Journal of Theoretical and Applied Information Technology
15st September 2026. Vol.104. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MULTI-LANDMARK EARLY WARNING FOR NON-SUCCESSFUL COMPLETION IN STEM ONLINE COURSES: TEMPORAL CONSISTENCY, EXPLAINABLE PREDICTION, AND CAPACITY CONSTRAINTS

TINGTING XIAO[1], NURULLIZAM JAMIAT[2*]
[1]Centre for Instructional Technology and Multimedia, Universiti Sains Malaysia, Penang, Malaysia
[2*]Centre for Instructional Technology and Multimedia, Universiti Sains Malaysia, Penang, Malaysia
Email: [1] xiaotingting98@student.usm.my, [2*]nurullizamj@usm.my

ABSTRACT

Most online-learning early-warning studies rely on a single prediction point and rarely consider changes in the risk population, the accumulation of stage-specific information, and the match between alert volume and support capacity. Using the Open University Learning Analytics Dataset, this study developed a multi-landmark framework for predicting non-successful completion in STEM online courses, defined as Fail or Withdrawn. After retaining each learner's earliest eligible record, 17,956 unique learners were included, with risk sets of 15,068, 14,411, and 13,906 learners at Days 30, 60, and 90. Learner assignments to training, validation, and test sets were fixed across landmarks.

Logistic regression, random forest, and XGBoost performed similarly, with no model consistently superior across all landmarks and metrics. XGBoost was selected for interpretation and decision analysis because of its competitive performance and compatibility with TreeSHAP; test ROC-AUC values were 0.804, 0.851, and 0.876. SHAP results showed that assessment participation and stage-specific attainment became increasingly influential, while recent activity and persistence remained complementary signals. Validation-selected F2 thresholds achieved recall above 0.92 but alerted 81.4%, 74.4%, and 64.8% of test learners. Capacity-constrained analysis revealed a trade-off between earlier intervention time and later identification efficiency. The strongest predictive signals were not necessarily the earliest or most actionable, and high recall did not guarantee a manageable alert list. A multi-landmark early-warning system should therefore update the eligible population, available information, and decision threshold at each stage, rather than interpret later performance as strict longitudinal improvement.

Keywords: *Academic Risk Prediction; STEM Online Learning; Multi-Landmark Prediction; Explainable Learning Analytics; SHAP; Capacity-Constrained Intervention*

1. INTRODUCTION

1.1 Research Background

With the development of machine learning and high-performance computing, the use of artificial intelligence (AI) to predict student academic performance has become an important topic in higher education research. In the field of educational data mining (EDM), machine learning methods such as decision trees, random forests, and gradient boosting have been widely applied to the prediction of academic performance and student risk [1], [2]. Previous research suggests that machine learning models can identify nonlinear relationships in large-scale learning behaviour data that may not be fully captured by traditional statistical methods, thereby supporting academic risk identification and early intervention. However, these predictive approaches also face substantial transparency challenges. Although complex models may achieve strong predictive performance, their internal decision-making processes are often difficult to interpret directly. For frontline educators, risk probabilities and classification results without an understandable basis may be difficult to translate into trustworthy and actionable teaching decisions, potentially widening the gap between algorithmic prediction and educational intervention [3].

Explainable artificial intelligence (XAI) provides methodological support for bridging the gap between complex predictive models and educational decision-making. Common post-hoc explanation methods include Local Interpretable Model-agnostic Explanations (LIME) and SHAP (SHapley Additive exPlanations) [4], [5]. Based on Shapley values from cooperative game theory,

SHAP decomposes model outputs into the relative contributions of different input features. It can be used both to analyse local attributions for individual learner predictions and to describe the overall predictive basis of a model by aggregating individual contributions. This property makes SHAP particularly suitable for interpreting tree-based models applied to heterogeneous educational data. It should be emphasised that SHAP identifies the predictive contribution of features to model outputs rather than causal relationships between those features and learning outcomes. Although existing educational XAI studies have used SHAP or LIME to explain the predictive roles of grades, attendance, and platform behaviour, most analyses remain focused on a single prediction point or an overall feature ranking. How the predictive basis of a model changes across course stages, and how local explanations can support practical early-warning decisions, remain insufficiently examined[6].

Academic risk in online learning commonly takes the form of course failure, withdrawal, or non-completion. Timely identification of learners who may not successfully complete a course can support learning assistance and the allocation of educational resources. This study focuses specifically on STEM online courses. STEM courses often involve cumulative knowledge development, continuous task completion, and substantial cognitive demands, meaning that learners' knowledge mastery and task performance at an earlier stage may affect their subsequent learning activities. Student engagement is generally understood as comprising interrelated behavioural, cognitive, and emotional dimensions, which together represent important aspects of learners' participation in the learning process [7], [8]. Although platform log data in OULAD cannot directly measure cognitive engagement, emotional states, or learning motivation, recorded information on resource access, continuity of activity, task submission, and stage-specific assessment can serve as operational indicators of the learning process. As a course progresses, these observable data accumulate and their predictive value may change. Therefore, models based only on complete-course data or a single prediction point may fail to capture changes in risk information across different course stages. They also provide limited insight into when an early warning should be issued and which information available at that time should inform the prediction [9], [10].

This challenge becomes more complex when early warning is conducted at multiple stages of a course. As the course progresses, the population still eligible for support changes, new behavioural and assessment information becomes available, and the practical meaning of a given risk threshold may also change. A model that performs well at one prediction point may therefore not remain equally appropriate for later decision-making, particularly when learners who have already withdrawn are no longer eligible for intervention and the same decision rule may produce a different alert burden at a later stage. Multi-landmark early warning should therefore be treated as a sequence of stage-specific decision problems rather than as repeated prediction on a fixed population.

Accordingly, this study develops a multi-landmark early-warning framework for the risk of non-successful completion in STEM online courses. Stage-specific risk sets are constructed at Days 30, 60, and 90, while learner assignments to the training, validation, and test sets remain fixed across landmarks. After comparing the predictive performance of logistic regression, random forest, and XGBoost, XGBoost is selected for subsequent interpretation and decision analysis because of its competitive performance and compatibility with TreeSHAP. The study examines changes across prediction landmarks in the ability to identify learners at risk of eventual non-successful completion, the information underlying model predictions, and the size of the resulting warning lists. It also examines how classification thresholds relate to available support capacity. By combining stage-specific risk sets with explainable prediction, the study provides a temporally consistent and practical basis for risk screening and support decisions at different stages of an online course.

1.2 Research Questions and Contributions

The overall aim of this study is to develop and evaluate a multi-landmark early-warning framework for the risk of non-successful completion in STEM online courses. The framework brings temporal consistency, explainable prediction, and practical decision-making into a single analytical process. Stage-specific risk sets are constructed at Days 30, 60, and 90, while each learner's assignment to the training, validation, and test sets remains fixed. This design helps maintain consistency between the prediction population, the information available at each stage, and the timing of prediction.

More specifically, the study has three objectives. First, it compares the risk discrimination and classification performance of logistic regression, random forest, and XGBoost across the three prediction points and examines whether any model performs consistently better across all stages. Second, it uses progressive feature settings and SHAP analysis to assess the added predictive value of assessment-process information and stage-specific academic attainment, and to examine how the basis of model predictions changes over the course. Third, it combines F2 thresholds selected on the validation set with simulated capacity constraints to analyse the trade-offs among the coverage of learners who do not successfully complete the course, classification errors, and alert-list size.

Based on these objectives, the study addresses the following research questions:

RQ1. How do logistic regression, random forest, and XGBoost differ in risk discrimination and classification performance across the Day 30, Day 60, and Day 90 prediction tasks, and does any model perform consistently better across all three points?

RQ2. What added predictive value is provided by assessment-process information and stage-specific academic attainment, and how do the relative contributions of individual features and feature groups to XGBoost predictions change across the three prediction points?

RQ3. Under F2 thresholds selected on the validation set and different simulated capacity constraints, what trade-offs emerge among Recall, Precision, Specificity, and Alert rate?

The contribution of this study is primarily methodological and decision-oriented. It provides a multi-landmark framework that maintains temporal consistency across prediction stages, examines how the basis of model prediction changes over the course, and links threshold selection with practical support capacity. In doing so, the study broadens the evaluation of academic early warning beyond predictive performance alone to include prediction timing, interpretability, and decision feasibility.

2. RELATED WORK

2.1 Identifying Academic Risk in Online Learning Environments

In recent years, the rapid growth of digital learning environments has generated large amounts of learning-process data, creating new opportunities to identify risks related to course completion and improve learner support[11]. Course failure and withdrawal can interrupt learners' academic progress, making course retention and learning support important management concerns in higher education. Learning analytics can use educational process data to identify learners who may face academic difficulty or withdrawal risk and provide a basis for timely and targeted feedback and intervention. As a result, the early identification of at-risk learners has become an important area of learning analytics in higher education [12], [13], [14].

Existing research in educational data mining has widely used machine learning methods to identify information related to course outcomes. Early studies mainly focused on online behavioural variables such as study time and frequency of platform access[15]. Some studies also included relatively stable factors, such as demographic characteristics, prior education, and learning background [16]. As online learning platforms have improved their ability to record detailed activity data, process variables such as clickstream data, resource access, forum participation, assignment submission, and stage-specific assessment scores have also become important sources of information for predicting course outcomes. In this context, the use of artificial intelligence and machine learning to predict student success, failure, and withdrawal has become an important part of intelligent education systems and academic early-warning research [17].

However, different types of predictive information become available at different times. Demographic characteristics and prior educational background are usually known before a course begins, whereas platform activity, assignment submission, and stage-specific assessment information are produced gradually as the course progresses. Many existing studies still rely on complete-course data or models built at a single prediction point, without clearly considering whether a given feature would actually be available at the time an alert is issued. Therefore, comparing the predictive performance of different variables or algorithms alone is not enough to show whether a model can provide a temporally consistent early warning during the course.

2.2 Retention in STEM Online Learning and the Need for Multi-Landmark Early Warning

Course completion and learner retention in STEM education have long attracted attention from both higher education researchers and practitioners [18], [19]. STEM courses often involve continuous knowledge development, regular task practice, and stage-based assessment. Learners' participation and task completion in the early stages may therefore be closely related to their later course performance. Although these features are not unique to STEM courses, the cumulative nature of STEM knowledge, the continuity of learning tasks, and the use of staged assessments make this context well suited to examining changes in course-process information and academic risk.

Predictive information related to online course completion includes both background factors, such as gender, socioeconomic status, prior education, and previous learning experience, and process information generated during the course, such as resource use, continued participation, assignment submission, and stage-specific academic performance. A relatively systematic framework for understanding online learning withdrawal has considered factors such as personal background, motivation, academic integration, social integration, and technical support [20]. However, concepts such as learning motivation and academic integration cannot always be measured directly through platform logs. In studies based on learning management system data, resource access, active days, assignment submission, and assessment scores are more appropriately treated as observable indicators rather than direct measures of learners' cognitive, emotional, or motivational states.

Traditional academic early-warning systems often rely on relatively stable indicators, including demographic information, attendance, and examination results. However, these indicators may not fully capture changes in learner participation during the course. As the course progresses, more behavioural and assessment information becomes available. At the same time, learners who have already withdrawn or experienced the target outcome are no longer part of the group that could benefit from support at that stage. Multi-landmark early warning should therefore involve more than running the same model on different dates. It should redefine the risk population at each prediction point and restrict the model to information that is genuinely available at that time.

Research on progressive early warning suggests that learning analytics systems can provide educators with updated risk information throughout a course, allowing more time for learner support and teaching intervention[9], [21]. Compared with evaluating student performance only at the end of a course, setting prediction points at both earlier and later stages makes it possible to compare differences in risk identification, available predictive information, and the timing of support[22]. Multi-landmark studies should therefore consider not only model performance, but also how the prediction population, information availability, and potential alert-list size change across course stages.

2.3 Applications of XAI in Educational Prediction

Learning analytics and EDM provide important methods for identifying students at risk. Ensemble learning models such as XGBoost can handle nonlinear relationships, mixed types of variables, and possible feature interactions. However, whether they offer a stable advantage over simpler models such as logistic regression still depends on the dataset, prediction task, and evaluation metrics. At the same time, the internal decision process of complex models is often difficult for educators to understand, which limits the use of predictions in teaching decisions and learner support.

Explainable artificial intelligence has become an important way to improve model transparency and make predictions easier to review [23], [24]. Existing studies in educational prediction have used LIME and SHAP to explain models of academic performance and dropout risk. The variables examined include prior grades, absence records, academic history, and entry background [1], [2], [25]. SHAP can describe the overall basis of a model by combining feature contributions across learners, while also explaining the risk prediction of an individual learner. It therefore provides a relatively consistent tool for examining different types of risk judgements.

However, current educational XAI research still has three main limitations. First, some studies mainly report global feature importance and give limited attention to whether the basis of prediction changes across course stages. Second,

analysis often focuses on overall rankings, with less attention to the combined contribution of different information groups and to the prediction patterns of correctly identified, missed, and falsely alerted learners. Third, some studies extend model explanations into explanations of learning processes. In fact, SHAP shows how input variables contribute to the output of a specific model, but it cannot prove a causal relationship between those variables and course outcomes.

XAI analysis in educational risk prediction should therefore go beyond asking which variables are most important. It should also examine how the relative contribution of different types of information changes across course stages, how feature values push predictions towards higher or lower risk, and why the model produces correct or incorrect classifications for some learners. Combining global features, feature groups, and individual cases can provide a more complete explanation, but the conclusions should remain focused on the basis of model prediction rather than on causal learning mechanisms.

2.4 Research Gaps

Previous studies have shown that student performance and academic risk can be predicted at multiple points by updating the learning data available as a course progresses[26]. At the same time, explainable AI methods such as SHAP have been used to improve the transparency of educational prediction models. However, systematic reviews indicate that SHAP is still used mainly for global feature-importance analysis, while its application at the individual level remains relatively limited[6]. Different explanation methods may also produce different conclusions about feature importance for the same prediction model[27].

Overall, three key issues remain unresolved. First, multi-point prediction studies often focus on changing the feature observation window and comparing model performance across stages, but give limited attention to which learners still remain relevant for prediction and possible support at each point. Second, explanation studies usually report feature importance for a fixed model or a single time frame, with less attention to how the contribution of different information groups and individual prediction attributions change over the course. Third, educational prediction research still places greater emphasis on model development and performance evaluation than on how predictions can be translated into action and under what conditions this can be done [28]. Within this prediction-to-action gap, there is still limited work that jointly considers classification thresholds, risk coverage, and the size of alert lists that institutions can realistically manage.

