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.

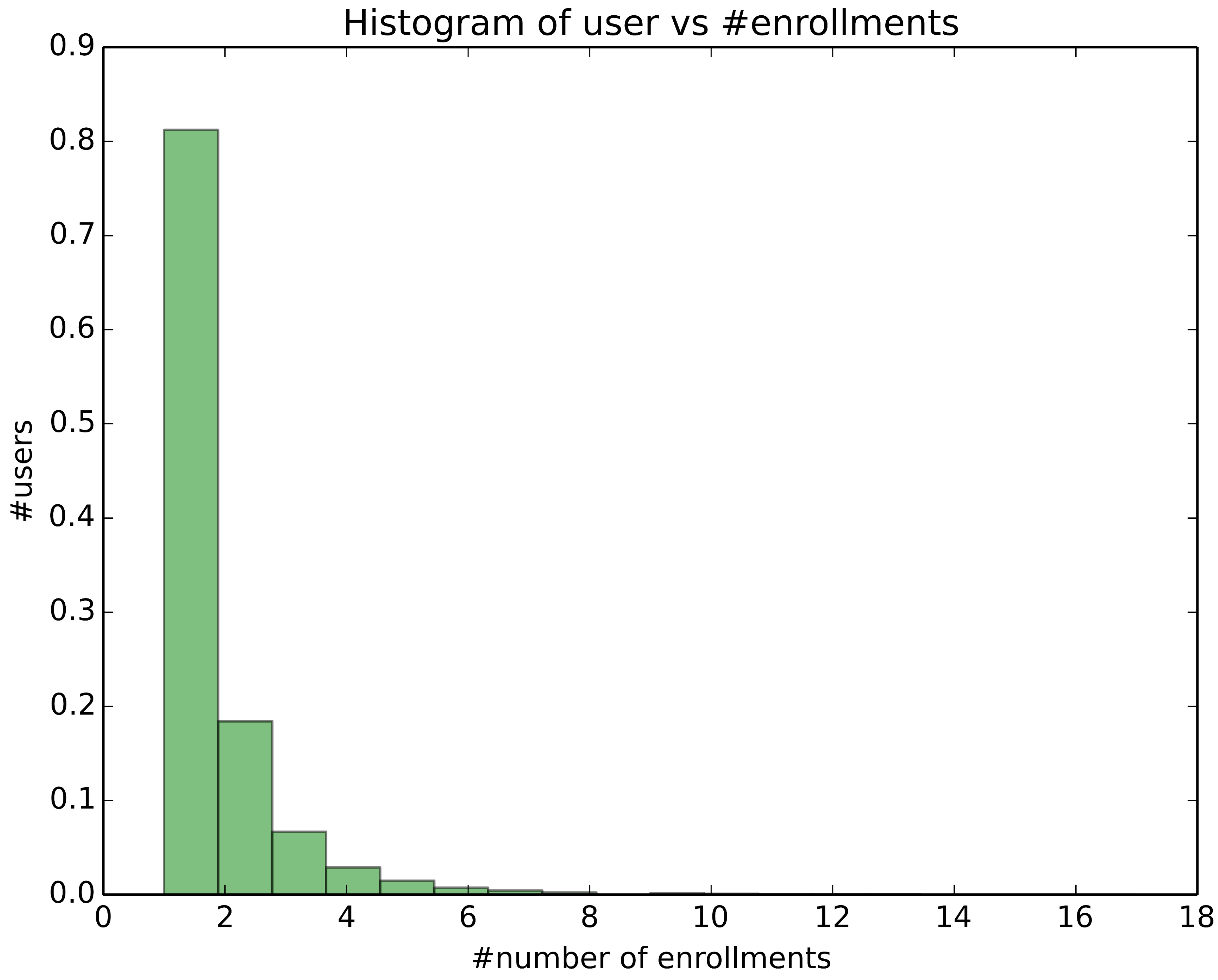
6.navigate - Navigation through the course;

7.page_close – Leaving the course’s web page.

