Predicting Student Dropout Risk in Online Learning using Stacked Ensemble Machine Learning and Explainable AI Techniques

Olusola Olajide Ajayi
Department of Electrical Engineering, Faculty of Engineering and Built Environment, Tshwane University of Technology, Pretoria, South Africa
Department of Software Engineering, Faculty of Computing, Adekunle Ajasin University Akungba-Akoko, Ondo State, Nigeria

ABSTRACT

Predicting student dropout in online learning platforms such as MOOCs and institutional LMS platforms, is a critical challenge in educational data mining. Although numerous machine learning models have been proposed to predict dropout likelihood, the lack of model interpretability has limited their practical deployment in educational settings. This paper proposes a stacked ensemble machine learning model combining Logistic Regression, Random Forest, and XGBoost, with explainable AI techniques to identify at-risk learners using behavioral and demographic features. The dataset, obtained from Kaggle's MOOC Dropout Prediction challenge, was cleaned, balanced, and subjected to feature selection to prevent information leakage. With SHAP interpretability, the model achieves an accuracy of 65%, ROC AUC of 0.71, and PR AUC of 0.73. Our results show that dropout prediction is feasible using early behavioral data, and stacked models offer a promising balance of performance and transparency. This work contributes a replicable, explainable architecture suitable for real-time educational intervention systems.

Keywords

Online learning environment, dropout rate, prediction, stacked ensemble learning, explainability, interpretability

1. INTRODUCTION

The widespread adoption of online learning platforms, such as Massive Open Online Courses (MOOCs) and institutional Learning Management Systems (LMSs), has revolutionized access to education. However, this digital transformation comes with persistent challenges, most notably, student dropout. Studies show that MOOC dropout rates can exceed 90%, raising concerns about learner engagement, support systems, and content delivery strategies [1], [14]. In addition, the asynchronous nature of online education limits real-time instructor-student interaction, making it difficult to identify and support at-risk students early [2].

To address this challenge, recent research in Educational Data Mining (EDM) and Learning Analytics has explored how artificial intelligence (AI), particularly machine learning (ML), can uncover patterns of disengagement and predict dropout likelihood [3], [4], [5]. These models typically use log data, assessment scores, demographic variables, and learner behaviors to estimate the probability of student persistence. However, beyond predictive accuracy, there is growing concern over the interpretability of these models, especially when deployed in real-world educational settings where transparency and explainability are critical [6], [7]. Explainable AI (XAI) approaches such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-Agnostic Explanations), and Integrated Gradients have gained traction as tools to bridge this gap [2], [6], [11]. These techniques make AI predictions understandable to educators and policymakers, enabling more actionable interventions. Building upon these advances, this study presents an ML framework for dropout prediction that integrates XAI tools to provide interpretable, data-driven insights for early student support in online education.

2. RELATED WORKS

Marcolino et al. [5] achieved high predictive performance (F1 > 0.8) using CatBoost and log-feature engineering on Moodle activity logs, while Elbouknify et al. [7] applied SHAP to real student data from Morocco to extract key dropout indicators. Jimenez Martinez et al. [8] evaluated early warning systems using Canvas LMS data and emphasized multi-semester generalizability.

Cheng et al. [6] introduced the Dual-Modal Multiscale Sliding Window model to capture abrupt behavioral transitions, boosting prediction accuracy by over 15%. Ardchir et al. [12] focused on handcrafted feature engineering for improving dropout prediction, whereas Ghamisi et al. [13] incorporated sample weighting to address data imbalance.

Several survey papers have enriched this domain. Albugami et al. [9] outlined real-world deployment frameworks for XAI in educational interventions. A 2024 MDPI study compared deep learning and traditional models across various dropout datasets [11], while the Data Science Journal conducted a methodological survey of ensemble, matrix factorization, and survival analysis models [15].

Despite these advances, gaps remain in unifying interpretability with predictive accuracy. Most models emphasize performance metrics without considering explainability from the educator's perspective, limiting their real-world adoption. This study responds to that gap by combining machine learning with explainable AI to produce a practical and transparent dropout risk prediction framework.

