

DOI: 10.5281/zenodo.23465775

Link: <https://zenodo.org/records/23465775>

PREDICTIVE LEARNING ANALYTICS FOR STUDENT RETENTION IN DIGITAL DISTANCE EDUCATION: MODELS AND ECONOMIC IMPLICATIONS

Yakhshiboev Rustam Erkinboy ugli

*PhD, Associate Professor, Tashkent State University of Economics, Tashkent, Uzbekistan
r.yaxshiboyev@tsue.uz*

Abstract – Student attrition remains a persistent weakness of distance education, causing losses for learners, universities and society. Predictive learning analytics (PLA), which uses institutional and digital trace data to identify students at risk of dropping out, is increasingly promoted as a remedy, yet its economic value is rarely assessed. This article aims to synthesise evidence on the models used to identify at-risk students in digital distance education and to determine the conditions under which their use is economically justified. The study combines an integrative review of 31 publications from 2013–2026, covering predictive modelling studies, intervention evaluations, systematic reviews and research on fairness and dropout economics, with a scenario-based cost–benefit model of an early warning system for a cohort of 1,000 students. The review shows that models based on learning platform activity and assessment results reach high discrimination, with AUC values mostly between 0.8 and 0.95, that accuracy increases as the course progresses, and that pre-entry data alone predict dropout poorly. Evidence on retention effects is mixed: teachers' use of predictions was associated with better pass rates, whereas two controlled evaluations found no effect of prediction-driven support on dropout. The economic model shows that prediction without an effective intervention generates only costs and that the break-even intervention effectiveness ranges from 3.7% to 13.9% of correctly identified at-risk students. The effectiveness and cost of interventions affect profitability more than marginal gains in model accuracy. Practical recommendations on data integration, model choice, intervention design and budgeting are formulated for distance education providers in Uzbekistan.

Keywords: *predictive learning analytics, student retention, dropout prediction, distance education, early warning system, machine learning, cost–benefit analysis, return on investment, algorithmic fairness, higher education economics*

INTRODUCTION

Distance education has moved from the periphery to the mainstream of higher education. Digital platforms, virtual learning environments (VLEs) and online assessment allow universities to enrol large numbers of working adults and students from remote regions. Yet distance education has a persistent weakness: many of the students who enrol never complete their studies. Comparing graduation rates across distance institutions, Simpson [31] found that many of them graduate fewer than a quarter of the proportion of students who graduate from conventional institutions – a gap he described as the distance education deficit. Dropout is the outcome of a long decision process shaped by individual, institutional and system-level factors [4, 9], and it represents a loss for the student, for the institution that has invested in teaching and for the economy as a whole.

The digitalisation of distance education has, at the same time, created a new resource for addressing this problem. Every login, page view, forum post and submitted assignment leaves a digital trace. Predictive learning analytics (PLA) uses such traces, together with administrative and

assessment data, to estimate the probability that a student will fail or withdraw and to alert teachers or advisers in time to intervene [28, 30]. Large open datasets, such as the Open University Learning Analytics Dataset with records of 32,593 students [25], have stimulated a rapidly growing body of research in which increasingly sophisticated machine learning models are reported to identify at-risk students with high accuracy [5]. Early warning systems based on such predictions have been introduced at distance and fully online universities, where the absence of face-to-face contact makes digital signals the main source of information about students' engagement [8, 22].

However, the accuracy of a prediction is not an end in itself. A prediction creates value only if it triggers an intervention that changes a student's trajectory, and every element of this chain – data infrastructure, model development, staff time for contacting students – has a cost. Systematic reviews note that rigorous, large-scale evidence of the effectiveness of learning analytics in supporting study success is still lacking [24, 34], and controlled evaluations have reported that prediction-driven support did not reduce dropout [14, 27]. The scientific problem therefore lies in the gap between the predictive and the economic perspectives: research on PLA has concentrated on improving model accuracy, while the question of when the use of PLA pays off for a university has received little attention. For university managers, the decisive question is not how accurate a model is, but whether the resources spent on prediction and follow-up are returned through the tuition revenue preserved by retained students.

This problem is directly relevant to Uzbekistan. The Law “On Education” recognises distance education as a form of education without interruption from work and defines it as learning delivered through information and communication technologies and the internet [1]; the Concept for the development of the higher education system until 2030 envisages the wide introduction of digital technologies and distance learning [2]; and the Strategy for the development of artificial intelligence technologies until 2030 sets the goal of developing human resource capacity in the application of AI [3]. As distance enrolment grows, universities need evidence-based tools for retaining students and clear criteria for deciding whether investments in such tools are justified.

The aim of this study is to synthesise evidence on the predictive models used to identify students at risk of dropout in digital distance education and to determine the conditions under which their use is economically justified. The objectives are: (1) to systematise the data sources, algorithms and performance of predictive models reported in recent research; (2) to summarise the evidence on the effects of prediction-driven interventions on retention; (3) to develop a cost–benefit model that links model quality and intervention effectiveness to the economic result of an early warning system; and (4) to formulate recommendations for distance education providers.

The object of the study is student retention in digital distance higher education; the subject is predictive learning analytics models and the economic conditions of their use. The study addresses three research questions: RQ1 – which data and models predict dropout in digital distance education, and how accurately? RQ2 – what evidence exists that prediction-driven interventions improve retention? RQ3 – under what conditions does an early warning system generate a positive economic return? It is hypothesised that the economic return of PLA depends more on the effectiveness and cost of interventions than on marginal improvements in model accuracy. The scientific novelty of the study lies in combining a synthesis of predictive and intervention evidence with an explicit cost–benefit model, in deriving break-even conditions for intervention effectiveness, and in applying these results to the regulatory context of distance education in Uzbekistan.

