**MOOC Drop out Prediction - Description**

The problem is to predict whether or not a user will drop a course in next 10 days based on his or her prior actions. We say a user *U* drops a course *C* in the next 10 days if *U* leaves no records of course *C* in the *log* during the next 10 days. For more details about *log*, please refer to Datasets Section.

**Datasets:**

1. **enrollment.csv** – Each line in this file is a course enrollment record with an enrollment id, a userid *U* and a course id *C*, indicating that *U* enrolled in course *C*.

- enrollment\_id – Enrollment ID

- username - Student ID.

- course\_id - Course ID.

1. **object.csv** - Each line in this file describes a module in a course with its category, children object and its release time. Those modules represent different online materials of the courses, e.g., chapters, videos, problem sets and so on. The modules of a course are organized as a tree, i.e., each course contains several chapters; each chapter contains several sections; each section contains several objects (videos, problem sets, etc.).

- course\_id - The course to which the module belongs.

- module\_id - The ID of a courseware module.

- category - The category of the courseware module. ( can be video , problem, etc)

-children - The children modules of the courseware module.

- start - The time that the module was released to students.

1. **log.csv** – each line is a behavior record called 'event'. Each event contains the following information: username, course\_id, time, event type, event\_sourse (server or browser), and object.

- enrollment\_id - Enrollment ID.

- user\_id - Student ID.

- course\_id - Course ID.

- time - Time of the event.

- source - Event source (server or browser).

- event - Event Type

- object - The object the student access or navigate to. (For navigate and access event only).

In terms of event type, we defined 7 different event types:

- problem - Operations on course problems

- video - Operation on course videos;

- access - Accessing other courseware objects (neither a video nor a problem);

- wiki - Accessing the course wiki;

- discussion - Accessing the course forum.

- navigate - Navigating to other part of the course;

- page\_close – Close the web page.

1. JoinedLogs\_TrainTest.tsv

The event logs and the object data can be joined to one single file.

Schema:

eid: enrollment\_id

cid: course\_id

uid: user\_id

time: time of event

src\_evt: concatenation of source and event columns

cat: category ( from object.csv file

obj: object

1. Truth\_train.tsv

Truth\_train.tsv contains the labels with this schema:

eid: enrollment\_id

churn\_label: the label. 1 = dropped out; 0 = stayed.

**Featurizing Strategy**

First we decide the different pivots that will be relevant to the problem. That is what are the keys are we going to use to create summary columns?

Some of the keys are –

1. User Id
2. Enrollment Id
3. Course Id
4. User Id, Enrollment Id
5. User Id, Course Id
6. Enrollment Id, Course Id

Summary features that we will create will be applicable for each of the above 6 pivots.

The features can be segregated into 3 different types – No grouping, Category, Source|Event