Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When Rankings Remain Useful but Thresholds Fail: Temporal Portability of LMS-Based Student Early-Warning Models

Muhammad Zeeshan Babar

m.z.babar@leeds.ac.uk

University of Leeds

Saif ur Rahman

Xi'an Jiaotong-Liverpool University (XJTLU)

Research Article

Keywords: learning analytics, early warning systems, temporal generalisation, probability calibration, student risk prediction, capacity-based prioritisation

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11285072/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

When Rankings Remain Useful but Thresholds Fail: Temporal Portability of LMS-Based Student Early-Warning Models

Muhammad Zeeshan Babar[1*] and Saif ur Rahman[2†]

[1*]Department of Computer Science, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT, England, United Kingdom.
[2]Xi'an Jiaotong-Liverpool University (XJTLU), Suzhou, China.

*Corresponding author(s). E-mail(s): m.z.babar@leeds.ac.uk;
Contributing authors: Saif.rahman@xjtlu.edu.cn;
†These authors contributed equally to this work.

Abstract

Learning management system (LMS) data are widely used to identify students who may require academic support. However, models developed from historical cohorts may encounter behavioural change in later academic years. This study examines whether early warning policies remain reliable under temporal change. LMS data from Accountancy and Global Economics courses across three academic years were analysed using observation windows from Week 4 to Week 12. A logistic regression model based on activity intensity and active day regularity was evaluated using chronologically ordered cohorts. The analysis considered discrimination, probability calibration, and the portability of decision thresholds selected to match support capacity. Discrimination remained useful in later cohorts, while calibration deteriorated in a cohort-dependent manner. Historical thresholds did not preserve their intended workload when applied to later cohorts. At Week 12, a threshold targeting 20 percent of students selected 4.0 percent in Accountancy, but 89.4 percent and 69.0 percent in the two later Global Economics cohorts. Fixed-capacity ranking kept the number of selected students close to the intended support capacity and continued to identify groups with a higher concentration of risk of not passing on the first attempt than random selection. These results distinguish model portability from policy portability. Temporal evaluation of early warning systems should therefore consider both predictive performance and the reliability of the decision rule used to allocate student support.

Keywords: learning analytics, early warning systems, temporal generalisation, probability calibration, student risk prediction, capacity-based prioritisation

1 Introduction

Learning management systems record detailed information about how students engage with online learning. These records can include access to course materials, study sessions, active study days, and other forms of digital activity. Such data can support early identification of students who may experience academic difficulty. This is useful when support can still be offered before an important assessment or before disengagement becomes difficult to reverse.

A large body of learning analytics and educational data mining research has therefore focused on predicting student performance, withdrawal, and academic risk. Many studies compare machine learning models using measures such as accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve. These measures are useful, but they do not fully describe how a model will behave when it is deployed with a later group of students. Student behaviour can change between academic years. Course delivery can also change. As a result, patterns learned from historical data may not remain stable when the same prediction model is used with a future cohort.

Recent studies have examined how well student risk models hold up across academic years. Tiukhova et al. studied the stability of student success prediction models across academic years. They found substantial changes in learning behaviour, model predictions, and feature importance across course runs [1]. Their findings show that historical relationships between learning behaviour and academic outcomes should not be assumed to remain unchanged. Sonnleitner et al. reached a related conclusion when studying student performance prediction under concept drift. They showed that simpler models can remain competitive when student behaviour changes, while greater model complexity does not guarantee better performance in a shifted cohort [2]. Together, these findings support temporal evaluation using later cohorts rather than relying only on randomly mixed train and test splits.

Another concern is the reliability of predicted probabilities. A model can rank students reasonably well even when its probability estimates no longer represent the observed level of risk. This distinction matters in an early warning system because institutions may use predicted probabilities to decide which students should receive support. Utomo et al. used a forward cohort evaluation framework and showed that calibration can deteriorate in future cohorts even when discrimination remains competitive [3]. Schwerter et al. also examined probability calibration when student risk models were transferred across courses and institutions. Their results showed that useful discrimination can coexist with weaker calibration when the target context differs from the setting in which the model was developed [4]. These findings suggest that predictive performance and probability reliability should be evaluated separately.

Prediction is only one part of an early warning system. Institutions also need a rule for deciding which students can realistically be supported. Universities do not

normally have unlimited capacity for student support. A teaching team may only be able to contact a fixed proportion of students each week. This has motivated the use of ranking-based policies. Pei and Xing evaluated Top K risk ranking as a way to direct limited support resources towards students with the highest predicted risk [5]. More recently, Wu et al. examined calibrated early warning models under explicit capacity constraints. Their Top x percent policy fixed the number of students selected for outreach and made the trade-off between workload, precision, and recall visible [6]. This work shows that the practical value of a student risk model depends not only on predictive accuracy, but also on the rule used to convert predictions into action.

These strands of research answer several important questions. Temporal changes in student behaviour can affect model stability. Probability calibration can deteriorate under distribution shift. Capacity-based ranking can help institutions operate within limited support resources. A further deployment question remains less well understood. An institution may select a probability threshold using a historical cohort because that threshold produces an acceptable support workload. The same numerical threshold may then be used with the next academic cohort. If the distribution of predicted risk changes, the threshold may no longer select the intended proportion of students.

This is a practical problem rather than only a statistical one. A threshold that previously selected 20 percent of students may select far fewer students in a later cohort. It may also select most of the class. In both cases, the historical operating policy no longer matches the support capacity for which it was chosen. This issue concerns the temporal portability of the decision policy itself. It is different from asking only whether a classifier retains acceptable discrimination or whether its probabilities remain calibrated.

The present study investigates this problem using longitudinal learning management system data from successive academic cohorts. Student risk is defined as failure to pass the course examination on the first attempt. Prediction is evaluated at several points during the teaching period using information available up to Weeks 4, 6, 8, 10, and 12. This creates a realistic early warning setting in which future learning activity is not used at an earlier prediction point. The main model uses two simple measures of engagement behaviour. These represent activity intensity and the regularity of active study days. The model is evaluated using chronologically ordered cohorts rather than randomly mixed observations.

The study connects three parts of the deployment problem. First, it examines how learning behaviour, discrimination, and probability calibration change across academic cohorts and prediction windows. Second, it tests the temporal portability of an operating threshold. The threshold is selected from a preceding cohort so that the number of alerts matches a predefined support capacity. That numerical threshold is then applied unchanged to a later cohort. The analysis measures whether the resulting alert volume still matches the intended workload. Third, the study evaluates fixed capacity risk ranking when the historical probability threshold becomes operationally misaligned. The same modelling design is then replicated in a second course.

The study addresses the following research questions.

1. How do behavioural distributions, predictive discrimination, and probability calibration evolve across successive academic cohorts and progressively expanding early warning windows?
2. To what extent do probability thresholds selected from a preceding cohort to match predefined support capacity preserve their intended alert volume when transferred to a later academic cohort?
3. When historical probability thresholds become operationally misaligned, does fixed capacity risk ranking continue to concentrate risk of not passing on the first attempt within the available support capacity, and does this pattern replicate in a second course?

The contribution of this study lies in examining the temporal portability of the decision rule attached to an early warning model. The analysis separates two questions that are often considered together: whether the model continues to rank students usefully, and whether a probability threshold chosen in an earlier cohort still produces a feasible number of alerts. This distinction allows the temporal reliability of the predictive model and the operating policy to be evaluated separately.

