Capstone I Project Proposal :
Predicting the registrant success rate of EdX (HarvardX / MITx) Courses

Sukanya Chandramouli

Background :

A joint research team from Harvard and MIT released the dataset containing details of course-by-course patterns of student interactions with courses from Harvard and MIT during the first year of edX.

With over hundred thousand registrants in total in the first year of courses , the average course completion rate varies from 1% to 7% for these courses. The joint research report focuses on improving the open online courses and approaches to quantifying course efficacy.

Dataset

The HarvardX-MITx dataset ('HMXPC13_DI_v2_5- 14-14.csv ') contains de-identified data from the first year (Academic Year 2013: Fall 2012, Spring 2013, and Summer 2013) of 13 MITx and HarvardX courses offered on the edX platform . This large scale dataset contains more than six hundred thousand aggregated records about diverse kinds of learning interactions and outcome measures for the learners who took courses on the platform. The dataset includes both administrative as well as user-provided data. Personally identifiable information has been de-identified in the dataset.

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

Objective

The goal of this capstone project is to evaluate if any predictions can be made on the course completion rate of the registrants. The dataset does not capture the intent of the participants – whether to complete the course certification or to get access to the course/teaching material. Also given that more than half the participants did not even click through the material, special considerations are to made while interpreting the data. The dataset that has been made public does not have finer details such as the time-wise break of course progress. It only provides aggregate statistics such as the the number of interactions , chapters reviewed , grades at the end of the course.

With this in mind, I want to address the following questions:

1. Is it possible to predict the course completion rate based on the student demographics - such as age ,educational background , region , course type etc ?
2. Is it possible improve the course completion based on early interactions in the course such as how many chapters were reviewed during the early weeks ?
3. Is it possible to suggest interventions to improve learning / ensure the completion rates based on the early trends ?

Approach

1. Understand the dataset and variables in the data
2. Clean data to exclude invalid entries
3. Explore and build visualization
4. Explore and construct prediction models
5. Validate the accuracy of the model

Deliverables

- Jupyter Notebooks with exploratory data analysis and prediction model
- Summary Report
- List of variables for data collection for MOOC courses to improve success rates or apply early intervention