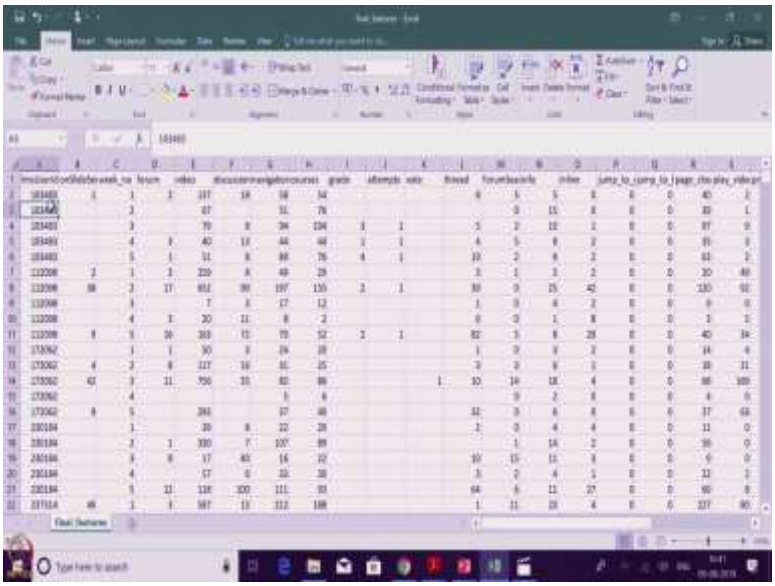
Introduction to Learning Analytics
Prof. Ramkumar Rajendran
Department of Interdisciplinary Programme in Educational Technology
Indian Institute of Technology, Bombay

Lecture - 17
MOOC data for Course Project

In this LeD we will talk about the Course Project, there is a small project in this course; before that we will describe what is the data and how we process the data. And we will explain the each columns on the data sheet, also what you have to predict. This course project is based on the MOOC data we collected from IIT Bombay x. So, we discussed this MOOC data like what data can be collected from MOOC in a week 2, just to recollect here is the data we can collect from MOOC such as the learners Id, student Id, session Id, their IP address, their discussion in the forums, number of upvotes, downvotes.

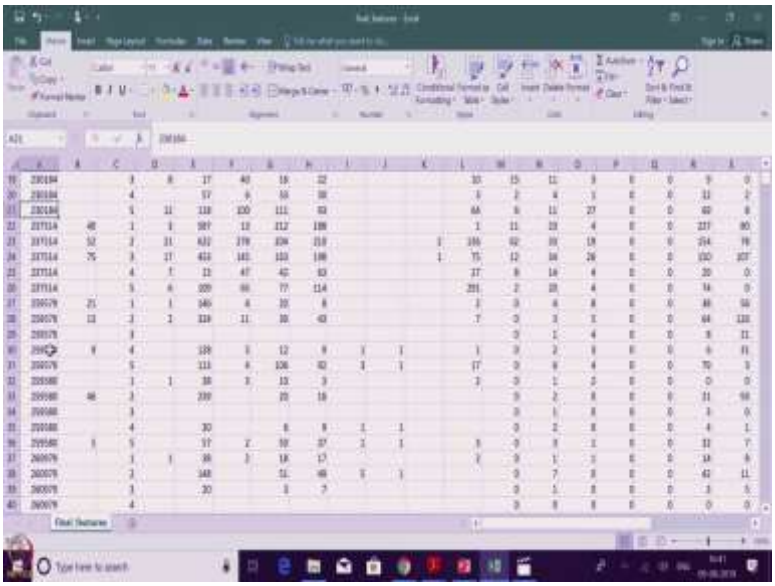
Also, their video watching behavior, there is seeking the video or they watching video in which speed 1 x speed, 1.5 x speed or using the caption, all this information can be collected. And also we saw the two set of data processed for example, this is a raw data we collect from the MOOC courses and the raw data consists of the student is seeking the video. And where the student is seeking the video also given here, also we saw a similar raw data for scrolling the book or scrolling the PDF in the MOOC. So, we have a data collected from MOOC and pre-processed, we will provide you the pre-processed data for the course project.