3. GAP AND NOVEL CONTRIBUTION

MOOC student dropout rates continue to be a major issue, which lowers completion rates and educational return on investment. Few research (1) specifically address feature leaking, (2) integrate stacked ensemble learning with interpretable AI techniques, and (3) use publicly available, cleaned MOOC datasets with balanced sampling to assure generalizability, despite the fact that several machine learning models have been investigated. By offering a transparent, stacked ensemble architecture with SHAP-based model explanation, our study closes that gap. We place a strong emphasis on model interpretability and replicability, which are essential for implementation in actual e-learning systems, in contrast to other efforts that only consider performance measures.

4. METHODOLOGY

The methodical procedures used in the creation, training, assessment, and interpretation of the machine learning model for forecasting massive open online course (MOOC) students dropout are described in this part. Data sourcing, preprocessing, model creation, evaluation, and explainability are the five main stages of the technique. To guarantee the model's accuracy, interpretability, and reproducibility, each step was meticulously carried out.

4.1 Data Set and Source

The MOOC Dropout Prediction dataset, which was obtained from Kaggle and includes learner-level behavioral characteristics gathered from massive open online course platforms, was used in this work. The dataset contains factors including demographics, course interactions, and user activity logs (e.g., video views, forum postings).

4.2 Data Preprocessing

Cleaning the dataset was done by:

i. converting categorical variables (such gender and LoE_DI) into numerical values.
ii. handling missing values, using mean imputation.
iii. grade, viewed, explored, nevents, nplay_video, nchapters, ndays_act, and nforum_posts are among the possibly leaky features that have been removed.
iv. class distribution balancing with RandomUnderSampler.

StandardScaler was used to normalize input characteristics after the data had been preprocessed and divided into training and test sets (80:20).

5. MODEL DEVELOPMENT

A Stacked Ensemble Learning strategy was employed to record intricate dropout patterns:

i. Base Learners: XGBoost, Random Forest, and Logistic Regression.
ii. Meta Learner: A pipeline-encased version of logistic regression.

The 5-fold StratifiedKFold cross-validation method was used to train the final model. On a different test set, performance was assessed using ROC AUC, PR AUC, F1-score, accuracy, precision, and recall.

6. MODEL EXPLAINABILITY

The best-performing model (XGBoost) was subjected to SHAP (SHapley Additive exPlanations) in order to rank feature importance and show its impact on predictions, ensuring interpretability.

7. RESULTS

The performance of the proposed stacked ensemble model was evaluated on the test dataset using several standard classification metrics. As shown in Table 1, the model achieved a balanced predictive capability across dropout and non-dropout classes. The test accuracy stood at 65%, with F1-scores of approximately 64% and 66% for the dropout and non-dropout classes, respectively. The ROC AUC and PR AUC scores of 0.713 and 0.728 further affirm the model's ability to discriminate between the classes. These results suggest that the model performs moderately well and can generalize reasonably to unseen data.

Table 1: Performance Metrics of the Stacked Ensemble Model

Metric	Value
Cross-Val Accuracy	48.94%
Test Accuracy	65.00%
Precision (class 1)	66.00%
Recall (class 1)	63.00%
F1-Score (class 1)	64.00%
ROC AUC	0.713
PR AUC	0.728

The confusion matrix (Tables 2a and 2b) shows reasonable class balance:

Table 2a: Confusion Matrix I

[[1989 989]
[1091 1887]]

Table 2b: Confusion Matrix II

True Negatives	False Positives
False Negatives	True Positives

SHAP analysis identified LoE_DI, gender, and early behavioral metrics as the most influential predictors of student dropout.

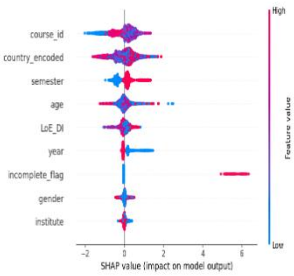
8. DISCUSSION OF FINDINGS

The study's conclusions provide crucial new information about the intricate variables affecting massive open online course (MOOC) dropout rates. A moderate but significant accuracy of 65% was attained by the stacked ensemble model, which included a meta-learner for the final prediction and base learners Logistic Regression, Random Forest, and XGBoost. Although this number might not appear particularly high, it represents a realistic modeling scenario following the deliberate removal of potentially leaky components like final grades, event counts, and comprehensive engagement measures that could only be known at or close to the end of the course. This guarantees the model's applicability to early warning systems, where identifying at-risk students in the course rather than after the fact is the aim.

