A Survey of Student Dropout Analysis Using Machine Learning Approach

Zhilu Wang *

Department of Mathematics, University College London, Gower Street, London, UK WC1E 6BT

* Corresponding author: zczl467@ucl.ac.uk

Abstract. Online learning and conventional learning are two important methods that students pursue their educational degrees or expand their knowledge. The enormous rate of dropout of online students and the fact that this rate is still increasing leads to the concern of raise student retention rate. And the large number of conventional student dropout create loss in economics, time and education resources. Therefore, it is important to seek the factor affecting student dropout and investigate effective machine learning-based models on prediction of student dropout. Different factors including the clickstream, academic information, family are discussed, followed by introductory information about current popular machine learning algorithms. According to this comprehensive review, the grade of online learners and the total number of students assessment to courses seem to be the most powerful features and conventional students are affected by social contact and attendance of social activities.

Keywords: Student dropout, Online learning, Conventional Learning, Massive Open Online Courses, machine learning.

1. Introduction

Online learning and conventional learning are two major study strategies in 21st centuries. Online learning applying Massive Open Online Courses (MOOCs) has received increasing attention and support because it allows users to assess up-to-date resources unlimitedly that would be impossible with the absence of invention and development of information and communication areas [1]. Coursera, Udacity, edX, etc are main MOOCs' management systems currently that collect information through hierarchical structure [2]. An explicit example of MOOCs is French national MOOC platform which collects over 1,000,000 registrations including user ids, time stamps and the ownness of certifications [3]. The reasons that MOOCs gets increasingly attractive come from the fact that more and more people pay attention to self-learning through MOOCs and choose a lifelong learning attitude [4]. The wide range of courses and high-quality materials provided by MOOCs makes it possible to reshape opportunities of learning and reaching higher education. Leading universities often organize MOOCs' programs that allow users to learn without cost (or with minimal cost) despite geographical and certification differences.

Unlike online learning, traditional learning requires students to take in-person lectures, tutorials, and exams. Students graduated from university with certification are more likely finding a job and have higher profit than students without a tertiary education [5]. However, the number of students who fail to earn a certificate is enormous and is still increasing. To illustrate, the estimated number of dropout students who are aged 7-17 is 5.1 million in 2017, with around 1.5 million of secondary students [6]. It has been reported that annual dropout rates among high school students from Tanzania raised from 3.8% to 4.2% [6]. Dropout of traditional student would cause costs mainly to individuals and institutions. Considerably impacts are education resources, time and money. Therefore, it is crucial to mitigate dropout rate of traditional students.

Similarly, the dropout rate of MOOCs students tends to be extremely high. For example, Massachusetts Institute of Technology discovered that around 96 percent of students tend to give up their MOOCs programs [7]. On the other hand, online students in developing countries may have higher retention rate than developed countries according to Brothers, for less than half of the students choose to quit [2]. Students' situations including engagement, certification, personality and lack of

wealth and time lead to the low retention rate. Apart from that, limits of techniques including bad arrangement of platforms, incomparable with mobile devices and poor communication between students and teachers may also cause high dropout rate [7]. Therefore, it is extremely important for institutions to mitigate the large dropout rate [8]. One of the most common ways is constructing Student Dropout Prediction Model, therefore institutions can take early interventions to the substantially failing students.

In this paper, different variables collected in student dropout dataset that affect the ML model will be discussed, followed by the comparison of different types of machine learning models. Hence one of the most comprehensive models will be deduced.

2. Review on dropout factors

Institutions decide to construct Student Dropout Predicting Model based on machine learning. Different classifications including DT, SVM, ANN and so on [8]. Because of the fact that traditional undergraduates have higher retention rates than online-learning students and the importance for online learners to obtain a degree certification, it is crucial for universities and institutions to evaluate the factors influencing online-learning and traditional student dropout.

2.1. Machine Learning Procedure

Figure 1 demonstrates the process of constructing ML model for predicting student dropout. First of all, the raw data is collected and cleaned, then important features are extracted from the dataset and used in pre-processing data. ML algorithm is then selected and processed with the chosen features to produce predictive model. With data visualization, the model will produce the prediction of student dropout rate. All of the discussed models based on ML algorithms follow similar approaches shown in Figure 1.

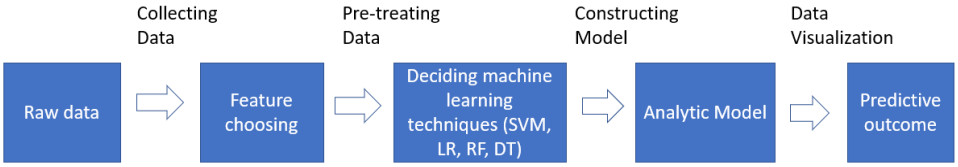

Figure 1. Process of constructing a ML model.

