Predicting Outcomes from Student Engagement Data

Intervening to support positive outcomes

Petter Lövehagen (05976423)

## Abstract

## Introduction

Higher Education Institutions (HEIs) are complex organisations offering educational services to a wide and diverse student body. HEIs have the delivery of educational success at their heart but they must also operate as viable businesses. The sucess of a HEI is measured by the success of its students - good outcomes for students equate with success for the HEI, including financial success, improved student recruitment and retention and reputational enhancement.

Learner Analytics is an ever growing field of research in Higher Education (HE). HE institutions are increasingly using data to inform decision making and to improve the student experience. In a sector report from 2017, it was found that HEIs “working towards implementation has nearly doubled from 34% to 66%” and that the focus has shifted “towards retention more than learning” (17% to 37%). (Newland, 2017).

This report evaluates the performance of several learning algorithms to predict student outcome using ‘student engagement’ data. It explores the potential for using data to establish whether there is an opportunity for early intervention in the the case of students who are predicted to fail or withdraw.

## Data

[Open University](https://www.open.ac.uk/about/main/) is one of the world’s largest distance learning providers, with over 200,000 students (Wikipedia, 2023.) All of its teaching takes place in virtual learning environments (VLEs) and student interactions are recorded in log files.

It is important to note that students to not have ‘in-person’ teaching as is provided by most HEIs. Analysis and conclusions in this report cannot be diretly extended to HEIs that do provide in-person teaching - they may want to consider additional or different ‘engagement’ data.

However, the analysis and conclusions are generalisable to the extend that engagement behaviour is a good predictor of student outcome and that VLEs form a critical element of modern HE teaching provision.

### Dataset

The [Open University Learning Analytics Dataset (OULAD)](https://analyse.kmi.open.ac.uk/open_dataset) contains data about courses, students and their interactions with Virtual Learning Environment (VLE) for seven selected courses (called modules). The dataset consists of tables connected using unique identifiers:

|  |
| --- |
| OULAD data model |

The dataset is rich, containing biographic and demographic characteristics about the students, including their gender, age, disability status, educational background, IMD band[[1]](#footnote-27), and region of origion. In addition to person details, the dataset contains information about students’ academic history, assessments and VLE interacations.

It includes registrations for seven modules across two academic years (2013/2014 and 2014/2015) with two intake months (February and October). In the raw data, there are 32593 student registrations from 28,785 uniqe students - 13,529 are active in 2013 and 19,064 in 2014.

Assessment data includes submissions throughout the course including assessment type, date, weight and score. VLE data includes type of count of interactions, date and time.

More details can be found here[[2]](#footnote-28).

### Exploratory Data Analysis

Following initial data processing and feature engineering, the dataset contains 31437 rows and 27 *potential* features.[[3]](#footnote-31) with an overall distribution of outcomes as follows - 46.6% of students passed, 31.4% withdrew and 22.0% failed.

|  |
| --- |
| Overall outcome distribution |

Some analysis was done to understand the dataset, identify issues and potential features. A selection of findings are presented below.[[4]](#footnote-35)

#### Students

These are the distributions of outcomes by gender (M/F) and disability status (Y/N):

|  |
| --- |
| Distribution by gender, disability |

Chi-square tests of independence indicate that there is a statistically significant () association between both gender (moderate) and disability status (high) and final\_outcome.[[5]](#footnote-39) This was also the case when looking at a binary outcome of ‘needs intervention’ (fail or withdraw) vs ‘does not need intervention’ (pass or distinction).[[6]](#footnote-40)

Chi-square tests for the other student characteristics (age\_band, region, highest\_education, imd\_band) also indicate statistically significant associations with final\_outcome. The plots below show clear differences between groups in terms of outcomes, for example, students with lower ‘highest\_eduction’ have much higher withdrawal rates, perhaps because they are less accustomed to the academic environment. It is also notable that students have more success (pass, distinction) in the older age bands.

#### Curriculum

The seven modules are from two subjects - STEM and Social Sciences - there are far more withdrawn students amongst STEM registrations.

|  |
| --- |
| outcome\_by\_subject |

When looking at outcomes by the year of study, there is a diffence in the distribution of the outomes. Aside from there being more students in the 2014 year, there are far more students who withdraw:

|  |
| --- |
| outcome\_by\_year |

#### Academic / Engagement

|  |
| --- |
| outcome\_by\_age |

The boxplots of engagement behaviour between the outcomes show very clearly that there is an association between the feature and the outcome. Students who have higher average scores, click more and spend more days in the VLE have a higher chance of passing and passing with distinction.

The Kruskal-Wallis test was used to statistically compare distributions of these continuous variables across the different final outcomes and they are all statistically significant (). The null-hypothesis that the differences between outcomes are due to chance alone is rejected in favour of the alternative.[[7]](#footnote-52)

This is a good indication that these features will be useful in predicting the outcome.

### Data Preparation

I considered different ways of splitting the data:

* by year - train on 2013 and predict on 2014
* by module - train and predict on each module separately
* complete - train and test datasets from complete dataset

In the end, opted for the complete dataset. Whilst training on 2013 and predicting on unseen 2014 data may mimic real-world scenario, there were some module\_presentations which were only represented in 2014 which would limit the opportunity. There was also appeared to be significant differences in distributions between years. Taking this into account was beyond the scope of this project.

* Module CCC only has 2014 data
* Module EEE only have 2013J data
* Module GGG only has 2013J data

## Analysis Type

## Learning Algorithms

### Selection

### Evaluation Metrics

## Discussion

## Limitations / Future Considerations

## Bibliography

* HE has a problem with student retention and success
* HE has a lot of data about students
* HE can collect in-time data about behaviour, engagement, performance

Can we use this data to predict student success or outcomes? Can it be used to help future students by putting interventions in place if they are predicted to fail or withdraw? \* \* HE institutions are increasingly using data to inform decision making and to improve the student experience. \* Learning Analytics solutions installed, big business \* Can we use data to predict student success or outcomes? \* Does the data we have available allow us to do this? \* Does

This aim of this report is to evaluate suitable learning algorithm(s) to address this research questions - early prediction of student success and failure. The report will also evaluate the performance of the selected algorithm(s) and discuss the results.

1. IMD is the Index of Multiple Deprivation which is a standard measure of relative deprivation of the student using multiple variables. [↑](#footnote-ref-27)
2. Additional details about OULAD preparation can be found here - <https://www.nature.com/articles/sdata2017171#Sec2> [↑](#footnote-ref-28)
3. See XXXX include github link [↑](#footnote-ref-31)
4. Student profile exploration via unsupervised machine learning techniques like clustering were considered and explored, but are not included in this report. Given the statistical significance between groups and outcomes, exploring ‘student profile’ based on characteristics, behaviour, academic history, background, or a combination could be worthwhile. This is ear-marked as a potential future project. See XXX for notes on this. [↑](#footnote-ref-35)
5. Gender: Chi-square statistic: 24.208, p-value: 2.260e-05 Disability: Chi-square statistic: 128.412, p-value: 1.189e-27 [↑](#footnote-ref-39)
6. Gender: Chi-square statistic: 24.018, p-value: 9.543e-07 Disability: Chi-square statistic: 104.779, p-value: 1.365e-24 [↑](#footnote-ref-40)
7. Average score: Kruskal-Wallis statistic: 22898.85, p-value: 0.0 Submission distance: Kruskal-Wallis statistic: 1745.16, p-value: 0.0 Total clicks: Kruskal-Wallis statistic: 14263.82, p-value: 0.0 Days active: Kruskal-Wallis statistic: 16091.56, p-value: 0.0 [↑](#footnote-ref-52)