MOOC Success Predictor

GA Data Science - Final Project Outline

Matthew Lichti

November 19, 2014

1. Problem to be solved

Massive Online Open Courses (MOOCs) can be a very affordable and flexible way to learn valuable skills. Unfortunately, MOOCs tend to have very low completion rates. I will analyze what attributes predict success and failure in MOOCs to help students and instructors be successful in their courses.

2. Description of dataset

I have a dataset from EdX in CSV form that contains 641,138 course records from 476,532 unique students and 13 classes (5 HarvardX & 8 MITx). The courses include a wide variety of topics including humanities, engineering, science, and CS. The features include basic information provided by the student like age, gender, country, and highest level of education completed. It also contains information about what the student did during the course including when they started, how many times and how many unique days they worked on the course, how many chapters of the course they looked at, how many lecture videos they watched, and how many times they posted on the discussion forums. The target is whether the student passed the course and their final grade.

Most of the features are continuous variables. Education level, gender, and country are categorical, so I will convert them into dummy variables. I would also like to look at whether the student has previously enrolled in EdX courses and if they completed them, which would be additional categorical variables. I can extract this features by looking at whether there is a course record with the same student ID from an earlier course.

3. Hypothesis

I think the amount that the student interacts with the course will be the biggest predictor of success. I assume students that work on the course over a lot of different days perform better than students that only work on the course once a week or so. I'm interested in the relative importance watching the videos, posting on the forum, or looking at the courseware chapters. I think education level will also be a an important predictor of success.

One of the biggest problems with the analysis is differentiating between students that didn't complete the course because it was too difficult, students that didn't complete it simply because they were busy with other things, and students that just wanted to explore part of the course and never intended to complete it. There will probably be considerable differences between the different courses, but it will be hard to determine if that is due to the different subject matter of the course or differences in the format and teaching style of the course. Since there is such variation between courses, the findings for these 13 courses may not apply as well to many of the hundreds of other MOOCs that are currently available.

4. Statistical methods I plan to use and why

This is primarily a classification problem. I think the most important way to evaluate success is whether the student earned a certificate for passing the class which means they earned the minimum grade that varies depending on the course between 50% and 80%. For that, I will try all of the classification ML techniques we've studied so far (KNN, logistic regression, SVM, naïve Bayes, decision trees, and ensemble methods like random forest). I think it's good to try a lot of different techniques in order to see which ones work best. I will experiment with low thresholds since less than 10% of total students passed each class. With a threshold of 50%, a classifier could just predict that all of the students fail and be over 90% accurate. I'll have to set it lower in order to capture more of the true positives even though it will mean a lot more false positives as well.

I could also do some linear regression to predict final scores for students that completed the course or to predict how far students that didn't finish the course made it in before quitting. I probably won't use unsupervised learning other than maybe PCA just for data exploration.

5. Applications the finding may have

The value of this analysis will be determining the important features for succeeding in MOOCs. This can help students better plan what they need to do to have a better chance of successfully completing a class. It could also help instructors and MOOC providers to know whether it's better to put more emphasis on different features of the class like videos, discussion forums, etc.

The deliverables will be a report on my findings, some visualization probably just from matplotlib, and python code.