International Journal of Research in Interdisciplinary Studies
Volume 4, Issue 11, November 2026
www.ijris.com | E-ISSN: 2584-1017

Leakage-Free, Explainable and Fairness-Aware Early Prediction of At-Risk Students: A Temporal Validation Study on the Open University Learning Analytics Dataset

Md Selim Mia[1*], Mohammed Abdullah Ahmed Baroom[1]
[1]*Department of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing, China*

Abstract—Early identification of students at risk of failing or withdrawing enables timely intervention, yet many published predictors are evaluated with random splits, test-set tuning and accuracy-centred metrics that overstate deployable performance. This study proposes a leakage-free, explainable and fairness-aware early-warning framework evaluated on the Open University Learning Analytics Dataset (32,593 enrolments). Models were trained on three earlier course presentations and tested on an unseen later presentation, using only information available on each prediction day. Eight classifiers were tuned with student-grouped cross-validation. At day 60, LightGBM achieved a ROC-AUC of 0.813 and a PR-AUC of 0.761 and identified 75.3% of at-risk students; ROC-AUC rose from 0.72 at course start to 0.90 at day 150. Assessment behaviour accounted for 47% of SHAP attributions. A fairness audit revealed disparities across deprivation and prior-education groups. A replication on the UCI Student Performance data confirmed that early academic signals outweigh demographic predictors.

***Index Terms*—Algorithmic fairness, At-risk students, Early warning system, Explainable AI, Learning analytics, Temporal validation.**

1. Introduction

Student attrition is one of the most persistent problems in higher education, and it is particularly acute in distance and online learning, where students study largely alone and signs of disengagement are easily missed by tutors. Failing or withdrawing from a module carries financial and psychological costs for the student and reputational and funding costs for the institution. Theoretical accounts of student departure have long emphasised that withdrawal is the end point of a gradual process of weakening academic and social integration rather than a sudden event (Tinto, 1975). If that process leaves observable traces early enough, institutions can intervene while there is still time to change the outcome.

The growth of virtual learning environments (VLEs) has made such traces measurable at scale. Every click on a learning resource, every assignment submission and every assessment score is recorded, and the learning analytics community has used these records to build early-warning systems that flag at-risk students for tutor follow-up (Arnold and Pistilli, 2012; Jayaprakash et al., 2014). The literature on predicting student performance with machine learning is now extensive (Romero and Ventura, 2020), and the Open University Learning Analytics Dataset (OULAD) has become a standard public benchmark for this task (Kuzilek et al., 2017).

Despite this progress, the evidence base for deployable early-warning models remains weaker than headline accuracies suggest. First, many studies split data randomly at the enrolment level, so that a model is trained and tested on students from the same course presentation. In practice, a model must be trained on past cohorts and applied to a future cohort whose conditions may have changed (Hlosta et al., 2017). Second, features are often computed over the whole course, or students who have already withdrawn are retained in the prediction cohort, which constitutes target leakage (Kaufman et al., 2012). Third, hyper-parameters and decision thresholds are frequently selected on the same data used to report performance, which produces optimistically biased estimates (Varma and Simon, 2006; Cawley and Talbot, 2010). Fourth, accuracy and weighted averages are commonly reported even though the at-risk class is the minority and the class that matters; precision-recall analysis and the Matthews correlation coefficient (MCC) are more informative under imbalance (Saito and Rehmsmeier, 2015; Chicco and Jurman, 2020). Finally, high-stakes educational decisions require that predictions be explainable to tutors and students (Khosravi et al., 2022) and that their error rates be examined across demographic groups, because predictive models can reproduce or amplify existing inequalities (Gardner et al., 2019; Baker and Hawn, 2022).

This study addresses these gaps with a single, reproducible framework that combines a realistic evaluation protocol with explainability and fairness analysis. The following research questions (RQs) guide the work:

- *RQ1*. How accurately can students who will fail or withdraw be identified under a train-on-past, predict-next-presentation protocol, and do model families differ significantly?

*Corresponding author: selim.shpi@gmail.com

- *RQ2.* How early in a course can a reliable warning be issued?
- *RQ3.* Which information sources (demographic, VLE engagement, assessment) and which class-imbalance strategies drive predictive performance?
- *RQ4.* Which factors explain the predictions at the global and the individual level?
- *RQ5.* Is the early-warning model equally reliable across demographic groups?
- *RQ6.* Do the main conclusions hold in a different educational context?

The study makes four contributions:

- A leakage-free early-prediction pipeline for OULAD that uses only information available on the prediction day, removes students who have already withdrawn, holds out the most recent course presentation as a test set and groups cross-validation folds by student.
- A statistically grounded comparison of eight classifiers, from logistic regression to gradient boosting and neural networks, with bootstrap confidence intervals, Friedman and Nemenyi tests and DeLong tests.
- An analysis of how predictive performance develops over the first 150 days of a course, together with feature-group and imbalance-strategy ablations.
- A SHAP-based explanation of global and individual predictions, a group-fairness audit of the deployed decision rule and a replication on a secondary-school dataset from a different country.

The remainder of the paper is organised as follows. Section 2 reviews the theoretical background and related work. Section 3 describes the data and methodology. Section 4 reports the results, which Section 5 discusses together with the limitations of the study. Section 6 outlines implications for research and practice, Section 7 concludes and Section 8 proposes directions for future research.

2. Literature and Theoretical Underpinning

A. Theoretical Background: Integration, Engagement and Departure

Tinto's (1975) model of student departure explains withdrawal as the outcome of a longitudinal interaction between the student's background characteristics and his or her academic and social integration into the institution. Background attributes such as prior education and socio-economic status shape initial commitment, but it is the subsequent experience of academic integration, reflected in performance and intellectual engagement, that most directly determines persistence. Two implications follow for predictive modelling. First, demographic attributes should carry some predictive signal but should be dominated by measures of what the student actually does once the course has begun. Second, because integration develops over time, predictive power should increase as behavioural evidence accumulates. In a VLE, click activity serves as a behavioural proxy for engagement, while assessment submission and scores capture academic integration directly. The feature design and the research questions of this study are organised around these three information sources.

B. Predicting Student Performance and Dropout

Early work in educational data mining showed that student outcomes in distance learning can be predicted from demographic and early-assessment data using standard classifiers (Kotsiantis et al., 2004). Cortez and Silva (2008) predicted secondary-school performance in Portugal and found that earlier period grades were by far the strongest predictors, while demographic and social variables yielded modest accuracy on their own. Subsequent research shifted towards learning management system (LMS) log data. Conijn et al. (2017) compared 17 blended courses in Moodle and found that the predictive value of LMS activity varied considerably between courses and was limited for early prediction once in-course assessment data were available. Gašević et al. (2016) similarly showed that the instructional conditions of each course moderate which activity features are predictive, cautioning against one-size-fits-all models. Gray and Perkins (2019) demonstrated that engagement measured in the first weeks of a programme can identify at-risk students early, and Riestra-González et al. (2021) analysed large-scale Moodle logs to predict course-agnostic performance at early stages of the course. Surveys of the field document the breadth of these approaches (Romero and Ventura, 2020).

C. Studies Using the Open University Learning Analytics Dataset

OULAD was released by the Open University (UK) to provide an open benchmark that combines demographic, registration, assessment and daily VLE interaction data (Kuzilek et al., 2017). Hlosta et al. (2017) highlighted that models trained on previous presentations may not transfer to a new one and proposed the Ouroboros approach, which learns from students in the current presentation who have already submitted an assessment. Aljohani et al. (2019) modelled weekly clickstream sequences with long short-term memory networks to predict at-risk students, Waheed et al. (2020) applied deep artificial neural networks to features derived from VLE big data, and He et al. (2020) combined static demographic data and sequential activity in joint recurrent neural networks. Adnan et al. (2021) evaluated a range of machine learning models at different percentages of course length, showing that performance improves as more of the course is observed.