2 Related Work

2.1 Early prediction of student risk from LMS data

The early identification of students who may experience academic difficulty has been studied for more than a decade. Early systems combined information about prior achievement, current grades, demographic characteristics, and activity recorded in learning management systems. Course Signals at Purdue was one of the early examples of this approach. It used academic and engagement information to provide students with timely feedback during a course [7]. The Open Academic Analytics Initiative later examined how similar early warning models could be used across several institutions and how predictive results could be linked with support for students [8].

Several studies then focused on improving the identification of students who were most likely to need support. Lakkaraju et al. proposed a machine learning framework for identifying students at-risk of adverse academic outcomes and evaluated the models using measures that were relevant to educational decision makers [9]. He et al. developed weekly models for identifying students at-risk of leaving a massive open online course [10]. Their work also recognised that predicted probabilities should be meaningful if they are to support intervention decisions. Marbouti et al. compared several classification methods for the early prediction of students at risk in an engineering course [11]. Their results showed that useful predictions could be obtained from information available during the course.

The availability of detailed virtual learning environment data further increased interest in behavioural prediction. The Open University Learning Analytics Dataset provided demographic information, assessment results, and large-scale records of student interactions with the virtual learning environment [12]. This dataset has supported a large number of studies on dropout and academic risk. Hlosta et al. proposed a self-learning approach that reduced the dependence on models trained using previous

course presentations [13]. Gardner and Brooks reviewed the wider literature on student success prediction in massive open online courses and identified several concerns related to model evaluation and practical use [14]. They noted that models should be evaluated using information that would actually be available when a prediction is made.

Later studies made greater use of clickstream data and progressive observation windows. Aljohani et al. used virtual learning environment clickstream information to predict students at risk [15]. Coussement et al. combined demographic, academic, cognitive, and behavioural information to predict dropout in an online learning environment [16]. Mubarak et al. used interaction logs for early dropout prediction and examined how behavioural information developed over time [17]. Adnan et al. evaluated prediction at several percentages of course completion and showed that performance generally improved as more information became available [18]. Pei and Xing later proposed a weekly prediction framework that incorporated information from previous weeks into the current prediction [5].

Taken together, these studies show that LMS activity can provide useful information for early student risk prediction. They also show that prediction performance changes as more information becomes available. However, much of this literature evaluates models within the available course data. It does not directly examine whether a decision rule developed in one academic cohort remains suitable when it is applied to a later cohort.

2.2 Temporal change and model generalisation

A common assumption in predictive modelling is that the data used for model development are representative of the data observed during deployment. This assumption may not hold when student populations and learning behaviour change over time. Dataset shift refers to changes between the distributions observed during model development and later use [19]. Concept drift provides a related view and considers changes in the statistical relationships that support prediction [20, 21]. If these changes are ignored, patterns learned from historical data may not remain valid.

The problem is particularly relevant in education because course delivery and student behaviour can change between academic years. Historical models may therefore perform differently when they are applied to a new cohort. Earlier work already raised concerns about the portability of learning analytics models. Jayaprakash et al. examined the portability of academic risk models across institutions [8]. Hlosta et al. developed their self-learning approach partly to avoid a strong dependence on legacy course data [13]. Gardner and Brooks also identified generalisation and realistic model evaluation as important issues in student success prediction [14].

More recent work has examined this problem directly across academic cohorts. Tiukhova et al. studied the stability of student success prediction models across three academic years [1]. Their analysis included changes in behavioural indicators, model predictions, predictive performance, and feature importance. The study used Accountancy and Global Economics courses and showed that learning behaviour can change substantially between course runs. Some indicators remained more stable than others, but historical patterns could not be assumed to remain unchanged.

Sonnleitner et al. examined early student performance prediction under concept drift [2]. Their results showed that more complex models did not consistently provide better generalisation when student behaviour changed. Simpler models could remain competitive in the shifted cohort. These findings support a careful evaluation of temporal reliability rather than relying only on model complexity or performance measured within the development data.

The present study builds on this work but considers a different operational question. The concern is not only whether predictive performance changes. It is also whether a decision policy selected from a previous cohort still produces a feasible support workload in a later cohort.

2.3 Probability calibration and decision thresholds

Many student prediction studies focus on discrimination. Measures such as AUC, precision, recall, and F1 score describe how well a model separates students with different outcomes. These measures do not show whether a predicted probability represents the observed level of risk. This distinction is important when predicted probabilities are used to make decisions.

The calibration of predictive models has been studied extensively in machine learning. Niculescu-Mizil and Caruana showed that classifiers with good discrimination can still produce distorted probability estimates [22]. Guo et al. later demonstrated that strong predictive models can be poorly calibrated and showed that post processing methods can improve probability estimates [23]. Van Calster et al. argued that calibration is an essential part of model evaluation whenever predicted risks are used for decision making [24].

This issue becomes more important under distribution shift. A model may preserve useful ordering of cases while the probability scale moves away from the observed outcome rate. Utomo et al. examined this problem in student retention prediction using a forward cohort evaluation design [3]. Their study separated training, validation, and test cohorts chronologically. The results showed that calibration could deteriorate in future cohorts even when discrimination remained competitive.

Schwerter et al. reported a related problem when student risk models were transferred across courses and institutions [4]. They examined both discrimination and calibration across weekly observation windows. They also considered the transfer of classification thresholds. A threshold selected in one context could become unsuitable in another context when the underlying risk distribution changed. Their results showed that threshold adjustment can improve classification performance when information about the target context is available.

These studies show that the probability scale deserves separate attention from ranking performance. They also suggest that a threshold should not automatically be treated as portable. In a real early warning setting, however, the final outcome rate of the new cohort is not known when support decisions must be made. This limits the use of approaches that require target outcome information for threshold adjustment.

2.4 Capacity-based student prioritisation

Student support is normally limited by the time and resources available to instructors or support teams. For this reason, several studies have considered not only who is predicted to be at-risk, but also which students should be prioritised when support cannot be provided to everyone.

Lakkaraju et al. developed their prediction framework with the practical needs of educational decision makers in mind [9]. Hlosta et al. used Top K measures when evaluating the identification of students who were most likely to require attention [13]. Pei and Xing examined Top K precision and recall across different weeks of a course [5]. They linked the value of K to the amount of learning support that could realistically be provided. Their analysis showed that the useful value of K can change with the course and the stage of prediction.

Wu et al. extended this idea by evaluating explicit capacity-based alerting policies [6]. They examined Top 10, Top 20, and Top 30 percent policies and reported the resulting trade-off between precision and recall. This makes the support workload explicit. It also avoids the assumption that every student above a fixed probability threshold can be contacted. Capacity-based ranking therefore provides a practical way to control the number of students selected for support. Its workload is fixed by design. The important empirical question is whether the students selected within that capacity still contain a useful concentration of academic risk.

2.5 Research gap

Previous work has established that LMS activity can support early identification of students at risk [5, 10, 15, 18], while research on dataset shift and concept drift has shown that predictive relationships may change over time [19–21]. Studies in learning analytics have also shown that student behaviour and model performance can change across cohorts and educational contexts [1, 2, 4]. More recent studies have also drawn attention to probability calibration and to the use of capacity-based decision policies in student support [3, 6].

What is less clear is how these issues interact during forward deployment. In practice, an institution may choose a probability threshold because it produces a manageable number of alerts in one academic cohort and then reuse the same threshold in the following year. If the distribution of student behaviour or predicted risk changes, that threshold may no longer generate a workload that matches the available support capacity.

