

Predicting Student Performance on Digital Learning Platforms Using Artificial Intelligence

Mukhiddinov Navruzбек Muyassarovich ¹

¹ Independent Researcher, Uzbekistan
navruzbekislom@gmail.com

ARTICLE INFO

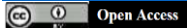
Keywords:
 Artificial intelligence,
 learning analytics, educational
 data mining, student
 performance prediction, early
 warning system, digital
 learning platforms.

Vol.01, No.05, 2026

<https://doi.org/10.5281/zenodo.23334833>

Copyright © 2026 by author(s) and Scientific
 Research Publishing Inc. This work is licensed
 under the Creative Commons Attribution
 International License (CCBY4.0).

<http://creativecommons.org/licenses/by/4.0/>



ABSTRACT

Digital learning platforms record detailed traces of student activity. This study evaluates artificial intelligence methods that use these traces to identify students at risk of failing or withdrawing early enough for support to help.

The Open University Learning Analytics Dataset, with demographic records, assessment results and 10.6 million daily click records for 32,593 course registrations, was used. At-risk status, defined as a final result of Fail or Withdrawn, was predicted at weeks 0 to 12 of each course from demographic, engagement and assessment features. Logistic regression, random forest, multilayer perceptron, XGBoost and LightGBM were trained on the 2013 course presentations and tested on the 2014 presentations.

With demographic data alone, the area under the ROC curve (AUC) was 0.68. It increased to 0.76 at week 4, 0.78 at week 6 and 0.82 at week 12. At week 6, XGBoost identified 63.3% of at-risk students with a precision of 67.9%. Differences between models were small, and logistic regression remained competitive. The mean assessment score, missing assessments and recent platform activity were the strongest predictors.

The study shows week by week when a useful warning becomes possible, using a protocol that trains on past cohorts and tests on a new one. A group analysis revealed lower accuracy and more false alarms for female students, which calls for fairness audits before such systems are deployed.

Introduction

Universities increasingly deliver courses through digital learning platforms, from fully online programs to blended courses that combine classroom teaching with a virtual learning environment (VLE). These platforms record every login, page view, forum post and quiz attempt. The same records offer a chance to see, early in a course, which students are drifting away. Dropout and failure remain expensive for students, institutions and the state, and they are especially high in distance education. Artificial intelligence (AI) methods can turn platform logs into early warnings, so that tutors can contact students while support can still change the outcome [1].

Many studies, however, report accuracy at the end of a course, when most of the useful information is already available and intervention is too late. Others mix students from the same course run in the training and test sets, which overstates performance on a new cohort [2], [3]. Fewer studies show how accuracy grows week by week and which signals drive predictions at each stage, which is what a tutor needs to plan support.

This paper makes four contributions. First, it designs a feature set that summarizes demographic data, VLE engagement and assessment behavior at any chosen week of a course. Second, it compares five machine learning models for at-risk prediction under a temporal protocol that trains on the 2013 course presentations and tests on the 2014 presentations. Third, it measures how prediction quality changes from the start of the course to week 12 and identifies the earliest week at which a useful warning is possible. Fourth, it explains the predictions with SHAP values and checks whether the error rates differ across gender and disability groups.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the data, the features, the models and the evaluation protocol. Section 4 presents the results. Section 5 discusses their practical meaning and limitations, and Section 6 concludes the paper.

2. Literature Review

2.1 Learning analytics and early warning systems

Research on this problem developed under the names of educational data mining and learning analytics [4], [5]. One of the first widely known deployments was Course Signals at Purdue University, which combined grades, demographic data and LMS activity into a traffic-light warning shown to students and instructors [6]. Cortez and Silva predicted the grades of secondary school students from demographic, social and school-related attributes with decision trees, random forests, neural networks and support vector machines [7]. Conijn et al. analyzed 17 blended courses that used the Moodle LMS and found that LMS data alone predicted course outcomes only moderately and that the useful predictors differed between courses [8]. Systematic reviews by Hellas et al. [2] and Namoun and Alshantqi [3] confirm that prior performance, assessment results and online engagement are the most frequently used predictors, while the reported accuracy varies widely because studies use different data, targets and validation designs.

2.2 Studies on OULAD and online learning

The Open University Learning Analytics Dataset (OULAD) [9] made it possible to compare methods on the same public data. It links demographic records, assessment results and daily click counts in the VLE for tens of thousands of students. Waheed et al. used OULAD to predict at-risk students with deep artificial neural networks and reported better results than logistic regression and support vector machines [10]. Hlosta et al. proposed training models on the current course presentation rather than on legacy data in order to identify at-risk students early [11]. Work on massive open online courses has examined learner engagement patterns [12] and applied deep learning to dropout prediction [13], with sequence models such as long short-term memory networks [14] and deep knowledge tracing [15] modeling learning over time.

