Summer 7-5-2020

Dropout Predictions of Ideological and Political MOOC Learners Based on Big Data

Yan Zhang
School of Marxism, China University of Geosciences, Wuhan,430074, China

Qian Zhang
Institute of Higher Education, China University of Geosciences, Wuhan,430074, China

Xu Liu
School of Marxism, China University of Geosciences, Wuhan,430074, China;Institute of Higher Education, China University of Geosciences, Wuhan,430074, China, xu.liu@cug.edu.cn

Dropout Predictions of Ideological and Political MOOC Learners Based on Big Data

Yan Zhang[1], *Qian Zhang* [2], *Xu Liu*[21]
[1]School of Marxism, China University of Geosciences, Wuhan,430074, China
[2]Institute of Higher Education, China University of Geosciences, Wuhan,430074, China

Abstract: The massive open online course (MOOC) has expanded rapidly, providing users with a low-cost, high-quality learning experience. High dropout rate is a serious obstacle that restricts the development of ideological and political MOOC. One of the ways to solve this obstacle is to use the rich data resources in MOOC to explore the relevant factors of dropout. Reduce dropout rates by building drop-out prediction models and establishing early-warning mechanisms. However, the ideological MOOC data is huge and complex, which is prone to problems such as loss of data value, mismatch between data and models, and poor research reproducibility. This paper uses a more mature logistic regression method of machine learning to transfer it to the field of education, providing a new path for data-driven MOOC dropout prediction research.

Keywords: MOOC, learner, behavior prediction, dropout

1. INTRODUCTION

Modern information technology has promoted the reform and development of education informatization. Massive Open Online Courses (MOOC) is a new teaching model emerged at the historic moment in the information age under the network environment [1]. The birth and development of MOOC has lowered the education threshold and promoted knowledge sharing [2]. Compared with traditional classroom education, MOOC has almost no registration threshold and the cost of dropping out is very low. Behind the huge number of users is a generally high dropout rate, of which the dropout rate for ideological and political courses is generally higher than 90% [3]. The high dropout rate severely restricts the development of ideological and political MOOCs. At present, some researchers use the increasingly mature learning behavior analysis technology to quantitatively study the MOOC dropout problem. Based on the existing behavior data of learners, analyze the behavior patterns of dropouts and non-dropouts, and predict possible dropouts. Some progress has been made [4].

How to use the MOOC system to record a wealth of learning process data, as a basis for early warning, intervention, optimization of the learning process of the MOOC, to help the MOOC to better develop and play its due role, it is particularly important [5]. However, the increasing volume of data and the complexity of records also make data processing and analysis more difficult. With the development of technology, machine learning is gradually applied to learning behavior analysis [6]. This study uses machine learning methods to transform raw data into meaningful feature data to improve the operability and interpretability of data analysis and modeling, and to explore the complementary role of human researchers and machine learning techniques in MOOC dropout prediction analysis [7].

2. MOOC DROPOUT FACTORS

Compared with the traditional teaching model, MOOC is considered to lack self-efficacy, and poor self-regulation is the main reason for MOOC learners to drop out of school [8]. In addition, study time, course setting, and course difficulty are also the main reasons for learners to drop out of class [9]. Some scholars have

[1] Corresponding author Email: xu.liu@cug.edu.cn

pointed out that teacher guidance and credit encouragement have significantly improved the completion rate of MOOC. Students with higher learning expectations and better grades are more likely to persist in learning [10]. Based on questionnaire surveys, Xu Zhenguo and others found that lack of self-control, lack of encouragement mechanisms, and regulatory measures were the main factors for school dropout [11]. The platform factor is also an important factor affecting learners, such as poor platform supervision measures, inconvenient interaction between teachers and students, etc., will affect the course completion rate [12].

Table1. Common MOOC dropout factors

Learner factors	Teacher factors	Curriculum factors	Platform factors
Lack of self-control	Lack of interactive Q & A	Long course period	Lack of incentives
Time is limited	Boring teaching	Content is not as expected	Lack of regulatory measures
Lack of motivation to complete	Poor language expression	Stale content or low quality	Inconvenient teacher-student interaction
Can't keep up with learning progress	Poor teacher image	Teaching focus is not prominent	High certificate fees

3. LOGICAL REGRESSION PREDICTION MODEL

Logical regression originates from the linear regression mathematical model in mathematical statistics, and is the most widely used classification algorithm model, which is often used to solve two classification problems. Nagrecha uses LR to predict dropouts, which is good. The main idea is to use the Sigmoid function to calculate the probability of a sample x belonging to class 1 h(x) [13].

