.712–0.857
roc_auc	0.936	0.901–0.968
pr_auc	0.915	0.869–0.955

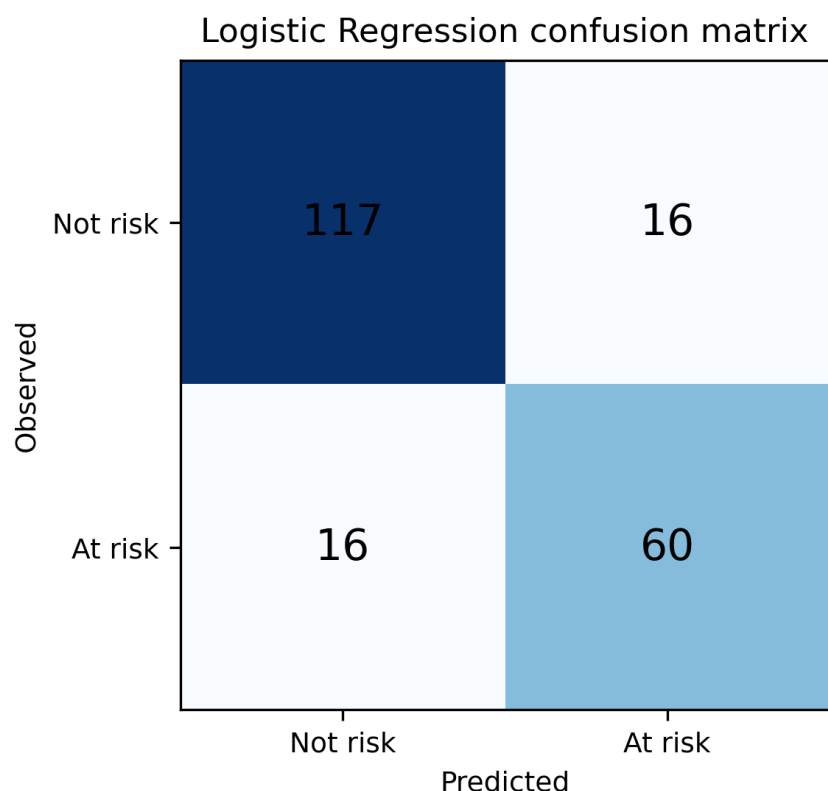


Fig. 11. Held-out confusion matrix for the logistic-regression reference model at the illustrative 0.50 threshold (n=209).

5.7 Feature ablation and explainable-AI results

Feature-family ablation shows where predictive information was concentrated when each prespecified family was modeled in the reported analysis. Demographic/context variables alone produced PR-AUC 0.515 ± 0.045 (only somewhat above the at-risk prevalence baseline). Engagement-only features produced PR-AUC

0.813 ± 0.053 , showing that early VLE behavior carries substantial standalone signal even without assessment scores. Assessment-only features produced PR-AUC 0.924 ± 0.036 , while the full early-window feature set produced 0.925 ± 0.031 . These point estimates do not support a large aggregate discrimination gain from adding the reported engagement/context features once assessment features are available, but because no

paired confidence interval was computed for the assessment-only versus full model difference and no one-variable rule such as early mean score alone was benchmarked, the incremental value of the full machine-learning specification over a very simple assessment rule remains unquantified. Table 10 should therefore be interpreted not as numerically paired with the separate model-comparison summary in Table 8 but as a self-contained feature-family comparison.

Random Forest was used for SHAP because it was the leading tree-based model in training set cross-validation (PR-AUC 0.929 ± 0.012) and remained closely competitive on the held-out set (PR-AUC 0.914). Global SHAP importance ranked `early_mean_score` first (mean $|\text{SHAP}| = 0.133$), then `assessment_participation` (0.082), `early_nonzero_mean_score` (0.078), `score_trend` (0.044), `recent_click_share` (0.031), `early_total_clicks` (0.029), and `early_active_weeks` (0.025). Aggregated

demographic/context variables were substantially smaller: IMD band 0.0128, region 0.0120, highest education 0.0111, gender 0.0061, age band 0.0049, and disability 0.0035. These attributions explain the fitted Random Forest, not the logistic-regression reference model.