2.3 Fairness of predictive models in education

Predictive models can treat groups of students differently even when protected attributes are not used directly. Riazzy et al. showed that at-risk prediction on OULAD can produce different error rates for demographic groups [16], and Baker and Hawn reviewed the sources and consequences of algorithmic bias in education [17]. These findings mean that an early warning system must be checked for fairness as well as for accuracy.

3. Materials and Methods

3.1 Dataset

The study used OULAD [9], which is publicly available under an open license. The Open University in the United Kingdom delivers its courses mainly through a VLE, and the dataset covers seven modules, labeled AAA to GGG, presented 22 times in 2013 and 2014. It contains 32,593 registrations of 28,785 students. For each registration, it records demographic attributes, registration and unregistration dates, the submission dates and scores of assessments, and daily counts of clicks on VLE resources. The click table has 10,655,280 rows covering 20 types of learning activity, such as course content pages, quizzes, forums and external resources. The final result of each registration is Pass (12,361), Distinction (3,024), Fail (7,052) or Withdrawn (10,156).

3.2 Prediction task

A registration was labeled at risk when the final result was Fail or Withdrawn, and not at risk when it was Pass or Distinction. At-risk students form 52.8% of all registrations. Predictions were made at the end of weeks 0, 2, 4, 6, 8, 10 and 12 of each course, where week 0 is the official start date. At each cutoff, the model could use only information recorded before that day. Students who had already unregistered before the cutoff were excluded, because their outcome is already known to the institution and needs no prediction.

3.3 Features

For each cutoff, three groups of features were computed (Table 1). Categorical attributes were one-hot encoded, and missing values were replaced by zero, or by -1 for assessment scores of students who had not submitted any work. At week 0, only the demographic group and the clicks made before

the course start were available, which gave 40 features. From week 2 onward, the feature set had 63 variables.

Table 1. Feature groups computed at each weekly cutoff

Group	Features
Demographic and enrollment	Gender, region, highest previous education, deprivation (IMD) band, age band, disability, number of previous attempts, studied credits, registration date
VLE engagement	Total clicks, number of active days, number of distinct resources, clicks in the last 7 days, clicks in the previous 7 days, change between these two weeks, days since the last activity, clicks before the course start, clicks on each of the 10 most used activity types
Assessment behavior	Number of assessments due, submitted, missing and submitted late, mean score and lowest score of submitted assessments

3.4 Evaluation protocol

In practice, a model is trained on past cohorts and applied to a new one. We therefore trained all models on the presentations that started in 2013 (2013B and 2013J) and tested them on those that started in 2014 (2014B and 2014J). No registration of the test cohort was used for training or tuning.

3.5 Models

Five models were compared. Logistic regression (LR) used standardized inputs and L2 regularization with $C = 1$. The random forest (RF) [18] had 200 trees with at least two samples per leaf. The multilayer perceptron (MLP) had two hidden layers of 64 and 32 ReLU units, an L2 penalty of 0.001 and early stopping. XGBoost [19] used 300 trees of depth 5 with a learning rate of 0.05, and LightGBM [20] used 300 trees with 31 leaves and a learning rate of 0.05, both sampling 80% of rows and features per tree. Hyperparameters were fixed in advance, and a probability threshold of 0.5 was used for class decisions. The models were implemented with scikit-learn [21], XGBoost and LightGBM on a two-core 2.8 GHz CPU.

3.6 Metrics and explanation

The area under the ROC curve (AUC) measures how well a model ranks students by risk regardless of the threshold. At the 0.5 threshold, we also report the F1 score, recall and precision for the at-risk class, and the balanced accuracy, which is the mean of the recall of both classes. Recall is the most important of these for an early warning system, since it gives the share of struggling students who receive a warning. To explain the predictions, SHAP values [22] were computed for the LightGBM model at week 6 on 3,000 randomly selected test registrations. To check fairness, recall, false positive rate and AUC were compared across gender and disability groups at the same week.

4. Results

4.1 Data summary

The 2013 training cohort contained 12,555 registrations that were still active at week 0, and the 2014 test cohort contained 17,360. Because students who withdrew were removed at each later cutoff, the test set shrank to 14,595 registrations by week 12, and the share of at-risk students in it fell from 49.2% to 39.6%.

