EdX Dataset : Data Wrangling

Sukanya Chandramouli

Dataset

The HarvardX-MITx dataset ('HMXPC13_DI_v2_5- 14-14.csv ') contains de-identified data from the first year of 13 MITx and HarvardX courses offered on the edX platform . This large scale dataset contains more than six hundred thousand aggregated records .

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147&version=10.0

Description of the Dataset

Variable	**Description**	***Data Source***
course_id	unique identifier with info about institution , course and semester	*administrative*
userid_DI	identifies user / course participant	*administrative*
registered	registered for course, (=1 for all records)	*administrative*
viewed	indicates the user has viewed the course content	*administrative*
explored	indicates user has accessed at least half of the chapters in the course	*administrative*
certified	indicates the user has completed the course with passing grade	*administrative*
final_cc_cname_DI	indicates the country / region of the user	*administrative*
LoE	level of education of the user	*user provided*
YoB	year of birth	*user provided*
gender	Possible values: m (male), f (female) and o (other)	*user provided*
grade	final grade of the user	*administrative*
start_time_DI	date of course registration	*administrative*
last_event_DI	date of last interaction with course	*administrative*
nevents	number of interactions with the course	*administrative*
ndays_act	number of unique days student interacted with course	*administrative*
nplay_video	number of play video events within the course	*administrative*
nchapters	number of chapters completed by the student	*administrative*
inconsistent_flag	identifies records that are internally inconsistent	*administrative*

Data Wrangling

Missing Values :

As a first step of data clean up , we look at the missing values in the data set . There are quite a few records for which the user provided data (Year of birth , education background , gender) is missing. We replace these values with the median value (for that particular course).Some of the administrative variables such as 'nchapters','ndays_act' etc which were missing are replaced with zero.

Data Types

The date of birth which is classified as a text is converted to a "date" format.Other numeric data types are set to appropriate data types (int64 , categorical , float). The grades which are specified between 0 and 1 is converted to percentage grades . The categorical order for educational level is set.

New Columns :
Using the Date of Birth information , we create a new column for the “age” of participants.The course-id is decoded to obtain the institution , course and semester/year. The course-id is mapped to the course title.

Renamed Columns:
Some of the column names are renamed for readability