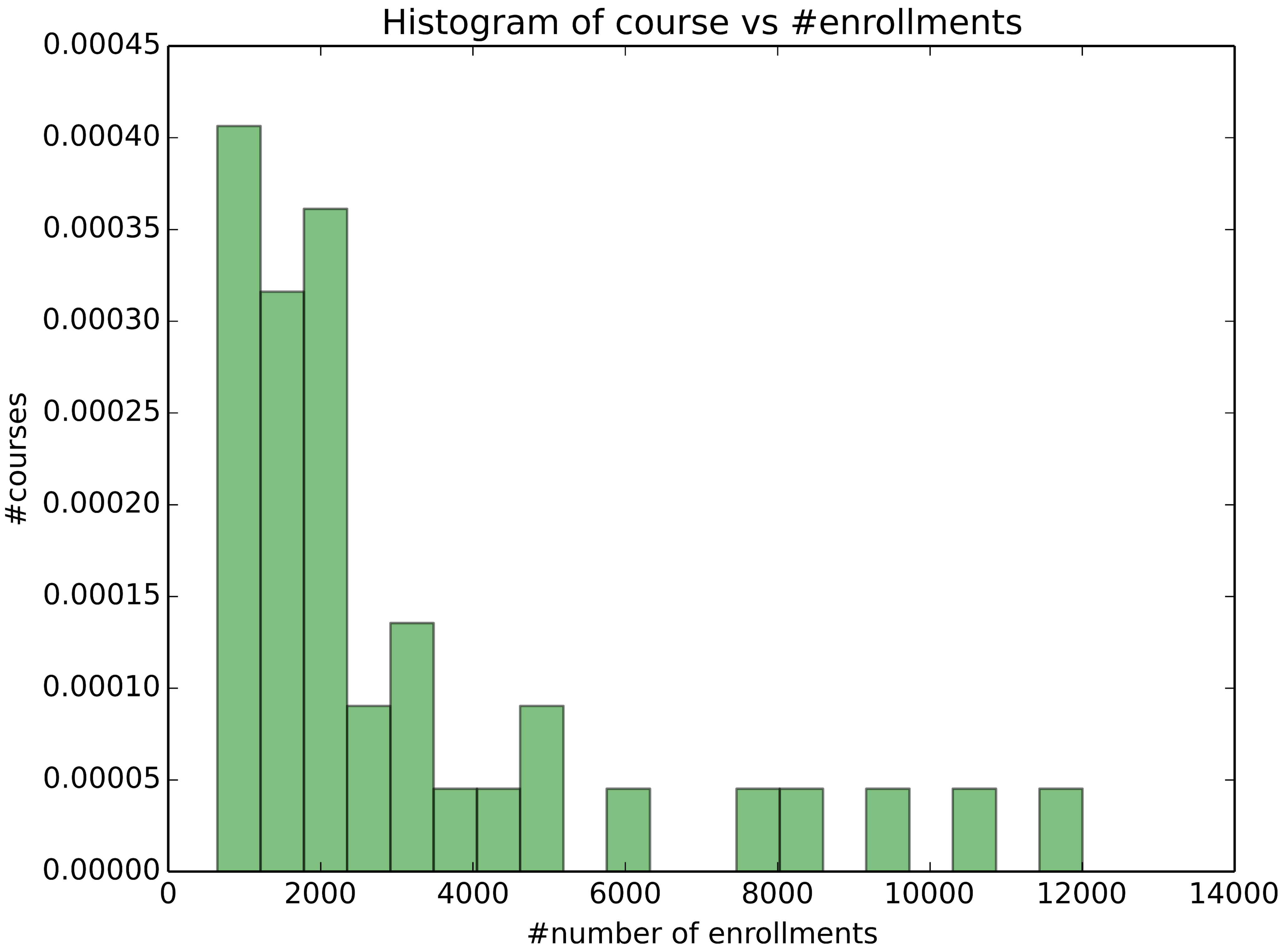
Data Statistics

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

