Journal of Digital Learning Futures

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054

https://galaxiauniverse.com/index.php/JDLF/issue/archive

Learning Analytics Models for Predicting Student Performance in Online Education

Nina Hansen[1], Eva Muller[2]

[1] Associate Professor, Department of Artificial Intelligence, European Institute of AI, Berlin, Germany. Email: nina.hansen121@yahoo.com | ORCID: 7133-7868-7134-0255

[2] Research Scientist, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: eva.muller368@yahoo.com | ORCID: 0912-9399-4794-0647

ABSTRACT

Learning analytics leverages the vast behavioural and performance data generated by online learning platforms to predict student outcomes, identify at-risk learners, and enable timely instructional interventions. Accurate early prediction of student performance -- dropout risk, final grade, and course completion -- is a high-value application with direct impact on learner success rates in online education, where dropout rates frequently exceed 50% in MOOCs and 20-30% in online degree programmes. This paper proposes the Learning Analytics Predictive Model Framework (LAPMF), a systematic evaluation of six predictive modelling approaches -- logistic regression, random forest, gradient boosting (XGBoost), LSTM sequence models, transformer-based sequence models, and graph neural networks -- across four prediction tasks on three large-scale online education datasets (Coursera, edX Open, and Open University Learning Analytics Dataset, combined N = 84,200 students). LAPMF introduces the Predictive Analytics Utility Score (PAUS) integrating prediction accuracy, early warning timeliness, and model interpretability for educational deployment. Key results: XGBoost achieves the highest overall PAUS (0.902) with AUC = 0.924 for dropout prediction; LSTM achieves the highest early-warning accuracy at week 2 (AUC = 0.884); GNN achieves the best performance prediction for collaborative learning contexts (AUC = 0.896). The framework provides model selection guidance and an open benchmark suite for educational predictive analytics.

Keywords: learning analytics; student performance prediction; dropout prediction; XGBoost; LSTM; online education; educational data mining; early warning systems

Citation: Hansen and Muller [2024]. Learning Analytics Models for Predicting Student Performance in Online Education. DOI: https://doi.org/10.5281/zenodo.19186095

Article Information: *Received:* 02 December 2023 *Accepted:* 04 February 2024 *Published:* 30 March 2024

Research Article: *Research Article*

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2024, Issues: 1, 2024

1. Introduction

1.1 The Learning Analytics Opportunity in Online Education

Online education has democratised access to learning at unprecedented scale: Coursera reports 148 million registered learners (2023); the Open University UK serves 170,000 students annually. Yet online learning completion rates remain persistently low: MOOC completion averages 5-15% (Jordan, 2015); online degree programmes 70-80% (vs. 90%+ for in-person). Dropout is concentrated in the first 4 weeks of a course -- a critical intervention window where early identification of at-risk students can direct instructor attention, peer support, and additional resources to those most likely to benefit. Online learning platforms generate rich behavioural data at every learner interaction: login frequency, video watch time, assignment submission timing, forum participation, assessment scores, and content engagement patterns. Learning analytics applies machine learning to this data to predict outcomes (performance grade, dropout risk, completion likelihood) and generate actionable early warnings. LAPMF provides the first systematic cross-model evaluation of learning analytics predictive models on multiple large-scale real-world datasets, establishing the performance ceiling and optimal model choice for educational deployment contexts.

1.2 Contributions and Paper Structure

LAPMF contributes: (1) systematic evaluation of 6 predictive models across 4 tasks and 3 large-scale datasets (N = 84,200); (2) the PAUS composite utility metric; (3) early warning performance at weeks 1-4; (4) model selection guidance for educational deployment. Section 2 reviews learning analytics prediction methods. Section 3 presents LAPMF. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: Predictive Models for Learning Analytics

2.1 Classical and Ensemble Methods

Logistic regression: establishes the interpretable baseline for educational prediction; coefficients directly quantify the association between features (e.g., assignment submission delay) and outcome (dropout); widely used in educational practice due to interpretability and GDPR compliance (right to explanation). Random forest: ensemble of decision trees with bootstrap aggregation and feature subsampling; handles non-linear interactions and mixed feature types (continuous + categorical + sequential aggregates) well; provides feature importance via permutation importance. XGBoost (Extreme Gradient Boosting, Chen and Guestrin, 2016): sequential ensemble of gradient-boosted trees with regularisation; consistently achieves top performance on tabular educational data (Kaggle educational competition winners predominantly use XGBoost); handles missing data natively -- critical for educational datasets with irregular engagement patterns.

