Beyond Prediction: First Steps Toward Automatic Intervention in MOOC Student Dropout

Jacob Whitehill, Joseph Williams, and Justin Reich
HarvardX, Harvard University
edX and HarvardX

• HarvardX:
  • Harvard’s strategic initiative for online education.

• edX:
  • Non-profit provider of massively open online courses (MOOCs).
HarvardX

• 2013-2014:
  • 27 online courses
  • 2M registrations, 60K certificates

• Subject areas:
  • Computer programming
  • Chinese history
  • Copyright law
  • RNA sequence analysis
  • American poetry
  • Greek mythology
  • Health policy
  • ...

Copyright law
Research at HarvardX

- Student drop-out:
  - Why do many students not complete a course? What can we learn from these students about how to improve online education?
Research at HarvardX

• Student drop-out:
  • Why do many students not complete a course? What can we learn from these students about how to improve online education?
  • In pre-course surveys, 58% of responders say they intend to complete the course and earn certificate.
  • But even within this group, only 22% of responders actually do so (Reich 2014).
Research at HarvardX

• Student drop-out:

  • Why do many students not complete a course? What can we learn from these students about how to improve online education?

  • In pre-course surveys, 58% of responders say they intend to complete the course and earn certificate.

    • But even within this group, only 22% of responders actually do so (Reich 2014).

  • Room for improvement in MOOCs? In students’ learning?
Research at HarvardX

• Student stop-out:
  • Why do many students not complete a course? What can we learn from these students about how to improve online education?
  • In pre-course surveys, 58% of responders say they intend to complete the course and earn certificate.
    • But even within this group, only 22% of responders actually do so (Reich 2014).
  • Room for improvement in MOOCs? In students’ learning?
You may submit up to a paragraph in response to the following questions. Your thoughts are very important to us, and a more complete response will help us understand your experience and views.

What were your favorite aspects of this course?
Post-course survey

- Post-course survey response rate is ~2%.
- Responses mostly from students who certified.
- Too time-consuming, too long after students stopped out.
Post-course survey

• Post-course survey response rate is ~2%.
  • Responses mostly from students who certified.
  • Too time-consuming, too long after students stopped out.

• Alternative: send students a “query” email after they stop out.
  • Short
  • Timely
Hi,

It seems like you haven't been to Visualizing Japan in a while, so we wanted to see how things were going.

When were you next planning to visit it, and why?

[ ] Not in the next few months, because...
  I already got what I intended to.
  The course takes too much time.
  It was not what I expected when I signed up.
  I am not happy with the quality of the course.
  I just want to 'bookmark' the course for future reference.

[ ] In a few days, because I am going to...
  Watch the videos.
  Do the exercises.
  Participate in the discussion forum.

Follow this link to provide another answer, give feedback on the course, or say more.

Click here to opt out of future emails or paste this URL into your browser: Click here
“Stop-out” Query

- To collect stop-out feedback, we need:
  - *Classifier* of whether a student has stopped out by time $t$ based on his/her history of events.
“Stop-out” Query

• To collect stop-out feedback, we need:
  
  • *Classifier* of whether a student has stopped out by time $t$ based on his/her history of events.

  • *Controller* that uses the classifier’s output to decide whether to survey a student at time $t$. 
Definition of “stop out”

• Suppose a MOOC has certification date $d$.

• A student $s$ has stopped out by week $t$ iff:
  • $s$ does not earn a certificate; and
  • $s$ has no events between time $t$ and $(d+12$ weeks).
  • Don’t want a single event years later to preclude classification as “stop-out”.
Stop-out classifier
Prior work

- Statistical and time-series analysis of event logs:
  - Halawa, Greene, and Mitchell (2013)
  - Taylor, Veeramachaneni, and O’Reilly (2014)
Prior work

• Statistical and time-series analysis of event logs:
  • Halawa, Greene, and Mitchell (2013)
  • Taylor, Veeramachaneni, and O’Reilly (2014)
    • Crowdsourced feature engineering
Prior work

• Statistical and time-series analysis of event logs:
  • Halawa, Greene, and Mitchell (2013)
  • Taylor, Veeramachaneni, and O’Reilly (2014)
    • Crowdsourced feature engineering
• Social network analysis of MOOC forums:
  • Yang, Sinha, Adamson, and Rosé (2013)
Prior work

- Statistical and time-series analysis of event logs:
  - Halawa, Greene, and Mitchell (2013)
  - Taylor, Veeramachaneni, and O’Reilly (2014)
    - Crowdsourced feature engineering
- Social network analysis of MOOC forums:
  - Yang, Sinha, Adamson, and Rosé (2013)
Classification approach

- Decide whether a student has stopped out by time $t$:
  - How many problem solutions, forum posts, video plays up to time $t$.
  -Elapsed time since last event.
  -1-D temporally-local band-pass (Gabor) filters.
  -Still need grade data (pending)!