This study examines that problem directly. A probability threshold is selected from the preceding cohort so that its alert rate approximately matches a predefined support capacity. The same numerical threshold is then applied unchanged to the later cohort without using later outcomes to adjust it. The resulting alert volume is compared with a fixed-capacity ranking policy, where the number of selected students is controlled directly. The focus is therefore not on proposing a new classification algorithm or a new ranking method. The contribution lies in examining whether an operating rule that was workable in one academic cohort remains workable in the next, even when the underlying model continues to provide useful ranking performance.

3 Methodology

3.1 Dataset and course cohorts

The study used the public KU Leuven learning analytics dataset described by Tiukhova et al. [25]. The dataset contains student activity and academic performance records from two first-year bachelor courses. These courses are Accountancy and Global Economics. Data are available for three academic years from 2018/19 to 2020/21. The dataset combines course participation records with detailed activity logs from the Blackboard learning management system. Each activity record contains a pseudonymous student identifier, a course identifier, a content identifier, a session identifier, and a timestamp. The participation files contain examination outcomes for each course registration. The public data also include course content and discussion forum information. The present study uses the course participation records and the detailed LMS activity logs.

The student identifiers are pseudonymous. They cannot be used to track the same person across academic years or courses. The analysis therefore treats each course and academic year as a separate cohort. It does not attempt to construct longitudinal records for individual students.

Exact duplicate rows were removed from the course participation files before feature construction. Students with no LMS activity were retained. This choice reflects the intended early warning setting. A registered student with no recorded activity may still be relevant to a support decision. Table 1 summarises the cohorts used in the analysis. The risk rate is the proportion of students who did not pass the course examination on the first attempt.

Table 1 Course cohorts used in the study.

Course	Academic year	Students	Risk rate
Accountancy	2018/19	752	0.464
Accountancy	2019/20	705	0.445
Accountancy	2020/21	723	0.368
Global Economics	2018/19	743	0.541
Global Economics	2019/20	681	0.383
Global Economics 1	2020/21	341	0.499
Global Economics 2	2020/21	335	0.600

The 2020/21 Global Economics data contain two separate course identifiers. These are recorded as Global Economics 1 and Global Economics 2. There is no overlap between the registered students in these two groups. Both groups share the same course calendar metadata. They are therefore evaluated as separate cohorts and are not pooled.

3.2 Prediction target and observation windows

The prediction target is first-attempt examination risk. The participation data provide a binary variable called `PASSED_FIRST_ATTEMPT`. The target used in this study is defined as

$$Y_i = 1 - \texttt{PASSED_FIRST_ATTEMPT}_i, \tag{1}$$

where $Y_i = 1$ identifies a student who did not pass on the first attempt. A value of $Y_i = 0$ identifies a student who passed on the first attempt. Features are constructed using progressively expanding observation windows. The cutoffs are Weeks 4, 6, 8, 10, and 12. Each window begins at Week 2. The Week 4 model therefore uses activity from Weeks 2 to 4. The Week 6 model uses activity from Weeks 2 to 6. The same rule is applied to the later cutoffs.

Week 1 was excluded because the 2019/20 Accountancy activity log does not contain LMS events for that week. Using Week 2 as the common starting point avoids giving one Accountancy cohort a longer observation period than another. The same observation rule was later retained for Global Economics. This keeps the feature design fixed during the cross-course replication.

Only activity observed within the relevant window is used to construct a prediction. Activity recorded after the cutoff is excluded. The target outcome is taken from the later examination result. This separation prevents later LMS behaviour from entering an earlier prediction window.

3.3 Feature construction

A broader set of LMS activity features was examined during the Accountancy development stage. These included total actions, number of sessions, number of unique content items, active days, active weeks, actions per session, and related proportions. Several of these variables contained similar information. Strong correlations were also observed between some activity measures.

The final model therefore uses two behavioural features. The first is `ACTIONS_PER_ACTIVE_DAY`. It measures activity intensity on days when a student used the LMS. It is calculated as

$$\text{Actions per active day}_i = \frac{\text{Total actions}_i}{\text{Active days}_i}. \tag{2}$$

The value is set to zero when no activity is observed in the window. The second feature is `ACTIVE_DAY_PROPORTION`. It measures the regularity of activity across the observation period. It is defined as

$$\text{Active day proportion}_i = \frac{\text{Active days}_i}{\text{Available days in the observation window}}. \tag{3}$$

The two variables capture different aspects of engagement. One represents intensity when the student is active. The other represents how regularly the student appears in the LMS. The final feature set was frozen before the Global Economics replication was carried out.

3.4 Model development

Accountancy was used as the main development setting. Several candidate feature representations were examined during this stage. For the model-family comparison, logistic regression, XGBoost, and CatBoost were evaluated using the five-feature relative behavioural representation described in Appendix A. The models were evaluated using fixed configurations rather than separately optimised hyperparameters. The purpose of the comparison was to assess whether additional model complexity provided a consistent advantage under temporal evaluation. The model configurations and comparison results are reported in Appendix A.

The final analysis uses logistic regression with the two features defined in Section 3.3. More complex models did not provide a consistent advantage across the later Accountancy cohort. Logistic regression was therefore retained because it provided competitive temporal performance while allowing a direct interpretation of the two behavioural features. The predictors were standardised using parameters estimated from the training cohort. The same transformation was then applied to the validation and evaluation cohorts. Model fitting used only the training cohort for the relevant experiment.

The 2020/21 Accountancy cohort was inspected during model and feature development. It is therefore treated as a later evaluation cohort rather than an untouched final test set. The Global Economics analysis was conducted after the feature and model design had been fixed. This provides a stronger replication of the final design in a second course.

3.5 Temporal evaluation design

The evaluation follows the chronological order of the academic cohorts. For Accountancy, the 2018/19 cohort is used for model training. The 2019/20 cohort is used for validation. The 2020/21 cohort is used for later temporal evaluation. The same structure is used for Global Economics. The model is trained on Global Economics 2018/19 and validated on Global Economics 2019/20. It is then evaluated separately on Global Economics 1 and Global Economics 2 from 2020/21.

The Global Economics analysis is a replication of the feature and model design. It is not a direct transfer of the fitted Accountancy coefficients. A new logistic regression model is fitted using the Global Economics 2018/19 cohort. This distinction allows the study to test whether the same simple representation and evaluation procedure produce a similar pattern in another course. The temporal design is repeated at every observation cutoff. A separate model is fitted for Weeks 2 to 4, Weeks 2 to 6, Weeks 2 to 8, Weeks 2 to 10, and Weeks 2 to 12. No future cohort is used to fit the model for an earlier cohort.

3.6 Evaluation of discrimination and calibration

Model performance is assessed using both discrimination and calibration. Discrimination is measured with ROC AUC and PR AUC. ROC AUC measures the ability of the model to rank a randomly selected at-risk student above a randomly selected student

who passed on the first attempt. PR AUC is also reported because the proportion of at-risk students changes across cohorts.

Calibration is assessed using the Brier score, the difference between mean predicted risk and observed risk, and calibration intercept and slope. The mean predicted risk is compared with the observed risk rate in each cohort. Their difference is reported as the calibration gap

$$\text{Calibration gap} = \bar{\hat{p}} - \bar{y}. \tag{4}$$

A negative value indicates that the model predicts less risk than is observed in the cohort. A positive value indicates higher predicted risk than the observed rate. Calibration intercept and calibration slope are also estimated as diagnostic measures. These quantities are fitted after outcomes are observed in the evaluation cohort. They are used to describe the form of miscalibration. They are not used to recalibrate predictions during deployment.

