SI 671: Data Mining
Dissecting Dropout Behavior in MOOCs

Andy Chen
techen@umich.edu

December 14, 2022

Abstract

Who, when, and more importantly—why do learners drop out of MOOCs? In this project, I attempted to dissect and answer these questions using several techniques. First, I noticed that there exist subtle differences in user characteristics and actions between dropouts and non-dropouts. Next, I further uncovered how action sequences impact users' odds of dropping out by implementing GLM and frequent sequence mining algorithms. Finally, using an optimized XGBoost model, I was able to predict potential dropouts with an F1 score of 0.901. These findings lay the foundation for reducing dropout rates and improving learning experiences in future MOOC courses.

1 Introduction

In the past two years, Massive open online courses (MOOCs) have witnessed rapid expansion across the globe as a byproduct of the pandemic. This growth does not appear to be ending soon, as the global MOOC market is projected to grow at an astounding 34.54% CAGR through 2027 (Research and Markets, 2022). However, our current understanding of learners' learning behavior in MOOCs is still limited despite collecting large amounts of data (Reich, 2022), hence the need for more research in this field.

Among learners' various behaviors in MOOCs, *dropping out*, i.e., failing to complete courses, is probably one of the most alarming yet intriguing. Researchers around the world have been approaching the dropout phenomenon from numerous perspectives (Goopio & Cheung, 2020), and in this report, I will consider all of them with the aim of better explaining dropouts. Who drops out? When do they drop out? And most importantly, what leads to dropout?

2 Data & Methodology

2.1 Datasets

The data I used for this report is collected from a MOOC platform initiated by Tsinghua University, XuetangX[1]. Founded in 2013, XuetangX is the first Chinese MOOC platform

[1]All datasets used in this report are publicly available at http://www.moocdata.cn.

and boasts over 50 million registered learners as of April 2022 (Wikipedia contributors, 2022). The analysis primarily surrounds the dropout prediction dataset (see Table 1), which consists of learners' activity logs from June 2015 to June 2017, as well as the enrollment and dropout records for each learner. Additional features are also extracted from the user profile (see Table 2) and course information (see Table 3) datasets to further characterize each enrollment.

Table 1: Dropout Prediction Dataset

Column	**Description**
enroll_id	The id of (user, course) pair
username	The id of the user
course_id	The id of the course
session_id	The id of the session
action	The type of user activity
object	The corresponding object of the action
time	The occurrence time of the action
truth	The label of user's dropout (1: dropout; 0: non-dropout)

Table 2: User Profile Dataset

Column	**Description**
user_id	The id of the user
gender	The gender of the user
education	The user's education level
birth	The user's birth year

Table 3: Course Information Dataset

Column	**Description**
id	The id of the course
start	The start time of the course
end	The end time of the course
course_type	The course mode (0: instructor-paced; 1: self-paced)
category	The category of the course

2.2 Related Work

Past quantitative research on dropouts has predominantly focused on predicting their occurrences via machine learning models. While Feng et al. (2019) designed and implemented multi-layered neural networks, others went with more traditional models such as SVM (Kloft et al., 2014). Despite the difference in modeling techniques, they all went into

detail about the model tuning and optimization process.

2.3 Objectives & Methodology

Instead of fixating on model performance and accuracy, I hope to focus on discovering patterns and explanations for dropouts. Throughout my analysis, I will attempt to address the following research questions:

1. What characteristics do these dropouts share?
2. Can we discern any patterns in their learning behaviors?
3. How might we identify learners that are potential dropouts?

3 Analysis & Results [2]

3.1 First Look at Dropouts

After mapping the user and course info datasets to the activity log, the merged dataset contains 157,943 enrollments in total. I further examined the distribution between dropouts (labeled as 1) and non-dropouts (labeled as 0). Detailed statistics are listed in Table 4.