To address these gaps, this study constructs landmark-specific risk sets at Days 30, 60, and 90 while maintaining fixed learner-level assignments to training, validation, and test sets. Each landmark reflects learners still eligible for support and information genuinely available at that stage. Model interpretation moves beyond static feature rankings by examining how the basis of prediction changes across course stages. Threshold analysis links alert-list size with simulated support capacity, allowing predictive performance to be considered alongside operational feasibility. Together, these elements extend existing multi-point early-warning approaches by integrating temporal consistency, stage-specific explainability, and capacity-sensitive decision analysis within one framework.

3. DATA SOURCE AND SAMPLE CONSTRUCTION

This section introduces the data source, the selection of STEM modules, the construction of the learner-level cohort, and the inclusion rules and outcome definitions for the three landmark risk sets.

3.1 Data Source

This study uses the Open University Learning Analytics Dataset (OULAD) as its data source [29]. The dataset was released by The Open University in the United Kingdom and is a widely used anonymised benchmark dataset in learning analytics and educational data mining. It has also been certified by the Open Data Institute (ODI). OULAD contains 32,593 records, each representing a student's enrolment in a specific module presentation. The dataset covers seven course modules and 22 presentations delivered between 2013 and 2014. The dataset includes student demographic information, registration and withdrawal records, stage-specific assessment results, and detailed virtual learning environment (VLE) interaction data. Although the data were collected several years ago, the complete multi-table structure, public availability, and wide use of OULAD make it suitable for reproducible methodological research in learning analytics. However, the relevance of the findings to current

online learning settings should be interpreted carefully in light of the age of the data and the course context.

OULAD consists of several linked data tables. The main tables used in this study are studentInfo, studentRegistration, courses, assessments, vle, studentVle, and studentAssessment. The studentInfo table contains learner background information and final course outcomes. The studentRegistration table records registration and withdrawal dates. The courses table provides information on course modules and presentations. The assessments table records assessment types, deadlines, and weights. The vle and studentVle tables contain information on platform resources and learners' VLE activity, respectively. The studentAssessment table records assessment submissions and scores.

The tables are linked through fields such as id_student, code_module, and code_presentation. The main OULAD tables and their relationships are shown in Fig. A1 in Appendix A.

3.2 STEM Sample Selection and Outcome Labelling

To focus on predicting non-successful completion in STEM online courses, this study selected four STEM related modules, CCC, DDD, EEE, and FFF, from the seven modules in OULAD based on their subject areas. The course modules, number of presentations, and corresponding record counts are shown in Table 1.

Table 1. Module summary and domain information (Data source: Kuzilek et al., 2017)

Module	Domain	Presentations	Students
AAA	Social Sciences	2	748
BBB	Social Sciences	4	7,909
CCC	STEM	2	4,434
DDD	STEM	4	6,272
EEE	STEM	3	2,934
FFF	STEM	4	7,762
GGG	Social Sciences	3	2,534

To avoid counting the same learner more than once because they studied multiple STEM modules or participated in different course presentations, this study used id_student as the unique identifier and retained only the earliest eligible STEM module and course presentation record for each learner. The resulting initial cohort included 17,956 unique learners. Among them, 8,376 achieved a Pass or Distinction and were coded as Success (0), while 9,580 received a Fail or Withdrawn outcome and were coded as non-successful completion (1). The study outcome was therefore defined as non-successful course completion, including both course failure and withdrawal. As these two outcomes may result from different processes, further outcome sensitivity analyses were conducted for Failure and Future Withdrawal.

Based on the initial cohort, stage-specific risk sets were constructed at Day 30, Day 60, and Day 90. At each prediction point, only learners who had registered and were still enrolled in the course were included. Learners with missing registration dates, those who registered after the prediction point, those who withdrew on or before the prediction point, and withdrawn records with missing withdrawal dates were excluded from the relevant risk set. At each point, VLE behaviour and assessment features were constructed using only information observed between the start of the course and the corresponding prediction point. This ensured consistency between feature generation and the timing of prediction. The detailed process of data selection and preparation is shown in Figure 1.

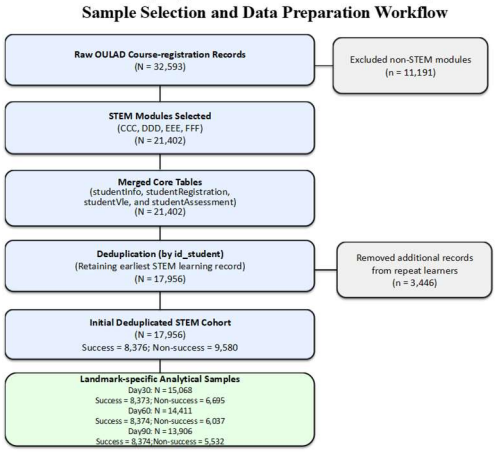


Fig. 1. Sample Selection and Data Preparation Workflow

4. FEATURE CONSTRUCTION AND DATA PREPROCESSING

Based on the three stage-specific risk sets constructed in Section 3.2, this study integrated learner background information, course registration records, VLE interactions, and assessment records using id_student, code_module, and code_presentation as the linking keys. final_result was converted into a binary outcome representing non-successful course completion. Pass and Distinction were coded as Success (0), while Fail and Withdrawn were coded as non-successful completion (1). date_unregistration was used only to determine eligibility for each risk set and was not

included as a predictor. final_result and any variables derived from it were used only to construct the outcome label and were excluded from the feature matrix.

Based on the observable information available in OULAD, the predictors were divided into four groups: background and course context, behavioural engagement, sustained engagement, and assessment participation and stage-specific academic attainment. Background and course context features included demographic characteristics, prior educational background, registration timing, and course information. Behavioural engagement features reflected the intensity of VLE interaction and patterns of resource use. Sustained engagement features described the continuity of learning activity, recent activity status, and trends in participation. Assessment related features represented task requirements, submission behaviour, and stage-specific academic performance up to each prediction point. All variables were treated as operational indicators recorded by the platform and were not interpreted as direct measures of learning motivation, emotional state, or cognitive processes.

All dynamic features were constructed separately for Day 30, Day 60, and Day 90 using only information observed from Day 0 to the corresponding prediction point, including the prediction day itself. VLE features included cumulative clicks, number of active days, range of resources accessed, daily interaction statistics, and the number and proportion of clicks across different resource types. To capture changes in participation within each observation period, the period was divided into an earlier stage and a later stage. Because the two stages did not contain exactly the same number of days, both click_trend_diff and click_trend_ratio were calculated using the average number of daily clicks in each stage. click_trend_diff represented the difference between the average daily clicks in the later stage and those in the earlier stage. click_trend_ratio represented the ratio between the smoothed average daily clicks in the two stages.

Assessment process features were defined according to assessment metadata and included only tasks that were due by the relevant prediction point. assessment_count represented the number of assessments that should have been completed, submitted_count represented the number actually submitted, and submission_rate was calculated as the ratio of the two. weighted_score_rate used the total weight of assessments due before the prediction point as the denominator, with unsubmitted assessments assigned a score of zero. Measures such as mean_score, pass_assessment_count, and mean_submission_delay were calculated only from submissions recorded before the prediction point. OULAD does not record the actual timing of score release or feedback. The main analysis therefore treated scores linked to submissions made before the prediction point as information available at that time. The assessment process feature setting without score variables was used to examine separately the added predictive value of submission and task completion information. This availability assumption is discussed further in the study limitations.

Missing value handling distinguished between structural absence and missing background information. When no VLE interaction had occurred, no assessment had been submitted, or no assessment was yet due, the corresponding process features were coded as zero according to their operational meaning. Missing values in demographic and background variables were retained until the modelling stage. Infinite values generated during calculation were converted to missing values. Numerical variables were imputed using the median from the training data and then standardised. Categorical variables were imputed using the mode from the training data and then converted using one hot encoding. All preprocessing steps were included within the modelling pipeline. Their parameters were estimated only from the training folds or the training set and were then applied to the corresponding validation and test data. The feature groups, representative variables, and their operational definitions are presented in Table 2.

Table 2. Summary of feature dimensions and representative variables

Predictor domain	Representative variables	Operational definition
Background characteristics and course context	gender, region, highest_education, imd_band, age_band, num_of_prev_attempts, studied_credits, disability, code_module, code_presentation, date_registration	Static learner profile, prior educational background, course context, and registration timing used as contextual predictors.
Behavioural engagement	total_clicks, active_records, active_sites, daily_click_mean,	Indicators of VLE interaction intensity and resource-use patterns observed up to

Predictor domain	Representative variables	Operational definition
Persistence	clicks_activity_*, ratio_activity_*	the corresponding landmark.
	active_days, first_active_day, last_active_day, clicks_first_half, clicks_second_half, click_trend_diff, click_trend_ratio	Indicators of continuity, recency, and temporal change in learning participation. Trend variables were calculated from the average daily clicks in the first and second observation periods.
Assessment participation and stage-specific academic attainment	assessment_count, submitted_count, submission_rate, mean_submission_delay, mean_score, pass_assessment_count, assessment_pass_rate, total_weight, weighted_score_sum, weighted_score_rate, count_cma, count_tma	Indicators describing assessments due by the landmark, task submission and timing, observed scores, pass status, and weighted attainment. Unsubmitted assessments remained in the denominators of submission_rate and weighted_score_rate.

Note. Dynamic features were constructed separately for Day30, Day60, and Day90 using only information observable up to each landmark. Students who had withdrawn on or before a landmark were excluded from that risk set. final_result, target-derived grouping variables, and date_unregistration were excluded from predictor inputs; date_unregistration was used only to determine landmark eligibility.

5. RESEARCH METHODS

The overall research process is presented in Fig. 2. It consists of sample and risk set construction, fixed data splitting, candidate model comparison, explanation analysis, threshold based decision analysis, and robustness and sensitivity analyses.

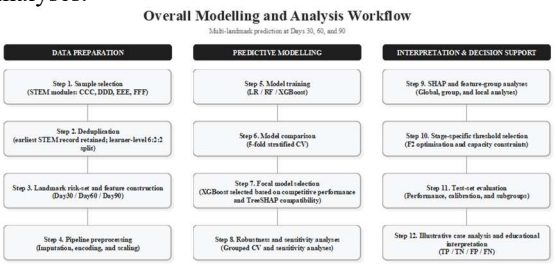


Fig. 2. Overall Modelling and Analysis Workflow

5.1 Multi-Landmark Modelling Design and Data Splitting

This study defined Day 30, Day 60, and Day 90 after the start of the course as three prespecified landmark prediction points, corresponding to the first, second, and third stage prediction tasks. At each point, only learners who had registered, had not withdrawn, and were still enrolled in the course were included. Information available on or before the prediction point was used to predict their final course completion status.

To prevent the same learner from appearing in different data subsets, the initial learner level cohort was divided once using a fixed assignment. The split was stratified by the combination of final course outcome and course module. Using a random seed of 42, learners were assigned to the training, validation, and test sets in proportions of 60%, 20%, and 20%, respectively. These subset assignments were then mapped to the Day 30, Day 60, and Day 90 risk sets. Therefore, when the same learner met the inclusion criteria at more than one prediction point, their assignment to the training, validation, or test set remained unchanged.

After the fixed assignments were mapped to the landmark risk sets, the training, validation, and test sample sizes at Day 30 were 9,037, 3,010, and 3,021, respectively. The corresponding sample sizes were 8,676, 2,837, and 2,898 at Day 60, and 8,368, 2,745, and 2,793 at Day 90. The training set was used for cross validation of candidate models and final model fitting. The validation set was used only to select stage-specific F2 thresholds and capacity constraint thresholds. The test set was used only for final performance evaluation, SHAP interpretation, calibration analysis, and uncertainty analysis.

Because course resources and assessment tasks became available or reached their deadlines gradually as the course progressed, the available feature spaces were not identical across the three prediction points. The main analysis retained all variables that were available at each actual prediction point to reflect the accumulation of risk information over time. In addition, candidate model comparisons were repeated using the same 83 predictors available at all three points. This analysis examined whether differences in performance across prediction points were mainly driven by variables added at later stages or by changes in the number of features.

5.2 Model Training and Selection of the Main Model

This study compared three models: logistic regression (LR), random forest (RF), and extreme gradient boosting (XGBoost). LR was used as a linear baseline to evaluate the ability of stage-specific variables to distinguish risk under a simple

linear decision structure. RF served as a tree ensemble baseline and provided another reference for nonlinear modelling. XGBoost was considered as the main candidate model for the subsequent SHAP interpretation and stage-specific decision analysis.

To ensure comparability across prediction points, all three models used predefined fixed parameter settings, and no separate tuning was conducted for individual points. The maximum number of iterations for LR was set to 2,000. RF included 300 decision trees. XGBoost used 300 estimators, a learning rate of 0.05, a maximum tree depth of 5, and logloss as the evaluation metric during training. No class weights were applied in the main analysis, and a random seed of 42 was used for all analyses involving random processes.

Candidate model comparison was conducted only within the fixed training set. Five fold stratified cross validation was performed separately at Day 30, Day 60, and Day 90. The complete preprocessing and modelling pipeline was fitted again within each fold. Model comparison reported the mean and standard deviation across folds for Accuracy, Precision, Recall, F1 score, ROC AUC, and Average Precision (AP). ROC AUC and AP were used as the main measures of risk ranking performance. Accuracy, Precision, Recall, and F1 score were also reported to describe classification performance at the default threshold of 0.50.

Neither the validation set nor the test set was involved in the cross validation comparison of candidate models. The main model was selected by considering the cross validation results within the training set, stability across folds, and the interpretive aims of the study. When XGBoost showed no clear or systematic loss in performance compared with LR and RF, it was used for the subsequent prediction interpretation, threshold selection, and capacity constraint analysis because of its direct compatibility with TreeSHAP. LR and RF were mainly used to assess whether adopting the XGBoost interpretation framework involved a clear loss of predictive performance, rather than to demonstrate that complex models necessarily outperform simpler models.