The SHAP distribution supplied direction as well as rank: high early mean scores, more positive early assessment scores, stronger conditional assessment performance, greater recent-click share, more clicks, and more active weeks generally shifted the Random Forest output toward lower predicted risk, while higher click variability shifted it toward higher risk. These model-based directions agree with the non-parametric group comparisons and Spearman correlations, providing triangulation across analytical methods. The agreement reinforces the descriptive interpretation but does not convert these variables into causal levers or calibrated intervention effects.

Table 10. Feature-family ablation with five-fold cross-validation

Feature set	PR-AUC	ROC-AUC	F1
demographic_context	0.515 ± 0.045	0.648 ± 0.019	0.540 ± 0.031
engagement_only	0.813 ± 0.053	0.832 ± 0.043	0.685 ± 0.057
assessment_only	0.924 ± 0.036	0.943 ± 0.022	0.823 ± 0.058
full_early_window	0.925 ± 0.031	0.941 ± 0.022	0.826 ± 0.059

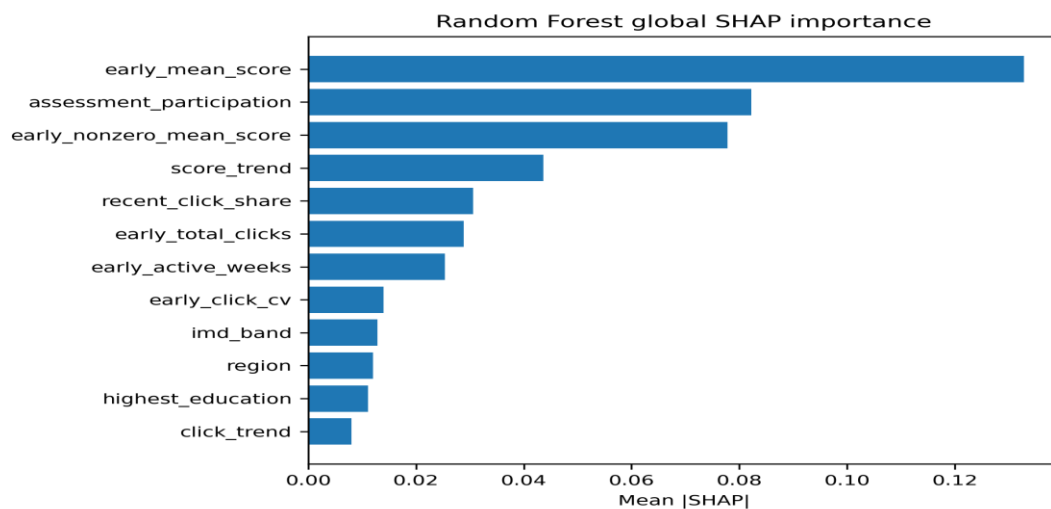


Fig. 12. Random-Forest global SHAP importance aggregated to source features.

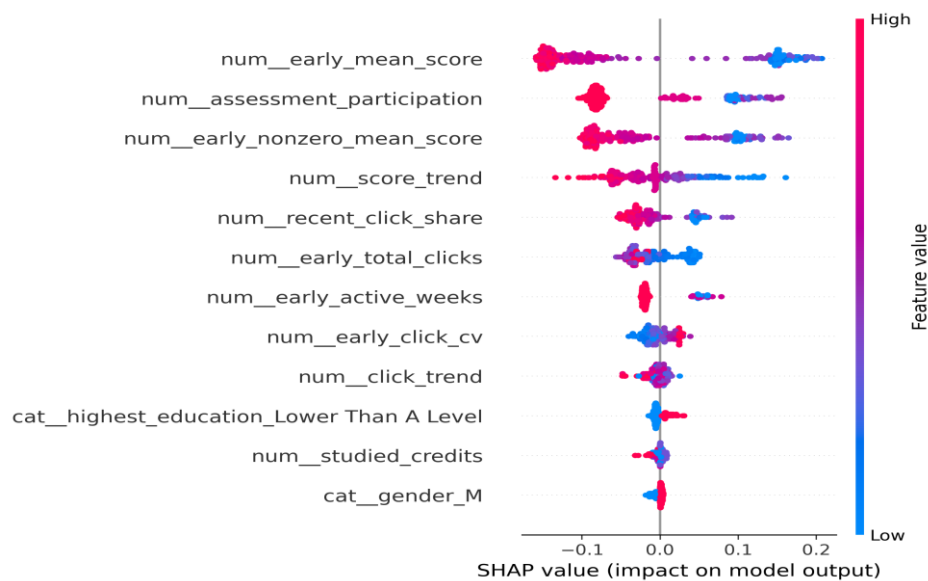


Fig. 13. SHAP summary for the Random-Forest explanation model. Color indicates feature value; horizontal position indicates contribution to the model output for the at-risk class.

