The 14th International Learning Analytics and Knowledge Conference
LAK 24
March 18 - 22, 2024

Mobile Learning Dropout Prediction: Identifying Key Features

Sylwia Macinska & Lakshmi Babu Saheer

Cambridge University Press & Assessment, Anglia Ruskin University

CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

INTRODUCTION

Despite its advantages, online learning faces higher attrition rates than traditional classrooms, influenced by factors like isolation, lack of community, and the need for strong self-regulation in learning.

While predictive models for detecting learners at risk of dropout have been proposed in the past [1, 2], mobile learning environments differ from traditional online courses as they tend to have limited features due to smaller screens, suffer from greater potential for distractions and differ in access to educational content.

Because of this, typical dropout predictors may no longer be relevant or simply unavailable in the context of mobile learning.

AIM

To address this issue, this paper investigates the dropout prediction problem specifically in mobile learning applications, using a language learning app as a case study.

The following research questions are explored:

1. What are the best predictors of engagement for mobile learning environments and whether they differ from engagement predictors used in traditional online learning platforms?
2. What is the minimum sample period required to reliably identify learners at risk of dropout in mobile learning environments?

METHODOLOGY

The study utilised data from **a self-study language learning app** employing exercise-based instruction. Learners are presented with bundles of four activities, one for each of the language skills: reading, listening, speaking, and writing.

A new bundle of activities becomes available every 24 hours, which increase in difficulty over time. The app is available to anyone, and it is free to use.

The dataset included learner data from a **12-month period**, considering only those who began using the app within this timeframe. The final dataset comprised **15,512 learners**.

Engagement was defined based on the completion of activity bundles. Learners were categorised into three groups:

1. early dropouts (completed up to three bundles, 80% of learners)
2. high risk (completed 4 to 12 bundles, 14%)
3. low risk (completed more than 12 bundles, 6%)

FEATURE ENGINEERING AND MODEL SELECTION

Three predictor categories were identified based on prior literature on traditional online learning environments [1, 2] and exploratory data analysis of the current dataset:

- **performance-related:** proportion of correct responses, response time, attempt rate, and proportion of skipped questions
- **session-related:** time to first activity, number of logins, weekend study, notification interactions, and session duration
- **self-regulated learning (SRL)-related**, reflecting planning, monitoring, and regulating activities: time on instructions, feedback, progress views, and learning interval.

Machine learning (ML) techniques were also selected based on prior literature and included: Logistic Regression (LR), Support Vector Machines (SVMs), Decision Trees (DTs), Random Forest (RF), K-Nearest Neighbours (KNN), and Extreme Gradient Boosting (XGBoost). Opaque models were not considered given the context.

The best hyperparameters for each model were identified through a Coarse-to-Fine informed search, and models were evaluated using stratified 5-fold cross-validation for generalisability. Performance accuracy was assessed on an unseen test set (a random holdout sample). Because of the imbalanced dataset, SMOTE was applied after the train-test split to balance class representation in the training set.

RESULTS

Figure 1 and Table 1 show models' performance as measured by accuracy score.

RF and XGBoost were found the most effective. However, XGBoost showed less variability in performance and the highest performance on an unseen test set with a score of **87%**. Further, the model was highly accurate (97%) in identifying early dropouts and performed well (86%) with the low-risk group. Misclassifications were more common in the high-risk group, often labelled as early dropouts.

Table 1: Final performance of each model on an unseen test set.