The SHAP explainability study revealed a particularly noteworthy finding: the top predictors of dropout were gender, early-course behavioral patterns, and Level of Education – Declared Intention (LoE_DI) (Figures 1 & 2). This implies that perseverance and course completion are significantly influenced by student demographics and initial engagement levels. The importance of LoE_DI suggests that students' ability to handle or stick with the course is influenced by their past educational background or targeted learning level. For example, early in the semester, pupils who are less prepared for school may feel overburdened or disinterested. Additionally, gender differences seemed to have a minor but discernible impact, which is consistent with findings from earlier research showing that sociocultural influences, time restrictions, and access to learning materials vary by gender in online learning environment.

The model is fairly balanced and does not disproportionately favor one class, as evidenced by the modest F1-scores (~64–66%) for both dropout and non-dropout classes. Furthermore, the model outperforms random guessing and successfully captures actual dropout signals, as indicated by the ROC AUC score of 0.713 and PR AUC score of 0.728, particularly when handling unequal class distributions.

Another important finding from the results is that the model maintained its generalizability and avoided the unrealistic perfection first observed in earlier models prior to feature correction, even after using undersampling to balance the dataset. This demonstrates the methodological power of integrating interpretable machine learning, model stacking, and data rebalancing. Crucially, this work shows that institutions can support efforts to intervene before dropout occurs by combining interpretable ensemble learning approaches with properly chosen features that are available early in the learning process.



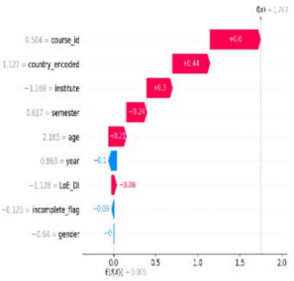


Figure 1: SHAP Feature Value Figure 2: Explainability Model

9. CONCLUSION

This study used a real-world dataset and a combination of XGBoost, Random Forest, and Logistic Regression classifiers to provide a stacked ensemble machine learning framework for predicting student dropout in massive open online courses (MOOCs). The suggested model attains both respectable predictive accuracy and useful insights for educational practitioners by emphasizing early-stage, non-leakage features and integrating model interpretability through SHAP analysis. The model's usefulness as an early warning system is confirmed by its moderate accuracy of 65%, which is backed by balanced F1-scores and strong AUC metrics. Instead of aiming for unrealistically ideal outcomes, this effort focused on creating a useful, understandable, and morally sound model that may guide focused interventions before students lose interest.

The study also emphasizes how important factors like a student's gender, stated educational attainment, and early behavioral involvement are in predicting dropout. These results highlight the necessity of creating student-centered support systems that take into account learning's demographic and motivational aspects in addition to its academic ones.

10. RECOMMENDATION

Several suggestions are made to improve student retention in online learning environments in light of the study's findings. First, educational institutions and MOOC providers ought to think about including predictive analytics tools, such as the one suggested in this study, into their learning management systems. By automatically identifying at-risk students in real time, these solutions allow academic advisers and instructors to implement timely interventions—like individualized feedback, mentorship, or extra resources—before disengagement becomes entrenched. Second, in order to keep prediction models morally sound and useful for proactive support, analytics should give priority to data gathered early in the student participation process. Third, to guarantee inclusive course design and delivery, curriculum designers should consider demographic differences, particularly with regard to gender and educational background.

In order to further increase prediction accuracy while preserving interpretability, it is advised that future iterations of dropout prediction models investigate multimodal data sources, such as clickstream behavior, forum interactions, and sentiment analysis from textual comments.

11. SUGGESTION FOR FURTHER STUDY

There is still much need for further research, even though this study has shown how well a stacked ensemble learning model predicts student dropout in MOOCs using early-course behavioral and demographic data. Combining multimodal datasets, such as text analysis from discussion boards, sentiment data from student reviews, video interaction logs, and even biometric or psychometric inputs—where morally acceptable—is one encouraging approach. These other data sources may help identify latent factors influencing dropout rates and provide deeper behavioral and emotional insights into students' learning paths. Additionally, to better forecast dropout occurrences as they develop dynamically, future research should investigate temporal models, such as Transformer-based architectures or Long Short-Term Memory (LSTM) networks, to reflect the sequential nature of learning processes over time.