5.8 Preliminary subgroup performance analysis

Held-out discrimination was similar by gender: ROC-AUC was 0.930 for women ($p < 0.0013$) and 0.936 for men ($p < 0.0014$), with PR-AUC 0.904 and 0.918, respectively. Threshold recall was 0.833 for women and 0.776 for men; false-positive rates were 0.132 and 0.116. These differences are descriptive only. The subgroup samples are small, confidence intervals were not computed for these metrics, and thus the analysis cannot establish parity or absence of disparity.

Age showed a larger threshold-performance contrast: learners aged 0-35 ($p < 0.0015$) had recall 0.810 and false-positive rate 0.160, whereas those aged 35-55 ($p < 0.0016$) had recall 0.692 and false-positive rate 0.058. Their risk prevalences also differed (0.438 vs. 0.200), so threshold metrics cannot be interpreted independently of base rates. Disability-positive

and oldest-age subgroups were too small in the held-out split to support stable estimates. The appropriate conclusion is not that the model is fair but that these descriptive signals identify groups requiring larger-sample monitoring, uncertainty intervals, and threshold review before deployment.

Ablation provides an additional governance insight: demographic/context-only prediction was weak while assessment and engagement carried most of the useful signal. This reduces the operational argument for reliance on sensitive context variables. Institutions could evaluate models that reserve protected characteristic variables for auditing rather than prediction, provided discrimination and calibration remain adequate. Such a design decision must be tested locally; removal of protected variables does not by itself guarantee fairness.

Table 11. Held-out subgroup performance for sufficiently represented groups (descriptive; no subgroup confidence intervals)

Attribute	Group	n	Prevalence	Recall	FPR	ROC-AUC	PR-AUC
gender	F	56	0.321	0.833	0.132	0.930	0.904
gender	M	153	0.379	0.776	0.116	0.936	0.918

age_band	0-35	144	0.438	0.810	0.160	0.932	0.927
age_band	35-55	65	0.200	0.692	0.058	0.922	0.862

6. Discussion

This study answers RQ1 by showing that early assessment behavior and VLE engagement are strongly associated with the score-defined at-risk outcome, but the magnitude of association differs materially across feature types. Assessment-related variables and early score level were dominant signals. The effect for early mean score (rank-biserial $r=-0.880$) was substantially larger than the effect for total clicks ($r=-0.614$), while click variability and recent activity retained moderate effects. The practical interpretation is not that clicking causes success but that sustained and recent engagement are useful behavioral markers, whereas direct evidence of recorded assessment performance is closer to the score-defined endpoint and thus more discriminative. Because 78.1% of the at-risk group had final_score=0, these associations should be interpreted against a target that is heavily influenced by zero final outcomes.

RQ2 is answered by strong but imperfect early-window discrimination. Logistic regression achieved ROC-AUC 0.936 and PR-AUC 0.915 using only weeks 1-4, SVM a slightly higher thresholded F1 and accuracy, and Random Forest nearly matching logistic regression in held-out ranking metrics while having the numerically highest training cross-validated PR-AUC. The leading cross-validation means and standard deviations closely overlap, and thus the evidence does not support a strong superiority claim for any one algorithm. The substantive result is that a regularized linear model can be competitive with kernel and ensemble methods for this engineered feature set, supporting model choice based on discrimination, stability, explanation burden, calibration, and governance rather than algorithmic complexity alone.

The comparison aligns with the broader OULAD literature, where many algorithm families can reach strong predictive performance depending

on target and window [11]. It also complements recent deployment-oriented OULAD work by Wu et al., who emphasize calibration, fairness, thresholding, and capacity-aware decisions rather than a leaderboard mentality [37]. The present results add a narrower empirical point: when early assessment participation-like information and performance is engineered explicitly, a logistic model can be competitive with modern boosting methods. However, the study did not benchmark a one-variable rule such as early mean score alone, and thus it cannot claim that multivariable machine learning adds substantial discrimination beyond a simple assessment-based heuristic. Simpler rules should therefore be mandatory comparators in future validation.

