Learning Analytics for MOOC
Business Understanding Document

Group No. : DS707-2017-10
ChellaPriyadharshini M (**MT2016041**)
Daminee Sao (**MT2016045**)
Jyotsana (**MT2016068**)
Kanika Narang (**MT2016069**)
Tehreem Ansari (**MT2016145**)

Determining Business Objectives

Background:

Learning data from open online courses hold great promise for research. In 2012, the Massachusetts Institute of Technology (MIT) and Harvard University launched open online courses. The series features detailed reports about individual courses; these reports reveal differences and commonalities among massive open online courses (MOOCs).

Massive Open Online Courses (MOOCs) provide massive amounts of data about learners and how they interact with an online learning environment. Due to its openness, MOOC students vary in their heterogeneity such as age, gender, educational background and location. Learners are not only limited to a single type path learning specialization.

Data Overview:

Open source data taken from Kaggle, it provides data on 290 Harvard and MIT online courses, 250 thousand certifications, 4.5 million participants, and 28 million participant hours on the edX platform since 2012.

Dataset contains 23 columns, the important ones are:

Column Name	Description	Data Type
Institution	online course holders	String
Course Number	the unique id of each course	String
Launch Date	the launch date of each course	DateTime
Course Title	the title of each course	String
Instructors	the instructors of each course	String

Course Subject	the subject of each course	String
Year	the last time of each course	Numeric
Participants	the number of participants who have accessed the course	Numeric
Audited	the number of participants who have audited more than 50% of the course	Numeric
Certified	the number of participants who have been certified	Numeric
Median Age	median age of the participants	Numeric
% Male	percentage of male students	Numeric
% Female	percentage of female students	Numeric
% Bachelor's Degree or higher	the percent of bachelor's degree or higher	Numeric

Stakeholders:

- **Institution**: We have two institutes, MIT and Harvard University. This includes instructors, content setters, teaching assistants and course providers.
- **Prospective Student**: New incoming students who are thinking of attending the course.
- **Current Student**: Students who are currently enrolled in a course.
- **Instructor**: Instructor for a course. Two or more courses may have same instructor.
- **Course Subject**: Under which subject does the course lie. Common subjects include: Science, Technology, History, Humanities, etc.

Business Objectives:

The main objectives are:

- Improve learning in MOOCs in general
- Enhance the completion rate in MOOCs
- Study learner patterns and predict at-risk students
- More specifically,
    - Which students are likely to drop out and which ones will reach till completion.

- Which course are more beneficial to the students and will provide a better result.
- Compare Universities and their data and see which institute (MIT/Harvard) is providing most liked course.
- Popularity of the courses; popularity of topic/technology/subject; popularity of Instructors.
- Demographics of the students attending the courses.
- Categorisation of students into groups such as: students who engage in every piece of coursework, only read text, only view videos, only take assessments or complete problem test, or who demonstrate combination of these behaviours.

Business Success Criteria:

- The Institute can get better understanding of students' online behaviour; which courses reap maximum benefits; what kind of students prefer what kind of courses; etc. And help increase the retention rate of students enrolled in a particular course. (**Stakeholder benefited: Institution**)
- A Student can select from a number of courses and decide the most beneficial. Among the universities that offer same course (MIT and Harvard) which one should they opt for? (**Stakeholder benefited: Prospective Student**)
- What is the behaviour of the students who have already enrolled in the course? Their demographics, completion rate, etc. (**Stakeholder benefited: Current Student**)
- The instructor can identify which courses he/she has taught were most liked by the students, which domain reaps maximum enrollments? The success rate of the instructor? etc. (**Stakeholder benefited: Instructor**)
- Under which subject of the course do students find maximum/minimum interest. Which instructor teaches maximum subject related course? (**Stakeholder benefited: Course Subject**)

Assessing situation

Assumptions and Risks:

- We are assuming that if a person has suspended the course then it is a dropout.
- It is possible that the person has just temporarily suspended the course with the intention of resuming it again in future. However, such a person will be considered as a dropout.
- We assume that any person who has registered for the course is a genuine user and is aiming for completion.We are not assuming that any enrollee has just started the course for seeing the content or viewing particular topic i.e. False starter.
- There may be people who register only for few video/text lectures but not for the complete course. This will indirectly affect the quality of the result of analysis of course.

Determining Data Mining Goals

Data Mining Goals:

- Use the data to determine how the courses from Harvard and MIT are faring among the online learners, including: student demographics, completion/drop-out rate, engagement rate, etc. (**Descriptive and Exploratory Analysis**)
- Based on how the students engage in the course we can come up with classes that depict their level of interaction. (**Classification**)
- Identify areas (like math, science, humanities), specific courses, specific instructors that have good conversion rates. These help us understand popularity of topics, popularity of courses and popularity of instructors, respectively. (**Clustering**)
- Detect specific student characteristics that help relate similar learning behaviour. For example: we need to identify rules such as: " a student who completes course X is more likely to complete course Y too". (**Association Rule Mining**)

Data Mining Success Criteria:

- Complete classification of students based on completion of the course.
- Accuracy of the prediction of student belonging to which class should be significant.
- Identifiable clusters based on courses, instructors and topics.

Project planning

Project Plan:

Phase	**Time**	**Resources**	**Risk**
Business Understanding	1 week	All analysts	Misinterpretation of domain level ideas
Data Understanding	1 week	All analysts	Data problems, Technology problems
Data Preparation	2 weeks	All analysts	Data problems, Technology problems
Modeling	2 weeks	All analysts	Technology problems, inability to find adequate models
Evaluation	1 week	All analysts	Inability to implement results

Initial assessments of tools and technologies:

- R Programming tool and it’s libraries
- Tableau for Visualization
- Excel/LibreOffice