(Refer Slide Time: 02:04)



Now, I will show you the pre-processed data in excel sheet and I also explain what is the goal of this project. So, that is what aim of this particular project and what you have to do. Here is the data, the goal of this course project is to predict when the user will drop out; the first column is the user Id, the user Id is anonymized. So, the user is taking the course for 5 weeks. So, 1 2 3 4 5, this each row in this sheet is weeks data 1 week data, like those users 1 week interaction data with the MOOC. And for the second user, the user also interacted with more than 5 weeks. So, course was only for 4 weeks so, the user is interacting with 5 weeks means he has completed all the weeks of courses. Similarly, for user 3 user 4 and everyone has interacted more than 5 weeks.

(Refer Slide Time: 03:00)



There are few users who interacted with the course for first two weeks and third week and they drop out. We want to predict which users will drop out on this week; for example, this user one has interacted with a MOOC for a 5 weeks and we completed the course; so, the student is not dropping out. So, the column 1 is the UserId and column 2 is onSlideSeek, like if the student is seeking their video. If seeking is when you are watching a video, you can move the time like timeline; you can say I want to watch video from 1 minute seek back to 2 minutes, you can seek the video.

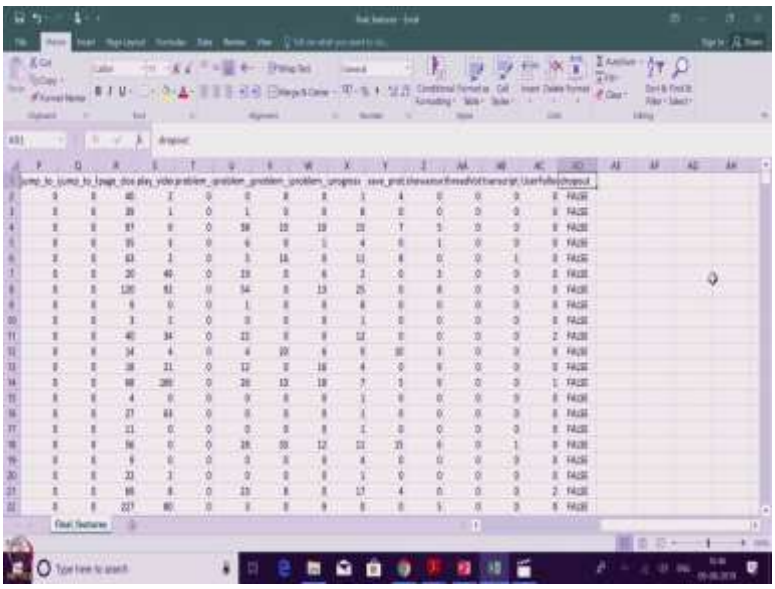
If they are seeking the video how many times they did in this particular week, that is a count, this is a column 1. Then column 2 this is week number like the user is participating in 5 weeks, 4 weeks, 1 or 2 weeks with each week number. So, we have data for each week. And how many times the user interacted in a forum? So, 2 times and how many how many minutes the student spent on the video that is 137 minutes here and 67 minutes, 79 minutes. And this particular column indicates how many times the student were actively participating in the discussion forum or discussions.

This one indicates how many times the students navigation moving the page from video to LxT or LxT to LxI or navigation happens or here this is a time the student spent on courses. So, these each thing is a time spent on each of these pages in the LMS model. Also, the grade student has taken a course assignment, this column is grades, the grades indicates students performance in the assignment for each week. We will indicate what is

a maximum marks for each week and this is the attempts how many times the student attempted on this particular assignment is a 1 time attempt or multiple times, the attempts numbers is there. And vote is the student upvoted someone's comments or someone's discussion in the thread.

And how many times user is participated in a thread and it sees is the user searched anything in the forum, the forum has lot of comments; is he searched for anything, any keyword in the forum and we look for that, the search number of searches also given in this column. And is looking for info's and is like looking for in inline comments or he is jumping to the courses or he jumped to particular html page. Or, how many times he closed the particular page or video playing time, a number of times he played the video and problem check failed and problem graded. So, we got all this information, we will provide a details of each column in the project.

(Refer Slide Time: 06:04)



So, what we have to predict? We have to predict the students drop out. So, here we have a user's, we have weekly data or interaction data in LMS and also we have the label to predict that is what you predict the students is dropping out or not so, only binary classification.

You need to use this data and apply the data on software Weka, you need to use this data and apply it on a ML software Weka and create a binary classifier model. And you have to use 10-fold cross validation and predict the students dropout rate from the given data.

In our course project we will provide you the details of each column and what to predict, then what software use and we will talk about also cross validation, that is all about the data.

Thank you.