

KDD CUP 2015

Problem

The competition participants need to predict *whether a user will drop a course* within next **10 days** based on his or her prior activities.

If a user leaves no records for course in the log during the next 10 days, we define it as dropout from course.

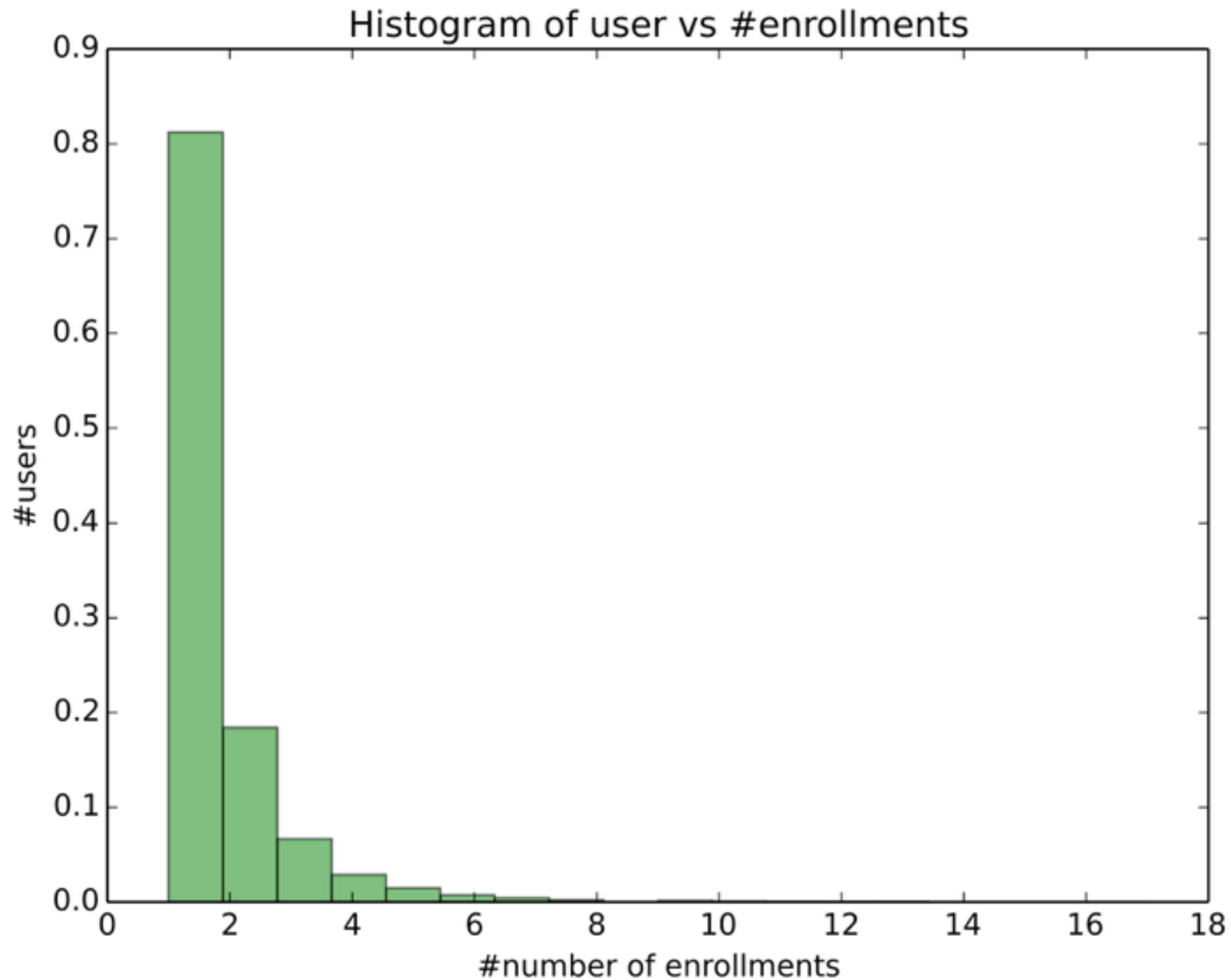
User Logs

1. problem – Interaction with the course's quiz;
2. video - Interaction with the course's video;
3. access – Interaction with other course objects (rather than videos or quizzes);
4. wiki – Interaction with the course wiki;
5. discussion – Interaction with the course forum.
6. navigate - Navigation through the course;
7. page_close – Leaving the course's web page.