These studies establish that OULAD contains a strong predictive signal, but their reported results are difficult to compare or to translate into deployment expectations. Most use random or stratified splits pooled across presentations, many report accuracy as the primary metric, some compute features over windows that extend to the end of the course, and few report confidence intervals, significance tests, or analyses of fairness and explainability. The present study does not aim to surpass a particular reported accuracy. It asks instead what performance can honestly be expected when these sources of optimism are removed.

Table 1
OULAD presentations and outcome distribution

Presentation	n	Modules	Pass	Distinction	Fail	Withdrawn	At-risk	Role
2013B	4,684	3	37.7%	7.0%	26.5%	28.8%	55.3%	Train
2013J	8,845	6	42.1%	8.5%	22.6%	26.8%	49.4%	Train
2014B	7,804	6	33.0%	10.0%	23.5%	33.5%	57.0%	Train
2014J	11,260	7	38.1%	10.3%	17.6%	34.0%	51.5%	Test

D. Machine Learning Methods

Ensembles of decision trees are strong baselines for tabular data. Random forests (Breiman, 2001) average decorrelated trees, while gradient-boosting implementations such as XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017) and CatBoost (Prokhorenkova et al., 2018) build trees sequentially and add regularisation and efficient split finding. Class imbalance can be handled by cost-sensitive weighting, by resampling such as SMOTE (Chawla et al., 2002) or by moving the decision threshold. Multiple classifiers should be compared with non-parametric tests such as the Friedman test followed by the Nemenyi post-hoc test (Friedman, 1937; Nemenyi, 1963; Demšar, 2006). Two correlated ROC curves evaluated on the same test set should be compared with the DeLong test (DeLong et al., 1988; Sun and Xu, 2014).

E. Explainability and Fairness in Learning Analytics

Explainable artificial intelligence is increasingly regarded as a prerequisite for the responsible use of predictive models in education, because tutors must understand why a student has been flagged in order to choose an appropriate intervention (Khosravi et al., 2022). SHapley Additive exPlanations (SHAP) attribute each prediction to the input features with game-theoretic guarantees (Lundberg and Lee, 2017), and TreeSHAP computes these values exactly and efficiently for tree ensembles (Lundberg et al., 2020). Fairness is an equally pressing concern. Baker and Hawn (2022) reviewed evidence that educational algorithms perform differently across demographic groups, and Gardner et al. (2019) proposed slicing analysis to evaluate predictive student models across subgroups. From the fairness literature, equality of opportunity requires equal true positive rates across groups (Hardt et al., 2016), and the disparate-impact ratio compares selection rates between groups (Feldman et al., 2015). For an early-warning system, a lower true positive rate in one group means that its at-risk students are systematically less likely to receive support.

F. Research Gap

To the best of our knowledge, no study on OULAD jointly (i) enforces a strict temporal and leakage-free protocol across multiple prediction days, (ii) compares a broad set of modern classifiers with appropriate statistical tests and imbalance-aware metrics, (iii) explains predictions at the global and the individual level and (iv) audits the fairness of the resulting decision rule, with (v) a replication on an independent dataset. This study fills that gap.

3. Methodology

A. Datasets

OULAD (Kuzilek et al., 2017) contains anonymised data on 32,593 student–module enrolments in seven modules taught in four presentations: February (B) and October (J) starts in 2013 and 2014. It comprises seven linked tables: courses, assessments, VLE resources, student information, registrations, assessment submissions (173,912 records) and daily VLE interactions (10,655,280 records). Each enrolment has one of four final results: Distinction, Pass, Fail or Withdrawn. Table 1 summarises the presentations and outcome distribution.

UCI Student Performance (Cortez and Silva, 2008) describes 649 students in a Portuguese-language course and 395 students in a mathematics course at two Portuguese secondary schools. It contains 30 demographic, social and school-related attributes and three period grades (G1, G2, G3) on a 0–20 scale. This dataset differs from OULAD in education level, delivery mode, country and size, and it therefore provides a demanding test of whether the conclusions generalise.

B. Problem Formulation and Leakage Control

For OULAD, an enrolment is labelled *at risk* (positive class) if its final result is Fail or Withdrawn and *not at risk* if it is Pass or Distinction. A prediction is issued on day t after the official module start, with $t \in \{0, 30, 60, 90, 120, 150\}$. Day 60, roughly the end of the first quarter of a typical 240–270-day presentation, is used as the main analysis point. Three rules prevent leakage. First, only VLE clicks recorded before day t (including pre-course activity) and only non-exam assessments submitted before day t are used. Second, students who had already unregistered before day t are removed from the prediction cohort at t, because their outcome is already known and including them would inflate performance trivially. Third, the final result, the unregistration date and the presentation identifier are never used as inputs. As a result, the cohort shrinks from 29,915 enrolments at day 0 to 24,095 at day 150, and the at-risk rate falls from 48.6% to 36.1%.

For UCI, a student is labelled as failing if the final grade G3 is below 10. Two realistic scenarios are defined: S1 (enrolment time) uses only demographic, social and school attributes, and S2 (after the first period) adds the first-period grade G1. G2 and G3 are always excluded.

C. Feature Engineering

Forty-six features were computed for every enrolment and cutoff, following the three information sources identified in Section 2.1 (Table 2). All cutoffs share the same feature schema. Ordinal attributes (deprivation band, highest education, age band) were encoded as integers. Within-

Table 2
Engineered feature groups (46 Features)

Group	No.	Features
Demographic and registration	11	Gender; disability; index of multiple deprivation (IMD) band; highest education; age band; number of previous attempts; studied credits; days registered before start; presentation length; module (categorical); region (categorical)
VLE engagement	24	Total clicks (raw and log); active days; distinct resources; pre-course clicks; clicks in the last 14 days; days since last activity; no-activity flag; weekly mean, standard deviation and slope of clicks; cohort percentile of clicks and of active days; clicks per activity type (ten most frequent types and 'other')
Assessment	11	Assessments submitted, due and missed; submission rate; mean, minimum and weight-weighted score; number of scores below 40; number of late submissions; mean days submitted before the deadline; number of banked assessments

presentation percentile ranks of total clicks and active days make engagement comparable across modules with different VLE designs. Weekly engagement statistics include weeks without any activity, and the engagement slope is the least-squares trend of weekly clicks over the observed weeks. Assessment features cover only assessments due or submitted before the cutoff.

D. Evaluation Protocol

A temporal hold-out design was adopted. Models were trained on the 2013B, 2013J and 2014B presentations and evaluated once on 2014J, the most recent presentation. At day 60 this gives 17,566 training enrolments (43.5% at risk) and 8,837 test enrolments (38.2% at risk). Because some students take several modules or repeat a module, all cross-validation inside the training set used five-fold StratifiedGroupKFold with the student identifier as the group, so that no student contributes to both the training and the validation folds. Preprocessing was fitted inside each fold. Numeric features were imputed with the median (with missing-value indicators) and standardised, and categorical features were one-hot encoded.