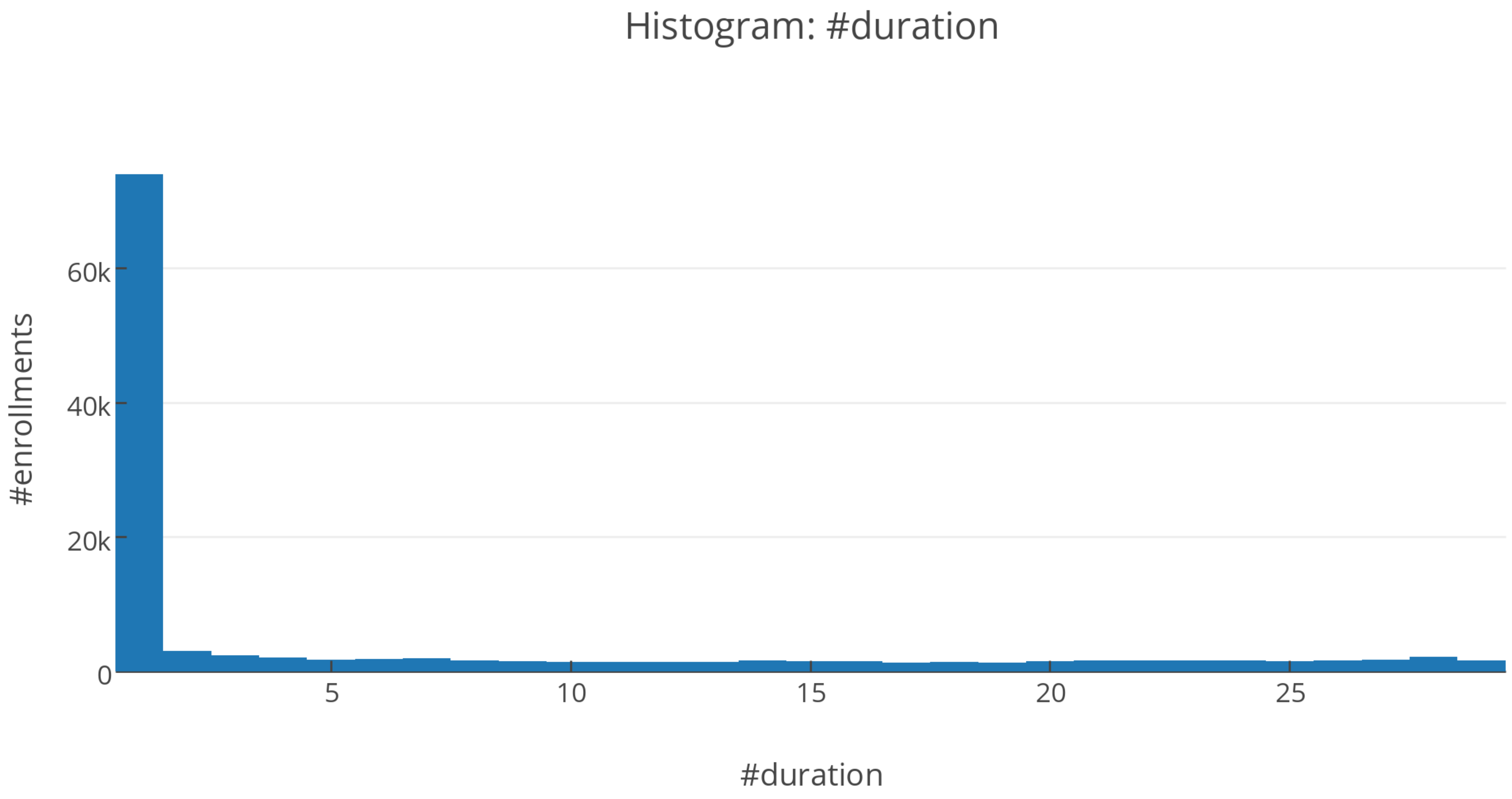
Data Statistics

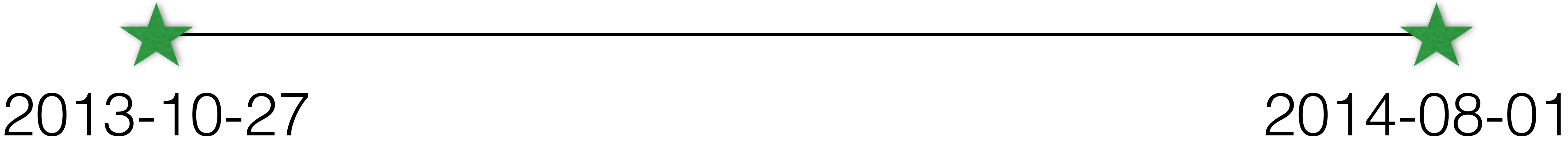


Data Statistics



Data Statistics