3.7 Threshold transfer and capacity-based prioritisation

The main deployment experiment compares two ways of converting predicted risk into a support decision. The first uses a numerical probability threshold. The second uses a fixed capacity ranking.

For the threshold policy, a separate numerical threshold was selected at each observation window using predictions from the validation cohort. For a target support capacity of 20 percent, the intended number of alerts was calculated as the ceiling of 0.20 times the validation cohort size. Students were ordered by predicted risk, and the threshold was set equal to the lowest predicted probability within this highest-risk group. Students with predicted risk greater than or equal to this threshold were flagged. The resulting validation alert rate was therefore approximately 20 percent, with small deviations possible because of rounding or tied predicted probabilities. The selected threshold was then frozen for that observation window.

The frozen threshold for each observation window was applied unchanged to the corresponding 2020/21 cohort. No 2020/21 outcome information was used to select or alter the threshold. The proportion of students flagged in the later cohort was then measured. The difference between this proportion and the intended capacity is defined as

$$\text{Capacity deviation} = \text{Observed flagged proportion} - \text{Target capacity}. \tag{5}$$

The capacity ratio is defined as

$$\text{Capacity ratio} = \frac{\text{Observed flagged proportion}}{\text{Target capacity}}. \tag{6}$$

A ratio of one indicates that the historical threshold produces the intended workload. A ratio below one indicates unused support capacity. A ratio above one indicates that the number of alerts exceeds the intended capacity. The fixed-capacity policy does not reuse a numerical probability threshold. Students in the later cohort are ranked by

predicted risk, and the number selected is the ceiling of 0.20 times the cohort size. The corresponding highest-risk students are then flagged. This keeps the support workload as close as possible to the intended 20 percent capacity given the discrete cohort size. The quality of the selected group is evaluated using precision, recall, and lift. Lift compares the risk rate within the selected group with the overall risk rate in the cohort

$$\text{Lift} = \frac{\text{Precision within selected group}}{\text{Overall risk prevalence}}. \tag{7}$$

A lift above one indicates that the selected group contains a higher concentration of first-attempt risk than random selection.

3.8 Uncertainty and sensitivity analysis

Bootstrap resampling is used to quantify uncertainty in the main later cohort results. Two thousand nonparametric bootstrap samples are drawn from each evaluation cohort. The fitted model is kept fixed during this procedure. ROC AUC, PR AUC, precision, recall, and lift are recalculated in each bootstrap sample. The 2.5 and 97.5 percentiles are used to form 95 percent confidence intervals. Because the fitted model was held fixed, these intervals reflect sampling uncertainty in the evaluation cohort only and do not incorporate uncertainty arising from model fitting.

A second sensitivity analysis changes the historical training period. The model is trained using both prior academic cohorts and is then evaluated on the 2020/21 cohort. This analysis tests whether the main findings depend strongly on using only the earliest cohort for model fitting. The main threshold transfer analysis assumes a 20 percent support capacity. A final sensitivity analysis repeats the Week 12 threshold experiment at 10, 20, and 30 percent capacity. For each capacity, the threshold is selected using the 2019/20 validation cohort and is then transferred unchanged to 2020/21. This analysis tests whether the observed workload instability depends on the particular choice of 20 percent.

4 Results

4.1 Changes in cohort risk and LMS behaviour

The proportion of students at risk changed across academic years. In Accountancy, the first-attempt risk rate decreased from 0.464 in 2018/19 to 0.445 in 2019/20 and 0.368 in 2020/21. The pattern was different in Global Economics. The risk rate was 0.541 in 2018/19 and 0.383 in 2019/20. In 2020/21, the two Global Economics cohorts had risk rates of 0.499 and 0.600. The cohort sizes and outcome rates are reported in Table 1.

The LMS activity distributions also changed substantially over time. The largest shift was observed in Accountancy 2020/21. At Week 12, the median number of actions per active day was 19.64 for students who passed on the first attempt in 2018/19. The corresponding median was 25.61 in 2019/20 and 65.40 in 2020/21. For at-risk students, the median increased from 17.38 to 23.49, then to 53.98.

A similar change was observed in the regularity of LMS activity. At Week 12, the median active day proportion among students who passed was 0.377 in 2018/19, 0.403

in 2019/20, and 0.571 in 2020/21. For at-risk students, the corresponding values were 0.260, 0.286, and 0.409.

In each Accountancy cohort, students who passed on the first attempt had higher median activity intensity and active-day proportion than students who did not pass on the first attempt. At the same time, the absolute levels of both behavioural measures changed substantially across academic years. The broad descriptive ordering between the two outcome groups was therefore similar, while the distributions of the behavioural predictors shifted over time.

4.2 Temporal discrimination across prediction windows

Figure 1 shows ROC AUC across the later evaluation cohorts. In Accountancy 2020/21, ROC AUC increased from 0.694 at Week 4 to 0.770 at Week 12. PR AUC increased from 0.554 to 0.664 over the same period.

Global Economics showed a similar but weaker increase in discrimination. For Global Economics 1, ROC AUC increased from 0.682 at Week 4 to 0.737 at Week 12. PR AUC increased from 0.713 to 0.768. For Global Economics 2, ROC AUC remained close to 0.68 across the observation windows. It was 0.672 at Week 4 and 0.683 at Week 12. PR AUC increased from 0.736 to 0.776.

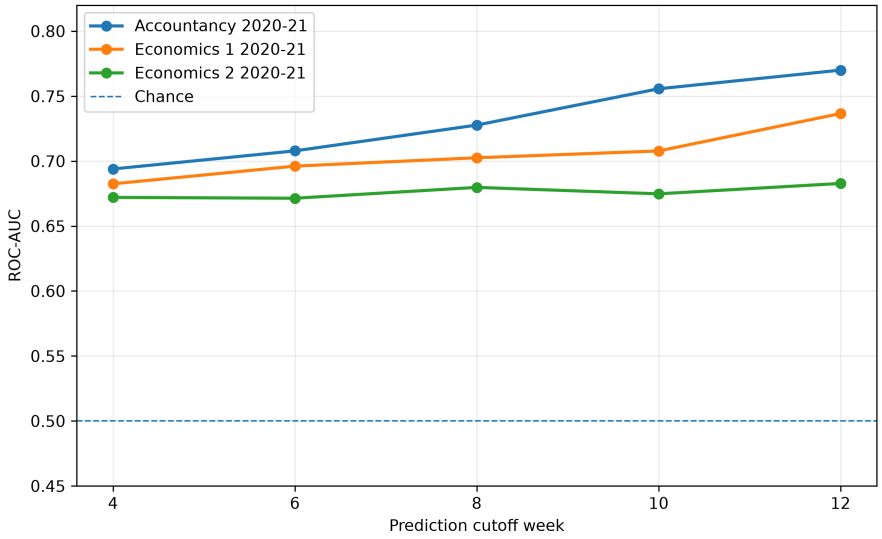


Fig. 1 Temporal discrimination across the later evaluation cohorts. The models were trained on the earliest cohort within each course.

Table 2 reports the Week 12 results with bootstrap confidence intervals. The lower confidence bound for ROC AUC remained above 0.5 in all three evaluation cohorts. The lower confidence bound for PR AUC also remained above the corresponding outcome prevalence.