RQ3 concerns what the models use. SHAP and feature-family ablation both show that assessment-related information dominates. The assessment-only feature family nearly matched the full model in the separate ablation analysis, whereas engagement-only features remained useful at PR-AUC 0.813. This supports two cautious implications. First, institutions with delayed or sparse assessment schedules may still obtain useful early-warning ranking information from VLE engagement, albeit at reduced discrimination. Second, where early assessments exist, their recorded scores and positive-score occurrence should be central to review. The present ablation does not demonstrate a large incremental PR-AUC gain from adding engagement after assessment information is available. Banihashem et al. show that learning analytics can strengthen feedback practices when analytic information is integrated into meaningful feedback processes rather than simply visualized [32]. Here, a low positive-score occurrence signal could prompt a check for submission or access barriers, while a declining score trajectory could prompt formative academic support.

RQ4 is addressed by convergence and contradiction across methods. Mann-Whitney

effects, Spearman correlations, feature ablation, and SHAP consistently highlighted assessment participation/performance, click volume/recency, and engagement variability. At the same time, model errors reveal why individual decisions cannot be reduced to these group patterns. False negatives often had strong early participation and reasonable scores yet ultimately became at risk, suggesting latent factors not visible in the trace data. Hlosta et al. similarly show that explaining algorithmic errors can reveal circumstances beyond logged behavior [10]. False positives showed weaker early trajectories but subsequently succeeded, which may reflect recovery, delayed support, or idiosyncratic learning strategies. These cases are precisely where human review adds value.

RQ5 concerns translation into practice. The empirical evidence supports a layered workflow: collect only the early data needed for the decision; compute a risk-ranking score; expose model-appropriate observable reasons; route the case to a tutor or support professional; select a proportionate intervention; and monitor response over time. If a probability is required for policy decisions, calibration should be demonstrated on appropriate local data. The model should not automatically label a learner as incapable, restrict access, or trigger punitive consequences. Human-centered learning-analytics research emphasizes that end users should shape the balance between automation and human control [16]. The present error analysis provides a concrete rationale for that principle.

The subgroup performance analysis should be read as a governance diagnostic rather than a certification of fairness. Similar gender ROC-AUC values are reassuring only in a limited descriptive sense and do not establish equal treatment, particularly because threshold error rates can vary even when ranking metrics are close. Age-group recall differed more substantially, small disability/older groups prevented stable estimates, and subgroup confidence intervals were not computed. Baker and Hawn's analysis of algorithmic bias in education warns that bias can emerge from the

measurement and decision pipeline even when sensitive features are omitted [18]. Accordingly, institutions should audit model behavior by relevant groups, quantify uncertainty, and test whether interventions themselves are equitably accessible and effective.

Several limitations constrain the conclusions. First, the present study evaluated only one OULAD-derived CCC cohort rather than all OULAD modules/presentations. Second, the source is a third-party learner-level weekly extract rather than a reconstruction from the official OULAD relational tables, which means that neither the timing of each event nor the timing of assessment release dates is independently verified, nor is the derivation of `final_score` fully characterized. This implies that the degree of mathematical overlap between early component scores and `final_score` cannot be characterized from the present manuscript alone. Thirdly, the endpoint `final_score < 40` is dominated by zero outcomes: 296 of 379 at-risk learners (78.1%) had `final_score = 0`. Since no sensitivity analysis distinguishing zero scores from scores 1-39 was reported, the model should be considered to predict a composite low/zero final-performance state rather than a homogenous failure mechanism. Fourth, the presence of positive-score occurrence cannot distinguish between non-submission and attempted zero mark in the supplied extract, and the article does not independently document all zero-denominator conventions for derived ratio features. Moreover, the ablation analysis does not report a paired uncertainty estimate for assessment-only versus full features, and no one-variable assessment baseline was evaluated. Furthermore, although Brier score was reported, formal calibration curves, calibration slope/intercept, and recalibration were not assessed, so the 0.50 threshold is illustrative rather than optimized for deployment. The random stratified hold-out estimates within-cohort generalization only: stronger temporal or institutional claims require external presentation- or module-level validation. Finally, the data are observational and historical: neither statistical effects nor SHAP attributions identify causal intervention targets.

