MET CS777 Big Data Analytics
Term Project Report

Data Analysis on Open University Learning Analytics Dataset

Site Li (Stella)
U53946229
Graduate School of Arts and Science, MSSP
Boston University

Abstract

The completion rate of Massive Open Online Courses is a controversially topic and many researches are trying to understand the reasons behind the low completion rates.[1] This project leverages data from Open University Learning Analytics Dataset [2] to explore the factors that affect the completion rates and construct models to predict the probability of a student to pass or fail a course. This study focused on factors including demographic information of students and the course materials students interacted with.
The accuracy of 2-class logistic regression is 82% and F1 score is 0.865.

Keywords: MOOC, big data, learning analytics, completion prediction, intervention

Introduction

Background

The completion rates of Massive Open Online Courses vary from 0.7% to 52.1% across multiple platforms. [1] The rates were sometimes regarded as “abysmally low” [3]. Many studies tried to analyze user behaviors and the factors that affect the completion rates.

This project leveraged the data provided by The Open University (UK) and focused on factors including demographic information of students and the course materials students interacted with. The research questions are: 1) comparing students with different features, are some certain type of students have higher probability of getting pass the course? 2) comparing students who interacted with different materials, are interaction with some materials associated with higher probability of getting pass the course? The results can be used to inform faculty about what kind of students may need early intervention and what materials they provided to students may be more important than others.

Data Source

The dataset contains seven relational data files. The detailed schema and descriptions can be found at https://analyse.kmi.open.ac.uk/open_dataset#description. This project used data in “vle.csv”, “studentinfo.csv”, “studentRegistration.csv” and “studentVle.csv” files. I intentionally did not include the data related to results of assessments, as they may have strong correlations with the final results that weaken the relationships between the final results and the variables of interest.

Method

Exploratory Data Analysis

The course BBB – 2014B module was selected for this project. The course lasted for 243 days, and provided 311 materials. There were 1613 students registered to the course, while 290 students withdrew before the course even started, and another 200 students withdrew during the semester.

Among the 1123 students who completed the course, 990 were females, and 98 of them declared a disability. There are 5 levels of education and 3 levels of age, and all the levels were transformed to numeric indexes in prediction models. (Table 1)

Index	Description	# of Students
0	No formal quals	19
1	Lower than A level	464
2	A Level or equivalent	504
3	HE qualification	134
4	Post Graduate Qualification	2

Index	Age group	# of Students
0	0 - 35	744
1	35 - 55	378
2	55 <=	1

Table 1. Distribution of student age and education background:
left: education; right: age

For the 311 materials, the interaction was defined as clicks of a certain material. Every material got used by some students, and the 1st most used 100 materials (most students had some clicks on the material) were selected. The most used material had 1086 students clicked while the 100th most used material had only 94 students clicked. And the total number of clicks of a certain material was calculated for each student.

Prediction Models

In prediction model, only the information of students who completed the course was used in training and testing the models. All the demographic information of students are categorical variables, and were transformed to numeric indexes. There are three levels of outcome of the course: fail, pass, and distinct. The information included for each student is as below:

Student ID	Final result	Female	Edu	Age	Disability	# of click on material 1	…	# of click on material 100
6 digits	Fail/Pass /Distinct	1/0	5-level index	3-level index	1/0	integer		integer

In the first model, both “pass” and “distinct” were treated as “pass” to build a 2-class logistic regression model with “LogisticRegressionWithLBFGS” from PySpark Mllib package. Each student has 1 indicating pass or 0 indicating fail of the class as final result. The student IDs were not included in the prediction model.

In the second model, “fail”, “pass” and “distinct” are coded as “0”, “1” and “2” respectively to build 3-class prediction model. “LogisticRegressionWithLBFGS” and “RandomForest.trainClassifier” from PySpark Mllib package were used to build two prediction models.

For both models, 70% of all students’ data were used to train the models and the rest were used to test it.

Withdrawn behavior analysis

For students who withdrawn during the semester, the day of withdrawn and the last materials they interacted with were analyzed to find the trend in withdrawn behavior.