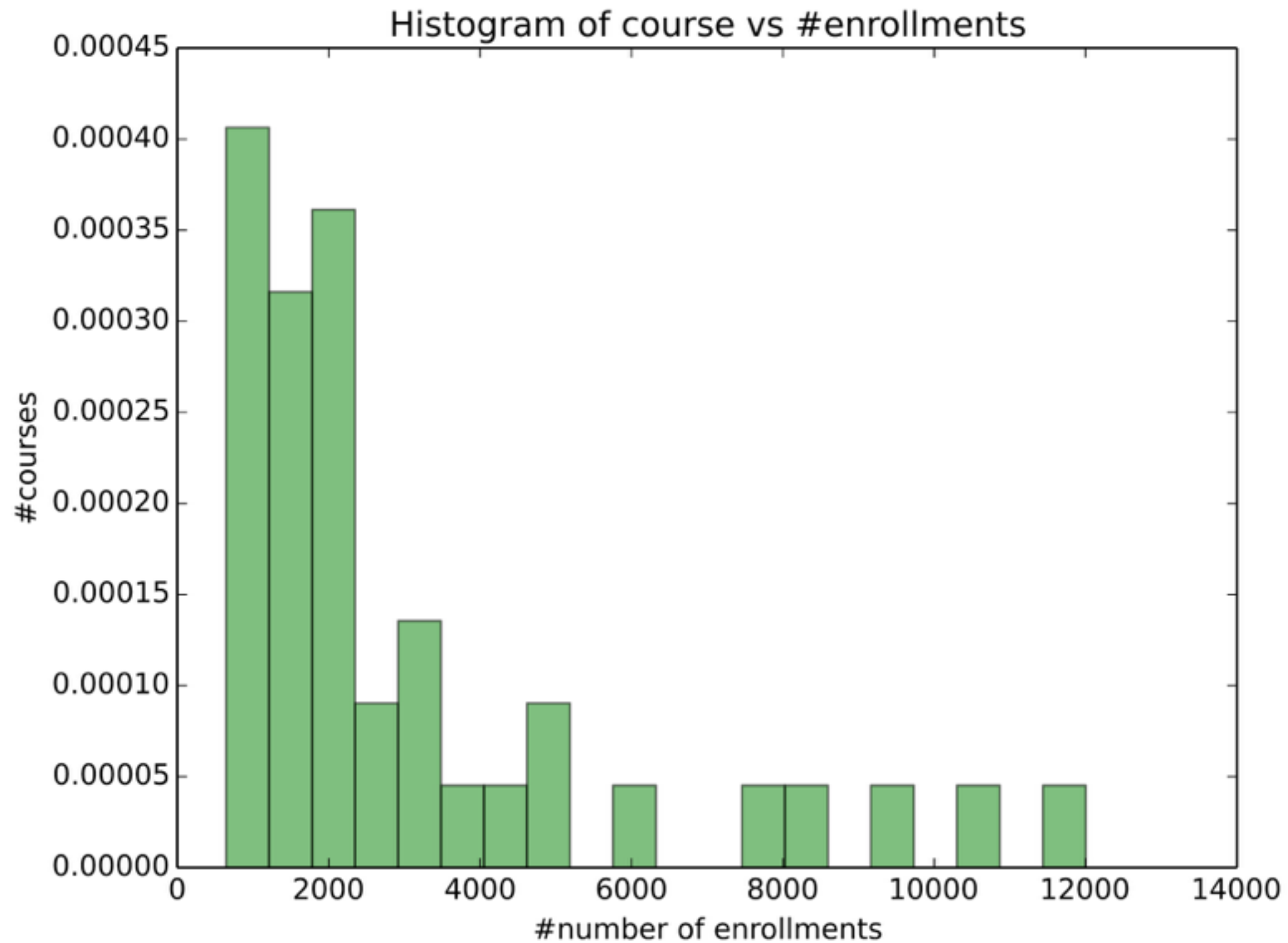
Data Statistics

- Start Time: 2013-10-27
- End Time: 2014-08-01
- Duration: 10 months

Data Statistics

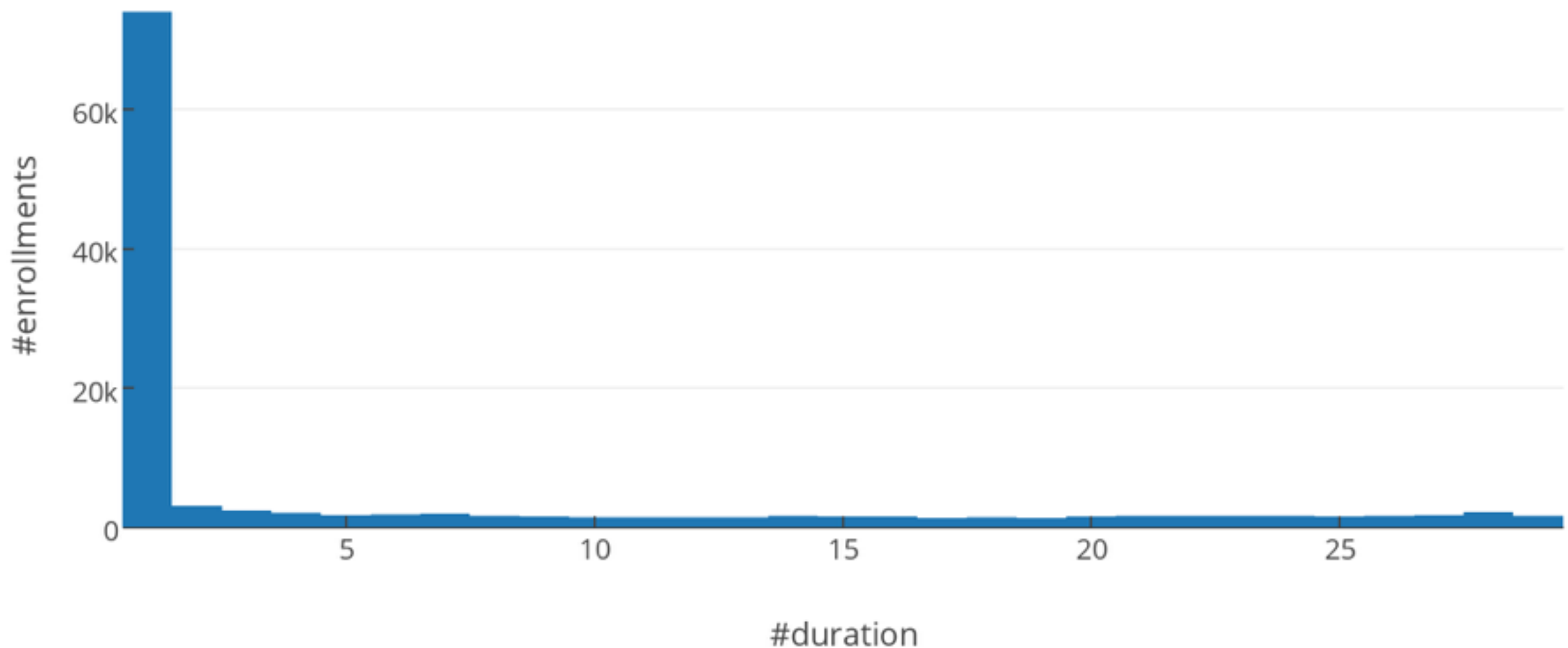


Data Statistics

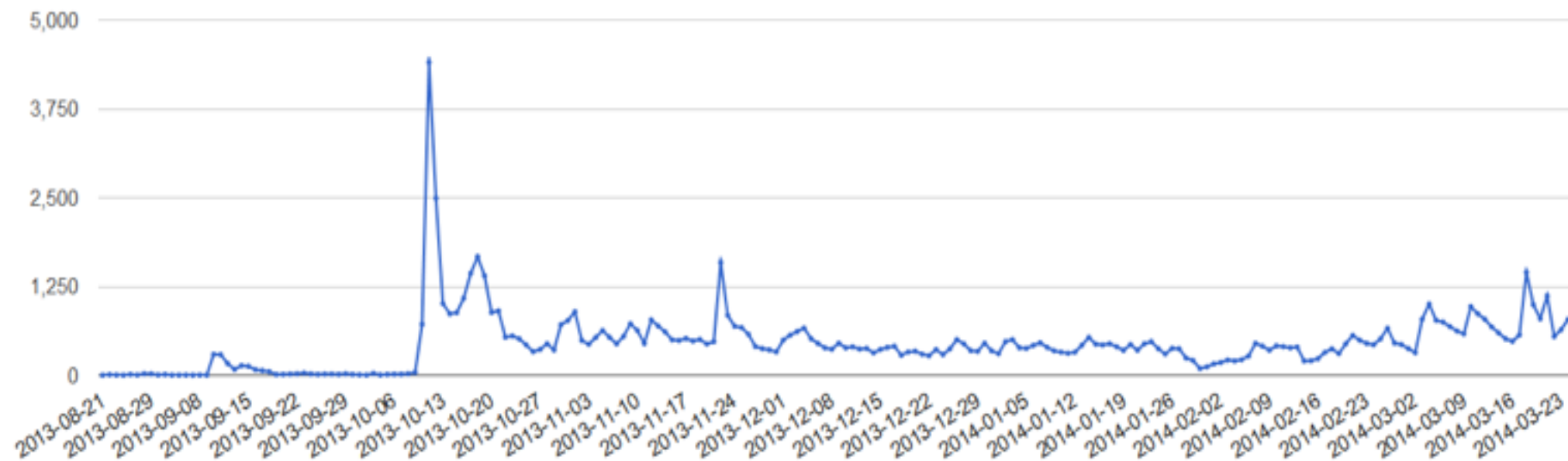


Data Statistics

Histogram: #duration



User enrolment in the past months

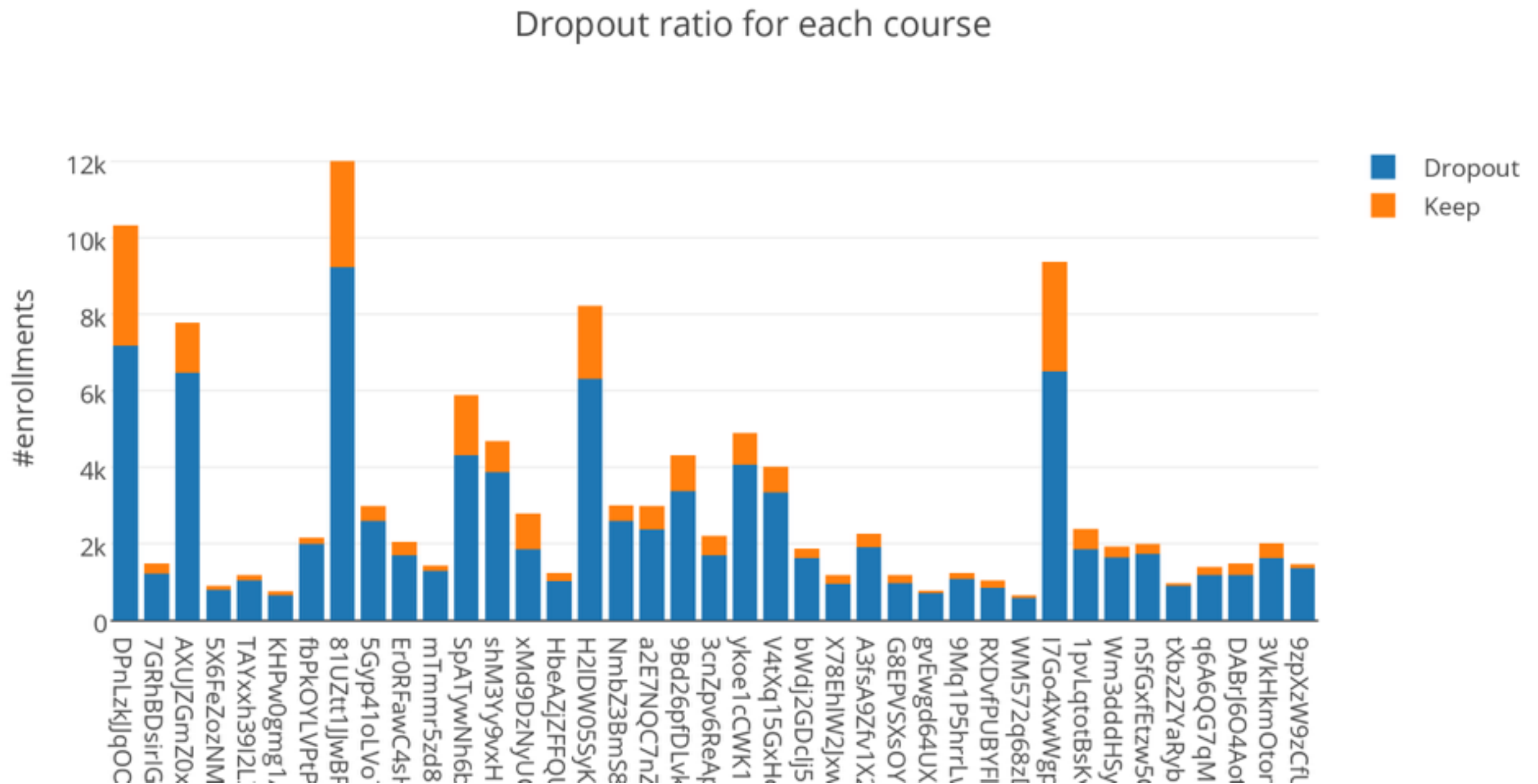


2013-10-27



2014-08-01

Dropout vs Keep per Course



Features #1: summary features

- **Duration** (days) of enrolment: from the start of the first event to the last one
- **Number of (*cumulated*) events** of enrolment: the count of event observed in the log
 - **Number of videos:** the count of observed video event in the log
 - **Number of problem:** the count of observed problem event in the log
 - **Number of wiki:** the count of observed wiki event in the log
 - **Number of navigate:** the count of observed navigate event in the log
 - **Number of access:** the count of observed access event in the log
 - **Number of discussion:** the count of observed discussion event in the log
 - **Number of page_close:** the count of the page_close event in the log