LITERATURE REVIEW

Research on student dropout has long emphasised that withdrawal is rarely a single event. In a review of the international literature, Behr et al. [9] describe dropout as the end point of a long decision process and group its causes into three areas: the national education system (for example, financing policy), the higher education institution (for example, institution type and teaching quality) and the students themselves, both before and during their studies (for example, type of secondary school and working while studying). The socio-economic review by Aina et al. [4] similarly shows

that persistence depends on a combination of individual, institutional and economic factors whose influence is mediated by the student's academic and social integration. The two reviews differ in emphasis – economic research stresses informational problems, sociological research relational ones – but both conclude that some determinants are fixed and can serve only for description, whereas others can be changed by institutional action. Behr et al. [9] explicitly recommend the use of data mining on comprehensive datasets to support early warning systems for students at risk.

The concept of dropout itself is operationalised differently across studies, which complicates comparison and has direct economic implications. In a distance technical course, a student was considered to have dropped out after six weeks without any interaction with the learning platform [29]; at a campus university, dropout was defined as not registering for any course for two consecutive years without completing the degree [18]; and in another study, as at least four consecutive terms without re-enrolment and without graduation [17]. Short, activity-based definitions are well suited to early warning, because they capture disengagement while a student can still be reached, whereas long, enrolment-based definitions correspond more closely to the actual loss of tuition revenue. An institution that plans an early warning system therefore has to decide which of these outcomes it seeks to predict and which it seeks to prevent.

For distance education these general findings acquire a specific character. Distance students are typically older, combine study with work and family obligations and have fewer opportunities for informal contact with teachers and peers. Simpson [31] argues that distance institutions have invested heavily in the production of teaching materials, especially online, but too little in motivating students, and that institutional attitudes to retention are a greater barrier than resources. Importantly for the economic perspective, he argues that proactive motivational contact with students, if carefully designed, can be financially viable and even generate surpluses for institutions. This argument links retention directly to institutional economics and provides the starting point for the present study.

Predictive learning analytics developed as the technological response to this challenge. A decade review by Sghir, Adadi and Lahmer [30] traces the evolution of PLA from 2012 to 2022, and the survey by Prenkaj et al. [28] systematises machine learning approaches to dropout prediction in online courses, proposing a hierarchical classification of studies according to their design choices and paying particular attention to deep sequential models, evaluation metrics, data collection and privacy. Albreiki, Zaki and Alashwal [5], reviewing 78 studies published in 2009–2021, found that decision trees, logistic regression, naive Bayes and support vector machines were the most common algorithms for performance prediction, that student interaction with e-learning platforms was the most frequently used type of attribute, and that class balancing, temporal features, deep learning and dynamic performance tracking remained under-researched. The release of the Open University Learning Analytics Dataset [25], which combines demographic information, assessment results and daily summaries of students' interactions with the VLE for 32,593 students, provided a common benchmark that has since been used to compare a wide range of models.

The way model quality is measured has direct economic consequences. Dropout prediction is usually an imbalanced classification problem: students who withdraw are a minority, so a model that predicts that every student will stay can achieve high overall accuracy while identifying no one at risk. For this reason, surveys of the field emphasise the choice of evaluation metrics [28] and the need for class balancing [5]. Recall (the share of actual dropouts that the model identifies) determines how many students who need help are reached, whereas precision (the share of flagged students who are actually at risk) determines how much of the intervention budget is spent on students who would have continued anyway. The area under the ROC curve (AUC) summarises the ability of a model to rank students by risk regardless of the chosen threshold, but it does not show the cost of a particular decision rule. An institution that uses PLA therefore has to choose a threshold that balances the cost of false alarms against the cost of overlooked dropouts – an economic rather than a purely statistical decision.

At the same time, several studies questioned how far predictive models can be generalised. Gašević et al. [16] showed that the predictive power of indicators of learning platform activity

depended on the instructional conditions of individual courses, so that models trained across courses could misrepresent the factors relevant to a particular course; the authors warned that learning analytics should not promote a “one size fits all” approach. Comparing 17 blended courses that used the Moodle platform, Conijn et al. [13] found that the predictive value of platform data differed substantially between courses and that assessment data collected during the course were more informative than platform activity for early intervention. Bird et al. [12] systematically compared modelling approaches on administrative data and showed that simple models performed about as well as complex machine learning models, while the ranking of individual students by predicted risk varied substantially across approaches. The last finding is particularly important for practice, because it means that the choice of model determines which students receive support.

Experience with institution-wide implementation is concentrated at a small number of universities, most notably the Open University in the United Kingdom, a distance learning institution that developed the OU Analyse system. The system uses machine learning to flag students at risk of not submitting or failing their next assignment and presents weekly predictions to teachers in interactive dashboards [20]. At the Universitat Oberta de Catalunya, a fully online university, an early warning system based on assessment grades was integrated into teacher dashboards and updated after each graded activity [8]. Interviews with 20 educational managers involved in its roll-out allowed Herodotou et al. [21] to formulate guidelines for implementing PLA at scale, and a longitudinal case study described how the system was scaled across the university [22]. Analysing students whose risk had been predicted incorrectly, Hlosta et al. [23] found that the errors were often caused by unexpected events in students' lives – changes in family or work responsibilities, sudden health problems or computer failures – that no model could foresee, and concluded that algorithmic predictions must be complemented by the judgement of teachers and the voice of students. Beyond the Open University, adoption remains limited: consultations with senior managers from 83 institutions in 24 European countries showed that learning analytics was usually small in scale, used mainly by teaching and support staff, and rarely involved students as stakeholders [35].

The studies also differ in who acts on a prediction and how. At the Open University, predictions were given to teachers, who decided whether and how to contact students [20]; at a Dutch university, risk information was given to student counsellors [27]; in a university retention programme, flagged students received early learning and remediation support [14]; and at a fully online university, predictions were combined with dashboards and feedback from teachers [8]. These interventions differ in intensity and therefore in cost, from an automated message to individual coaching [11]. Because the literature on PLA rarely reports what the follow-up actually consisted of, it is difficult to distinguish the contribution of the prediction from that of the support that followed it.