Course start $\rightarrow$ $t$ $\rightarrow$ Time
Classification approach

• We use multinomial logistic regression (MLR):

\[ P(Z_t = 1 \mid x_t, \phi) = \frac{1}{1 + \exp(-x_t^\top \phi)} \]

where:
- \( Z_t = 1 \) if student has stopped out by time \( t \),
- \( x_t \) is the feature vector at time \( t \), and
- \( \phi \) is the parameter vector.
Target MOOC

• Target course: USW30x 2014, “Tangible Things”

• 8 week course

• 14236 students for training+testing:
  • 1090 certified
  • 11853 stopped out
Classifier training

- Student event logs sampled every 7 days
- 52587 data points for training
- Compute $\phi$ using maximum-likelihood estimation.
Example
Example
Example

Classifier output

# days into course

Events
Example

Classifier output

Events

# days into course
Example
Example

Classifier output vs. # days into course with events.
Example

# days into course

Classifier output

Events
Accuracy

AUC = 0.82

Halawa, Greene, and Mitchell (2013)
Accuracy

AUC = 0.82

Most salient feature: time-since-last-event (AUC = 0.79).
Query controller
Control problem

- We now have a prototype classifier of student stop-out.
- How do we decide, based on the classifier’s output, *which* students to query and *when* to query them?
Trade-offs

• Trade-off 1:

  • The longer we wait, **the more confident** we become that the student has truly stopped out.
Classifier output vs. 
# weeks since stop-out

Average Classifier Output versus Weeks Since Dropout
Trade-offs

• Trade-off 1:
  • The longer we wait, **the more confident** we become that the student has truly stopped out.
  • The longer we wait, **the less likely** the student will respond to the query.
Post-course survey response rates

Figure based on post-course survey.

End-of-course survey response for stopped-out students
Trade-offs

• Trade-off 2:
  • The lower the threshold, **the more responses** we might get to the survey.
  • The lower the threshold, **the more spam** we send and reputational damage we inflict.
Problem formulation

• Once each week, the controller decides which students to query and sends a survey email.
Problem formulation

• Once each week, the controller decides which students to query and sends a survey email.

• The cost $C_{st}$ for student $s$ at week $t$ is defined as:
  
  $C_{st} = -1$ if we survey student $s$ who had stopped out and answers the survey.
  
  $C_{st} = D$ if we survey student $s$ who had not stopped out.
  
  $C_{st} = 0$ otherwise.
Problem formulation

• Once each week, the controller decides which students to query and sends a survey email.

• The cost $C_{st}$ for student $s$ at week $t$ is defined as:
  - $C_{st} = -1$ if we survey student $s$ who had stopped out and answers the survey.
  - $C_{st} = D$ if we survey student $s$ who had not stopped out.
  - $C_{st} = 0$ otherwise.

• We want to minimize:

$$f(\theta) = \mathbb{E} \left[ \sum_{s=1}^{S} \sum_{t=1}^{T} C_{st} \mid \theta \right]$$
Problem formulation

• Once each week, the controller decides which students to query and sends a survey email.

• The cost $C_{st}$ for student $s$ at week $t$ is defined as:
  $C_{st} = -1$ if we survey student $s$ who had stopped out and answers the survey.
  $C_{st} = D$ if we survey student $s$ who had not stopped out.
  $C_{st} = 0$ otherwise.

• We want to minimize:

$$f(\theta) = \mathbb{E} \left[ \sum_{s=1}^{S} \sum_{t=1}^{T} C_{st} \mid \theta \right]$$

In practice, $D$ is a “dial” to adjust response rate versus spam rate.
Control architectures

• Let $y$ denote the classifier’s output, and let $p(y, t)$ denote the probability of giving a survey at time $t$ based on $y$. 
Control architectures

• Let $y$ denote the classifier’s output, and let $p(y, t)$ denote the probability of giving a survey at time $t$ based on $y$.