Hyper-parameters were selected by randomised search (Bergstra and Bengio, 2012) with 30 configurations per model, maximising the area under the precision-recall curve (PR-AUC) in grouped cross-validation. The decision threshold was then chosen to maximise the F1 score of the at-risk class on out-of-fold training predictions. The test presentation was not used for any of these choices. For the early-prediction analysis, each model was refitted at every cutoff with the hyper-parameters selected at day 60, so that the curve isolates the effect of information availability.

E. Classifiers and Imbalance Handling

Eight classifiers were compared: logistic regression (LR), k-nearest neighbours (KNN), decision tree (DT), random forest (RF), XGBoost (XGB), LightGBM, CatBoost and a multilayer perceptron (MLP). Where supported, class imbalance was handled by cost-sensitive learning (balanced class weights, or a positive-class weight equal to the negative-to-positive ratio for XGBoost). An ablation compared this choice with no correction, SMOTE (Chawla et al., 2002), random under-sampling and threshold tuning alone, using the imbalanced-learn library (Lemaître et al., 2017).

F. Performance Measures and Statistical Testing

The primary metrics are ROC-AUC and PR-AUC, which are threshold-independent. PR-AUC is particularly informative for the minority class (Saito and Rehmsmeier, 2015). At the selected threshold, the study reports balanced accuracy, F1 of the at-risk class, macro-F1, MCC (Chicco and Jurman, 2020), recall, precision, specificity and the Brier score as a measure of probability calibration (Niculescu-Mizil and Caruana, 2005). Percentile 95% confidence intervals (CIs) were obtained from 2,000 bootstrap resamples of the test set. To compare models, the PR-AUC of each tuned model was computed on 15 folds (three repetitions of grouped five-fold cross-validation) and analysed with the Friedman test and the Nemenyi post-hoc test (Demšar, 2006). On the test set, the ROC-AUC of the best model was compared with that of every other model using the DeLong test (DeLong et al., 1988), implemented with the fast algorithm of Sun and Xu (2014), and the p-values were adjusted with the Holm procedure (Holm, 1979).

G. Explainability

TreeSHAP (Lundberg et al., 2020) was applied to the best tree-based model on a random sample of 2,000 test enrolments. Global importance was measured as the mean absolute SHAP value per feature. Values of one-hot encoded columns and missing-value indicators were summed into their source feature and aggregated by information source. Dependence plots and an individual waterfall explanation for a correctly flagged student illustrate local behaviour.

H. Fairness Audit

The deployed decision rule (best model and its out-of-fold threshold) was audited on the test presentation across five attributes: gender, disability, age band, IMD quintile and highest education. Groups with fewer than 30 test enrolments were omitted. For each group, the study reports the base rate, selection rate, true positive rate (TPR), false positive rate (FPR), precision and ROC-AUC. Three disparity measures were derived: the equal-opportunity gap (the maximum difference in TPR; Hardt et al., 2016), the equalised-odds gap (the larger of the TPR and FPR gaps) and the disparate-impact ratio (the minimum over the maximum selection rate; Feldman et al., 2015).

I. Cross-Dataset Replication

Because the UCI datasets are small, a fixed split would give unstable estimates. Nested cross-validation was therefore used: an outer three-times-repeated stratified five-fold loop for evaluation (15 folds) and an inner stratified three-fold loop for randomised hyper-parameter search (15 configurations) and threshold selection. No information from an outer test fold was used for model selection. The same eight classifiers and metrics were used, and models were compared with the Friedman test.

Table 3
Hold-Out performance at Day 60 (2014J, N = 8,837)

Model	ROC-AUC [95% CI]	PR-AUC [95% CI]	Bal. Acc.	F1	MCC	Rec.	Prec.
LR	0.807 [0.798, 0.816]	0.752 [0.739, 0.766]	0.726	0.668	0.444	0.709	0.631
KNN	0.794 [0.785, 0.803]	0.729 [0.716, 0.743]	0.718	0.667	0.424	0.770	0.588
DT	0.785 [0.775, 0.795]	0.724 [0.710, 0.739]	0.707	0.652	0.402	0.734	0.586
RF	0.809 [0.800, 0.818]	0.760 [0.747, 0.773]	0.725	0.673	0.438	0.768	0.599
XGB	0.812 [0.803, 0.821]	0.758 [0.746, 0.772]	0.730	0.676	0.450	0.742	0.621
LightGBM	0.813 [0.804, 0.822]	0.761 [0.749, 0.774]	0.731	0.678	0.451	0.753	0.616
CatBoost	0.810 [0.801, 0.819]	0.757 [0.745, 0.770]	0.727	0.675	0.442	0.765	0.604
MLP	0.811 [0.802, 0.820]	0.757 [0.744, 0.771]	0.730	0.675	0.449	0.740	0.621

Rec. = recall; Prec. = precision.

J. Implementation

All experiments were implemented in Python 3.13 with scikit-learn 1.6.1 (Pedregosa et al., 2011), XGBoost 3.4.1, LightGBM 4.6.0, CatBoost 1.2.10, imbalanced-learn and SHAP 0.52.0, and were run on a Kaggle CPU environment with a fixed random seed (42). The complete pipeline, which automatically downloads both datasets from the UCI Machine Learning Repository and regenerates every table and figure, is publicly available (see Data and Code Availability).

4. Results and Findings

A. Descriptive Analysis

Outcomes are almost balanced overall but vary over time. The at-risk rate in the training presentations is 43.5% at day 60, compared with 38.2% in the test presentation, which indicates a degree of distribution shift that a temporal evaluation captures and a random split would hide. Fig. 1 shows the average weekly VLE activity per student by final outcome. Students who later achieved Distinction or Pass were consistently more active from the first week onwards, while students who failed or withdrew showed low and declining activity. The separation is visible within the first few weeks, which motivates early prediction.

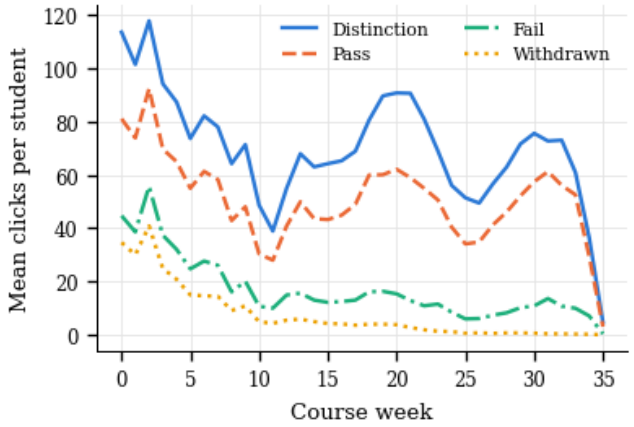


Fig. 1. Weekly VLE clicks by final outcome.

B. Hold-out Performance at Day 60 (RQ1)

Table 3 reports the performance of all models on the unseen 2014J presentation at day 60. The four strongest models (LightGBM, XGB, MLP and CatBoost) achieved ROC-AUC values between 0.810 and 0.813 and PR-AUC values between 0.757 and 0.761, with overlapping confidence intervals. LightGBM obtained the highest ROC-AUC (0.813, 95% CI 0.804–0.822), PR-AUC (0.761, 95% CI 0.749–0.774), balanced accuracy (0.731), F1 of the at-risk class (0.678) and MCC (0.451). Logistic regression remained competitive (ROC-AUC 0.807), whereas KNN and the single decision tree were clearly weaker. Brier scores ranged from 0.168 (MLP) to 0.185 (DT). At its out-of-fold threshold, LightGBM flagged 4,132 of the 8,837 test enrolments and correctly identified 2,546 of the 3,380 students who eventually failed or withdrew (recall 0.753, precision 0.616).

