Predicting Outcomes from Student Engagement Data

Early Intervention to Support Positive Outcomes

Petter Lövehagen (05976423)

## Abstract

## Introduction

Higher Education Institutions (HEIs) are complex organisations offering educational services to a diverse student body. HEIs have students’ learning success at heart but they must also operate as viable businesses.

An HEI’s success is measured by its students’ success - good outcomes for students equate with improved student recruitment, continuation and retention, enhanced reputation and financial security.

Learner Analytics is a growing field of research in Higher Education (HE). HEIs are increasingly using data to inform decision-making and improve 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 in predicting student outcome using ‘student engagement’ data. It explores the 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).

### 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. It consists of tables connected using unique identifiers:

|  |
| --- |
|  |

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

It includes February and October registrations across two academic years (2013 and 2014). There are 32,593 student registrations (28,785 unique students) - 13,529 are active in 2013 and 19,064 in 2014.

Assessment data includes 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-3) - 46.6% of students passed, 31.4% withdrew and 22.0% failed.

|  |
| --- |
|  |

A selection of findings is presented below.

#### Students

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

|  |
| --- |
|  |

Chi-square tests of independence indicate that there is a statistically significant () association between both gender (moderate) and disability status (high) and final\_outcome.[[4]](#footnote-4) This was also the case when looking at a binary outcome of ‘intervention’ (fail or withdraw) vs ‘no intervention’ (pass or distinction).[[5]](#footnote-5)

Chi-square tests for the other student characteristics (age\_band, region, highest\_education, imd\_band) also suggest 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 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

There are seven modules across two subjects - STEM and Social Sciences; STEM students are more likely to withdraw. There are more students in 2014.

|  |
| --- |
|  |

Average assessment scores vary between modules. If a student’s subject or intake *potentially* affects their outcome, this might have merit as a predictive feature.

|  |
| --- |
|  |

#### Academic / Engagement

The main goal of this report is to evaluate a generalised model predicting outcome from engagement *behaviour* - that is, excluding student person and curriculum features. Thus, the models be built with below features.

Boxplots of engagement behaviour show clear association between features and the outcome. Students who have higher average scores, click more and spend more days using the VLE have a higher likelihood of passing.

|  |
| --- |
|  |

Kruskal-Wallis test was used to statistically compare distributions of continuous variables. 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.[[6]](#footnote-6)

This is a good indication that these features will be useful.

#### Dates

Students withdraw throughout their course.

There are two spikes towards the beginning of the course - on day 0 and a couple of weeks into the course. There may be too little engagement data to confidently predict outcomes but a different research question may involve other data and techniques to predict early withdrawals. Students who did not start the course are not included in the dataset.

|  |
| --- |
|  |

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 these students can be identified with precisoin, there may be opportunities to affect 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

As the 2013 and 2014 subsets are different, 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 combination.[[7]](#footnote-7)

* training - 60%
* validation - 20%
* test - 20% - unseen until final model

In the real-world, theses models *would* predict 2015 students.

#### Feature Selection and Engineering

As the scenario is to predict outcome from engagement, all student features were removed from the dataset[[8]](#footnote-8) to focus on engagement and assessment as predictors.

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

Features were engineered from the original dataset:

Numerical features (scaled):

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

Categorical features (one-hot encoded):

* subject

The target is: final\_result

Models predicting multivariate outcome (‘Pass’, ‘Distinction’, ‘Withdrawn’, ‘Fail’) were compared with binary classifiers (‘intervention’ (fail or withdraw) vs ‘no intervention’ (pass or distinction)).

#### Dimensionality Reduction

Principal Component Analysis as a dimensionality reduction technique was explored but ultimately not pursued - the results were similar.[[9]](#footnote-9)

## Analysis Type

This is a multivariate binary classification problem. HEIs want to predict students who are likely to discontinue their studies as early as possible using availble engagement data. 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. For example, students who engage on more days and have higher average scores may be more likely to pass or pass with distinction.

Predicting multiclass outcome (pass, distinction, fail, withdraw) may have some use-cases but the primary goal is to address non-continuation.

## Learning Algorithms

### Selection

Several models where considered for evaluation:

* **Logistic Regression** - LR is easy to interpret with optional probabilistic output. LRs assume a linear relationship between features and target. It is less prone to overfitting.
* **Decision Tree** - DTs are simple 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 robust to outliers, noisy data, high-dimensionality and non-linear relationships.
* **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.
* **K-Nearest Neighbours** - KNNs are 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 explored - Gaussian Naive Bayes. GNB assumes independence between features and can struggle with imbalanced datasets. This dataset suffers both - there are less well-represented module\_presentations and features are not independent they have been engineered from the same underlying data.