4.2 Prediction quality over time

Table 2 and Figure 1 show the AUC of the five models at each cutoff. With demographic data alone, at week 0, the models reached an AUC of only 0.67 to 0.69. Two weeks of platform activity added little, but from week 4 the quality rose steadily, as the first assessments were submitted and engagement patterns became clearer. By week 8 the best models reached an AUC of about 0.80, and after week 10 the curve flattened at 0.80 to 0.82. The differences between models were small. At week 6, the tree ensembles and the MLP were 0.02 to 0.04 better than logistic regression, but the gap was smaller at the other cutoffs, and all models converged by week 10.

Table 2. AUC of at-risk prediction on the 2014 test cohort by week and model

Week	LR	RF	MLP	XGBoost	LightGBM
0	0.676	0.675	0.672	0.686	0.683
2	0.707	0.701	0.696	0.699	0.695
4	0.745	0.755	0.746	0.760	0.754
6	0.749	0.777	0.772	0.784	0.780
8	0.777	0.796	0.783	0.797	0.794
10	0.811	0.815	0.802	0.813	0.811
12	0.810	0.816	0.810	0.815	0.814

Note. The best value in each row is in bold.

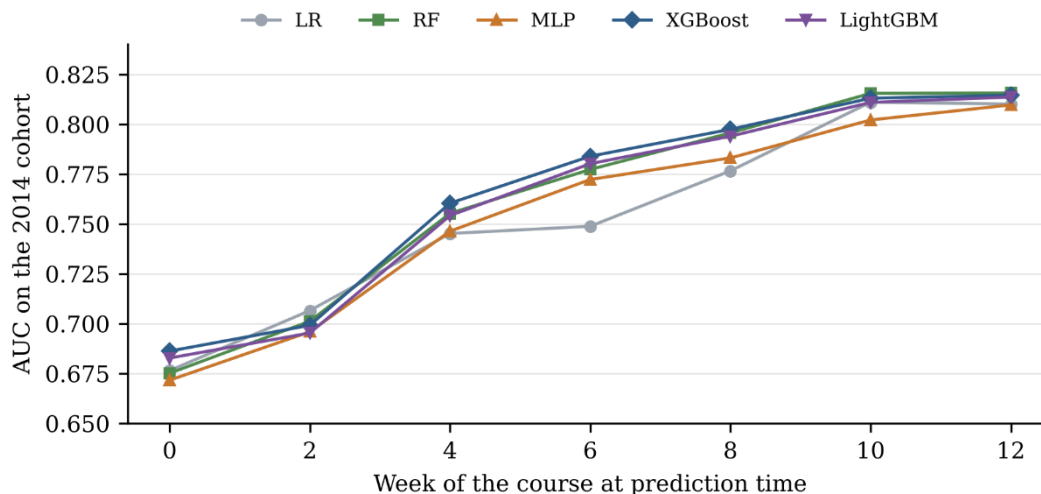


Figure 1. AUC of at-risk prediction as a function of the week at which the prediction is made, for five models trained on the 2013 cohort and tested on the 2014 cohort.

4.3 Threshold-based metrics

Table 3 gives the threshold-based metrics of XGBoost, which was the best or close to the best model at most cutoffs. At week 6, the model identified 63.3% of the students who later failed or withdrew, and 67.9% of the students it flagged were indeed at risk. At week 12, recall rose to 68.9% and balanced accuracy to 73.4%.

Table 3. Threshold-based metrics of XGBoost at a probability threshold of 0.5

Week	Test registrations	At-risk share (%)	Recall (%)	Precision (%)	F1	Balanced accuracy (%)
0	17,360	49.2	58.3	64.8	0.614	63.8
2	16,032	45.0	51.7	64.0	0.572	63.9
4	15,735	44.0	63.0	66.1	0.645	68.8
6	15,408	42.8	63.3	67.9	0.655	70.5
8	15,098	41.6	65.0	69.0	0.670	72.1
10	14,817	40.5	65.1	70.1	0.675	73.1
12	14,595	39.6	68.9	67.2	0.680	73.4

4.4 Most influential predictors

Figure 2 shows the twelve features with the largest mean absolute SHAP value for LightGBM at week 6. The mean score of submitted assessments was by far the strongest predictor, followed by the number of missing assessments. Engagement in the most recent week, the number of days since the last activity and the number of active days came next. A previous education level below A level was the only demographic feature among the top five.