Future research should focus on target validation, transportability, and intervention evaluation rather than adding new algorithms. Immediate extensions could include sensitivity analyses distinguishing $\text{final_score}=0$ from scores 1-39 and comparing the score-defined endpoint with OULAD final_result wherever reconstructable, verification of the derived extract against official OULAD tables, simple one-variable and rule-based assessment baselines, fixed-fold or nested resampling for model and ablation comparison, calibration curves and local recalibration, validation across modules and presentations, prospective threshold selection under tutor capacity constraints, privacy-preserving feature minimization, larger subgroup analyses with uncertainty intervals, and randomized or quasi-experimental evaluation of support triggered by risk signals. Such work will advance the field from within-cohort predictive performance to evidence that AI-supported decisions are valid, transportable, fair, and beneficial to learners.

7. Opportunities And Challenges

7.1 Empirically supported opportunities

Early identification is the foremost opportunity supported by the data. At the illustrative 0.50 threshold, the logistic-regression reference model retrieved 78.9% of held-out at-risk learners with 88.0% specificity. The model can identify a cohort for further review before knowing the final outcome. However, since formal calibration was not established, institutions should initially consider the model output to be a ranking score rather than an actual probability and set thresholds only after local validation against tutor capacity and consideration of the relative cost of missing versus unnecessary outreach.

Personalized support can be reason-informed rather than score-only. Random-Forest SHAP reveals which recorded features drove that model's output, such as early scores, positive-score occurrence, score trajectory, recent engagement, and click variability. Those patterns could suggest different questions for a tutor: whether assessments were submitted, if there were access or scheduling barriers, if assessment material was understood, or if study-planning

support is needed. The explanation should drive inquiry rather than dictate a causal remedy, and the explanation method should match the model used operationally.

Institutional decision support is another opportunity. The feature-family analysis shows that assessment variables carry most of the predictive information, while VLE behavior has substantial standalone signal and may be especially useful before graded evidence accumulates. The study does not show a large incremental discrimination gain from adding engagement once assessment features are available, so collection of additional trace data should be dictated by operational need rather than assumed value. Human-centered analytics frameworks explicitly emphasize translating analytic results into context-sensitive actions [33].

7.2 Challenges and safeguards

Privacy and data governance are primary constraints, since longitudinal traces can reveal patterns of participation, performance, socioeconomic context, and disability. Data minimization is preferable to collecting every available variable. Weak demographic-only performance and small SHAP contribution of sensitive context features suggests that institutions should test whether those variables can be excluded from prediction and retained only for fairness auditing. Access controls, purpose limitation, retention schedules, and transparent learner notices should accompany any production system [19].

Fairness cannot be reduced to a single parity statistic: different group prevalences and sample sizes alter precision, recall, and false-positive rates. Small groups may be precisely those for whom an institution most needs protection, but their metrics are the least statistically stable. The present subgroup results are descriptive because confidence intervals were not reported: responsible deployment should use uncertainty intervals, minimum subgroup-size rules, repeated audits, and qualitative investigation of disparities rather than a declaration of fairness from one split [18].

Explainability has limits: SHAP can show how the model formed a prediction but cannot establish why the learner behaved that way or what intervention will work. Explanations should be framed as observable evidence (for example, 'few assessment attempts in weeks 1-4') rather than psychological diagnoses ('low motivation'). Khosravi et al. emphasize that explanations in education must be aligned with stakeholder needs and context [12].

Generalizability, calibration, and data drift remain substantial challenges. ODL modules differ in assessment schedules, learning design, learner population, and VLE structure. A threshold tuned on one course may produce unacceptable workloads or error rates elsewhere, and a well-ranked model can still give poorly calibrated probabilities. Models should undergo course-level validation, calibration assessment, threshold review, and monitored re-evaluation before extension. Predictive learning analytics at

institutional scale also requires tutor engagement and organizational workflows, not just software integration [8].