Table 2 Week 12 predictive performance in the later evaluation cohorts.

Cohort	Risk rate	ROC AUC	ROC 95% CI	PR AUC	PR 95% CI
Accountancy 2020/21	0.368	0.770	0.733 to 0.807	0.664	0.600 to 0.728
Global Economics 1 2020/21	0.499	0.737	0.683 to 0.790	0.768	0.701 to 0.828
Global Economics 2 2020/21	0.600	0.683	0.625 to 0.741	0.776	0.716 to 0.836

4.3 Probability calibration under temporal change

The discrimination results were not matched by stable probability calibration. Figure 2 shows the difference between mean predicted risk and observed risk in each later cohort.

For Accountancy 2020/21, the observed risk rate was 0.368. The mean predicted risk was 0.344 at Week 4 and decreased to 0.174 at Week 12. The calibration gap therefore changed from minus 0.024 to minus 0.194. The Brier score increased from 0.210 to 0.230 while ROC AUC improved over the same period. For Global Economics 1, the calibration gap changed from minus 0.020 at Week 4 to minus 0.133 at Week 12. The same pattern was stronger in Global Economics 2. The calibration gap changed from minus 0.097 to minus 0.304. The direction and size of miscalibration therefore differed across cohorts.

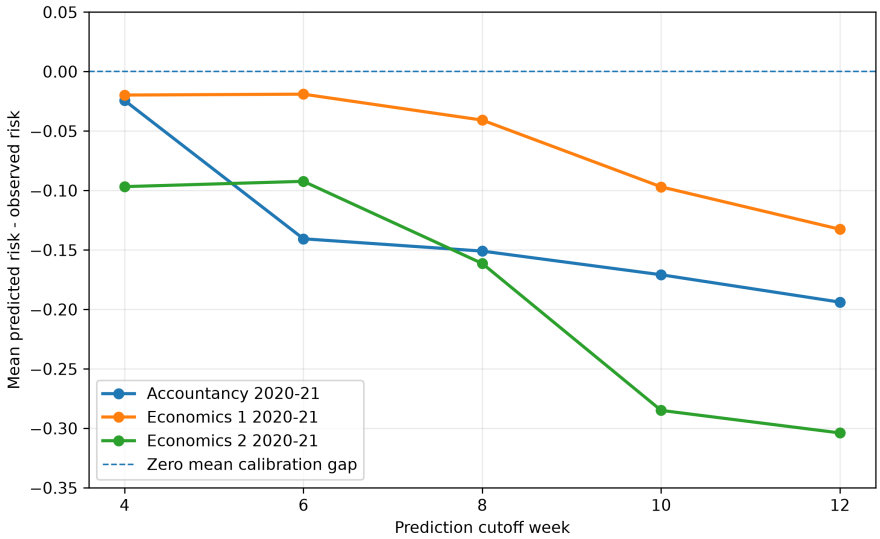


Fig. 2 Calibration gap across prediction windows in the later evaluation cohorts. The gap is defined as mean predicted risk minus observed risk.

The calibration intercept and slope provided the same general picture. At Week 12, the Accountancy 2020/21 calibration intercept was 1.385, and the slope was 1.086.

The positive intercept indicates systematic underprediction of risk. The slope remained close to one. This suggests that the main calibration problem was a shift in the overall probability level rather than a complete loss of ordering.

The Global Economics cohorts showed a more mixed pattern. At Week 12, the calibration intercept and slope were 0.893 and 1.457 for Global Economics 1. They were 1.308 and 0.872 for Global Economics 2. The results therefore indicate cohort-specific calibration changes rather than one common form of temporal miscalibration.

4.4 Temporal portability of probability thresholds

The central deployment experiment examined whether a probability threshold selected in the validation cohort continued to match the intended support capacity in 2020/21. The main analysis used a target capacity of 20 percent.

The selected thresholds matched the intended capacity in the validation cohort by design. Their behaviour changed markedly when they were transferred to the later cohort. At Week 12, the Accountancy threshold selected only 4.0 percent of the 2020/21 cohort. The intended capacity was 20 percent. The resulting capacity ratio was 0.201. The direction of the mismatch was reversed in Global Economics. The transferred Week 12 threshold selected 89.4 percent of Global Economics 1 and 69.0 percent of Global Economics 2. The corresponding capacity ratios were 4.472 and 3.448.

Figure 3 shows how the flagged proportion changed across prediction windows. The fixed capacity policy remained close to 20 percent by construction. In contrast, the workload generated by the transferred probability threshold varied strongly across both time and cohort.

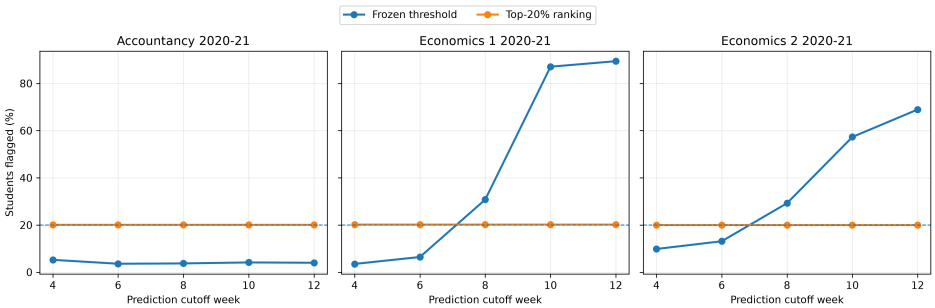


Fig. 3 Alert volume produced by a transferred probability threshold and by fixed capacity ranking. The fixed-capacity policy selects approximately 20 percent of each cohort by construction, subject to integer rounding.

Table 3 compares the Week 12 operating policies. In Accountancy, the transferred threshold produced high precision but very low recall because only a small part of the available support capacity was used. In the two Global Economics cohorts, the threshold selected a large majority of students. Recall was high, but lift moved much closer to one.

Table 3 Week 12 comparison of the transferred threshold and fixed 20 percent capacity ranking.

Cohort	Policy	Flagged	Precision	Recall	Lift
Accountancy	Transferred threshold	0.040	0.828	0.090	2.249
Accountancy	Top 20 percent	0.201	0.752	0.410	2.043
Global Economics 1	Transferred threshold	0.894	0.521	0.935	1.046
Global Economics 1	Top 20 percent	0.202	0.884	0.359	1.773
Global Economics 2	Transferred threshold	0.690	0.675	0.776	1.126
Global Economics 2	Top 20 percent	0.200	0.896	0.299	1.493

4.5 Performance of fixed capacity prioritisation

The fixed capacity policy continued to concentrate risk above the cohort baseline. At Week 12, the Top 20 percent group in Accountancy had a precision of 0.752 and a recall of 0.410. The corresponding lift was 2.043. This means that the selected group contained slightly more than twice the risk concentration expected from random selection.

The same pattern was observed in Global Economics. Global Economics 1 produced a precision of 0.884, recall of 0.359, and lift of 1.773. Global Economics 2 produced a precision of 0.896, recall of 0.299, and lift of 1.493. Bootstrap intervals supported the same conclusion. In Accountancy, the 95 percent confidence interval for lift was 1.816 to 2.241. The corresponding interval was 1.618 to 1.995 for Global Economics 1 and 1.335 to 1.641 for Global Economics 2.