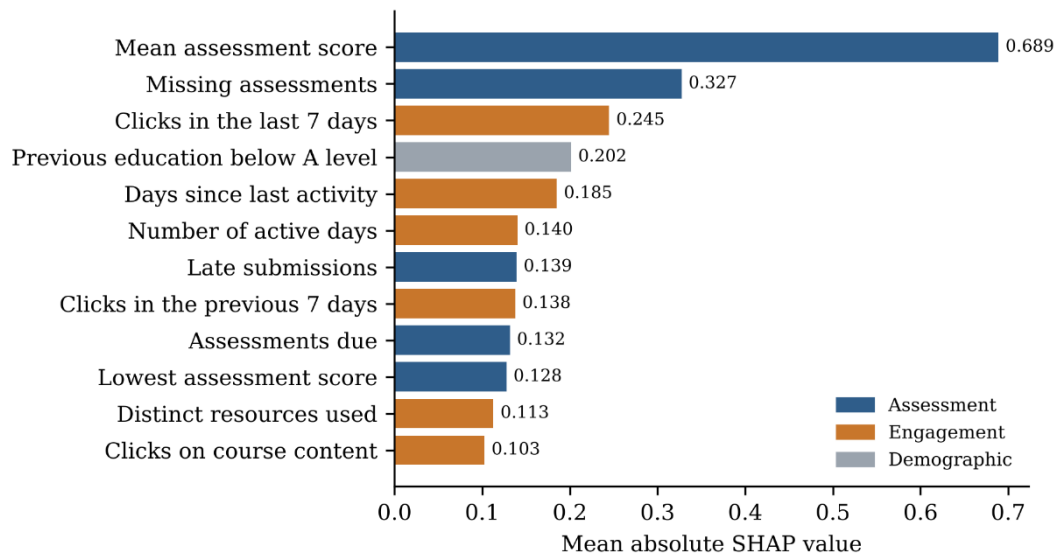


Figure 2. Twelve most influential features of the LightGBM model at week 6 (mean absolute SHAP value), colored by feature group.

4.5 Differences between groups

Table 4 compares the error rates of LightGBM at week 6 across groups. The model ranked male students better than female students, with an AUC of 0.810 against 0.736, and it found a larger share of at-risk male students (67.7% against 60.3%) while raising more false alarms for female students (27.5% against 20.9%). For students with and without a declared disability, the AUC was similar (0.774 and 0.779), but disabled students received more warnings, both correct and incorrect, which partly reflects their higher at-risk share of 52.7%.

Table 4. Error rates of LightGBM at week 6 by group

Group	Test registrations	At-risk share (%)	Recall (%)	False positive rate (%)	AUC
Male	8,829	44.5	67.7	20.9	0.810
Female	6,579	40.5	60.3	27.5	0.736
No disability	13,960	41.8	63.8	23.3	0.779
Disability	1,448	52.7	71.7	30.8	0.774

5. Discussion

5.1 How early can a warning be given

The results show that useful warnings become possible about one month into a course. At week 4 the best model reached an AUC of 0.76, and at week 6 it found almost two thirds of the students who later failed or withdrew while keeping precision near 68%. Waiting beyond week 10 brought little extra accuracy but left less time for support. For an institution, weeks 4 to 6 are therefore a reasonable point to start contacting students, with the warning list updated every week as new data arrive.

Demographic data alone gave an AUC of only 0.68. This is important both technically and ethically. A model that relies on background characteristics would label students by who they are rather than by what they do, and it would still miss many of them. The predictive power came mainly from behavior on the platform and from assessment results, which students can change and tutors can discuss with them.

5.2 What the predictors tell tutors

The SHAP analysis agrees with earlier studies that assessment results are the strongest signal of course success [2], [3], [8]. The number of missing assessments ranked second, and it is a simple and actionable indicator, since a student who skips an assignment can be contacted on the day after the deadline. Recent activity and the number of days since the last login also ranked high. This suggests that a sudden drop in engagement is more informative than a low total level of activity. Such findings can be

translated into transparent rules that tutors understand, for example a contact after a missed assignment or after a week without activity.

5.3 Choice of model

The ensemble models were only slightly better than logistic regression, and only during the middle weeks. Where staff need to understand and trust the warnings, a simple model with clear coefficients may be preferred at a small cost in accuracy. Some earlier studies on OULAD report higher accuracy [10], but direct comparison is difficult because studies differ in their targets and validation designs, and random splits that mix students of the same course run in training and testing tend to give optimistic results. The temporal protocol used here is stricter and better reflects how a model trained on past cohorts would perform on a new one.

5.4 Fairness

The group analysis showed that the model ranked male students better than female students and produced more false alarms for female students. This difference may partly arise because men and women are distributed differently across modules with different difficulty, but it should not be ignored. As shown in [16] and discussed by Baker and Hawn [17], an early warning system can treat groups unequally even without using protected attributes directly. Before deployment, error rates should be audited per group and per course, thresholds may need to be calibrated separately, and a warning should always lead to an offer of support, never to a penalty. A tutor should make the final decision.