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

|  |
| --- |
|  |

### Evaluation Metrics

To evaluate the binary classification models, the following metrics were considered:

* **Accuracy** - proportion of correct predictions:

Accuracy is the most common evaluation metric but it is not suitable for imbalanced datasets.

* **Precision** - proportion of correctly predicted positive instances (TP) out of all predicted positive instances (TP + FP):

Precision is useful when the cost of false positives (misclassifying a student as needing intervention when they do not) is high.

* **Recall** - proportion of correctly predicted positive instances (TP) out of all actual positive instances (TP + FN):

Recall is particularly useful when the cost of false negatives (misclassifying a student as not needing intervention when they do) is high. This might be a metric for HEIs to consider.

* **F1 Score** - the harmonic mean of precision and recall:

F1 score balances precision and recall and is useful when there is an uneven class distribution or when both metrics are important.

The trained models were optimised using accuracy due to its ubiquity. Whilst important, precision and recall are better considered with more developed models and requirements.

## Models - Results and Interpretation

### Initial Results

‘PCA-transformed’ and ‘non-PCA’ datasets were compared across all evaluation metrics. There is very little difference between datasets, all else equal:

|  |
| --- |
|  |

The best performing model was the Random Forest model on the non-PCA dataset with an accuracy of 0.93.[[10]](#footnote-10)

|  |
| --- |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Dataset** | **Val Accuracy** | **Val Precision** | **Val Recall** | **Val F1 Score** |
| LogisticRegression | X\_train\_transformed | 0.918 | 0.921 | 0.918 | 0.919 |
| DecisionTreeClassifier | X\_train\_transformed | 0.902 | 0.902 | 0.902 | 0.902 |
| RandomForestClassifier | X\_train\_transformed | 0.930 | 0.932 | 0.930 | 0.930 |
| SVC | X\_train\_transformed | 0.921 | 0.927 | 0.921 | 0.921 |
| KNeighborsClassifier | X\_train\_transformed | 0.917 | 0.920 | 0.917 | 0.917 |
| GaussianNB | X\_train\_transformed | 0.913 | 0.916 | 0.913 | 0.914 |
| LogisticRegression | X\_train\_pca | 0.917 | 0.920 | 0.917 | 0.917 |
| DecisionTreeClassifier | X\_train\_pca | 0.879 | 0.879 | 0.879 | 0.879 |
| RandomForestClassifier | X\_train\_pca | 0.915 | 0.918 | 0.915 | 0.915 |
| SVC | X\_train\_pca | 0.917 | 0.923 | 0.917 | 0.917 |
| KNeighborsClassifier | X\_train\_pca | 0.911 | 0.914 | 0.911 | 0.911 |
| GaussianNB | X\_train\_pca | 0.901 | 0.906 | 0.901 | 0.901 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| In the confusion matrix, the ‘actual positives’ are students who do not require intervention (passed, distinction) and ‘actual negatives’ are students who do may benefit from intervention (withdrawn, fail).   |  |  |  | | --- | --- | --- | |  | **Predicted Negative** | **Predicted Positive** | | Actual Negative | 3036 | 321 | | Actual Positive | 116 | 2815 | |

The model correctly identified 3036 students for intervention and failed to predict intervention for 321 students who went on to fail or withdraw. It incorrectlly predicted that 116 students should have intervention when in actual fact they passed.

### Tuned Results

#### Hyperparameter Tuning

The non-PCA data was used to optimise the models using GridSearchCV to find the best combination of hyperparameters.[[11]](#footnote-11)

The models were improved by 0-2% with the best performing model still being the Random Forest model with an accuracy of 0.932.