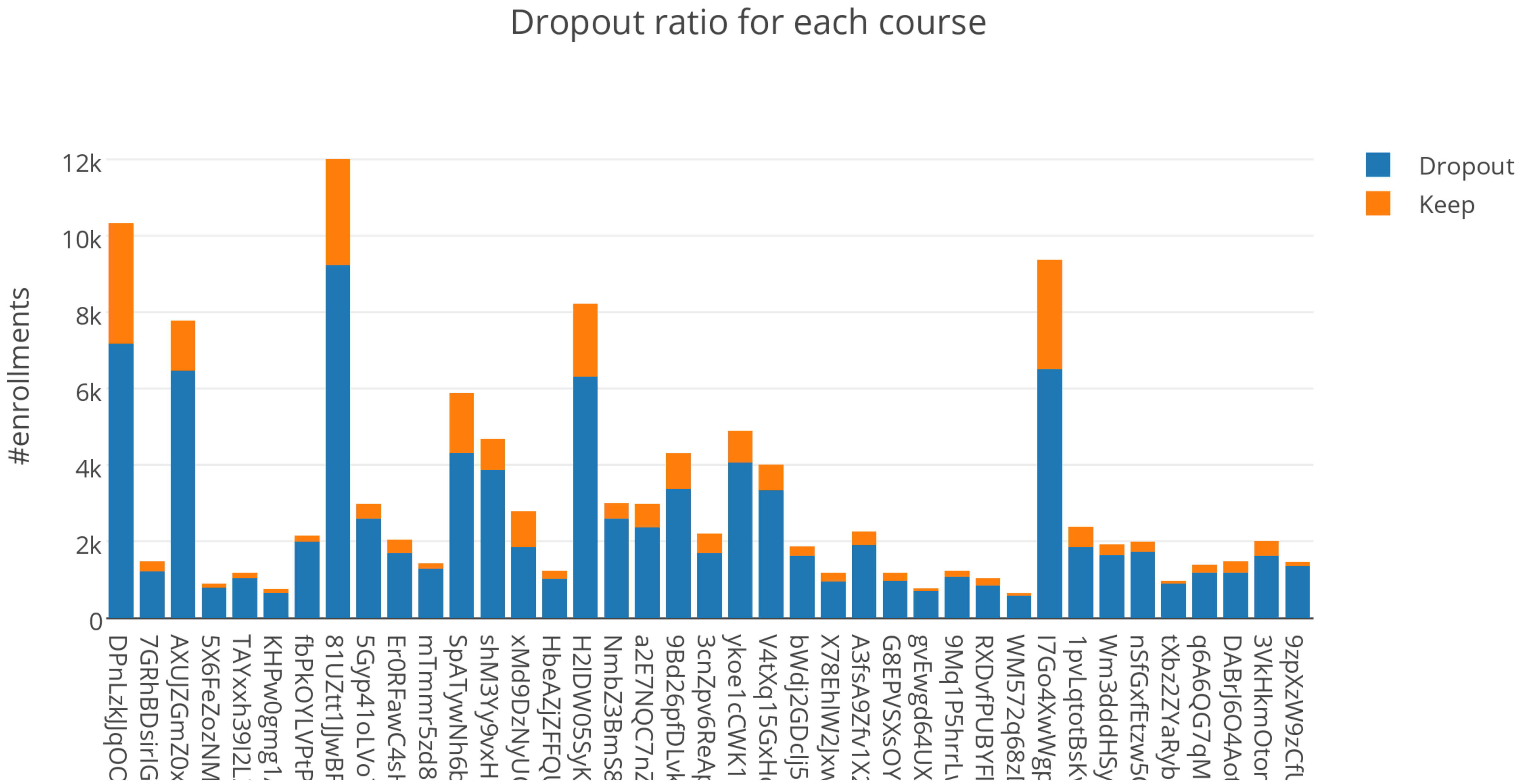




2013-10-27

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