The use of predictive models also raises ethical questions. Slade and Prinsloo [32] identified challenges related to the collection and interpretation of student data, consent, privacy, de-identification and the classification of students, and proposed six principles grounded in a sociocritical perspective that takes account of power, surveillance and transparency. Baker and Hawn [7] reviewed evidence that predictive models in education can be less accurate for some demographic groups than for others and discussed the causes of such algorithmic bias and ways to address it. To make such differences measurable, Gardner, Brooks and Baker [15] proposed slicing analysis and the ABROCA statistic, which quantifies the difference in a model's performance between groups of students.

Evidence on the economics of retention measures is much scarcer than evidence on prediction. In a randomised experiment, Bettinger and Baker [11] found that individual student coaching increased persistence and was more cost-effective than increases in financial aid, but the coaching was not based on risk prediction. Methods of economic evaluation in education are well established [26]. Cost-effectiveness analysis compares alternative interventions by their cost per unit of effect – for example, the cost per additional retained student – whereas benefit–cost analysis expresses the effects in monetary terms and compares them with costs. Both approaches require that all resources used by an intervention, including staff time and infrastructure, be identified and valued, not only the

direct expenditure. Applied to PLA, this means that the cost of an early warning system includes not only software and data infrastructure but also the time of the teachers and advisers who contact flagged students – including those flagged incorrectly. Such methods have, however, rarely been applied to PLA. The literature therefore reveals a gap: studies of predictive models report accuracy but not costs, studies of interventions report effects but rarely link them to the quality of prediction, and the economic conditions under which an early warning system pays off have not been formalised. The present study addresses this gap.

METHODOLOGY

The study combined two methods. The first was an integrative literature review [33], chosen because the research field includes predictive modelling studies, intervention evaluations, systematic reviews and conceptual work on ethics and the economics of dropout, which differ too much in design and outcome measures to be pooled in a statistical meta-analysis. The second was scenario-based cost–benefit modelling [26], chosen because primary studies of PLA rarely report costs, so that the economic conditions of its use can be examined only by combining effect and accuracy parameters from the literature with explicit cost assumptions.

The literature search was conducted in October 2026 using academic search engines and the publisher platforms ScienceDirect, Wiley Online Library, SpringerLink, the ACM Digital Library, SAGE Journals, MDPI and the Journal of Learning Analytics. Combinations of the following terms were used: (“predictive learning analytics” OR “dropout prediction” OR “early warning system” OR “at-risk students”) AND (“distance education” OR “online learning” OR “higher education”) AND (retention OR intervention OR cost). The search was complemented by backward and forward snowballing of the reference lists of recent reviews, and the bibliographic data of every retained publication were verified against Crossref records and publisher pages.

Publications were included if they (a) were peer-reviewed journal articles, full conference papers or dataset descriptors published between 2013 and 2026; (b) addressed the prediction of dropout, withdrawal or failure in higher education, the evaluation of interventions with students at risk, the fairness and ethics of predictive models, or the determinants and economics of dropout; and (c) were written in English. Studies from campus-based universities were included when they addressed general methodological questions such as the choice of data, the timing of prediction or the comparison of models. Studies of school education, reports without peer review and conference posters without full text were excluded.

The final corpus comprised 31 publications in five groups: predictive modelling studies and data resources ($n = 12$) [6, 8, 10, 12, 13, 16–18, 23, 25, 29, 36]; studies of implementation and interventions ($n = 7$) [11, 14, 19–22, 27]; systematic reviews and institutional surveys ($n = 6$) [5, 24, 28, 30, 34, 35]; studies of fairness and ethics ($n = 3$) [7, 15, 32]; and studies of the determinants and economics of dropout ($n = 3$) [4, 9, 31]. For predictive studies, the context, data sources, algorithms, performance metrics and timing of prediction were extracted; for intervention studies, the design, sample, intervention and effect on retention or performance. Performance metrics were recorded as published; because the studies used different metrics (accuracy, AUC, precision, recall and true positive rate), they were compared descriptively and not pooled.

The cost–benefit model considered a cohort of N students with a baseline annual dropout rate d . A predictive model with recall R and precision P flags $N \cdot d \cdot R/P$ students, of whom $TP = N \cdot d \cdot R$ are correctly identified at-risk students. Each flagged student receives an intervention at cost c , and the early warning system incurs a fixed annual cost K for data infrastructure, model maintenance and analysis. The intervention prevents dropout for a share e of the correctly identified at-risk students (intervention effectiveness), and each retained student preserves tuition revenue for L years at an annual fee F . The net annual benefit was therefore calculated as $NB = e \cdot TP \cdot L \cdot F - (c \cdot TP/P + K)$, and the break-even effectiveness, at which $NB = 0$, as $e^* = (c \cdot TP/P + K) / (TP \cdot L \cdot F)$. All monetary values were expressed in units of the annual fee F , so that the results do not depend on currency or on the fee level of a particular university. The parameter values and their sources are given in Table 1.

Table 1.

Parameters of the cost–benefit model

Parameter	Value	Justification and source
N	1,000 students	Normalised cohort size
d	0.30 (sensitivity: 0.20–0.40)	Dropout of 24.2% and 30.1% per year in a distance course [29]; graduation deficit of distance institutions [31]
R, P	Early-term 0.60/0.60; mid-term 0.75/0.75; late-term 0.90/0.90	Weekly precision rising from 0.59 to 0.93 and recall from 0.61 to 0.76 during a course [6]; true positive rate of 58.8–93.6% [8]
e	0–25%; anchors 0% and 12.4%	Null effects in controlled evaluations [14, 27]; retention gain of 5.2 percentage points against a 42% baseline loss in a randomised trial of coaching ($5.2/42 = 12.4\%$) [11]
c	0.05F per flagged student (sensitivity: 0.05–0.20F)	Assumption: several hours of tutor or adviser contact per student
K	10F per year	Assumption: data infrastructure, model maintenance and analyst time
L	1 year; 2.5 years	Conservative case: only the next year's fee is preserved; alternative: average remaining study period

Three model-quality scenarios were defined to reflect the increase in prediction accuracy during a course reported in the literature. For the mid-term model, a sensitivity analysis varied the baseline dropout rate ($d = 0.2; 0.3; 0.4$) and the cost of an intervention ($c = 0.05F; 0.10F; 0.20F$). The calculations were performed in Python, and the model equations were applied to each combination of parameters.