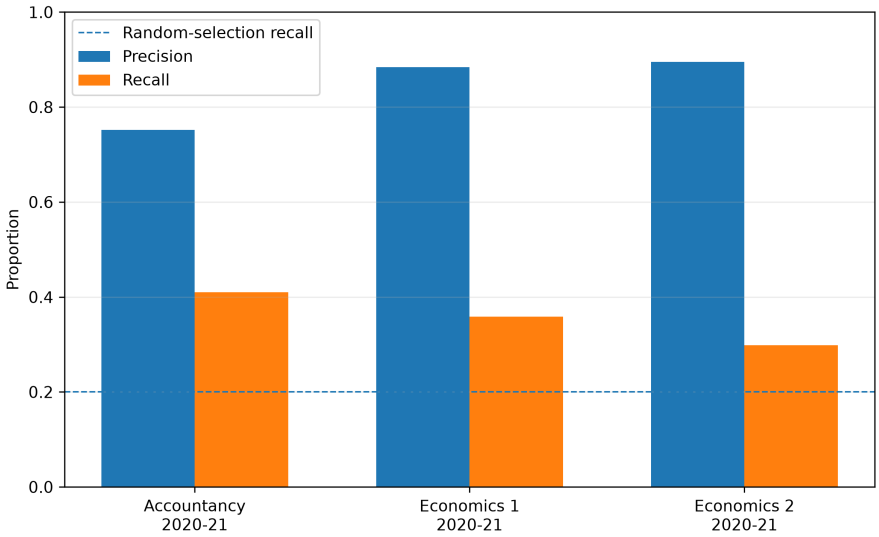


Fig. 4 Precision and recall for fixed 20 percent capacity prioritisation. The random selection recall level is 0.20.

These results do not imply that the fixed capacity policy is unaffected by temporal change. The workload is stable because the number of selected students is fixed. The empirical result is that the selected group continued to contain a higher concentration of first-attempt risk than random selection.

4.6 Sensitivity analyses

The main findings were similar when both earlier academic cohorts were used for model training. At Week 12, Accountancy produced a ROC AUC of 0.772 and a PR AUC of 0.662. The corresponding values were 0.727 and 0.764 for Global Economics 1. They were 0.696 and 0.789 for Global Economics 2.

The probability level remained lower than the observed risk rate in all three later cohorts. Mean predicted risk was 0.194 in Accountancy compared with an observed rate of 0.368. It was 0.434 in Global Economics 1 compared with 0.499. It was 0.427 in Global Economics 2 compared with 0.600. Using an additional prior cohort therefore did not remove the calibration shift.

The fixed capacity results also changed only modestly. At 20 percent capacity, lift was 1.968 in Accountancy, 1.773 in Global Economics 1, and 1.443 in Global Economics 2. The threshold portability result was also tested at 10, 20, and 30 percent capacity. Table 4 reports the Week 12 capacity ratios. Accountancy consistently used much less capacity than intended. Both Global Economics cohorts consistently generated more alerts than the intended capacity.

Table 4 Week 12 sensitivity of transferred threshold workload across support capacities.

Cohort	10% capacity	20% capacity	30% capacity
Accountancy 2020/21	0.235	0.201	0.189
Global Economics 1 2020/21	4.604	4.472	3.304
Global Economics 2 2020/21	3.075	3.448	3.154

The direction of threshold failure was therefore unchanged across the three capacity assumptions. The Accountancy threshold remained too restrictive. The Global Economics threshold remained too permissive.

5 Discussion

5.1 Temporal change does not affect every part of the model in the same way

The results show that temporal change can affect different parts of an early warning system in different ways. Student behaviour changed substantially across academic years, especially in the later Accountancy cohort. At the same time, the two-feature logistic regression model retained useful discrimination. This pattern is important

because it shows that a model can remain useful for ranking students even when the scale of the underlying behavioural variables changes.

This finding is consistent with earlier work on temporal instability in learning analytics. Tiukhova et al. showed that student behaviour, prediction patterns, and feature importance can change across repeated course deliveries [1]. Sonnleitner et al. also showed that concept drift can reduce the advantage of more complex models and that simpler models can remain competitive under changing conditions [2]. The present results support both observations. The scale of LMS behaviour changed strongly, but a simple model based on engagement intensity and regularity continued to separate higher and lower risk students.

The result should not be interpreted as evidence that the predictive relationship was fully stable. The probability scale changed substantially, and the analysis does not establish that the conditional relationship between the predictors and the outcome remained unchanged across cohorts. The clearest empirical finding is that the behavioural predictor distributions changed markedly while useful ranking performance was retained. This pattern is compatible with temporal covariate shift, but the analysis does not establish that covariate shift alone explains the observed changes.

The use of progressively expanding observation windows adds another point. In Accountancy, discrimination improved as more activity became available. A similar but smaller improvement was observed in Global Economics 1. Global Economics 2 remained more stable across the observation windows. This suggests that additional behavioural information can improve ranking in some cohorts, but the benefit is not uniform across courses or years. The timing of an early warning model should therefore be evaluated within the deployment context rather than assumed to improve in the same way for every cohort.

5.2 Useful discrimination can coexist with poor probability calibration

A central finding is that discrimination and calibration moved in different directions. In Accountancy 2020/21, ROC AUC improved across the observation windows while the mean predicted risk moved further below the observed risk rate. The same broad pattern appeared in both Global Economics cohorts, although the size of the calibration error differed.

This distinction is well established in predictive modelling. A model may rank cases correctly while producing probability estimates that are systematically too high or too low [22–24]. The current results show why this distinction matters in learning analytics. If the model is used only to rank students, useful ordering may remain. If the predicted probabilities are used directly to determine support decisions, miscalibration becomes an operational problem.

The findings also align with the forward cohort results reported by Utomo et al. [3]. Their study showed that calibration can deteriorate under temporal change even when discrimination remains useful. The present study adds a different perspective by examining what this means for the decision rule attached to the model.

The form of calibration change was not identical across cohorts. In Accountancy, the Week 12 calibration slope remained close to one while the intercept shifted strongly.

This indicates that the main problem was a change in the overall probability level. Global Economics showed a more mixed pattern. This variation is important. It suggests that temporal miscalibration should not be treated as a single process that will appear in the same form across courses.

5.3 Historical probability thresholds were not temporally portable

The most important operational result concerns the transfer of a probability threshold from one academic cohort to the next. The threshold was chosen in the validation cohort so that the alert rate matched a known support capacity. It was then applied unchanged to the later cohort.

The resulting workloads changed sharply. In Accountancy, the transferred threshold became too restrictive and selected only a small fraction of the intended number of students. In Global Economics, the same procedure produced the opposite problem. The transferred threshold selected a large majority of the later cohort.

These results show that a threshold can lose its operational meaning even when the underlying model still provides useful ranking. The observed calibration and score distribution changes were accompanied by severe changes in alert volume. The analysis does not establish that calibration drift alone caused the threshold failure. The safer interpretation is that changes in behaviour and predicted risk distributions were associated with poor temporal portability of the historical operating threshold.

Schwerter et al. reported a related problem in cross-course prediction [4]. They showed that a threshold selected in one context can become unsuitable in another context and that threshold adjustment can improve performance when target prevalence is available. The present study considers a complementary deployment setting. The future cohort outcome rate is not known when support decisions are made. A historical threshold must therefore operate without access to the later outcome distribution.