After the candidate model comparison, an XGBoost model was fitted again using the complete training set at each prediction point. The validation set was used to determine the classification decision rules, while the independent test set was used to report final risk discrimination, probability quality, and classification performance under the selected thresholds.

5.3 Model Interpretation and Incremental Analysis

To examine the information underlying model risk judgements at different course stages, this study conducted SHAP analysis at the levels of individual features, feature groups, and individual cases. Progressive feature settings were also used to assess the added predictive value of assessment information.

First, TreeExplainer was used to calculate SHAP values for XGBoost on the independent test set at each prediction point. For global interpretation, the mean absolute SHAP value of each encoded variable was used to measure its overall contribution to the model output. SHAP bar plots and beeswarm plots were used to present the importance and direction of the main variables. The explainer calculated SHAP values based on the raw model output. The results therefore indicate whether each feature pushes the model risk output in a positive or negative direction and can be understood on the log odds scale. However, they cannot be interpreted directly as specific changes in predicted probability.

Second, to compare the overall roles of different information sources, the predictors were divided into four feature groups: background and course context, VLE behavioural engagement, sustained learning engagement, and assessment participation and stage-specific academic attainment. The contribution of each group was calculated by summing the absolute SHAP values of all encoded variables within that group and expressing the result as a proportion of the total absolute SHAP contribution. Each encoded variable was assigned to only one feature group, and checks were conducted to identify missing or duplicate assignments. Because the groups differed in both the number of variables and their correlation structures, feature group contributions were used only to describe the relative composition of predictive attribution within the model. They were not interpreted as direct measures of underlying learning constructs or as evidence of causal effects.

To assess the added predictive value of assessment information, three progressive feature settings were developed. Setting A included only background and course context, VLE behavioural

engagement, and sustained learning engagement variables. Setting B added assessment process variables, including the number of assessment tasks, submission behaviour, submission rate, and submission delay. Setting C further added stage-specific scores, assessment pass indicators, and weighted attainment measures to form the complete feature model. All three settings used the same XGBoost parameters and the same five fold stratified cross validation procedure within the training set. Changes in ROC AUC, AP, Recall, and F1 score were compared to evaluate the added predictive value of assessment process information and stage-specific academic attainment.

For local interpretation, four types of illustrative cases were selected from the independent test set at each prediction point according to predefined rules. These included the true negative case with the lowest predicted risk, representing a learner who successfully completed the course and was not included in the alert list; the true positive case with the highest predicted risk, representing a learner who did not successfully complete the course and was included in the alert list; the false negative case with a predicted probability below and closest to the selected threshold; and a false positive case with a relatively high predicted probability. These cases were used to illustrate the structure of model judgements under different classification outcomes. They were not intended to represent the average or typical learner within the corresponding groups.

5.4 Stage-specific Alert Thresholds and Probability Quality Evaluation

To construct a highly sensitive early warning scenario that gives priority to reducing missed cases among learners who ultimately do not successfully complete the course, this study searched for the threshold that maximised the F2 score on the validation set at each prediction point. The search range was from 0.05 to 0.95, with an interval of 0.01. The F2 score gives greater weight to Recall than to Precision and is therefore suitable for early warning situations in which the cost of missed cases is relatively high.

The threshold selected at each prediction point was then fixed and applied to the corresponding independent test set. Precision, Recall, Specificity, F1 score, F2 score, and Alert rate were reported for the test set. Alert rate was defined as the proportion of learners in the test set who were included in the alert list and was used as an approximate measure of the potential alert list size. The threshold that maximised the F2 score was used to demonstrate the level of identification that the model could achieve under a high risk coverage scenario. It was not treated as a single optimal threshold suitable for all institutions or course stages.

To simulate practical situations in which educational institutions have limited capacity to support learners, five target intervention capacities were examined: 10%, 20%, 30%, 40%, and 50%. For a target capacity ccc, the classification threshold was determined according to the top ccc proportion of predicted probabilities in the validation set. This threshold was then fixed and applied to the independent test set.

Because the thresholds were determined entirely from the probability distribution in the validation set, and the probability distributions in the validation and test sets could differ, the actual Alert rate in the test set could be slightly different from the stated capacity. The thresholds were not readjusted according to the actual number of learners in the test set. The capacity constraint analysis reported Precision, Recall, Specificity, F1 score, F2 score, and the actual Alert rate under each scenario. These measures were used to compare risk coverage and false alert control across prediction points when the number of learners who could receive support was fixed.

The quality of the predicted probabilities was evaluated using calibration curves, the Brier score, the Null Brier score, and the Brier Skill Score. Calibration curves were constructed using 10 equal frequency groups of predicted probabilities in the test set. The Null Brier score used the proportion of learners with non-successful completion in the training set at the corresponding prediction point as a constant predicted probability. The Brier Skill Score was calculated as follows:

$$BSS = 1 - \frac{BS_{model}}{BS_{null}} \quad (1)$$

A value greater than 0 indicates that the model has a lower probability prediction error than the baseline model. The Brier score was used to evaluate the overall error of the predicted probabilities, while the calibration curve was used to describe the agreement between the predicted probabilities and the observed proportion of non-successful completion. This study evaluated the quality and calibration of the original predicted probabilities and did not apply additional

calibration methods such as Platt scaling or isotonic regression.

5.5 Robustness, Sensitivity, and Uncertainty Analyses

This study examined the robustness of the main findings in relation to the feature space, course context, class handling methods, outcome definitions, and uncertainty in the test sample.

First, model comparisons were repeated using the same 83 variables available at all three prediction points. This analysis examined whether differences in performance across the prediction points were mainly driven by variables added at later stages or by differences in the number of features.

Second, code_module × code_presentation was used as the grouping unit for course context. Five fold stratified group cross validation was conducted within the fixed training set. This ensured that the same combination of course module and presentation did not appear in both the training fold and the validation fold, allowing the sensitivity of model performance to differences in course context to be evaluated.

Although none of the three prediction points showed severe class imbalance, SMOTE was included as an additional sensitivity analysis. Oversampling was restricted to the training fold within each cross validation round, while the validation fold retained its original class distribution. This analysis examined whether risk coverage and the trade-off between Precision and Recall were sensitive to the method used for class handling.

Because Fail and Withdrawn may result from different processes, separate prediction tasks were also constructed for Failure and Future Withdrawal. The Failure analysis excluded learners whose final outcome was Withdrawn. Fail was defined as the positive outcome, while Pass and Distinction were defined as negative outcomes. In the Future Withdrawal analysis, learners who withdrew after the relevant landmark were defined as positive cases. Learners who did not withdraw after the landmark, including those whose final outcomes were Pass, Distinction, or Fail, were defined as negative cases. Learners who had withdrawn on or before the landmark were excluded from the corresponding risk set. Both analyses retained the fixed data subset assignments, the same time boundaries, and the same XGBoost configuration. ROC AUC and AP were used to evaluate prediction performance for each outcome.

Uncertainty in the main test set metrics was estimated using stratified Bootstrap sampling. At each prediction point, sampling with replacement was conducted separately within the true outcome groups of Success and non-successful completion. The procedure was repeated 2,000 times using a random seed of 42. In each sample, the fixed test set predictions and the F2 threshold selected on the validation set were used to recalculate ROC AUC, AP, Precision, Recall, Specificity, F1 score, F2 score, Alert rate, and Brier score. The percentile method was used to obtain 95% confidence intervals. The models were not trained again and the thresholds were not selected again during the Bootstrap procedure. The resulting intervals therefore represent only the uncertainty in the fixed test predictions under repeated sampling of the test sample.

Finally, exploratory subgroup performance audits were conducted according to gender, disability status, and IMD group. These analyses were used to identify differences in risk prevalence, alert coverage, classification errors, and probability error across groups. The subgroup analysis was intended to describe the limits of model performance across different groups and was not treated as a complete assessment of algorithmic fairness. All analyses were conducted in Python.

6. RESULTS

This section reports the descriptive characteristics of the three landmark risk sets, the predictive performance of the candidate models and XGBoost, changes in the sources of risk information across course stages, alert outcomes under different thresholds and capacity constraints, individual prediction attributions, and the results of the robustness and sensitivity analyses. These analyses address model discrimination, the structure of predictive attribution, decision trade-off, and the conditions under which the findings remain applicable.

6.1 Descriptive Characteristics of the Stage-specific Risk Sets

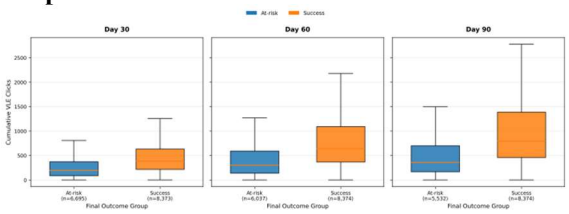


Fig. 3. Distribution of cumulative VLE clicks by final outcome group across the Day30, Day60, and Day90 landmark risk sets.

The Day 30, Day 60, and Day 90 risk sets included 15,068, 14,411, and 13,906 learners, respectively. Among them, 6,695, 6,037, and 5,532 learners ultimately did not successfully complete the course, corresponding to rates of 44.43%, 41.89%, and 39.78%.

The decreases in sample size and the proportion of non-successful completion across later prediction points mainly reflect updates to eligibility for the stage-specific risk sets. In particular, learners who had already withdrawn on or before a given prediction point were excluded from that point and all subsequent risk assessments. Therefore, these changes should not be interpreted as a natural decline in the risk of non-successful completion among the same group of learners over time.

As shown in Fig. 3, the distribution of cumulative VLE clicks for the Success group was generally higher than that of the non-successful completion group at all three prediction points. The distributions for both groups were clearly right skewed and included some extreme values. This result only describes differences in platform participation between groups with different final outcomes. It does not indicate that cumulative clicks have an independent or causal effect on course outcomes. Descriptive results by gender, prior education, course module, and IMD group are presented in Figs. B1 to B3.

6.2 Prediction Performance Across Multiple Landmarks

Table 3 reports the prediction performance of LR, RF, and XGBoost under five fold stratified cross validation within the fixed training set. The point estimates of ROC AUC and AP increased at later prediction points for all three models. For XGBoost, ROC AUC was 0.7994 at Day 30, 0.8431 at Day 60, and 0.8681 at Day 90. The corresponding AP values were 0.7804, 0.8257, and 0.8475. LR and RF showed the same overall pattern.

Table 3. Predictive performance of LR, RF, and XGBoost across the three landmark risk sets

Time	Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	AP
Day 30	LR	0.7346± 0.0101	0.7392± 0.0152	0.6227± 0.0202	0.6757± 0.0140	0.8061± 0.0079	0.7883± 0.0099
Day 30	RF	0.7321± 0.0094	0.7433± 0.0149	0.6070± 0.0218	0.6680± 0.0144	0.7957± 0.0108	0.7738± 0.0114
Day 30	XGBoost	0.7321± 0.0107	0.7301± 0.0162	0.6304± 0.0211	0.6763± 0.0144	0.7994± 0.0094	0.7804± 0.0111
Day 60	LR	0.7782± 0.0092	0.7847± 0.0254	0.6540± 0.0152	0.7130± 0.0088	0.8482± 0.0074	0.8302± 0.0062
Day 60	RF	0.7742± 0.0091	0.7805± 0.0239	0.6466± 0.0130	0.7069± 0.0087	0.8436± 0.0069	0.8240± 0.0081
Day 60	XGBoost	0.7736± 0.0053	0.7734± 0.0191	0.6554± 0.0192	0.7091± 0.0066	0.8431± 0.0074	0.8257± 0.0057
Day 90	LR	0.8078± 0.0102	0.8027± 0.0117	0.6885± 0.0218	0.7411± 0.0158	0.8747± 0.0055	0.8537± 0.0067
Day 90	RF	0.8043± 0.0032	0.8076± 0.0085	0.6703± 0.0166	0.7323± 0.0075	0.8688± 0.0043	0.8467± 0.0066
Day 90	XGBoost	0.8051± 0.0042	0.7998± 0.0065	0.6837± 0.0146	0.7371± 0.0079	0.8681± 0.0053	0.8475± 0.0063

Because the three landmark risk sets differed in learner composition, the proportion of non-successful completion, and the available feature space, these results represent performance differences across three stage-specific prediction tasks. They should not be interpreted directly as a longitudinal improvement in predictive ability among the same group of learners. In addition, AP is affected by the proportion of positive cases. Comparisons across prediction points should therefore also consider the proportion of non-successful completion in each risk set.

The overall performance of LR, RF, and XGBoost was similar. LR achieved the highest point estimates for ROC AUC and AP at all three prediction points, but the absolute differences between XGBoost and LR were below 0.01. XGBoost showed slightly higher Recall than LR at Day 30 and Day 60, while RF did not show a consistent advantage across prediction points.

These results therefore do not indicate that XGBoost consistently outperformed LR or RF across all prediction points and evaluation metrics. However, XGBoost maintained competitive risk discrimination at all three points and directly supported the subsequent TreeSHAP analysis. It was therefore selected as the main model for interpretation. This choice was made to examine the basis of model predictions and stage-specific decision performance, rather than to claim that

complex models are generally superior to linear models.

6.2.1 XGBoost performance on the independent test set and uncertainty

The ROC AUC values of XGBoost on the independent test sets were 0.8041 at Day 30, 0.8513 at Day 60, and 0.8763 at Day 90. The corresponding 95% Bootstrap confidence intervals were [0.7886, 0.8194], [0.8374, 0.8647], and [0.8622, 0.8894]. The AP values were 0.7861, 0.8330, and 0.8533, respectively.

The test set results showed the same stage-specific pattern as the cross validation results from the training set. However, the three prediction points differed in risk set membership, the proportion of positive cases, and the information available for prediction. Therefore, the higher point estimates at later prediction points should be understood as differences among separate stage-specific prediction tasks. They should not be interpreted directly as a longitudinal improvement in model performance for the same model or the same group of learners as the course progressed. The complete Bootstrap results are presented in Table C1.

6.3 Changes in the Sources of Risk Information Across Course Stages

6.3.1 Added predictive value of assessment information

To evaluate the added predictive value of assessment information at different course stages, this study compared three progressive feature settings within the fixed training set using the same data preprocessing procedure, XGBoost parameters, and five fold stratified cross validation process, as shown in Table 4. Setting A included only background and course context, VLE behavioural engagement, and sustained learning engagement features. Setting B added assessment process information to Setting A, including the number of assessment tasks, task submission, submission rate, and submission delay. Setting C further added stage-specific scores, assessment pass indicators, and weighted attainment measures to form the complete feature model.