Digital inequality creates another interpretive risk: lower platform activity may reflect device constraints, connectivity, disability-related access patterns, work schedules, or legitimate offline study rather than disengagement. Since the present engagement-only model is informative but weaker than assessment-based prediction, institutions should avoid using click counts as behavioral surveillance. Contact initiated from a risk flag should be supportive and allow the learner to contextualize the trace.

Table 12 relates these opportunities and challenges directly to the empirical evidence. The central design principle is proportionality: stronger within-cohort predictive evidence can justify prioritized human review but not stronger automation or unvalidated probability-based decisions.

Table 12. AI opportunities, empirical evidence, and ODL implementation implications

Opportunity/challenge	Empirical evidence in this study	Implication
Early warning	LR recall 0.789 at illustrative 0.50 threshold	Use risk ranking for timely tutor review; validate thresholds locally
Assessment support	Positive-score occurrence $r=-0.773$; early score $r=-0.880$	Check non-submission/access barriers separately from low recorded performance
Engagement monitoring	Engagement-only PR-AUC 0.813; full-model increment not established	Use as contextual/early signal; justify added trace collection
Personalization	Random-Forest SHAP identifies model-specific attribution patterns	Present model-matched reason codes; avoid causal wording
Model simplicity	Logistic regression remained competitive with complex models	Compare with simple models and rules before adding complexity
False negatives	16; often showed strong early participation	Do not use AI as sole gatekeeper for support

Opportunity/challenge	Empirical evidence in this study	Implication
False positives	16; often showed declining early score trajectory	Use low-cost supportive outreach rather than punitive action
Subgroup performance	Age/gender threshold metrics differed; small groups unstable	Quantify uncertainty and audit disparities repeatedly
Privacy	Demographic-only PR-AUC 0.515	Minimize sensitive inputs; retain protected data for auditing where lawful
Generalization	Single OULAD-derived CCC cohort	Validate by module/presentation and under cohort shift before deployment
Outcome validity	296/379 at-risk learners had final_score=0	Treat target as low/zero final-performance proxy; run outcome sensitivity analyses
Calibration	Brier score reported; formal calibration not assessed	Calibrate and validate before interpreting scores as probabilities

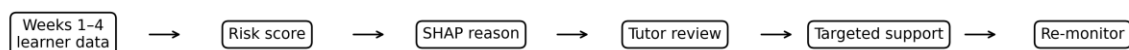


Fig. 14. Evidence-based AI-enabled ODL intervention framework: early learner data -> risk-ranking score -> model-appropriate explanation/reason code -> tutor review -> targeted support -> re-monitoring. SHAP is one model-specific explanation option rather than a universal explanation layer.

8. Conclusion

This empirical study shows that the first four recorded weekly assessment and engagement variables in an OULAD-derived CCC extract contain strong within-cohort signal for a score-defined low/zero final-performance outcome without using weeks 5-8 or final_score as predictors. At-risk learners had substantially lower VLE activity, lower recent engagement, fewer positive early assessment scores, lower early scores, and more variable engagement. Assessment-related information was the strongest signal: early mean score, positive-score

occurrence, conditional positive-score performance, and score trajectory dominated both statistical effects and Random-Forest SHAP attribution. Engagement was still useful as a standalone signal, while demographic/context variables contributed comparatively little predictive information.

Machine-learning results showed strong but imperfect discrimination: logistic regression achieved held-out ROC-AUC 0.936 and PR-AUC 0.915, while Random Forest and SVM were similarly competitive on other metrics, and the study does not claim that logistic regression is

statistically superior. Logistic regression was used as a transparent reference case for error and threshold analysis, and Random Forest was used for TreeSHAP explanation. Error analysis showed that some genuinely at-risk learners looked successful early and some flagged learners later reached the non-risk outcome, establishing a clear role for human review. Preliminary subgroup analysis provided monitoring signals rather than evidence sufficient to certify fairness. Formal calibration, a simple one-variable assessment baseline, and zero-score sensitivity analyses were not established in the reported pipeline.

The practical implication is that AI can most responsibly support ODL when it narrows attention, explains observable evidence, and enables timely, proportionate learner support. Institutions should validate outcome definitions, compare against simple rules, assess calibration before probability-based decisions, use transparent reason codes that match the operational model, maintain tutor oversight, minimize unnecessary sensitive data, quantify subgroup uncertainty, and validate locally before deployment. The value of AI in ODL therefore lies less in automating educational judgment than in making large-scale learner-support decisions earlier, more evidence-informed, and more accountable.