ANALYSIS AND RESULTS

The 31 publications in the corpus were published between 2013 and 2026; 24 of them (77%) appeared in 2019 or later. Of the twelve predictive modelling studies and data resources, four concerned distance or fully online universities [6, 8, 23, 29] and one described an open dataset of a distance university [25]. Table 2 summarises the predictive studies that reported quantitative performance or a systematic comparison of models.

Table 2.

Predictive models of student risk: data, methods and reported performance

Study	Context and data	Methods	Reported performance
Aljohani et al., 2019 [6]	Distance university (OULAD), 22,437 students; weekly VLE clickstream	LSTM vs logistic regression and ANN	Precision 0.59 → 0.93, recall 0.61 → 0.76, accuracy 0.70 → 0.95 from week 1 to the final week
Waheed et al., 2020 [36]	VLE clickstream data with handcrafted features	Deep ANN vs logistic regression and SVM	Accuracy 84–93% (logistic regression 80–86%, SVM 80–89%)
Queiroga et al., 2020 [29]	Distance technical course, Brazil; Moodle interaction counts only	Five classifiers tuned by a genetic algorithm	Mean AUC 0.85; AUC above 0.7 from week 4
Bañeres et al., 2020 [8]	Fully online university, Spain; assessment grades	Gradual at-risk model with teacher dashboards	True positive rate 58.8–93.6%; predictions from the first assessment activity
Bird et al., 2021 [12]	Colleges, USA; administrative data	Simple models vs complex machine learning	Similar performance; risk rankings of students varied across methods
Goren et al.,	Campus university, Israel;	XGBoost, neural	AUC up to 0.94; pre-entry data

2024 [18]	8,267 students; pre-entry, LMS and grade data	network	only: AUC 0.61–0.63
Glandorf et al., 2024 [17]	Campus university, USA; 33,133 students; administrative data	Random forest and other models	AUROC 0.83 after year 1 and 0.91 after year 2; modest differences between student groups
Bettahi et al., 2025 [10]	University, Morocco; 17,883 records; external replication	Nine algorithms; SHAP explanations	AUC 0.993 and F1 0.911 (institutional data); AUC 0.922 (replication)

Data and models (RQ1). Three types of data were used: pre-entry and demographic data, learning platform activity, and assessment results. Models based on platform activity and assessments reached AUC values mostly between 0.8 and 0.95 [17, 18, 29] and accuracy of up to 93–95% [6, 36], whereas pre-entry data alone yielded an AUC of only 0.61–0.63 [18]. In two studies, deep neural networks outperformed logistic regression and support vector machines [6, 36]; in a systematic comparison, simple models performed about as well as complex ones, but the ranking of students by risk varied across methods [12]. Accuracy increased with time: precision rose from 0.59 in the first week to 0.93 in the final week [6], the AUROC rose from 0.83 after the first year to 0.91 after the second [17], and the assessment-based model updated its predictions after each graded activity [8]. The importance of predictors also shifted over time: pre-entry variables lost their predictive value quickly, grade point average became the strongest predictor during the first year and the number of enrolled terms dominated later [17]. Course achievements were the strongest group of predictors, whereas adding platform activity data to achievement data had little effect on overall metrics [18]. In the study with external replication, performance on an independent public dataset was lower than on institutional data (F1 0.792 vs 0.911), and SHAP explanations were used to identify risk and protective factors and to flag ambiguous cases for human review [10]. In the distance course studied by Queiroga et al. [29], an AUC above 0.7 was reached in week 4, and most first-year dropouts occurred in the first 20 weeks. The predictive value of platform data differed between courses [13, 16]. Incorrect predictions were frequently associated with unforeseen events in students' lives [23]. Differences in predictability between demographic groups were modest in one large study [17].

Effects of interventions (RQ2). The evidence on the effects of prediction-driven interventions is summarised in Table 3.

Table 3.

Evidence on the effects of prediction-driven and related interventions

Study	Setting and design	Intervention	Effect
Herodotou et al., 2019 [20]	Distance university; 15 courses, over 14,000 students, 559 teachers (189 with access); comparison with the previous presentation	Weekly risk predictions for teachers (OU Analyse)	Students of teachers with average use performed significantly better
Herodotou et al., 2019 [19]	Distance university; large-scale implementation	Same system	Greater teacher use associated with better completion and pass rates
Dawson et al., 2017 [14]	University retention programme, 2012–2014; 11,160 students, 1,868 flagged	Early learning and remediation support	Positive association in conventional analysis; no effect on retention in mixed-effects models
Plak et al., 2022 [27]	Dutch university; randomised field experiment with first-year students	Risk information for student counsellors	No reduction in first-year dropout; no improvement in performance
Bettinger & Baker, 2014 [11]	Eight US institutions; randomised trial; 13,555 students	Individual coaching (not prediction-based)	Retention +5.2 percentage points after six months (63% vs 58%); more cost-effective than increased financial aid

At the Open University, students of teachers who used OU Analyse with average intensity – accessing it in roughly 10% to 40% of course weeks and intervening with flagged students – performed significantly better than students of the same teachers in the previous presentation of the course, after controlling for prior academic performance [20], and greater teacher use was associated with better completion and pass rates [19]. In contrast, a three-year retention programme showed no effect on retention once mixed-effects models were applied [14], and in a randomised field experiment, counselling supported by an early warning system did not reduce first-year dropout [27]. A randomised trial of individual coaching without prediction increased retention by 5.2 percentage points [11]. Systematic reviews found that only 11 of 689 identified publications evaluated the effectiveness of learning analytics interventions [34] and that rigorous large-scale evidence of effectiveness is still lacking [24].