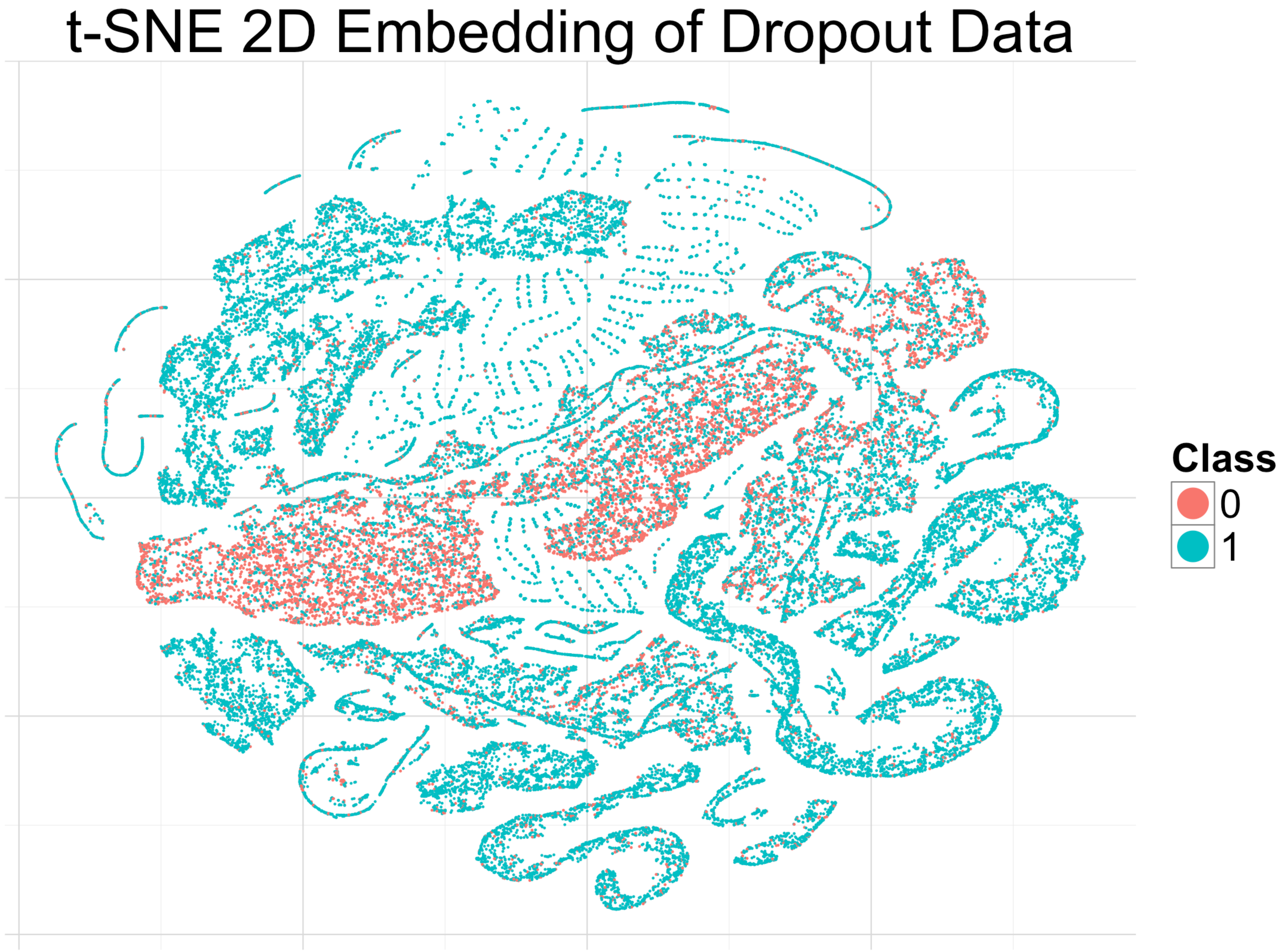
- The **stay time** for every active days
    - max/min/mean/variance stay time