This distinction is practically important. A probability threshold is often treated as if it were a stable part of the model. The present results suggest that it should instead be treated as a policy parameter that may require separate validation. A threshold that produced a manageable workload in one cohort may not preserve that workload in the next.

5.4 Fixed capacity ranking provided workload control

The fixed capacity policy behaved differently because the number of selected students was controlled directly. Its workload stability is therefore a consequence of the policy design and should not be interpreted as an empirical advantage of the model.

The empirical finding lies elsewhere. The students selected within the fixed capacity group continued to contain a higher concentration of first-attempt risk than would be expected from random selection. This was observed in Accountancy and in both Global Economics cohorts. The bootstrap intervals for lift remained above one in all three later cohorts.

This result is consistent with earlier work on resource-constrained student prioritisation. Lakkaraju et al. considered prediction in relation to the needs of educational

decision makers [9]. Hlosta et al. used Top K measures when evaluating at-risk identification [13]. Pei and Xing linked Top K selection directly to the amount of support available to instructors [5]. Wu et al. later evaluated explicit Top 10, Top 20, and Top 30 percent policies [6].

The present study does not claim that capacity-based ranking is a new idea. Its contribution is to compare this policy with a historical probability threshold under forward temporal change. The results suggest that fixed capacity ranking can remain operationally useful when the probability scale is unstable. The institution controls the workload directly, while the model is used only to order students within that workload.

The sensitivity analysis provides further support for this interpretation. The qualitative threshold failure remained at 10, 20, and 30 percent capacity. Accountancy continued to underuse the intended capacity, while both Global Economics cohorts continued to exceed it. The result was therefore not limited to the illustrative 20 percent setting.

5.5 Replication across courses

The Global Economics analysis was designed as a replication of the feature and model design rather than a direct transfer of the Accountancy model. This distinction matters. The coefficients were estimated again using the earliest Global Economics cohort. The purpose was to test whether the same modelling approach produced a similar deployment pattern in another course.

The replication produced both similarities and differences. Useful discrimination was retained in both Global Economics cohorts. The direction of the two feature coefficients was also consistent with Accountancy. More regular activity was associated with lower first-attempt risk, and greater activity intensity was generally associated with lower risk.

The calibration behaviour was less consistent. The size of underprediction differed substantially between Global Economics 1 and Global Economics 2. This reinforces the view that probability reliability is context-dependent. A model design can remain useful for ranking while the operating scale of its predictions changes across cohorts.

The threshold transfer result was nevertheless clear in both Global Economics cohorts. The historical threshold generated substantially more alerts than the intended capacity. This is important because it shows that the main operational finding was not restricted to the Accountancy development setting.

The replication should still be interpreted cautiously. It does not demonstrate that the same fitted model can be transferred directly across courses. It shows that the same simple feature representation, model class, and temporal evaluation procedure produced a similar pattern in a second course family.

5.6 Implications for early warning system design

The results suggest that temporal monitoring should extend beyond discrimination. A model may continue to rank students reasonably well while its probability scale

changes enough to make a previously selected threshold operationally unsuitable. Calibration therefore becomes particularly important when predicted probabilities are used directly to decide which students should receive support.

The information available when predictions are used also needs to be considered. The outcome prevalence of a new cohort is not known when early support decisions are being made. Approaches that require outcome information from the target cohort for threshold adjustment may therefore be difficult to use at that point. Where the number of students who can realistically be supported is known, capacity-based ranking offers a more direct way of incorporating this constraint into the decision rule.

This does not remove the need for continued model monitoring. Ranking performance may also deteriorate as student behaviour changes. The practical implication is that the predictive model and the rule used to turn its predictions into support decisions should be monitored separately as new cohorts become available. A model may remain useful for prioritisation even when the operating threshold used in an earlier cohort no longer produces a suitable workload.

5.7 Limitations and future work

The findings should be considered in light of the scope of the study. The analysis covers two course families from one institution. Although the main pattern was observed in both Accountancy and Global Economics, broader validation is needed before the findings can be generalised to other subjects, institutions, or learning platforms. Future work should therefore examine whether the same threshold portability problem appears across a wider range of educational settings.

The Accountancy analysis also informed the development of the final feature set. The Global Economics analysis provides the stronger replication because the feature and model design was fixed before that analysis was carried out. A useful next step would be to evaluate the full procedure prospectively on later cohorts without further model development. The present study also focuses on prediction and support prioritisation rather than the effect of the support itself. Future research could combine the proposed capacity-based policy with actual student support interventions and examine whether improved targeting leads to better academic outcomes. It would also be useful to study adaptive recalibration and threshold updating when suitable information from a new cohort becomes available.

6 Conclusion

This study examined whether an early warning decision rule developed for one academic cohort remains workable when it is carried forward to the next. The results show that this cannot be assumed. Student LMS behaviour changed substantially across academic years, yet the two-feature logistic regression model continued to provide useful discrimination in the later cohorts. The probability scale was less stable, however, and this had important consequences for the use of historical decision thresholds.

The clearest evidence came from the threshold transfer experiment. At Week 12, a threshold selected to flag approximately 20 percent of the validation cohort flagged only 4.0 percent of the later Accountancy cohort, but 89.4 percent and 69.0 percent

of the two later Global Economics cohorts. The direction of the mismatch differed between courses, but in each case the historical threshold no longer represented the support workload for which it had originally been selected. This pattern also remained when the assumed support capacity was changed to 10 percent, 20 percent, and 30 percent.

Fixed-capacity ranking avoided this workload instability because the number of students selected was controlled directly. More importantly, the students prioritised within that capacity continued to show a higher concentration of risk of not passing on the first attempt than would be expected from random selection. This does not mean that ranking performance will always remain stable, and continued monitoring is still required. It does show that model portability and policy portability are not the same problem. A model may remain useful for ordering students by risk while a probability threshold taken from an earlier cohort becomes impractical for allocating support. Early warning systems should therefore evaluate not only whether predictions remain informative, but also whether the decision rule attached to those predictions remains suitable for the cohort in which it is being used.

Appendix A Model comparison during development

During the Accountancy development stage, logistic regression, XGBoost, and CatBoost were compared using a five-feature relative behavioural representation. The features were `ACTIONS_PER_SESSION`, `ACTIONS_PER_ACTIVE_DAY`, `ACTIVE_DAY_PROPORTION`, `ACTIVE_WEEK_PROPORTION`, and `ZERO_ACTIVITY`. These variables represented engagement intensity, regularity, and absence of LMS activity without relying directly on raw activity counts.

The candidate models were evaluated using fixed configurations across all observation windows. Logistic regression used standardised predictors with a maximum of 1000 iterations and a random seed of 42. XGBoost used 300 trees, a learning rate of 0.03, maximum tree depth of 3, subsampling of 0.8, and column subsampling of 0.8. CatBoost used 300 iterations, a tree depth of 4, and a learning rate of 0.03. Random seeds were fixed at 42 for all stochastic models. No separate hyperparameter search was performed for this development comparison.

Logistic regression provided competitive validation performance and generally stronger later-cohort performance across the observation windows. More complex models did not provide a consistent advantage. Logistic regression was therefore retained as the model family for the subsequent analysis. The feature set was then reduced further during development, resulting in the final two-feature model based on `ACTIONS_PER_ACTIVE_DAY` and `ACTIVE_DAY_PROPORTION`. The model-family comparison is reported in Table A1.

Table A1 Comparison of candidate model families using the five-feature relative behavioural representation during Accountancy development.