Make $z = x^{\theta}$,e is the sample input, θ is the model parameter,

$$g(z) = \frac{1}{1 + e^{-z}}$$

The model output is:

$$h = \frac{1}{1 + e^{-x\theta}}$$

The value of h (x) is between [0,1]. The larger the value, the higher the probability that the sample belongs to class 1. The smaller the value, the higher the probability that the sample belongs to class 0. This model was chosen mainly to realize the LR by using the Logistic Regression function of the free open source resource sklearn.linear-model library in Python.

4. PREDICTION of DROPOUT RATE IN IDIELOGICAL MOOC

4.1 Study design

Learning is a complicated process. Under the MOOC environment, a large number of learners form a more complex learning system, with rich but difficult to use data resources and urgent demand for dropout prediction [14]. The ideological MOOC dropout prediction framework includes four modules: data acquisition, feature data selection, predictive modeling, and result analysis.

4.1.1 Data acquisition

In data-driven education research, data acquisition and storage are the primary tasks [15]. In terms of data acquisition, the MOOC learner behavior data used in this study mainly comes from the MOOC platform of Chinese universities. MOOC log data is often in JSON format, which needs to be parsed and dumped into tabular data. In addition, the study needs to supplement the course cycle, videos and number of tasks. And other metadata to explore feature weights and predictive effects in different curriculum contexts [16].

4.1.2 Feature data selection

MOOC learner behavior data is heterogeneous and complex. It needs to combine learning scientific theories to extract data features from various aspects of certain behaviors. It also needs data cleaning, feature transformation, featureless dimensioning, and samples Balance and other processing [17].

4.1.3 Predictive modeling

Predictive modeling needs to compile the data into training and test sets. In order to effectively reduce the possible bias and imbalance when randomly splitting the sample data, a ten-fold cross-validation method is used to generate the training and test sets.

4.1.4 Result analysis

Here is a simple and effective, explanatory traditional machine learning classifier LR for the practice of thinking MOOC drop-out prediction practice, monitoring the number of learners visiting course content, the number of videos viewed, the number of tasks submitted, the number of visits to post content, and the impact of the total number of active days in drop-out rates[18].

4.2 Data source

This research data is derived from the "School Online" platform, which is a future-oriented Internet University initiated by Tsinghua University, bringing together more than 600 well-known institutions around the world to provide learners with a full range of online education services ranging from prestigious school courses and academic degrees to practical skills. The subjects chose the course of introduction to the basic principles of Marxism, and the courses in which the evaluation scores were in the top 5 were selected for analysis. These courses are week-by-week and include videos, courseware, tests, assignments, discussions, and final exams. These courses were chosen because they were a quality course with earlier and more iterations, teaching design was recognized by peers, the quality of teaching was recognized by the students, and the "cliff-cut" drop-out caused by the quality of the course itself could be effectively avoided, the course was dominated by video, and the unit tasks included tests, assignments and discussions. The teaching elements are fully equipped, the course data is large, and the learners are active, the data quality is relatively good, suitable as an exploratory predictive modeling sample.

4.3 Feature selection

In the learning environment, MOOC learners interact with a series of learning content or learning tools to produce specific learning behaviors, obtain corresponding learning effects, and then form individual learning trajectories. We select corresponding features from them for modeling.

4.3.1 Learning environmental characteristics

MOOC online learning environment mainly includes social environment and network environment. Social environment includes group environment, organization environment, and family environment. The network environment is mainly hardware or software that supports learners' smooth learning, which is divided into infrastructure, learning resources and teaching platforms. This study is based on the same courses on the same learning platform, and it can be considered that the network facilities, the level of resource construction, and the nature of the courses are the same. These are not the reasons that cause learners to show different behavior patterns. Therefore, this study only considers the group environment in the social environment.

The group environment is specifically the learning participation of teachers and other learners in the learning community. It mainly includes teachers' MOOC teaching. Teachers usually measure their participation in the learning process from two aspects: teaching guidance and evaluation. Other learners in the MOOC teaching, the consideration of other learners is mainly the learning atmosphere of the learning group, and the

interaction between the learning members and learners in the group as a measure of participation.