Models

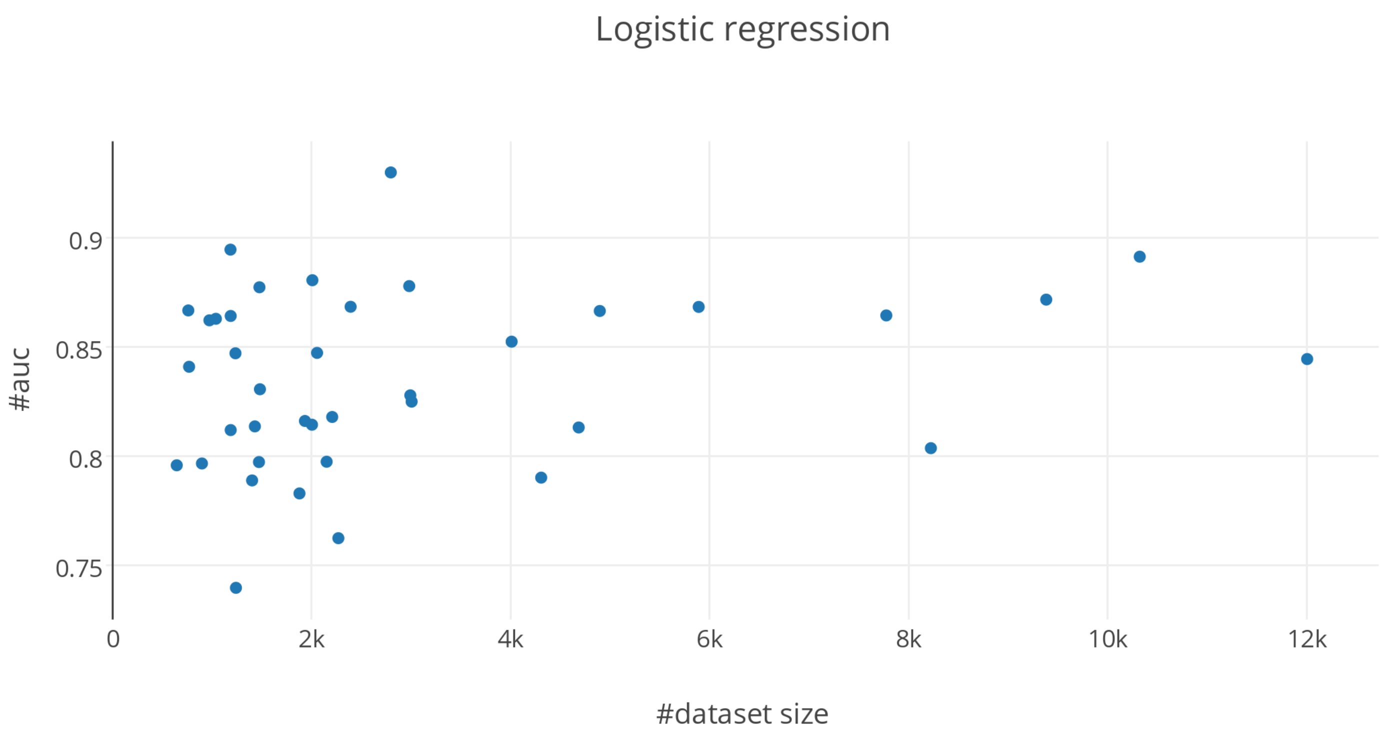
- Logistic Regression
- Gradient Boosting Tree (xgboost)
- Random Forest
- Deep learning