Table 4: Summary Statistics of Datasets

Measure	**Value**
Total Enrollments	157,943 (1: 75.9%; 0: 24.1%)
Unique Users (Labeled)	69,823
Unique Courses (Labeled)	1,454
Gender (F/M)	1: 65%/35%; 0: 68%/32%
Age	1: 27.5; 0: 27.9
Average Total Actions	1: 93; 0: 472

One of the most intriguing observations comes from their education levels (see Figure 1). Surprisingly, more doctorate learners are dropouts rather than non-dropouts, whereas high school graduates actually have the lowest dropout-to-non-dropout ratio among all education levels (see Figure 2).

Aside from user characteristics, we can also get a clearer picture of users' behaviors by grouping their actions into the following action groups:

[2] Full analysis and code can be found at `https://github.com/andy-techen/mooc-dropouts`.

Figure 1: Education Level Among Dropouts

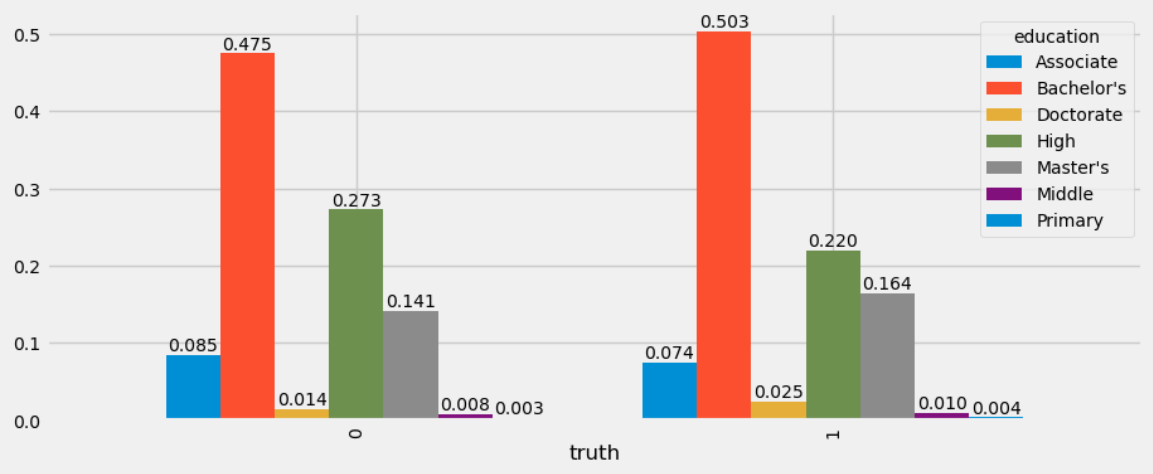
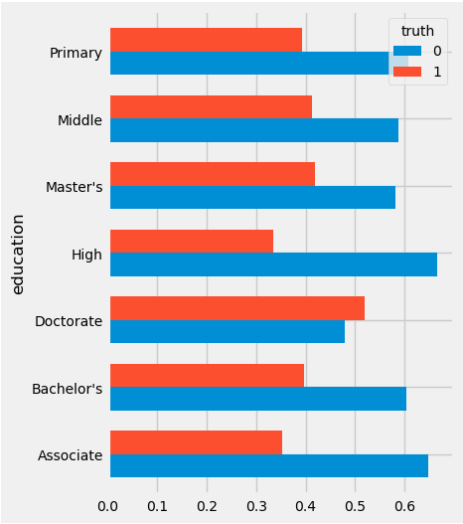


Figure 2: Education Level Among Dropouts (Cont.)