Table2. Learning environmental characteristic indicators

Group environment	Index
Teacher（T）	Teacher guidance times（T1）
	Number of teacher evaluations（T2）
Student（S）	Class average participation in discussions（S1）
	Class average postings（S2）

4.3.2 Learning behavior characteristics

MOOC learning activities are based on a series of interactions with learning content or learning tools that occur in MOOC courses, which correspond to the construction of different cognitive levels of learners. Therefore, we divide a series of learning behaviors from cognitive participation. The first behavior is environmental interaction behavior. Learners understand the behavior of the learning environment, including browsing teaching dynamics, browsing teaching resources, etc., with a low level of cognitive participation. The second type of behavior is learning interactive behaviors. Everyone in the learning community conducts learning activities, including reading posts, finding information, establishing learning groups, etc., involving some collaborative behaviors of learners. The degree of cognitive participation is higher than that of operational interaction. The third behavior is interactive behavior, which focuses on negotiation, communication and sharing, such as creating discussion topics, asking and answering questions, sharing and recommending learning resources, and summing up learning gains. The required cognitive participation is the highest.

Table3. Learning behavioral indictors

Learning behavior	Index
Environmental interaction	Number of signed-in courses（E1）
	View course progress times（E2）
	Number of browse course resources（E3）
Learn interactive behavior	Number of videos watched（L1）
	Watch the duration of the video（L2）
	Views of forum posts（L3）
Interactive behavior	Participation in discussions（I1）
	Questions(I2)
	Posts(I3)
	Number of replies(I4)

4.3 Analysis of results

The sample was randomly divided into 10 parts by the ten-fold cross-validation method. Each time, one of the samples was used as the test set, and the remaining nine were used as the training set. The logistic regression algorithm was used to train the prediction model. This study uses test set data as input data to predict whether learners will eventually complete the course. Because the dependent variable Y of both models is a final grade with a range of 0-100, in order to facilitate the analysis of the accuracy of subsequent predictions, this study uniformly defines the final grade less than 60 as "dropout", and the final grade is greater than or A score equal to 60 is defined as “passed the course”, and the numerical proportion of the results of “predicted dropout” and "predicted pass" in "real dropout" and "real pass".

Table4. Test results of 5 courses

Course code	C1	C2	C3	C4	C5
Number of people who passed the test	448	816	749	1422	3356
Number of dropouts in the test set	4346	2883	2953	3730	6259
Actual to passed and predicted to pass	65.7%	83.7%	75.1%	39.7%	42.3%
Actual dropouts and predicted dropouts	87.6%	86.6%	88.2%	81.7%	83.8%

The average accuracy rate of the LR model in correctly "predicting dropouts" (that is, actual dropouts and predicted dropouts) is 85.6% on average. This is largely related to the usual learning habits of dropout learners. Generally, most learners drop out because they have less learning participation and less learning behavior, and their behavior patterns are relatively simple and easier to predict. The analysis of the accuracy of the correct "prediction pass" found that the accuracy rate of the 5 courses is high or low (up to 83.7%, and the lowest is only 39.7%)-this aspect has more behaviors with learners who have passed the course. The model is more complicated, and on the other hand it is related to the sparse number of people who pass the course in the data set.

5. CONCLUSIONS

In the MOOC environment, the number of learners is much larger than teachers and education administrators. This makes the “human” resources that provide learning support for learners relatively scarce. It is impossible to rely on manual tasks to identify problem learners and provide targeted timely intervention.[19] Inadequate intervention sits unable to cope with the serious problem of dropout, and unnecessary intervention is a waste of resources and more likely to affect learners who are learning normally. Dropout prediction can help teachers and education administrators monitor learners’ dropout risk, accurately identify which learners need help, analyze what problems learners have, or what factors contribute to learning risk, in order to select the right resources and methods for timely intervention [20]. This study helps:

5.1 Categorize learners according to the prediction results

According to the forecast results, learners are divided into two categories. Learners who have a high risk of dropping out and need help can be divided into learners who have left the course and learners who have not yet quit but have shown problematic behavior according to the disengagement node. The former needs to use means other than MOOC, such as email to remind them to continue the class. The latter needs to be intervened during the MOOC learning process. Learners who are expected to adhere to the course content, and these learners can provide good guidelines for learning behaviors and provide suggestions for radically improving MOOC teaching.