Economic results (RQ3). The results of the cost–benefit model are presented in Table 4 and Figure 1. With the parameters used, each scenario flagged 300 students, and the annual cost of the system was 25F. At an intervention effectiveness of zero, the net annual benefit was –25F in all scenarios. The break-even effectiveness ranged from 13.9% (early-term model, $L = 1$) to 3.7% (late-term model, $L = 2.5$). At the anchor value of 12.4%, the net benefit was negative only for the early-term model with a one-year value horizon (–2.7F) and positive in the other five scenarios, ranging from 2.9F to 58.7F. The corresponding return on investment ranged from –11% to 235%.

Table 4.

Results of the cost–benefit model by scenario ($N = 1,000$; $d = 0.30$; $c = 0.05F$; $K = 10F$)

Scenario (R/P)	TP	Flagged (false positives)	Annual cost	Break-even e^* ($L = 1 / 2.5$)	Net benefit at $e = 12.4\%$ ($L = 1 / 2.5$)	ROI at $e = 12.4\%$ ($L = 1 / 2.5$)
Early-term (0.60/0.60)	180	300 (120)	25.0F	13.9% / 5.6%	–2.7F / 30.8F	–11% / 123%
Mid-term (0.75/0.75)	225	300 (75)	25.0F	11.1% / 4.4%	2.9F / 44.7F	12% / 179%
Late-term (0.90/0.90)	270	300 (30)	25.0F	9.3% / 3.7%	8.5F / 58.7F	34% / 235%

At the anchor effectiveness of 12.4%, the system retained 22.3 students per 1,000 in the early-term scenario, 27.9 in the mid-term scenario and 33.5 in the late-term scenario. The resulting cost per additional retained student was 1.12F, 0.90F and 0.75F, respectively, compared with a preserved revenue of 1F per retained student with a one-year value horizon and 2.5F with a horizon of 2.5 years. Moving from the early-term to the late-term model reduced the break-even effectiveness from 13.9% to 9.3% with a one-year value horizon and from 5.6% to 3.7% with a horizon of 2.5 years.

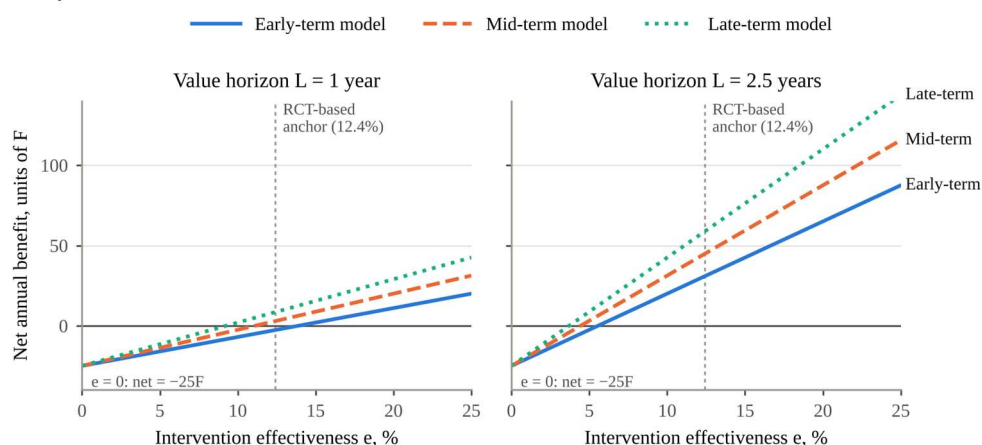


Figure 1. Net annual benefit of an early warning system as a function of intervention effectiveness for three model-quality scenarios

Extending the value horizon from one to 2.5 years reduced the break-even effectiveness by a factor of 2.5 in every scenario. The sensitivity analysis (Table 5) showed that increasing the cost of an intervention from 0.05F to 0.20F per flagged student raised the break-even effectiveness of the mid-term model from 11.1% to 31.1% ($L = 1$), whereas changing the baseline dropout rate from 0.20 to 0.40 changed it from 13.3% to 10.0%.

Table 5.

Sensitivity of the break-even effectiveness e^* to dropout rate and intervention cost (mid-term model; $L = 1 / L = 2.5$)

Baseline dropout rate d	$c = 0.05F$	$c = 0.10F$	$c = 0.20F$
0.20	13.3% / 5.3%	20.0% / 8.0%	33.3% / 13.3%
0.30	11.1% / 4.4%	17.8% / 7.1%	31.1% / 12.4%
0.40	10.0% / 4.0%	16.7% / 6.7%	30.0% / 12.0%

DISCUSSION

With regard to RQ1, the results confirm that dropout in digital distance education can be predicted with high accuracy once students have generated activity and assessment data, which agrees with earlier reviews [5, 30]. Two qualifications, however, are essential. First, very high reported values – such as an AUC of 0.993 [10] – should be interpreted with caution, because they may rely on predictors that become available late in a course or are closely related to the outcome itself. Second, accuracy and timeliness are in conflict: predictions become reliable precisely when there is less time left to act. Since most dropouts in the distance course examined by Queiroga et al. [29] occurred during the first 20 weeks, a model that is accurate only at the end of the course has little practical value. The variation of predictive value between courses [13, 16] and of risk rankings between methods [12] further implies that models must be calibrated locally and that a simple, transparent model is often a better starting point than a complex one.

With regard to RQ2, the evidence shows that a prediction is not an intervention. Positive results were obtained where predictions were actively used by teachers [19, 20], but these results come from comparisons with previous cohorts rather than from randomised designs. Controlled evaluations of prediction-driven support found no effect on retention [14, 27], whereas a randomised trial of individual coaching that did not rely on prediction did improve retention [11]. Taken together, these findings suggest that the quality of the intervention matters more than the quality of the prediction, and they agree with the conclusions of systematic reviews that the effectiveness of learning analytics interventions has not yet been rigorously demonstrated [24, 34]. The authors of the null-effect studies themselves point to the reasons: Plak et al. [27] suggest that addressing poor academic performance requires more actionable feedback and recommended counselling practices, and Dawson et al. [14] conclude that more data on individual differences are needed to design more effective interventions. In other words, knowing who is at risk does not tell a counsellor what to do. The analysis of prediction errors [23] adds that even an accurate model cannot see unexpected events in students' lives, so that human contact remains indispensable.