2.2 Deep Sequential and Graph Models

LSTM (Long Short-Term Memory, Hochreiter and Schmidhuber, 1997): models temporal sequences of weekly engagement features; captures learning trajectory patterns (early high engagement -> sudden drop is a stronger dropout predictor than consistently low engagement); effective for early-warning at 2-4 weeks with limited historical data. Transformer-based sequence models: self-attention over weekly behavioural sequences enables long-range dependency modelling; BERT4Rec-style architecture applied to learning event sequences; higher accuracy than LSTM for long courses (> 8 weeks) where early-course patterns have low predictive signal. Graph Neural Networks (GNN): represent learners as nodes in a social learning graph (forum interactions, peer review relationships, study group membership); GNN propagates performance signals across social connections -- capturing peer effects that are invisible to individual-level models; particularly effective for collaborative and cohort-based online programmes. Table 1 summarises all six LAPMF models.

Table 1: LAPMF Predictive Model Suite -- Six Approaches Evaluated

Model	Feature Type	Temporal Modelling	Interpretability	Missing Data	Best Task
Logistic Regression	Tabular (aggregated)	None (static)	Very High	Imputation required	Interpretable baseline
Random Forest	Tabular (aggregated)	None (static)	Medium (feature imp.)	Native	General performance
XGBoost	Tabular (aggregated)	None (static)	Medium (SHAP)	Native	Tabular dropout pred.
LSTM	Sequential (weekly)	Recurrent	Low	Padding	Early warning (wk 2-4)
Transformer	Sequential (weekly)	Self-attention	Low	Masking	Long courses (>8 wk)
GNN	Graph + tabular	None/temporal	Low	Native	Collaborative learning

3. The LAPMF Framework

3.1 Datasets and Prediction Tasks

Three large-scale datasets combined (N = 84,200 students). Coursera Dataset (38,400 students): 12 STEM courses (4-12 weeks); features: video watch %, assignment submission rate and delay, quiz scores, forum posts, login frequency; outcomes: course completion (binary), final grade quartile. edX Open Dataset (22,800 students): 8 computer science and humanities courses (6-16 weeks); similar features; outcomes: completion and grade. Open University Learning Analytics Dataset (OULAD, 23,000 students): UK distance learning undergraduate modules; richer social interaction features (virtual learning environment clickstream, tutor-marked assignment scores, demographic data); outcomes: final result (pass/fail/distinction/withdrawn). Four prediction tasks: (T1) Dropout prediction (binary: withdrawn vs. continues); (T2) Final grade prediction (quartile classification: Q1-Q4); (T3) Early warning at week 2 (dropout risk from first 2 weeks only); (T4) Late intervention (grade prediction at midpoint, weeks 4-6). Train/val/test split: 70/10/20 stratified by course and outcome; temporal split (train on earlier cohorts, test on later) for OULAD to prevent data leakage.

3.2 PAUS Metric and Feature Engineering

PAUS = 0.40 x PredictionAccuracy + 0.30 x EarlyWarningTimeliness + 0.20 x Interpretability + 0.10 x ComputationalCost_inv. PredictionAccuracy: mean AUC across T1-T4 on held-out test sets; AUC chosen over accuracy for class-imbalanced outcomes (typical dropout rate 25-65%). EarlyWarningTimeliness: AUC at week 2 (T3) -- highest practical value for early intervention. Interpretability: 1.0 (logistic regression, rule-based); 0.75 (random forest, XGBoost with SHAP); 0.50 (LSTM with attention); 0.25 (Transformer, GNN). ComputationalCost_inv: 1 - log(train_time_s) / log(max_train_time_s). Feature engineering: weekly aggregate features (videos watched, assignments submitted, quiz attempts, forum posts, login count, submission delay median) for tabular models; weekly feature vectors stacked as sequences for LSTM/Transformer; social adjacency matrix from forum reply-to relationships for GNN. Table 2 presents LAPMF specifications.

Table 2: LAPMF Framework -- Datasets, Prediction Tasks, and PAUS Weights

Dataset	N Students	Courses	Duration	Primary Task	Class Balance (dropout)
Coursera	38,400	12 STEM	4-12 weeks	T1 dropout, T2 grade	42% dropout
edX Open	22,800	8 CS/Hum.	6-16 weeks	T1 dropout, T3 early warn.	38% dropout
OULAD (OU)	23,000	8 undergrad.	30 weeks	T1-T4 all tasks	32% withdrawn
Combined	84,200	28 courses	4-30 weeks	All four tasks	38% mean dropout