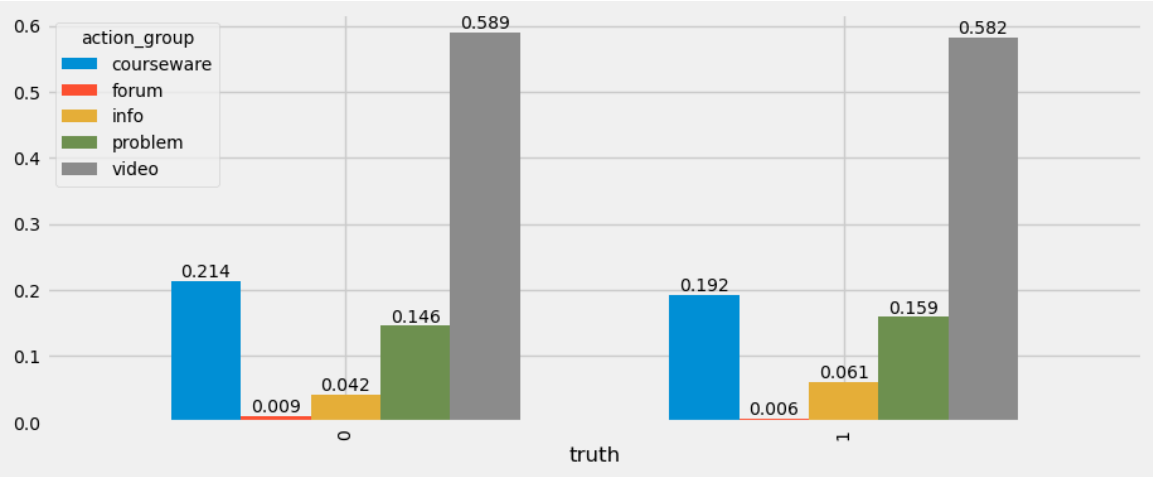


- Video: interacting with the video (e.g., load, play, stop...)
- Courseware: opening and closing the course
- Problem: tackling problems and checking answers
- Info: checking the course info and progress
- Forum: creating threads and comments in the forum

While there is some dissimilarity between the two groups (see Figure 3), the difference is not apparent at first glance. That being said, we already know that the average number of actions differs significantly between them (see Table 4), so next, I will investigate deeper

to uncover the cause for this variance.

Figure 3: Action Distribution Among Dropouts



3.2 Understanding Dropouts

3.2.1 Feature Engineering

To better encompass and explain the variance between enrollments, I derived the following features from our preexisting datasets:

- Action frequency: Frequency of each action for each enrollment. I also tried using the action groups as features, but they were less effective than the ungrouped features
- Action occurrences: First and last actions of each enrollment
- Session duration: Total number of sessions and average session duration (seconds) of each enrollment
- User attributes: Education level, age, and gender of each user
- Course attributes: Course category (e.g., engineering, math, history)

3.2.2 GLM & Interpretation

With the features in place, we can fit a GLM model to have a closer look at them. While the model fit is suboptimal (Pseudo R-square of 0.2419), we can spot a few intriguing features from the summary report (Table 5) by sorting them according to their coefficients.

It appears that a few course categories tend to have more dropouts, and on the other hand, a higher number of sessions might be an indicator of non-dropouts (odds of dropping out decrease by 0.74 for every additional session). Users also have lower odds of dropping out if their first action was 'click_progress' or 'close_courseware'. This then raises the question: Does the sequence of actions matter?

Table 5: GLM Summary Report

No. Observations: 42631	Model Family: Binomial
Df Model: 82	Pseudo R-squ. (CS): 0.2419

Feature	**Coef**	**P > |z|**
category_foreign language	1.188387	1.362914e-75
category_computer	0.964839	1.024122e-69
...		
first_action_close_courseware	-0.207932	1.239197e-02
sessions	-0.296089	3.189101e-270
first_action_click_progress	-1.829305	1.591993e-03

3.3 Examining Action Sequences

3.3.1 Weighted Random Sampling

Our entire activity log data has 29,165,540 rows, and while some enrollments only have one single action, some have up to 120,000 (see Table 6) actions. To further examine the action sequences, we will need to obtain action samples from each enrollment. Here, I implemented the weighted random sampling algorithm proposed by PS Efraimidis (Efraimidis, 2015) (see Figure 4) with an m of 138, which is the third quartile of actions per enrollment.