With regard to RQ3, the economic model makes these qualitative conclusions quantitative. A prediction without an effective intervention generates only costs (–25F per 1,000 students in all scenarios). Improving the model from early-term to late-term quality lowered the break-even effectiveness by only 4.6 percentage points with a one-year horizon, whereas a fourfold increase in the cost of an intervention almost tripled it. The economic result of an early warning system is therefore determined primarily by the effectiveness and cost of interventions and by how long a retained student continues to study, and only secondarily by marginal gains in accuracy. This supports the hypothesis of the study. It also echoes Simpson's argument [31] that proactive contact with distance students can be financially viable, while specifying the conditions under which this is the case. Because the model does not account for the shorter time available for intervention after late predictions, it probably overstates the advantage of the late-term scenario; in practice, the economic

optimum is likely to lie in mid-term predictions combined with low-cost, scalable interventions.

The cost per additional retained student provides a practical yardstick in the sense of cost-effectiveness analysis [26]. In the mid-term scenario it was slightly below one annual fee, which means that an early warning system pays for itself even under the conservative assumption that a retained student pays for only one more year, provided that the intervention reaches the effectiveness observed in the coaching trial [11]. The same yardstick allows prediction-based support to be compared with alternatives such as universal proactive contact with all students, recommended by Simpson [31]: when interventions are cheap, contacting everyone may be competitive with targeting, whereas expensive interventions make accurate targeting worthwhile. A promising design for distance education is therefore a two-tier approach in which low-cost, universal contact in the first weeks is followed by more intensive, individual support for students flagged by a mid-term model.

These relationships can be summarised as a value chain of predictive learning analytics (Figure 2). Value is created only if every link works – data, prediction, action, effect and economic value – and because the links multiply, a weak link cannot be compensated by strengthening another. An excellent model combined with a weak intervention produces the same result as no early warning system at all, but at a higher cost.

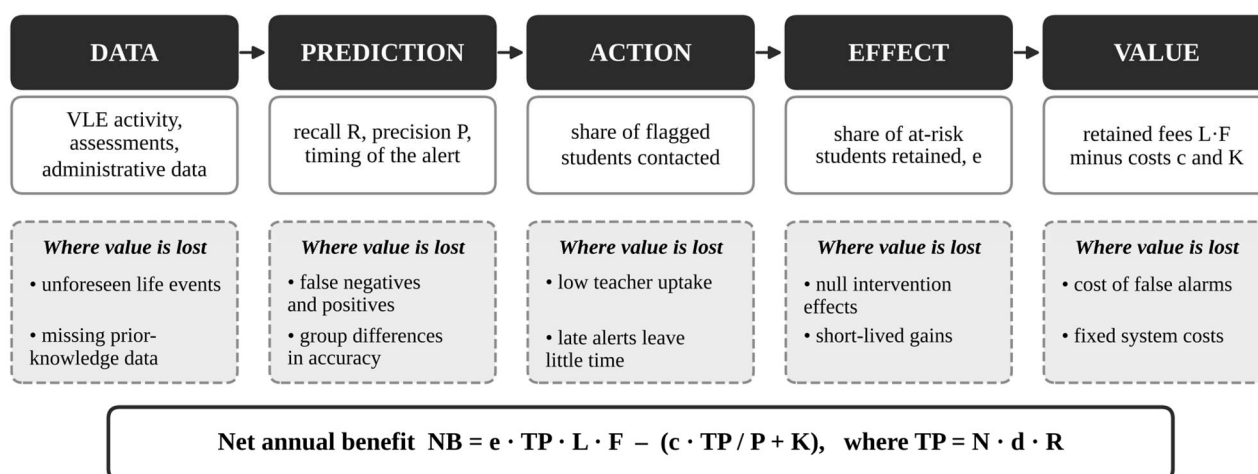


Figure 2. Value chain of predictive learning analytics for student retention

The value chain also shows where costs can be reduced. Most implementations deliver predictions to staff, and students are rarely treated as stakeholders [35]. Giving students themselves access to transparent information about their progress, combined with clear recommendations, could lower the cost of the first level of intervention and leave scarce staff time for students who need personal contact. Such an approach would also respond to the call to respect students' agency in the use of their data [32]. Its economic effect would appear in the model as a lower cost per flagged student c, which, as Table 5 shows, is one of the most influential parameters of the break-even condition.

Fairness and ethics are part of the same chain. False positives impose costs on the institution and may stigmatise students, and differences in model accuracy between groups mean that some students are more likely to be overlooked than others [7]. Regular slicing analysis of model performance [15], transparency about the use of student data and respect for students' agency [32] should therefore be built into any early warning system from the outset.

For distance education providers in Uzbekistan, where distance education is a legally recognised form of study [1] and is expected to expand [2], the findings lead to several recommendations. First, learning platform and assessment data should be integrated into a single student-level dataset before any investment in complex algorithms is made. Second, institutions should begin with simple, interpretable models calibrated on their own data and monitored for differences between student groups. Third, the larger part of the budget should be allocated to the design of interventions – tutor contact, coaching and timely feedback – and their effectiveness should

be tested in randomised pilots before scaling. Fourth, the break-even condition derived in this study can be used to decide in advance what minimum effectiveness an intervention must achieve at a given cost. Fifth, a common operational definition of dropout for distance students – for example, a period without platform activity combined with non-re-enrolment – would make results comparable between faculties and universities and would link early warning directly to revenue loss. Finally, teachers need training in interpreting and acting on predictions, in line with the national priority of developing human resource capacity in the application of AI [3].