ML technique	Accuracy score
LR	79%
SVM	82%
KNN	76%
DT	77%
RF	86%
XGB	87%

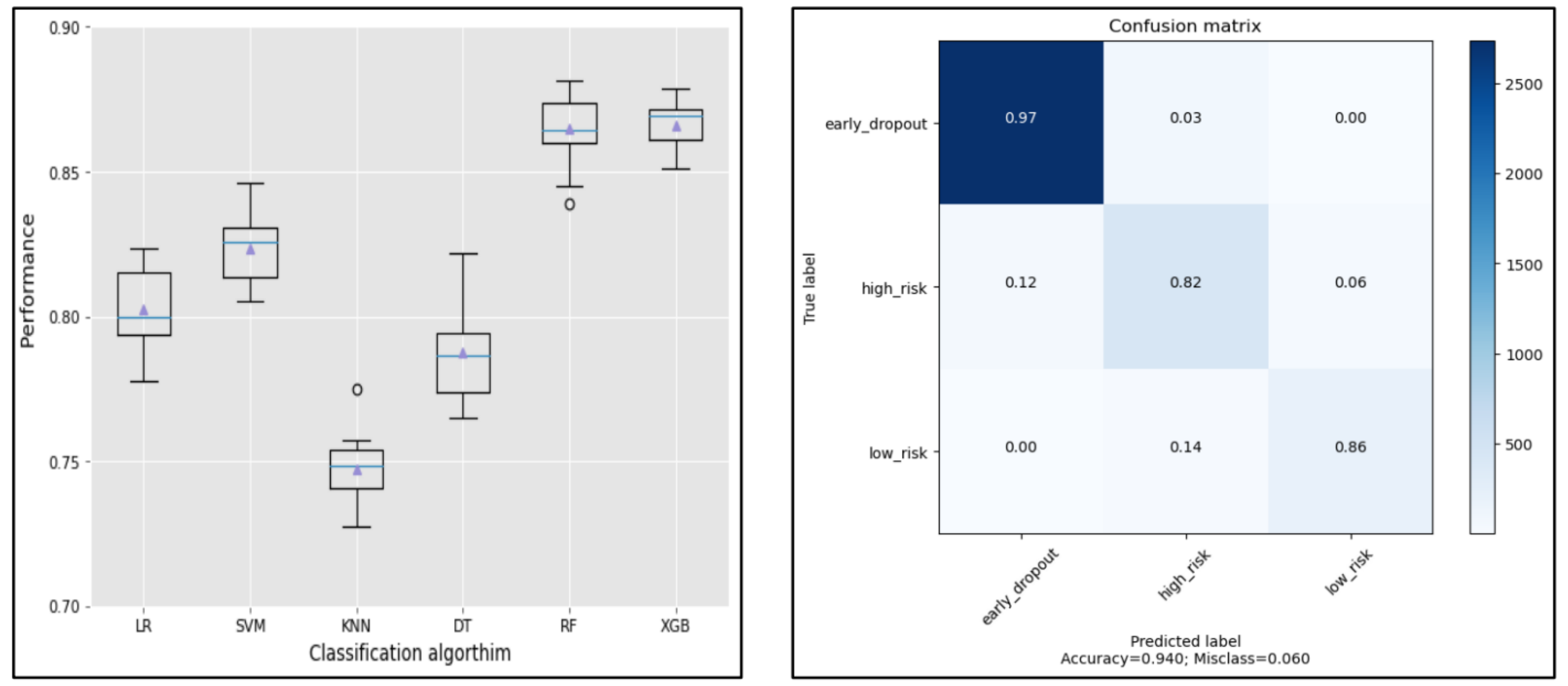

Figure 1: Models' performance based on 5-fold cross-validation (left) and confusion matrix for XGBoost algorithm (right)

To understand the impact of different feature categories, the models were retrained using only performance-related, session-related, or self-regulated learning (SRL) features.

Session-related features had the most significant impact (81%), followed by SRL (67%) and performance features (60%). Individual feature importance was also analysed using a drop column technique.

The number of logins, learning interval, and median session time were among the most impactful features.

The study examined different time periods, ranging from 2 days to 6 weeks, to determine the shortest amount of data needed to identify early dropouts [3].

The model's performance improved as more data became available (see: Figure 2). A seven-day sample period was enough to correctly classify 69% of learners, increasing to 77% after two weeks.