Features #1

- **Active days** of enrolment: the total number of days that the user access the course
- **Active days per week**: the average active days every week
- For the last 3 months from 05-13-2014 - End (12 week)
- **Active days in week [1-12]**: the active days in the #-th week

Features #2: sessions

- **Number sessions:** the number of sessions included in the enrolment logs
 - The time gap between sessions is 30 minutes
- **Avg requests per session:** #video, #problem, #access, #navigate, #discussion
- **Avg video per session**
- **Avg problem per session**
- **Avg access per session**
- **Avg navigate per session**
- **Avg discussion per session**

Features #3: behaviour time-pattern

- **Daytime** vs **Nighttime**
 - *Daytime*: 07:00 - 19:00
 - *Nighttime*: others
- **Weekday** vs **Weekend**

Features #4: temporal features

- Summary features in last **{1, 2}** week
- Session features in last **{1, 2}** week
- The **number request happens in time slots**:
 - 0am-6am
 - 6am-9am
 - 8am-12am
 - 12am-18pm
 - 17pm-20pm
 - 19pm-24pm
- The **count/mean/variance hour** of requests

Features #5: lagging

- **Lag:** the time gap (in unit day) between active days
- Min/Max/Mean/Std lags
- Number of lags > 3 days
- Number of lags > 1 week
- Number of lags > 2 weeks

Features #6: module level features

- The lag (in unit of day) between the release time of the accessed module and the access date
- The median days of the lags for 1st/last access
- The 25% and 75% percentage days of lags for 1st/last access

Features #7: stay time

- The **stay time** for every active days
 - max/min/mean/variance stay time

New Features

- Enrolment Features
 - month of first access
 - month of last access
 - stay time per active day (min/max/mean/variance)
- Course Features
 - the audience of the enrolment course
 - the number of dropout of the enrolment course
 - the dropout ratio of the enrolment course
- Module Features
 - the number of times that this student navigated backwards from a learning sequence to another one released earlier in the term (e.g., from Lecture 6 to Lecture4), divided by the total number of sequences visited by this student.
 - the sequence order of the modules that user access (Kendall-tau with the release order)