Table 6: Actions Per Enrollment

Measure	**Value**
min	1.0
25%	8.0
50%	29.0
75%	138.0
max	128992.0

Figure 4: Efraimidis and Spirakis Weighted Random Sampling

Input: A population V of n weighted items
Output: A weighted random sample of size m
1: For each $v_i \in V$, $u_i = random(0, 1)$ and $k_i = u_i^{1/w_i}$
2: Select the m items with the largest keys k_i

3.3.2 Sequence Pattern Mining

After sampling 138 actions from each enrollment, we now have a more reasonable size of events to work with. Next, I discovered a Python package [3] that implements PrefixSpan (Pei et al., 2002), a frequent sequence pattern mining algorithm, to help us compute frequent patterns in our action sequences.

If we compare the most frequent sequence patterns of dropouts and non-dropouts (see Table 7), unfortunately, the difference between them is not apparent. However, it should be noted that although the patterns are mostly similar, 3-item sequences are much more frequent among non-dropouts, which makes sense since non-dropouts record much more actions on average.

Table 7: Frequent Sequence Patterns (Dropouts vs. Non-Dropouts)

Rank	Dropouts (119,817 users)		Non-Dropouts (38,126 users)	
	Sequence Pattern	**Frequency**	**Sequence Pattern**	**Frequency**
1	load_video, load_video	70,745	load_video, load_video	31,757
2	pause_video, load_video	70,662	load_video, pause_video	31,094
3	load_video, close_courseware	65,567	close_courseware, load_video	30,840
4	pause_video, pause_video	64,182	pause_video, close_courseware	29,643
5	click_courseware, click_courseware	60,842	load_video, load_video, load_video	29,244
6	click_about, click_about	60,814	pause_video, load_video, load_video	29,113

3.4 Predicting Dropouts

3.4.1 Model Selection

Besides exploring the dropouts' characteristics and their action sequences, it would also be helpful if we could identify them ahead of time. The user and course info data can only be mapped to one-fourth of the enrollments, so I decided to fit the models using both 1) only the behavioral data and 2) all relevant features, including user and course attributes.

[3] Package details can be found at `https://github.com/chuanconggao/PrefixSpan-py`

The models I selected are Logistic Regression, Decision Tree, Multilayer Perceptron, Random Forest, and XGBoost. Logistic Regression and Decision Tree models serve as good baseline models, while XGBoost generally performs well on medium-sized datasets like the one we have.

3.4.2 Hyperparameter Tuning

After cross-validating all five models, XGBoost generated the best accuracy and F1 results (see Table 8). The models all performed better when only using the behavioral features, but this could be attributed to the high missingness of the user and course info datasets.

Then, I optimized the XGBoost hyperparameters using a grid search, obtaining a slightly improved F1 score of 0.902 on the validation set. The parameter set unfortunately had to be limited due to computational constraints. The selected and optimized hyperparameters are listed in Table 9.

Table 8: Model Performance Comparison (Cross Validation)

Model	**Accuracy**	**F1**
XGBoost	**0.843114**	**0.901305**
Random Forest	0.837986	0.897105
Decision Tree	0.765441	0.844923
Multilayer Perceptron	0.736709	0.815960
Logistic Regression	0.615850	0.677881

Table 9: XGBoost Optimized Hyperparameters

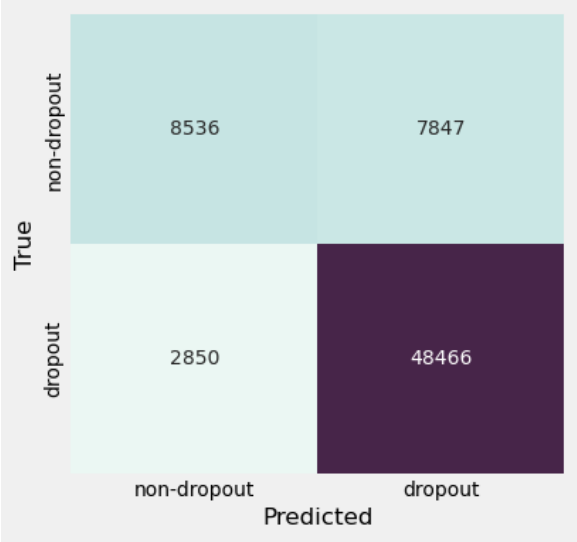
Hyperparameter	**Value**
gamma	0.1
max_depth	5
min_child_weight	5