References

- J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics dataset," *Scientific Data*, vol. 4, 170171, 2017. doi:10.1038/sdata.2017.171.
- The Open University, "Open University Learning Analytics Dataset: Data description," *OU Analyse*, accessed Oct. 5, 2026.
- H. Crompton and D. Burke, "Artificial intelligence in higher education: the state of the field," *International Journal of Educational Technology in Higher Education*, vol. 20, art. 22, 2023. doi:10.1186/s41239-023-00392-8.
- M. Bond et al., "A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour," *International Journal of Educational Technology in Higher Education*, vol. 21, art. 4, 2024. doi:10.1186/s41239-023-00436-z.
- C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, e1355, 2020. doi:10.1002/widm.1355.
- D. Ifenthaler and J. Y.-K. Yau, "Utilising learning analytics to support study success in higher education: a systematic review," *Educational Technology Research and Development*, vol. 68, pp. 1961–1990, 2020. doi:10.1007/s11423-020-09788-z.
- C. Herodotou et al., "The scalable implementation of predictive learning analytics at a distance learning university: Insights from a longitudinal case study," *The Internet and Higher Education*, vol. 45, 100725, 2020. doi:10.1016/j.iheduc.2020.100725.
- C. Herodotou, C. Maguire, N. McDowell, M. Hlosta, and A. Boroowa, "The engagement of university teachers with predictive learning analytics," *Computers & Education*, vol. 173, 104285, 2021. doi:10.1016/j.compedu.2021.104285.
- C. Herodotou, G. Naydenova, A. Boroowa, A. Gilmour, and B. Rienties, "How can predictive learning analytics and motivational interventions increase student retention and enhance administrative support in distance education?" *Journal of Learning Analytics*, vol. 7, no. 2, pp. 72–83, 2020. doi:10.18608/jla.2020.72.4.
- M. Hlosta, C. Herodotou, T. Papathoma, 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, 100108, 2022. doi:10.1016/j.caeai.2022.100108.
- L. Jin, Y. Wang, H. Song, and H.-J. So, "Predictive modelling with the Open University Learning Analytics Dataset (OULAD): A systematic literature review," in *Artificial Intelligence in Education: AIED 2024*, CCIS 2150, pp. 477–484, 2024. doi:10.1007/978-3-031-64315-6_46.
- H. Khosravi et al., "Explainable Artificial Intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 3, 100074, 2022. doi:10.1016/j.caeai.2022.100074.