Fig. 2 shows the ROC and precision-recall curves. The curves of the ensemble models and the MLP are almost indistinguishable, and all models lie well above the no-skill baselines. Fig. 3(a) shows that the class-weighted models tend to overestimate the probability of risk in the middle of the probability range, an expected consequence of re-weighting. Ranking quality is unaffected, but the probabilities should be recalibrated if they are to be communicated as risk estimates. Fig. 3(b) gives the confusion matrix of LightGBM.

C. Statistical Comparison of Models

The Friedman test on 15 cross-validation folds rejected the hypothesis that all models perform equally ($\chi^2 = 91.78$, $p = 5.3 \times 10^{-17}$). The average ranks were XGB 1.80, CatBoost 2.33, LightGBM 2.67, MLP 3.40, RF 5.27, LR 5.53, KNN 7.00 and DT 8.00. The Nemenyi critical-difference diagram (Fig. 4) shows that the three gradient-boosting models and the MLP form a top group that is not significantly different internally, while DT and KNN are significantly worse than the top group. Table 4 reports DeLong tests on the test presentation. LightGBM's advantage over LR, KNN, DT, RF and CatBoost is statistically significant after Holm correction, but its differences from XGB ($p = 0.141$) and the MLP ($p = 0.185$) are not. All differences among the top four models are smaller than 0.003 in ROC-AUC.

Table 4
DeLong Tests: Light GBM versus the other models

Comparison	ΔROC-AUC	p-value	Holm-adjusted p
LightGBM vs LR	0.0063	0.0027	0.0107
LightGBM vs KNN	0.0195	< 0.0001	< 0.0001
LightGBM vs DT	0.0287	< 0.0001	< 0.0001
LightGBM vs RF	0.0042	0.0015	0.0076
LightGBM vs XGB	0.0017	0.0706	0.1412
LightGBM vs CatBoost	0.0028	0.0114	0.0341
LightGBM vs MLP	0.0024	0.1849	0.1849

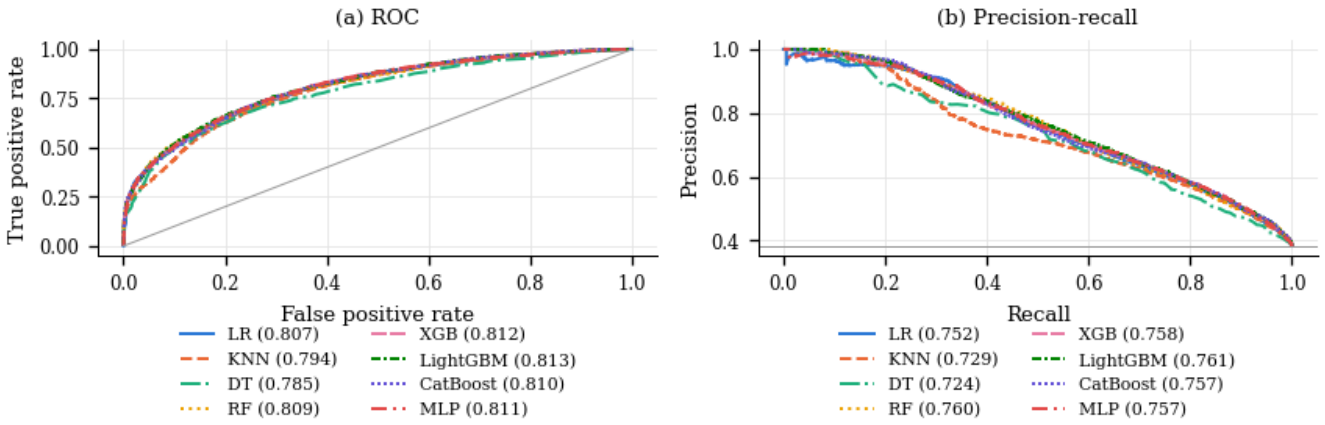


Fig. 2. ROC (a) and precision-recall (b) curves at day 60.

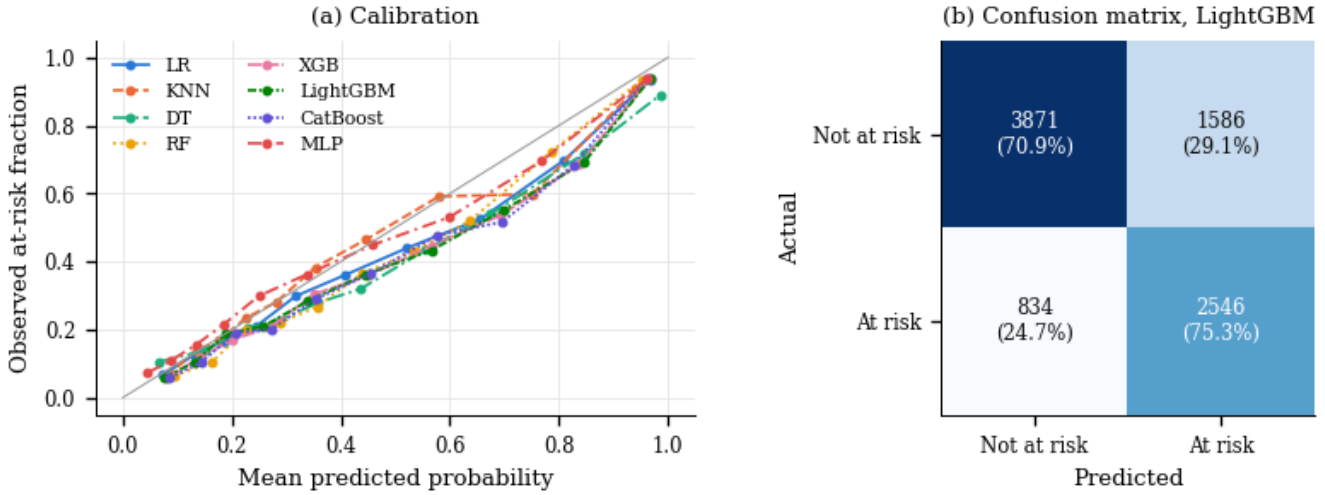


Fig. 3. (a) Calibration, (b) LightGBM confusion matrix.

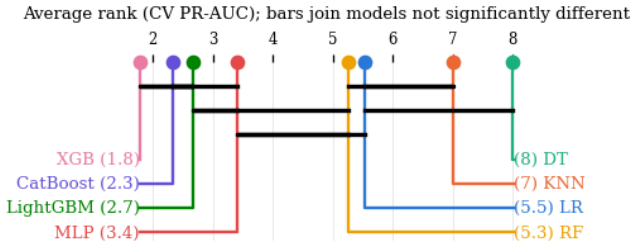


Fig. 4. Critical-difference diagram (Nemenyi test, α = 0.05).

Table 5
Hold-Out ROC-AUC by prediction day

Model	Day 0	Day 30	Day 60	Day 90	Day 120	Day 150
LR	0.722	0.735	0.807	0.833	0.869	0.897
KNN	0.711	0.736	0.794	0.822	0.855	0.884
DT	0.681	0.713	0.785	0.804	0.852	0.879
RF	0.717	0.745	0.809	0.830	0.871	0.900
XGB	0.727	0.753	0.812	0.831	0.875	0.902
LightGBM	0.724	0.747	0.813	0.834	0.877	0.904
CatBoost	0.727	0.754	0.810	0.837	0.872	0.902
MLP	0.725	0.737	0.811	0.838	0.869	0.899

D. How Early Can At-Risk Students Be Identified? (RQ2)

Table 5 and Fig. 5 show the hold-out performance at each prediction day (the shaded band in Fig. 5 is the bootstrap 95% CI for LightGBM). At course start (day 0), when only demographic, registration and pre-course activity data are available, the best models reached a ROC-AUC of about 0.72. Performance rose steadily as behavioural evidence accumulated: by day 60 every ensemble exceeded 0.80, and by day 150 LightGBM reached 0.904. The largest single gain occurred between day 30 and day 60 (approximately +0.06 ROC-AUC for LightGBM), when the first assessment results become available.