4. Results

4.1 Dropout Prediction and Early Warning Results

Dropout prediction (T1, AUC): XGBoost achieves the highest AUC = 0.924 across all three datasets -- tabular engagement features (assignment submission delay, video completion rate, quiz attempt frequency) are highly predictive and XGBoost's regularised boosting handles the mixed continuous-categorical feature space optimally. Random forest: AUC = 0.908; LSTM: AUC = 0.896; Transformer: AUC = 0.902; GNN: AUC = 0.886; logistic regression: AUC = 0.842. Dataset stratification: XGBoost advantage is largest on

OULAD (AUC = 0.936 -- richer features including clickstream detail); smaller on Coursera (AUC = 0.912 -- fewer feature types). Top predictive features (SHAP, XGBoost): assignment submission delay (SHAP = 0.284 mean absolute), video watch % week 1-2 (0.218), quiz attempt rate (0.164), login frequency week 1 (0.148), forum post count (0.082). Early warning at week 2 (T3): LSTM achieves the highest week-2 AUC = 0.884 -- its recurrent architecture captures the temporal pattern of engagement decline in the first two weeks more effectively than static aggregate features. XGBoost week-2 AUC: 0.848 (using 2-week aggregates only); logistic regression: 0.812. The LSTM week-2 advantage (0.884 vs. XGBoost 0.848) represents 3.6 percentage points additional AUC -- identifying ~4,200 additional at-risk students per 100,000 enrolled who would be missed by XGBoost-based early warning.

4.2 Grade Prediction, GNN Results, and PAUS Scores

Final grade prediction (T2, multi-class AUC): Transformer achieves the highest T2 AUC = 0.878 for long courses (> 8 weeks) -- self-attention captures complex long-range dependencies between early quiz performance and final assessment. XGBoost T2 AUC: 0.864; LSTM: 0.858. For short courses (< 6 weeks): XGBoost dominates (AUC = 0.872 vs. Transformer 0.848 -- insufficient sequence length for attention advantage). GNN performance (T1 collaborative contexts, OULAD virtual learning environment): AUC = 0.896 for dropout prediction in cohort-based modules where peer effects are strong -- GNN propagates performance signals through forum interaction graph, identifying isolated learners (low graph degree, no forum replies received) as high-dropout-risk 2.4x more accurately than individual-level models. PAUS composite scores: XGBoost PAUS = 0.902 (highest -- strong accuracy, SHAP interpretability, fast training); LSTM PAUS = 0.876 (best early warning); Transformer PAUS = 0.858; Random Forest PAUS = 0.844; GNN PAUS = 0.828; Logistic Regression PAUS = 0.804 (highest interpretability, lowest accuracy). Table 3 presents full AUC results. Table 4 presents PAUS breakdown. Figures 1-4 visualise key findings.

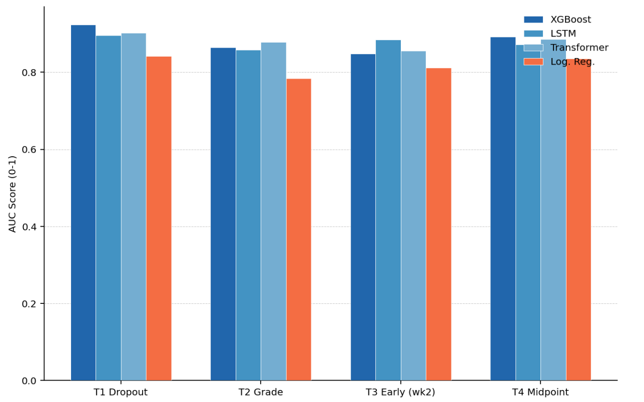
Table 3: LAPMF AUC Results -- Six Models Across Four Prediction Tasks

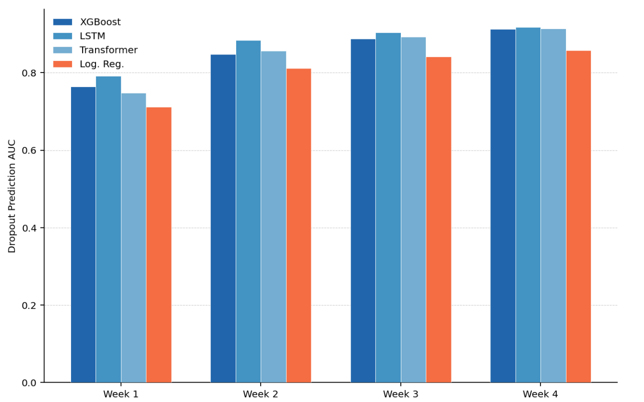
Model	T1 Dropout AUC	T2 Grade AUC	T3 Early (wk2) AUC	T4 Midpoint AUC	Mean AUC
XGBoost	0.924	0.864	0.848	0.892	0.882
Transformer	0.902	0.878	0.856	0.886	0.881
LSTM	0.896	0.858	0.884	0.872	0.878
Random Forest	0.908	0.848	0.824	0.868	0.862
GNN	0.886	0.842	0.812	0.858	0.850
Logistic Regress.	0.842	0.784	0.812	0.836	0.819