Results

2-class Prediction Models

1. Confusion matrix: The prediction confusion table is showed below.

		Reference	
		Pass	Fail
Prediction	Pass	183	36
	Fail	21	82

Accuracy = 82%

Table 2. Confusion table of 2-class logistic regression model

2. Coefficients analysis

The four demographic features all have negative coefficients. For example, the coefficient for "Female" variable is 1.65. It indicates that holding other variables constant, the female students are expected to have lower probability of passing the course. However, there are no standard errors information that we cannot calculate p-values to determine whether it is statistically significant.

The coefficients of the features were sorted decreasingly and the IDs of 10 materials with highest positive coefficients are collected. According to my result, materials 768401', '768469', '768772', '768620', '768409', '768499', '768655', '768345', '768477', and '768607' have highest coefficients, and can be regarded as "important" materials.

3-class Prediction Models

The confusion matrices of the two models are showed below.

	Logistic Regression				Random Forest			
		Reference				Reference		
		Distinct	Pass	Fail		Distinct	Pass	Fail
Pred	Distinct	10	22	8	Distinct	4	5	1
	Pass	38	112	27	Pass	50	138	38
	Fail	6	16	83	Fail	0	7	79
	Accuracy = 64%				Accuracy = 69%			

Table 3. Confusion matrices for 3-class prediction models

Comparing the two results, random forest algorithm performs better than logistic regression. However, comparing with the two-class model, the accuracy of these two models are both not as good.

Withdrawn behavior analysis

Histogram function is used to show the trend of withdrawn date.

```
StudentWithdrawnTime.values().histogram(32)
```

I drew 32 bins to make each bin represents about 1-week period and got results as following:

```
[18, 13, 8, 3, 7, 6, 13, 7, 3, 5, 7, 16, 5, 4, 6, 9, 3, 3, 4, 3, 3, 2, 7,
3, 6, 6, 2, 6, 3, 5, 6, 8]
```

Approximately, each number in this list represents the count of student withdrawn in a certain week. The first two weeks had a lot of withdrawn, while in the 7th week and 12th week the number of withdrawn increased significantly. Therefore, instructors of the course should look into the content of week 7 and 12, to figure out why there were such increases, and in the future when they teach the course again, they should come up with some intervention actions prior to the two weeks to keep students staying in the course.

I also analyzed the last materials that students interacted before they quit. It is interesting that 84 students withdrew after interacting with the material "768351". The number is significantly higer than those of other materials (1 to 11 students). It indicates that the instructors should investigate the timing, content and difficulty level of this material and find out the reason.

Discussion

1. In this project, I only analyzed one course and the sample size of students is around 1,000. However, there are hundreds of courses across MOOC platforms and thousands of users of online courses. Therefore, there is high potential that the methods in this project to be used on large scale of data.
2. The accuracy rates of two-class and three-class prediction models are not very high, but it is what I expected. We all understand that there are many factors that can explain students' performances of a course other than their demographic information and what materials they used. For example, some students may already have some understanding of the topic of this course, or a student may be too busy to spend enough time in preparing for tests. However, the dataset did not include such information.
3. There are 11 tests in this course and all test results are available. However, in the analysis, I did not use the test results, as they would be highly correlated with the final scores.
4. Based on the result of logistic regression model, one can analyze the coefficient of each material, and estimate their "importance" according to the coefficients: materials with larger coefficients are more important. Though we cannot draw causal conclusions from the results, it can provide a starting point for instructors to review and revise the materials they provided to students. Also, the course platform or instructors can remind a student to read a certain material multiple times if it is regarded as "important".
5. The analysis on withdrawn students can help instructors to understand when to intervene to keep students in the program and to reconsider the usage of certain materials.

Reference:

[1] Jordan, K. (2015). Massive open online course completion rates revisited: Assessment, length and attrition. International Review of Research in Open and Distributed Learning, 16(3) pp. 341–358.

[2] Kuzilek J., Hlosta M., Zdrahal Z. (2017) Open University Learning Analytics dataset Sci. Data 4:170171 doi: 10.1038/sdata.2017.171.

[3] Ahearn, A. (2017) The Flip Side of Abysmal MOOC Completion Rates? Discovering the Most Tenacious Learners. https://www.edsurge.com/news/2017-02-22-the-flip-side-of-abysmal-mooc-completion-rates-discovering-the-most-tenacious-learners