After assessment process information was added, AP for Setting B increased by 0.0148 at Day 30, 0.0224 at Day 60, and 0.0308 at Day 90 compared with Setting A. Recall increased by 0.0120, 0.0239, and 0.0490, respectively. These results indicate that assessment deadlines, task submission, submission rate, and submission timing provided additional information for risk discrimination beyond learner background and VLE behaviour, with larger gains at later prediction points.

Table 4. Incremental Contribution Of Assessment-Related Feature To Xgboost Performance

Time	Setting	Features	Accuracy	Precision	Recall	F1-score	ROC-AUC	AP
Day 30	A: No assessment	63	0.7016± 0.0124	0.6970± 0.0165	0.5808± 0.0202	0.6335± 0.0169	0.7607± 0.0142	0.7363± 0.0148
Day 30	B: Assessment process	73	0.7097± 0.0151	0.7068± 0.0211	0.5928± 0.0195	0.6447± 0.0190	0.7723± 0.0145	0.7511± 0.0160
Day 30	C: Full model	83	0.7321± 0.0107	0.7301± 0.0162	0.6304± 0.0211	0.6763± 0.0144	0.7994± 0.0094	0.7804± 0.0111
Day 60	A: No assessment	65	0.7337± 0.0058	0.7308 ± 0.0108	0.5825± 0.0214	0.6480± 0.0122	0.7940± 0.0046	0.7697± 0.0061
Day 60	B: Assessment process	75	0.7478 ± 0.0036	0.7477± 0.0149	0.6064± 0.0175	0.6693± 0.0066	0.8113± 0.0051	0.7921± 0.0038
Day 60	C: Full model	85	0.7736± 0.0053	0.7734± 0.0191	0.6554± 0.0192	0.7091± 0.0066	0.8431± 0.0074	0.8257± 0.0057
Day 90	A: No assessment	65	0.7549± 0.0082	0.7588± 0.0087	0.5668± 0.0186	0.6488± 0.0151	0.8140± 0.0126	0.7866± 0.0120
Day 90	B: Assessment process	75	0.7796± 0.0080	0.7865± 0.0084	0.6158± 0.0178	0.6907± 0.0138	0.8399± 0.0067	0.8174± 0.0058
Day 90	C: Full model	85	0.8051± 0.0042	0.7998 ± 0.0065	0.6837± 0.0146	0.7371± 0.0079	0.8681± 0.0053	0.8475± 0.0063

Note. Only XGBoost results are shown. All feature-setting comparisons were conducted by five-fold stratified cross-validation within the fixed training subset.

After stage-specific scores, assessment pass indicators, and weighted attainment measures were added to Setting B, AP for Setting C increased further by 0.0293 at Day 30, 0.0336 at Day 60, and 0.0301 at Day 90. Recall increased further by 0.0376, 0.0490, and 0.0679, respectively. These findings show that stage-specific academic attainment provided additional predictive value beyond assessment process information. The increase in Recall became larger at later prediction points, while the increase in AP remained relatively

stable across the three points. Setting B also provided a process based comparison that did not rely on score variables, helping to distinguish the contribution of the submission process from that of stage-specific attainment when the timing of score release was not observable.

6.3.2 Changes in SHAP contributions of feature Groups across course stages

To further compare the overall roles of different information sources, the encoded predictors were assigned to four feature groups: background conditions, behavioural engagement, sustained engagement, and assessment participation and stage-specific academic attainment. The contribution of each group was calculated as the sum of the absolute SHAP values of all variables within that group, expressed as a proportion of the total absolute SHAP values across all features. Detailed results are presented in Table 5.

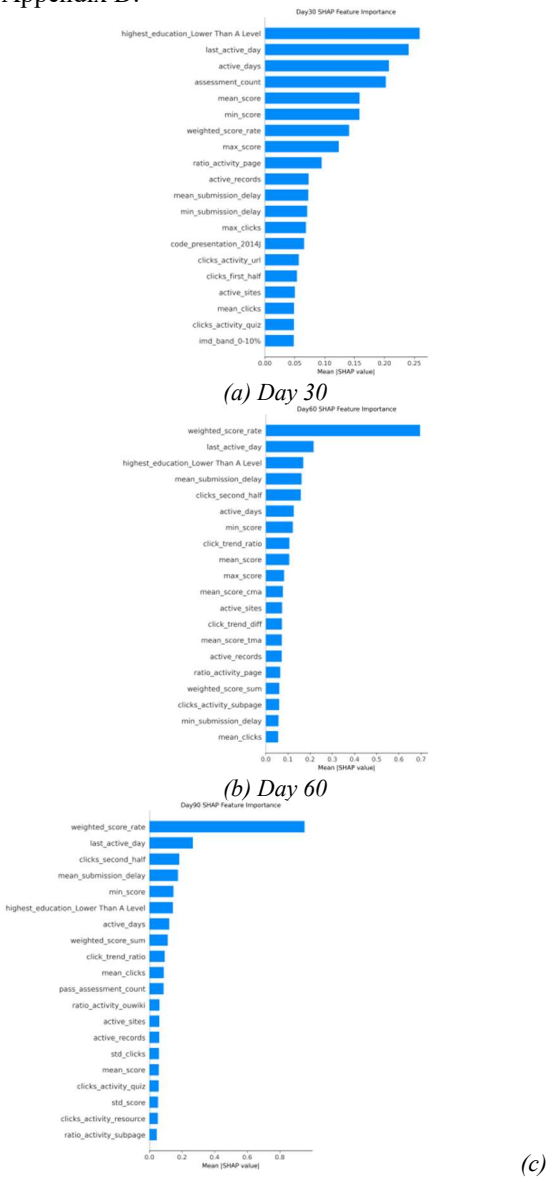
Table 5. Normalised SHAP contributions of feature groups across landmark models

Group	Day30 Day60 Day90
Background	21.23%13.86%11.9%
Behavioral	33.42%29.23%28.18%
Assessment participation and attainment	28.71%39.08%42.68%
Persistence	16.64%17.83%17.24%

Overall, model attribution gradually shifted from multiple information sources at Day 30 towards assessment participation and stage-specific academic attainment as the main source, while behavioural engagement and sustained participation continued to play a supporting role. This pattern reflects differences in the information structures used by the models at different course stages and should be understood in relation to the composition of each risk set and the results of the separately fitted models.

6.3.3 Changes in SHAP contributions of individual features across course stages

The results for individual features were consistent with the feature group analysis. At Day 30, the features with high contributions were distributed across prior educational background, last_active_day, active_days, assessment_count, and early assessment scores. At Day 60 and Day 90, weighted_score_rate became the feature with the highest contribution, while indicators related to submission timing and recent engagement, such as mean_submission_delay and clicks_second_half, also ranked highly. last_active_day and active_days remained among the leading features at all three prediction points, indicating that sustained and recent activity continued to provide additional predictive information even as more assessment information became available. Fig. 4 summarises the ranking of mean absolute SHAP values at each prediction point. The beeswarm plots showing the direction of feature effects are presented in Appendix D.



(a) Day 30

(b) Day 60

(c) Day 90

Fig. 4. Mean absolute SHAP feature importance across the three landmark models: (a) Day 30; (b) Day 60; and (c) Day 90.

6.4 Stage-specific Early Warning Decisions and Probability Quality

6.4.1 High sensitivity early warning under the F2 optimal thresholds

To compare the trade-off among risk coverage, false alert control, and alert list size across different prediction points, this study

conducted stage-specific threshold analysis using the main XGBoost model. The F2 optimal threshold at each prediction point was determined using only the independent validation set and was then fixed and applied to the corresponding independent test set. Table 6 reports Precision, Recall, Specificity, F1 score, F2 score, and Alert rate under each selected threshold. Unlike Table 3, which compares candidate models through five-fold cross-validation within the training set, Table 6 presents the actual classification performance on the independent test set under the decision rules selected from the validation set.

Table 6. Test-Set Performance Under Stage-Specific F2-Optimised Thresholds

Time	Threshold	Precision	Recall	Specificity	F1-score	F2-score	Alert rate
Day 30	0.1600	0.5203	0.9509	0.2965	0.6726	0.8159	0.8136
Day 60	0.1400	0.5371	0.9476	0.4045	0.6856	0.8220	0.7440
Day 90	0.1500	0.5699	0.9230	0.5358	0.7047	0.8213	0.6477

Note. *Thresholds were selected exclusively on the validation sets by maximising F2-score and were subsequently fixed for evaluation on the independent test sets. Alert rate represents the proportion of test students classified for alert at the validation-selected threshold*

The F2 optimal thresholds selected on the validation set were 0.16 at Day 30, 0.14 at Day 60, and 0.15 at Day 90. Recall on the corresponding test sets exceeded 0.92 at all three prediction points, while the F2 score ranged from 0.8159 to 0.8220. At the same time, Precision increased from 0.5203 to 0.5699, Specificity increased from 0.2965 to 0.5358, and Alert rate decreased from 81.36% to 64.77%.

These results show that high coverage screening at earlier prediction points involved more false alerts and larger alert lists, whereas later prediction points reduced the alert list size while maintaining high Recall. The F2 thresholds represent high sensitivity scenarios that give priority to reducing missed cases and should not be treated as a single rule suitable for all courses and institutions. The corresponding confusion matrices for the test sets are shown in Fig. E1.

6.4.2 Early warning tradeoffs under intervention capacity constraints

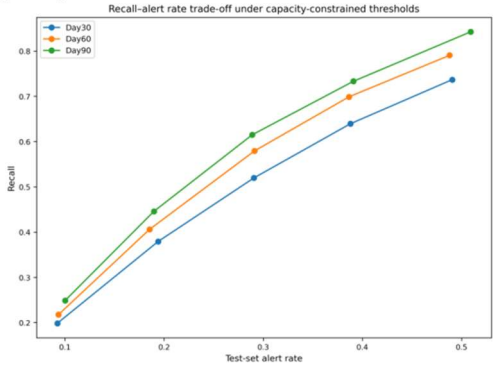


Fig. 5. Recall–alert rate trade-off under capacity-constrained thresholds across the three landmark models.

The alert list produced by the F2 optimal threshold may exceed the support capacity available in practice. This study therefore simulated target intervention capacities of 10%, 20%, 30%, 40%, and 50%, with thresholds determined from the distribution of predicted probabilities in the validation set, as shown in Fig. 5. These thresholds were then fixed and applied to the independent test set. As a result, the actual Alert rate was close to, but not required to equal, the target capacity.

Under the 40% target capacity, the actual Alert rates at Day 30, Day 60, and Day 90 were 38.80%, 38.65%, and 39.10%, respectively. The corresponding Recall values were 0.6394, 0.6989, and 0.7332. With a similar level of support capacity, later prediction points covered a larger proportion of learners who ultimately did not successfully complete the course, while earlier prediction points provided a longer potential intervention period. The complete results are presented in Table E1.

Capacity constraints and F2 optimisation represent different decision goals. Capacity constraints address the question of which learners should receive priority when support resources are limited, whereas F2 optimisation focuses on reducing missed cases as much as possible. Threshold selection should therefore consider the course stage, available support capacity, and the relative costs of missed cases and false alerts.

6.4.3 Predicted probability quality and calibration performance

Table 7. Probability Calibration Metrics Across The Landmark Models

Time	Train prevalence	Test prevalence	Brier score	Null Brier	Brier skill
Day 30	0.4443	0.4452	0.1775	0.2470	0.2814
Day 60	0.4210	0.4217	0.1516	0.2439	0.3783
Day 90	0.3997	0.3999	0.1356	0.2400	0.4348

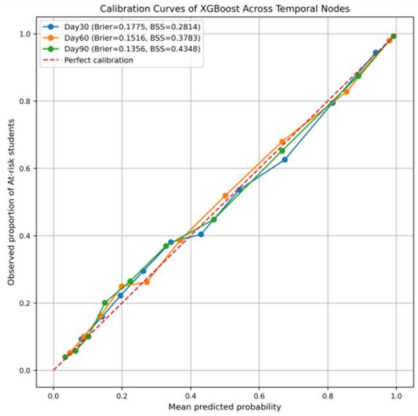


Fig. 6. Calibration curves of XGBoost across the three landmark risk sets.

As shown in Table 7, the Brier scores at Day 30, Day 60, and Day 90 were 0.1775, 0.1516, and 0.1356, respectively. All were lower than the corresponding Null Brier scores based on the proportion of non-successful completion in the training set, which were 0.2470, 0.2439, and 0.2400. The corresponding Brier Skill Scores were 0.2814, 0.3783, and 0.4348.

The original predicted probabilities at all three points performed better than the constant probability baseline, and the later prediction points showed lower overall probability errors in their respective risk prediction tasks. Fig. 6 shows that some deviation remained in the middle and high risk ranges at Day 30. The model is therefore more suitable for risk ranking and capacity based screening, and individual outputs should not be interpreted directly as precise probabilities of non-successful completion without local calibration.

6.5 Individual Prediction Patterns and Interpretation of Classification Errors

The local SHAP analysis was used only to illustrate how the model produced different classification outcomes. The predicted probabilities for the three false negative cases were 0.1599, 0.1391, and 0.1498, respectively. All were close to the corresponding thresholds of 0.16, 0.14, and 0.15, mainly reflecting uncertainty near the decision boundary. By contrast, the predicted probabilities for the three false positive cases were 0.9811, 0.9685, and 0.9902. This indicates that their observable features at the relevant course stages closely matched the high risk patterns identified by the model, although their final outcomes were Success.

False negative cases may therefore be more suitable for manual review or continued monitoring, while false positive cases show that stage-specific risk signals should not be treated as certain evidence of non-successful completion. Model outputs should be used to support teacher judgement and resource allocation. The case selection rules, predicted probabilities, and complete waterfall plots are presented in Appendix F, including Table F1 and Figs. F1 to F3.

6.6 Robustness, Sensitivity, and Scope of Applicability

6.6.1 Robustness to feature space, course context, and class handling

The robustness analyses examined class handling, the shared feature space, and the separation of course contexts. All comparisons were conducted only within the fixed training set. Complete results are provided in Appendix G.