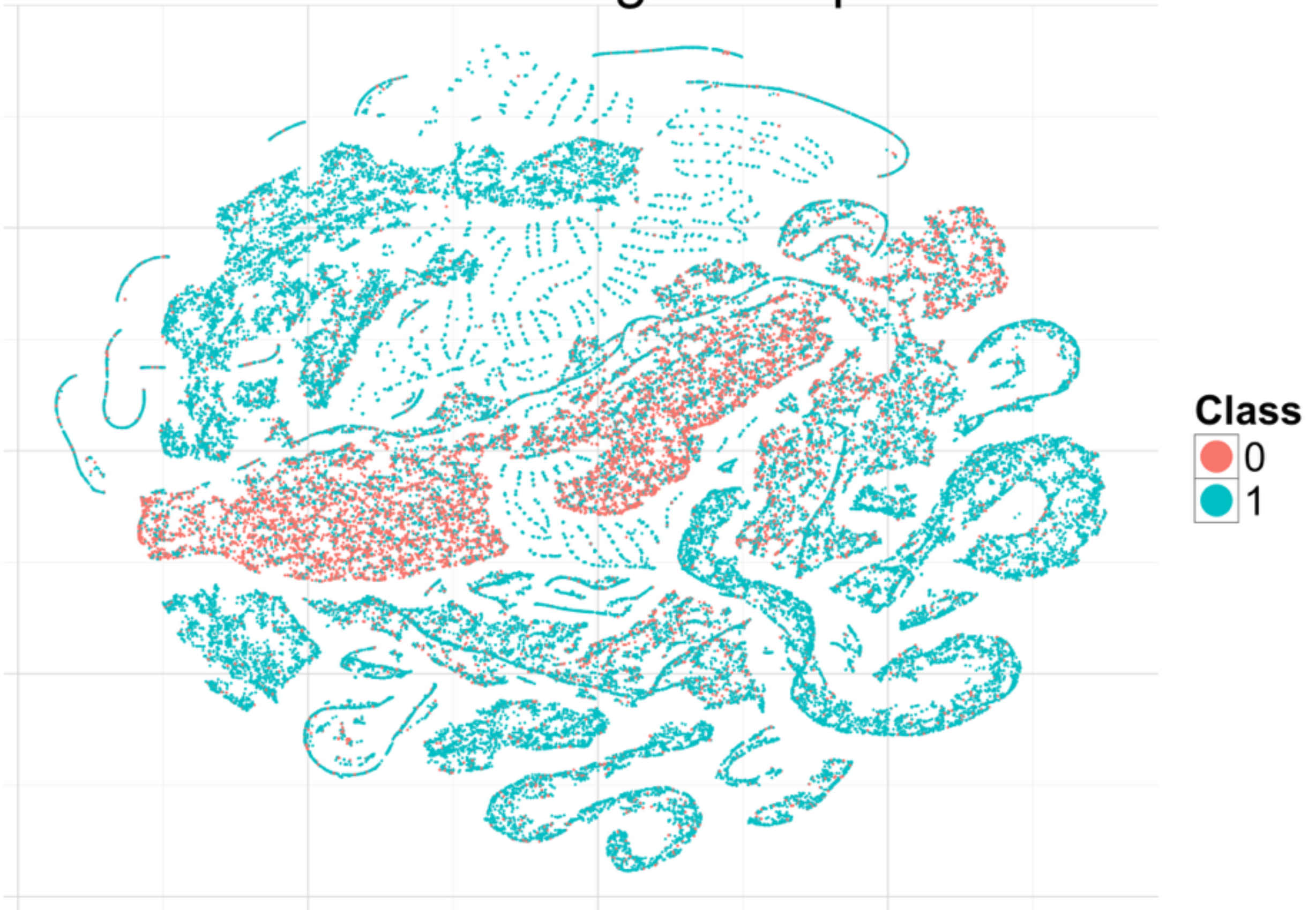
Models

- Logistic Regression
- Gradient Boosting Tree (xgboost)
- Random Forest
- Deep learning
- LibFM (<http://www.slideshare.net/hustwj/modeling-dynamic-social-networkslearning-from-users-and-prediction>)