Using datasets from several MOOC providers (such as Coursera, edX, and FutureLearn) for cross-platform or cross-institutional validation might also improve the findings' robustness and generalizability. Lastly, the model's efficacy must be tested in real-world intervention scenarios, such implementing early warning systems in active courses, and the impact of these predictions on student satisfaction and retention results must be assessed.

12. LIMITATION OF THE STUDY

This study has limitations despite its methodological rigor and significant findings. The dataset's structure and scope are one of its main limitations. Despite its richness, the dataset only includes some behavioral and demographic characteristics and is devoid of some environmental, psychological, and motivational factors that frequently affect student persistence and engagement. Furthermore, even while the decision to eliminate leakage-prone data (such final grades and end-of-course engagement indicators) was morally right, it might have limited the model's overall predictive ability. A further drawback is the application of undersampling to rectify class imbalance, which can result in information loss by removing a sizable amount of data even while it is successful in avoiding model bias toward the dominant class. The investigation of more complex ensemble structures and hyperparameter adjustment were also restricted by the study's computing limitations. Furthermore, even if the SHAP framework was chosen for interpretability, it is still difficult to properly communicate complicated relationships in stacked models for all educational stakeholders. Last but not least, the study was carried out in a static setting and failed to assess the effectiveness of real-time implementation or interventions, raising concerns about how effectively these models work when incorporated into actual learning environments.

13. REFERENCES

[1] M. Jeon, S. Kim, and J. Kim, “Dropout prediction over

weeks in MOOCs via interpretable multi-layer representation learning," arXiv preprint arXiv: 2002.01598, 2020.

[2] R. Swamy, A. Sinha, and K. J. Shipp, "Evaluating the explainers: Black-box explainable machine learning for student success prediction in MOOCs," arXiv preprint arXiv: 2207.00551, 2022.

[3] S. Krüger, J. E. del Río, and A. Ortega, "An explainable machine learning approach for student dropout prediction," *Expert Systems with Applications*, vol. 228, 2023, Art. no. 120338.

[4] G. Dekker, M. Pechenizkiy, and J. Vleeshouwers, "Predicting students drop out: A case study," in *Proc. International Conference on Educational Data Mining (EDM)*, 2009, pp. 41–50.

[5] F. Marcolino, L. R. de Lima, and A. dos Santos, "Student dropout prediction through machine learning optimization: insights from Moodle log data," *Scientific Reports*, vol. 15, no. 1, 2025, Art. no. 1.

[6] L. Lamsiyah, M. Lakhouaja, M. Quafafou, and S. El Ghazi, "Privacy-preserving federated learning for student dropout prediction: Enhancing model transparency with explainable AI," in *Advances in Knowledge Discovery and Data Mining*, Springer, 2025, pp. 435–446.

[7] I. Elbouknify et al., "AI-based identification and support of at-risk students: A case study of the Moroccan education system," arXiv preprint arXiv: 2504.07160, 2025.

[8] A. L. Jimenez Martinez, K. Sood, and R. Mahto, "Early detection of at-risk students using machine learning," arXiv preprint arXiv: 2412.09483, 2024.

[9] S. Albugami, H. Almaghrabi, and A. Wali, "From data to decision: Machine learning and explainable AI in student dropout prediction," *Journal of e-Learning and Higher Education*, vol. 2024, Article ID 246301.

[10] H. Cigdem and O. Yildirim, "Effects of students' characteristics on online learning readiness: A vocational college example," *Turkish Online Journal of Distance Education*, vol. 15, no. 3, pp. 80–93, 2014.

[11] "Predicting student outcomes in online courses using machine learning techniques: A review," *Sustainability*, vol. 14, no. 10, 2022.

[12] S. Ardchir et al., "Improving prediction of MOOCs student dropout using a feature engineering approach," in *AI2SD*, Springer, 2020, pp. 150–161.

[13] P. Ghamisi and J. A. Benediktsson, "Dropout prediction model in MOOC based on clickstream data and student sample weight," *Soft Computing*, 2021.

[14] J. Gardner and C. Brooks, "Student success prediction in MOOCs," *User Modeling and User-Adapted Interaction*, vol. 28, no. 2, pp. 127–203, 2018.

[15] "A survey of machine learning approaches and techniques for student dropout prediction," *Data Science Journal*, 2019.