The study has limitations. The review is integrative rather than exhaustive and includes only English-language sources. Performance metrics and definitions of dropout differ between studies, so the results could not be pooled. The economic model is static and covers one academic year; it assumes that all flagged students are contacted, considers only tuition revenue and not the wider social benefits of completion, and relies on assumed values of the intervention and fixed costs, which is why the results are presented as scenarios and break-even conditions rather than as forecasts. Finally, no data from Uzbek universities were available, so the parameters should be re-estimated on local data.

CONCLUSION

This study examined predictive learning analytics for student retention in digital distance education from both a predictive and an economic perspective. The synthesis of 31 publications shows that dropout can be predicted with high accuracy once platform activity and assessment data are available, that accuracy grows as a course progresses, and that pre-entry data alone are insufficient. Simple, transparent models perform about as well as complex ones, but predictions must be calibrated locally, because the value of platform data differs between courses. The evidence that prediction-driven interventions improve retention remains mixed: positive associations were found where teachers actively used predictions, whereas controlled evaluations of prediction-driven support found no effect.

The cost–benefit model demonstrates that an early warning system without an effective intervention is a pure cost and that its break-even point depends above all on the effectiveness and cost of interventions and on the value horizon of retained students. Under the scenarios examined, an intervention must prevent dropout for between 3.7% and 13.9% of correctly identified at-risk students to cover the costs of the system. With the anchor effectiveness taken from a randomised trial, the cost per additional retained student was close to or below one annual fee. The hypothesis that the economic return of predictive learning analytics depends more on interventions than on marginal improvements in model accuracy is therefore supported.

The contribution of the study lies in linking the predictive and economic perspectives through an explicit model with break-even conditions and in proposing a value chain that shows where the value of predictive analytics is created and lost. Its practical significance lies in recommendations that allow universities, including distance education providers in Uzbekistan, to plan investments in learning analytics on the basis of transparent economic criteria rather than model accuracy alone. For distance education, the most promising design is a two-tier system that combines low-cost contact with all students in the first weeks with more intensive support for students flagged by a locally calibrated mid-term model. Such a system should be transparent to students, regularly checked for differences in accuracy between groups of students and evaluated not by the accuracy of its predictions but by the number of students it actually retains and the cost of each retained student. For policy makers, the results imply that funding for learning analytics in distance education should be linked to evidence of retention effects and accompanied by investment in the people who act on predictions. Future research should test early warning systems in Uzbek universities using randomised designs, collect data on the actual costs of interventions, extend the model to multi-year and social returns, and examine the fairness of predictions for different groups of distance students.

REFERENCES

1. On Education: Law of the Republic of Uzbekistan No. ZRU-637 of 23.09.2020 [Electronic resource]. – URL: <https://lex.uz/docs/5013009>
2. On approval of the Concept for the development of the higher education system of the Republic of Uzbekistan until 2030: Decree of the President of the Republic of Uzbekistan No. UP-5847 of 08.10.2019 [Electronic resource]. – URL: <https://lex.uz/docs/4545887>
3. On approval of the Strategy for the development of artificial intelligence technologies until 2030: Resolution of the President of the Republic of Uzbekistan No. PP-358 of 14.10.2024 [Electronic resource]. – URL: <https://lex.uz/docs/7158606>
4. Aina C., Baici E., Casalone G. [et al.] The determinants of university dropout: A review of the socio-economic literature // *Socio-Economic Planning Sciences*. – 2022. – Vol. 79. – Art. 101102. – DOI: 10.1016/j.seps.2021.101102.
5. Albreiki B., Zaki N., Alashwal H. A systematic literature review of student' performance prediction using machine learning techniques // *Education Sciences*. – 2021. – Vol. 11, No. 9. – Art. 552. – DOI: 10.3390/educsci11090552.
6. Aljohani N.R., Fayoumi A., Hassan S.-U. Predicting at-risk students using clickstream data in the virtual learning environment // *Sustainability*. – 2019. – Vol. 11, No. 24. – Art. 7238. – DOI: 10.3390/su11247238.
7. Baker R.S., Hawn A. Algorithmic bias in education // *International Journal of Artificial Intelligence in Education*. – 2022. – Vol. 32, No. 4. – P. 1052–1092. – DOI: 10.1007/s40593-021-00285-9.
8. Bañeres D., Rodríguez M.E., Guerrero-Roldán A.E. [et al.] An early warning system to detect at-risk students in online higher education // *Applied Sciences*. – 2020. – Vol. 10, No. 13. – Art. 4427. – DOI: 10.3390/app10134427.
9. Behr A., Giese M., Tegum Kamdjou H.D. [et al.] Dropping out of university: a literature review // *Review of Education*. – 2020. – Vol. 8, No. 2. – P. 614–652. – DOI: 10.1002/rev3.3202.
10. Bettahi A., Belouadha F.-Z., Harroud H. A modular and explainable machine learning pipeline for student dropout prediction in higher education // *Algorithms*. – 2025. – Vol. 18, No. 10. – Art. 662. – DOI: 10.3390/a18100662.
11. Bettinger E.P., Baker R.B. The effects of student coaching: An evaluation of a randomized experiment in student advising // *Educational Evaluation and Policy Analysis*. – 2014. – Vol. 36, No. 1. – P. 3–19. – DOI: 10.3102/0162373713500523.
12. Bird K.A., Castleman B.L., Mabel Z. [et al.] Bringing transparency to predictive analytics: A systematic comparison of predictive modeling methods in higher education // *AERA Open*. – 2021. – Vol. 7. – DOI: 10.1177/23328584211037630.
13. Conijn R., Snijders C., Kleingeld A. [et al.] Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS // *IEEE Transactions on Learning Technologies*. – 2017. – Vol. 10, No. 1. – P. 17–29. – DOI: 10.1109/TLT.2016.2616312.
14. Dawson S., Jovanović J., Gašević D. [et al.] From prediction to impact: Evaluation of a learning analytics retention program // *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK'17)*. – New York: ACM, 2017. – P. 474–478. – DOI: 10.1145/3027385.3027405.
15. Gardner J., Brooks C., Baker R. Evaluating the fairness of predictive student models through slicing analysis // *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK19)*. – New York: ACM, 2019. – P. 225–234. – DOI: 10.1145/3303772.3303791.
16. Gašević D., Dawson S., Rogers T. [et al.] Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success // *The Internet and Higher Education*. – 2016. – Vol. 28. – P. 68–84. – DOI: 10.1016/j.iheduc.2015.10.002.
17. Glandorf D., Lee H.R., Orona G.A. [et al.] Temporal and between-group variability in college dropout prediction // *Proceedings of the 14th Learning Analytics and Knowledge Conference (LAK '24)*. – New York: ACM, 2024. – DOI: 10.1145/3636555.3636906.
18. Goren O., Cohen L., Rubinstein A. Early prediction of student dropout in higher education using