PR-AUC behaves differently between day 0 and day 30, where it decreases slightly before rising steadily from day 60 onwards (Fig. 5(b)). This is not a loss of discriminative ability, since ROC-AUC increases over the same interval. It is a consequence of the leakage-control rule: students who withdraw within the first 30 days are the easiest to recognise and are removed from the cohort at day 30, which lowers the at-risk rate from 48.6% to 44.1% and makes the remaining positive cases harder to detect. Studies that keep already-withdrawn students in the cohort would not observe this effect and would overstate early performance.

E. Ablation Studies (RQ3)

Table 6 isolates the contribution of each information source

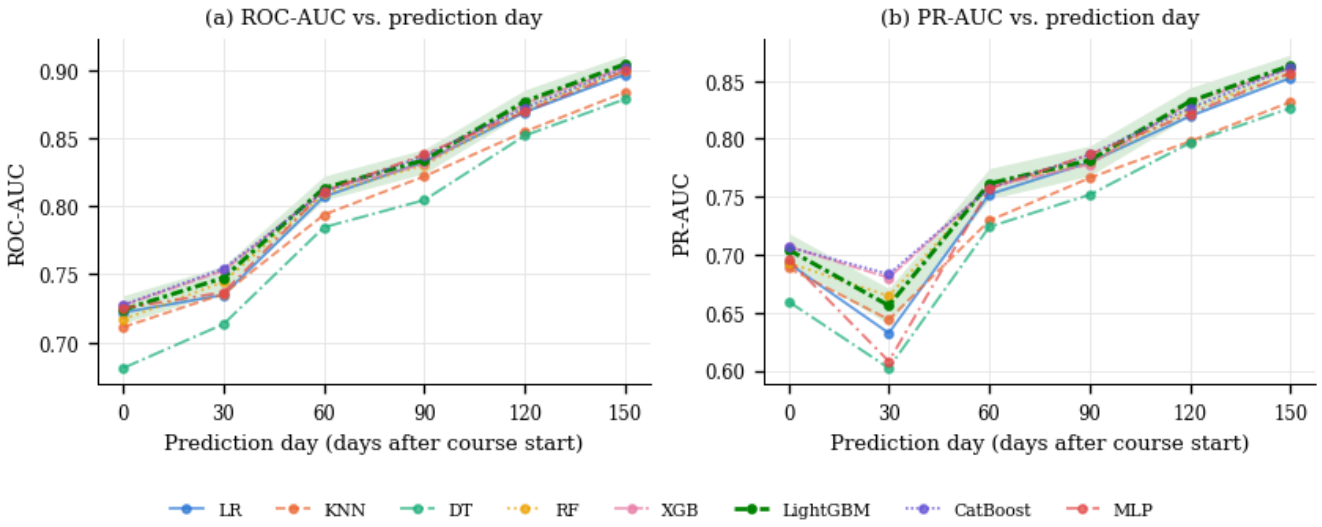


Fig. 5. (a) ROC-AUC, and (b) PR-AUC, by prediction day.

using LightGBM at day 60. Demographic and registration data alone gave only modest discrimination (ROC-AUC 0.639). Assessment data alone (0.784) and VLE engagement alone (0.759) were far more informative. Combining the two behavioural sources gave 0.806, and adding demographics yielded the full model (0.813). Every combination with more information sources outperformed its subsets, and the confidence intervals of the full model and of the demographic-only model do not overlap.

Table 7 compares imbalance-handling strategies. The threshold-independent metrics were practically identical across strategies (PR-AUC 0.756–0.762), which indicates that imbalance correction does not change how well the model ranks students. The strategies differed mainly in where they place the operating point. At the default threshold of 0.5, the uncorrected model had the lowest recall (0.642), while cost-sensitive weighting (0.688) and random under-sampling (0.702) raised it. Tuning the threshold on out-of-fold predictions raised recall the most (0.747) without any resampling. For an early-warning system, where a missed at-risk student is usually more costly than an unnecessary check-in, the choice of threshold matters more than the choice of resampling method.

F. Explaining the Predictions (RQ4)

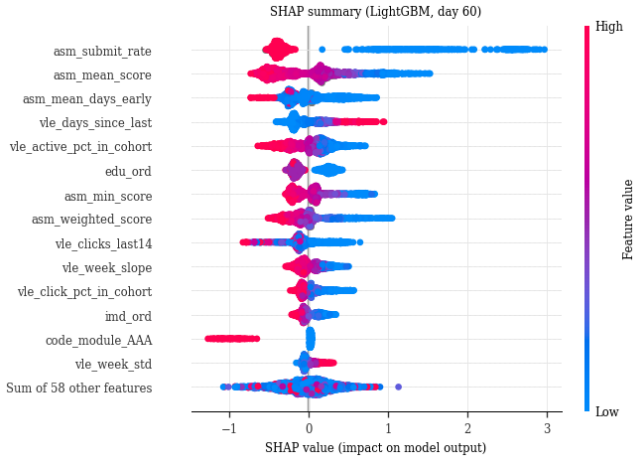


Fig. 6. SHAP summary of Light GBM at day 60.

Table 6
Feature-Group Ablation (Light GBM, Day 60)

Feature set	No. of features	ROC-AUC [95% CI]	PR-AUC [95% CI]
Demographic only	11	0.639 [0.626, 0.651]	0.516 [0.498, 0.532]
VLE only	24	0.759 [0.749, 0.769]	0.691 [0.675, 0.707]
Assessment only	11	0.784 [0.774, 0.794]	0.719 [0.706, 0.734]
Demographic + VLE	35	0.775 [0.765, 0.785]	0.707 [0.692, 0.722]
Demographic + Assessment	22	0.797 [0.788, 0.807]	0.737 [0.724, 0.751]
VLE + Assessment	35	0.806 [0.797, 0.815]	0.753 [0.740, 0.766]
All (proposed)	46	0.813 [0.805, 0.822]	0.761 [0.748, 0.774]

Table 7
Imbalance-handling strategies (Light GBM, Day 60)

Strategy	Thr.	PR-AUC	Bal. Acc.	F1	MCC	Rec.	Prec.	Spec.
None	0.500	0.762	0.730	0.664	0.467	0.642	0.687	0.819
None + OOF-tuned threshold	0.378	0.762	0.732	0.678	0.453	0.747	0.621	0.718
Cost-sensitive weighting	0.500	0.761	0.731	0.671	0.458	0.688	0.654	0.774
SMOTE	0.500	0.757	0.727	0.661	0.455	0.655	0.668	0.798
Random under-sampling	0.500	0.756	0.730	0.671	0.453	0.702	0.642	0.758

Thr. = threshold; Spec. = specificity.

Fig. 6 summarises the SHAP values of LightGBM. The single most influential feature was the assessment submission rate: students who had not submitted the assessments due by day 60 received large positive (risk-increasing) contributions. It was followed by the mean assessment score, the mean number of days by which assessments were submitted before the deadline, the number of days since the last VLE activity and the student's percentile of active days within the cohort. Lower highest education, higher deprivation (lower IMD band) and a declining weekly-engagement slope also increased predicted risk, and the module identifier captured module-specific base rates. Aggregated by information source, assessment features accounted for 46.6% of the total attribution, VLE engagement for 34.0% and demographic and registration features for 19.4%. This ordering mirrors the ablation results.