2.2. Clickstream and Forum

Clickstream and Forum represents the total number of students activities recorded in the dataset. Students' activities include the total number of course interactions, posts in the forum. To illustrate, Harvard's Dataverse repository stored one dataset that considers nevents (number of interactions with the course), nplay_video (number of studies the recordings with the course), nchapters (number of chapters user read) and nforum_posts (number of posts to the discussion forum) [1]. The study shows that the correlation values of nevents and nchapters against target variable (certificate: 0/1) are around 0.70, which are seen as related. On the other hand, the correlation value of nplay_video is much smaller (0.41), while the correlation value of norum_posts is less than 0.01, so both are considered to be uncorrelated [1]. Another dataset provided by KDD Cup 2015 includes activities. Each activity consists of ID of enrolment record, student name, ID of the course, event type and so on [8]. The activities were further decomposed to the numbers of problem events, video events, access events, wiki events and discussion events [8]. These features clarify Dropout Prediction Model and are helpful in construction prediction models.

2.3. Academic Information

Academic information includes Education level, assignment, and test grades. Education level is a very important criteria in determining the probability of learning situation of MOOCs students. A

survey carried out at Politecnico di Milano considers the types of previous studies (possible values: Scientific, Classic, Technical, Other) and admission test grade as important features in constructing ML Models [9]. The importance of assignment is announced by a survey carried out by Janka and Martin [10]. Another study which uses the data from the Hellenic Open University takes two written assignments as variables (possible values: no, fail, good, excellent) when applying ML techniques [11]. The educational data provided by VLE databases collects a total score from different types of LMS Moodle activities and online tests as assignments, which is further chosen as a variable in different types of machine learning models [10]. Apart from the above two features, the exam grades are taken into consideration. For instance, the number of students absenting from final exam for two consecutive semesters is viewed as one of the important features collected from the dataset stored in the Open University of China [11]. The score of final written exam which is collected in the dataset from Hawassa University Student Information System Portal is considered as an important feature in developing ML models [12]. The grades of important exams are commonly set as variables in ML models, possible reasons may be that the score influences the confidence of the MOOCs users.

2.4. Family and Surrounding

Family and surrounding factor includes parents' income, student satisfaction to surrounding, social interaction, and social presence. According to Nurmalitasari et al., patents' income would play a great role in lower the retention rate in Indonesia, since 1. If the family has no profit to support student's study, then the student will directly quit from the program; 2. If the student tries to study while working, then time must be divided and lead to hardship in following learning process [13]. Contini and Zotti state that shortage of parents' income could create solid barrier to the process of study [14]. Apart from economic reasons, students may also be affected by their academic satisfaction. The relationship between lecturers/tutors and student satisfaction towards pedagogic quality (possible values: very content, content, indifferent, uncontent, very uncontent) play an important role on students dropping out [15]. Thirdly, in-person teaching provides better opportunities for students to communicate with tutors and schoolmates. According to Lee and Choi, social contact and being in-person activities influence student's purpose [16]. The study from Zhang et al. also supports this idea. With the fact that encouragement and care from kins, friends, and other students can affect student's willing of complete the programs [17], it is important to take social interaction and social presence as one of the key factors.

3. Discussion

ML models are very effective in predicting the student dropout rate. Table 1 provides the introductory information about recent studies of different classifiers. Following are brief introduction of ML algorithms used.

SVM is a popular supervised algorithm, which can discover a hyperplane that maximizes the gap between groups to create classifications. In the studies of predicting MOOCs student dropout, SVM has been applied when choose grade-related data as features. For instance, Kabathova and Drlik build a SVM model with the score of projects, final exam, along with the score of assessment [10]. Bujang et al. also build a SVM model with further classified grade groups, including the grade of tutorial, quiz, total grade of the study year and so on [12].

LR is a classification algorithm that can predict the results of dichotomous dependent classes. In the studies of predicting MOOCs student retention rate, LR has been widely used in a variety of datasets. For example, Kumar, Amar, and Ashok use LR algorithm to predict the student dropout and evaluate the correlation between features and target values [1]. On the other hand, Kabathova and Drlik build a LR model in order to find the relationship between the grade and student dropout [10].

DT are supervised learning algorithms that aim to predict the outcome through learning simple decision rules from the variables. For example, the study carried out by Mnyawami et al. uses DT to predict the student dropout rate in secondary school [5].

RF applies multiple DT algorithms that are created to predict the outcome using bootstrapping, resampling, along with averaging [10]. In the studies of predicting MOOCs student dropout, RF is normally carried out with DT.

Table 1. Introductory information of different classifiers.

Reference	Description and methodology	Highlights	Performance
Kumar, Amar, and Ashok [1].	To select the significant features highly correlated towards students' dropout using • LR • DT • RF • KNN.	Find and discuss the correlation between features and target value.	Accuracy rate ranges from 0.95 to 0.96.
Kabathova and Drlik [10].	Use the assignments, midterm test and project assignment score to predict the student dropout rate using. • NB • RF • NN • LR • SVM • DT.	The score of projects is also taken into consideration since this score is also directly connected to complete of the course.	Accuracy rate ranges from 0.77 to 0.98.
Tan and Shao [12].	To predict the student dropout rate based on the dataset from the Open University of China, using. • ANN • DT • BN	Introduce the use of the confusion matrix to show the modelling results.	Accuracy rate ranges from 0.9392 to 0.9463.
Bujang et al. [13].	To predict student dropout using • LR • DT • RF • SVM	• Divide grade into several features, including grade from quiz, tutorial, total grade of the study year. • Classify the group value according to student performance.	Accuracy rate ranges from 85.9 to 99.8.