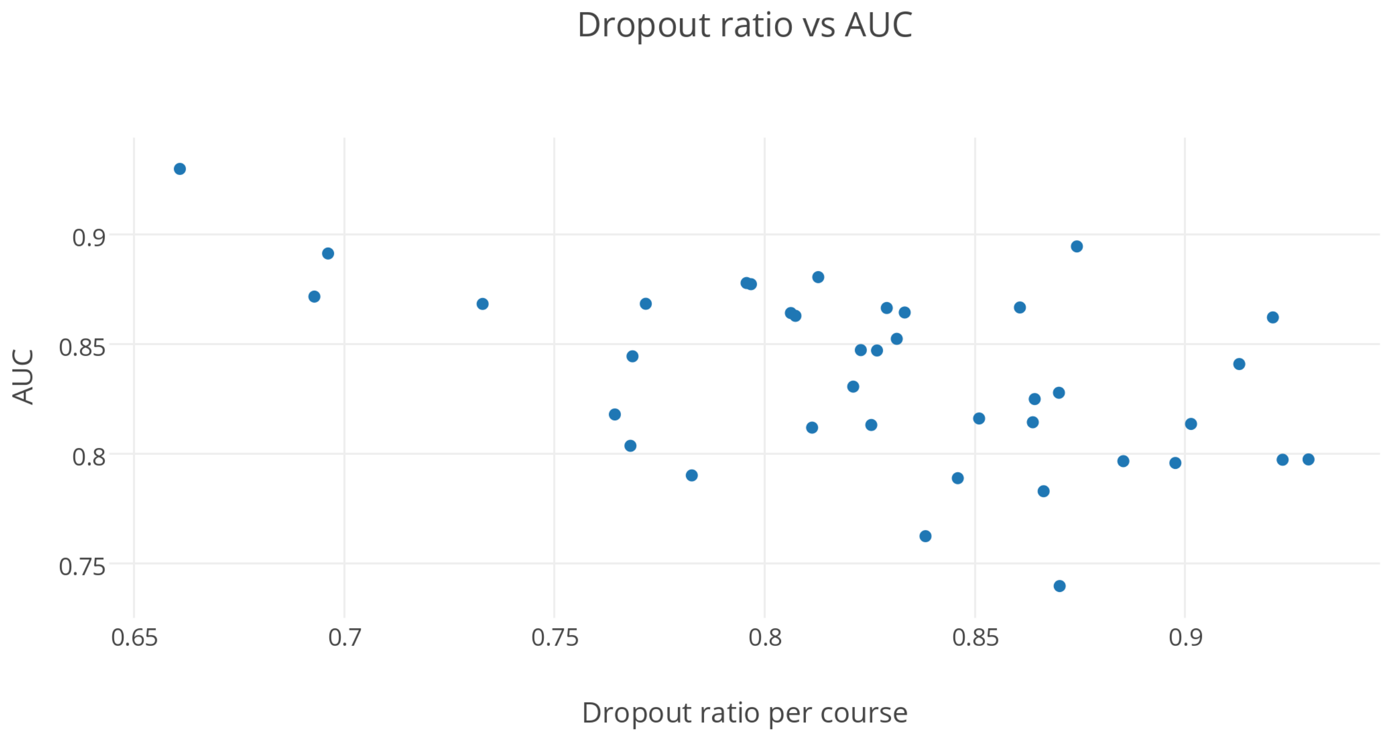
Improvement Directions

- The **correlations between users** who enrol the same course
    - similarities between users
- 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)