SMOTE produced small increases in Recall and F1 score at all three prediction points, but reduced Precision. Changes in ROC AUC and AP were no greater than 0.003, indicating that oversampling mainly changed the classification tradeoff without substantially improving risk ranking. When model comparisons were repeated using the same 83 variables available at all three prediction points, the AP values of XGBoost were 0.7804, 0.8254, and 0.8473, respectively. These results were almost identical to those of the main analysis, suggesting that differences across prediction points were not mainly caused by variables added at later stages or by increases in the number of features.

Under five-fold cross-validation grouped by module and presentation, the AP values of XGBoost were 0.7654, 0.8184, and 0.8424, respectively. Average performance was slightly lower than that obtained through conventional stratified cross validation, and variation across folds was greater at Day 30. However, the overall pattern of stronger performance at later prediction points remained. The results also suggest that model performance was sensitive to differences across

course modules and presentations, as shown in Tables G1 to G3.

6.6.2 Sensitivity analysis for failure and future withdrawal outcomes

To examine whether the combined outcome of non-successful course completion concealed differences between Failure and Future Withdrawal, the two outcomes were analysed separately as positive outcomes. Detailed results are presented in Table 8.

Table 8. Sensitivity Analyses For Failure And Future Withdrawal Outcomes

Time	Outcome	ROC-AUC	AP
Day 30	Failure	0.8143	0.6993
Day 30	Future Withdrawal	0.6891	0.3672
Day 60	Failure	0.8596	0.7881
Day 60	Future Withdrawal	0.6926	0.3119
Day 90	Failure	0.8841	0.8238
Day 90	Future Withdrawal	0.7165	0.2870

When Failure was treated as the positive outcome, ROC AUC values at Day 30, Day 60, and Day 90 were 0.8143, 0.8596, and 0.8841, respectively, while AP values were 0.6993, 0.7881, and 0.8238. For the Future Withdrawal task, ROC AUC values were 0.6891, 0.6926, and 0.7165, while AP values were 0.3672, 0.3119, and 0.2870. Overall, performance for Future Withdrawal was lower than that for Failure.

The proportion of positive cases for Future Withdrawal decreased from 19.86% to 13.32%, so part of the decline in AP was related to changes in the class prevalence. The two outcomes showed clear differences in predictability. Combining them provides a single entry point for early warning, but it may conceal the differences between course failure and future withdrawal.

6.6.3 Subgroup performance audit

The exploratory subgroup audit did not show a clear gender difference in Recall. Recall among learners with disabilities was not lower than that among learners without disabilities, but the sample size ranged from only 217 to 252 learners, which resulted in considerable uncertainty.

More noticeable differences in risk prevalence and alert burden were observed across IMD groups. Learners in more economically disadvantaged groups generally had higher proportions of non-successful completion, higher Alert rates, and higher Recall, together with lower Specificity. These differences were jointly influenced by risk prevalence, sample size, and feature distribution. They should therefore be treated only as indicators for further review before deployment and should not be interpreted directly as evidence of algorithmic bias or causal social differences, as shown in Table H1.

7. DISCUSSION

The main finding of this study is not simply that later prediction points produced higher performance metrics, but that the early warning task in STEM online courses changes in structure as the course progresses. As learning activities continue, the prediction population is updated, available information accumulates, and the evidence used by the model shifts from multiple early signals towards more direct measures of stage-specific academic attainment. Multi-landmark early warning should therefore not be understood as repeatedly applying the same static model on different dates. Instead, it should be defined as a stage-specific prediction process in which the risk set, information conditions, and decision rules are updated together.

7.1 Information Accumulation, Risk Set Updates, and the Intervention Window

The findings support the view that accumulating information can improve stage-specific risk discrimination. However, this interpretation requires an important qualification. Differences across prediction points arise not only from the addition of new features, but also from changes in the prediction population. Learners who have already withdrawn before a prediction point are no longer potential recipients of subsequent support, while learners who register later may enter the risk set only at a later point. The prediction points therefore do not involve exactly the same groups of learners. Differences in performance are influenced by information accumulation, prediction horizon, outcome prevalence, and changes in sample composition.

Previous studies of prediction across course stages have often attributed stronger performance at later points to the continued accumulation of behavioural, submission, and assessment information [9], [30], [31]. The present findings show that the key issue in multi-landmark research is not simply whether the observation window is shortened or extended, but whether the learners included at each prediction point remain relevant for support at that time. Eligibility for the risk set is therefore not merely a routine data cleaning step. It forms part of the definition of the early warning task. Compared with repeating

predictions in a fixed sample, this study reconstructed a risk set for each landmark and kept learners assigned to the same data subset. This design improved temporal consistency among the prediction population, the information window, and the final outcome.

The similar performance of the three models provides further support for this interpretation. Under the fixed parameter settings used in this study, greater algorithmic complexity did not produce a consistent advantage. This suggests that defining the risk set, controlling the time boundary, and constructing stage-specific features may be at least as important as model complexity.

These changes in the prediction task also explain the practical tradeoffs across prediction points. Earlier prediction points provide a longer potential period for learner support, but the available risk information is more widely distributed. Later prediction points provide more concentrated risk rankings, but they are also closer to the final course outcome. The findings therefore do not support the search for a single best prediction point. Instead, different points may serve different purposes. Early points may support initial screening, while later points may be used to update support priorities. This study did not examine how the risk of individual learners changed across prediction points. The findings therefore support stage-specific reassessment, but do not directly demonstrate the effectiveness of a complete continuous early warning process.

7.2 Changes in the Basis of Risk Prediction Across Course Stages

At Day 30, model predictions were based on a combination of background conditions, platform behaviour, sustained participation, and early assessment information. At Day 60 and Day 90, the contributions of assessment participation and stage-specific academic attainment increased, while recent activity and sustained engagement continued to provide additional information. The progressive feature settings and SHAP analysis produced supporting evidence. Assessment process information added predictive value beyond background and behavioural variables. When scores, pass indicators, and weighted attainment measures were added, predictive performance increased further. The information structure used by the model therefore did not simply shift from behaviour to scores. Instead, stage-specific attainment gradually became the main source of predictive information, while sustained behavioural engagement continued to play a supporting role.

Educational XAI studies often report global feature importance at a single prediction point, which can present model explanations as a stable ranking of variables. Previous research has shown that time and learning context may change both model performance and the structure of feature importance[32]. The present study further shows that these changes occur not only in the ranking of individual variables, but also in the overall contributions of different categories of information.

However, a high predictive contribution does not necessarily mean that a feature is more suitable as a basis for intervention. Stage-specific scores are closer to the final course outcome and therefore often have strong predictive value. However, they may mainly confirm learning difficulties that have already become visible and may leave less time for subsequent support. In contrast, declining activity, interrupted submissions, and changes in participation may have lower SHAP contributions, but they may provide earlier and more actionable signals before the problem has fully developed. Educational early warning should therefore not select support targets based only on feature importance rankings. It should also consider when information becomes available and whether it can support meaningful action.

These findings further suggest that the value of educational XAI should go beyond identifying the variables used by a model. Explanations should also be connected to the timing of prediction and to actionable feedback, allowing model outputs to support timely and targeted educational action [33], [34].

7.3 Early Warning Decisions Under Capacity Constraints

The findings show that high Recall does not necessarily lead to practical decision making. The F2 optimal thresholds identified most learners who ultimately did not successfully complete the course, but also included between 64.77% and 81.36% of test learners in the alert lists. A model with good discrimination can therefore still produce a list that is too large to manage and can include many learners who ultimately complete the course successfully. ROC AUC describes the ability to

rank risk, but it cannot replace decision evaluation based on Precision, Specificity, and Alert rate.

Previous studies have attempted to connect dynamic risk signals or score thresholds with educational intervention and have noted that alert rules need to be adjusted to specific course and programme contexts[14], [35]. However, risk prediction studies still often focus on ROC AUC, Accuracy, or classification metrics at a single threshold, with less attention to whether the resulting alert list exceeds the number of learners that an institution can realistically support.

The simulated capacity scenarios further revealed the tension between the timing of an alert and the efficiency of risk identification. With similar alert list sizes, the Day 60 and Day 90 tasks covered a greater proportion of learners who did not successfully complete the course, while Day 30 provided a longer potential period for support. Earlier prediction points may therefore be more suitable for low cost support with broad coverage, whereas later points may allow limited resources to be focused on learners with higher priority. However, this possible match between prediction point and support approach is a practical implication derived from the current results rather than evidence of a tested intervention effect.

The study therefore extends model evaluation from a focus on predictive metrics to an assessment of decisions under capacity constraints. A classification threshold is not an inherent technical property of the model. It is a context specific rule shaped by the costs of missed cases and false alerts, the course stage, the type of support, and the manageable alert list size. The capacity levels used in this study are illustrative scenarios only. Alert rate provides an approximate measure of alert list size at a single prediction point and should not be treated as equivalent to actual staff workload or the total resource demand across a full course.

7.4 Implications for Multi-Landmark Early Warning

The findings indicate that multi-landmark early warning should be understood as a stage-specific decision process rather than as repeated prediction at different time points. Earlier landmarks provide a longer opportunity for support, whereas later landmarks benefit from richer behavioural and assessment information. Their value therefore lies in serving different functions within the course: earlier prediction can support broader screening, while later prediction can help refine support priorities as stronger evidence of risk becomes available. Similarly, features with high predictive contribution are not necessarily the most useful signals for intervention, because some may become informative only after learning difficulties are already well established. The timing and actionability of predictive information should therefore be considered alongside feature importance.

For practical implementation, prediction thresholds should also reflect the level of support that can realistically be provided. A highly sensitive threshold may improve risk coverage but create an alert list that exceeds available resources. Institutions could therefore update the eligible risk population and risk ranking at each landmark, using broader, lower-intensity support at earlier stages and more selective support when stronger risk evidence is available later in the course. These implications do not demonstrate the effectiveness of any specific intervention, but they provide a practical basis for aligning prediction timing, model interpretation, and support capacity in a multi-landmark early-warning system.

7.5 Limitations and Future Research

Several limitations should be considered when interpreting the findings. OULAD comes from an earlier period and represents a single open-university setting, so internal validation across course contexts cannot substitute for external validation across institutions, platforms, and time periods. The use of fixed Day 30, Day 60, and Day 90 landmarks also provides only an approximate representation of course progression, because the same calendar day may correspond to different instructional stages across courses. In addition, the actual timing of score release and feedback was unavailable, and the main analysis combined Failure and Withdrawal into a single outcome of non-successful completion, although these outcomes may differ in their predictability and underlying processes.

Other limitations concern the scope of modelling and potential deployment. Fixed parameter settings were used for model comparison, and the study did not assess the maximum performance achievable through systematic tuning. The analysis also did not track individual risk trajectories across landmarks, continuous alerting, or repeated use of support resources. Variables such as gender, disability

status, and socioeconomic background would require further ethical and fairness evaluation before practical use. Most importantly, the study was based on offline prediction and simulated decision scenarios rather than actual intervention delivery, and it therefore cannot establish whether model-triggered support improves learner outcomes.

Future research should build on these limitations by testing the framework in more recent and diverse institutional settings and by using prediction landmarks that are more closely aligned with course structure or instructional progress. Greater attention should also be given to changes in individual risk over time and to the distinction between failure and withdrawal where these outcomes reflect different pathways. Ultimately, prospective deployment under real staffing and support conditions will be necessary to determine whether stage-specific early-warning decisions translate into measurable educational benefits.

8. CONCLUSION

Using OULAD, this study developed an explainable multi-landmark early-warning framework for predicting non-successful completion in STEM online courses. Separate risk sets were defined at Days 30, 60, and 90 to include learners who remained eligible for support, while learner assignments to the training, validation, and test sets were kept fixed across landmarks. This design improved consistency between the prediction population, the information available at each stage, and the timing of prediction. This indicates that multi-landmark early warning involves more than adding new information over time; the prediction task itself changes as the population still eligible for support is updated.

Under the fixed parameter settings used in this study, logistic regression, random forest, and XGBoost showed similar overall performance, with no model consistently outperforming the others across all landmarks and evaluation metrics. Progressive feature analysis and SHAP showed that the basis of risk prediction changed across course stages. Assessment participation and stage-specific attainment became increasingly influential, while recent activity and sustained engagement remained complementary sources of information. These findings also indicate that predictive contribution should not be equated directly with intervention value, because highly informative features may become available only after learning difficulties are already well established.

The validation-based thresholds and simulated capacity scenarios further showed that high risk coverage can produce alert lists that are difficult to manage. These findings support a broader view of multi-landmark early warning as a sequence of stage-specific decision tasks rather than repeated prediction at different time points. Academic early-warning evaluation should therefore consider not only model discrimination, but also whether learners remain eligible for support, whether the required information is available at the relevant stage, and whether the resulting decision rule is compatible with available support capacity. This stage-specific perspective provides a more operationally relevant basis for coordinating prediction timing, model interpretation, and intervention feasibility across the course.

Future research should validate the framework across institutions and contemporary online-learning environments, examine individual changes in risk between landmarks, and evaluate through prospective deployment whether stage-specific warning and support strategies lead to measurable improvements in learner outcomes.

References

[1]Johora, F. T., Hasan, M. N., Rajbongshi, A., Ashrafuzzaman, M., & Akter, F. (2025). An explainable AI-based approach for predicting undergraduate students academic performance. *Array*, *26*, 100384.

[2]Hidayatulloh, W., Mahardika, F., & Junaedi, D. I. (2026). Explainable Artificial Intelligence-Based Model for Student Academic Performance Prediction. *Journal of Information System Exploration and Research*, *4*(1), 31-40.

[3]Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, *6*, 52138-52160.

[4]Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

[5]Lundberg, S. M., & Lee, S. I. (2017). A unified

approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

[6]Choi, W. C., Lam, C. T., Pang, P. C. I., & Mendes, A. J. (2025). A systematic literature review of explainable artificial intelligence for interpreting student performance prediction in computer science and STEM education. In Proceedings of the 30th ACM Conference on Innovation and Technology in Computer Science Education (Vol. 1, pp. 221–227).

[7]Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research*, *74*(1), 59-109.

[8]Bond, M., & Bedenlier, S. (2019). Facilitating student engagement through educational technology: towards a conceptual framework. *Journal of Interactive Media in Education*, *2019*(1).

[9]Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., ... & Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *Ieee Access*, *9*, 7519-7539.

[10]Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2016). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies*, *10*(1), 17-29.