- M. Nagy and R. Molontay, "Interpretable dropout prediction: Towards XAI-based personalized intervention," *International Journal of Artificial Intelligence in Education*, vol. 34, pp. 274–300, 2024. doi:10.1007/s40593-023-00331-8.
- D. Bañeres, M. E. Rodríguez-González, A.-E. Guerrero-Roldán, and P. Cortadas, "An early warning system to identify and intervene online dropout learners," *International Journal of Educational Technology in Higher Education*, vol. 20, art. 3, 2023. doi:10.1186/s41239-022-00371-5.
- K. Alalawi, R. Athauda, and R. Chiong, "An extended learning analytics framework integrating machine learning and pedagogical approaches for student performance prediction and intervention," *International Journal of Artificial Intelligence in Education*, vol. 35, pp. 1239–1287, 2025. doi:10.1007/s40593-024-00429-7.
- R. Alfredo et al., "Human-centred learning analytics and AI in education: A systematic literature review," *Computers and Education: Artificial Intelligence*, vol. 6, 100215, 2024. doi:10.1016/j.caeai.2024.100215.
- I. Rets, C. Herodotou, and A. Gillespie, "Six practical recommendations enabling ethical use of predictive learning analytics in distance education," *Journal of Learning Analytics*, vol. 10, no. 1, pp. 149–167, 2023. doi:10.18608/jla.2023.7743.
- R. S. Baker and A. Hawn, "Algorithmic bias in education," *International Journal of Artificial Intelligence in Education*, vol. 32, pp. 1052–1092, 2022. doi:10.1007/s40593-021-00285-9.
- A. Nguyen, H. N. Ngo, Y. Hong, B. Dang, and B.-P. T. Nguyen, "Ethical principles for artificial intelligence in education," *Education and Information Technologies*, vol. 28, pp. 4221–4241, 2023. doi:10.1007/s10639-022-11316-w.
- S. Joarder and M. A. Chatti, "Human-centred development of indicators for self-service learning analytics: A transparency through exploration approach," *Journal of Learning Analytics*, vol. 13, no. 1, pp. 163–189, 2026. doi:10.18608/jla.2026.8921.
- R. da Silva, J. Eicher, and G. Longo, "A unified survival benchmark for temporal dropout risk prediction in learning analytics," *arXiv:2604.08870*, ver. 3, 2026. doi:10.48550/arXiv.2604.08870.
- E. Howard, "ouladFormat R package: Preparing the Open University Learning Analytics Dataset for analysis," *arXiv:2501.08366*, ver. 2, 2025. doi:10.48550/arXiv.2501.08366.
- S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, 100804, 2023. doi:10.1016/j.patter.2023.100804.
- S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, pp. 56–67, 2020. doi:10.1038/s42256-019-0138-9.
- S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi:10.1145/2939672.2939785.
- G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- 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, e0118432, 2015. doi:10.1371/journal.pone.0118432.
- B. Van Calster et al., "Calibration: the Achilles heel of predictive analytics," *BMC Medicine*, vol. 17, 230, 2019. doi:10.1186/s12916-019-1466-7.
- T. Cerratto Pargman and C. McGrath, "Mapping the ethics of learning analytics in higher education: A systematic literature review of empirical research," *Journal of Learning Analytics*, 2021. doi:10.18608/jla.2021.1.
- A. Palanci, R. M. Yilmaz, and Z. Turan, "Learning analytics in distance education: A systematic review study," *Education and Information Technologies*, vol. 29, pp. 22629–22650, 2024. doi:10.1007/s10639-024-12737-5.

- S. K. Banihashem, O. Noroozi, S. van Ginkel, L. P. Macfadyen, and H. J. A. Biemans, "A systematic review of the role of learning analytics in enhancing feedback practices in higher education," *Educational Research Review*, vol. 37, 100489, 2022. doi:10.1016/j.edurev.2022.100489.
- F. Zhao, G.-Z. Liu, J. Zhou, and C. Yin, "A learning analytics framework based on human-centered artificial intelligence for identifying the optimal learning strategy to intervene in learning behavior," *Educational Technology & Society*, vol. 26, no. 1, pp. 132–146, 2023. doi:10.30191/ETS.202301_26(1).0010.
- F. Ouyang and L. Zhang, "AI-driven learning analytics applications and tools in computer-supported collaborative learning: A systematic review," *Educational Research Review*, vol. 44, 100616, 2024. doi:10.1016/j.edurev.2024.100616.
- B. Ujkani, D. Minkovska, and N. Hinov, "Course success prediction and early identification of at-risk students using explainable artificial intelligence," *Electronics*, vol. 13, no. 21, 4157, 2024. doi:10.3390/electronics13214157.
- S. Gunasekara and M. Saarela, "Explainable AI in Education: Techniques and Qualitative Assessment," *Applied Sciences*, vol. 15, no. 3, 1239, 2025. doi:10.3390/app15031239.
- S. Wu, J. Duan, and M. Luo, "Calibrated early-warning models with fairness auditing and selective prediction for course withdrawal risk: Evidence from OULAD," *PLOS ONE*, vol. 21, no. 7, e0352867, 2026. doi:10.1371/journal.pone.0352867.
- M. Adnan et al., "Predicting at-risk students at different percentages of course length for early intervention using machine learning models," *IEEE Access*, vol. 9, pp. 7519–7539, 2021. doi:10.1109/ACCESS.2021.3049446.
- Tran Uoc, "Dataset_MLA: OULAD course CCC learner extract," GitHub repository, https://github.com/TranUoc/Dataset_MLA, accessed Oct. 7, 2026.
- The Open University, "Module Results Determination Policy," ver. 1.2, Dec. 2023, p. 11 (undergraduate fully graded credit: pass mark 40), accessed Oct. 7, 2026. <https://help.open.ac.uk/documents/policies/module-results-determination-policy/files/239/module-results-determination-policy%20%28DEC%2023%29.pdf>.