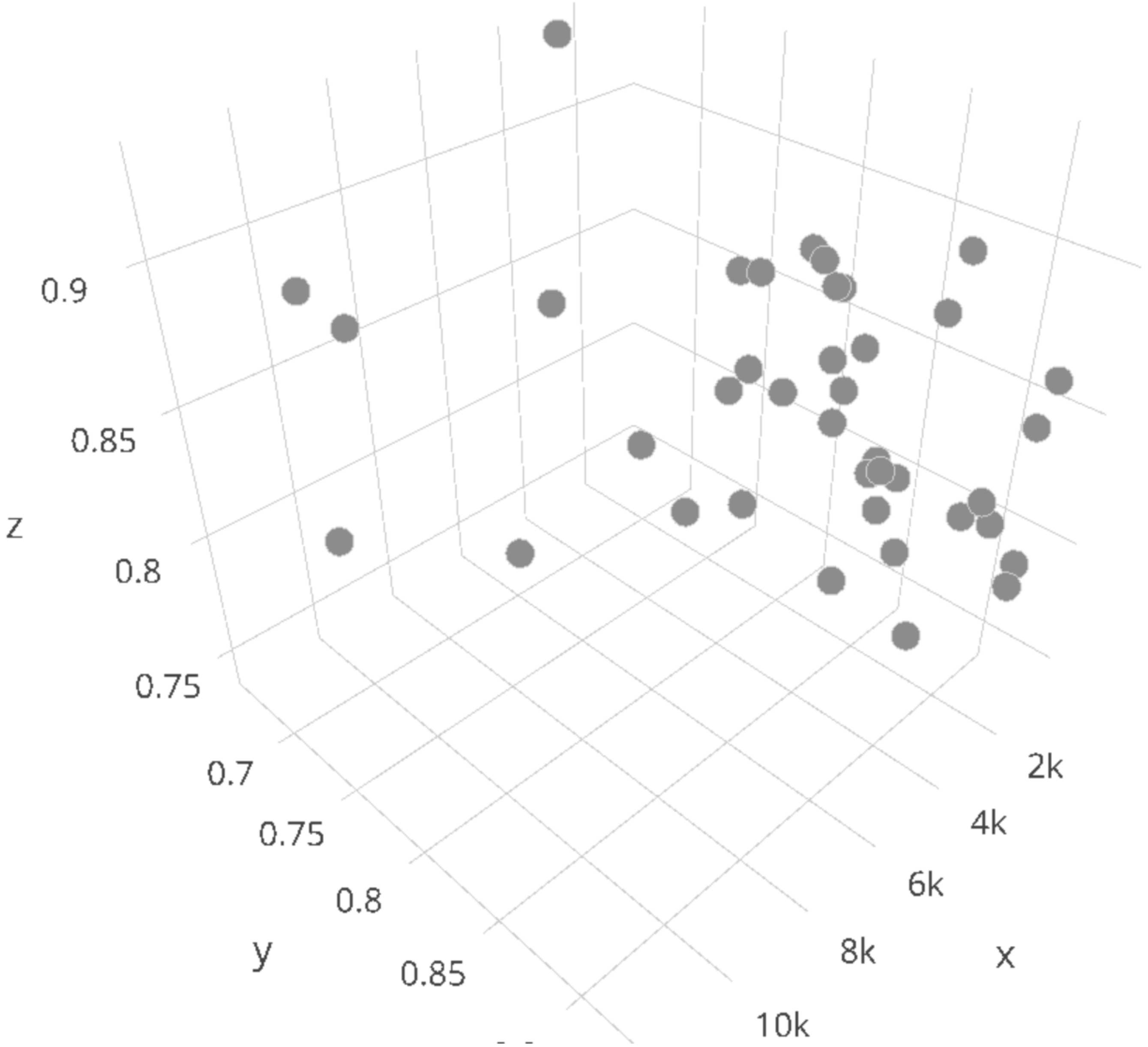

t-SNE 2D Embedding of Dropout Data
Class
0
1

Course-independent Model





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


0.9
0.85
z
0.8
0.75
0.7
0.75
0.8
y
0.85
2k
4k
6k
8k
x
10k