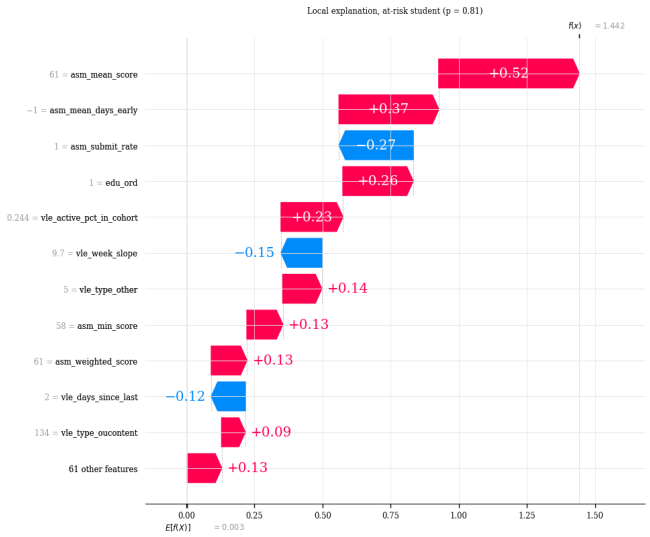


Fig. 7. SHAP explanation for one at-risk student.

Fig. 7 shows an individual explanation for a correctly flagged student with a predicted risk of 0.81. The prediction was driven by a mediocre mean assessment score (61), submissions close to or after the deadlines, a lower prior-education level and activity in the lower quarter of the cohort. A complete submission record and recent VLE activity partly offset this risk. Explanations of this kind give tutors a concrete starting point for a conversation with the student, for example about time management or assessment preparation.

G. Fairness Audit (RQ5)

Table 8 and Fig. 8 summarise the fairness audit of the deployed LightGBM rule. Discrimination (ROC-AUC) was similar across groups for disability (0.809–0.812), age band (0.793–0.823) and IMD quintile (0.796–0.819). TPRs were almost equal for women and men (0.762 and 0.748; equal-opportunity gap 0.014), but women had a markedly higher FPR (0.370 versus 0.229) and a lower ROC-AUC (0.776 versus 0.841). Larger disparities appeared along socio-economic and educational lines. The TPR decreased monotonically from 0.826 in the most deprived IMD quintile to 0.667 in the least deprived, and the selection rate fell from 0.593 to 0.386 (disparate-impact ratio 0.381 when the 'unknown' IMD group is included). Students with lower prior education were flagged far more often than those with higher education qualifications, in line with differences in base rates (0.456 versus 0.314). Older students (35–55 and over 55) had lower TPRs than younger students (0.675 and 0.600 versus 0.786), so at-risk older students were more likely to be missed.

H. Cross-Dataset Replication on UCI Student Performance (RQ6)

The failure rates were 15.4% for Portuguese and 32.9% for mathematics. Table 9 reports the nested cross-validation results. With enrolment-time attributes only (S1), discrimination was moderate: the best ROC-AUC was 0.835 (RF) for Portuguese and 0.701 (RF) for mathematics, and the PR-AUC values (0.509 and 0.569) were only moderately above the respective base rates. Adding the first-period grade (S2) produced a large and consistent improvement: ROC-AUC rose to 0.943 for Portuguese (RF) and 0.926 for mathematics (XGB), and PR-AUC to 0.726 and 0.859. The Friedman test detected significant differences between models in every scenario ($p \leq 0.004$), but the best models were again tree ensembles and LR, with small differences among them (Fig. 9). As in OULAD, early

Table 8
Fairness summary of the deployed model

Attribute	Groups	TPR range	FPR range	Equal-opp. gap	Equalised-odds gap	Disparate-impact ratio
Gender	F, M	0.748–0.762	0.229–0.370	0.014	0.140	0.852
Disability	No, Yes	0.746–0.806	0.283–0.389	0.060	0.106	0.762
Age band	3	0.600–0.786	0.179–0.313	0.186	0.186	0.617
IMD quintile	6 (incl. unknown)	0.561–0.826	0.100–0.395	0.265	0.295	0.381
Highest education	5	0.500–0.875	0.162–0.516	0.375	0.375	0.364

Table 9
UCI nested cross-validation results (ROC-AUC / PR-AUC)

Model	Mathematics S1	Mathematics S2	Portuguese S1	Portuguese S2
LR	0.686 / 0.549	0.923 / 0.852	0.810 / 0.451	0.929 / 0.698
KNN	0.652 / 0.537	0.874 / 0.761	0.781 / 0.444	0.874 / 0.564
DT	0.654 / 0.495	0.915 / 0.817	0.795 / 0.431	0.917 / 0.672
RF	0.701 / 0.569	0.924 / 0.852	0.835 / 0.509	0.943 / 0.726
XGB	0.678 / 0.525	0.926 / 0.859	0.808 / 0.469	0.935 / 0.718
LightGBM	0.671 / 0.521	0.919 / 0.846	0.799 / 0.462	0.930 / 0.689
CatBoost	0.678 / 0.550	0.923 / 0.851	0.820 / 0.503	0.938 / 0.703
MLP	0.614 / 0.490	0.894 / 0.807	0.748 / 0.410	0.805 / 0.500

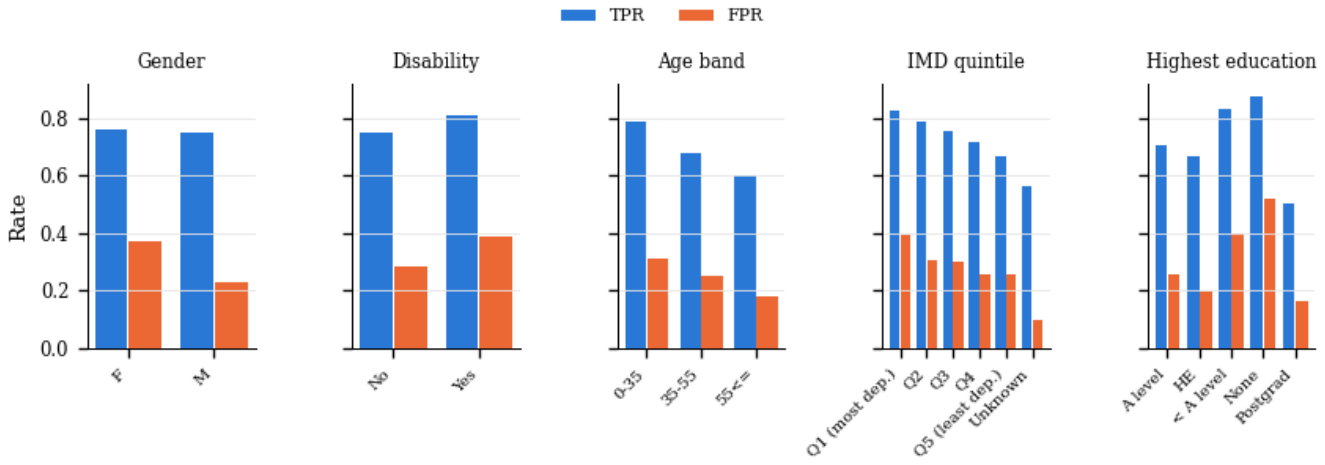


Fig. 8. TPR and FPR by demographic group.