• We explore two different control architectures…
Control architectures

• Controller 1 (stationary, deterministic):

\[ p_\theta(y) = \mathbb{I}[y > \theta] \]
Control architectures

- Controller 1 (stationary, deterministic):
  \[ p_\theta(y) = \mathbb{I}[y > \theta] \]

- Controller 2 (non-stationary, stochastic):
  \[ \theta = (s, b_1, \ldots, b_T) \]
  \[ p_\theta(y, t) = \sigma(sy - b_t) \]
  \[ \sigma(z) = \frac{1}{1 + \exp(-z)} \]
Control architectures

- Controllers 1 and 2:
  - To reduce the likelihood of spam disaster, we allow the controller to survey each student at most once.
Experiments

• To compare these two controllers, we need to know $P(\text{SurveyResponse} \mid \text{TimeSinceStopout})$.

• Won’t know until we run live experiment.
Experiments

• To compare these two controllers, we need to know $P(\text{SurveyResponse} \mid \text{TimeSinceStopout})$.
  • Won’t know until we run live experiment.

• But we have everything else we need:
  • Stop-out classifier with known accuracy.
  • Data from >14000 students of USW30x course.
Experiments

- We estimate $P(\text{SurveyResponse} \mid \text{TimeSinceStopout})$ as 10x values from *post-course* survey.
- Vary $D$ over \{ 0.025, 0.030, …, 0.250 \}.
Experiments

- We estimate \( P(\text{SurveyResponse} \mid \text{TimeSinceStopout}) \) as 10x values from post-course survey.
- Vary \( D \) over \{ 0.025, 0.030, \ldots, 0.250 \}.
- For each student in training set:
  - For each week \( t \):
    - Run classifier to compute \( y \).
    - Sample controller’s action from \( p_\theta(y, t) \).
    - If “query”, then sample student’s response from \( P(\text{SurveyResponse} \mid \text{TimeSinceDropout}) \).
    - Compute \( C_{st} \) from action & response.
Optimization

• Controller 1:
  • 1D “grid” search over \{ 0, 0.01, \ldots, 0.99, 1 \}.

• Controller 2:
  • Policy gradient (“REINFORCE”, Williams 1992) and stochastic gradient descent (2000 steps) with learning rate=0.1, momentum=0.5.
Simulation results

Each mark corresponds to the minimum cost for some $D$. 
Simulation results

• Example of learned parameters $\theta$ (for $D=0.175$):

  • $s = 27.97$, $b_1$, ..., $b_T$ shown below:

\[
p_\theta(y, t) = \sigma(sy - b_t)
\]

Controller: probability of survey is

Controller: probability of survey is $p_\theta(y, t) = \sigma(sy - b_t)$
Simulation results

• Simulation suggests that expected survey response rates may be meager.
  • Need to craft the query emails to be “mild”.

• Simulation provides mild evidence that more powerful, non-stationary controller is useful.

• However, numerical optimization is slower and more complicated (convergence detection, hyper-parameter selection, etc.).
Next steps

• In January we will run a pilot “stop-out query” experiment on real MOOC students.

• With new data, we can re-train the stop-out classifier, and re-optimize the controller.

• We are excited to learn why MOOC students are stopping out!
Next next steps

• Based on survey responses, modify course content, presentation style, etc.

• Design first interventions, e.g.:
  • If student is on the verge of stopping out, appoint him/her help from a teaching assistant.
Completion and Retention Rates in the Context of Student Intent

19.5% of survey takers who intend to complete go on to earn a certificate, compared with 5.4% of survey takers who do not intend to complete that go on to earn a certificate.


Computer-Assisted Reading and Discovery for Student Generated Text in MOOCs

Advances in topic modeling approaches show promise for helping instructional teams make meaning of student writing contributions to MOOCs

http://harvardx.harvard.edu/harvardx-working-papers
Staggered Versus All-At-Once Content Release

MOOC students show a revealed preference for flexible course designs, and ontrackness seems only weakly correlated with measures of student performance. Course developers should consider potential benefits of flexible course designs.

http://harvardx.harvard.edu/harvardx-working-papers
End