[11]Nie, H., Wen, Y., Cao, B., & Liang, B. (2023, August). MOOC Dropout Prediction Using Learning Process Model and LightGBM Algorithm. In *CCF Conference on Computer Supported Cooperative Work and Social Computing* (pp. 121-136). Singapore: Springer Nature Singapore.

[12]Ifenthaler, D., & Yau, J. Y. K. (2020). Utilising learning analytics to support study success in higher education: a systematic review. *Educational technology research and development*, *68*(4), 1961-1990.

[13]Foster, E., & Siddle, R. (2020). The effectiveness of learning analytics for identifying at-risk students in higher education. *Assessment & Evaluation in Higher Education*, *45*(6), 842-854.

[14]Bañeres, D., Rodríguez-González, M. E., Guerrero-Roldán, A. E., & Cortadas, P. (2023). An early warning system to identify and intervene online dropout learners. *International Journal of Educational Technology in Higher Education*, *20*(1), 3.

[15]Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, *1*(1), 6-47.

[16]Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of educational technology & society*, *17*(4), 49-64.

[17]Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators?. *International journal of educational technology in higher education*, *16*(1), 39.

[18]Alkhasawneh, R., & Hargraves, R. H. (2014). Developing a hybrid model to predict student first year retention in STEM disciplines using machine learning techniques. *Journal of STEM Education: Innovations and Research*, *15*(3).

[19]Li, C., Herbert, N., Yeom, S., & Montgomery, J. (2022). Retention factors in STEM education identified using learning analytics: A systematic review. *Education Sciences*, *12*(11), 781.

[20]Jun, J. (2005). Understanding E-dropout?. In *International Journal on e-learning* (Vol. 4, No. 2, pp. 229-240). Association for the Advancement of Computing in Education (AACE).

[21]Baneres, D., Rodríguez-Gonzalez, M. E., & Serra, M. (2019). An early feedback prediction system for learners at-risk within a first-year higher education course. *IEEE Transactions on Learning Technologies*, *12*(2), 249-263.

[22]Jokhan, A., Sharma, B., & Singh, S. (2019). Early warning system as a predictor for student performance in higher education blended courses. *Studies in Higher Education*, *44*(11), 1900-1911.

[23]Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web*, *2*(2), 1.

[24]Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115.

[25]Nagy, M., & Molontay, R. (2024). Interpretable dropout prediction: Towards XAI-based personalized intervention. *International Journal of Artificial Intelligence in Education*, *34*(2), 274-300.

[26]Bertolini, R., Finch, S. J., & Nehm, R. H. (2021). Enhancing data pipelines for forecasting student performance: integrating feature selection with cross-validation. *International Journal of Educational Technology in Higher Education*, *18*(1), 44.

[27]Swamy, V., Radmehr, B., Krco, N., Marras, M., & Käser, T. (2022). Evaluating the explainers: black-box explainable machine learning for student success prediction in MOOCs. *arXiv preprint arXiv:2207.00551*.

[28]Alalawi, K., Athauda, R., & Chiong, R. (2025). An extended learning analytics framework integrating machine learning and pedagogical approaches for student performance prediction and intervention. *International Journal of Artificial Intelligence in Education*, *35*(3), 1239-1287.

[29]Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific data*, *4*(1), 170171.

[30]Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in Finnish higher education. *Technology in Society*, *76*, 102474.

[31]Huang, Q., & Chen, J. (2024). Enhancing academic performance prediction with temporal graph networks for massive open online courses. *Journal of Big Data*, *11*(1), 52.

[32]Tiukhova, E., Vemuri, P., Flores, N. L., Islind, A. S., Óskarsdóttir, M., Poelmans, S., ... & Snoeck, M. (2024). Explainable Learning Analytics: Assessing the stability of student success prediction models by means of explainable AI. *Decision Support Systems*, *182*, 114229.

[33]Susnjak, T. (2024). Beyond predictive learning analytics modelling and onto explainable artificial intelligence with prescriptive analytics and ChatGPT. *International Journal of Artificial Intelligence in Education*, *34*(2), 452-482.

[34]Afzaal, M., Zia, A., Nouri, J., & Fors, U. (2024). Informative feedback and explainable AI-based recommendations to support students' self-regulation. *Technology, Knowledge and Learning*, *29*(1), 331-354.

[35]Smit, N., Osler, Z., & Van Der Merwe, L. (2025, October). At-risk student identification and interventions for data science programs at a South African university. In *Frontiers in Education* (Vol. 10, p. 1501796). Frontiers Media SA

Appendix

Appendix A

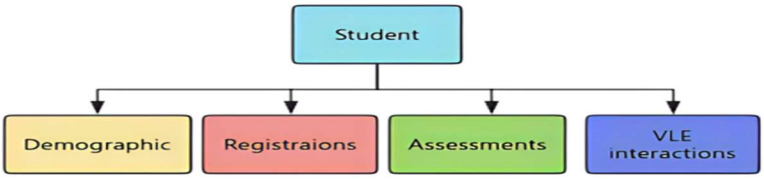


Fig. A1. Overall dataset structure (Source: Kuzilek et al., 2017)

Appendix B

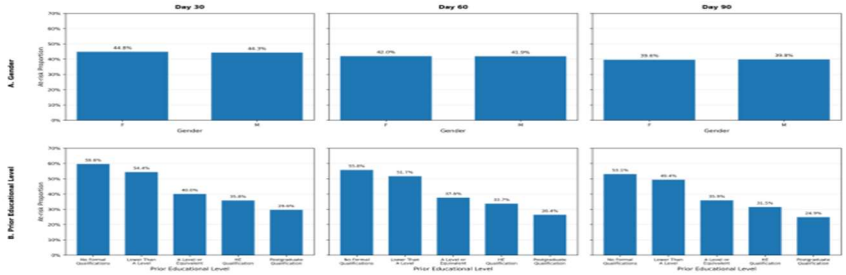


Fig. B1. Non-Successful Proportions By Gender And Prior Educational Level Across The Day30, Day60, And Day90 landmark risk sets.

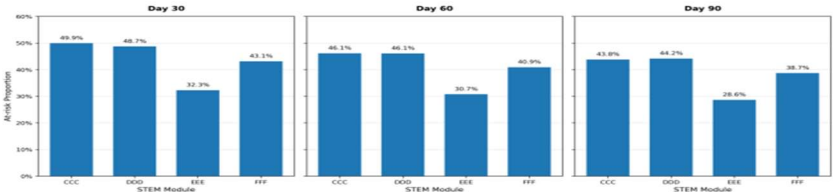


Fig. B2. Non-Successful Proportions Across STEM Modules In The Day30, Day60, And Day90 Landmark Risk Sets.

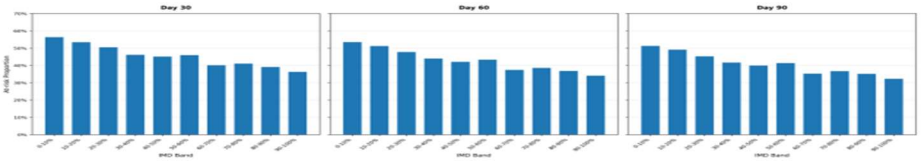


Fig. B3. Non-Successful Proportions Across IMD Bands In The Day30, Day60, And Day90 Landmark risk sets.

Appendix C

Table C1. Bootstrap Estimates And 95% Confidence Intervals For Xgboost Test-Set Performance Across The Three Landmark Risk Sets.

Time	Threshold	Metric	Estimate	95% CI lower	95% CI upper
Day 30	0.1600	ROC-AUC	0.8041	0.7886	0.8194
Day 30	0.1600	AP	0.7861	0.7679	0.8028
Day 30	0.1600	Precision	0.5203	0.5123	0.5289
Day 30	0.1600	Recall	0.9509	0.9390	0.9621
Day 30	0.1600	Specificity	0.2965	0.2745	0.3180
Day 30	0.1600	F1-score	0.6726	0.6642	0.6812
Day 30	0.1600	F2-score	0.8159	0.8065	0.8249
Day 30	0.1600	Alert rate	0.8136	0.8004	0.8272
Day 30	0.1600	Brier score	0.1775	0.1705	0.1847
Day 60	0.1400	ROC-AUC	0.8513	0.8374	0.8647
Day 60	0.1400	AP	0.8330	0.8170	0.8479
Day 60	0.1400	Precision	0.5371	0.5267	0.5478
Day 60	0.1400	Recall	0.9476	0.9345	0.9599
Day 60	0.1400	Specificity	0.4045	0.3807	0.4284
Day 60	0.1400	F1-score	0.6856	0.6753	0.6958

Time	Threshold	Metric	Estimate	95% CI lower	95% CI upper
Day 60	0.1400	F2-score	0.8220	0.8115	0.8318
Day 60	0.1400	Alert rate	0.7440	0.7295	0.7588
Day 60	0.1400	Brier score	0.1516	0.1444	0.1591
Day 90	0.1500	ROC-AUC	0.8763	0.8622	0.8894
Day 90	0.1500	AP	0.8533	0.8382	0.8678
Day 90	0.1500	Precision	0.5699	0.5576	0.5832
Day 90	0.1500	Recall	0.9230	0.9078	0.9382
Day 90	0.1500	Specificity	0.5358	0.5125	0.5597
Day 90	0.1500	F1-score	0.7047	0.6927	0.7166
Day 90	0.1500	F2-score	0.8213	0.8086	0.8336
Day 90	0.1500	Alert rate	0.6477	0.6323	0.6634
Day 90	0.1500	Brier score	0.1356	0.1281	0.1434

Appendix D

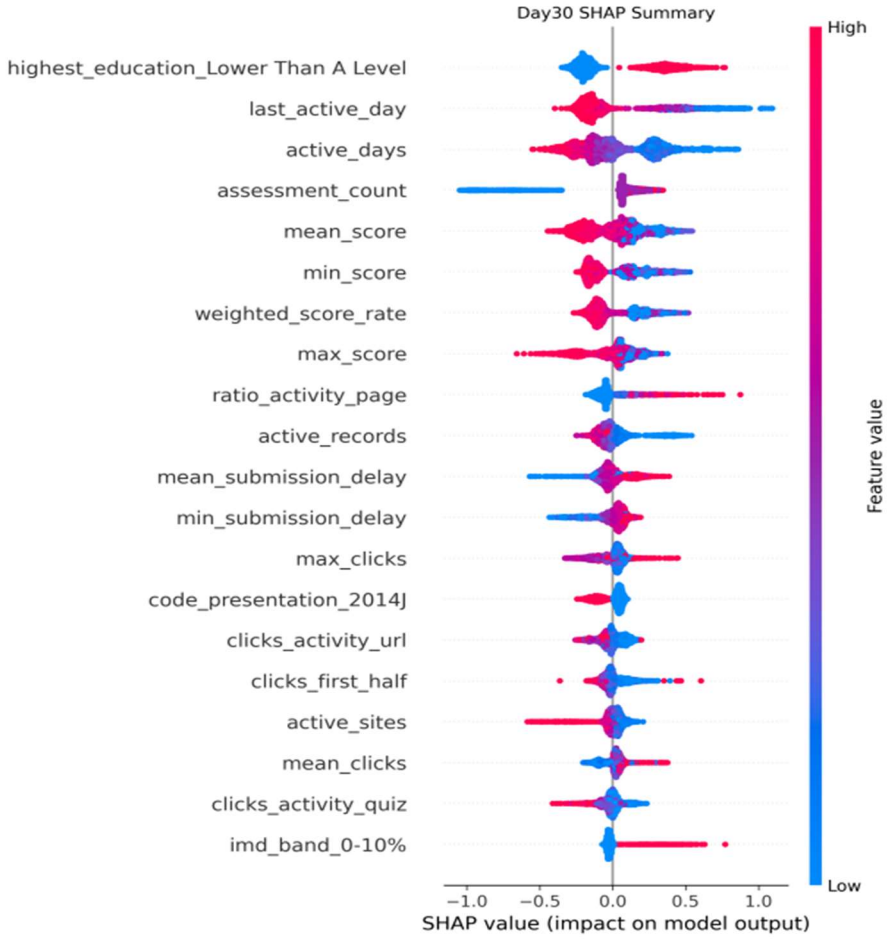


Fig. D1. SHAP Summary Plot For The Day 30 Model.

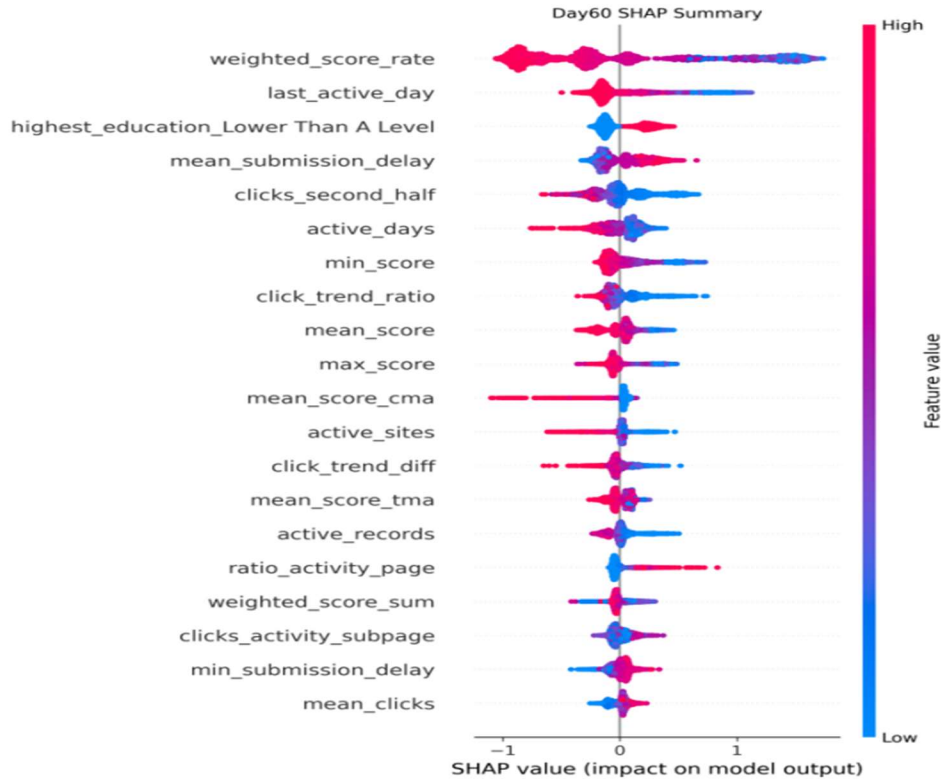


Fig. D2. SHAP Summary Plot For The Day 60 Model.