[‘LR’, ‘DT’, ‘RF’, ‘SVC’, ‘KNN’, ‘GNB’] {‘LR’: LogisticRegression(C=10.0, max\_iter=10000), ‘DT’: DecisionTreeClassifier(max\_depth=10, min\_samples\_leaf=2, min\_samples\_split=10), ‘RF’: RandomForestClassifier(min\_samples\_split=5, n\_estimators=200), ‘SVC’: SVC(C=10.0), ‘KNN’: KNeighborsClassifier(metric=‘manhattan’, n\_neighbors=7), ‘GNB’: GaussianNB()} [0.9185750636132316, 0.9223918575063613, 0.931615776081425, 0.9243002544529262, 0.9225508905852418, 0.9134860050890585]

#### Ensemble Modelling

For the sake of completeness, an ensemble model was created using the best performing models for each outcome. The model was created using the VotingClassifier with the following models:

* Random Forest
* Logistic Regression
* Support Vector Classifier
* K-Nearest Neighbours
* Gaussian Naive Bayes
* Decision Tree

The model performed similarly to the Random Forest model with an accuracy of 0.926.

A stacking model was also created using the StackingClassifier with the same models. It was fractionally behind the Random Forest model with an accuracy of 0.932.

## Limitations / Future Considerations

### Conclusion

This project is an initial exploration of a very rich learning analytics dataset. It is a proof-of-concept demonstrating the potential of using ‘engagement data’ to predict student outcomes.

It should be noted that:

* HEIs with face-to-face teaching may have different engagement patterns, datasets and outcomes.
* models do not consider student demographics, etc.
* models do not explicitly consider differences between modules, presentations, etc.

In addition to the above, there is much potential for further research:

* temporal nature of the data - for example, the number of days between engagements, the time of day, etc.
* impact of VLE and/or assessment type.
* student profiles using unsupervised learning techniques.
* impact of new/more data
* model performance on substantial changes to behaviour - e.g. student behaviour pre and post COVID-19?
* impact of missing data
* generalised models compare v module-specific models
* real-time data ingesting to improve prediction
* interventions as a feature
* intervention impact

### Final Assessment

In the real-world, HEIs would only be able to predict using existing data up to the prediction point. A mechanism to prepare data based on a prediction\_point was developed - assessments and VLE engagements after the prediction point are removed. In addition to this, students who have already withdrawn are also removed, as their outcome is already known.

The data changes in several ways:

* the number of students is reduced (withdrawals removed)
* engagement and assessment values are changed (future interactions are removed before calculations)
* where no engagement or assessment but student is not withdrawn, the value is set to 0

Predicting on *completely unseen* test data (prepared for 50, 100, 150 days after start) using the best model (Random Forest) produced favourable results, suggesting the identification of students for early intervention is possible using assessment and VLE engagement data with accuracies of around 97%.

|  |
| --- |
|  |

## 

## Bibliography

Abu Zohair, L.M. (2019) Prediction of Student’s performance by modelling small dataset size. *International Journal of Educational Technology in Higher Education*. [online]. 16 (1), p.27. Available from: https://doi.org/10.1186/s41239-019-0160-3 [Accessed 14 May 2023].

Al Mayahi, K. and Al-Bahri, M. (2020) *Machine Learning Based Predicting Student Academic Success*In: *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*. [online]. 2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT) 264–268.

Blankers, M., Koeter, M.W.J. and Schippers, G.M. (2010) Missing Data Approaches in eHealth Research: Simulation Study and a Tutorial for Nonmathematically Inclined Researchers. *Journal of Medical Internet Research*. [online]. 12 (5), p.e54. Available from: http://www.jmir.org/2010/5/e54/ [Accessed 22 March 2023].

Bujang, S.D.A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H. and Ghani, N.A.Md. (2021) Multiclass Prediction Model for Student Grade Prediction Using Machine Learning. *IEEE Access*. [online]. IEEE Access 9, pp.95608–95621.

‘English indices of deprivation 2019’ (2019) *GOV.UK*. [online]. Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 [Accessed 29 May 2023].

Fraley, C. and Raftery, A.E. (2002) Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*. [online]. 97 (458), Taylor & Francis, pp.611–631. Available from: https://doi.org/10.1198/016214502760047131 [Accessed 8 March 2023].

Géron, A. (2022) *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* 3rd ed. O’Reilly Media, Inc.

Harris, C.R. *et al.* (2020) Array programming with NumPy. *Nature*. [online]. 585 (7825), pp.357–362. Available from: https://www.nature.com/articles/s41586-020-2649-2 [Accessed 18 May 2023].

Hashim, A.S., Awadh, W.A. and Hamoud, A.K. (2020) Student Performance Prediction Model based on Supervised Machine Learning Algorithms. *IOP Conference Series: Materials Science and Engineering*. [online]. 928 (3), IOP Publishing, p.032019. Available from: https://dx.doi.org/10.1088/1757-899X/928/3/032019 [Accessed 14 May 2023].

