CS 773/873: Data Mining and Security

Summer 2017

Course project

Points: 100

Analyzing Open University Learning Analytics Dataset

This is a team project with 1 or 2 team members

Reference: https://archive.ics.uci.edu/ml/datasets/Open+University+Learning+Analytics+dataset#

Complete data and data description: https://analyse.kmi.open.ac.uk/open_dataset

Open University Learning Analytics Dataset (OULAD) contains data about courses, students and their interactions with Virtual Learning Environment (VLE) for seven selected courses (called modules)----AAA, BBB, CCC, DDD, EEE, FFF, GGG. Presentations of courses start in February and October - they are marked by 'B' and 'J' respectively (2013B, 2013J, 2014B, 2014J). The dataset consists of tables connected using unique identifiers. Dataset is stored in several csv files.

Course	2013-Offering 1	2013-Offering 2	2014-Offering 1	2014-Offering 2
AAA		2013J		2014J
BBB	2013B	2013J	2014B	2014J
CCC			2014B	2014J
DDD	2013B	2013J	2014B	2014J
EEE		2013J	2014B	2014J
FFF	2013B	2013J	2014B	2014J
GGG		2013J	2014B	2014J

Data is available in 7 files all in csv format. Here are the details of the data available.

Files available (all in csv format)

Description of files available is in OULAD.names

Data files (numbered F1-F7):

assessements.csv (F1)

```
"code_module","code_presentation","id_assessment","assessment_type","date","weight"

"AAA","2013J","1752","TMA","19","10"
```

courses.csv (F2)

code_module	code_presentation	module_presentation_length
AAA	2013J	268

Assessment.csv (F3)

"id_assessment","id_student","date_submitted","is_banked","score"

"1752","11391","18","0","78"

studentinfo.csv (F4)

code_module","code_presentation","id_student","gender","region","highest_education","imd_band","age_band","num_of_prev_attempts","studied_credits","disability","final_result"

"AAA","2013J","11391","M","East Anglian Region","HE Qualification","90-100%","55<=","0","240","N","Pass"

studentRegistration.csv (F5)

"code_module","code_presentation","id_student","date_registration","date_unregistration"

"AAA","2013J","11391","-159","?"

studentVle.csv (F6)

"code_module","code_presentation","id_student","id_site","date","sum_click"

"AAA","2013J","28400","546652","-10","4"

vle.csv (F7)

"id_site","code_module","code_presentation","activity_type","week_from","week_to"

"546943","AAA","2013J","resource","?","?"

This information is summarized in Table 1.

Table 1. Summary of attributes available in the seven data files

	Attribute	Description	F1	F2	F3	F4	F5	F6	F7
A1	code_module	Course ID	X	X		X	X	X	X
A2	code_presentation	Year and offering (e.g., 2013J)	X	X		X	X	X	X
A3	id_assessment	Unique ID assigned to the assessment for each course/offering	X		X				
A4	assessment_type	Type of assessment	X						
A5	date	Days from the start	X						
A6	weight	Weightage of that assessment	X						
A7	module_presention_length	Total length of the offering		X					
A8	id_student	Student ID			X	X	X	X	
A9	date_submitted	When score was submitted for an assessment for a student			X				
A10	is_banked	??????????????			X				
A11	score	Obtained score			X				
A12	gender	Student gender				X			
A13	region	From which region student comes				X			
A14	highest_education	Highest education obtained				X			
A15	imd_band	??????????????????				X			
A16	age_band	Age range				X			
A17	num_of_prev_attempts	How many times the same course was taken earlier?				X			
A18	studied_credits	How many credits accrued already				X			
A19	disability	Any disability?				X			
A20	final_result	Final result Distinction/Pass/Fail/Withdrawn				X			
A21	date_registration	When registered prior to the start of the course					X		
A22	date_unregistration	When withdrawn					X		
A23	date	When course was accessed by each student						X	
A24	sum_click	Number of clicks in each login						X	
A25	id_site	Where course was accessed							X
A26	activity_type	What activity							X
A27	week_from	???????							X
A28	week_to	??????????????????							X

What is the problem to solve?

The given student data for the 7 courses (with multiple offerings) is to be used to identify at-risk students so suitable interventions can be taken to help students succeed.

This is a capstone project and not a homework. So you are not provided with all the information that you may hope to have such as what should be the training the set, what should be the test set, etc. Instead, these decisions are left to the student. You may use clustering, classification, and regression techniques to achieve the objective.

The data has the demographic data, the VLE interaction data, and the assessment data. Each of these pieces of data is useful by itself as well as when combined. You area sked to experiment with this data, in all combinations and discover which combination works the best. Clearly state your metrics for measuring the goodness of a method and how well it does better than the other.

In this process, you should also employ suitable data mining techniques we have studied such as clustering, classification, and regression. For example, you may want to cluster students based on demographics, and then use the new group data for further analysis.

You will be graded on the rigor of your data analysis, data mining, and inferences drawn from the results. Use graphs, tables, figures etc. to tabulate your results. You will submit a formal report of your project with the following sections. Simply filling the report with tables or figures is not sufficient. You must be able to describe the tables and draw appropriate inferences from them.

1. Executive summary (1/2 to 1 page)
2. Introduction
3. Problem statement
4. Solution methodology
5. Experimental setup and data used
6. Results
7. Conclusions