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 delivery of learning success at 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 improve the student experience. In a sector report from 2017, it was found that HEIs “working towards [learner analytics systems] 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).

Analysis and conclusions in this report cannot be directly extended to HEIs that provide in-person teaching - they may want to consider additional or different ‘engagement’ data such as attendance, face-to-face engagements, library usage, etc.

However, the questions, methods and conclusions are generalisable to the extent that engagement behaviour is a good predictor of student outcome and that VLEs form a critical element of most 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 Virtual Learning Environment (VLE) interactions for seven modules. The dataset consists of tables connected using unique identifiers:

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

The dataset is rich, containing student biographic and demographic characteristics, including gender, age, disability status, educational background, IMD band[[1]](#footnote-27), and region. In addition to person details, the dataset contains information about students’ academic history, assessments and VLE interactions[[2]](#footnote-28) (Kuzilek, et al, 2017).

It includes registrations for two subjects across two academic years (2013/2014 and 2014/2015) with two possible intake months (February and October). In the raw data, there are 32593 student registrations (28,785 unique 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.

### Exploratory Data Analysis

Following initial data processing, 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 |

Analysis was peformed 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.

|  |
| --- |
| Distribution by Age, IMD, Education, Region |

#### Curriculum

There are seven modules from two subjects - STEM and Social Sciences - there are far more withdrawn students amongst STEM registrations and also in 2014 compared to 2013.

|  |
| --- |
| outcome\_year\_subject |

Average assessment scores vary between modules. STEM subjects are likely assessed differently to Social Science subjects but if a student’s subject potentially affects their outcome, this might have merit as a predictive feature.

|  |
| --- |
| score\_by\_module |

#### Academic / Engagement

The main goal of this report is to consider a generalised model which predicts outcome from engagement behaviour - that is, without student person and curriculum features. Thus, the models will make primary use of the features below.

The boxplots of engagement behaviour 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.

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

The Kruskal-Wallis test was used to statistically compare distributions of these continuous variables across the different final outcomes. They are all statistically significant () and the null-hypothesis that the differences between outcomes are due to chance alone is rejected in favour of the alternative.[[7]](#footnote-55)

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

#### Dates

Dates are important for this analysis as students withdraw throughout their course.

When looking at withdrawals over time, there are two spikes towards the beginning of the course - on day 0 and a couple of weeks into the course. It may be too early to predict these students with engagement data but a different research question might be clustering student profiles based on data other than behaviour.

|  |
| --- |
| withdrawals\_over\_time |

However, there is a steady stream of withdrawals throughout the course. The cumulative withdrawal plot shows that approximately 50% of students withdraw by day 100 - if we can predict these students, we can intervene and potentially alter their outcome.

### Data Preparation for Modeling

#### Splitting

Several approaches to splitting the data for training was considered:

* 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

Given that the 2013 and 2014 subsets are not comparable, the models were trained on a subset of the complete dataset, split with stratification. This was to ensure that the training data contained examples of each module and presentation.[[8]](#footnote-62)

* training - 60%
* validation - 20%
* test - 20% - this dataset was not used during training, tuning or validation

In the real-world, this model *would* be used to predict 2015 students which would differ from 2013 and 2014 students.

#### Feature Selection and Engineering

As the the business scenario is to predict outcome from engagement, all student features were removed from the dataset[[9]](#footnote-64) to focus on engagement and assessment.

Data was cleaned to remove missing values, impute missing values as per dataset author notes. (Kuzilek, et al, 2017)

Features were engineered from the original dataset resulting in these features:

Numerical features were scaled:

* number\_of\_previous\_attempts
* studied\_credits
* proportion\_submissions
* average\_score
* submission\_distance
* activity\_count
* total\_clicks
* days\_active

Categorical features were one-hot encoded:

* subject

The target variable is: final\_result

Modeling was done with a multivariate outomce - ‘Pass’, ‘Distinction’, ‘Withdrawn’ and ‘Fail’ and a binary outcome - ‘Needs intervention’ (fail or withdraw) vs ‘Does not need intervention’ (pass or distinction) for comparison.

#### Dimensionality Reduction

Principal Component Analysis as a dimensionality reduction technique was explored and compared but ultimately not pursed as the dataset is probably not large enough to warrant it and the model results were similar.[[10]](#footnote-66)

## Analysis Type

The analysis type is a binary classification problem. HEIs want an early identification system to predict students who are likely to discontinue their studies by the end of the course. They are interested in predicting which students will either fail or withdraw from their course using engagement data.

Whilst predicting multiclass outcome (pass, distinction, fail, withdraw) was explored, HEIs are initially interested in predicting non-continuation so that they can intervene and affect the outcome.

The analysis is multivariate since it considers multiple engagement features simultaneously to predict the outcome. This allows the model to detect complex interactions and dependencies between features and the outcome to make accurate predictions. For example, students who engage on many days and have higher average scores may be more likely to pass or pass with distinction.

## Learning Algorithms

### Selection

Several models where considered and evaluated:

* Logistic Regression - LR is simple and easy to interpret model with optional probabilistic output. LRs assume a linear relationship between features and target. It is less prone to overfitting.
* Decision Tree - DTs are simple and easy to interpret but prone to overfitting and sensitive to small variations in the data; therefore they may create overly-complex trees which do not generalise well.
* Random Forest - RFs are known for their good performance and robustness to outliers, noisy data, high-dimensionality and non-linear relationships. T
* Support Vector Classifier - SVCs are good for high-dimensional feature spaces, handle linear and non-linear relationships and are robust to outliers. They can be computationally expensive and sensitive to tuning. T
* K-Nearet Neighbours - KNNs are simple and easy to interpret but computationally expensive and sensitive to k neighours.

All models are suitable for both categorical and numerical features.

A less suitable model was included - Gaussian Naive Bayes. GNB assumes independence between features and can struggle with imbalanced datasets. This dataset has both characteristics - there are less well represented module\_presentations and the features are not independent as several are correlated and engineered from the same underlying data.

The highly-correlated features on the bottom right corner are those included in the model.

|  |
| --- |
| selected\_correlation |

### Evaluation Metrics

## Results and Discussion

In the real-world, a HEI would only be able to predict using existing data up to the prediction point. Therefore, an approach was developed which allows for this - where the data is split based on a prediction\_point.

## Limitations / Future Considerations

factor in intervention data - was it successful, did it impact outcome, is there a threshold for intervention

## 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. Modules CCC is only represented in 2014 and modules EEE and GGG are only represented in October 2013. [↑](#footnote-ref-40)
7. See xxx for more information [↑](#footnote-ref-55)
8. Modules CCC is only represented in 2014 and modules EEE and GGG are only represented in October 2013. [↑](#footnote-ref-62)
9. See xxx for more information [↑](#footnote-ref-64)
10. See xxx for more information [↑](#footnote-ref-66)