Jayaprakash, S., Kathiresan, V., Shanmugapriya, N. and Dadhich, M. (2021) Opportunities and challenges in transforming higher education through machine learningIn: *Machine Learning for Sustainable Development*. [online]. Walter de Gruyter GmbH & Co KG, 17–30. Available from: https://books.google.co.uk/books?id=Lo89EAAAQBAJ&lpg=PA17&dq=machine%20learning%20predicting%20higher%20education%20applicant&lr&pg=PA17#v=onepage&q=machine%20learning%20predicting%20higher%20education%20applicant&f=false.

Kuzilek, J., Hlosta, M. and Zdrahal, Z. (2017) Open University Learning Analytics dataset. *Scientific Data*. [online]. 4 (1), Nature Publishing Group, p.170171. Available from: https://www.nature.com/articles/sdata2017171 [Accessed 18 May 2023].

*Machine learning in higher education | McKinsey*. (no date) [online]. Available from: https://www.mckinsey.com/industries/education/our-insights/using-machine-learning-to-improve-student-success-in-higher-education [Accessed 26 March 2023].

Newland, D.B. (2017) Learning Analytics in UK HE 2017.

‘Open University’ (2023) *Wikipedia*. [online]. Available from: https://en.wikipedia.org/w/index.php?title=Open\_University&oldid=1157323102 [Accessed 29 May 2023].

Oqaidi, K., Aouhassi, S. and Mansouri, K. (2022) Towards a Students’ Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms. *International Journal of Emerging Technologies in Learning (iJET)*. [online]. 17 (18), pp.103–117. Available from: https://online-journals.org/index.php/i-jet/article/view/25567 [Accessed 14 May 2023].

Orji, F. and Vassileva, J. (2020) *Using Machine Learning to Explore the Relation Between Student Engagement and Student Performance*In: *2020 24th International Conference Information Visualisation (IV)*. [online]. 2020 24th International Conference Information Visualisation (IV) 480–485.

Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J. and Phasinam, K. (2023)‘Classification and prediction of student performance data using various machine learning algorithms’ *Materials Today: Proceedings*.*SI:5 NANO 2021* [online]. 80, pp.3782–3785. Available from: https://www.sciencedirect.com/science/article/pii/S221478532105241X [Accessed 14 May 2023].

Pedregosa, F. *et al.* (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 12, pp.2825–2830.

Sekeroglu, B., Dimililer, K. and Tuncal, K. (2019)‘Student Performance Prediction and Classification Using Machine Learning Algorithms’In: *Proceedings of the 2019 8th International Conference on Educational and Information Technology*.*ICEIT 2019* [online]. New York, NY, USA, Association for Computing Machinery, 7–11. Available from: https://doi.org/10.1145/3318396.3318419 [Accessed 14 May 2023].

Team, T.P.D. (2023) *pandas-dev/pandas: Pandas* [online]. Zenodo. Available from: https://zenodo.org/record/3509134 [Accessed 18 May 2023].

Yadav, S.K., Bharadwaj, B. and Pal, S. (2012) *Mining Education Data to Predict Student’s Retention: A comparative Study* [online]. arXiv. Available from: http://arxiv.org/abs/1203.2987 [Accessed 14 May 2023].

1. IMD is the Index of Multiple Deprivation which is a standard measure of relative deprivation of the student using multiple variables. (Gov.uk, 2019) [↑](#footnote-ref-1)
2. Additional details about OULAD preparation can be found here - <https://www.nature.com/articles/sdata2017171#Sec2> [↑](#footnote-ref-2)
3. See XXXX include github link [↑](#footnote-ref-3)
4. **Gender**: Chi-square statistic: 24.208, p-value: 2.260e-05 **Disability**: Chi-square statistic: 128.412, p-value: 1.189e-27 [↑](#footnote-ref-4)
5. **Gender**: Chi-square statistic: 24.018, p-value: 9.543e-07 **Disability**: Chi-square statistic: 104.779, p-value: 1.365e-24 [↑](#footnote-ref-5)
6. **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-6)
7. Modules CCC is only represented in 2014 and modules EEE and GGG are only represented in October 2013. [↑](#footnote-ref-7)
8. See xxx for more information [↑](#footnote-ref-8)
9. See xxx for more information [↑](#footnote-ref-9)
10. For reference, the best accuracy for the multivariate outcome was 0.71 for the Random Forest model. [↑](#footnote-ref-10)
11. [↑](#footnote-ref-11)