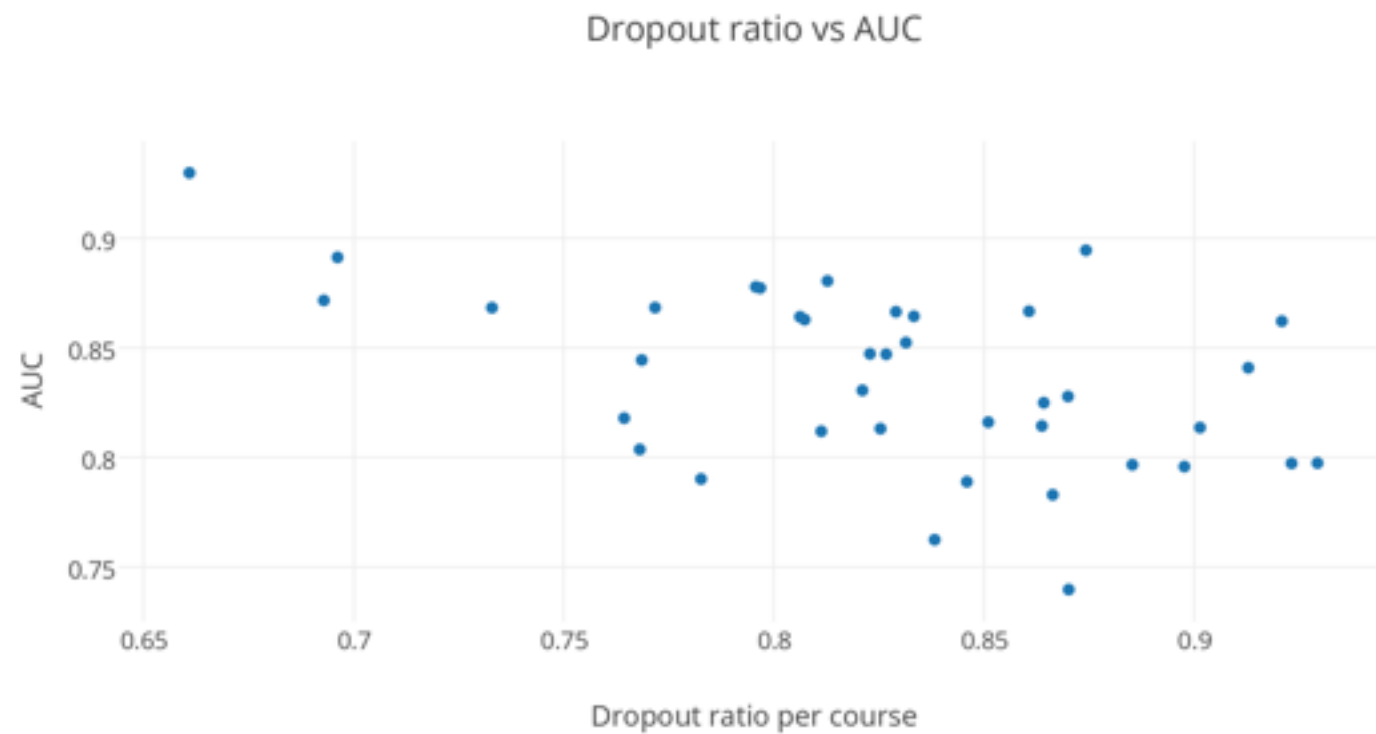
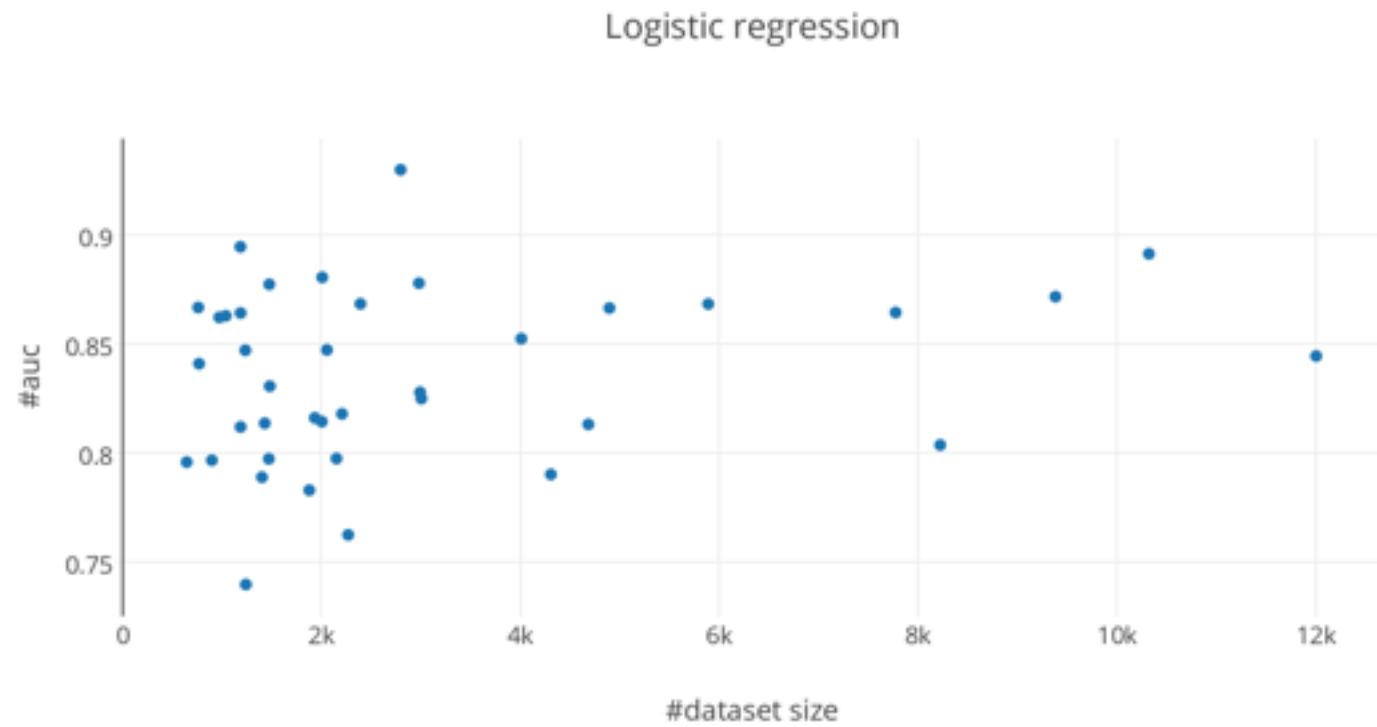
Improvement Directions

- The **correlations between users** who enrol the same course
 - similarities between users
 - the preference regarding to the courses' module
- Feature selections/normalise/scaling
- t-SNE dimension reduction (would benefit the logistic regression/neural network classifier)
- Different perspective of this problem
 - ranking/recommendation problem
 - pair-wise (positive vs negative enrolment) model (address the imbalance problem)

