Dissecting Dropout Behavior in MOOCs

School of Information, MSI
Andy Chen

Motivation & Preface

- The global MOOC market is projected to grow at an **34.54%** CAGR through 2027
- Our current understanding of learners' learning behavior in MOOCs is still limited despite collecting large amounts of data
- **Dropping out**, i.e., failing to complete courses, is one of the most alarming yet intriguing user behaviors
- Data collected from a MOOC platform initiated by Tsinghua University, **XuetangX**. Datasets include learners' activity log, user profile, and course information

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?

First Look at Dropouts

User Profile

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

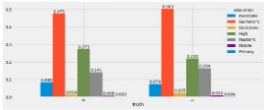
User Behavior

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

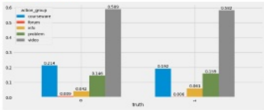
Total Actions (avg) - 1: 93; 0: 472

➡ Further investigate action patterns among dropouts

Education Level



User Action



Truth: 1 - dropout; 0 - non-dropout

Understanding Dropouts

Feature Engineering

Action Frequency: frequency of each action (action group less effective) for each enrollment.
Action Occurences: first and last action of each enrollment.
Session Duration: number of sessions and average session duration of each enrollment.
User Attributes: education level, age, and gender of each user.
Course Attributes: course category (e.g., engineering, math, history)

GLM & Interpretation

With the features in place, I fit a GLM model to have a closer look at them. While the model fit is suboptimal, we can spot a few intriguing features from the summary report.

It appears that a few course categories tend to have more dropouts, and on the other hand, some first actions and a higher number of sessions (1 session, odds * 0.74) might be indicators of non-dropouts.

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

feature	coef	P>|z|
category_foreign language	1.188387	1.367914e-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	-1.829385	1.591993e-03

Examining Action Sequences

Weighted Random Sampling (PS Efraimidis)

Our entire activity log data has 29,165,540 rows, so while some enrollments only have one single action, some have up to 120,000 actions. To further examine the action data stream, we'll need to obtain action samples for each enrollment. Here, I implemented the weighted random sampling algorithm proposed by PS Efraimidis with an initial k of **138**, which is the third quartile of actions per enrollment.

Sequence Pattern Mining (PrefixSpan)

After sampling 138 actions from each enrollment, we now have a more reasonable size of events to work with. I found a Python package that implements **PrefixSpan**, a frequent sequence pattern mining algorithm, to help us discover patterns in our action sequences.

If we compare the frequent sequence patterns of dropouts and non-dropouts, the difference is not apparent.

Dropout	Non-Dropout
['load_video', 'pause_video'] ['load_video', 'close_courseware'] ...	
['click_courseware', 'click_courseware'] ['click_about', 'click_about']	Mostly similar, but **3-item sequences** are much more frequent

Predicting Dropouts

Model Selection

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. The models I selected are Logistic Regression, Decision Tree, Multilayer Perceptron, Random Forest, and XGBoost.

Hyperparameter Tuning

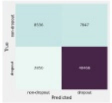
After cross validating all five models, **XGBoost** generated the best accuracy and F1 results. I further optimized the XGBoost hyperparameters using GridSearchCV and obtained an average F1 of **0.902** on the validation set. The param set unfortunately had to be limited due to computational constraints.

- **'min_child_weight':** [1, 3, 5]
- **'gamma':** [0.1, 0.5]
- **'max_depth':** [3, 5]

Model	Accuracy	F1
XGB	**0.843114**	**0.901305**
RF	0.837986	0.897105
DT	0.765441	0.844973
MLP	0.736709	0.815960
LR	0.615850	0.677881

Model Performance

Using the finalized XGBoost model, I obtained an accuracy of **0.842** and an F1 of **0.901** on the held-out training set. The results are just slightly lower than the best F1 score achieved in previous works (F1 of 0.9095). Incentives or notifications can be sent to these potential dropouts and potentially reducing the dropout rate. Among the features, the **number of sessions** has a feature importance of 0.5, which is by far the highest.



Discussion & Future Work

More Complete Data & Exhaustive Features

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

Revise Reservoir Sampling

My implementation of weighted random sampling is flawed in a sense that it doesn't prioritize more recent events, and I couldn't find off-the-shelf packages for this. I'll need to make some revisions to ensure everything works as planned.

Consider Time-Related Attributes

Currently, I've only evaluated the data as a sequence, and non of the model features considered the "time" element in the activity log apart from session duration. This should be worth exploring moving forward.

Alternative Models & Neural Networks

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