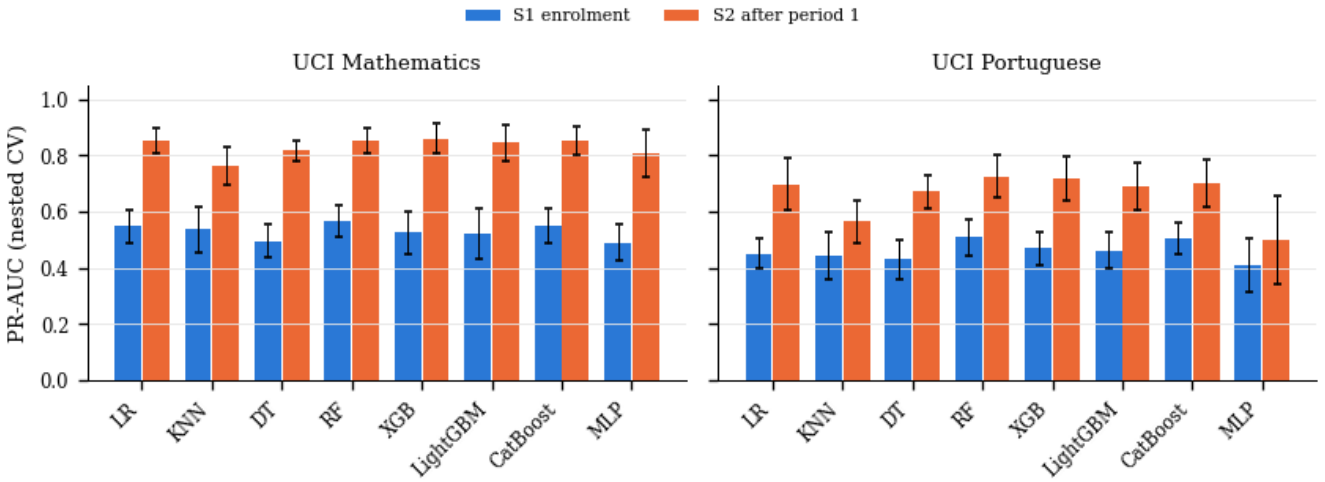


Fig. 9. UCI PR-AUC under scenarios S1 and S2.

academic evidence dominated demographic and social attributes.

5. Discussion

A. Realistic Performance and the Role of the Model

Under a strict temporal and leakage-free protocol, the best model identified three out of four students who would fail or withdraw, at about one quarter of the course, with a ROC-AUC of 0.81. These figures are lower than some accuracies reported on OULAD in the literature, but they are obtained under conditions that match deployment: the model never sees the target presentation, never uses information from after the prediction day and never evaluates students whose outcome is already known. The gap between training and test at-risk rates (43.5% versus 38.2%) also shows that presentation-level drift exists and must be expected in practice (Hlosta et al., 2017).

Regarding RQ1, gradient-boosting models and the MLP formed a statistically indistinguishable top group, and their test-set differences did not exceed 0.003 ROC-AUC. Even the significant DeLong differences are small in practical terms; for example, LightGBM exceeded LR by only 0.006. The best model also differed between cross-validation (XGB) and the hold-out presentation (LightGBM). The main determinants of performance in this task are therefore the information available on the prediction day and the evaluation design, not the choice among modern tabular learners. Logistic regression, which is fully transparent, is a defensible choice where simplicity matters more than a marginal gain in discrimination.

B. Early Prediction and Information Sources

The early-prediction curve (RQ2) and the ablations (RQ3) support the theoretical expectations of Section 2.1. Demographic attributes carry a real but limited signal (ROC-AUC 0.64 alone), which is consistent with the role of background characteristics in Tinto's (1975) model. Behavioural evidence of academic integration dominates once it becomes available: assessment features provide most of the information, VLE engagement adds complementary information, and the largest jump in performance coincides with the first graded assessments between day 30 and day 60. This agrees with Cortez and Silva (2008), who found early grades to be the strongest predictor in secondary education, and with Conijn et al. (2017), who found that assessment data

reduce the added value of LMS activity. The UCI replication (RQ6) reproduces the same pattern in a different country, education level and delivery mode. Practically, day 60 offers a reasonable compromise: performance has passed 0.80 ROC-AUC while about three-quarters of the course remain for intervention.

The temporary decrease in PR-AUC between day 0 and day 30 is a methodological finding in its own right. It shows that the population being predicted changes over time as students withdraw. Comparing performance across cutoffs without accounting for this change, or retaining withdrawn students, can produce misleading conclusions about how early prediction is possible.

C. *Explanations and Fairness*

The SHAP analysis (RQ4) identifies actionable signals (missed or late submissions, low scores and a growing gap since the last login) rather than fixed traits. This matters for practice, because tutors can act on these signals. However, SHAP values describe the model's associations, not causal effects. A higher risk attributed to students from deprived areas does not mean that deprivation itself should be targeted by a tutor, and such attributions must be interpreted with care.

The fairness audit (RQ5) shows that a model with good average performance can still distribute its errors unevenly. Students from more deprived areas and with lower prior education are flagged more often, partly because their base rates are higher. Higher selection rates for these groups can be desirable if support is beneficial and non-stigmatising, but the higher FPR for women and the lower TPR for older students and for less deprived students mean that some groups receive more unnecessary contacts while at-risk students in other groups are more likely to be missed. These findings echo concerns raised by Gardner et al. (2019) and Baker and Hawn (2022) and show that fairness should be reported alongside accuracy for any early-warning model.

D. *Limitations*

Several limitations should be noted. First, OULAD covers one institution and the years 2013–2014, and learning behaviour may have changed since then. Second, the hyper-parameters selected at day 60 were reused at the other cutoffs. This isolates the effect of information availability but may slightly underestimate performance at the earliest and latest cutoffs. Third, 1,149 students in the test presentation had also appeared in a training presentation, through module repetition or enrolment in another module. Because the presentations are different, this is not leakage, but it reflects the real deployment population and could be examined in a sensitivity analysis. Fourth, some fairness subgroups (for example, students over 55 or with postgraduate qualifications) are small, so their estimates are uncertain. Fifth, the probabilities of the class-weighted models were not recalibrated. Finally, the UCI replication uses the same methodology on a different dataset rather than transferring a trained model, because the two datasets do not share a common feature space.

6. Implications for Research and Practice

A. *Implications for Research*

For researchers, the study shows that evaluation design can affect conclusions as much as the choice of algorithm. Future OULAD studies should (i) evaluate on a later, unseen presentation; (ii) compute features strictly from data available before the prediction day and remove students who have already withdrawn; (iii) group cross-validation folds by student; (iv) select hyper-parameters and thresholds without access to the test data; (v) report PR-AUC, MCC and recall for the at-risk class with confidence intervals; and (vi) support model comparisons with appropriate statistical tests. Explainability and fairness analyses should be routine rather than optional. The complete, openly available pipeline accompanying this paper is intended to serve as a reproducible baseline for such work.

B. *Implications for Practice*

For institutions, the results suggest a practical operating model. A first warning can be issued from enrolment data and pre-course activity, but it should be treated as low-confidence. A more reliable warning becomes possible by about day 60, after the first assessments. The decision threshold should be set according to the institution's intervention capacity. At the threshold used here, about 47% of students would be contacted, and six in ten of them would truly be at risk. A higher threshold would reduce the workload at the cost of recall. Each flag should be accompanied by an individual explanation that points to concrete behaviours, such as missed submissions or a long absence from the VLE, and should lead to supportive rather than punitive contact. Finally, group-wise error rates should be monitored after deployment, and group-specific thresholds or recalibration should be considered where unacceptable gaps appear.