5.2 Provides learner behavior analysis

This study found that dropouts and non-dropouts have different learning behavior characteristics when they visit pages, watch videos, submit tasks, and collaborate on forums. Based on the learning behavior patterns of non-dropouts, it is able to identify the problem behaviors of dropouts and find that their dropout factors, such as insufficient learning investment, such as low video and task completion rates, lack of self-control, and activity gradually decline over time .Base on these learning difficulties to develop targeted interventions.

5.3 Advise on data-based support for improving ideological and political MOOC

Early warning system can be constructed on the basis of predictions of dropouts, feedback of learning risk levels to learners in the form of “signals”, and the prediction results are presented to teachers or teaching assistants in a visual form, and teaching improvements or learning are provided Intervention recommendations, manual intervention by teachers or teaching assistants combined with educational theory.

REFERENCES

[1] Kennedy, J.(2014).Characteristics of Massive Open Online Courses(MOOCs) : A Research Review,2009-2012. Journal of Interactive Online Learning. (13): 1.

[2] K Park, MC Nguyen, H Won.(2015).Web-based collaborative big data analytics on big data as a service platform. In Advanced Communication Technology. 564-567.

[3] Xie Luyan, Zheng Mingjiu. (2015).Using MOOC Platform to Construct a New Network Teaching Model of Ideological and Political Education Theory Courses in Colleges and Universities. Research in Ideological Education. (9):61.

[4] Rai L,& Chunrao D.(2016).Influencing Factors of Success and Failure in MOOC and General Analysis of Learner Behavior. International Journal of Information and Education Technology. 6(4):262.

[5] Li Liang. (2014).MOOC and the Innovation of teaching Mode of Ideological and Political Theory. Ideological Theory Education. (01)：65.

[6] Hao Siyuan, Xie Taifeng.(2019).Application of machine learning methods in the prediction of MOOC learners' academic completion rate. The Practice and Understanding of Mathematics. (21)：85.

[7] Li Hailin.(2014).The teaching exploration of data mining course in the environment of big data. The Computer Age. (02)：54-55.

[8] Li Manli,Zhang Yu,Ye Guifu.(2013).Decoding MOOC: An Educational Survey of Large-Scale Online Open Courses .Beijing: Tsinghua University Press.36-37.

[9] Wang Xueyu, Zou Gang, Li Xiao.(2017). Study on the prediction of learner dropout based on MOOC data. Modern educational technology. (6)：98.

[10] Zhang Qianfan,Wang Chengyu, Zhang Yajun. (2015).An Empirical Study of the Factors Affecting College Students' MOOC Learning Intention. Higher Education Exploration. (8):66-70.

[11] Xu Zhenguo.(2017). Research on the Influencing Factors of MOOC Learners' Dropout Behavior in Modern Educational Technology. Modern educational technology. (27)：100.

[12] Gao Di.(2014). Hot and Cold Thinking of MOOC—Reviewing Six Major Issues in Teaching MOOCs in the World. Distance education journal. (2):39-47.

[13] Nagrecha S, Dillon J Z, Chawla N V. (2017). MOOC Dropout Prediction: Lessons Learned from Making Pipelines Interpretable. International Conference on World Wide Web Companion: 351-359.

[14] Jiang Lin, Han Xibin, Chen Jiangang.(2013). Study on the Characteristics of MOOCs Learners and Their Learning Effects. China Electrification Education. (11):54-59.

[15] Wang Yanming.(2015). Teaching Reform in the Age of MOOCs: A Rational Thinking. Educational Science Research. (07)：59-64.

[16] Zhang Zhe, Wang Yining, Chen Xiaohui.(2016). MOOC Empirical Study on Influencing Factors of Intention of Continuous Learning—Based on Improved Expectation Confirmation Model. Research on electro-education. (5):30-36.

[17] Wang Guoyin, Liu Qun, Yu Hong.(2017). Big Data Mining and Application. Beijing:Tsinghua University Press.

[18] Ruochen Li.(2019). Behavioral Analysis and Dropout Prediction of MOOC Learners Based on Feature Engineering. M Thesis. Shanghai:East China Normal University.

[19] Xu Meiling, Zhang Lihua, Guo Bu.(2019). Research and Practice of Educational Big Data Mining Based on MOOC. China Information Technology Education. (05)：101.

[20] Yan Beibei.(2014). Application of MOOC in the Teaching of Ideological and Political Theory in Colleges. Party building and ideological education in schools. (04):59.