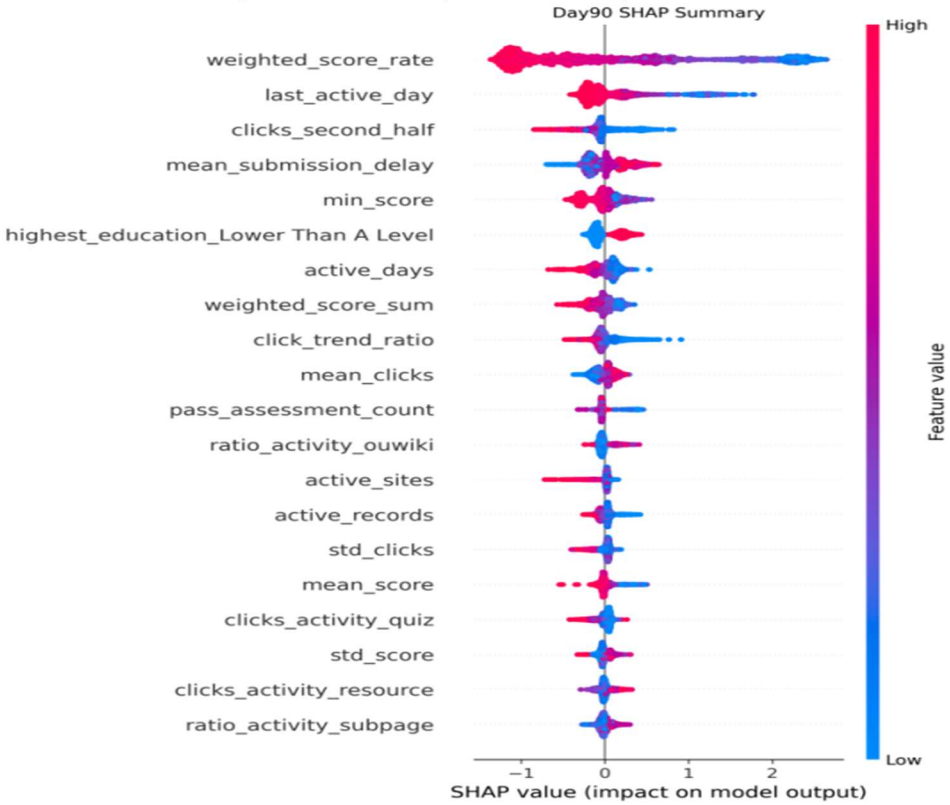


Fig. D3. SHAP Summary Plot For The Day 90 Model.

Appendix E

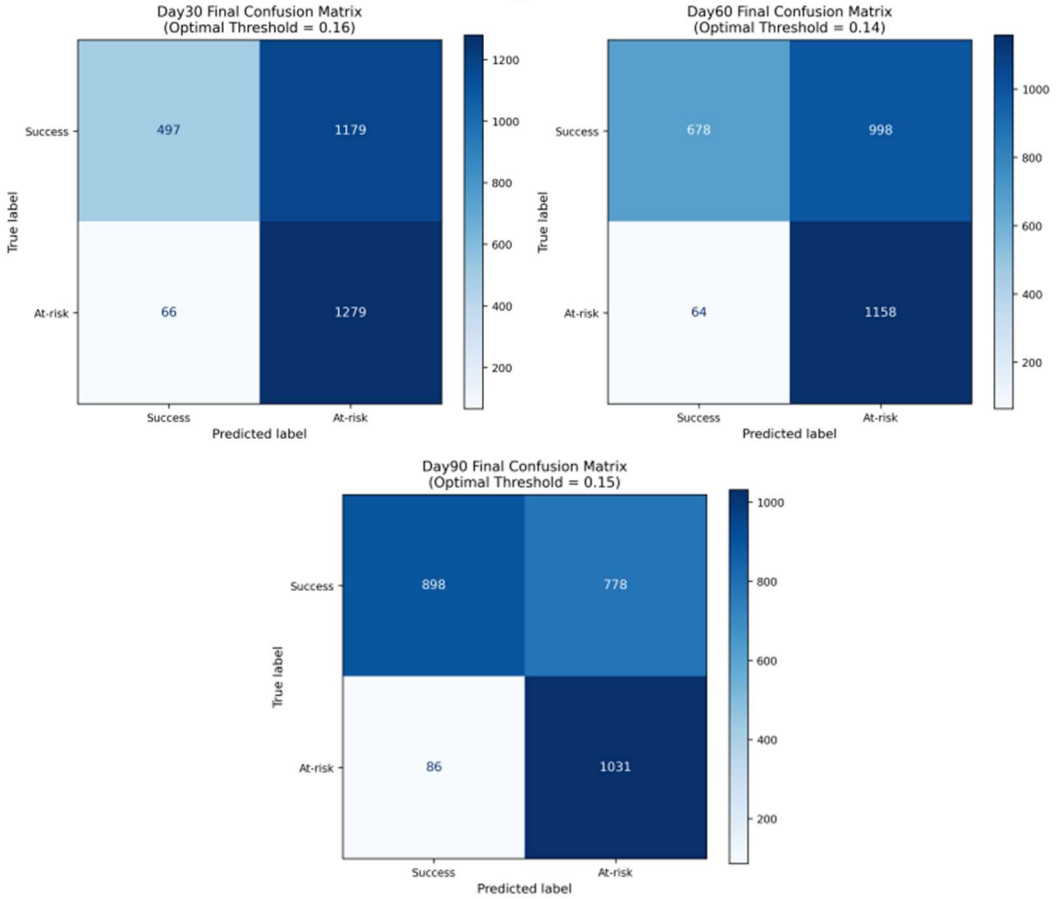


Fig. E1. Confusion Matrices On The Fixed Test Set Using The Validation-Selected F2 Thresholds At Day 30, Day 60, And Day 90.

Table E1. Test-Set Performance Under Capacity-Constrained Intervention Thresholds.

Time	Target capacity	Validation threshold	Test alert rate	Precision	Recall	Specificity	F1-score	F2-score	TN	FP	FN	TP
Day 30	0.1	0.8868	0.0924	0.9570	0.1985	0.9928	0.3288	0.2359	1,664	12	1,078	267
Day 30	0.2	0.7533	0.1940	0.8703	0.3792	0.9547	0.5282	0.4274	1,600	76	835	510
Day 30	0.3	0.6159	0.2906	0.7961	0.5197	0.8932	0.6289	0.5585	1,497	179	646	699
Day 30	0.4	0.4957	0.3880	0.7338	0.6394	0.8138	0.6834	0.6563	1,364	312	485	860
Day 30	0.5	0.3935	0.4906	0.6687	0.7368	0.7070	0.7011	0.7221	1,185	491	354	991
Day 60	0.1	0.9478	0.0935	0.9815	0.2177	0.9970	0.3563	0.2578	1,671	5	956	266
Day 60	0.2	0.7922	0.1853	0.9236	0.4059	0.9755	0.5640	0.4571	1,635	41	726	496
Day 60	0.3	0.5926	0.2912	0.8389	0.5794	0.9189	0.6854	0.6176	1,540	136	514	708
Day 60	0.4	0.4452	0.3865	0.7625	0.6989	0.8413	0.7293	0.7107	1,410	266	368	854

Time	Target capacity	Validation threshold	Test alert rate	Precision	Recall	Specificity	F1-score	F2-score	TN	FP	FN	TP
Day 60	0.5	0.3284	0.4879	0.6832	0.7905	0.7327	0.7329	0.7664	1,228	448	256	966
Day 90	0.1	0.9738	0.1003	0.9929	0.2489	0.9988	0.3980	0.2928	1,674	2	839	278
Day 90	0.2	0.7996	0.1898	0.9396	0.4458	0.9809	0.6047	0.4982	1,644	32	619	498
Day 90	0.3	0.5794	0.2889	0.8513	0.6150	0.9284	0.7141	0.6512	1,556	120	430	687
Day 90	0.4	0.4009	0.3910	0.7500	0.7332	0.8371	0.7415	0.7365	1,403	273	298	819
Day 90	0.5	0.2635	0.5091	0.6617	0.8424	0.7130	0.7412	0.7988	1,195	481	176	941

Appendix F

Table F1. Selected Correctly Classified And Misclassified Cases At Day 30, Day 60, And Day 90

Time	Case type	Student ID	True label	Predicted decision	Predicted risk probability	Threshold
Day 30	Correct non-alert	895524	Success	No alert	0.0158	0.16
Day 30	Correct alert	494335	Non-successful completion	Alert	0.9955	0.16
Day 30	False negative	158575	Non-successful completion	No alert	0.1599	0.16
Day 30	False positive	597319	Success	Alert	0.9811	0.16
Day 60	Correct non-alert	2088765	Success	No alert	0.0125	0.14
Day 60	Correct alert	435294	Non-successful completion	Alert	0.9986	0.14
Day 60	False negative	167557	Non-successful completion	No alert	0.1391	0.14
Day 60	False positive	652333	Success	Alert	0.9685	0.14
Day 90	Correct non-alert	2088765	Success	No alert	0.0076	0.15
Day 90	Correct alert	577401	Non-successful completion	Alert	0.9990	0.15
Day 90	False negative	554038	Non-successful completion	No alert	0.1498	0.15
Day 90	False positive	363122	Success	Alert	0.9902	0.15

Journal of Theoretical and Applied Information Technology
15st September 2026. Vol.104. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

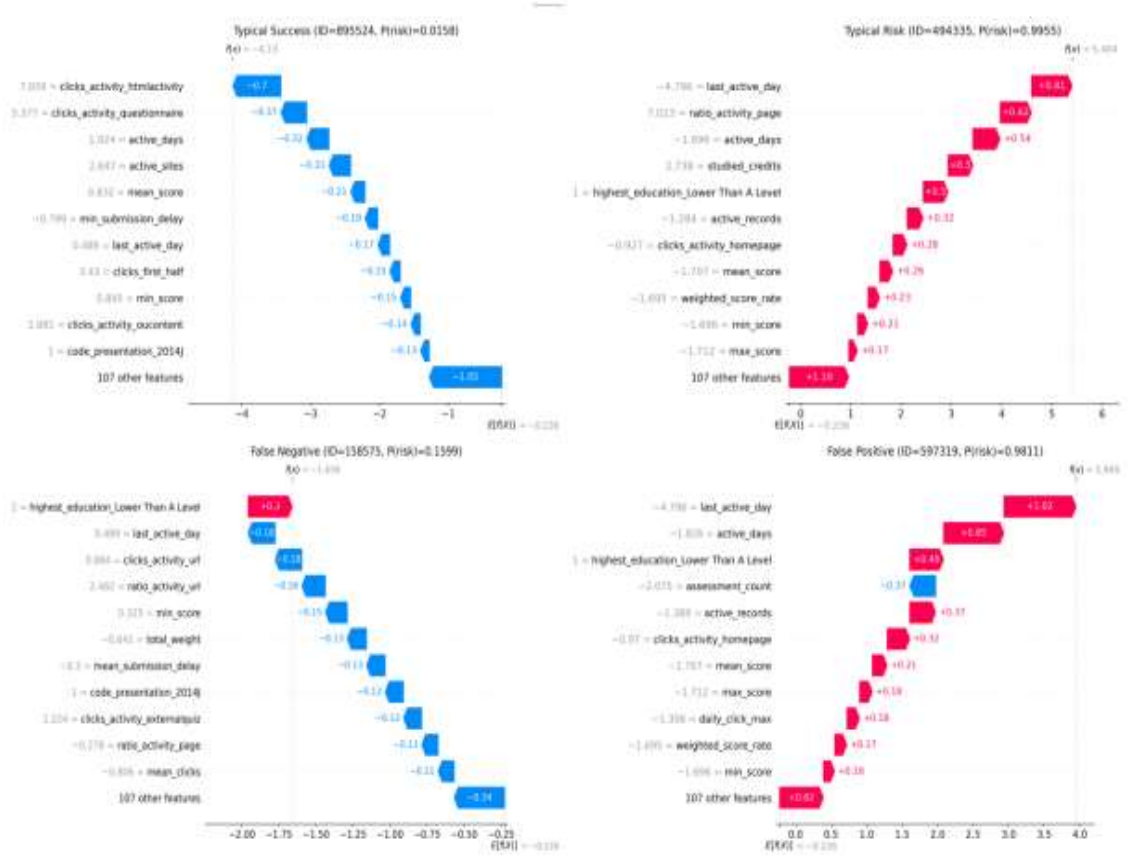


Fig. F1. SHAP-Based Local Explanations For Illustrative Cases At Day 30.

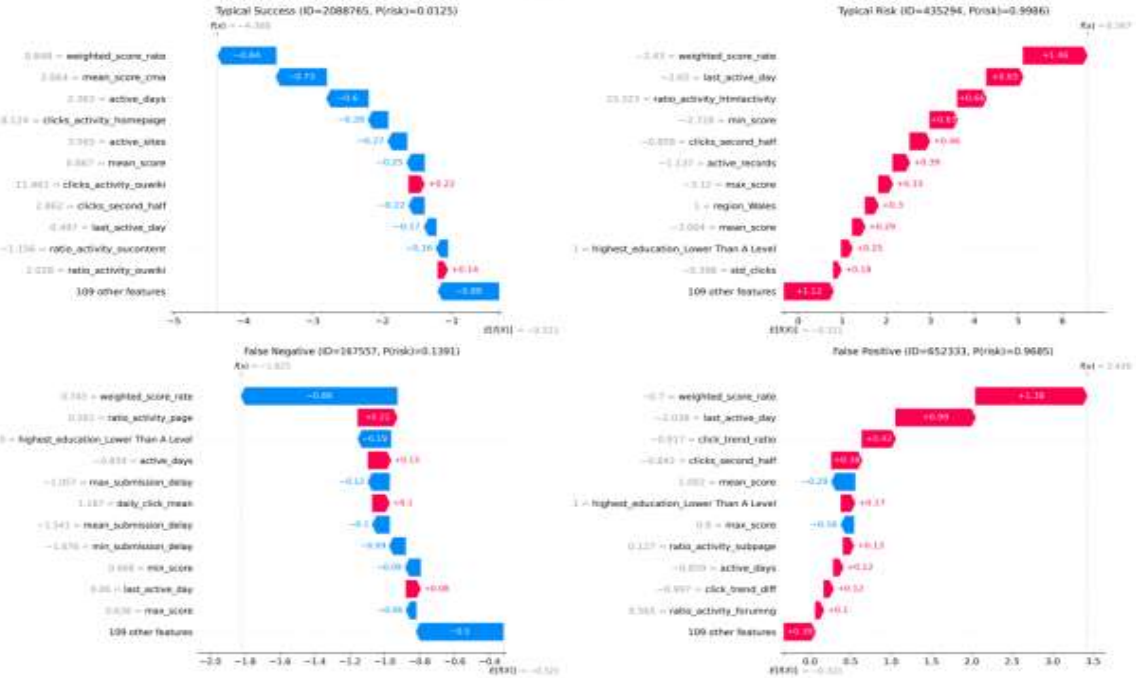


Fig. F2. SHAP-Based Local Explanations For Illustrative Cases At Day 60.