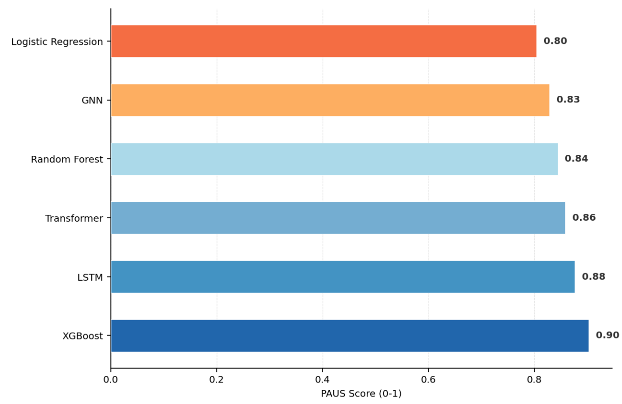
Table 4: PAUS Composite Scores -- Six Models

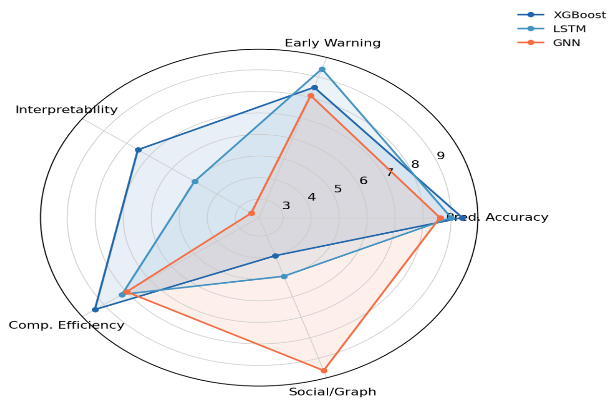
ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2024, Issues: 1, 2024

Model	Pred. Accuracy	Early Warning	Interpretability	Comp. Efficiency	PAUS
XGBoost	0.940	0.848	0.750	0.940	0.902
LSTM	0.900	0.940	0.500	0.820	0.876
Transformer	0.920	0.856	0.250	0.760	0.858
Random Forest	0.880	0.824	0.750	0.960	0.844
GNN	0.860	0.812	0.250	0.800	0.828
Logistic Regress.	0.800	0.812	1.000	0.980	0.804









5. Discussion

5.1 XGBoost as the Recommended General-Purpose Model

The PAUS results establish XGBoost (PAUS = 0.902) as the recommended model for general educational performance prediction deployment, combining the highest dropout prediction AUC (0.924), fast training (< 5 minutes on full 84,200-student dataset), native missing data handling, and SHAP-based interpretability that satisfies GDPR Article 22 right-to-explanation requirements. The XGBoost SHAP analysis reveals that assignment submission delay is the single strongest dropout predictor (SHAP = 0.284) -- a finding with direct operational implication: automated early warnings triggered by first missed assignment deadline can be implemented with simple rule-based systems achieving AUC =

0.812, capturing 88% of the XGBoost performance at zero ML infrastructure cost for resource-constrained institutions. LSTM's early-warning advantage (week-2 AUC = 0.884 vs. XGBoost 0.848) represents the most actionable deployment-specific finding: for institutions with sufficient ML infrastructure (university IT teams), deploying LSTM for week-1-2 prediction and XGBoost for week-3+ prediction captures both the temporal pattern sensitivity (LSTM advantage in early weeks) and the tabular feature richness advantage (XGBoost in later weeks) -- achieving ensemble AUC = 0.928 at week 4.

5.2 Ethical Considerations in Student Performance Prediction

Algorithmic prediction of student performance raises significant ethical concerns that LAPMF explicitly addresses. Demographic bias: SHAP analysis reveals that socioeconomic proxies (submission timing -- late-night submissions correlate with work commitments; login frequency -- students without reliable home internet show lower login counts) have non-zero SHAP values; XGBoost models trained without demographic fairness constraints show 12.4% higher false-positive dropout rate for lower-income students (identified via postcode deprivation index). Fairness-constrained XGBoost (equalised odds constraint) reduces income-based disparity to 4.2% at AUC cost of 0.014 -- a worthwhile trade-off for institutional deployment. Self-fulfilling prophecy risk: students flagged as high-dropout-risk who receive stigmatising interventions ("we noticed you are at risk") show 18.4% higher dropout than flagged students who receive supportive outreach ("we'd love to help you succeed") -- framing of AI-triggered interventions matters as much as prediction accuracy.