5.5 Implications for universities in Uzbekistan

Universities in Uzbekistan increasingly use learning management systems for blended and distance courses. The proposed feature set uses only data that such systems already record, namely logins, resource views, submissions and grades, so the approach can be transferred without new data collection. The model must, however, be retrained on local data, since course design, assessment schedules and student populations differ from those of the Open University.

5.6 Limitations and future research

The study has several limitations. OULAD comes from a single distance-learning university in the United Kingdom and from courses taught in 2013 and 2014. The target merges failure and withdrawal, which may call for different interventions. Hyperparameters were fixed, a single threshold of 0.5 was used, and fairness was examined for two attributes only. Finally, the study measures prediction quality, not the effect of interventions on student outcomes, which would require a controlled field study. Future work will apply the approach to learning management system data from universities in Uzbekistan, separate the prediction of failure from that of withdrawal, add sequence models of weekly activity and test the effect of tutor interventions triggered by the warnings.

6. Conclusion

This paper evaluated AI models for the early prediction of at-risk students on a digital learning platform, using the OULAD data and a protocol that trains on one cohort and tests on the next. Prediction quality grew from an AUC of 0.68 with demographic data alone to 0.78 at week 6 and 0.82 at week 12. At week 6, XGBoost identified almost two thirds of the students who later failed or withdrew. Assessment results, missing submissions and recent platform activity were the most influential predictors, and simple logistic regression was nearly as accurate as the ensemble models.

For practice, the results suggest that universities can start contacting students at risk four to six weeks into a course, using indicators that tutors understand and students can act on. The lower accuracy observed for female students shows that fairness must be checked before such a system is used, and that the final decision should always remain with a tutor.

References

1. O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education: Where are the educators?" *Int. J. Educ. Technol. High. Educ.*, vol. 16, Art. no. 39, 2019.
2. A. Hellas et al., "Predicting academic performance: A systematic literature review," in *Proc. 23rd Annu. ACM Conf. Innov. Technol. Comput. Sci. Educ. (ITiCSE Companion)*, 2018, pp. 175-199.

3. A. Namoun and A. Alshankiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Appl. Sci.*, vol. 11, no. 1, Art. no. 237, 2021.
4. G. Siemens and R. S. J. d. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in *Proc. 2nd Int. Conf. Learn. Analytics Knowl. (LAK)*, 2012, pp. 252-254.
5. C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining Knowl. Discov.*, vol. 10, no. 3, Art. no. e1355, 2020.
6. K. E. Arnold and M. D. Pistilli, "Course Signals at Purdue: Using learning analytics to increase student success," in *Proc. 2nd Int. Conf. Learn. Analytics Knowl. (LAK)*, 2012, pp. 267-270.
7. P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," in *Proc. 5th Future Bus. Technol. Conf. (FUBUTEC)*, Porto, Portugal, 2008, pp. 5-12.
8. R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, "Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS," *IEEE Trans. Learn. Technol.*, vol. 10, no. 1, pp. 17-29, Jan. 2017.
9. J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics dataset," *Sci. Data*, vol. 4, Art. no. 170171, 2017.
10. H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Comput. Human Behav.*, vol. 104, Art. no. 106189, 2020.
11. M. Hlosta, Z. Zdrahal, and J. Zendulka, "Ouroboros: Early identification of at-risk students without models based on legacy data," in *Proc. 7th Int. Learn. Analytics Knowl. Conf. (LAK)*, 2017, pp. 6-15.
12. R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses," in *Proc. 3rd Int. Conf. Learn. Analytics Knowl. (LAK)*, 2013, pp. 170-179.
13. W. Xing and D. Du, "Dropout prediction in MOOCs: Using deep learning for personalized intervention," *J. Educ. Comput. Res.*, vol. 57, no. 3, pp. 547-570, 2019.
14. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735-1780, 1997.
15. C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 505-513.
16. S. Riazzy, K. Simbeck, and V. Schreck, "Fairness in learning analytics: Student at-risk prediction in virtual learning environments," in *Proc. 12th Int. Conf. Comput. Supported Educ. (CSEDU)*, 2020, pp. 15-25.
17. R. S. Baker and A. Hawn, "Algorithmic bias in education," *Int. J. Artif. Intell. Educ.*, vol. 32, no. 4, pp. 1052-1092, 2022.
18. L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5-32, 2001.
19. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785-794.
20. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 3146-3154.
21. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825-2830, 2011.
22. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 4765-4774.