Predicting at-risk students using weekly activities and Assessments

Jawthari Moohanad[1], **Veronika Stoffová** [1,2]

[1] Eötvös Loránd University, Department of Media & Educational Informatics
[2] Trnava University, Department of Mathematics and Computer Science
moohrash@yahoo.com

Introduction

Higher education institutes face challenges caused by low course enrolments and beyond lower course completion rates. The issue higher dropping out rates is fast becoming a priority, and universities are seeking for strategies to improve students' retention rates. OCED reported that in Australia just 31% of students' completed a 4-year degree programme, US had 49% completions while UK is on top with 71% completions [1]. Lower retention rates are a serious threat to universities long term financial security. Hence, Universities are focusing on identifying strategies which ensure students successes and that can provide proactive actions to support students in their course work. Having some analytical strategy that can enable predictions on students' performances will help those institutes to make timely interventions for improving students' performance.

The Common use of tools like Student Management Systems (SMS) and Learning Management Systems (LMS) have supported higher education institutes in providing seamless online communication, in delivery of learning and teaching resources, designing interactive learning activities and managing academic assessments. In addition, they provide them with large datasets that are related to students' demographics student academic records and log files. These logs are based upon students' interactions with the LMS and have offered us with new research directions that can help in improving students' academic performance [2][3].

In the research, OU Learning Analytics Dataset (OULAD). The dataset comprises student demographics, clickstream history, and assessment submission information of 32,593 students over a course duration of 9 months, from 2014 to 2015. The data is composed of several courses, with each course being taught at different intervals in a year. Four distinct performance classes were defined: distinction, pass, fail, and withdrawal. A course belonging to Engineering and technology category was chosen with 1303 students. The OULAD comprised students' information regarding their interaction with the Virtual Learning Environment (VLE)—their assessments, quizzes, and course performances. VLE interaction was classified into 20 different activity types with each activity referring to a specific action, such as downloading or viewing lectures, course content, or quizzes. The current course has 536,837 records in the VLE log, and 10373 record in the assessments log. The research has three questions: does aggregating students weekly clicks and per week assessment when available predict at-risk student early? Other question was: does accumulating previous weeks assessments provide better prediction results compared to previous question results? Third question was: Which demographic features are affective in first weeks, before assessments availability, beside weekly activities?

Methodology and challenges

OULAD data is provided as tables, so it can be used directly with machine learning models. Logs were maintained daily; a "data" column represents a day - 0 to the end of the course. First step was to aggregate the events into weekly format. Then students VLE table and assessments were processed and divided into 34 sub datasets; a week dataset will have all the student who had an activity during that week. For first question, we followed this strategy: $w_n = \sum_{i=0}^{n} W^i$ to obtain the dataset for week number n. Predictor variables for a given week is the count of total online activities performed that week. Second variable could be assessment scores (if accessible at that time). Second question datasets were accumulating previous assessments in the subsequent weeks as predictors in addition to the available assessment score at that week.

To answer question three: the first 3 weeks subsets predictors further analysed to select the best features by Information Gain.

To classify students in two group (either at-risk or not-at-risk of failing the course), those machine learning algorithms were used: Random Forest classifier (RF), Naive Bayes (NB), Logistic regression (LR), Linear Discriminating analysis (LDA). 10-fold cross validation was used to train all the classifiers, and F-measure was used as an evaluation metric.

Results

Strategy 1:

Results showed that LDA outperformed other models with score of 0.69 in first week- using just clickstream data. In second week, LR was the best scoring 0.715, and LDA was close. RF was leading in third week with 0.777 score. All models scored improved 4-5% in this week as the first assessment grade was available. All models scores were 74% above starting from forth week.

Strategy 2:

The first 3 weeks data was similar to strategy 1 because during those weeks, only engagement data was available. In forth week and onward, the improvement was 3% and more in the models performances. The most noticeable improvement was in RF; Rf scored 0.692 in 4th week in strategy 1, while it scored 0.79 I in 4th week subset of strategy 2.

For answering research question 3, we applied feature importance and ranking techniques on the first three weeks datasets after combining demographic features with and engagement data. The purpose was to select just the affective features that would improve the performance.

To sum up, this research proposed two analytics strategies to predict at-risk students based on their weekly activities. Experiments shows that strategy two is better and was able to predict struggling students in early weeks. Also, feature selection techniques improved the prediction based on behaviour and demographic features.

References

[1] O. Indicators, “Education at a glance 2016,” Editions OECD, 2012.

[2] C. R. Graham, “Blended learning systems,” The handbook of blended learning: Global perspectives, local designs, pp. 3-21, 2006.

[3] N. Cavus and T. Zabadi, “A comparison of open source learning management systems,” Procedia-Social and Behavioral Sciences, vol. 143, pp. 521-526, 2014.