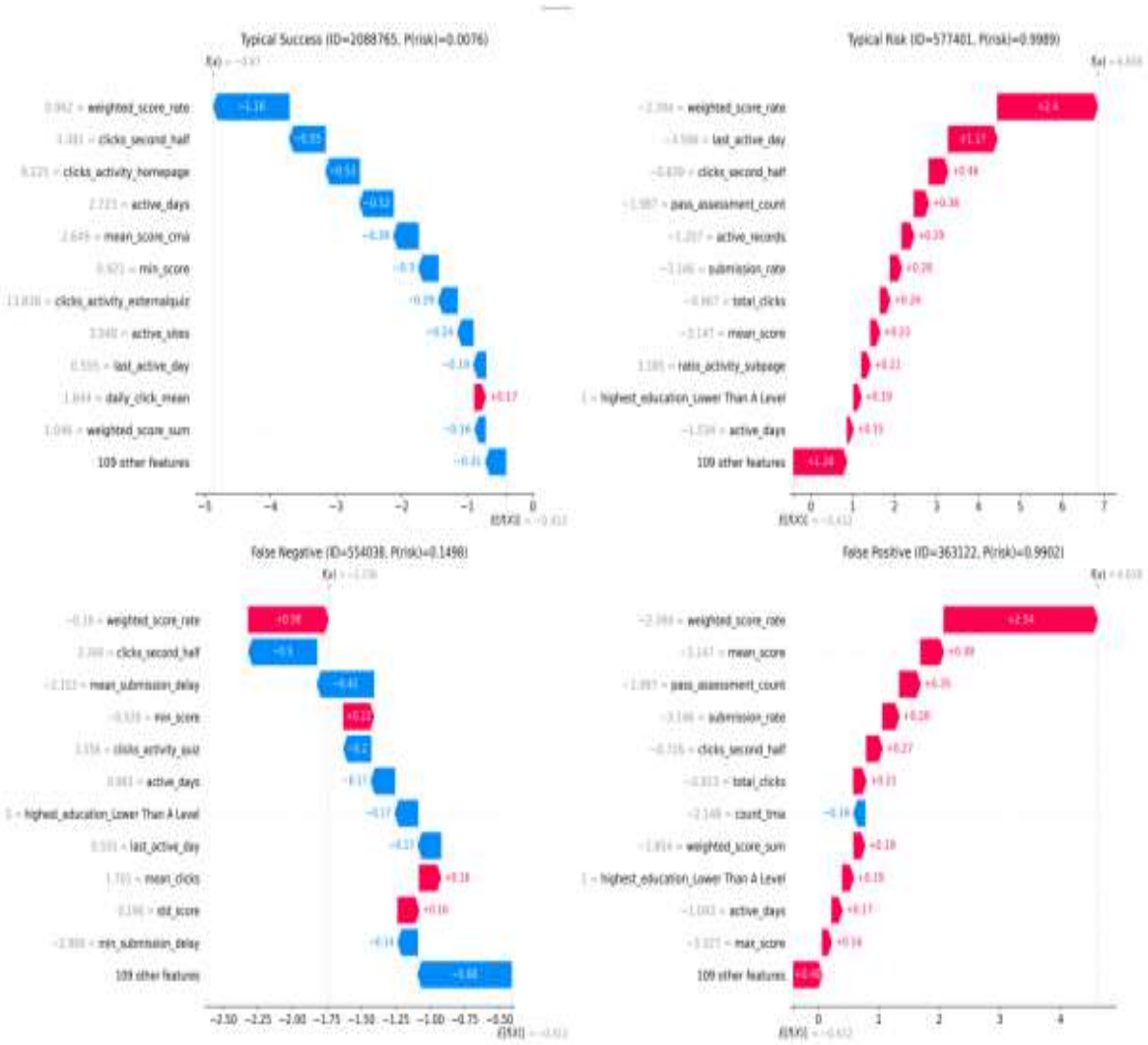


Fig. F3. SHAP-Based Local Explanations For Illustrative Cases At Day 90.

Appendix G. Robustness and Sensitivity Analyses

Table G1. Cross-Validation Performance Of Xgboost With And Without SMOTE.

Time	Setting	Train pool	Accuracy	Precision	Recall	F1-score	ROC-AUC	AP
Day 30	Original	9,037	0.7321 ± 0.0107	0.7301 ± 0.0162	0.6304 ± 0.0211	0.6763 ± 0.0144	0.7994 ± 0.0094	0.7804 ± 0.0111
Day 30	SMOTE	9,037	0.7309 ± 0.0105	0.7223 ± 0.0117	0.6406 ± 0.0247	0.6788 ± 0.0161	0.8010 ± 0.0101	0.7830 ± 0.0115
Day 60	Original	8,676	0.7736 ± 0.0053	0.7734 ± 0.0191	0.6554 ± 0.0192	0.7091 ± 0.0066	0.8431 ± 0.0074	0.8257 ± 0.0057
Day 60	SMOTE	8,676	0.7718 ± 0.0047	0.7609 ± 0.0148	0.6688 ± 0.0119	0.7116 ± 0.0036	0.8440 ± 0.0052	0.8265 ± 0.0042
Day 90	Original	8,368	0.8051 ± 0.0042	0.7998 ± 0.0065	0.6837 ± 0.0146	0.7371 ± 0.0079	0.8681 ± 0.0053	0.8475 ± 0.0063
Day 90	SMOTE	8,368	0.8058 ± 0.0045	0.7881 ± 0.0105	0.7037 ± 0.0178	0.7433 ± 0.0081	0.8697 ± 0.0036	0.8481 ± 0.0064

Table G2. Robustness Check Using A Common Cross-Time Feature Set

Time	Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	AP
Day 30	LR	0.7346 ± 0.0101	0.7392 ± 0.0152	0.6227 ± 0.0202	0.6757 ± 0.0140	0.8061 ± 0.0079	0.7883 ± 0.0099
Day 30	XGBoost	0.7321 ± 0.0107	0.7301 ± 0.0162	0.6304 ± 0.0211	0.6763 ± 0.0144	0.7994 ± 0.0094	0.7804 ± 0.0112
Day 30	RF	0.7291 ± 0.0068	0.7390 ± 0.0126	0.6040 ± 0.0212	0.6644 ± 0.0118	0.7967 ± 0.0102	0.7753 ± 0.0099
Day 60	LR	0.7778 ± 0.0089	0.7847 ± 0.0258	0.6526 ± 0.0160	0.7121 ± 0.0082	0.8481 ± 0.0072	0.8302 ± 0.0060
Day 60	RF	0.7773 ± 0.0112	0.7871 ± 0.0276	0.6474 ± 0.0117	0.7101 ± 0.0108	0.8446 ± 0.0065	0.8241 ± 0.0080
Day 60	XGBoost	0.7740 ± 0.0060	0.7717 ± 0.0174	0.6589 ± 0.0180	0.7105 ± 0.0079	0.8433 ± 0.0072	0.8254 ± 0.0060
Day 90	LR	0.8068 ± 0.0096	0.8006 ± 0.0103	0.6879 ± 0.0222	0.7398 ± 0.0154	0.8746 ± 0.0054	0.8535 ± 0.0066
Day 90	RF	0.8037 ± 0.0042	0.8077 ± 0.0094	0.6682 ± 0.0179	0.7311 ± 0.0088	0.8685 ± 0.0039	0.8464 ± 0.0062
Day 90	XGBoost	0.8051 ± 0.0050	0.8007 ± 0.0082	0.6825 ± 0.0202	0.7367 ± 0.0103	0.8683 ± 0.0054	0.8473 ± 0.0066

Table G3. Model Performance Under Stratified Grouped Five-Fold Cross-Validation Across Module–Presentation Groups

Time	Model	n groups	Accuracy	Precision	Recall	F1-score	ROC-AUC	AP
Day 30	LR	13	0.7209 ± 0.0195	0.7139 ± 0.0367	0.5965 ± 0.1570	0.6350 ± 0.0970	0.7806 ± 0.0276	0.7460 ± 0.0666
Day 30	RF	13	0.7277 ± 0.0168	0.7300 ± 0.0333	0.5810 ± 0.0786	0.6440 ± 0.0566	0.7828 ± 0.0204	0.7563 ± 0.0455
Day 30	XGBoost	13	0.7307 ± 0.0187	0.7227 ± 0.0328	0.6080 ± 0.0737	0.6577 ± 0.0489	0.7872 ± 0.0206	0.7654 ± 0.0443
Day 60	LR	13	0.7711 ± 0.0193	0.7754 ± 0.0409	0.6431 ± 0.0489	0.7007 ± 0.0241	0.8362 ± 0.0179	0.8183 ± 0.0199
Day 60	RF	13	0.7749 ± 0.0090	0.7837 ± 0.0321	0.6359 ± 0.0431	0.7011 ± 0.0296	0.8403 ± 0.0077	0.8191 ± 0.0238
Day 60	XGBoost	13	0.7711 ± 0.0090	0.7680 ± 0.0218	0.6470 ± 0.0523	0.7009 ± 0.0319	0.8387 ± 0.0082	0.8184 ± 0.0222
Day 90	LR	13	0.7978 ± 0.0237	0.7956 ± 0.0438	0.6806 ± 0.0973	0.7269 ± 0.0438	0.8649 ± 0.0158	0.8405 ± 0.0303
Day 90	RF	13	0.8006 ± 0.0204	0.8020 ± 0.0328	0.6681 ± 0.0565	0.7275 ± 0.0400	0.8649 ± 0.0159	0.8431 ± 0.0354
Day 90	XGBoost	13	0.8001 ± 0.0179	0.7899 ± 0.0294	0.6811 ± 0.0482	0.7309 ± 0.0370	0.8642 ± 0.0171	0.8424 ± 0.0356

Appendix H

Table H1. Exploratory Subgroup Audit

Time	Threshold	Subgroup	Category	N	Non-success N	Non-success rate	Precision	Recall	FPR	Specificity	Alert rate	Brier score
Day 30		Gender	F	792	346	0.4369	0.5092	0.9566	0.7152	0.2848	0.8207	0.1749
Day 30		Gender	M	2,229	999	0.4482	0.5243	0.9489	0.6992	0.3008	0.8111	0.1784
Day 30		Disability	N	2,769	1,196	0.4319	0.5065	0.9498	0.7038	0.2962	0.8100	0.1778
Day 30		Disability	Y	252	149	0.5913	0.6651	0.9597	0.6990	0.3010	0.8532	0.1745
Day 30		IMD group	Less deprived (70–100%)	807	303	0.3755	0.4487	0.9373	0.6925	0.3075	0.7844	0.1771
Day 30	0.16	IMD group	Middle (30–70%)	1,230	568	0.4618	0.5488	0.9507	0.6707	0.3293	0.8000	0.1839
Day 30		IMD group	Missing	180	47	0.2611	0.3203	0.8723	0.6541	0.3459	0.7111	0.1695
Day 30		IMD group	More deprived (0–30%)	804	427	0.5311	0.5806	0.9696	0.7931	0.2069	0.8868	0.1699
Day 60		Gender	F	757	311	0.4108	0.5158	0.9453	0.6188	0.3812	0.7530	0.1596
Day 60		Gender	M	2,141	911	0.4255	0.5448	0.9484	0.5870	0.4130	0.7408	0.1488
Day 60		Disability	N	2,668	1,095	0.4104	0.5277	0.9470	0.5900	0.4100	0.7365	0.1530
Day 60		Disability	Y	230	127	0.5522	0.6335	0.9528	0.6796	0.3204	0.8304	0.1359
Day 60		IMD group	Less deprived (70–100%)	775	271	0.3497	0.4866	0.9373	0.5317	0.4683	0.6735	0.1498
Day 60	0.14	IMD group	Middle (30–70%)	1,180	518	0.4390	0.5428	0.9421	0.6208	0.3792	0.7619	0.1564
Day 60		IMD group	Missing	176	43	0.2443	0.3495	0.8372	0.5038	0.4962	0.5852	0.1488
Day 60		IMD group	More deprived (0–30%)	767	390	0.5085	0.6013	0.9744	0.6684	0.3316	0.8240	0.1467
Day 90		Gender	F	731	285	0.3899	0.5549	0.9228	0.4731	0.5269	0.6484	0.1407
Day 90		Gender	M	2,062	832	0.4035	0.5753	0.9231	0.4610	0.5390	0.6474	0.1338
Day 90		Disability	N	2,576	1,003	0.3894	0.5595	0.9192	0.4615	0.5385	0.6398	0.1365
Day 90		Disability	Y	217	114	0.5253	0.6770	0.9561	0.5049	0.4951	0.7419	0.1260
Day 90		IMD group	Less deprived (70–100%)	761	257	0.3377	0.5188	0.9105	0.4306	0.5694	0.5926	0.1411
Day 90	0.15	IMD group	Middle (30–70%)	1,138	476	0.4183	0.5847	0.9139	0.4668	0.5332	0.6538	0.1408
Day 90		IMD group	Missing	172	39	0.2267	0.3875	0.7949	0.3684	0.6316	0.4651	0.1315
Day 90		IMD group	More deprived (0–30%)	722	345	0.4778	0.6199	0.9594	0.5385	0.4615	0.7396	0.1228