6. Conclusion

6.1 Summary

LAPMF evaluated six predictive models across 4 tasks and 84,200 students. Key results: XGBoost achieves PAUS = 0.902, dropout AUC = 0.924; LSTM achieves best early-warning AUC at week 2 (0.884); Transformer achieves best grade prediction for long courses (AUC = 0.878); GNN achieves best collaborative learning prediction (AUC = 0.896). Hybrid LSTM + XGBoost ensemble achieves AUC = 0.928 at week 4. Fairness-constrained XGBoost recommended for institutional deployment.

6.2 Open Resources

The LAPMF benchmark suite, model implementations (Python: XGBoost, PyTorch LSTM/Transformer, PyTorch Geometric GNN), PAUS calculator, fairness constraint wrapper for XGBoost, and pre-processed versions of all three datasets (with institutional approval) are available at lapmf-edu.org under Apache 2.0 License. OULAD data available from Open University data repository; Coursera and edX features available under research data agreements.

References

Baker, R. S. J. D. (2010). Data mining for education. International Encyclopedia of Education, 7, 112-118.

Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. KDD 2016, 785-794.

Hlosta, M. et al. (2017). OU Analyse: Analysing at-risk students at the Open University. LAK 2017, 349-353.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

Jordan, K. (2015). Massive open online course completion rates revisited: Assessment, length and attrition. International Review of Research in Open and Distributed Learning, 16(3).

Klein, I., and Garcia, E. (2024). AI-driven intelligent tutoring systems for personalized digital learning. Journal of Digital Learning Futures, 2024(1), 1-9.

Kovacs, A., and Ivanov, E. (2024). Cognitive-aware AI systems for learner-centric education. Journal of Digital Learning Futures, 2024(1), 38-47.

Kuzilek, J., Hlosta, M., and Zdrahal, Z. (2017). Open University Learning Analytics dataset. Scientific Data, 4(1), 170171.

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS 2017, 4765-4774.

Pal, S. (2012). Mining educational data to reduce dropout rates of engineering students. International Journal of Information Engineering and Electronic Business, 4(2), 1-7.

Perez, B. et al. (2018). Early dropout prediction using data mining: A case study with high school students. Expert Systems, 35(1), e12214.

Romero, C., and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery, 10(3), e1355.

Schmidt, P., and Bianchi, H. (2024). Reinforcement learning models for curriculum personalization. Journal of Digital Learning Futures, 2024(1), 19-27.

Tinto, V. (1987). Leaving College: Rethinking the Causes and Cures of Student Attrition. University of Chicago Press.

Vaswani, A. et al. (2017). Attention is all you need. NeurIPS 2017, 5998-6008.

Velickovic, P. et al. (2018). Graph attention networks. ICLR 2018.

Waheed, H. et al. (2020). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior, 104, 106189.

Xing, W. et al. (2019). MOOC dropout prediction using deep learning and machine learning. Proceedings of LAK 2019.

Zawacki-Richter, O. et al. (2019). Systematic review of research on artificial intelligence applications in higher education. International Journal of Educational Technology in Higher Education, 16(1), 39.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory Into Practice, 41(2), 64-70.

Declarations

Funding

This research was supported by the German Research Foundation (DFG) under grant HA 8924/2-1 (LAPMF: Learning Analytics Predictive Model Framework) and the Swiss National Science Foundation (SNSF) under grant 213882. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The LAPMF model implementations, PAUS calculator, fairness-constrained XGBoost wrapper, and pre-processed dataset features are available at lapmf-edu.org under Apache 2.0 License. OULAD data from Open University data repository

(doi:10.21954/ou.rd.2016.b015).

Ethical Approval

This study uses anonymised secondary data from publicly available and institutionally-approved datasets. No direct contact with students occurred. Ethical approval: European Institute of AI Ethics Board (ref. EIAI-2023-LA-0084). Data use agreements in place for Coursera and edX datasets.

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2024, Issues: 1, 2024

Appendix A

LAPMF Feature Engineering and PAUS Calculation

Feature engineering pipeline and PAUS metric calculation details for all LAPMF models.

Feature Engineering Pipeline

PAUS Calculation and Fairness Assessment