7. Conclusion

This study developed and evaluated a leakage-free, explainable and fairness-aware framework for the early prediction of students at risk of failing or withdrawing. On OULAD, models trained on earlier presentations and tested on an unseen later presentation reached a ROC-AUC of 0.813 and a PR-AUC of 0.761 at day 60 and identified 75.3% of at-risk students. Gradient-boosting models and an MLP performed best, but the differences among them were small and partly non-significant. Predictive performance increased from a ROC-AUC of about 0.72 at course start to 0.90 at day 150. Assessment behaviour was the most informative information source, followed by VLE engagement and demographic attributes, and SHAP explanations highlighted actionable behaviours. The fairness audit revealed meaningful disparities in error rates across deprivation, education, age and gender groups. A replication on the UCI Student Performance data confirmed that early academic evidence outweighs demographic information. Together, these results provide a realistic benchmark and a reproducible methodology for building early-warning systems that are accurate, transparent

and accountable.

8. Future Research

Several directions follow from this work. First, the framework should be validated on more recent data from other institutions and countries, and with external validation across institutions. Second, sequence models and transformers that learn directly from daily clickstreams could be compared with engineered features under the same temporal protocol. Third, fairness-aware learning, post-processing with group-specific thresholds and probability recalibration should be evaluated to reduce the disparities identified here without sacrificing recall. Fourth, causal and counterfactual methods could distinguish modifiable risk factors from mere correlates and support the design of interventions. Finally, field studies that deploy the model with tutors are needed to establish whether earlier and explained warnings actually improve retention and achievement.

Data and Code Availability

Both datasets used in this study are publicly available. The Open University Learning Analytics Dataset (Kuzilek et al., 2017) can be obtained from the Open University (https://analyse.kmi.open.ac.uk/open_dataset) and from the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/349/open+university+learning+analytics+dataset). The UCI Student Performance dataset (Cortez and Silva, 2008) is available from the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/320/student+performance). No new data were collected for this study.

The complete source code is openly available on GitHub at https://github.com/miamdselim/Imbalance-Aware-Early-Prediction-of-At-Risk-Students. The repository contains a single Jupyter notebook that downloads both datasets, builds the leakage-free features, tunes and evaluates all models, performs the statistical, explainability and fairness analyses, and regenerates every table and figure reported in this paper. All experiments use a fixed random seed (42), and the software versions are listed in Section 3.10.

References

[1] M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. A. Raza, M. Abid, M. Bashir, and S. U. Khan, "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.

[2] N. R. Aljohani, A. Fayoumi, and S.-U. Hassan, "Predicting at-risk students using clickstream data in the virtual learning environment," *Sustainability*, vol. 11, no. 24, Art. no. 7238, Dec. 2019.

[3] K. E. Arnold and M. D. Pistilli, "Course Signals at Purdue: Using learning analytics to increase student success," in *Proc. 2nd Int. Conf. Learn. Anal. Knowl. (LAK '12)*, Vancouver, BC, Canada, 2012, pp. 267–270.

[4] R. S. Baker and A. Hawn, "Algorithmic bias in education," *Int. J. Artif. Intell. Educ.*, vol. 32, no. 4, pp. 1052–1092, Dec. 2022.

[5] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.

[6] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[7] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, Jul. 2010.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD '16)*, San Francisco, CA, USA, 2016, pp. 785–794.

[10] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, Art. no. 6, Jan. 2020.

[11] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, "Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS," *IEEE Trans. Learn. Technol.*, vol. 10, no. 1, pp. 17–29, Jan.–Mar. 2017.

[12] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in *Proc. 5th Future Bus. Technol. Conf. (FUBUTEC 2008)*, Porto, Portugal, 2008, pp. 5–12.

[13] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, Sep. 1988.

[14] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.

[15] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD '15)*, Sydney, NSW, Australia, 2015, pp. 259–268.

[16] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Stat. Assoc.*, vol. 32, no. 200, pp. 675–701, Dec. 1937.

[17] J. Gardner, C. Brooks, and R. Baker, "Evaluating the fairness of predictive student models through slicing analysis," in *Proc. 9th Int. Conf. Learn. Anal. Knowl. (LAK '19)*, Tempe, AZ, USA, 2019, pp. 225–234.

[18] D. Gašević, S. Dawson, T. Rogers, and D. Gašević, "Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success," *Internet High. Educ.*, vol. 28, pp. 68–84, Jan. 2016.

[19] C. C. Gray and D. Perkins, "Utilizing early engagement and machine learning to predict student outcomes," *Comput. Educ.*, vol. 131, pp. 22–32, Apr. 2019.

[20] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, Barcelona, Spain, 2016, pp. 3315–3323.

[21] Y. He, R. Chen, X. Li, C. Hao, S. Liu, G. Zhang, and B. Jiang, "Online at-risk student identification using RNN-GRU joint neural networks," *Information*, vol. 11, no. 10, Art. no. 474, Oct. 2020.

[22] M. Hlosta, Z. Zdrahal, and J. Zendulka, "Ouroboros: Early identification of at-risk students without models based on legacy data," in *Proc. 7th Int. Learn. Anal. Knowl. Conf. (LAK '17)*, Vancouver, BC, Canada, 2017, pp. 6–15.

[23] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979.

[24] S. M. Jayaprakash, E. W. Moody, E. J. M. Lauría, J. R. Regan, and J. D. Baron, "Early alert of academically at-risk students: An open source analytics initiative," *J. Learn. Anal.*, vol. 1, no. 1, pp. 6–47, 2014.

[25] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 4, Art. no. 15, Dec. 2012.

[26] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, CA, USA, 2017, pp. 3146–3154.

[27] H. Khosravi, S. Buckingham Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, and D. Gašević, "Explainable artificial intelligence in education," *Comput. Educ.: Artif. Intell.*, vol. 3, Art. no. 100074, 2022.

[28] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," *Appl. Artif. Intell.*, vol. 18, no. 5, pp. 411–426, May 2004.

[29] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics dataset," *Sci. Data*, vol. 4, Art. no. 170171, Nov. 2017.

[30] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017.

[31] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, CA, USA, 2017, pp. 4765–4774.

[32] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020.

[33] P. B. Nemenyi, "Distribution-free multiple comparisons," Ph.D. dissertation, Dept. Math., Princeton Univ., Princeton, NJ, USA, 1963.

[34] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Mach. Learn. (ICML '05)*, Bonn, Germany, 2005, pp. 625–632.

[35] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[36] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, Montréal, QC, Canada, 2018, pp. 6638–6648.

[37] M. Riestra-González, M. d. P. Paule-Ruíz, and F. Ortin, "Massive LMS log data analysis for the early prediction of course-agnostic student performance," *Comput. Educ.*, vol. 163, Art. no. 104108, Apr. 2021.

[38] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Min. Knowl. Discov.*, vol. 10, no. 3, Art. no. e1355, May/Jun. 2020.

[39] 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. no. e0118432, Mar. 2015.

[40] X. Sun and W. Xu, "Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves," *IEEE Signal Process. Lett.*, vol. 21, no. 11, pp. 1389–1393, Nov. 2014.

[41] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Rev. Educ. Res.*, vol. 45, no. 1, pp. 89–125, 1975.

[42] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, Art. no. 91, Feb. 2006.

[43] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Comput. Hum. Behav.*, vol. 104, Art. no. 106189, Mar. 2020.