- machine learning models // Proceedings of the 17th International Conference on Educational Data Mining (EDM 2024). – Atlanta: International Educational Data Mining Society, 2024. – P. 349–359. – DOI: 10.5281/zenodo.12729834.
19. Herodotou C., Rienties B., Boroowa A. [et al.] A large-scale implementation of predictive learning analytics in higher education: the teachers' role and perspective // Educational Technology Research and Development. – 2019. – Vol. 67, No. 5. – P. 1273–1306. – DOI: 10.1007/s11423-019-09685-0.
20. Herodotou C., Hlosta M., Boroowa A. [et al.] Empowering online teachers through predictive learning analytics // British Journal of Educational Technology. – 2019. – Vol. 50, No. 6. – P. 3064–3079. – DOI: 10.1111/bjet.12853.
21. Herodotou C., Rienties B., Verdin B. [et al.] Predictive learning analytics “at scale”: Guidelines to successful implementation in higher education based on the case of the Open University UK // Journal of Learning Analytics. – 2019. – Vol. 6, No. 1. – P. 85–95. – DOI: 10.18608/jla.2019.61.5.
22. Herodotou C., Rienties B., Hlosta M. [et al.] The scalable implementation of predictive learning analytics at a distance learning university: Insights from a longitudinal case study // The Internet and Higher Education. – 2020. – Vol. 45. – Art. 100725. – DOI: 10.1016/j.iheduc.2020.100725.
23. Hlosta M., Herodotou C., Papathoma T. [et al.] Predictive learning analytics in online education: A deeper understanding through explaining algorithmic errors // Computers and Education: Artificial Intelligence. – 2022. – Vol. 3. – Art. 100108. – DOI: 10.1016/j.caeai.2022.100108.
24. Ifenthaler D., Yau J.Y.-K. Utilising learning analytics to support study success in higher education: a systematic review // Educational Technology Research and Development. – 2020. – Vol. 68, No. 4. – P. 1961–1990. – DOI: 10.1007/s11423-020-09788-z.
25. Kuzilek J., Hlosta M., Zdrahal Z. Open University Learning Analytics dataset // Scientific Data. – 2017. – Vol. 4. – Art. 170171. – DOI: 10.1038/sdata.2017.171.
26. Levin H.M., McEwan P.J., Belfield C. [et al.] Economic evaluation in education: Cost-effectiveness and benefit-cost analysis. – 3rd ed. – Los Angeles: SAGE, 2018. – ISBN 978-1-4833-8180-0.
27. Plak S., Cornelisz I., Meeter M. [et al.] Early warning systems for more effective student counselling in higher education: Evidence from a Dutch field experiment // Higher Education Quarterly. – 2022. – Vol. 76, No. 1. – P. 131–152. – DOI: 10.1111/hequ.12298.
28. Prenkaj B., Velardi P., Stilo G. [et al.] A survey of machine learning approaches for student dropout prediction in online courses // ACM Computing Surveys. – 2021. – Vol. 53, No. 3. – P. 1–34. – DOI: 10.1145/3388792.
29. Queiroga E.M., Lopes J.L., Kappel K. [et al.] A learning analytics approach to identify students at risk of dropout: A case study with a technical distance education course // Applied Sciences. – 2020. – Vol. 10, No. 11. – Art. 3998. – DOI: 10.3390/app10113998.
30. Sghir N., Adadi A., Lahmer M. Recent advances in predictive learning analytics: A decade systematic review (2012–2022) // Education and Information Technologies. – 2023. – Vol. 28, No. 7. – P. 8299–8333. – DOI: 10.1007/s10639-022-11536-0.
31. Simpson O. Student retention in distance education: are we failing our students? // Open Learning: The Journal of Open, Distance and e-Learning. – 2013. – Vol. 28, No. 2. – P. 105–119. – DOI: 10.1080/02680513.2013.847363.
32. Slade S., Prinsloo P. Learning analytics: Ethical issues and dilemmas // American Behavioral Scientist. – 2013. – Vol. 57, No. 10. – P. 1510–1529. – DOI: 10.1177/0002764213479366.
33. Snyder H. Literature review as a research methodology: An overview and guidelines // Journal of Business Research. – 2019. – Vol. 104. – P. 333–339. – DOI: 10.1016/j.jbusres.2019.07.039.
34. Sønderlund A.L., Hughes E., Smith J. The efficacy of learning analytics interventions in higher education: A systematic review // British Journal of Educational Technology. – 2019. – Vol. 50, No. 5. – P. 2594–2618. – DOI: 10.1111/bjet.12720.
35. Tsai Y.-S., Rates D., Moreno-Marcos P.M. [et al.] Learning analytics in European higher education – Trends and barriers // Computers & Education. – 2020. – Vol. 155. – Art. 103933. – DOI:

10.1016/j.compedu.2020.103933.

36. Waheed H., Hassan S.-U., Aljohani N.R. [et al.] Predicting academic performance of students from VLE big data using deep learning models // Computers in Human Behavior. – 2020. – Vol. 104. – Art. 106189. – DOI: 10.1016/j.chb.2019.106189.