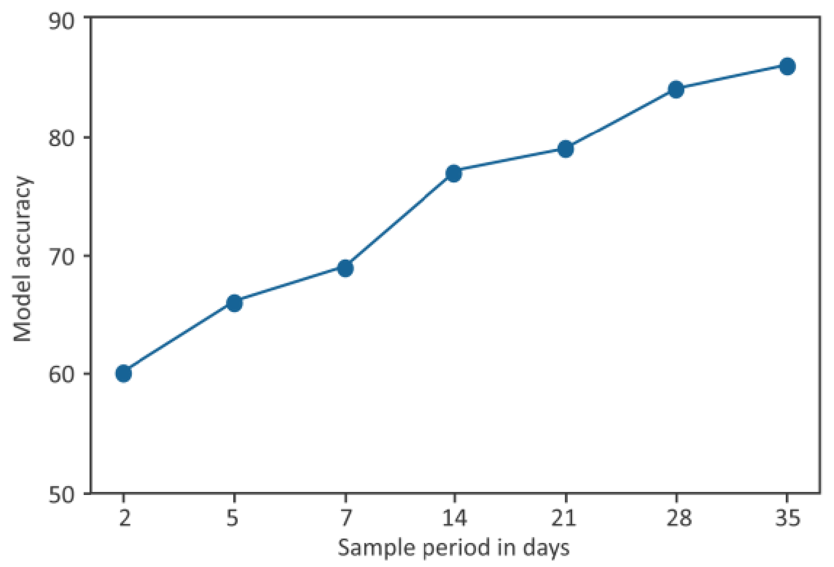

Figure 2: Model accuracy as a function of a sample period length

DISCUSSION

Despite the absence of some features like forum activity and video-related activity, common in traditional learning platforms, **the resulting model still achieved an 87% accuracy**, in line with models developed for the traditional online platforms (e.g., [4]). Unsurprisingly, a trade-off was observed between the accuracy and length of the sample period. Yet, a satisfactory 77% accuracy was achieved with as little as two weeks' worth of data.

Session-related features emerged as key predictors, underlining the importance of frequent interactions with the app. A model based solely on these features attained 81% accuracy, reinforcing the link between regular app usage and lower dropout risk. This aligns with traditional learning environments, where attendance is a strong engagement and performance indicator. Surprisingly, SRL features, yielded only a 66% accuracy when used alone. However, the SRL indicators used in this study were rather superficial and might have not represented SRL behaviours sufficiently.

The findings suggest that **encouraging regular app usage can reduce dropout rates**. This can be achieved by incorporating features that address learners' educational needs and increase affective responses like interest and enjoyment (e.g., personalised content) and use of positive reinforcement (e.g., badges, positive feedback) to create a sense of accomplishment [5].

This research contributes to understanding dropout prediction in mobile learning, suggesting that even simple mobile platforms can provide sufficient data for effective dropout prediction.

REFERENCES

1. Dalipi, F., Imran, A. S., & Kastrati, Z. (2018, April). MOOC dropout prediction using machine learning techniques: Review and research challenges. In *2018 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1007-1014). IEEE.

2. Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., & Kloos, C. D. (2018). Prediction in MOOCs: A review and future research directions. *IEEE transactions on Learning Technologies, 12*(3), 384-401.

3. Moreno-Marcos, P. M., Muñoz-Merino, P. J., Maldonado-Mahauad, J., Pérez-Sanagustín, M., Alario-Hoyos, C., & Kloos, C. D. (2020). Temporal analysis for dropout prediction using selfregulated learning strategies in self-paced MOOCs. *Computers & Education, 145*, 103728.

4. Liang, J., Li, C., & Zheng, L. (2016, August). Machine learning application in MOOCs: Dropout prediction. In *2016 11th International Conference on Computer Science & Education (ICCSE)* (pp. 52-57). IEEE.

5. Fan, M., Ndavi, J., Qalati, S., Huang, L., & Pu, Z. (2021). Applying the time continuum model of motivation to explain how major factors affect mobile learning motivation: a comparison of SEM and fsQCA. *Online Inf. Rev., 46*, 1095-1114.

CONTACT

Sylwia Macinska: sylwia.macinska@cambridge.org