4. Conclusion

In this paper, the strategy in choose features that are important in affecting both online and conventional student dropout from a large dataset to provide better ML models is discussed, followed by discussion in the comparison between several studies using different ML algorithms. It turns out that the grade of students and the total number of students assessment to courses content are often taken into consideration and further set as features, which may inspire the idea that future data collection for MOOCs student need to include grade and assessment of courses. For traditional students, one of the most feasible and effective ways is improving teaching quality and provide a great communicative atmosphere. In the section of discussion, the most popular ML techniques are introduced, followed by comparison between recent studies. Most of the mentioned studies suggest that DT may be the necessary model since it provides high accuracy rate in a variety of different dataset. In the future, researchers may develop more specific models in predicting student dropout.

Finding ways to mitigate the student dropout from the limitation of MOOCs platform may also need to be researched in the future.

References

[1] Kumar, G., Singh, A. and Sharma, A., "To evaluate the performance of machine learning algorithms in predicting student dropout on MOOC platforms", Journal of Physics: Conference Series, doi:10.1088/1742-6596/2327/1/012063 (2022).

[2] Gupta, K. P., "Investigating the adoption of MOOCs in a developing country. Application of technology-user-environment framework and self-determination theory", Interactive Technology and Smart Education, doi: 10.1108/ITSE-06-2019-0033 (2019).

[3] Wintermute, E. H., Cisel, M. and Linder, A. B., "A survival model for course-course interactions in a Massive Open Online Course platform", PLoS ONE 16(1), doi: https://doi.org/10.1371/journal.pone.0245718 (2021).

[4] Steffens, K., "Competences, Learning Theories and MOOCs: Recent Developments in Lifelong Learning", European Journal of Education, 1(50), doi: 10.1111/ejed.12102 (2015).

[5] Srairi, S., "An analysis of factors affecting student dropout: the case of Tunisian Universities", International Journal of Educational Reform, 2(31), pp. 168-186, doi: 10.1177/10567879211023123 (2022).

[6] Mnyawami, Y. N., Maziku, H. H. and Mushi, J. C., "Enhanced model for predicting student dropouts in developing countries using automated machine learning approach: a case of Tanzanian's secondary schools", Applied Artificial Intelligence, 1(36), doi: 10.1080/08839514.2022.2071406 (2022).

[7] Rajkumar, R. and Ganapathy, V., "Bio-inspiring learning style chatbot inventory using brain computing interface to increase the efficiency of e-learning", Digital Object Identifier, 8(2020), 67377-67395 (2020).

[8] Jin, C., "Dropout prediction model in MOOC based on clickstream data and student sample weight", Soft Computing, doi: https://doi.org/10.1007/s00500-021-05795-1 (2021).

[9] Cannistra, M. et al," Early-predicting dropout of university students: an application of innovative multilevel machine learning and statistical techniques", Studies in Higher Education, pp. 1935-1956, doi: 10.1080/03075079.2021.2018415 (2022).

[10] Kabathova, J. and Drlik, M., "Towards predicting student's dropout in university courses using different machine learning techniques", Applied Science, 11(2021), doi: https://doi.org/10.3390/app11073130 (2021).

[11] Tan, M. and Shao, P., "Prediction of student dropout in e-learning program through the use of machine learning method", iJET, 1(10), pp. 11-17, doi: http://dx.doi.org/10.3991/ijet.v10i1.4189 (2015).

[12] Bujang, S. D. A., Selamat, A. and Krejcar, O., "A predictive analytics model for students grade prediction by supervised machine learning", Materials Science and Engineering, doi:10.1088/1757-899X/1051/1/012005 (2021).

[13] Nurmalitasari, Long, Z. A. and Noor, M. F. M., "Factors influencing dropout students in higher education", Educational Research International, doi: https://doi.org/10.1155/2023/7704142 (2023).

[14] Contini, D. and Zotti, R., "Do financial conditions play a role in university dropout? New evidence from administrative data", in Teaching, Research and Academic Careers, Contini, D., Jappelli, T. and Uricchio, A., Eds., pp. 39-70, Springer, Cham (2022).

[15] Behr, A. et al., "Early prediction of university dropouts – A random forest approach", Journal of Economics and Statistics, pp. 743-789, doi: https://doi.org/10.1515/jbnst-2019-0006 (2019).

[16] Lee, Y., Choi, J. and Kim, T., "Discriminating factors between completers of and dropouts from online learning courses", British Journal of Educational Technology, pp. 328-337, doi:10.1111/j.1467-8535.2012. 01306.x (2013).

[17] Zhang, Q. et al., "Exploring the communication preferences of MOOC learners and the value of preference-based groups: is grouping enough?", Education Tech Research Dev, pp. 809-837, doi: 10.1007/s11423-016-09439-4 (2016).