Week	Model	Val. ROC	Val. PR	Later ROC	Later PR
4	Logistic regression	0.669	0.602	0.707	0.569
4	XGBoost	0.648	0.597	0.662	0.531
4	CatBoost	0.660	0.608	0.693	0.545
6	Logistic regression	0.699	0.631	0.698	0.568
6	XGBoost	0.672	0.617	0.618	0.524
6	CatBoost	0.691	0.629	0.691	0.554
8	Logistic regression	0.709	0.657	0.713	0.595
8	XGBoost	0.658	0.611	0.635	0.526
8	CatBoost	0.684	0.629	0.637	0.533
10	Logistic regression	0.717	0.684	0.740	0.625
10	XGBoost	0.696	0.672	0.705	0.586
10	CatBoost	0.705	0.673	0.743	0.616
12	Logistic regression	0.718	0.699	0.752	0.655
12	XGBoost	0.682	0.670	0.704	0.607
12	CatBoost	0.692	0.678	0.718	0.611

Declarations

- **Funding Declaration:** The authors received no specific funding for this study.
- **Competing interests:** The authors declare no competing interests.
- **Author contributions:** Muhammad Zeeshan Babar conceived the study, developed the methodology, conducted the data analysis, interpreted the results, and drafted and revised the manuscript. Saif ur Rahman contributed to the interpretation of the results and critically revised the manuscript. Both authors reviewed and approved the final manuscript.
- **Data availability:** The data analysed in this study are publicly available through the KU Leuven learning analytics dataset described by Tiukhova et al. [25]. The dataset is available from Zenodo at https://doi.org/10.5281/zenodo.17087849.
- **Code availability:** The analysis code used in this study is available from the corresponding author upon reasonable request.
- **Ethics, Consent to Participate, and Consent to Publish:** Not applicable. This study involved secondary analysis of a publicly available, de-identified dataset. No participants were recruited, and no new personal data were collected.

References

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

[2] Sonnleitner, B., Madou, T., Deceuninck, M., Theodosiou, F., Sagaert, Y.R.: Evaluation of early student performance prediction given concept drift. Computers and Education: Artificial Intelligence **8**, 100369 (2025) https://doi.org/10.1016/j.caeai.2025.100369

[3] Utomo, A.P., Purwanto, P., Surarso, B.: Measuring predictive reliability under temporal class-prior drift: A forward cohort-based evaluation framework with application to student retention prediction. Measurement **280**, 121884 (2026) https://doi.org/10.1016/j.measurement.2026.121884

[4] Schwerter, J., Sabel, L., Bose, J., Bernacki, M.L., Xu, D., Schmellenkamp, M., Zeume, T., Doebler, P.: Cross-course generalizability of SRL-aligned predictive models using digital learning traces. Computers & Education **253**, 105670 (2026) https://doi.org/10.1016/j.compedu.2026.105670

[5] Pei, B., Xing, W.: An interpretable pipeline for identifying at-risk students. Journal of Educational Computing Research **60**(2), 380–405 (2022) https://doi.org/10.1177/07356331211038168

[6] Wu, S., Duan, J., Luo, M.: Calibrated early-warning models with fairness auditing and selective prediction for course withdrawal risk: Evidence from OULAD. PLOS ONE **21**(7), 0352867 (2026) https://doi.org/10.1371/journal.pone.0352867

[7] Arnold, K.E., Pistilli, M.D.: Course signals at purdue: Using learning analytics to increase student success. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, pp. 267–270 (2012). https://doi.org/10.1145/2330601.2330666

[8] Jayaprakash, S.M., Moody, E.W., Lauría, E.J.M., Regan, J.R., Baron, J.D.: Early alert of academically at-risk students: An open source analytics initiative. Journal of Learning Analytics **1**(1), 6–47 (2014) https://doi.org/10.18608/jla.2014.11.3

[9] Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., Addison, K.L.: A machine learning framework to identify students at risk of adverse academic outcomes. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1909–1918 (2015). https://doi.org/10.1145/2783258.2788620

[10] He, J., Bailey, J., Rubinstein, B.I.P., Zhang, R.: Identifying at-risk students in massive open online courses. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29, pp. 1749–1755 (2015). https://doi.org/10.1609/aaai.v29i1.9471

[11] Marbouti, F., Diefes-Dux, H.A., Madhavan, K.: Models for early prediction of at-risk students in a course using standards-based grading. Computers & Education **103**, 1–15 (2016) https://doi.org/10.1016/j.compedu.2016.09.005

[12] Kuzilek, J., Hlosta, M., Zdrahal, Z.: Open university learning analytics dataset. Scientific Data **4**, 170171 (2017) https://doi.org/10.1038/sdata.2017.171

[13] Hlosta, M., Zdrahal, Z., Zendulka, J.: Ouroboros: Early identification of at-risk students without models based on legacy data. In: Proceedings of the Seventh International Learning Analytics and Knowledge Conference, pp. 6–15 (2017). https://doi.org/10.1145/3027385.3027449

[14] Gardner, J., Brooks, C.: Student success prediction in moocs. User Modeling and User-Adapted Interaction **28**(2), 127–203 (2018) https://doi.org/10.1007/s11257-018-9203-z

[15] Aljohani, N.R., Fayoumi, A., Hassan, S.-U.: Predicting at-risk students using clickstream data in the virtual learning environment. Sustainability **11**(24), 7238 (2019) https://doi.org/10.3390/su11247238

[16] Coussement, K., Phan, M., De Caigny, A., Benoit, D.F., Raes, A.: Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. Decision Support Systems **135**, 113325 (2020) https://doi.org/10.1016/j.dss.2020.113325

[17] Mubarak, A.A., Cao, H., Zhang, W.: Prediction of students' early dropout based on their interaction logs in online learning environment. Interactive Learning Environments **30**(8), 1414–1433 (2022) https://doi.org/10.1080/10494820.2020.1727529

[18] Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A.A., Abid, M., Bashir, M., Khan, S.U.: Predicting at-risk students at different percentages of course length for early intervention using machine learning models. IEEE Access **9**, 7519–7539 (2021) https://doi.org/10.1109/ACCESS.2021.3049446

[19] Moreno-Torres, J.G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N.V., Herrera, F.: A unifying view on dataset shift in classification. Pattern Recognition **45**(1), 521–530 (2012) https://doi.org/10.1016/j.patcog.2011.06.019

[20] Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., Bouchachia, A.: A survey on concept drift adaptation. ACM Computing Surveys **46**(4), 1–37 (2014) https://doi.org/10.1145/2523813

[21] Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., Zhang, G.: Learning under concept drift: A review. IEEE Transactions on Knowledge and Data Engineering **31**(12), 2346–2363 (2019) https://doi.org/10.1109/TKDE.2018.2876857

[22] Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning, pp. 625–632 (2005). https://doi.org/10.1145/1102351.1102430

[23] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330 (2017)

[24] Van Calster, B., McLernon, D.J., Smeden, M., Wynants, L., Steyerberg, E.W.: Calibration: The achilles heel of predictive analytics. BMC Medicine **17**, 230 (2019) https://doi.org/10.1186/s12916-019-1466-7

[25] Tiukhova, E., Van Landuyt, D., Baesens, B., Snoeck, M.: Open data, private learners: A de-identified student activity and performance dataset for learning analytics. Scientific Data **13**, 548 (2026) https://doi.org/10.1038/s41597-026-06821-3