3.4.3 Model Performance

Using the finalized XGBoost model, I was able to obtain an accuracy of 0.842 and an F1 score of 0.901 (see Figure 5) on the held-out training set. The results are only slightly lower than the best F1 score in previous works (F1 score of 0.9095). Among the features, it should be noted that the number of sessions has a feature importance of 0.5, which is by far the highest. This aligns with our previous observations.

Using this prediction model, incentives or notifications can be sent to the predicted dropouts

Figure 5: Confusion Matrix of Final Model Predictions



based on their preferences and characteristics. In the long run, this mechanism could potentially reduce the dropout rate and improve learners' learning experiences.

4 Future Work & Discussion

4.1 Data & Feature Completeness

The user and course info datasets are fairly limited and contain lots of data that is either missing or erroneous, making it difficult to construct robust and accurate explanatory models. This should be a priority for improvement in future research.

4.2 Reservoir Sampling Improvements

My current implementation of weighted random sampling is flawed in the sense that it does not prioritize more recent events, and I was not able to find off-the-shelf packages for this particular use case. I will need to iterate through some revisions to ensure the sampling algorithm works correctly and efficiently.

4.3 Time-Related Attributes

In my current analysis, I only evaluated the datasets as sequences. Furthermore, non of the features in the final model considered the time-related attributes apart from session duration. Moving forward, I should explore methods to better integrate the sequential

and temporal nature of learner actions into the models.

4.4 Alternative Models & Neural Networks

I initially considered testing out SVM, RNN, and other more complex neural networks for the prediction model, but I eventually chose to leave them out due to computational and time constraints. This could be another place for improvement in the future.

References

Efraimidis, P. S. (2015). Weighted Random Sampling over Data Streams. *Algorithms, Probability, Networks, and Games*, 183–195. doi: 10.1007/978-3-319-24024-4_12

Feng, W., Tang, J., Liu, T. X., Zhang, S., & Guan, J. (2019). Understanding Dropouts in MOOCs. *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*.

Goopio, J., & Cheung, C. (2020). The MOOC Dropout Phenomenon and Retention Strategies. *Journal of Teaching in Travel & Tourism*, *21*(2), 177–197. doi: 10.1080/15313220.2020.1809050

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC Dropout over Weeks Using Machine Learning Methods. *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*. doi: 10.3115/v1/w14-4111

Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., & Hsu, M.-C. (2002). Prefixspan,: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. *Proceedings 17th International Conference on Data Engineering*. doi: 10.1109/icde.2001.914830

Reich, J. (2022). Terabytes of Data, Little New Insight. In *Failure to Disrupt: Why Technology Alone Can't Transform Education* (p. 38–40). Harvard University Press.

Research and Markets. (2022, Aug). *Global Massive Open Online Courses Market Report (2022 to 2027) - Featuring Coursera, edX, XuetangX and Futurelearn Among Others.* Retrieved from https://www.globenewswire.com/en/news-release/2022/08/22/2501997/28124/en/Global-Massive-Open-Online-Courses-Market-Report-2022-to-2027-Featuring-Coursera-edX-XuetangX-and-FutureLearn-Among-Others.html

Wikipedia contributors. (2022, Jun). *Xuetangx — Wikipedia, the free encyclopedia.* Retrieved 2022-11-01, from https://en.wikipedia.org/wiki/XuetangX