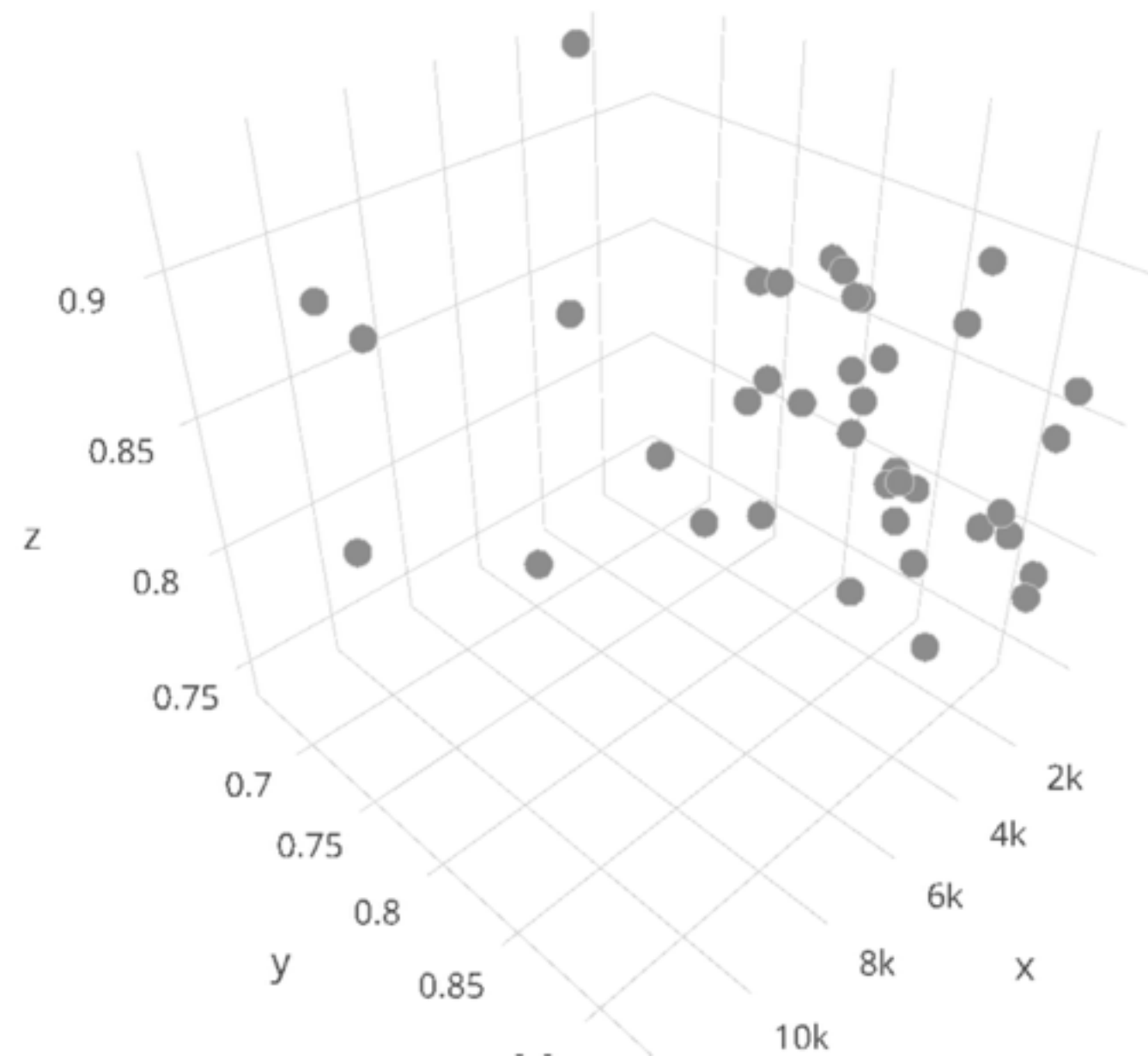
t-SNE 2D Embedding of Dropout Data



Course-independent Model



<https://plot.ly/~numb3r3/181>



T-SNE & Meta bagging

- Meta bagging: <https://kaggle2.blob.core.windows.net/forum-message-attachments/80068/2554/metabag.pdf?sv=2012-02-12&se=2015-06-06T06%3A28%3A42Z&sr=b&sp=r&sig=ZeOdKotoMq6RhXB6R%2BhSei%2Bo9KjfozfwoVRCFGVrV3Y%3D>
- <https://www.kaggle.com/c/otto-group-product-classification-challenge/forums/t/14295/41599-via-tsne-meta-bagging>
- <https://www.kaggle.com/c/otto-group-product-classification-challenge/forums/t/14134/deep-learning-h2o-0-44>